INFORMATION, COMMUNICATION & SOCIETY, 2016
http://dx.doi.org/10.1080/1369118X.2016.1218528

Routledge
Taylor & Francis Group

Social media approaches to modeling wildfire smoke dispersion: spatiotemporal and social scientific investigations

Sonya Sachdeva[a], Sarah McCaffrey[a] and Dexter Locke[b]

[a]US Forest Service, Evanston, IL, USA; [b]Graduate School of Geography, Clark University, Worcester, MA, USA

ABSTRACT

Wildfires have significant effects on human populations, economically, environmentally, and in terms of their general well-being. Smoke pollution, in particular, from either prescribed burns or uncontrolled wildfires, can have significant health impacts. Some estimates suggest that smoke dispersion from fire events may affect the health of one in three residents in the United States, leading to an increased incidence of respiratory illnesses such as asthma and pulmonary disease. Scarcity in the measurements of particulate matter responsible for these public health issues makes addressing the problem of smoke dispersion challenging, especially when fires occur in remote regions. Crowdsourced data have become an essential component in addressing other societal problems (e.g., disaster relief, traffic congestion) but its utility in monitoring air quality impacts of wildfire events is unexplored. In this study, we assessed if user-generated social media content can be used as a complementary source of data in measuring particulate pollution from wildfire smoke. We found that the frequency of daily tweets within a 40,000 km^2 area was a significant predictor of PM2.5 levels, beyond daily and geographic variation. These results suggest that social media can be a valuable tool for the measurement of air quality impacts of wildfire events, particularly in the absence of data from physical monitoring stations. Also, an analysis of the semantic content in people's tweets provided insight into the socio-psychological dimensions of fire and smoke and their impact on people residing in, working in, or otherwise engaging with affected areas.

ARTICLE HISTORY
Received 1 April 2016
Accepted 26 July 2016

KEYWORDS
Crowdsourcing; air quality; wildfire; smoke; risk perception

Wildfires pose a challenging conundrum for managers. Fires are an essential component of many forest and prairie ecosystems, yet they can also have adverse impacts on human well-being by affecting homes, infrastructure, and air quality. As populations move into fire prone areas and global climate change elongates the wildfire season (Calkin, Thompson, & Finney, 2015), mitigating the negative impacts to humans is increasingly important. Smoke from wildfires is an especially critical public health concern: studies suggest that a third of households have a member with health concerns that can be exacerbated by wildland fire smoke (McCaffrey & Olsen, 2012). It is therefore of increasing interest to fire

CONTACT Sonya Sachdeva sonyasachdeva@fs.fed.us US Forest Service, 1033 University Pl., Ste. 360, Evanston, IL 60201, USA

Supplemental data for this article can be accessed at http://dx.doi.org/10.1080/1369118X.2016.1218528.

managers and information officers to have better models for identifying the extent and range of impact of smoke dispersion from wildfire events. The goals of this paper are two-fold: (1) to assess whether information gleaned from social media sites, such as Twitter, has the potential to fill in estimates of air quality where physical monitoring stations may not and (2) to understand the most important (and social media-relevant) issues in people's minds as they experience wildfire events, including the prevalence of smoke concerns.

Background

Wildfire smoke

Smoke from wildland fires, prescribed or otherwise, can have a substantial impact on air quality through particle emissions. Particulate matter, composed of a mixture of microscopic solids and liquid in the air, is one of the most dangerous types of pollution for human health. While particulate matter larger than 10 micrometers (PM10) in diameter can be filtered through the throat and nose, fine particles, especially those smaller than 2.5 micrometers (PM2.5), can get deeply embedded in the lungs and may even pass through the bloodstream to other organs (Kinney, 2008; US EPA, 2015). Long-term exposure, over the course of years, to high PM2.5 levels has been associated with severe respiratory illnesses and premature death, while short-term exposure can exacerbate existing lung conditions such as asthma and bronchitis. Much research has found that emergency room visits due to acute asthmatic or other respiratory ailments tend to increase when high levels of PM2.5 concentration are found in the air (Bowman & Johnston, 2005; Dominici et al., 2006; Mott et al., 2002; Ram, Zhang, Williams, & Pengetnze, 2015; Schwartz, Slater, Larson, Pierson, & Koenig, 1993).

In California, approximately a third of fine particulate pollution can be attributed to wildfires (Rittmaster, Adamowicz, Amiro, & Pelletier, 2006). Globally, researchers have attributed over 300,000 deaths to wildfires every year (Johnston et al., 2012). The USDA Forest Service and other federal agencies are increasingly underscoring the need to consider the economic, health, and social implications of wildfire smoke (Kochi, Donovan, Champ, & Loomis, 2010; Richardson, Champ, & Loomis, 2012). However, the concern over PM2.5 from wildfires is even more pressing now as climate change worsens drought conditions and elongates the wildfire season across the world (Liu, Stanturf, & Goodrick, 2010; Morgan et al., 2010). However, current air quality estimates and forecasts are limited by the number of physical stations in a geographic region. Currently, there are 4000 Environmental Protection Agency's (EPA) monitoring stations across the continental United States with the majority of stations concentrated on the eastern seaboard and close to metropolitan centers. This leaves substantial data gaps, particularly in remote or rural areas, where no estimates are available. Social media may be a complementary data source: by effectively using humans as air quality monitors, there is an opportunity to close some of these existing gaps.

Social media as a social tool

User-generated content (UGC) and volunteered geographic information (VGI) from web sources, whether through microblogging sites like Twitter or Weibo, or through social networking sites like Facebook, are increasingly powerful tools in the wake of natural disasters and extreme weather events (Goodchild, 2007; Kent & Capello, 2013; Shelton, Poorthuis, Graham, & Zook, 2014). For instance, both the 2010 earthquake in Haiti and the 2011 earthquake and tsunami in Japan instigated the dissemination of massive amounts of crowdsourced information. Reports coming in directly from the public helped relief organizations locate specific requests for help and provide resources more effectively (Barrington et al., 2012; Cassa, Chunara, Mandl, & Brownstein, 2013; Gao, Barbier, & Goolsby, 2011; Sutton, League, Sellnow, & Sellnow, 2015; Vieweg, Hughes, Starbird, & Palen, 2010). Social media networks also provide a platform where people can share emotional experiences to help cope with crisis events such as earthquakes or hurricanes (Genes, Chary, & Chason, 2014; Veer, Ozanne, & Hall, 2015). These websites can provide rich insight into the psychological processes of coping and nearly real-time observations of people's mental states as they experience these events.

