



# Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges <sup>☆</sup>



Taha H. Rashidi <sup>a,\*</sup>, Alireza Abbasi <sup>b</sup>, Mojtaba Maghrebi <sup>d</sup>, Samiul Hasan <sup>c</sup>, Travis S. Waller <sup>a</sup>

<sup>a</sup> School of Civil and Environmental Engineering, UNSW, Australia

<sup>b</sup> School of Engineering and Information Technology, UNSW, Australia

<sup>c</sup> Department of Civil, Environment, and Construction Engineering, University of Central Florida, United States

<sup>d</sup> Department of Civil Engineering, Ferdowsi University of Mashhad, Mashhad, Khorasan Razavi, Iran

## ARTICLE INFO

### Article history:

Received 30 May 2016

Received in revised form 12 December 2016

Accepted 15 December 2016

Available online 30 December 2016

### Keywords:

Travel diary survey

Social media

Travel demand modelling

Mobility behaviour

## ABSTRACT

In the past few years, the social science literature has shown significance attention to extracting information from social media to track and analyse human movements. In this paper the transportation aspect of social media is investigated and reviewed. A detailed discussion is provided about how social media data from different sources can be used to indirectly and with minimal cost extract travel attributes such as trip purpose, mode of transport, activity duration and destination choice, as well as land use variables such as home, job and school location and socio-demographic attributes including gender, age and income. The evolution of the field of transport and travel behaviour around applications of social media over the last few years is studied. Further, this paper presents results of a qualitative survey from travel demand modelling experts around the world on applicability of social media data for modelling daily travel behaviour. The result of the survey reveals positive view of the experts about usefulness of such data sources.

© 2016 Elsevier Ltd. All rights reserved.

## 1. Introduction

The digital age accelerated the evolution of online social networks. Social media has become an emerging industry with massive input and output cash flow. As a result, massive data sources have been created as a result of such massive market. Harnessing such big data has become an interesting topic for researchers, scientists, practitioners and governments. Fields such as computer science, mathematics, social sciences, economics and management have invested considerable effort in developing understanding about various aspects of social networks and media data. It has been only recently that transport engineering, urban planners and travel demand modellers have noticed the richness of such big data and have started exploring the capacity of such data source for planning, management and operating purposes.

Initial movements towards understanding social media and their impact on the transport system started with descriptive analysis on mobility using location based social networks (Onnela et al., 2011). As the potentials of such data sources further explored, transport modellers pushed the frontiers of applications of social media data for modelling transport related issues (Hasan and Ukkusuri, 2015; Hasan et al., 2016). Nonetheless such efforts are still at their infancy and the community is not

<sup>☆</sup> This article belongs to the Virtual Special Issue on “Social Network Analysis in Future Transportation Systems: Contributions on Observability, Behaviour and Structure”.

\* Corresponding author.

E-mail addresses: [rashidi@unsw.edu.au](mailto:rashidi@unsw.edu.au) (T.H. Rashidi), [a.abbasi@unsw.edu.au](mailto:a.abbasi@unsw.edu.au) (A. Abbasi), [Mojtabamaghrebi@um.ac.ir](mailto:Mojtabamaghrebi@um.ac.ir) (M. Maghrebi), [samiul.hasan@ucf.edu](mailto:samiul.hasan@ucf.edu) (S. Hasan), [s.waller@unsw.edu.au](mailto:s.waller@unsw.edu.au) (T.S. Waller).

yet convinced about full potential of such cheaply available but costly to prepare dataset (Wu et al., 2014), where privacy of users must be maintained through aggregate or anonymized parsing analysis (Smith et al., 2012).

There are several complications associated with using social media data, especially if analysing the content of such a big data is of importance in understating the observations. For example, Twitter<sup>1</sup> data (tweets) typically contain normal text, hash-tag(s), and/or check-in data. Check-in data include location of tweets, making it associated with activities happening at that location (e.g., all tweets linked to a stadium, by users who provided checked-in data, are more likely to be related recreational activities). Similarly, hash-tag (#) messages are associated with an activity, event, location, etc. Therefore, it is relatively easier to work with check-in and hash-tag data as they are already associated with an event or location (Katakis et al., 2008). In particular, when check-in data is used for analysis of the destination/origin of the activity, determining trip purpose is relatively easy (Cheng et al., 2011). More information about applications of Twitter data can be found in a review paper by Steiger et al. (2015) where transport is completely excluded from their study. If check-in data or hash tag data is not of interest and more general information is used, extracting meaningful information can be challenging. More importantly, there are several biases and issues highlighted affecting the research on human mobility behaviour in different ways for some of which solutions have been proposed in fields such as epidemiology, statistics, and machine learning (Ruths and Pfeffer, 2014).

This study presents an overview of transport related studies which used social media for transportation planning and management. A special focus is given to the application of social media data in travel demand modelling studies. Relevant studies focusing on applications of social media on the following categories that are related to transport research are discussed in Section 2: (i) travel demand modelling, (ii) mobility behaviour (iii) individuals' activity pattern, (iv) assessing public transport and (v) traffic condition, (vi) and incidents and natural disasters. Section 3 presents a discussion about the evolution of evolution of social media use for transportation applications. This section is followed by a more detailed discussion about the capacity of social media data through results of an online survey in which travel demand modelling experts declared their opinions about usefulness of different social media data sources for planning, management and operation purposes. Finally a summary of the discussion and recommendations for future directions of using social media data in the field is discussed.

