# Complementing Travel Diary Surveys with Twitter Data: Application of Text Mining Techniques on Activity Location, Type and Time

Mojtaba Maghrebi[1*], Alireza Abbasi[2], Taha Hossein Rashidi[1], S.Travis Waller[1]
[1]School of Civil and Environmental Engineering, University of New South Wales (UNSW), Sydney, Australia
[1]School of Engineering and IT, University of New South Wales (UNSW), Canberra, Australia
{maghrebi, a.abbasi, rashidi, s.waller}@unsw.edu.au

***Abstract*— A growing body of literature in social science has been devoted to extracting new information from social media to assist authorities in manage crowd projects. In this paper geolocation (or spatial) based information provided in social media is investigated to utilize intelligent transportation services. Further, the general trend of travel activities during weekdays is studied. For this purpose, a dataset consisting of more than 14,517 tweets in south and west part of the Sydney metropolitan area is utilized. After a data processing effort, the tweets are clustered into seven main categories using text mining techniques, where each category represents a type of activity including shopping, recreation, and work. Unlike the previous studies in this area, the focus of this work is on the content of the tweets rather than only using geotagged data or sentiment analysis. Beside activity type, temporal and spatial distributions of activities are used in the classification exercise. Categories are mapped to the identified regions within the city of Sydney across four time slots (two peak periods and two off-peak periods). Each time slot is used to construct a network with nodes representing people, activities and locations and edges reflecting the association between the nodes. The constructed networks are used to study the trend of activities/locations in a typical working day.**

**Keywords—component; travel diary survey, twitter, social media**

## I. INTRODUCTION

Travel demand modelling, and analyzing and/or managing the operation of the transport network require availability of detailed information of several types of agents playing role in generation of trips use the transport network. These agents include but not limited to: individuals, households, vehicles and firms for each of which information about their attributes should be collected, collated, processed and analyzed. This information includes socio-demographic and economic attributes of people and firms, and the pattern of trips generated by them.

Data is generally a valuable product which exhausts a large portion of the provided financial resources for planning and operating the transport system. As a result, not necessarily all metropolitan areas can afford collecting data on a monthly or yearly basis. This has resulted in emergent of innovative approaches to temporally or /and spatially transferring data and models [1] or indirectly imputing the required data from other readily accessible data source [2]. Traditionally, data for demand modelling has been collected using two major methods called: i) revealed preference (RP) surveys and ii) stated preference (SP) surveys. These two major methods are used to collect data about a) household/individual travel diary ([3]), b) attitudes or opinions of people about the system and service [4] and c) counting agents (people or vehicles) using the transport system [5] Conventional data collection techniques for a and b include face-to-face, telephone, mail-out-mail-back, web-based, on-board (on transit for example) surveying methods. Count (c) data has been traditionally collected using roadside, GPS, on-board, smart card techniques. The significantly large cost associated with the data collection methods for data types of a and b does not require further discussion as the average cost of one complete household travel survey is more than $175 [6]. As a result technology has been employed to collect household travel survey data (or even count data) in a cost effective manner. For example the capacity of web-based surveys (apps), social network software, smart phones (accelerometers) and personal health sensors have been explored. Nonetheless, the practical inherent capacity of these emerging technology-based methods is yet to be explored.

A significant source of data which has been barely considered and its capacity for providing household travel information has been only trivially examined is the social network data provided by social media platforms such as Twitter. The main challenge before using Twitterdata is the significant noise existing in it which requires advanced text mining and linguistic techniques to extract the information that can be related to travel of people. When appropriate linguistic and text mining techniques are applied to the tweeter data, then, travel time, travel duration, travel mode, origin and destination of trip and many other travel related attributes can be potentially approximated with some level of accuracy.

This paper presents a proof of concept for application of social media (e.g., Twitter) data for replacing (if better data is not available) or complementing travel diary data using text mining and linguistic techniques. When the noise in the Twitter

978-1-4673-6596-3/15 $31.00 © 2015 IEEE
DOI 10.1109/ITSC.2015.43

data is gauged, three major applications are proposed by this study for which data is readily accessible from Twitter.