More recently, UGC and VGI data sources have been generating potentially actionable knowledge. Crowdsourced data from social media sites have been shown to be most useful in addressing events in real time. For instance, many natural disasters occur quite rapidly and require a fast response. In these situations, UGC has become a powerful predictive and information dissemination tool. For example, one study shows that Twitter-based earthquake detection systems can indicate earthquake events within 30 seconds, compared to several minutes from the US Geological Survey's National Earthquake Information Center (Earle et al., 2010). Similarly, VGI from status posts about smoke and haze on the Chinese microblogging site, Sina Weibo, has been shown to accurately predict air quality levels in China (Jiang, Wang, Tsou, & Fu, 2015; Mei, Li, Fan, Zhu, & Dyer, 2014). The available UGC can be an especially useful tool in the absence of more reliable sources. Combining search engine results related to asthma with air quality data from official sources and data on emergency room visits, Ram et al. (2015) found a reliable pattern suggesting that an increase in respiratory illness related searches on bad air quality days was followed by an uptick in respiratory condition related emergency room visits. The combination of multiple publicly available (or generated) databases can help to alleviate noise from anomalous events which may affect any singular source and cross-validate results by comparing across different data sources.

Modeling smoke dispersion

There are a number of systems that currently provide data to inform the public about potential smoke exposure during wildfire events. The BlueSky Smoke Modeling Framework (BlueSky) is perhaps one of the most comprehensive and widely used in the United States to predict smoke emissions and changes in air quality from prescribed and uncontrolled burns (Larkin et al., 2009; Strand et al., 2012). BlueSky integrates existing datasets and models (e.g., fuel loadings, fire consumption, plume rise) into a coherent structure and requires only the input of meteorological information and fire size/location information. It is also readily available to the public via the web ('BlueSky Modeling Framework | AirFire, n.d.,' accessed January 2016). The National Oceanic and Atmospheric Administration has also developed a smoke forecasting system that models PM2.5 concentrations in the United States from large wildfires and agricultural burning. This system relies on remote sensing data to detect smoke using imagery and then estimates PM2.5 emission using BlueSky's framework (Rolph et al., 2009). BlueSky also gets input from remote sensing, but it combines satellite detections with ground-based reporting to help refine the fire input information. Both models have two distinct advantages over the EPA's air quality information. They can both predict air quality impacts, unlike the EPA's monitors, and fill in where one of the 4000 stations may not be present.

Current study

The first objective of our study was to assess whether data obtained from social media sites, such as Twitter, could be used to ascertain air quality impacts from wildfire events. This type of data, effectively utilizing local people as on-the-ground monitors, may be particularly useful for fires which occur in more remote or rural areas where mechanical sensors may not be present. Additionally, the availability and relatively low-cost nature of this type of data makes it appealing for emergency responders who may need to intervene rapidly while maximizing limited budgets.

The second objective of our study was to assess whether the semantic content of people's posts on these platforms can provide insight into the socio-psychological dimension of fire and smoke, its relative importance to them compared to other tweeted topics, and how this may vary based on proximity to the fire.

King fire

Our analysis focused on the air quality impacts of the 2014 King fire in northern California. This was the second largest wildfire in the state in 2014 and engulfed over 97,000 acres of land near Pollock Pines in El Dorado County (Lac, 2014). The fire started on 13 September and took close to a month to contain fully. It destroyed 80 residential structures. Arson is thought to be primarily responsible for the fire ('InciWeb the Incident Information System: King Fire, n.d.,' accessed January 2016)

Objective 1 – spatiotemporal modeling of air quality via King fire tweets

Social media data

Data were purchased from Gnip, Twitter's enterprise API (application programming interface) platform, on the basis of several keywords or hashtags (e.g., King AND Fire, King AND Smoke, #KingFire). All tweets originated from the United States between 1 September and 15 October 2014 and were geocoded using either the location of the tweet, when available, or extracting the user's profile to extrapolate a location. These criteria yielded approximately 14,000 tweets.

PM2.5 data

Ground-based monitoring of PM2.5 levels was obtained from the EPA's AirData air quality database. Measurements are collected by monitoring stations nationwide which then send hourly or daily aggregates to the EPA's database.

Geographic analytic area

A bounding box was superimposed on the area of the King fire to isolate tweets and air quality data that could be more directly linked to the fire event. An area of about 40,000 km^2 around El Dorado County was identified (see Figure 1). Defining the study area with the bounding box had the additional advantage of filtering out noise from large cities like San Francisco and Los Angeles, which heavily skewed both the Twitter and air quality data (i.e., more Tweets and greater PM2.5 concentrations that are not necessarily related to the fire). The area within the bounding box was then divided into 16 quadrats of 2500 km^2 each. There were five monitoring stations and a total of 1297 tweets within the bounding box.

Statistical model

A generalized additive model (GAM) was fit to the mean daily PM2.5 levels. Tweet counts, geographic quadrats, and days since the start of the fire were used as predictors. GAMs have been widely used for effectively modeling time-series health impact data of air pollution (Dominici, McDermott, Zeger, & Samet, 2002). An autoregressive integrated moving average process was applied to the residuals to account for temporal autocorrelation in the data. We controlled for geographically variable factors such as population, wind patterns, temperature, and other factors shown to influence PM2.5 levels (Preisler et al., 2015) by incorporating the geographic quadrats in the statistical model. Only quadrats which had a non-zero value for tweets and an active monitoring station were used in the analysis.