## 2. Use of social media in transport research

### 2.1. Travel demand modelling studies

The history of planning the transport system infrastructure goes back to the time the wheel was invented followed by the construction of the first paved road in Sumer in 500 BCE. At the same time, Darius I the Great, 500 BCE, started construction of an extensive road system for Persia including the famous Royal Road which was one of the first highways. About the same time, Roman roads were constructed with advanced technologies of stone-paved and metaled, cambered for drainage and were flanked by footpaths, bridleways and drainage ditches. Same road structure was later used by the Great Britain in the 18th century to establish the first toll system which included 250 miles of road and 40 bridges. All of these early transport system planning and network design efforts inspired transport engineers of the 20th century to develop a systematic procedure for policy appraisal and network design purposes. It was in the 1950s when the first prototypes of the conventional four-step models developed in Chicago and Detroit in USA. Since then, many metropolitan areas adopted a similar structure to evaluate the short, medium and long term consequences of different designs and policies. The 4-step modelling paradigm, which is a trip-based approach, led to the tour-based scheme in which individual level travel information is regarded for modelling purposes. Tour-based models were later evolved to activity based model in which individual/household level data is used to model individual/household level travel attributes (Rashidi and Kanaroglou, 2013).

Travel demand modelling techniques target modelling the mobility (movement) of people and vehicles (including passenger and commercial vehicles) in cities to understand their (mainly short distance) travel behaviour. Models are developed based on individual level data sources, in which behaviour of travellers is reflected, have been argued to dominate aggregate level models in terms of policy appraisal (Rashidi and Kanaroglou, 2013).

The evolution of travel demand modelling techniques developed the need for high resolution databases in which socio-demographic and economic attributes of people are used to model their day-to-day travel behaviour. Such data sources encompass travel diary of a sample of people representing the population. Having access to such an individual level travel diary is crucial to develop several components of the advanced behavioural modelling frameworks like tour-based and activity-based. The most important travel attributes considered in these modelling frameworks are: (a) trip purpose, (b) departure time, (c) mode of transport, (d) activity duration, (e) activity location, (f) travel route, (g) party composition, and (h) traffic condition

Other than travel data, information about long-term household decisions should be collected and modelled to be used as an important input to travel demand models. The major household decision for which commonly data is collected and models are developed are: residential location, job location and vehicle ownership. Among these three, vehicle ownership has been modelled more in travel demand frameworks. Housing and job search behaviour have been mainly considered exogenously in the travel modelling structures (Rashidi et al., 2012).

<sup>1</sup> [www.twitter.com](http://www.twitter.com).

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 (Rashidi and Mohammadian, 2011) or indirectly imputing the required data from other readily accessible data source (Miller et al., 2014).

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 (Rashidi et al., 2010), (b) attitudes or opinions of people about the system and service (Beirão and Cabral, 2007), and (c) counting agents (people or vehicles) using the transport system (Francis et al., 2003). 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 and smart card techniques. The significantly large cost associated with the data collection methods for data types of a and b is quite clear as the average cost of one complete household travel survey is more than \$200 (Zhang and Mohammadian, 2010). 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 networking sites or applications, smart phones (accelerometers) and personal health sensors have been explored (Wilde et al., 2015). Nonetheless, the practical inherent capacity of these emerging technology-based methods is yet to be explored.

The capacity of social media platforms such as Facebook,<sup>2</sup> Twitter LinkedIn,<sup>3</sup> Instagram,<sup>4</sup> Foursquare,<sup>5</sup> and Yelp<sup>6</sup> to provide information on household daily travel has been minimally examined (Golder and Macy, 2014; Yin et al., 2015). Tasse and Hong (2014) presented a wide range of possible ways of using geotagged social media to develop understanding of urban areas instead of using traditional ways of data collections. They categorized the opportunities (i) for city planner (such as: understanding the mobility pattern, understanding average distance travelled), (ii) for small business owners (such as: understanding customers demographic and customers before and after activities) and (iii) for individuals (such as: understating socially constructed places and understanding social flows in cities).

Social media platforms have a feature known as location-based services, which enable people to share their activity related choices (check-in) in their virtual social networks. Through location-based services, users can share their activity-locations when they visit restaurants, shopping malls, movie theatres and so on. Location-based data has received increasing attention, for travel demand modelling as the data can provide further knowledge about travel behaviour. However, the amount of check-in information using such services is less than the geo-tagged associated 'text' data available on people's posts on social media platforms such as Twitter. However, the main challenge before using such rich data is the significant noise existing in them which requires advanced text mining, natural language processing and data mining techniques to extract useful information that can be related to travel behaviour of people (Cramer et al., 2011; Maghrebi et al., 2015).

