7-31-2014

Ambient Geographic Information and Risk Mapping

Todd Barr

Follow this and additional works at: https://digitalcommons.du.edu/geog_ms_capstone

Part of the Geographic Information Sciences Commons, and the Human Geography Commons

Recommended Citation

Barr, Todd, "Ambient Geographic Information and Risk Mapping" (2014). Geography & the Environment Graduate Student Capstones. 28.
https://digitalcommons.du.edu/geog_ms_capstone/28

This Capstone Project is brought to you for free and open access by the Department of Geography & the Environment at Digital Commons @ DU. It has been accepted for inclusion in Geography & the Environment Graduate Student Capstones by an authorized administrator of Digital Commons @ DU. For more information, please contact jennifer.cox@du.edu, dig-commons@du.edu.
Ambient Geographic Information and Risk Mapping

Todd Barr

University of Denver Department of Geography and the Environment

Capstone Project

For

Master of Science in Geographic Information Science

July 31, 2014

GEOG 4993
Abstract

Research on social media as an information source for emergency managers and epidemiologists is expanding at a rapid pace. When combined with the spatial predictive analysis methodology known as Risk Terrain Mapping and the capability of the local health agencies to respond, the predictions of outbreaks and their affects can be modeled. This study attempts to create a spatial Risk Terrain Model outbreaks of influenza, and influenza like illnesses within the continental United States. The Risk Terrain Model will be built upon accepted epidemiological models from the Centers for Disease Control and Prevention, as well as Google’s Flu index and lastly the Robert Wood Johnson’s foundation study of the county level health agencies.
Dedication

I dedicate this work to my daughter, who has taught me patience, compassion and who has reminded me of why we learn.
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATA CLEANING</td>
<td>39</td>
</tr>
<tr>
<td>GEOCODING</td>
<td>40</td>
</tr>
<tr>
<td>Methodology</td>
<td>43</td>
</tr>
<tr>
<td>Results</td>
<td>54</td>
</tr>
<tr>
<td>IRTM</td>
<td>54</td>
</tr>
<tr>
<td>Social Media Data</td>
<td>56</td>
</tr>
<tr>
<td>Fixing the IRTM</td>
<td>57</td>
</tr>
<tr>
<td>Conclusion</td>
<td>60</td>
</tr>
<tr>
<td>References</td>
<td>62</td>
</tr>
<tr>
<td>Appendix One: Tweet Modifiers</td>
<td>65</td>
</tr>
<tr>
<td>Appendix Two: Risk Groups</td>
<td>67</td>
</tr>
</tbody>
</table>
The past few years have seen the proliferation of Social Media, mobile technologies and user generated content. Humans now generate and share more information and data than ever before, and Twitter with its 140 character limited format (tweets), is at the forefront of this movement. Most Social Media sites, including Twitter, are accessed via mobile devices and use HTML 5 location technology to allow users to share their location. Searching and harvesting shared information from Social Media frameworks allows people to find the best sushi restaurant or where people are gathering for a social "tweet-up" or demonstration.

However, an additional utility has been found in this platform. During events or incidents, geo-tagged tweets and other Social Media posts have provided insight that emergency managers and first responders never had before. These "citizens as sensors," allow for real time information and updates during and after an event, providing emergency professionals a nearly up to the minute information about the situation in the affected area.

This was first seen on a wide scale during the Haitian Earthquake in 2010. Haitian cellular towers remained mostly intact, making mobile phones the only reliable form of communication. This allowed Twitter users tweeting about current conditions and potential hazardous situations to assist emergency responders to find people, reconnect families and save lives.
Hurricane Sandy in October of 2012 generated more than 20 million tweets. Twitter was the primary source of information about the explosion of the Con Edison substation in the East Village, which knocked out most of the power to lower Manhattan. When false information was transmitted, it was quickly debunked by other users.

Since 2010, Twitter feeds have become a source of "on the spot" information during both large-scale and local disasters. Twitter is now often cited in newscasts, situational awareness briefings in Emergency Management Centers, and by researchers looking to determine public activity and interaction.

Risk Terrain Mapping (RTM) is a geographic statistical methodology that uses GIS to determine which areas are prone to certain events. RTM is used extensively in law enforcement for "Intelligence Led Policing" to assist in determining which crimes occur in what areas. The RTM methodology is now expanding into other areas, including Public Health and Humanitarian relief.

Within this report an RTM model will be created to assist in the prediction of an influenza outbreak. Once that model is created, both CDC data collected during the study’s timeframe, and social media data will be compared to the RTM map. Through the use of these two data sources will
determine the accuracy of the RTM, and if social media and an RTM can be used as a Bio surveillance tool.

**GIS and Event Detection and Response**

Since the events of September 11, 2001, GIS technology has been at the forefront of the Emergency Response and Disaster Response communities. As long as it has a location element, GIS creates a common framework, for data analysis and data visualization. When the event or scenario data is placed on a map, it is easily understood, and can project the potential impacts on the environment and infrastructure.

GIS allows users and decision makers at all levels access to the same data and analysis, referred to as the Common Operational Picture (COP). During an emergency response situation, Local, State and Federal actors can all look at the same data, and have the same information.

Using the Haitian example, a web-based COP, based on OpenStreetMap, of the affected areas, was used by multiple international agencies, non-profits, and Defense agencies. This map displayed the ad hoc infrastructure, where assists were located, what their status was and where there were calls and tweets for help.
John Snow is often cited as the father of both epidemiology, and mapping the physical vectors of disease; his study of the water pumps in London with their correlation to cholera is mentioned in both Public Health and Geographic/GIS texts and academia.

GIS and Spatial Statistics offer public health providers the capability to express the results of a study in a geographically visualized format. GIS can also facilitate the rapid discovery of clusters, pathogen origins and allows the epidemiologist the ability to visualize the correlation between the environment and health issues. Data obtained from GIS also influences policy decisions by allowing decision makers to see where disparities exist within the health care systems of various political entities.