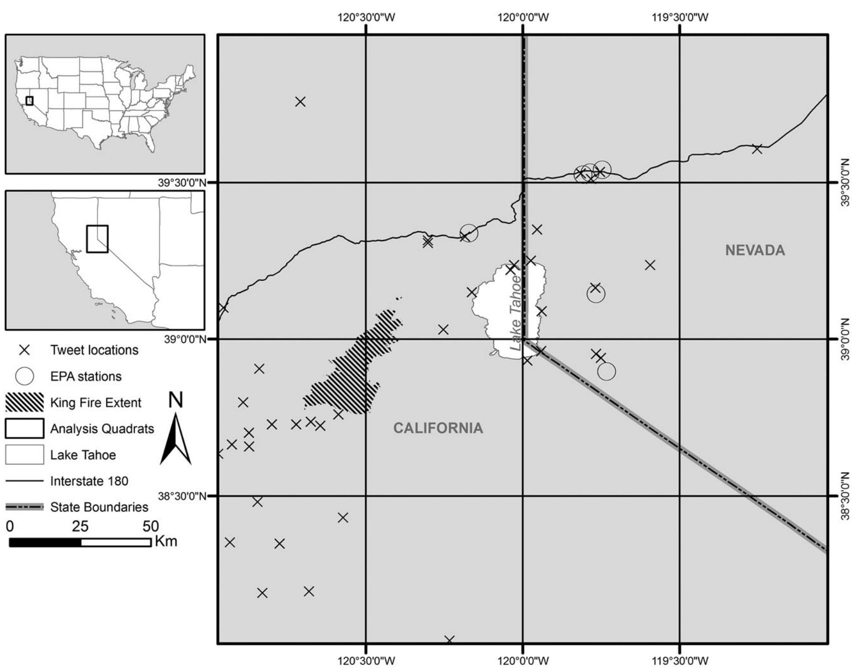

Figure 1. Map of analytic area depicting locus of King fire, quadrats, Tweet locations, and EPA monitoring stations.

The final model included 3 of the original 16 quadrats and 705 tweets across 37 days from 9 September 2014[1] to 15 October 2014. All models were fit using the 'mgcv' package in the R version 3.1.2 (R Core Team, 2014) following the equation:

$$Y_i = \beta_0 + s(day_i) + s(TC_i) + \gamma_{Q_i} + \varepsilon_i,$$

where Y_i is the daily mean PM2.5 levels, β_0 the intercept of the regression line, s(day) the smooth spline function of day of the 37-day analytic time period, s(TC) the smooth spline function of tweet count, γ_{Q_i} the categorical variable indicating quadrat of measurement, and ε_i the first-order autoregressive error to account for potential serial correlation in the daily PM values.

Results

PM2.5 concentrations over the 37-day period across our analytic region ranged from 3.15 to 100.76 μg/m^3 with an average level of 11.71 μg/m^3 as measured by the five EPA monitors.[2] The number of tweets in the same period and area ranged from 0 per day to 77 tweets a day, averaging about seven tweets a day. There was a high degree of correspondence between the two variables (Figure 2). The statistical model described 51% of the deviance in daily mean PM2.5 concentrations and additionally, the smoothed term for tweet count was found to be significantly different from zero ($p < .001$) and thus contributed to the model fit.

Discussion

The purpose of the model presented in this section was to assess whether crowdsourced data in the form of tweets could be used as a way to approximate daily mean PM2.5 levels.

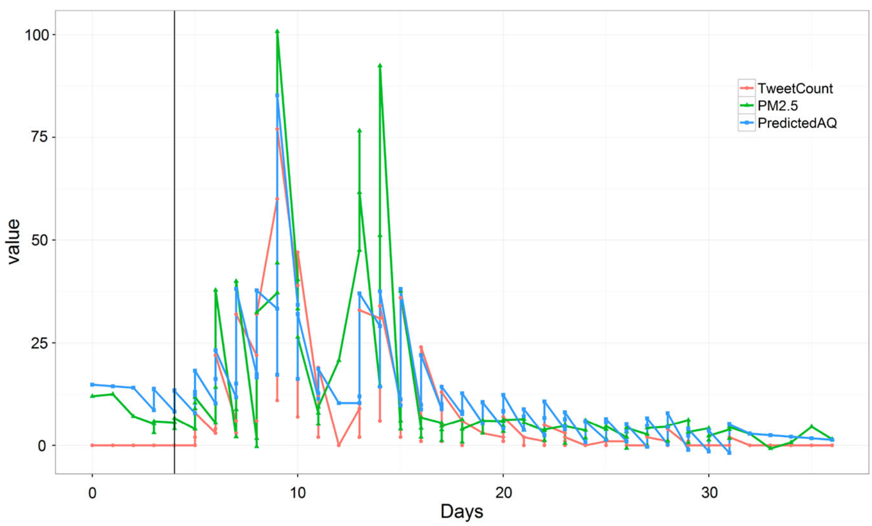

Figure 2. Relationship between number of tweets and PM2.5 concentrations over days since the start of analytic time period (9 September–15 October 2014). Vertical line indicates the start of the King fire on 13 September at approximately 6 pm.

We found that the frequency of daily tweets within a 40,000 km^2 area was a significant predictor of PM2.5 levels, beyond daily and geographic variation. These results suggest that social media can be a useful tool for the measurement of air quality impacts of wildfire events, particularly in the absence of data from physical monitoring stations.

Objective 2 – semantic content analysis of King fire tweets

Results from the first part of the project indicate that user-generated social media data can be used for estimating where there might be elevated levels of air pollutants. These data can also provide deeper, qualitative insight into how people are experiencing these sorts of events. Analyzing the content of their tweets can shed light on how people think about wildfire and the relative importance of smoke compared to other topics.

Topic models

Automated text analyses are emerging as a valuable way of inferring mental and social processes from unstructured, user-generated data (Dehghani, Sagae, Sachdeva, & Gratch, 2014). These new tools allow analysis of open-ended data without relying on resource-intensive, albeit more precise, manual human coding (Iliev, Dehghani, & Sagi, 2015). Given the abundance of text data being generated by internet users globally, automated text analysis techniques are crucial for distilling the themes and ideas present in virtual exchanges of information. There are two commonly used approaches to model topics in text: Latent Dirichlet Allocation (LDA; Blei, Ng, & Jordan, 2003) and the Structural Topic Model (STM) (Roberts et al., 2014). Both are generative approaches, built on the assumptions that documents are comprised of a distribution of topics and that topics are made up of a semantically coherent distribution of words. Topic models in both approaches result in the most probable structure to explain the collection of documents (Chen, 2011) and are both bottom-up, unsupervised approaches, in the sense that they infer rather than assume the content of topics. Both have been applied to a number of fields such as health research (e.g., tagging patient records), education research (e.g., quickly identifying commonalities in student-generated text), and political science (e.g., differences in content by party affiliation) (Blei et al., 2003; Grimmer, 2010; Quinn, Monroe, Colaresi, Crespin, & Radev, 2010; Wang & Blei, 2011). LDA and STM differ in the probability distributions they use to generate topics. LDA uses a Dirichlet distribution and STM uses a logistic-normal distribution and allows for topic prevalence to vary based on document metadata. In other words, STM allows researchers to examine the prevalence of particular topics in a body of text and examine how it varies based on other factors of interest. Therefore, STM was a better fit for the current research objective, as it allowed us to examine how the content of users' tweets may vary by time and geographic area.

