Twitter Interactions as a Data Source for Transportation Incidents

Eric Mai*
Berkley Transportation Systems, Inc. An Iteris Company
2150 Shattuck Ave., Suite 200
Berkeley, CA 94704
Phone: (510) 984-1474
Fax: (510) 540-7612
mai@bt-systems.com

Rob Hranac
Berkley Transportation Systems, Inc. An Iteris Company
2150 Shattuck Ave., Suite 200
Berkeley, CA 94704
Phone: (510) 540-4816
Fax: (510) 540-7612
rob@bt-systems.com

For Presentation and Publication:
92nd Annual Meeting
Transportation Research Board
January 2013
Washington, D.C.

Submitted: July 28, 2012

Word count: 5,078 (excluding Figure and Tables)
Plus 6 Figures and 3 Tables: 2,250
TOTAL: 7,328

* Corresponding author
ABSTRACT

Twitter is a microblogging platform that contains a large amount of publically accessible user generated content. This content consists of short social interactions between users. These interactions often describe day-to-day events, and can include location information, making them potentially suitable for use in transportation-related analysis. This paper evaluates the use of data from public social interactions on Twitter as a potential complement to traffic incident data. We compare incident records from the California Highway Patrol with Twitter messages related to roadway events over the same time period. Relationships between the two datasets are evaluated by visualizing the density of incidents and tweets that coincide near the same location. Additionally, the content of Twitter messages is weighted by its relevance to traffic incidents. This weighting is then compared to the time and space proximity of the message to an incident record to determine if more vivid Twitter messages may correspond to the presence of incidents. Twitter information is interesting because it is inexpensive, readily accessible, has broad geographic coverage, and provides a uniquely passenger-centric perspective. It is expected that this research will lead to a better understanding of the potential for information from Twitter to add context to other traffic measurements as a supplemental data source.

Keywords: incident detection, incident analysis, traffic incident, performance measurement, intelligent transportation systems, social media, crowdsourcing,
INTRODUCTION

Understanding the day-to-day operating conditions on a roadway requires monitoring. When a roadway is sufficiently monitored, trends in that roadway’s performance over time can be examined. However, the causes of the roadway’s performance (e.g., a severe weather event) cannot be uncovered with traffic sensors alone. External data gives context to roadway performance measurement. When available, information on things like weather conditions, traffic accidents, or lane closures can be merged with freeway performance data to answer questions from “Why are speeds low in this area?” to “How much delay would a rainstorm at this time and location likely cause?”.

Similarly, richer meaning can be gained on the effects of freeway performance characteristics through the integration of user-focused contextual information. Those traveling in vehicles along the roadway are the system’s users, and their experience is most often evaluated in aggregate metrics such as vehicle-hours of delay. These aggregate measures can only ever represent the experiences of many users across space and time. Roadway performance data could be made richer with the addition of information generated directly by the users themselves. Moreover, the users of a system are an important source for information on that system’s performance; any information they generate about the system can be useful as supplemental data, adding value to existing data collection sources.

Lifecycle of an Incident

To understand how Twitter information may be used in traffic incident analysis, it is important to understand the stages that an incident goes through from the time it occurs, to its detection, to its documentation. In this study, an incident is any traffic event under the jurisdiction of the California Highway Patrol (CHP) which warrants the attention of an officer and results in the generation of an incident report. This includes traffic accidents, congestion events, disturbances on the road, weather events, and lane closures.

The first timepoint in the lifecycle of an incident from roadway to data record is the time when the incident actually occurs. Individual traffic incidents are unpredictable events, however they follow patterns in aggregate over certain time ranges. In California, the total number of incidents exhibits little seasonal variation, as shown in Figure 1. In 2011, most months had roughly 60,000 traffic incidents recorded by the CHP. On the other hand, a clear daily pattern for roadway incidents exists. Weekdays have the most incidents, with many incidents occurring during the AM and PM commuting periods on those days. Friday is the most dangerous day for traffic incidents, particularly during the PM commute. Fewer incidents tend to occur on Saturdays and Sundays, and those that do happen near the middle of the day.