1- In-home activity data: There is challenge in the area of travel demand modelling to obtain data about in-home activities of people. This is important to travel demand modelers, specifically activity-base modelers, because there is a tradeoff between hours people spend for some types of activities like eating in home and out of home. If the activity is scheduled to happen at home, one out-of-home activity is cancelled which results in less number of travels happening on the transport network which is of great importance to travel demand modelers and planners.

2- Tour formation: tour-based models are among advanced demand modelling approaches which require collecting information about trips forming a tour of activities typically starting from home and ending to home. Twitter users often provide information about their daily activities which can be mined to extract information about the location, time and purpose of different activities, especially if it is linked with land use data. Using Twitter data for modelling tour formation behaviour can significantly complement the models that are developed using household travel surveys.

3- When the Twitter data is mined using linguistic techniques, it becomes possible to forecast potential activities to happen in future. In other words, if future tense is used in a tweet, and a location is stated about an activity to happen soon in future (in less than a week), it can imply that the person is likely to be at that location in a short time to be determined. When a model processes tweets' contents and approximates number of trips to happen in a short run in future, operation and management of the transport system can be facilitated. This has a significant impact on evacuation management and managing any disruption in the network which can be the result of an accident or a large event.

The abovementioned three applications of the Twitter are introduced by this study for the first time and the authors are in the process of developing such models to demonstrate the inherent capacity of the proposed approaches for complementing the existing demand models.

## II. Literature Review

Social networking sites or Web 2.0 applications, such as Facebook, Twitter and YouTube, have revolutionized the way information is produced, shared and stored: anyone can provide information; access and comment on the information; resulting in information to be propagated timely to more audiences. Therefore, such applications or social networking sites are also referred to as social media. The rapid development of telecommunication technology and devices such as smart phones and compatibility of such applications have helped to increase the number of users using such services. These are getting more popular and millions of documents (e.g., texts, photos, videos) are published and propagated widely everyday. For instance, Facebook's statistics recorded [1]"890 million daily active users on average for December 2014" with 745 million users using mobile devices. Similarly Twitter claimed[2] "500 million Tweets are sent per day" and "288 million monthly active users" with 80% of active users using mobiles. This therefore creates a great opportunity for businesses and governmental departments to gain benefit from the free available information provided by the public to improve their services.

Crowdsourcing social media for disaster or emergency management [7] is one of the widely used samples of using social media data (e.g., Twitter's post -- tweets) to facilitate response and relief operation by emergency response organizations. The main aim of such studies is to enhance emergency situation awareness using social media [8]. Among different approaches, some attempted to develop tools to track the information provided in the social media for predicting a likely event [9].

This study attempts to investigate how social media can be used to facilitate and enhance transportation management. Among very few existing relevant studies, Gao et al [10] analyzed user's social behaviourr from a spatial-temporal aspect using location-based data tracked by "check-in" applications which enables social media users to share the locations of their trips. Later, Gao and his colleagues [11] used similar approach to propose a location-based recommendation system based on the temporal properties of user movement tracked using the same "check-in" data. Such approaches facilitate a variety of services such as traffic forecasting, advertisement, and disaster relief [10].

In literature there are a few studies that have focused on applications of social media and social networks in transportation. The existing studies can be divided into 4 categories: (i) assessing the public transportation service, (ii) evacuation and natural disaster, (iii) user activity pattern based on geotagged data and (iv) traffic incidents.

Collins et al. [12] used Twitter data to evaluate transit rider satisfaction in order to assess public transportation. They proposed a two-sided(?) assessment model by considering people opinions along with metrics which are typically measured by authorities. Similarly, Lung et al [13] studied public opinions and attitudes about light rail transit service in Los Angles by looking at Twitter data instead of traditional survey and interview. NikBakht et al [14] used Twitter data and news for assessing public involvement in transportation planning. They picked Eglinton Crosstown transit project in Toronto as case study because this project was mostly re-designed after public consultations.

