JOURNAL OF RISK RESEARCH
https://doi.org/10.1080/13669877.2019.1569098

Assessing wild fire risk in the United States using social media data

Yaojie Yue[a,b,c], Kecui Dong[b], Xiangwei Zhao[d] and Xinyue Ye[e]

[a]Key Laboratory of Environmental Change and Natural Disaster, Ministry of Education, Beijing Normal University, Beijing, China; [b]School of Geography, Beijing Normal University, Beijing, China; [c]Faculty of Geographical Science, Beijing Normal University, Beijing, China; [d]Geomatics College, Shandong University of Science and Technology, Qingdao, China; [e]Department of Informatics, New Jersey Institute of Technology, Newark, USA

ABSTRACT

Massive Geo-tagged social media data provide new opportunities for disaster risk assessment, prevention, and management. This article presents a proof of concept for assessing wildfire risk using Geo-tagged social media data, by taking wildfire risk as a function of wildfire hazard and social–ecological vulnerability. The case study of the United States shows that the regions with the highest wildfire hazard are concentrated in the Western, while the most vulnerable areas are mainly distributed in the Eastern, the Western Coast, and the Southern parts of the nation. Areas with high wildfire risk are mainly located in the Northwestern and Southeastern United States. It shows that the wildfire risk level has significant linear relationship with population density. Massive and vulnerable population might result in significant increase in wildfire risk perception. We conclude that Geo-tagged social media data have great potential in disaster risk studies.

ARTICLE HISTORY
Received 14 March 2018
Accepted 26 November 2018

KEYWORDS
Disaster risk; Twitter; wildfire; social–ecological system vulnerability

Introduction

Social networking based on social media forms a virtual community and communication platform for people to share and exchange information and viewpoints (Kietzmann et al. 2011; Wu et al. 2016; Li et al. 2017). Because of the popularity of social media platforms and devices (Albuquerque et al. 2015; Gao, Barbier, and Goolsby 2011; Miles and Morse 2007; Goodchild and Li 2012; Hong and Ye 2017), the amount of geo-tagged social media data has increased dramatically in recent years (Nicholas and Rowlands 2011; Java et al. 2007; Zhao et al. 2011; Cohen, Hughes, and White 2007; Burger et al. 2013; Ye et al. 2016; Wang et al. 2016; Wang, Ye, and Tsou 2016; Wang et al. 2017). Some geo-tagged social media data contain abundant information of disaster, which provides opportunities for disaster assessment and post-hazard countermeasures (Sakaki, Okazaki, and Matsuo 2012; Smith et al. 2015; Guan and Chen 2014; Wang and Ye 2017). Thus, research on how to use geo-tagged social media data to assess the risk of abrupt disasters, such as wildfire, is not only essential but also a challenging scientific issue for the application of big data in disaster risk management.

Mapping wildfire risk is important because wildfires are known to pose a hazard to landscape, property, and life (Arago et al. 2016). There are many methods for assessing wildfire risk at

CONTACT Xinyue Ye xye@njit.edu Department of Informatics, New Jersey Institute of Technology, Newark, NJ 07102, USA.

present, but every method has its pros and cons. Wildfire risk can be evaluated via wildfire-spread model based on historical wildfire point data (Cutter, Boruff, and Shirley 2003; Reid et al. 2009). However, such data are difficult to obtain, as well as costly and time-consuming (Peters et al. 2013; Wang et al. 2016; Wang, Ye, and Tsou 2016). Therefore, wildfire risk mapping based on statistical model by analyzing such factors as solar radiation, topography humidity, altitude, vegetation types, and population density, which are closely related to wildfire, has been put forward (Lein and Stump 2009; Parisien and Moritz 2009; Keane et al. 2010). But the researchers need to evaluate the site conditions constantly and carefully to guarantee the statistical model works, because such model is usually strictly restricted to the case study site (Peters et al. 2013). Furthermore, maps of wildfire risk based on statistical models are usually too coarse for decision makers (Arago et al. 2016). A third method is to use a wildfire ignition model to determine wildfire occurring possibility and the consequent risks (Calkin et al. 2010; Thompson et al. 2011). However, the simulation results are sometimes problematic due to many unobserved factors that might affect wildfire. Moreover, most of these studies neglect human perception on the wildfire hazard (Slavkovikj et al. 2014). In contrast, social media data transmit and collect the information related to disaster risk perception (Slavkovikj et al. 2014). Therefore, the introduction of social media data into wildfire risk assessment makes it possible to cope with wildfire timely and effectively. In addition, it can also shed light on the issues such as simulation model uncertainty and ignorance of human elements in previous studies.

Social media data have been successfully applied to many disaster management cases (Gao, Barbier, and Goolsby 2011; Miles and Morse 2007; Ghosh and Guha 2013). For example, social media data analytics has provided real-time monitoring and management for wildfire in California (Sutton, Palen, and Shklovski 2008). Emergency rescue troops and institutions achieve great success by applying social media data to monitor the earthquake in real time and make decisions during Haiti earthquake in 2010 (Yates and Paquette 2011). In Japan, social media data play an important role in interpersonal communication and post-hazard management by delivering emergency alert and searching for missing people during the tsunami in 2011 (Gao, Barbier, and Goolsby 2011). However, there are still limitations in the use of social media data in disaster management. First, social media data are unevenly distributed in space under the influence of population density, income level, and social economic status (Xiao, Huang, and Wu 2015). Meanwhile, social media data have uncertainties of both locational information (Gao, Barbier, and Goolsby 2011) and message contents (Alexander 2014). Moreover, most studies focus on disaster emergency response instead of risk assessment. Though there are many challenges in hazard mapping using social media contents, it can still provide ample opportunities to the rapid assessment of disaster risks (Tsou 2015). However, what is urgently needed is to give more in-depth discussions regarding the aspect of concepts and methods of disaster risk assessment based on social media.

Therefore, we use Tweets as an example to present a proof of concept about assessing hazard-inducing risks and social–ecological vulnerability using social media data. We adopt the United States as the case study to evaluate the spatial wildfire risk. The structure of the article follows: Part I outlines the motivation; Part II gives a detailed introduction of the data sources and research methods; Part III illustrates the research results; Part IV discusses the findings; Part V presents the conclusion.

Data and method

Data sources

The Tweets in this research refer to those related to wildfire recorded on Twitter from 29 July 2015 (the first wildfire occurrence) to 29 August 2015 (the most severe wildfire occurrence period) within the United States. To minimize the impact of uncertainties of tweets location and contents on the study results, we have formulated the following principles in the data collection.

First of all, the principle of high-density data collection is considered. Considering that wildfire and flame spread are hazard processes with strong dynamic nature and timeliness, it is necessary to collect high-density data for the sake of applying social media data to monitor the wildfire. In this article, we consider one minute as an interval to continuously collect the related records on Twitter. Second, the principle of geo-tagged tweets priority is highlighted. Many users refuse to reveal their geo-position information after taking privacy and other factors into account, therefore most tweets do not contain geographical coordinates. As such, we only use those with geographical coordinate information. Third, keyword-based filtering is essential in retrieving the relevant information (Bahir and Peled 2013). To further reduce the content uncertainty of wildfire on Twitter, we apply multiple keywords to the collected data. These keywords include: (1) "Wildfire," (2) "Rocky fire" OR "rocky wildfire," (3) "Fire evacuation," (4) "Fire closure." These keywords are selected on the basis of knowledge related to wildfire and associated government response following wildfire occurrence. In addition, we inspect every record to confirm that the data obtained from Twitter are related to wildfire and each tweet is unique.

