EVERY TWEET COUNTS: EXAMINING SPATIAL VARIABILITY OF TWITTER DATA REPRESENTATIVENESS

by

MASRUDY OMRI

A THESIS

Presented to the Department of Geography
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Master of Science

June 2016

THESIS APPROVAL PAGE

Student: Masrudy Omri

Title: Every Tweet Counts: Examining Spatial Variability of Twitter Data Representativeness

This thesis has been accepted and approved in partial fulfillment of the requirements for the Master of Science degree in the Department of Geography by:

Dr. Amy Lobben	Chairperson
Dr. Christopher Bone	Member

and

Scott L. Pratt	Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded June 2016

THESIS ABSTRACT

Masrudy Omri

Master of Science

Department of Geography

June 2016

Title: Every Tweet Counts: Examining Spatial Variability of Twitter Data Representativeness

The growing global Twitter population has prompted social scientists to examine the potential of Twitter-generated sentiments to serve as an alternative to public opinion polls. This thesis intends to study this potential by evaluating the variability of sampled data representativeness that is voluntarily submitted through Twitter. This research examines a case study: President Barack Obama's public approval as viewed by the United States population. The sentiments generated from Twitter were compared to the sentiments from public opinion polls in order to measure the degree of representativeness at both national and state level. The results show that Twitter data are not representative at the U.S. national level. At the state level, Twitter data representativeness is highly varied and such variability can be linked to individual state’s total population.

CURRICULUM VITAE

NAME OF AUTHOR: Masrudy Omri

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
University of Wisconsin-Madison, Madison

DEGREES AWARDED:

Master of Science, Geography, 2016, University of Oregon
Bachelor of Science, Geography, Cartography and Geographic Information Systems, 2013, University of Wisconsin-Madison

AREAS OF SPECIAL INTEREST:

Geographic Information Science
Social Media and Volunteered Geographic Information
Cartographic Design
Interactive Mapping and Geovisualization

PROFESSIONAL EXPERIENCE:

Graduate Research Fellow, InfoGraphics Lab,
University of Oregon, September 2015 – June 2016

Graduate Teaching Fellow, Department of Geography,
University of Oregon, September 2014 – June 2015

Cartographer and GIS Analyst, Center for Sustainability and the Global Environment, University of Wisconsin-Madison, December 2012 – June 2015

GIS Data Intern, University of Wisconsin-Madison & University of California-Berkeley, June 2014 – September 2014

GRANTS, AWARDS, AND HONORS:

Bill Loy Award for Excellence in Cartographic Design & Geographic Visualizations, First Place, Department of Geography, University of Oregon, 2016

Arthur Robinson Award for Best Printed Map, Honorable Mention, Cartography & Geographic Information Society (CaGIS), 2016

David Woodward Award for Best Electronic Map, Honorable Mention, Cartography & Geographic Information Society (CaGIS), 2016

Master's Research Grant, Cartography & Geographic Information Society (CaGIS), 2016

Poster Session Award, Graduate Research Forum, University of Oregon, 2016

MAPP Web Mapping & App Challenge Award, Department of Geography, University of Oregon, 2015

Travel Grant, North American Cartographic Information Society (NACIS), 2015

Cartography Specialty Group (CSG) Master's Thesis Award, Association of American Geographers (AAG), 2015

Bill Loy Award for Excellence in Cartographic Design & Geographic Visualizations, Honorable Mention, Department of Geography, University of Oregon, 2015

Graduate Teaching Fellowship, University of Oregon, 2014-2016

Wisconsin Alumni Association (WAA) & International Student Services (ISS) Academic Achievement Award, University of Wisconsin-Madison, 2013

Barbara Bartz Petchenik Undergraduate Award for Cartographic Design, University of Wisconsin-Madison, 2012

PUBLICATIONS:

Weisse, M., Omri, M., White, G., Roth, R., & Naughton-Treves, L. (2015). Tambopata Transformed: Using Web Mapping to Enhance a Geography Course Exercise About Forest Conservation. *The Journal of Maps*, *11*(3), 525-533.

ACKNOWLEDGMENTS

What can I say, I had a very wonderful time in Eugene. I am eternally grateful for my advisor, Dr. Amy Lobben for her encouragement and support of my work, and Dr. Chris Bone for agreeing to be part of my thesis committee. I am very appreciative for extensive review of my work and invaluable advice that they have given.

I am thankful to everyone in the Department of Geography, InfoGraphics Lab and Spatial Computation, Cognition and Complexity (S3C) Lab for their support and inspiration. Thanks for being part of my journey; it has been a profound and meaningful experience, for which I am deeply grateful. I would also like to thank Cartography and Geographic Information Society (CaGIS) and Cartography Specialty Group (CSG) of Association of American Geographers (AAG) for their generous research grants.

Finally, of course, a huge thank you to my parents, my family, and my friends for their boundless love and encouragement. Being 8,000 miles away from some of you was not easy. Well, home is where the heart is, but my heart had to roam.

To my parents and family.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

The overarching goal of this thesis is to evaluate the variability of representativeness of sampled data that are voluntarily submitted through social media, particularly Twitter. For this thesis, the term representativeness refers to the degree to which the opinion of Twitter users in a particular place reflects the opinion of the entire population of that place. Specifically, this thesis examines the variability of representativeness of Twitter data through a comparative analysis between opinions expressed towards President Barack Obama on Twitter, and opinions derived from presidential performance approval data collected from public opinion surveys.

The increase in the use of Twitter and the utility of Twitter's spatial elements like geotagging and user-defined locations have shed light on the practicality of tweets as a source of public opinion data. Data about Twitter users and their behavior provide grounds for social scientists to understand the way people voice out their opinions on the Internet differ from that in real world.

Due to this potential, scientists are now actively exploring and experimenting with Twitter data to find out if specific patterns in tweets might be able to reflect real world events. Twitter maps are an example of such exploration. Every so often Twitter maps are circulated online, depicting Twitter users' response towards a social phenomenon. Typically, these maps would display individual tweets as point data, suggesting places with high and low response density. Occasionally, these points are aggregated to larger spatial units, like counties.

While these maps can be captivating, they could also be somewhat misleading because most of the time they are only highlighting "digital divide" such as urban versus rural, and high versus low population. The pattern and point distribution in the Twitter map examples in Figure 1.1, 1.2, and 1.3 clearly show the visual similarities of Twitter maps with the population map of the United States in Figure 1.4. These Twitter maps often end up not representing overall public opinion accurately as they are merely mirroring population patterns. What we are still unable to deduce from these maps is whether it is true that majority of the people in these highlighted area really care about the subject that is being mapped, as suggested by the maps.

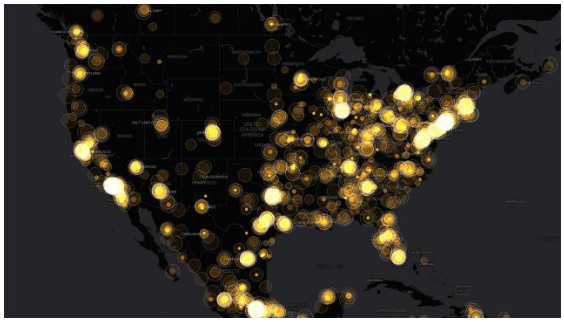

Figure 1.1. Twitter map showing tweets mentioning #Oscar, during Academy Awards 2013. Source: Fast Co Create

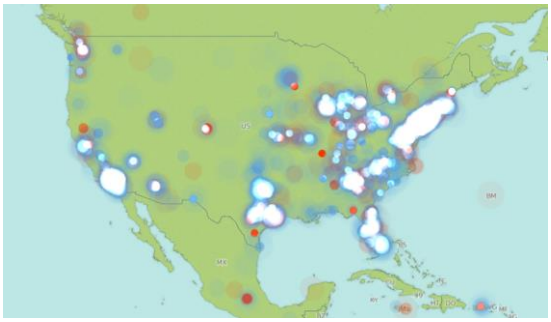

Figure 1.2. Twitter map showing tweets mentioning #WorldCup, during USA vs. Ghana World Cup match in 2014. Source: metro.us

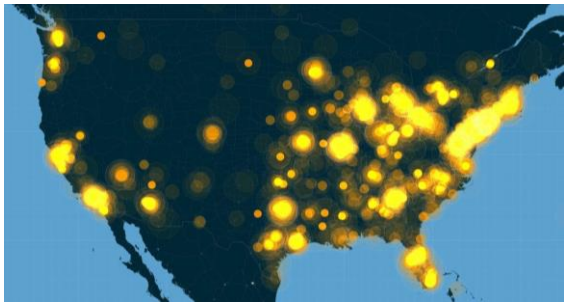

Figure 1.3. Twitter map showing tweets mentioning #Ferguson, created on August 12 2014. Source: CityLab

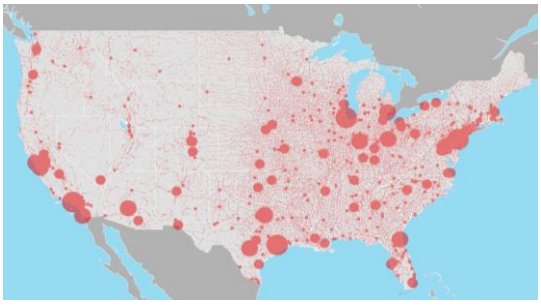

Figure 1.4. Proportional symbol map showing U.S. cities and their population.

Other than Twitter maps, social scientists are also aware of the potential of Twitter sentiment analysis as a less laborious technique to gather public opinion more extensively compared to traditional surveys. However, evidence to support this potential has yet to be recognized. Previous studies examining links between Twitter and public opinion did not come to a consensus; these studies settled with mixed results in their attempt to prove the accuracy of tweets in reflecting the opinion of the mass. Results from several studies did demonstrate a correlative relationship between Twitter sentiments and public opinion derived from polls (O'Connor, Balasubramanyan, Routledge & Smith, 2010; Tumasjan, Sprenger, Sandner & Welpe, 2010; Shi, Agarwal, Agrawal, Garg & Spoelstra, 2012), but others have found out that such correlation is almost absent (Hong & Nadler, 2010; Gayo-Avello, 2012).

It is important to note that these studies were conducted at the national level where the data were aggregated to represent the entire American population. It is very likely that the spatial variation of the sentiments could have been lost due to the aggregation, leading to inconsistency in the reported nationwide representativeness. Moreover, we do not know if national level representativeness is applicable and easily extrapolated to characterize representativeness at the state level. Is Twitter data representativeness for two different states comparatively and statistically similar?

Therefore, the main reason why analysis of spatial variability of Twitter data representativeness is inherently important is because studying Twitter data and public opinion polls at the national level do not tell us much about the *variation* of representativeness; it requires the analysis to be conducted at a finer spatial scale to reveal whether Twitter population could speak on behalf of the actual population. It is

almost impossible to evaluate the potential of tweets to be representative of population based on correlation at the national level alone since representativeness could vary across space, and could also be influenced by various socioeconomic and demographic factors. Furthermore, inspecting representativeness solely at the national level could lead to the modifiable areal unit problem, where statistical bias is more likely to occur when point-based measures of spatial phenomena like tweets, are aggregated into a very large spatial unit. Examining representativeness at the state level would be worthwhile in minimizing this problem.

Quoting Mislove et al. (2012), social media data exploration initiatives like Twitter maps are capable of prompting mass interest on the utility of Twitter data, but considering that the results from the previously mentioned studies are so inconsistent, we still have one big question to ponder: *Are Twitter users a representative sample of a particular population?* This thesis attempts to answer this main inquiry, driven by three research questions:

1. What is the national level Twitter data representativeness in the context of President Barack Obama presidential performance approval in the summer of 2015?
2. What is the spatial and statistical variability of Twitter data representativeness at the state level?
3. What geographic characteristics can explain this variability?

The goal of the first research question is to determine the degree of Twitter data representativeness of the United States overall. I anticipate the representativeness to be

low with no direct correlation between Twitter data and the ground truth data. The second research question then extends the representativeness analysis to a finer spatial scale, state level. In addition, through question two, spatial and statistical variability between states is investigated. I would expect the state level variability of representativeness to be high, suggesting a lack of direct connection between state level and national level representativeness. The focus of the third research question is to explore the potential causal variables affecting representativeness variability. These variables include: overall state population, urban population, non-white population, political preferences, educational attainment, and state median income. My hypothesis is that the overall state population will be the most distinct predictor of the variability.

CHAPTER II

BACKGROUND

Social Media and Geography

Despite the growing body of literature on social media, little attention has been paid to potential spatial variability of social media demographics. The bridge between social media, Big Data, and geography has been present for a while, but the array of academic literature is mostly available only in fields like computer sciences and media studies, with little to no geographic connotation. However, geographers recognize the utility of social media data because location-based social networks provide real-time spatiotemporal data (De Longueville, Smith & Luraschi, 2009).

Twitter has a more substantial presence within geographers' research domain compared to other social media platforms like Facebook, Instagram, and Foursquare. Unlike most social network sites, Twitter enables instant access to their public data by making a significant portion of their data available through an application programming interface (API). Twitter API is an open source interface that allows developers to pull through incoming tweet data from around the world and use them for analysis. Such availability and simplicity have made Twitter one of the leading data sources in social communication research. Google Scholar listed more than 3.2 million academic manuscripts citing the API (Leetaru, Wang, Cao, Padmanabhan & Shook, 2013). Relative to other social media sites' APIs, Twitter is widely known among developers for its characteristics related to both time and space, highlighting its utility in spatiotemporal

research. The Twitter API provides timing of tweets (i.e. the individual occurrences of Twitter data) to one second accuracy (De Longueville et al., 2009).

Relative to its temporal counterpart, the spatial component of Twitter offers more retrieval options in which geographic location of a tweet can be specified in two different ways. First, user location can be provided manually on their profile page. In addition, individual tweets may contain geographic coordinates or sometimes geocoded place names, which are usually more accurate and less subjective.

This twofold way of determining locations has become a dilemma for geographers in Twitter research, deliberating on the issue of precision and accuracy. Shi et al. (2012) advised that researchers should not ignore non-geotagged tweets in their studies. This is because geotagged tweets only account for extremely small percentage of the entire Twitter data: only less than 1.5% of the tweets contain geographic coordinates, making it necessary to develop an algorithm that can determine the location of as many tweets as possible. On the other extreme, Li, Goodchild and Xu (2013) ultimately discarded non-georeferenced tweets, only taking tweets that have point locational information with relatively high precision into account. Crampton et al. (2013), however took an intermediate position by suggesting that researchers should look "beyond geotagging" which can be done by including "ground truth" data to support an analysis.

Social Media and Volunteered Geographic Information (VGI)

The utility of Twitter in geography is further established through the use of tweets as volunteered geographic information (VGI). The term VGI is coined by Michael F. Goodchild in 2007, which he described as the process of "harnessing of tools to create,

assemble, and disseminate geographic data provided voluntarily by individuals" (Goodchild, 2007). The increased popularity of VGI is assisted by the expansion of Internet coverage, the advent of Web 2.0, as well as increased individual accessibility to computers and smartphones. VGI is a perfect example of the transition from passive Internet users to active participants who are not merely using data, but also generating them. Every individual's contribution and participation have led to the emergence of citizen-based or crowd-sourced science like VGI where scientific research and data collection are conducted by amateur scientists.