To study possible ways of extracting additional information about natural disasters from social networks, Hasan and Ukkusuri [15] considered the social network influences on evacuation decision. Ukkusuri et al [16] studied the potential influences of social media during natural disasters to more effectively understand people behavior when a crisis happens. They particularly focused on Twitter data posted about the

[1] http://newsroom.fb.com/company-info/

[2] https://about.twitter.com/company

tornado in Moore, Oklahoma and applied a sentiment analysis. Similarly Kaigo [17] studied the role of social networks and particularly Twitter during Tsukuba 2011 earthquake in Japan where power outage immediately after the earthquake limited users access to media and social networks via smartphones became the primary way of access to media.

Also, a few studies have been conducted to extract human activity pattern using geotagged data such as Hasan et al [18] who conducted a research on location-based data collected from social networking sites to study human mobility and activity patterns. They used users "check-in" data which contains user activity and geo-location information and also Hasan et al [19] used similar data for extracting weekly activity pattern of individuals and user-specific activity pattern.

Finally, using social media for more effectively managing traffic incidents, Fu et al [20] attempted to study the feasibility of detecting traffic incidents from tweets and also proposed a way to manage incidents more effectively based on extra information that can obtain from related Twitter data. They only focused on tweets that contain incident related keywords and evaluated their achievements by comparing with the real-world incident data. They showed that tweets are useful for early incident detection and can be used as additional source of information for incident management. Similar approach was taken by Mai and Hranac [21] by comparing recorded incidents by California Highway Patrol with related tweets via visualizing the density of incidents and tweets coincide near the same location. Steur [22] conducted a similar approach but for highways in Netherland.

In this paper, we focus on extracting information about users' activities from tweets using text mining techniques rather than only considering the geotagged data. It is aimed to introduce an automated method to analyze social media data for travel data extraction purposes.

## III. METHODOLOGY

As it was briefly mentioned in the introduction, extracting data from social media for transportation purposes is much cheaper than traditional data collection methods. Therefore, this paper aims to extract users' activities pattern in an aggregated level by directly analyzing the context of tweets. First step to do so is data gathering. We randomly targeted a population size of near 1500 of Twitter' users in Sydney metropolitan area and similar to [18, 19] only used geotagged tweets. Data cleaning was conducted on over 20,000 extracted tweets (posted by the sample users over five years period) before any further process. The tweets obtained from Twitter API converted to an uniform format as shown in Fig. 1.

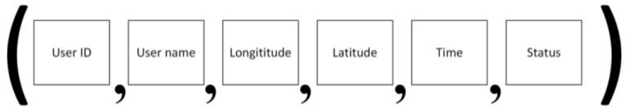

Fig. 1. The structure of cleaned tweets obtained from Twitter API

### A. Text Mining

In this paper, unlike [17, 21] where only sentiment analysis was implemented, we used text mining techniques associated with statistical and linguistic analysis to discover content of tweets. To do so, the content of all available tweets were analyzed to build a dictionary of unique keywords. Then the status of tweets were extensively searched to find groups of keywords that mostly come together. We came up with a bunch of over 40 groups of keywords. On the other hand, similar to [19] we assumed that users' activities can be one of the activities listed below:

- Eating
- Entertainment
- Home
- Recreation
- Shopping
- Social
- Work

So, an activity type was assigned to each group of words which frequently come together and then by running another search the tweets were tagged.

Rather than activity types, temporal and spatial information were also taken into account. South and west parts of Sydney metropolitan area were divided into 7 regions: CBD; Eastern Surburb; Burwood; Hurstvile; Bankstwon; Castle Hill and Penrith. Each tweet was linked to one of these zones and received the appropriate tag.