## 2.2. Aggregate mobility behaviour

Several studies have investigated how social media data can be used for understanding human mobility behaviour for a large number of people. These studies discovered universal laws for mobility behaviour of people at aggregate levels across different geographical scales (Noulas et al., 2012; Cheng et al., 2011; Jurdak et al., 2015). For instance, Cheng et al. (2011) analysed 22 million check-ins and observed Lévy Flight patterns and periodic behaviours in mobility behaviour of social media users. Cho et al. (2011) investigated the relationship between human mobility and social relationship using location-based check-in data. They found that social relationships can explain up to 30% of all human movements, while periodic behaviour explains 50–70%. Hasan et al. (2013) used a dataset of Foursquare check-ins to analyse urban human mobility and activity patterns. They determined the spatial distributions of visiting different places for various activity purposes by counting the number of purpose-specific visits within each cell and computed the proportion of visits to each cell for each activity category. Zhu et al. (2014) discussed an alternative way for household travel survey using location-based social networks (LBSNs). It was tried to predict Puget Sound Travel Survey (PSRC) using geotagged Foursquare data. To do so, they extracted (i) demographic features (age and work status), (ii) temporal features (proportion of a day), (iii) spatial features (using Foursquare API) from social media data. They analysed 13 million geotagged tweets over a period of 1 year to investigate crowd movements (spatio-temporal) pattern in New York (Manhattan).

Several studies have investigated if aggregate patterns, suitable for transportation planning, can be obtained from social media data. In particular, social media data has been used to estimate Origin-destination (OD) matrix. Cebelak (2013) and Jin et al. (2014) investigated the feasibility of using the location-based social media data to estimate travel demand using a doubly-constrained gravity model. They evaluated their result against the OD matrix generated by an existing singly-constrained gravity model and a reference matrix from the local metropolitan planning organization. They found significant improvement in reducing estimation errors caused by the sampling bias from the OD estimation method based on the

<sup>2</sup> [www.facebook.com](http://www.facebook.com).

<sup>3</sup> [www.linkedin.com](http://www.linkedin.com).

<sup>4</sup> [www.instagram.com](http://www.instagram.com).

<sup>5</sup> [www.foursquare.com](http://www.foursquare.com).

<sup>6</sup> [www.yelp.com](http://www.yelp.com).

singly-constrained gravity model. In another study, Lee et al. (2015, 2016) used geo-tagged Twitter data to understand its relationship with traditional travel demand model. Based on greater Los Angeles metropolitan area, they compared the Twitter based OD matrix with a recent OD matrix provided from a 4-step model output and estimated regression models to measure the correlations between the ODs provided traditional travel demand model and Twitter-based method. Their preliminary results show the added value of large-scale location-based social media data for modelling travel demand.

Although the above studies show the potential of social media data for modelling aggregate travel behaviour, these studies have limited scopes and hence further research is needed to utilize the full potential of this kind of data. Most of the studies, related to discovering mobility patterns, actually analysed the visiting patterns of the users to different places in a city. While such information is valuable, for modelling purposes we also need the origins and destinations of the movements and modal preferences. Methods to estimate O-D matrix can help us to resolve the problem of identifying the origins and destinations of movements. However, it is not clear how much error is introduced in the aggregate patterns due to the lack of sample representativeness and the biases present in the data. More comparative analysis between traditional survey-based and social media data is needed to measure and correct the biases.

### 2.3. Individual-based activity behaviour

Geo-tagged social media data and particularly check-in data has been utilized to infer activity purposes. Using venue category information from check-in data, studies from social science, computer science, and transportation science have used innovative ways to extract meaningful activity behaviour patterns and model behaviours with diverse applications. These studies include activity recognition (Lian and Xie, 2011), activity choice patterns (Pianese et al., 2013; Coffey and Pozdnoukhov, 2013; Hasan and Ukkusuri, 2014), predicting next place to check-in and friendship (Chang and Sun, 2011) and inferring life-style behaviour from activity-location choices patterns (Hasan and Ukkusuri, 2015).

Lian and Xie (2011) developed a conditional random fields model which predicts user activities given the location, time, identification and check-in history of the user. Chang and Sun (2011) analysed Facebook check-in data to predict next check-in place using a logistic regression model. Coffey and Pozdnoukhov (2013) compared Foursquare data with CapitalBikeShare report in Washington DC to predict the behaviour of people who use Bike Sharing facilities. Using probabilistic topic models, they analysed bikeshare user movement as well as finding relationship between bikeshare user activities before, during and after using bike sharing facilities. Hasan and Ukkusuri (2014) analysed Foursquare check-in data from social media for extracting individual weekly activity patterns using probabilistic topic models. Lee et al. (2016) used geo-tagged tweets to create individual activity spaces based on minimum bounding geometry (convex hull). By creating density maps of activity space, they found clear differences between weekday and weekend activity spaces. They used a clustering model to classify activity patterns. However, social media data contains rich information on activity types but this study could not differentiate activity types as found in several earlier studies. Davis and Goulias (2015) presented an ordered probit model to explain the attractiveness and opportunities of places perceived by the residents of Santa Barbara, California. They combined information from a place perception survey, geo-tagged tweets from Twitter and business establishment data from Yelp. This study has found improved explanatory power for models because of social media data showing a promising direction towards developing better activity-travel behaviour models.