From a data perspective, GIS allows the public health practitioner to find a common ground for the utilization of multiple data formats. If the data include any geographic information, such as zip codes or county names, they can be linked with any other data source that also share geographical information. Similar to the COP, this provides the common framework to use and share their data in a common environment. This common framework allows the researcher to share their data easily with decision makers, peers or in the case of emergency response, mission partners.
While a great deal of research is being done on the use of Social Media during unexpected events, such as hurricanes or earthquakes, "slow burn" events, like an influenza outbreak, are generally not the focus for research. It was not until the flu season of 2012/13 when the traditional media began to understand and integrate social media into their analysis of the epidemic.

Through the utilization of GIS, as well as assorted mapping technologies, Public Health practitioners can develop a COP, which will display the relevant geo-tagged social media data. Emergency Management Professionals can then assess this dynamic and real time data against their current geospatial models and algorithms.

Through the combination of Social Media and GIS applied it to a Public Health Emergency the following topics are addressed:

- Assessing Risks
- Evaluating Current Threats
- Predicting Future Threats
- Tracking Current Threats and Outbreaks
- Providing a Common Operational Picture (COP)
- Intelligence Lead application of resources (e.g., vaccines, personnel)
Investigation into this topic will merge research from the disciplines of Disaster Response/Emergency Management, Risk Terrain Modeling (RTM), Public Health and Big Data/Social Media analysis, and fuser generated geographic data from Twitter.

**Problem Statement**

A number of studies focus on the use of Social Media during the "response" phase of a rapid onset event, and how to integrate this data into an analysis of the event. Modelers use this harvested data to recreate the situation to gain a greater understanding of what occurred during the event. With these lessons learned, similar problems in response and process can be understood and mitigated against during the next event.

During a public health event, social media allow researchers to see the "real time" spread of a disease or event. As more people post their symptoms on social media outlets, researchers may then harvest this data and gain a greater understanding of the "hot spots" for the illness.

Traditional bio-surveillance is an expensive and lengthy process; collection and analysis of data can take days or weeks. Traditional bio-surveillance only collects data on those patients who are seeking treatment at a medical facility. This medical data is then collected, and sent to the CDC for aggregation to the state level. This methodology ignores those people who do not go to a traditional medical facility. By leverage social
media researchers may collect data from those who are not seeking medical attention, but are exhibiting symptoms of the outbreak. Data can then be collected quickly and analysis take place in near real time.

Using geo-tagged Social Media, analysts can conceivably trace the outbreak back to its physical origin, and by researching the user's social network, we may determine which of their friends could have been the potential disease vector. Researchers and Investigators could then perform a deeper analysis and increase the likelihood of finding patient zero.

Influenza was chosen due to its temporal predictability and the availability of background data and information. Since Influenza is a yearly event, and it affects people in every state, it was determined that this pathogen would be a strong case study.

By using Risk Terrain Modeling (RTM) and comparing that data to where the Social Media hot zones are located, an analyst can estimate the movement of the disease. This analysis assists first responders, who can be proactive, and not simply react to the event as it unfolds. The question becomes, "Can Risk Terrain Modeling principles assist with the prediction of a communicable disease?"
Adds to Existing Body of Knowledge

Research of Social Media in Emergency Management is in its infancy, however, with events like Sandy and the earthquake in Japan, research in the field is growing. Researchers are now beginning to focus on the "citizen as sensor," and including this data in their overall model for predicting the situation during an event, or the effects of a cascading event. The collection of Social Media provided health data specifically within the field of Bio-Surveillance and Biosecurity. Through the analysis of this ambient data, trends and metrics can be developed and established, which will then allow analysts and managers to view social media as a tool to determine trends and uncover threats and hazards while they are still small.

This project will add to the body of knowledge by using RTM and Social Media to help predict the areas at greatest risk for an outbreak of influenza. RTM will provide the baseline data on how susceptible a state could be to an influenza outbreak, by identifying those states that tend to have a high level of flu activity, and a health care system that is unprepared or simply lacks the capacity to react to a large Biosecurity event. Social Media data harvesting will capture data from those states where traditional Bio-Surveillance may fail, and could alert the government to a possible outbreak allowing the relevant authorities to bring the proper resources to bear.

The analysis will test for the statistical significances of the CDC data and the harvested social media data separately to determine if a Risk Terrain
Map could be used to find where an influenza outbreak could occur. If the results of this study show a strong correlation between the Risk Terrain Map and the CDC and Social Media data, then it will demonstrate that this Risk Map can be used by planners to predict influenza outbreaks. If the Risk Terrain Map shows a strong correlation between it and the CDC data and not the Social Media data, then that will show that social media was not a viable form of Bio Surveillance at the time of this study.

A strong correlation between the social media data and a weak correlation to the CDC data, the Risk Terrain Map will show that Social Media is a viable method of Bio Surveillance. In this scenario, the viability of the Risk Terrain Map will be questioned.

Lastly, if both datasets show a weak correlation to the Risk Terrain Map, this will indicate that the Risk Terrain Map methodology was flawed, or that Risk Terrain Mapping cannot be used to predict influenza outbreaks.


**Literature Review**

**Risk Terrain Modeling**

The core of this process is the development of the health RTM. RTM is used extensively by both Criminologists and Human Geographers. The theory of RTM is outlined in the *Risk Terrain Modeling Manual*, written by Caplan and Kennedy, a work focusing on research on the application of RTM to crime. (Caplan and Kennedy, 2012)

Caplan and Kennedy theorize that crimes are linked geographically. Through the aggregation of data provided by the City of Philadelphia they were able to generate crime "hot spots" where certain crimes tended to happen. Upon further investigation they found overlapping hot spots, indicating that certain crimes tended to happen where other crimes were also occurring. (Caplan and Kennedy, 5, 2011)

As an example, the authors describe the conditions that indicate an increased risk for aggravated assault. These spatial factors include proximity to bars and nightclubs, entertainment venues, areas of known gang activity, drug trade and areas known for drug and alcohol use. (Caplan and Kennedy, 35–37, 2012)

By taking these factors into consideration the authors describe the process of the development of intelligence led policing. Production of Risk Terrain Maps provides a spatial understanding and intelligence to the officers...
which in turn leads to strategic decision making. Leadership then can use the information to effectively target resources where they will be most effective. (Caplan and Kennedy, 21-22, 2012)

The technical approach to RTM is straightforward; it requires the analyst to utilize empirical methods, literature reviews and professional/practitioner experience to determine which risk or risks they wish to assess. Once the determination of the risk is made, multiple GIS layers are applied to connect common factors by geography. The Risk Terrain Map is produced once all these layers are combined, and will show the composite risks that include all the factors associated with a particular outcome. This map then displays the areas of interest, and indicates the likelihood of a particular event to take place in a specific area.