As a result, 64,990 tweets with geographic information reflect the spatial distribution of wildfire hazards. Every tweet record includes the following contents: (1) Tweet ID (the unique identification of a tweet); (2) User_from_ID (the author ID of a tweet); (3) Text (the tweet content which includes 140 bytes at most); (4) CREATED_AT (the time when the tweet is posted on Twitter); (5) FOLLOWERS_COUNT(the number of followers); (6) COORDINATE (X and Y coordinates of the posted tweet); (7) CITY (the city of the posted tweet); (8) GEOCODE_TYPE, including GPS (it reflects the accurate geographic position where the tweet is posted), null (the system fails to get any position information of the posted tweet), and user Profile (the user provides the geographic location).

MODIS Normalized Difference Vegetation Index (NDVI) is selected as the indicator to reflect the vegetation growth state and evaluate the ecological system's vulnerability, because vegetation is an essential condition for the formation and spread of wildfires. In this article, the MODIS NDVI Monthly L3 Global 1 km product from 2010 to 2014 is used. First, the annual average NDVI is calculated. Second, the average NDVI data during 2010–2014 are obtained by applying the weighting method (calculation via field calculator) to enclose it in the wildfire risk assessment grid.

In this article, the grid serves as the assessment unit. Usually, the size of the grid is determined according to the spatial size of the study area, the characteristics of driving factors, and the accuracy of data sources (Jordaan, Jordaan, and Procter 2011; Alam 2011). In essence, it is used to determine a threshold which can achieve better balance between data spatial precision and data volume (Cutter, Boruff, and Shirley 2003). By analyzing the computational workload and mapping effects of grid at three scales, namely 1 km $\times$ 1 km, 10 km $\times$10 km, and 100 km $\times$ 100 km in the United States, we find that 1 km $\times$ 1 km grids will consume a large amount of computation, but such a size is much smaller than the average spatial range of wildfire. However, the area of 100 km $\times$ 100 km will exaggerate the impact range of a wildfire hazard. Relatively speaking, a 10 km $\times$ 10 km grid has moderate computational workload, and it can also ideally reflect the spatial distribution characteristics of wildfire hazards. Albuquerque et al. (2015) found that tweets messages near (up to 10 km) severely flooded areas have a much higher probability of being related to floods. Given this, we select a 10 $\times$ 10 km grid to serve as the assessment unit in the space. Table 1 introduces the datasets used in this research.

Wildfire risk evaluation method

The United Nations office for Disaster Risk Reduction (UN/ISDR 2002) defines the disaster risk as a negative outcome or loss possibility caused by the interaction between hazard and social ecological vulnerability, namely, Risk = Hazard $\times$ Vulnerability (Yates and Paquette 2011; Dilley et al.

Table 1. Data sources.

Data name	Contents	Source
Tweets	Wildfire data of USA from 29 July to 29 August 2015	http://vision.sdsu.edu/hdma/smart/
Administrative division	State boundary of USA in 2014	USA Census Bureau (https://www.census.gov/geo/maps-data/data/cbf/cbf_counties.html, accessed 29 December 2015)
Population	Population estimate of USA in 2014	USA Census Bureau, Population Division (https://www.census.gov/popest/data/datasets.html, accessed 29 December 2015)
NDVI	MODIS NDVI 2012–2014 (1 km × 1 km)	MODIS Web (http://modis.gsfc.nasa.gov/data/dataprod/mod13.php, accessed 23 January 2015)

2005; Hardy 2005; Chen, Blong, and Jacobson 2003). In this article, we follow the model proposed by UN/ISDR to build the following wildfire risk model:

$$Ri = Hi \times Vi \tag{1}$$

where, R_i refers to the wildfire risk index within the grid i, H_i refers to the result after wild fire hazard-inducing risk index is normalized within the grid i, and V_i refers to the social–ecological vulnerability index within the grid i.

Natural breaks method (Jenks 1967) is applied to classify R_i into five levels: extremely slight wildfire risk, slight wildfire risk, moderate wildfire risk, severe wildfire risk, and extremely severe wildfire risk. Relative risk is widely used when data are inadequate for understanding the absolute level of risk (Dilley et al. 2005; Zhou et al. 2015). The division of wildfire risk into subsets gives the possibility of obtaining relatively homogeneous classes in terms of the level of wildfire risk. Furthermore, Kernel Density Estimation (Silverman 1986; Sheather 2004) is conducted according to R_i at the grid scale to reveal the spatial pattern of wildfire risk.

Wildfire hazard evaluation

Hazard refers to the source of danger causing the disaster (UN/ISDR 2002; Chen, Blong, and Jacobson 2003). In terms of wildfire, there are two kinds of hazard evaluation methods. One is to estimate fire hazard by simulating wildfire occurrence frequency (Peters et al. 2013; Chuvieco et al. 2014). However, this method proves to be data-intensive and time-consuming (Wang et al. 2016; Wang, Ye, and Tsou 2016; Lein and Stump 2009). Another one considers the fire occurrence possibility as the standard for wildfire hazard (Scott, Thompson, and Calkin 2013). Some studies directly apply the regional wildfire occurrence times in a specific period to indicate wildfire hazard (Jordaan, Jordaan, and Procter 2011). The the geographical location plays an important role as this may allow for assessing the density and physical boundaries of the disaster (Bahir and Peled 2013). Considering that the tweets in every grid reflect the wildfire occurrence possibility, scale, and intensity, we adopt the quantity of wildfire as the index for wildfire hazard evaluation.

However, wildfire hazard is not just a kind of natural hazard. In other words, human factors will inadvertently make a significant contribution to wildfire occurrence. According to Syphard et al. (2007), people can affect the wildfire frequency and spatial distribution rules. Moreover, wildfires accident is more likely to be found and reported in regions with large population density. Therefore, many records are published on Twitter due to the larger population and easier information retrieval and exchange regardless of the physical distance from the site where a wildfire occurs. According to Guan and Chen (2014), the tweeting activities are remarkably positively related to the population density. Therefore, we apply the standardized wildfire hazard index to reflect the wildfire danger, by dividing the quantity of wildfire records in every grid by the corresponding population number. The formula about the standardized wildfire hazard index (*Hsd*) is shown as follows

$$Hi = \frac{Tri - Trmin}{Trmax - Trmin}(i = 1, 2, \ldots n) \tag{2}$$

where, *Hi* refers to the wildfire hazard index in the grid *i*; *Tr*max refers to the maximum value after record quantity on Twitter is revised in the grid *i*; and *Tr*min refers to the minimum value after records quantity on Twitter is revised in the grid *i*;

$$Tri = \frac{Ti}{Pi}(i = 1, 2, \ldots n) \tag{3}$$

where *Tri* refers to the result after record quantity on Twitter is revised in the grid *i*; *Ti* refers to the record quantity on Twitter in the grid *i*; *Pi* refers to the population number in the grid *i*.

Social–ecological vulnerability

As one of the core variables in the risk assessment, the social–ecological vulnerability mainly refers to the evaluation for the social and economic factors and potential losses of ecological assets (Chen, Blong, and Jacobson 2003; Zhou et al. 2014). Syphard et al. (2007) point out that biological variable is the most important index in the wildfire risk assessment. Other vulnerable factors, such as terrain, house, population, and infrastructure, are also included (Chen, Blong, and Jacobson 2003; Xu et al. 2005). We argue that, compared with property factors, population and vegetation are the main life entities. To evaluate the vulnerability of population and ecological system, we select population density and vegetation index as the main indicators. Therein, the population density can reflect both population concentration and exposure levels in a certain region, and the vegetation index will be used for inspecting the vegetation growth state and coverage degree.

In this article, we apply the NDVI and the normalized population density to serve as the evaluation index for social and ecological vulnerability. The formula is as follows:

$$Vi = \frac{1}{2}(Psdi + \mathrm{NDVI}sdi)(i = 1, 2, ..n) \tag{4}$$

where *Vi* refers to the social–ecological system vulnerability in the grid *i* and *Psdi* refers to the result when the population quantity is normalized in the grid *i*. *NDVIsdi* refers to the result when NDVI is normalized in the grid *i*. Considering that weight of index has little influence on the disaster risk, these two indexes will be combined via equal weights; as a result, the final calculation formula is obtained (Cutter, Boruff, and Shirley 2003; Shook 1997; Johnson et al. 2012; Reid et al. 2009).