Applications of VGI have been proven to be useful and exploratory. A large fraction of prior social media research has incorporated VGI for emergency responses, particularly in the events of natural disasters (Ashktorab, Brown, Nandi & Culotta, 2014; Palen, Anderson, Mark & Martin, 2010; Pozdnoukhov & Kaiser, 2011; Starbird & Palen, 2010) and disease mitigation (Chunara, Andrews & Brownstein, 2012; Moorhead et al., 2013; Salathé & Khandelwal, 2011; Scanfeld, Scanfeld & Larson, 2010; Schmidt, 2012). Social media has also been used to study the well-being of society (Frank, Mitchell, Dodds & Danforth, 2013; Gruzd, Doison & Mai, 2011; Mitchell, Frank, Harris, Dodds & Danforth, 2013; Quercia, Ellis, Capra & Crowfort, 2012). In the past few years, social media websites play an important role in social activism by making it easier for people to acquire related information (Bonilla & Rosa, 2015; Crampton et al., 2013; Tremayne, 2013). Other uses of social media VGI are relatively less common but still highly practical, such as transportation research (Mai & Hranac, 2012; Sasaki, Nagano, Ueno & Cho, 2012) as well as financial and economic analysis (Bollen & Mao, 2011; Zhang, Fuehres & Gloor, 2011).

Social Media and Representativeness Studies

Regarding representativeness, there are wide-ranging studies that concentrate on bias on the sampled tweets generated by the API, but most of these studies put more weight on demographic bias. A vast amount of literature in social media studies are mostly concerned about demographic representativeness, especially socioeconomic characteristics of social media users (Longley, Adnan & Lansley, 2013; Mislove et al., 2012; Sloan et al., 2013). A few studies did have some discussion on spatial variability of social media data representativeness, but only as a peripheral subtopic.

The most relevant study about social media representativeness is perhaps the one conducted by Ceron, Curini, Iacus and Porro (2013). They found out that Internet users, particularly on Twitter, are not essentially representative of the entire population of a specific region, but the data still have remarkable ability to show correlation to some degree. Other than that, Graham, Stephens and Hale (2013) attempts to explain the spatial variability of Twitter data representativeness at the international level by comparing number of geocoded tweets with each country's Internet population. The study revealed that there is a high variation in the relationship between number of tweets and Internet population, inconsistent with the notion that the Global North – predominantly North America and Europe – was thought to be the biggest producers of Twitter data.

Social Media and Public Opinion

Other than the practices listed above, VGI from social media is also used to gather public opinion. This practice is often used exclusively in politics, where researchers use

tweets to forecast elections and predict citizens' political inclination (Barbera & Rivero 2014; Gayo-Avello, 2012; Hong & Nadler, 2010; O'Connor et al., 2010).

Prior to the emergence of social media networks, it was challenging to gather public opinion data. Questionnaires, surveys, and polls used to be the only method used to gather public opinion. These methods are still used today – despite the fact that the process is rather slow and labor-intensive – often with limited sample sizes. In the late 1990s and early 2000s, public opinion data mining began to become less arduous. This is partly due to the expansion of Internet coverage and the emergence of online messageboards, forums, and blogs (Wesolowski, 2014; Zhang, 2008). However, opinions were often hidden in long chains of posts and threads, and there were no automated ways available to retrieve them easily. Beginning the late 2000s, social media websites have dramatically changed the common practices in public opinion research. Social media websites have made public opinion retrieval process easier by providing continuous stream of opinion data like tweets and introducing various methods like APIs to retrieve such data.

At a broad spatial scale like national level, numerous studies have attempted to compare sentiments on Twitter with sentiments derived from public opinion polls to assess Twitter's potential in measuring public opinion. Both O'Connor et al. (2010) and Hong & Nadler (2010) compared Twitter data at the U.S. national level with results from public opinion polls to look for possible correlative relationships between the two. The results from both studies differ significantly. O'Connor et al. (2010) discovered that the correlations can be as high as 80%. They stressed that such correlations are able to capture substantial macro-scale trends to some degree, despite that variation across

multiple datasets still exists. On the other hand, Hong and Nadler (2010) found out that the relationship between public opinion polls and Twitter data is not statistically significant.

At a finer spatial scale like counties, election results are more often used as baseline data and the studies mostly focused on political discourse instead of general public opinion. Some of these studies asserted that the content of a tweet tends to reflect political landscape in real life, and therefore can be used to forecast actual election outcomes (Shi et al., 2012; Tumasjan et al., 2010). These studies reveal that the number of tweets closely match the vote proportion in election outcomes, proving the feasibility of using social media to predict voting behavior. Nonetheless, there are also studies that argue that the representativeness of Twitter data, especially in terms of political preference, has been significantly inflated (Gayo-Avello, 2012).

Significance of the Thesis

Based on this literature review, majority of social media representativeness studies are prediction analyses that take two basic forms and this thesis attempts to incorporate both of them. The first form of representative studies is done through comparisons of Twitter data with political election outcome at a reasonably fine spatial level, such as counties. Political election results are a much less subjective ground truth data compared to public opinion polls, but they do not take temporal changes into account. Election results are generated in one particular day i.e. the Election Day, and usually they do not capture variations across time. In order to consider temporal differences, some representativeness studies utilize public opinion polls as ground truth

data because these polls are conducted continuously throughout the year. However, most of these polls do not have spatial resolution fine enough for comparisons across space. This thesis combines these two forms by looking at the variation of representativeness spatially while retaining possible time-sensitive effects.

This literature review also reveals a gap within social media representativeness research: most of the previous studies have made an attempt to assess Twitter's potential in measuring public opinion, but only by studying representativeness at the national level and disregarding possible variations that might be present across the states. This thesis attempts to narrow this gap by comparing Twitter and public opinion sentiments at the state level, and eventually examining the variability that may exist.

CHAPTER III

METHODOLOGY

Figure 3.1 shows a flowchart of the thesis methodology. This section organizes the methods sequentially. Comparative analysis subsection organizes the statistical analyses used based on research questions.

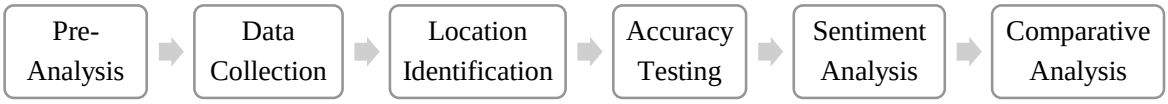

Figure 3.1. Streamlined methodology flowchart.

Pre-Analysis

Selection of Social Media

Due to the fact that social media is still developing and growing, spatiotemporal social media research in geography is fairly new. Amongst social networking websites that exist today, Facebook was launched in 2004, followed by Twitter in 2006, and Instagram in 2010. Despite being relatively new, these sites have garnered a massive amount of users in a very short time. Twitter, for example, grew 1,460 percent between 2008 and 2009, with approximately 44.5 million users globally (Scanfeld et al., 2010). Preliminary testing of each social media site's API methods led to the conclusion that the Twitter API would be the best choice for mining social media data. Due to the large volume of tweets published on Twitter, social media studies nearly exclusively utilize Twitter (Leetaru et al., 2013).

From a technical standpoint, the Twitter API has the most comprehensive API documentation and offers greatest flexibility on data access compared to APIs of other social media websites. The existence of online community of developers who continuously post sample scripts and coding tutorials to Stack Overflow and Github helps advancing the documentation even further. In other words, Twitter offers the most straightforward development tool and the most extensive support management, making it the most practical to be used by non-programmers.

The format of the streamed Twitter data also makes Twitter an obvious choice for this project; the data are structured in a format that is easy to read and understand. All things considered, Twitter provides the best means for spatiotemporal social media research as it allows geographic location and user sentiment to be identified easily, while at the same time yielding relatively great amount of data.

Selection of Case Study

The case study chosen for this project is President Barack Obama's job performance as viewed by Americans. This topic is selected based on the availability of ground truth data, which are the poll results derived from nationwide public opinion surveys. Almost all social survey research centers include a specific question about President Obama's job performance in their monthly surveys. For instance, both Pew Research Center and CNN/ORC regularly feature "*Do you approve or disapprove of the way Barack Obama is handling his job as president?*" as one of the questions in their polls.

This case study is also selected for temporal consistency reasons; the ground truth data need to be gathered at the same time the tweets were streamed and collected. These social survey research centers conduct their surveys as often as every week, hence making it possible to match the temporal resolution of the two datasets (Twitter data and ground truth data) together.

Data Collection

Ground Truth Data Collection

"Ground truth" is a set of measurements that is known to be highly or perfectly accurate. The selection of political survey data was made based on three criteria: survey question, temporal resolution and spatial resolution. Initially, six different social survey research centers were considered: CNN/ORC, Pew Research Center, Gallup, Rasmussen Reports, YouGov/The Economist, and Reuters/IPSOS. In the end, only two surveys satisfied all three conditions: CNN/ORC and Pew Research Center. Gallup polls, while they fulfilled these criteria, were eventually discarded due to high subscription fees.

To be included in this research, each poll must inquire at least one question that address about respondents' opinion regarding President Obama's job performance. The question can be worded differently, but it has to be synonymous to "*What do you think of President Obama's performance as president?*" All surveys include this question in their questionnaires.

To ensure temporal match of the polling and Twitter data, the poll must also be done within the time the Twitter data were collected i.e. May 2015 to September 2015. Temporal match is important since events and actions taken by President Obama may

affect people's sentiment towards the President's job performance. For this research, the assumption is that the temporal patterns of Presidential job performance would be reflected in both the ground truth and the Twitter data. All political surveys have conducted at least one poll that address the question mentioned above within the said timeframe.

The third condition requires that the poll dataset can be aggregated based on state names. Aggregation can be accomplished in a variety of ways, as long as the dataset includes any of the following: complete state name (for example Oregon, Wisconsin, Arkansas), state name abbreviation (OR, WI, AR) or Federal Information Processing Standard (FIPS) state numeric code (41 for Oregon, 55 for Wisconsin, 05 for Arkansas). Only two polls satisfy this condition: CNN/ORC and Pew Research Center. CNN/ORC dataset consists of a spreadsheet that includes a state abbreviation field, while Pew Research Center uses the FIPS state numeric code.

The ground truth datasets are generally a .zip file that includes a questionnaire and a .csv file of detailed tabular poll results (Appendix A). These tables typically contain question numbers, respondents' answers, and their geographic region (Midwest, South, Pacific Northwest etc.). Some polls would also include socioeconomic background of the respondents such as gender, ethnicity, highest education qualification, occupational sector, and annual household income. Most questions provide either binary (yes or no) or multiple choices of answers. Questions regarding President Obama's presidential performance are often followed by four answer choices: approve, disapprove, neutral, and prefer not to answer. All polls anonymize their respondents.

In terms of methodology consistency and validity, both Pew Research Center and CNN/ORC can be considered reliable. Pew Research Center uses weighting methods to statistically adjust their samples to make sure that they reflect the population as closely as possible, particularly in terms of demographic characteristics such as educational background and income. Meanwhile, FiveThirtyEight's Pollster Ratings (2014) rated CNN/ORC with an A- for having one of the lowest errors and least political bias relative to other public opinion polls.

Twitter Data Collection

Twitter produces more than 200 million tweets a day and allows developers' access to its freely available streaming API (Morstatter, Pfeffer & Huan, 2013). For this project, I employed one of the most common methods used to retrieve sample of public tweets, called "Tweetcrawling". This method involves the application of Twitter API, mediated by a Python script. The script collects only tweets that contain the word "Obama" for 20 weeks from May 2015 to September 2015, irrespective of location (Appendix B). This method retrieves around 100,000 to 400,000 tweets per day. The collected Twitter data are organized in MySQL database tables. Each table contains rows of strings of user ID, date, time of the day, tweet, location, and geographic coordinates. Only user ID (e.g. 35362823) is recorded, which implies that the user's actual username will not be saved and they remain anonymous. Twitter API streams tweets in real-time, which means Twitter data collection is a continuous, uninterrupted process.

Every tweet returned by Twitter API is structured as nested key-value pairs (Figure 3.2). Each pair contains a property with its associated values and holds the

metadata about the *user* such as name, user-defined location, username, profile description, and user ID. The tweet also contains metadata about the *status update* such as time, date, geographic coordinates, device, and language.

{
Tweet ID: **27585738348495,**
User Information: {
User ID: **255673849,**
Username: **@rudyomri,**
Profile description: "**Graduate student at @Univ_Of_Oregon."**
User-defined location: **"Eugene, OR",**
Image: **"/profile_images/639461666929774593",**
Number of followers: **144,**
Number of followings: **279,**
Number of tweets posted: **422,**
Time zone: **UTC – 8,**
Language: **English,**
}.
Tweet Information: {
Device used: **iPhone 6,**
Link to tweet:
"https://twitter.com/rudyomri/status/646123037037756417",
Date and time the tweet was posted: **2015-09-22 08:45:49,**
Tweet: **"Finally signed up for #nacis2015! Now the tricky part... flights and hotel rooms...",**
Location of the tweet: **[44.044241, -123.073812],**
Retweet count: **1,**
Favorite count: **4,**
},
}

Figure 3.2. A simplified version of a tweet's metadata. This study only utilizes the properties in red.

Subsequently, the collected tweets are "sliced" and grouped together within the threshold of the dates where the poll was conducted. For instance, for the CNN/ORC poll that took place from May 29 to May 31, the collected tweets that were posted on these dates are grouped together to form a set of tweets. These sets of tweets are then analyzed

so that the total aggregated sentiments of Twitter data for that period of time can be determined.

Location Identification

Location Parsing Algorithm

Dealing with a large volume of tweets is a messy task and location parsing is an essential step to clean up those data. Location parsing is important for two purposes: first, to tag as many tweets as possible with state names, and second, to ensure that the collected Twitter data only take public opinion of Americans into account. This step is implemented in two different ways: through identification of geographic coordinates (Figure 3.3) and through retrieving location information specified by users (Figure 3.4).

Figure 3.3. Example of a geotagged tweet.

Figure 3.4. Example of a user-defined location.

Geotagged tweets (Figure 3.3) have latitudes and longitudes embedded to them which allow the exact location of the tweets to be identified easily. The individual state

name of the geotagged tweets can be determined by projecting the geographic coordinates of tweets as points on top of a Shapefile of U.S. states in ArcGIS. Geotagged tweets that were not posted from the United States are discarded.