Caplan et al’s further work on RTM, “Analyzing and Visualizing Worldwide Spatial Data: an Application of Risk Terrain Modeling” further explains how RTM can be applied to national data sets outside of crime, and includes events such as flooding, civil unrest, and health related events. (Caplan et al, 37-42)

Within the field of Public Health, Caplan believes that RTM can be used to indicate outbreaks of diseases. By mapping the outbreaks of West Nile Virus and Malaria and then associating those with the areas with large bodies of water, researchers can better prepare their region for a possible
outbreak. In the case of influenza, Caplan believes that urban centers are at a much greater risk than rural settings or smaller cities, simply because people interact with more humans, and as such, expose themselves to many more disease vectors.

RTM provides decision makers with the capability to make decisions based on years of previous data, research and analysis through a simple visualization. Through the use of RTM and GIS, decision makers can easily overlay any geospatial data on the Risk Map, and receive near instant feedback in an easy to understand format.

Within this project the Risk Terrain Model will be created at a county level. Data to determine the health conditions and quality of healthcare will be based on the methodologies utilized within the Robert Wood Johnson Foundation funded report, “County Health Rankings and Roadmaps: a Healthier Nation, County by County.”

Researchers focused on 32 influencers of health and health care quality. Using quantifiable data, including, among other criteria, the number of uninsured, access to physician care and proximity to health care facilities. Counties are then ranked against other counties in the same state. On a national level, counties are placed into quartiles, in order to compare the system across the national spectrum. ("County Health Rankings", 2012)
Mobile technologies and social media have changed the lives of millions around the world. The amount of data that flows from social media platforms on a daily basis is staggering. Some of these data are tied to coordinates which support mapping and increase our understanding of trends within society. By utilizing the Twitter platform, we can discern both localized and national trends and researchers and analysts then can utilize these data to predict possible problems, and to achieve near real-time situational awareness.

In Stefanidis et al’s work, “Harvesting ambient geospatial data from social media feeds”, the authors describe a methodology to harvest user-generated geo-data, and then discuss uses of this data in two case studies - determining hotspot emergence and tracing information within social networks. (Stefanidis, 1-2)

Stefanidis et al, indicate their methodology is dependent on specialized taxonomies which need to be developed for each new study. These taxonomies allow the researcher to filter through the “fire hose” of information being generated, reduce the amount of data to aggregate and focus on what they wish to research. (Stefanidis, 4-7)

The Arab Spring was chosen by the authors as their case study for hotspot analysis. By harvesting social media from multiple platforms, the
researchers were able to determine and map hot spots of protest and police activity. (Stefanidis, 9-12)

The authors also chose the Japanese earthquake of 3/11/11 as the case study for tracing information within social networks. Tracing geo-located tweets and retweets, the authors followed the flow of information from the affected areas to the rest of Japan and the world. (Stefanidis, 12-14)

Stefanidis, et al conclude with a discussion of ambient geographic information (AGI) as opposed to volunteered geographic information (VGI). The authors compare social media feeds to flood gauges. The information simply flows, and the researcher consumes or harvests those data they wish to use. The users are not focused on just the geography of the event but the attributes around it. Through the harvesting of AGI the researchers gain a greater understanding of groups and group dynamics. (Stefanidis, 15-17)

There are a number of other works dealing with AGI. Crooks et al. in their work, “Earthquake Twitter as a Distributed Sensor System” describe the tweets after the August 23, 2011 earthquake in Virginia. In this study they leveraged the harvested Twitter data to determine how “far out” people could feel the quake. (Crooks et al 1-3) The authors also describe how the “tweetepers” operate in similar fashion to a sensor. (Crooks et al 4-5) From
the data gathered from Twitter, the authors created a "shake map", a map of recent earthquakes, and determined it could be felt as far away as Quebec, to the North and Atlanta to the South, a fact confirmed by the USGS Earthquake sensors confirmed. (Crooks et al., 13)

In the Song et al. work, "Discovering More Meaningful Regions: a Regularized Geographical Topic Model", the authors, using a regularization method known as the Regularized Geographic Topic Model (RGTM), discovered that social media has a "regional" effect - that regardless of the social media platform, Twitter, Flickr or Facebook, posts that originated geographically near each other often described the same situations. (Song et al., 34) The regional effect can be used as a "check" against bad data being posted by the public, much like a control group. (Song et al., 37)

AGI, GIS and Public Health

In Cinnamon and Schuurman's work, "Confronting the data-divide in a time of spatial turns and volunteered geographic information", they describe the advancement of both GIS and social media within the field of Public Health. (Cinnamon and Schuurman, 3) Epidemiologists and other Public Health Practitioners have only recently begun to use GIS and social media to assist them with analysis and reactions to epidemics and other public health events. (Cinnamon and Schuurman, 5)
The authors attribute this change to the reduction in the cost of bio-surveillance data accomplished through harvesting health related postings from social media platforms. With this data being posted publicly by individuals there are no privacy concerns, and the researchers have real-time access to the information, which does include some location data. (Cinnamon and Schuurman, 7-8)

GIS and Geo-visualization are also seeing wider adoption in the Public Health community. The authors attribute this to the increase in the amount of geographically referenced health data. Also, the growth of web based GIS applications makes spatial analysis cost effective, and extends the reach of the technology. (Cinnamon and Schuurman, 9)

The authors do express a concern about what they call "the data divide." The data divide is built upon the concept of "the digital divide," where developing nations neither had ready access to information technology, nor the education or infrastructure to support it. The digital divide concept was discussed in both academia and the non-profit sector in the late 20th and early 21st century. While the issue has not been completely resolved, the internet reaches further today than it did in the year 2000. (Cinnamon and Schuurman, 10)

