Shaping Urban Resilience: Whether Social Media Data Can Aid in Improving Disaster Management

A Thesis Presented to the Faculty of Architecture and Planning

COLUMBIA UNIVERSITY

In Partial Fulfillment of the Requirements for the Degree

Master of Science in Urban Planning

by

JIACHENG ZHOU

Advisor: Professor Leah Meisterlin

Reader: Anthony Vanky

May 2019

Abstract

In order to shape urban resilience, it is necessary to understand disaster risks to get better disaster response. Twitter allows people to collect abundant real-time or historical social media data via its API, which gradually make it a repository for disaster-related information collection.

This study has two research objectives. The first is to evaluate whether Twitter data can reflect the emergency and vulnerability and thus aid in disaster response when Hurricane Harvey struck Houston in 2017. The second is to evaluate whether Twitter data can be used to perform damage assessment after Hurricane Harvey.

Three new conceptions are defined to perform evaluation: tweet awareness (TAw), tweet activity (TAc) and tweet focus (TFo). By comparing with other variables such as normalized average proximity (NAP), social vulnerability index (SVI) through spatiotemporal analysis, main conclusions are drawn as the following.

First, the temporal distribution of tweets is periodic: the tweets at night are much more than that in the daytime and there exists the “outbreak” time of the tweets.

Second, when the hurricane is getting closer to the land, the TAw is increasing and vice versa, which reflect the emergency situation of hurricane temporally.

Third, there is no statistically significant relationship between TAw and NAP based on county-level data.

Fourth, the relationship between TFo and SVI is not statistically significant and thus, the twitter data could not reflect the social vulnerability.

Next, there is no significant relationship between them spatially and it is not feasible to perform rapid assessment of damage loss.

Last but not least, the highest clustered point (the points share the same coordinate) of hurricane-related tweets is located in the University of Houston Downtown, which indicates that main active users of Twitter might be college school students.

Acknowledgements

First, I would give my grateful thanks to my instructor, Professor Leah Meisterlin, for her dedication, insight, and patient feedback throughout the research process. This thesis will not be finished without her. Another appreciation is to my reader, Professor Anthony Vanky, whose Urban Informatics inspired many aspects of this thesis, especially on how to use python (Jupyter Notebook) to retrieve Twitter data via an API.

My academic career in GSAPP, Columbia is compact and full of challenges, and I would like to thank my professors for their patience for answering my all kinds of questions and thank my colleagues for their suggestions on revision.

Last but not least, enormous appreciation to my parents. I am not able to continue this master degree without their support, either physical or mental.

Contents

List of Figures

List of Abbreviations

ATSDR … Agency for Toxic Substances and Disease Registry

API … Application Programming Interface

CRT … Climate Resilience Toolkit

DCPC … Damage Claim Per Capita

DHS … Department of Homeland Security

DIRR … Disaster-related Ratio

FEMA … Federal Emergency Management Agency

H-GAC … Houston-Galveston Area Council

HHDC … Hurricane Harvey Data Call

ICT … Information Communication Technology

NAP … Normalized Average Proximity

NFIP … National Flood Insurance Program

NHC … National Hurricane Center

NOAA … National Ocean and Atmospheric Administration

NYCDCP… New York City Department of City Planning

SMART … Social Media Analytics and Reporting Toolkit

SVI … Social Vulnerability Index

TAw … Tweet Awareness

TAc … Tweet Activity

TFo … Tweet Focus

TDI … Texas Department of Insurance

TNRIS … Texas Natural Resources Information System

UNISDR … The United Nations Office for Disaster Risk Reduction

100RC … 100 Resilient Cities

1. Introduction and Background

1.1 Urban Resilience to Disasters

Resilience is the “ability to withstand shocks and recover from the failure” (Yamagata & Maruyama, 2016, p.3). Due to a lack of resilience, many problems and crises can emerge in cities when disasters strike. The concept of resilience was initially used in ecology, and then in economics and social and political science. Urban resilience, as an urban planning terminology, has only been proposed in the last few decades. “100RC” (100 resilient cities, a program created by the Rockefeller Foundation) defines urban resilience as “the capacity of individuals, communities, institutions, businesses, and systems within a city to survive, adapt, and grow no matter what kinds of chronic stresses and acute shocks they experience” (Rockefeller Foundation, 2016, para.2). Urban resilience to disasters is of great significance as it measures to what extent cities, where more than 54.82% of population (World Bank, 2018) on earth dwell, could recover after a disaster hits. In order to shape urban resilience, it is necessary to understand disaster risks and then perform a good management on risk reduction. According to “Sendai Framework for Disaster Risk Reduction (2015-2030)” drafted by the United Nations Office for Disaster Risk Reduction (UNISDR) (2015), disaster risk reduction needs “inclusive risk-informed decision-making” (UNISDR, 2015, p.13) via “real-time access to reliable data” (UN, 2015, p.15). Thus, attaining real-time data has become a foundational prerequisite to shaping urban resilience. However, the collection of timely data is always a big challenge, especially when a severe disaster strikes.

1.2 Social Media Data

Fortunately, with the development of information communication technology (ICT), nowadays social media can provide public safety institutions with netizen's real-time personal updates, including messages, pictures and reposts, all of which could be obtained legally.

Twitter is one of the most widely used in the world. Compared to other social media, Twitter allows people to collect abundant social media data, including geo-located data via its API (application programming interface), a set of communication protocols which enables developers to collect open data from other users more easily (Hoffman, 2018). Naturally, such openness allows it to be not only a platform for communication, but also a repository for disaster-related information collection. Such use of social media in recent disasters (e.g. Hurricane Sandy in 2012) has been well documented by scholars (Kate, 2014 & Hughes et al., 2014). The Department of Homeland Security Science and Technology Center for Excellence (2018) has developed a system, Social Media Analytics and Reporting Toolkit (SMART) to enhance the real-time decision-making ability based on Twitter and other social media platforms. In the future, retrieving real-time Twitter data from the affected areas might play a much more significant role in the natural disaster relief management when a natural disaster strikes.

1.3 Hurricane Harvey

On the early morning of August 26, 2017, Hurricane Harvey made landfall and, in the following several days, people located in Texas suffered storm rain and flood and were desperate for instant rescue from the outside. It was a Category 4 storm and the second costliest hurricane (NHC, 2018) on record in U.S history (Fig.1). Total damage loss is approximately $125 billion (NHC, 2018). More than 56,000 calls into 911 within 15 hours overwhelmed the official emergency response system during this crisis (Yang, 2017). According to The Washington Post

(Wax-Thibodeaux, 2018), the storm lasted for nearly five days and destroyed more than 300,000 building structures and 500,000 automobiles (Fig.2), and even worse, nearly 42 percent of the residents in affected areas have declared that they have not received any aid to rebuild their homes. The reason why people who were trapped in affected areas could not attain on-time rescue is not only because of the lack of rescue teams, but also the lack of location information of trapped people within the affected areas. Thus, using rapid and real-time response platforms such as Twitter is of great imperativeness.

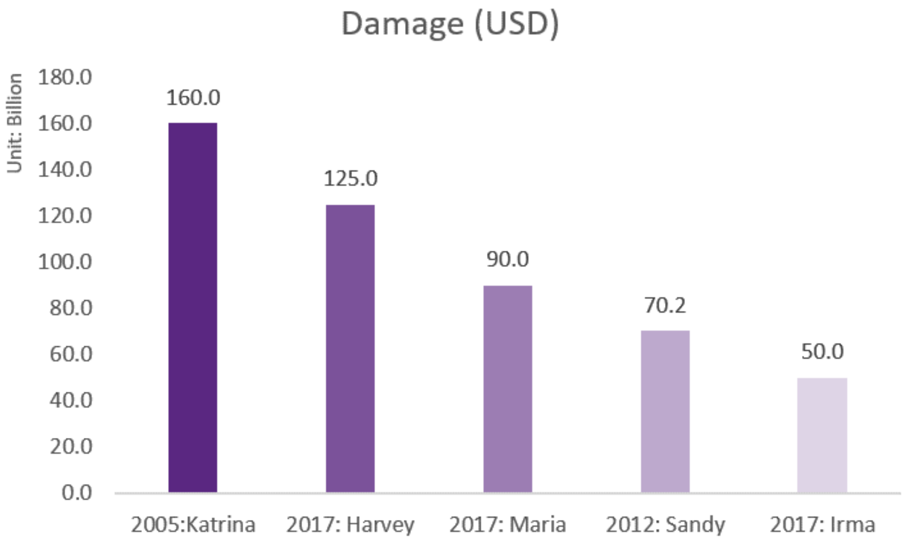


Fig. 1 Top 5 costliest hurricanes in US history (inflation-adjusted to 2017 USD)
Data Source: National Centers for Environmental Information (2017)

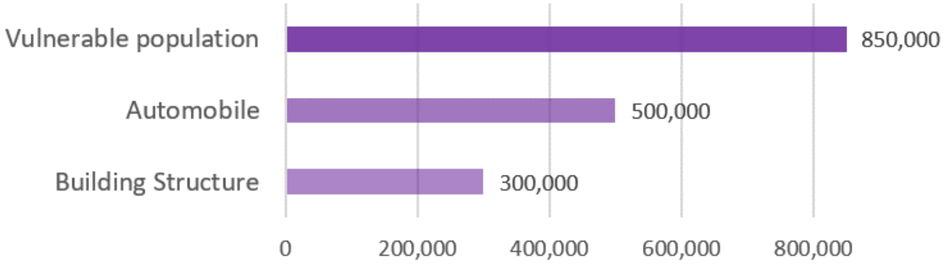


Fig. 2 Physical Damage loss in Hurricane Harvey
Data Source: Washington Post (2018)

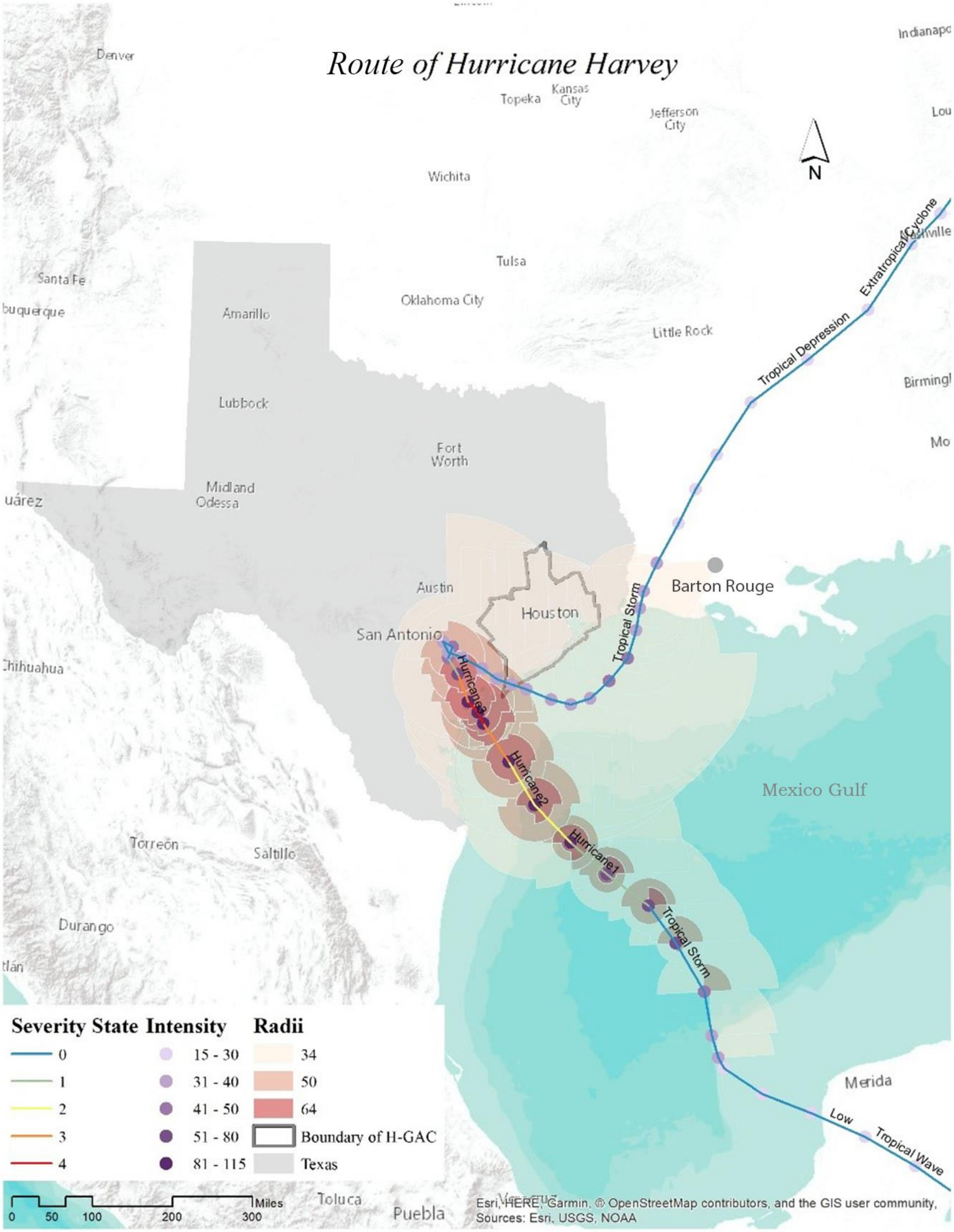


Fig. 3 Affected areas during the Hurricane Harvey

Data Source: NOAA (2017)

1.4 Study Area

Figure 3 mainly shows the route of Hurricane Harvey and the position of each point in this figure (both dark and light purple) represents the location of the hurricane center. The location of the hurricane center is estimated by National Ocean and Atmospheric Administration (NOAA) every six hours. The index of intensity and severity was becoming larger when it was marching towards the land from the Mexico Gulf. Not until the hurricane arrived in Barton Rouge did the radius of it become 0. During the process from landing on to arriving at Barton Rouge, Houston area was totally covered within the buffer of the hurricane based on radii. As is known, Houston is the most populous city in Texas and the fourth most populous city in the U.S., with a census-estimated population of 2.328 million in 2017 (City of Houston, Planning and Development Dept., 2018). Considering the current data sources and the statistical approach, Houston-Galveston Area Council (H-GAC) is selected as my study area. H-GAC is used instead of 'Houston-Galveston Area Council' in the following for convenience. H-GAC is a 13-county region with an area of 12,500 square miles and 6,862,641 people.

Geographically, H-GAC is located on a costal prairie in Texas (Fig.4) and the soil on the top of the ground is clay-based, which causes the city to be prone to flooding (USDA, 2008). As the rapid urbanization continues (Fig.5), the impermeable roads, mainly asphalt and concrete, diminish the run-off ability of the surface. The light red area in Figure 5 is the urban area.

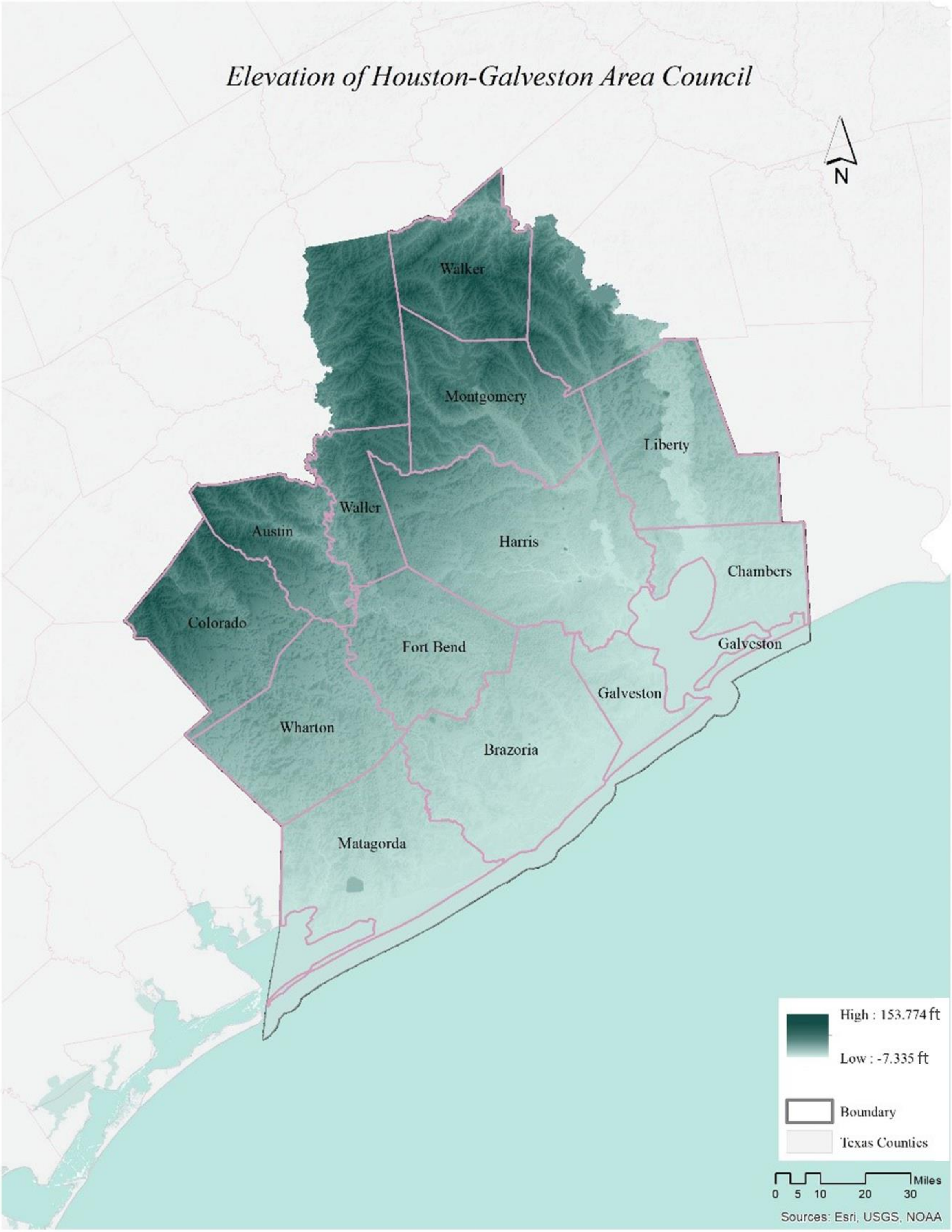


Fig. 4 Elevation of Houston

Data Source: H-GAC (2018)

