Nat Hazards (2015) 79:1663–1679
DOI 10.1007/s11069-015-1918-0

ORIGINAL PAPER

Understanding social media data for disaster management

Yu Xiao[1] · Qunying Huang[2] · Kai Wu[1]

Received: 12 March 2015 / Accepted: 27 July 2015 / Published online: 21 August 2015

Abstract Social media data are increasingly being used in disaster management for information dissemination, establishment of situational awareness of the "big picture" of the disaster impact and emerged incidences over time, and public peer-to-peer backchannel communications. Before we can fully trust the situational awareness established from social media data, we need to ask whether there are biases in data generation: Can we simply associate more tweets with more severe disaster impacts and therefore higher needs for relief and assistance in that area? If we rely on social media for real-time information dissemination, who can we reach and who has been left out? Due to the uneven access to social media and heterogeneous motivations in social media usage, situational awareness based on social media data may not reveal the true picture. In this study, we examine the spatial heterogeneity in the generation of tweets after a major disaster. We developed a novel model to explain the number of tweets by mass, material, access, and motivation (MMAM). Empirical analysis of tweets about Hurricane Sandy in New York City largely confirmed the MMAM model. We also found that community socioeconomic factors are more important than population size and damage levels in predicting disaster-related tweets.

Keywords Social media · Disaster management · Digital divide · Hurricane Sandy

Electronic supplementary material The online version of this article (doi:10.1007/s11069-015-1918-0) contains supplementary material, which is available to authorized users.

✉ Yu Xiao
yuxiao@tamu.edu

Qunying Huang
qhuang46@wisc.edu

Kai Wu
hardy011@tamu.edu

[1] Department of Landscape Architecture and Urban Planning, Hazard Reduction and Recovery Center, Texas A&M University, 3137 TAMU, College Station, TX, USA

[2] Department of Geography, University of Wisconsin-Madison, Madison, WI 53706, USA

1 Introduction

The past decade saw a surge in the use of social media, such as blogs, chat rooms, wikis, Facebook, Twitter, Flickr, LinkedIn, YouTube Channels, and Yelp, for information sharing among social groups. Social media data are increasingly being used in disaster management for disseminating critical information to the public about the hazard event, relief, and recovery, for establishing situational awareness of the "big picture" of the disaster impact and emerging incidences overtime, and for grassroots-level, peer-to-peer backchannel communications to gather, verify, and disseminate information (Sutton et al. 2008; Lindsay 2011; Houston et al. 2014).

While social media has gained popularity as a promising channel to expand the horizon of disaster management, the social inequality in the usage of social media data should make us cautious about the use of these tools for such purposes. Due to the "digital divide," referring to the gap between those who do and do not have access to information and communication technologies (van Dijk 2006), certain groups (i.e., low income, low education, and elderly) may lack the tools and skills to access social media and therefore may be left out of information sharing through social media. The situational awareness information extracted from social media data may be biased because certain areas may be severely damaged by the disaster, and therefore, these areas may have extremely low participation in social media usage. As a result, their needs can be significantly underestimated.

In this study, we utilize a mass–material–access–motivation (MMAM) model to understand the social and spatial inequities in the generation of tweets after a disaster. This model goes beyond the "digital divide" to explain the spatial heterogeneity in the usage of social media. Different from previous studies (i.e., Ames and Naaman 2007; Li et al. 2013) that explain an individual's participation in social media, this study uses community, an aggregation of individuals in a certain geographic area, as the unit of analysis. Specifically, we answer the following research questions: What factors affect the number of tweets generated from a geographic area after a disaster? What socioeconomic factors explain the spatial variation in the number of tweets posted after the disaster? We analyzed tweets generated from census tracts in New York City after the 2012 Hurricane Sandy to empirically test our model.

2 Use of social media in disaster management

Social media can be used for enhancing communications before, during, and after a disaster (Houston et al. 2014). In recent years, social media has evolved from being a passive outlet of information (i.e., disseminating static information on how to prepare for disasters) to an emergency management tool that is capable of distributing real-time warning information, receiving requests for assistance, and establishing situational awareness based on user activities (Lindsay 2011). Social media can also be used for peer-to-peer backchannel communications that increase the social capacity of information generation and dissemination.

2.1 Real-time dissemination of information

Traditionally, crisis and risk communications have heavily relied on mass media, such as radio and television. In recent years, social media has emerged as another important source for information dissemination. Compared with the one-way communication of the traditional mass media, social media breaks down the traditional sender/receiver model. Users of social media can both receive and post messages. Instead of waiting for professional news reporters to arrive on-site to report the situation, individuals can gather first-hand information and disseminate it through social media in real time. Messages can be quickly forwarded to many people via social media channels through the users' social networks. Studies show that during the immediate aftermath of the 2010 Haiti earthquake, information about the quake was first released through social media sources (Keim and Noji 2011). Local communities can also use social media to enhance emergency responses. For instance, Texas A&M University implemented a CodeMaroon system to communicate emergency information to its students, faculty, and staff. Campus members can sign up for this service with their university ID and password. Community members not affiliated with Texas A&M University can follow "TAMUCodeMaroon" on Twitter. The CodeMaroon system was used to disseminate warning messages to thousands of people shortly after the onset of emergencies, such as a fire at a chemical plant close to campus, an on-campus chemistry laboratory explosion, and bomb threats (Villarreal and Sigman 2010; The Dallas Morning News 2013).

2.2 Establish situational awareness

Besides being a channel for pushing information to community members, emergency responders and policy makers can pull social media data to monitor evolving situations. Citizens in the disaster area are "sensors"; they can provide real-time, geo-referenced information to supplement crisis information generated by professional sources (De Longueville et al. 2010a, b). For example, Schnebele and Cervone (2013) showed that volunteered geographic information (VGI) from social media sources such as Flickr, YouTube, and Wikipedia can be fused with images from remote sensing (i.e., Landsat TM) and digital elevation model (DEM) to create flood hazard maps. Even a small amount of VGI can dramatically improve the quality of hazard mapping.

