Enhancing Traffic Incident Detection by Using Spatial Point Pattern Analysis on Social Media

Shen Zhang, Jinjun Tang, Hua Wang, and Yinhai Wang

Expedient incident detection and understanding are important in traffic management and control. Social media as important information venues have immense value for increasing an awareness of traffic incidents. In this paper, an attempt is made to assess the potential of using harvested social media for traffic incident detection. Twitter in Seattle, Washington, was chosen as a representative sample environment for this work. A hybrid mechanism based on latent Dirichlet allocation and document clustering was proposed to model incident-level semantic information, while spatial point pattern analysis was applied to explore the spatial patterns and to assess the spatial dependence between incident-topic tweets and traffic incidents. A global Monte Carlo *K*-test indicated that the incident-topic tweets were significantly clustered at different scales up to 600 m. The nearest neighbor clutter removal method was used to separate feature tweet points from clutter; then a density-based algorithm successfully detected the clusters of tweets posted spatially close to traffic incidents. In multivariate spatial point pattern analysis, *K*-cross functions were investigated with Monte Carlo simulation to characterize and model the spatial dependence, and a positive spatial correlation was inferred between incident-topic tweets and traffic incidents up to 800 m. Finally, the tweet intensity as a function of distance from the nearest traffic incident was estimated, and a log-linear model was summarized. The experiments supported the notion that social media feeds acted as sensors, which allowed enhancing awareness of traffic incidents and their potential disturbances.

Social media have drastically altered the concepts of information contribution, dissemination, and exchange (*1*). Their power in situational awareness and crisis response have also been effectively demonstrated during the past few years (*2, 3*). During recent events such as earthquakes, bushfires, and terrorist attacks, people have been accustomed to sharing their knowledge through openly accessible online social media. By empowering the general public to publish and distribute user-generated content, social media have made virtually every citizen a potential contributor, which is in many ways more convenient, timely, and information-rich than traditional media. In addition to the rich semantic and multimedia content, members of location-based social media services can also be seen as "semantic sensors" with the ability to report and describe incidents by sending messages along with geographic footprints (*4*). They therefore provide emerging opportunities for assessing the feasibility of harvesting semantic and spatial information from social media feeds for traffic incident detection and analysis. In this study, Twitter is considered in particular as a representative sample social media environment; it has more than 200 million monthly active users, while an average of 500 million tweets are sent daily.

Although social media surpass many of the limitations in traditional sensors, such as area coverage and the costs of installation and maintenance, there are still several challenging issues that need to be addressed. First, the detection of traffic incidents can be challenged by possible large volumes of irrelevant tweets and a much lower number of relevant ones. For that reason, most current studies gather tweets for target events by identifying keywords but without convincing evidence for sufficiency and the necessity of these event-term candidates. In addition, the information obtained from social media may be unstructured, which hampers modeling incident-related semantic information. Moreover, unlike typical traffic sensors that always collect specific types of measurements, the novel semantic sensors may operate in a wide range of the sociocultural spectrum. For example, a tweet with the content "@dummy: ??? I'm trippen hard about my phone right now. There is an 11 car collision on 1-5 northbound, and I can hardly back home for dinner on time." has multiple topics across phone, traffic, and dinner. Therefore, the traditional document clustering can hardly deal with the semantic retrieval of topic-specific tweets. Second, a tweet can also be seen as a realization of a spatial stochastic process, since it comprises semantic features as well as spatial coordinates. Quantifying the spatial pattern of a tweet and revealing its dependence on incident position are fundamental to understanding tweet dynamics and detecting incidents. However, few studies have used spatial statistical analysis to describe or to explain the underlying spatial mechanisms of tweets (e.g., tweets are often hypothesized as having clustered distributions without any formal testing of their spatial patterns).

To address these challenges, a hybrid semantic filter based on latent Dirichlet allocation (LDA) and document clustering is first proposed; this method allows for an ad hoc analysis of incident-topic tweets concerning the topic distribution (*5*). As the statistical method for the spatial point process is rather well developed, tweets in the form of the realization of a spatial point process in a finite spatial region are then considered (*6, 7*). As a result, the spatial process of tweets could be analyzed to make inferences about the spatial distribution, which implies the random, dispersion, or attraction pattern. Since the ultimate objectives of the analysis are

S. Zhang, J. Tang, and H. Wang, School of Transportation Science and Engineering, Harbin Institute of Technology, No. 73, Huang-He Street, Harbin 150090, China. Y. Wang, Department of Civil and Environmental Engineering, University of Washington, Seattle, WA 98195-2700. Corresponding author: Y. Wang, yinhai@u.washington.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2528, Transportation Research Board, Washington, D.C., 2015, pp. 69–77.
DOI: 10.3141/2528-08

to understand and to model the potential spatial association between incident-topic tweets and traffic incidents, a distance-based algorithm is further applied to distinguish between features and clutter in the tweet point process. Moreover, the feature points are clustered by means of a density-based algorithm to detect traffic incidents. Finally, a multivariate point pattern analysis is applied to assess the scale and extent of tweet aggregation affected by the spatial distribution of incidents (*8*).

RELATED WORK

Social Media for Event Detection

Social media feeds have been used quite extensively for event detection. Notable applications addressed the early detection of contagious disease outbreaks by monitoring influenza-related blogging trends during the emergence of the U.S. 2008 flu season (*9*); the analysis of the structure of social media networks can improve the prediction of the spread of these outbreaks (*10*). Influenza epidemic outbreaks were also detected in the same application domain of bioinformatics (*11*). Earle et al. attempted to assess how fast tweeters reacted to the small and localized earthquake of Morgan Hill, California, in March 2009 (*12*), while Crooks et al. also tried to identify spatial characteristics of this information dissemination avenue (*2*).

As stated above, most current approaches to event detection in social media streams focus on trending topic detection, in which the event to be detected is on a large, often global scale and receives widespread coverage on social media sites (*13*). However the focus here is on a novel scenario concerned with detecting and understanding traffic incidents, many of which are of a fairly small scale and thus are often covered by only a few tweets.

Spatial Point Pattern Analysis

Spatial point pattern analysis first became commonplace in plant ecology, where it is widely used for detecting spatial patterns of species distribution, understanding interactions between plants and the environment, and inferring important ecological processes or mechanisms of plant population dynamics (*14, 15*). In the past few years, researchers further applied first- and second-order analyses of spatial point patterns to study the spatial structure of forest fires in Florida and Missouri (*16, 17*); the characteristics of preincident, postincident, and nonincident traffic conditions on freeways were also investigated with second-order analysis (*18*). However, the literature review undertaken for this paper did not find this theory being applied to explore the spatial structure of tweets.