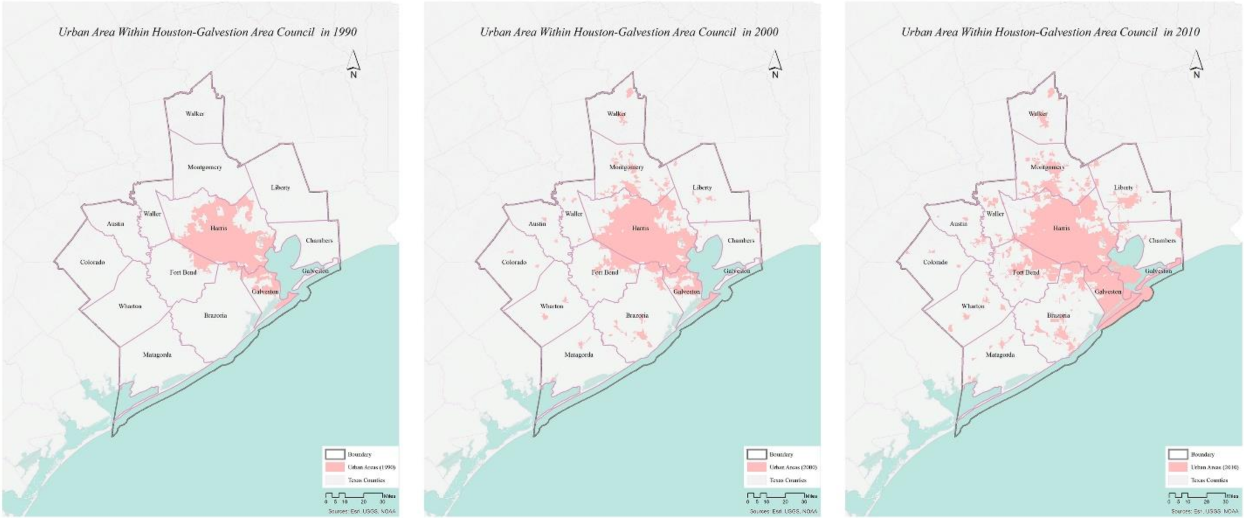

Urban Area Within Houston-Galvestion Area Council in 1990
Urban Area Within Houston-Galvestion Area Council in 2000
Urban Area Within Houston-Galvestion Area Council in 2010
N
Walker
Montgomery
Liberty
Austin
Waller
Harris
Chambers
Colorado
Fort Bend
Galveston
Wharton
Brazoria
Matagorda
Boundary
Urban Areas (1990)
Urban Areas (2000)
Urban Areas (2010)
Texas Counties
0 5 10 20 30 Miles
Sources: Esri, USGS, NOAA

Fig. 5 Change of Urbanized Area from 1990 to 2010

Data Source: H-GAC (2018)

Some of the roads that were designed to drain rain and storm-water are now actually landfills, obstructing the evacuation of people inside the affected areas and the rescue of people coming from outside.

Grossman and Maclean (2018, para.2) point out that Houston has especially weak zoning rules and regulations. Scott (2017, para.1) argues that deregulation, has led Houston to a chaotic, even ugly, city and its people to be vulnerable. Twitter data generated during and after the hurricane might remind people of the places that are most serious and are in need of help most, thus guiding to lower these places' vulnerability in the future planning.

1.5 Research Question

This paper examines if and how Twitter data will be beneficial in the disaster management during and after a natural disaster, using Hurricane Harvey as a study case. More specifically, three research questions will be focused in this paper:

1. Could Twitter data reflect the emergency and vulnerability to aid in disaster response during Hurricane Harvey?

2. Could Twitter data be used to perform damage assessment after Hurricane Harvey?

3. If the answers of previous two questions are negative, how can we mine the value of twitter data?

Compared to other researchers (Yury et al., 2016; Yuan & Liu, 2018) who focus purely on using Twitter data to improve disaster response and perform damage assessment, this paper also uses data to offer recommendations for the future resilience planning. The goal ultimately is to promote what Yamagata and Maruyama (2016, p. v) prefer to call “transformation” rather than simply recovery, which means a city will regenerate itself to become stronger after the strike based on proper resilience strategies.

2. Literature Review

The literature review for this thesis is organized into four subsections: urban resiliency in relation to natural disaster management, innovative use of social media data in disaster management, social vulnerability and resilience-oriented planning.

2.1 Urban Resilience in Relation to Disaster Management

Generally, disaster management can be measured as the efficiency, effectiveness and seamlessness of managing various resources and multi-source information when a disastrous incident happens (Modh, 2010). The concept of resilience has emerged in the literature related to disaster management since the 1980s (Wildavsky, 1988). ‘Resilience’ used to be one of important criteria to measure sustainability. Gradually, huge loss of communities caused by disasters such as Hurricane Katrina require it to be an essential one (Boin, Comfort & Demchak, 2010). Longstaff (2005) contributes to mining the essence of resilience in relation to disaster management. She argues that in times of dangerousness, all individuals are eager to obtain information on risk and damage assessment to help them reduce uncertainty, implement the feasible resistance strategies and thus improve their individual resilience. Dufty (2012) proposes the concept “community disaster resilience” (p.40) to link the urban resilience and disaster management. He sets up a framework between resilience and disaster in social aspects, via predicting possible prospects for emergency managers to use social media data to build such a “community disaster resilience”.

2.2 Innovative use of social media data in disaster management

As social media becomes popular, those who are witnessing or experiencing the disasters can timely provide public safety institutions with geo-located information through updating posts and reposts. Therefore, as the report “Innovative Uses of Social Media in Emergency Management” (DHS, 2013) says, public safety organizations can leverage the power of these social platforms to enhance emergency management performance. However, Alexander (2014) offers a review of the negative sides of social media in disasters. He points out that rumor propagation, dissemination of false or misleading information is a huge problem although this might be done unintentionally.

In recent years more scholars with computer science and information-technology background begin to do research in this field with the perspective of big data analytics. It has become an interdisciplinary combination of spatiotemporal real-time analysis, machine learning, and disaster management. Yury et al. (2016) focuses on whether Twitter data is helpgul in the rapid assessment of disaster damage. They analyze the Twitter activity at multiple scales before, during, and after Hurricane Sandy, and prove that there exists a strong relationship between Twitter activity related to hurricanes and proximity to the routine of Hurricane Sandy.

Nazer et al. (2017) proposes four stages in disasters: warning, impact, response, and recovery. Social media posts during the four periods are dense enough for machine learning methods to achieve reliable results. They argue that topics, trend and memes are the three important results of tracking disasters via monitoring changes in data statistics, clustering similar messages, and automatic translation. Martín et al. (2017) examine the exchange of hurricane-related information based on Twitter data through both spatial and temporal analysis and estimate the population that are successfully evacuated during Hurricane Matthew. Matthew-related Twitter Activity (MTA), the ratio between the number of Matthew-related tweets and the

Twitter population is adopted in the study. Huang and Xiao (2015) investigate the nature of tweet content generated during different disaster phases of Hurricane Sandy and present a new coding schema to categorize tweets into 47 themes for establishing geographic situational awareness, and a framework that can be applied to separate tweets into those categories. Yuan and Liu (2018) do semantic analysis to pick up hurricane-related tweets generated during Hurricane Matthew by creating the index dictionary. They also collect insurance claim data from the Florida Office of Insurance Regulation and use it as the reference of the damage assessment. Both correlation analysis and comparative analysis of the spatial distribution of Twitter data and insurance data at the county level are performed to verify the feasibility to adopt social media data. Other scholars are interested in finding out how to rescue the affected people based on nature language processing (NPL), supporting vector machine (SVM) and priority scheduling algorithm, such as Sakaki et al. (2010) and Yang Zhou et al. (2017).

2.3 Social Vulnerability and Resilience

U.S. Climate Resilience Toolkit (CRT, 2019) points out that human suffering and loss of properties can be mitigated by lowering the "social vulnerability index" (SVI), an approach which employs the American Community Survey data to help identify which communities are in need of improvement to shape higher resilience.

Centers for Disease Control and Prevention (CDC) declares that SVI can help officials and planners get better preparedness for and response to the emergency events like hurricanes, disease outbreaks, or exposure to dangerous chemicals (CDC & ATSDR, 2018). Generally, SVI ranks each county (or census tract) based on four themes (or sectors) (CDC, 2018), 15 social factors. Flanagan and Gregory, et al., develops the SVI from those 15 census variables and conclude that planners are able to target and aid more effectively community-based efforts to mitigate and prepare for disaster events by knowing the location of socially vulnerable communities. Cutter, Boruff and Shirley (2003) use the "hazards-of-place" (p. 244) model of vulnerability to find out if the dimensions of social vulnerability are of significance. They conclude that social vulnerability is a multidimensional concept and it is a feasible approach to enabling communities to recover from environmental hazards.

In this research, SVI is compared with the tweet awareness to find out whether the Twitter data can reflect the vulnerability of one place, considering that many elderly and uneducated people might not know how to use social media data to save themselves, which may cause a bias when rescuing the people within affected areas. The concept of "tweet awareness" is defined in the Methodology section.

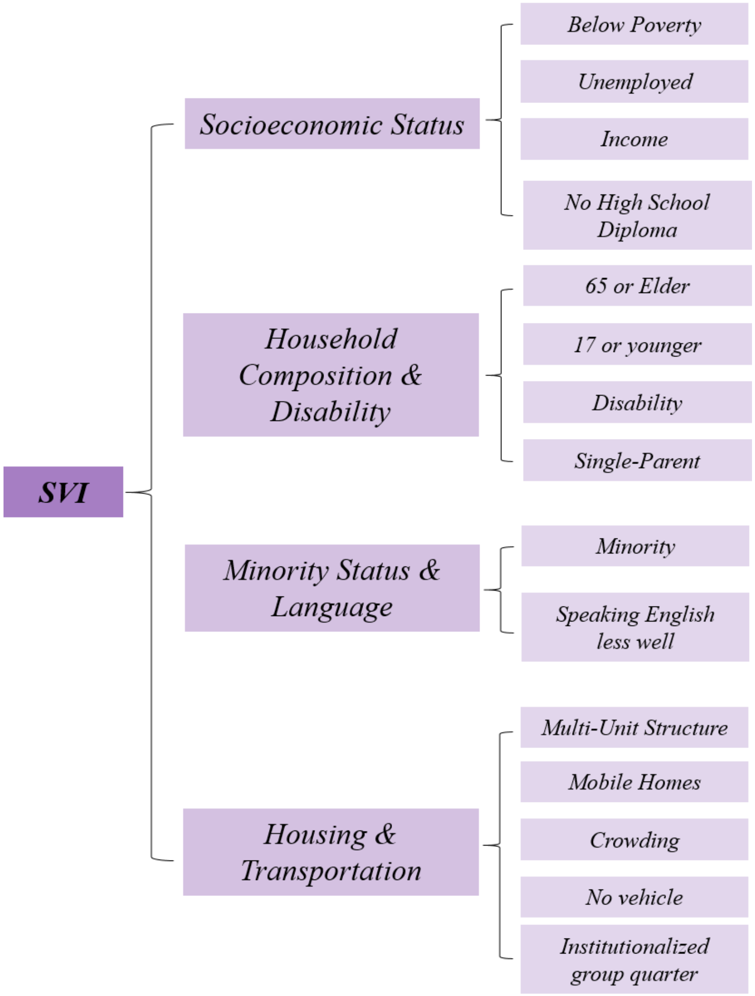


Fig. 6 Composition of Social Vulnerability Index

Data Source: CDC & ATSDR (2018)

2.4 Resilience-oriented Planning

After Hurricane Sandy hit the New York City, the Department of City Planning (2013) examined strategies for designing buildings more resilient to the future flooding crisis, including changing restrictions on the building elevation, modifying floor area regulations and adjusting use requirements of the ground-floor. A set of specific parameters has been taken into consideration, such as first-floor elevation and distance from center street line. It should be recognized that limited resources forced us to choose them to change based on the urgent level.

Adopting social media data could be beneficial to the decision-making process since more feedbacks on the current situations will be collected without holding meetings frequently.

Laundry et al. (2016) hold the view that open data help build knowledge, capacity, and outcomes that strengthen urban resilience. They suggest the government and communities cooperate together to develop a flexible approach to improving resilience, such as mapping real-time flooding and launching toolsets to quicken community disaster response.

3. Methodology

3.1 Definition of critical conceptions

Considering Yury's (2016) "Twitter Activity" and Martin's (2017) "MTA", to answer the question of whether twitter data could aid in the disaster response and the damage assessment, the research performs spatiotemporal analysis to investigate the relationship among tweet awareness (TAw), tweet activity (TAc), tweet focus (TFo), proximity and damage claim per capita (DCPC). Proximity means the Euclidean distance from centroid of the spatial unit (by H-GAC, county or census tract) to the hurricane center. It is used to measure the emergency situation of the hurricane. An assumption is set up that nearer the distance is, more severe the situation is. TAw refers to the ratio between the number of hurricane-related tweets and the total population, while TAc is the ratio between the number of hurricane-related tweets and twitter population within a spatial unit (county or census tract). Here, 'twitter population' is defined as the number of active twitter users within a spatial unit (county or census tract) during study period. TFo focuses on the ratio between the number of hurricane-related tweets and the number of general tweets within a spatial unit. DCPC refers to the ratio of number of damage claims and the total population within a spatial unit. Twitter data can be retrieved from the server of Twitter Inc. via an API. Damage claims can be retrieved from Hurricane Harvey data call (HHDC) released by Texas Department of Insurance (TDI). The selection principle of time period will be referred in the following part.

$$TAw = \frac{The\ number\ of\ hurricane\text{-}related\ tweets}{Total\ Population} * 100 \qquad (1)$$

$$TAc = \frac{The\ number\ of\ hurricane\text{-}related\ tweets}{Twitter\ \ Population} \qquad (2)$$

$$TFo = \frac{The\ number\ of\ hurricane\text{-}related\ tweets}{The\ number\ of\ general\ tweets} \qquad (3)$$

$$DCPC = \frac{The\ number\ of\ damage\ claims\ on\ HHDC}{Total\ \ Population} \qquad (4)$$

In Yury's research (2016), "Twitter activity" is defined as "the number of daily messages (Sandy-related) divided by the number of local users active on the topic" (Yury, et.al., 2016, p. 3). They find that Twitter activity has a sharp decline as the distance between Hurricane Sandy and urban area increases. They also find that there is a significant relationship between Twitter activity and the per-capita economic damage caused by the hurricane. Thus, this conception is adopted and here the essence of TAc and "Twitter activity" are actually the same.

In Guan and Liu's research (2018), disaster-related ratio (DIRR) is defined as the ratio of the number of disaster-related tweets divided by the number of general tweets. It describes the percentage of tweets focusing on the hurricane-related topics and the correlation coefficient is 0.469 between DIRR and DR (Damage Rate). DR is the ratio between the number of claims and total population in each county (or other spatial unit). In this paper, TFo and DIRR are the same as well as DCPC and DR are. To be specific, TFo is used to replace DIRR, and DCPC is used to replace DR.

The limitation of Twitter makes it hard to collect the entire general tweets within whole H-GAC, which means it is impossible to obtain all the general tweets with 13 counties. Thus, county-level analysis cannot be finished by TAc and TFo. Some indicators such as DCPC are counted by census tract, so TAw is proposed by myself to make sure the county-level can be still analyzed.

3.2 Data preparation

3.2.1 Basic GIS Datasets

A direct problem is the level of the spatial unit should be chosen. According to other scholars' findings (Martin, Li & Cutter, 2017), county level is a feasible choice when conducting temporal analysis while spatial analysis requires a finer degree of resolution than the temporal one. Therefore, both county and census tract will be used as the spatial unit when conducting the spatial analysis. But considering that damage claim is investigated based on county level, thus, when answering the second research question, the lowest spatial unit level should be the county level. There are 1109 census tracts within H-GAC in total. The average area of each census tract is 12.56 square miles while the standard deviation is 39.24352, which indicate that the there is a huge difference among the area of these census tracts. From Figure 7, the average census tract area of Colorado and Wharton are much larger than that of Harris County, and Harris County has the most census tracts, which means the resolution will be more precise if Harris County is selected to do the census-tract level analysis.

