Applied Methodology for Identifying Hurricane-Induced Social Media Signal Changes in Vulnerable Populations

Rachel Samuels[1] and John E. Taylor, Ph.D.[2]

[1]Graduate Student, Dept. of Civil and Environmental Engineering, Georgia Institute of Technology, 790 Atlantic Dr. NW, Atlanta, GA 30332-0355. E-mail: rsamuels3@gatech.edu
[2]Professor, Dept. of Civil and Environmental Engineering, Georgia Institute of Technology, 790 Atlantic Dr. NW, Atlanta, GA 30332-0355. E-mail: jet@gatech.edu

ABSTRACT

Current crisis informatics research using social media to identify human location and behavior has neglected to identify which populations may be prevented from using social media during a disaster. Unfortunately, most crisis informatics focuses on either individual postings or large-scale analyses of signal changes and is potentially overlooking or misrepresenting vulnerable populations. To assess this, we utilize scaled spatial nets and disaggregated population and vulnerability index data in order to identify relationships between social media signal deviation and vulnerable populations within areas that experienced substantial infrastructural damage from Hurricane Harvey. Many studies have identified a strong positive correlation between areas experiencing high amounts of hurricane damage and increased Twitter activity; however, we find that highly damaged areas with more elderly people, disabled people, and people without access to vehicles instead have a significant negative correlation with Twitter activity during disaster. In defining this relationship, we show that some vulnerable populations have decreased instead of increased social media visibility during disasters. These findings identify demographic inequalities in the scalar representation of data from humans-as-sensors in disaster.

BACKGROUND

As new and varied forms of information become available to researchers during crises, there has been a substantial push towards finding ways of applying that information to emergency responder priorities on the ground and at higher levels of decision-making. More people than ever are living in areas susceptible to catastrophic disasters because of urban sprawl (Allen, 2006) and increased extreme weather patterns from climate change (Adachi et al., 2017). As such, our ability to effectively and accurately utilize all forms of available information will be critical to reducing loss of human life and increasing the resilience of our cities. While worsening extreme events are becoming more of a certainty than a possibility (Hauer et al., 2016), the extent of the impact on humans and society can be mitigated with improved resource planning and resource agility through increased real-time information on human location, activity, and responsiveness (Roshan et al., 2016). Ultimately, more efficient distribution of our resources will depend on what we know about the people in the path of these extreme events.

One source of data on human behavior and the distribution of need during crises is the data generated through human interaction with communication networks. These data sources, such as Twitter (Spence et al., 2015), FourSquare (Aubrecht et al., 2017), and cellular data (Jennex, 2012), are particularly useful as they each can have unique user identifiers, a location attribute, and sometimes a topical attribute, such as the text of a Tweet or the type of store someone has visited. The incorporation of these data attributes has been useful for tracking individuals'

mobility and the influence of a disaster on that mobility (Wang et al., 2017), the change of individuals' sentiment in response to different disaster impact levels (Wang and Taylor, 2018), the need and availability of resources (Huang and Xiao, 2015). Research using spatiotemporal aggregation to compare two spatial datasets has shown that bursts of social media behavior and disaster-related posts can indicate areas of relatively higher hurricane damage (Kryvasheyeu et al., 2016) and the location of flooding (de Albuquerque et al., 2015).

