THE SPATIAL DISTRIBUTION OF HEALTH NARRATIVES IN TWITTER AND THE RELATIONSHIP TO CORRESPONDING CANCER RATES ACROSS THE UNITED STATES: A CASE STUDY OF CANCER-RELATED COMMUNICATIONS

by

David A. Novak Jr.

A Thesis
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Master of Science
Geoinformatics and Geospatial Intelligence

Committee:

_____	Dr. Anthony Stefanidis, Thesis Director
_____	Dr. Andrew Crooks, Committee Member
_____	Dr. Arie Croitoru, Committee Member
_____	Dr. Anthony Stefanidis, Department Chairperson
_____	Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science
_____	Dr. Peggy Agouris, Dean, College of Science
Date: _____	Summer Semester 2017 George Mason University Fairfax, VA

The Spatial Distribution of Health Narratives in Twitter and the Relationship to Corresponding Cancer Rates Across the United States: A Case Study of Cancer-related Communications

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

by

David A. Novak Jr.
Bachelor of Science
Old Dominion University, 2004

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

Summer Semester 2017
George Mason University
Fairfax, VA

DEDICATION

This is dedicated to my loving & supportive wife, and my wonderful son.

ACKNOWLEDGEMENTS

I would like to thank the members of my committee for their guidance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS AND/OR SYMBOLS

Breast Cancer Awareness Month.....BCAM
Geographic Information Systems .....GIS
Application Programming Interface .....API

ABSTRACT

THE SPATIAL DISTRIBUTION OF HEALTH NARRATIVES IN TWITTER AND THE RELATIONSHIP TO CORRESPONDING CANCER RATES ACROSS THE UNITED STATES: A CASE STUDY OF CANCER-RELATED COMMUNICATIONS

David A. Novak Jr., M.S.

George Mason University, 2017

Thesis Director: Dr. Anthony Stefanidis

National Breast Cancer Awareness Month (BCAM) and Movember health campaigns in Twitter from the years 2015 and 2016 were studied to understand how tweets formed around these campaigns relate to cancer incidence ground truth data. Geolocated tweets were collected to characterize the spatial distribution at the state level of breast and prostate cancer related tweets, and comparisons were made between tweets and cancer incidence data to assess the relationship between tweet rate and state cancer incidence rates in the United States. It was hypothesized that states which participate the most in these cancer campaigns would exhibit higher cancer incidence rates; contrariwise, there was no correlation found between tweet rate and state cancer incidence rate for all four campaigns studied, indicating that these two variables did not exhibit a relationship in this study. A better understanding of health campaign participation and the relationship to

cancer affected populations in Twitter can assist health professionals determine the effectiveness and impacts of health campaigns in Twitter.

1. INTRODUCTION

Social media analysis is another method to sample and derive knowledge from the thoughts, discussions, and sentiments of large numbers of people. The survey of large numbers of people was limited to traditional survey methods such as questionnaires and interviews before the advent of Web 2.0. The fact that many people use Twitter extensively to discuss heath issues [1] and that Twitter provides geolocation data [2] enables a study of cyber communities and their translation to physical communities by comparing the location of cancer related narratives with ground truth data. Participation in health awareness campaigns and associated health narratives in social media and the relationships of these narratives with ground truth data may serve as proxies for the geographic distribution of health and disease issues.

People may discuss cancer in social media for many reasons, maybe because they have been diagnosed, because of a family member or friend, or simply out of empathy for others. Studying health campaigns like BCAM and Movember, as they are discussed and shared in a public social media platform like Twitter may provide information about their effectiveness, and about the population affected by disease and other health issues.

By developing an understanding of the mechanisms which drive participation in social media and techniques to measure or characterize it, we can advance our ability to analyze crowdsourced health content and design health information campaigns.

Some examples of the healthcare and medical profession realizing the potential of studying health topics in social media like Twitter are:

- The collaborative effort between Stanford University's Medicine X academic medical program and the healthcare social media analytics company Symplur[1]. One facet of this collaboration is the Healthcare Hashtag Project which seeks to vet, categorize, and manage a database of relevant health related hashtags to assist health researchers take advantage of health narratives occurring in Twitter.
- Health topic chat groups on Twitter, known as tweet chats, are being used to connect doctors, patients, and researchers around health topics and supplement the traditional use of the in-person focus group. Many health chats are regularly scheduled and are centered around a range of topics. Users can find these chats using established hashtags and participate in moderated, structured discussions in real-time and read compiled transcripts [3].
- A recent study of topics and themes occurring in Twitter focused health research literature enabled the development of a taxonomy of Twitter data use with 6 categories: content analysis, surveillance, engagement, recruitment, intervention, and network analysis which researchers and health professionals can use as a guide for designing health studies using Twitter [4].

[1] https://www.symplur.com/healthcare-social-media-research/

In order to raise awareness for diseases, the public health community has developed targeted campaigns that address specific diseases in an effort to inform the public, form relevant communities, and often drive fundraising efforts. Breast and prostate cancers are the most commonly diagnosed among women and men respectively, and the second leading death causing cancers[2]. There are many health awareness campaigns associated with these cancers; two prevailing ones are Breast Cancer Awareness Month (BCAM) and Movember which are the main focus of this thesis. BCAM is an effort promoted and sponsored by the American Cancer Society and other breast cancer charities to fundraise and educate women about early detection and increase awareness about breast cancer[3]. The concept of Movember began simply as a mustache growing contest in Australia. Inspired by breast cancer fund raising, the contest turned into a fundraising campaign for prostate cancer, and is managed by the Movember Foundations and several other men's health organizations. Movember is a more generalized campaign, focusing on several men's health issues such as prostate and testicular cancers, mental health issues (depression, suicide prevention), and concern with physical inactivity[4].

The overall research challenge addressed in this thesis is whether local disease rates (at the state level) drive participation in social media narratives associated with these diseases. To pursue this question, we study the spatial distribution of geolocated tweets across the United States and compare them to ground truth data of corresponding

[2] https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2016.html
[3] https://www.cancer.org/
[4] https://us.movember.com/about/foundation

cancer rates across the United States. This will be achieved by analyzing four datasets of collected tweets to determine the number of tweets originating from each state regarding breast and prostate cancer during BCAM and Movember, and comparing the tweet rates to breast and prostate cancer incidence rates. Through this analysis we pursue insights to the following research questions:

- Which states tweeted the most during the campaigns?
- How did the campaigns differ from year to year, and how did the two campaigns differ from each other in terms of participation and spatial distribution?
- Is there a relationship between the number of tweets (tweet rate) and local state cancer incidence rate?

Collectively, these research questions will address the following hypothesis: "States with higher breast and prostate cancer incidence rates participate more in online breast and prostate cancer-related Twitter activities."

1.1 Thesis Organization

This thesis is organized as follows: Section 2 is the literature review, Section 3 describes the Twitter and cancer incidence data and provides some exploratory analysis of the Twitter data, Section 4 presents the analysis results of the Twitter data and cancer incidence data, and Section 5 offers discussion of the findings, future considerations, and the conclusion.

2. LITERATURE REVIEW

This section provides an overview of topics regarding social media in health study: Twitter metadata and how it is used for analysis, the nature of health campaigns, breast and prostate cancer and campaign study, and the thesis contribution.

2.1 Social Media in Health Study

There are many platforms available for the study of health and medicine in social media, in the form of blogs (WebMD) and microblogs (Twitter), social and professional networking applications (Facebook, LinkedIn), wikis (Wikipedia) and many others. Social media offers non-profit and authoritative health organizations a means of inexpensive electronic advocacy, and health researchers a source of digital media research [5].

2.2 Twitter

Within social media platforms, Twitter not only offers a platform for interaction through recruitment of patients for studies, intervention, and health messaging, but importantly generates a rich data source for research. During review of research literature produced from 2010 to 2015 found in well recognized databases such as Web of Science, Google Scholar, and PubMed, researchers found the number of Twitter focused health research publications approximately doubled each year, and continues to rapidly increase [4]. They characterized the current state of health research in Twitter by defining a

taxonomy of Twitter data use with 6 categories: content analysis, surveillance, engagement, recruitment [3], [6], intervention, and network analysis [4].

The Twitter platform itself can used for research, but often the Twitter application programming interface (API) is accessed to obtain tweet metadata. The 1% streaming API is free, and has been found to be adequate where a general overview of the most characteristic aspects of a dataset is required. Examples of these aspects are the extremely popular tweets and retweets, and tweets containing geolocation information. The query parameters used when querying the API affect how representative the 1% sample stream is of the entire data set. For example, if a bounding box is specified in the query, mostly all geolocated content may be pulled, which seems to imply a priority or selectiveness for geolocation information in Twitter's API sampling mechanism. An encouraging aspect of the 1% streaming API sample is that one is likely to get much more than 1% of the possible available data by using it [7]. The metadata record for a tweet [8] can contain dozens of data features, with the most notable ones being the geographic information, username, user ID, timestamp, text content, and retweet information. Study of the information provided in tweet metadata enables the study of those who tweet and many health phenomena to include the spread of epidemic outbreaks, health campaigns, and many others.

2.2.1 Geographic Information

Regarding the geographic information provided in tweets, social media communities, such as Twitter, have both a cyber and a physical presence. When geolocation information (geographic coordinates, toponyms) accompanies metadata or

tweet content, it is often referred to as geosocial media and can be mapped to understand the geographic representation of localized events [9]. The granularity of geolocation information included with tweets differs, and can provide precise location information if it is provided by the user's electronic device, or coarse where the geographic information is based upon geolocated toponyms [2]. An example of this might be comparing geolocated tweets of people discussing a topic to a geographic dataset of addresses to see the degree which the cyber-community represents the physical one. The relationship between geolocated tweets and the location of addresses would be dependent on the level of spatial aggregation, for example, there may be a strong correlation between tweets and addresses at the state level, but not at the zip code level [10].

2.2.2 Content analysis

Content analysis was found to be the most popular form of research in cancer related social media [11], and content analysis often focuses on user content and tweet text content by searching with keywords of interest.

An elusive aspect of analyzing health narrative content are the associated demographics which characterize groups of people (e.g. age, race, sex, ethnicity, education level, income level etc.), as different demographics are affected differently by health issues and differ in their use of social media. However, estimating the demographic characteristics of Twitter data is difficult [12], does not represent the general population, and the demographics of the Twitter population that may tweet about health concerns is not known and difficult to estimate [13]. The lack of demographic information associated with social media presents a significant challenge of working with

it in social science research [14]. However, some researchers have used methods of gender identification from first names in profiles, and social class and age from vocabulary in the tweets [15] and supervised classification techniques in combination with Census data to detect user attributes [16]. One source of statistics about the demographics of social media use from the Pew Research Center [17] indicates that certain age groups, for example 18-49, use Twitter the most. This must be considered when trying to understand how well health narratives in reflect who tweets about cancer [18].

Regarding user information and content which can be associated with the user, social media research comes with a responsibility to treat user information with care. Ethical issues can arise because potentially personal information is being collected without the users knowing it, or in the case of health study recruitment in social media, patients are being asked to participate in a discussion or study [19].

2.2.3 Temporal analysis

The available time and date (timestamp) data available in social media facilitates temporal analysis of how health narratives evolve and enables event-based surveillance; a potentially valuable resource for understanding the emergence of disease and evolution of epidemics and pandemics. This event-based surveillance is important for vaccine preventable disease, pandemics, emergence of pathogens, food issues, and bioterrorism among others. Further study of health information communication will help public health agencies incorporate event-based methods into their surveillance programs [20], [21].

2.2.4 Social network analysis

Social network analysis in Twitter enables study of communication and the spread of health or disease information. The structure of health narratives can be described as tweet text, patterns of re-tweeting which show networking and community formation amongst people discussing the topic, and the spatial patterns associated with geolocated tweets. As an example, a study of vaccination discussion following the 2015 measles outbreak found the following: retweet patterns and the statistics of the tweet content showed that tweets often referenced terminology or headlines associated with mainstream media news stories, and that mainstream media seemed to have a significant impact on the social media narratives regarding vaccinations and measles, in contrast to the impact associated with authoritative health organizations [18].

Social media and general public participation play an important role in dissemination of health information (accurate or inaccurate) and may be more influential than that of formal press and health agencies. Poor understanding of social media participation limits our ability to use readily available, very large data sets to understand the top-down and bottom-up effectiveness of health agencies and the public respectively to evaluate and optimize public health campaigns [22].