For non-geotagged tweets (Figure 3.4), I retrieved the location based on the "Location" field that was typed into the profile by the user. User-defined locations are then matched with collection of census places programmatically to identify the states they belong to. MySQL database tables that are used to store collected tweets are parsed thrice, through three different comma-delimited file (CSV) spreadsheets – cities, states, and countries – using a short algorithm written in Python programming language. In the first run, tweets are parsed through the list of state names. Tweets with state names identified are transferred to the "final table", while the ones with no identified state names are retained. In the second run, the retained tweets are parsed through the list of city names, attempting to determine the state names based on those cities. Similar to the first run, if the state names are identified, those tweets will be transferred to the "final table". In the third run, the "final table" is then parsed through the list of international place names to filter out possible non-US locations such as India (often conflict with India and Indianapolis) and Alexandria, Egypt (often conflict with Alexandria, Virginia).

As an example, in the first run the script will identify the state name in user-defined location of "Denver, Colorado", tag the tweet with "Colorado" and move it to the "final table". But if the tweet has a location that reads "Denver", the tweet is retained. In the second run, the script will then try to find the city name "Denver" in the list of city names, look for its matching state, register the state name "Colorado" to the tweet and move it to the "final table". By the end of this step, the "final table" will contain tweets

with geographic locations identified and is ready for sentiment analysis. Figure 3.5 shows the workflow of state name identification process for non-geotagged tweets.

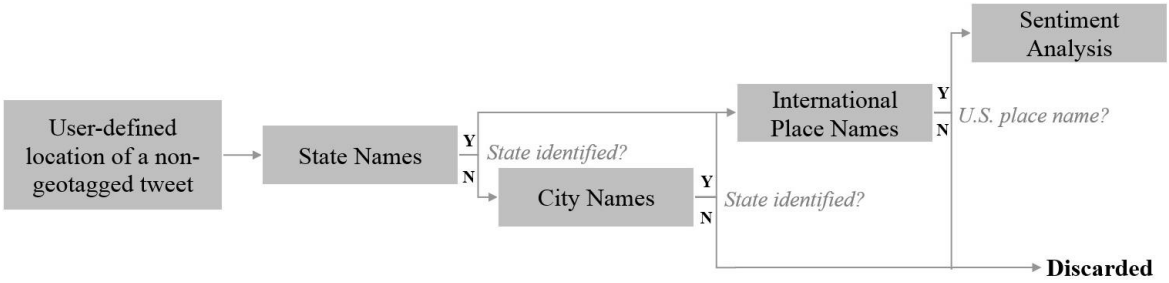

Figure 3.5. Simplified workflow of the location parsing algorithm.

Accuracy Testing

Using a sample dataset containing 1,000 tweets, I conducted an accuracy test for the location parsing algorithm based on a modified version of "precision and recall" approach. "Precision and recall" is typically used to evaluate classifier output quality by measuring relevance and accuracy of a particular system. The first step of the test is to manually go through each tweet in the sample dataset and tag each tweet as either true negative, true positive, false negative, and false positive. *True negative* denotes that the locations are appropriately left unidentified. *True positive* indicates that the locations are correctly identified by the algorithm. *False negative* represents locations that should be identified but the algorithm failed to do so. *False positive* means that the locations are incorrectly identified by the algorithm.

The equation used for the accuracy test is:

$$ACC = \frac{TN + TP}{n} \times 100\% \qquad (1)$$

where,

ACC = accuracy,

TN = number of true negative observations,

TP = number of true positive observations,

n = number of tweets in the sample dataset

Sentiment Analysis

Nasukawa and Yi (2003) writes that the fundamental goal of sentiment analysis is to identify how sentiments are expressed in short text and to find out whether the expressions indicate positive (favorable, approval or agreement) or negative (unfavorable, disapproval or disagreement) opinions toward the subject. Sentiment analysis uses natural language processing, text analysis, and computational linguistics to identify and extract subjective information in tweets. Sentiment analysis of tweets usually involves three components: identification of sentiment expressions in a tweet, polarity and strength of those expressions, and their relationship to the entire tweet.

Typically, sentiment analysis relies on "prebaked dictionaries" or "lexicons" that contain subjective, opinion-related words in English language where each word has its own predetermined polarity score. To put it simply, final polarity scores for each tweet are computed by an algorithm, which determines the sentiment of that tweet (Appendix

B). Negative scores suggest negative sentiments, scores close to zero indicate neutrality, and positive scores represent positive opinions.

For this study, I used AFINN-111 subjectivity lexicon (Figure 3.6) published by Nielsen (2011) to determine the sentiments. The lexicon consists of 2,477 English words encoded with sentiment polarity scores between minus five (negative) and plus five (positive). This lexicon is chosen due to its accuracy and simplicity, as Koto and Adriani (2015) acknowledged in their study. They evaluated the effectiveness of nine different lexicons regarding their applications with Twitter data. The study focuses on two sentiment domains: polarity and subjectivity, and tests out each lexicon with four different datasets using four different experiment methods: Naive Bayes, Neural Network, SVM, and Linear Regression. The study reveals that AFINN-111 scores among the highest for all four methods, with accuracy rate between 58.7% and 75.2%, making it one of the most accurate lexicons in the study to be used as a baseline for Twitter sentiment analysis.

```
"abandon -2"            "abused -3"
"abandons -2"           "abuses -3"
"abandoned -2"          "accept 1"
"absentee -1"           "accepting 1"
"absentees -1"          "accepts 1"
"aboard 1"              "accepted 1"
"abducted -2"           "accident -2"
"abduction -2"          "accidental -2"
"abductions -2"         "accidentally -2"
"abuse -3"              "accidents -2"
```

Figure 3.6. First 20 words in Nielsen (2011) subjectivity lexicon. Each word is paired with a polarity score

The sentiment of each individual tweet is calculated using this equation:

$$s = \frac{\Sigma x}{\sqrt{N}} \qquad (2)$$

where,

s = sentiment of a tweet,

x = polarity score of each word,

N = length of the tweet

This equation, adopted from Nielsen (2011), normalizes the sum of the individual words' sentiment. The purpose of normalization is to adjust values and make them comparable to each other. Sentiment of each tweet can be scaled in various ways and the most common way is by dividing the total sentiments for each tweet by its *N* (normalizing by mean) or 1 (no normalization). The formula that I used in this study implements square root normalization, which is a compromise between the two. Square root normalization minimizes extreme individual sentiment values that otherwise would occur in the two methods.

Consider these two tweets; one is fairly long and another one is relatively short:

1. *"Obama is great!"*
2. *"My opinion on President Barack Obama is that he is a great leader throughout his presidency."*

If these tweets are normalized by mean, with the word "great" encoded with +3, short text often scores extreme values (here is +1) while longer text scores closer to neutral

(here is +0.23) despite both tweets appear to have similar overall sentiment. On the other hand, without normalization, these tweets have exactly the same sentiment value (both are +3). Square root normalization is a solution that can minimize these discrepancies by assigning values that seem more reasonable. For these two tweets, normalizing the sentiments by square root normalization gives an overall sentiment value of +0.63 to the longer tweet and +1.73 to the shorter tweet. Figure 3.7 shows an example of how the algorithm determines the overall sentiment of a tweet.

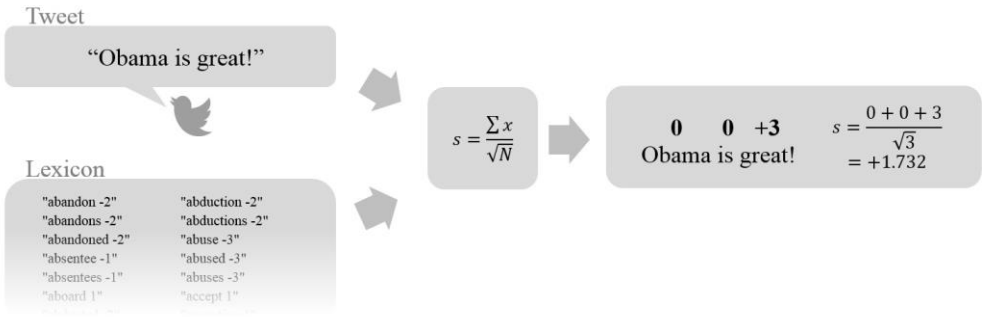

Figure 3.7. Simplified workflow of the sentiment analysis process.

Comparative Analysis

In order to quantify the representativeness, I computed the mean absolute error (MAE) using this equation:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\ \hat{\theta}_i - \theta_i| \qquad (3)$$

where,

N = number of observations,

$\widehat{\theta}$ = predicted value,

θ = actual value

For this study, this formula is modified to:

N = number of dataset pairs i.e. 5 (five rounds of public opinion polls),

$\widehat{\theta}$ = percentage of positive sentiment in Twitter dataset,

θ = percentage of positive sentiment in ground truth dataset

MAE, along with its variations like mean absolute scaled error and mean squared error, is frequently used in predictive models to determine a forecast's accuracy (Hyndman & Koehler, 2005). MAE computes the average magnitude of the errors in a set of forecasts, which eventually measures the accuracy of the predictions. For this study, I modified the concept of MAE so that the value characterizes the degree of Twitter data representativeness. Twitter data are assumed to be the "predictor," while ground truth data convey the "actual" values. Low MAE is associated with high degree of representativeness and low bias, while high MAE indicates otherwise. In other words, degree of representativeness is inversely proportional to the value of MAE. Table 3.1 shows an example of MAE calculation for Texas.

Table 3.1. Example of MAE calculation for Texas.

Date Range	Twitter, $\widehat{\theta}$	Ground Truth, θ	Absolute Difference, $\lvert \widehat{\theta} - \theta \rvert$
May 12 - 18	47.15%	37.30%	9.85%
May 29 - 31	41.12%	42.50%	1.38%
June 26 - 28	53.37%	55.71%	2.34%
July 14 - 20	57.83%	43.87%	13.96%
July 22 - 25	40.57%	33.33%	7.24%
N = 5			$\sum \lvert \widehat{\theta} - \theta \rvert =$ 34.77%

$$MAE_{TX} = \frac{1}{N}\sum_{i=1}^{N} |\ \hat{\theta}_i - \theta_i|$$

$$= \frac{34.77}{5}$$

$$= 6.954\%$$

Research Question 1: What is the national level Twitter data representativeness in the context of President Barack Obama presidential performance approval in the summer of 2015?

High correlative relationship between Twitter and ground truth dataset is not anticipated and I generated scatter plots and Pearson's r correlation coefficients to test this expectation. MAE for the entire dataset is also calculated to determine the nationwide Twitter data representativeness. The representativeness measured here serves as the baseline data in evaluating the variability of state level representativeness in research question two.

Research Question 2: What is the spatial and statistical variability of Twitter data representativeness at the state level?

The second research question brings the outcome from the first question further down to a more detailed spatial resolution, which is at the state level. The goal of this section is to examine the variability of the representativeness of Twitter data based on spatial relationships and statistical significance.

For *spatial* variability, the MAE for each state is displayed in a choropleth map. Variability is assessed based on homogeneity and visual pattern or clusters. High variability is implied through heterogeneity of values displayed by the states. On the other hand, for *statistical* variability, I used histogram and range of the MAE to study the dispersion of MAE across the individual states. I also used Chi-Squared Test for the Variance to test whether the variance of the MAE is significant enough to account for high variability in Twitter data representativeness.

Research Question 3: What geographic characteristics can explain this variability?

Analysis of the third research question involves comparing the values of MAE with the adjusted values of geographic characteristics to identify a causal variable that can best explain the variability of state level representativeness. The geographic variables chosen are total state population, percentage of urban population, percentage of non-white population, percentage of Democrat votes in the 2012 election, percentage of college graduates, and median state income. Selection of these variables are made based

on a study done by Pew Research Center, "*Who Votes, Who Doesn't, and Why: Regular Voters, Intermittent Voters, and Those Who Don't*" (2009).

The rationale behind selecting total state population and urban population as comparative variables is to assess possible high versus low population bias and urban versus rural bias in the representativeness measures. The other four variables are merely socioeconomic factors that are assumed could potentially affect Twitter data representativeness in various ways, for instance political and sociological bias in Twitter sentiments (political preferences and non-white population), degree of mobile phone ownership and Internet accessibility (median income) and level of technology literacy (educational attainment). The data for total state population, non-white population, and median income have to be logarithmically transformed to remove skewness.

I used Pearson's r correlation to examine the relationship between individual state representativeness and those variables. I also used multiple regression analysis to inspect the degree of influence of those geographic variables on individual state representativeness collectively. The Test on Individual Regression Coefficients (*t*-test) results and standard coefficients from multiple regression analysis are used to identify a geographic variable that has the strongest relationship with the degree of representativeness. Other than statistical measures, I also created individual choropleth maps to allow for visual comparison between values of the MAE and the geographic variables.

CHAPTER IV

RESULTS

This section organizes the results parallel to the steps in the methodology.

Data Collection and Location Identification

As shown in Table 4.1, Twitter API collected 4,968,809 tweets within the five opinion polls' timeframes. The location parsing algorithm successfully identified geographic locations for 28% (1,392,426) of those tweets. The remaining 72% (3,576,383) are discarded for having either international or ambiguous user-defined locations. 99.5% (1,385,305) of these located tweets are non-geotagged tweets that have their respective state name identified through location parsing algorithm. The remaining 0.5% (7,121) are geotagged tweets with precise latitude and longitude pair. Accuracy test of 1,000 tweets for location parsing shows that the algorithm is considerably accurate. The algorithm successfully determined correct locations for 953 tweets (95.3%) and mistakenly identified the remaining 47 (4.7%).

Table 4.1. Number of tweets collected, grouped by the polls they belong to.

Dates	Poll	Number of collected tweets	Number of non-geotagged tweets with location identified	Number of geotagged tweets	Number of tweets considered for sentiment analysis
May 12 – 18	Pew	717,655	182,107	1,221	183,328
May 29 – 31	CNN/ORC	392,289	105,106	747	105,853
June 26 – 28	CNN/ORC	654,238	156,040	802	156,842
July 14 – 20	Pew	2,043,360	684,895	2,327	687,222
July 22 – 25	CNN/ORC	1,161,267	257,157	2,024	259,181
Total		**4,968,809**	**1,385,305**	**7,121**	**1,392,426**

Sentiment Analysis

Out of 1,392,426 tweets that were prepared for sentiment analysis, approximately 65% (897,310) of the tweets were discarded due to neutral classification. One of a few factors that leads to a high percentage of neutral tweets is the limited vocabulary of the lexicon. While Koto and Adriani (2015) did prove that AFINN-111 is highly accurate, AFINN-111 does not include most non-colloquial terms in the list. For example, AFINN-111 does not contain the word "espouse" and thus the tweet "*I espoused Obama throughout his presidency*" is classified as neutral, despite its obvious positive sentiment. The inability of the algorithm to detect hidden sentiments could also cause a tweet to appear neutral. For instance, the tweet "*Obama speaking at SXSW, not attending Nancy Reagan's funeral*" in spite of being critical of President Obama, is categorized as neutral due to the absence of negative words.