The data divide, within the scope of this topic, relates to the lack of health data from within these developing states. Unlike the developed
world, which has decades of data available for trend analysis and clustering, the underdeveloped world lacks both the data and the infrastructure to collect them. The authors believe that the proliferation of social media will assist these nations in the collection, analysis and dissemination of health data, at a low cost. As part of their research, the authors outfitted the local Emergency Medical Services (EMS) staff with low cost tablet computers to enter information. The EMS personnel collected location as well as data from the patient or accident, synced their devices and uploaded the data to a central database server. (Cinnamon and Schuurman, 12)

The authors found that the collection of social media AGI data from the field was not enough to generate hotspot analysis, however, the addition of the EMS personnel and the data they collected, they were able generate sound data. Using the AGI, and Authoritative Data, the authors were able to determine hotspots and generate trend analysis. (Cinnamon and Schuurman, 14)

In "A Model for Mining Public Health Topics from Twitter", Paul and Dredze discuss the Ailment Topic Aspect Model (ATAM). The ATAM provides researchers with taxonomy of key search terms in order to "reduce the noise of twitter." (Paul and Dredze, 1-2) Their methodology was to classify the tweets into 3 different groups - General, Symptoms and Treatment. General provides the researcher with an insight to what the ailment is (i.e., sick, tooth, back). Symptoms are a modifier to the general taxonomy though a
By leveraging the ATAM, researchers can now discern between a single tweet about an individual being sick, as opposed to inferring someone or something is “sick” or other uses of slang in tweets. The authors found ancillary words such as “Doctor”, “Hospital”, or the name of drug used to treat the ailment in the tweets increased the accuracy between the Twitter data, and the data gathered using classic bio surveillance techniques. (Paul and Dredze, 6-7)

The researchers studied Twitter data for the flu season between September 2009 and April 2010. They aggregated the weekly data into months and discovered that the accuracy of the tweets against the CDC’s bio surveillance data varied from month to month. When the data was aggregated to the full influenza season, the researchers found a statistically significant connection between the tweets and the CDC Influenza data.

Sadilek et al.’s work “Predicting Disease Transmission from Geo-Tagged Micro-Blog Data” attempts to predict the spread of the disease, and locate patient zero or a “typhoid Mary.” (Sadilek et al, 138-9) The researchers step past simple location analytics by including social connections of the users “Followers” or “Friends” in their analysis. This resulted in determining the spread of a virus or illness across the various
social groups, and finding the connections with one another. (Sadilek et al., 143)

Burton et al. researched using Twitter to determine public health outbreaks and found that tweets are indeed an accurate representation of the various ailments affecting the different regions of the country, even though only a small percentage of the country uses Twitter, and only 3% of those who do geo-tag their tweets. (Burton et al., 3)

Kriec et al. found that while Twitter can be used as a resource for collecting data on common disease surveillance such as influenza, for more advanced diseases, like Q-Fever, tweets lack the information needed and reduce accuracy. (Kriec et al., 4) The authors concluded that Twitter could be used to describe general trends of sickness, but in cases that did require a medical diagnosis, Twitter was unable to provide relevant data. (Kriec et al., 6-7)

Chunara et al.’s work with the Cholera outbreak in Haiti counters Kriec’s claim. Through the use of Twitter and other social media outlets, the researchers were able to predict the movement of Cholera through Haiti. Public Health officials used Twitter to focus their assets, and slow the spread of the disease. The data collected from Social Media were found to be accurate, and could be analyzed instantaneously. This reduced the epidemiologists’ reporting time, from two weeks, to a single day. (Chunara et al., 39-42)
Chunara et al. acknowledge that informal data collection methods do contain an innate sample bias, but that using social media as a resource, researchers can better detect outbreaks, and can respond more rapidly.

According to the authors the difference between these two studies is an informed public. In Kriek’s work, the public was not given the proper instructions on how to react to the symptoms, while in Chunara’s work they were. When the public is informed of what symptoms to look for, they provide a more accurate picture of what is occurring “on the ground.” (Chunara et al. 43–45)

Lastly, Dredze describes in his work how social media will forever change the science of epidemiology. Building on the examples of other disasters and the use of real-time updates from Twitter and other social media platforms, researchers have access to unprecedented amounts of data. Dredze argues that the use of Twitter as a real-time disease sensor with its ability to generate metrics in real-time are simply “a boon” for the field of Public Health. (Dredze, 81) The use of social media collection will augment bio-surveillance, and assist researchers to fill in the gaps in their data.

Social media is not a single source of data, it works with other sources of data to show the larger picture. Dredze argues there is an inherent bias within the use of social media. People who tend to use social media are not
an accurate representative sample of the public at large. (Dredze, 83)

Dredze concedes that even though this is the case, it is another tool to provide insight into the population that officials have not had in the past. (Dredze, 85)

The Study’s Geography, Population, Timeframe and Privacy

Geography

The continental United States (CONUS) was selected for the geographic area of this study, since on a per capita basis CONUS has the most Twitter users. There is a great deal of research and data about health care conditions at the local level. From a geographic data standpoint, the United States has easily accessible spatial datasets to provide for the cartography and data analysis for this project.

Population

The study’s population consisted of Twitter users who live in the continental United States during the timeframe of the study and share location data. At the end of 2012 the United States had an estimated 28 Million Twitter users, or roughly one in every ten citizens.

The demographics of the “average” Twitter user is as follows: Twitter users tend to be male, between the ages of 18 and 49, live in an urban area, have some college education, but have not graduated, and are ethnically of
European decent. Lastly, their household income is more than $75,000 US dollars per year.

While demographically Twitter users are not diverse, but usage of the platform requires a smart phone and a data plan. As well as a level of technical knowledge to create, and maintain a Twitter account. It is not surprising then that Twitter users are primarily affluent. In similar studies the demographics of the sample size were discussed, but the development of a model and method to control for this disparity is not available at this time.

Timeframe

The timeframe for this study spanned the Weeks of January 5, 12, 19, 26 and February 2nd of 2013. Flu data collected over the past 30 years indicate that the peak 3 months for flu are January, February, and March. January was chosen because it is the beginning of the peak flu season, so there would be areas not affected at the beginning of the month that would be later. In effect, it is the "ramp-up" month, so the data will increase as the month goes on.