The formula of the index normalization is shown as follows:

$$Xsdi = \frac{Xi - Xmin}{Xmax - Xmin}(i = 1, 2, ..n) \tag{5}$$

where *Xsdi* refers to the result when certain index is normalized in the grid *i*; *Xi* refers to certain index in the grid *i*; *X*min refers to the minimum value of certain index in the grid *i*; *X*max refers to the maximum value of certain index in the grid *i*; and *Psdi* and *NDVIsdi* refer to the results when the population number and NDVI are normalized.

In summary, the research framework as specified in the article is shown in Figure 1.

Results

Wildfire hazard

The distribution of wildfire hazard is shown in Figure 2. The wildfire hazard level is divided into extremely severe, severe, moderate, slight, and extremely slight, of which their area ratios account for 2.38%, 4.75%, 7.77%, 18.44%, and 66.66% respectively. It shows that

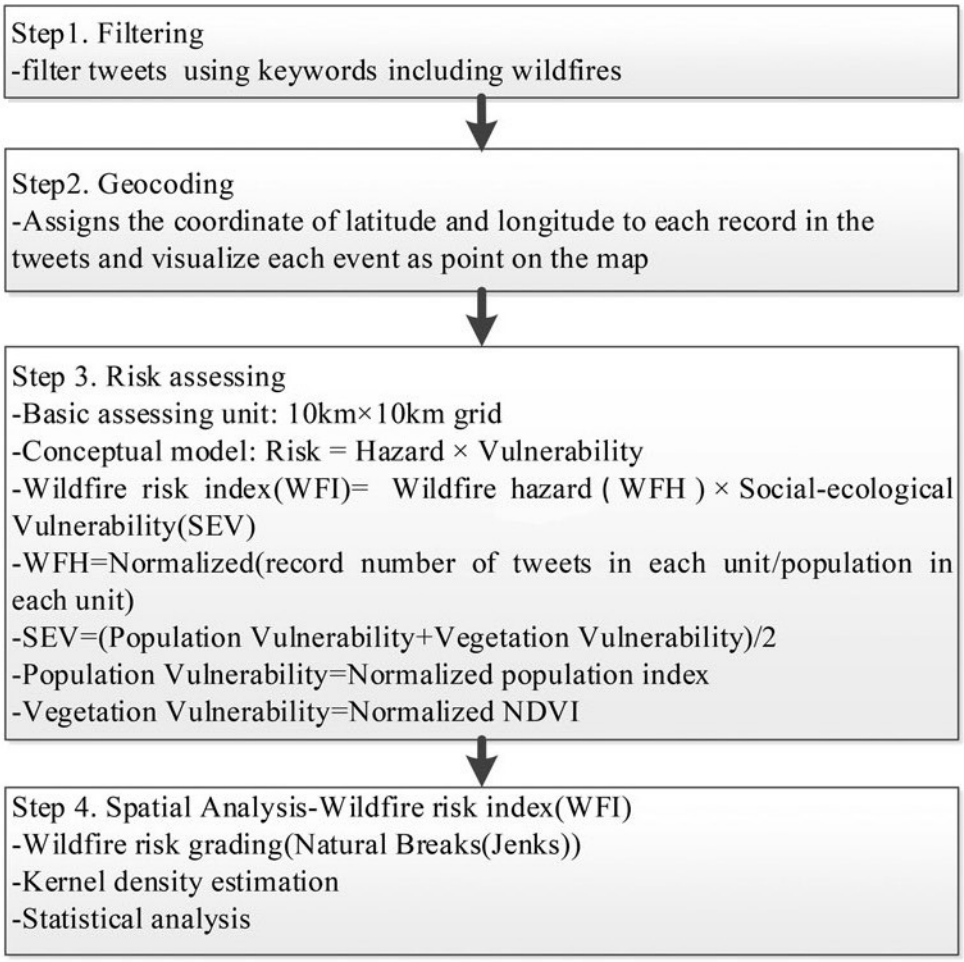

Figure 1. Research framework.

there is an overall relatively low wildfire hazard level. However, the regions of wildfire hazard over moderate level hold 14.90% of total areas, and spatial distribution are highly concentrated.

The regions facing with moderate wildfire hazard level and above are mainly located in the states such as Washington, Montana, Idaho, Oregon, Nebraska, Texas, Arizona, and California. On the whole, the regions with the highest hazard level are relatively concentrated in the West. Our results highly coincide with prior research and observations (Westerling et al. 2006; Westerling 2016; NIFC 2017). This proves that the wildfire disaster records based on twitter can effectively reveal the spatial patterns of wildfire hazard.

But our results, to a certain extent, are different from some other studies. For example, Fitzmorris (2010) finds that Southern California in particular is at high risk of wildfires, but our results show that the wildfire hazard level in Southern California is not as expected. These differences might be due to the fact that the wildfire records of tweets used in our study is very limited (only one month) to reveal the distribution pattern of wildfire. But our results indicate that tweets have great potential for revealing the spatial–temporal patterns of wildfire hazard. So if more tweets are collected, we can reveal more robust spatial distribution of wildfires in the United States.

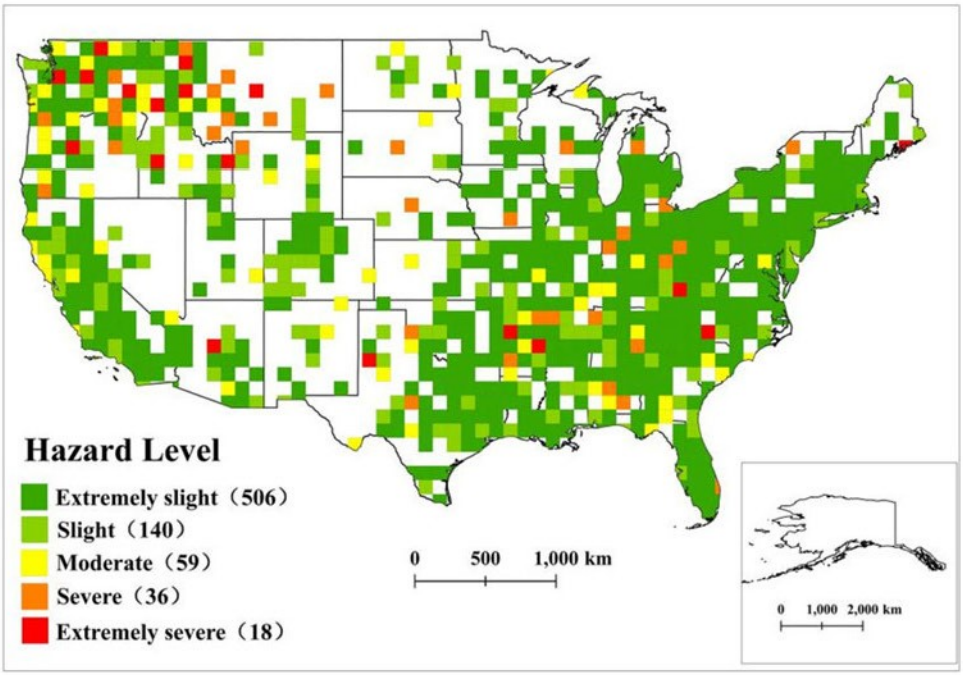

Figure 2. Spatial distribution of wildfire hazard.