For this analysis, neutral sentiments are discarded because they constitute an extremely high percentage of the tweets. For some states, neutral sentiments could entail as much as 75% of the entire pool of tweets. The high amount of tweets with neutral sentiments significantly deflates the percentage of positive and negative sentiments. Using Colorado's first round dataset pair as an example, 64.95% of the tweets are neutral, with 16.31% as positive and 18.74% as negative. This shows that such deflation could obscure the variation that we would like to see in the MAE calculation; the massive amount of neutral sentiments has reduced the difference between the two polarities. In order to avoid the deflation, tweets with neutral sentiments are ultimately discarded, leaving 46.54% tweets as positive and 53.46% as negative.

Once the tweets with neutral sentiments are removed, the sum of positive and negative sentiments is now complementary to each other, adding up to 100% for every Twitter and ground truth dataset. Following the abovementioned example for Colorado, 46.54% of Twitter data and 61.54% of ground truth are identified as positive. That makes negative sentiments to account for 53.46% (100 - 46.54) for Twitter and 38.46% (100 - 61.54) for ground truth, mirroring the lesser half of the entire pool. The Twitter-ground truth difference is the same for both sentiments, which is 15%, suggesting that it is somewhat redundant to consider the direction of polarity. Furthermore, such directionality cannot be retained in the MAE calculation. For these reasons I decided to only include positive sentiments in the comparative analysis (Table 4.2).

Table 4.2. Percentage of tweets mentioning "Obama" with positive sentiments and percentage of public approval of President Obama's job performance according to results from five different public opinion polls.

	Round 1, Pew (May 12 – May 18)		Round 2, CNN/ORC (May 29 – May 31)		Round 3, CNN/ORC (June 26 – June 28)		Round 4, Pew (July 14 – July 20)		Round 5, CNN/ORC (July 22 – July 25)	
State	Twitter	Ground Truth	Twitter	Ground Truth	Twitter	Ground Truth	Twitter	Ground Truth	Twitter	Ground Truth
AK	50.27	50.00	37.93	0.00	56.04	0.00	56.81	100.00	45.15	0.00
AL	45.88	35.14	43.42	27.78	52.70	52.38	59.00	44.44	40.71	44.44
AR	43.94	29.41	40.37	37.50	50.80	66.67	59.86	15.79	43.43	38.46
AZ	44.35	45.95	41.75	42.86	54.72	25.00	57.02	58.14	43.32	38.46
CA	48.67	63.04	41.96	45.74	57.45	55.81	55.68	62.71	42.29	56.48
CO	46.54	61.54	41.85	30.77	52.68	56.52	57.19	58.62	43.70	56.25
CT	42.99	52.38	48.05	80.00	53.94	63.64	57.10	44.44	44.98	50.00
DC	50.64	75.00	42.25	100.00	59.81	100.00	54.08	83.33	46.49	66.67
DE	46.86	50.00	45.83	50.00	48.53	33.33	56.95	50.00	41.74	33.33
FL	45.32	51.75	42.43	30.51	53.04	64.71	57.92	56.57	40.35	45.45
GA	48.46	40.00	45.28	61.54	53.82	56.45	56.45	43.14	40.23	44.44
HI	42.46	66.67	42.95	0.00	57.45	0.00	55.53	50.00	46.97	0.00
IA	49.74	48.00	42.90	33.33	58.99	23.08	51.05	53.57	48.14	14.29
ID	43.43	50.00	38.25	54.55	47.48	14.29	60.96	45.00	44.14	44.44
IL	50.11	59.42	42.28	57.14	56.51	66.67	55.47	56.72	39.02	61.90
IN	45.35	42.22	43.60	29.03	54.24	42.86	57.99	41.54	41.93	30.00
KS	42.79	48.28	41.94	0.00	62.19	15.38	57.66	50.00	39.16	40.00
KY	45.07	17.39	41.67	22.22	52.47	46.67	58.67	33.33	40.61	37.50
LA	45.98	43.75	41.98	25.00	56.76	53.33	55.14	48.00	44.99	6.67

Table 4.2 (continued).

	Round 1, Pew (May 12 – May 18)		Round 2, CNN/ORC (May 29 – May 31)		Round 3, CNN/ORC (June 26 – June 28)		Round 4, Pew (July 14 – July 20)		Round 5, CNN/ORC (July 22 – July 25)	
State	Twitter	Ground Truth	Twitter	Ground Truth	Twitter	Ground Truth	Twitter	Ground Truth	Twitter	Ground Truth
MA	46.73	53.85	43.11	62.50	55.60	69.23	57.37	48.78	42.93	53.33
MD	51.03	67.57	39.34	75.00	52.82	79.17	53.76	58.62	43.18	64.29
ME	45.59	58.33	37.35	60.00	60.00	25.00	56.73	33.33	44.60	50.00
MI	48.38	51.92	45.05	48.28	52.71	51.22	57.99	46.81	45.15	37.50
MN	45.22	65.85	44.20	50.00	56.47	46.15	55.34	55.32	41.91	57.14
MO	46.51	38.64	41.04	37.50	55.63	57.14	58.69	51.35	39.63	45.00
MS	45.87	33.33	46.00	20.00	57.36	52.94	58.72	66.67	39.27	50.00
MT	46.07	28.57	39.29	33.33	53.75	16.67	55.42	50.00	43.80	16.67
NC	50.09	45.76	42.81	37.04	55.00	64.00	57.20	48.68	41.49	55.26
ND	62.56	33.33	43.14	0.00	59.42	0.00	55.06	50.00	43.95	25.00
NE	50.43	41.67	50.57	33.33	61.07	40.00	56.09	31.25	43.78	50.00
NH	47.60	71.43	40.65	0.00	59.71	50.00	56.25	25.00	43.51	0.00
NJ	44.38	50.00	43.76	60.87	55.05	63.33	58.04	73.91	40.06	60.00
NM	53.11	72.22	41.22	42.86	49.55	66.67	53.57	35.29	39.47	50.00
NV	47.40	27.78	39.34	22.22	56.98	50.00	55.56	18.18	40.45	37.50
NY	47.70	68.24	44.85	47.69	56.42	62.75	56.57	65.35	43.40	60.66
OH	46.65	24.73	44.31	46.51	55.97	61.11	58.20	43.24	40.71	54.39
OK	46.98	28.57	37.40	26.92	56.16	53.33	56.28	48.57	37.53	58.82
OR	44.48	44.00	46.67	40.00	55.72	57.14	56.25	47.62	39.60	36.84
PA	47.21	48.10	42.17	41.18	57.52	59.32	55.69	38.82	39.63	43.90

Table 4.2 (continued).

	Round 1, Pew (May 12 – May 18)		Round 2, CNN/ORC (May 29 – May 31)		Round 3, CNN/ORC (June 26 – June 28)		Round 4, Pew (July 14 – July 20)		Round 5, CNN/ORC (July 22 – July 25)	
State	Twitter	Ground Truth	Twitter	Ground Truth	Twitter	Ground Truth	Twitter	Ground Truth	Twitter	Ground Truth
RI	52.15	83.33	47.57	33.33	62.61	33.33	52.90	66.67	52.16	16.67
SC	40.86	32.00	42.18	36.84	50.20	58.06	59.65	50.00	38.48	28.57
SD	55.00	42.86	38.71	33.33	54.10	25.00	54.76	20.00	43.42	100.00
TN	49.02	42.31	45.74	30.77	54.39	48.00	58.69	38.24	39.72	47.06
TX	47.15	37.30	41.12	42.50	53.37	55.71	57.83	43.87	40.57	33.33
UT	54.05	50.00	36.32	33.33	57.76	33.33	58.19	40.00	35.26	50.00
VA	46.55	48.28	43.33	30.43	55.48	63.16	56.84	43.08	41.22	39.29
VT	54.78	60.00	30.91	100.00	61.80	75.00	52.84	80.00	39.41	66.67
WA	47.07	35.19	45.44	40.00	54.44	63.64	55.27	66.67	43.07	56.52
WI	48.49	51.35	43.54	65.22	55.35	40.91	56.65	61.54	41.85	40.00
WV	51.82	46.67	45.45	37.50	57.69	40.00	58.15	66.67	37.31	27.27
WY	42.22	33.33	54.17	0.00	51.40	25.00	59.49	33.33	34.35	0.00

Comparative Analysis

Research Question 1: What is the national level Twitter data representativeness in the context of President Barack Obama presidential performance approval in the summer of 2015?

At a broader nationwide scale, the relationship between Twitter and ground truth dataset is statistically non-significant. Scatter plots and Pearson's r correlation coefficients for each dataset pair suggest that the relationship is fairly weak, ranging from 0.10 to -0.24. As previously mentioned, correlation between these two datasets is not anticipated and this is evidenced by the findings from statistical analyses. Table 4.3 shows the correlation coefficients while Figure 4.1 displays the scatter plots for every dataset pair. The MAE for the entire country is 14.95%, which is fairly distant from the perfect representativeness of 0%, consistent with the results from Pearson's r correlation analysis.

Table 4.3. Pearson's r correlation coefficients for each Twitter – ground truth dataset pair.

Dataset pair	Correlation coefficient
May 12 – May 18	0.2113
May 29 – May 31	-0.1824
June 26 – June 28	-0.0948
July 14 – July 20	-0.2360
July 22 – July 25	-0.1853

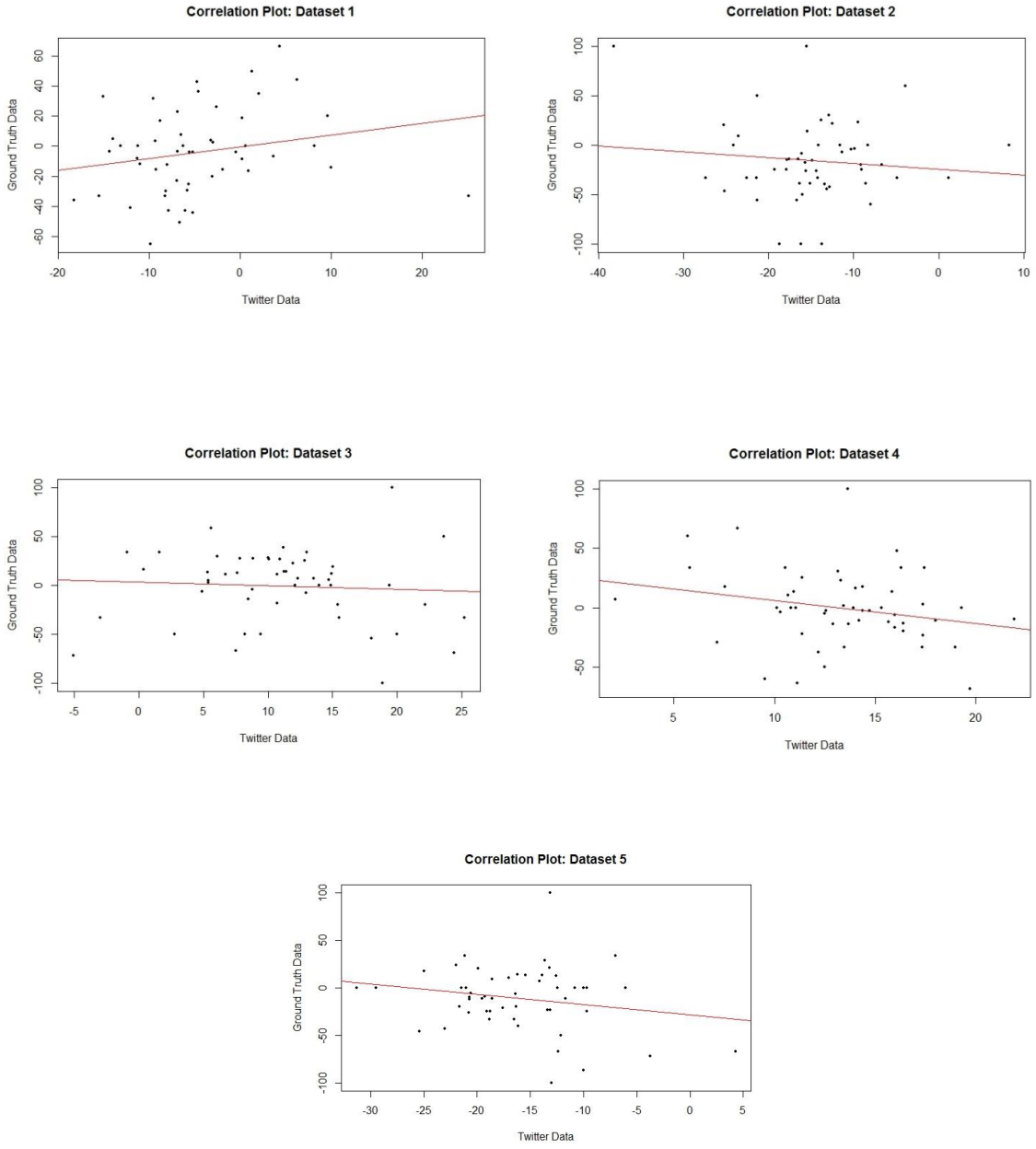

Figure 4.1. Scatter plots for Twitter – ground truth dataset pairs.

Research Question 2: What is the spatial and statistical variability of Twitter data representativeness at the state level?

Statistical Variability of Representativeness

The variability of the representativeness is assessed by measuring the degree of representativeness at a finer spatial scale, state level. High range of MAE reveals that the statistical variability of Twitter data representativeness in regard to President Obama's job performance is considerably high. The MAE range is 27.64% and the variance is 51.0585, with Arkansas having the highest error (lowest degree of representativeness), and Arizona the lowest (highest degree of representativeness). Table 4.4 and Figure 4.2 shows the MAE values for every state.

Table 4.4. MAE for every state.