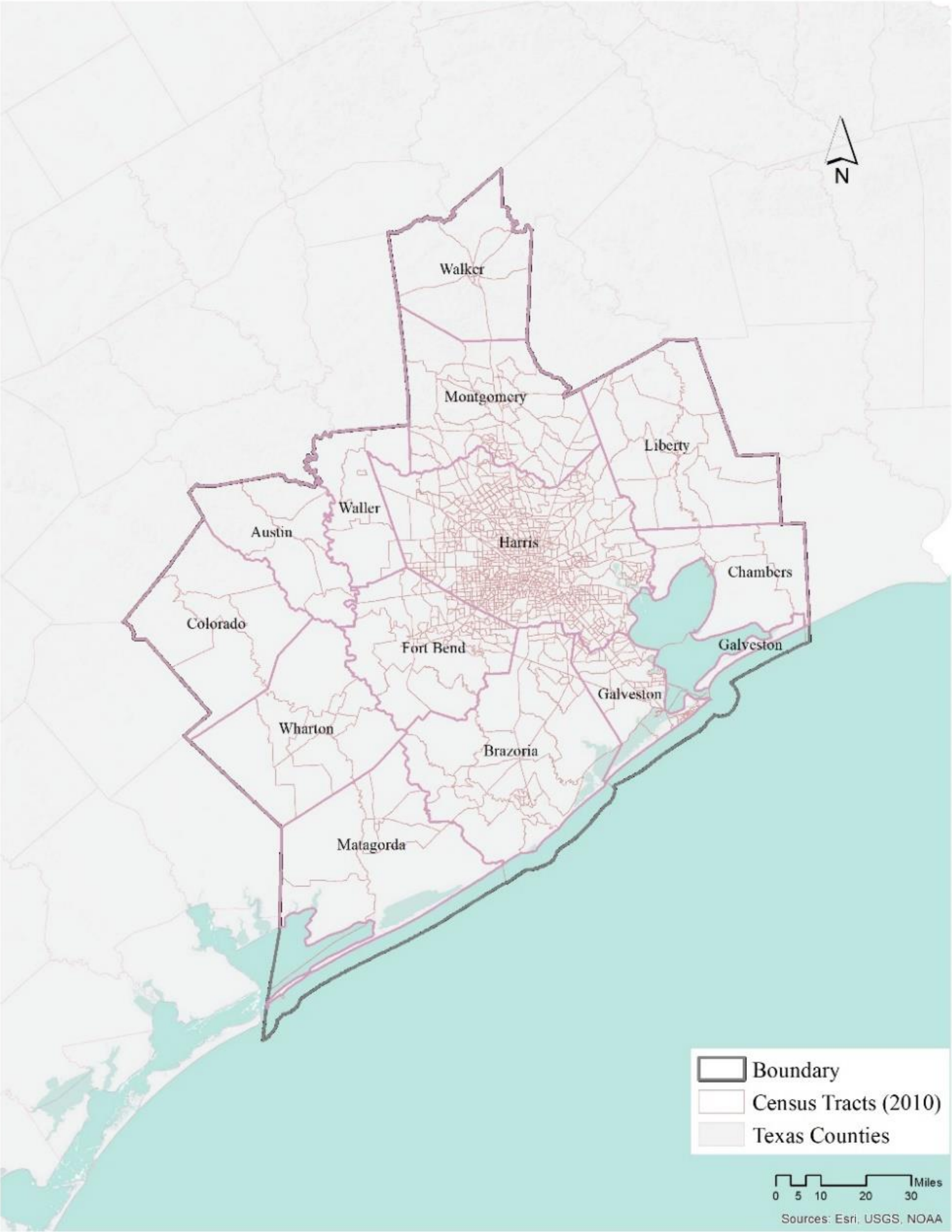
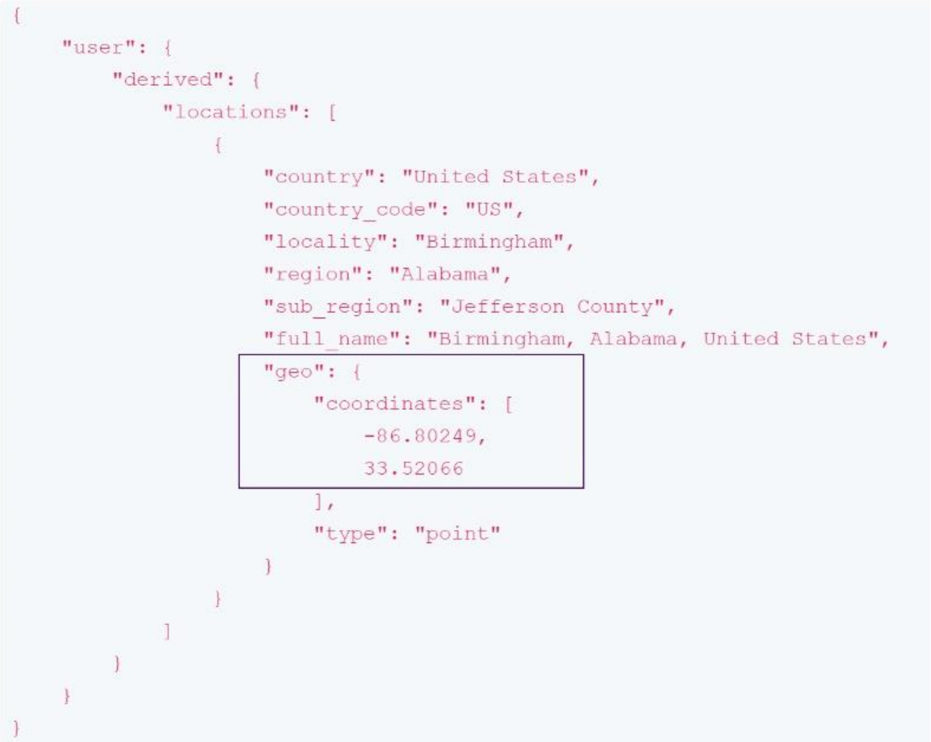


Fig. 7 Counties and Census Tracts within Houston

Data Source: H-GAC (2018)

3.2.2 Twitter Data

Then the next step is Twitter data collection for each county in the Houston. Data collection and cleansing is evitable. A premium Twitter API has been acquired to collect hurricane related tweets within Houston which have “profile geo”. According to the explanation on the platform of Twitter developer (Twitter developer, 2019), the ‘profile geo’ (the contents within the purple box in Figure 8) provides latitude/longitude coordinates relevant to the user derived location and it is one of the privileges of using a premium API.



```
{
    "user": {
        "derived": {
            "locations": [
                {
                    "country": "United States",
                    "country_code": "US",
                    "locality": "Birmingham",
                    "region": "Alabama",
                    "sub_region": "Jefferson County",
                    "full_name": "Birmingham, Alabama, United States",
                    "geo": {
                        "coordinates": [
                            -86.80249,
                            33.52066
                        ],
                        "type": "point"
                    }
                }
            ]
        }
    }
}
```

Fig. 8 Illustration of profile geo

Data Source: Twitter developer (2019)

Python 3 (Jupyter notebook) is used in this study. The main packages that are imported to python include ‘searchtweets’ (Jeffakolb & Binaryaaron, 2013) and ‘pandas’ (Pydata, 2018).

The study period spans from August 25, 2017 to September 6, 2017, from right before the beginning of Harvey’s land on to a week after the completeness of official rescue task. Due to the limitation of Twitter, including the rate limit per minute, total requests per month and tweets per call, while obtaining the general tweets, the total requests are easy to overflow when trying the entire Houston and at most one county can be considered in the study. That is to say, according to the definition of tweet activity and tweet focus, they cannot be calculated in a county-level but census tract level. Harris County is selected in this research because it has the top 3 largest urbanized rate (The ratio between urban area and total area within each county, Fig.9) as well as the largest population (Fig.10). Also, as mentioned in 4.1, it has more refined spatial resolution.

Figure 11 shows the tweets that retrieved from the server, including the five variables pertinent to this study. The field "name" is used to calculate the number of twitter population and "date" helps to do the temporal analysis. Based on "lon" and "lat", feature points of each tweet can be generated in ArcMap. The user name of each tweet in Figure 11 is covered for the privacy policy.

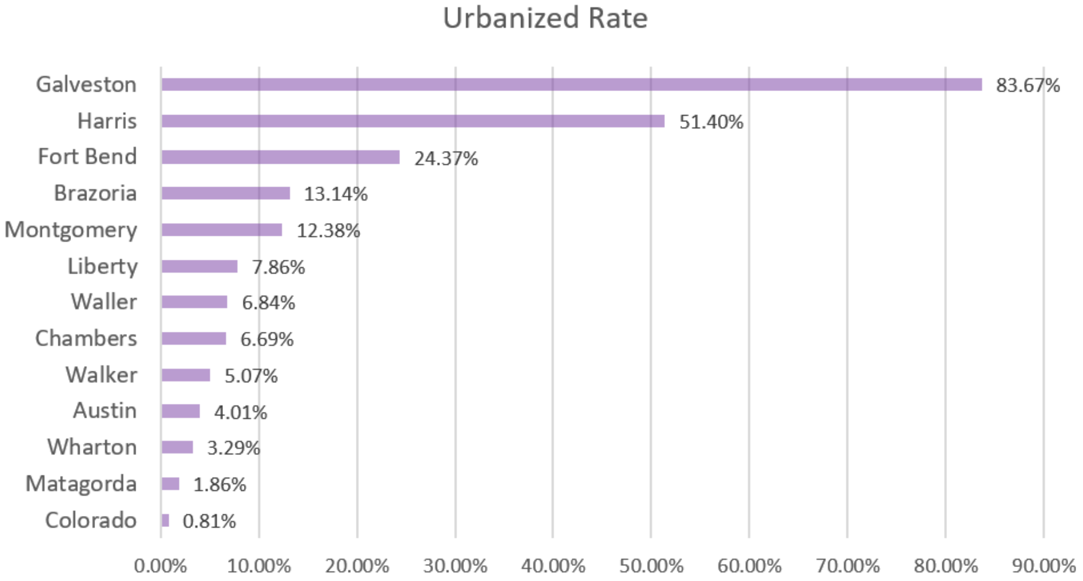


Fig. 9 Urbanized Rate of each county

Data Source: H-GAC (2018)

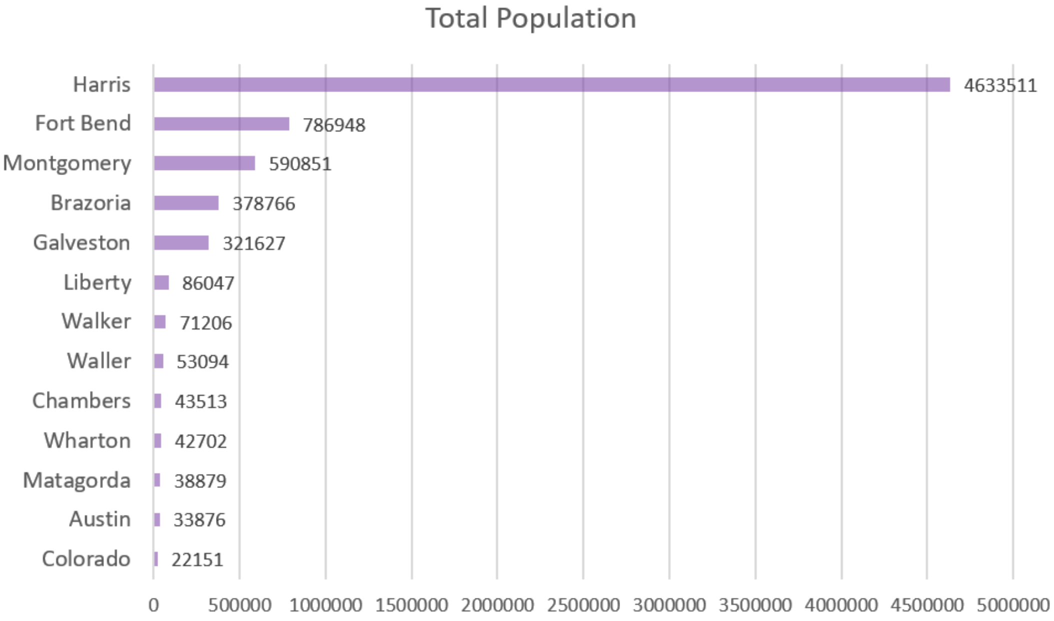
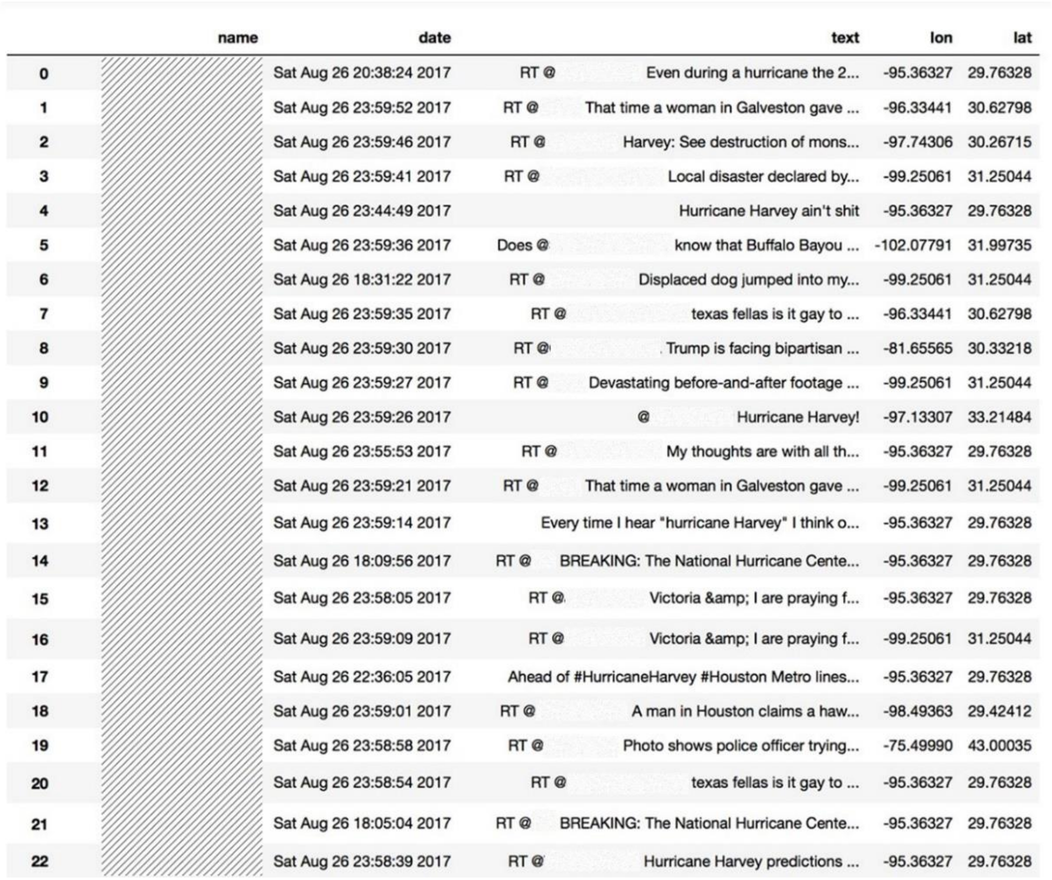


Fig. 10 Total population in each county

Data Source: H-GAC (2018)



	name	date	text	lon	lat
0	[redacted]	Sat Aug 26 20:38:24 2017	RT @[redacted] Even during a hurricane the 2...	-95.36327	29.76328
1	[redacted]	Sat Aug 26 23:59:52 2017	RT @[redacted] That time a woman in Galveston gave ...	-96.33441	30.62798
2	[redacted]	Sat Aug 26 23:59:46 2017	RT @[redacted] Harvey: See destruction of mons...	-97.74306	30.26715
3	[redacted]	Sat Aug 26 23:59:41 2017	RT @[redacted] Local disaster declared by...	-99.25061	31.25044
4	[redacted]	Sat Aug 26 23:44:49 2017	Hurricane Harvey ain't shit	-95.36327	29.76328
5	[redacted]	Sat Aug 26 23:59:36 2017	Does @[redacted] know that Buffalo Bayou ...	-102.07791	31.99735
6	[redacted]	Sat Aug 26 18:31:22 2017	RT @[redacted] Displaced dog jumped into my...	-99.25061	31.25044
7	[redacted]	Sat Aug 26 23:59:35 2017	RT @[redacted] texas fellas is it gay to ...	-96.33441	30.62798
8	[redacted]	Sat Aug 26 23:59:30 2017	RT @[redacted] . Trump is facing bipartisan ...	-81.65565	30.33218
9	[redacted]	Sat Aug 26 23:59:27 2017	RT @[redacted] Devastating before-and-after footage ...	-99.25061	31.25044
10	[redacted]	Sat Aug 26 23:59:26 2017	@[redacted] Hurricane Harvey!	-97.13307	33.21484
11	[redacted]	Sat Aug 26 23:55:53 2017	RT @[redacted] My thoughts are with all th...	-95.36327	29.76328
12	[redacted]	Sat Aug 26 23:59:21 2017	RT @[redacted] That time a woman in Galveston gave ...	-99.25061	31.25044
13	[redacted]	Sat Aug 26 23:59:14 2017	Every time I hear "hurricane Harvey" I think o...	-95.36327	29.76328
14	[redacted]	Sat Aug 26 18:09:56 2017	RT @[redacted] BREAKING: The National Hurricane Cente...	-95.36327	29.76328
15	[redacted]	Sat Aug 26 23:58:05 2017	RT @[redacted] Victoria &amp; I are praying f...	-95.36327	29.76328
16	[redacted]	Sat Aug 26 23:59:09 2017	RT @[redacted] Victoria &amp; I are praying f...	-99.25061	31.25044
17	[redacted]	Sat Aug 26 22:36:05 2017	Ahead of #HurricaneHarvey #Houston Metro lines...	-95.36327	29.76328
18	[redacted]	Sat Aug 26 23:59:01 2017	RT @[redacted] A man in Houston claims a haw...	-98.49363	29.42412
19	[redacted]	Sat Aug 26 23:58:58 2017	RT @[redacted] Photo shows police officer trying...	-75.49990	43.00035
20	[redacted]	Sat Aug 26 23:58:54 2017	RT @[redacted] texas fellas is it gay to ...	-95.36327	29.76328
21	[redacted]	Sat Aug 26 18:05:04 2017	RT @[redacted] BREAKING: The National Hurricane Cente...	-95.36327	29.76328
22	[redacted]	Sat Aug 26 23:58:39 2017	RT @[redacted] Hurricane Harvey predictions ...	-95.36327	29.76328

Fig. 11 The sample tweets

3.2.3 Census Data

To calculate the SVI, it is necessary to collect American Community Survey data from American Factfinder according to Figure 6. After selecting all the census tracts within the Harris County, topics Age Group, Disability, Employment, Poverty, Language and Housing. are selected to support the calculation of SVI.