As briefly outlined above social media analysis and Twitter data offer a wide range of applications and techniques for research of health issues. We shift our focus to the study of health campaigns.

2.3 Health campaigns

It is not well understood why people want to or are comfortable using social media to discuss health issues. A study found that four main themes motivate people to

use social media to discuss health issues on Twitter: sense of community, raising awareness and eliminating stigma, safe space for expression, and coping empowerment [23]. Health campaigns are often viewed as positive vehicles to improve the health of the general population, but it has also been shown that corporate hashtag campaigns may also serve as a public health concern [24].

Twitter was found to be the second most used social media platform by nonprofit human service organizations, and mostly used to promote and advertise services or engage with the respective community. It was also found that organizations had no specific goals or strategy when using social media [25]. Further understanding of this may help organizations with defining goals for health campaigns.

Health organizations use Twitter differently to promulgate health information during campaigns. To understand these differences, tweet content was analyzed to assess the organizations' implementation of the Health Belief Model (HBM), a framework for explaining why people may participate in efforts to prevent and detect diseases. The HBM construct consists of the following attributes: perceived susceptibility, perceived severity, perceived benefits, perceived barriers, cues to action, and self-efficacy. Of the organizations studied, perceived barrier was the attribute noted the most in the tweet content. This indicates that the organizations focused on appealing to what individuals may perceive as barriers, for example a belief that a threat does not exist, or perception that screenings or treatments are out of reach [26].

2.3.1 Opportunities for health campaigns

Twitter has a significant reach and tends to be used by many teens and young adults, potentially making it a tool for communicating health or other behavior intervention information [27], especially where health issues prone to affect these groups are concerned. Research involving indoor tanning found much discussion about tanning (e.g. tanning beds, indoor tanning), sunburns, and other consequences of tanning, but found little discussion about the health issues associated with tanning, such as skin cancer [28]. The absence of cancer discussions in this example presents an opportunity for a health messaging campaign.

Although Twitter is a popular social media platform for health narratives in the United States, preference for a particular social media platform varies throughout the world. In the country of Ghana, not only were young people interested in seeing more health messaging in social media, but WhatsApp was identified as the preferred platform from which they would like to receive health information [29].

2.3.2 Changing attitudes and behavior

A feasibility study to determine the extent to which Twitter could be used to influence attitudes in Northern Ireland about skin cancer found that attitudes improved regarding UV exposure and skin cancer following a skin cancer campaign. It was estimated that 23% of Northern Irelands population was reached by the campaign, and that people tended to share information-related tweets more often than tweets containing humor or shocking content [30].

Following a review of general health campaign analysis in Twitter, next is a review of breast and prostate cancer specific campaign research.

2.4 Breast and prostate cancer campaigns

2.4.1 Breast Cancer Awareness Month (BCAM)

The following attributes were studied in Twitter during BCAM: tweet frequency, who tweets, message reach, and typical content. It was found that most tweets occurred in the beginning of the campaign and sharply tapered off, and that organizations and celebrities focused on fundraising and prevention, whereas individuals focused on events and apparel. Most tweets involving individuals were one-way communications, and most messages did not promote prevention. At that time, 2013, the scarcity of research regarding the study of health awareness campaigns in social media was noted [31].

2.4.2 Movember

Analysis of Movember campaigns shows that most narratives discussed during these campaigns are not focused on health and cancer related aspects of men's health (e.g. prostate and testicular cancer), but are focused on topics related to fundraising, general well-being, and growing moustaches [32], [33].

A major function of health campaigns is to raise money. It was shown that while some moderate to significant correlations were found between Movember website visits and tweets, there was not significant correlation between donations and tweets. Similar results were found with the Twitter data separated into two separate sets: health topic tweets and social tweets [34].

2.4.3 Comparison of breast and prostate cancer

Community structure analysis of breast and prostate cancer narratives on Twitter showed core communities which discussed these cancers often and were associated with breast or prostate cancer specific sources, and visiting communities which discussed

these cancers less often and were associated with general cancer information sources, such as the American Cancer Society. The core communities for prostate cancer were smaller than breast cancer, possibly for a few reasons: sources of breast cancer are more well-known and the information is more established, and women interact with each other more than men interact with each other [35].

A recent study reinforces that breast cancer has a higher presence in Twitter than prostate cancer, but that Twitter activity for both cancers has substantially increased for both cancers since 2012. Patients, celebrities, and the breast cancer industry were the most influential for breast cancer discussion, where physicians and media personalities were the top influencers for prostate cancer discussion [36].

2.4.4 Statistical relationships between disease and incidence

Spearman correlation was used to assess the relationship between prevalence for 22 different diseases and associated mentions on Twitter. It was shown that the correlation coefficient could be increased by adjusting for Twitter population and validating tweet content; for example, validating that "heart attack" used in a tweet referred to the condition, and not the idiom for extreme surprise. Regarding cancer prevalence, Twitter data indicated significantly more cancer prevalence than ground truth prevalence data did, implying heavy "over-tweeting" about cancer with respect to actual prevalence [37]. This example raises questions about how useful health narratives occurring in social media can be used as proxies for what may be occurring in the geographic space.

The relationships between different populations, state cancer incidence, tweet count/rate variables, and cancer centers and doctors were studied using bivariate correlation (e.g. Pearson correlation coefficient). There was no correlation found between tweet rate and cancer incidence, but relationships were found between cancer tweet frequency and the number of cancer centers and doctors per state population [38].

2.5 Thesis Research Contribution

The purpose of this thesis is to study the spatial distribution of campaign participation across the United States via Twitter during distinct health campaigns, and to understand the relationship between participation and cancer incidence in the United States. A focus on specified health campaign months and measurement of campaign participation and its relationship to cancer incidence is lacking in the literature.

As discussed in [38], Murthy and Eldredge collected Twitter data over the course of six months from December 2010 to May 2011 using specific cancer search terms (chemo, lymphoma, melanoma, cancer survivor). This thesis focused on tweets collected the during the specific campaign months of October and November for BCAM and Movember respectively, with cancer terms associated with the specific campaigns. Health campaigns should be a good time to collect data; it is known when they happen and theoretically it should be a time of heightened discussion about these topics where public narratives are on display for analysis. This differs from [38] by targeting specific campaigns and time periods where cancer discussions might be expected to be maximized. Also, the number of tweet records used for this study was higher: 561,586, 231,049, 88,948, and 41,801 for the four campaigns, in contrast to the approximately

91,000 total Tweets analyzed by [38]. Murthy and Eldredge also studied the relationship between tweet rate and cancer incidence rate, and did not find a correlation between the two. A study of this relationship was revisited in this thesis as well, but with four different data sets, and from two different years, following their example of using Pearson correlation to compare tweet rate and cancer incidence to see if different results can be obtained.

3. METHODOLOGY

3.1 Introduction

This section provides a description of the data and methods used for analysis. The data used for this project consists of tweets from Twitter, and cancer incidence data from the State Cancer Profiles web site [39], a collaboration initiative between the National Cancer Institute and the Center for Disease Control. The general approach for this study was to obtain cancer related tweets from the United States during BCAM and Movember awareness campaign months, count the tweets which originated in each state to assess the spatial distribution, and compare the tweet counts to the cancer incidence rates for those states to understand the relationship between the two and determine if states with higher occurrence of cancer also participate more in Twitter during the campaigns.

The tools used or this project were as follows: Microsoft Excel (data management), ArcGIS 10.5 (GIS analysis, data visualization), Jupyter Notebook (programming environment, data visualization), Python programming language and data analysis modules (data management and analysis), and Tableau (data visualization).

3.2 Data

3.2.1 Twitter

The tweets were collected from the 1% Twitter streaming application programming interface (API) during BCAM (October 1 through 31) and Movember (November 1 through 31), in 2015 and 2016, using the following keywords for breast

cancer: breast cancer, pinkribbonwalk2015, breastcancer, pink ribbon, pinkribbon, advancedbc, americancancersociety, bcastrength, bcaware, bckills, and nflpink. The keywords used for prostate cancer were: prostate cancer: prostate cancer, prostatecancer, blue cure, bluecure, and Movember. The Twitter dataset was obtained using the Geosocial Gauge System. This system interfaces with the Twitter API, queries it, and receives tweets and metadata from which geographic information and other pertinent tweet content can be parsed and readied for further analysis [40], [41].

Table 1 shows a tweet example and the data structure of a sample record. A username is shown in the author field for demonstration, in this case a genetic testing lab. This field normally contains an individual's username. Therefore, usernames are removed from the analysis, and the data are treated as anonymous contributions. Of particular interest to this project are the tweets which have geographic coordinates which can be mapped to the state where they occurred. The collected set consisted of tweets from all over the world. The data set was sorted based on the *country* field and records with a value of 'us' were chosen. Table 2 shows the tweets which met this criterion and were used for analysis of BCAM and Movember health communications in the United States.

Table 1: An example of a tweet and associated data structure. The data was stored in a tab-separated values (.tsv) text file.

id	7.82E+17
location	Aliso Viejo, CA
country	us
state	ca
zip	92656
x	-117.727804
y	33.567372
published_at	2016-10-01 00:00:01 UTC
author	AmbryGenetics (Ambry Genetics)
coords_from	location
mood	-1
lang	en
text	5-10% of breast cancers are inherited. Get more resources on breast cancer genetics here: https://t.co/ybwjCp0COzâ€¦ https://t.co/A1sg62ltQs

Table 2: Number of tweets collected from each campaign.

Health Campaign	Total United States tweets (all geolocated)	Tweets which fell inside the United States GIS boundaries
BCAM 2015	615,290	561,586
BCAM 2016	251,397	231,049
Movember 2015	98,582	88,948
Movember 2016	46,024	41,801

The tweets were loaded into ArcGIS 10.5 to perform a spatial join operation with polygons of the continental United States, Alaska, and Hawaii which facilitated counting

of geolocated tweets and associating them with each state polygon. This operation revealed an unforeseen but sensible drawback of using the GIS join operation to count geolocated tweets; only tweets which were located inside the polygons used in the GIS were accounted for. There were many Tweets which were just outside of the polygon boundaries which were not captured in the join operation. For example, these "lost" tweets were right outside the perimeter of the polygons, and in some cases, may have represented people in coastal areas, boaters off the coast, or imprecise representation of geographic boundaries in the GIS polygons. See Figure 1 for an example.

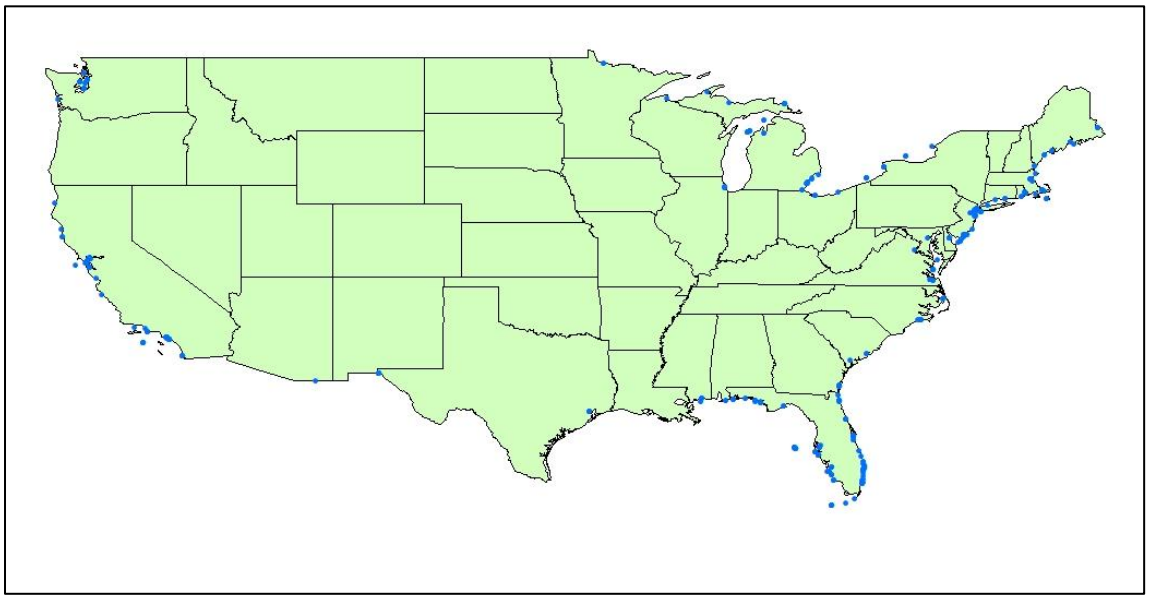

Figure 1: Movember 2015 tweets (blue points) just outside of state boundaries, near water bodies.

