#INDYREF: A GEOGRAPHICAL COMPARISON OF TWEETS RELATED TO THE SCOTTISH INDEPENDENCE REFERENDUM OF 2014

by

Tom Holby
A Thesis
Submitted to the Graduate Faculty of George Mason University in Partial Fulfillment of The Requirements for the Degree of Master of Science Geography and GeoInformation Science

Committee:

Dr. Anthony Stefanidis, Thesis Director
Dr. Arie Croitoru, Committee Member
Dr. Andrew Crooks, Committee Member
Dr. Richard Medina, Committee Member
Dr. Anthony Stefanidis, Department Chair
Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science
Dr. Peggy Agouris, Dean, College of Science

Date: Dec 9, 2014

Fall Semester 2014 George Mason University Fairfax, VA
#IndyRef: A Geographical Comparison of Tweets Related to the Scottish Independence Referendum of 2014

A Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at George Mason University

by

Tom Holby
Bachelor of Arts
Flagler College, 2002

Director: Anthony Stefanidis, Professor, Chair
Department of Geography and GeoInformation Science

Fall Semester 2014
George Mason University
Fairfax, VA
This work is licensed under a creative commons attribution-noderivs 3.0 unported license.
DEDICATION

This is dedicated to my loving family, thanks for all the support along the way.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Tables</td>
<td>vi</td>
</tr>
<tr>
<td>List of Figures</td>
<td>vii</td>
</tr>
<tr>
<td>List of Abbreviations</td>
<td>ix</td>
</tr>
<tr>
<td>Abstract</td>
<td>x</td>
</tr>
<tr>
<td>Chapter One</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Introduction</td>
<td></td>
</tr>
<tr>
<td>Chapter Two</td>
<td>3</td>
</tr>
<tr>
<td>2.1 Background</td>
<td>3</td>
</tr>
<tr>
<td>2.1.1 History and Geography</td>
<td>3</td>
</tr>
<tr>
<td>2.1.2 Growth of the Scottish National Movement</td>
<td>5</td>
</tr>
<tr>
<td>2.1.3 The Scottish Independence Referendum of 2014</td>
<td>6</td>
</tr>
<tr>
<td>2.1.4 Yes Scotland and Better Together</td>
<td>7</td>
</tr>
<tr>
<td>2.2 Appropriateness of the event for this research</td>
<td>9</td>
</tr>
<tr>
<td>2.3 Related Literature</td>
<td>11</td>
</tr>
<tr>
<td>2.3.1 Social Media</td>
<td>11</td>
</tr>
<tr>
<td>2.3.2 Social Media and Elections in Scotland</td>
<td>13</td>
</tr>
<tr>
<td>2.3.3 Volunteered Geographic Information</td>
<td>13</td>
</tr>
<tr>
<td>2.3.4 Ambient Geographic Information</td>
<td>15</td>
</tr>
<tr>
<td>Chapter Three</td>
<td>18</td>
</tr>
<tr>
<td>3.1 Objective and Hypothesis</td>
<td>18</td>
</tr>
<tr>
<td>Chapter Four</td>
<td>19</td>
</tr>
<tr>
<td>4.1 Data and Methods</td>
<td>19</td>
</tr>
<tr>
<td>4.1.1 Data Preparation</td>
<td>19</td>
</tr>
<tr>
<td>4.1.2 Selected Campaign Issues</td>
<td>23</td>
</tr>
<tr>
<td>4.1.3 Demographic Information</td>
<td>25</td>
</tr>
<tr>
<td>4.1.4 Methods</td>
<td>30</td>
</tr>
</tbody>
</table>
Chapter Five ...................................................................................................................................... 32
  5.1 Analysis and results .................................................................................................................. 32
    5.1.1 Number of tweets ............................................................................................................ 32
    5.1.2 Demographic comparisons .............................................................................................. 35
    5.1.3 Standard deviational ellipses ........................................................................................ 40
    5.1.4 Referendum Results ...................................................................................................... 56
Chapter Six ...................................................................................................................................... 65
  6.1 Conclusion and Outlook ....................................................................................................... 65
References ..................................................................................................................................... 69
LIST OF TABLES

Table | Page
--- | ---
Table 1: Number of tweets and users from 12 - 18 September | 21
Table 2: Tweets collected per council area and referendum result percentages per council area | 22
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Map of Scotland including major cities</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>Sample Ballot Form (Electoral Commission, 2014)</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>Population of Scotland per square kilometer by council area</td>
<td>27</td>
</tr>
<tr>
<td>4</td>
<td>Percent of pensionable age people per council area of Scotland, shown as a percentage of council area population</td>
<td>28</td>
</tr>
<tr>
<td>5</td>
<td>Percent of working age people per council area of Scotland, shown as a percentage of council area population</td>
<td>29</td>
</tr>
<tr>
<td>6</td>
<td>Number of tweets per day</td>
<td>33</td>
</tr>
<tr>
<td>7</td>
<td>Comparison of council area population and tweets</td>
<td>36</td>
</tr>
<tr>
<td>8</td>
<td>Tweet from a 72 year old user</td>
<td>38</td>
</tr>
<tr>
<td>9</td>
<td>Prime Twitter users in Scotland per council area as a percentage of population. Prime Twitter users are defined as people between the ages of 15 and 34 for this thesis.</td>
<td>39</td>
</tr>
<tr>
<td>10</td>
<td>Distribution of #VoteYes tweets in Scotland for 12 – 18 September, showing standard deviational ellipse.</td>
<td>41</td>
</tr>
<tr>
<td>11</td>
<td>Distribution of #VoteNo tweets in Scotland for 12 - 18 September, showing standard deviational ellipse.</td>
<td>42</td>
</tr>
<tr>
<td>12</td>
<td>Distribution of #VoteYes and #VoteNo tweets for Scotland for 12 - 18 September, showing standard deviational ellipses.</td>
<td>43</td>
</tr>
<tr>
<td>13</td>
<td>Distribution of #VoteYes and #VoteNo tweets in Scotland for 12 – 18 September excluding Glasgow and Edinburgh, showing standard deviational ellipses.</td>
<td>45</td>
</tr>
<tr>
<td>14</td>
<td>Breakdown of Northern Area, Population Belt, and Southern Area used for analysis in this thesis.</td>
<td>47</td>
</tr>
<tr>
<td>15</td>
<td>Distribution of #VoteYes and #VoteNo tweets in the Population Belt of Scotland for 12 - 18 September, showing standard deviational ellipses.</td>
<td>49</td>
</tr>
<tr>
<td>16</td>
<td>Distribution of #VoteYes and #VoteNo tweets north of the Population Belt in Scotland for 12 -18 September, showing standard deviational ellipses.</td>
<td>51</td>
</tr>
<tr>
<td>17</td>
<td>Distribution of #VoteYes and #VoteNo tweets south of the Population Belt of Scotland for 12 - 18 September, showing standard deviational ellipses.</td>
<td>53</td>
</tr>
<tr>
<td>18</td>
<td>Distribution of #VoteYes and #VoteNo tweets outside of the Population Belt of Scotland for 12 - 18 September, showing standard deviational ellipses.</td>
<td>55</td>
</tr>
<tr>
<td>19</td>
<td>Percentage of #VoteYes, #VoteNo, and #IndyRef tweets in the sub-national level dataset.</td>
<td>57</td>
</tr>
<tr>
<td>20</td>
<td>Political control of Scotland showing largest party by council seats.</td>
<td>58</td>
</tr>
<tr>
<td>21</td>
<td>Scottish independence referendum, results by council area.</td>
<td>60</td>
</tr>
</tbody>
</table>
Figure 22: #VoteYes and #VoteNo tweets in council areas of Scotland for 12 - 18 September .................................................................62
Figure 23: Normalized comparison of #VoteYes and #VoteNo tweets by council area of Scotland for 12 - 18 September, #VoteYes tweets normalized by over-representative magnitude .............................................................64
# LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGI</td>
<td>Ambient Geographic Information</td>
</tr>
<tr>
<td>MAUP</td>
<td>Modifiable Areal Unit Problem</td>
</tr>
<tr>
<td>SNP</td>
<td>Scottish National Party</td>
</tr>
<tr>
<td>SDE</td>
<td>Standard Deviational Ellipse</td>
</tr>
<tr>
<td>UK</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>VGI</td>
<td>Volunteered Geographic Information</td>
</tr>
</tbody>
</table>
#INDYREF: A GEOGRAPHICAL COMPARISON OF TWEETS RELATED TO THE SCOTTISH INDEPENDENCE REFERENDUM OF 2014

Tom Holby, BA
George Mason University, 2014
Thesis Director: Dr. Anthony Stefanidis

This thesis examines the 2014 Scottish independence referendum through the lens of social media. In September of 2014, a vote was held in Scotland to determine whether or not it should sever ties with the United Kingdom and become an independent country. During the lead up to this referendum, people took to Twitter and other forms of social media to discuss the upcoming vote and issues surrounding the event. Using locational information in Twitter feeds, a comparison is made between the digital discussion surrounding the vote and the physical happenings. Data for the week leading up to the referendum was collected and analyzed for geographic differences between the two main campaigns.
CHAPTER ONE

1.1 Introduction

On September 18, 2014, people from all over Scotland went to the polls to cast a vote on the future of their country’s independence. This Scottish independence referendum has been dubbed a once in generation chance for independence (Lockhart, 2014). Scotland is a part of the United Kingdom, and this vote was held to determine whether Scotland would continue to be a part of the UK, or separate off into an independent nation. In the end, the votes were counted and the people of Scotland ultimately decided to remain with the UK.

