The Workshops of the Tenth International AAAI Conference on Web and Social Media
Social Web for Environmental and Ecological Monitoring: Technical Report WS-16-20

Using Social Media to Detect and Locate Wildfires

Chris A. Boulton
College of Life and Environmental Sciences
University of Exeter
c.a.boulton@exeter.ac.uk

Humphrey Shotton
College of Engineering, Mathematics and Physical Sciences
University of Exeter

Hywel T. P. Williams
College of Life and Environmental Sciences
University of Exeter
h.t.p.williams@exeter.ac.uk

Abstract

Methods for detecting and tracking natural hazards continue to increase in coverage, resolution and reliability. However, information on the social impacts of natural hazards is often lacking. Here we test the feasibility of using social media data (Twitter and Instagram) to detect and map an important class of natural hazard: wildfires. We analyse social media posts associated with wildfires over several time periods and compare them with wildfire occurrence data derived from satellite-based remote sensing data and on-the-ground observations. For the whole of the contiguous United States, we find significant temporal correlations between wildfire-related social media activity and wildfire occurrence, but also that there is substantial variation in the strength of this relationship at smaller spatial scales (states and counties). We then explore the utility of social media for location of wildfire events, finding good evidence to support further development of such methods. We conclude by discussing several challenges and opportunities for application of this novel data resource to provide information on impacts of natural hazards.

Background

Advances in measurement and remote sensing are giving an increasingly detailed and accurate record of natural hazards. Yet methods for systematically recording and quantifying the ways in which environmental hazards impact on human society are relatively sparse. There remains a large disparity between knowledge of the physical phenomena and knowledge of how these events will affect people. Meanwhile, the increasing digitisation of human behavior through online communication and the Web is creating rich datasets that are being successfully mined for insights into many different social processes (Lazer et al. 2009). The massive volume and global reach of online data flows offers an opportunity to observe the impacts of environmental hazards from a human-oriented perspective; since the data is a by-product created by the online communication of billions of individual webusers, it will inevitably reflect those aspects of natural hazards that have most relevance to human activity. The creation of the Internet and its rapid growth in use worldwide has inadvertently created a global sensor network that can be harnessed as a social observatory for environmental hazards and events.

Social media have already been successfully used to detect extreme weather (Kirilenko, Molodtsova, and Stepchenkova 2015) and earthquakes (Sakaki, Okazaki, and Matsuo 2010), while mobile phone data have been used to map migratory flows following natural disasters (Lu, Bengtsson, and Holme 2012) and to predict disease spread (Wesolowski et al. 2012; 2015). Widespread social media use by affected individuals in disaster scenarios has lead to its integration into software tools for humanitarian response management which are now routinely used by aid agencies (e.g. Ushahidi http://www.ushahidi.com/). Various studies (e.g. (Elgesem, Steskal, and Diakopoulos 2015; Schafer 2012; Kirilenko and Stepchenkova 2014; Kirilenko, Molodtsova, and Stepchenkova 2015; Olteanu et al. 2015; Williams et al. 2015; O'Neill et al. 2015) have shown a high volume of social media communication around the broad topic of climate change, which is predicted to increase the frequency and severity of several kinds of natural hazard. Olteanu et al. (2015) showed that social media can be used to detect climate-related events and that the events detected by social media show only partial overlap with those reported in mainstream news media. Kirilenko, Molodtsova and Stepchenkova (2015) showed US social media activity about climate change was correlated with local incidence of extreme temperatures. These initial findings suggest that social media may have utility for detecting and mapping environmental hazards and climate-related impacts, but a robust methodology has yet to be defined and validated.

This paper explores the feasibility of creating a "social observatory" that uses social media data to detect and characterise environmental hazards and associated social

impacts. Focusing on wildfires in the USA as a case study, we demonstrate that social media activity related to wildfires is temporally and spatially correlated with known wildfire events. We then show that wildfire events can be located with reasonable accuracy using social media data alone. Together these findings establish the preconditions for use of social media data to augment conventional observations of wildfires with previously unavailable information on social impacts that is hard to gather elsewhere.

Below we describe the data collection, methods and results, followed by some discussion of various challenges and opportunities presented by this novel use of social media data.

Data Collection and Methods

We collected social media posts related to wildfires from the popular platforms Twitter and Instagram. Twitter is a social messaging platform with 288 million active monthly users sending 500 million tweets per day (Statista 2015). Instagram is a social photo-sharing platform with 300 million active monthly users sharing photos from smartphones (Statista 2015). Both Twitter and Instagram offer public Application Programming Interfaces (APIs) by which their databases can be queried. Data collection used hashtags and keywords associated with wildfires. Instagram posts were collected using hashtags {#wildfire, #bushfire}. Twitter posts were collected using the Search API with keywords {"wildfire", "wild fire", "grassfire", "grass fire", "wildland fire", "brush fire", "bushfire", "bush fire", "forest fire", "forestfire", "fire science", "firewise", "fire danger"}.