A spatial point pattern can be described by first-order statistics that include the intensity of a point process, its large-scale variation, and nearest neighbor distances. Second-order statistics include neighborhood characteristics, such as Ripley's *K*-function, that provide information on the correlation of points over a range of spatial scales, which allows for the detection of randomness, clustering, or dispersion at any given scale. Furthermore, recently developed methods allow for a more comprehensive analysis of point patterns; these methods additionally consider covariates that explain their spatial distribution (*7*). In the present study, full advantage is taken of these recent developments for the analysis of potential spatial associations between incident-topic tweets and traffic incidents.

METHODOLOGY

Integrating Topic Modeling and Document Clustering for Semantic Filter

In classic document clustering approaches, documents are usually represented with a bag-of-words model, which is based purely on raw terms and is insufficient to capture all semantics. Topic models are able to put words with similar semantics into the same group called a "topic," in which synonymous words are treated as the same. Therefore, integration of tweet topic modeling and clustering can make them mutually promote each other to achieve better performance for incident-related tweet retrieval and filtering. In this study, a hybrid process is designed to first use a topic model–LDA to project tweets into a topic space, reducing the noise of similarity measure, and then perform clustering algorithms in the topic space to obtain topic clusters. Since traditional algorithms, such as hierarchical clustering and *K*-means, are commonly used for document clustering, here the LDA model will only be formulated and discussed.

Latent Dirichlet Allocation

LDA has been demonstrated as a highly effective unsupervised probabilistic method for finding distinct topics in tweets and various other collections of documents. An advantage of LDA is that without any prior knowledge, it can identify a topic, which is characterized by key words that may never appear next to each other in the same document.

Therefore, LDA was applied to model lane-blocking topic and detected various tweet topics in this study. In LDA, the corpus contains tweets $T_i = (\omega_{i1}, \dots, \omega_{in})$ of length n_i. Each word ω_{ij} comes from a vocabulary that consists of V terms. The term distribution for each topic is modeled by

$$\beta_i \sim \text{Dirichlet}(\eta)$$

where Dirichlet(η) denotes the Dirichlet distribution for parameter η.

The proportion of topic distribution for each document is distributed as

$$\omega_i \sim \text{Dirichlet}(\alpha)$$

Each word ω_{ij} is associated to a topic z_{ij} that follows

$$z_{ij} \sim \text{multinomial}(\omega_i)$$

where multinomial (ω_i) denotes the multinomial distribution with one trial (*5*).

In brief, specifying the LDA model consists of three steps: (*a*) draw *K* topics from a symmetric Dirichlet distribution, (*b*) for each tweet *t*, draw topic proportions from a symmetric Dirichlet distribution, and (*c*) for each word *n* in each tweet *t*: (*c*i) draw a topic assignment from the topic proportions and (*c*ii) draw the word from a multinomial probability distribution conditioned on the topic.

Spatial Pattern Analysis

A spatial point process X is a stochastic model governing the locations of events $\{x_i\}_{i=1\dots n}$ in a bounded region $A \subset R^2$ with area $|A| > 0$. In practice, A is usually built as an observed window (sampling window). In this work, events are mainly tweet points, while traffic incidents will be modeled as a spatial point process in multivariate spatial point pattern analysis. The method is outlined briefly as described in the subsections below.

Complete Spatial Randomness Test

The first step in analyzing the tweet spatial pattern is to test the complete spatial randomness (CSR) hypothesis. The null hypothesis indicates the absence of spatial interaction between tweet points, while a departure from CSR indicates that a regular or clustered spatial pattern exists among the tweets. It is possible to detect incidents only within a clustered tweet spatial process. Therefore, in the present study, a Monte Carlo test is applied to evaluate the CSR hypothesis based on the K-function, which can characterize the second-order structure of a spatial point process. For a stationary tweet point process the test is defined as

$$K(r) = \lambda^{-1} E(n \text{ of tweets within distance } r \text{ of arbitrary tweet})$$

and is estimated by

$$\hat{K}(r) = \left(n(n-1)\right)^{-1} |A| \sum_{x_i \in X \cap A} \sum_{x_j \in (X \cap A) \setminus \{x_i\}} I\left(\|x_i - x_j\| \le r\right) w_{ij}^{-1} \tag{1}$$

where w_{ij} is Ripley's edge correction factor, x_i and x_j are pairs of spatial points i and j, and $\hat{\lambda} = (n-1)/|A|$; for a homogeneous Poisson process in R^2, $K(r) = \pi r^2$.

The Monte Carlo K-test is implemented as follows: (a) the K-function of the observed pattern $\hat{K}_1(r)$ is estimated; (b) $s-1$ homogeneous Poisson processes of the same size as the observed pattern are simulated and K-functions are estimated; (c) the upper and lower envelopes of the simulations are defined as $U(r) = \max_{i=2,\dots,s}\{\hat{K}_i(r)\}$, $L(r) = \min_{i=2,\dots,s}\{\hat{K}_i(r)\}$; and ($d$) the clustered or regular spatial pattern is identified, depending on whether values of the empirical K-function are higher than the upper envelope or lower than the lower envelope.

Feature Extraction for Spatial Filter

Considering the limited impact distance of most lane-blocking incidents, sparse tweet spatial points are of little use for incident detection. The nearest neighbor clutter removal algorithm is adopted to extract tweet feature points from the clutter (*19*). The algorithm calculates for each point i of a spatial pattern the two-dimensional Euclidean distance to its Kth nearest neighbors (Kth-NN), while clutter and tweet feature points are considered as two distinct groups distributed as a Poisson point process with feature points superimposed over the clutter and scattered along the study area. Under this assumption, for a set of n random points, the distribution of the distances D_K to their Kth-NN becomes highly bimodal, displaying a strong separation between feature and clutter groups. For each given Kth-NN, the maximum likelihood estimator of the intensity ($\hat{\lambda}$) equals the number of events ($\sim K$) falling inside a circle of radius r_i around a randomly chosen tweet i of the point process, divided by the area of relevance

$$\hat{\lambda} = \frac{K}{\pi \sum_{i=1}^{n} r_i^2} \tag{2}$$

The expectation maximization algorithm can be applied to fit a mixture distribution (the feature and the clutter) to the nearest neighbor distances (D_k), which is used to classify each point as belonging to the class feature or clutter. The algorithm estimates the intensities of mixture distribution and the probability of a tweet point being assigned to the class feature.