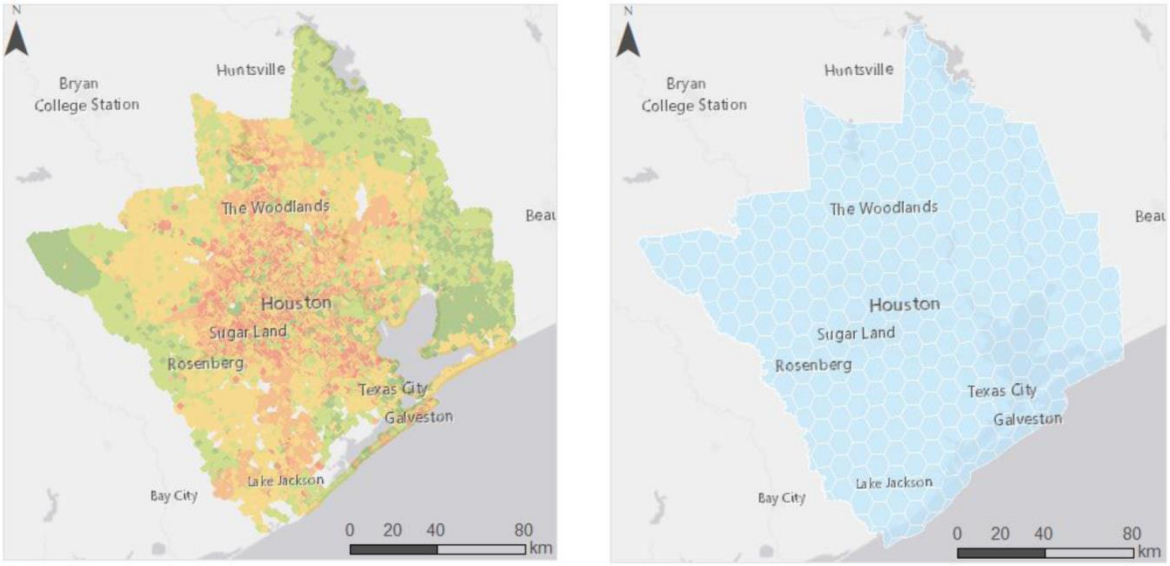

Figure 1. (Left) The plotted redistribution of Houston's population at 30mx30m scale. Green indicates less population density; red indicates higher population density. (Right) A hexagonal grid covering the majority Houston, TX.

As compelling as these findings are, big data research has often been critiqued for overlooking human variability and for mistaking big data for complete data (Blumenstock, 2018; Gandomi and Haider, 2015). These two fallacies can also be found intertwined in some aspects of existing crisis informatics, as one of the critical dilemmas with humans-as-sensors analyses is that humans are not reliable sensors. We do not transmit consistent, coordinated, or comparable information through public data channels that can be continuously accessed by connected emergency responders or data analysts. The rush to utilize information produced by humans-as-sensors in disasters has neglected to incorporate some of the distinctions of diverse human response and capabilities in its analysis. One study found that 50% of deaths from Hurricane Sandy occurred in an area with a complete lack of Twitter activity (Shelton et al., 2014). Another study focused on Hurricane Harvey found that some areas' decreases in Twitter activity during a hurricane correlate as strongly with damage as others areas' increases (Samuels et al., 2018). We do not currently understand what factors could contribute to some populations being represented by social media in an emergency while others simply disappear.

In examining why some people stop Tweeting during disaster, it is critical to incorporate human variability into the analysis to shed light on how variations in populations is associated with Tweet deviations. We chose to incorporate demographic data as a proxy for human variability. Previous research has shown that the people most endangered by disasters, such as lower socioeconomic groups, the elderly, and the disabled, are those with the least robust resources (Bian and Wilmot, 2017). Lower socioeconomic groups are more likely to lose internet access during power loss, as internet usage by people in a lower income bracket is often

facilitated by free Wi-Fi hotspots located at places of employment or cafes (Khan et al., 2016). During a disaster, those hotspots are no longer available due to closures or travel impedances.