Following the GIS data exploration and join operation, the final tweet counts used for analysis of spatial distribution are reflected in the third column of Table 2., and the counts were exported for further analysis. Figure 2 shows the raw tweets counts and distribution of tweets across the United States for each campaign, with California having

the most, and Wyoming with the least. There is a marked difference in counts between 2015 and 2016; the counts from 2016 are much lower. Although not proven herein as an effect for this study, it has been suggested[5] that political content in Twitter significantly increased during October and November 2016 leading up to the United States Presidential election, and that overall Twitter content only grew by approximately 6% between 2015 and 2016[6], implying a significant influence of political discussion on Twitter content during this time frame.

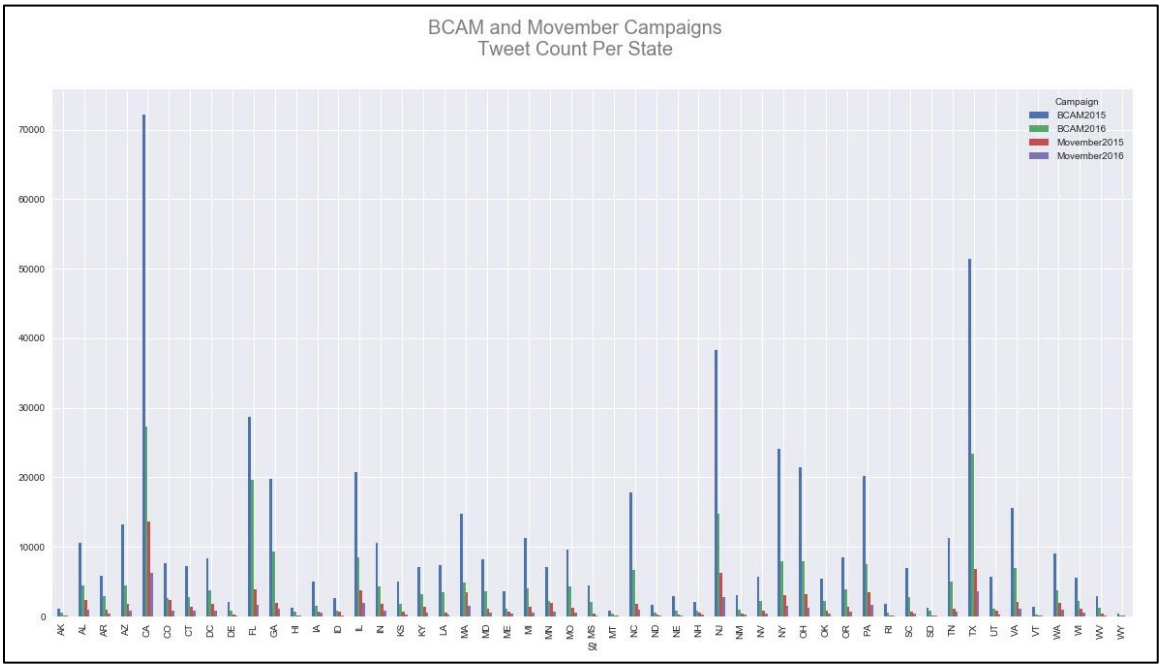

Figure 2: Graph showing the number of tweets collected during BCAM (October 2015/2016) and Movember (November 2015/2016) for all 50 states, and District of Columbia. States are ordered alphabetically on the x-axis, and the tweet counts are shown on the y-axis.

[5] http://elections.ap.org/buzz
[6] http://www.smartinsights.com/internet-marketing-statistics/happens-online-60-seconds/

Whereas Figure 2 shows raw tweet counts, Figure 3 shows health campaign participation, which incorporates state internet user [42] population information into a measure of participation. According to these data, internet usage across states is similar, varying from 86-97% for each state. The term "participation" used here is obviously also a function of the amount of Twitter data collected during the campaigns, which can vary based on many factors, including how it was collected or sampled from a larger cancer tweet data set. This variable for each state was calculated as follows, and indicates the number of tweets per 100,000 internet users:

Equation 1: Health Campaign Participation Variable

$$Participation = \left(\frac{State\ tweet\ count}{State\ internet\ user\ count}\right) * 100{,}000$$

Comparisons between state participation during the campaigns can be made using Figure 3 by looking at differing bar heights between states, or the bars for a single state. For example, DC participated significantly more than any state in all campaigns, and New Jersey is a distant second. This graph also illustrates the effects of adjusting and accounting for population (state internet users) with the number of tweets from each state. For example, California is clearly seen as having the most tweets in Figure 2, but has much less impressive participation as seen in Figure 3 once population is accounted for.

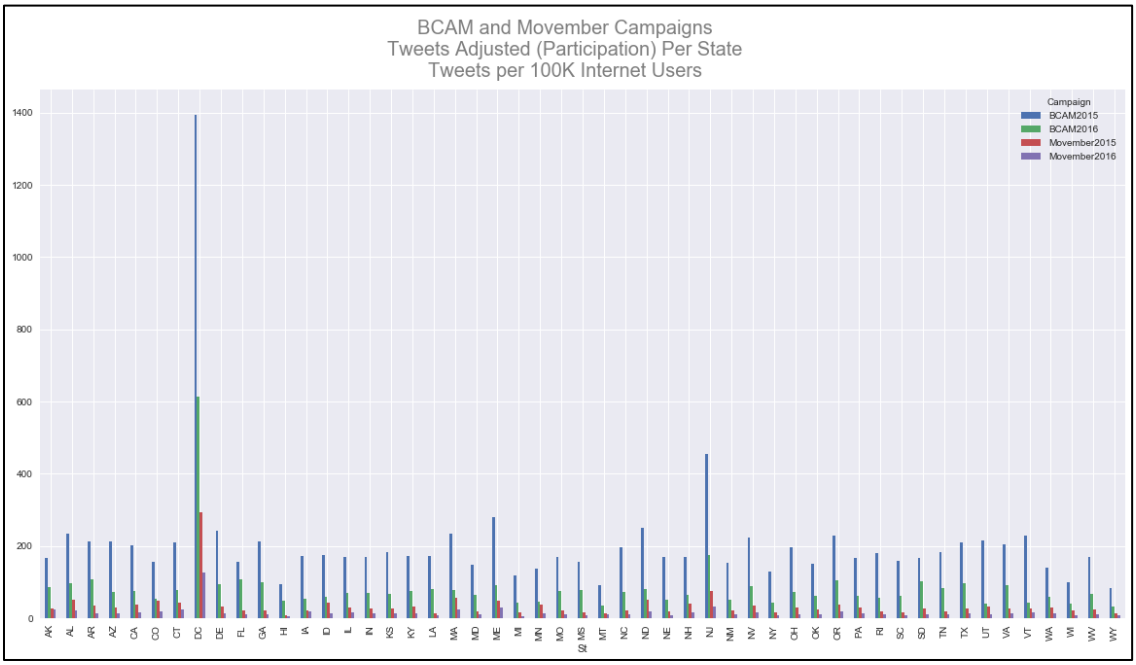

Figure 3: Bar graph showing the campaign participation during BCAM (October 2015/2016) and Movember (November 2015/2016) for all 50 states, and District of Colombia.

3.2.2 State Cancer Profiles

Cancer incidence rates for each state were obtained from the State Cancer Profiles website [23]. Cancer incidence represents new diagnoses of cancer within a specified time frame. The cancer incidence rate data used for this study is a five-year average, covering 2009-2013 and represents the number of cases per 100,000 population per year.

3.3 Tweet content analysis

3.3.1 Geographic/geolocation Content

As shown if Table 1, the data structure of the tweets as provided by the Geosocial Gauge System contained a field called *coords_from*. Unless blank, this column was populated with three possible attributes: ‘location’, ‘Twitter’, or ‘coords’, each specifying

the nature of how the geolocation information was assigned to each tweet. For example, a value of ‘location’ means the geolocation information was based on geocoding via gazetteer of a place name from the location column. A value of ‘Twitter’ means that Twitter provided the coordinates, and a value of ‘coords’ means that the geolocation information was provided along with the tweet from the user’s device, such as a computer or cell phone [2]. For this Twitter data corpus, the tweets were designated as shown in Table 3. Those designated as ‘coords’ represented a very small proportion (1%) of tweets with user provided location data. This small number of tweets provided from personal devices shows the limitations of using the data collected from the streaming Twitter 1% API data to determine the individual locations of where tweets come from.

The datasets were initially sorted based on the *country* field for records from the US. There were many tweets tagged as US, but actually located outside the country in the following US territories: Guam, Puerto Rico, Virgin Islands, American Samoa, and Northern Mariana Islands. Additionally, there were tweets which fell outside of the GIS state polygons, but within 100 miles of the US boundaries. These represent the “lost” tweets referred to in section 3.2 and unfortunately this data was not included with the cancer incidence rate study. Ultimately, these tweets could have been accounted for and added to the state counts manually.

Table 3: Tweet geolocation content. A small percentage of tweets, those with the 'coords' attribute, had geolocation information provided the device (e.g. cell phone).

Health Campaign	'coords'	'location'	'twitter'
BCAM 2015	6,394	596,113	12,783
BCAM 2016	2,667	241,055	7,675
Movember 2015	1,272	94,425	2,885
Movember 2016	555	43,662	1,807

3.3.2 Popular Terms Content

Hashtags present in tweets represent characteristic terms, often convey conviction, and serve to categorize and group similar messages in social media. The campaign tweet data corpora were analyzed for hashtag content. The 50 most frequently occurring hashtags for each campaign are shown in word clouds, where the size of the text represents the frequency in which the hashtag appeared; Figures 4-7 refer. Comparison of the word clouds show that different terms rise to prominence during the campaigns. Table 4 provides a summary. Comparison of the word clouds show that different terms rise to prominence during the campaigns.

A review of BCAM 2015 and 2016 show the following: heavy influence of the Twibbon campaign tool in 2015, professional sports references (NFL, football), fitness events (walks), and various terms associated with breast cancer awareness (pink, breast, BCAM, etc.). The Twibbon tool[7] allows social media users to embellish their profile picture with a symbol showing support for a campaign or cause (e.g. pink ribbon for breast cancer). Specific terms regarding health, science or prevention are sparse

[7] https://twibbon.com/

(#medical research, #physical). Chevrolet's #iDriveFor[8] and variations of this hashtag are prominent in these campaigns. Given that Chevy donates $5 per hashtag to the American Cancer Society, and there were approximately 4000 of these hashtags in BCAM 2015 (a sample of the total tweets for BCAM), this give some idea of the fundraising utility of a simple Twitter post.

The Movember campaigns show the following: many masculine and "bro culture" themes (beards, moustaches, beer, Jeep, No shaving), a different set of professional sports references (FIFA, NHL, HUT), possibly due to Movember's international stature and reach. Regarding health topics, prostate cancer terms and phrases dominated, with little to no mention of the other tenets of Movember, such as testicular cancer, mental health, or physical inactivity.