![Monthly CHP Incidents](image)

<table>
<thead>
<tr>
<th>Month</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incidents</td>
<td>50000</td>
<td>55000</td>
<td>60000</td>
<td>65000</td>
<td>70000</td>
<td>75000</td>
<td>80000</td>
<td>85000</td>
<td>90000</td>
<td>95000</td>
<td>100000</td>
<td>105000</td>
</tr>
</tbody>
</table>

![Average CHP Incidents by Time of Day](image)

<table>
<thead>
<tr>
<th>Time of Day</th>
<th>Sunday</th>
<th>Weekdays M-F</th>
<th>Saturday</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incidents</td>
<td>50</td>
<td>100</td>
<td>150</td>
</tr>
</tbody>
</table>

Figure 1: Incident trends over time

After a roadway incident occurs, the incident’s detection is the next meaningful event in terms of the data record. In California, incidents can be detected by passing motorists who notice something on the roadway, whether it is debris or a traffic accident, and notify law enforcement. Traffic incidents are also often detected by CHP officers as they patrol certain beats throughout the day. Regardless of the type of incident, early detection is critical so that actions can be taken quickly to ensure public safety and efficient roadway operations.

Once an incident has been detected, formal agency documentation of the incident can begin. Documentation is important because it reveals patterns in incident rates by location over time. This information allows agency officials to identify and address conditions (e.g., roadway geometry, weather events) which may...
contribute to larger numbers of incidents. Incident documentation is conducted by CHP dispatchers at various communication centers across California. The initial CHP record made of a traffic incident after it has been detected consists of the elements given in Table 1 below.

Table 1: Incident record elements

<table>
<thead>
<tr>
<th>Element</th>
<th>Example</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Timestamp</td>
<td>9:58am July 27, 2012</td>
<td>The time at which a CHP unit is initially assigned to the incident, given in local time</td>
</tr>
<tr>
<td>Type</td>
<td>Trfc Collision-No Inj</td>
<td>Freeform categorization of incident, often uses shorthand and highway patrol codes</td>
</tr>
<tr>
<td>Location</td>
<td>Sr78 / Sr67</td>
<td>Freeform location of the incident along the roadway, often uses freeway and cross street names</td>
</tr>
<tr>
<td>Area</td>
<td>El Cajon</td>
<td>Geographic area of the incident (e.g., city, if applicable)</td>
</tr>
<tr>
<td>Coordinates</td>
<td>(33.042033, -116.868555)</td>
<td>Latitude and longitude, when available</td>
</tr>
</tbody>
</table>

This first record is made when a CHP unit is assigned to the incident by the dispatcher. Additional details about the incident are added to the record as the incident unfolds. Updates are made when the CHP unit is en route, and when the CHP unit reaches the scene, and continue on as appropriate depending on the incident. Each update contains a timestamp and a freeform text description. No publically available record is made of the incident’s end time, however this can be inferred. The CHP maintains a public incident feed that displays incidents currently active throughout the state. Incidents are removed from the list when they are no longer active, and this removal can be observed to determine the incident’s approximate end time.

Twitter

Twitter is a widely-used platform for the real-time social sharing of short text-based messages. Users of Twitter generate content in the form of 140-character posts (called tweets). Users can choose to broadcast their tweets publically or keep them private, and can interact socially by subscribing to feeds of tweets from other users. Twitter’s character limit for tweets encourages the service to be used for short, frequent exchanges of information. Twitter’s growing popularity has coincided with a rise in smartphone usage, resulting in Twitter being used frequently on mobile devices. Given this ease of use, many Twitter users post tweets throughout the day wherever they might be as they engage in other activities. This gives Twitter data good spatial and temporal coverage, and results in content that often concerns the users’ day-to-day experiences.

Twitter delivers all of this content in real-time. This means that a tweet written by a particular Twitter user is made accessible and pushed to that user’s followers instantaneously. Twitter’s real-time delivery enables Twitter content to be consumed as it is happening. When the tweets from a large number of users are aggregated, real-time information about current occurrences can be extracted. This makes Twitter a valuable source for live information on events as they progress.

Twitter has a large volume of activity, giving analysis conducted on Twitter data increased relevance to the global population. In 2012, 8% of online adults in the United States used Twitter regularly, a rate four times as high as it was in 2010. This rise in usage was exponential, doubling in 2011 and again in 2012 [1]. In March of 2012, Twitter reported having 500 million active users worldwide posting one billion tweets every three days. With such a large amount of content, Twitter data can be filtered to only include publically available tweets that contain location information and refer to transportation incidents while still retaining a large quantity of data with good coverage.