3.3 Spatiotemporal Analysis

From the methodology framework (Fig. 12), temporal hurricane-related dataset contains the 13 counties within Houston during and after the Hurricane Harvey while the spatial hurricane-related dataset also contains all the census tracts.

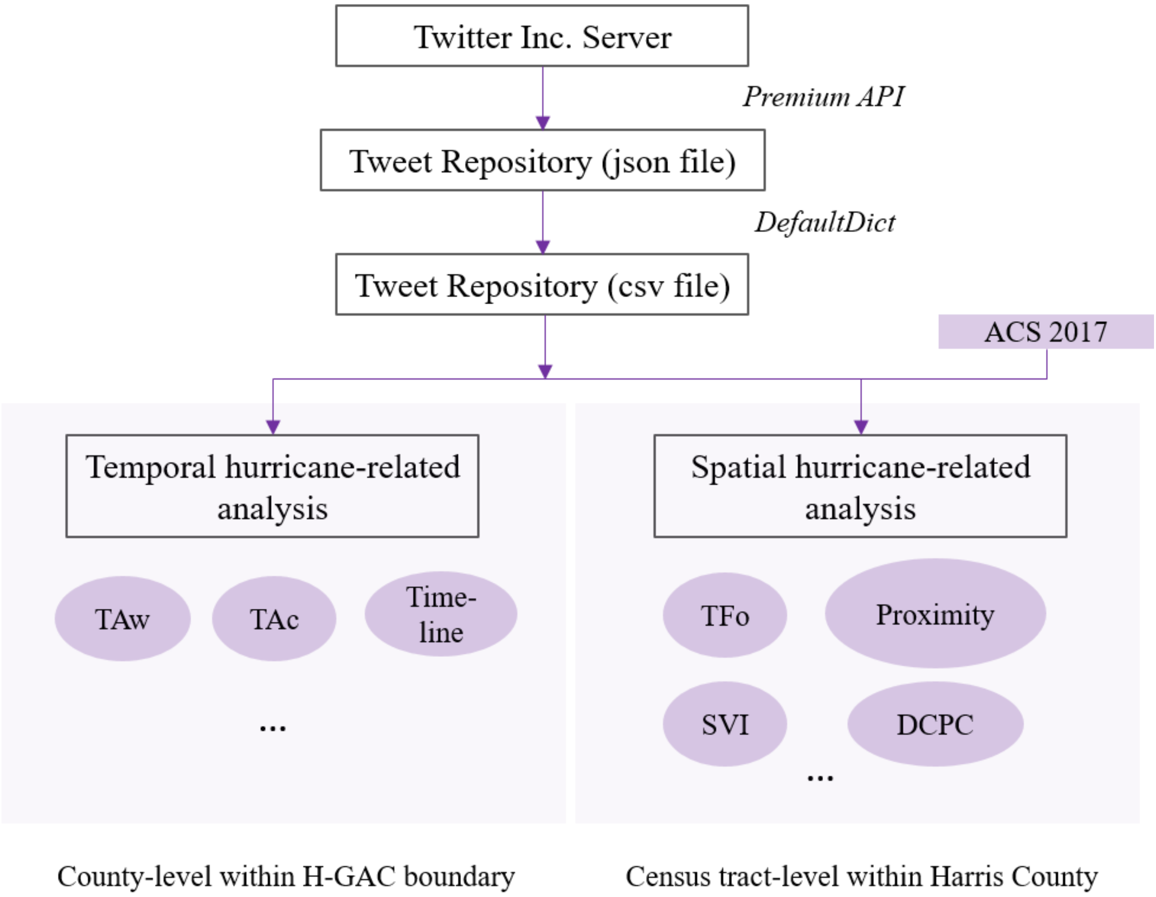

Twitter Inc. Server
Premium API
Tweet Repository (json file)
DefaultDict
Tweet Repository (csv file)
ACS 2017
Temporal hurricane-related analysis
TAw
TAc
Time-line
...
Spatial hurricane-related analysis
TFo
Proximity
SVI
DCPC
...
County-level within H-GAC boundary
Census tract-level within Harris County

Fig. 12 Methodology Framework

3.3.1 Temporal Analysis

Through inspecting the temporal trend of the number of hurricane-related by hour and tweet awareness by day, when is the peak hour for twitter users is shown to find out when people are in need of help the most, and it is also helpful to see if tweet awareness will gradually fall off or continue growing after the hurricane leaves.

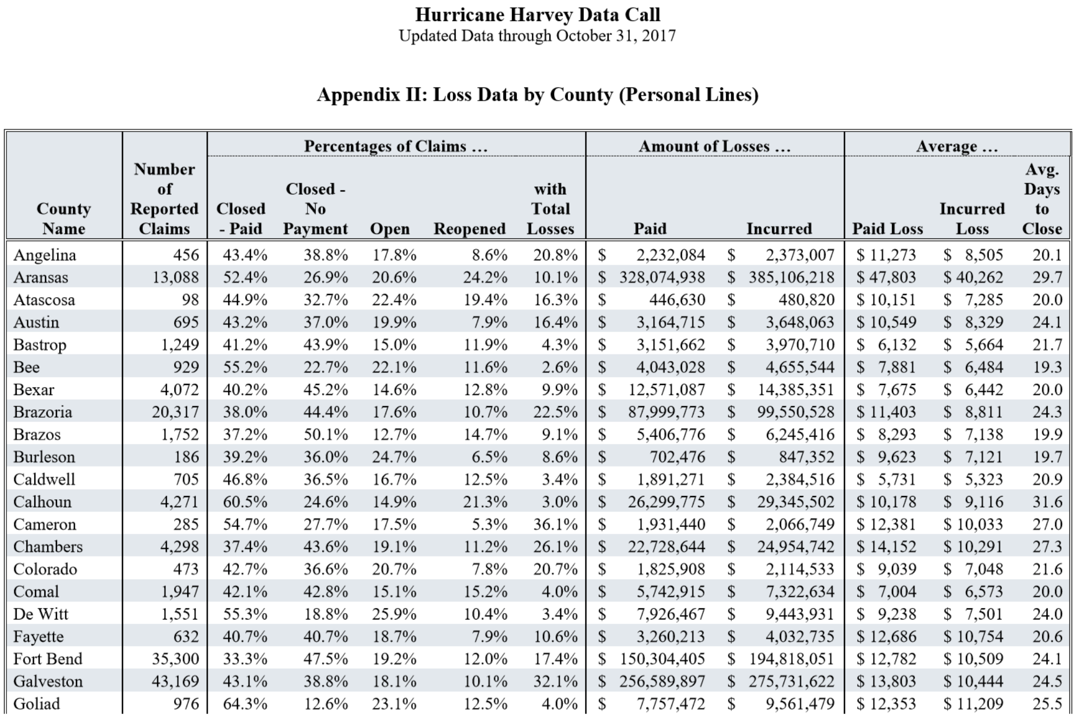
3.3.2 Spatial Analysis

Firstly, in order to verify whether twitter data could reflect the emergency situation, scatter plots graphs are output to see the general trend of the point and check if tweet awareness increases and declines with the proximity. In this part, according to Figure 8, there are too many census tracts, to be specific, 1109, in Harris County, which means distance between the centroid of some census tracts and the centroid of the hurricane are very close, which might cause skewness. Thus, county-level data are used. Correlation and covariance analysis then are conducted to figure out to what extent the proximity can be of significance to the tweet awareness. The correlation output is between -1 and 1. The negative result indicates an inverse relationship while the positive means closer relationship. Covariance measures the strength of the correlation between two or more random variates. Another important index, p-value, is also adopted to analyze the relationship. If the significance (p-value) is below than 0.05, then it can be concluded that twitter data could reflect the emergency situation spatially and thus will be beneficial to rescue response.

Then, to verify whether social media data can reflect the vulnerability, on the census tract level, the relationship between tweet activity, tweet focus and SVI will be analyzed. Similarly, correlation, covariance analysis will be conducted and p-value of the result will be checked. To reduce bias, census tracts with at least 10 tweets are selected to do the analysis. In Martin, Li & Cutter's study (2017), those with more than ten tweets in the same city represents as "active".

Next, how could we verify that these tweets can be useful in damage assessment?

Tweet awareness in Houston at the county level and Hurricane Harvey data call released by Texas Department of Insurance (TDI) are used. In this study, the number of reported claims based on personal lines will be used. Personal lines contain homeowners' insurance, residential dwelling insurance, mobile owners' insurance and personal automobile insurance (TDI, 2017).



Hurricane Harvey Data Call
Updated Data through October 31, 2017

Appendix II: Loss Data by County (Personal Lines)

		Percentages of Claims …					Amount of Losses …		Average …		
County Name	Number of Reported Claims	Closed - Paid	Closed - No Payment	Open	Reopened	with Total Losses	Paid	Incurred	Paid Loss	Incurred Loss	Avg. Days to Close
Angelina	456	43.4%	38.8%	17.8%	8.6%	20.8%	$ 2,232,084	$ 2,373,007	$ 11,273	$ 8,505	20.1
Aransas	13,088	52.4%	26.9%	20.6%	24.2%	10.1%	$ 328,074,938	$ 385,106,218	$ 47,803	$ 40,262	29.7
Atascosa	98	44.9%	32.7%	22.4%	19.4%	16.3%	$ 446,630	$ 480,820	$ 10,151	$ 7,285	20.0
Austin	695	43.2%	37.0%	19.9%	7.9%	16.4%	$ 3,164,715	$ 3,648,063	$ 10,549	$ 8,329	24.1
Bastrop	1,249	41.2%	43.9%	15.0%	11.9%	4.3%	$ 3,151,662	$ 3,970,710	$ 6,132	$ 5,664	21.7
Bee	929	55.2%	22.7%	22.1%	11.6%	2.6%	$ 4,043,028	$ 4,655,544	$ 7,881	$ 6,484	19.3
Bexar	4,072	40.2%	45.2%	14.6%	12.8%	9.9%	$ 12,571,087	$ 14,385,351	$ 7,675	$ 6,442	20.0
Brazoria	20,317	38.0%	44.4%	17.6%	10.7%	22.5%	$ 87,999,773	$ 99,550,528	$ 11,403	$ 8,811	24.3
Brazos	1,752	37.2%	50.1%	12.7%	14.7%	9.1%	$ 5,406,776	$ 6,245,416	$ 8,293	$ 7,138	19.9
Burleson	186	39.2%	36.0%	24.7%	6.5%	8.6%	$ 702,476	$ 847,352	$ 9,623	$ 7,121	19.7
Caldwell	705	46.8%	36.5%	16.7%	12.5%	3.4%	$ 1,891,271	$ 2,384,516	$ 5,731	$ 5,323	20.9
Calhoun	4,271	60.5%	24.6%	14.9%	21.3%	3.0%	$ 26,299,775	$ 29,345,502	$ 10,178	$ 9,116	31.6
Cameron	285	54.7%	27.7%	17.5%	5.3%	36.1%	$ 1,931,440	$ 2,066,749	$ 12,381	$ 10,033	27.0
Chambers	4,298	37.4%	43.6%	19.1%	11.2%	26.1%	$ 22,728,644	$ 24,954,742	$ 14,152	$ 10,291	27.3
Colorado	473	42.7%	36.6%	20.7%	7.8%	20.7%	$ 1,825,908	$ 2,114,533	$ 9,039	$ 7,048	21.6
Comal	1,947	42.1%	42.8%	15.1%	15.2%	4.0%	$ 5,742,915	$ 7,322,634	$ 7,004	$ 6,573	20.0
De Witt	1,551	55.3%	18.8%	25.9%	10.4%	3.4%	$ 7,926,467	$ 9,443,931	$ 9,238	$ 7,501	24.0
Fayette	632	40.7%	40.7%	18.7%	7.9%	10.6%	$ 3,260,213	$ 4,032,735	$ 12,686	$ 10,754	20.6
Fort Bend	35,300	33.3%	47.5%	19.2%	12.0%	17.4%	$ 150,304,405	$ 194,818,051	$ 12,782	$ 10,509	24.1
Galveston	43,169	43.1%	38.8%	18.1%	10.1%	32.1%	$ 256,589,897	$ 275,731,622	$ 13,803	$ 10,444	24.5
Goliad	976	64.3%	12.6%	23.1%	12.5%	4.0%	$ 7,757,472	$ 9,561,479	$ 12,353	$ 11,209	25.5

Fig. 13 The Hurricane Harvey Data Call (TDI, 2017. p.50)
Image source: TDI (2017)

With the number of reported claim it is easy to obtain DCPC per county and it is shown in the Figure 14. Galveston County has the highest DCPC (0.134) while Walker has the smallest (0.012).

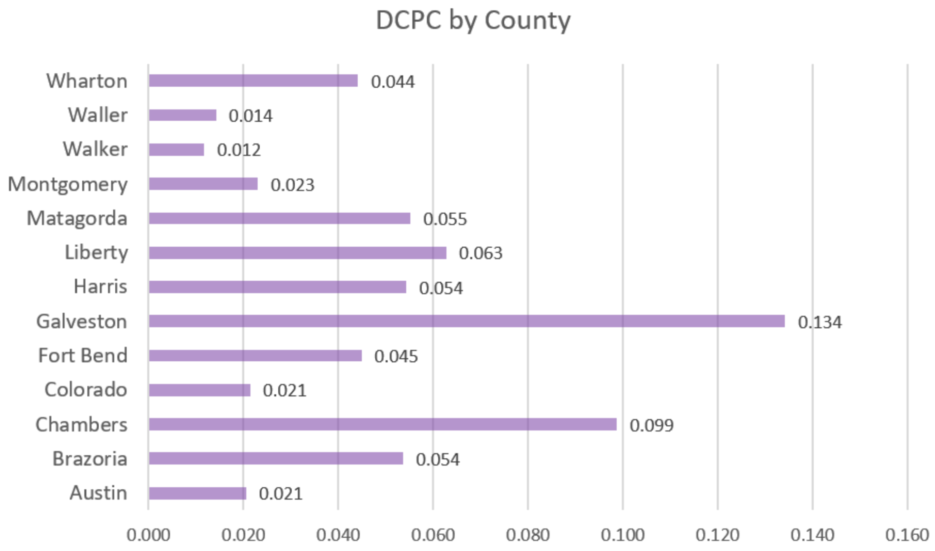


Fig. 14 DCPC by county

Data Source: H-GAC (2018)

Correlation between Twitter awareness and damage per capita in H-GAC, at the county level will be examined.

Maps of Spatial distribution of the tweet awareness, tweet activity, tweet focus, SVI, and DCPC are exported in ArcMap. The results will show which census tract has the highest Taw, TAC and TFo. If there is a significant relationship between TAw and DCPC, top 5 census tracts with the highest TAw will be selected and be considered as the "critical area", where should be set as the priority rescue area.

If there exists the critical areas, suitable sites will be selected to build more LID measurements, such as permeable roads, roof gardens and rain gardens and retrofit the local buildings using land suitability analysis.

3.3.3 Approaches to Calculate SVI

In this research, how to obtain more precise SVI is not the key part. CDC's method (CDC, 2016) is adopted to calculate the SVI of each census tract. First is to calculate the index of each sector (theme). To calculate each theme, it is essential to find data from the ACS data to calculate each indicator (15 in total). For example, institutionalized group quarters can be calculated as the following:

$$\frac{\sqrt{(MOE\ Persons\ in\ group\ quarters)\ ^2 - (Estimated\ proportion\ persons\ in\ group\ quarters)^2 * MOE\ Total\ population^2}}{Total\ population\ estimate * 100}$$

MOE Persons in group quarters and other parameters can all be found directly in the ACS data.

4. Findings and Discussions

4.1 Temporal Analysis

In total, 52010 hurricane-related tweets within H-GAC and 30420 general tweets within Harris County are obtained from the Twitter.