We repeat this process for four critical time slots which include two picks and two off-pick periods as follows:

- 6am-9am
- 9am-3pm
- 3pm-6pm
- 6pm-9pm

Then four networks were built for the aforementioned four time slots. In the networks nodes are users, activities and locations where an arc between users and activities or locations represents location and activity tags associate with the user in each tweet. The main purpose of this part of study is introducing an automated process to analyze location-activity pattern of social media users.

## IV. RESULT

Due to space limitation in this section only time slots (6am-9am) and (9-3pm) are selected and illustrated however, all time slots are statistically compared together.

Two networks were constructed Fig.2 and Fig 3 which are respectively illustrate time slot (6am-9am) and (9am-3pm) The activity-location pattern of Twitter users between 6am-9am is illustrated in Fig. 2: the location nodes in green and activity nodes in red. Fig. 2 and Fig. 3 clearly demonstrate that most of users' activities are social which is very hard to extract any transportation related information.

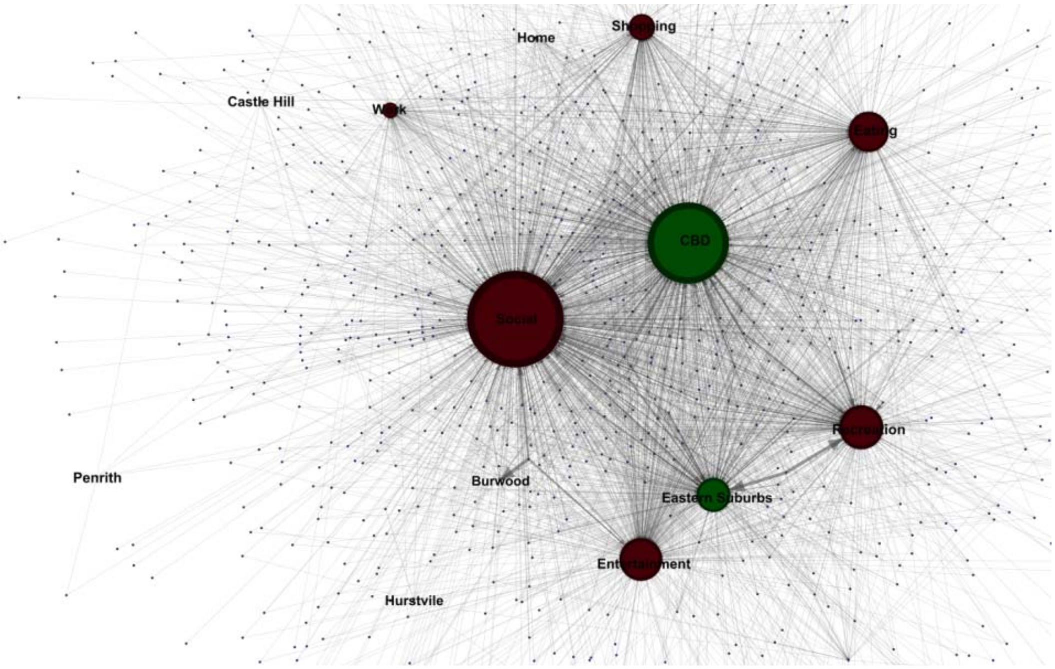

Fig. 2. Activity location pattern 6am-9am

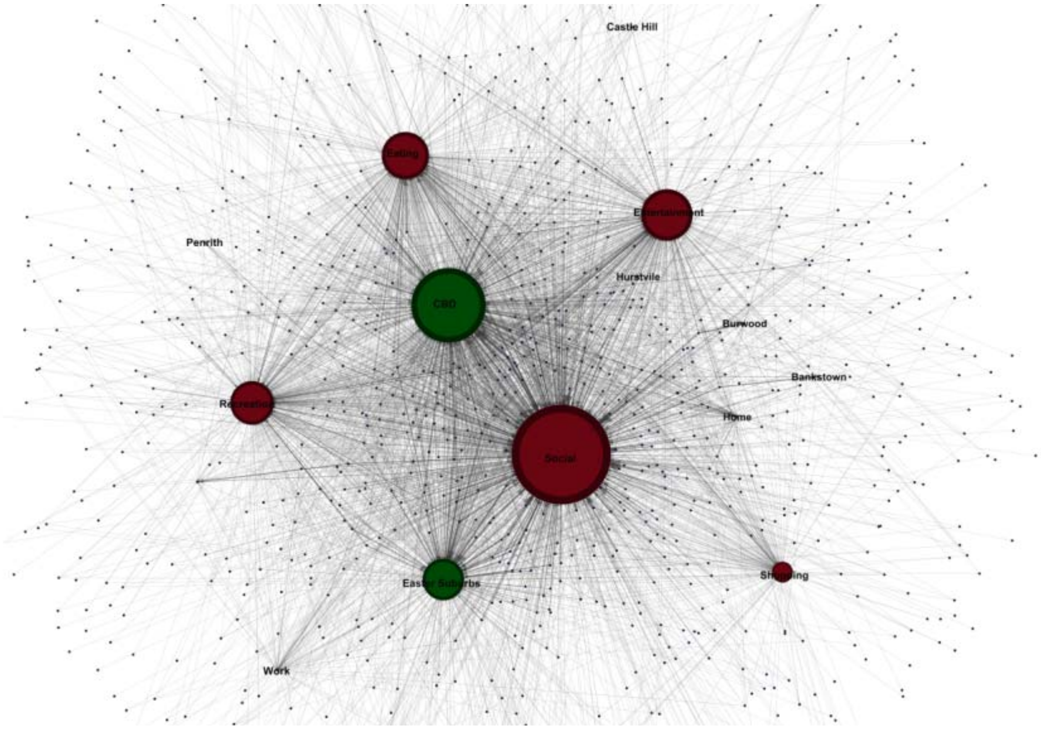

Fig. 3. Activity location pattern 9am-3pm