Similar to incident data, the quantity of interactions on Twitter follow certain patterns. Most tweets are posted in the evening hours, with 6pm to 8pm containing the largest volumes of tweets. The early morning hours (midnight to 3am) typically only see 20% as many tweets posted as during peak hours (see Figure 2). When looking across a typical week, Thursday and Friday have the largest average volumes of tweets. These trends match the incident data trends seen in Figure 1 well. It appears that the evening hours at the end of the week offer the greatest likelihood for tweets about roadway incidents to be posted.
Twitter information is particularly well suited for data mining for several specific reasons. First, because of Twitter’s focus on the public broadcasting of content, only a small percentage of Twitter accounts are private. Additionally, tweets can be encoded with a location. Users can enable a setting to have all of their tweets automatically tagged with their coordinates, or they can choose to attach their location to certain tweets. Perhaps due to privacy concerns, the vast majority of tweets (98.6% of those analyzed in this study) do not have any location information attached. As discussed above, much of the content on Twitter concerns day-to-day events which can be relevant to transportation analysis. Finally, Twitter maintains a free real-time streaming API through which a sample of all tweets can be accessed. The quantity of data, relevance to transportation analysis, and ease of access all make Twitter an appealing source for supplemental transportation data.

The Potential in Comparing Tweets with Incident Records

Twitter provides the opportunity to compare the things people are saying with official records of events such as traffic incidents. If the content of a tweet appears to concern a traffic incident, and the tweet was made near the same time and location as an incident was officially recorded, that tweet likely represents contextual information about that incident. Comparing tweets and traffic incident records could be interesting for three reasons: (1) Twitter offers the potential for consistent analysis and measurement nationwide, (2) tweets represent a purely passenger-centric perspective, and (3) as an test case for using publicly broadcast user generated content to learn about roadway conditions.

Part of the appeal of Twitter as a supplemental data source for traffic incidents is its broad geographic usage. This means that Twitter content could be analyzed in a repeatable way to provide similar information anywhere with sufficiently dense Twitter usage. This universality contrasts with current traffic incident data collection practices. While national systems such as the Fatality Analysis Reporting System (FARS) collect standardized records of fatality incidents, there is no nationwide standard for the collection of data related to general traffic incidents. The types of incident information recorded, the times it is recorded, the mechanism through which it is recorded and the way it is stored and retrieved can all vary by agency.

Twitter is a platform for the generation and sharing of small pieces of text-based information. Because this information is generated by the Twitter’s users themselves, Twitter is a source of user-centric data. While traditional hardware sensor-based performance measures can sufficiently describe the performance of a transportation system, the passenger’s experience in that system can tell an entirely different story. Passenger effects can be often overlooked in transportation analysis, perhaps due to a lack of reliable information on the passenger’s experience. Twitter offers an opportunity for the passenger’s experience to be monitored at low cost.

Finally, this paper discusses using publicly broadcast user generated content to learn more about roadway conditions. The same analysis techniques employed here can be useful for mining other sources of user generated content for meaning as well. Other patterns of activity or communication could potentially be monitored and treated similarly to supplement existing transportation data sources and add meaning to data.

OTHER EFFORTS TO MINE TWITTER DATA

The information on Twitter consists of the collective social interaction of a large number of people, and thus is a valuable resource for predicting various large-scale trends. In the health profession, Twitter data has been...
used to improve models for predicting flu trends. Twitter’s advantage in this arena is that it provides near-real time
location-encoded content that can prove useful in supplementing other flu prediction models [2]. Another study
shows that the overall mood of content on Twitter can reliably and accurately predict daily stock market fluctuations
[3]. To accomplish this, this study measures and quantifies mood across six dimensions. Both of these studies
considered the content of a large number of tweets in aggregate to uncover underlying trends in a sample of Twitter
users that accurately represented the global population.

Twitter has also proven to be a very fast resource for information on discrete events. Because Twitter
enables users to easily post content in real time, and from anywhere with internet access, it has become common for
news of notable events to be reported on Twitter before appearing on more established media channels. Users can
rebroadcast (retweet) each other’s tweets, which enables the rapid spread of interesting information to a wide
audience. Twitter is particularly suited to quickly spreading information about real-time events, as evidenced by its
being a prominent early source of information on the 2009 United Airways Hudson River plane crash. Similarly,
other studies have treated Twitter as a “social sensor” for earthquakes, showing that tweets about earthquakes can be
observed to detect seismic activity in an area [4]. These examples show that certain events can be detected in real-
time through Twitter. This is possible because Twitter users often share information instantaneously, and
information can travel quickly throughout the Twitter user network.