Check-in data from social media has an enormous potential of improving our knowledge in activity participation behaviour. Approaches so far used to understand activity participation from social media mainly come from machine learning and data mining fields. Probabilistic models such as conditional random fields, logistic regression, and probit models have been used to predict various aspects of activity participation. Different classification techniques such as probabilistic topic and  $k$ -nearest neighbour models have been to classify activity choice patterns and cluster users based on their activity patterns. However, the full potential of check-in data for activity-based modelling is yet to be realized. It is not clear how the derived activity patterns can be explained since very limited socio-demographic variables are available from social media data. Researchers have also identified the challenges of modelling activity generation and sequences/scheduling using social media data due to missing activities (Hasan, 2013). Complex probabilistic models accounting for missing observations and inferring socio-demographic characteristics will be needed.

### 2.4. Public transportation assessment

There are a few papers in this area that mostly used sentiment analysis and keyword search for assessing public transportation (Schweitzer, 2014). Public transport has benefited from a solid review paper discussing applications of social media data in domains related to public transport by Pender et al. (2014) which is a standalone exercise of its kind with a special focus on transit. Collins et al. (2013) used Twitter data to evaluate transit rider satisfaction in Chicago train lines. They proposed a two-side assessment model by considering people opinions along with metrics that are typically measured by authorities. This paper is recognised as one of the pioneers of using social media data in public transport analysis. Similarly, Luong and Houston (2015) studied public opinions and attitudes about light rail transit service in Los Angeles by looking at Twitter data instead of traditional survey and interview. Nik Bakht et al. (2015) used Twitter data and only news sources to assess 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. Steiger et al. (2014) assessed public transportation flows using geotagged social media (from Twitter, Foursquare, Instagram and Flickr) and validated it using

real data obtained from OpenStreetMap. They applied density-based spatial clustering (DBSCAN) to LDA to cluster the topics related to “Train”. Then it was tried to segment the geotagged social media data to railways.

For agencies and city planners, knowing people's opinion about public transport projects/service is crucial. As it has been discussed in the literature, social media can be used as a possible source of information to obtain public opinion about the system. This approach can be used at a low cost level at any time the information about public opinion is required.

## 2.5. Traffic conditions

There are some recent efforts trying to extract traffic condition data from social media which are mainly useful for network operation and management purposes. Tian et al. (2016) assessed the validity of traffic incidents reported in social media by comparing field camera data from Austin, Texas and social media posts. The study found that citizens tweet more often about true incidents compared to false incidents and tweet more often about major severe incidents compared to minor incidents such as traffic hazards and stalled vehicles. However, they also found that social media incident reports have low quality as around half of the verifiable incidents in their sample turned out to have limited information to the traveling public. Steur (2015) showed in a particular highway in the Netherlands, there is a meaningful correlation between number of accidents and frequency of tweets near that area. Wanichayapong et al. (2011) proposed a rule-based content analysis for extracting traffic related information from tweets related to either points or links in Bangkok urban network. This paper has been significantly cited in the literature as it was one of the first papers introducing social media as a means for early accident identification for traffic management purposes. They used an approach to detect tweets including place and traffic related information of accidents. Ribeiro et al. (2012) illustrated that there is a meaningful correlation between real traffic conditions and tweets talking about traffic conditions in Belo Horizonte (Brazil). They searched a predefined list of words in the content of tweets reflecting traffic conditions such as movement (e.g. “slow”) or traffic status (e.g. “accident”). Then to match proper location a gazetteer was used to find street and neighborhood names as described in contents. Kosala and Adi (2012) developed a method for monitoring traffic condition of roads in Jakarta by real-time analysis of tweets. They also did keywords search among the tweets' contents to extract the traffic conditions. Their results were enhanced with confidence level of traffic information. Gao et al. (2012) attempted to investigate how social media data can be used to facilitate and enhance transportation management. Later, Gao et al. (2013) used a 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.

It has been addressed in the literature that social media contents can be used for traffic monitoring. This approach might be considered as a supplement for the ever growing transport monitoring platforms. However, reliability of social media data can be questioned due to low response rate for specific modes of transport at different time of day. Nonetheless, when sufficiently large data is in hand by social media, it can be considered a supplementary source of data to extract information about traffic conditions.

## 2.6. Interventions: Incidents and natural disasters

Pender et al. (2014) reviewed the literature of unplanned transit network disruptions with a focus on social media applications. This paper discussed how social media can be used to inform people and collect data during disruptions when other types of media are not necessarily accessible. Lindsay (2011) addressed the potential advantages of using social media in case of nature disasters. Hasan and Ukkusuri (2013) considered the social network influences on evacuation decisions. Ukkusuri et al. (2014) studied the potential influences of social media during natural disasters to more effectively understand people behaviour when a crisis happened. They particularly applied a sentiment analysis on Twitter data posted about the tornado in Moore, Oklahoma. Similarly, Kaigo (2012) 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. In this situation, social networks via smartphones became the primary way of access to media. Sakaki et al. (2010) focused on tweets related to earthquake/typhoon to extract real time information about a disaster and constructing an earthquake reporting system in Japan. They developed a platform that can notify public much faster than authorities and agencies.

Another application of social media which also has received attention in the literature is acquiring real information about incidents. In those papers using social media for more effectively managing traffic incidents was discussed. Fu et al. (2015) studied the feasibility of detecting traffic incidents from tweets. They 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. It was 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 (2013) 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 (2015) did a similar approach but for highways in the Netherlands.