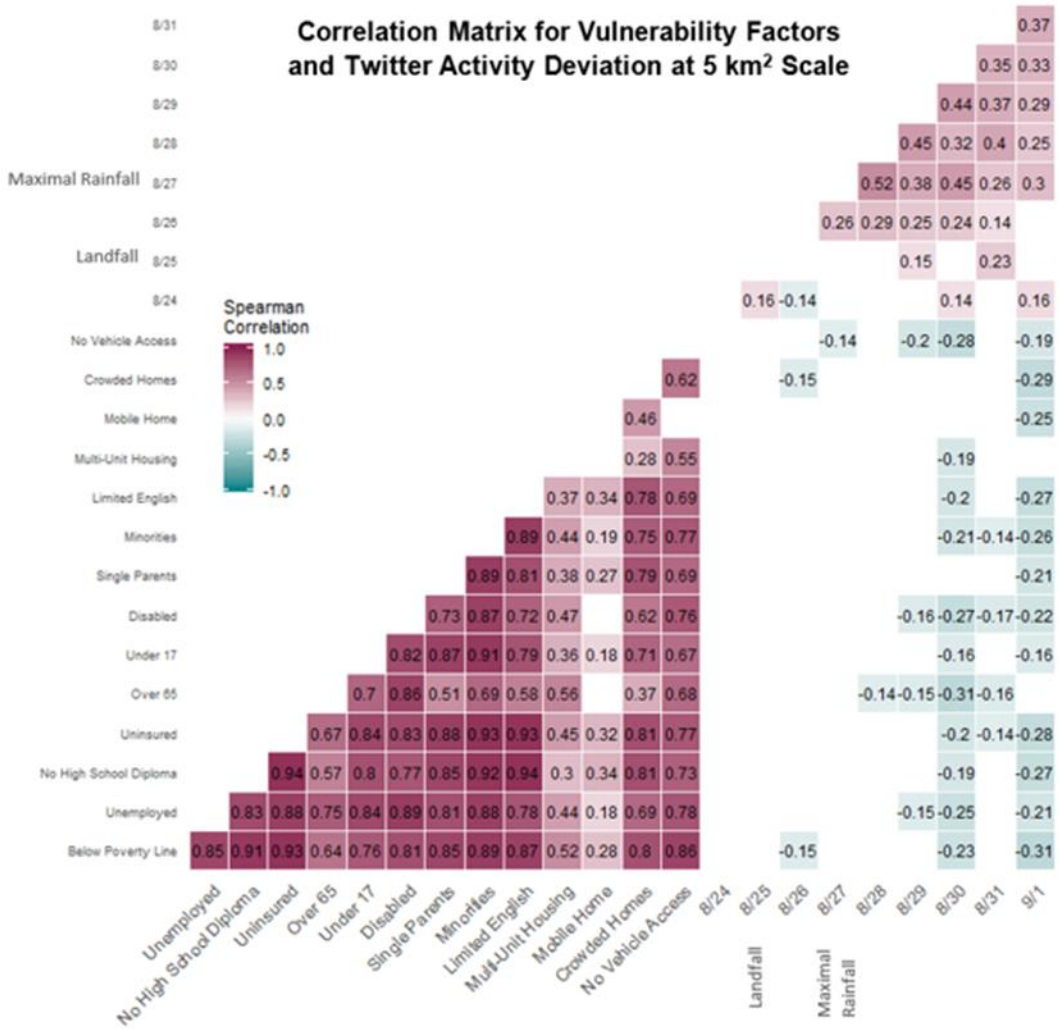

Figure 2. Correlation matrix showing the Spearman rank correlation coefficient between the estimated vulnerable populations and the daily social media signal deviation during Hurricane Harvey where p<0.05. Blank squares indicate a p>0.05.

Increases in Twitter activity are more common than decreases in Twitter activity in a disaster; however, those decreases do not appear to be random, and studies that assume the decreases to be negligible may be overlooking vulnerable populations. Understanding how social factors influence individuals' interactions with sensors and technologies in a disaster is critical to understanding which endangered populations can be identified through social media analyses. As our cities become smarter and most people become more connected to technology, the technological data signal from vulnerable populations, especially in disasters, is necessary for understanding what populations could be left behind in our future cities. Our research objective is to identify the demographic and infrastructural vulnerability factors related to discrepancies in the humans-as-sensors signal response. We examine this in the specific context of Houston, Texas during and after Hurricane Harvey. Through the identification of demographic factors that correlate with social media silence, we can start to identify which kinds of people crisis

informatics analyses tend to exclude.

METHODS

Social Media Data Acquisition: The geolocated Twitter data for the greater metropolitan area of Houston for three weeks prior to and one week following Hurricane Harvey's landfall were streamed through the Twitter API (Wang and Taylor, 2015). These Tweets were aggregated by day and plotted in ArcGIS. Tweets from the days prior to the recognition of Harvey's formation in the Atlantic were designated as "steady state" (August 3rd to August 17th), and Tweets from one day prior to landfall and seven days following landfall were designated as "perturbed state" (August 24th to September 1st).

Distributing Demographic Data: Vulnerability indices are based on foundational social research addressing the socioeconomic, mobility, disability, and resource-availability factors that cause discrepancies in the abilities of people to rebuild after disaster. Most of these vulnerability indexes are defined at the county level and use census data and other social indicator data, such as school performance and community connectedness. However, recent research has shown that hurricane damage can vary at small spatial scales (Wurman and Kosiba, 2018), and we theorized that discrepancies between social media signal behavior responses would occur at the sub-county level. Thus, in order to understand the societal response to hurricane damage, we needed to reduce the scale at which we could analyze human demographic data from the census tract and Zip Code Tabulation Area (ZCTA) scales. The ZCTA shapefiles that were used to match the demographic census data were downloaded from the Harris County GIS data portal. First, we redistributed the population within each census tract to a more granular scale, and then we assigned the attributes of a social vulnerability index to the redistributed population points.