Impact Analysis Between Incidents and Tweets

It is important to have a deep insight into the incident impact on the spatial pattern of incident-topic tweets. Therefore, a multitype point process test based on the K-cross function, which is an extension of the K-function, is proposed for this study. Given a multivariate process with constant intensity for tweets and incidents, the K-cross function in an observed area in region A is given by the expression (*8*)

$$\hat{K}_{ij}(r) = \frac{1}{\hat{\lambda}_i \hat{\lambda}_j |A|} \sum_i \sum_j w_{x_i x_j}^{-1} I\left(\|x_i - x_j\| \le r\right) \tag{3}$$

where $w_{x_i x_j}$ is the edge effect correction and $\hat{\lambda}_i$ and $\hat{\lambda}_j$ are the empirical estimators of the intensity for the tweet and incident point process. Both tweet and incident point processes are homogeneous Poisson processes. $\hat{\lambda}_i$ is the maximum likelihood estimator for the tweet point process, while $\hat{\lambda}_j$ is the maximum likelihood estimator for the incident point process.

A Monte Carlo K-cross test is implemented in a manner similar to that described in the subsection on LDA. First, the marginal intensities $\hat{\lambda}_i$ and $\hat{\lambda}_j$ are estimated, and the incident and tweet point processes are combined to generate a single process. Then this procedure is repeated $s-1$ times, and $\hat{K}_{ij,k}(r)$ is computed for each simulation. The maximum and minimum values of $\hat{K}_{ij,k}(r)$, $k = 2, \dots, s$ are the upper and lower envelopes, respectively. If $\hat{K}_{ij}(r)$ is greater than the upper envelope or lower than the lower envelope for the observed process, a respective attraction or inhibition will be detected between the incidents and tweets.

EXPERIMENT DESIGN

Data

To illustrate the approach, three sources of data are involved in the experiment: the Washington State Incident Tracking System (WITS), a Seattle Police Department 911 Incident Response data set, and an open Twitter stream.

The information pertaining to the traffic incidents collected by the Washington State Department of Transportation Incident Response team is stored in WITS, in which the incident responders usually make notes and leave a number of details, such as the location, duration, and artificial comments. Since the incident comments are similar to tweets in text length and idiomatic vocabulary, the incident comments from between 2003 and 2012 are chosen to model the incident-related semantic topic in this study. Although incidents are commonly classified into seven main categories with blocking or not blocking symbols in WITS, only the lane-blocking incidents are included in the present experiment (*20*).

The 911 incident response data set includes all the police responses to 911 calls in the city and is refreshed periodically every 4 h. In this data set the transportation-related incidents can also be divided into seven categories. However, to be consistent with the data source of WITS, only the blocking incident records with the code "415" between April and June of 2014 are chosen for validation analysis.

Since the Twitter platform provides direct access to the public live stream but restricts access to its history data in the last few days, this study also collected tweets tagged with Geo coordinates falling within the city limits of Seattle between April and June in 2014. There are two Twitter sources that commonly report traffic incidents in Seattle. WITS can be treated as an official source of information provided by the Washington State Department of Transportation or traffic service companies (e.g., wsdot_traffic, WaTransITTest, and TotalTrafficSPO). Information from these sources is usually gathered by the traditional sensors and posted as tweets by desktop computers and, therefore, provide few useful on-site data items. The Seattle Police Department 911 Incident Response data set is shared by individual users. These tweets may come directly from on-site users of the traffic network, meeting the research objectives well. Retweets are also filtered out for the same purpose. The core of the analytical inspection is fully executed in the R 3.1.0 environment for statistical computing in the study.

Topic Modeling and Clustering

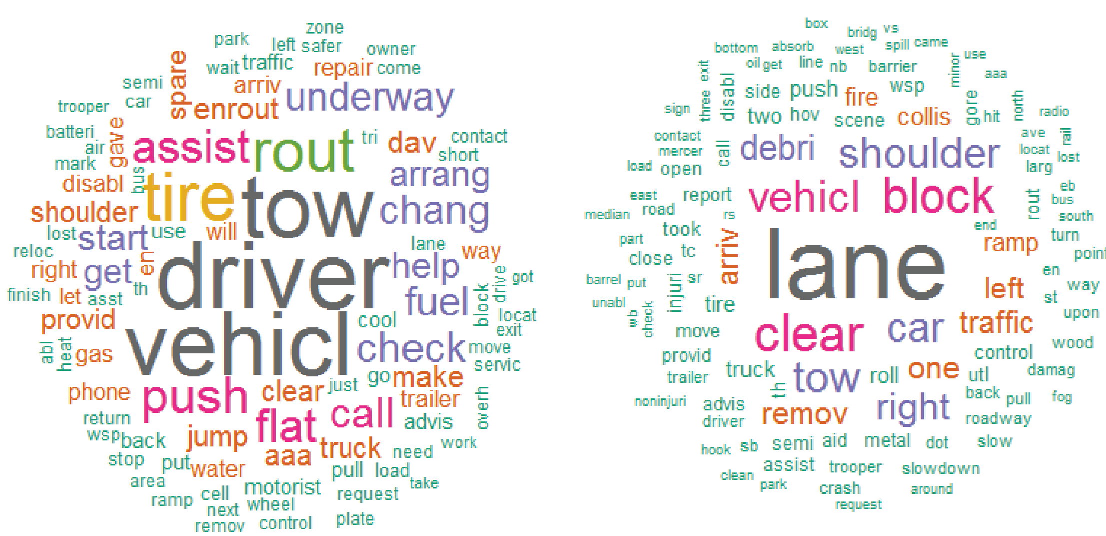
To filter out the irrelevant tweets, two corpora will be established for the experiment. One is the incident corpus based on the comments from the WITS database, and the other is the tweet corpus. These corpora are first processed with a couple of transformations, including changing letters to lowercase and removing punctuation, numbers, and stop words. Then the words are stemmed to retrieve their radicals for the purpose of topic modeling. LDA models are generated to decompose these corpora into their salient topics and determine the specific distributions over the tokens for each topic and distribution of topics over each comment or tweet. In the above process the optimal number of topics needs to be decided to fit the LDA model in advance.

For the incident corpus, the log likelihood for the LDA model with between two and 50 topics is calculated to identify the optimum number of topics from the WITS database. The model with the highest log likelihood value indicates the optimum value that is the best fit for the data. In this experiment, the incident comments seem concise while only two topics are identified.