In short, case of emergency situations in large cities, particularly when natural disasters happen, having real-time information about the current network is crucial for managing the system in an efficient way. Social media, given its popularity during disasters, can supplement other data sources to reflect the situation in a more holistic way. The main drawback is



related to the part of data being posted by people who are not in the affected areas but reporting about the event, which requires appropriate data processing and filtering mechanism.

### 3. Emerging area of transport

This section presents a descriptive discussion about the evolution of literature on social media with a focus on transport. After testing the result for different keywords in Scopus, as one of the main widely used bibliographical databases, we have used the following query to search for the keywords/terminologies in the 'Title' and 'Abstract' of publications index by Scopus to retrieve a reasonable list of publications on transport using social media.

**("Social media" OR Twitter OR foursquare OR facebook OR yelp OR instagram) AND ("travel" OR "transport" OR "mobility" OR "geo")**

Originally 935 records have been extracted but 874 records have been selected for further analysis after cleaning the data (e.g., removing some records with no 'author' and 'abstract' information). As shown in Table 1, six different types of publications have been found where 'conference' and 'journal' publications (with 461 and 377 records respectively) cover more than 95% of the records. As expected all the relevant records have been published after the development of social media platforms such as Facebook and Twitter in early 2000s. Please note that the results do not covers the complete list of all the publications in 2015 as the data collection has been conducted in late 2015. This result clearly shows that this field of study is relatively new with fairly more attention in academia since 2012 and as seen it is still growing with a steadier pace.

Table 2 shows more information about the number of publications and the total number of citations received by the publications in the top 25 sources which are mainly conferences and journals. As shown, 'PLOS One', which is a multi-disciplinary and open access journal, is ranked 1st in terms of both number of publications and citations followed by three Tourism related journals (i.e., 'Tourism management', 'Travel and Tourism Marketing' and 'Vacation Marketing') and two information science journals (i.e., 'Communications in Computer and Information Science' and 'Cartography and Geographic Information Science'). ICWSM, Advances in GIS and WWW are the main conferences with the more number of publications. In terms of quality, considering the average number of citations, 'Tourism Management' and 'Journal of Vacation Marketing' can be ranked 1st and 2nd with an average one citation per publication followed by Advances in GIS and WWW conferences with almost the same ratio.

In order to identify the main topics of study, the keywords (considering the multi-word terms specified by the authors) are analysed where 'social media', 'Twitter', 'social networks', 'Facebook', 'data mining' and 'location-based social networks' have been found as the most frequently used keywords with 170, 69, 35, 25, 22 and 21 respectively. Fig. 1 visualizes the frequency of the most repeated keywords (after removing 'social media') in a word-cloud form to give a bigger picture of the research in this domain.

In order to show the evolution of the field, we compared the frequency of the keywords between the early stage of the development of the field (2007–2011) and the recent years with more publications (2012–2015). As shown in Table 3, most of the commonly used keywords by researchers are the same for the two periods but their frequency has been increased dramatically and also their order has been changed (e.g., 'Web 2.0' was the 4th popular keyword during 2007–11 but the 11th between 2012 and 2015). However, some terminologies such as 'location-based social networks', 'big data' and 'human mobility' are quite new and just recently used by researchers during the past four years.

The temporal distribution of the keywords as an evolving procedure can be further examined by looking at the journal citation network of keywords over time and clustering the top cited keywords based on their co-citation correlation. Using a bibliometric analysis package, CiteSpace III (Chen, 2006), initial citation networks are generated for before 2005 and consecutive year until 2015. In these networks nodes are keywords and links are the co-occurrence of them citations. A pruning and merging technique (Chen, 2006) is used to cluster the keywords with stronger correlations. Then an alluvial diagram (Edler and Rosvall, 2015), is developed to track the changes of the keywords areas over time (Rosvall and Bergstrom, 2010). Alluvial diagram in Fig. 2 shows the keywords included in the data. As it can be seen from this figure, most of the transport related keywords started appearing in the literature since 2012. Transport related keywords are still occupying a small portion of what is of significance in these studies. Derivatives of "Mobility" such as mobility and human-mobility have been consistently used in the literature since its introduction in 2012. Travel and travel planning are also consistently

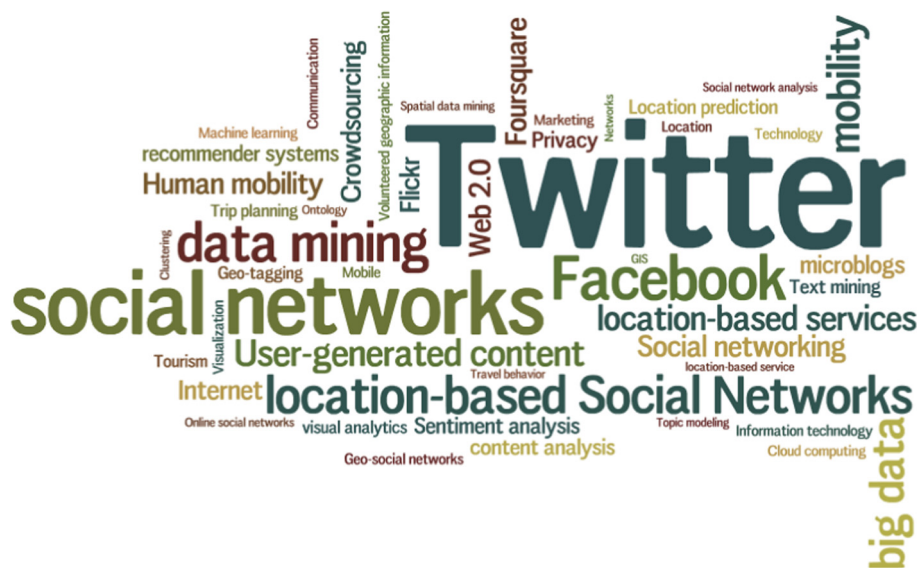
**Table 1**  
Publications statistics of the domain.