The rise in social media has provided a new way to examine the conversations surrounding events around the world. Vast amounts of user-created data are constantly being produced. Twitter reports 284 million active users generating 500 million tweets each day (Twitter, 2014). Many researchers have taken note of the potential value of data derived from social media. The speed at which this data is available, combined with the large number of data points is enticing to researchers, especially social scientists. Another factor that adds to the allure of social media data stems from its geographic diversity. It is comparatively easy to gather social media data about a far away place that is not easily accessible otherwise. However, when gathered from afar, important trends in the information may not be apparent to researchers. Also, the idea of gathering distant data
does not need to be a function of geographic distance. Researchers may find it easier to collect data on issues that are more locally focused, but with groups with whom they do not have regular interaction. With this ability to research from afar, attention should be paid to how inherent tendencies of the data may influence the results.

This thesis will address the question of whether or not social media data can be used effectively for research surrounding a major event. We will compare social media data about the Scottish independence referendum to demographic data and other traditional data sources. The comparison will shed light on how real life events play out in the digital sphere. Data were collected in the lead up to the referendum to see how the Twitter representation of the data compared to the situation on the ground. This has important applications due to the increase in studies utilizing social media as a data source. Comparing these results will allow future researchers to have a more complete picture of what biases may be present in social media data. Specifically, issues regarding the age of a population and differences in geographic distribution of data classes will enable scholars to have a better understanding of some of the challenges associated with social media data.

This thesis will begin by providing background information on the referendum, and examining literature related to social media, volunteered geographic information, and ambient geographic information. Following this, specifics about the data will be discussed as well as the methods used for analysis. Chapter five contains the results of the analysis, and we will conclude with an eye toward the future outlook and commentary.
CHAPTER TWO

This chapter presents background information related to the Scottish independence referendum, as well as discussion on why it was an appropriate event for this thesis. Previous literature related to social media, volunteered geographic information, and ambient geographic information is also discussed.

2.1 Background

2.1.1 History and Geography

The Scottish Independence Referendum, held on September 18, 2014, was a vote to decide on whether Scotland should become a country independent from the United Kingdom. Historically, the Kingdoms of Scotland and England joined to form the United Kingdom of Great Britain with the Treaty of Union in 1707, where the parliament of Scotland was merged with the English parliament. Now known as the United Kingdom of Great Britain and Northern Ireland, it is composed of England, Scotland, Wales, and Northern Ireland as well as dependent areas around the globe such as Bermuda, the British Indian Ocean Territory, the Cayman Islands and others (Central Intelligene Agency, 2014). The capital of Scotland is Edinburgh, which is also the second largest city in the country with 486,000 people. 75 kilometers away is Glasgow, the largest city in Scotland, with almost 600,000 people. These and other major population centers in the country can be seen in Figure 1, including Aberdeen, Dundee, Inverness, and Stirling.
Figure 1: Map of Scotland including major cities
2.1.2 Growth of the Scottish National Movement

Throughout the more than 300 years of union, the central UK government has applied policies that allowed the identity of the Scottish people to flourish, without attempts to force conformity of customs and culture (Dardanelli & Mitchell, 2014). This allowed people in Scotland to self-identify as Scottish, British, or both. As of the 2011 census, 62% of the population identified as ‘Scottish only’, 8% identified as ‘British identity only’ and 18% responded with ‘Scottish and British identities only’ (“Scotland Census 3A,” 2013). There was and remains a uniqueness to Scottish culture that is visible to those who are from there, as well as people outside of the area. Whether someone lived in Scotland, another part of the UK, or elsewhere around the world, they could say they were Scottish and people recognized their ties to Scotland either personally or ancestrally. Recently, these ties to the Scottish homeland by its diaspora were examined. Researchers looked at how the reestablishment of Scottish Parliament in 1999 impacted the views of national identity by Scottish people outside of Scotland (Sim, 2012). This continual foundational Scottish cultural collective allowed for notions of independence to form over the years. In 1934 the Scottish National Party (SNP) was formed but remained a peripheral political player until 1974 when it won 11 seats in UK Parliament (Dardanelli & Mitchell, 2014). In 1997, Scots voted on a referendum in favor of the creation of their own parliament, to better address the specific needs of Scotland within the political structure of the United Kingdom. The first elections to the new Scottish Parliament were held in 1999.
The SNP competed with other political parties for control of Scottish Parliament, and in 2007, edged out Scottish Labour, and became the largest party, though still without an overall majority (The Scottish Parliament, 2007). Four years later, to the surprise of many, the SNP won an outright majority in Scottish Parliament, the first since the creation of the institution (Baxter & Marcella, 2013; Dardanelli & Mitchell, 2014).

2.1.3 The Scottish Independence Referendum of 2014

The concept of a referendum on independence was a standing component of SNP politics and after capturing an outright majority, the referendum was all but inevitable. The UK government did not oppose this move and the referendum was set for September 18, 2014. With the referendum on the way, decisions were made regarding who would be allowed to vote. British or other Commonwealth (UK) citizens living in Scotland, along with European Union Citizens living in Scotland would be allowed to participate if they were 16 or older on 18 September 2014 (The Electoral Commission, 2014). Those living outside Scotland, including citizens living in the rest of the UK and people around the world who considered themselves Scottish, did not have a say in the decision. The age of eligibility for voters was lowered from 18 to 16 for this referendum. Researchers examined the perceptions of the referendum by 14 – 17 year olds in Scotland and found that from 2013 to 2014 there was an increase in the percentage of eligible voters in this age range who supported independence for Scotland (Eichorn, Paterson, MacInnes, & Rosie, 2014). When asked about their national identity, most of them reported an identity that was Scottish to some degree and a large proportion identified as a mixture of British and Scottish (Eichorn et al., 2014).
National Identity, economics, natural resources, defense, and international relations were all prominent issues during the campaigns. In the current arrangement, some issues such as defense and international relations are handled at the UK level, while other aspects are run by Scotland itself through the Scottish Parliament. Some of these include education, transport and environmental issues (Werth, 2014). If the independence movement won, this link would have dissolved, and all of these matters would have become the responsibility of Scotland.

2.1.4 Yes Scotland and Better Together

When the notion of a referendum started taking hold, two main camps quickly formed. In May of 2012, the independence campaign officially commenced through the formation of the Yes Scotland organization (“Yes campaign seeks million Scots,” 2012). At the forefront of this effort was Alex Salmond, the head of the SNP and Scotland’s First Minister. A month later, in June of 2012, the Better Together campaign launched to encourage people to vote against independence from the UK. Lead by Alistair Darling, a former UK chancellor, the Labour, Liberal Democrat, and Conservative parties backed Better Together (“Independence is ‘one-way ticket,’” 2012).

The core of the Yes Scotland campaign revolved around the notion that Independence for Scotland would allow the people who lived there to have more control over how they are governed. By cutting governmental ties with Westminster in the UK, independence advocates proposed that decisions made about Scotland would be made by an entirely Scottish government, and thus be more aligned with Scotland’s goals. As it is
now, in the view of the independence supporters, Scotland makes up a small part of the government of the UK and is limited in what can be accomplished through this setup.