Figure 1 shows the top-ranked tokens associated with each of the two topics. Since a topic is defined as a probability distribution over all words in the corpus that capture the salient themes that run through the corpus, these topics provide a characterization of the traffic incidents identified by the token frequency and association analysis. Topic 1 (Figure 1*a*) clearly shows that "vehicl," "driver," and "tow" are the top word stems, which embody information about incident response, while some secondary terms such as "call," "check" "assist," and "push" further clarify this topic. The right panel illustrates representative word stems, such as "lane," "vehicl," and "block," indicating that Topic 2 (Figure 1*b*) covers "lane-blocking" incidents. In summary, these two topics together constitute a WITS incident report involving situational awareness and incident response. As a result, both topics will be used for the semantic filter in the following study.

In contrast, the topics in the tweet corpus in Seattle are diverse and uncertain, while the optimal number of tweet topics ranges from 18 to 29. A reason to explain this phenomenon could be that the tweet topics can be generally classified into "persistent" topics and "temporary" topics; the former cover daily life, while the latter topics include short-term cases, for example, the Seattle campus shooting and the 2014 World Cup. To simplify the topic clustering for the 2-month tweet data set, an average of 23 tweet topics are selected for the following experiments.

Each tweet in the corpus is represented as a probability distribution over some of the topics, so a topic space with a total of 25 topics (plus two topics from WITS) is established. A hierarchical clustering algorithm is then adopted to filter out the irrelevant tweets. Hierarchical clustering first assigns each topic to its own cluster. Then, at each iteration, the two most similar clusters are joined in a single cluster. Distances between topic clusters are then recomputed, and further iterations of cluster joins are executed. A threshold is also defined in advance to identify the tweet topics that are most similar to incident topics. For the sake of space, the clustering results are not presented in this paper.



(a) (b)

FIGURE 1 Word clouds of two incident topics from WITS.

Complete Spatial Randomness Test

After the above steps are performed, the incident-topic tweets can be identified as candidates for spatial pattern analysis. Figure 2 depicts the semantic filter results compared with the raw data, and about 95% of semantic irrelevant tweets are filtered out as shown in Figure 2*b*. The theory of spatial point pattern assumes that a point process extends throughout two-dimensional space, but is observed only inside a region *W*, named the "sampling window." In this section, the window is constructed according to the Seattle boundary. The projection of the coordinate reference system is essential for calculating distances between observations, especially necessary for integrating multisource spatial data for the same study area. Therefore, the tweet point process is first converted to the state plane coordinates for Washington State.

Then the Monte Carlo *K*-test, which is demonstrated over a large number ($n = 99$) of simulation envelopes under homogeneity hypotheses, is applied to the tweet point process to test the CSR hypothesis. The results across two random days are shown in Figure 3. All the observed *K*-functions within 2 days remain above the envelope, providing evidence of clustering. The *K*-function is a global cluster indicator; it provides useful information about the distance scale at which clusters take place. These results show that the incident-topic tweets are not randomly distributed over the study area; rather they tend to concentrate in clusters, a finding that indicates these tweet point processes are feasible to be spatially clustered in theory.

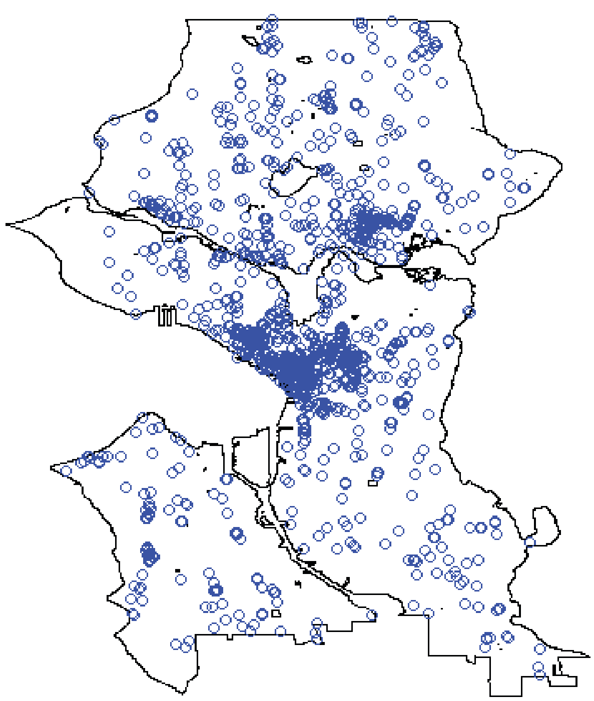
Spatial Filter and Clustering

Since sparse tweet spatial points are of little use for incident detection, the nearest neighbor clutter removal algorithm is applied to distinguish between the noise (i.e., clutter) and the feature points in the tweet point process. Intuitively, the tweet points inside regions of higher density (the downtown and university area) have a smaller *K*th-NN distance than points inside regions of lower density (Figure 4). Therefore a distance-based spatial clustering algorithm can be applied to extract target features, which are helpful in providing a reliable detection result with high precision.

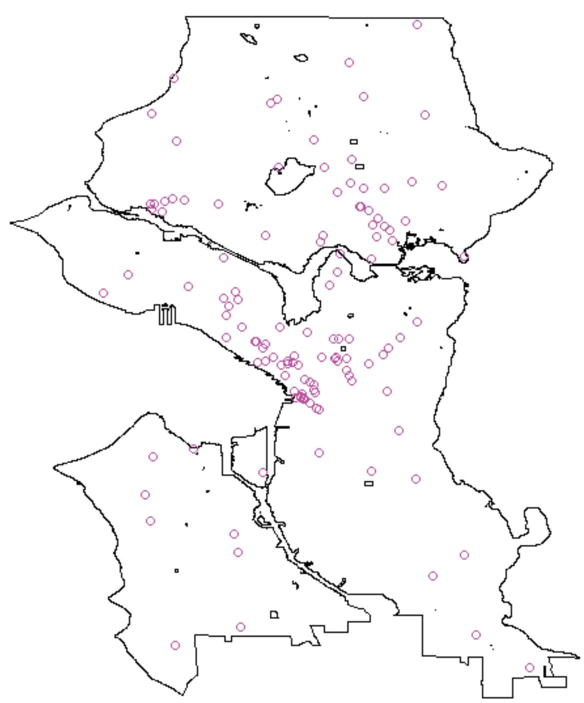
The probability for each tweet point is computed with the expectation-maximization algorithm, while the points with a probability less than a threshold are recognized as noise. The result of the nearest neighbor clutter removal algorithm is shown in Figure 4, which shows intense clusters of tweets against the background "noise."

