Do Twitterstorms Reflect Real Storms? Locating Areas Impacted by Hurricane Harvey Through Analysis of Related Tweets

Ben Barrett

October 24, 2017

Introduction

Ever since its widespread introduction in the mid-2000s, social media has played a role in shaping the lives of individuals - with a notable influence in social capital acquisition (1), health and consumption-related actions (2; 3), and contemporarily, governmental elections and policy decisions. Given its far-reaching impact, social media has been utilized in natural disaster response, with a variety of platforms being used to assist in disaster relief coordination and information spread (4; 5). Beyond the benefit of boosting situational awareness among individuals experiencing a natural disaster "on the ground", social media output has been used to map natural disasters in real-time in order to assess the spread of damage (6), drawing upon the understanding that the content of social media is related to the proximity of a natural disaster - that is, the closer one is to a natural disaster, the more likely it is that social media output is related to that signficant event (7). This concept was recently put to use in the analysis of Twitter activity produced before, during, and after Hurricane Sandy, which found that the amount of hurricane-related tweets was associated with proximity to locations damaged by Sandy, and per-capita tweet frequency was strongly correlated with the extent of economic damage experienced by an area (8).

Building upon this established framework, this paper sought to analyze Twitter data related to Hurricane Harvey, which struck Southeastern Texas as a Category 4 hurricane on August 25, 2017 and continued to travel up into Southern Louisiana, finally ending around September 1, 2017 (9). Geostatistical analyses were performed to assess whether any clustering was present within the Hurricane Harvey Twitter data and identify areas that were hardest hit by the storm. Findings indicated that Hurricane Harvey tweets did cluster together, with the greatest tweet intensity occurring in Houston, Texas.

Data and Methods

Data

The Python library 'Tweepy' was used to connect to the Twitter Streaming API and download relevant tweets (code section "Streaming in tweets via the Python program twitter_streaming.py"). The Python program, used to initiate the live stream was adapted from code provided by Mikael Brunila (10). The stream was set to search for the hashtags 'HurricaneHarvey' and 'HurricaneHarveyRelief', and was started at 9:10AM on 9/1/2017. The live tweet stream was stopped at 9:56AM on 9/4/2017, which resulted in a program run time of 36 hours, 46 minutes and a total of 3,289,336 KB (3.290 GB) of Twitter data collected. This corresponds to 1,491.086 KB of data per minute, on average. All of the output was saved as a JSON file, twitter_data.json, and stored in Dropbox.

Because of the large data file size, the twitter_data.json file was first streamed into R with the R packages jsonlite (11) and rdrop2 (12), using a handler to randomly sample 25% of each page of JSON lines (code section "Loading and Processing", lines 113-137). The import procedure found 420,294 records, each corresponding to a tweet collected during the Twitter live stream. The random sample yielded a study dataset of 22,689 tweets.

Methods

Data cleaning was performed with the assistance of the R packages tibble (13), dplyr (14), tidyr (15), stringr (16), and data.table (17). Tweets were restricted to only those that had a user location recorded, which gave a sample of 15,535 tweets (code section "Loading and Processing", lines 139-143). Then, tweets were further restricted to those either with geographic coordinates already saved, those made in Texas or Louisiana (tweet location), those with a user location set within Texas or Louisiana, or those with a user location of geographic coordinates (code section "Data Cleaning", lines 185-620). This left 1,842 tweets, 21 of which had geographic coordinates already saved, and 80 of which had a location of tweet creation recorded (state and city). Next, user locations set as geographic coordinates were extracted as user geographic coordinates (code section "Data Cleaning", lines 622-626). Tweet locations and user locations that were assigned a state (Texas or Louisiana), but were not assigned a city, were removed from the dataset if a corresponding tweet or user geographic coordinate did not exist (code section "Data Cleaning", lines 642-650). Then, tweet locations that were assigned a city and state had their geographic coordinates imputed by assigning a random latitude and random longitude, bound between the respective city's most extreme border points, as identified by Google Maps. These imputed tweet location geographic coordinates were assigned as a tweet's coordinates if coordinates did not already exist from when the tweet was originally made (code section "Data Cleaning", lines 652-993). Following this, Google Maps was used to assess whether user geographic coordinates that had originally been set as the user location fell within Texas or Louisiana, and if not, these observations were deleted if a tweet geographic coordinate was not already assigned (code section "Data Cleaning", lines 995-1053). User locations that were assigned a state and city then had their geographic coordinates imputed using the same method as tweet locations, and the user geographic coordinates from the original user location, as well as the imputed user location geographic coordinates, were assigned as a tweet's coordinates if coordinates were not already assigned (code section "Data Cleaning", lines 1055-2248). This process left a final sample size of 1,256, with each tweet having geographic coordinates and a city and state associated with it.

