Identifying and Displaying Issues in Public Health through the use of Social Media and an Interactive Web Map

Wesley DeWitt, Therese Norman, Sean Phayakapong

Master of Science in Geographic Information Science (MSGISci)

Department of Geography, California State University, Long Beach

Introduction

An open source approach for analyzing social media data and disseminating results through an interactive web map was developed. This web map allows the user to explore different variables in Los Angeles County. We display geographical patterns of tweets and examine how sentiments expressed on Twitter regarding individuals' physical and emotional health relates to the surrounding environment (Figure 1).

This research enhances our understanding of how to use data from Social Media to analyze a location's impact on public health.

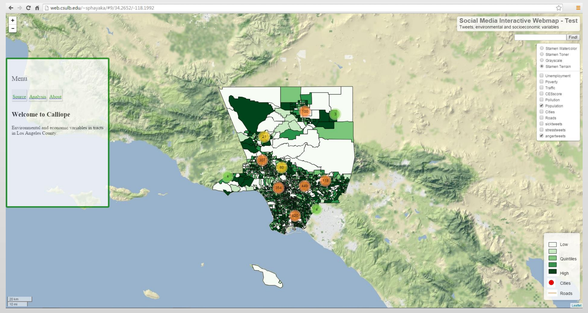

Figure 1. *Snapshot of the web map displaying geotagged Tweets with a background of population per Census tract in LA County*

Methodology

The collected Twitter data as well as socioeconomic and environmental variables were converted into shapefiles using the QGIS leaflet plug-in. These shapefiles were then converted into GeoJSON, so they could be incorporated into a web page using HTML, CSS, JavaScript and the Leaflet API. We also used jQuery to implement other pages for the project's web site.

We performed spatial statistical analysis using ArcMap. The statistical approaches used to analyze the geographical patterns of the Twitter data include heat maps, Getis-Ord Hot Spots and geographically weighted regression.

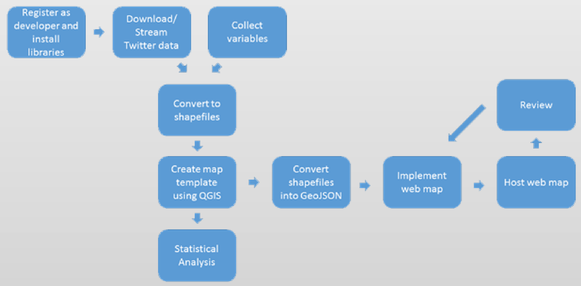

Figure 2. *Workflow diagram*

Geostatistical Analysis

The spatial analysis revealed some statistically significant spatial patterns of the Twitter data. The heat maps showed that there is a higher density of tweets near downtown Los Angeles and almost no "sick" tweets by the coastline (Figure 4). There is a high concentration of "anger" tweets east of Los Angeles and a low concentration close to Malibu/Santa Monica as can be seen in the Getis-Ord Hot Spots maps (Figure 5). The geographically weighted regressions show a significant relationship where the share of anger tweets is low and where the socioeconomic and environmental conditions are better, which mostly occurs in the western part of the county close to the coast (Figure 6).

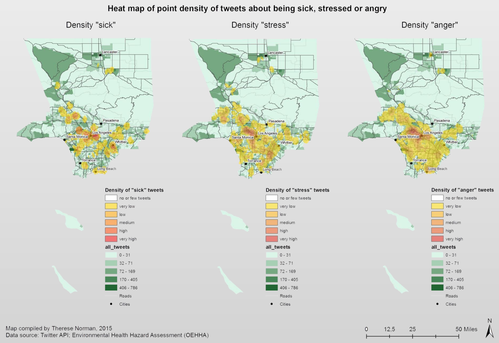

Figure 4. Heat map of point density of tweets about being sick, stressed or angry

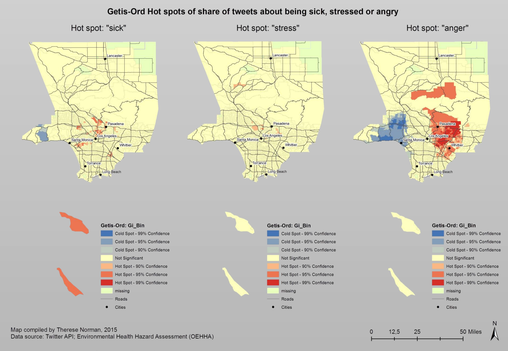

Figure 5. Hot spots of standardized tweets per tract about being sick, stressed or angry

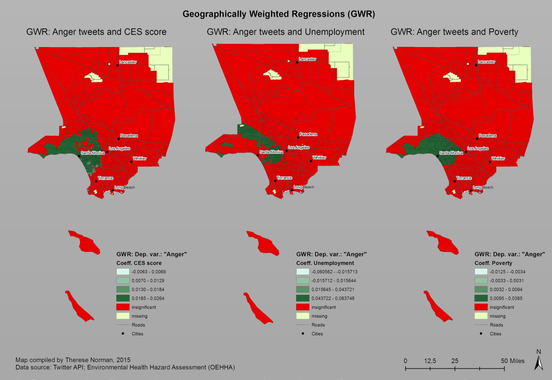

Figure 6. Geographically weighted regressions where share of anger tweets is explained by CES score (vulnerable populations), Unemployment or Poverty

Discussion

One of the limitations of our findings is the inherent bias in the data generated by social media platforms since its users are heavily skewed towards young, affluent whites. Additional bias was generated during the data processing stage since the classification into "sick", "stressed", and "anger" tweets using keywords correctly categorized about 75 percent of the tweets only. Using machine learning with linguistic analysis would have quantified the sentiments, leading to more robust results.