The heart of the Better Together campaign was that the UK and Scotland were both better off continuing the union. Campaign materials expressed the uncertainty of what would happen if Scotland voted for independence and also warned that, it would be irreversible (“The Facts You Need,” n.d.). Interestingly, though Better Together was pushing for continuation of the union, they were not necessarily advocating for things to remain the same. The promise of more powers for Scotland, if it remained a part of the UK, was assured at times throughout the campaign. This was especially the case as the referendum drew near and polls showed a much closer than anticipated contest. Early on, the possibility of having a highly popular middle of the road position on the ballot was discussed, but strongly opposed by the UK. (Dardanelli & Mitchell, 2014; Renton, 2014). This would have provided the option to vote for more powers to be afforded to the Scottish Parliament, without voting for independence.

In the end it was decided to provide a ballot with a single question, the wording for which was agreed upon by both sides, and was written in easy to interpret language. The actual ballot simply stated: “Should Scotland be an independent country?” People responded by checking a box next to “Yes” or “No”. Figure 2 is a sample ballot for the Scottish independence referendum that appeared in a voting guide produced by the Electoral Commission (2014).
2.2 Appropriateness of the event for this research

Twitter data has been used to analyze many issues across a variety of subjects. Many times, an event occurred and the Twitter data was examined as a response. This has been applied in studies ranging from earthquakes (Acar & Muraki, 2011; Crooks, Croitoru, Stefanidis, & Radzikowski, 2013) to sports games (Zhao, Zhong, Wickramasuriya, & Vasudevan, 2011). A key factor in these studies was that the event was unexpected, and the response was what was analyzed using Twitter. Since the event was unknown, data was not collected before the event. Our research focused on the other side of the event timeline, the lead up to an event.
The Scottish independence referendum provided an excellent case to examine for a variety of reasons. Some of these reasons were that it was a planned event, an event with a great deal of participation, an easy to assess stance (Yes/No), in a region with quality demographic data available, and conducted mostly in English.

First and foremost, the event was planned for, and took place on a specific date. Since the event required considerable planning and preparation, and it was not just sprung on voters one afternoon, the ability to plan our collection of data around a specific date was possible. This enabled data to be collected a week prior to the event to look for trends and changes in the pre-event Twitter data. Many events examined using Twitter data are unpredictable, and it is appropriate to examine the response. We chose to focus on the lead up to the event, to help facilitate the comparison of differences between pre and post event profiles within Twitter data.

The referendum also had an extremely high profile domestically, as well as prominence in the international arena. This was a major event at the forefront of the public conscience. Constant news coverage and preparatory events beforehand ensured that the people of Scotland were well aware of the significance of the upcoming referendum. These events included debates, news conferences, publication of positions, rallies, as well as local, national, and international news coverage of these actions. The significance was also seen during the actual vote through the record participation rates.

Another positive factor associated with the choice of this event was the amount of quality demographic data in the study area. Twitter studies are done throughout the world, and in some areas, obtaining high-quality demographic data is a major hurdle, if
not impossible. Due to the demographics of Twitter users, it was useful to be able to access this type of data, age in particular, for the region being studied. Scotland provides this data through the National Records of Scotland, and previously, the General Register Office for Scotland. Additionally, age had added interest, because as mentioned in section 2.1.3, the eligibility age to vote was lowered from 18 to 16 for this referendum.

The binary nature of this referendum (Voters could either vote “Yes” or vote “No”) added to the suitability of this event for analysis. While any event has more than one position, and even binary choices have subtleties and nuances to account for, the stark differences between a “Yes” and “No” present well within Twitter data.

Finally, the event occurred in a location where a single language is spoken by a majority of the people. In Scotland 93% of people over three years old speak only English and only 1.4% of people are not able to speak English (Scotland Census 2A, 2013) Other reported languages are Gaelic 1.1%, Polish 1%, and British Sign Language 0.2% (Scotland Census 2A, 2013). Having one dominant language simplified analysis and eased concerns associated with language bias in the data. This along with the other factors mentioned above contributed to the suitability of this referendum as an event to examine using Twitter data.

2.3 Related Literature

2.3.1 Social Media

Social media websites and applications such as Facebook, Twitter, Flickr, Youtube and others allow for the interaction of real people in virtual space. Though individual platforms have different methods, for example microblogging provides a
different type of interaction than can occur within a virtual gaming world, they all share
the overarching feature of personal interaction. The popularity of these services,
combined with the ability to access current and historical data from some of these sites,
has enabled researchers to use social media data for a wide variety of subjects. Looking at
some examples, researchers have used social media data to examine the connections of
virtual communities that transcend traditional geopolitical borders (Stefanidis et al.,
2013). Another area where social media data is being applied by researchers is in the field
of disaster response, an example of which can be seen in the use of geotagged photos,
tweets and other data to assess transportation infrastructure after Hurricane Sandy
(Schnebele, Cervone, & Waters, 2014).

Data from Twitter was used in this thesis, so a limited explanation of Twitter is
warranted. Twitter is a popular microblogging service that allows users to create
messages of 140 characters or less, known as tweets, that contain text, videos, photos,
and links. Tweets are linked to the user that shared it, and can be viewed by others who
‘follow’ that user, and the public in general. Private messages are possible with Twitter,
but that is out of the scope of this thesis, as they are not collected, or even available to be
collected for research. Tweets can contain hashtags, to signify the topic of the message
(Twitter 2014, n.d.). These are represented by the character ‘#’ followed by text. This
method allows people to explore tweets related to topics of interest. Another commonly
used element of Twitter relevant to this thesis is a retweet. This is when someone reposts
a tweet from another user. This thesis will examine topics using hashtags as well as look
at the number of tweets versus retweets in the data.
2.3.2 Social Media and Elections in Scotland

Social media use in elections related to Scotland has been examined before (Baxter & Marcella, 2012, 2013). These studies, which focused on the 2010 UK General Election and 2011 Scottish Parliamentary Election respectively, looked at social media use by the political parties themselves, and some of their members. The general results of these studies were that, at that time, candidates wanted to be seen as using social media, but were wary of using it for much more than one-way communications, as an extension of their normal media usage.

2.3.3 Volunteered Geographic Information

In 2007, Goodchild coined the term Volunteered Geographic Information, or VGI, to describe user-generated web content created for geographic purposes (Goodchild, 2007). This data differed from past geographic sources in that it was created and edited by users instead of larger, geographically centered organizations such as mapping agencies. Sites such as Open Street Map and Wikimapia allow users to contribute geographic information through a web or mobile device. This information is viewable by others who can use it or edit it, with the overall goal of improving accuracy over time through crowd sourcing.

VGI has been used for a variety of tasks, one of the most prominent of which occurred after the earthquake in Haiti on 12 January 2012. Zook et al. describe how VGI was used by people around the world to rapidly assist in creating accurate reference and damage maps to aid disaster relief operations following the magnitude 7 earthquake that decimated Port-au-Prince (Zook, Graham, Shelton, & Gorman, 2010). In another
example, maps of wildfires in California were updated by volunteers who integrated data from Twitter, television news, and official reports with satellite data and online map platforms, enabling residents to make more informed decisions about evacuation (Goodchild & Glennon, 2010). A major benefit of VGI is the speed at which it can be created and disseminated, which is why much of the focus involves sudden, time critical events. However, research on VGI is not limited to disaster response. An example of VGI being used for environmental research is the RinkWatch project. Scientists at Wilfrid Laurier University in Canada created a website were people input the conditions of their local ice rinks, and the data is used in climate change research as well as for providing a map of ice skating rink conditions ("RinkWatch," n.d.). In both rapid response and long-term projects, the added value of VGI is being implemented. In these examples, as in all cases of VGI, data is provided by a large group of people, not a traditional authoritative source, which raises the questions of data quality.

Volunteered geographic information, like any form of geographic information, is subject to data quality issues; though some issues apply specifically to VGI. Researchers have noted that VGI utilizes different methods of quality control than standard geographic information, relying on a larger number of non-expert peers to review data in VGI contrasted to fewer reviews by authoritative experts in traditional geographic settings (Elwood, Goodchild, & Sui, 2012). The quality of VGI in relation to standard methods was analyzed using data from Open Street Map in England and comparing it to Ordnance Survey datasets showing that VGI can achieve high data quality, though it is not necessarily as complete as traditional geographic sources, and the quality is not
consistent in all places (Haklay, 2010). Looking specifically at the accuracy and completeness of VGI point features, researchers found VGI to be effective in capturing many of the features in a traditional database (Jackson et al., 2013). In that particular study, it was shown that when VGI is used in conjunction with traditional methods, the results are even more favorable. The quality of data accuracy for VGI is a continuing research area, and the many advantages VGI offers specifically timeliness, contribute to its extensive adoption for research. However; due to the general lack of metadata and standard methods for expressing accuracy, though potentially accurate, the quality of specific VGI data is generally unknown to users (Elwood et al., 2012).