The above tweet geocoding methods primarily rely on the user location (entered once a user first start Twitter), rather than the location associated with an individual tweet, as most users turn the tweet location data off. In this case, however, these methods should be appropriate - as people who fled the impacted area before Hurricane Harvey hit will still have their user location linked to the impacted area.

Spatial statistical analyses were performed with the assistance of the R packages maps (18), spatstat (19), splancs (20), maptools (21), rgeos (22), and rgdal (23). These analyses were restricted to the spatial area of Texas and Louisiana, as this was the area most impacted by Hurricane Harvey. Maps of Texas and Louisiana were first loaded into R as map objects, and then converted to spatial polygon objects, projected as Lambert Cylindrical Equal Area using kilometers as the units, and merged to form one common spatial polygon (code section "Spatial Statistics", lines 2345-2360). The Hurricane Harvey tweet coordinates were then converted to a spatial points dataframe, projected as Lambert Cylindrical Equal Area using meters as the units, and then transformed into a ppp object - necessary for spatial point pattern analyses with the spatstat package (19) - using the border coordinates of the common spatial polygon of Texas and Louisiana as a window, while applying an internal conversion to shift the units from meters to kilometers (code section "Spatial Statistics", lines 2362-2384). During this process, 29 tweet coordinates were rejected as lying outside the specified window, which is the result of some tweet locations falling just outside of Texas or Louisiana during the imputation procedure.

Once data was set in the proper format, the spatial intensity of the collected Hurricane Harvey tweets was estimated using the kernel approach, with the bandwidth selected via the (1989) Berman and Diggle method (24), designed to minimize the mean squared error of the kernel smoothing estimator (code section "Spatial Intensity", line 2440). Spatial intensity of point pattern data is simply the expected number of events in an area. The kernel approach estimates intensity by calculating the spatial intensity (λ) within a buffer around a given location (s), with the radius of the buffer being termed the bandwidth (τ), and the entire intensity estimation being weighted by the distance event points are to the location of estimate (s) - with the weight estimator being termed the kernel (κ). This equation can be represented by:

$$\hat{\lambda}(s) = \frac{1}{\delta(s)} \sum_{i=1}^{n} \frac{1}{\tau^2} \kappa \left(\frac{(s_i - s)}{\tau} \right)$$

where, in addition to the parameters listed above, n is the number of events in a given area, s_i designates an event, and $\delta(s)$ is an edge correction factor (25).

Following the estimation of spatial intensity, the K-function for the Hurricane Harvey tweets was calculated (code section "Spatial Clustering", line 2456). The K-function is a means of quantifying spatial clustering among point pattern data, which can be conceptualized by event locations occurring close to other events. The K-function can be defined as:

$$\kappa(h) = \frac{E}{\lambda}$$

where $\kappa(h)$ is the K-function, reliant on distance (h), E is the number of events within distance h of an arbitrary event, and λ is the spatial intensity, which for these purposes is assumed to be constant. In the case of the Hurricane Harvey tweets, the constant spatial intensity (λ), was calculated to be 0.001527626 expected tweets per kilometer. In practice, the K-function is estimated by:

$$\hat{\kappa}(h) = (\hat{\lambda}^{-1})(\frac{1}{n})\sum_{i=1}^{n}\sum_{j\neq i} w_{ij}^{-1} I(d_{ij} \leq h)$$

where, in addition to the parameters listed above, n is the number of events i, j denotes any other point besides event i, d_{ij} is the radius of a circle centered at point s_i within the area of interest, w_{ij} is a weight - defined as the proportion of circumference of the circle centered at point s_i that falls within the area of interest, and $I()$ is the indicator function. The estimated K-function (y-axis) is plotted against distance (x-axis), and compared to the K-function expected under complete spatial randomness (CSR), where events occur independently of one another. An estimated K-function that is greater than the K-function expected under CSR suggests the presence of clustering (26).