Privacy

Twitter allows users to make their streams private, where only the people they follow can see their tweets. Since these are not public tweets, they were not harvested; only public tweets with the users volunteering location were used in this study. Additionally, the application that collected
the tweets did not store the user's name, only the user's Twitter handle, and only for that day. When the script began to run, its first function was to delete the user's handles from the database. The user's handle was only used to ensure the script did not collect more than one tweet per user per day.

**Public Health Risk Terrain Map**

**Software**

ArcGIS for Desktop was utilized as the software system to produce the map, utilizing both the Spatial Analyst Extension as well as the Risk Terrain Toolbox. The Risk Terrain Tool (RTT) box is a custom extension developed for the ArcGIS Framework by Rutgers University in New Jersey. The RTT either generates a raster heat map, that shows both hot areas and cold areas, or creates a tabular data output which can then be tied to pre-existing vector data. For the purposes of this study, that would be States.

Once the RTT generated the data, the tabular data were then converted into a Microsoft Excel Spreadsheet, which was then spatially joined with the State vector data.

**Methodology**

Development of the Influenza Risk Terrain Map (IRTMap) utilized data from three sources. First, for historical trends and analysis, we look to the
Centers for Disease Control and Prevention (CDC) in conjunction with the Department of Health and Human Services (HHS). Our second data source to create the IRTM is the Robert Wood Johnson Foundation’s County Health Rankings Data, from 2012. The third and final data source is Google’s Flu Trends.

**CDC Data**

These two organizations work together to monitor and track Influenza Like Illness (ILI) through the United States Outpatient Influenza Like Illness Surveillance Network (ILInet).

Monitoring Influenza in this fashion is a classic Bio-surveillance methodology. Data is aggregated from the county to the state and up to the Federal level at the end of each week. Using the ILInet data puts the analysts up to 8 days behind the virus, whereas with a social media solution, it could be daily if not hourly.

The total size of the ILInet network is well over 2700 clinics. Data are collected weekly from each of the ILInet clinics. Each clinic reports patients who arrive with a fever in excess of 100 degrees, cough and sore throat.

This percentage of patients who arrive with an ILI is then compared against a national baseline average of 2.2%. This percentage is then compared to data of the last three flu seasons of the same week and the severity of the flu is then determined by the number of deviations from the
mean that the data quantify. The data is then aggregated from the clinics to the state level.

Table One depicts the number of deviations from the mean, and their severity levels.

Table One

<table>
<thead>
<tr>
<th>Deviations from the Mean</th>
<th>Severity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-3</td>
<td>Sporadic</td>
</tr>
<tr>
<td>4-5</td>
<td>Local</td>
</tr>
<tr>
<td>6-7</td>
<td>Regional</td>
</tr>
<tr>
<td>8-10</td>
<td>Widespread</td>
</tr>
</tbody>
</table>

At the end of each Flu Season, the CDC aggregates all their data into a final report. Here the CDC indicates which states had No Activity, Sporadic, Regional or Widespread Flu outbreaks. For the purposes of the IRTM, we will use CDC data from 2003 onward.

Health Rankings

County health data were downloaded from the Robert Wood Johnson Foundation’s website (www.countyhealthrankings.org). The site provides a full dataset in Excel format. This study measures the health outcomes of
each county in the United States based on a number of factors. The data utilized for this study were the 2012 County Health Rankings data.

County geometry was obtained from the Census Bureau’s 2010 tiger dataset, in a shapefile format.

County Health Rankings

Both county geometry and the county health rankings spreadsheet were loaded into Quantum GIS. The health data and geometry were merged on a common unique attribute, FIPS code. The county health rankings data had already aggregated the datasets into quartiles.

The first quartiles are those counties that are considered to have the best health care systems (i.e., best doctor to patient ratio, access to parks) based on the report. These areas tend to align with major metro areas. The fourth quartiles are those that are considered to have the worst health care
systems in the United States. These areas tend to align with the poor and rural areas. There are 190 counties in which data was not available for this report.

To align this data with both the CDC and Google’s aggregations to a state level, the same was done for the county level data. Each state’s ranking was determined by an average of all the counties. The states then were divided into Quadrilles, determined by the average number of counties in the state. Once this aggregation was complete the map changed.

Google Flu Trends

It is well known that the search platform Google collects metrics on searches conducted through its website Google.com. Starting in 2003, Google began to look at searches for health search queries, with "The Flu" and other flu related terms spiking during the flu season. With this query selection method, Google is able to aggregate its data down to the city, or township level. However, to maintain the same aggregation level as the CDC, the data for state level were obtained and utilized.

This flu data are freely available at Google.org, and easily downloaded in either Excel, comma delimited or fusion table format. Once the data were obtained in a comma delimited format, they were simply spatially joined with the geometry of the U.S.
Google Flu data are considered by many in the bio-surveillance field to be "operational" or raw data - data that are not vetted by Public Health professionals. Regardless of this view, Google Flu data are used by many academics as well as the CDC for monitoring ILI within the United States.
Development of the Risk Terrain Base Map

The IRTM model utilizes the CDC's ILI Reports, The County Health Rankings, and Google Flu, using data from 2004, the first year Google started tracking results, until 2012, the year before the study's window. This assists us in creating a map showing which states are prone to "stronger" outbreaks of ILI, and will assist us in determining risk. Each state will be divided one of five categories - no activity, sporadic, local, regional, and widespread. These five categories match groups used by both the CDC and Google Flu. We used the CDC in order to maintain consistency across the various data sources.

Once the data were generated, they were then loaded into a PostGIS database, where they could be quickly manipulated into multiple spatial formats.

The following formula was utilized to determine the total risk level for each state for each year of the study. The total for the 9 years was then averaged.
**Barr - 30**

\[ CD * .7 + GF * .2 + CH * .1 = RA \]

This formula is the basis for the SpatialSQL query. This query was then run against the State data stored in a PostGIS database that generated temporal results and composite results. These results in turn were then visualized within QGIS in order to generate the final maps, which were used to generate the cartography, the legend, and the final visualization.

The RA totals end between 1 and 5. A full listing of the states and their Risk Group is located in Appendix B. Once these totals were computed, Jenks Natural Breaks were applied to generate the RTBM. The resulting analysis is a generation of 5 groups - Risk Group 1 being the highest total risk, and Risk Group 5 being the lowest total risk.

