Sergio Hernandez

GIS 501 Final Project

Jim Thatcher Ph.D.

Abstract:

When flu and cold season arrives, health care providers and public health officials look to any tool available in order to spot and control outbreaks as early as possible. Any edge to their efforts is in the public interest and should be explored. In the age of social media and online posting, personal information is constantly being uploaded and shared. This wealth of data could potential serve as indicators for a number of patterns. The purpose of this project will be to spatially display concentrations of potential seasonal illnesses based on the collection of social media data. This will be accomplished through the creation of a computer program in python that collects data and then produces a map through ArcGIS (Blake, 2011). The program will use methods already established on the collection of Twitter data and Mongo DB data (Bruns & Yuxian, 2012). No other form of social media will be used for this project and geographic consideration will only be given to data for Tacoma and Seattle, Washington. This project makes no assumptions as to the potential outcome but rather seeks to ask if such indicators could serve the greater good.

Introduction:

During the course of this project it came to my attention that there is a large body of work by academics from a wide range of fields on the role of social media as predictor for health issues and a range of other topics. It light of this I chose to place greater emphasis on the tools being used rather than the academic viability of the data analysis. In this section I discuss the objectives and the data collected in order to achieve those objectives.

*Research Tool/ Objectives*

The primary mode of data collection was through a python script and the Twitter API. The script used the API to gather data posted using a set of hash tag mentioned in the methodology section of this paper. ArcGis 10.2 was used as the geospatial engine for the spatial analysis portion, and .kmz file via HTML for result publication. The object of the script was to automate as much of the process as possible and to set the script to run once every 24 hours. The objective of the entire effort was to provide twitter data and mortality in a web based map.

*Data*

Two primary types of data were necessary for this project. The first consisted of captured twitter feed based on keywords, location and time. The second data came from the Centers for Disease Control (CDC) in the form of mortality statistics from influenza and pneumonia in Seattle and Tacoma for the 49th week of the 2014 calendar year. This mortality data was downloaded as comma separated value and viewed in Microsoft Excel.

Secondary data consisted of a base map provided by ESRI and a municipalities layer obtained from the state of Washington.

Methodology:

The python script searches the twitter feed for the following keywords and is geographically limited to a 50-mile radius from Tacoma. It is further limited by timeframe to the 2014 calendar year but the API restricts places additional temporal limitations on the data passed to the API user.

1. Cold
2. Flu
3. Cough
4. Pneumonia
5. Bronchitis
6. Fever
7. Chills

Once the data is returned to the script, a shapfile is created from the passed data for use in ArcGIS. TwitterData.shp is the converted to a layer by using the “Make Feature Layer Management” tool. The product is a layer file stored in the same location as the original shapefile output of the script. One final file must be created in order to achieve the desired outcome. By using the “layer to KML” conversion, TwitterData.lyr is converted to TwitterData.kmz. This final product can then be sent via FTP to a pre-designated location to be rendered on a webpage and viewed by the Google Earth plugin. The Python script handles all of these steps.

Discussion:

The deviation of output from an image file, (.jpg) was made in order to reduce the number of steps necessary for the desired outcome. The goal was to present a map with easily understood data to a broad audience through a webpage. By producing a .kmz file and sending it to a webpage, a user could interact with the map features. Additionally this would, when effective, result in enhanced understanding by consumers.

All of these steps worked with the exception of the file upload to a website. I am continuing to work on this part of the script and will do so until a successful file transfer occurs.

Script aside, we can see from the data presented that there does not appear to be any correlation of Twitter posts as collected in this exercise, and mortality and morbidity data (see figure 1). The data provided by the CDC is updated for major cities on a weekly basis and is collected locally. This represents the complexity of the underlying mechanisms that exist when one tries to achieve predictive success through social media (Schoen, Gayo-Avello, Metaxas, Mustafaraj, & Strohmaier, 2013). There are simply too many variables to attempt prediction from one isolated medium.