Moreover, the low count of "sick" and "stress" tweets create additional bias of the statistical analysis. For this reason, the regression analysis is focused on the anger tweets. Collecting tweets during a longer time period would have reduced much of the limitations of the data.

Although Twitter data has many limitations, the type of data we collected could be used to uncover information which would previously have been impossible. In particular, policy makers and researchers can use our results to stay informed regarding the population's physical and emotional health and increase their understanding of how a location impacts an individual's health. As a result of our findings, more targeted measures could be implemented to improve the health of the population.

We successfully achieved our goal of showing the viability of web mapping and the use of open source products for data dissemination. This is significant since it allows for a cheap and relatively easy way to disseminate important information to a wide audience.

Data and Data Sources

Geotagged Twitter data were obtained using a Python 2.7 script which incorporates a Streaming API. Public tweets were collected for five weeks between March 17, 2015 to April 9, 2015.

Socioeconomic and environmental variables were collected from the US Census Bureau and the California Communities Environmental Health Screening Tool. In addition, cities and roads were used as reference layers. These were also provided by the US Census Bureau.

Table 1. *List of data and data sources used in the project*

Dataset	Source
Socioeconomic	US Census Bureau
Environmental	California Communities Environmental Health Screening Tool, version 2.0
Reference layers	TIGER / US Census Bureau
Tweets: "sick"	Keywords: am sick, am ill, feeling sick, feeling ill, cough, a cold, runny nose, the flue, a fever, influenza, sore throat, throw up, puke, under the weather, etc.
Tweets: "stress"	Keywords: stress, panic, anxiety, miserable, nervous, sleepless, restless, uneasy, etc.
Tweets: "anger"	Keywords: angry, anger, pissed, annoyed, hate, hating, bitter, enraged, furious, irritated, awful, sucks, horrible, loath, resentment, dread, excruciating, distasteful, painful, terrible, disturbing, gruesome, appalling, unpleasant, etc.

Interactive Web Map

The website has four different tabs (Source Analysis, Map and About) that the user can click on. Each layer in the web map can be toggled, and overlays on top of each other as each is selected. The user can choose between different base maps as well. Tweets are shown as clusters, but as the user clicks on a cluster or zooms in, the individual tweets will be shown as points. If the user clicks on one of the point features of the tweets, a pop-up will display the text of that tweet as well as the date and time the tweet was sent (Figure 3).

The count of "stress" and "sick" tweets collected were relatively low (142 and 210 respectively), while we collected 2060 "anger" tweets and 90,076 total tweets for the county.

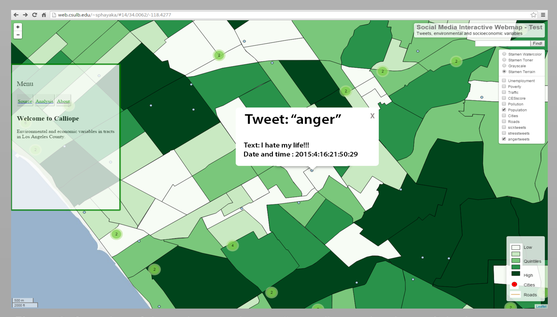

Figure 3. Snapshot of the web map when an "anger" tweet has been clicked on

Conclusion

The project showed the feasibility of collecting and analyzing geotagged social media data generated in Los Angeles County for the purpose of identifying epidemiological patterns that could be presented through an interactive web map. Our results from the spatial analysis indicate that areas with better socioeconomic and environmental conditions have a lowers share of anger tweets.

Additionally, our project can be developed to track the outbreak and spread of certain diseases. Using our method could help health workers identify spatial patterns where the number of individuals experiencing similar symptoms are more concentrated. These patterns can be investigated in order to find out the possible source of the sickness.

The web map's functionality is quite extensive, but our hopes were that it would have featured a live feed with which a user could interact. However, this did not prove possible because we were not able to configure a server for storing the data and using PHP.

Submitted in partial fulfillment of the requirements of the Masters of Science in Geographic Information Science(MSGISci), August 15, 2015.

For additional information please contact:

Wesley DeWitt wesrdewitt@gmail.com
Therese Norman norman.theresec@gmail.com
Sean Phayakapong seanpha22@gmail.com
Web site: http://www.csulb.edu/~sphayaka