Finally, in order to assist with interpretability, the estimated constant spatial intensity for the Hurricane Harvey tweets was multiplied by the estimated K-function in order to quantify the observed number of Hurricane Harvey tweets by distance, compared to the expected number of Hurricane Harvey tweets by distance under CSR (26) (code section "Spatial Clustering", lines 2466-2472).

Results

Table 1 displays the frequency of tweets made using the hashtag #HurricaneHarvey or #HurricaneHarveyRelief, collected between 9:10AM on September 1, 2017 and 9:56AM on September 4, 2017, with tweet locations restricted to only those made in Texas and Louisiana. From the table it can be observed that more Hurricane Harvey tweets came out of Texas across all days of data collection as compared to Louisiana. The two most popular days of Hurricane Harvey tweeting were September 1st (N = 350 tweets) and September 2nd (N = 499 tweets), which intuitively makes sense, as these were the dates that were closest to the final landfall of the storm. It should be noted that the low tweet counts for September 4th should be interpreted with caution, as the Twitter live stream was ended on 9:56AM of that day.

Figure creation was made possible through the use of the R packages ggplot2 (27), ggmap (28), gridExtra (29), and grid (30). Figure 1 presents the locations of Hurricane Harvey tweets, aggregated from all days of data collection. From this figure, it is clear that there are several pockets of high tweet density - primarily occurring around major cities, with notable clusters in Dallas, Austin, San Antonio, Houston, and New Orleans. For an even more refined visualization of the Hurricane Harvey tweet locations, the geocoded tweets were stratified by day and are presented in figure S1 of the supplementary material.

Table 1: Frequencies of Hurricane Harvey tweets, categorized by day and state of creation.

Day of Tweet Creation	Texas	Louisiana
September 1, 2017	335	15
September 2, 2017	473	26
September 3, 2017	315	16
September 4, 2017	71	5
All Days	1194	62