If one of the data sources did not have data for a particular state, that state is considered to have "No Data." This was only the case with the 2005 analysis of, where no CDC data was discovered.

Comment [SRH5]: ??? of what?
As this is the first time that RTM is being used to predict a public health event or ILI outbreak, there is no model, or best practice to create a model from. As such, this model is a "model in progress." The results of this study will assist with the further development and accuracy of future models.

The following maps are the Risk Terrain maps from each year, 2004-2012, and the final map is the aggregated results based on the formula above.

Influenza Like Activity for 2004
Influenza Like Activity for 2011

Influenza Like Activity for 2012
The initial analysis indicates that Risk Group 1, tends to be less populated states and in states where the major urban centers are separated by many miles. While, on the other end of the study, those states in Risk Group 5 tend to have large urban centers, and have multiple dense population centers. Since influenza is passed from person to person via an airborne/waterborne pathogen it goes to reason that areas where people live closer together are more susceptible to the flu.

Data Collection

This project required the collection of two datasets. The first dataset to be determine the quality of each county’s health care, and taking that
data and joining them to the geometry of a map. The second dataset to harvest all the tweets within the US for the timeframe of the study.

Twitter data was obtained using Twitter’s API and a Python script that searched for hashtags, particular words within the tweet, both key and modifier words. The keyword within a tweet determined the topic, or if they had a flu. The modifier word provided the study with collaborating data to determine if the user was actually sick, or referencing “flu” for some other reason. Appendix A contains a list and analysis of the modifiers used by users over the course of this study.

Additionally, the Python script validated that each tweet was unique from a single user, and not a retweet, or a user tweeting multiple times. If the tweet was generated by the same user, or was a retweet, that tweet was not saved into the data store.

This harvester collected an average of about 40,000 tweets a night. Of these tweets less than 5% had location data attached to them.

For the duration of the study this data were collected by the script, and migrated to an XML file. At the end of the study, the XML file was imported into an Excel spreadsheet for data cleaning and geocoding.
### Table 2

<table>
<thead>
<tr>
<th>Element</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>The Date, Hour, Minute and Second of the Tweet</td>
</tr>
<tr>
<td>User_handle</td>
<td>The Twitter user’s handle who posted the tweet</td>
</tr>
<tr>
<td>Tweet_Text</td>
<td>The actual text of the tweet</td>
</tr>
<tr>
<td>City</td>
<td>The City in which the Tweet originated</td>
</tr>
<tr>
<td>State</td>
<td>The State in which the Tweet originated</td>
</tr>
<tr>
<td>Key_Word</td>
<td>The key word used in the Tweet to determine if it was harvested</td>
</tr>
<tr>
<td>Modifier</td>
<td>The Modifier used to categorize it as a flu as a sickness related tweet, or to determine if it was someone referencing the flu for another reason</td>
</tr>
</tbody>
</table>

### Table 3

<table>
<thead>
<tr>
<th>Key word from Tweet</th>
<th>Total Times Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flu</td>
<td>24,573</td>
</tr>
<tr>
<td>Influenza</td>
<td>370</td>
</tr>
<tr>
<td>H1N1</td>
<td>30</td>
</tr>
</tbody>
</table>
As the above table and graph demonstrate, the most commonly used one was flu, followed by influenza. The remaining keywords combined are less than 50 tweets over the course of the month of this study.

**DATA CLEANING**

The health data had already been cleaned and aggregated; the Twitter data were raw and needed to be thoroughly cleaned in order to geocode the tweets to a city level. The first step was to remove the tweets that did not
have any location data attached. The second step was to remove any tweets from outside of the study area.

The third step required looking through the data line by line to determine if city and state were populated correctly. In some tweets the city was replaced by a drug store, or a medical center; these locations were searched on the internet, and the proper city populated. In some cases the record was thrown out because the location was too vague, such as "at CVS."

The fourth and final step was sorting the remaining entries based on their modifier. This eliminated the tweets where users were discussing getting a flu shot, or advertisers tweeting about the flu, or where the public could go to get a flu shot.

After the data cleaning process was completed, roughly 37,500 tweets were still viable.

**Geocoding**

The target accuracy for geolocation of the tweets was at a city level. The data were geocoded by two different geocoders, the ArcGIS Online Geocoder and CartoDB. Not only would this provide a metric to determine which of the geocoders was more accurate, but also function as a quality check against the data.
The ArcGIS Online Geocoder was called from ArcGIS Desktop, and the accuracy placed at 50%. Since this was not address specific, 50% accuracy should be enough to place the location within the city. Those records that were discarded could then be reviewed and scrubbed another time.

CartoDB's (www.cartoDB.com) free service limits file sizes to 5MB or less. In order to meet this limitation, the dataset was exported as a CSV to reduce its size and uploaded to CartoDB to be geocoded. Those records that did not match were unable to be viewed at the end of the process, and what was returned is a GeoJson file, which could be transformed into an ESRI compatible format, using an OGR services available at the web at http://www.shape.ly.

The ArcGIS Online Geocoder crashed a number of times with the full dataset. Through trial and error it was discovered that the ArcGIS online service could only reliably geocode record sets of 5000 rows or less. CartoDB did not have any row limitation, just the 5MB file size. However, if you close the browser window where it is operating, the geocoding process stops. The final results were as follows:

The ArcGIS Online Geocoders had a 70% match. The CartoDB geocoding process had a 90% match. The data were inspected to determine a cause for the discrepancy and it was discovered that the ESRI data could not locate records where the state name was used, rather than the
abbreviation. Once this was corrected, and the state name abbreviated, the ArcGIS Online Geocoder had the same accuracy as the CartoDB geocoding process. This data were then subdivided by week and transformed into shapefiles.

The reason for the variance in the accuracy of the two services was a simple data format discrepancy. When the state name was fully spelled out, the ArcGIS Online service could not geocode the location.

Since these data were aggregated to the state level, they may not be geographically accurate, as an urban center may have a large number of flu cases, while the rural areas do not. Even with this limitation, the ILI Report is considered by the CDC to be the best judge of the geographic movement and intensity of the flu.
**Methodology**

Data collection began on December 30, 2012 and ended on February 2, 2013. The entire study lasted for 35 days, broken down into 5 separate weeks. The following sections will describe both the tweets as individual data points, and on a per capita analysis of the Twitter activity captured within a state. This will then be compared against the CDC data, to determine if the Twitter data is accurate or not.