Only tweets with coordinates are collected. Many twitter users refuse to share their exact location so results might be biased because of the incompleteness of the tweet collection. On the other hand, collecting tweets with coordinates ensures that the tweets are generated within the study area and exclude those from beyond the study area.

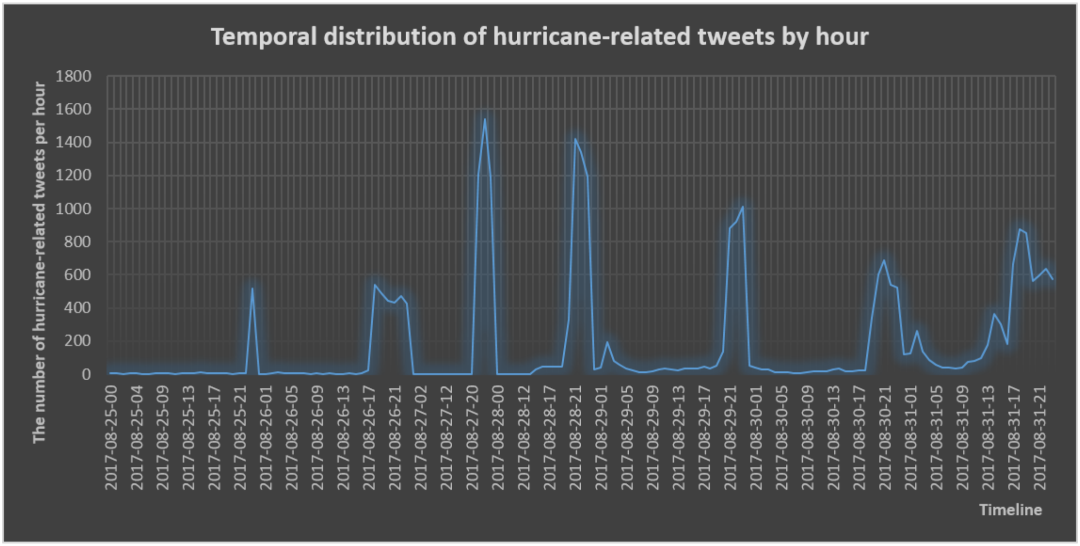


Fig. 15 The temporal distribution of hurricane-related tweets by hour

Figure 15 shows how the number of tweets changes by hour during Hurricane Harvey from August 25 to August 3 within H-GAC. From the figure, it can be obtained that hurricane-related tweets started to emerge in the evening of August 25th, as the consequence of the landfall of the hurricane. Based on Harvey's route data, it landed on the continent on the midnight of August 26 (00:00), which is one hour after the "outbreak" of the tweets. This implies local people's scare and also good awareness when the hurricane strikes. A possible scenario is that people were sharing their geolocations to warn their friends. Later, the number of tweets suddenly decreased. One hypothesis is that most twitter users could not do anything useful so they fell asleep and waited for the new day and the coming rescue. Most of them did not realize, or underestimate the severity of this hurricane might bring. Another speculation is that people realized that it was a hurricane and Twitter posting is not a priority when a disaster strikes.

From Figure 15, it can also be obtained that the temporal distribution of tweets is periodic and the tweets at night are much more than that in the daytime. Some of high points are more than 1000 tweets. Then all the time periods when the number of tweets are more than 1000 is shown in Figure 16 to inspect the most active periods of Twitter using.

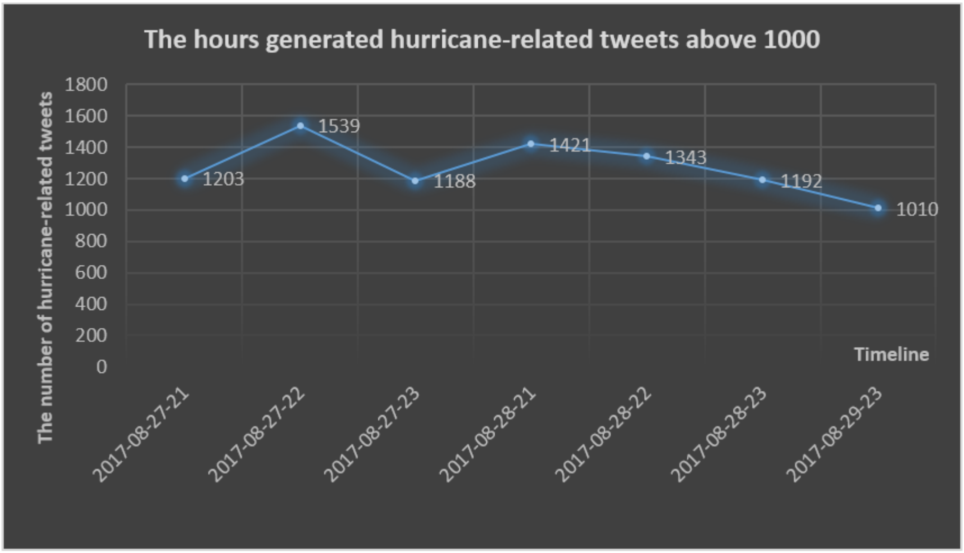


Fig. 16 The hours that hurricane-related tweets are above 1000

From Figure 16, the result shows that all of them are at night between 21-23pm. Without considering the bias (e.g. incompleteness of tweets), it can be inferred that people do not post hurricane-related tweets frequently during daytime for that they might be dealing with its impact, which would be during the days of the storm and while they are trying to reach help of cleanup during the day. Also, people may care more about local news on government rescue agencies in the daytime during the disaster.

The relationship between the change of the TAw and the proximity is also examined. Since temporal analysis is analyzed in the scale of H-GAC, according to the definitions of four indicators given before and the limitation of a premium API, tweet awareness (TAw) is the only one which allows to make analysis in a large scale, i.e. in the scale of whole H-GAC. Here, proximity is defined as the Euclidean distance from the centroid of H-GAC to the hurricane center. From Figure 17, temporal distribution of tweet awareness is similar to Gaussian distribution that the highest point is almost in the middle (on August 28) and the distribution is symmetrical around its highest point. Considering the route of the hurricane Harvey, whether the temporal distribution is correlated with the proximity is examined.

The position of hurricane center is recorded every six hours by NOAA. NAP (normalized average proximity) is adopted to measure the average distance from the hurricane center changing every six hours to the centroid of Houston in a single day. In order to see the relationship between TAw and proximity more clearly, proximity is normalized by 100,000 to make sure all of values of proximity are smaller than 1 so that the value of both tweet awareness and NAP are located between 0 and 1. Figure 17 tells that when the proximity is closer, the tweet awareness is increasing and vice versa. This distribution implies that the tweet awareness, during a large-scale disaster, can reflect the proximity of the hurricane, which means it could reflect the emergency situation of hurricane temporally.

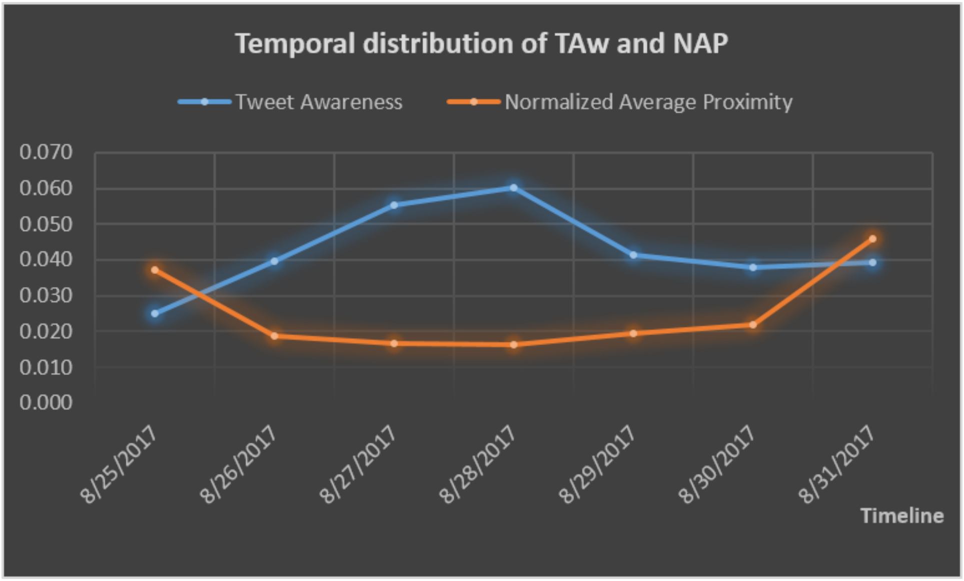


Fig. 17 Tweet Awareness of Houston and normalized average distance

4.2 Spatial Analysis

Firstly, based on county, the map is drawn to show the spatial distribution of TAw (Fig. 18). The Harris county and Walker county has the highest tweet awareness while Matagorda, Colorado and Liberty have the lowest. Considering that Harris County has the largest population among all the counties, it is reasonable to have the highest TAw while it is interesting to see that Walker County also has the highest TAw.

Secondly, the scatter plot of TAw and NAP and that of –log (TAw) and –log(NAP) are also drawn based on each county. Here, proximity is defined as the Euclidean distance from the centroid of each county to the hurricane center and thus NAP is different from the previous one because of the change of the spatial unit. According to the trend line in Figure 19, the slope on the left figure is almost 0 while the slope on the right indicates that there might be a positive relationship between them. However, the R-square is only 0.0703, which means the –log(TAw) can only 7.03% explain the result of –log(NAP). This is too weak to prove that there is a significant relationship between them.

Then correlation and covariance analysis are conducted. The correlation output should be between -1 and 1. Covariance measures the strength of the correlation between two or more random variates. The result show that the correlation is 0.168 while the covariance is 0.00132. The correlation shows that the relationship between TAw and DCPC are positive. The covariance indicates that when the tweet awareness changes, NAP almost remains still. Then from Figure . 20, the p-value (0.583) shows that there is no statistically significant relationship between TAw and NAP. Overall, the conclusion is drawn that there is no statistically significant relationship between TAw and NAP spatially based on county-level data.

Since at county level the result is not satisfying, then how about the situation at the census tract level? The relationship between TFo and SVI are used. Spatial distribution of SVI and its different sectors are shown in the Figure 21 and Figure 22.

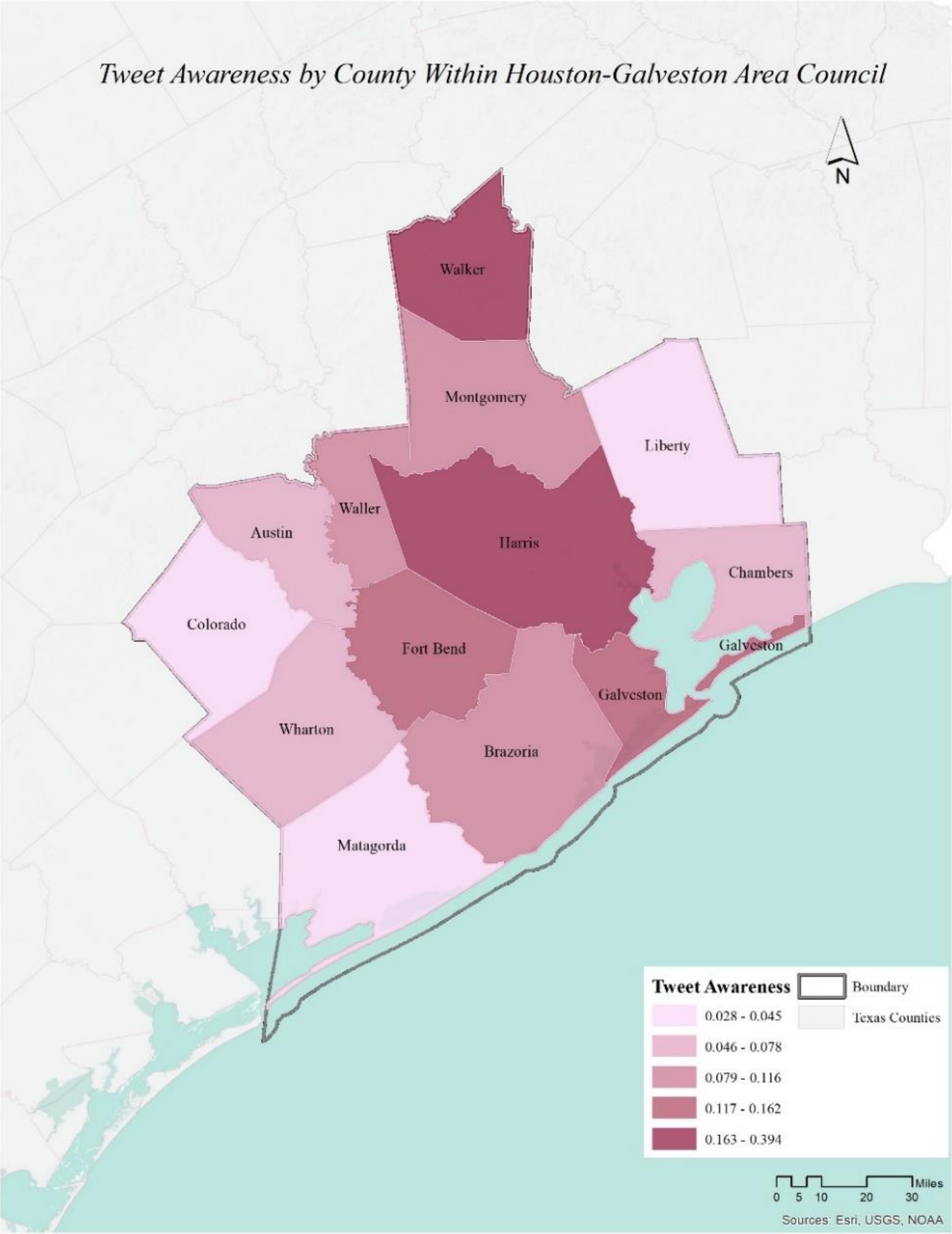


Fig. 18 Spatial Distribution of tweet awareness

Data Source: H-GAC (2018)

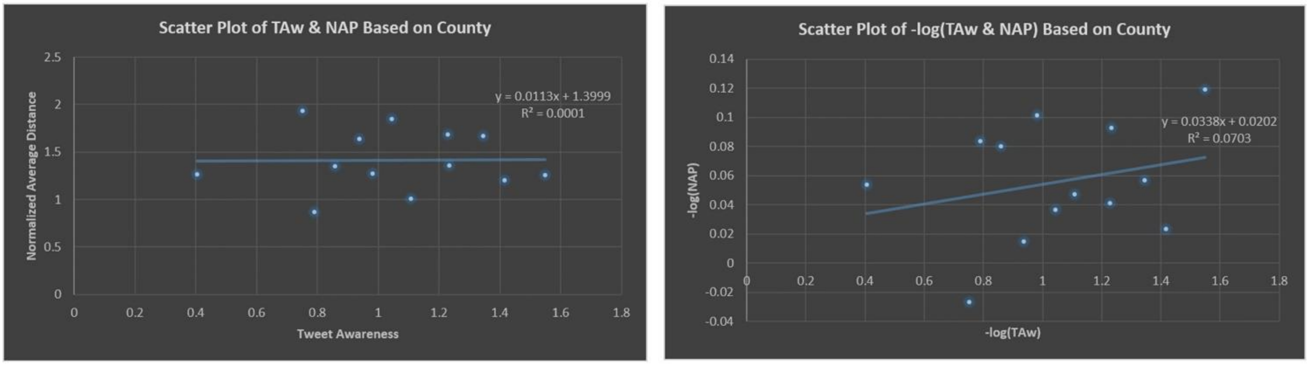
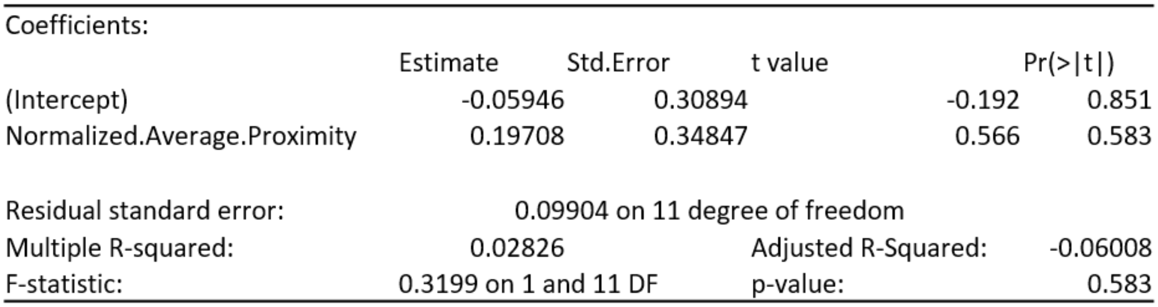


Fig. 19 Scatter plot of TAw and NAP and of –log(TAw and NAP)



Coefficients:				
	Estimate	Std.Error	t value	Pr(>|t|)
(Intercept)	-0.05946	0.30894	-0.192	0.851
Normalized.Average.Proximity	0.19708	0.34847	0.566	0.583
Residual standard error:	0.09904 on 11 degree of freedom			
Multiple R-squared:	0.02826		Adjusted R-Squared:	-0.06008
F-statistic:	0.3199 on 1 and 11 DF		p-value:	0.583