The geographic information systems field has historically utilized National Land Cover Database (NLCD) data to increase the granularity of the census data with substantial accuracy (Reibel and Agrawal, 2007). The NLCD contains a raster file with 30 meters (m) by 30 m cells that have been classified, through satellite imagery, as one of 16 classes. The classification includes four classes of developed land: open space, low intensity, medium intensity, and high intensity. We extracted the raster cells from the 2011 dataset that were classified as developed and were located within the greater metropolitan Houston Area. Using ArcGIS' Raster to Point function, we then transformed each of the raster cells into points located at the center of each cell and spatially joined these points by count into the census tracts for Houston. Using the counts of each type of NLCD class and the population record for each census tract, we used multiple linear regression to determine the contributing coefficients of each land type with respect to population. The results have an adjusted R-squared of 0.8317 and a model p-value of <0.001.

For the vulnerable population assignment, we used the Social Vulnerability Index (SVI) developed at the Centers for Disease Control (CDC) (Flanagan et al., 2011). This data is available at the census tract level as well as the county level, and the individual demographic factors utilized in the CDC's analysis are available in their component parts. These factors, and the thorough documentation of the incorporation of those components, made the CDC's SVI ideal for our analysis. We multiplied the percentages of the population ascribing to each of the salient factors identified by the SVI by the identified population of the 30mx30m representative points by that percentage, accounting for mean error of the assignations. When we spatially aggregated the population points, we also aggregated the ZCTA-specific demographic data into the aggregation polygons. Using this method, we developed an estimate for the number of people or buildings ascribing to the following 14 categories within each area: people without vehicle

access, people with limited English skills, minorities, single parents, disabled people, people over 65, people under 17, people without a high school diploma, unemployed people, people below the poverty line, crowded homes, mobile homes, and homes within multi-unit complexes.

Scalar Aggregation Nets: We designed six spatial nets to catch the population data and the Twitter stream for each day. The nets are composed of a series of equal-area, hexagonal polygons that cover the greater Houston area. Using ZCTAs can exacerbate the modifiable areal unit problem due to their varying sizes and shapes (Jelinski and Wu, 1996). The census tracts within Houston range in size from 0.16 square kilometers (km^2) to 677.20 km^2. For this analysis, we developed nets that could be deployed across a large area, and could be scaled according to the intended research design. Hexagons are better suited for tiling large geospatial areas because of their scalability and the reduction in sampling bias from edge effects (Carr et al., 1992). The six hexagonal nets consist of hexagons with square areas of 1 km^2, 2 km^2, 5 km^2, 10 km^2, 15 km^2, and 20 km^2. A comparison of the redistributed census data and one of the hexagonal nets is presented in Figure 1**.**

Population and Social Signal Analysis: We spatially joined the daily Tweet sets via count and the redistributed population data with the socially vulnerable population attributes via sum to each of the hexagonal nets. In order to be able to focus on areas experiencing significant amounts of infrastructural hurricane damage, we spatially joined the Hurricane Harvey Federal Emergency Management Agency (FEMA) Building Level damage assessments to our hexagonal nets. This dataset contains a list of building locations and a damage rating from 1 (minor damage) to 4 (destroyed). Using the previously described sets of steady and perturbed state data, we standardized the Tweet counts of each day of the perturbed state using the mean and standard deviation of the steady state. From these datasets, we extracted the polygons that had nonzero Twitter activity during the steady state period and the areas that contained at least one FEMA Building Level damage assessment of "4", or "destroyed". From this data pool, we generated a series of rank correlation matrices. To generate the matrices, we calculated Spearman's rho and incorporated a threshold p-value of 0.05 to identify any statistically significant relationships between the variation in vulnerable populations and the variation in signal response at increasing scales to catastrophic hurricane damage.

RESULTS

The correlation between Twitter activity signal changes and demographic vulnerability was most pronounced at the 5 km^2, so the matrix for the 5 km^2 hexagonal grid is presented as Figure 2. As could be expected, the presence of one vulnerability factor in an area has a strong positive relationship with the presence of another. Equally, the signal behavior of an area following the day of maximal infrastructure damage (August 27th, two days following landfall), correlates with the other days' behavior, but not the days prior to the 27th. Within the correlations between deviation and the vulnerability factors, we see some clear distinctions. Primarily, we see that the significant relationships are all negative, indicating a relationship between vulnerable populations' decreasing microposts during a disaster. Additionally, the correlations between vulnerability indicators (excepting populations without vehicle access) and activity deviations are only significant following the day of maximal rainfall. The factors with the most consistent significant correlations with activity decreases are: lacking vehicle access, being disabled, and being over the age of 65. The strongest negative correlations ($\rho \geq -0.28$) appear 3-5 days following the day of maximal rainfall within populations that are over 65, lack vehicle access, are in crowded homes, and/or are below the poverty line. With regard to differences between scalar

aggregations, there was an identifiable increase in the number of significant decreasing correlations with vulnerable populations and social media signal deviation with increasing areal size. The correlations also tended to strengthen with increasing scale up to the 15km^2 nets.