State	MAE (%)	State	MAE (%)	State	MAE (%)
AK	21.73	KY	26.50	NY	14.66
AL	12.65	LA	4.68	OH	18.44
AR	29.30	MA	7.85	OK	13.06
AZ	1.36	MD	10.70	OR	4.55
CA	10.70	ME	18.07	PA	8.88
CO	8.21	MI	7.36	RI	22.47
CT	11.02	MN	10.33	SC	9.25
DC	26.80	MO	7.61	SD	23.45
DE	5.05	MS	10.24	TN	13.58
FL	3.89	MT	11.46	TX	11.90
GA	10.89	NC	6.42	UT	11.12
HI	14.87	ND	17.14	VA	7.74
IA	2.13	NE	16.80	VT	16.19
ID	11.26	NH	27.54	WA	11.64
IL	5.28	NJ	10.75	WI	3.88
IN	9.79	NM	18.70	WV	6.84
KS	6.57	NV	28.50	WY	17.52

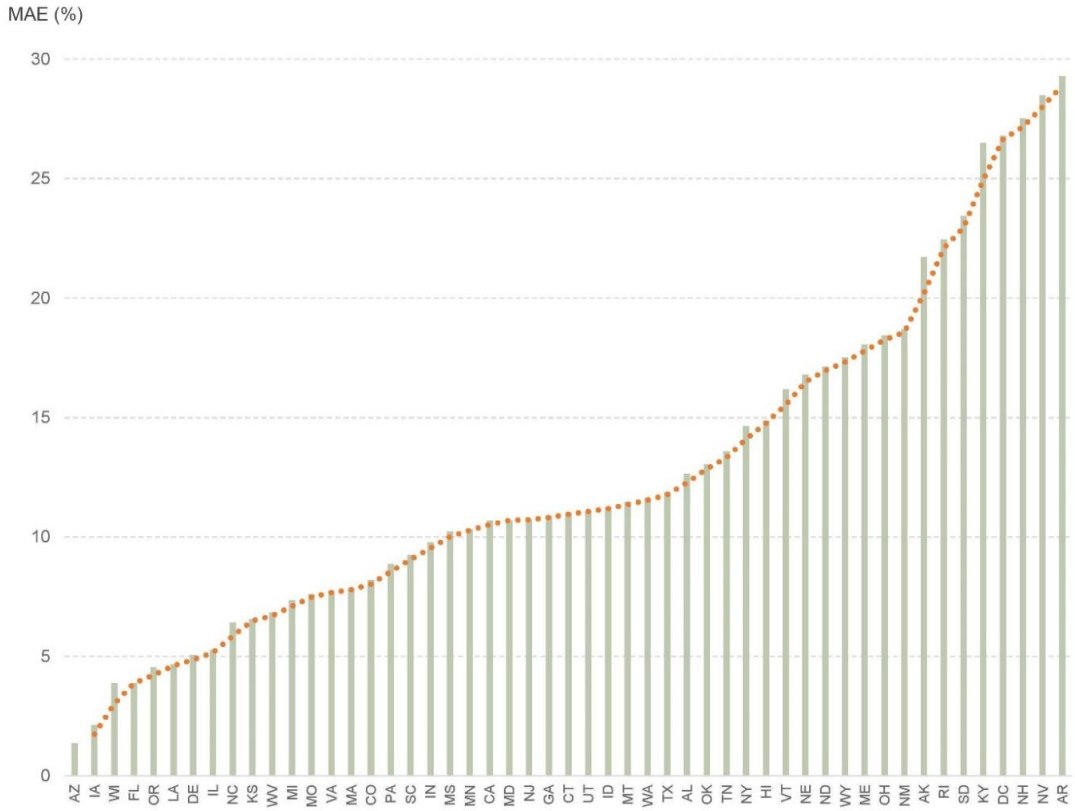

Figure 4.2. MAE for every state, ordered from lowest (highest representativeness) to highest (lowest representativeness).

The histogram in Figure 4.3 shows the distribution of MAE values across different states. Based on the right-skewed pattern of the histogram, it is apparent that majority of the states have low MAE errors or high degree of representativeness. 19 states (37.3%) have MAE lower than 10%, and 31 (60.8%) have MAE lower than 12%. On the other hand, only eight states (15.7%) have MAE higher than 20%.

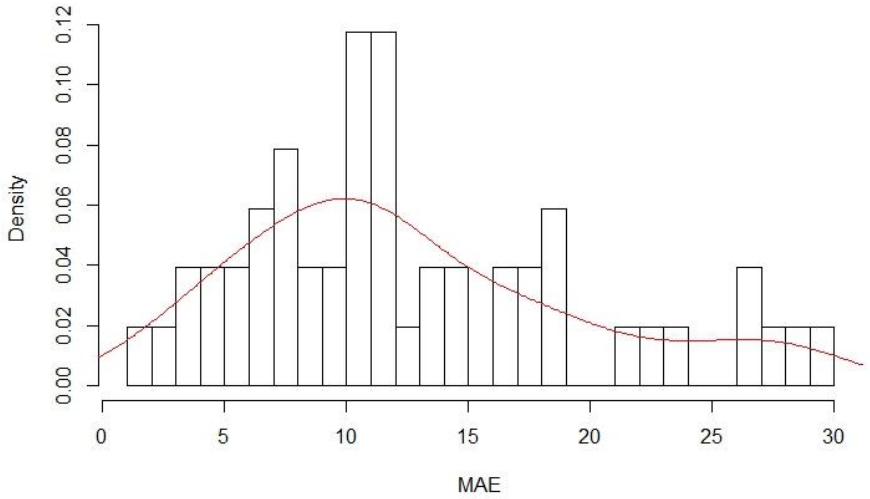

Figure 4.3. Histogram of individual state's MAE.

Table 4.5. Results from Chi-Squared Test for the Variance of MAE.

H_0: $\sigma^2 = x$ H_a: $\sigma^2 > x$	Test statistic, *T*	Critical values		
		$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
$x = 15$	170.1949	33.523	37.818	40.415
$x = 10$	255.2924			
$x = 5$	510.0585			

Degrees of freedom: N – 1 = 50
Accept H_a if T > critical values

Chi-Squared Test for the Variance is used to test the significance of MAE variance. Three values of theoretical variance are chosen for the null hypotheses: 5, 10 and 15. These values are selected as one-tailed thresholds of variances, serving as multiple hypothetical boundaries between low versus high variability. Since we already found out that the variance is 51.0585, this test is done to test the significance of the variance, whether it is remarkably high or not. Based on the results, the test statistics, *T*, are much larger than the critical values, α, across all confidence levels, and across all variance thresholds, *x*, indicating that the statistical variability of Twitter data is significantly high.

Spatial Variability of Representativeness

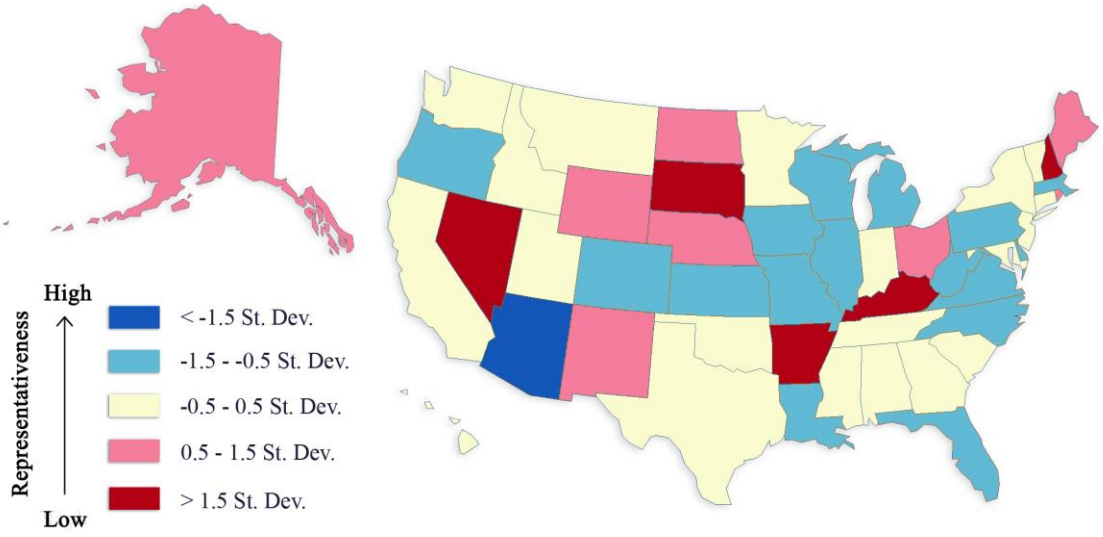

Figure 4.4. Choropleth map showing individual state's MAE, categorized using standard deviation classification.

The choropleth map in Figure 4.4, as well as the distribution of representativeness level in both Table 4.6 and Figure 4.5, illustrate the MAE of every individual state, classified based on standard deviation. This classification scheme shows the MAE variability across the states and identifies states that are above, below or close to the average MAE value. The further the MAE of a particular state departs from the *mean of MAE* of all states towards zero, the higher the representativeness is, and vice versa. In this map, the middle class "-0.5 – 0.5 St. Dev." Consists of states that are closest to the mean of MAE of all states. The upper two classes in shades of blue are states that have MAE values relatively more distant from the mean, but closer to the perfect representativeness of 0%. The lower two classes in shades of red are states that also have MAE values further from the mean, but closer to no representativeness of 100%.

Using this classification method, Arizona is the only state that exhibits perfect or near perfect representativeness, compared to six states that have very low representativeness on the lower extreme. However, there are more states in the perfect and high representativeness groups than in low and very low groups combined.

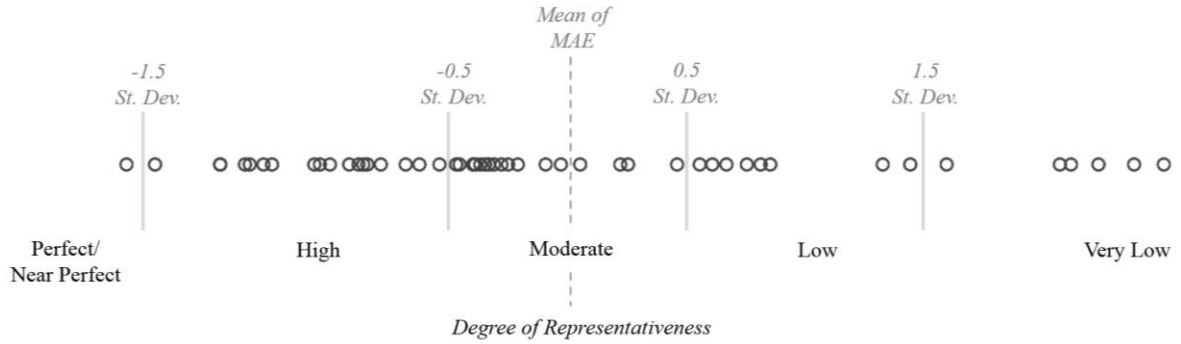

Figure 4.5. Distribution of representativeness level based on standard deviation classification.

Table 4.6. Distribution of representativeness level based on standard deviation classification.

Degree of Representativeness	Departure from mean of MAE	State
Perfect/Near Perfect	< -1.5 St. Dev.	Arizona (1)
High	-1.5 – -0.5 St. Dev.	Iowa, Wisconsin, Florida, Oregon, Louisiana, Delaware, Illinois, North Carolina, Kansas, West Virginia, Michigan, Missouri, Virginia, Massachusetts, Colorado, Pennsylvania (17)
Moderate	-0.5 – 0.5 St. Dev.	South Carolina, Indiana, Mississippi, Minnesota, Maryland, California, New Jersey, Georgia, Connecticut, Utah, Idaho, Montana, Washington, Texas, Alabama, Oklahoma, Tennessee, New York, Hawaii, Vermont (20)
Low	0.5 – 1.5 St. Dev.	Nebraska, North Dakota, Wyoming, Maine, Ohio, New Mexico, Alaska, Rhode Island (8)
Very Low	> 1.5 St. Dev.	South Dakota, Kentucky, District of Columbia, New Hampshire, Nevada, Arkansas (6)

Visually, high spatial variability can be observed from the map based on lack of homogeneity and presence of random pattern or observable clusters. Homogeneity in the map is implied by a uniform and consistent pattern across the states, suggesting that the variable being mapped – in this case, the MAE – has relatively similar magnitudes for all states. The choropleth map in Figure 4.4 clearly shows some sort of a random pattern with no homogeneity, indicating that the spatial variability of representativeness is high.

Research Question 3: What geographic characteristics can explain this variability?

Findings from the second research question inform us that the spatial variability of Twitter data representativeness is high. Research question three explores potential causal variables that are influencing this variability. This process involves comparing the values of MAE for each state with their geographic characteristics such as total state population, urban population, political preferences, median income, educational attainment, and non-white population. To allow for valid comparison and to remove major skewness, the data for some variables are adjusted through logarithmic transformation. The values used for the analysis are listed in Table 4.7. The raw data values for these variables can be found in Appendix C.

Table 4.7. Adjusted values of geographic variables for every individual state. The distribution of the actual values for total population, non-white population, and median income are highly skewed. These values have been transformed logarithmically to minimize the skewness.

State	Total Population[1]	Non-white Population[2]	Urban Population[3]	Political Preference[4]	Educational Attainment[5]	Median Income[6]
AL	6.66	1.48	59.04	38.36	23.80	4.61
AK	5.82	1.51	66.02	40.81	31.50	4.92
AZ	6.76	1.20	89.81	44.59	26.60	4.67
AR	6.44	1.30	56.16	36.88	32.00	4.58
CA	7.55	1.42	94.95	60.24	33.10	4.96
CO	6.68	1.08	86.15	51.49	27.40	4.73
CT	6.54	1.26	87.99	58.06	25.10	4.91
DE	5.92	1.46	83.30	58.61	29.60	4.93
DC	5.75	1.75	100.00	90.91	28.50	4.72
FL	7.24	1.34	91.16	50.01	26.90	4.67
GA	6.96	1.57	75.07	45.48	25.80	4.67
HI	6.10	1.87	91.93	70.55	25.10	5.08
ID	6.14	0.80	70.58	32.62	27.40	4.62
IL	7.11	1.35	88.49	57.60	25.70	4.74
IN	6.80	1.14	72.44	43.93	29.50	4.66
IA	6.47	0.88	64.02	51.99	31.00	4.69
KS	6.44	1.11	74.20	37.99	35.90	4.67
KY	6.62	1.06	58.38	37.80	29.20	4.60
LA	6.66	1.56	73.19	40.58	38.20	4.63
ME	6.12	0.68	38.66	56.27	35.60	4.68
MD	6.74	1.60	87.20	61.97	23.90	4.94
MA	6.81	1.23	91.97	60.65	35.70	4.77
MI	7.01	1.30	74.57	54.21	24.60	4.68
MN	6.72	1.14	73.27	52.65	26.40	4.78
MS	6.47	1.60	49.35	43.79	24.10	4.56
MO	6.76	1.21	70.44	44.38	28.70	4.64
MT	5.96	1.02	55.89	41.70	34.50	4.68
NE	6.25	1.01	73.13	38.03	48.50	4.69
NV	6.37	1.37	94.20	52.36	25.20	4.64
NH	6.12	0.76	60.30	51.98	22.50	4.79
NJ	6.94	1.42	94.68	58.38	34.00	4.98
NM	6.28	1.23	77.43	52.99	30.60	4.64
NY	7.29	1.46	87.87	63.35	22.70	4.73
NC	6.94	1.45	66.09	48.35	27.90	4.65
ND	5.81	1.02	59.90	38.69	25.30	4.70
OH	7.06	1.23	77.92	50.67	32.40	4.65
OK	6.55	1.39	66.24	33.23	30.50	4.61
OR	6.55	1.08	81.03	54.24	26.50	4.69
PA	7.09	1.23	78.66	51.97	25.60	4.70
RI	6.03	1.16	90.73	62.70	27.50	4.73

Table 4.7 (continued).