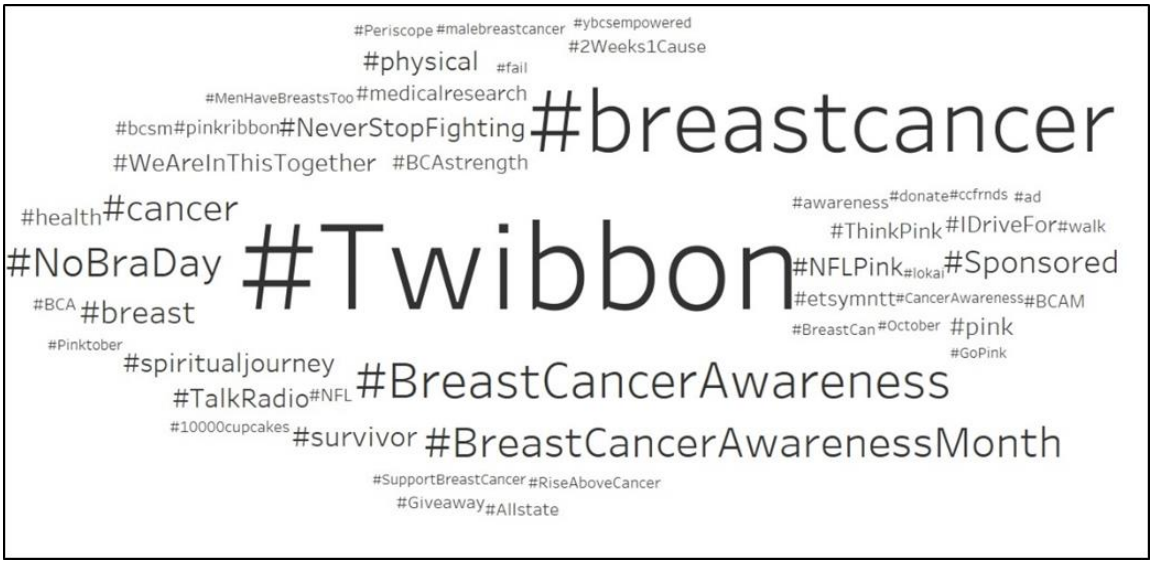

Figure 4: BCAM 2015 top 50 hashtag content.

[8]http://media.chevrolet.com/media/us/en/chevrolet/news.detail.html/content/Pages/news/us/en/2015/oct/1005-chevy-fights-cancer.html

Figure 5: BCAM 2016 top 50 hashtag content.

Figure 6: Movember 2015 top 50 hashtag content.

Figure 7: Movember 2016 top 50 hashtag content.

Table 4: Top ten hashtags for BCAM and Movember 2015 and 2016.

Top Ten Hashtags for BCAM Campaigns			
2015	Ct.	2016	Ct.
#Twibbon	82189	#breastcancer	24126
#breastcancer	40677	#BreastCancerAwarenessMonth	6417
#BreastCancerAwareness	11315	#IDriveFor	5433
#BreastCancerAwarenessMonth	8030	#BreastCancerAwareness	3928
#NoBraDay	7698	#Twibbon	2684
#cancer	6191	#NFLPink	2284
#breast	3900	#cancer	2184
#Sponsored	3790	#breast	1271
#survivor	2847	#SaveTheTatas	1157
#TalkRadio	2644	#NoBraDay	1098
Top Ten Hashtags for Movember Campaigns			
2015	Ct.	2016	Ct.
#Movember	32084	#Movember	12178
#menshealth	1692	#ProstateCancer	2175
#prostatecancer	1625	#menshealth	759
#NoShaveNovember	1407	#cancer	724
#prostate	1139	#InternationalMensDay	641
#Movember2015	866	#NoShaveNovember	605
#health	760	#jeepstache	380
#cancer	691	#Soda	377
#mustache	611	#movember2016	373
#HUT	560	#prostate	262

4. RESULTS

4.1 Introduction

This section presents the results of the spatial distribution analysis of adjusted tweet rates (participation) and cancer incidence. State tweet rates (participation) and incidence data for the states were compared to assess their relationship and find answers to the following research question from Section 1: *Is there a relationship between the number of tweets (tweet rate) and cancer incidence rate?* and hypothesis: "*States with higher breast and prostate cancer incidence rates participate more in online breast and prostate cancer-related Twitter activities.*" Comparisons were made using scatter plots and Pearson correlation.

4.2 Data analysis and visualization

4.2.1 Tweet Rate and Cancer Incidence Graphs

Figure 8-14 are graphs showing the following: state tweet rates and state incidence. These bar graphs and scatter plots allow for comparison of all 50 states and District of Columbia to each other regarding tweet rate, and incidence. The tweet rate for each state was calculated using Equation 1, and represents the number of tweets per 100,000 internet users of each state. Cancer incidence as provided by the State Cancer Profiles website is the number of female breast or prostate cancer cases per number of females or males respectively in the state, and represents the number of females or males diagnosed with cancer per 100,000 population. Additionally, the state cancer incidence

data uses the age-adjusted populations of women and men for breast and prostate cancer in the state respectively.

Figures 8 and 9 show the cancer incidences and state tweet rates to allow for comparison plotted on a bar graph. Cancer incidence is ordered from the state with the highest cancer incidence rate to the lowest, along with the accompanying state tweet rate. There was no cancer incidence data available for Nevada, therefore it is shown as last on both bar graphs. These graphs show how states with the highest cancer incidence rates do not exhibit the highest levels of campaign participation, as the tweet rate bars do not show similarly ordered (greatest to least) patterns.

Differences between the BCAM and Movember campaigns can be seen, with some examples: DC has the highest breast and prostate cancer incidence and participates the most in both campaigns. Louisiana has the second highest incidence rate of prostate cancer, and some of the lowest participation in the Movember campaign.

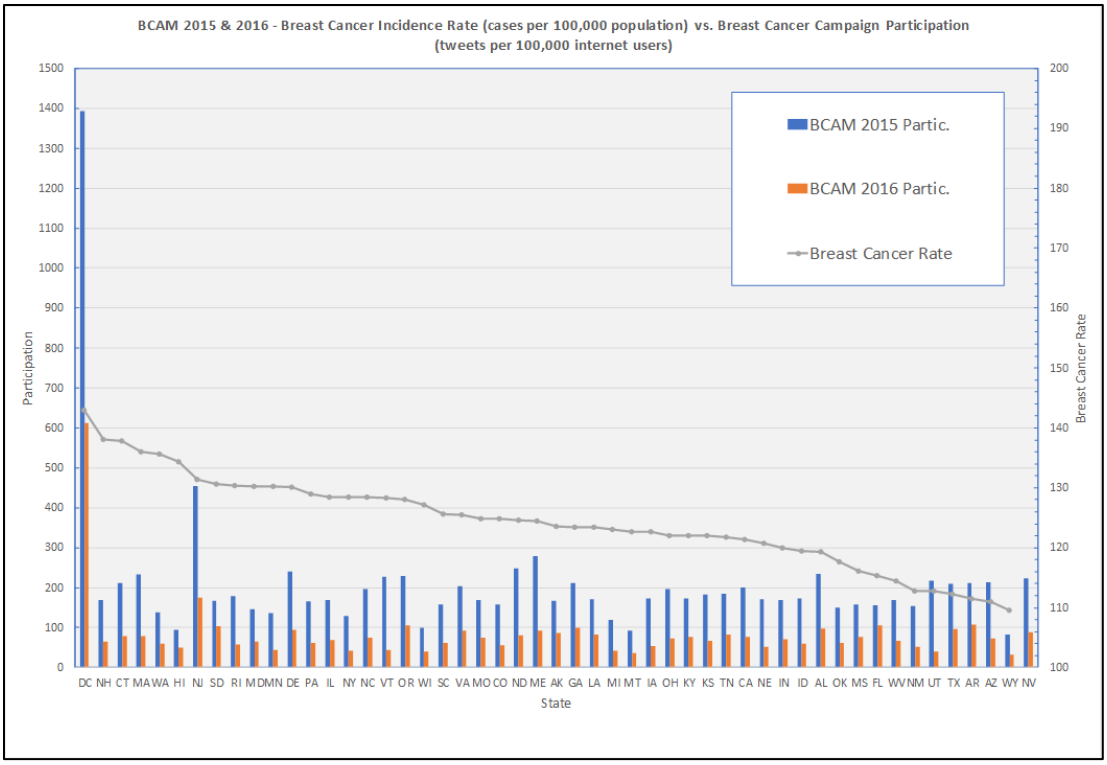

Figure 8: BCAM 2015 & 2016 Breast Cancer Incidence Rate and BCAM Campaign Participation.

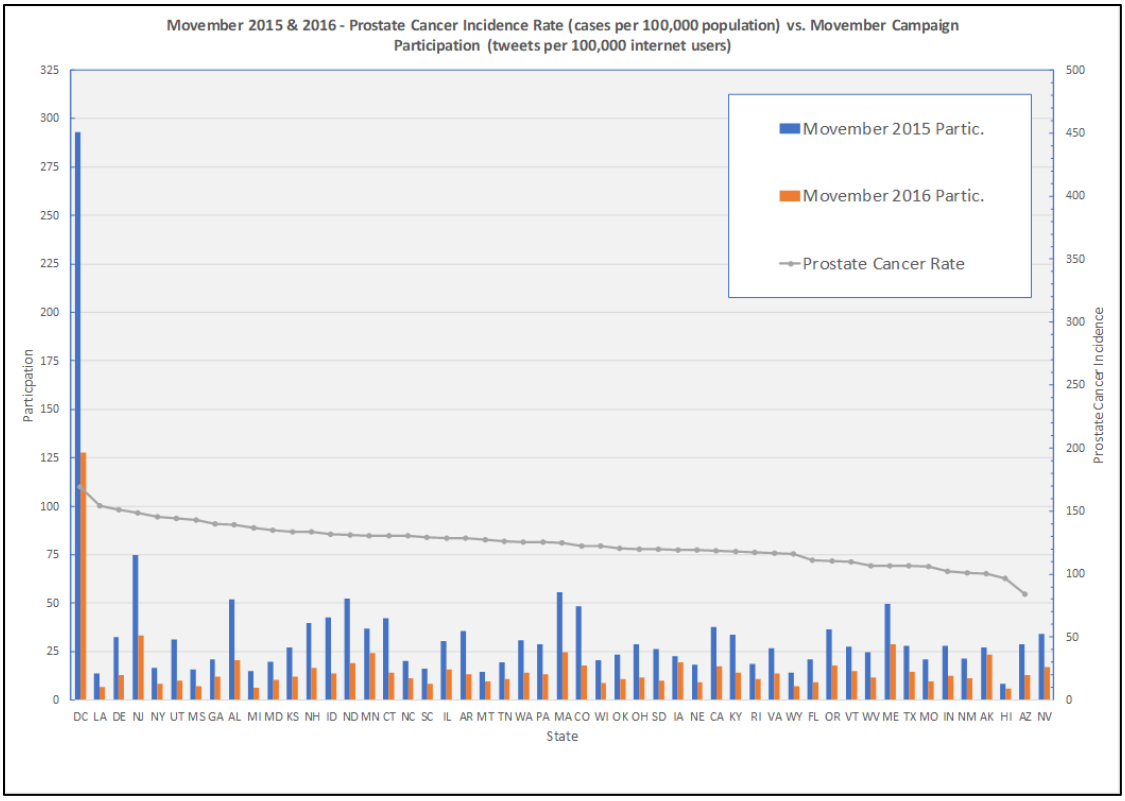

Figure 9: Movember 2015 & 2016 Prostate Cancer Incidence Rate and Movember Campaign Participation.

4.2.2 Scatter Plots

As shown in Section 3.21. and 4.2.1, DC and NJ appeared significantly different from other states in terms of participation. To account for this, the states' participation numbers were assessed for the presence of mild and extreme outliers using a basic "boxplots with fences" method[9]. DC and NJ were found as outliers in all four campaigns, and were removed prior to producing the Pearson correlation scatterplots. Figure 10 shows an example of the skewing effects of DC and NJ before removal

Figure 11-14 are scatter plots of the following: state tweet rate (participation) vs. state incidence rate. These plots were used to measure the relationships and measure correlation between state tweet rate variables and cancer incidence variables. Each scatter

[9] http://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm

plot shows two variables, a histogram for each variable, a regression line, and the Pearson correlation coefficient for each plot. The Pearson correlation coefficient ranges from -1 to 1, and describes the relationship between x and y variables. A positive 1 would indicate a perfect linear relationship between x and y, where y increases as x increases in value. A value of zero would indicate no relationship between the two variables, and value of -1 would indicate a perfect linear relationship between x and y where the y variable decreases as x increases.