Publication Type	2007	2008	2009	2010	2011	2012	2013	2014	2015	Total
Book						3	4	2	1	10
Book chapter		1	1	1	2	4	5	4	4	22
Conference paper		3	10	30	56	78	116	112	56	461
Editorial						1	1		1	3
Journal article	2	1	4	6	16	48	67	97	136	377
Short Survey							1			1
Total	2	5	15	37	74	134	194	215	198	874

Table 2

Source titles of the Publications including the number of publications and citations.

	Journal/Conference Names	# of Pubs	Sum of citations
1	PLoS ONE	22	12
2	Proceedings-Int. Conference on Weblogs and Social Media, ICWSM 2012–2014	10	4
3	Proceedings-ACM Int. Symposium on Advances in Geographic Information Systems, GIS	9	8
4	Proceedings-Int. Conference on World Wide Web, WWW 2010–14	9	8
5	Tourism Management	8	8
6	Journal of Travel and Tourism Marketing	8	5
7	Proceedings-ACM SIGSPATIAL Int. Workshop on Location-Based Social Networks, LBSN 2010–13	8	4
8	Proceedings-Int. Conference on Advances in Social Networks Analysis and Mining, ASONAM 2013–14	7	5
9	Proceedings-Int. Society for Optical Engineering, SPIE	7	3
10	Proceedings-Int. Conference on Data Engineering, ICDE	7	2
11	Proceedings-Int. Conference on Information Systems for Crisis Response and Management, ISCRAM 2010–15	7	2
12	Proceedings-Int. Conference on Information and Knowledge Management, CIKM	6	3
13	Communications in Computer and Information Science	6	-
14	Journal of Vacation Marketing	5	5
15	Proceedings-Conference on Human Factors in Computing Systems	5	4
16	Proceedings-IEEE Int. Conference on Mobile Data Management	5	3
17	Cartography and Geographic Information Science	5	3
18	Proceedings-ACM Conference on Ubiquitous Computing, UbiComp 2012–13	5	3
19	Proceedings-ACM Multimedia Conference and Co-Located Workshops, MM 2011–12	5	3
20	Proceedings-IEEE Conference on Visual Analytics Science and Technology, VAST 2011–14	5	3
21	Cornell Hospitality Quarterly	4	4
22	ACM Transactions on Intelligent Systems and Technology	4	4
23	IEEE Transactions on Multimedia	4	4
24	Proceedings-ACM Int. Conference on Multimedia, MM 2012	4	2
25	Transportation Research Record	4	2



**Fig. 1.** The word cloud of the keywords of publications between 2007 and 2015 with five or more repeats.

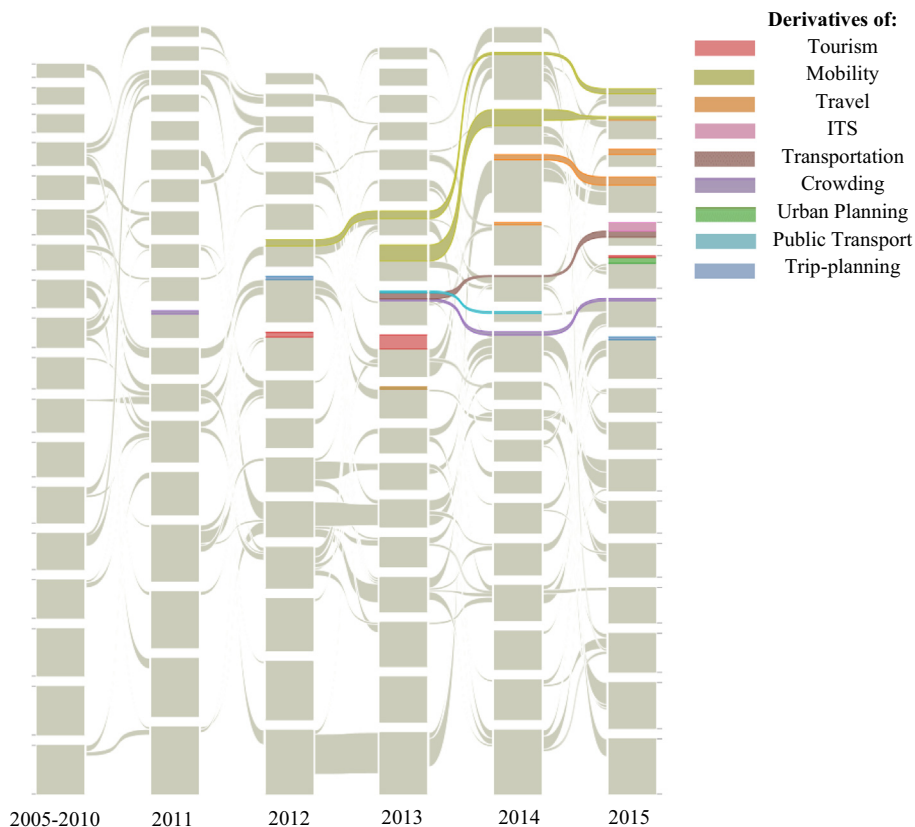
used in the literature while there are sudden appearance for public transport and tourism among the highly cited keywords. ITS and urban planning are the most recent applications of social media data.