State	Total Population[1]	Non-white Population[2]	Urban Population[3]	Political Preference[4]	Educational Attainment[5]	Median Income[6]
SC	6.63	1.50	66.33	44.09	21.80	4.64
SD	5.89	1.15	56.65	39.87	24.30	4.68
TN	6.77	1.32	66.39	39.08	23.00	4.62
TX	7.36	1.29	84.70	41.38	25.30	4.68
UT	6.38	0.92	90.58	24.75	18.90	4.74
VT	5.80	0.68	38.90	66.57	21.40	4.71
VA	6.88	1.47	75.45	51.16	22.00	4.77
WA	6.79	1.27	84.05	56.16	21.00	4.70
WV	6.26	0.79	48.72	35.54	29.90	4.57
WI	6.74	1.08	70.15	52.83	19.60	4.70
WY	5.70	0.86	64.76	27.82	25.50	4.74

[1] Total Population by State, U.S. Bureau of the Census, 2014, logarithmically transformed.
[2] Non-White Population Percentage by State, U.S. Bureau of the Census, 2013, logarithmically transformed.
[3] Urban Percentage of the Population for States, U.S. Bureau of the Census, 2014
[4] Percentage of Democrat votes by state, U.S. Presidential Election Results, 2012.
[5] Educational Attainment by State, U.S. Bureau of the Census, 2009.
[6] Median Household Income (In 2009 Inflation-Adjusted Dollars), U.S. Bureau of the Census, 2009, logarithmically transformed.

As presented in Table 4.8 and Figure 4.6, Pearson's r correlation coefficients for most of the variables with the MAE are somewhat weak. Total state population is the only variable that shows a relatively strong correlation. With -0.46, state population would more likely influence the variability of representativeness than other variables.

Table 4.8. Pearson's r correlation coefficient between the MAE and six selected geographic variables.

Geographic characteristics	Correlation coefficient
Total population	-0.460
Non-white population	-0.051
Urban population	-0.166
Political preferences	0.058
Educational attainment	0.029
Median income	-0.038

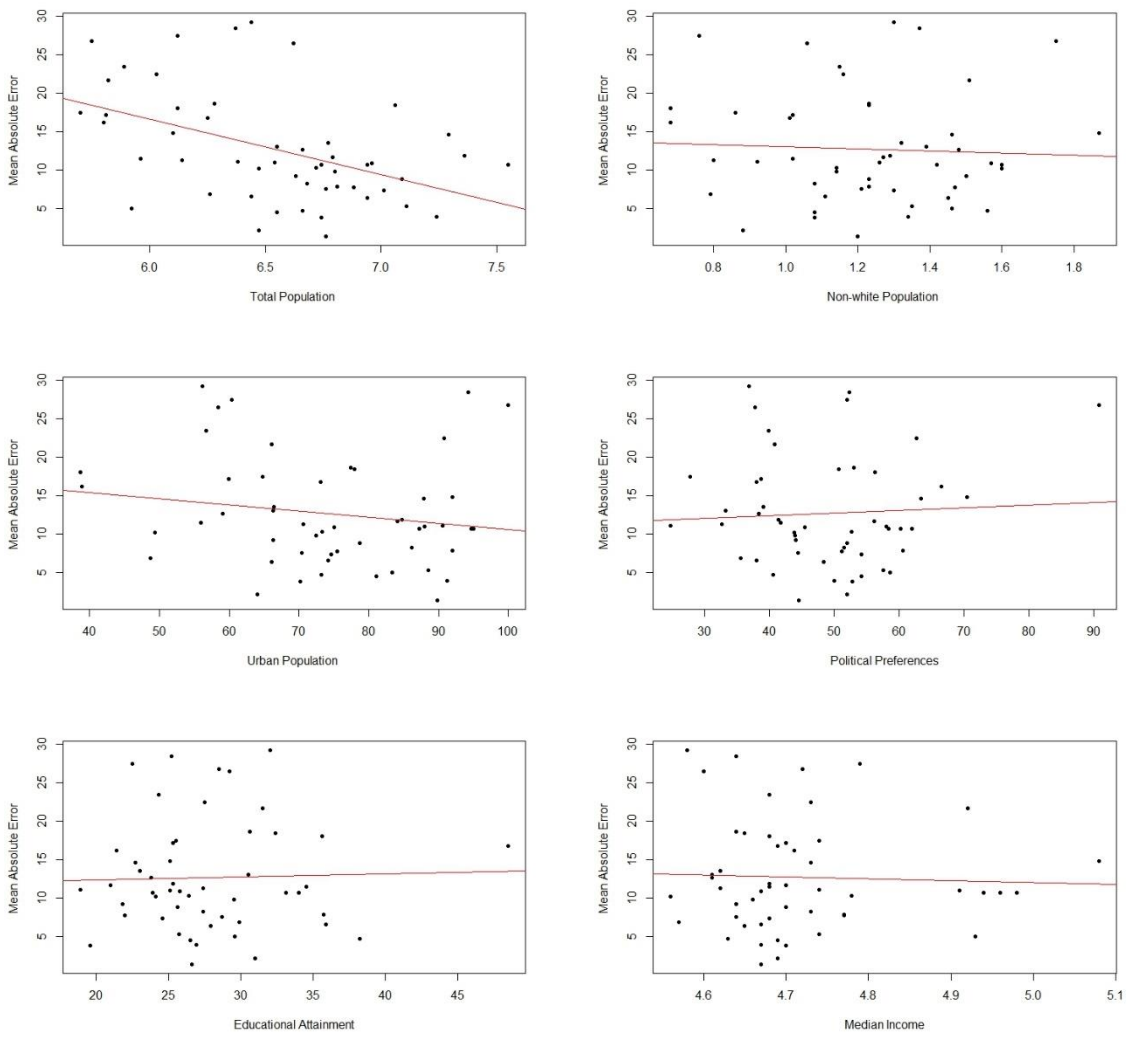

Figure 4.6. Scatter plots for the MAE and each of the geographic variables.

Figure 4.7 shows the individual choropleth maps of MAE values and the adjusted values of chosen geographic variables. The visual pattern in the choropleth maps is fairly consistent with the Pearson's r correlation coefficients in which the total state population map somewhat appears like an opposite reflection of the map of MAE, where highly populated states tend to demonstrate low MAE values. On the other hand, the maps for five other variables do not show any associative pattern in regard to the map of MAE.

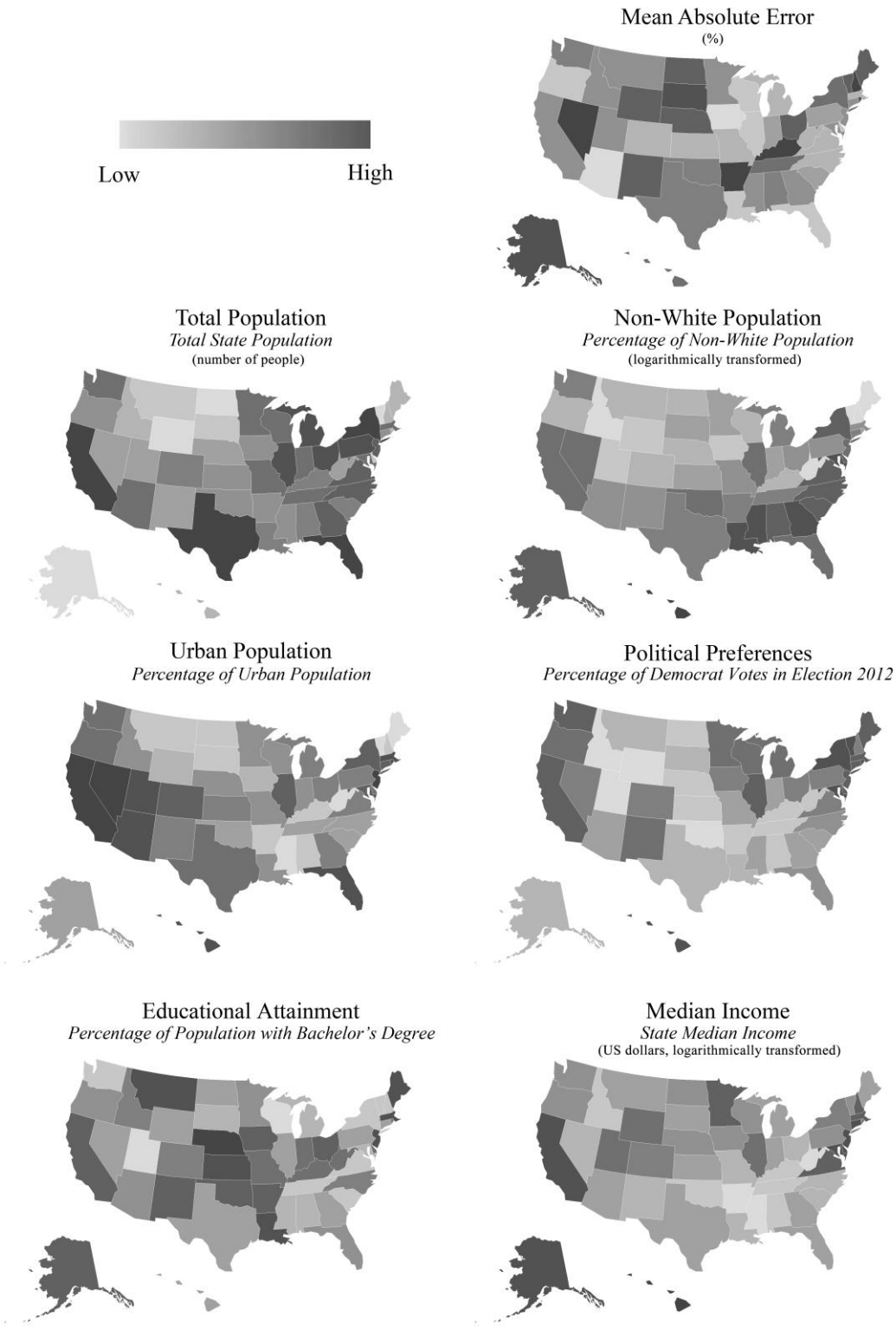

Figure 4.7. Choropleth maps of the MAE and the geographic variables.

I performed a multiple regression analysis to test the significance of all geographic variables collectively within a model, as shown in Table 4.9. Based on the outcome of the *F*-test, the model is statistically significant (*p*-value = 0.0494 < 0.05, *F*-statistic = 2.32). However, the six-predictor model is able to account for only 24% of the variance, as suggested by the *R*-squared value. This shows that the combination of these variables do not explain much variation in the MAE despite being statistically significant.

Table 4.9. Results from multiple regression analysis for all six geographic variables with the MAE.

	Standard Coefficient	*t*-value	*p*-value	Significance
Total population	-0.507	-3.320	0.002	**
Non-white population	0.112	0.716	0.478	ns
Urban population	-0.001	-0.003	0.998	ns
Political preferences	0.123	0.772	0.444	ns
Education attainment	0.008	0.059	0.953	ns
Median income	-0.150	-0.904	0.371	ns

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 'ns' 1
Min: -11.966, 1st Quartile: -3.757, Median: -0.304, 3rd Quartile: 2.837, Maximum: 15.151
Residual standard error: 6.639 on 44 degrees of freedom
Multiple *R*-squared: 0.2403
F-statistic: 2.32 on 6 and 44 DF
p-value: 0.0494

The variable with the greatest standardized coefficient is usually the best predictor among other variables, and in this case, it appears to be total state population. Standardized coefficient informs us the number of standard deviation change on the MAE that will be produced by a change of one standard deviation on the independent geographic variable concerned. For this particular analysis, a change of one standard

deviation in population will produce standard deviation that is closer to one compared to other three variables.

The individual *t*-tests from the multiple regression further attests that total state population is the best predictor variable for the MAE. The large *t*-value with a small *p*-value indicates that the correlation for total state population is significantly different from zero hence providing support to reject any possibilities that there is a weak relationship between total state population and the MAE. Figure 4.8 shows the correlation matrix plot of these relationships.

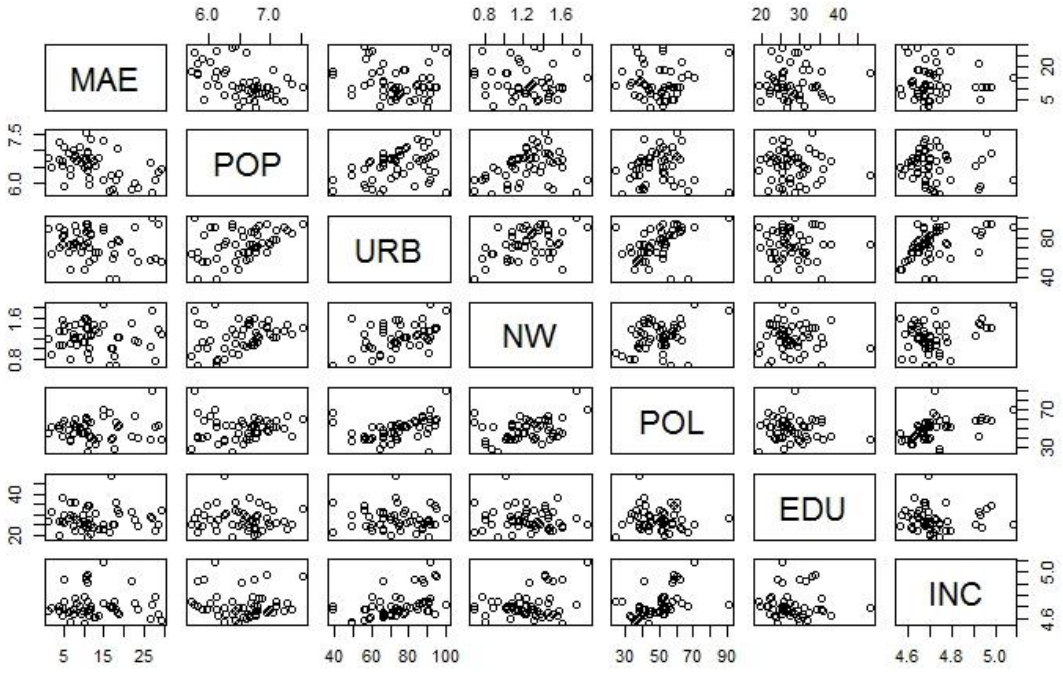

Figure 4.8. Matrix plot of correlation analysis.

CHAPTER V

CONCLUSION

This section begins with a comprehensive discussion of the analyses from the results section, followed by a broader conclusion in regards to the goal of the thesis which includes the limitations and implications of the thesis as well as suggestions for future social media representativeness studies.

Discussion of the Results

With respect to my first research question, I found out from the countrywide MAE that the overall outlook on the national representativeness is very low. Low nationwide MAE indicates that Twitter data are not able to "predict", or in this case, to reflect, the results of public opinion polls, suggesting that the whole collection of tweets are not representative of the entire American population. Plotting Twitter sentiments and ground truth sentiments on a scatter plot shows that Twitter and public opinions do not substantially reflect each other (Figure 4.1), which is evidenced by the low correlation coefficients ($-0.2 < p < +0.2$) (Table 4.3). However, such a weak correlation is not surprising at all since aggregating data to a larger set typically obscures the variation in the data and increases the possibility of errors, which is evidenced by the nationwide MAE. From these analyses, I concluded that at the national level Twitter data are not representative of the United States population.