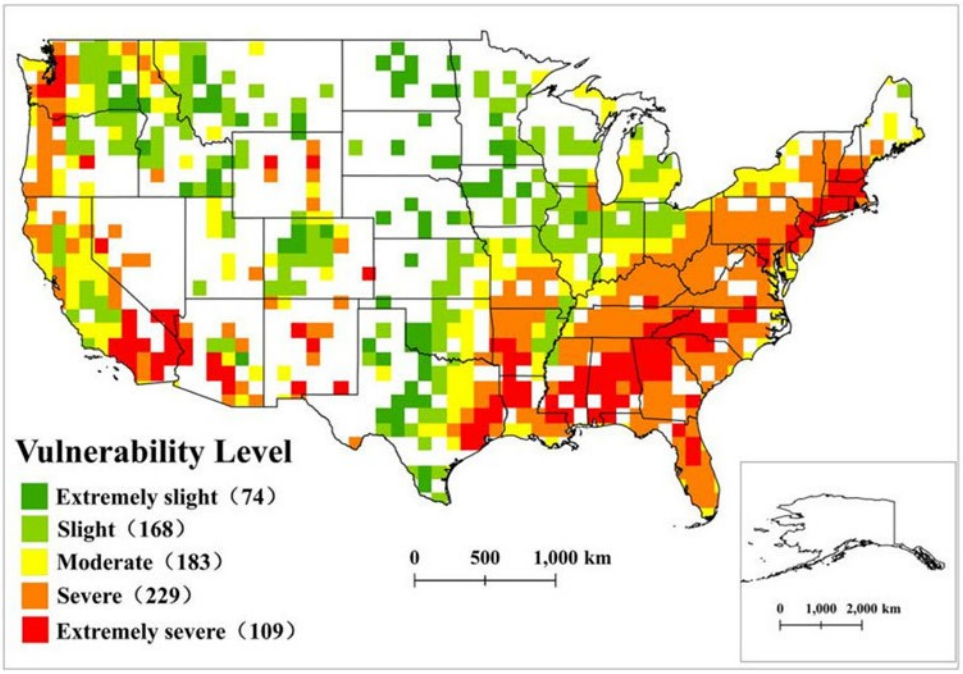

Figure 3. Social–ecological vulnerability distribution of wildfire disaster.

Social–ecological vulnerability to wildfire

The spatial distribution of social–ecological vulnerability to wildfire is shown in Figure 3. Therein, areas with extremely severe, severe, and moderate social–ecological vulnerability levels account for 14.3%, 30.1%, and 24% respectively. Areas with slight and extremely slight

social–ecological vulnerability levels account for 31.7% totally. Indubitably, the social–ecological vulnerability level is relatively high. The most vulnerable areas are mainly distributed in the East, the Western Coast, and the South.

Since only population density and NDVI are applied to serve as the vulnerability indicators, the spatial pattern of social–ecological vulnerability level reflects the potential exposure of population and vegetation to wildfire. The wildfires mainly distribute in the West and rare occur in the East. However, because of the high social–ecological vulnerability in the eastern region, if a wildfire occurs, it may lead to a great loss of population, property, and ecological assets. For example, NASA (2017) has reported several wildfires in the southeastern and eastern United States in 2017. The West Mims Fire on Florida/Georgia Border has been reported on 6 April and fully out on 22 June, leading to dramatic damages to property and ecological assets.

Spatial distribution of wildfire risk

The distribution of wildfire risk is shown in Figure 4. Areas with extremely severe, severe, moderate, slight, and extremely slight wildfire risk levels account for 0.66%, 2.5%, 7.5%, 15.55%, and 73.78%. It is obvious that the overall level of wildfire risks is relatively low. The regions faced with moderate risk level and above are mainly distributed in Washington, Oregon, North Carolina, and California. Therein, the extremely severe and severe wildfire risk level regions include Washington, Arizona, Texas, Arkansas, North Carolina, Kentucky, New York, and Alabama. These results are basically consistent with Thompson et al. (2011)

To better reflect the spatial distribution of wildfire risk level, kernel density analysis is conducted based on the wildfire risk index as shown in Figure 5. It shows that areas with high wildfire risk are mainly located in the Northwest and Southeast. However, a higher risk level does not necessarily lead to more disasters. The East is more humid than the West, so it not conducive to wildfires. As a result, no major wildfires have occurred in the East. While in the West, much more wildfire disasters occurred because of the dry climate. For example, Alaska,

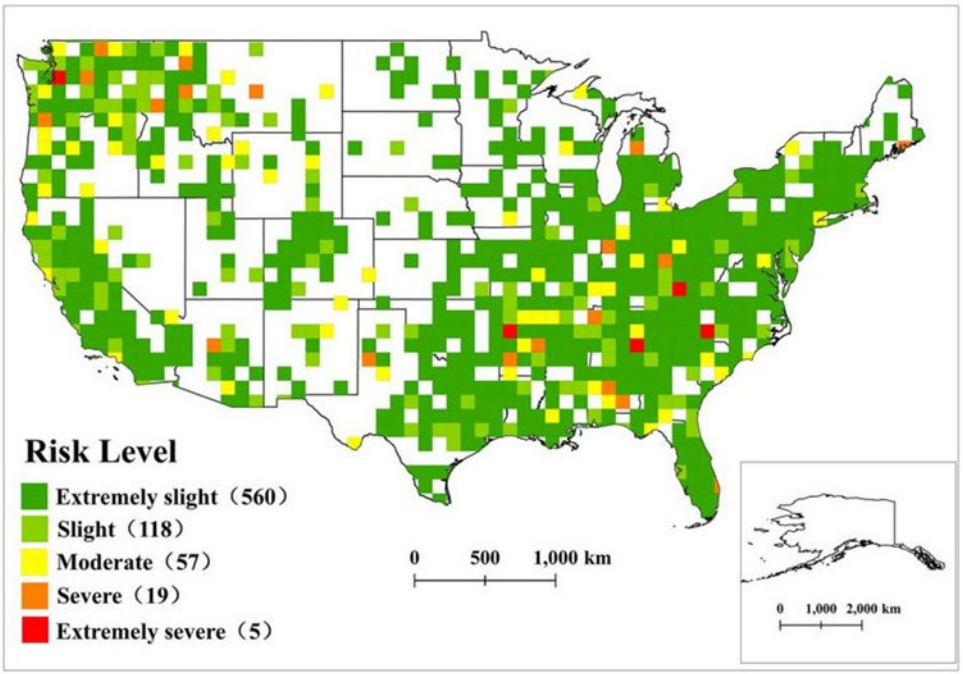

Figure 4. Spatial distribution of wildfire risk.

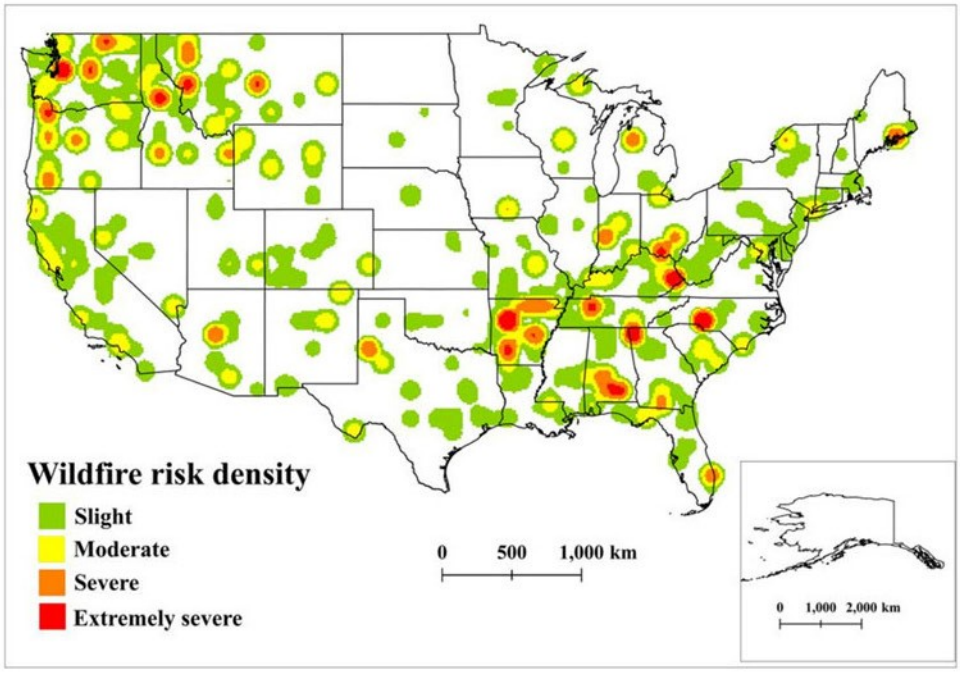

Figure 5. Kernel density distribution of wildfire risk.