User adoption of Twitter is one of the largest data accuracy risks for a study of this type. While Twitter may be used heavily in major urban centers and “progressive” states, other states where there could be a large flu outbreaks, Twitter will not help simply because of lack of user adoption.

The final analysis will determine the accuracy of the IRTM, on both an ILI level, and at a social media level. If the state appears in the correct categories for ILI and the per capita social media group, then it is marked as a positive. If the state is above or below its risk group, then it is marked as zero; these results are binary. For example, if Texas falls in the top tier of ILI activity, but has a low social media score, a positive result will appear in the ILI column, but a zero will appear in the social media column.

Social media activity is ranked by the per capita number of geolocated tweets within a state compared to that state’s population. So, if New
York State falls into the highest tier of Social Media Activity per capita, they receive a positive result. If they fall into the second tier, they receive a negative result.

Once the data are compiled for the whole of the study the statistical significance will be tested multiple layers of the data. First all the data, a combination of both the IRTM and Social Media data, will be combined and tested for significance. Then the data will be broken down into subcategories, where the significant of just the IRTM will be tested, along with just the social media data.

Lastly, the significance of each of the risk groups will be tested. This will assist in the overall analysis, and determine the accuracy of the data on the most granular level gathered for this study.
Maps Week One

Individual Tweets on the Risk Map

Per Capita Tweets by State

Comment [SRH12]: The legend subheader (variable) is misleading.
CDC ILI Map for the Week of January 5, 2013

Comment [SRH13]: I prefer this project, but why change now?
Maps Week Two

Individual Tweets on the Risk Map

Per Capita Tweets by State

Legend
Per-Capita By Population
- Less than 1 per 10,000
- 1 to 9 per 10,000
- 10 to 24 per 10,000
- 25 to 49 per 10,000
- 50 to 74 per 10,000
- 75 to 99 per 10,000
- 100 or more per 10,000
Are you getting this map from CDC or are you regenerating it?
Maps Week 3

Individual Tweets on the Risk Map

Per Capita Tweets by State

Legend
- Per Capita for Population
- Low (0-15,000)
- Medium (15,001-40,000)
- High (40,001-75,000)
- Very High (75,001+)
- Exceptionally High (100,001+)
Maps Week Four

Individual Tweets on the Risk Map

Per Capita Tweets by State

Legend
Per Capita By Population
Low (<1 tweeter per 100k)
Medium (<1 tweeter per 500k)
High (<1 tweeter per 1M)
Very High (>1 tweeter per 1M)
Highest (>1 tweeter per 2M)
Maps Week Five

Individual Tweets on the Risk Map

Per Capita Tweets by State
Results

From the macro view none of the data present a statistically significant indication that IRTM can be used as a predictive tool for the outbreak of ILI in the United States. Testing for statistically significance, the combined social media and IRTM data show a 3.2 deviation from the mean. This indicates that there is no correlation on the macro level between the flu outbreaks, the IRTM, and social media data. The following paragraphs will discuss the computation and the IRTM and social media results separately.

IRTM

Breaking down the data into just the IRTM shows that there is no further macro correlation between the IRTM and the ILI flu data. Analyzing just the IRTM data yields a 2.8 deviation from the mean. This indicates no statistical significance between the IRTM and the flu data at a macro level.
Breaking the IRTM data down further into the weeks shows a similar result. None of the 5 weeks of this study show a statistically significant relationship between the IRTM and the flu data. It is at this point where a pattern within the assorted risk groups became clear.

While on the macro and temporal levels there was no statistically significant relationship, the analysis began to show that the higher the number of the risk group, the less accurate the results became. Breaking down the analysis by risk levels yielded a new outlook on the IRTM. In states with the highest level of risk, risk Group 1, the data showed that there was only 1.42 deviation from the mean, indicating that there is a statistical relationship of .97 between the IRTM and the flu data.

In Risk Group 4, there was a greater deviation from the mean, that of 1.61. This still indicates a statistically significant correlation of .95 between the IRTM and the flu data. So, the IRTM is statistically accurate for the states in the top two risk groups.

Risk Group 3 is where the IRTM begins to lose its accuracy. The deviation from the mean for this risk group is 3.2. This shows no correlation between the IRTM and the flu data.

In Risk Group 2, the deviation from the mean is further removed at 3.7. This shows a less than average chance of the state appearing in the risk group where it was assigned. The only interesting results from this data
set are that during the first week, all of the states were in their risk group, and then they became scattered.

The final Risk Group, Risk Group 1, yielded a deviation from the mean of 4.25. As such, there is no statistical significance between Risk Group 1 and the flu data. Risk Group 5 is interesting because the only state that appears in this risk group, appearing every week, was Montana. None of the other states ever appeared in this risk group.

Social Media Data

The analysis of the social media data against the flu data did not prove any statistical significance. At the macro level, the data showed a 2.25 standard deviation from the mean, showing that there is no statistically significant relationship between the collected social media data, and the CDC's ILI data.

Further breakdown of the collected Social Media data, both temporally and geographically, shows no statistical significance with the CDC data. The lowest deviation from the mean results were in Risk Group 2 during Week 3, where the standard deviation from the mean was 1.92.

When the Risk Groups are removed, social media data becomes more accurate. Comparing and contrasting just the state level tweets and the CDC ILI data, shows a greater correlation. From a macro perspective, without the Risk Groups, the social media data's standard deviation from the
mean is 1.73. While not statistically significant, it does show a greater accuracy. This reinforces the idea that the IRTM model is flawed, and should be engineered in future tests. It also shows that while social media is not a replacement for traditional bio-surveillance, it can lend itself to the practitioner to see trends and assist with further analysis.

**Fixing the IRTM**

The IRTM model for this study was flawed, as the analysis of the data demonstrated, with the Risk Groups losing accuracy past the higher risk states and the social media data becoming more statistically significant when the Risk Groups were removed. These factors all point to a flaw in the IRTM's development and the model itself.