Fig. 20 p-value test of TAw-NAP

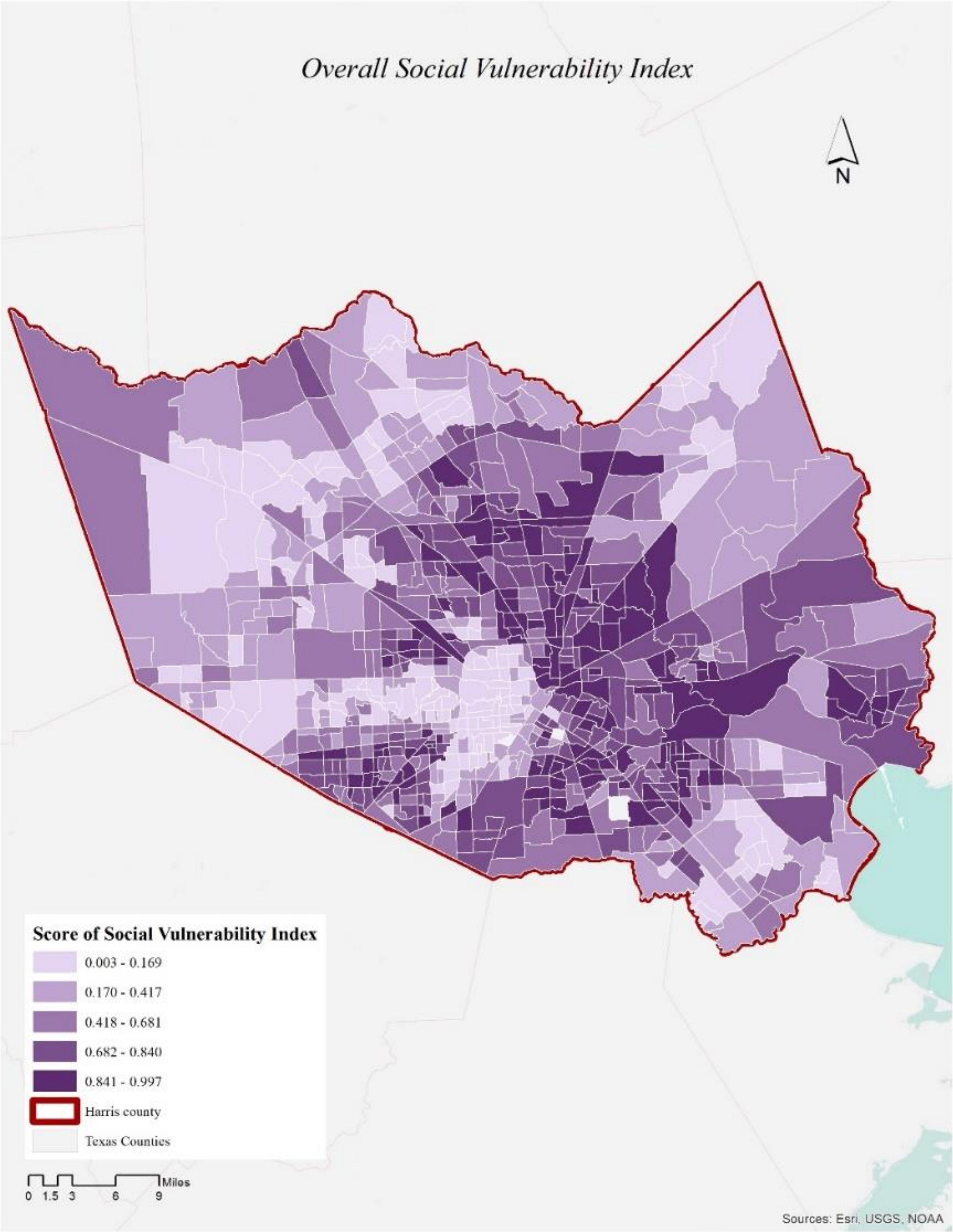


Fig. 21 Overall Social Vulnerability Index

Data Source: H-GAC (2018)

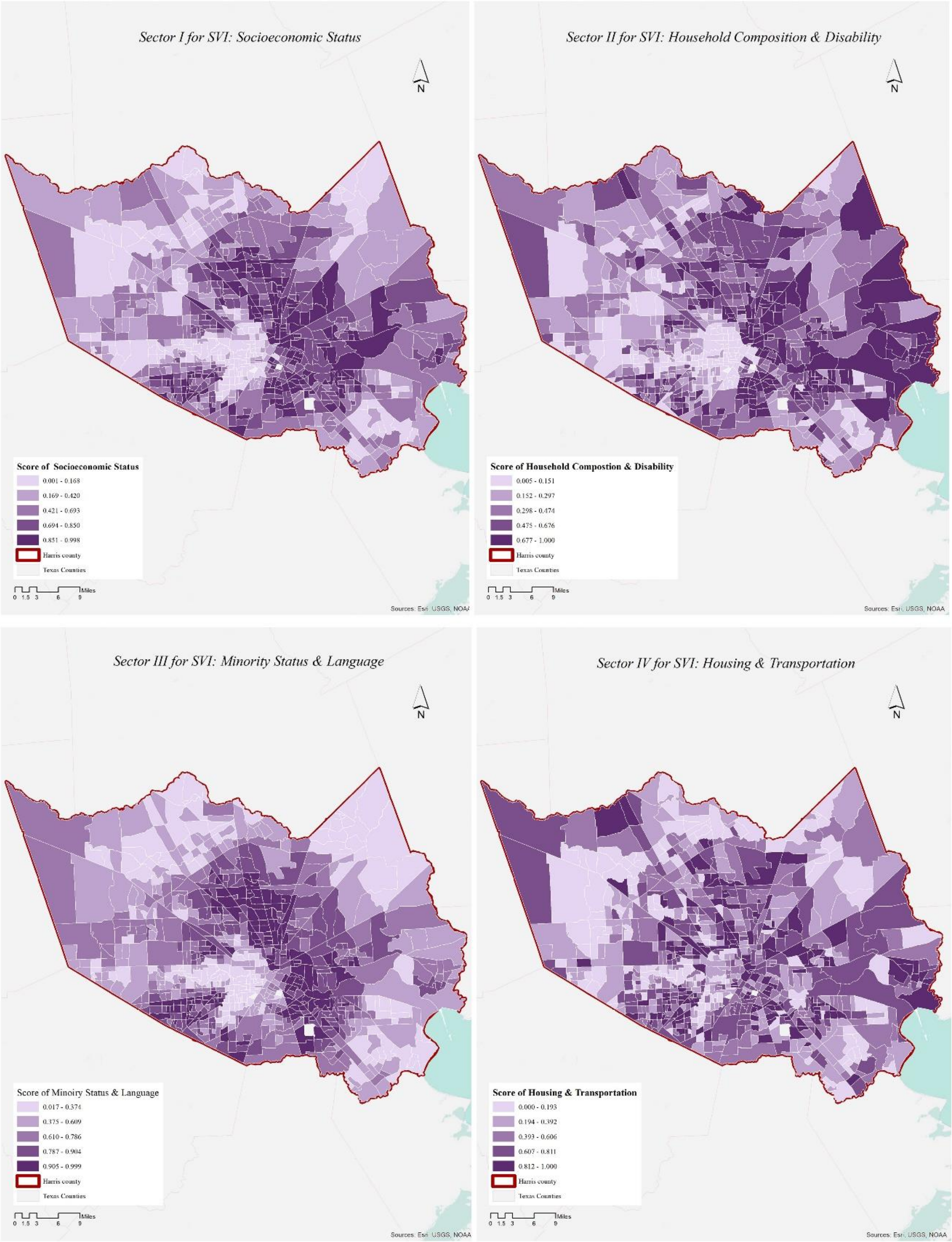


Fig. 22 Four distinct sectors to support the SVI

Data Source: H-GAC (2018)

Based on the previous constraints (at least 20 general tweets within the census tract), the scatter plot of TFo and SVI and –log (TFo & SVI) (Fig. 23) are drawn to make a comparison. Only 12 census tracts meet the constraints so it is not necessary to show the spatial distribution of TFo at the Harris county level. On every single day, each tweet has a distinct username and twitter population (30409 unique records) is almost the same as the number of general tweets (30420 unique records) in H-GAC. Thus, in this study, tweet activity can be regarded equal to twitter focus. In the following part, simply twitter focus (TFo) is used to do related spatial analysis.

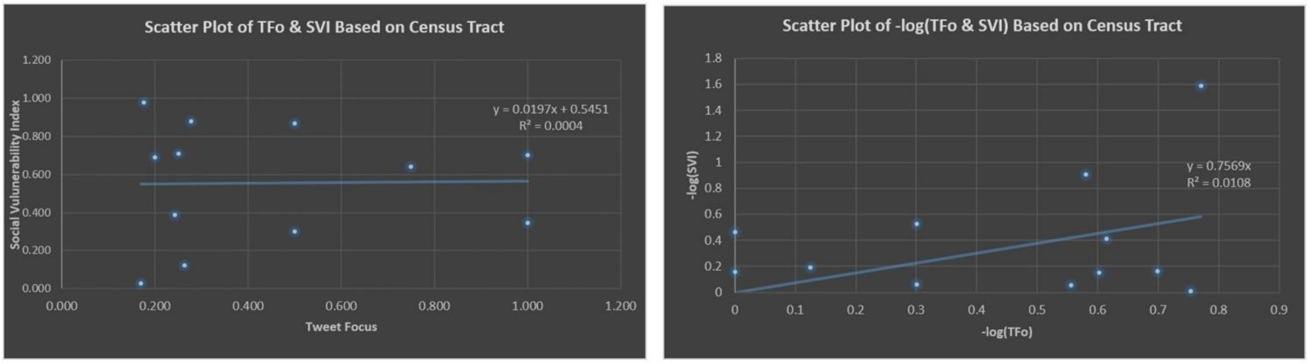
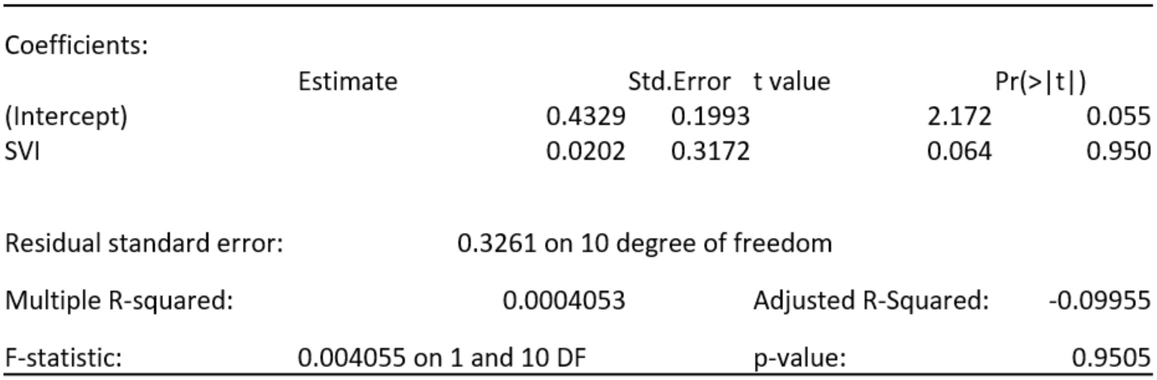


Fig. 23 Scatter plot of TFo and SV and of –log(TFo and SVI)

Correlation result between TFo and SVI is 0.02 and covariance is 0.002, which shows that the relationship between them hardly exists. The summary of p-value continues to prove this point.



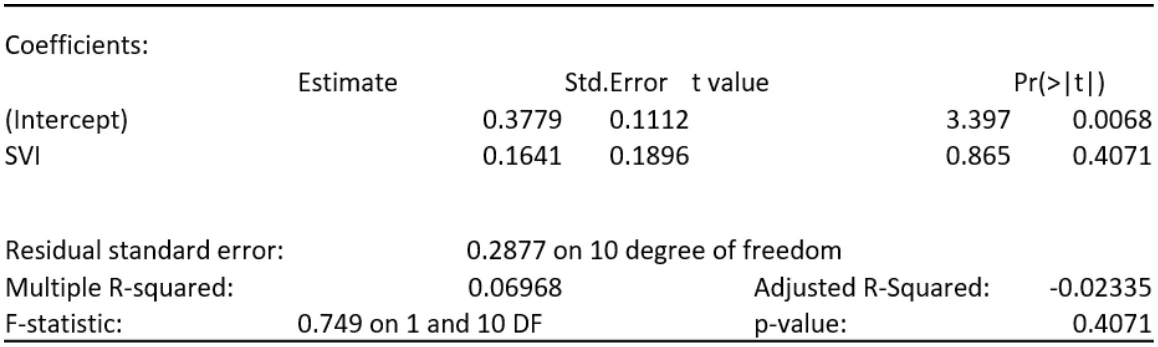
Coefficients:

	Estimate	Std.Error	t value	Pr(>|t|)
(Intercept)	0.4329	0.1993	2.172	0.055
SVI	0.0202	0.3172	0.064	0.950

Residual standard error:	0.3261 on 10 degree of freedom		
Multiple R-squared:	0.0004053	Adjusted R-Squared:	-0.09955
F-statistic:	0.004055 on 1 and 10 DF	p-value:	0.9505

Fig. 24 p-value test of TFo-SVI

How about the correlation, covariance and p-value of –log(TFo) and –log(SVI)? It looks like there is some relationship between them according to Figure 23. The correlation is 0.26 and the covariance is 0.034. The p-value is 0.41 which indicates the relationship between them is still not statistically significant (Fig. 25).



Coefficients:

	Estimate	Std.Error	t value	Pr(>|t|)
(Intercept)	0.3779	0.1112	3.397	0.0068
SVI	0.1641	0.1896	0.865	0.4071

Residual standard error:	0.2877 on 10 degree of freedom		
Multiple R-squared:	0.06968	Adjusted R-Squared:	-0.02335
F-statistic:	0.749 on 1 and 10 DF	p-value:	0.4071

Fig. 25 p-value test of –log (TFo-SVI)

Therefore, according to the comprehensive analysis, the twitter data could not reflect the social vulnerability.

Then, to answer the second question, relationship between tweet awareness and damage claim per capita is examined. Firstly, the spatial distribution of DCPC within Houston is drawn (Fig. 26). The largest locates in Galveston area, Matagorda and Liberty. The figure shows the spatial divergence between counties. According to the phenomenon, low-elevation coastal areas are more heavily impacted by storm surge flooding than inland areas, and that higher-value areas and higher-population areas will sustain greater damage loss.

The correlation result is 0.102 and the covariance is 0.0003. The p-value is 0.740, which indicates that there is no significant relationship between TAw and DCPC. Therefore, it cannot be used to do rapid assessment of damage loss.

Different from other research papers that all have an anticipated outcome, here nothing is related to each other and p-value is always much larger than 0.1.

In previous study, only census tract level data within Harris country is selected as the research area when conducting the TFo-related analysis while other researchers' spatial unit are either the state-level or county-level. From the perspective of the quantity, the number of tweets is much smaller than theirs, which might cause such a different finding. Also, there might be something different among people in different cities. For example, people living in the New York City use social media much more often than other cities.