California, Oregon, and Washington have suffered from the wildfires most in 2015 (The Telegraph 2015). More recently, the most destructive wildfires have been observed in California, Oregon, Washington, Idaho, and Montana in the West (The Reuters 2017). All these wildfire distribution patterns are consistent with our results except for the East.

At the same time, the overall wildfire risk level is relatively low. However, Westerling et al. (2006) discover that the large-scale forest wildfire frequency has been unceasingly increasing since 1970s compared with that in the last 10 years, and Westerling (2016) also finds that there has been an obviously increasing frequency of wildfire from 1973 to 2012, as well as the size of the burned region. Thus, wildfire has become an urgent problem to be solved. With the warming climate, the risk level of wildfire may not be lower than our results.

The reason for above differences is that the tweets data about wildfire risk are only based on a short time span, so the result cannot better reflect the spatial distribution of wildfire risk level for the long run. Despite this, the research results still have an important and realistic significance for revealing the wildfire risk distribution. Once the tweets about wildfire could be accumulated for a longer period in the future, the results will become more reliable to provide the scientific and powerful supports for the effective prevention and response of wildfire risk.

Population and vegetation at risk

The population at risk of wildfire hazard is obtained (Figure 6). Population under extremely severe and severe wildfire threat only accounts for 7.13% totally and dispersedly distributed. However, wildfire is also likely to cause serious life losses because many large and medium size cities are located in these regions, such as Fayetteville, Columbus, and Seattle. Population at the moderate wildfire risk level account for 7.77% with relatively dispersed distribution, where Washington DC is included. Population at the extremely slight and slight wildfire risk accounts for 18.44% and 66.66%, respectively. Although the wildfire risk is relatively low in these regions, the population density is higher than the other wildfire affected regions. Moreover, large

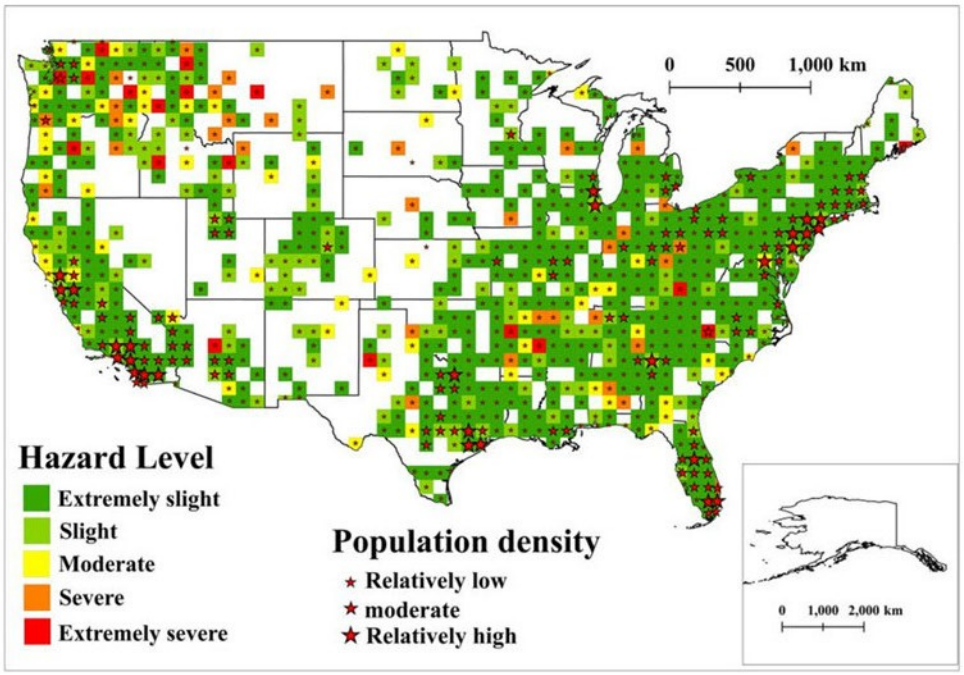

Figure 6. Population distribution influenced by wildfire hazard.

populous cities, such as Los Angeles, Santiago, Las Vegas, San Antonio, Houston, Chicago, Newhaven, and Miami, are located in these regions, which has significantly increased the possibility of potential life and property losses.

Fitzmorris (2010) mention that because of urban sprawl, every year countless homes in the greater Los Angeles area are damaged or destroyed by wildfires. More recently, it reports that 27 large fires have burned nearly 180,000 acres across the West, forceing thousands of local residents to relocate (USA TODAY 2017). In California, the La Tuna fire near Burbank has burned nearly 7200 acres, becoming the largest fire ever recorded in Los Angeles in terms of area size (The Atlantic 2017). All these wildfires pose a great threat to the safety of human beings and property. Therefore, massive and vulnerable population might result in significant increase in potential life and property losses.

We further analyze the correlation between wildfire risk level and population density. The partial correlation coefficient is 0.520, larger than the correlation coefficient (0.517). Thus, linear relationship exists between wildfire risk level and population density. The probability value of partial correlation coefficient is 0, which is smaller than the significance level (0.05 or 0.01). It shows that the wildfire risk level has significant linear relationship with population density. Syphard et al (2007) also find highly significant relationships exist between population density and fire. This suggests that the level of wildfire risk may also be determined by the vulnerability degree of demographic factors, in addition to wildfire hazard itself (UN/ISDR 2002; Cutter, Boruff, and Shirley 2003; Zhou et al. 2014; Zhou et al. 2015).

The vegetation distribution under wildfire threat is obtained through spatial overlay of the wildfire hazard layerwith vegetation coverage (Figure 7). Vegetation affacted by the extremely severe and severe wildfire hazard level is distributed in a dispersed way with a small area proportion. However, regions under the moderate wildfire hazard level are distributed in a more continuous way. Some regions have a relatively high vegetation coverage level, such as Tallahassee, Olympia, and Medford. Regions with the slight hazard level are distributed in a dispersed and continuous way where 21 grids have a higher vegetation coverage level, such as