Then the DBSCAN algorithm, which has been widely used for density-based clustering in spatial databases, is proposed to verify the feasibility of incident detection with a spatial distribution of tweets. This algorithm allowed labeling the more dense aggregations of tweet feature points as belonging to a single traffic incident and removing the residual noise features. For each cluster, the neighborhood of a given radius (eps) has to contain at least a minimum number of points (MinPts). These clusters resulting from this unsupervised algorithm strongly depend on the parameters MinPts and eps, so this method cannot be completely unsupervised. In this study, the Seattle downtown area, which contains feature points with the highest density, is chosen as a representative sample social media environment for the following studies. Figure 5 illustrates the clustering results after the DBSCAN algorithm is run on the data set, with eps = 400 m and MinPts = 5, which means that all detected incidents will contain more than five tweets within a 400-m radius when this algorithm is used.

In practice it works well and provides good results. For example, a total of two clusters are found on May 22, 2014 (Figure 5*b*), indicating there are two possible areas where traffic incidents have taken



(a)



(b)

FIGURE 2 Semantic filter results compared with raw data of May 16, 2014: (*a*) spatial distribution of tweets and (*b*) semantic filter results.

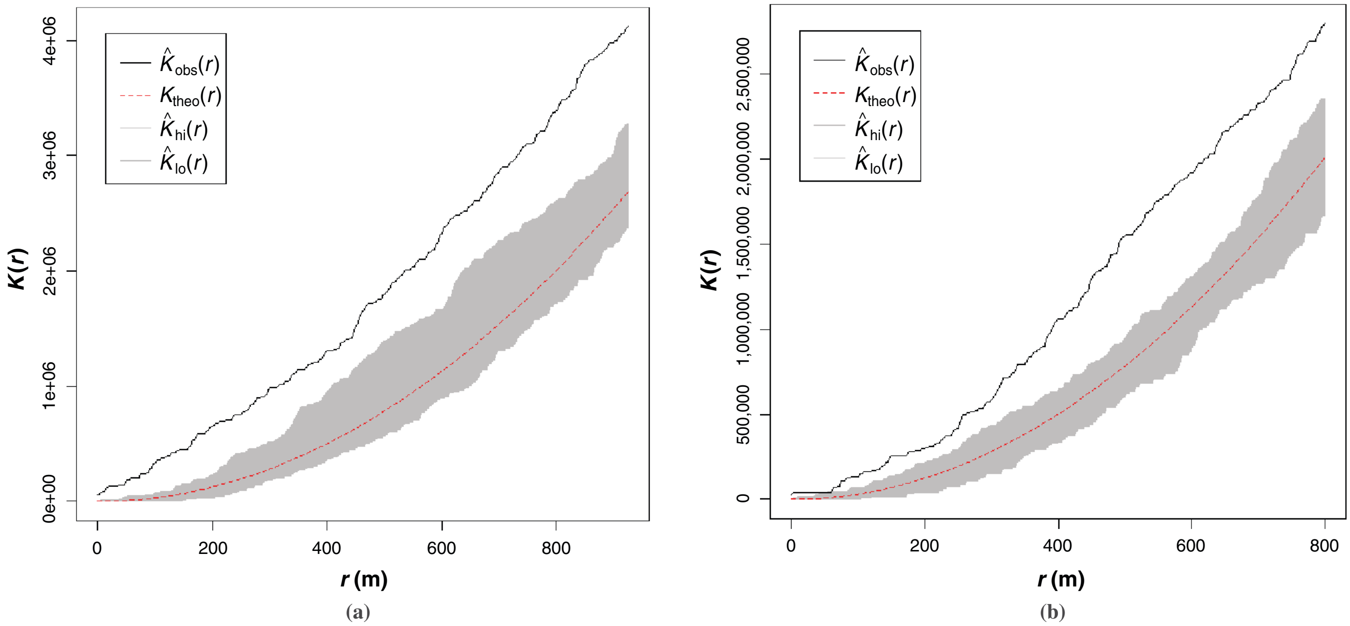


FIGURE 3 *K*-test for incident-topic tweets across two random days: (*a*) May 16, 2014, and (*b*) May 22, 2014 ($\hat{K}_{obs}$ = observed *K*-function; K_{theo} = theoretical *K*-function; $\hat{K}_{hi}$ = upper envelope; and $\hat{K}_{lo}$ = lower envelope).