DISCUSSION

These results show that some demographic factors correspond with the direction of social signal response in the face of disaster. The clearest links to decreased signal are with the elderly, the disabled, and those who are unable to easily access transportation away from the disaster area. Although many would assume the elderly in general do not contribute to the social signal data stream and thus do not produce a negative signal deviation, some of these populations—or people living less than a mile in radius from those populations, consistently and throughout Houston—were Tweeting prior to the hurricane and then stopped. For those without access to a vehicle and for the disabled, evacuating in the face of any oncoming disaster is difficult (Flanagan et al., 2011). This is doubly true for Harvey, for which an evacuation notice for Houston was not sent until almost immediately prior to the storm, and then only for a few counties (Houston-Galveston Area Council, 2017). Data shadows (areas in which no data is available) do not indicate whether the lack of signal is due to evacuation or due to a human behavior change; however, for more immobile populations, the data shadows are more likely due to human behavior change in response to damaged infrastructure or nonfunctioning technology.

The increase in correlation strength with increasing time (especially more than four days beyond the day of maximum flooding) could indicate either extended evacuation by vulnerable populations (unlikely due to the reasons stated above) or ongoing, substantially damaged energy or housing infrastructure. Both potential reasons indicate a lack of urban resilience surrounding the vulnerable populations, further indicating a need to adjust existing social media tool analyses.

This study's limitations are primarily centered on the distribution of vulnerable populations to very small, sub-ZCTA scales. Based on our regression analysis, the population distribution using NLCD data was mostly accurate; however, the social vulnerability factors were distributed as a percentage of the population and not weighted by other factors. Social media itself is notoriously noisy and less than reliable, and recent changes in metadata transitions from Instagram to Twitter have made that worse; however, by aggregating the data and focusing on both large-scale trends and signal deviation, we hope to have more thoroughly isolated humans and behaviors affected by the storm. Additionally, the NLCD distribution is limited to nighttime and not daytime accuracy. Based on the Twitter content we reviewed, many businesses were closed and many people stayed in their homes regardless. We are also unable to isolate the causes, social or technical, of the behavioral differences in vulnerable populations. Finally, the vulnerability factors were isolated in our covariance matrix; clearly, many of those factors are themselves correlated, and more accurate relationships would be determined through regression analyses. We will be focused on modeling this behavior and the root causes for distinctions between scalar aggregations in our future work.

CONCLUSIONS

To date, research has not investigated what socioeconomic factors might influence the increase or decrease of signal data from humans-as-sensors in a disaster. Ultimately, tools seeking to use social media data (and any form of data from humans-as-sensors that may develop) need also to understand what populations do and do not continue to contribute to certain technological data streams during a disaster. We developed and applied a methodology

for identifying demographic and infrastructural vulnerability factors related to neighborhood-scale discrepancies in humans-as-sensors behavioral response. We found that, in Houston, TX during Hurricane Harvey, there was a significant decrease in social media signals from vulnerable populations such as the elderly, disabled, and those who lack access to a vehicle. The results of our research indicate that humans-as-sensors data are not sufficient for identifying crises specific to these populations. As urban analytics and decision-making begin to utilize more big data produced by the interactions between people and technology, we need to incorporate the discrepancies between data produced by general and vulnerable populations such that the people in the most danger are not the ones with the least visibility.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Division of Information and Intelligent Systems (IIS) Grant No. 1760645. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Adachi, S. A., Nishizawa, S., Yoshida, R., … Tomita, H. (2017) Contributions of changes in climatology and perturbation and the resulting nonlinearity to regional climate change, *Nature Communications*, 8, 1.

Allen, K. (2006) Community-based disaster preparedness and climate adaptation: local capacity building in the Philippines, *Disasters*, 30, 1, 81–101.