Methods

We used the R implementation of STM to derive a topic model of tweets related to the King fire. Our sample consisted of 14,093 tweets posted since the beginning of the King fire, starting on 14 September through 15 October. A topic model with 20 topics was fit to the data. An analysis of topic semantic coherence, exclusivity, and number of iterations required for model convergence showed the 20-topic model to be the best fit to the corpus compared to alternate versions with 10, 15, 30, 40, or 50 topics (please see Supplementary Information for a table with key measures). Each tweet was also tagged with whether it belonged to the region contained by the bounding box described above (i.e., 40,000 km^2 region around El Dorado County) so we could analyze how topic distributions varied by distance to the locus of the King fire. We also retained information about when the tweet was posted to assess any longitudinal trends in the semantic content of tweets.

Results

The 15 most common words for the 20 topics generated in the model are depicted in Table 1. As expected, some of the most prevalent topics in users' tweets were related to fire characteristics such as the size of the fire and how many acres were contained or active. Another topic often tweeted about was the number of homes or residential structures that had been destroyed or were threatened as well as the areas under evacuation. As expected, concerns about the impact of smoke and air quality were another topic in tweets about the King fire. For instance, users posted tweets such as, 'last day of summer. cough cough cough. gag choke. ugh. king fire still burning. thank you summer' or 'air quality past unhealthy lord have mercy and grant containment and safety #kingfire #carsoncity.' Users also discussed the more aesthetic features of the smoke such as the size, shape, and color lent to the sunsets ('looks like a cloud but it's a massive smoke plume. serious. #kingfire'; 'smoke from the king fire made for quite a spectacular sunset tonight as seen from the carson high soccer field').

In addition, safety, particularly of firefighters, was a common topic. This topic was characterized by messages such as, 'holding safety of friends & neighbors evacuating from the #kingfire in my consciousness. praying for rain tomorrow' or 'thank you & godspeed to the brave firefighters & support crews!!! #kingfire.' The relationship between topics revealed a distinction between clusters of topics that provided factual characteristics of the fire (e.g., size, containment efforts, number of structures destroyed) and those that described the more emotional and personal aspects of the fire (praying for the community, health impact of air quality, concern for firefighting personnel) (Figure 3).

Geographic variation

The location information present in the tweets also allowed us to assess differences in topic distribution by distance to the locus of the King fire. Using the same geographic analytic area as described in the first analysis, we tagged tweets with whether they were within the bounding box or outside of it. We then computed the average prevalence for each topic for tweets within the bounding box versus outside. Several topics appeared to vary considerably across geographic information (Figure 4). Tweets about air quality and its potential health impact were more likely to originate from the region closer to the fire. In fact, all smoke-related topics seemed to be more pervasive inside the bounding box than outside. Similarly, messages about thoughts and prayers for firefighters and area families tended to be more frequent in the tweets from inside the bounding box than outside of it. However, tweets about the arson suspect, the apparent cause of the fire, were more frequent in the distal region than within the bounding box as were topics about the number of structures that were threatened or destroyed by the fire.

Table 1. Labeled topics with top 15 words in each topic.

Topic	Assigned label	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10	Word 11	Word 12	Word 13	Word 14	Word 15
1	Spread of Wildfire	'wildfir'	'via'	'spread'	'retard'	'drop'	'use'	'slow'	'record'	'across'	'water'	'drought'	'year'	'fireâ'	'extrem'	'amount'
2	Fire Frontline Personnel	'kingfir'	'today'	'line'	'httptcovqvqqp'	'httptcofkiseejv'	'rfirenew'	'assign'	'zone'	'personnel'	'front'	'hold'	'fuel'	'calfirenew'	'folk'	'complet'
3	Firefight Crew	'firefight'	'rain'	'crew'	'help'	'big'	'mani'	'infrar'	'control'	'gain'	'sent'	'sept'	'ground'	'return'	'injur'	'falsecolor'
4	Smoke Plume Visible	'smoke'	'look'	'mile'	'can'	'plume'	'north'	'away'	'sky'	'cawx'	'move'	'bad'	'west'	'yesterday'	'afternoon'	'think'
5	Thanks Safety Help	'amp'	'thank'	'work'	'last'	'night'	'safe'	'laketaho'	'head'	'deploy'	'team'	'photo'	'pleas'	'hour'	'strike'	'wow'
6	Evacuation Community Shelter	'evacu'	'cafir'	'communiti'	'shelter'	'school'	'burger'	'info'	'need'	'bolesfir'	'center'	'pollockpin'	'busi'	'pictur'	'camino'	'south'
7	Official Newscast Smoke	'say'	'video'	'offici'	'newsca'	'live'	'size'	'cloud'	'doubl'	'report'	'watch'	'peopl'	'pyrocumulus'	'meadow'	'huge'	'cbs'
8	Fire Progress Conditions	'sacramento'	'contain'	'grow'	'battl'	'weather'	'wind'	'continu'	'percent'	'crew'	'blaze'	'expect'	'progress'	'chang'	'aid'	'reach'
9	Unrelated to King Fire	'keep'	'way'	'portland'	'state'	'sacbe'	'blvd'	'martin'	'luther'	'local'	'region'	'emerg'	'block'	'next'	'medic'	'despit'
10	Casual Conversation	'get'	'will'	'good'	'make'	'today'	'tonight'	'hope'	'great'	'much'	'got'	'meet'	'realli'	'take'	'love'	'right'
11	Lake Tahoe Event Cancellation	'taho'	'lake'	'morn'	'cancel'	'check'	'via'	'placervill'	'ironman'	'vintag'	'etsi'	'famili'	'back'	'set'	'due'	'anchor'
12	Acres Contained Update	'acr'	'contain'	'updat'	'counti'	'burn'	'dorado'	'now'	'pine'	'pollock'	'near'	'calfir'	'east'	'amp'	'overnight'	'calif'
13	Massive Fire Fight	'fire'	'king'	'massiv'	'fight'	'still'	'whip'	'injuri'	'break'	'sustain'	'four'	'monster'	'gtgt'	'stockton'	'lodi'	'cvbtnew'
14	Arson Suspect Arrest	'arson'	'arrest'	'man'	'suspect'	'start'	'set'	'kingfir'	'huntsman'	'wayn'	'charg'	'accus'	'call'	'allen'	'made'	'held'
15	California Homes Threatened	'california'	'home'	'threaten'	'northern'	'structur'	'forest'	'destroy'	'time'	'rage'	'growth'	'nation'	'thousand'	'eldorado'	'los'	'rapid'
16	Air Quality Affected	'air'	'day'	'see'	'latest'	'qualiti'	'reno'	'affect'	'trucke'	'know'	'valley'	'spot'	'hot'	'advisori'	'prussia'	'number'
17	Roads Closed Residences	'close'	'just'	'like'	'hwi'	'highway'	'view'	'one'	'road'	'hous'	'resid'	'still'	'due'	'stay'	'flame'	'pic'
18	Area Map Information	'area'	'map'	'cal'	'come'	'team'	'incid'	'inform'	'high'	'command'	'bay'	'possibl'	'mop'	'heavi'	'perimet'	'god'
19	NASA Satellite Damage	'show'	'photo'	'new'	'nasa'	'damag'	'imag'	'devast'	'seen'	'space'	'warn'	'red'	'amaz'	'burn'	'satellit'	'issu'
20	News Sunset Photos	'news'	'citi'	'fox'	'even'	'sierra'	'sunset'	'post'	'shot'	'stop'	'smokey'	'carson'	'leav'	'nevada'	'fighter'	'alarm'