Wildfire occurrence data were derived from two sources. The MODIS Active Fire Detections data product supplied by the US Department of Agriculture (USDA) Forest Service Remote Sensing Applications Center (http://activefiremaps.fs.fed.us/) is remote sensing data derived from the MODIS satellites that combines imagery and thermal anomalies to detect fires at 1km spatial resolution (USDA Forest Service 2015). The Fire Program Analysis dataset (FPA) available from the USDA Forest Service Research Data Archive (http://www.fs.usda.gov/rds/archive/Product/RDS-2013-0009.3/) is compiled from observational reports by federal, state and local fire organisations in the US and gives a spatial database of 1.73 million wildfires, located at minimum 1 mile spatial resolution, over a 22-year period from 1992 to 2013 (Short 2015).

Due to restrictions on extracting historical data from the Twitter API, and the time period covered by the FPA dataset, we compare datasets over several different time periods (detailed in Table 1). While fire ocurrences from

Dataset	Period 1	Period 2	Period 3
Instagram	39,564 (2,163)	114,307 (6,898)	27,883 (2,943)
Twitter	-	-	905,611 (2,798)
MODIS	513,451	833,276	83,443
FPA	289,652	-	-

Table 1: Volumes of fire occurrences and social media posts in the datasets analysed. Three different time periods were studied due to differing data availability (Period 1: Jan 2011-Dec 2013; Period 2: Jan 2011-Sept 2015; Period 3: April-Sept 2015). Social media volumes are given as total number of posts collected; number in brackets indicates the number of geotagged posts located within the US, i.e. the number of posts used for further analysis. MODIS fire occurrences indicate the number of 1km squares where fire is detected. FPA fire occurrences are number of wildfires reported.

MODIS and FPA are spatially located, only 0.51% of the Twitter posts and 12.04% of the Instagram posts we collected were geotagged, preventing the majority of these from being linked to specific fire activity. We restrict our analysis to geotagged posts originating from within the contiguous USA.

We first analyse daily frequencies of social media posts and fire occurrences over time. Time series were created for the whole US, for each state and for each county. These spatial scales are used since fire management strategies vary between administrative units at these levels, creating a natural scale for spatial analysis and future exploitation of results. We compare correlations between the derived time series for different datasets and spatial units.

We then focus our analysis on the California-Nevada region (defined here as a rectangular box with coordinate ranges 114-126ºW and 33-42ºN) to determine the spatial association of social media posts and wildfire events. We test the simple hypothesis that social media posts about wildfires are triggered by nearby wildfire events, by comparing the observed distances between social media posts and wildfires against null expectations derived from three different null models. For each social media post, we measure the distance (km) to the closest observed wildfire (using MODIS data) on the same day. We also calculate the equivalent distances when the social media posts for that day are randomly re-located according to each of three null models:

- *Null Model 1*: Chooses a new location for each post by selecting new coordinates with uniform probability from the whole land surface of the study region.
- *Null Model 2*: Chooses a new location for each post by random selection from the set of all

locations from which social media posts have originated over the whole study period. This controls for spatial variation in social media usage based on the spatial distribution of posts in our dataset.

- *Null Model 3*: Similar to Null Model 1 but uses population density data (CIESIN/CIAT 2005) to weight the probability distribution towards locations that have higher populations. This controls for spatial variation in social media usage based on an assumed tendency for more social media posts to originate in areas where there are more people.

The statistics used for comparison are the mean/median distances across the set of all closest distances (i.e. collating all posts on all days). We calculate these values for the observed data and for 100 randomisations using each null model. A single null randomisation is created by applying the chosen null model to randomly re-locate the social media posts for each day separately, preserving the number/location of wildfires and the number of social media posts, and then combining the derived distances for each day into a single set to allow calculation of mean/median values.

Finally, we use the spatial distribution of social media posts on a given day to create a "heat map" of social media activity about wildfires, in order to explore the utility of such data for locating and detecting wildfire events. To do this, we first compiled the frequency distribution of all observed closest distances between social media posts and observed fires (Figure 3). This distribution has a folded normal form since only the magnitude of the distance is recorded (direction is ignored). We used the standard deviation of this distribution to set the radius of a multivariate Gaussian function; a spatial density field for social media activity for a given day is then created by summing across a set of such Gaussian functions centred on each social media post. Peaks (hotspots) in the resulting activity field, which may occur between the actual locations of social media posts due to the summation, may indicate the locations of wildfire events.