It might be a fact that the main purpose of people in using social media applications such as Twitter is to do social activities and being engaged with their community easily. Between 6am-9am, ‘entertainment’, ‘recreation’, ‘eating’, ‘shopping’ and ‘work’ related activities are the most popular recorded activities after ‘social’ and surprisingly ‘home’ activities are the least popular. Moreover, in terms of location, the graph shows a big portion of tweets has been recorded activities of people in CBD (Central Business District).

In the second time slot (9am-3pm), which is an off-pick period, again most of activities are identified as ‘social’ and the interesting point is having a few number of ‘work’ and ‘shopping’ activities in comparison with the first time slot (6am-9am).

Fig. 4 compares Twitter users’ activities among different four time periods. As shown, in all the four periods ‘social’ is the main activity recorded in tweets following by ‘Eating’, ‘Entertainment’ and ‘Recreation’. All the activities have more records for the second time slot (between 9 am and 3 pm) except for ‘Recreation’, ‘Shopping’ and ‘Work’ which surprisingly are depicted to have more records between 6 am and 9 am while it does not make sense at least for ‘Shopping’ and ‘Work’ which most of the shops are closed and official work hours are often 9 am and in some cases 8 am. It might be that the people are busy during that peak period to report their activities in Twitter. When Fig 4 is compared to the actual travel survey distribution of trips in Sydney in a typical in 2012/2013 it can be concluded that the Twitter data cannot be used to replace household travel surveys.