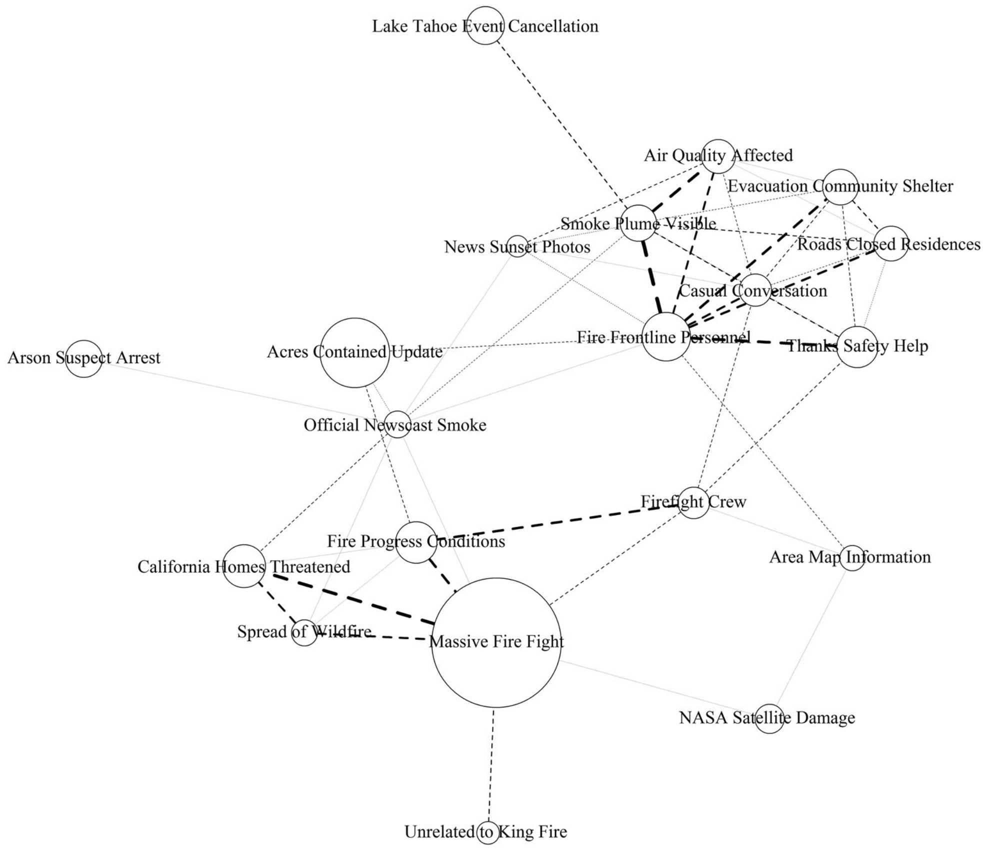

Figure 3. Topic map of King fire-related Tweets. Topic prevalence is depicted through the size of the nodes and the strength of correlation between topics is depicted by the thickness of the edges.

Temporal variation

We were also interested in assessing how topic distribution might vary over time. That is, did certain topics become more important as the fire persisted or did they show an initial spike in activity followed by a decline? We selected the two most distinctive topics from within the bounding box ('air quality' and 'smoke plume visible') and outside of it ('arson suspect arrested' and 'California homes threatened') and assessed how the prevalence of these topics changed over the course of the fire.

Discussion of the arson suspect peaked within a week of the start of the fire and markedly dissipated soon afterward. This indicates that within the overall scope of the fire, the primary cause of arson and the ensuing legal consequences were not of much interest to Twitter users. Similarly, there was a steep increase in coverage of the number of homes threatened in northern California which soon leveled off. The other two topics, while not as prevalent, were more persistent. Discussions of air quality peaked four to seven days after the start of the fire and then flattened out (though remaining at elevated levels) through the next couple of weeks (Figure 5). This timeline coincides well with the onset of heavy rains and wet weather which helped turn the tide in the suppression effort (Lac, 2014).