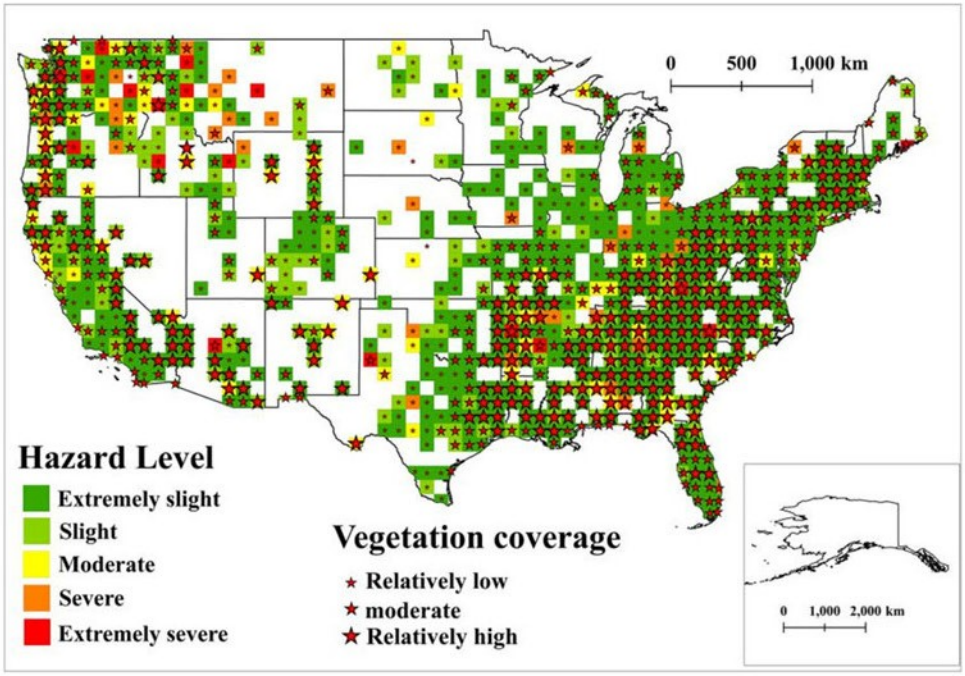

Figure 7. Vegetation distribution influenced by wildfire hazard.

Lino, Missoula, and Jacksonville. It shows that the wildfire hazard levels in the vast majority of high-vegetation-covered areas are not high, except the Northwest. Although the vegetation coverage level in the East is very high, there are few fires as a result of the plentiful rain and damp air. But in some regions of the West where the weather is hot, the air is dry, and the rainfall is small, although the vegetation coverage level is not very high, the wildfires are very frequent.

We further examine the correlation between wildfire hazard level and the vegetation coverage level, finding that the partial correlation coefficient is 0.051. It is larger than the correlation coefficient (0.047), which indicates the linear relationship between the wildfire hazard and vegetation coverage. However, the probability value of partial correlation coefficient is only 0.156, larger than the significance level (0.05 or 0.01). Given this, the wildfire hazard level has nonsignificant linear relationship with the vegetation coverage level.

Population and vegetation are not only the targets of wildfire but also play a decisive role in the occurrence and development of wildfires. According to Syphard et al. (2007), people can affect the wildfire frequency and spatial distribution rules, but such land surface features as vegetation determines the spread of wildfire. In conclusion, for the occurrence of wildfire disasters, human factors are of vital importance. Thus, how to strengthen the awareness about wildfire hazards and guide people to reduce the improper behaviors that will cause wildfire is of great importance to wildfire prevention and alleviation.

Discussion

Through applying geo-tagged social media big data to assess the risk of disasters, this article develops a conceptual framework. By taking wildfire in the United States as an example, we demonstrate the feasibility of the proposed framework and associated methods, by adopting one-month tweets data. Results of this article demonstrate that social media data have great potential in revealing the spatial patterns of disaster risk.

Two key steps are included in the wildfire hazard evaluation based on tweets. First of all, tweets associated with specific wildfires are gleaned based on keywords which are places where wildfires occurr. Through the multi-level keyword searching, high density screening is made for tweets to maximize a more complete representation of wildfire hazard distribution. Furthermore, we adopt the methods dealing with the tweets screening rules and population distribution effects used by Wang et al. (2016) and Wang, Ye, and Tsou (2016). Such methods can to some extent deal with the poor authenticity of wildfire frequency simulation (Peters et al. 2013; Chuvieco et al. 2014), as well as the data acquisition difficulty of wildfire occurrence. Our results show that the geo-tagged social media data have strong timeliness and good authenticity in evaluating disaster risks.

The vulnerability is one of the core variables in the risk assessment. Although a lot of vulnerability factors are used in the wildfire risk assessment (Chen, Blong, and Jacobson 2003, 2011; Xu et al. 2005), the safety of human and ecological assets are more crucial. Therefore, population density and NDVI are applied to evaluate the social–ecological vulnerability. However, our proposed social–ecological vulnerability index may have some limitations. As for the population, the mere use of population density may not reveal its spatial distribution accurately. For example, Guan and Chen (2014) point out that the residential areas and traffic networks are regions with large flow density. So, Jiao et al. (2015) make a meticulous treatment of the population distribution according to the spatial distribution of population activity. In addition, Alexander (2014) argue that the economic conditions, age, education level, and capacity factors will also exert important influence on the population vulnerability. As for ecological systems, we select NDVI as the index without distinguishing the influence of vegetation type and flammability on the vulnerability of ecological system. However, Jiao et al. (2015) state that the vegetation type illustrates a great difference upon the wildfire risk, show that the vegetation type is of most significance on wildfire occurrence. For example, some kinds of vegetation are more inflammable. Therefore, it is necessary to take more factors into account besides population and vegetation. For example, Cutter, Boruff, and Shirley (2003) develop a social vulnerability index at the county level for the United States using 42 variables, while Wigtil et al. (2016) select 26 socioeconomic and demographic variables to create the social vulnerability index for the wildfire risk assessment.

Our research reveals that population has larger impact on the wildfire risks than NDVI, which means that different weights should be assigned to NDVI and population. However, Dong show that the variation of factor weights has little influence on the disaster risk assessment. Many other studies also apply the equal index weights to calculate the vulnerability index based on multiple indexes (Cutter, Boruff, and Shirley 2003; Shook 1997; Johnson et al. 2012; Reid et al. 2009). Given this, we apply the calculation method of index weight in this article.

In summary, our research can be improved in the following aspects. Although the results of the wildfire risk distribution in the United States largely coincide with the other researchers', some differences still exist. This is mainly due to spatial and temporal coverage of tweets data used by different scholars. Therefore, it is very important to accumulate long-term wildfire data on Twitter to dig deeper into the spatial–temporal pattern of wildfire risks. Considering that the tweets are point data and wildfire has the characteristics of both point and sphere, it is needed to combine wildfire spread models to make more accurate evaluation for the influence range of wildfire in the future studies. From the perspective of vulnerability, it is necessary to strengthen the analysis of the formation mechanism from the wildfire hazard bearing body, such as population, vegetation, and buildings.

Conclusion

The geo-tagged social media data are increasingly being used for enhancing disaster risk assessment and assisting disaster management. After putting forward the framework applying social media data to assess wildfire risks, we discuss the wildfire hazard, social–ecological vulnerability,

and the methods about wildfire risk assessment in this article. The following findings are concluded.

The spatial–temporal pattern of wildfire risks can be assessed by our proposed framework. The wildfire risk level has significant linear relationship with population density via considering the population density influence. On the contrary, the linear relationship between wildfire risk level and NDVI is not significant. Hence, the wildfire risk level is higher with a larger population size under the same disaster-inducing hazard level.

The social–ecological vulnerability level of wildfire is relatively high, and the main regions involved are the states such as Washington, Oregon, Arizona, Texas, Arkansas, North Carolina, and Kentucky. While the overall level of wildfire risks is relatively low, the regions which are faced with moderate risk level and above are mainly distributed in Washington, Oregon, North Carolina, and California. Therein, the severe wildfire risk level and above are mainly distributed in Washington, Arizona, Texas, Arkansas, North Carolina, Kentucky, New York, and Alabama.

Our results show that social media data have a huge potential for disaster risk assessment and management. More accurate and detailed wildfire risk assessment results can be obtained with more tweets about wildfires for a longer period. This research can help promote the development of disaster science and risk prevention as long as we further explore and integrate the multi-source big data.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Key Research and Development Program (No. 2016YFA0602402), the National Natural Science Foundation of China (No. 41271515), the National Basic Research Program of China (No. 2012CB955403), and the National Science Foundation of the USA (No. 1416509, 1637242, and 1739491).