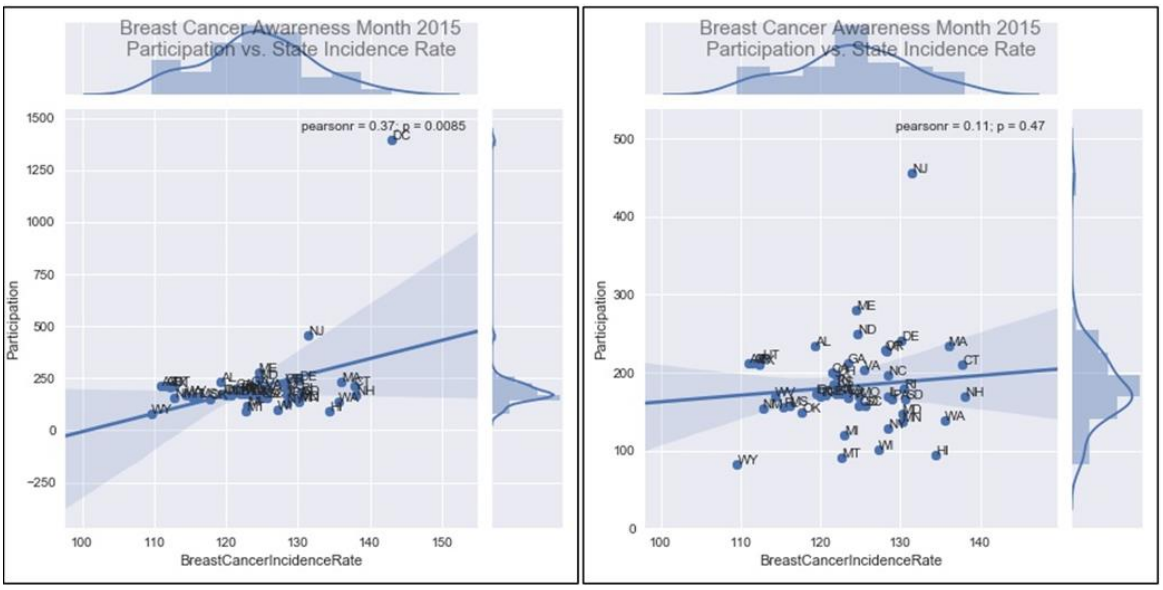

Figure 10: Skewing effects of DC and NJ data points example. These data points were removed prior to correlation.

Figure 11 and 12 shows BCAM 2015 and BCAM 2016. The plots look similar to each other, and the Pearson correlation coefficient is 0 and -0.044 respectively, which indicates no linear correlation between the two variables. For BCAM 2015, the histogram on the top of plot shows that most state cancer incidence rates range from 115-135 cases per 100,000 population. The histogram on the right side of the graph shows that most

state participation ranges from 125-225 tweets per 100,000 internet users. For BCAM 2016, the histogram on the top of plot is the same as it should be, and the histogram on the right side of the graph shows 50-90 tweets per 100,000 internet users.

Figure 13 and 14 shows Movember 2015 and Movember 2016. The plots also look similar to each other, and the Pearson correlation coefficient is 0.044 and -0.15 respectively, which indicates no linear correlation between the two variables. For Movember 2015, the histogram on the top of plot shows that most state cancer incidence rates range from 110-140 cases per 100,000 population. The histogram on the right side of the graph shows that most state participation rates range from 15-35 tweets per 100,000 internet users. For Movember 2016, the histogram on the top of plot is the same as it should be, and the histogram on the right side of the graph shows that state participation rates range from 15-35 tweets per 100,000 internet users.

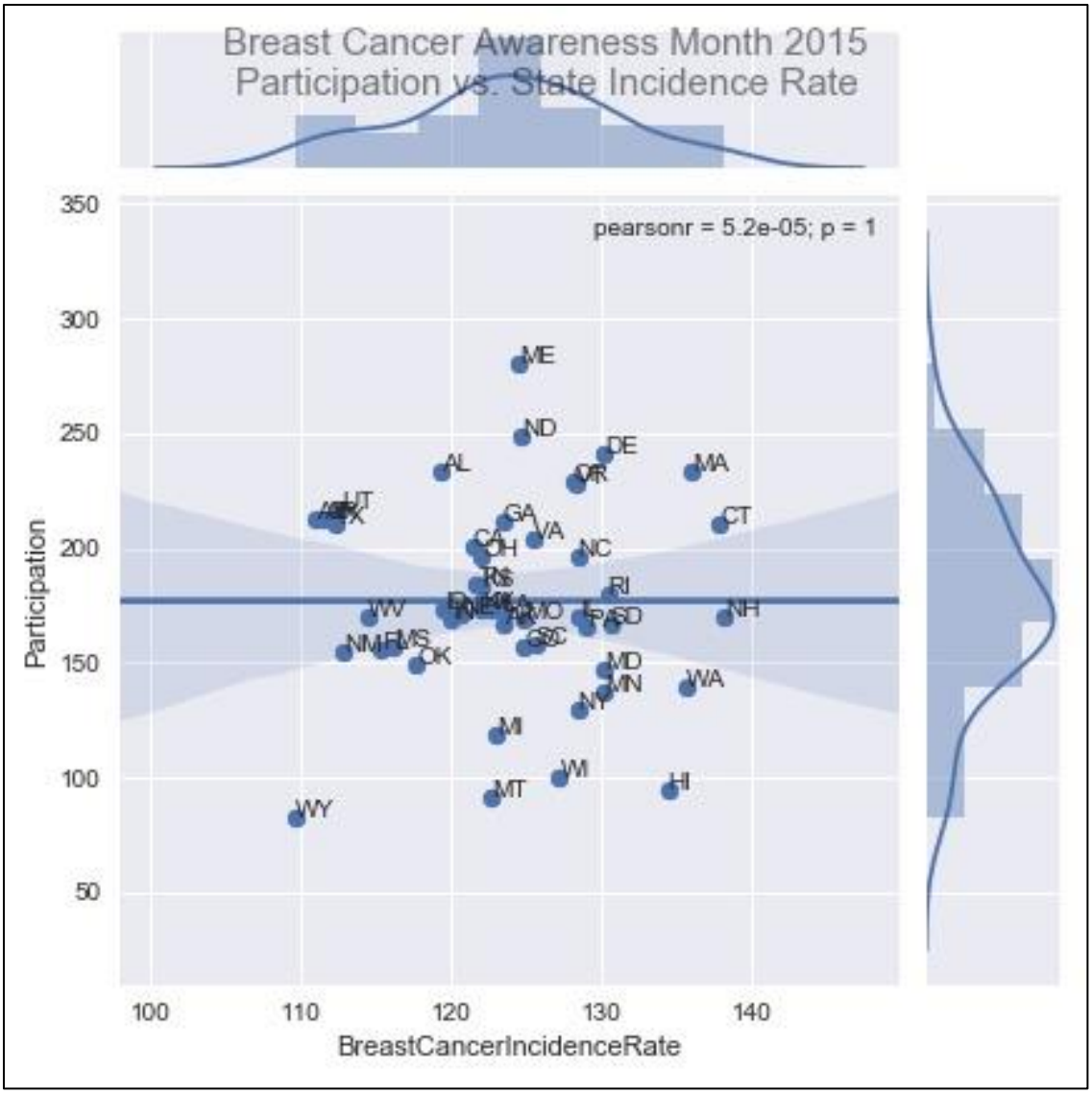

Figure 11: BCAM 2015 Campaign Participation vs. Breast Cancer Incidence Rate.

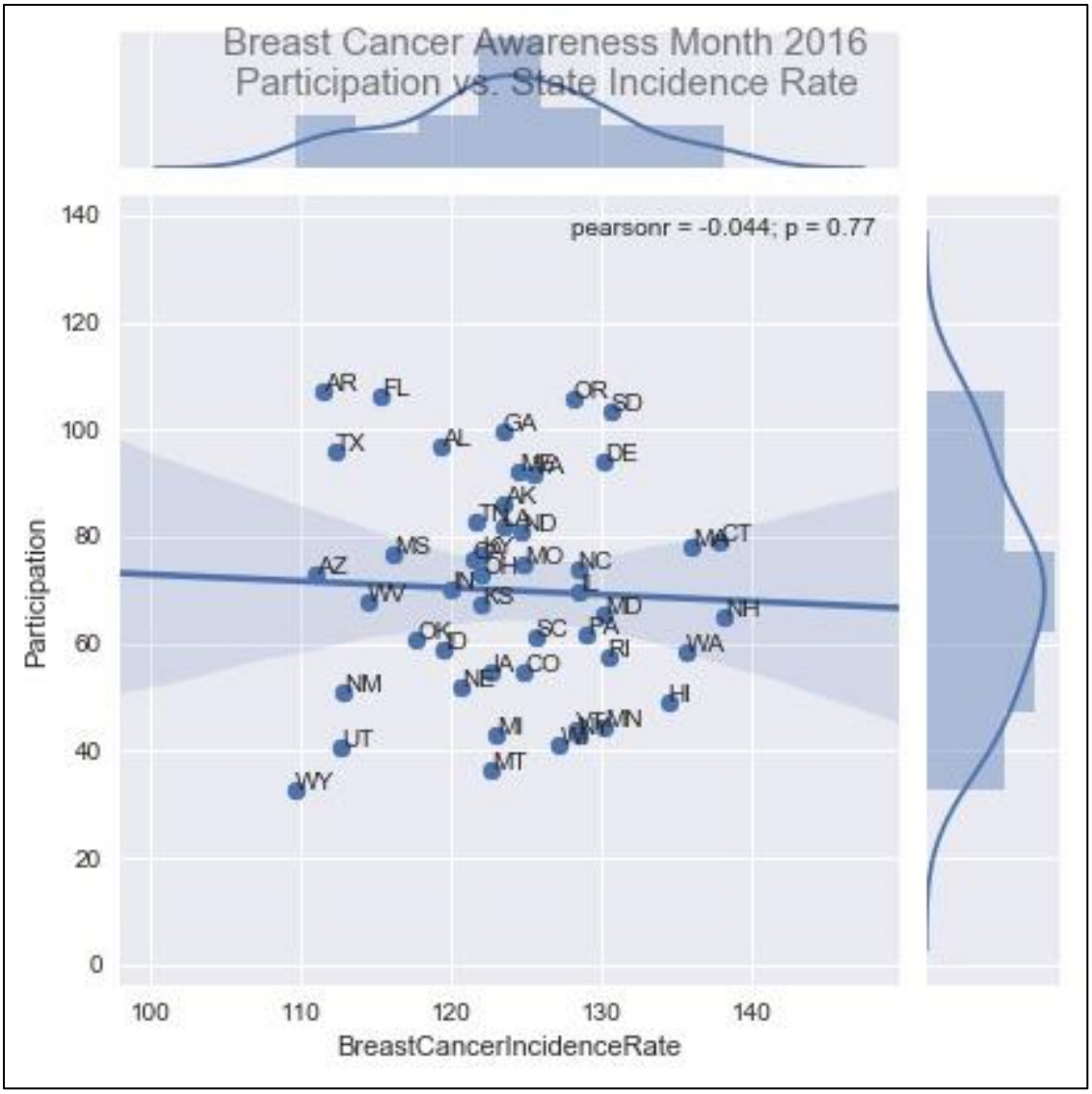

Figure 12: BCAM 2016 Campaign Participation vs. Breast Cancer Incidence Rate.