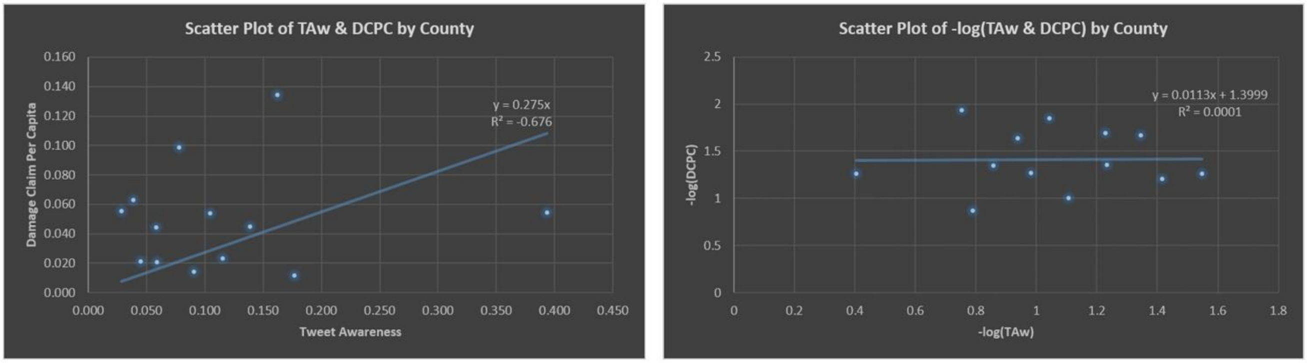


Fig. 26 Scatter plot of TAw and DCPC by county and of –log(TAw and DCPC) by county

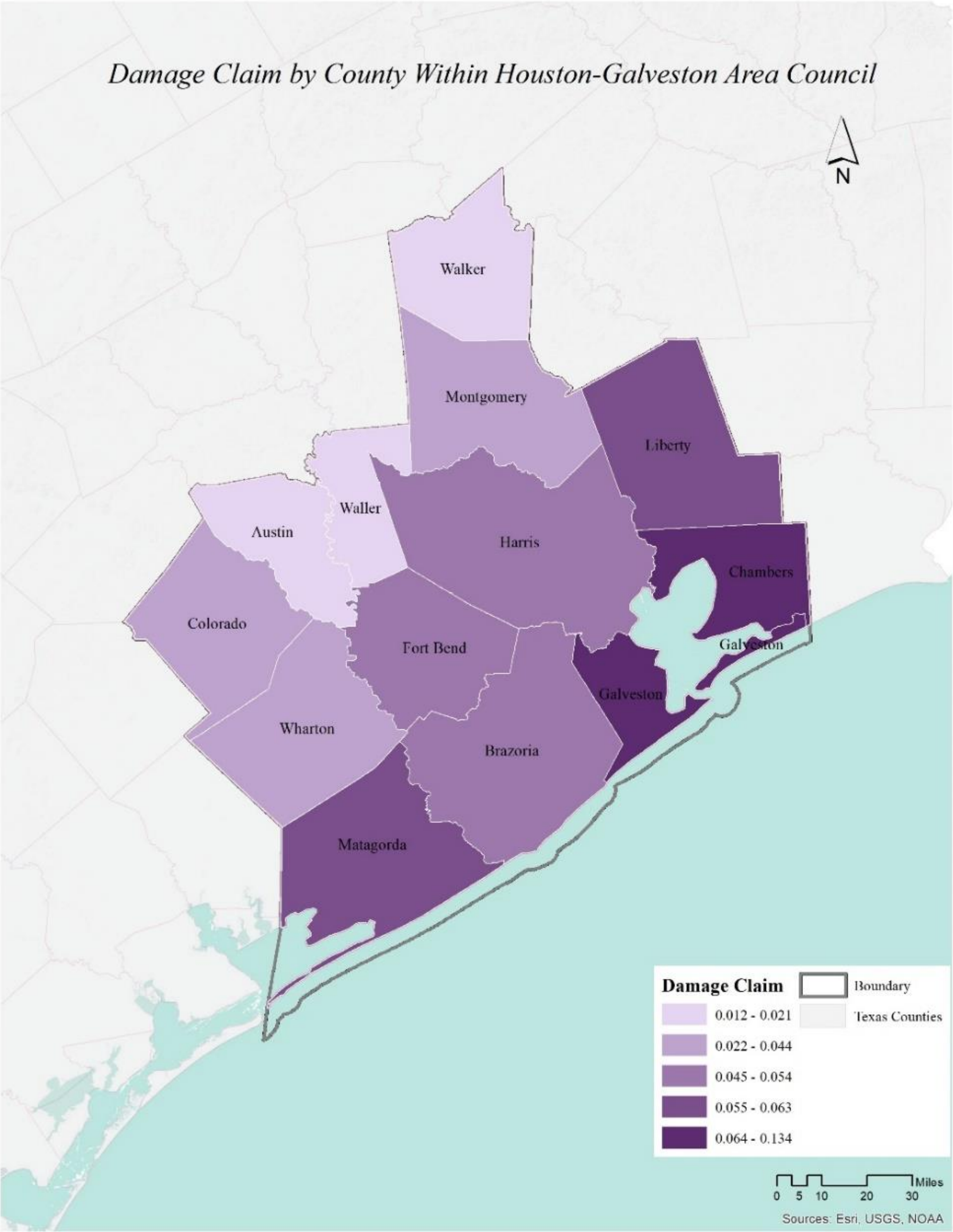


Fig. 27 Spatial distribution of DCPC by county

Data Source: H-GAC (2018)

4.3 Clustered Point Analysis

Most of the correlation test shows that the relationship is not of significance. One possible reason is that the points are too clustered. Here, those share the same coordinate (both latitude and longitude) are defined as the clustered point. 21244 feature points (hurricane-related tweets) are obtained within H-GAC and it accounts 40.85% of the total hurricane-related tweets. The top 1 clustered point, whose coordinate is (-95.36327, 29.76328), has 16930 points stacked at this location. From Figure 28, it is obtained that the University of Houston-Downtown locates here. The active Twitter users here post the most hurricane-related tweets, so it can be inferred that college student is mainstream of the active twitter users and they are better at using social media data to get support from outside than other age groups.

Considering the conclusion that the twitter data could not reflect the social vulnerability, simply relying on twitter data to rescue people can save active social media users but might not save “real vulnerable” population. For instance, those age above 65 years old or below 10 who do not know how to post a tweet.

Also, the location of the campus is adjacent to the rivers, which might be a big problem when a hurricane strikes. Thus, it can also be inferred that the site might be one of the most influenced places by Hurricane Harvey.

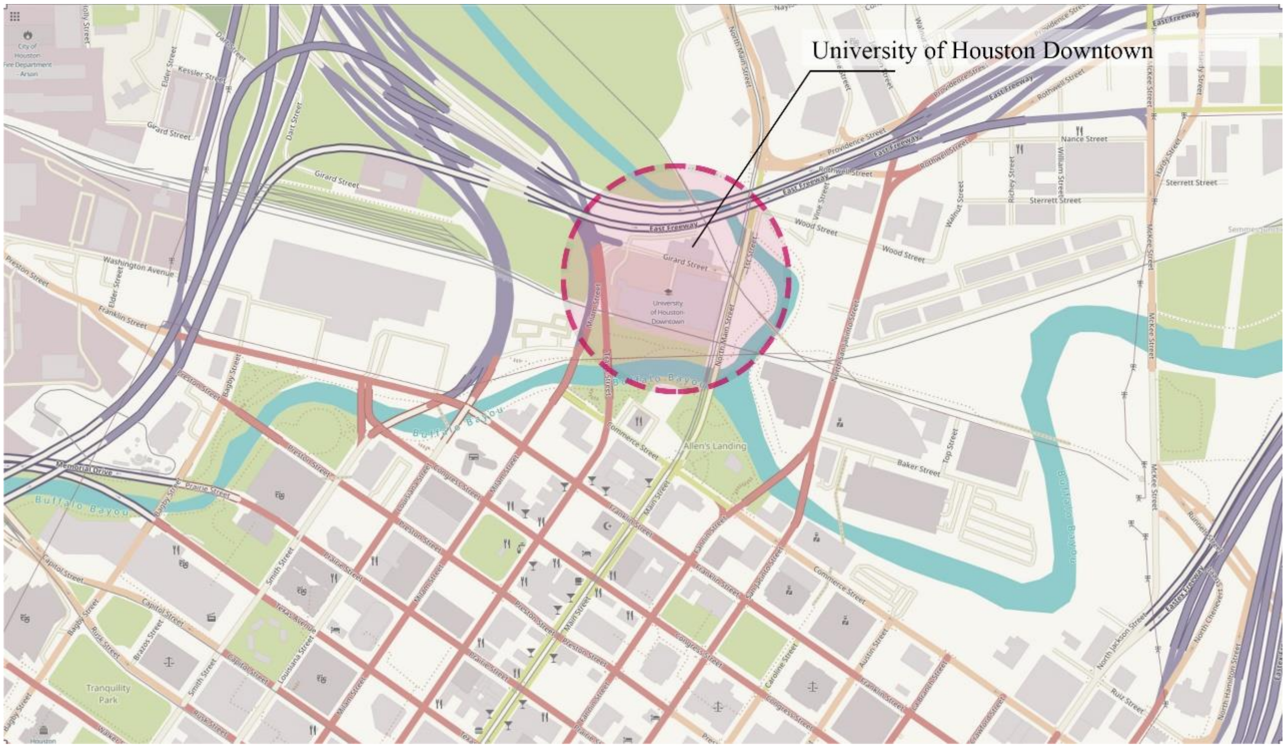
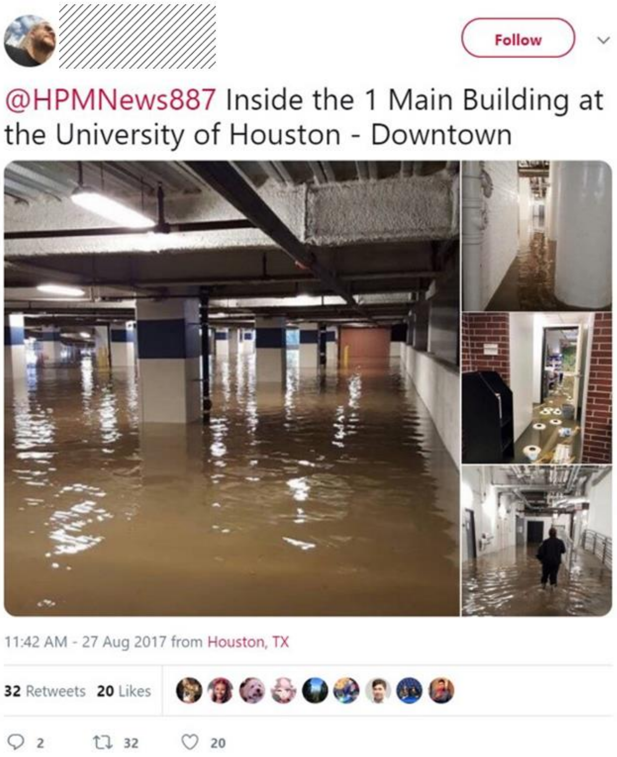


Fig. 28 Location of the University of Houston-Downtown

Data Source: OpenStreetMap

Then, based on the keyword “University of Houston”, “Downtown”, all the related tweets and extract the “urls” in the “text”. Here’s the sample tweets that are obtained (Fig. 29). These photos show the severity of the University of Houston-Downtown during the Hurricane Harvey and few people managed to evacuate.

What’s more, 96.27% of the hurricane-related tweets are located within the floodplain. There are nine counties where all the tweets within the county are located in the floodplain. They are Austin, Chambers, Colorado, Fort Bend, Harris, Liberty, Montgomery, Walker and Waller.



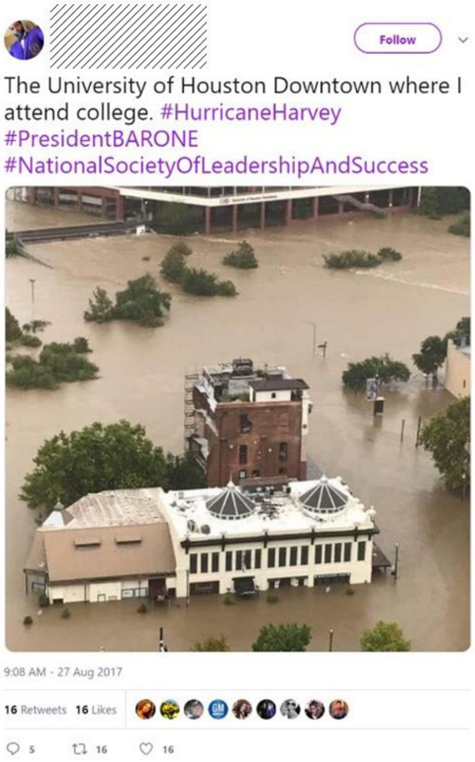
[redacted] Follow

@HPMNews887 Inside the 1 Main Building at the University of Houston - Downtown

11:42 AM - 27 Aug 2017 from Houston, TX

32 Retweets 20 Likes

2 32 20



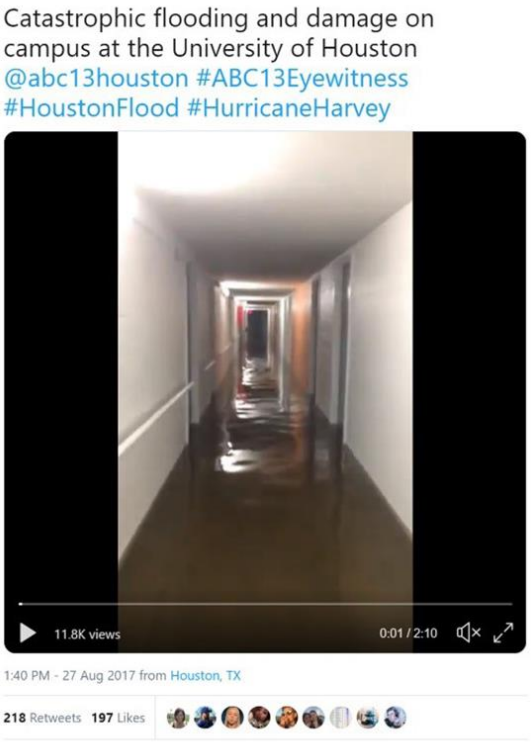
[redacted] Follow

The University of Houston Downtown where I attend college. #HurricaneHarvey #PresidentBARONE #NationalSocietyOfLeadershipAndSuccess

9:08 AM - 27 Aug 2017

16 Retweets 16 Likes

5 16 16

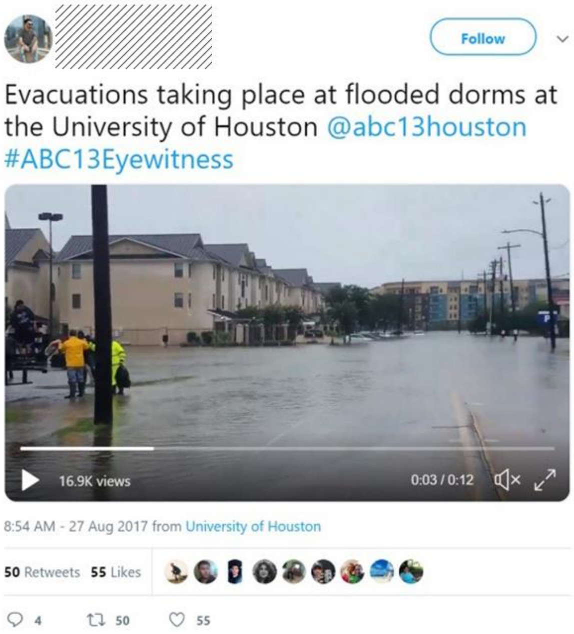


Catastrophic flooding and damage on campus at the University of Houston @abc13houston #ABC13Eyewitness #HoustonFlood #HurricaneHarvey

11.8K views 0:01 / 2:10

1:40 PM - 27 Aug 2017 from Houston, TX

218 Retweets 197 Likes



[redacted] Follow

Evacuations taking place at flooded dorms at the University of Houston @abc13houston #ABC13Eyewitness

16.9K views 0:03 / 0:12

8:54 AM - 27 Aug 2017 from University of Houston

50 Retweets 55 Likes

4 50 55

Fig. 29 Sample tweet related to University of Houston-Downtown

Data Source: www. twitter.com

5. Conclusions

5.1 Basic Conclusions

Generally, Twitter is a pervasive platform where active users create, repost and comment on tweets, which makes it easy to be a free tool for monitoring public activities. On the bright side, experiences in recent years have shown that it has the potential to support the disaster management. According to the spatiotemporal analysis, the main conclusions are drawn as the following.

The first is that the temporal distribution of tweets is periodic and the tweets at night are much more than that in the daytime and there exists the “outbreak” time of the tweets.

The first outbreak implies local people’s scare and also good awareness when the hurricane strikes, which means twitter data can be of a mechanism to alert larger population in less time.

Secondly, when the proximity is closer, the tweet awareness is increasing and vice versa. This distribution implies that the tweet awareness, during a large-scale disaster, can reflect the proximity of the hurricane, it can reflect the emergency situation of hurricane temporally.

Thirdly, there is no statistically significant relationship between TAw and NAP based on county-level data. Tweet awareness measure the ratio of hurricane-related tweets to the total population.

Fourthly, the relationship between TFo and SVI is not statistically significant and thus, the twitter data could not reflect the social vulnerability. It is not difficult to predict, as those who are old or young, those who are not good at speaking English or those who cannot afford a smartphone, they do not use Twitter or they do not even have a Twitter account. From this perspective, active Twitter users seems to be less vulnerable than those who do not use and here comes up with a contradiction, we want to help those with higher vulnerability while twitter users are those with lower vulnerability.

Next, there is no significant relationship between them spatially and it is not feasible to perform rapid assessment of damage loss.

What's more, the spatial distribution of clustered point of hurricane-related tweets is also an aspect that worth noticing. Using the highly-clustered hurricane-related tweets (those who share the same coordinates) to guide the rescue is liable. But if we can rank the rescue priority based on the number of the clustered points, it is another interesting topic that worth studying.

Last but not least, main active users of Twitter might be college school students and they are usually less vulnerable than those who do not use Twitter. Those who are not able to own a Twitter account might be of higher vulnerability and are more in need of help.

5.2 Limitation

5.2.1 Limitation of Research

In this study, when doing the spatial analysis, only correlation, covariance analysis and p-value test are conducted. More advanced spatial techniques can be used. For instance, hotspot analysis and network analysis shall be adopted to have a deeper insight of the Twitter data. New machine learning methods are also worth trying but many technical problems emerged when trying to deal with grammatical errors, special characters and emojis.

5.2.2 Limitation of Twitter Data

As mentioned before, only a small fraction of the Twitter users is willing to share their geo-location (coordinates with latitude and longitude) with others. The incompleteness of tweet collection leads the potential bias to the final result. Also, when obtaining twitter data there are many request and rate limitations, which stop people from getting enough tweets in time.

Twitter data itself also has some problems. As each tweet's coordinates (If it has) have only five digit numbers after the decimal point, too many tweets share the same location, which causes the cluster of the tweet activity. When spatial joining the feature points to the different census tracts, many census tracts do not even have a single general tweet or hurricane-related tweet. In this case, it is highly likely that no significant relationship can be found.

5.3 Recommendation

First, fewer request limitations in the future are anticipated when retrieving the data from the server of Twitter Inc. Compared to traditional emergency management approach, the biggest merit of Twitter data is that people can obtain them real-time data. In rescue time, the public sectors are often fully scheduled and more common volunteers are needed to join in the rescue. Volunteers need Twitter data to do spatial analysis to schedule their rescue planning but not everyone are affordable to a premium or an enterprise API. No one would like to see people prevented from being saved because vulnerable population don't use Twitter.