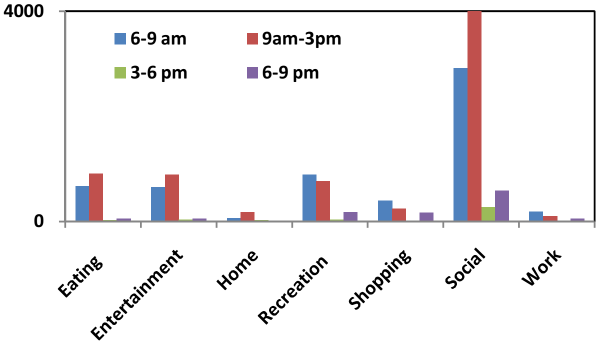

Fig. 4: Twitter users’ activities

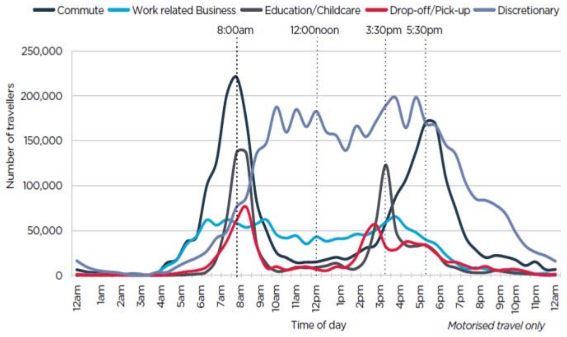

Fig 5. Persons travelling on motorized modes for selected purposes by time of day, average weekday in 2012/13 (Bureau of Transport Statistics Household Travel Survey Report: Sydney 2012/13 [23]).

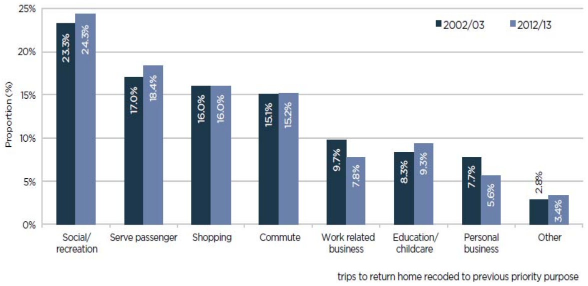

Fig 6. Distribution of trips by purpose on an average weekday (Bureau of Transport Statistics Household Travel Survey Report: Sydney 2012/13 [23] )

Nonetheless, when complementing household travel surveys is the aim of using social media data, it can be concluded that Twitter data is a suitable source of information for extracting information about social and recreational activities (figure 6). Further research is required to explore the inherent capacity of Twitter data for in-home activities, activity forecasting and evacuation management.

## V. Conclusion

This paper presented a proof of concept for application of social media (e.g., Twitter) data for replacing or complementing travel diary data using text mining and linguistic techniques. Unlike similar approaches, text mining techniques associated with statistical and linguistic techniques were used to automate the analysis the contents of tweets for extracting travel related activities. It was initially attempted to find the location-activity pattern of the data in different time slots (peak and off-peak). To do so, a dataset of more than 20,000 tweets in south and west part of Sydney metropolitan area was collected. After a data cleaning process, the tweets are clustered into 7 main categories using text mining techniques, where each category represents a type of activity such as social, shopping, recreation, and work. Unlike similar studies in this area, we focus on content of the tweets rather than only using geotagged data or sentiment analyses. Also other than the activity type, temporal and spatial information is used to tag tweets. To do so the longitude and latitude associated with the tweet are used. The studies area is divided into 7 regions and each tweet is linked to one of these zones. This process is repeated for 4 time slots (2 peak periods and 2 off-peak periods) and for each time slot, a network is built that its nodes are people, activities and locations. The constructed networks are used to study the trend of activities/locations in a typical working day.

Future research in this area includes using advanced linguistic techniques to extract the overall concepts of sentences and the tense of the verb used. This is an essential task for mining the text of Twitter data to be used for real time forecasting trips. Once the tense and location of the activity is determined in a precise way (other than using the reported location and time) it can be determined whether the activity was in-home or out-of-home and whether an in-home activity can be replacing an out-of-home activity. Having these two tasks working, tour formation behavior can be modeled for sequential tweets posted in a specific timeframe by one individual.