Similar research in the field of transportation monitoring has been carried out as well. Sentiment analysis
has been used to reveal public opinions regarding transit agencies and airlines [5, 6]. This can be done at minimal
cost, and in real-time. Additionally, because Twitter content is generated through social interaction, it comes from
the Twitter user’s perspective. This perspective distinguishes Twitter data from hardware-based sensor data, giving
it value as a supplemental data source.

DATA

Incident Data

Incident data was collected from the California Highway Patrol’s live incident feed [7]. This feed is publically
accessible and displays all active traffic incidents in California at any given time. Only live incidents are available;
there is no way to access historical incident records through this data portal. However California’s Freeway
Performance Measurement System (PeMS), an Archived Data User Service (ADUS), maintains a historical record
of activity on this feed which was accessed for this study.

Not all incident records contained coordinates identifying their location. Instead, those incidents’ locations
were labeled with a freeway name and absolute postmile. Absolute postmile is a linear reference that marks a
location based on the distance down in the roadway’s centerline that it lies from the freeway’s ultimate termination
point. In order to compare the locations of incidents and tweets, locations marked with only freeway and absolute
postmile had to be translated to latitude and longitude. This translation was carried out through the use of a table
(available in PeMS ) which relates roadway name (e.g., I80-S) and absolute postmile to latitude and longitude. The
latitude and longitude closest to the incident’s postmile was assigned to the incident record.

Some types of roadway incidents are excluded from this data source. Incidents that could be
considered dangerous or confidential, such as those involving police actions,
are not included in the interest of
public safety. Additionally, only
incidents that fall under the
jurisdiction of the CHP are included.
Incidents that fall under the
responsibility of city police or county
sheriffs are not included.

For this study, information on
11,752 incidents was obtained from
PeMS. The incidents took place
throughout California between
Monday July 16 and Tuesday July 24,
2012. The data elements included in
these records are those given in Table
1 with the exception that latitude and

Figure 3: Geographic distribution of one day of (a) tweets and (b) incidents
longitude were not available and had to be computed using the lookup table as described above. Due to ambiguous location information, 9.9% of all incidents were unable to be assigned coordinates and were discarded. As can be seen in Figure 3 (b), the incidents are well distributed over California’s freeways, with a larger concentration in the urban areas of Los Angeles, the Bay Area, San Diego, and Sacramento.

**Twitter Data**

Twitter data was collected from Twitter’s real-time streaming API, a resource available to the general public to access Twitter’s global stream of data. The Twitter streaming API is used by both application developers and researchers to access tweets. The streaming API encompasses three endpoints that provide streaming content from one user, a group of users, or all public users. Various filters can be set on these data streams to only capture tweets within a geographic area or only those containing certain terms. Data for this study was filtered to contain keywords that relate to traffic incidents. These keywords are “accident”, “crash”, “traffic”, “road”, “freeway”, “highway”, “lane”, “wreck”, “car”, and “cars”.

When accessing data from Twitter’s public real-time streaming feed, data is limited to 1% of total tweet volume at any given time. However, the keyword filter is applied to the global Twitter stream before the 1% sample takes place. This means that filtered public Twitter data, as we are dealing with here, can potentially contain the global set of all matching tweets as long as that result is less than 1% of total tweet volume.

Only tweets which included location information could be used for analysis. This dramatically reduced the size of the dataset as only 1.3% of all collected tweets contained location information. Other relevant information considered included the content of the tweet, the tweet’s UTC timestamp, and the author’s Twitter username. A large amount of additional information is also available such as the author’s connectedness within the network, and the number of times each tweet has been retweeted or favorited. This information was discarded for the purposes of this analysis.

In total, 5,106,682 unique tweets (13.6 GB) from July 16, 2012 through July 27, 2012 were obtained for analysis. Each of these tweets contained at least one of the incident-related keywords listed above. Only 65,236 of the original 5.1 million tweets contained location information, and those tweets could come from anywhere worldwide (see Figure 4). Twitter’s public streaming API does have the capability to filter content based on location, however multiple filters are applied independently, meaning that any tweet matching the criteria of any of the filters would get included. Adding a filter for tweets from California would dilute the traffic focus of the dataset so only the keyword filter was used.