Results

Over the whole of the US, social media activity is significantly positively correlated with fire occurrence (Figure 1). Instagram posts show reasonable correlations with fire occurrence as reported by FPA (*Instagram-FPA, 2011-2013: r=0.305, p<2e-16*) and by MODIS (*Instagram-MODIS, 2011-2015: r=0.366, p<2e-16*). We find stronger correlations in our later Instagram/Twitter/MODIS comparison (*Instagram-MODIS, April-Sept 2015: r=0.716, p<2e-16; Twitter-MODIS, April-Sept 2015: r=0.529, p<2e 11*). We attribute the weaker correlations found in the longer 2011-2013 and 2011-2015 comparisons to the substantial increase in Instagram usage since 2011, which implies much lower numbers of users and posts in the early stages. Interestingly, the two fire occurrence datasets show strong, but not perfect, correlation (*MODIS-FPA, 2011-2013: r=0.594, p<2e-16*). This highlights uncertainties in fire detection and reporting. The two social media datasets also show strong but imperfect correlation (*Instagram-Twitter, April-Sept 2015: r=0.520, p<4e-11*), suggesting that the two data sources offer distinct information and are not redundant.

When we look for temporal correlations between the daily time series at the scale of US states, we find a mixed signal (Table 2 & Figure 2). Some states have strong and significant positive correlations between fire occurrence and related social media activity, while others show no correlation or have too few observations to calculate the correlation. No states showed a significant negative correlation. Those states that have the highest frequency of fire occurrence in our dataset (California, Idaho, Montana, Oregon, Washington), which all fall in the fire-prone West/North-West region, also show strong correlations with fire-related activity on both Instagram and Twitter. Other states with less fire activity typically show weaker correlations and/or only have significant correlations for one of the social media platforms. At the scale of US counties, we find that 54 out of 336 (Twitter) and 64 out of 298 (Instagram) counties where both fire and social media data was recorded show significant positive correlations between social media (either Instagram or Twitter) and fire occurrence (MODIS); these are shown in Figure 2. We could not calculate correlations for the remaining counties because they lacked social media posts and/or fire occurrences in our datasets. No counties showed a significant negative correlation. Overall we find that the correlation between social media activity and fire occurrence is strongest in fire-prone states/counties.

We next examine the spatial localisation of social media activity about wildfires in relation to the positions of actual wildfires, using the California-Nevada region (126-114ºW, 33-42ºN) and a comparison between Twitter and MODIS (April-Sept 2015) as an example. We measure the distance (km) between each tweet and the closest recorded fire event on the same day (see Data Collection and Methods). The distribution of observed closest distances across the whole study period is shown in Figure 3. This distribution shows that most tweets about wildfire originate from locations close to current wildfires, but some tweets are from distant locations; this is consistent with a model whereby tweets about wildfire events are most often made by users who are potentially affected by them, but where