In summary, the analysis shows that the literature base is evolving rapidly and there is hardly any transport journals in the top 25 sources presented in Table 2. Table 3 also expresses valuable information about the sources of data (e.g., Twitter, Flickr), methodologies and tools (e.g., text/data mining, sentiment/content analysis), and transport related topics and issue (e.g., mobility, trip planning, location prediction, and privacy). This reflects the importance of social media data and relevant methods for the highlighted transport challenges. But relatively few numbers of publications, surprisingly published out of the transport related journals, can suggest a gap in the transport field which requires further attention by the scholars in this field. To further examine the capacity of this gap in the literature based on the opinion of transport experts, a survey is designed and discussed in this paper in the next section.

**Table 3**

Comparing the commonly used keyword by scholars and their frequency.

Author keywords	Frequency between 2007 and 2011	Frequency between 2012 and 2015
Social media	18	152
Twitter	13	56
Social networks	11	24
Facebook	2	23
Data mining	3	19
Location-based social networks	0	21
Big data	0	19
Mobility	3	15
User-generated content	1	14
Location-based services	1	13
Web 2.0	7	5
Social networking	2	10
Foursquare	1	12
Human mobility	0	11
Crowdsourcing	1	10
Internet	1	11
Flicker	0	10
Microblogs	4	6
Privacy	3	8
Sentiment analysis	2	6
Recommender systems	1	7
Content analysis	1	8
Text mining	1	7
Location prediction	1	7
Tourism	1	6
Trip planning	1	6

**Fig. 2.** Evolution of research categories considered in social media and transportation related topics from 2012 to 2015.



#### 4. Survey results and discussion

Social media data can be available at small cost, nonetheless, processing the data and preparing it for transport related analyses is a challenging task especially due to two major issues: (1) the data is bias toward social media users and (2) linguistic and text mining techniques should be used to extract the majority of the useful information from such sources. These two challenges have been the main reasons less attention has been dedicated to usage of social media for travel demand modelling purposes.


To develop understanding about expert opinions about usefulness of social media for travel demand modelling purpose an online survey was designed where invited experts could report their opinion about advantages and disadvantages of different sources of social media data as well as conventional data collection methods. A snapshot of the survey is provided in Fig. 3.

14 experts from all over the world participated in this survey whose opinion has been summarized in Fig. 3. These experts are all from academia working on different aspects of travel demand modelling whose journal or conference articles have been cited in this paper. Seven data sources were included in the survey for each of which experts reported their opinion about the advantages and disadvantages listed below. The data sources included: (i) Conventional Household Travel Surveys, (ii) Prompted Recall Travel Surveying with GPS, (iii) Twitter, (iv) LinkedIn, (v) Instagram, (vi) Foursquare and (vii) Yelp.


The advantages and disadvantages considered in the survey about which respondents provided their Likert type feedback are listed below.

1. This data type is easily available
2. Expensiveness of data acquisition
3. Challenges in preparing, processing and cleaning
4. Capability to complement Household Travel Surveys
5. Capability to be used in zone-based demand modelling
6. Capability to determine trip purposes
7. Capability to determine mode of transport
8. Capability to determine departure time, activity duration and location
9. Capability to determine travel route
10. Usefulness to manage disruptions in the network such as disaster management
11. Usefulness for planning
12. Capability to determine socio-demographic of people
13. Necessity of advanced linguistic and text mining techniques for data extraction
14. Capability to Fig. out residential and job location of travelers

It can be seen from Fig. 4 that the conventional household travel survey and prompted recall travel surveying with GPS are unanimously argued to be capable of providing the required information for various travel attributes, however, the cost of conducting such surveys is quite high.



### Usefulness of Social Network/Media Data



Faculty of Engineering

S. No.	Question	Conventional Household Travel Surveys	Prompted Recall Travel Surveying with GPS	Twitter	LinkedIn	Instagram	Foursquare	Yelp
1.	This data type is easily available	Strongly Disagree						
2.	The cost of obtaining this type of data is negligible	Disagree						
3.	Preparing, processing and cleaning this type of data is challenging	Can't Say						
4.	This data type can be used to complement Household Travel Surveys	Agree						
5.	This data type can be used in zone-based demand modelling	Strongly Agree						
6.	This data type can be used to determine							

Survey Coordinator: Dr. Taha H. Rashidi

A study conducted by rCITI group, UNSW

Fig. 3. Snapshot of the first 4 questions in the online survey.