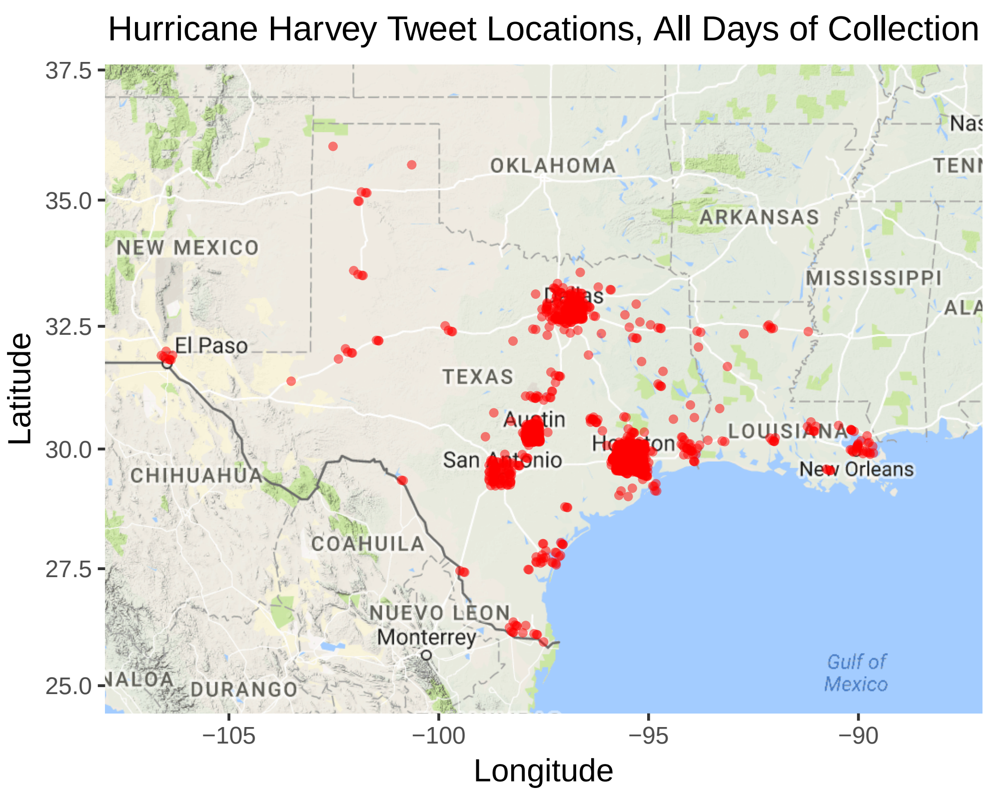


Figure 1: Locations of tweets made with #HurricaneHarvey or #HurricaneHarveyRelief, collected between 9:10AM on 9/1/2017 and 9:56AM on 9/4/2017.

Figure 2 presents the results of the spatial intensity estimation of Hurricane Harvey tweets. With spatial intensity hotspots being marked by brighter, more yellow areas, it can be observed that there are defined areas of high spatial intensity around Dallas, Austin, San Antonio, and Houston, Texas. The spatial intensity of Hurricane Harvey tweets is by far the greatest around Houston, which is to be expected, as this was one of the areas worst impacted by the storm (31).

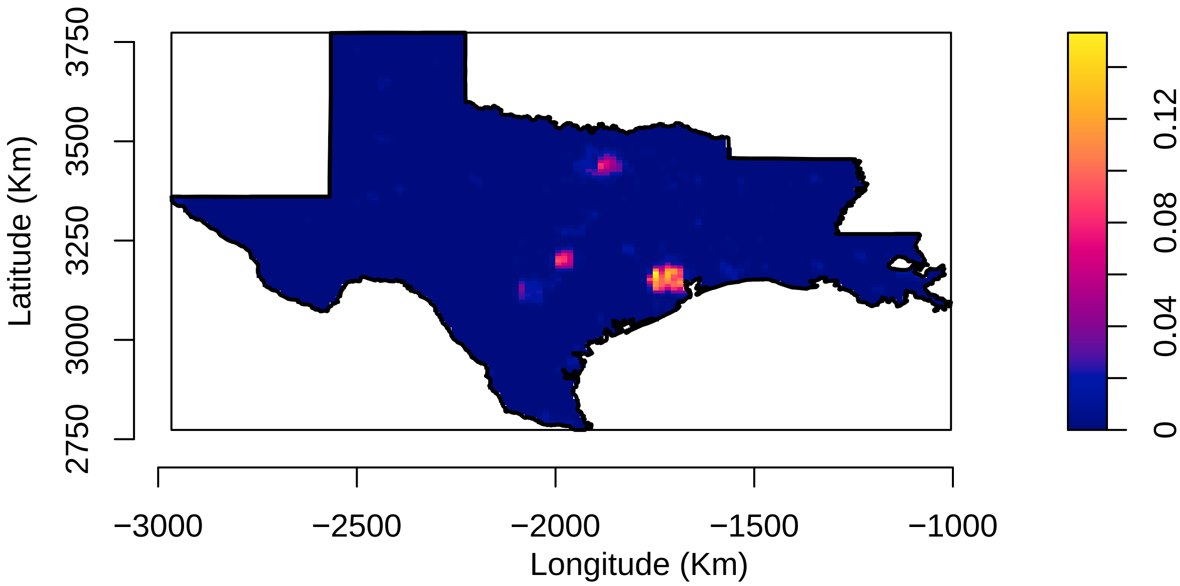


Figure 2: Spatial intensity of Hurricane Harvey tweets, calculated using kernel estimation with a bandwidth of 6.606245 kilometers.