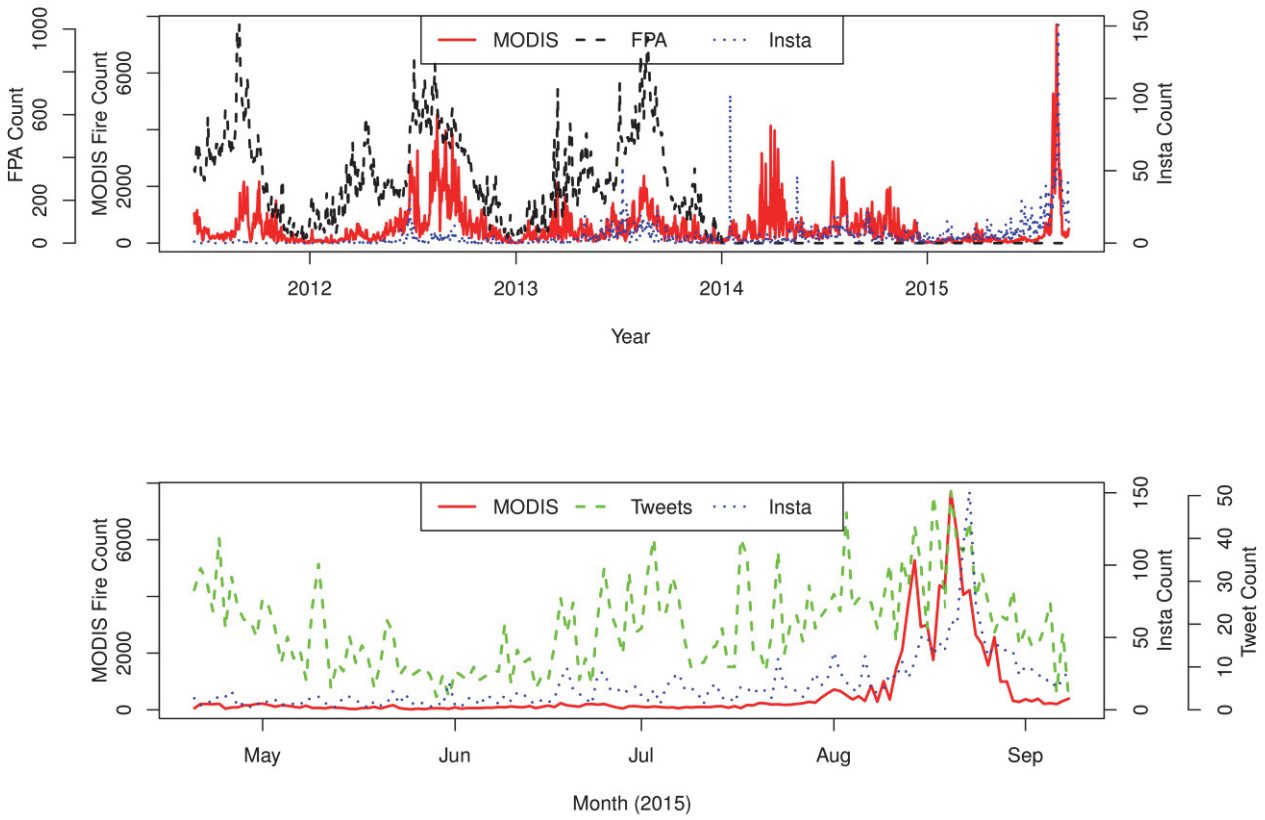

Figure 1: Time series of social media activity and fire occurrence for the whole US over various time periods. Top: Instagram and MODIS over period 2011-2015, FPA over period 2011-2013. Bottom: Instagram, Twitter and MODIS over period April-Sept 2015.

State	Instagram	Twitter	MODIS	Instagram-MODIS		Twitter-MODIS		Instagram-Twitter	
				r	p	r	p	r	p
Arizona	72	106	978	0.195	0.020	0.194	0.021	0.292	<5e-4
California	778	678	12450	0.522	<3e-11	0.381	<3e-6	0.550	<2e-12
Colorado	53	301	107	0.305	<3e-4	0.067	0.426	-0.006	0.943
Florida	39	71	1759	0.249	<3e-3	0.099	0.240	0.102	0.023
Idaho	108	92	11274	0.683	<2e-16	0.636	<2e-16	0.524	<3e-11
Maine	1	3	16	-0.021	0.801	0.182	0.030	-0.012	0.884
Massachusetts	6	34	31	-0.016	0.849	0.191	0.023	-0.026	0.756
Montana	165	112	4780	0.623	<2e-16	0.483	<2e-9	0.587	<2e-14
New Jersey	14	11	75	0.220	<9e-3	-0.055	0.517	-0.007	0.930
New Mexico	0	17	177	-	-	0.305	<3e-4	-	-
Oregon	317	132	10139	0.367	<8e-6	0.610	<9e-16	0.707	<2e-16
South Dakota	2	4	54	0.303	<3e-4	0.060	0.471	0.341	<4e-5
Texas	39	160	1959	-0.063	0.459	0.279	<8e-4	0.007	0.930
Utah	44	12	161	0.303	<3e-4	-0.030	0.723	0.165	<5e-9
Washington	387	224	25688	0.707	<2e-16	0.379	<4e-6	0.467	0.003
Wisconsin	10	26	238	0.242	<4e-3	0.084	0.321	0.180	0.504

Table 2: US States showing significant positive correlation ($p<0.05$) between social media post frequency and fire occurrence (April-Sept 2015). There were no states showing significant negative correlations.