## References

1. Rashidi, T.H. and A. Mohammadian, *Household travel attributes transferability analysis: application of a hierarchical rule based approach.* Transportation, 2011. **38**(4): p. 697-714.
2. Miller, E., et al., *A Framework for Urban Passenger Data Collection*, in *10th International Conference on Transport Survey Methods*. 2014: Leura, Australia.
3. Rashidi, T.H., A. Mohammadian, and Y. Zhang, *Effect of Variation in Household Sociodemographics, Lifestyles, and Built Environment on Travel Behavior.* Transportation Research Record: Journal of the Transportation Research Board, 2010. **2156**(1): p. 64-72.
4. Beirão, G. and J.S. Cabral, *Understanding attitudes towards public transport and private car: A qualitative study.* Transport policy, 2007. **14**(6): p. 478-489.
5. Francis, R.C., et al., *Object tracking and management system and method using radio-frequency identification tags*. 2003, Google Patents.
6. Zhang, Y. and A. Mohammadian, *Bayesian updating of transferred household travel data.* Transportation Research Record: Journal of the Transportation Research Board, 2008. **2049**(1): p. 111-118.
7. Abedin, B., A. Babar, and A. Abbasi. *Characterization of the Use of Social Media in Natural Disasters: A Systematic Review*. in *Big Data and Cloud Computing (BdCloud), 2014 IEEE Fourth International Conference on*. 2014. IEEE.
8. Yin, J., et al., *Using social media to enhance emergency situation awareness.* IEEE Intelligent Systems, 2012. **27**(6): p. 52-59.
9. Cameron, M.A., et al. *Emergency situation awareness from twitter for crisis management.* in *Proceedings of the 21st international conference companion on World Wide Web*. 2012. ACM.
10. Gao, H., J. Tang, and H. Liu. *Exploring Social-Historical Ties on Location-Based Social Networks*. in *ICWSM.* 2012.
11. Gao, H., et al. *Exploring temporal effects for location recommendation on location-based social networks*. in *Proceedings of the 7th ACM conference on Recommender systems*. 2013. ACM.
12. Collins, C., S. Hasan, and S.V. Ukkusuri, *A Novel Transit Rider Satisfaction Metric.* 2013.
13. Luong, T.T. and D. Houston, *Public opinions of light rail service in Los Angeles, an analysis using Twitter data.* iConference 2015 Proceedings, 2015.
14. Nik Bakht, M., S.N. Kinawy, and T.E. El-Diraby. *News and Social Media as Performance Indicators for Public Involvement in Transportation Planning: Eglinton Crosstown Project in Toronto, Canada*. in *Transportation Research Board 94th Annual Meeting*. 2015.
15. Hasan, S. and S.V. Ukkusuri, *Social contagion process in informal warning networks to understand evacuation timing behavior.* Journal of Public Health Management and Practice, 2013. **19**: p. S68-S69.
16. Ukkusuri, S.V., et al., *Use of Social Media Data to Explore Crisis Informatics.* Transportation Research Record: Journal of the Transportation Research Board, 2014. **2459**(1): p. 110-118.
17. Kaigo, M., *Social media usage during disasters and social capital: Twitter and the Great East Japan earthquake.* Keio Communication Review, 2012. **34**: p. 19-35.
18. Hasan, S., X. Zhan, and S.V. Ukkusuri. *Understanding urban human activity and mobility patterns using large-scale location-based data from online social media*. in *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*. 2013. ACM.
19. Hasan, S. and S.V. Ukkusuri, *Urban activity pattern classification using topic models from online geo-location data.* Transportation Research Part C: Emerging Technologies, 2014. **44**: p. 363-381.
20. Fu, K., R. Nune, and J.X. Tao. *Social Media Data Analysis for Traffic Incident Detection and Management*. in *Transportation Research Board 94th Annual Meeting*. 2015.
21. Mai, E. and R. Hranac. *Twitter Interactions as a Data Source for Transportation Incidents*. in *Proc. Transportation Research Board 92nd Ann. Meeting*. 2013.
22. Steur, R., *Twitter as a spatio-temporal source for incident management.* 2015.
23. BTS, *Household Travel Survey Report: Sydney 2012/2013*. 2013.