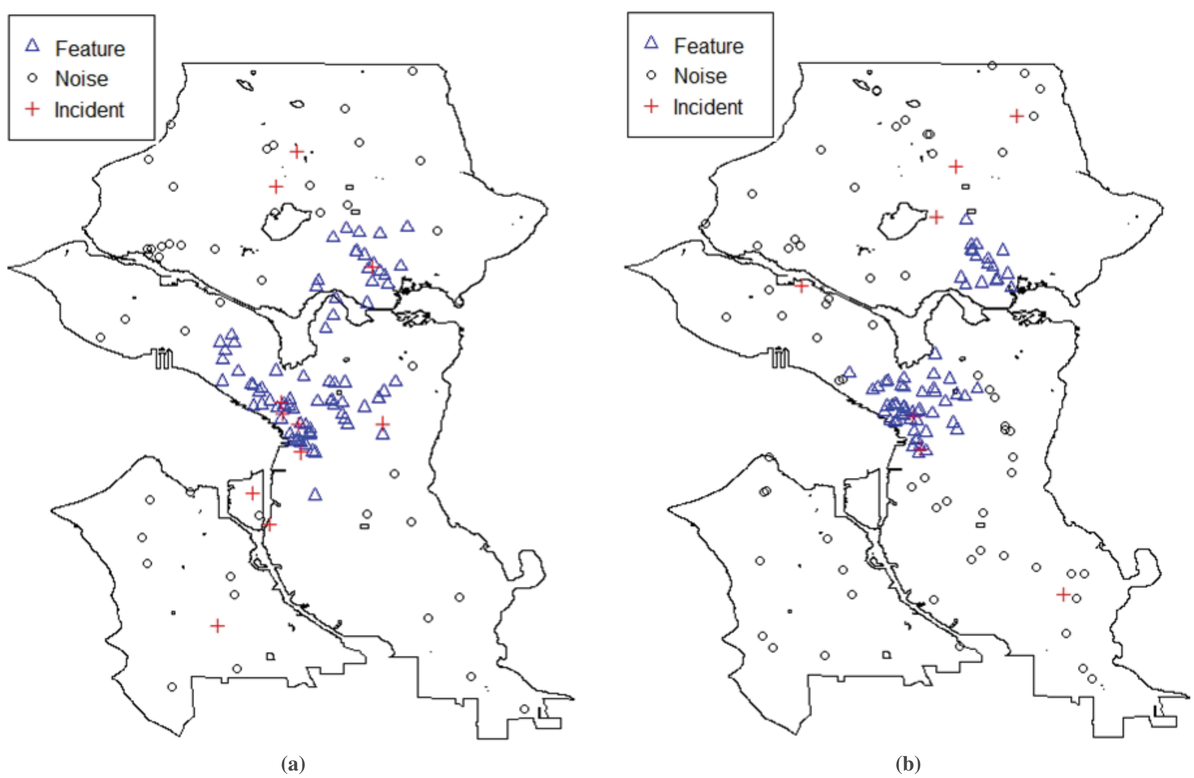
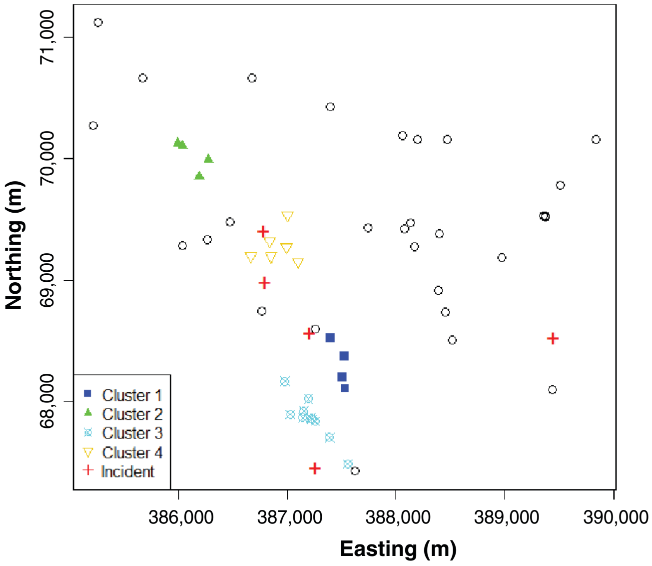


FIGURE 4 Features and clutter of tweet point process: (*a*) May 16, 2014, and (*b*) May 22, 2014.



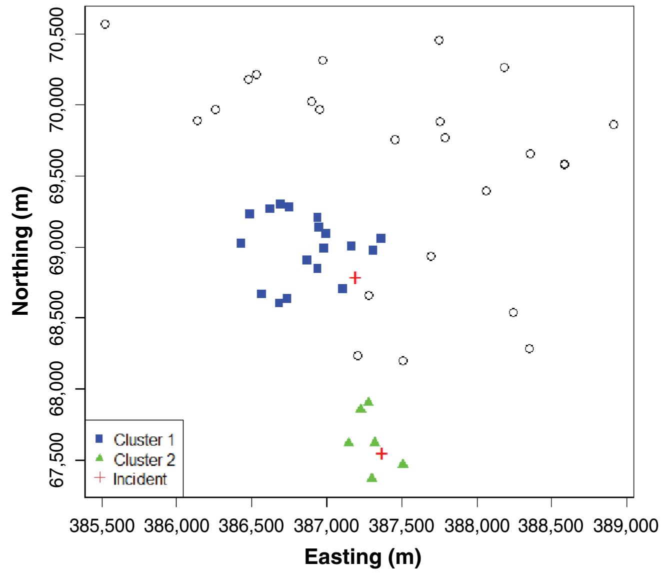


FIGURE 5 Spatial clustering results for traffic incident detection in Seattle downtown area: (*a*) May 16, 2014, and (*b*) May 22, 2014.

place. Most traffic incidents are located within tweet point clusters except one incident that occurred on May 16, 2014 (Figure 5*a*), and some of the incidents can even be found near the cluster center. The clustering results prove that the algorithm is useful for classifying each tweet point as belonging to a single traffic incident, and so it was possible to identify the approximate positions of the incidents. Figure 5 also shows little spatial overlap between traffic incidents in a 1-day interval, so a more complicated temporal–spatial clustering method was not adopted, rather the basic DBSCAN algorithm was used. It is efficient and of sufficient capacity for the initial purpose. Future work will focus on the development of a new temporal–spatial clustering method for the characterization of these incident areas in space and time. Moreover, there are still some undetected incidents or some tweet clusters without immediate use because some tweets are posted by on-site users who do not call 911 and certain incidents have too few eye witnesses for there to be shared messages. Even though, such social sensors still show their potential use for traffic incident detection in areas of high Twitter population density.

In the 2-month data set, more than 90% of the clusters consisted of fewer than 15 feature points. Their frequency distribution covered a range between five and 84 points (tweets), with most of the traffic incidents composed of five to 10 feature points. That finding means most traffic incidents were reported by relatively few Twitter users. Therefore, more efficient strategies should be developed to encourage citizens to report and describe traffic-related events in their region by sending messages along with geographic footprints.