References

Albuquerque, J. P. D., B. Herfort, A. Brenning, and A. Zipf. 2015. "A Geographic Approach for Combining Social Media and Authoritative Data towards Identifying Useful Information for Disaster Management." *International Journal of Geographical Information Science* 29 (4): 667–689. doi:10.1080/13658816.2014.996567.

Alexander, D. E. 2014. "Social Media in Disaster Risk Reduction and Crisis Management." *Science and Engineering Ethics* 20 (3): 717–733. doi:10.1007/s11948-013-9502-z.

Arago, P., P. Juan, C. Diaz-Avalos, and P. Salvador. 2016. "Spatial Point Process Modeling Applied to the Assessment of Risk Factors Associated with Forest Wildfires Incidence in Castellon, Spain." *European Journal of Forest Research* 135 (3): 451–464. doi:10.1007/s10342-016-0945-z.

Bahir, E., and A. Peled. 2013. "Identifying and Tracking Major Events Using Geo-Social Networks." *Social Science Computer Review* 31 (4): 458–470. doi:10.1177/0894439313483689.

Burger, J., M. Gochfeld, C. Jeitner, T. Pittfield, and M. Donio. 2013. "Trusted Information Sources Used during and after Superstorm Sandy: TV and Radio Were Used More Often than Social Media." *Journal of Toxicology and Environmental Health. Part A* 76 (20): 1138–1150.

Calkin, D. E., A. Ager, J. Gilbertson-Day, J. H. Scott, M. A. Finney, C. Schrader-Patton, J. Strittholt, and J. Kaiden. 2010. "Wildfire Risk and Hazard: Procedures for the First Approximation." USDA Forest Service – General Technical Report RMRS-GTR 235 (235): 1–62.

Chen, K., R. Blong, and C. Jacobson. 2003. "Towards an Integrated Approach to Natural Hazards Risk Assessment Using GIS: with Reference to Bushfires." *Environmental Management* 31 (4): 546–560. doi:10.1007/s00267-002-2747-y.

Chuvieco, E., I. Aguado, S. Jurdao, M. L. Pettinari, M. Yebra, J. Salas, S. Hantson., et al. 2014. "Integrating Geospatial Information into Fire Risk Assessment." *International Journal of Wildland Fire* 23 (5): 606–619. doi:10.1071/WF12052.

Cohen, E., P. Hughes, and P. B. White. 2007. "Media and Bushfires: A Community Perspective of the Media During the Grampians Fires 2006." *Environmental Hazards* 7 (2): 88–96. doi:10.1016/j.envhaz.2007.07.007.

Cutter, S. L., B. J. Boruff, and W. L. Shirley. 2003. "Social Vulnerability to Environmental Hazards." *Social Science Quarterly* 84 (2): 242–261. doi:10.1111/1540-6237.8402002.

Dilley, M., R. S. Chen, U. Deichmann, A. L. Lerner-Lam, M. Arnold, J. Agwe, P. Buys, et al. 2005. "Natural Disaster Hotspots: a Global Risk Analysis." Uwe Deichmann 20 (4):1–145.

Fitzmorris, P. K. 2010. "Wildfire Management in Los Angeles' Wildland-Urban Interface: Identifying Better Strategies for Reconciling Wildfires with LA's Communities."

Gao, H., G. Barbier, and R. Goolsby. 2011. "Harnessing the Crowdsourcing Power of Social Media for Disaster Relief." *IEEE Intelligent Systems* 26 (3): 10–14. doi:10.1109/MIS.2011.52.

Ghosh, D. D., and R. Guha. 2013. "What Are we 'twittering' about Obesity? Mapping Twitters with Topic Modeling and Geographic Information System." *Cartography and Geographic Information Science* 40 (2): 90–102. doi:10.1080/15230406.2013.776210.

Goodchild, M. F., and L. N. Li. 2012. "Assuring the Quality of Volunteered Geographic Information." *Spatial Statistics* 1: 110–120. doi:10.1016/j.spasta.2012.03.002.

Guan, X., and C. Chen. 2014. "Using Social Media Data to Understand and Assess Disasters." *Natural Hazards* 74 (2): 837–850. doi:10.1007/s11069-014-1217-1.

Hardy, C. C. 2005. "Wildland Fire Hazard and Risk: Problems, definitions, and Context." *Forest Ecology and Management* 211 (1–2): 73–82. doi:10.1016/j.foreco.2005.01.029.

Hong, X., and X. Ye. 2017. "Exploring the Influence of Land Cover on Weight Loss Awareness." *GeoJournal* 83 (5): 935–947. doi:10.1007/s10708-017-9806-7.

Java, A., X. Song, T. Finin, and B. Tseng. 2007. "Why We Twitter: understanding Microblogging Usage and Communities." *Webkdd and Sna-Kdd 2007 Workshop on Web Mining and Social Network Analysis ACM* 43: 56–65.

Jenks, G. F. 1967. "The Data Model Concept in Statistical Mapping." *International Yearbook of Cartography* 7 (1): 186–190.

Jiao, L. L., Y. Chang, D. Shen, Y. M. Hu, C. L. Li, and J. Ma. 2015. "Using Boosted Regression Trees to Analyze the Factors Affecting the Spatial Distribution Pattern of Wildfire in China." *Chinese Journal of Ecology* 34 (8): 2288–2296. (in Chinese with English abstract)

Johnson, D. P., A. Stanforth, V. Lulla, and G. Luber. 2012. "Developing an Applied Extreme Heat Vulnerability Index Utilizing Socioeconomic and Environmental Data." *Applied Geography* 35 (1–2): 23–31. doi:10.1016/j.apgeog.2012.04.006.

Jordaan, A. D., A. J. Jordaan, and M. Procter. 2011. "Wildfire Risk Assessment for the Northern Cape, South Africa." Accessed 22 January 2017. http://disaster.co.za/pics/PrADJordaan_etalDMISA2011WildfireRiskAssessment.pdf

Keane, R. E., S. A. Drury, E. C. Karau, P. F. Hessburg, and K. M. Reynolds. 2010. "A Method for Mapping Fire Hazard and Risk across Multiple Scales and Its Application in Fire Management." *Ecological Modelling* 221 (1): 2–18. doi:10.1016/j.ecolmodel.2008.10.022.

Kietzmann, J. H., K. Hermkens, I. P. Mccarthy, and B. S. Silvestre. 2011. "Social Media? Get Serious! Understanding the Functional Building Blocks of Social Media." *Business Horizons* 54 (3): 241–51. doi:10.1016/j.bushor.2011.01.005.

Lein, J. K., and N. I. Stump. 2009. "Assessing Wildfire Potential within the Wildland–urban Interface: A Southeastern Ohio Example." *Applied Geography* 29 (1): 21–34. doi:10.1016/j.apgeog.2008.06.002.

Li, Q., W. Wei, N. Xiang, D. Feng, X. Ye, and Y. Jiang. 2017. "Social Media Research, Human Behavior, and Sustainable Society." *Sustainability* 9 (3): 384. doi:10.3390/su9030384.

Miles, B., and S. Morse. 2007. "The Role of News Media in Natural Disaster Risk and Recovery." *Ecological Economics* 63 (2–3): 365–373. doi:10.1016/j.ecolecon.2006.08.007.

NASA. Accessed 22 September 2017. https://www.nasa.gov/image-feature/goddard/2017/fires-dot-the-eastern-united-states.

Nicholas, D., and I. Rowlands. 2011. "Social Media Use in the Research Workflow." *Learned Publishing* 24 (3): 183–195. doi:10.1087/20110306.

NIFC (the National Interagency Fire Center). Accessed 20 September 2017. https://www.nifc.gov/fireInfo/nfn.htm

Parisien, M. A., and M. A. Moritz. 2009. "Environmental Controls on the Distribution of Wildfire at Multiple Spatial Scales." *Ecological Monographs* 79 (1): 127–154. doi:10.1890/07-1289.1.