2.3.4 Ambient Geographic Information

Though Open Street map and Wikimapia have an overtly geographic focus, creating a more accurate map, other services contain geographic information, without it being the main goal. In fact, many forms of social media have messages where geographic content is attached, but this information is not the main intent of the message. Researchers have termed this evolution of VGI Ambient Geographic Information, or AGI (Stefanidis, Crooks, & Radzikowski, 2013). Though similar, AGI and VGI diverge in the area of intended use. Whereas the goal within VGI is explicitly geographic in nature, locating specific content in a specific place, in AGI the locational aspects are secondary, an afterthought, or not even considered in relation to the content. An example of this can be seen in the comparison of the use of AGI and VGI in response to an earthquake in the eastern United States (Crooks et al., 2013). In this study the authors examined information harvested through feeds related to the earthquake as well as information from
‘Did You Feel It?’ an official United States Geological Survey VGI effort. They concluded that both forms of information are useful in the study of events that impact large numbers of people.

As in VGI, one of the most appealing factors of AGI is the speed at which information becomes available. Immediately after an event occurs, people respond to that event through social media. Facebook statuses are updated, pictures are taken, tagged and shared, and messages are tweeted and retweeted. With geographic data embedded in many of these messages, a picture can emerge about the conversation taking place, and where it is occurring. The rapidness of this conversation can be seen the example from Crooks et al. (2013) related to the earthquake in the eastern United States. Immediately after this event, tweets with the hashtag #earthquake were sent by people who experienced the quake. These tweets were mapped in time intervals to show the rapid geographic expansion of the discussion on Twitter.

Another feature of AGI that interests researchers is the ability to collect data in areas or groups of people that would not be accessible in other circumstances. Distant locations can be targeted for data collection that would be difficult, dangerous, or cost prohibitive to reach. For example, Stefanidis et al. used AGI to look at data from Tahrir Square in Cairo during the Arab Spring in 2011 (Stefanidis et al., 2013). In another example of this, Crooks et al. examined the Syrian civil war using AGI and found social media being used to empower citizens by becoming an avenue of global expression that can counter state-driven messages (Crooks et al., 2014). Along with physically difficult places to reach, AGI has been used to examine groups of people and topics that are
socially inaccessible. This can be seen in the examination of the impact of events attributed to the terrorist groups Boko Haram and Al-Shabaab using AGI (Rodriguez, 2014). The ability of AGI to be used across broad geographic and social distances, combined with potential for real time or near real time analysis are important reasons behind researchers use of AGI.

Data collected for this thesis falls under the category of AGI, as the locational information provided within the Twitter data was not the focus of the message, but a secondary aspect. Residents across Scotland took to Twitter and other forms of social media to discuss the referendum and the issues, and geographic information was present in much of this content. This research will delve into some of this geographic social media data and explore the conversation surrounding the Scottish independence referendum through the use of AGI.
CHAPTER THREE

3.1 Objective and Hypothesis

The objective of this thesis is to apply geospatial analysis to geographically referenced Twitter data. By doing this we will be able explore distinctions between the geographic dispersion of varying data classes. This will be combined with demographic data and other techniques to investigate differences within social media data. The idea behind this is to offer insight into potential biases within social media data, with the overarching goal of improving social media analysis.

Our hypothesis is that differences will be present when examining different groups of users. Previous social media research described a vocal minority/silent majority dynamic present in Twitter data relating to contested elections in the United States (Mustafaraj, Finn, Whitlock, & Metaxas, 2011). We intend to explore whether or not these differences are present when examining the VoteYes and VoteNo tweets from the Scottish independence referendum. We will also examine the geographic dispersions of the VoteYes and VoteNo tweets, with the expectation that the distributions will differ. To analyze the geographic dispersions, we plan on comparing standard deviational ellipses, which will be discussed more in the following chapter, for both the VoteYes tweets and the VoteNo tweets.
CHAPTER FOUR

4.1 Data and Methods

4.1.1 Data Preparation

Before analyzing the tweets, some refinement of the dataset was necessary. Initial refinements focused on the geographic area, geographic scale, and temporal aspects of the individual tweets.

The harvesting and ingesting of the social media data was completed by the George Mason University Center for Geospatial Intelligence. The collection was done using a purpose built system for analysis of data feeds from social media sites such as Twitter, Flickr, and YouTube which is discussed in an article by Croitoru et al. (Croitoru, Crooks, Radzikowski, & Stefanidis, 2013). The system connects to the Application Programming Interface of Twitter (as well as other social media services such as Flickr and YouTube, however this thesis only uses data from Twitter) and is able to submit parameters for collection. The results are geocoded and processed for storage in a dedicated database. The resulting data was provided to the author for analysis as a tab-separated values (.tsv) file.

Tweets were collected during the two and a half weeks surrounding the Scottish Independence Referendum in September 2014. The hashtags #indyref, #scotlanddecides, #VoteYes, and #VoteNo were used as a basis for collection as well as @TogetherDarling
(Twitter handle of Alistair Darling) and @AlexSalmond (Twitter handle of Alex Salmond).

The results provided to the author included 5,422,026 tweets collected worldwide, based on the above terms. This number includes tweets with and without locational information; so further filtering of the dataset was required. After removing the tweets without locational information associated, 3,154,433 individual tweets remained, again at the worldwide level. To refine this to a dataset composed of only tweets from Scotland, the data was put into a GIS for analysis using boundary files downloaded from Ordnance Survey (Ordnance Survey, 2014). Tweets were plotted as individual points and those that were located within the boundary of Scotland remained, while the others were removed. After filtering the tweets to include only those that were located within Scotland, 1,042,768 tweets remained for the dataset.

Though this dataset included only tweets geolocated within Scotland, it included tweets at all geographic levels. This presented a problem with tweets that were geolocated at the national level. Points geocoded to ‘Scotland’ were geolocated to a specific point within the Highland council area, while tweets geolocated to ‘Great Britain’ fell within Dumfries & Galloway. These tweets were removed, as the analysis for this research was done at the council area scale. The primary reason this scale was chosen was the availability of demographic and voting data at this level. Finally, with the referendum vote held on September 18th, the dataset was narrowed down further to include only those tweets that occurred within a week before and after the vote. Tweets that occurred between September 12th and September 25th were kept, while others were
removed, resulting in 749,473 individual tweets in the dataset. A majority of the analysis was done using data from the period leading up to the vote. This data, from September 12th through voting day, September 18th, was composed of 466,427 geolocated tweets from 50,802 users as can be seen in Table 1 and Table 2, which will be discussed further in Chapter 5.

Table 1: Number of tweets and users from 12 - 18 September

<table>
<thead>
<tr>
<th>12 - 18 September</th>
<th>Users</th>
<th>Tweets</th>
<th>Average Number of Tweets per user</th>
</tr>
</thead>
<tbody>
<tr>
<td>All tweets in dataset</td>
<td>50,802</td>
<td>466,427</td>
<td>9.2</td>
</tr>
<tr>
<td>#VoteYes Tweets</td>
<td>19,896</td>
<td>132,378</td>
<td>6.7</td>
</tr>
<tr>
<td>#VoteNo Tweets</td>
<td>5,918</td>
<td>17,045</td>
<td>2.9</td>
</tr>
</tbody>
</table>
Table 2: Tweets collected per council area and referendum result percentages per council area