Multivariate Pattern Analysis

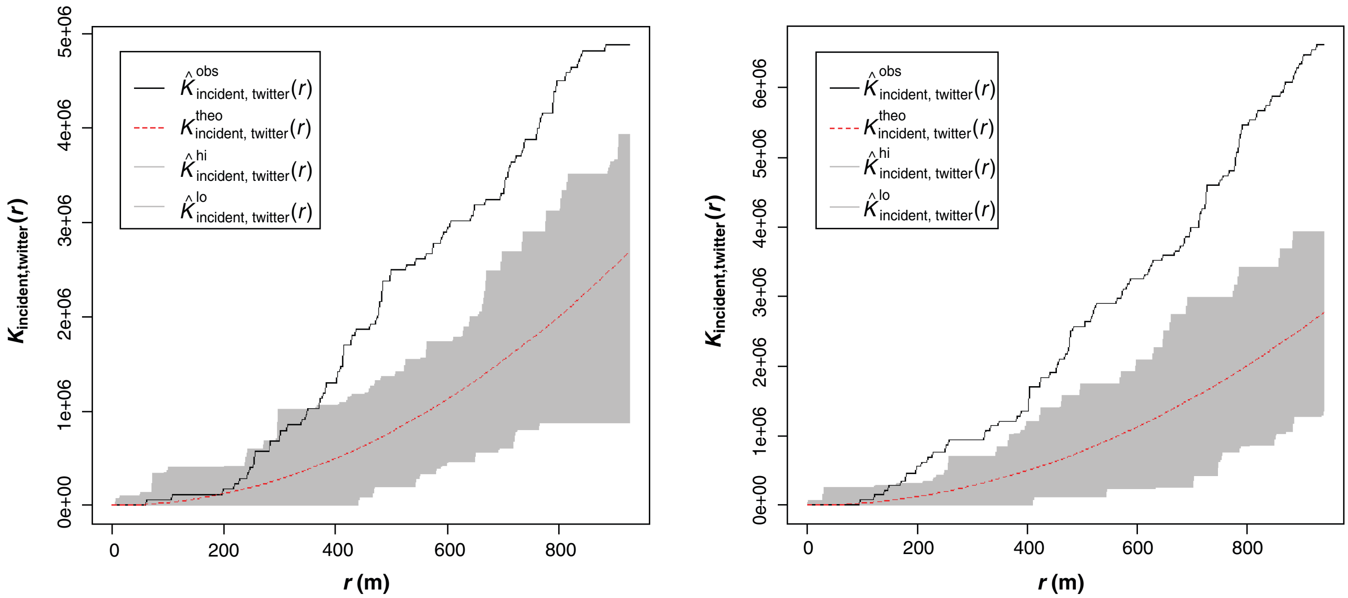
In the multivariate point pattern analysis, the spatial dependence pattern was examined by using *K*-cross function. The *K*-cross function returns $1/\lambda$ times the number of tweets within a distance r of a traffic incident. Figure 6 shows the results of the Monte Carlo tests based on the *K*-cross function for spatial dependence between incident-topic tweets and traffic incidents in the Seattle downtown area. In Figure 6, the observed *K*-cross values are above the upper envelope for most r-values and up to 800 m, evidencing positive dependence between tweets and traffic incidents. Moreover, the tweets can be observed to be clustered around incidents; that observation also provides a practical proof for the experiment in the previous subsection.

It is conjectured that incident-related tweets are more likely to be reported close to the incident. Therefore, at any spatial location u, let $\lambda(u)$ be the intensity of the tweet point process, and the covariate $Z(u)$ equal the distance from u to the nearest traffic incident. Then the function ρ in $\lambda(u) = \rho(Z(u))$ is investigated to determine how the incident-related tweet intensity depends on the distance to incidents. In the present study, kernel smoothing is used to estimate the function ρ with methods of relative distribution, as explained in Jones (*21*).

Figure 7 shows the estimated intensity for the tweet point pattern data as a function of distance from the nearest traffic incident. The estimated standard deviation is a generally decreasing function of distance because the tweet density of the spatial distribution of the distance covariate is decreasing. Figure 7 suggests ρ is approximately an exponential function. If nonparametric estimation is used as a preliminary step before an appropriate parametric model is chosen, then Figure 7 would support the initial choice of a log-linear model such as $\lambda(u) = \exp(\alpha + \beta Z(u))$, where α, β are parameters to be fitted.

CONCLUSIONS AND FUTURE WORK

In this study an attempt was made to assess the feasibility of harvesting semantic and spatial information from social media feeds for traffic incident detection and understanding. The combination of LDA and document clustering allows for semantic filtering of incident-topic tweets concerning the topic distribution. The spatial point pattern analysis is used to investigate the spatial pattern of incident-topic tweets in a case study region in Seattle, where a considerable clustering pattern is observed at different scales up to 600 m. A distance-based spatial clustering algorithm able to extract features from the tweet point

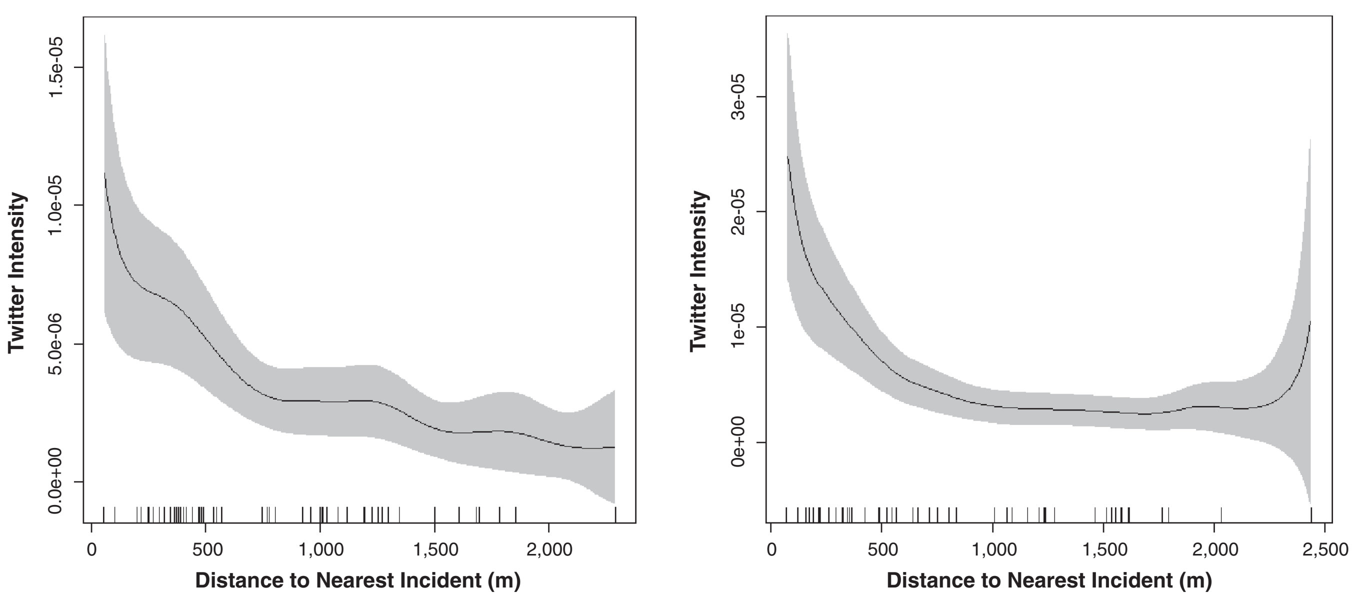


(a) (b)