Situational information generation can be active or passive, depending on whether users are conscious of the uses of their social media information. Active information generation refers to social media users actively reporting incidences or requesting help in the hope that relief organizations can respond to urgent cases immediately. In the case of the Haiti earthquake, community crisis maps were generated based on near real-time incident and status reports from social media users. These crisis maps were then used by relief organizations to coordinate, plan, and execute responses (Gao et al. 2011).

Passive information generation refers to data mining of existing social media data to establish situational awareness without the users actively seeking or requesting responses from relief organizations. For instance, Vieweg et al. (2010) coded microblogged information about the April 2009 Oklahoma Grassfires and 2009 Red River Floods from Twitter into categorizes such as warnings, hazard extents, evacuations, volunteering, animal management, and damage/injury reports. Using Hurricane Sandy as an example, Huang and Xiao (2015) coded social media messages into different themes within different disaster phases during a time-critical crisis, and a classifier based on logistic regression is trained and used for automatically classifying the social media messages into various topic categories during various disaster phases. Imran et al. (2013) utilized machine learning techniques to extract information from disaster-related messages posted on Twitter into several categorizes including warnings, casualties and damage, donations, and information sources. The coded information can be further analyzed over space and time to inform the situational awareness of the incidences as they unfold. Moreover, Liu et al. (2008) demonstrated that the activities of the disaster-specific Flickr groups could be analyzed to document disaster impact, response, and recovery efforts over time. The Australian Government developed an Automated Web Text Mining (ESA-AWTM) system that analyzes Twitter messages to provide incidence identification, near real-time notification, and monitoring (Cameron et al. 2012). Similarly, Kumar et al. (2011) designed an application called "TweetTracker" to track, analyze, and monitor tweets for disaster relief. This application can report separately geo-referenced and non-geo-referenced tweets, support keyword search, and generate and display trends of keywords specified by the user. Information extracted from social media data can help policy makers understand the big picture of the emergency situation in near real time.

2.3 Backchannel communications

Social media can support informal public peer-to-peer backchannel communications that travel parallel to official channels. Backchannel communications represent citizen power in the acquisition and sharing of information in emergency situations when traditional media outlets provide insufficient information about local conditions or lag behind in their responses. It should be noted that Twitter itself as a communication platform is not a "backchannel"; rather, it can support backchannel communications among users. An example of backchannel communications was the use of social media in the 2007 Southern California wildfires (Palen 2008; Sutton et al. 2008). The news coverage about the wildfires was primarily focused on urban areas. As such, residents in rural areas were frustrated with the lack of information in their localities. They turned to social media to obtain information from each other (Sutton et al. 2008). Another example was the 2007 Virginia Tech campus shooting. Right after the shooting, students checked the safety of their friends via text messages and instant messages. They also relied on Facebook activities to detect whether their friends were safe. A list of victims was compiled by the public even before the university officially released the names (Palen 2008). Public peer-to-peer backchannel communications supported by social media serve as important information outlets in situations where there is an "information dearth" after disasters. It enables citizens to generate and share otherwise unavailable information.

3 Reliability of social media information

Although social media has great potential for improving risk communication and information dissemination for disaster management, several concerns over the reliability of the quality of information derived from social media sources exist (Goodchild and Glennon 2010; Goodchild and Li 2012). The first is the *accuracy of information*. Although information from social media is generally accurate, there have been incidences of inaccurate or outdated information disseminated by social media. For example, it was reported that during the 2011 Tohoku earthquake and tsunami, the spread of tweets for assistance continued even after the victims had been rescued (Lindsay 2011). Using geo-tags of tweets to identify the locations of incidences can be problematic because the reporter may be reporting something he saw earlier at a different location (Gao et al. 2011).

Second is the *malicious use of social media*. Social media can be misused for pranks or terrorist attacks (Lindsay 2011). Falsified calls for help can be entered on social media sites from places without real emergencies. Nefarious groups can create an initial attack and then use social media to issue calls for assistance to draw first responders to the area and harm them in a secondary attack. Therefore, when first responders and officials respond to emergencies, they should be aware of the potential for the malicious use of social media (Lindsay 2011). Because tweets and messages are uncensored, rumors and falsified information can propagate through social media. Efforts have been made to separate truths from rumors on social media. Based on an analysis of how information spread through the Twitter network, Mendoza et al. (2010) found that rumors tended to be questioned much more than truths.

The third concern is *bias in data generation*. Before we can fully trust the situational awareness established from social media data, we need to ask whether there are biases in data generation. As discussed in the previous section, technology allows us to extract information from social media data to generate near real-time crisis maps. But can we simply associate more tweets with more severe disaster impacts and therefore higher needs for relief and assistance in that area? If we rely on social media for real-time information dissemination, who can we reach and who has been left out? Due to the uneven access to social media and heterogeneous motivations in social media usage, situational awareness based on social media data may not reveal the true picture.

Studies show that participation in social media was uneven across social groups, over space and time. Austin et al. (2012) used the social-mediated crisis communication (SMCC) model to explain how information is distributed through social media directly and indirectly. They found that besides convenience and personal involvement, third-party influence, such as personal recommendations, promotes social media usage. Li et al. (2013)'s work disclosed the spatial and temporal heterogeneity in social media use. They found two peaks in tweets among Los Angeles users, one around 1:00 to 2:00 pm and the other around 8:00 to 9:00 pm, while Flickr users are more active during the weekends. They also explored the relationships between locations of social media data and socioeconomic characteristics of local people. They found that well-educated people in the occupations of management, business, science, and arts are more likely to be involved in the generation of geo-referenced tweets and photographs (Li et al. 2013). It should be noted that many of the existing studies (i.e., Dutta-Bergman 2004, 2006; Austin et al. 2012) focused on explaining individual's participation in social media. Scant research has studied participation in social media at the community level or in aggregated geographic areas. Because spatial situational awareness relies on data aggregated at certain geographic scales, it is important to understand the generation of data at aggregate spatial levels.

In this paper, we do not address the accuracy of information and malicious use of social media data. Instead, we fill the void by examining the mechanism of data generation across aggregated spatial units in natural hazard situations, which directly affects the accuracy of situational awareness established based on social media data for effective disaster management.