<table>
<thead>
<tr>
<th>Council Area</th>
<th>Number of #Vote Yes Tweets</th>
<th>Number of #VoteNo Tweets</th>
<th>Actual Referendum Results Percent Yes</th>
<th>Actual Referendum Results Percent No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aberdeen City</td>
<td>3,107</td>
<td>825</td>
<td>41.4</td>
<td>58.6</td>
</tr>
<tr>
<td>Aberdeenshire</td>
<td>1,664</td>
<td>245</td>
<td>39.6</td>
<td>60.4</td>
</tr>
<tr>
<td>Angus</td>
<td>1,534</td>
<td>110</td>
<td>43.7</td>
<td>56.3</td>
</tr>
<tr>
<td>Argyll and Bute</td>
<td>2,243</td>
<td>108</td>
<td>41.5</td>
<td>58.5</td>
</tr>
<tr>
<td>City of Edinburgh</td>
<td>25,218</td>
<td>5,021</td>
<td>38.9</td>
<td>61.1</td>
</tr>
<tr>
<td>Clackmannashire</td>
<td>661</td>
<td>61</td>
<td>46.2</td>
<td>53.8</td>
</tr>
<tr>
<td>Dumfries and Galloway</td>
<td>835</td>
<td>284</td>
<td>34.3</td>
<td>65.7</td>
</tr>
<tr>
<td>Dundee City</td>
<td>2,515</td>
<td>141</td>
<td>57.3</td>
<td>42.7</td>
</tr>
<tr>
<td>East Ayrshire</td>
<td>929</td>
<td>130</td>
<td>47.2</td>
<td>52.8</td>
</tr>
<tr>
<td>East Dunbartonshire</td>
<td>477</td>
<td>84</td>
<td>38.8</td>
<td>61.2</td>
</tr>
<tr>
<td>East Lothian</td>
<td>1,316</td>
<td>118</td>
<td>38.3</td>
<td>61.7</td>
</tr>
<tr>
<td>East Renfrewshire</td>
<td>669</td>
<td>54</td>
<td>36.8</td>
<td>63.2</td>
</tr>
<tr>
<td>Falkirk</td>
<td>1,892</td>
<td>291</td>
<td>46.5</td>
<td>53.5</td>
</tr>
<tr>
<td>Fife</td>
<td>5,849</td>
<td>769</td>
<td>45.0</td>
<td>55.0</td>
</tr>
<tr>
<td>Glasgow City</td>
<td>52,003</td>
<td>5,267</td>
<td>53.5</td>
<td>46.5</td>
</tr>
<tr>
<td>Highland</td>
<td>2,633</td>
<td>177</td>
<td>47.1</td>
<td>52.9</td>
</tr>
<tr>
<td>Inverclyde</td>
<td>1,570</td>
<td>128</td>
<td>49.9</td>
<td>50.1</td>
</tr>
<tr>
<td>Midlothian</td>
<td>634</td>
<td>49</td>
<td>43.7</td>
<td>56.3</td>
</tr>
<tr>
<td>Moray</td>
<td>1,094</td>
<td>124</td>
<td>42.4</td>
<td>57.6</td>
</tr>
<tr>
<td>Na h-Eileanan Siar</td>
<td>308</td>
<td>49</td>
<td>46.6</td>
<td>53.4</td>
</tr>
<tr>
<td>North Ayrshire</td>
<td>1,623</td>
<td>139</td>
<td>49.0</td>
<td>51.0</td>
</tr>
<tr>
<td>North Lanarkshire</td>
<td>4,213</td>
<td>532</td>
<td>51.1</td>
<td>48.9</td>
</tr>
<tr>
<td>Orkney Islands</td>
<td>80</td>
<td>14</td>
<td>32.8</td>
<td>67.2</td>
</tr>
<tr>
<td>Perth and Kinross</td>
<td>5,186</td>
<td>379</td>
<td>39.8</td>
<td>60.2</td>
</tr>
<tr>
<td>Renfrewshire</td>
<td>1,850</td>
<td>242</td>
<td>47.2</td>
<td>52.8</td>
</tr>
<tr>
<td>Scottish Borders</td>
<td>1,283</td>
<td>163</td>
<td>33.4</td>
<td>66.6</td>
</tr>
<tr>
<td>Shetland Islands</td>
<td>66</td>
<td>55</td>
<td>36.3</td>
<td>63.7</td>
</tr>
<tr>
<td>South Ayrshire</td>
<td>2,149</td>
<td>235</td>
<td>42.1</td>
<td>57.9</td>
</tr>
<tr>
<td>South Lanarkshire</td>
<td>3,527</td>
<td>540</td>
<td>45.3</td>
<td>54.7</td>
</tr>
<tr>
<td>Stirling</td>
<td>2,017</td>
<td>206</td>
<td>40.2</td>
<td>59.8</td>
</tr>
<tr>
<td>West Dunbartonshire</td>
<td>887</td>
<td>66</td>
<td>54.0</td>
<td>46.0</td>
</tr>
<tr>
<td>West Lothian</td>
<td>2,346</td>
<td>439</td>
<td>44.8</td>
<td>55.2</td>
</tr>
</tbody>
</table>
It should be noted that our dataset was based on individual tweets, which were aggregated at the council area level. This represented the political conversation taking place. It was not meant to represent or replicate a poll of the voters. Studies have been done relating to the feasibility of election prediction using social media, but that is not the intent of this thesis. The literature has many articles discussing the plausibility of election prediction using Twitter, without a consensus reached on whether or not it is possible. A paper which examined the 2009 election in Germany (Tumasjan, Sprenger, Sandner, & Welpe, 2010) indicated that the potential exists for Twitter to be used predictively in this fashion. This paper was followed up by both similar research (Sang & Bos, 2012), and responses criticizing the methods used and offering counter arguments (Gayo-Avello, Metaxas, & Mustafaraj, 2011; Jungherr, Jürgens, & Schoen, 2012). Future studies will no doubt continue to address the viability of using social media for election prediction due to the tantalizing nature of the potential results. This thesis did not attempt to predict the outcome of the referendum; however, other valuable information can be gleaned when examining Twitter around a vote.

4.1.2 Selected Campaign Issues

For example, as mentioned in the background, this referendum provided many topics of debate for the citizens of Scotland. Each party discussed varying ideas on defense, monetary issues, and healthcare. To see if these topics were present in the data, a basic search for terms was conducted. Though in this somewhat simplistic form, it was only a quick representation of the degree to which topics were discussed, further research is possible using the associated locational data.
One issue that was noted by commentators was about currency. Had the referendum been victorious, each side had different views on what would happen. Those in favor of the referendum said they would continue to use the pound; however, the possibility of this was put into question by the UK government (Martin, 2014). To examine whether or not this issue was present in our data we searched for tweets containing the terms: ‘pound’, ‘currency’, and ‘banks’. The first two terms produced similar numbers of tweets, 1,031 and 1,281 respectively; however, the number of tweets containing ‘banks’ was much higher at 4,466.

Defense issues also played a role in the campaigns of both sides. Had Scotland become independent, uncertainties surrounding national security issues would have required quick decisive attention. During the campaign both sides offered different accounts of what would happen, with the independence movement stating Scotland would become a part of NATO while the anti-independence movement fueled the idea that nothing was certain if independence occurred (Martin, 2014). In our dataset, an initial search for the term ‘defense’ resulted in only 10 tweets, which was unexpected. Upon further investigation, it was recalled that the British spelling for defense contains a ‘c’ not an ‘s’. When ‘defence’ was applied to the search, 379 tweets were associated. While this mistake was somewhat humorous, it is a good example of how variations in language and terms can complicate analysis. A specific defense issue that was discussed in the campaign was the basing of the Trident nuclear program and submarines in Scotland. The VoteYes movement campaigned promising to remove nuclear weapons from Scotland, while the VoteNo campaign warned of negative consequences of that
action. Searching for specific terms related to this provides some results from our dataset. ‘Trident’ was found in 192 tweets, ‘nukes’ occurred in 184 and ‘disarm’ was found in 70 tweets.

Two other issues that appeared in the conversation captured by our dataset were related to oil and healthcare, considered two key issues in the referendum (Martin, 2014). A search for ‘oil’ returned 2,907 associated tweets while searching for ‘NHS’ resulted in 4,711 associated tweets. Though this method is not an in-depth form of investigation, it provides a starting point for future analysis, as well as a means to verify that the collected tweets are indeed related to the topics being researched.

**4.1.3 Demographic Information**

In considering the method of information (Twitter), along with the demographics of both the users and the locations, insights can be gained into how the digital discussion of the event compares to the results of the referendum. To begin this analysis, some demographic information will be presented.

Scotland is composed of 32 council areas and has a population of 5.3 million people (National Records of Scotland, 2014), that is not evenly dispersed across its area. This can be seen clearly in Figure 3. A central corridor that includes Glasgow and Edinburgh as well as the surrounding council areas is much more densely populated than the rest of the country. We will refer to this area as the population belt. Age is another factor that is not evenly distributed across Scotland. As might be expected, council areas with a high percentage of pensionable people tend to lie outside of the population belt, as
shown in Figure 4, while high percentages of working age people tend to be in council areas around major cities, shown in Figure 5.
Figure 3: Population of Scotland per square kilometer by council area
Figure 4: Percent of pensionable age people per council area of Scotland, shown as a percentage of council area population.
Figure 5: Percent of working age people per council area of Scotland, shown as a percentage of council area population.
4.1.4 Methods

Mean geographic centers and standard deviational ellipses computed using ArcGIS were used to analyze the data. The mean geographic center is similar to the mean from general statistics, but in the spatial sense, it is the center of gravity for the data points. If all of the points were weights on a piece of wood, the mean center would be the specific point used to balance the whole contraption.