Figures 3 and 4 display the results of the K-function calculation for Hurricane Harvey tweets. In Figure 3, the observed K-function for Hurricane Harvey tweets is clearly greater across all distances as compared to the K-function expected if the tweets were distributed with complete spatial randomness. This suggests that clustering is present in the Hurricane Harvey tweet data. Figure 4 presents the same graphical distribution as that in Figure 3 with an altered y-axis scale to assist with interpretability. From this figure, it can be seen that the observed number of tweets at every distance is far greater than the number expected under complete spatial randomness, indicating that event locations occur close to one another and there is clustering among the Hurricane Harvey tweets.

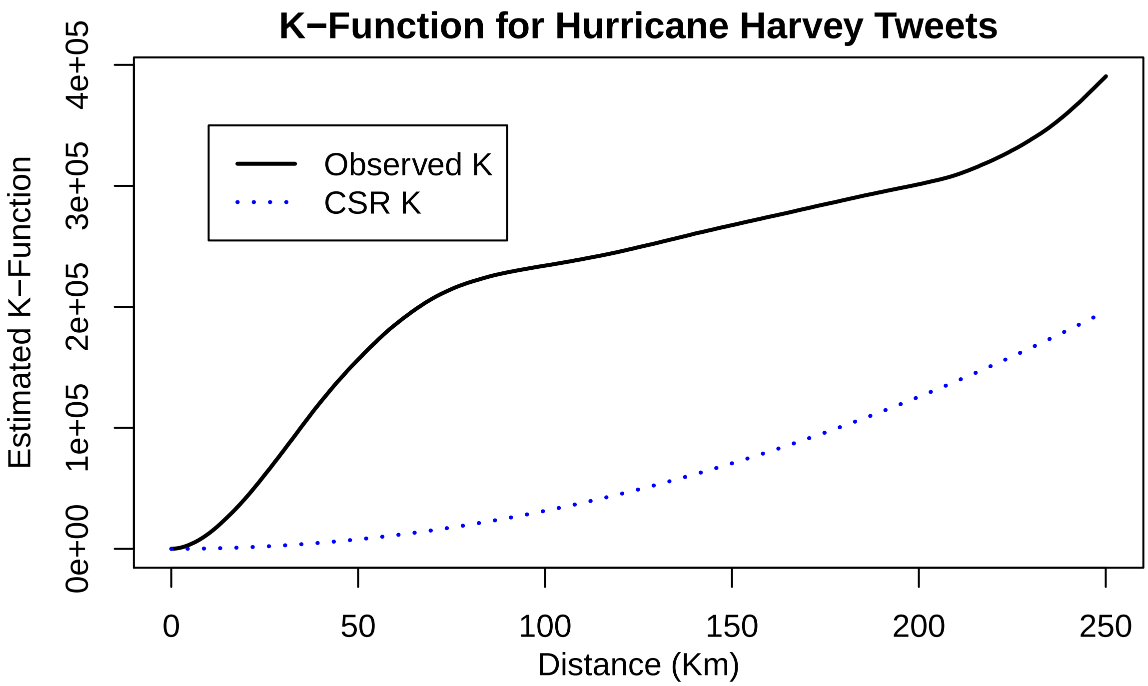


Figure 3: Estimated K−function for Hurricane Harvey tweets, comparing the observed K of tweets (black) to the K that would be expected under complete spatial randomness (CSR; blue).