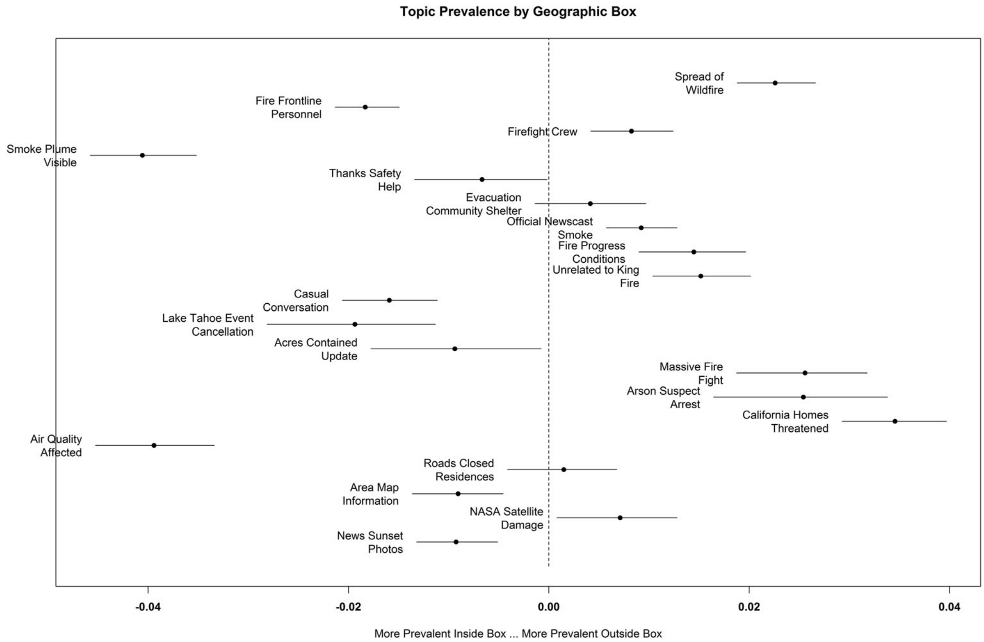

Figure 4. Plot showing differences in topic prevalence by geographic bounding box versus outside.

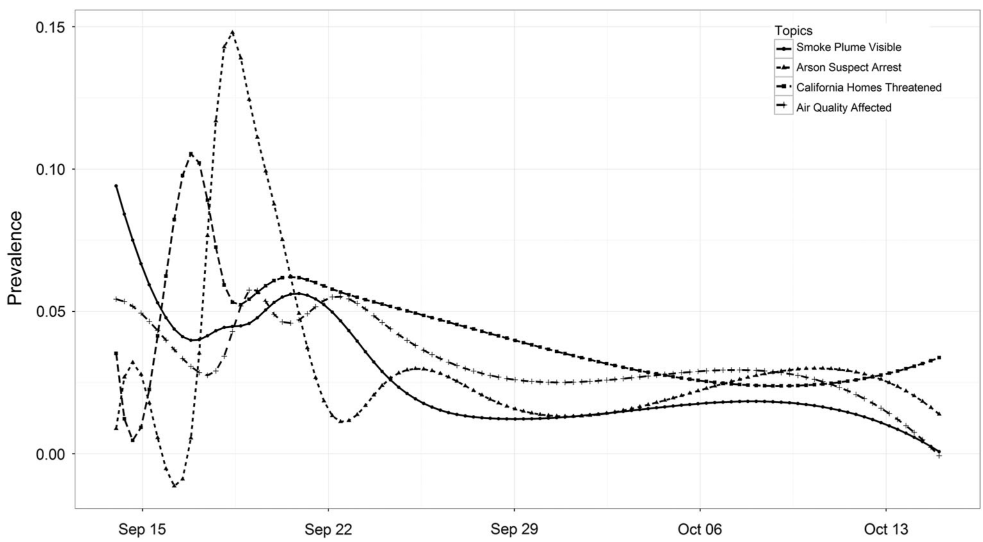

Figure 5. Topic prevalence across days since the start of King fire on 13 September 2014.

Discussion

The second analysis allowed us to explore tweet content about the King fire to understand what users were most concerned about and whether people closest to the fire attended more to certain issues than people further away from it. In addition to topics bearing information about the progress of the fire and containment efforts, several noteworthy topics emerged from the topic modeling approach. For instance, a significant topic about praying for firefighters and area families emerged from the model and was more dominant for tweets closest to the fire. Complementing the statistical model described in the first objective, we found that discussion of smoke and the air quality impacts of the King fire was at the forefront of users' tweets. Again, this topic was more characteristic of tweets within the geographic bounding box surrounding El Dorado County.

General discussion

This project combines two valuable facets of social media data. Using geographic information to model spatiotemporal patterns of air quality impacts after a wildfire event, we found a substantial correspondence between a number of tweets and measured PM2.5, when controlling for spatial and serial autocorrelation. This shows potential for using tweets as a rough approximation for air quality in areas where monitoring data is limited. In addition to the tweet frequency, analyzing the user-generated text component helped us understand social media users' perceptions of a wildfire event: specifically, what people find most noteworthy and concerning about the event. As expected, social media posts are an important means of disseminating information. For instance, important themes in the tweets include how far the fire has spread, updates on containment efforts, evacuation notices or even posts about school and other local institutions closing. However, in addition to these information-based tweets, there are significant instances of people posting their prayers and safety concerns for fighters as well as a sense of the frustration that accompanies a natural disaster with profound health impacts ('smoke again. i can't take this for too much longer'). The topic map derived through this approach and illustrated in Figure 3 can be compared to a concept map of Twitter users' discussion of the King fire. It reveals clusters of inter-related and isolated concepts (e.g., damage that can be seen via NASA satellites or the arrest of the arsonist responsible for the fire), showing that distance from the locus of the fire is a major determinant of how people think about wildfires (or at the very least, tweet about them). This is a novel contribution of the current project, illustrating that crowdsourced data can be a time-sensitive estimate of air quality impacts of wildfires as well as a means of understanding how people conceptualize and are affected by resultant smoke and rescue efforts.

Limitations and future directions

Despite these promising steps, it is worth noting that the current project was limited in scope to examining only the King fire in 2014. This relatively narrow context constrained the results reported here in at least two ways. (1) In the first part of the project, where we estimate PM2.5 levels using tweets in a geographically bounded region, we had only several hundred tweets, and (2) throughout the entire project we looked at a relatively small geographic region northeast of Sacramento, California. Both of these constraints limit the generalizability of this study as it does not answer the question of whether crowdsourced data can be a useful mapping tool for wildfire smoke in other regions. To address this concern, we have begun extending the current methodology to wildfires nationwide and are now assessing correspondence between air quality estimates and VGI from fire-related tweets in the 2015 wildfire season.