FIGURE 6 *K*-cross tests for incident-topic tweets and traffic incidents across two random days in Seattle downtown area: (*a*) May 16, 2014, and (*b*) May 22, 2014.

process was also presented, and the Seattle downtown area was chosen as a representative sample environment with feature points of high density, proving that one can reliably detect clusters of tweets posted spatially close to traffic incidents. Furthermore, the multivariate point pattern analysis was applied to measure the dependence relationships between point patterns, and the incident-topic tweets were found to be positively spatially correlated with traffic incidents. Results indicate that the spatial distribution of tweets is clustered mostly at distances up to 800 m around traffic incidents. Further, the tweet point intensity was estimated as a function of distance from the nearest traffic incident, and a log-linear model such as $\lambda(u) = \exp(\alpha + \beta Z(u))$ was inferred; the model indicated a highly significant decrease in the density of tweet points with increasing distance to nearest incidents (<1 km). The result indicates that social media may be further fused with physical sensors spread across the city (e.g., traffic sensors, video cameras, and voice recorders), improving the accuracy of incident detection.

Future research will consider an inhomogeneous model including spatial and temporal pattern analysis to characterize and model the



(a) (b)

FIGURE 7 Estimates of ρ for tweet intensity as function of distance to nearest incident: (*a*) May 16, 2014, and (*b*) May 22, 2014. Solid lines are estimates of ρ. Gray shading indicates ±2 standard deviation (nominally 95% pointwise confidence) intervals.

spatial–temporal variability of tweets; the future work can provide a better understanding of the spatiotemporal dependence between incident-topic tweets and traffic incidents.

ACKNOWLEDGMENTS

This research was supported by the National Natural Science Foundation of China, the National High Technology Research and Development Program of China, and the Natural Science Foundation of Heilongjiang Province, China.

REFERENCES

1. Kaplan, A.M., and M. Haenlein. Users of the World Unite! The Challenges and Opportunities of Social Media. *Business Horizons,* Vol. 53, No. 1, 2010, pp. 59–68.
2. Crooks, A., A. Croitoru, A. Stefanidis, and J. Radzikowski. #Earthquake: Twitter as a Distributed Sensor System. *Transactions in GIS,* Vol. 17, No. 1, 2012, pp. 124–147.
3. Croitoru, A., A. Crooks, J. Radzikowski, and A. Stefanidis. Geosocial Gauge: A System Prototype for Knowledge Discovery from Social Media. *International Journal of Geographical Information Science,* Vol. 27, No. 12, 2013, pp. 2483–2508.
4. Goodchild, M. Citizens as Sensors: The World of Volunteered Geography. *GeoJournal,* Vol. 69, No. 4, 2007, pp. 211–221.
5. Blei, D., A. Ng, and M. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research,* Vol. 3, 2003, pp. 993–1022.
6. Diggle, P.J. *Statistical Analysis of Spatial Point Patterns.* Edward Arnold, London, 2003.
7. Illian, J., A. Penttinen, H. Stoyan, and D. Stoyan. *Statistical Analysis and Modelling of Spatial Point Patterns.* John Wiley and Sons, Chichester, United Kingdom, 2008.
8. Gelfand, A.E., P.J. Diggle, P. Guttorp, and M. Fuentes. *Handbook of Spatial Statistics.* CRC Press, Boca Raton, Fla., 2010.
9. Corley, C.D., D.J. Cook, A.R. Mikler, and K.P. Singh. Text and Structural Data Mining of Influenza Mentions in Web and Social Media. *International Journal of Environmental Research and Public Health,* Vol. 7, 2010, pp. 596–615.
10. Christakis, N.A., and J.H. Fowler. Social Network Sensors for Early Detection of Contagious Outbreaks. *PLOS ONE,* Vol. 5, No. 9, 2010.
11. Ginsberg, J., M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, and L. Brilliant. Detecting Influenza Epidemics Using Search Engine Query Data. *Nature,* Vol. 457, 2009, pp. 1012–1014.
12. Earle, P., M. Guy, R. Buckmaster, C. Ostrum, S. Horvath, and A. Vaughan. OMG Earthquake! Can Twitter Improve Earthquake Response? *Seismological Research Letters,* Vol. 81, No. 2, 2010, pp. 246–251.
13. Walther, M., and M. Kaisser. Geo-Spatial Event Detection in the Twitter Stream. *Advances in Information Retrieval,* Vol. 7814, 2013, pp. 356–367.
14. Arevalo, J.R., and J.M. Fernandez-Palacios. Spatial Patterns of Trees and Juveniles in a Laurel Forest of Tenerife, Canary Islands. *Plant Ecology,* Vol. 165, No. 1, 2003, pp. 1–10.
15. Schurr, F.M., O. Bossdorf, S.J. Milton, and J. Schumacher. Spatial Pattern Formation in Semi-Arid Shrubland: A Priori Predicted Versus Observed Pattern Characteristics. *Plant Ecology,* Vol. 173, No. 2, 2004, pp. 271–282.
16. Yang, J., H.S. He, S.R. Shifley, and E.J. Gustafson. Spatial Patterns of Modern Period Human-Caused Fire Occurrence in the Missouri Ozark Highlands. *Forest Science,* Vol. 53, No. 1, 2007, pp. 1–15.
17. Hering, A.S., C.L. Bell, and M.G. Genton. Modelling Spatio–Temporal Wildfire Ignition Point Patterns. *Environmental and Ecological Statistics,* Vol. 16, No. 2, 2009, pp. 225–250.
18. Ishak, S., and C. Alecsandru. Analysis of Freeway Traffic Incident Conditions by Using Second-Order Spatiotemporal Traffic Performance Measures. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1925,* Transportation Research Board of the National Academies, Washington, D.C., 2005, pp. 20–28.
19. Byers, S., and A.E. Raftery. Nearest-Neighbor Clutter Removal for Estimating Features in Spatial Point Processes. *Journal of the American Statistical Association,* Vol. 93, No. 442, 1998, pp. 577–584.
20. Wang, Y.H., M. Hallenbeck, and P. Cheevarunothai. *Quantifying Incident-Induced Travel Delays on Freeways Using Traffic Sensor Data.* TNW2008-02. Washington State Department of Transportation, Olympia, 2008.
21. Jones, M.C. Kernel Density Estimation for Length-Biased Data. *Biometrika,* Vol. 78, No. 3, 1991, pp. 511–519.

The Standing Committee on Geographic Information Science and Applications peer-reviewed this paper.