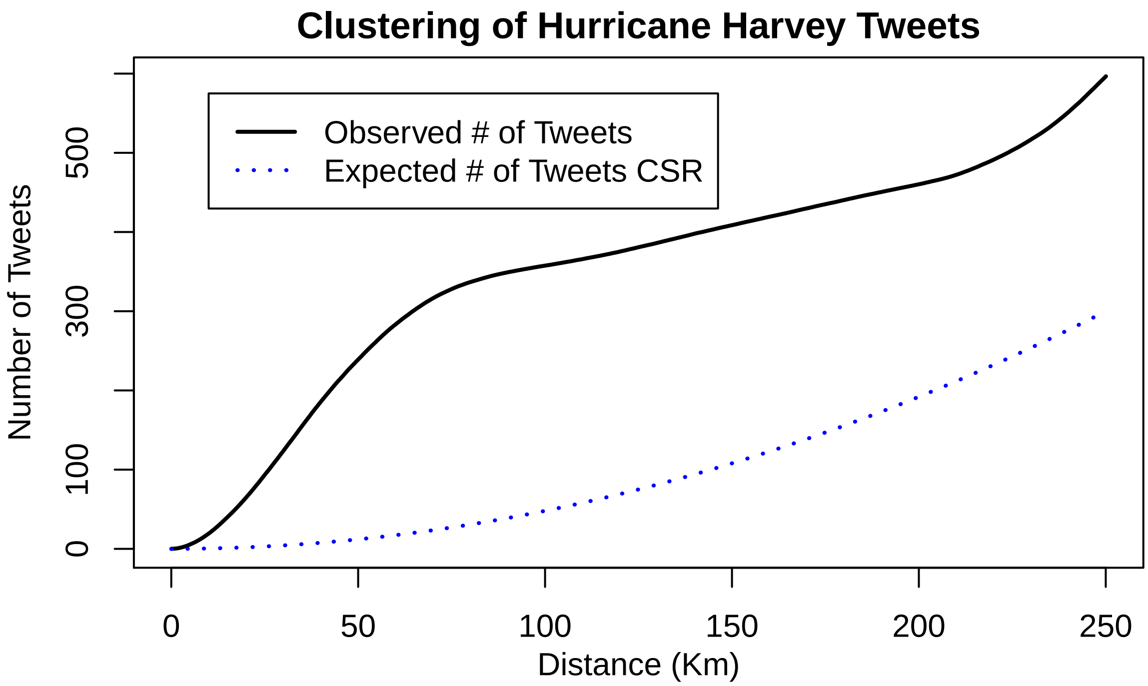


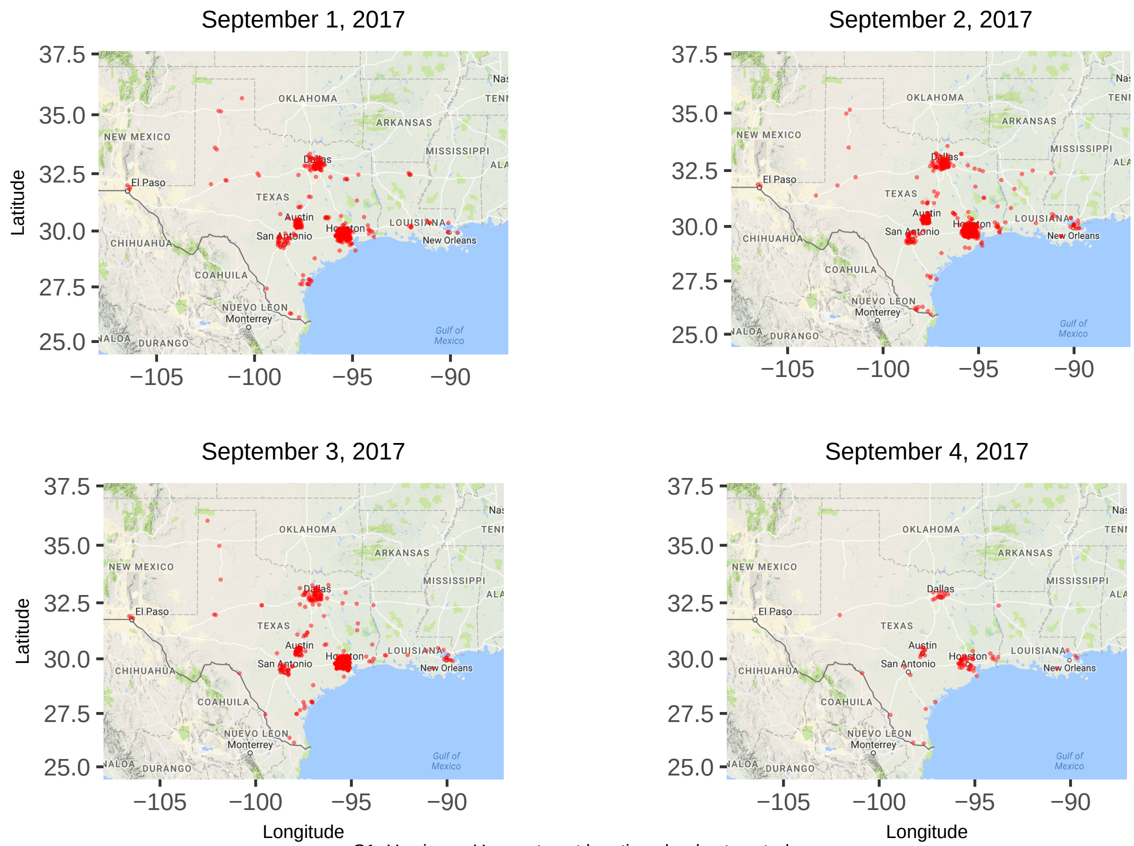
Figure 4: Observed number of Hurricane Harvey tweets (black), as compared to the number of tweets as would be expected under complete spatial randomness (CSR; blue).