Standard deviational ellipses are an extension of standard distance (Wong, 1999). The standard distance can be compared to the standard of deviation in general statistics. It is a geographic circle that measures the spatial distribution of points around the mean center (“ESRI ArcGIS Help File,” 2014). The standard deviational ellipse is similar, but accounts for variations in the general orientation of data, which is not shown when using the standard distance (Wong, 1999). Early work on the standard deviational ellipse was presented in the literature by Lefever (1926), which prompted a reply shortly thereafter by Furfey (1927) who clarified that the shape is not an actual ellipse. The interest in the actual shape has continued, with a more recent proposition for the use of standard deviation curve as a more accurate term (Gong, 2002). Depending on the software used, the standard deviational ellipse is created using a slightly different formula (Mamuse, Porwal, Kreuzer, & Beresford, 2009). Our research used ArcGIS, which applies the terms standard deviational ellipse and directional distribution for the processes (“ESRI ArcGIS Help File,” 2014). These terms will be used in this research to avoid confusion.

Standard deviational ellipses are useful for examining the spatial distribution of point data, especially if there are a large number of points (Raine, 1978). As a
representation of dispersion, this measure can be useful when comparing two or more classes of points. Levine et al. (1995), used standard deviational ellipses to examine differences in the distribution of car crashes with varying classes of injury levels in Honolulu as well as differences due to the number of involved vehicles. More recently, standard deviational ellipses were compared to standard distance measures for use in criminal geographic profiling (Kent & Leitner, 2007). An interesting conclusion of that study was that underlying landscape contributed to the dispersion of the data points, and this was better accounted for using the standard deviational ellipse. Studies using this method have also been done in the United Kingdom in general, and Scotland in particular. Marr (2012) used standard deviational ellipses to look at changes in the spatial distribution of the British Motorcycle Industry between 1896 and 2004, and earlier, Buttimer (1972) used the method in her analysis of residential areas in Glasgow. Our analysis used standard deviational ellipses to compare the distribution of Twitter data based on two classes of data, #VoteYes and #VoteNo.
CHAPTER FIVE

5.1 Analysis and results

5.1.1 Number of tweets

The dataset is composed of the geographically referenced tweets from 12 – 25 September 2014. When the data is presented in chart form, as can be seen in Figure 6, some initial observations are evident. Two of these observations that are prominent are the decline in related tweets after the vote and the overall numbers of #VoteYes and #VoteNo tweets. First, a sharp decrease in tweets with related hashtags is noticeable following the referendum for all three data classes. This can be interpreted in a variety of ways. It may show the transient tendency of Twitter users to move along to the topic of the day. With the event completed and over, there is no need to dwell on it any longer, as there are new topics about which to tweet. Another possibility is that it shows activism on Twitter. After the event occurred, there is less reason to use a hashtag expressing how the user thinks someone should vote.
A phenomenon that has been examined within social media is the vocal minority versus the silent majority (Mustafaraj, Finn, Whitlock, & Metaxas, 2011). This research discussed how during hotly contested elections in the United States, users of the minority party used social media differently than those of the majority. Members of the minority enhanced the impact of their messages through specific measures, such as more retweets, outside links, and hashtags (Mustafaraj et al., 2011). We propose that a similar phenomenon occurred during the Scottish independence referendum, though instead of focusing on the minority vs. majority dichotomy, we believe it is a matter of stability vs. change.

In our dataset, there are more tweets proposing that Scotland vote Yes on Independence than vote No. The #VoteNo tweets represent a much lower number of
tweets within the dataset, both in pure numbers and relative to #VoteYes tweets. Based solely on viewing the numbers of tweets for each side, one might incorrectly assume that those in favor of independence won the referendum. We propose that this is due, in part, to the vocalization of those seeking change. People who are urging change have more of a reason to be vocal, because if nothing happens, the status quo continues. As was shown in Table 1, the average number of tweets per user for #VoteNo tweets was 2.9 which is less than half the average number for #VoteYes tweets. Whether this difference is intentional, such as being a result of successful mobilization or organization of the #VoteYes campaign, or whether it can be attributed to other factors, requires further research. Others have examined a similar phenomenon related to elections and incumbency (Livne, Simmons, Adar, & Adamic, 2011). We propose that people in agreement with the status quo are less likely to be passionately vocal about it, as the default is in their favor. This difference could be augmented through variations in social media mobilization and strategy, and enthusiasm for a particular outcome. In the week before the referendum, #VoteYes tweets were represented 7.7 times more than #VoteNo tweets. Included in the potential impacts that led to that variation in numbers are demographics; age in particular may have had a bearing on the totals.

Another theme we explored within the data was undecided voters. To do this we looked for tweets with co-occurrences of #VoteYes and #VoteNo. We also searched for the term ‘undecided’ within the data. A total of 2211 tweets contained both #VoteYes and #VoteNo. Upon examination of these tweets, some supported the idea that co-occurrence of these terms was related to undecided voters. An example can be seen in
the following message: “Swaying towards yes today. The actions of better together and attitude in London is disgusting. Can't make my mind up. #VoteYes #VoteNo.” A similar, though more concise tweet simply stated: “#voteyes #voteno #confused.” Though these and similar examples expressing indecision were present within the co-occurrence tweets, voters who appeared to have already made their decision were also prevalent. In many of these tweets, a statement or position is attributed to one side, and a rebuttal is expressed using the other hashtag. An example of this can be seen in the following tweet: “How can someone #VoteNo basing it on the promise of change when a #VoteYes will guarantee it? #indyref.” The term ‘undecided’ was more prevalent than co-occurrences of #VoteYes and #VoteNo within our dataset. 4646 tweets contained the term. These results are important because our dataset contains tweets from the week before the vote. In post referendum polling 9% of No voters stated they made their decision on how to vote within the last week, while the number was even higher for Yes Voters at 21% (Lord Ashcroft, 2014).

5.1.2 Demographic comparisons

Social media data for the UK shows that 56% of people use social media in general and 18% of the population accesses Twitter in particular (Ipsos MediaCT, 2014). Users between the ages of 15 and 34 make up 59% of Twitter users, though they only account for 33% of the population (Ipsos MediaCT, 2014). Users in this age bracket will be referred to as prime Twitter users for the remainder of this thesis due to the large percentage of Twitter users that they represent. A map of Scotland’s prime Twitter users
can be found in Figure 9. This overrepresentation of young people on Twitter plays a part in the discussions taking place. To examine this, we compared the total number of tweets collected in our dataset to the population in each council area as well as to the population of young people in each council area. We found that the total number of tweets in our dataset correlated with the council area population. A scatter plot of these two variables is shown in Figure 7. This was expected, as areas with more people are likely to have more Twitter traffic.

![Council Area Population and Tweets](image)

**Figure 7: Comparison of council area population and tweets**

Next we compared the population of each council’s prime Twitter users, those people between the ages of 15 and 34, to the total number of tweets in our dataset for each council area. These values were also correlated, and at a higher level. As mentioned
before, people who use Twitter tend to be younger, which can shape the conversation that plays out on Twitter. With an overrepresentation of young people guiding the conversation, some groups may be overshadowed, particularly older people. Though we assume that our dataset is probably underrepresenting older Scots, we wanted to add weight to our assumptions through data.

In the UK, people 55 and older (the highest age bracket in the study) represent only 8% of Twitter users (Ipsos MediaCT, 2014). The National Records of Scotland provides data on the percentage of each council area that are pensionable; in Scotland, the pensionable age was 65 for men and approximately 62 for women (National Records of Scotland, 2014). We took this figure and compared it to the total number of tweets in our dataset for each council area. The result was a negative correlation. Council areas with higher percentage of pensionable people have lower tweets in our dataset. This result adds weight to the idea that content discussed on Twitter may underrepresent some groups of people, specifically older people who tend be less likely to be Twitter users. In a post referendum poll, 73% of voters over 65 stated they voted ‘No’. With the small representation for this age group in Twitter, it is likely that this group was underrepresented in our dataset.

Though they were likely underrepresented, older Scots were not absent from our data. In fact, the age of one particular Twitter user probably impacted the spread of her message though Twitter. A user claiming to be 72 expressed her support of the #VoteYes campaign, and this message was retweeted numerous times. This is an example of how the age of Twitter users can be hard to interpret. If every user began by stating their age,
as this user did, it would be much easier for researchers to account for discrepancies in the data due to age. Barring this (admittedly very unlikely) scenario, researchers should attempt to account for age differences within the data in other ways, such as through the use of quality demographic data. This is not to say that social media data should not be used, only that researchers should acknowledge and consider that the demographics of an area might impact the content of the data.