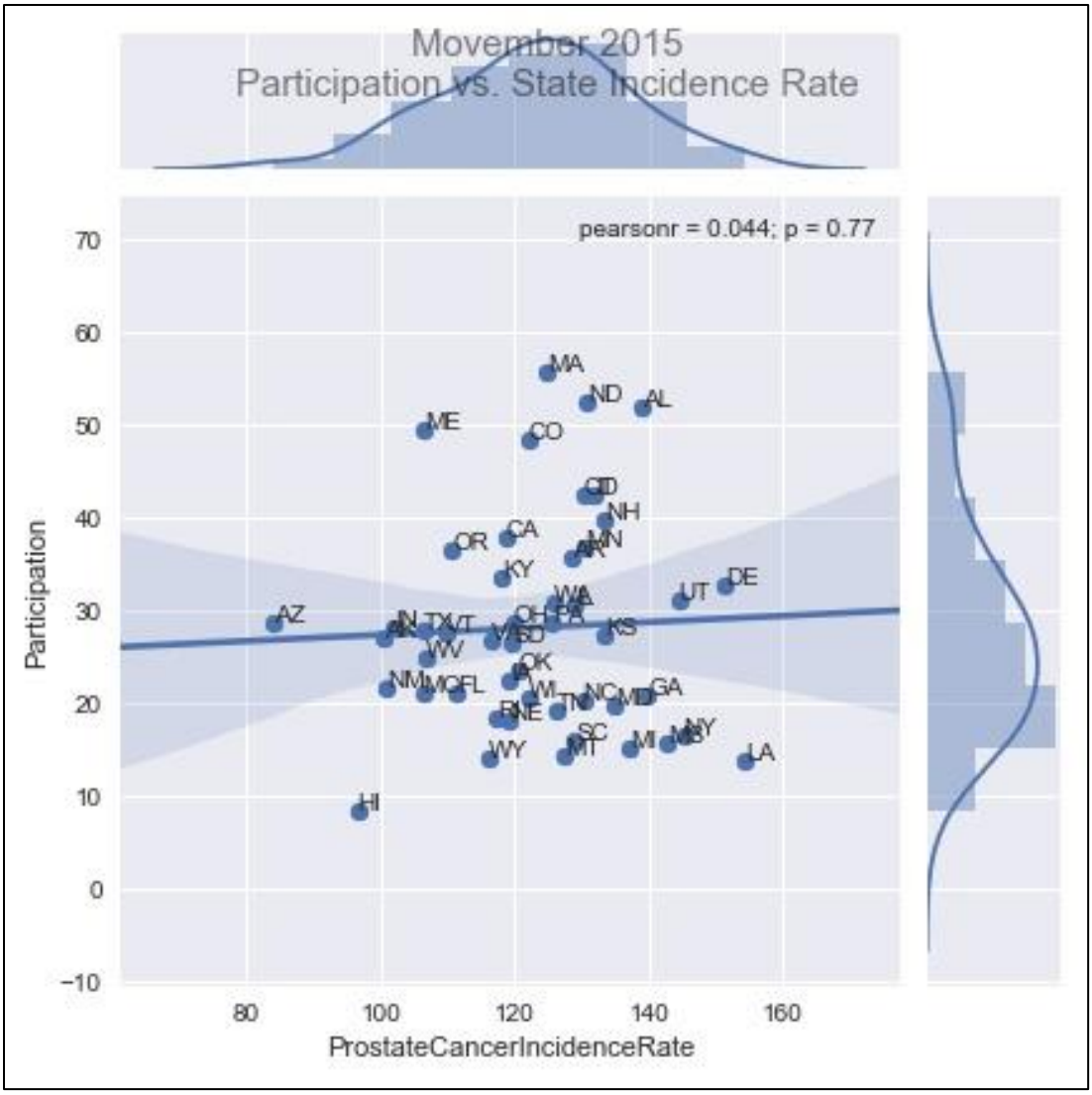

Figure 13: Movember 2015 Campaign Participation vs. Prostate Cancer Incidence Rate.

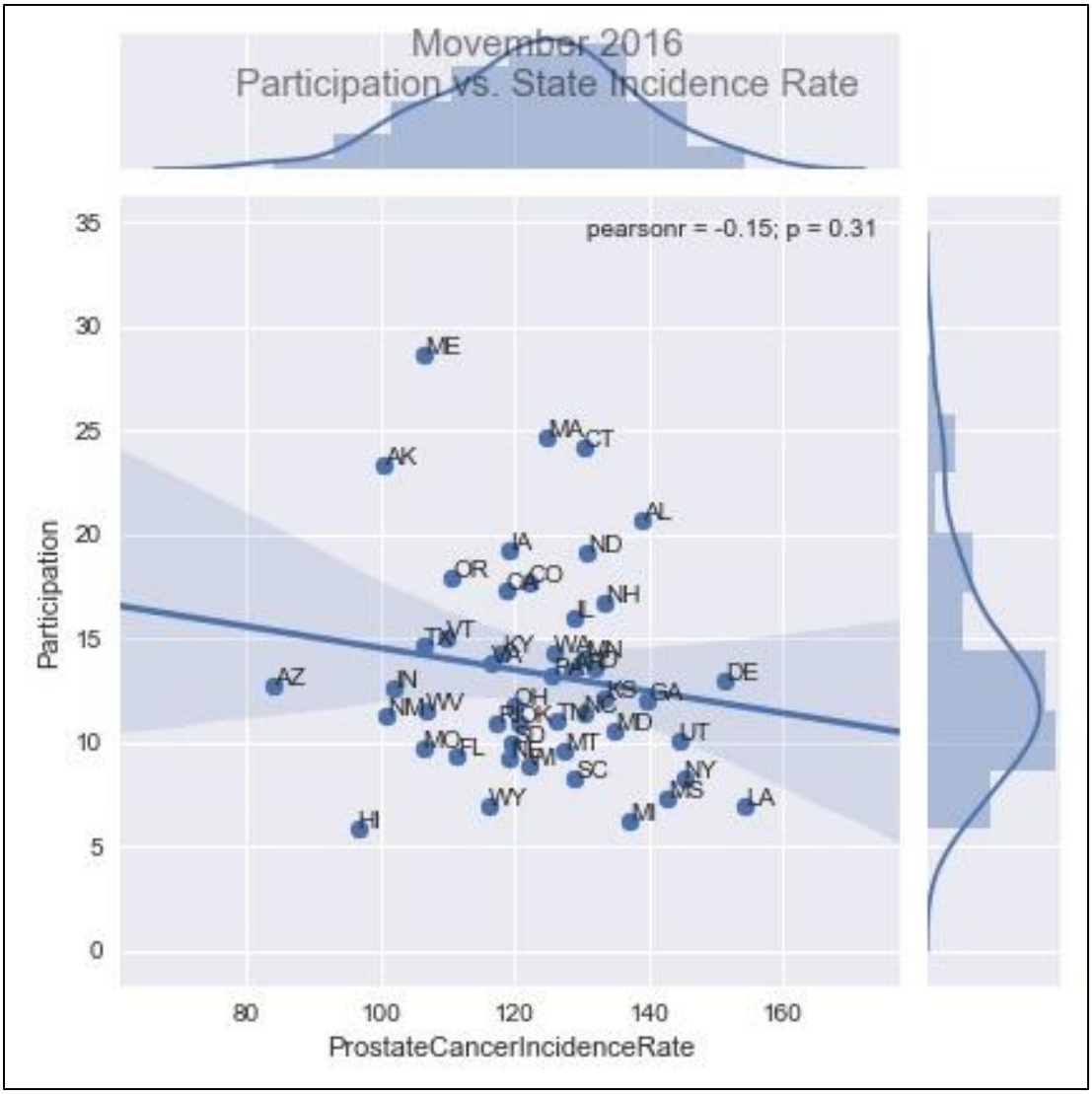

Figure 14: Movember 2016 Campaign Participation vs. Prostate Cancer Incidence Rate.

Figures 15-18 are choropleth maps for the four campaigns showing participation. These choropleth maps were created using natural breaks (Jenks) classification, by classing the data into five classes; three to classify the bulk of the data, and a separate class each for the minor and extreme outliers in each set [43].

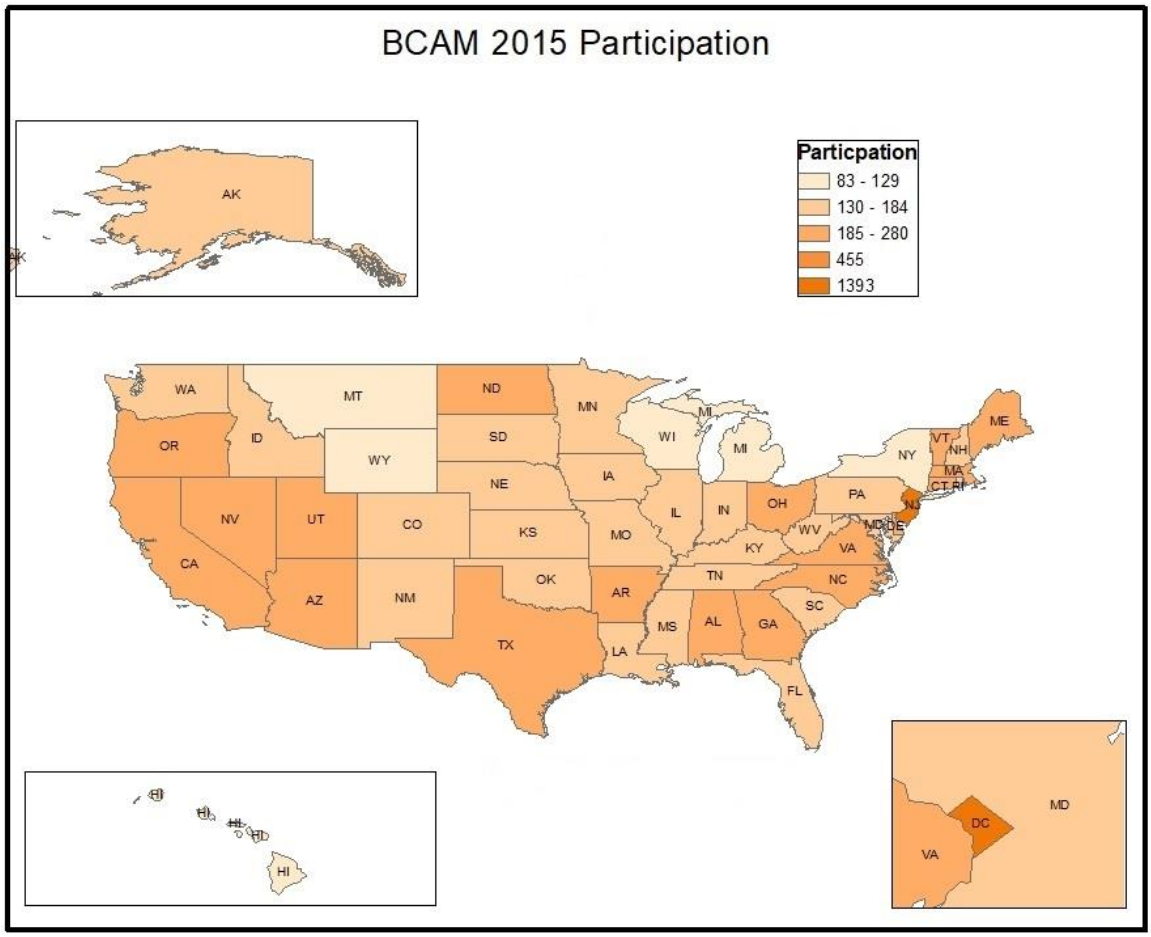

Figure 15: BCAM 2015 Participation Map.

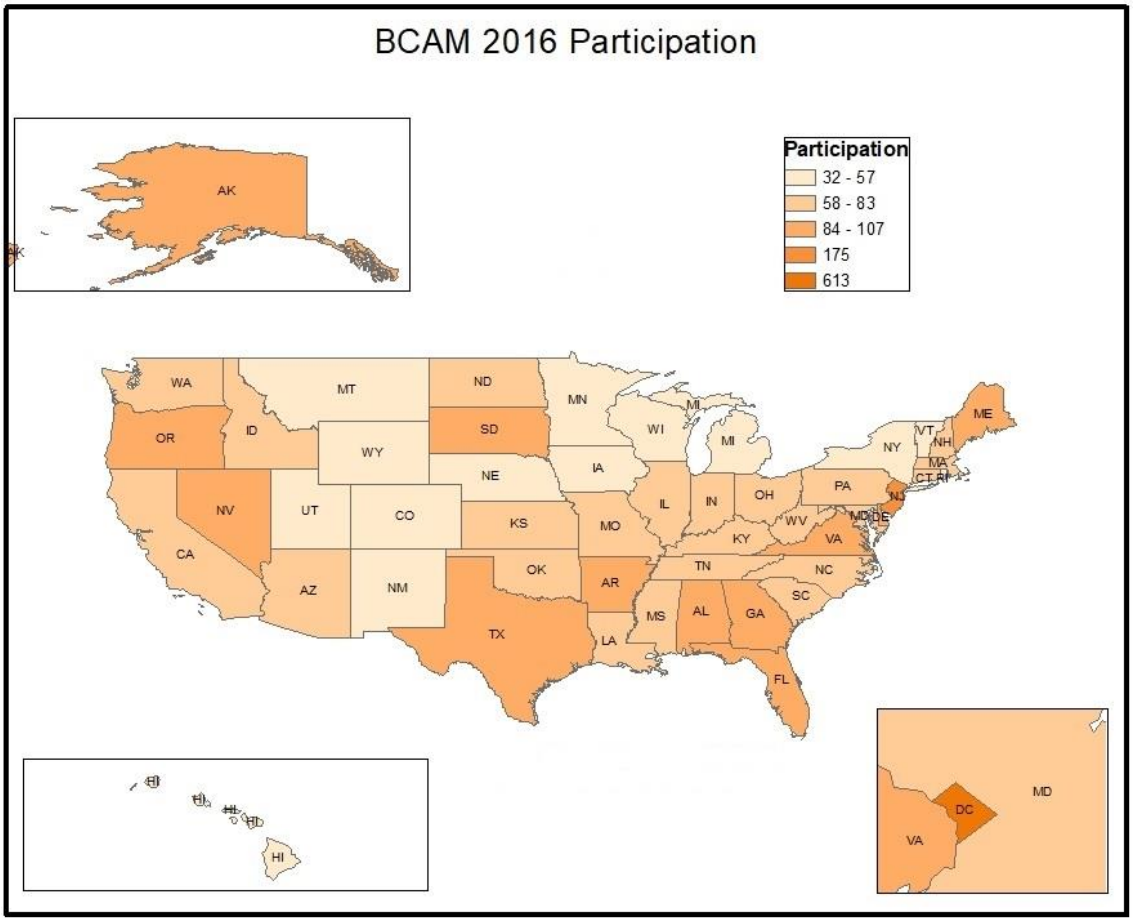

Figure 16: BCAM 2016 Participation Map.