4 Conceptual framework

We explain the generation of social media data from certain geographic areas by the mass–material–access–motivation (MMAM) model. First, the generation of social media data is associated with the *mass*, or population size, in an area. Other things being equal, the larger the population, the larger the number of messages that could be generated in the social media because more people can potentially make posts from that geographic location.

The availability of *material* for reporting (i.e., flood inundation, damaged buildings, broken trees, and search and rescue actions) also affects the number of social media messages. We expect the number of disaster-related messages to have a quadratic relationship with the level of damage. In areas with zero or little damage, fewer messages will be generated because of the low exposure to disaster impact, and therefore, less material exists to share on social media. As the level of damage increases, the number of messages generated increases because there are more visual cues about damage and more discussions of disaster impact. The number of messages will reach a peak point in areas with medium levels of damage. After that, the number of messages drops as the damage increases because people may not have access to the area or may be evacuated from the severely damaged areas for longer durations.

Aside from variables related to the possibility of generating tweets, we also expect socioeconomic factors to affect the number of disaster-related tweets, which reflect the "digital divide" hypothesis. Social groups differ in their *access* to technologies. In this work, we hypothesize that people with lower incomes who are less educated, minorities, and women are less likely to have access to technologies, such as smart phones; therefore, they are less likely to share and receive information via social media. College education is an important factor to affect Twitter usage. Studies found those with some college or above education represent the majority on Twitter (Bennett 2011; Skelton 2012). Younger people are more likely to embrace new technology and, therefore, are more likely to tweet more. It should be noted that these hypotheses are based on the experiences of the developed countries. The "digital divide" caused by differences in socioeconomic status may be lessened in the case of the developing countries where mobile phones are often shared among family members and friends (Kalba 2008).

The last factor we will look at is *motivation*. Unlike in prior studies (i.e., Li et al. 2013), we hypothesized a nonlinear correlation between the number of tweets and measurements of wealth, such as income and housing value, because the *motivation* of social media use may differ across income levels. We expected the amount of tweets to increase as income and housing values increase because higher incomes can increase access to communication technologies. However, after a certain tipping point, the correlation between tweets and income/housing values becomes negative. Specifically, the wealthier the neighborhood, the fewer the number of tweets that will be posted because the elite class may have less motivation to post on public forums and platforms about natural disaster events. They may have privacy concerns and/or different focus on important matters in life.

5 Research methodology

5.1 Event of study

The event of study is Hurricane Sandy, the second-costliest cyclone to hit the USA since 1900 (Blake et al. 2013). Sandy made landfall near Brigantine, New Jersey, on October 29, 2012, causing tremendous damage to the northeastern states. The total loss associated with Sandy was estimated around $50 billion, and the death toll in the USA was 72 (Blake et al. 2013).

New York City was hit extremely hard by Hurricane Sandy. Immediately after the storm, New York City experienced a widespread blackout that affected approximately

2 million people. In the hardest hit areas, power was not restored until several months after the storm (Gibbs and Holloway 2013). New York City's subway system suffered its worst flooding in a century. Airports, tunnels, and other transportation facilities also experienced extensive damage. It took more than 2 weeks to return many of these damaged facilities to normal operations (Henry et al. 2013). The storm also forced thousands of people out of their homes. Sandy caused damage to about 620 homes and interrupted essential lifeline services to approximately 8500 more homes, creating a high demand for temporary housing and shelter services. According to New York City officials, about 6800 evacuees were housed in 73 shelters as a result of the storm. Many others stayed with friends and family (Gibbs and Holloway 2013).

In addition to relying on traditional media, the New York City government also disseminated critical information about the storm through social media channels such as Twitter and YouTube channels. New York City government sent out more than 2000 messages via Twitter and gained more than 175,000 social media followers during the storm (Gibbs and Holloway 2013).

5.2 Data

Table 1 listed all variables used in this analysis and the data sources. The dependent variable of this analysis is the number of tweets (*TweetNum*). These data came from Twitter, a popular microblogging service. We retrieved messages posted on Twitter during October 10 and November 27, 2012, from Gnip,[1] by sending a geographic query with the boundary of the selected study area in New York. A total of 1,763,141 tweets were collected. In addition to the message text content, each tweet includes metadata, such as the timestamp of posting, geo-tag (location), and author profile information, which includes author location, profile description, number of tweets, and number of followers and friends.

We first selected tweets related to Hurricane Sandy. We started by detecting a set of hashtags related to Hurricane Sandy from the collected data. The following hashtags were identified as the top ones related to Hurricane Sandy:

> beprep, blackoutnyc, breakingstorm, franken-storm, frankenstorm, frankenstormsupplies, hurricane, hurricaneny, hurricanenyc, hurricaneprep, hurricanepreparation, hurricanerelief, hurricanes, hurricanesandy, hurricanesandyaftermath, hurricanesandyproblems, hurricanesandysuppprt, newyorkhurricane, newyorksandy, njpower, nychurricane, nycsandy, nycsandyneeds, nycstorm, nyhurricane, nysandy, nystorm, sandy, sandyaftermath, sandyaid, sandycommute, sandyhelp, sandyhuracan, sandyinny, sandyisknockingatmydoor, sandylove, sandyny, sandynyc, sandyprep, sandypreparation, sandyproblems, sandyrecovery, sandyregistry, sandyrelief, sandyshurricane, sandysucks, sandyvolunteer, storm, stormprep, storms, superstorm, superstorms

We then used those hashtags to filter out messages not relevant to the disaster. If a tweet text did not contain any predefined hashtag keyword in the hashtags or message text content, it was not included in the following analysis.

Lastly, we selected all Hurricane Sandy-related tweets with available geo-tag information (a total of 35,751 tweets). We then overlaid them with census tract to extract the number of tweets by census tract. This allowed us to test the relationships between the number of tweets and socioeconomic characteristics of the census tract.