Figure 8: Tweet from a 72 year old user

My first ever tweet, I'm a 72 year old pensioner - Vote yes!! #indyref #VoteYes
Figure 9: Prime Twitter users in Scotland per council area as a percentage of population. Prime Twitter users are defined as people between the ages of 15 and 34 for this thesis
The potential impacts to this study can be seen in Figure 9. In Scotland, the prime Twitter users are located in and around the major population centers of Glasgow, Edinburgh, Dundee, and Aberdeen City. The high percentages of young people in these areas, coupled with the larger than average population mentioned in section 4.1.3, leave the major cities of Scotland in a prominent position when it comes to Twitter content.

5.1.3 **Standard deviational ellipses**

Tweets with the hashtags #VoteYes and #VoteNo for the week leading up to and including the referendum were mapped and standard deviational ellipses were applied. The results can be seen in Figure 10 and Figure 11 separately, or combined in Figure 12. The size and shape of the ellipse represent the dispersion of tweets for the individual hashtags. Both ellipses are centered in the Population Belt, though differences can be noted. The #VoteYes ellipse is more compact than the #VoteNo ellipse, indicating that more of the tweets advocating for Scotland’s independence occurred within this more urban Population Belt. In contrast, the reach of the #VoteNo tweets extends further, signifying a less condensed core. After calculating the land area covered by each ellipse, we determined that the #VoteNo ellipse was indeed larger, by over 1200 square kilometers.
Figure 10: Distribution of #VoteYes tweets in Scotland for 12 – 18 September, showing standard deviational ellipse.
Figure 11: Distribution of #VoteNo tweets in Scotland for 12 - 18 September, showing standard deviational ellipse
Figure 12: Distribution of #VoteYes and #VoteNo tweets for Scotland for 12 -18 September, showing standard deviational ellipses
As noted in the literature (Kent & Leitner, 2007), the profile and direction of the ellipse relates to the underlying landscape. In our case, the direction is associated with the geographic distribution of the population.

As mentioned before, the cities potentially play a major role in determining the content of Twitter data. In the standard deviational ellipses, the high number of tweets from cities and the surrounding areas anchor the ellipses in the population belt. To see if the variations in dispersal between #VoteYes and #VoteNo occurred across the country, analysis was done without these areas included. The resulting map can be see in Figure 13. The ellipses were recreated without data from the two largest cities in Scotland, Glasgow and Edinburgh. The range of the dispersion increased dramatically; with a land cover increase of 78% for the #VoteYes ellipse and 81% for the #VoteNo ellipse. The relational layout of the #VoteYes and #VoteNo ellipses remained similar at this expanded range.
Figure 13: Distribution of #VoteYes and #VoteNo tweets in Scotland for 12 – 18 September excluding Glasgow and Edinburgh, showing standard deviational ellipses
This indicates that the difference in geographic dispersions between #VoteYes and #VoteNo tweets is impacted by the cities, but not necessarily driven by the cities. Meaning that in other areas of the country, when analyzed without these dense urban areas, there likely will have been geographic differences between the spatial distribution of the two sides’ tweets. The ellipses would tend to pull towards urban areas when they are included in the analysis. To see if this was the case, we divided the country into three main areas shown in Figure 14. To create these sections, we started by defining the population belt. To do this we looked at the breakdown of population per square kilometer and took any council area with a population per square kilometer of over 140 and grouped them together. Geographically this worked well, aside from Aberdeen City. All of the council areas with over 140 people per square kilometer were contiguous, aside from Aberdeen City. Since Aberdeen City was geographically separate from the rest of these council areas, we removed it from the population belt group and added it into the North Scotland group with all of the other council areas that are north of the population belt. In doing this we assumed that the ellipses for the area north of the population belt would trend toward this populated area. The remaining council areas were grouped into the Southern Scotland group.
Starting with the population belt, the standard deviational ellipses were completed again using the new divisions. The shape of the ellipses, shown in Figure 15, are much
more narrow when looking only at tweets in the population belt, due to all of the tweets falling within a smaller, nearly horizontal strip of land. The width shows the inclusion of both Glasgow and Edinburgh. Though similar, the independence tweets tend to edge more toward Glasgow whereas the dispersion of #VoteNo tweets tends to favor Edinburgh. These results are interesting given the conventional characterization of Glasgow as a working class/liberal city compared to Edinburgh as an upper class/more conservative location. Glasgow was also one of only four council areas to vote in favor of independence during the referendum, with two of the remaining council areas voting in favor being adjacent to it. After examining the land area covered by each ellipse, we found that when looking at only the population belt, the land area covered by the #VoteYes ellipse was larger, by 440 square kilometers.
Figure 15: Distribution of #VoteYes and #VoteNo tweets in the Population Belt of Scotland for 12 - 18 September, showing standard deviational ellipses
When examining tweets north of the population belt as is shown in Figure 16, the range became much larger showing a lower tweet density than within the Population Belt. The #VoteYes ring again edges to the South and West compared to the #VoteNo ellipse. The area of Dundee was the only council area outside of the population belt to vote in favor of independence. This concentration of people in favor of independence could contribute to the #VoteYes ellipse being more compact. The Orkney Islands and Shetland Islands off the northern coast of the Scottish mainland voted strongly in favor of keeping the union during the actual referendum and within our dataset, these locations were comparatively active in the conversation to #VoteNo. This along with the strong #VoteNo leanings of Aberdeen City within our dataset likely contribute to the pull of the #VoteNo ellipse to the North and East. When the land area covered by each ellipse was examined, the #VoteYes ellipse covered 18% more land area, even though the #VoteNo ellipse was bigger. This is likely due to the geographic pull of the islands mentioned earlier in this section. The pull of data from these islands, likely contributed to the area of the circle being larger, though much of this increase in size is over water. In future analysis using this technique, removing the islands and analyzing them independently may be prudent.
Figure 16: Distribution of #VoteYes and #VoteNo tweets north of the Population Belt in Scotland for 12 - 18 September, showing standard deviational ellipses
Next we looked at the Southern Scotland ellipses. As shown in Figure 17, differences between the dispersion of #VoteYes and #VoteNo were evident. The #VoteYes distribution was more compact, and edged to the west. The #VoteNo distribution was less compressed, and edged South and East toward England. In the actual referendum results, both council areas that border England voted strongly against independence, with over 66% of each of the council areas voting ‘No’. In our dataset, Dumfries and Galloway strongly led the #VoteNo conversation, but Scottish Borders fell slightly on side of #VoteYes. This probably contributed to the southern pull of the #VoteNo ellipse. The land area covered by the #VoteNo ellipse was 7,649 square kilometers, while the #VoteYes ellipse covered 6,063 square kilometers of land.
Figure 17: Distribution of #VoteYes and #VoteNo tweets south of the Population Belt of Scotland for 12 - 18 September, showing standard deviational ellipses
Finally, all council areas outside of the population belt were examined. This map shown in Figure 18 showed a similar pull toward the South and East for the #VoteNo dispersion, and also included the larger north-south range encompassed by this ellipse. The #VoteYes ellipse has a similar orientation, though it was more compact than the #VoteNo ellipse. The land area covered by the #VoteNo ellipse was only slightly larger than that covered by the #VoteYes ellipse, less than 5%. As mentioned earlier, relatively strong #VoteNo areas in our dataset included Dumfries and Galloway, pulling the ellipse to the south, as well as Aberdeen City, and the islands off of the northern coast. In our dataset, the Highlands conversation was dominated by #VoteYes, and likely contributes to the pull in that direction of the #VoteYes ellipse.
Figure 18: Distribution of #VoteYes and #VoteNo tweets outside of the Population Belt of Scotland for 12 - 18 September, showing standard deviational ellipses
5.1.4 Referendum Results