Discussion

This study examined the geospatial distribution of Texas and Louisiana tweets made with the hashtag #HurricaneHarvey and #HurricaneHarveyRelief in a short period immediately following landfall of the iconic storm. Results indicated that there was clustering among the Hurricane Harvey tweets, and that tweets occurred with the greatest intensity in Houston, Texas. These results suggest that social media data can be used as a tool to track and quantify the severity of a natural disaster in real-time.

While this study possessed several strengths, it was not without its limitations and areas for improvement. Firstly, a sensitivity analysis could be performed in the future that assigns new random coordinates within the coordinate imputation procedure, and then observes whether the same spatial clustering and intensity results are obtained. Such an analysis would provide an indication as to the robustness of the spatial dependence within the Hurricane Harvey tweet data. Second, when the tweet coordinates were imputed, the city borders were treated as rectangles, which usually is not the case in reality. Future analyses could impute the coordinates with shapefiles of the city borders for more accurate estimates. Third, a per-capita correction could be performed on the tweet data to help account for the fact that most of the identified areas of high spatial intensity occurred around cities, and it could be observed whether this has any influence on results. Finally, the majority of this data was collected after Hurricane Harvey had already struck. In future analyses, it would be quite interesting to collect Twitter data before, during, and after a major storm to 'live-track' the storm and make predictions about locations worst impacted.

Supplementary Material



S1: Hurricane Harvey tweet locations by day tweeted

References

1: Ellison, N.B., Steinfield, C., & Lampe, C. (2007). The benefits of Facebook "friends:" Social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication*, 12(4), 1143-1168.

2: Chou, W.Y.S., Hunt, Y.M., Beckjord, E.B., Moser, R.P., & Hesse, B.W. (2009). Social media use in the United States: Implications for health communication. *Journal of Medical Internet Research*, 11(4), e48.

3: Rayat, A., Rayat, M., & Rayat, L. (2017). The impact of social media marketing on brand loyalty. *Annals of Applied Sport Science*, 5(1), 73-80.

4: Gao, H., Barbier, G., & Goolsby, R. (2011). Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems*, 26(3), 10-14.

5: Yin, J., Lampert, A., Cameron, M., Robinson, B., & Power, R. (2012). Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems*, 27(6), 52-59.

6: Middleton, S.E., Middleton, L., & Modafferi, S. (2014). Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems*, 29(2), 9-17.

7: De Albuquerque, J.P., Herfort, B., Brenning, A., & Zipf, A. (2015). A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *International Journal of Geographical Information Science*, 29(4), 667-689.

8: Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Van Hentenryck, P., Fowler, J., & Cebrian, M. (2016). Rapid assessment of disaster damage using social media activity. *Science Advances*, 2(3), e1500779.

9: The Weather Channel. (2017). Historic Hurricane Harvey's Recap. https://weather.com/storms/hurricane/news/tropical-storm-harvey-forecast-texas-louisiana-arkansas [accessed October 15, 2017].