[1] http://gnip.com/

Table 1 Variables and data source

Variable name	Description	Data source
Dependent variable		
TweetNum	Number of tweets	Twitter, downloaded from Gnip
Independent variables		
Mass		
Population	Population in 1000	2012 American Community Survey, 5-year estimate, US Census Bureau
Material		
%InundatedArea	Percent of land area inundated by surge	FEMA MOFT
Access		
%White	Percent white population	2012 American Community Survey, 5-year estimate, US Census Bureau
MedianAge	Median age	2012 American Community Survey, 5-year estimate, US Census Bureau
SexRatio	Male-to-female ratio	2012 American Community Survey, 5-year estimate, US Census Bureau
%CLGEdu	Percent of population 25 and over with at least college education	2012 American Community Survey, 5-year estimate, US Census Bureau
Access and motivation		
MedianHHInc	Median household income in $1000	2012 American Community Survey, 5-year estimate, US Census Bureau
MedianHousingValue	Median housing value in $100,000	2012 American Community Survey, 5-year estimate, US Census Bureau

The independent variables are all from the 2012 American Community Survey (ACS), 5-year estimates, reported by the US Census Bureau, with the only exception of the variable on the extent of disaster damage. The socioeconomic variables used in this analysis include population size, percent white population, median age, male-to-female sex ratio, percent of population 25 and over with at least college education, median household income, and median housing value, all measured at the census tract level. To reduce the decimal points in the estimated coefficients, we adjusted the levels of a few numerical variables. Population was measured in thousands, median household income was measured in thousands of dollars, and median housing value was measured in $100,000.

The raw data for calculating the extent of disaster damage by census tract came from the Hurricane Sandy storm surge data reported by the FEMA Modeling Task Force (MOTF). The FEMA MOTF data were generated from a combination of multi-hazard loss modeling and "ground-truth" from the US Geological Survey surge sensor data, field observations, and aerial photograph imagery assessments (FEMA 2014). As shown in Fig. 1, the storm surge toppled neighborhoods in the lower south side and waterfront areas of Brooklyn and Queens. It also inundated a ring of low-lying waterfront areas in Manhattan. We overlaid the boundary of the Hurricane Sandy storm surge with census tracts to calculate the percentage of land area in each census tract toppled by the Sandy storm surge.

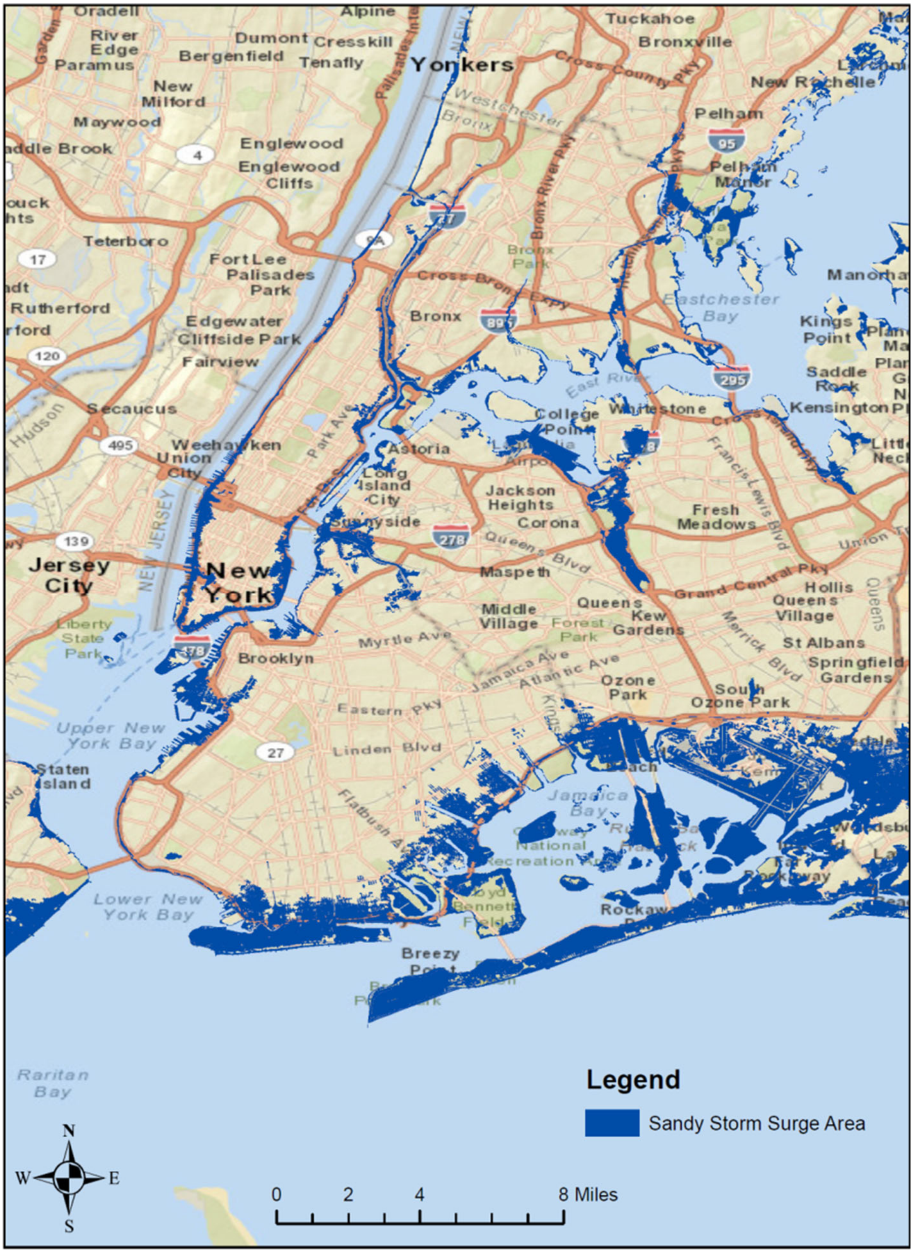

Fig. 1 Hurricane Sandy storm surge in New York city. *Data source*: FEMA (2014)

5.3 Regression analysis

To examine the explanatory powers of variables related to the possibility to generate tweets and those related to the digital divide and motivation, we compared four sets of models:

Model 1 $TweetNum = \alpha + \beta_1 Population + e,$

Model 2 $TweetNum = \alpha + \beta_1 Population + \beta_2\ \%InundatedArea + \beta_3\ \%InundatedArea^2 + e,$

Model 3 $TweetNum = \alpha + \beta_1 \%White + \beta_2 MediumAge + \beta_3 SexRatio + \beta_4 \%CLGEdu + \beta_5 MedianHHInc + \beta_6 MedianHHInc^2 + \beta_7 MedianHousingValue + \beta_8 MedianHousingValue^2 + e$, and

Model 4 $TweetNum = \alpha + \beta_1 Population + \beta_2 \%White + \beta_3 MediumAge + \beta_4 SexRatio + \beta_5 \%CLGEdu + \beta_6 MedianHHInc + \beta_7 MedianHHInc^2 + \beta_8 MedianHou\sin gValue + \beta_9 MedianHou\sin gValue^2 + \beta_{10} \%InundatedArea + \beta_{11}\ \%InundatedArea^2 + e$

Models 1 and 2 include variables related to the possibility to generate tweets. Model 1 is the base model that explains the number of tweets by population size. Model 2 added two damage variables to Model 1, namely the percent of land area inundated by the Hurricane Sandy surge and its square term. Model 3 includes only the set of socioeconomic variables as independent variables. We can compare Model 3 with Models 1 and 2 to derive the relative explanatory power of socioeconomic variables versus the population and damage variables in explaining the variations in the number of tweets generated from the census tract. Model 4 is the full model of the analysis. It includes all the independent variables used in the previous models.

Because the landmark sites in New York City attract many visitors year round, at these locations, many tweets were not generated by the residents, but the visitor population. Therefore, we excluded census tracts primarily composed of landmarks (i.e., the Central Park, LaGuardia Airport, and John Kennedy Int'l Airport) and parks and other vegetation land use areas from the regression analysis. As a limitation of this study, the non-landmark census tracks may still have tweets generated by the visitor or floating population, i.e., someone tweeted about hurricane Sandy from his workplace. We cannot fully match tweets with users' residential locations and demographics. If we assume error associated with such geographic mismatch is randomly distributed across all non-landmark census tracks, results from regression analysis should still be valid.

Because the dependent variable of this analysis—the number of tweets—is count data that are discrete in nature, ordinary least square (OLS) regression is not the best estimation method (Greene 2003). Instead, Poisson's regression has been widely used to fit such data (Greene 2003). Therefore, we fit the models by Poisson's regression.

6 Results

6.1 Descriptive analysis

The number of tweets by census tract is shown in Fig. 2. A visual scan of Fig. 2 shows that many landmark sites (i.e., airports, major parks, and green spaces) had a high number of tweets. Table 2 shows the comparison of landmark and non-landmark census tracts in number of tweets and registered residents. The average number of tweets of all census tracts was 19.2 or 5.1 tweets per 1000 people. In contrast, the John F. Kennedy Int'l Airport, LaGuardia Airport, and Prospect Park had no registered residents; however, they had over 100 tweets. Central Park had a registered population of only two, but it had 380 tweets, a density of 190,000 tweets per 1000 people. The Greenwood Cemetery and Flushing Meadows Corona Park also had a low population, but many tweets. Their respective tweet densities were 1391 and 771 tweets per 1000 people. The other smaller parks and green spaces on average had 162.9 tweets per 1000 people, which was more than 30 times higher than the non-landmark census tracts (Table 2). Tweets were generated by a large floating and visitor population to these landmark sites.

Table 3 presents descriptive statistics of all variables used in the regression analysis. As discussed earlier, census tracts containing landmark sites were excluded from the analysis. Therefore, Table 3 only reports the descriptive statistics of non-landmark census tracts used in the regression models. On average, about 19 Hurricane Sandy-related tweets were posted from each census tract between October 10 and November 27, 2012. There was wide variation in the number of tweets by census tract. Some census tracts had zero tweets, and some had 382 tweets. The census tracts also varied considerably in population size. The average population of a census tract was 3900, with the lowest being 500 and highest being 26,910. On average, the storm surge from Hurricane Sandy toppled 5 % of the land area of census tracts. The spatial distribution of damage was uneven. Some census tracts in the low-lying areas were completely inundated, and those on higher ground suffered no damage at all.

New York City neighborhoods are well known for their great diversity in socioeconomic conditions. The racial composition of census tracts ranged from having all minorities to all whites. The census tracts around the Central Park in Manhattan, on the west end of the Rockaways, around the downtown Brooklyn areas, and in pockets of Queens had a majority white population. Many other census tracts in Brooklyn and Queens had a white population of less than ten percent. The median household income in the census tract was also on a wide swing, ranging from $22,560 to $318,650. Median housing values in the poorest neighborhoods was below $10,000 and in the richest was well above $1 million.[2] On average, the median age of census tract was 36 years old, and the male-to-female sex ratio was 0.93. Some census tracts were very young (with more than half of the population <15 years old), and some were fairly old (with more than half of the population older than 60). Some had almost three times more men than women, while others had four times more women than men. On average, 32.5 % of population aged 25 and older had at least college degree. Some census tracts had no residents with a college degree or above, and some had all residents with a college degree or above. The mosaic of census tracts provided an ideal case to study how socioeconomic conditions affect public participation in social media.

6.2 Regression models

Results from Poisson's regression models are reported in Table 4. Population was the only independent variable in Model 1. Population size is positively and significantly correlated with the number of Hurricane Sandy-related tweets. For every 1000 people increase, the estimated number of tweets increases by a factor of 1.15 (or $e^{0.143}$). The scale of the

[2] The ACS dataset marked median housing values lower than $10,000 as "<$10,000" and higher than 1 billion as ">$1 million." Because no exact value was reported, we used $10,000 or $1 million as the median housing value for those census tracts.