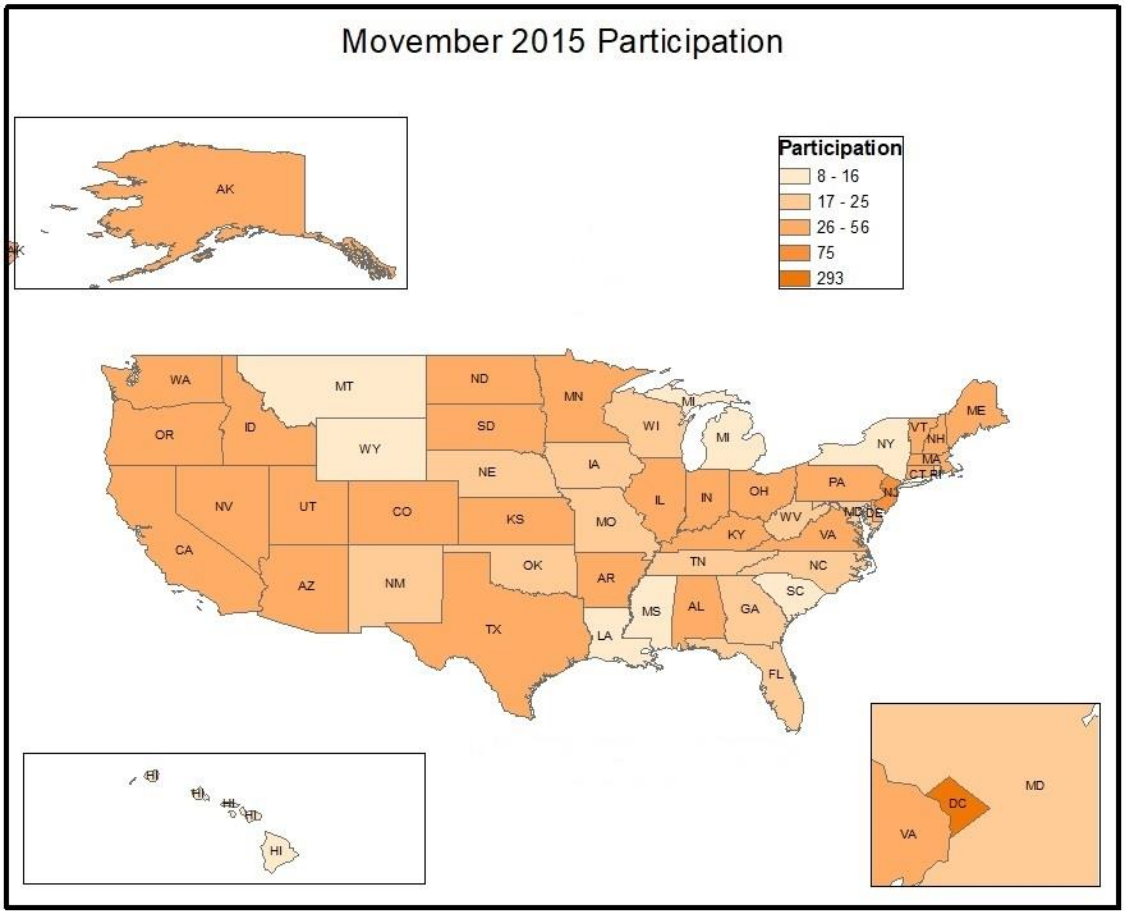

Figure 17: Movember 2015 Participation Map.

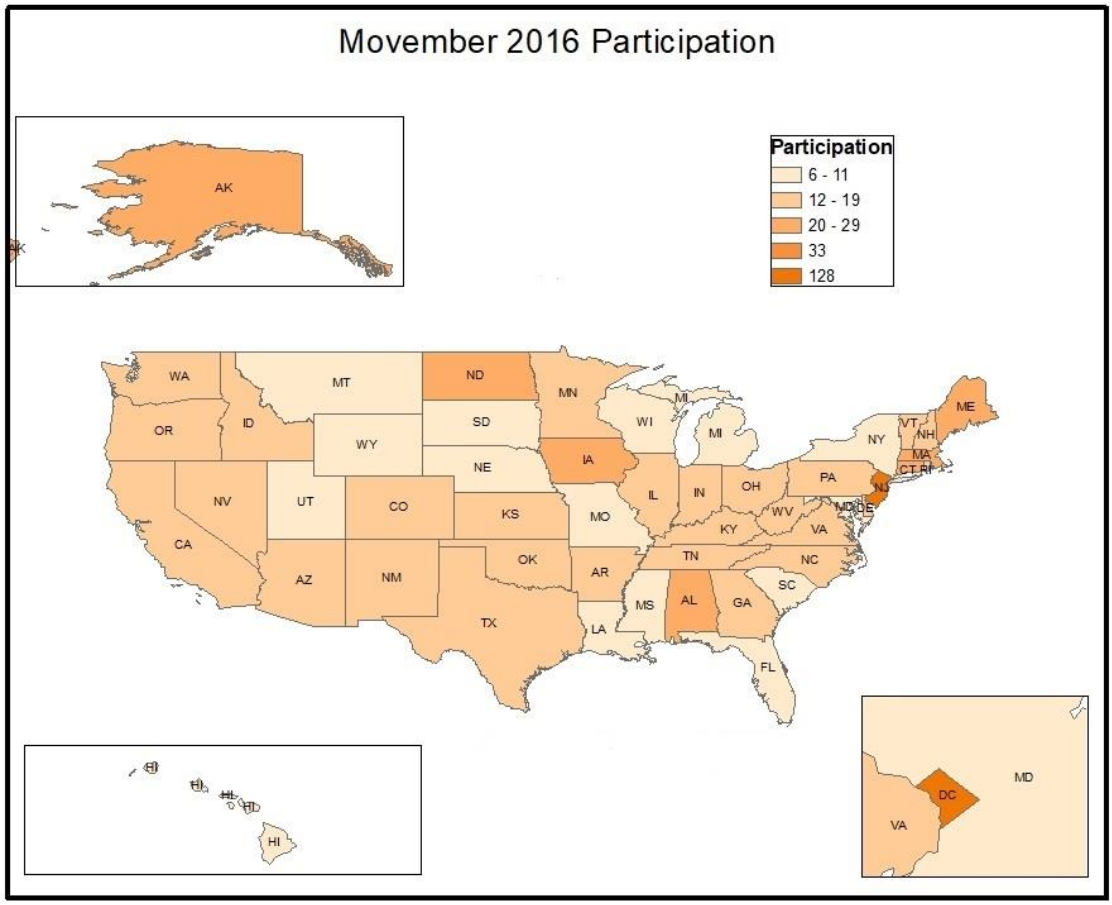

Figure 18: Movember 2016 Participation Map.

5. SUMMARY OF FINDINGS, CONCLUSIONS, AND OUTLOOK

5.1 Summary of Findings

The main focus of this study was to assess the spatial distribution of breast cancer and Movember tweets across the United States, and compare the associated tweet rates (participation) to cancer incidence. The spatial distribution of geolocated tweets across the United States was determined by counting how many tweets came from each state, and is shown in Figure 2. Adjusted tweet rate, also referred to herein as participation was determined in accordance with equation 1 of Section 3.2.1. and was shown in Figure 3. Tweet rate was compared with cancer incidence to look for a correlation between the two, as shown in the scatter plots of Figures 11-14. The scatterplots and Pearson correlation show there is really no correlation between adjusted tweet rate (participation) of the BCAM and Movember campaigns in this study and the breast and prostate cancer incidence data.

As shown in Figure 19, when plotted against population data, the tweet counts are closely related to state population, which implies that campaign participation is heavily dependent on population size; the larger the population, the more internet users (a subset and large percentage of state population) and people using Twitter there are likely to be.

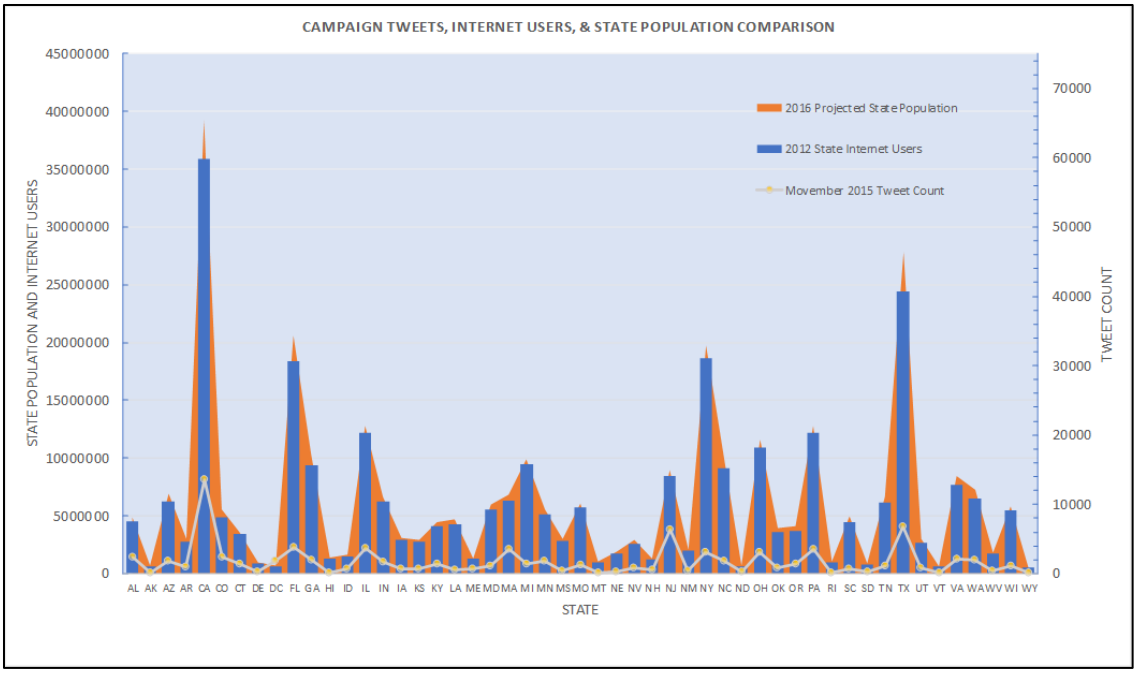

Figure 19: State population, Internet users, campaign tweet relationships.

Although not shown here, correlation between state population and cancer incidence was assessed and there was almost no correlation between the two. The campaign participation variable is heavily influenced by population, and the cancer incidence variable is not.

Given these results, the hypothesis of: "States with higher breast and prostate cancer incidence rates participate more in online breast and prostate cancer-related Twitter activities" is not true or inconclusive at best, as there was no discernible trend observed where states with high levels of participation also had high cancer incidence rates as well.

5.2 Discussion/ Considerations for future study

In this study, assumptions were made to keep its scope manageable. Each tweet was treated as though it came from a different user; that is, the same person was not tweeting multiple times, and that tweet content was similar between all states and generally related to breast and prostate cancer and the campaigns.

As discussed in [35] and [36], it is known that men and women use social media differently, and may account for the difference in tweet volume and participation for the campaigns, as breast cancer affects women and prostate cancer affects men.