Following the findings from the first research question, I sought to reveal the variation in the representativeness that was lost while aggregating the data to the national

level. This is done through my second research question where I discovered that the variability of Twitter data representativeness at the state level is high, both statistically and spatially. The high range and variance of the MAE implies that the state representativeness are highly varied and well dispersed. The results from Chi-Squared Test for the Variance (Table 4.5) prove that the variance of the MAE is significantly high, confirming the high variability.

In geography, spatial variability is exhibited when quantity that is measured at different places shows values that differ across those places. In this study, the spatial variability of Twitter data representativeness is noticeably high based on the presence of random visual pattern and the lack of homogeneity in the map (Figure 4.4). The analysis of representativeness at the state level also informs us that high representativeness is still exhibited by some states despite the low overall nationwide pattern of representativeness. In fact, there are more states showing higher than low or moderate degrees of representativeness.

With reference to the third research question, I discovered that total state population is the most fitting variable to predict the degree of Twitter data representativeness. Correlation coefficients (Table 4.8) primarily serve as the first pass in identifying such variable, with state population exhibiting the highest correlation. The results from multiple regression (Table 4.9) further validate the findings from the correlation analysis, showing that the relationship between total state population and Twitter data representativeness is much more consistent than any other tested variables.

Conclusion

This project has a few limitations which could have possibly affected the accuracy of the outcome. The first drawback is specifically related to the application of public opinion polls as "ground truth". There are many public opinion research centers available out there and often times they generate different results for the exact same topic. This suggests that even the ground truth data used in this study is subject to some sort of bias. Moreover, aggregating sentiments based on states might not be detailed enough spatially, but most polls only have states as their finest spatial resolution.

The sentiment analysis algorithm and the lexicon used in this thesis are fairly basic. Many language components such as sarcasm, connections between words, expressions, emoticons, and non-English tweets are not taken into account. A more complex sentiment analysis algorithm could be used to increase the accuracy, but it would require a higher level of understanding in machine learning and natural language processing. While the results could possibly get more accurate, the process is definitely more complicated. The AFINN-111 lexicon and the simple sentiment analysis algorithm that I used are perhaps the most optimum options for this level of study and this decision was made based on the balance between simplicity and accuracy that they offer.

There are a few takeaways that this thesis could offer for future social media representativeness research. Firstly, at the national level, Twitter data certainly do not always reflect opinions of the majority of Americans. Accumulating all tweets posted from the United States and determining their cumulative countrywide sentiments – like what most previous studies have done – is not an ideal approach to measure public opinion. For instance, if 80% of the tweets posted from the United States are critical

about Barack Obama, they do not necessarily indicate that a large percentage of American population would have the same judgment.

The high variability exhibited at the state level shows that there are more dimensions within the idea of Twitter data representativeness than what the past studies have proclaimed. Tweets may be representative for one state, but they are not necessarily the same for the other. This suggests that regardless how representative tweets are at the national level, the state level representativeness could be a complete hodgepodge: for some states, tweets could be almost perfectly representative of the population, while for some others, tweets could be completely insignificant. To put it in another way, Twitter data may reflect the opinion of the residents of a particular state, like Arizona, Iowa, and Florida, but it is not the case for some other states like New Hampshire, Nevada, and Arkansas (refer to Table 4.4). Such inconsistency provides another reason to doubt in the reliability of tweets to be used as main data source in public opinion research.

From the literature review, I learned that most previous studies examined Twitter data representativeness exclusively at the national level and thus neglecting potential variability that could be present at a finer spatial scale. Due to this, expectations on Twitter's potential as an alternative to public opinion polls were simply based on the results that are measured at the national level i.e. the United States as a whole. The findings from this thesis corroborates that such assessment can be fallacious since the representativeness is highly varied at the state level.

This study also reveals that the pattern of state population is fairly consistent with the pattern of state's Twitter data representativeness, in which highly populated states tend to have higher representativeness; hence providing grounds for a discussion that

representativeness is influenced by how populous a particular state is. Going into this research, I was already aware of the popular belief where number of tweets tends to follow population distribution despite the fact that the way Twitter API streams tweets is truly random with no preset biases, but I did not know if the degree of representativeness is also related to population. This study reveals that just like the number of tweets, degree of representativeness also has a close relationship with total population, especially individual state population. Simultaneously, the study also proves that Twitter data representativeness is not substantially influenced by how urbanized a state is, nor by how rich and educated the residents might be.

Twitter users are truly just a subset of a larger group of people. The findings from this thesis corroborate that those who voice out their opinion through their tweets are not necessarily representative of the group or place they belong to. However, since we also found out that representativeness go hand in hand with state population, we can deduce that maps that display Twitter sentiments can still be a reliable visualization medium of a particular phenomenon. These maps do not only depict *where most tweets are*, but they could possibly display both *where most tweets are* and *where tweets are a representative sample of the people in that place*. That is to say that we can perhaps trust Twitter data only if the tweets are used to represent people in highly populated areas, but not for places where population is sparse.

In a nutshell, the degree of representativeness of Twitter data is highly varied, implying that it is important for us to keep in mind that Twitter data are too volatile and complicated for public opinion measures. Nevertheless, this thesis perhaps has shed some

light to help determine the directions of future Twitter representativeness studies, especially in geography.

APPENDIX A

SAMPLE PUBLIC OPINION POLL QUESTIONNAIRES AND RESULTS

ALL RESPONDENTS:

First of all, (sir/ma'am) . . .I want to talk to you about a few issues that impact the United States of America.

PROGRAMMER NOTE: PLEASE INCLUDE THE FOLLOWING PROBE AFTER EACH QUESTION EXCEPT FOR DEMOS:

(IF DON'T KNOW/UNDECIDED/REFUSED IS THE VOLUNTEERED RESPONSE THEN INTERVIEWER PROBES) If you had to choose one of the response options I read to you, which would it be? (RE-READ QUESTION ONLY IF NECESSARY)

Q1 Do you approve or disapprove of the way Barack Obama is handling his job as president? (3301)

1 Approve
2 Disapprove
9 DON'T KNOW/UNDECIDED/REFUSED (vol.)

Q2 Do you approve or disapprove of the way Barack Obama is handling: (RANDOM ORDER)

1 Approve
2 Disapprove
9 DON'T KNOW/UNDECIDED/REFUSED (vol.)

a. The economy (3302)
b. ISIS (PRONO: EYE-siss), the Islamic militant group that controls some areas of Iraq and Syria (3303)
c. Race relations (3304)

(IF VERSION A: ASK)
d. Climate change (3305)
e. Illegal immigration (3306)
f. Government surveillance of U.S. citizens (3307)

(IF VERSION B: ASK)
g. Health care policy (3308)
h. Foreign affairs (3309)
i. Terrorism (3310)

Q3 Which of the following is the most important issue facing the country today? (RANDOM ORDER) (3311,3312)

01 The economy
02 The federal budget deficit
03 Health care
04 Education
05 The situation in Iraq and Syria
06 Terrorism
07 Energy and environmental policies
08 Illegal immigration
09 OTHER (vol.)
99 DON'T KNOW/UNDECIDED/REFUSED (vol.)

USORCCNN2015-005 Page 8

Sample questionnaire from a CNN/ORC poll. Question 1 asks respondents whether they approve or disapprove President Barack Obama's performance as the president of the United States

SEPARATELY FOR EACH VERSION OF THE INTRODUCTION FOR THE LANDLINE AND CELL FRAMES SEPARATELY. PLEASE RANDOMIZE INTRO LANGUAGE WITH ONE TREATMENT PER PHONE NUMBER NOT PER CALL.]

RANDOMIZE Q.1 AND Q.2
ASK ALL:
Q.1 Do you approve or disapprove of the way Barack Obama is handling his job as President? **[IF DK ENTER AS DK. IF DEPENDS PROBE ONCE WITH:** Overall do you approve or disapprove of the way Barack Obama is handling his job as President? **IF STILL DEPENDS ENTER AS DK]**

1 Approve
2 Disapprove
9 Don't know/Refused **(VOL.)**

RANDOMIZE Q.1 AND Q.2
ASK ALL:
Q.2 All in all, are you satisfied or dissatisfied with the way things are going in this country today?

1 Satisfied
2 Dissatisfied
9 Don't know/Refused **(VOL.)**

ASK ALL:
Q.3 Right now, which is more important for President Obama to focus on -- domestic policy or foreign policy?

1 Domestic policy
2 Foreign policy
3 Neither **(VOL.)**
4 Both **(VOL.)**
9 Don't know/Refused **(VOL.)**

NO QUESTIONS 4-9

ASK ALL:
Now thinking about how Barack Obama is handling some issues …
Q.10 Do you approve or disapprove of the way Barack Obama is handling **[INSERT ITEM, RANDOMIZE; OBSERVE FORM SPLITS]**? How about **[NEXT ITEM]**? **[REPEAT INTRODUCTION AS NECESSARY]**

a. Race relations
b. The economy
c. The nation's immigration policy

ASK FORM 1 ONLY:
d.F1 The nation's foreign policy
e.F1 International trade issues

ASK FORM 2 ONLY:
f.F2 Dealing with Iran
g.F2 The threat of terrorism

RESPONSE CATEGORIES:
1 Approve
2 Disapprove
9 Don't know/Refused **(VOL.)**

Sample questionnaire from a Pew Reserch Center poll. Question 1 asks respondents whether they approve or disapprove President Barack Obama's performance as the president of the United States.

STATE	CENDIV	REGION	CNNREG	METRO	URBAN	VERSION	DAY	TIMEDAY	Q1	Q2A	Q2B
MA	1	1	1	2	2	2	53115	1252	1	1	2
AK	1	1	1	2	2	2	53015	1135	2	1	2
CA	1	1	1		1	1	52915	1941	2	2	1
FL	1	1	1	2	2	2	53015	1303	1	1	2
CA	1	1	1	2	2	2	53015	1712	2	2	2
IL	1	1	1	2	2	2	53015	2056	1	1	2
ID	1	1	1	2	2	2	53115	1348	1	1	2
OR	1	1	1	2	2	1	53115	1321	2	2	2
ME	1	1	1		2	1	53015	1322	1	1	1
CA	1	1	1		99	2	53015	1312	2	2	2
NY	1	1	1	2	2	2	53115	1325	1	1	1
GA	1	1	1		2	1	52915	2053	9	2	2
OH	1	1	1		2	2	52915	2110	1	1	1
WI	1	1	1	2	2	1	53015	1346	2	2	2
ND	1	1	1		1	1	53015	1426	1	1	9
NY	1	1	1		1	2	53015	1152	1	1	1
TX	1	1	1		1	1	52915	2031	1	1	2
NE	1	1	1		1	2	52915	2018	2	1	1
TX	1	1	1		1	2	53015	1259	1	1	2
MA	1	1	1		2	1	53015	1450	1	1	1

Sample results from a CNN/ORC poll. Column highlighted in blue contains geographic information at the state level, while column shaded in yellow indicates the answer to the question related to President Barack Obama's job performance.

psraid	sample	int_date	fcall	attempt	refusal	ilang	usr	cregion	state	density	scregion	sstate	susr	igender	irace	form	llitext0	q1
100002	1	51215	150512	1	0	1	S	3	51	4	3	51	S	2	2	1	2	2
100004	1	51215	150512	1	0	1	S	1	34	2	1	34	S	1	1	2	2	2
100005	1	51215	150512	1	0	1	U	2	39	3	2	39	U	1	2	2	2	1
100006	1	51215	150512	1	0	1	S	2	17	2	2	17	S	1	1	1	2	2
100012	1	51815	150512	5	1	1	S	3	37	2	3	37	S	2	3	1	1	1
100016	1	51215	150512	1	0	1	S	3	5	1	3	5	S	2	3	1	1	1
100023	1	51215	150512	1	0	1	S	2	39	5	2	39	S	1	1	1	2	2
100025	1	51215	150512	1	0	1	R	2	26	1	2	26	R	1	2	2	2	2
100030	1	51215	150512	1	0	1	R	2	19	1	2	19	R	1	1	2	1	2
100031	1	51315	150512	2	0	1	R	3	21	1	3	21	R	1	2	1	2	9
100035	1	51515	150512	2	1	1	U	2	19	2	2	19	U	1	1	2	1	2
100039	1	51215	150512	1	0	1	S	3	1	1	3	1	S	1	2	2	2	2
100048	1	51515	150512	2	1	1	S	2	26	3	2	26	S	2	2	2	2	1
100056	1	51215	150512	1	0	1	U	2	31	3	2	31	U	1	2	1	2	1
100060	1	51315	150512	2	0	2	S	3	12	4	3	12	S	2	3	2	2	2
100062	1	51215	150512	1	0	1	R	2	27	1	2	27	R	1	2	1	2	2
100065	1	51515	150512	4	0	1	S	3	12	4	3	12	S	1	2	2	2	1
100069	1	51715	150512	3	1	1	R	2	26	2	2	26	R	1	3	2	1	9
100070	1	51315	150512	2	0	1	U	1	36	5	1	36	U	1	2	1	1	1

Sample results from a Pew Research Center poll. Column highlighted in blue contains geographic information at the state level, while column shaded in pink indicates the answer to the question related to President Barack Obama's job performance.