The model was heavily weighted on past ILI outbreaks, which are directly tied to reporting of these outbreaks. The affected individuals would have to go into a medical facility or see a doctor to be included in this data. Therefore, the addition of the county level health service data to the model was not necessary. The rating of the local Public Health system did not have any impact on the final analysis, as this was already addressed in the CDC ILI data. In the next iteration of this model, the health care ranking will not be included.

According to the CDC data, New York City maintained a high level of ILI activity during the course of this study, which was different from most of
the state, but the same as New Jersey and Connecticut. A similar effect occurred in District of Columbia, Maryland and Virginia, where each of these states tended to reflect each other’s ILI rates.

In the next iteration of this model, the Metropolitan Statistical Areas (MSAs) with their commuter data will be included. If a MSA has a large number of out of state commuters, then the IRTM model will include this as a weighted parameter. In addition to out of state commuter data, population density, both in the state and the MSA will be taken into consideration.

The CDC’s ILI data would not be so heavily weighted. In this IRTM, it was weighted at .8 of the total model. In future iterations, the CDC ILI data would only be weighted at .5. A greater emphasis will be put on the population data.

A larger sample of social media data would be harvested, extending the total time of the study to a full season of influenza. Additionally, with changes in the Twitter API, it is now feasible to find a tweeter’s followers. Using this new function, and using the same taxonomy, determining which of their friends also indicated influenza like symptoms will be easier.

To further strengthen the social media data the taxonomy would be changed to include geographic names. This can be accomplished through use of the GeoNames database. Looking for cities, rivers and other
landmarks would allow us to harvest a larger sample of tweets. Combining more tweets with a larger network group, would give us a much stronger case to use social media resources in lieu of classic bio-surveillance.
Conclusion

In analyzing the spread of disease, there are thousands of factors and variables to take into consideration. Building upon the CDC’s Data, Google Flu and the Health care rankings in order to create the IRTM seemed like solid science, as these analyses had been tested over the past decade, with millions of dollars of support. However, the IRTM model failed this time, and there was no statistically significant correlation between the IRTM and predicting ILI. Additionally, the failure of social media data to support the IRTM pointed directly to flaws in the model. While this first attempt at the model failed, the data did point us in a direction where the model could be altered.

A better understanding of the spatial dynamics of ILIs and advances in both social media research and social media technologies, namely the Twitter API, and a wider adoption of social media, could create a viable IRTM based on adoption of the following steps:

- The Integration of Commuter Data between the various states, and MSAs
- Breaking down the states with large MSAs
- Inclusion of population density, within both MSAs as well as the states themselves
- Focus on the friends of an identified user to determine if they are ill
• Inclusion of the GeoNames database into the taxonomy to capture geographic information, when the posting is not “tagged”

• A greater emphasis on Google Flu, weight it on par with the CDC data

The implementation of these changes should lead to an increase in the accuracy of the IRTM. As in all new fields, the development of accurate models is an iterative process that builds upon the failure of prior models.
References


Song, Zhao, Martin Ester, and Binay Bhattacharya. "Discovering More Meaningful Regions: A Regularized Geographical Topic Model."


## Appendix One: Tweet Modifiers

<table>
<thead>
<tr>
<th>Modifiers from Tweet</th>
<th>Total Times Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>have</td>
<td>1232</td>
</tr>
<tr>
<td>work</td>
<td>403</td>
</tr>
<tr>
<td>Not</td>
<td>332</td>
</tr>
<tr>
<td>has</td>
<td>322</td>
</tr>
<tr>
<td>sick</td>
<td>309</td>
</tr>
<tr>
<td>Got</td>
<td>256</td>
</tr>
<tr>
<td>ill</td>
<td>209</td>
</tr>
<tr>
<td>I feel</td>
<td>182</td>
</tr>
<tr>
<td>I think</td>
<td>108</td>
</tr>
<tr>
<td>had</td>
<td>101</td>
</tr>
<tr>
<td>don't</td>
<td>87</td>
</tr>
<tr>
<td>shots</td>
<td>84</td>
</tr>
<tr>
<td>better</td>
<td>58</td>
</tr>
<tr>
<td>dr</td>
<td>34</td>
</tr>
<tr>
<td>I hope</td>
<td>33</td>
</tr>
<tr>
<td>health</td>
<td>22</td>
</tr>
<tr>
<td>meds</td>
<td>27</td>
</tr>
<tr>
<td>Term</td>
<td>Count</td>
</tr>
<tr>
<td>--------------</td>
<td>-------</td>
</tr>
<tr>
<td>doctor</td>
<td>25</td>
</tr>
<tr>
<td>Hospital</td>
<td>17</td>
</tr>
<tr>
<td>death</td>
<td>13</td>
</tr>
<tr>
<td>kid</td>
<td>12</td>
</tr>
<tr>
<td>Vaccine</td>
<td>8</td>
</tr>
<tr>
<td>child</td>
<td>7</td>
</tr>
<tr>
<td>coming down</td>
<td>7</td>
</tr>
<tr>
<td>isn't</td>
<td>7</td>
</tr>
<tr>
<td>aint</td>
<td>6</td>
</tr>
<tr>
<td>died</td>
<td>6</td>
</tr>
<tr>
<td>School</td>
<td>5</td>
</tr>
<tr>
<td>daughter</td>
<td>4</td>
</tr>
<tr>
<td>drug</td>
<td>4</td>
</tr>
<tr>
<td>Nurse</td>
<td>4</td>
</tr>
<tr>
<td>Physician</td>
<td>1</td>
</tr>
</tbody>
</table>
Appendix Two: Risk Groups

Risk Group Five
New York
Connecticut
Rhode Island
Virginia
Pennsylvania
District of Columbia

Risk Group Four
North Carolina
New Jersey
Maryland
Indiana
Massachusetts
Kentucky
Michigan
Illinois
Iowa
Minnesota
**Risk Group Three**
California
Nevada
Utah
Tennessee
Idaho
Colorado
Georgia
South Carolina

**Risk Group Two**
Delaware
Maine
Kansas
New Mexico
Louisiana
Alabama
Florida
Ohio
Risk Group One

Oklahoma
New Hampshire
Vermont
Wisconsin
Washington
Oregon
North Dakota
South Dakota