Aubrecht, C., Özceylan Aubrecht, D., Ungar, J., Freire, S., & Steinnocher, K. (2017) VGDI – Advancing the Concept: Volunteered Geo-Dynamic Information and its Benefits for Population Dynamics Modeling, *Transactions in GIS*, 21, 2, 253–276.

Bian, R. & Wilmot, C. G. (2017) Measuring the vulnerability of disadvantaged populations during hurricane evacuation, *Natural Hazards*, 85, 2, 691–707.

Blumenstock, J. (2018) Don't forget people in the use of big data for development, *Nature News*.

Carr, D., Olsen, A., & White, D. (1992) Hexagon Mosaic Maps for Display of Univariate and Bivariate Geographical Data, *Cartography and Geographic Information Systems*, 19, 4, 228–236.

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.

Flanagan, B. E., Gregory, E. W., Hallisey, E. J., Heitgerd, J. L., & Lewis, B. (2011) A Social Vulnerability Index for Disaster Management, *Journal of Homeland Security and Emergency Management*, 8, 1.

Gandomi, A. & Haider, M. (2015) Beyond the hype: Big data concepts, methods, and analytics, *International Journal of Information Management*, 35, 2, 137–144.

Hauer, M. E., Evans, J. M., & Mishra, D. R. (2016) Millions projected to be at risk from sea-level rise in the continental United States, *Nature Climate Change*, 6, 7, 691–695.

Houston-Galveston Area Council (2017) Brazoria, Chambers, Galveston, Harris and Matagorda Hurricane Evacuation Zip-Zones, *Houston Emergency Operations Center*, , 1.

Huang, Q. & Xiao, Y. (2015) Geographic Situational Awareness: Mining Tweets for Disaster Preparedness, Emergency Response, Impact, and Recovery, *ISPRS International Journal of*

Geo-Information, 4, 3, 1549–1568.
Jelinski, D. E. & Wu, J. (1996) The modifiable areal unit problem and implications for landscape ecology, *Landscape Ecology*, 11, 3, 129–140.
Jennex, M. E. (2012) Social Media – Viable for Crisis Response?, *International Journal of Information Systems for Crisis Response and Management*, 4, 2, 53–67.
Khan, A. S., Rahman, A. ur, & Qazi, L. T. (2016) The Relationship Between Internet Usage, Socioeconomic Status, Subjective Health and Social Status, *Business & Economic Review*, 8, Special Edition, 67–82.
Kryvasheyeu, Y., Chen, H., Obradovich, N., … Cebrian, M. (2016) Rapid assessment of disaster damage using social media activity, *Science Advances*, 2, 3, 1–11.
Reibel, M. & Agrawal, A. (2007) Areal interpolation of population counts using pre-classified land cover data, *Population Research and Policy Review*, 26, 5–6, 619–633.
Roshan, M., Warren, M., & Carr, R. (2016) Understanding the use of social media by organisations for crisis communication, *Computers in Human Behavior*, 63, 350–361.
Samuels, R., Taylor, J., & Mohammadi, N. (2018) The Sound of Silence: Exploring How Decreases in Tweets Contribute to Local Crisis Identification, *Proceedings of the 15th ISCRAM Conference*, , May, 1–9.
Shelton, T., Poorthuis, A., Graham, M., & Zook, M. (2014) Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of 'big data', *Geoforum*, 52, 167–179.
Spence, P. R., Lachlan, K. A., Lin, X., & del Greco, M. (2015) Variability in Twitter Content Across the Stages of a Natural Disaster: Implications for Crisis Communication, *Communication Quarterly*, 63, 2, 171–186.
Wang, Q. & Taylor, J. E. (2015) Process Map for Urban-Human Mobility and Civil Infrastructure Data Collection Using Geosocial Networking Platforms, *Journal of Computing in Civil Engineering*, 30, 2, 04015004-1–11.
Wang, Y. & Taylor, J. E. (2018) Coupling sentiment and human mobility in natural disasters: a Twitter-based study of the 2014 South Napa Earthquake, *Natural Hazards*, 92, 2, 907–925.
Wang, Y., Wang, Q., & Taylor, J. E. (2017) Aggregated responses of human mobility to severe winter storms: An empirical study, *PLoS ONE*, 12, 12, 1–15.
Wurman, J. & Kosiba, K. (2018) The role of small-scale vortices in enhancing surface winds and damage in Hurricane Harvey (2017), *AMS: Monthly Weather Review*, , 2017, MWR-D-17-0327.1.