R Spatial Analysis of Geopolitical Twitter Data

Point Pattern and Sentiment Analysis of Tweets about Donald Trump | Allison Cahanin GEOG 665 | 15 FEB 21

Introduction

The purpose of this project is to explore methods of extracting information from big data sources and evaluate how this will impact sociological research. Unlike traditional methods of social research, which have established rules and methodologies to maintain the quality of the results, gathering social information from digital sources is largely an uncharted field (*Diaz-Bone, 2020*). By documenting challenges with analysis and areas where results could be misleading, this project aims to create discussion about effective methods of tracking and measuring cultural trends using the statistical programming language R.

In order to test this methodology for social media analysis, this research will perform point pattern and sentiment analysis on Twitter data gathered during the attempted Capitol siege, Donald Trump's 2nd impeachment, and President Biden's inauguration. These historic events produced trending hashtags that polarized Twitter users into two fairly clear cut groups: supporters and non-supporters of Donald Trump. This research will operate on the assumption that Tweets collected with these hashtags can be categorized as either positive or negative Tweets, although in reality the topic is certainly more nuanced than this (*Bail, 2014*).

Background

This research will evaluate the pattern of positive and negative Tweets using spatial analysis functions available in R that are suitable for point data mapped in a projection that preserves area or distance. The categorization will be based on a list of positive and negative hashtags and sentiment will be evaluated using the NRC Emotion Lexicon available in R. While more advanced methods such as supervised machine learning exist, NRC analysis is suitable for this exploratory project.

Twitter Data

- Collected from **Jan 13-21st 2021**
- Downloaded Tweets based on trending hashtags surrounding the attempted Capitol siege, Donald Trump's 2nd impeachment, and President Biden's inauguration
- **~50,000** total tweets with hashtags and text
- **1245 negative** geolocated Tweet points
- **1412 positive** geolocated Tweet points

#AmericaOrTrump
#WorstPresidentEver
#ByeByeTrump
#TrumpsLastDay #GTFO
#HesGone

#MAGAForever #TeamTrump
#TrumpTrain #ElectionFraud
#TrumpWon
#ThankYouTrump

Preparation and Cleaning

- Extracted lat long coordinates of Tweets with location metadata
- Reprojected points and boundary shapefile into EPSG 2136 (US National Atlas Equal Area) using spTransform() function
- Clipped positive and negative Tweet point data against generalized US boundary spatial polygon using subset() function
- Cleaned text data using functions from the tidyverse package to remove stop words, punctuation, hyperlinks, etc
- Reorganized text into word frequency data frames using unnest() function

Hypotheses

H_0: The pattern conforms to what would be expected from a CSR process.

H_1: The pattern is significantly different from a CSR process.

Nearest Neighbor ✕ **Lhat** ✕ **D(d) Khat$_1$-Khat$_2$** ✕ **Moran's I**

Methodology

Pattern Description

- Compare mean centers and standard distance ellipses of geolocated positive and negative Tweets

library(maps)
library(mapproj)
library(sp)
library(ggplot2)

First Order Neighbors

- Run a Monte-Carlo simulation based hypothesis test to examine pattern
- Compare expected mean nearest neighbor distance of 999 simulated means with observed mean

library(splancs)
library(ggplot2)
library(hrbrthemes)

Higher Order Neighbors

- Run a Monte-Carlo simulation based hypothesis test to examine pattern
- Compare observed Lhat against significance envelope from 99 simulations

library(splancs)

Bivariate Point Pattern

- Determine if positive Tweets are a random subset of all Tweets about Trump using D(d) Khat$_1$ - Khat$_2$ Monte Carlo Simulation test

library(splancs)

Spatial Autocorrelation

- Create choropleth maps of positive and negative Tweets
- Undertake a Moran's I spatial autocorrelation test

library(rgdal)
library(GISTools)
library(spdep)
library(RColorBrewer)
library(classInt)

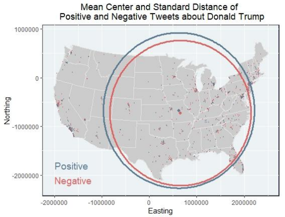
Sentiment Analysis

- Create NRC emotion lexicon based sentiment analysis bar plots
- Build word cloud visualizations of most frequently used words

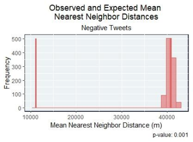
library(tidyverse)
library(Rtools)
library(wordcloud2)
library(syuzhet)

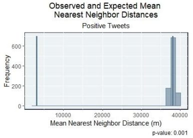
Results

Pattern Description

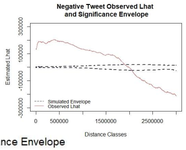


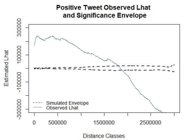
First Order Neighbors



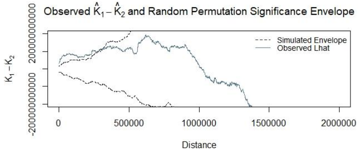


Higher Order Neighbors



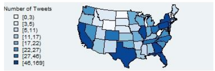


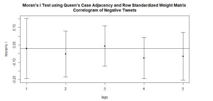
Bivariate Point Pattern



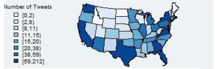
Spatial Autocorrelation

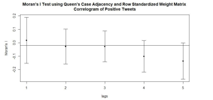
Choropleth of Negative Tweets



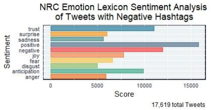


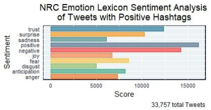
Choropleth of Positive Tweets



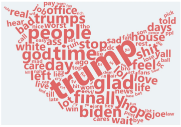


Sentiment Analysis

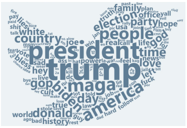




Negative Tweet Word Frequency



Positive Tweet Word Frequency



Conclusions

Although the results are significantly impacted by the lack of point data and clustering around metropolitan areas, performing this research has highlighted these limitations for future consideration. Since not all Tweets have location metadata, Tweets should be gathered using broader search terms or over a longer time period. Additional limitations include the difficulty to characterize sentiment using a lexicon that does not understand satire or cynicism (*Weichbold, 2020*). Future research should consider more sophisticated methods such as supervised machine learning.