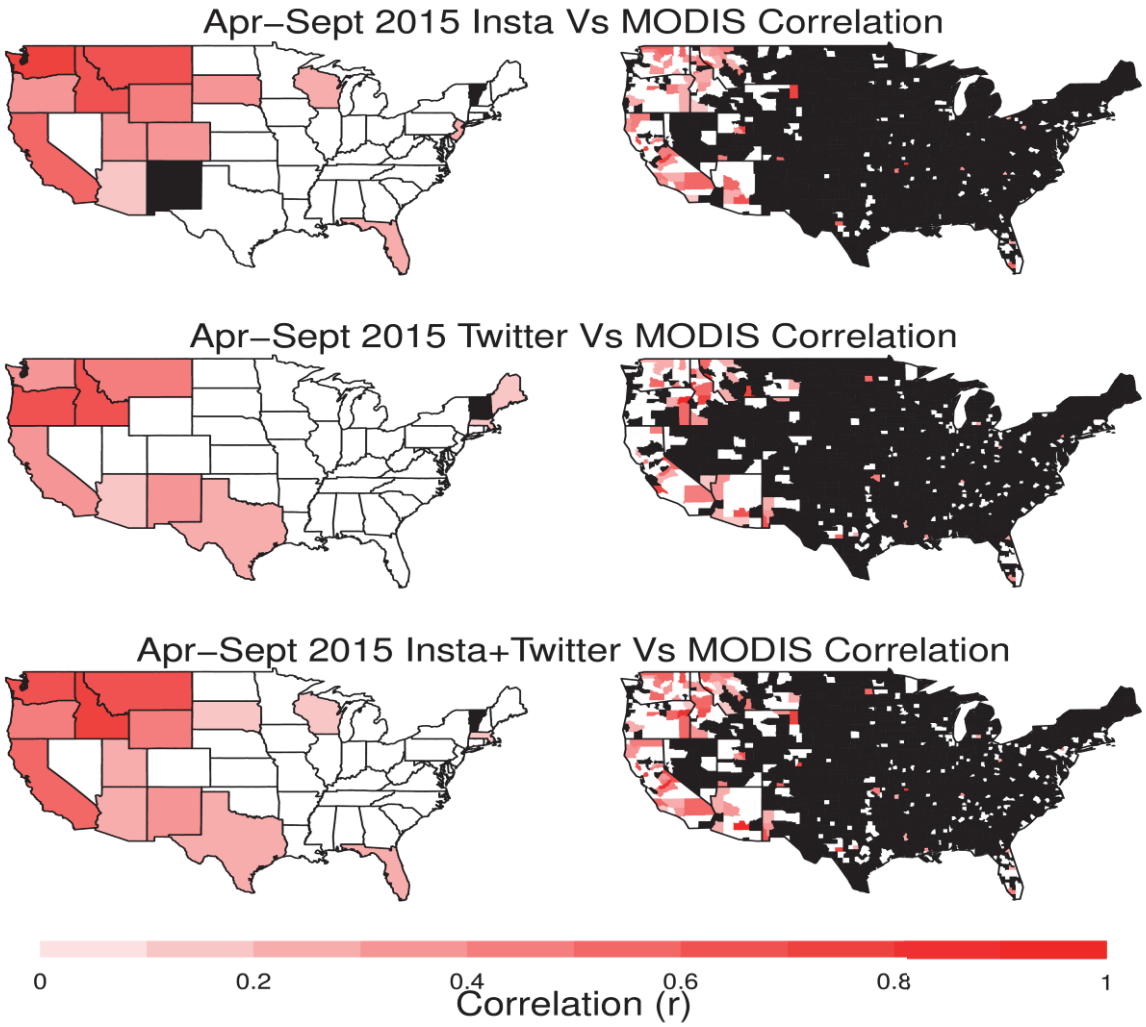

Figure 2: Spatial variation in temporal correlation between social media and fires (April-Sept 2015). Colour scale shows the strength of significant correlations. White indicates a correlation that is not significant ($p>0.05$). Black indicates a lack of data (zero social media posts and/or zero fire occurrences) such that correlations could not be calculated. Top: Correlation between MODIS fire occurrence and Instagram activity. Middle: Correlation between MODIS fire occurrence and Twitter activity. Bottom: Correlation between MODIS fire occurrence and summed Instagram/Twitter activity.

long-distance tweets sometimes occur due to media reporting of large wildfires and subsequent social media commentary from distant locations.

We confirm the spatial correlation of social media posts with wildfire events by comparing the observed closest distances from posts to fires to distances generated by three null models (Figure 4; see Data Collection and Methods). For Twitter, the mean observed distance to the closest fire is 201.5km and the median distance is 162.4 km, with a standard deviation of 166.8 km. Interestingly, we find that Instagram posts, which require an image to be uploaded, are closer on average than Twitter posts (mean 140.4 km, median 73.7 km and standard deviation 157.5 km). With respect to the null distributions of closest distances derived from the three null models, we find that mean/median observed distances are significantly smaller than expected, for both Instagram and Twitter ($p<0.01$ in all cases). This is perhaps unsurprising for Null Model 1, which re-locates social media posts with uniform probability across the whole study region (including desert areas). However, the strong signal seen with Null Models 2 and 3, which control for social media usage and population density, confirms that social media activity about wildfires is spatially correlated with actual wildfire events.

Our final analysis concerns the ability of social media data to effectively locate wildfire events in the absence of

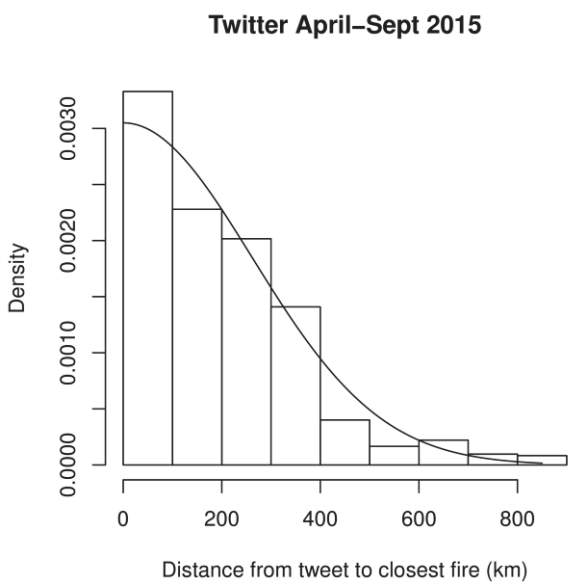

Figure 3: Distribution of distances between Twitter posts and the closest fire to each post on the same day, based on MODIS fire occurrence data, for the period April-Sept 2015.

other information. We created heat maps of Twitter activity in relation to wildfires in the California-Nevada region (see Data Collection and Methods). Example heatmaps are shown in Figure 5. Using a simple heuristic whereby hotspots of Twitter activity predict the locations of wildfires, we find that some fires are well predicted (e.g. hotspots are close to fires) while others are not. These mixed results suggest that this method has potential but needs further refinement.