![Traffic-related tweets on July 14, 2012](image)
METHODOLOGY

After both datasets are obtained, they are compared to determine the relationship between the two, if any. If traffic incidents are accompanied by tweets that could be used to identify them or give them context, then we would expect there to be more incident-related tweets at the times and locations incidents occurred. This is tested in the volume-based comparison. Additionally, if tweets can be used to supplement incident data with user-centric perspectives of traffic incidents, we would expect tweets that closely match the time and location of an incident record to contain more vivid incident-related content. This is tested in the semantic comparison.

Volume-Based Comparison
To conduct the volume-based comparison, the timestamps and coordinates of all incident-related tweets are compared with the start times and coordinates of the CHP traffic incidents. Tweets and incidents follow a many-to-many relationship model in that many tweets could concern a single incident. Because a new record is created whenever a CHP incident is updated, a single tweet could also reasonably correspond to multiple incident records (e.g., when the CHP officer was dispatched and again when the officer was en route).

In an effort to capture all valid pairings, any tweet and incident that occurred less than 9 hours apart and within 50 miles of each other are counted as a match. This wide range would account for situations in which a motorist observes an incident on the roadway but (exercising good judgment) waits until completing their trip before posting about the incident on Twitter. By convention, the time difference between the tweet and incident is calculated as:

\[ \Delta t = t_{\text{incident}} - t_{\text{tweet}} \]

such that \( \Delta t > 0 \) indicates a tweet that was posted before the CHP incident record was created.

With all pairings identified according to these rules, it should be possible to determine whether incident-related tweets are more common when coinciding with traffic incidents. This can be done by plotting the bivariate distribution of the differences in timestamp and distance between the incidents and tweets. If the density of the distribution increases as the time and distance differences decrease, then some portion of the tweets are likely directly related to incidents. Note that this comparison treats all tweets with equal weight, regardless of their content (beyond matching the incident-related filter keywords).

Semantic Comparison
The volume-based comparison described above does not take differences in the content of the tweets into account, but treats all tweets matched to incidents equally. This may not be accurate because some tweets included in this analysis are unrelated to traffic incidents (even though they matched the incident-related filter terms). A different method of analysis is necessary to extract meaning from the content of the twitter messages. This can be achieved through a semantic analysis, which assigns values to words with different connotations, and then computes a measure of overall mood or sentiment on a subject based on those values. The goal of this ranking is to highlight tweets that are more likely to concern the observation of a traffic incident.

Table 2: Semantic weighting rules

<table>
<thead>
<tr>
<th>Category</th>
<th>Tweet Content</th>
<th>Intensity Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incident</td>
<td>“car wreck”, “car crash”, “traffic accident”, “crashed into”</td>
<td>10</td>
</tr>
<tr>
<td>Incident</td>
<td>“car wreck”, “car crash”, “traffic accident”, “crashed into” AND “saw”, “just”</td>
<td>20</td>
</tr>
<tr>
<td>Congestion</td>
<td>“traffic”</td>
<td>3</td>
</tr>
</tbody>
</table>

The content of the tweets is translated into an intensity score, as described in Table 2 above where terms separated by commas indicate an OR relationship. This collection of rules attempts to assign greater weight to tweets that contain a personal observation of a traffic incident. The rules break tweets down into reports of incidents
and reports of congestion. Additional terms could be added to take other incident types such as roadway debris or severe weather into account as well. In assigning the intensity score, it is possible for a tweet to receive points for both incident and congestion-related subject matter. For example, the message “There was a car wreck on the freeway and now traffic is really bad!” would earn an intensity score of 15.

The intensity score becomes a third variable to include in analysis with the time difference and distance between the matched incident and tweet. If incident-tweet pairs that occur close together and near the same time also tend to have higher intensity scores, it would indicate a potential for twitter data to be used as a source for passenger-centric contextual information which could supplement official incident records.

RESULTS

Volume-Based Comparison

After generating the set of all possible pairs of tweets and incidents with timestamps less than 9 hours apart and straight-line distances between their locations constrained to 50-miles or less, the density of the relationship between the two variables is plotted in Figure 5. If the tweets and incidents were uncorrelated, this density plot would be consistently flat due to noise. Clearly this is not the case; a single region of interest exists when the time difference $\Delta t$ is between 0 and -300 minutes, and the tweets and incidents are between 10 and 25 miles apart.