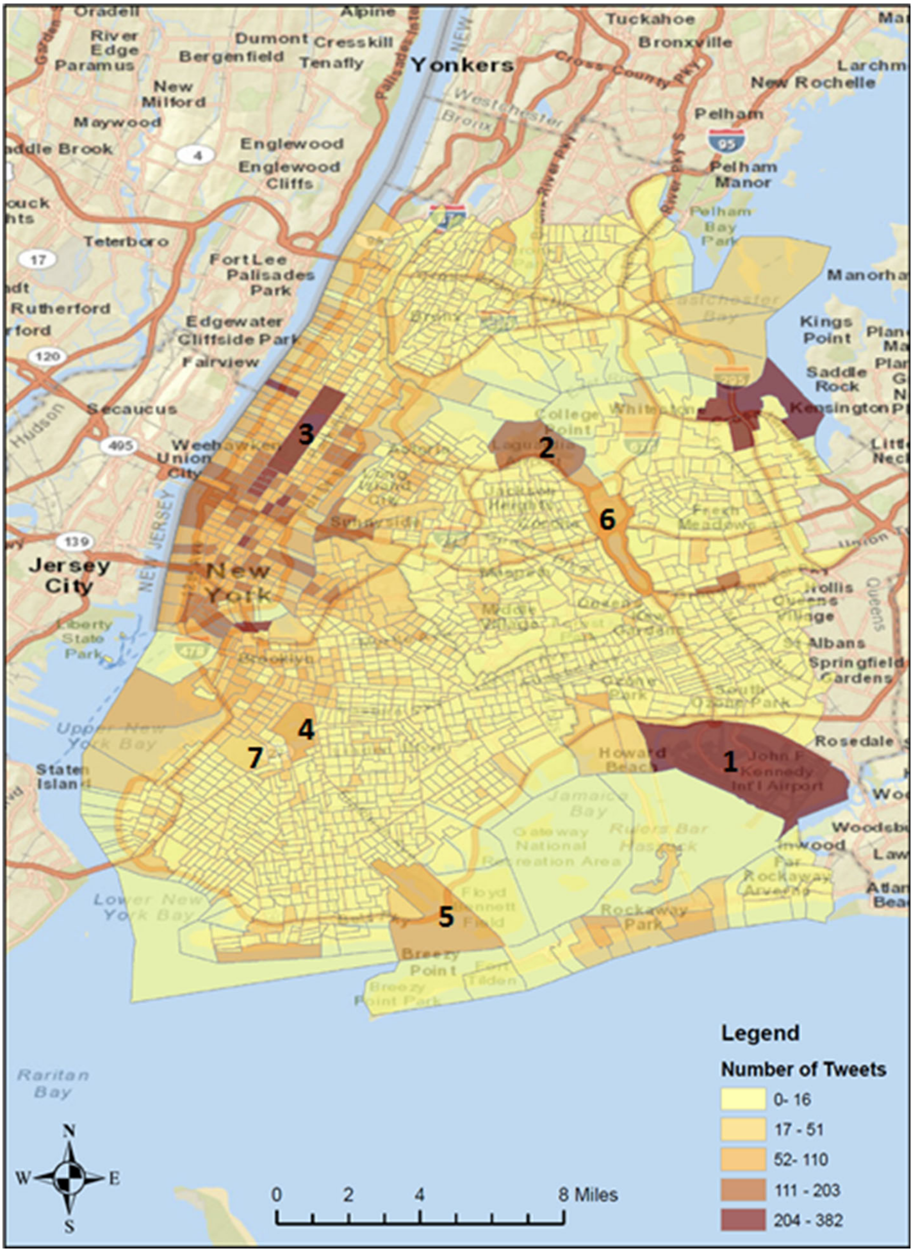

Fig. 2 Number of tweets by Census Tract. *Note*: *1* John F. Kennedy Int'l Airport, *2* LaGuardia Airport, *3* Central Park, *4* Prospect Park, *5* Brooklyn Marine Park, *6* Flushing Meadows Corona Park, and *7* Greenwood Cemetery

coefficient on the population variable barely changes after damage variables are added to the regression (see Model 2 for details). Moreover, both *%InundatedArea* and its quadratic form are statistically significant at the 0.01 level, indicating the extent of storm surge

Table 2 Number of Tweets: landmark locations versus non-landmark locations

	Number of tweets	Population count	Number of tweets/(1000 pop)
All census tracts (average)	19.2	3784.1	5.1
Airports			
LaGuardia Airport	127.0	0.0	–
John F. Kennedy Int'l Airport	260.0	0.0	–
Major parks and green spaces			
Central Park	380.0	2.0	190,000.0
Prospect Park	105.0	0.0	–
Brooklyn Marine Park	54.0	0.0	–
Flushing Meadows Corona Park	81.0	105.0	771.4
Greenwood Cemetery	32.0	23.0	1391.3
Other green spaces (average)	7.4	45.2	162.9
Non-landmark census tracts (average)	18.9	3901.3	4.8

Table 3 Descriptive statistics of variables

Variable name	Obs.	Mean	Std. dev.	Min	Max
TweetNum	1701	18.85	35.80	0.00	382.00
Population (1000)	1701	3.90	2.12	0.50	26.91
%InundatedArea	1701	5.35	17.73	0.00	100.00
%White	1701	45.73	30.12	0.00	100.00
MedianAge	1701	35.97	6.19	12.80	61.70
SexRatio	1701	0.93	0.17	0.43	2.78
%CLGEdu	1701	32.53	20.95	0.00	100.00
MedianHHInc ($1000)	1701	61.40	36.87	22.56	318.65
MedianHousingValue ($100,000)	1701	5.53	2.00	0.10	10.00

inundation being a significant factor to explain the number of tweets. The negative coefficient on the quadratic form of *%InundatedArea* implies a reversed U-shaped relationship between the level of damage and the number of tweets. Before the inundation area reaches 34.6 % of the total land area in the census tract, as the damage-level increases, more tweets are produced. After the damage exceeds 34.6 % of the land area, as the damage-level increases, fewer tweets are produced. These results confirm our hypotheses about population and damage levels being factors that influence the ability to generate tweets. The pseudo-R^2 for Models 1 and 2 are 0.086 and 0.114, respectively.