10: Brunila, M. (2017). Scraping, extracting and mapping geodata from Twitter. http://www.mikaelbrunila.fi/2017/03/27/scraping-extracting-mapping-geodata-twitter/ [accessed September 1, 2017].

11: Jeroen Ooms (2014). The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. arXiv:1403.2805 [stat.CO] URL https://arxiv.org/abs/1403.2805.

12: Karthik Ram and Clayton Yochum (2017). rdrop2: Programmatic Interface to the 'Dropbox' API. R package version 0.8.1. https://CRAN.R-project.org/package=rdrop2.

13: Kirill Müller and Hadley Wickham (2017). tibble: Simple Data Frames. R package version 1.3.4. https://CRAN.R-project.org/package=tibble.

14: Hadley Wickham, Romain Francois, Lionel Henry and Kirill Müller (2017). dplyr: A Grammar of Data Manipulation. R package version 0.7.4. https://CRAN.R-project.org/package=dplyr.

15: Hadley Wickham and Lionel Henry (2017). tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions. R package version 0.7.1. https://CRAN.R-project.org/package=tidyr.

16: Hadley Wickham (2017). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.2.0. https://CRAN.R-project.org/package=stringr.

17: Matt Dowle and Arun Srinivasan (2017). data.table: Extension of `data.frame`. R package version 1.10.4. https://CRAN.R-project.org/package=data.table.

18: Original S code by Richard A. Becker, Allan R. Wilks. R version by Ray Brownrigg. Enhancements by Thomas P Minka and Alex Deckmyn. (2016). maps: Draw Geographical Maps. R package version 3.1.1. https://CRAN.R-project.org/package=maps.

19: Adrian Baddeley, Ege Rubak, Rolf Turner (2015). Spatial Point Patterns: Methodology and Applications with R. London: Chapman and Hall/CRC Press, 2015. URL http://www.crcpress.com/Spatial-Point-Patterns-Methodology-and-Applications-with-R/Baddeley-Rubak-Turner/9781482210200/.

20: Barry Rowlingson and Peter Diggle (2016). splancs: Spatial and Space-Time Point Pattern Analysis. R package version 2.01-39. https://CRAN.R-project.org/package=splancs.

21: Roger Bivand and Nicholas Lewin-Koh (2017). maptools: Tools for Reading and Handling Spatial Objects. R package version 0.8-41. https://CRAN.R-project.org/package=maptools.

22: Roger Bivand and Colin Rundel (2017). rgeos: Interface to Geometry Engine - Open Source (GEOS). R package version 0.3-23. https://CRAN.R-project.org/package=rgeos.

23: Roger Bivand, Tim Keitt and Barry Rowlingson (2017). rgdal: Bindings for the Geospatial Data Abstraction Library. R package version 1.2-6. https://CRAN.R-project.org/package=rgdal.

24: Berman, M. & Diggle, P. (1989) Estimating weighted integrals of the second-order intensity of a spatial point process. *Journal of the Royal Statistical Society, series B*, 51, 81-92.

25: Curriero, F. (2017). Statistical analysis of spatial point pattern data in public health. Lecture given on February 14, 2017, as a component of the Johns Hopkins Bloomberg School of Public Health course, Spatial Analysis III: Spatial Statistics.

26: Curriero, F. (2017). Statistical analysis of spatial point pattern data in public health. Lecture given on February 16, 2017, as a component of the Johns Hopkins Bloomberg School of Public Health course, Spatial Analysis III: Spatial Statistics.

27: H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.

28: D. Kahle and H. Wickham. ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161. URL http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf.

29: Baptiste Auguie (2016). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.2.1. https://CRAN.R-project.org/package=gridExtra.

30: R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

31: Carlsen, A. & Lai, K.K.R. (2017). Where Harvey Hit Hardest Up and Down the Texas Coast. *The New York Times*. https://www.nytimes.com/interactive/2017/09/01/us/hurricane-harvey-damage-texas-cities-towns.html [accessed October 16, 2017].