Peters, M. P., L. R. Iverson, S. N. Matthews, and A. M. Prasad. 2013. "Wildfire Hazard Mapping: exploring Site Conditions in Eastern US Wildland–Urban Interfaces." *International Journal of Wildland Fire* 22 (5): 567–578.

Reid, C. E., M. S. O'Neill, C. J. Gronlund, S. J. Brines, A. V. Diez-Roux, D. G. Brown, and J. D. Schwartz. 2009. "Mapping Community Determinants of Heat Vulnerability." *Environmental Health Perspectives* 117 (11): 1730–1736. doi:10.1289/ehp.0900683.

Sakaki, T., M. Okazaki, and Y. Matsuo. 2012. "Tweet Analysis for Real-time Event Detection and Earthquake Reporting System Development." *IEEE Transactions on Knowledge and Data Engineering* 25 (4): 919–993. doi:10.1109/TKDE.2012.29.

Scott, J. H., M. P. Thompson, and D. E. Calkin. 2013. A wildfire risk assessment framework for land and resource management. USDA Forest Service – General Technical Report RMRS-GTR (315)

Sheather, S. J. 2004. "Density Estimation." *Statistical Science* 19 (4): 588–597. doi:10.1214/088342304000000297.

Shook, G. 1997. "An Assessment of Disaster Risk and Its Management in Thailand." *Disasters* 21 (1): 77–88.
Silverman, B. W. 1986. *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall, 34–72.
Slavkovikj, V., S. Verstockt, S. V. Hoecke, and R. V. D. Walle. 2014. "Review of Wildfire Detection Using Social Media." *Fire Safety Journal* 68 (8): 109–118. doi:10.1016/j.firesaf.2014.05.021.
Smith, L., Q. Liang, P. James, and W. Lin. 2015. "Assessing the Utility of Social Media as a Data Source for Flood Risk Management Using a Real-time Modelling Framework." *Journal of Flood Risk Management* 10 (3): 370–380. doi:10.1111/jfr3.12154.
Sutton, J. N., L. Palen, and I. Shklovski. 2008. Emergency Uses of Social Media in the 2007 Southern California Wildfires. University of Colorado, 3: 668–669
Syphard, A. D., V. C. Radeloff, J. E. Keeley, T. J. Hawbaker, M. K. Clayton, S. I. Stewart, and R. B. Hammer. 2007. "Human Influence on California Fire Regimes." *Ecological Applications: A Publication of the Ecological Society of America* 17 (5): 1388–1402.
The Atlantic. Accessed 20 September 2017. https://www.theatlantic.com/photo/2017/09/wildfires-rage-across-the-american-west/538977/
The Reuters. Accessed 22 September 2017 https://www.reuters.com/article/us-usa-wildfires/u-s-wildfire-preparedness-raised-to-highest-level-idUSKBN1AR07F
The Telegraph. Accessed 22 September 2017. http://www.telegraph.co.uk/news/worldnews/northamerica/usa/11932329/2015-becomes-worst-US-wildfire-year-on-record.html
Thompson, M. P., D. E. Calkin, M. A. Finney, A. A. Ager, and J. W. Gilbertson-Day. 2011. "Integrated National-scale Assessment of Wildfire Risk to Human and Ecological Values." *Stochastic Environmental Research and Risk Assessment* 25 (6): 761–780. doi:10.1007/s00477-011-0461-0.
Tsou, M. H. 2015. "Research Challenges and Opportunities in Mapping Social Media and Big Data." *Cartography and Geographic Information Science* 42 (Supp1): 70–74. doi:10.1080/15230406.2015.1059251.
UN/ISDR 2002. "Living with Risk: a Global Review of Disaster Reduction Initiatives." *Bmc Public Health* 7 (4): 336–342.
USA TODAY. Accessed 20 September 2017. https://www.usatoday.com/story/news/nation/2017/06/29/thousands-flee-wildfires-roar-through-west/438518001/
Wang, Y., W. Jiang, S. Liu, X. Ye, and T. Wang. 2016. "Evaluating Trade Areas Using Social Media Data with a Calibrated Huff Model." *ISPRS International Journal of Geo-Information* 5 (7): 112. doi:10.3390/ijgi5070112.
Wang, Z., and X. Ye. 2017. "Social Media Analytics for Natural Disaster Management." *International Journal of Geographical Information Science* 32 (1): 49–72. doi:10.1080/13658816.2017.1367003.
Wang, Z., X. Ye, and M. Tsou. 2016. "Spatial, temporal, and Content Analysis of Twitter for Wildfire Hazards." *Natural Hazards* 83 (1): 523–540. doi:10.1007/s11069-016-2329-6.
Wang, Y. D., X. K. Fu, W. Jiang, T. Wang, M. H. Tsou, and X. Ye. 2017. "Inferring Urban Air Quality Based on Social Media." *Computers, Environment and Urban Systems* 66: 110–116. doi:10.1016/j.compenvurbsys.2017.07.002.
Westerling, A. L. 2016. "Increasing Western US Forest Wildfire Activity: sensitivity to Changes in the Timing of Spring." *Philosophical Transactions of the Royal Society B: Biological Sciences* 371 (1696): 20150178. doi:10.1098/rstb.2015.0178.
Westerling, A. L., H. G. Hidalgo, D. R. Cayan, and T. W. Swetnam. 2006. "Warming and Earlier Spring Increase Western US Forest Wildfire Activity." *Science* 313 (5789): 940–943. doi:10.1126/science.1128834.
Wigtil, G., R. B. Hammer, J. D. Kline, M. H. Mockrin, S. I. Stewart, D. Roper, and V. C. Radeloff. 2016. "Places Where Wildfire Potential and Social Vulnerability Coincide in the Coterminous United States." *International Journal of Wildland Fire* 25 (8): 896–908. doi:10.1071/WF15109.
Wu, C., X. Ye, R. F. Wan, Y, P. Ning, and Q. Du. 2016. "Spatial and Social Media Data Analytics of Housing Prices in Shenzhen, China." *PLoS One* 11 (10): e0164553. doi:10.1371/journal.pone.0164553.
Xiao, Y., Q. Huang, and K. Wu. 2015. "Understanding Social Media Data for Disaster Management." *Natural Hazards* 79 (3): 1663–1679. doi:10.1007/s11069-015-1918-0.
Xu, D., L. M. Dai, G. F. Shao, L. Tang, and H. Wang. 2005. "Forest Fire Risk Zone Mapping from Satellite Images and GIS for Baihe Forestry Bureau, Jilin, China." *Journal of Forestry Research* 16 (3): 169–174.
Yates, D., and S. Paquette. 2011. "Emergency Knowledge Management and Social Media Technologies: A Case Study of the 2010 Haitian Earthquake." *International Journal of Information Management* 31 (1): 6–13. doi:10.1016/j.ijinfomgt.2010.10.001.
Ye, X., S. Li, X. Yang, and C. Qin. 2016. "Use of Social Media for Detection and Analysis of Infectious Disease in China." *ISPRS International Journal of Geo-Information* 5 (9): 156. doi:10.3390/ijgi5090156.
Zhao, W. X., J. Jiang, J. Weng, J. He, E. P. Lim, H. Yan, and X. Li. 2011. "Comparing Twitter and Traditional Media Using Topic Models." *Lecture Notes in Computer Science* 6611 (2011): 338–349.
Zhou, L., Y. J. Yue, J. Li, M. Qiu, and Y. R. Shang. 2014. "Review of Research on Vulnerability Assessment of Hail Hazard Bearing Body." *Chinese Journal of Agrometeorology* 35 (3): 330–337. (in Chinese with English abstract)
Zhou, Y., Y. Liu, W. Wu, and N. Li. 2015. "Integrated Risk Assessment of Multi-hazards in China." *Natural Hazards* 78 (1): 257–280. doi:10.1007/s11069-015-1713-y.