This can be interpreted to mean that tweets concerning incidents tend to be posted within 5 hours of the incident. This is certainly far from real-time, but also includes the time it takes the Twitter user to reach the incident after it occurs, and perhaps to complete the trip before tweeting their observation. It also appears that the region of interest could extend into the positive time difference range, which would indicate tweets occurring before incidents are recorded. This is likely the result of noise due to incorrect incident-tweet matches in almost all cases.

Figure 5: Density relationship between incident-tweet distance and time difference
Semantic Comparison

The initial time- and location-based matching process revealed several incidences of twitter content appearing to reference specific CHP incidents. Three examples are shown in Table 3 below. These messages qualify as passenger-centric contextual information, and add depth to the CHP records.

Table 3: Tweets and incidents

<table>
<thead>
<tr>
<th>Tweet Content</th>
<th>CHP Incident Description</th>
<th>Time difference</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Just saw a dead deer in the middle of the road....</td>
<td>ANIMAL - Live or Dead Animal</td>
<td>-3 hr, 51 min</td>
<td>17.9 mi</td>
</tr>
<tr>
<td>Saw a car crash on my way tu school .Carlooked busted up from the front .Did'nt see the other car tho ..He mustof hit nd run ! :o .. Crazy. !</td>
<td>1182-Trfc Collision-No Inj</td>
<td>14 min</td>
<td>2.6 mi</td>
</tr>
<tr>
<td>This girs car caught on fire it was so sad seeing her struggle getting outta the car hope she is ok #scary #60E</td>
<td>CFIRE-Car Fire</td>
<td>-28 min</td>
<td>41.3 mi</td>
</tr>
</tbody>
</table>

Figure 6: Tweet intensity by incident-tweet distance and time difference

To determine whether Tweets with greater intensity rankings also tend to coincide more closely with incidents, a scatterplot is made (see Figure 6). Each point represents a pairing between an incident and a tweet, and is shaded between blue and red based on the intensity score of the tweet’s content. While the trend is not dramatic, it does appear that the highest intensity tweets contribute to the shape of the density function shown in Figure 5. However, there are also many high-intensity tweets that appear to be randomly distributed, indicating a large amount of noise in the data.

DISCUSSION

The methodology used here of matching all tweets and incidents within 9 hours and 50 miles of each other naturally leads to noisier results. The large number of tweets and traffic incidents occurring concurrently necessitates a more targeted approach to incident-tweet mapping, for example by leveraging information about the roadway network. Calculating the most direct driving route between the tweet and incident coordinates could reveal how realistic a potential match really is. Also, many incident-related tweets reference specific freeways by name, which should be taken advantage of when matching tweets to incident reports.

Efforts could also be made to increase the quality of Twitter content. Local agencies interested in a low-cost supplemental data source could encourage citizen incident reporting using Twitter’s established platform. A
coordinated effort like this could encourage standardized message formats or the use of certain hashtags (a Twitter convention used to identify a tweet’s subject) to make the messages easier to match to roadway incidents.

It is only possible to gain additional information about roadway incidents from Twitter for areas with sufficient density of Twitter usage. The proportion of Twitter users who might post messages about roadway incidents is small; a large population of Tweet users is more likely to generate more meaningful content. Twitter is used by a larger proportion of the population in cities than rural areas. Thus, this methodology may produce results that are too sparse to be useful outside of urban areas.

Finally, it is important to note that tweeting while driving is unsafe in any circumstances and, as with text messaging, the risk inherent in tweeting while driving is never worthwhile.

CONCLUSION

Twitter use appears to correlate with CHP traffic incident records. Tweets can be matched to traffic incidents by examining the content of the tweets for key words and comparing the timestamps and locations of the tweets and incidents. When compared broadly, it appears that tweets tend to be posted within 5-hours of the incident that they refer to, and are most often sent between 10 and 25 miles of the incident’s location.

Because tweets are authored by the general public, tweets concerning roadway incidents give agencies access to a previously unavailable perspective. The user-centric perspective which characterizes Twitter data can give context and emotional meaning to roadway incident reports.

However, the large volume of content on Twitter means that sophisticated filtering based on content and location is needed to maximize the conclusions which can be drawn by the data. However, Twitter offers a low-cost, readily available data source for agencies interested in uncovering trends in incidents or in gaining information on incidents on their roadways and their effects on the population.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the California Highway Patrol for the use of their incident records as well as Twitter for the use of their streaming API. The authors are grateful to OpenHeatMap for providing a free heatmap generation tool (openheatmap.com).

REFERENCES