APPENDIX B

CODES ASSOCIATED WITH THESIS

Twitter Data Collection

```
from tweepy.streaming import StreamListener
from tweepy import OAuthHandler
from tweepy import Stream
import time
import MySQL database
import os
import twitter
import json
from twitter import *
import logging

ckey = 'xxx'
csecret = 'xxx'
atoken = xxx'
asecret = 'xxx'

db = MySQL database.connect(host="127.0.0.1", port= 3306, user="root",
passwd="xxx", db="twitter")
db.set_character_set('utf8')
data_list = []
count = 0
curr=db.cursor()

class listener(StreamListener):
   def on_data(self, data):
      global count
      if count <= 10000000000:
         json_data = json.loads(data)
         coords = json_data.get("coordinates")
         tweet = data.split(',"text":"')[1].split('","source')[0]
         time = data.split('"created_at":"')[1].split('","id')[0]
         userid = data.split('"id_str":"')[1].split('","text')[0]

         if coords is not None:
            xy = coords["coordinates"]
            lon = coords["coordinates"][0]
            lat = coords["coordinates"][1]
            location = "NA"
```

```
            curr.execute("""INSERT INTO june_pres (UserID, Date, X, Y, Location,
Tweet) VALUES
                (%s, %s, %s, %s, %s, %s);""",(userid, time, lon, lat, location, tweet))
            db.commit()
            count += 1

          if coords is None:
            location = data.split (',"location":"')[1].split('","url')[0]
            lon = "NA"
            lat = "NA"

            curr.execute("""INSERT INTO june_pres (UserID, Date, X, Y, Location,
Tweet)                          VALUES
                (%s, %s, %s, %s, %s, %s);""",(userid, time, lon, lat, location,  tweet))
            db.commit()

            count += 1

          return True
        else:
          return False

    def on_error(self, status):
        print status

auth = OAuthHandler(ckey, csecret)
auth.set_access_token(atoken, asecret)

def start_stream():
    while True:
        try:
            twitterStream = Stream(auth, listener(), timeout=30.0)
            twitterStream.filter(track=["obama"], async=False)
        except:
            logging.exception('There is an exception.')
            continue

start_stream()
```

Location Parsing

```
import csv
import string
import math
import MySQL database
import re

db=MySQL database.connect(host="127.0.0.1", port= 3306, user="root", passwd="xxx",
db="twitter_loc")
db.set_character_set('utf8')
curr=db.cursor()
curs=db.cursor(MySQL database.cursors.DictCursor)

file = 'C:/Users/PlaceNames.txt'

with open(file, mode='r',) as infile:
   reader = csv.reader(infile, delimiter='\t')
   mydict = {rows[0]:rows[1] for rows in reader}

if __name__ == '__main__':
   curs.execute("""SELECT Location FROM database_name """)
   result_set = curs.fetchall()
   x = 0
   for k in result_set:
      x += 1
      loc = k.get('Location')
      for eachkey in mydict.keys():
         if str(eachkey) in loc:
            eachvalue = mydict.get(eachkey)
            print "Value: " + str(x) + ", Location: " + loc + ", Key: " + eachkey + ", State: "
+ eachvalue
            curr.execute ("""
            UPDATE database_name
            SET LocationNew=%s
            WHERE id=%s
         """, (eachvalue, x))

         db.commit()

   print "Done"
```

Sentiment Analysis

```
import math
import MySQL database
import re

db=MySQL database.connect(host="127.0.0.1", port= 3306, user="root", passwd="xxx",
db="twitter_loc")
db.set_character_set('utf8')
curr=db.cursor()
curs=db.cursor(MySQL database.cursors.DictCursor)

filenameWords = 'words/AFINN-111.txt'
wordlist = dict(map(lambda (w, s): (w, int(s)), [
          ws.strip().split('\t') for ws in open(filenameWords) ]))

pattern_split = re.compile(r"\W+")

def sentiment(text):
   words = tweet.split(' ')
   sentiments = map(lambda word: wordlist.get(word, 0), words)
   if sentiments:
      sent = float(sum(sentiments))/math.sqrt(len(sentiments))
   else:
      sent = 0
   return sent

if __name__ == '__main__':
   curs.execute("""SELECT Tweet FROM database_name """)
   result_set = curs.fetchall()
   #print result_set
   count = 1
   for k in result_set:
      tweet = k.get('Tweet')
      sentval = sentiment(tweet)
      sentvaldec = "%.3f" % sentval
      print count, sentval
      if sentval > 0:
         sentimentcon = "1"
      if sentval < 0:
         sentimentcon = "-1"
      if sentval == 0:
         sentimentcon = "0"
      curr.execute ("""
        UPDATE database_name
        SET Sentiment=%s
```

```
        WHERE id=%s
    """, (sentvaldec, count))
    db.commit()
    count +=1
print "All sentiments have been identified."
```

APPENDIX C

RAW VALUES OF GEOGRAPHIC VARIABLES BEFORE ADJUSTMENT

State	Total State Population[1]	Percentage of Non-White Population[2]	Percentage of Urban Population[3]	Percentage of Democrat Votes in 2012 Election[4]	Percentage of Population with at least a Bachelor's Degree[5]	State Median Income[6]
AK	666,489	30.20	66.02	40.81	23.8	$40,554
AL	4,532,319	32.70	59.04	38.36	31.5	$84,035
AR	2,779,891	16.00	56.16	36.88	26.6	$47,085
AZ	5,746,102	20.10	89.81	44.59	32.0	$38,134
CA	35,874,307	26.50	94.95	60.24	33.1	$90,967
CO	4,758,713	12.00	86.15	51.49	27.4	$53,514
CT	3,475,008	18.40	87.99	58.06	25.1	$81,333
DC	566,957	28.90	100.00	90.91	29.6	$85,591
DE	831,181	56.60	83.30	58.61	28.5	$52,746
FL	17,463,048	21.90	91.16	50.01	26.9	$46,253
GA	9,052,085	37.50	75.07	45.48	25.8	$46,597
HI	1,249,329	73.40	91.93	70.55	25.1	$97,317
IA	2,984,964	6.30	64.02	51.99	27.4	$41,452
ID	1,394,540	22.30	70.58	32.62	25.7	$55,062
IL	12,802,243	13.70	88.49	57.6	29.5	$45,888
IN	6,286,453	7.50	72.44	43.93	31.0	$48,730
KS	2,773,227	12.90	74.20	37.99	35.9	$47,292
KY	4,161,887	11.50	58.38	37.8	29.2	$40,267
LA	4,594,531	36.50	73.19	40.58	38.2	$42,367
MA	6,448,526	4.80	91.97	60.65	35.6	$47,448
MD	5,543,441	39.50	87.20	61.97	23.9	$87,080
ME	1,323,312	16.80	38.66	56.27	35.7	$59,365
MI	10,203,863	19.90	74.57	54.21	24.6	$47,950
MN	5,215,595	13.80	73.27	52.65	26.4	$59,948
MO	5,778,759	40.20	70.44	44.38	24.1	$36,338
MS	2,918,131	16.30	49.35	43.79	28.7	$43,424
MT	917,958	10.50	55.89	41.7	34.5	$47,804
NC	8,701,483	10.30	66.09	48.35	48.5	$48,576
ND	643,802	23.30	59.90	38.69	25.2	$43,753

State	Total State Population[1]	Percentage of Non-White Population[2]	Percentage of Urban Population[3]	Percentage of Democrat Votes in 2012 Election[4]	Percentage of Population with at least a Bachelor's Degree[5]	State Median Income[6]
NE	1,766,814	5.80	73.13	38.03	22.5	$61,369
NH	1,316,126	26.60	60.30	51.98	34.0	$95,470
NJ	8,705,645	17.10	94.68	58.38	30.6	$43,531
NM	1,915,323	29.10	77.43	52.99	22.7	$53,914
NV	2,366,908	28.30	94.20	52.36	27.9	$44,670
NY	19,316,116	10.40	87.87	63.35	25.3	$49,907
OH	11,473,289	16.80	77.92	50.67	32.4	$45,114
OK	3,570,756	24.60	66.24	33.23	30.5	$40,926
OR	3,581,202	11.90	81.03	54.24	26.5	$49,136
PA	12,439,617	16.80	78.66	51.97	25.6	$49,889
RI	1,070,807	14.40	90.73	62.7	27.5	$54,124
SC	4,226,172	31.70	66.33	44.09	21.8	$43,329
SD	784,601	14.10	56.65	39.87	24.3	$47,451
TN	5,928,850	20.90	66.39	39.08	23.0	$41,567
TX	22,859,965	19.70	84.70	41.38	25.3	$47,548
UT	2,413,090	8.40	90.58	24.75	18.9	$55,109
VA	7,510,923	4.80	75.45	51.16	21.4	$51,731
VT	630,054	29.20	38.90	66.57	22.0	$59,562
WA	6,185,300	18.80	84.05	56.16	21.0	$50,082
WI	5,553,428	6.20	70.15	52.83	29.9	$37,060
WV	1,823,000	11.90	48.72	35.54	19.6	$50,578
WY	502,223	7.30	64.76	27.82	25.5	$55,212

[1] Total Population by State, U.S. Bureau of the Census, 2014
[2] Non-White Population Percentage by State, U.S. Bureau of the Census, 2013
[3] Urban Percentage of the Population for States, U.S. Bureau of the Census, 2014
[4] U.S. Presidential Election Results, 2012
[5] Educational Attainment by State, U.S. Bureau of the Census, 2009.
[6] Median Household Income (In 2009 Inflation-Adjusted Dollars), U.S. Bureau of the Census, 2009

REFERENCES CITED

Ashktorab, Z., Brown, C., Nandi, M., & Culotta, A. (2014). Tweedr: Mining twitter to inform disaster response. *Proceedings of the 11th International ISCRAM Conference.* University Park, PA.

Barberá, P., & Rivero, G. (2014). Understanding the Political Representativeness of Twitter Users. *Social Science Computer Review*, 1-18.

Bollen, J., & Mao, H. (2011). Twitter Mood as a Stock Market Predictor. *Computer, 44*(10), 91-94.

Bonilla, Y., & Rosa, J. (2015). #Ferguson: Digital protest, hashtag ethnography, and the racial politics of social media in the United States: #Ferguson. *American Ethnologist, 42*(1), 4-17.

Ceron, A., Curini, L., Iacus, S. M., & Porro, G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society, 16*(2), 340-358.

Chunara, R., Andrews, J. R., & Brownstein, J. S. (2012). Social and News Media Enable Estimation of Epidemiological Patterns Early in the 2010 Haitian Cholera Outbreak. *American Journal of Tropical Medicine and Hygiene, 86*(1), 39-45.

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), 130-39.

De Longueville, B., Smith, R. S., & Luraschi, G. (2009). OMG, From Here, I Can See the Flames!: A Use Case of Mining Location Based Social Networks to Acquire Spatio-temporal Data on Forest Fires. *Proceedings of the 2009 International Workshop on Location Based Social Networks* (pp. 73-80). Seattle: ACM.

FiveThirtyEight's Pollster Ratings . (2014, September 25). Retrieved May 1, 2016, from http://fivethirtyeight.com/interactives/pollster-ratings/

Frank, M. R., Mitchell, L., Dodds, P. S., & Danforth, C. M. (2013). Happiness and the Patterns of Life: A Study of Geolocated Tweets. *Scientific Reports, 3*(2625), 1-9.

Gayo-Avello, D. (2012). "I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper"–A Balanced Survey on Election Prediction using Twitter Data. *arXiv preprint arXiv:1204.6441*.

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

Graham, M., Stephens, M., & Hale, S. (2013). Featured graphic: Mapping the geoweb: a geography of Twitter. *Environment and Planning A, 45*(1), 100-102.

Gruzd, A., Doiron, S., & Mai, P. (2011). Is happiness contagious online? A case of Twitter and the 2010 Winter Olympics. *System Sciences (HICSS), 2011 44th Hawaii International Conference on System Sciences* (pp. 1-9). Hawaii: IEEE.

Hong, S., & Nadler, D. (2011). Does the Early Bird Move the Polls? The Use of the Social Media Tool 'Twitter' by U.S. Politicians and its Impact on Public Opinion., (pp. 182-186). College Park, MD.

Hyndman, R., & Koehler, A. (2005). *Another look at measures of forecast accuracy.*

Koto, F., & Adriani, M. (2015). A Comparative Study on Twitter Sentiment Analysis: Which Features are Good? *The 20th International Conference on Applications of Natural Language To Information Systems* (pp. 453-457). Passau, Germany: Springer.

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).

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.

Longley, P. A., Adnan, M., & Lansley, G. (2015). The geo-temporal demographics of Twitter usage. *Environment and Planning A, 47*(2), 465-484.

Mai, E., & Hranac, R. (2013). incidents, Twitter interactions as a data source for transportation. *Proceedings of Transportation Research Board 92nd Annual Meeting.* Washington, DC.

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. N. (2012). Understanding the Demographics of Twitter Users. *Association for the Advancement of Artificial Intelligence (AAAI).* ICWSM.

Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., & Danforth, C. M. (2013). The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place. *PLoS ONE, 8*(5).

Moorhead, S. A., Hazlett, D. E., Harrison, L., Carroll, J. K., Irwin, A., & Hoving, C. (2013). A New Dimension of Health Care: Systematic Review of the Uses,

Benefits, and Limitations of Social Media for Health Communication. *Journal of Medical Internet Research, 15*(4).

Morstatter, F., Pfeffer, J., & Huan, L. (2014). When is it Biased? Assessing the Representativeness of Twitter's Streaming API. *arXiv:1401.7909v1*.

Nasukawa, T., & Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. *Proceedings of the 2nd International Conference on Knowledge Capture* (pp. 70-77). Sanibel Island, FL: ACM.

Nielsen, F. Å. (2011). A new ANEW: evaluation of a word list for sentiment analysis in microblogs. *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages.*, (pp. 93-98).

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *Proceedings of the International AAAI Conference on Weblogs and Social Media* (pp. 122-129). Washington, DC: ICWSM.

Palen, L., Anderson, K. M., Mark, G., & Martin, J. (2010). A vision for technology-mediated support for public participation & assistance in mass emergencies & disasters. *Proceedings of the 2010 ACM-BCS visions of computer science conference.* Edinburgh, Scotland.

Pew Research Center: Sampling. (n.d.). Retrieved May 1, 2016, from http://www.pewresearch.org/methodology/u-s-survey-research/sampling/

Pozdnoukhov, A., & Kaiser, C. (2011). Space-time dynamics of topics in streaming text. *roceedings of the 23rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks* (pp. 1-8). New York, NY: ACM.

Quercia, D., Ellis, J., Capra, L., & Crowcroft, J. (2012). Tracking gross community happiness from tweets. *Proceedings of the ACM 2012 conference on computer supported cooperative work* (pp. 965-968). Seattle: ACM.

(2006). *Regular Voters, Intermittent Voters, and Those Who Don't: Who Votes, Who Doesn't and Why.* Washington, D.C.: The Pew Research Center.

Salathé, M., & Khandelwal, S. (2011). Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control. *PLoS Computational Biology, 7*(10), 1-7.

Sasaki, K., Nagano, S., Ueno, K., & Cho, K. (2012). *Feasibility Study on Detection of Transportation Information Exploiting Twitter as a Sensor.* Association for the Advancement of Artificial Intelligence.

Scanfeld, D., Scanfeld, V., & Larson, E. L. (2010). Dissemination of health information through social networks: Twitter and antibiotics. *American Journal of Infection Control, 38*(3), 182-188.

Schmidt, C. W. (2012). Using social media to predict and track disease outbreaks. *Environmental Health Perspectives, 120*(1), A30-A33.

Shi, L., Agarwal, N., Agrawal, A., Garg, R., & Spoelstra, J. (2012). Predicting US primary elections with Twitter. *Social*.

Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P., & Rana, O. (2013). Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter. *Sociological Research Online, 18*(3), 1-11.

Starbird, K., & Palen, L. (2010). Pass it on?: Retweeting in mass emergency. *Proceedings of the 7th International ISCRAM Conference* (pp. 1-10). Seattle, WA:

International Community on Information Systems for Crisis Response and Management.

Tremayne, M. (2014). Anatomy of Protest in the Digital Era: A Network Analysis of Twitter and Occupy Wall Street. *Social Movement Studies, 13*(1), 110-126.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (pp. 178-185). ICWSM.

Weisse, M., Omri, M., White, G., Roth, R., & Naughton-Treves, L. (2015). Tambopata Transformed: Using Web Mapping to Enhance a Geography Course Exercise About Forest Conservation. *The Journal of Maps, 11*(3), 525-533.

Weslowski, P. (2014). Using Forums and Message Boards to Recruit Study Participants in Qualitative Research. *The Qualitative Report, 19*(39), 1-15.

Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear". *Procedia - Social and Behavioral Sciences, 26*, 55-62.

Zhang, Y. (2008). Blogs as a new form of public participation in Mainland China. *Intercultural Communication Studies, 17*(3), 44-56.