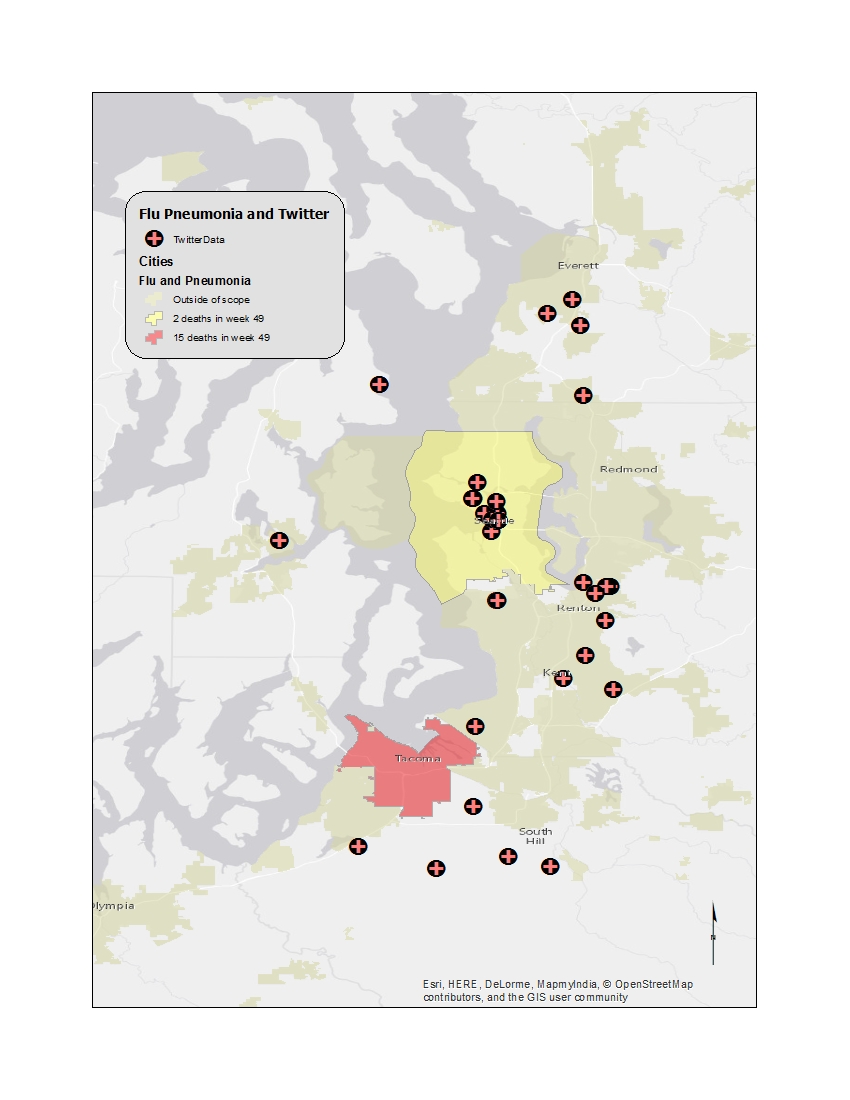


Figure 1.

In Tacoma in the 49th week of December 2014 there have been 15 deaths versus 2 in Seattle. Tweets that contain the keywords are skewed in the opposite direction from mortality rates with far greater number in Seattle than in Tacoma.

Conclusion:

The outcome of the scripting portion of this project, although modified, was able to accomplish most of the initial desired outcomes. With further effort and increased skills with python, full automation can be achieved. However, the data gathered does not reflect any spatial correlation with cdc data. Exploration of existing academic research, suggests that such a correlation might exist but also corroborate the lack of evidence in this type of comparative model. It is evident that with greater data collection from a wider variety of social media there is the potential for indicator development. Expansion of the tools used in this project to diverse social media modalities through API access would lend itself to broader data collection. In the future I will expand the script to incorporate these modalities. However it is my opinion that my efforts are better spent exploring the skills learned in this project and others as a means to an end and that it is ultimately the outcome that matters rather than the tool used. In this instance the tool being solely Twitter data is inadequate. Additionally further pursuit of this particular topic would simply add to an already large body of work.

# Bibliography

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S. H., & Liu, B. (2011, April). Predicting flu trends using twitter data. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on* (pp. 702-707). IEEE.

Blake, R. J. (2011). Current Trends in Online Language Learning. (C. Polio, Ed.) *Annual Review of Applied Linguistics* *, 31*, 19-35.

Bruns, A., & Yuxian, E. L. (2012). Tools and methods for capturing Twitter data during natural disasters. *First Monday* *, 17* (4).

Department of Health and Human Services. (2014). *Morbidity and Mortality Weekly Report*. Retrieved 11 10, 2014, from Centers for Disease Control: <http://wonder.cdc.gov/mmwr/mmwrmort.asp>

Harald Schoen, Daniel Gayo-Avello, Panagiotis Takis Metaxas, Eni Mustafaraj, Markus Strohmaier, Peter Gloor, (2013) "The power of

prediction with social media", Internet Research, Vol. 23 Iss: 5, pp.528- 543. 10.1108/IntR-06-2013-0115

Hutwagner, L. C., Maloney, E. K., Bean, N. H., Slutsker, L., & Martin, S. M. (1997). Using laboratory-based surveillance data for prevention: an algorithm for detecting Salmonella outbreaks. *Emerging infectious diseases*, *3*, 395-400

Lim, D., & Peters, T. (2014, November 13). *Use Social Media In The Fight Against Communicable Diseases*. (A. inc., Producer, & AOL Inc.) Retrieved December 10, 2014, from Tech crunch: http://techcrunch.com/2014/11/13/social-media-and-the-fight-against-ebola/

Mandeville, K. L., Harris, M., Thomas, H. L., Chow, Y., & Seng, C. (2014). Using Social Networking Sites for Communicable Disease Control: Innovative Contact Tracing or Breach of Confidentiality? *Public Health Ethics* *, 7* (1), 47-50.