Having looked at the overall patterns of the tweets, we next examined them at a council area level. As mentioned in section 4.1.1, the dataset was reduced to the sub-national level for our analysis. A concern in this method was that by removing tweets geolocated at the national level one side would be favored more than the other. To look closer at this, the daily total of each hashtag in the sub-national dataset was compared to the daily total of the same hashtag including the national level data. This is shown in Figure 19. When the data is examined, it appears as though the #VoteNo campaign took a harder hit with the removal of national level data points. The sub-national dataset for both #VoteYes and #Indyref hover around 70% of the dataset that includes national geolocation for the week prior to the vote. In the same time period, the #VoteNo tweets average just under 60%. This phenomenon has been described in other areas using geographic information as the Modifiable Areal Unit Problem (MAUP), an overview of which can be found by Charlton (2008). The potential discrepancy of geographic scale for a particular data group represents an issue that should be considered and addressed by researchers using geolocated social media data, as well as a potential research area. It is possible that this thesis, with a particular topic of research dealing with issues of nationality and independence, may be more prone to this type of influence due to the connection of the issue to a geographic region. However, more analysis will be needed to address this.
Before examining the Twitter results by council area, the political control of Scotland was mapped. This representation, displayed in Figure 20, shows each council area shaded according to which party had the most seats. As stated earlier, the Labour, Conservative, and Liberal Democratic parties backed the Better Together campaign (“Independence is ‘one-way ticket,’” 2012), which was in favor of keeping the union between the UK and Scotland. The Scottish National Party was in favor of Independence.
Figure 20: Political control of Scotland showing largest party by council seats

On 18 September, the referendum was held with record turnout. 84.6% of eligible voters took part in the vote with the final results of 55.3% in favor of continuing the
union with the UK and 44.7% voting for independence (Scottish Parliament Information Centre, 2014). Four council areas, Glasgow, Dundee, North Lanarkshire and West Dunbartonshire voted in favor of independence, while the other 28 voted against. These results are shown in Figure 21, with the blue council areas representing those that voted in favor of independence and the red areas showing those areas that voted against independence. The darker shades for each color represent the strength of the winning vote in each council area.
We looked at the number of tweets in each council area and it was evident that based on pure number of messages, the independence movement controlled the
conversation on Twitter. This is shown in Figure 22. The pie symbols are divided to show a comparison of the number of #VoteYes and #VoteNo tweets in each council area. The size of each pie is related to the total number of #VoteYes and #VoteNo tweets in each council area. Every council area in Scotland had more #VoteYes tweets than #VoteNo tweets. As mentioned before, the silent majority and incumbency issues probably played a major role in these results, but age is another factor that should be considered. In post referendum polling (Lord Ashcroft, 2014), prime aged Twitter users said they voted in favor of independence. 51% of 16-24 year olds said they voted in favor of independence and that number rose to 59% when looking at 25 – 34 year olds. Older people, who as mentioned before were much less likely to be active Twitter users, voted against independence; 73% of people 65 and older said they voted ‘No’.
With the conversation dominated by the #VoteYes side, we wanted to account for this factor to examine the relative differences in conversation at the council area level. To
do this we took the overrepresentation factor of #VoteYes tweets for the entire data set, and normalized the #VoteYes tweets in individual council areas using this factor. The results, shown in Figure 23, illustrate the relative conversation occurring at the council area level. The four council areas that voted in favor of independence led the discussion in their respective areas, though interestingly not necessarily all were at an overwhelmingly strong rate. In Dundee and West Dunburtonshire, the #VoteYes led the conversation accounting for 70% and 63% of the normalized tweets respectively. In Glasgow #VoteYes accounted for 56% of the normalized tweets, while in North Lanarkshire, the #VoteYes tweets only represented 51% of the normalized dataset.

Though many of the council areas that voted strongly against independence were leading the discussions in our dataset, some were not. In particular Scottish Borders voted strongly against independence, but based on our dataset, it appeared that the conversation was led by the #VoteYes side. Urban centers including Glasgow, Edinburgh, Dundee, and Aberdeen City all voted with the side leading the conversation in our dataset.
Figure 23: Normalized comparison of #VoteYes and #VoteNo tweets by council area of Scotland for 12 - 18 September, #VoteYes tweets normalized by over-representative magnitude
CHAPTER SIX

6.1 Conclusion and Outlook

This thesis presented a study of how the referendum played out on social media, specifically Twitter, in comparison to how the event was recorded through the polls. In essence we wanted to see how data from Twitter and the digital sphere, compared to data from the actual event. More specifically, we focused on the study of the geographic distributions of #VoteYes and #VoteNo tweets in Scotland in the days leading to the poll. So how does Twitter data relate to what is happening on the ground? Twitter analytics differs from traditional data mining tasks due to the very nature of Twitter: the data reflect expressions of public interest, and as such they are the subject of a complex social process that involves not only opinion formation, but also the dynamics that govern the patterns of expression of such opinions. For example, different political causes and/or opinions on the same topic may display substantially different levels of mobilization (Diani, 2009), which in turn can lead to substantial variations in the level of participation in social media by the Yes supporters as compared to the No supporters.

For example, incumbency has been considered correlated with depressed voting turnouts (Schaffner et al., 2001) and, one can extrapolate, lower levels of mobilization. One could argue that #VoteNo represented incumbency in this referendum, and as such
its supporters tended to mobilize late and at lower levels of excitement than their opponents.

Furthermore, when doing research of this type, researchers must be cognizant of how other factors can influence the conversations on Twitter. Age plays a major role, so in areas where quality demographic data exists, research may benefit from its use. For example, our data probably underrepresented older Scots, and knowing this is important for further analysis.

Using basic text searches, we examined whether topics present within the campaign were found within our dataset. This opens up the door to future analysis using terms, perhaps though geographic analysis. Clustering or geographic spread of individual terms is a possible next step, along with comparing the locations of term clusters to demographics. This could provide details on issues that are important to specific locations. As mentioned in section 4.12, the possibility of removing the Trident nuclear program from Scotland was discussed. By examining the locations of these tweets, trends may emerge showing whether the event is being discussed more heavily in specific locations, perhaps near the current facilities where jobs and economic impacts would be felt more intensely, or whether it was an issue discussed equally across the country. More research is necessary to make conclusions in this area.

The geographic spread of classes of data was examined as well, showing that different topics are not confined to the same geographic space. This has the potential to impact research. For instance, our data noted a difference in twitter data between Glasgow and Edinburgh. If a sample of data were to be collected from Scotland, this
variation could come into play, depending on what areas were used to collect data. Also, in analyzing topics in Twitter at various geographic scales, differences in content may be observed. Though this has been reported before in relation to geographic information (Charlton, 2008), the effects on social media data specifically have not been studied in detail. Whether or not this phenomenon is more acutely at play with topics explicitly linked to geographic regions, such as was used for this thesis, requires more research. Our data heavily favored densely populated urban areas. When conducting research using social media data, accounting for the urban/rural divide needs to occur in order to limit bias within a dataset. To see how this data impacted our sample, we conducted analysis both with and without these urban areas.

Using social media data also comes with some limitations. Age was discussed in this thesis; however, other demographic factors including race and ethnicity as well as gender impact the data collected (Mislove et al., 2011). The locations used for analysis can also be questionable. The ability to lie, or misrepresent your location is a possibility, as well as the ability to input made-up places. In the United States, Twitter users generally report their location at the city level (Hecht, Hong, Suh, & Chi, 2011); however this does not necessarily mean users in other countries do the same. In fact, the amount of research comparing Twitter users in different countries to each other is limited (Poblete et al., 2011). Digital divide issues also have the potential to impact results (Stefanidis et al., 2011).

Even with these limitations, social media data are extremely useful for research. Data can be collected from afar, in near real time. Social media data is also building a
history of the public’s conversations at this point in time. With the increase in storage for digital data, soon it would not be hard to imagine all tweets recorded and stored in a time-capsule like reference for future generations to give context to events occurring today.

Though Scotland ended up remaining a part of the United Kingdom, other areas around the world face similar situations. Quebec in Canada, Catalonia and the Basque Country in Spain, and Flanders, Belgium all have independence movements that may progress in the future (Dardanelli & Mitchell, 2014). Comparing the conversations that occur on Twitter before any potential referendums or major events in these regions could provide additional information on the way Twitter is used during independence movements. Comparisons of our data set to other events in Scotland could also help identify trends in the use of Twitter and social media in general for that particular area.

The data available to researchers today through social media is extremely useful. It is available quickly and in regions and topics that would not have been practically feasible to access regularly in the recent past. New methods to access and analyze this data are constantly being created and implemented, and more and more research is being conducted using this type of data. Researchers must keep in mind that variations in the data can occur due to demographics, geographic scale and other issues. Being conscious of these issues, and addressing them as completely as possible, will allow for superior social media data analysis.
REFERENCES


BIOGRAPHY

Tom Holby graduated from Lehigh Senior High School, in 1998. He received his Bachelor of Arts from Flagler College in 2002.