Model 3 explains the variation in number of Hurricane Sandy-related tweets only with socioeconomic variables. All the variables are significant at the 0.05 level except for *%White*, indicating that minority status does not significantly correlate with social media usage. Our model largely supports the digital divide hypothesis that the census tracts with higher percentages of young, male, and educated populations are more likely to have more tweets. Specifically, a 1-year increase in median age decreases the number of tweets by a factor of 0.973 (or $e^{-0.028}$). An increase of one in the male-to-female sex ratio is correlated with an increase in tweets by a factor of 1.738 (or $\mathrm{e}^{0.553}$). A one percentage point increase in people 25 years and older with at least college education is associated with an increase

Table 4 Poisson's regression results (dependent variable = *TweetNum*)

	Model 1		Model 2		Model 3		Model 4	
	Coef.	Sig.	Coef.	Sig.	Coef.	Sig.	Coef.	Sig.
Constant	2.315136	***	2.276756	***	0.9630914	***	0.374563	***
Population	0.143118	***	0.138923	***			0.095919	***
%InundatedArea			0.047423	***			0.017336	***
%InundatedArea^2			−0.000686	***			−0.000215	***
%White					−0.000283		−0.000442	
MedianAge					−0.027522	***	−0.027528	***
SexRatio					0.552754	***	0.818850	***
%CLGEdu					0.026955	***	0.026806	***
MedianHHInc					0.022165	***	0.017193	***
MedianHHInc^2					−0.000062	***	−0.000048	***
MedianHousingValue					0.033630	**	0.087176	***
MedianHousingValue^2					−0.002949	***	−0.006528	***
N	1701		1701		1701		1701	
Pseudo R^2	0.0861		0.1139		0.5006		0.5332	

* Significant at 0.1 level; ** significant at 0.05 level; and *** significant at 0.01 level

in estimated tweets by a factor of 1.027 (or $e^{0.027}$). For instance, compared with a community with 35 % of its residents aged 25 years and older with college and above education, those with 55 % of residents achieving such educational attainment will generate 1.704 (or 1.027^{20}) times more tweets, holding all other variables constant.

The model confirms a statistically significant nonlinear relationship between tweets and wealth (measured by income and median housing value), which reflects both the digital divide and the motivation hypotheses. The negative coefficients on $MedianHHInc^2$ and $MedianHousingValue^2$ indicate reversed U-shaped relationships between tweet generation and the wealth variables. Starting from a relatively low community wealth level, increasing in income and housing value increases the number of tweets generated from the community; after reaching a maximum point, the number of tweets goes down as community wealth continues to go up. More specifically, before the median household income reaches $178,170 (which is about three times the average median household income in all census tracks in New York City), increases in income are positively related to increases in tweets, which reflect the digital divide hypothesis. This relationship turns negative after the median household income reaches $178,170, whereas the number of tweets decreases as the household income increases. The turning point for median housing values is $570,280, which is slightly higher than the average median housing value in all census tracts in New York City. For census tracts with less than $570,280 in median housing values, increases in median housing values increase the number of tweets; after that, the number of tweets decreases as the housing value increases, which reflects a lack of motivation in social media participation among the wealthier class during natural disasters. The pseudo-R^2 for Model 3 is 0.5006, much higher than those of Models 1 and 2.

Model 4 is the full regression model with all relevant independent variables included. The coefficients on all independent variables are very similar in levels compared with those in Models 1–3. The reversed U-shaped correlations between the number of Sandy-related tweets and damage levels, median household income, and median housing values still exist. The turning points of change from positive to negative associations are 40.34 % of land area in the census tracts inundated by storm surge, $180,599 in median household income, and $667,726 in median housing value. The coefficient on population decreases slightly, from 0.139 in Model 2 to 0.096 in Model 4, while the coefficient on sex ratio increases from 0.553 in Model 3 to 0.819 in Model 4. For a 1000 increase in population and one point increase in the male-to-female sex ratio, the number of tweets increases by a factor of 1.101 (or $e^{0.096}$) and 2.268 (or $e^{0.819}$), respectively. The pseudo-R^2 for Model 4 is 0.533, the highest among all models.

7 Discussion and conclusions

In this study, we examined the spatial heterogeneity in the generation of tweets in natural disaster situations. We proposed the MMAM model to explain the number of tweets by mass, material, access, and motivation. Empirical analysis of tweets about Hurricane Sandy in New York City largely confirmed the MMAM model.

As expected, we found the number of tweets is significantly correlated with population size. As the rule of thumb, the larger the residential population, the more tweets there are. It should be noted that landmark sites, such as airports, major public parks, and green spaces, can have no or low numbers of residents, but many tweets. These landmark locations have a large floating and visitor population who had free time to tweet.

We also found that damage and tweets exhibit a reversed U-shaped relationship. The highest number of tweets is found in census tracts with 34.6–40.3 % of land areas inundated by the storm surge. As the damage increases from zero to about 34.6–40.3 %, the number of Hurricane Sandy-related tweets increases, probably because more information and damaged material, such as fallen trees and traffic signs or long gas waiting lines, are available for the tweets as damage increases. After that, more severe damage is associated with fewer tweets, probably because of population displacement caused by the severe damage.

Our findings also confirm the difference in access or the digital divide hypothesis. The census tracts with higher percentages of young, male, and educated people are found to be more likely to have more tweets. We also found reversed U-shaped relationships between the number of tweets and the average income/median housing value. With increases in wealth, participation in social media initially increases, probably due to the increased access to computers and mobile devices. After reaching a turning point, the participation in social media decreases with wealth, probably because of a lack of motivation among the wealthy class.

It should be noted that compared with the damage and population variables, the socioeconomic variables have higher explanatory power in explaining the variations in the tweet generation. To use social media for real-time information dissemination and/or backchannel communications in emergency situations, we need to consider the abilities to share information among different social groups. Also, the situational awareness extracted from social media data may be potentially biased due to the unbalanced participation among social groups. Future research should compare the situational awareness information extracted from social media with information collected by ground truthing, such as field observation and surveys, for validation.