The Twitter data files used for this study were a subset of a larger file of worldwide tweets. That is, the data was not collected from the Twitter API based on the specification of a bounding box as might be frequently be done if interested in a specific geographic area. The sorting of tweets from a larger file of collected data into those which were geolocated in the states was more complex than anticipated for two reasons: not all records had information in the state field, meaning they could not be sorted based on that field, and performing a spatial join of the tweets with GIS polygons had its limitations as discussed in sections 3.2 and 3.3. If deemed necessary to have every available geolocated tweet, a better way to sort out the tweets with geographic coordinates in the United States would be to specifying some type of bounding box or range of coordinate values when reading and searching the files, or manually (using GIS) locate the tweets which fell outside the polygons and update the state tweet counts.

During the content analysis phase of this project, all available tweets were used in the analysis, and the tweet text field content was overlooked. That is, the data was not cleaned to remove duplicates or other unwanted tweets, with the exception of tweets

located at the geographic center of the United States by default. It was assumed that while not all tweets contain personal discussions or expressions of being impacted by breast or prostate cancer, the aggregation of tens and hundreds of thousands of collected tweets still represented active participation at some level in the BCAM and Movember campaigns. Cleaning the data to remove tweets which did not add value to assessment of the campaign narratives would characterize these campaigns more effectively and represent the tweet counts for each state differently.

This thesis looked at the relationship between participation and incidence as a whole; all 51 states as a large group. If the states are looked at as smaller groups where correlation is found, different relationships may be found.

The State Cancer Profiles incidence data uses age-adjusted cancer rates[10], where a weight factor is determined for each age group based on population. The tweet rate (participation) variable used in this study was not adjusted in a similar manner, as data to support such adjustment was not available. The tweet rate calculation is more similar to a crude cancer incidence rate calculation[11].

Only one cancer health narrative associated with Movember cancer was studied in this thesis; prostate cancer. It may be interesting to compare mental health, testicular cancer, and physical inactivity data with the campaign Twitter data as well. Additionally, Twitter data specific to the Movember campaign month was used; prostate cancer has its own dedicated campaign month (September) from which Twitter data could be collected.

[10] https://seer.cancer.gov/seerstat/tutorials/aarates/step1.html
[11] https://seer.cancer.gov/seerstat/tutorials/aarates/step2.html

The results of this study showed no correlation between campaign participation and cancer incidence in Twitter; different results may be found in other social media platforms. Furthermore, using cancer mortality rates instead of cancer incidence rates could yield different results and relationships with campaign participation.

5.3 Conclusions

In summary, this paper presented an example of how geosocial analysis can be used to understand the behavior of BCAM and Movember health campaigns, as public discussion can be collected from Twitter, mapped, and compared to ground truth data in the form of breast cancer and prostate cancer incidence for the states. It can add to an evolving body of research regarding social media and health campaigns. People may discuss cancer for many reasons, maybe because they have been diagnosed, because of a family member or friend, or simply out of empathy for others. Studying health campaigns[12] as they are discussed and shared in a public forum like Twitter may provide information about their effectiveness, and about the population affected by disease and other health issues.

[12] http://www.healthline.com/health/directory-awareness-months

REFERENCES

[1] M. Paul and M. Drezde. 2011. You Are What You Tweet: Analyzing Twitter for Public Health. *Icwsm* 20 (2011), 265-272.

[2] J. Radzikowski, H. Hollen, and S. Fuhrman. 2015. Using Twitter Content to Crowdsource Opinions on Tanning in the United States. *Kartographisce Nachtrichten*, 3 (2015), 131-138.

[3] P. Chai, R. Rosen, D. Lewis, M. Ranney, and E. Boyer. 2017. Crowd-Sourced Focus Groups on Twitter: 140 Characters of Research Insight. *Proceedings of the 50th Hawaii International Conference on System Sciences* (2017).

[4] L. Sinnenberg, A. M. Buttenheim, K. Padrez, C. Mancheno, L. Ungar, and R. M. Merchant. 2017. Twitter as a Tool for Health Research: A Systematic Review. *American Journal of Public Health* 107, 1 (2017), 143–143.

[5] F. J. Grajales III, S. Sheps, K. Ho, H. Novak-Lauscher, and G. Eysenbach. 2014. Social Media: A Review and Tutorial of Applications in Medicine and Health Care. *Journal of Medical Internet Research* 16, 2 (2014).

[6] K. Rabarison et al. 2017. Measuring Audience Engagement for Public Health Twitter Chats: Insights From #LiveFitNOLA. *JMIR Public Health and Surveillance* 3, 2 (August 2017).

[7] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley, "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose," CoRR, vol. abs/1306.5204, (2013).

[8] "Twitter Development Documentation." [Online]. Available: https://dev.twitter.com/overview/api/tweets. [Accessed: 29-Jul-2017].

[9] A. Croitoru, N. Wayant, A. Crooks, J. Radzikowski, and A. Stefanidis. 2015. Linking cyber and physical spaces through community detection and clustering in social media feeds. *Computers, Environment and Urban Systems* 53 (2015), 47–64.

[10] X. Lu, A. Croitoru, J. Radzikowski, A. Crooks, and A. Stefanidis. 2013. Comparing the spatial characteristics of corresponding cyber and physical communities. Proceedings of the 6th ACM SIGSPATIAL *International Workshop on Location-Based Social Networks* – LBSN '13 (2013).

[11] A. Koskan et al. 2014. Use and Taxonomy of Social Media in Cancer-Related Research: A Systematic Review. *American Journal of Public Health* 104, 7 (2014).

[12] A. Mislove, S. Lehmann, Y. Y. Ahn, J.P. Onnela, and J. N. Rosenquist. 2011. Understanding the Demographics of Twitter Users. *ICWSM, 11*, 5th.

[13] A. Signorini, A. Maria Segre, and P. M. Polgreen. 2011. The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. *PloS ONE* 6, 5 (April 2011).

[14] L. Sloan et al. 2013. Knowing the tweeters: Deriving Sociologically Relevant Demographics from Twitter. *Sociological Research Online* 18, 3 (2013).
[15] L. Sloan, J. Morgan, P. Burnap, and M. Williams. 2015. Who tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data. *Plos One* 10, 3 (February 2015).
[16] E. Mohammady and A. Culotta. 2014. Using County Demographics to Infer Attributes of Twitter Users. *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media* (2014).
[17] Pew Research Center (2016). Washington, DC: Pew Research Center's Internet & American Life Project. Social Media Update 2016.
[18] J. Radzikowski, A. Stefanidis, K. H. Jacobsen, A. Croitoru, A. Crooks, and P. L. Delamater. 2016. The Measles Vaccination Narrative in Twitter: A Quantitative Analysis. *JMIR Public Health and Surveillance* 2, 1 (April 2016).
[19] J. Bender, A. Cyr, L. Arbuckle, and L. Ferris. 2017. Ethics and Privacy Implications of Using the Internet and Social Media to Recruit Participants for Health Research: A Privacy-by-Design Framework for Online Recruitment. *Journal of Medical Internet Research* 19, 4 (June 2017).
[20] E. Velasco, T. Agheneza, K. Denecke, G. Kirchner, and T. Eckmanns. 2014. Social Media and Internet-Based Data in Global Systems for Public Health Surveillance: A Systematic Review. *Milbank Quarterly* 92, 1 (2014), 7–33.
[21] David M. Hartley. 2014. Using Social Media and Internet Data for Public Health Surveillance: The Importance of Talking. *Milbank Quarterly* 92, 1 (2014), 34–39.
[22] K. H. Jacobsen et al. 2016. Lessons from the Ebola Outbreak: Action Items for Emerging Infectious Disease Preparedness and Response. *EcoHealth* 13, 1 (2016), 200–212.
[23] N. Berry, F. Lobban, M. Belousov, R. Emsley, G. Nenadic, and S. Bucci. 2017. #WhyWeTweetMH: Understanding Why People Use Twitter to Discuss Mental Health Problems. *Journal of Medical Internet Research* 19, 4 (May 2017).
[24] L. Laestadius and M. Wahl. 2017. Mobilizing social media users to become advertisers: Corporate hashtag campaigns as a public health concern. *Digital Health* 3 (2017),
[25] J. Young. 2016. Facebook, Twitter, and Blogs: The Adoption and Utilization of Social Media in Nonprofit Human Service Organizations. *Human Service Organizations: Management, Leadership & Governance* 41, 1 (2016), 44–57.
[26] P. Diddi and L. Lundy. 2017. Organizational Twitter Use: Content Analysis of Tweets during Breast Cancer Awareness Month. *Journal of Health Communication* 22, 3 (2017), 243–253.
[27] M. Moran et al. 2017. Why Peer Crowds Matter: Incorporating Youth Subcultures and Values in Health Education Campaigns. *American Journal of Public Health* 107, 3 (2017), 389–395.
[28] M. R. Wehner et al. 2014. Twitter: an opportunity for public health campaigns. *The Lancet* 384, 9938 (2014), 131–132.
[29] R. Bannor, A. K. Asare, and J. N. Bawole. 2017. Effectiveness of social media for communicating health messages in Ghana. *Health Education* 117, 4 (May 2017),

342–371.
[30] A. Gough et al. 2017. Tweet for Behavior Change: Using Social Media for the Dissemination of Public Health Messages. *JMIR Public Health and Surveillance* 3, 1 (2017).
[31] R. Thackeray, S. H. Burton, C. Giraud-Carrier, S. Rollins, and C. R. Draper. 2013. Using Twitter for breast cancer prevention: an analysis of breast cancer awareness month. *BMC Cancer* 13, 1 (2013).
[32] C. Bravo and L. Hoffman-Goetz. 2015. Tweeting About Prostate and Testicular Cancers: What Are Individuals Saying in Their Discussions About the 2013 Movember Canada Campaign? *Journal of Cancer Education* 31, 3 (2015), 559–566.
[33] C.A. Bravo and L. Hoffman-Goetz. 2015. Social Media and Men's Health: A Content Analysis of Twitter Conversations During the 2013 Movember Campaigns in the United States, Canada, and the United Kingdom. *American Journal of Men's Health* (2015).
[34] N. Prasetyo, C. Hauff, D. Nguyen, T. Van Den Broek, and D. Hiemstra. 2015. On the Impact of Twitter-based Health Campaigns: A Cross-Country Analysis of Movember. *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis* (2015).
[35] I. Himelboim and J. Han. 2013. Cancer Talk on Twitter: Community Structure and Information Sources in Breast and Prostate Cancer Social Networks. *Journal of Health Communication* 19, 2 (October 2013), 210–225.
[36] S. Loeb et al. 2017. Tweet this: how advocacy for breast and prostate cancers stacks up on social media. *BJU International* (July 2017).
[37] C. Weeg, H. Schwartz, S. Hill, R. M. Merchant, C. Arango, and L. Ungar. 2015. Using Twitter to Measure Public Discussion of Diseases: A Case Study. *JMIR Public Health and Surveillance* 1, 1 (2015).
[38] D. Murthy and M. Eldredge. 2016. Who tweets about cancer? An analysis of cancer-related tweets in the USA. *Digital Health* 2 (2016).
[39] State Cancer Profiles. [Online]. Available: https://statecancerprofiles.cancer.gov/. [Accessed: 24-April-2017].
[40] A. Croitoru, A. Crooks, J. Radzikowski, and A. Stefanidis. 2013. Geosocial gauge: a system prototype for knowledge discovery from social media. *International Journal of Geographical Information Science* 27, 12 (2013), 2483–2508.
[41] A. Stefanidis, A. Crooks, and J. Radzikowski. 2011. Harvesting ambient geospatial information from social media feeds. *GeoJournal* 78, 2 (April 2011), 319–338.
[42] U.S. Census Bureau, Current Population Survey, October 2012. Table 2. Reported Internet Usage for Individuals 3 Years and Older, by State: 2012. [Online]. Available: https://www.census.gov/topics/population/computer-internet/data/tables.html. [Accessed: 24-April-2017]
[43] Cynthia A. Brewer. 2006. Basic Mapping Principles for Visualizing Cancer Data Using Geographic Information Systems (GIS). *American Journal of Preventive Medicine* 30, 2 (2006).

BIOGRAPHY

David A. Novak Jr. received his Bachelor of Science from Old Dominion University in 2004.