Discussion

Here we have shown that social media activity about wildfires is both temporally and spatially correlated with wildfire events. These findings establish the preconditions for further research that seeks to derive information on social impacts of wildfires from social media datasets, to augment and complement more conventional observations that focus on the physical aspects, and to support decision making by fire managers and public agencies. Our results show the promise of social media as a sensor for natural hazards and highlight some of the methodological challenges associated with this kind of analysis.

Our methods could be improved in a number of ways. Firstly, here we analysed all data returned by the Twitter and Instagram APIs based on our search terms, meaning that our datasets contain an amount of irrelevant content. Filtering our social media datasets for relevance using some kind of automated classifier is likely to strengthen the observed correlations. In particular, inspection shows that our datasets contain a small number of irrelevant posts that occurred a long distance away from any fires; we retained these in our analysis in order to show the quality of the unfiltered datasets, but note that their presence reduces the apparent spatial correlation between posts and wildfire events. Secondly, detrending the Instagram and Twitter datasets for changes in overall usage of these platforms over time would likely improve the temporal correlations we observed. Instagram in particular has seen a large increase in usage during the 2011-2015 time period studied here, which affects the level of wildfire-related activity we might expect to observe. Third, application of automated methods to infer the locations of non-geotagged social media posts might increase the size of our datasets. While use of estimated locations may carry associated methodological challenges, it is possible that the overall performance of the social media datasets at detecting and locating wildfire events would be improved. We are addressing some of these improvements in ongoing work.

So far our research on the relationship between social media and wildfires has focused on measuring their spatial and temporal correlation. However, if social media is to become a useful data source with which to document and characterise wildfires and other natural hazards, it is necessary to perform more rigorous evaluations. If the challenge is framed as using social media to detect wildfire events, then we might judge performance in terms of precision (detection only of genuine wildfire events and avoidance of Type 1 “false positive” errors) and recall (detection of all wildfire events and avoidance of Type 2 “false negative” errors). While a perfect correlation between wildfire occurrence and social media reporting of wildfire would deliver both precision and recall, imperfect correlation (as reported here) may be driven by failure in either aspect. Our data exhibits days without social media posts where wildfires occurred (false negatives), as well as days where social media posts occur but there are no fires (false positives). These errors may arise from temporal lags (e.g. posts about fires that have ended on the previous day) or from inaccuracies in wildfire observations (e.g. uncertainties in satellite detection of fires or failure to record a wildfire event in a remote area). Our methods here currently do not attempt to measure precision and recall directly, but future development will address this aspect, beginning with development of content-based methods for associating social media posts with particular fire events, beyond simple spatial co-location as used here.

An important factor with any geographical “social sensing” tool using social media data is population density. We note that the California-Nevada region we chose to analyse in more depth shows high variation in population density (containing large Californian cities as well the Nevada desert). When we control for this aspect using null