References

Ames M, Naaman M (2007) Why we tag: Motivations for annotation in mobile and online media. SIGCHI conference on human factors in computing systems, San Jose, CA

Austin L, Liu BF, Jin Y (2012) How audiences seek out crisis information: exploring the social-mediated crisis communication model. J Appl Commun Res 40(2):188–207

Bennett S (2011) Twitter users better educated than Facebook users, but both dumb compared to LinkedIn. http://www.adweek.com/socialtimes/pew-social-network-education/453137. Retrieved 24 May 2015

Blake ES, Kimberlain TB, Berg RJ, Cangialosi JP, Beven II JL (2013) Tropical cyclone report-Hurricane Sandy (AL182012), National Hurricane Center. http://www.nhc.noaa.gov/data/tcr/AL182012_Sandy.pdf. Retrieved 20 Oct 2014

Cameron MA, Power R, Robinson B, Yin J (2012) Emergency Situation awareness from twitter for crisis management. WWW 2012—SWDM'12 Workshop, Lyon, France

De Longueville B, Annoni A, Schade S, Ostlaender N, Whitmore C (2010a) Digital earth's nervous system for crisis events: real-time sensor web enablement of volunteered geographic information. Int J Digit Earth 3(3):242–259

De Longueville B, Luraschi G, Smits P, Peedell S, De Groeve T (2010b) Citizens as sensors for nautural hazards: a VGI integration workflow. Geomatica 64(1):41–59

Dutta-Bergman MJ (2004) Complementarity in consumption of news types across traditional and new media. J Broadcast Electron Media 48(1):41–60

Dutta-Bergman MJ (2006) Community participation and internet use after September 11: complementarity in channel consumption. J Comput Med Commun 11(2):469–484

FEMA (2014). Hurricane Sandy impact analysis, FEMA modeling task force (MOTF). http://www.arcgis.com/home/item.html?id=307dd522499d4a44a33d7296a5da5ea0. Retrieved 15 Feb 2014

Gao H, Barbier G, Goolsby R (2011) Harnessing the crowdsourcing power of social media for disaster relief. IEEE Intell Syst 26(3):10–14

Gibbs LI, Holloway CF (2013) Hurricane Sandy after action: report and recommendations to Mayor Michael R. Bloomberg. http://www.nyc.gov/html/recovery/downloads/pdf/sandy_aar_5.2.13.pdf. Retrieved 1 Dec 2014

Goodchild MF, Glennon JA (2010) Crowdsourcing geographic information for disaster response: a research frontier. Int J Digit Earth 3(3):231–241

Goodchild MF, Li L (2012) Assuring the quality of volunteered geographic information. Spat Stat 1:110–120

Henry DK, Cooke-Hull S, Savukinas J, Yu F, Elo N, Arnum BV (2013) Economic impact of Hurricane Sandy: potential economic activity lost and gained in New Jersey and New York U.S. Department of Commerce. http://www.esa.doc.gov/sites/default/files/reports/documents/sandyfinal101713.pdf. Retrieved 1 Dec 2014

Houston JB, Hawthorne J, Perreault MF, Park EH, Hode MG, Halliwell MR, McGowen SET, Davis R, Vaid S, McElderry JA, Griffith SA (2014) Social media and disasters: a functional framework for social media use in disaster planning, response, and research. Disasters. doi:10.1111/disa.12092

Huang Q, Xiao Y (2015) Geographic situational awareness: mining tweets for disaster preparedness, emergency response, impact, and recovery. Int J Geo-Inf 4(3):1549–1568. doi:10.3390/ijgi4031549

Imran M, Elbassuoni S, Castillo C, Diaz F, Meier P (2013) Extracting information nuggets from disaster-related messages in social media. 10th international ISCRAM conference, Baden-Baden, Germany

Kalba K (2008) The global adoption and diffusion of mobile phones. Havard University, Cambridge

Keim ME, Noji E (2011) Emergent use of social media: a new age of opportunity for disaster resilience. Am J Disaster Med 6(1):47–54

Kumar S, Barbier G, Abbasi MA, Liu H (2011) TweetTracker: an analysis tool for humanitarian and disaster relief, Association for the Advancement of Artificial Intelligence. www.aaai.org

Li L, Goodchild MF, Xu B (2013) Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. Cartogr Geogr Inf Sci 40(2):61–77

Lindsay BR (2011) Social media and disasters: current uses, future options, and policy considerations, congressional research service, report number: R41987

Liu SB, Palen L, Sutton J, Hughes AL, Vieweg S (2008) In search of the bigger picture: the emergent role of on-line photo sharing in times of disaster. 5th international ISCRAM conference, Washington, DC

Mendoza M, Poblete B, Castillo C (2010) Twitter under crisis: can we trust what we RT? 1st workshop on social media analytics (SOMA'10), Washington, DC

Palen L (2008) Online social media in crisis events. Educ Q 31(3):76–78

Schnebele E, Cervone G (2013) Improving remote sensing flood assessment using volunteered geographical data. Nat Hazards Earth Syst Sci 13:669–677

Skelton A (2012) Social demographics: who's using today's biggest networks. http://mashable.com/2012/03/09/social-media-demographics/. Retrieved 24 May 2015

Sutton J, Palen L, Shklovski I (2008) Backchannels on the front lines: emergent uses of social media in the 2007 southern California wildfires. 5th international ISCRAM conference, Washington, DC

The Dallas Morning News (2013) Texas A&M issues 'code maroon' alert, evacuates buildings due to Kyle Field bomb threat. http://www.dallasnews.com/news/state/headlines/20130220-texas-am-issues-code-maroon-alert-evacuates-buildings-due-to-kyle-field-bomb-threat.ece. Retrieved 1 Mar 2015

van Dijk JAGM (2006) Digital divide research, achievements and shortcomings. Poetics 34:221–235

Vieweg S, Hughes AL, Starbird K, Palen L (2010) Microblogging during two natural hazards events: what twitter may contribute to situational awareness. CHI 2010: Crisis Informatics, Atlanta, GA

Villarreal S, Sigman A (2010) Explosion at Texas A&M Chemistry Annex Building. http://www.kbtx.com/home/headlines/93421299.html. Retrieved 1 Mar 2015