A large-scale expansion of this project will be the next step in the potential development of a predictive tool for fire and land managers, as well as for people living in proximity of wildfire-prone regions. We hope that we can use social media related to fire and smoke to build concrete estimates of PM2.5 levels that can be monitored and measured virtually in real time. Though these tools cannot be viewed as a substitute for physical measurements of air quality, they may be able to provide people with more dynamic data to make decisions regarding their health and well-being. It is also probable that this tool can become a more direct means for people to provide feedback regarding smoke conditions, thereby opening up a two-way avenue of communication where managers can disseminate information to the public and the public can respond.

Finally, although a new set of computer tools make it possible to collect, process, and analyze the large amount of data being generated by users, computers are not and may never be as proficient at understanding the gradations of human speech as are humans themselves. For instance, humor, pain, sarcasm, and other emotional data cannot be captured satisfactorily through automated text approaches. Crowdsourced data trawled en masse may also have a much lower signal to noise ratio through more controlled studies which ask targeted questions and have more direct interaction with participants. Yet, this type of data promises to give voice to regions that were previously underrepresented and create new ways to provide disaster and crisis relief. Wildfires, increasing in frequency and severity from global climate change, and the resultant smoke have become an urgent public health issue. Once validated, UGC data could make a significant contribution to understanding the patterns of smoke concentration and their effects on livelihoods.

Notes

1. The analytic time period began five days before the start of the fire to provide a baseline of PM2.5 levels. The results of the model are even stronger if we exclude this period as there were no tweets about the King fire in these days.
2. For reader reference, the EPA's most recent guidelines hold the cutoff for a 'good' level of PM2.5 concentration at 12 $\mu g/m^3$ and levels above 55 $\mu g/m^3$ are considered unhealthy for all individuals (http://www3.epa.gov/airquality/particlepollution/2012/decfsstandards.pdf, accessed January 2016).

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Sonya Sachdeva is a computational social scientist with the US Forest Service and an adjunct faculty member in the Environmental Policy and Culture program at Northwestern University. Her research examines the impact of cultural factors, such as sacred values, scarcity, or self-concepts, on conservation behavior and attitudes [email: sonyasachdeva@fs.fed.us].

Sarah McCaffrey, Ph.D, is a Research Social Scientist with the Northern Research Station. Her research focuses on understanding the social dynamics of wildfire management [email: smccaffrey@fs.fed.us].

Dexter Locke is a PhD candidate in the Graduate School of Geography at Clark University. He has degrees in Natural Resources Planning, Environmental Science, and Geography. His research focuses on human-environment interactions, frequently in urban areas [email: dexter.locke@gmail.com].

References

Barrington, L., Ghosh, S., Greene, M., Har-Noy, S., Berger, J., Gill, S., … Huyck, C. (2012). Crowdsourcing earthquake damage assessment using remote sensing imagery. *Annals of Geophysics*, *54*(6). doi:10.4401/ag-5324

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

BlueSky Modeling Framework | AirFire. (n.d.). Retrieved January 25, 2016, from http://www.airfire.org/bluesky/

Bowman, D. M. J. S., & Johnston, F. H. (2005). Wildfire smoke, fire management, and human health. *EcoHealth*, *2*(1), 76–80. doi:10.1007/s10393-004-0149-8

Calkin, D. E., Thompson, M. P., & Finney, M. A. (2015). Negative consequences of positive feedbacks in US wildfire management. *Forest Ecosystems*, *2*(1), 1–10. doi:10.1186/s40663-015-0033-8

Cassa, C. A., Chunara, R., Mandl, K., & Brownstein, J. S. (2013). Twitter as a sentinel in emergency situations: Lessons from the Boston Marathon explosions. *PLoS Currents*, *5*. doi:10.1371/currents.dis.ad70cd1c8bc585e9470046cde334ee4b

Chen, E. (2011). *Introduction to latent Dirichlet allocation*. Retrieved from http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/

Dehghani, M., Sagae, K., Sachdeva, S., & Gratch, J. (2014). Analyzing political rhetoric in conservative and liberal weblogs related to the construction of the 'Ground Zero Mosque'. *Journal of Information Technology & Politics*, *11*(1), 1–14. doi:10.1080/19331681.2013.826613

Dominici, F., McDermott, A., Zeger, S. L., & Samet, J. M. (2002). On the use of generalized additive models in time-series studies of air pollution and health. *American Journal of Epidemiology*, *156*(3), 193–203. doi:10.1093/aje/kwf062

Dominici, F., Peng, R. D., Bell, M. L., Pham, L., McDermott, A., Zeger, S. L., & Samet, J. M. (2006). Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *JAMA*, *295*(10), 1127–1134. doi:10.1001/jama.295.10.1127

Earle, P., Guy, M., Buckmaster, R., Ostrum, C., Horvath, S., & Vaughan, A. (2010). OMG earthquake! Can Twitter improve earthquake response? *Seismological Research Letters*, *81*(2), 246–251. doi:10.1785/gssrl.81.2.246

Gao, H., Barbier, G., & Goolsby, R. (2011). Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems*, *26*(3), 10–14. doi:10.1109/MIS.2011.52

Genes, N., Chary, M., & Chason, K. (2014). Analysis of Twitter users' sharing of official New York storm response messages. *Medicine 2.0*, *3*(1), e1. doi:10.2196/med20.3237

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, *69*(4), 211–221. doi:10.1007/s10708-007-9111-y

Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, *18*(1), 1–35. doi:10.1093/pan/mpp034

Iliev, R., Dehghani, M., & Sagi, E. (2015). Automated text analysis in psychology: Methods, applications, and future developments. *Language and Cognition*, *7*(2), 265–290. doi:10.1017/langcog.2014.30

InciWeb the Incident Information System: King Fire. (n.d.). Retrieved January 22, 2016, from http://inciweb.nwcg.gov/incident/4108/

Jiang, W., Wang, Y., Tsou, M.-H., & Fu, X. (2015). Using social media to detect outdoor air pollution and monitor air quality index (AQI): A geo-targeted spatiotemporal analysis framework with Sina Weibo (Chinese Twitter). *PLoS ONE*, *10*(10), e0141185. doi:10.1371/journal.pone.0141185

Johnston, F. H., Henderson, S. B., Chen, Y., Randerson, J. T., Marlier, M., DeFries, R. S., … Brauer, M. (2012). Estimated global mortality attributable to smoke from landscape fires. *Environmental Health Perspectives*, *120*(5), 695–701. doi:10.1289/ehp.1104422

Kent, J. D., & Capello, H. T. (2013). Spatial patterns and demographic indicators of effective social media content during the Horsethief Canyon fire of 2012. *Cartography and Geographic Information Science*, *40*(2), 78–89. doi:10.1080/15230406.2013.776727

Kinney, P. L. (2008). Climate change, air quality, and human health. *American Journal of Preventive Medicine*, *35*(5), 459–467. doi:10.1016/j.amepre.2008.08.025

Kochi, I., Donovan, G. H., Champ, P. A., & Loomis, J. B. (2010). The economic cost of adverse health effects from wildfire-smoke exposure: A review. *International Journal of Wildland Fire*, *19*(7), 803–817.