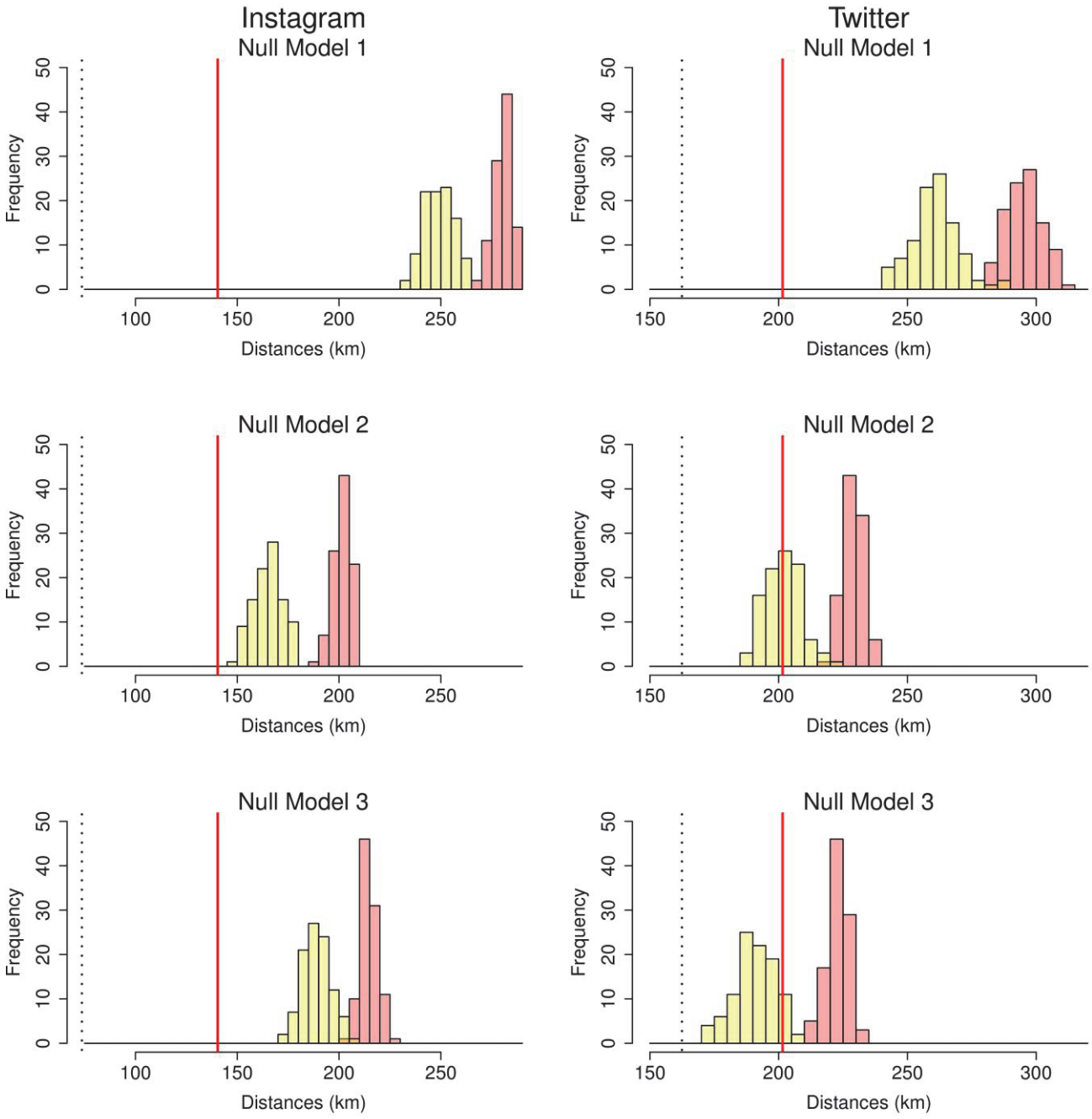

Figure 4: Social media location to closest fire distance mean (red or grey) and median (yellow or white) values are compared to 3 null models (see Data Collection and Methods). Distributions of mean and median distance to closest fire from each null model on each social media dataset are shown as semi-transparent histograms with the values of the mean and median from the actual datasets shown as red/grey and black dotted lines respectively.

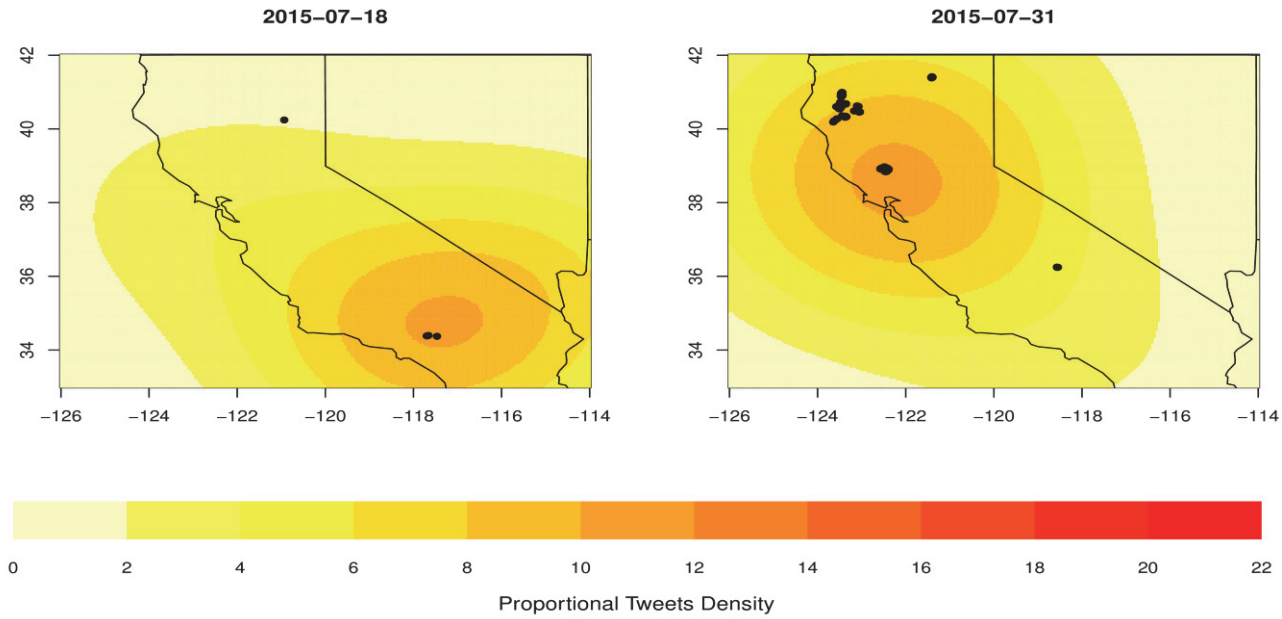

Figure 5: Heat map showing Twitter activity (contour plots) in relation to observed fire events (black points) for the California-Nevada region, using MODIS fire occurrence data. Two examples are shown of typical days in our dataset. Contour plots are normalised such that a single tweet will have a peak height of 1 unit.

models that take into account social media usage and population density, the statistical significance of the results we achieve is affected. While all null models show that social media posts are closer to wildfire events than the null expectation, the variation in "effect size" between different null models suggests that this form of population density bias is likely to affect efficacy of social sensing and needs to be carefully corrected. Other biases may be harder to control.