Also, according to my study, although social media data like Twitter data are very prevalent, it cannot be relied on too much when conducting a rescue or designing a resilience-oriented planning. It might bring bias and make people overlook those who are more vulnerable than others. Facing the several of natural and human-made disasters, shaping urban resilience is still a long way to go.

Bibliography

Alexander, D. E. (2014). Social media in disaster risk reduction and crisis management. *Science & Engineering Ethics, 20*(3), 717-733. Retrieved from https://link.springer.com/journal/11948

Boin, A., Comfort, L., & Demchak, C. (2010). The rise of resilience, designing resilience: preparing for extreme events. In *Designing Resilience: Preparing for Extreme Events*. University of Pittsburgh Press Pittsburgh, PA.

Cameron, D., Meko, T., Alcantara, C. and Lu, D. (2017). Seven million Texans, flood-prone land and oil refineries stand in Hurricane Harvey's path. *The Washington Post.* Retrieved from https://www.washingtonpost.com/graphics/2017/national/harvey/?utm_term=.3e5f1009f333

Carter, W. Nick. (2008). Notes on evacuation. In *Disaster Management: A Disaster Manager's Handbook* (pp. 254). Retrieved from https://www.think-asia.org/bitstream/handle/11540/5035/disaster-management-handbook.pdf?sequence=1

Centers for Disease Control and Prevention (CDC). (2016). *SVI 2014 Documentation*. Retrieved from https://svi.cdc.gov/Documents/Data/2014_SVI_Data/SVI2014Documentation.pdf.

Christian, K. (2013). *TwitterSearch*. Retrieved from https://github.com/ckoepp/TwitterSearch

City of Houston, Planning and Development Dept. (2018). *DEMOGRAPHIC DATA.* Retrieved from https://www.houstontx.gov/planning/Demographics/

Cutter, S. L., Boruff, B. J., & Shirley, W. L. (2003). Social vulnerability to environmental hazards. *Social science quarterly*, *84*(2), 242-261. https://doi.org/10.1111/1540-6237.8402002

Department of Homeland Security. (2013). *Innovative Uses of Social Media in Emergency Management.* Retrieved from https://www.dhs.gov/sites/default/files/publications/Social-Media-EM_0913-508_0.pdf

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).* https://doi.org/10.2202/1547-7355.1792.

Grossman, D. & Maclean, A. (2018). Lessons from Hurricane Harvey. *Pulizter Center.* Retrieved from https://pulitzercenter.org/reporting/lessons-hurricane-harvey

Guan, X., & Chen, C. (2014). Using social media data to understand and assess disasters. *Natural Hazards*, 74(2), 837-850. https://doi.org/10.1007/s11069-014-1217-1

Guo, J. C., & Urbonas, B. (2009). Conversion of natural watershed to kinematic wave cascading plane. *Journal of Hydrologic Engineering*, *14*(8), 839-846. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000045

Henrique, J. (2016). *GetOldTweets-Python.* Retrieved from https://github.com/Jefferson-Henrique/GetOldTweets-python

Herfort, B., Schelhorn, S. J., Albuquerque, J. P. D., & Zipf, A. (2014, May). Does the spatiotemporal distribution of tweets match the spatiotemporal distribution of flood phenomena? A study about the River Elbe Flood in June 2013. In *International Conference* *on Information Systems for Crisis Response and Management, 11*. The Pennsylvania State University. Retrieved from http://www.producao.usp.br/bitstream/handle/BDPI/46102/2482951.pdf?sequence=1&isAllowed=y

Hoffman, C. (2018). *What is an API?* Retrieved from https://www.howtogeek.com/343877/what-is-an-api/

Homeland Security Science and Technology. (2018). SMART: Social Media Analytics and Reporting Toolkit. Retrieved from https://www.dhs.gov/sites/default/files/publications/841_R-Tech_SMART-Social-Media-Analytics-Reporting-Toolkit_FactSheet-180405-508.pdf

Houston. (n.d.). In *Wikipedia*. Retrieved December 7, 2018 from https://en.wikipedia.org/wiki/Houston

H-GAC. (2018). *GIS Datasets*. [zip]. Retrieved from http://www.h-gac.com/rds/gis-data/gis-datasets.aspx

Houston-Galveston Area Council. (2018). *Our Great Region 2040*. Retrieved from http://www.ourregion.org/download/OurGreatRegion2040-FINAL.pdf

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. https://doi.org/10.3390/ijgi4031549

Hughes, A. L., St Denis, L. A., Palen, L., & Anderson, K. M. (2014, April). Online public communications by police & fire services during the 2012 Hurricane Sandy. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1505-1514). ACM.

Jeff, K., Aaron, G. & Fiona, P. (2013). *Search-tweets-python*. Retrieved from https://github.com/twitterdev/search-tweets-python

Kaminska, K. & Rutten,B. (2014). Social media in emergency management: Capability assessment. *Defence Research and Development Canada*. Retrieved from http://cradpdf.drdc-rddc.gc.ca/PDFS/unc157/p800316_A1b.pdf

Landry, J. N., Webster, K., Wylie, B., & Robinson, P. (2016). How Can We Improve Urban Resilience with Open Data? *Open Data Institute: London, UK*. Retrieved from https://data.gov.ru/sites/default/files/documents/print_version_report-resilient-cities-03-web.pdf

Luna, S. & Pennock, M. J. (2018). Social media applications and emergency management: a literature review and research agenda. *International Journal of Disaster Risk Reduction*. doi:10.5055/jem.2015.0242

Martín, Y., Li, Z., & Cutter, S. L. (2017). Leveraging Twitter to gauge evacuation compliance: spatiotemporal analysis of hurricane matthew. *Plos One, 12*(7), e0181701. https://doi.org/10.1371/journal.pone.0181701

Modh, S. (2009). *Introduction to disaster management. Macmillan.* Retrieved from https://www.researchgate.net/publication/277327554_Introduction_to_Disaster_Management

Nazer, T. H., Xue, G., Ji, Y., & Liu, H. (2017). Intelligent disaster response via social media analysis a survey. *Explorations Newsletter, 19*(1), 46-59.

National Hurricane Center. (2018). *Costliest U.S. tropical cyclones tables updated.* Retrieved from https://www.nhc.noaa.gov/news/UpdatedCostliest.pdf

National Hurricane Center. (2018). *Tropical Cyclone Report Hurricane Harvey.* Retrieved from https://www.nhc.noaa.gov/data/tcr/AL092017_Harvey.pdf

NOAA National Centers for Environmental Information. (2017). *State of the Climate: Hurricanes and Tropical Storms for Annual 2017.* Retrieved on December 7, 2018 from https://www.ncdc.noaa.gov/sotc/tropical-cyclones/201713

NYCDCP. (2013). *Coastal Climate Resilience: Designing for Flood Risk.* Retrieved from http://www.sustainablenyct.org/news/NYCDCP_DESIGNING%20FOR%20FLOOD%20RISK_DRAFT-LOW.pdf

NYCDCP. (2014). *Coastal Climate Resilience: Retrofitting Buildings for Flood Risk.* Retrieved from https://www1.nyc.gov/assets/planning/download/pdf/plans-studies/retrofitting-buildings/retrofitting_complete.pdf

Phipps, C., Levin, S., Lartey, J., Weaver, M. & Russell, G. (2017). *Thousands await rescue amid 'catastrophic' flooding in Texas – as it happened.* The Guardian. Retrieved from https://www.theguardian.com/us-news/live/2017/aug/28/ex-hurricane-harvey-houston-flooded-as-catastrophe-unfolds-in-texas-latest-updates

Rockfeller Foundation. (2016). *What is Urban Resilience? 100RC*. Retrieved from http://100resilientcities.org/resources/#section-1

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010, April). Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World Wide Web* (pp. 851-860). ACM. doi:10.1145/1772690.1772777

Scott, B. (2017). *Did Houston Flood Because of A Lack of Zoning?* The Forbes. Retrieved from https://www.forbes.com/sites/scottbeyer/2017/08/30/did-houston-flood-because-of-a-lack-of-zoning/ #7e267bb15580

TDI. (2018). Hurricane Harvey Data Call. *TDI Presentation to the Senate Business and Commerce Committee.* Retrieved from https://www.tdi.texas.gov/reports/documents/Harvey-20180123.pdf

TNRIS. (2018). *Texas Statewide Imagery and GIS Data*. Retrieved from https://tnris.org/data-download/#!/statewide

UNISDR. (2015). *Sendai framework for disaster risk reduction 2015–2030.* United Nations. Retrieved from https://www.unisdr.org/files/43291_sendaiframeworkfordrren.pdf

U.S. Department of Agriculture (USDA). (2008). *General Soil Map of Texas.* Retrieved from https://legacy.lib.utexas.edu/maps/texas/texas-general_soil_map-2008.pdf

U.S. Climate Resilience Toolkit. (2019). *Social Vulnerability Index*. Retrieved from https://toolkit.climate.gov/tool/social-vulnerability-index

Wax-Thibodeaux, E. (2018). Hurricane Harvey survey: 30 percent say lives are still upended a year after storm hit Texas. *The Washington Post.* Retrieved from https://www.washingtonpost.com/national/hurricane-harvey-survey-30-percent-say-lives-are-still-upended-a-year-after-storm-hit-texas/2018/08/22/0251a688-a610-11e8-8fac-12e98c13528d_story.html?utm_term=.611140ae844b

Wildavsky, A. B. (1988). Bowling Green State University. *Social Philosophy & Policy Center. Searching for Safety. New Brunswick, NJ: Transaction Books.*

World Bank. (2018). Urban population (% of total). *United Nations Population Division*. World Urbanization Prospects: 2018 Revision. Retrieved from https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS

Wukich, C. (2015). Social media use in emergency management. *Journal of Emergency Management, 13*(4), 281-294. doi: 10.5055/jem.2015.0242

Yamagata, Y., & Maruyama, H. (2016). *Urban Resilience*. Springer. Retrieved from https://link.springer.com/content/pdf/10.1007/978-3-319-39812-9.pdf

Yuan, F., & Liu, R. (2018). Feasibility study of using crowdsourcing to identify critical affected areas for rapid damage assessment: Hurricane Matthew case study. *International journal of disaster risk reduction*, *28*, 758-767. https://doi.org/10.1016/j.ijdrr.2018.02.003

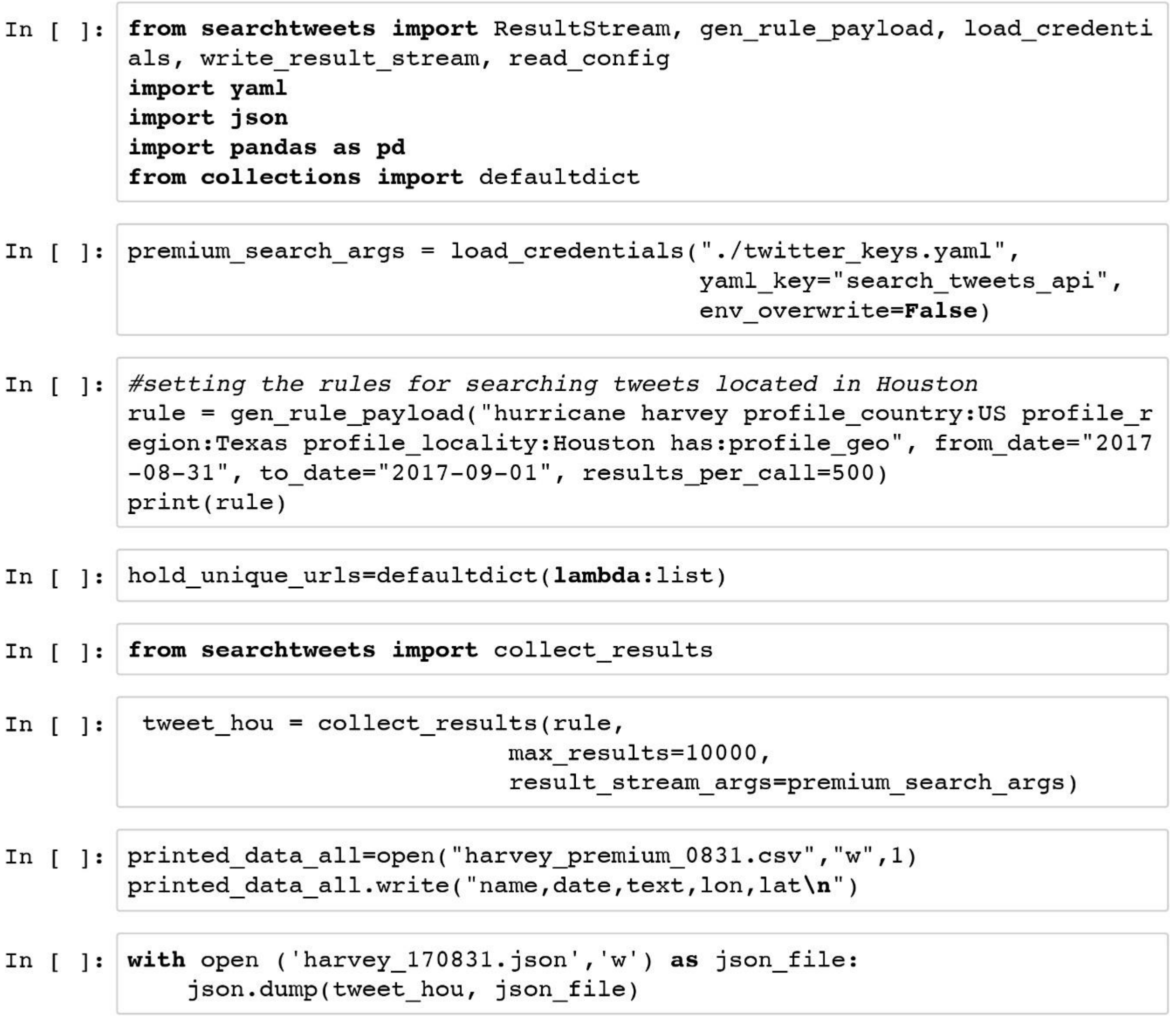
Yang, Z., Long, H. N., Stuve, J., Cao, G., & Jin, F. (2017). Harvey flooding rescue in social media. *IEEE International Conference on Big Data* (pp.2177-2185). IEEE. Retrieved from http://myweb.ttu.edu/fjin/papers/Harvey_Rescue_Scheduling.pdf

Yury, K., Chen, H., Moro, E., Van, H. P., & Cebrian, M. (2015). Performance of social network sensors during hurricane sandy. *Plos One, 10*(2), e0117288. https://doi.org/10.1371/journal.pone.0117288

Yury, K., Chen, H., Nick, O., Esteban, M., Pascal, V. H., & James, F., et al. (2016). Rapid assessment of disaster damage using social media activity. *Science Advances, 2*(3), e1500779. doi:10.1126/sciadv.1500779

Appendix

Sample Code



```
In [ ]: from searchtweets import ResultStream, gen_rule_payload, load_credenti
        als, write_result_stream, read_config
        import yaml
        import json
        import pandas as pd
        from collections import defaultdict

In [ ]: premium_search_args = load_credentials("./twitter_keys.yaml",
                                               yaml_key="search_tweets_api",
                                               env_overwrite=False)

In [ ]: #setting the rules for searching tweets located in Houston
        rule = gen_rule_payload("hurricane harvey profile_country:US profile_r
        egion:Texas profile_locality:Houston has:profile_geo", from_date="2017
        -08-31", to_date="2017-09-01", results_per_call=500)
        print(rule)

In [ ]: hold_unique_urls=defaultdict(lambda:list)

In [ ]: from searchtweets import collect_results

In [ ]:  tweet_hou = collect_results(rule,
                                  max_results=10000,
                                  result_stream_args=premium_search_args)

In [ ]: printed_data_all=open("harvey_premium_0831.csv","w",1)
        printed_data_all.write("name,date,text,lon,lat\n")

In [ ]: with open ('harvey_170831.json','w') as json_file:
            json.dump(tweet_hou, json_file)
```

In [ ]:

```
#Write the data into a csv
count=0
error=0
for tw in tweet_texas:
    try:
        name=tw['user']['name']
        tweet_txt= str(tw['text']).replace(",","").replace("\n"," ").r
eplace("\r"," ")
        date= str(tw['created_at']).replace(" +0000","").replace(","," 
")
        lon=str(tw['user']['derived']['locations'][0]['geo']['coordina
tes'][0])
        lat=str(tw['user']['derived']['locations'][0]['geo']['coordina
tes'][1])
        hold_unique_urls[name]=[tweet_txt,date,lon,lat]
        printed_data_all.write("%s,%s,%s,%s,%s\n" % (name,date,tweet_t
xt,lon,lat))
        count=0
    except:
        error+=1
printed_data_all.close()

printed_data=open("harvey_pr_0831.csv","w",1)
printed_data.write("name,date,text,lon,lat\n")

for tweet_name,list_elements in hold_unique_urls.items():
        printed_data.write("%s,%s,%s, %s, %s\n" % (tweet_name,list_ele
ments[1],list_elements[0], list_elements[2], list_elements[3]))
printed_data.close()
```

In [ ]:

```
df_0831=pd.read_csv("zhou/harvey_pr_0831.csv", engine='python',encodin
g='utf-8', error_bad_lines=False)
df
```

In [ ]:

```
df.to_csv(r 'h_0831.csv', encoding='utf-8')
```