Lac, J. F. du. (2014, September 25). California's King fire is the size of a major city. These are the firefighters battling the big blaze. *The Washington Post*. Retrieved from https://www.washingtonpost.com/news/post-nation/wp/2014/09/25/californias-king-fire-is-as-big-as-a-major-city-these-are-the-firefighters-battling-the-enormous-blaze/

Larkin, N. K., O'Neill, S. M., Solomon, R., Raffuse, S., Strand, T., Sullivan, D. C., … Ferguson, S. A. (2009). The BlueSky smoke modeling framework. *International Journal of Wildland Fire*, *18*(8), 906–920. doi:10.1071/WF07086

Liu, Y., Stanturf, J., & Goodrick, S. (2010). Trends in global wildfire potential in a changing climate. *Forest Ecology and Management*, *259*(4), 685–697. doi:10.1016/j.foreco.2009.09.002

McCaffrey, S., & Olsen, C. (2012). *Research perspectives on the public and fire management: A synthesis of current social science on eight essential questions*. JFSP synthesis reports. Retrieved from http://digitalcommons.unl.edu/jfspsynthesis/17

Mei, S., Li, H., Fan, J., Zhu, X., & Dyer, C. R. (2014). *Inferring air pollution by sniffing social media*. 2014 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), pp. 534–539). doi:10.1109/ASONAM.2014.6921638

Morgan, G., Sheppeard, V., Khalaj, B., Ayyar, A., Lincoln, D., Jalaludin, B., … Lumley, T. (2010). Effects of bushfire smoke on daily mortality and hospital admissions in Sydney, Australia. *Epidemiology*, *21*(1). Retrieved from http://journals.lww.com/epidem/Fulltext/2010/01000/Effects_of_Bushfire_Smoke_on_Daily_Mortality_and.9.aspx

Mott, J. A., Meyer, P., Mannino, D., Redd, S. C., Smith, E. M., Gotway-Crawford, C., & Chase, E. (2002). Wildland forest fire smoke: Health effects and intervention evaluation, Hoopa, California, 1999. *Western Journal of Medicine*, *176*(3), 157–162.

Preisler, H. K., Schweizer, D., Cisneros, R., Procter, T., Ruminski, M., & Tarnay, L. (2015). A statistical model for determining impact of wildland fires on Particulate Matter (PM2.5) in Central California aided by satellite imagery of smoke. *Environmental Pollution*, *205*, 340–349. doi:10.1016/j.envpol.2015.06.018

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, *54*(1), 209–228. doi:10.1111/j.1540-5907.2009.00427.x

Ram, S., Zhang, W., Williams, M., & Pengetnze, Y. (2015). Predicting asthma-related emergency department visits using big data. *IEEE Journal of Biomedical and Health Informatics*, *19*(4), 1216–1223. doi:10.1109/JBHI.2015.2404829

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Richardson, L. A., Champ, P. A., & Loomis, J. B. (2012). The hidden cost of wildfires: Economic valuation of health effects of wildfire smoke exposure in Southern California. *Journal of Forest Economics*, *18*(1), 14–35. doi:10.1016/j.jfe.2011.05.002

Rittmaster, R., Adamowicz, W. L., Amiro, B., & Pelletier, R. T. (2006). Economic analysis of health effects from forest fires. *Canadian Journal of Forest Research*, *36*(4), 868–877. doi:10.1139/x05-293

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., … Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, *58*(4), 1064–1082. doi:10.1111/ajps.12103

Rolph, G. D., Draxler, R. R., Stein, A. F., Taylor, A., Ruminski, M. G., Kondragunta, S., … Davidson, P. M. (2009). Description and verification of the NOAA smoke forecasting system: The 2007 fire season. *Weather and Forecasting*, *24*(2), 361–378. doi:10.1175/2008WAF2222165.1

Schwartz, J., Slater, D., Larson, T. V., Pierson, W. E., & Koenig, J. Q. (1993). Particulate air pollution and hospital emergency room visits for asthma in Seattle. *American Review of Respiratory Disease*, *147*(4), 826–831. doi:10.1164/ajrccm/147.4.826

Shelton, T., Poorthuis, A., Graham, M., & Zook, M. (2014). Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of 'big data'. *Geoforum*, *52*, 167–179. doi:10.1016/j.geoforum.2014.01.006

Strand, T. M., Larkin, N., Craig, K. J., Raffuse, S., Sullivan, D., Solomon, R., … Pryden, D. (2012). Analyses of BlueSky Gateway PM2.5 predictions during the 2007 southern and 2008 northern California fires. *Journal of Geophysical Research: Atmospheres*, *117*(D17), D17301. doi:10.1029/2012JD017627

Sutton, J., League, C., Sellnow, T. L., & Sellnow, D. D. (2015). Terse messaging and public health in the midst of natural disasters: The case of the Boulder floods. *Health Communication*, *30*(2), 135–143. doi:10.1080/10410236.2014.974124

US EPA. (2015). *Basic information | fine particle (PM2.5) designations | US EPA*. Retrieved January 22, 2016, from http://www3.epa.gov/pmdesignations/basicinfo.htm

Veer, E., Ozanne, L. K., & Hall, C. M. (2015). Sharing cathartic stories online: The internet as a means of expression following a crisis event. *Journal of Consumer Behaviour*. doi:10.1002/cb.1569

Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). Microblogging during two natural hazards events: What Twitter may contribute to situational awareness. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1079–1088). New York, NY: ACM. doi:10.1145/1753326.1753486

Wang, C., & Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 448–456). New York, NY: ACM. doi:10.1145/2020408.2020480