It seems intuitively likely that social media posts are more likely to mention fires which are bigger and have a larger impact on society, than smaller fires which are far from populated areas. This relationship could be explored by utilising data on wildfire impacts. Fire trackers (such as at http://firetracker.scpr.org/) provide further information on fire events in California, such as how many structures were damaged and how many injuries occurred, and also provide names for larger fires. While such data is not always available, careful historical analysis of available data could reveal which kinds of fires create most social media activity. Names of fires could also provide additional search terms with which to collect social media data. However, it should be remembered that one of the most important properties of social media data for providing information to fire managers and the public lies in its timeliness; historical analysis may help to model the underlying interaction of society with wildfires, but is less useful for planning management actions around ongoing events.

Acknowledgements

The authors were supported by a Research on Changes of Variability and Environmental Risk (RECoVER) grant funded by EPSRC (EP/M008495/1).

References

CIESIN/CIAT. 2005. Gridded population of the world, version 3 (gpwv3): Population density grid, future estimates. Technical report. NASA Socioeconomic Data and Applications Center (SEDAC), Center for International Earth Science Information Network (CIESIN) and Centro Internacional de Agricultura Tropical (CIAT). Date of access: 14th December 2015.

Elgesem, D.; Steskal, L.; and Diakopoulos, N. 2015. Structure and content of the discourse on climate change in the blogosphere: The big picture. *Environmental Communication* 9(2):169–188.

Kirilenko, A. P., and Stepchenkova, S. O. 2014. Public microblogging on climate change: One year of twitter worldwide. *Global Environmental Change* 26:171–182.

Kirilenko, A. P.; Molodtsova, T.; and Stepchenkova, S. O. 2015. People as sensors: Mass media and local temperature influence climate change discussion on twitter. *Global Environmental Change* 30:92–100.

Lazer, D.; Pentland, A.; Adamic, L.; Aral, S.; Barabasi, A.-L.; Brewer, D.; Christakis, N.; Contractor, N.; Fowler, J.; Gutmann, M.; Jebara, T.; King, G.; Macy, M.; Roy, D.; and Van Alstyne, M. 2009. Computational social science. *Science* 323(5915):721–723.

Lu, X.; Bengtsson, L.; and Holme, P. 2012. Predictability of population displacement after the 2010 Haiti earthquake. *Proceedings of the National Academy of Sciences of the United States of America* 109:11576–11581.

Olteanu, A.; Castillo, C.; Diakopoulos, N.; and Aberer, K. 2015. Comparing events coverage in online news and social media: The case of climate change. In Proceedings of the Ninth International AAAI Conference on Web and Social Media, number EPFL CONF-211214.

O'Neill, S.; Williams, H. T.; Kurz, T.; Wiersma, B.; and Boykoff, M. 2015. Dominant frames in legacy and social media coverage of the IPCC Fifth Assessment Report. *Nature Climate Change* 5(4):380–385.

Sakaki, T.; Okazaki, M.; and Matsuo, Y. 2010. Earthquake shakes Twitter users: Real-time event detection by social sensors. In Proceedings of the 19th International Conference on World Wide Web (WWW), 851–860. ACM.

Schafer, M. S. 2012. Online communication on climate change and climate politics: A literature review. *Wiley Interdisciplinary Reviews: Climate Change* 3(6):527–543.

Short, K. C. 2015. Spatial wildfire occurrence data for the United States, 1992-2013 [fpa-fod-20150323], 3rd edition. Technical report, Fort Collins, CO. Forest Service Research Data Archive.

Statista. 2015. Global social networks ranked by number of users (March 2015). Technical report. Date of access: 12th June 2015.

USDA Forest Service, R. S. A. C. 2015. MODIS active fire detections for the conus (2015). Technical report, Salt Lake City, USA.

Wesolowski, A.; Eagle, N.; Tatem, A. J.; Smith, D. L.; Noor, A. M.; Snow, R. W.; and Buckee, C. O. 2012. Quantifying the impact of human mobility on Malaria. *Science* 338(6104):267–270.

Wesolowski, A.; Qureshi, T.; Boni, M. F.; Sundsoy, P. R.; Johansson, M. A.; Rasheed, S. B.; Engo-Monsen, K.; and Buckee, C. O. 2015. Impact of human mobility on the emergence of dengue epidemics in Pakistan. *Proceedings of the National Academy of Sciences* 112(38):11887–11892.

Williams, H. T.; McMurray, J. R.; Kurz, T.; and Lambert, F. H. 2015. Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global Environmental Change* 32:126–138.