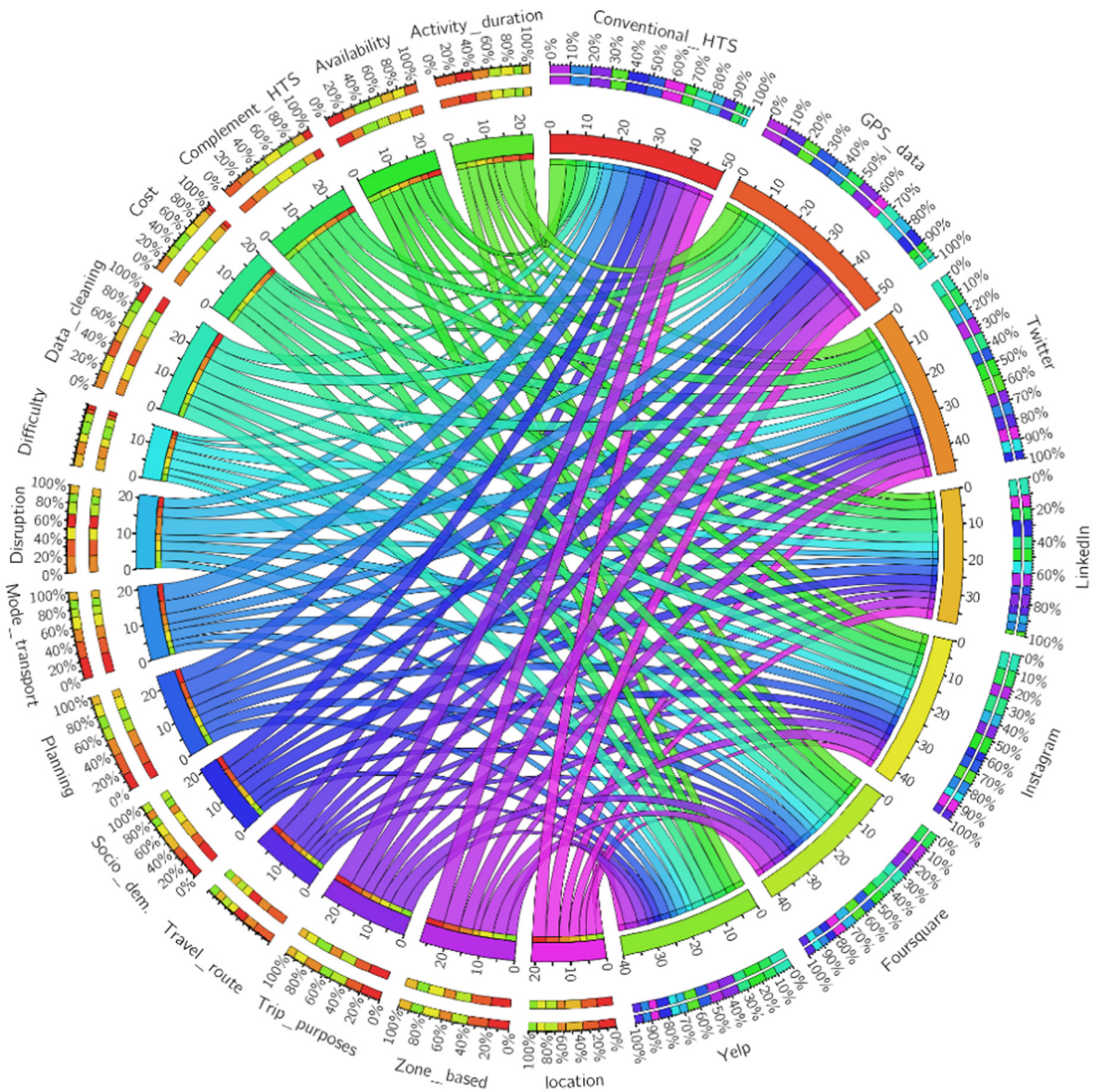


Fig. 4. Relationship between different data sources and travel attributes and land use variables.

Generally, the cost of obtaining the data from social media data is trivial, but processing such massive databases to extract travel information is a challenging task, especially for attributes such as trip purpose. As a result the accuracy of the outcome is not expected to be high unless advanced data mining and linguistic techniques are used. Nonetheless, the true potential of these techniques in extracting information from social media data is yet to be explored.

To better observe the opinion of the expert regarding these data sources, advantages and disadvantages have been categorized into six categories: (i) availability, (ii) cost, (iii) preparation, (iv) land use, (v) socio-demographic and (vi) planning. LinkedIn as a business-oriented social networking platform seems to be a good source of data to extract information about demographic and economic status of people. Yelp provides information about locations people visited but the temporal aspect is lost as people post after the fact. This data might be useful for destination choice behaviour. Instagram and Foursquare provide geocoded information which are useful for modelling overall pattern of people. Similarly, Twitter data, if geocoded, can be used for studying the mobility pattern of people. While more information can be extracted from the content of tweets, advanced text mining techniques is required to analyse the text (content) and extract transport related data. This will be discussed further in the next section (see Fig. 5).

The literature of travel survey methods has already started the discussion about how smart phones and apps can be developed to collect travel information with minimum distraction to travellers (Byon et al., 2009; Williams and Currid-Halkett, 2014). The literature has identified some capacity for such techniques to indirectly collect some travel information

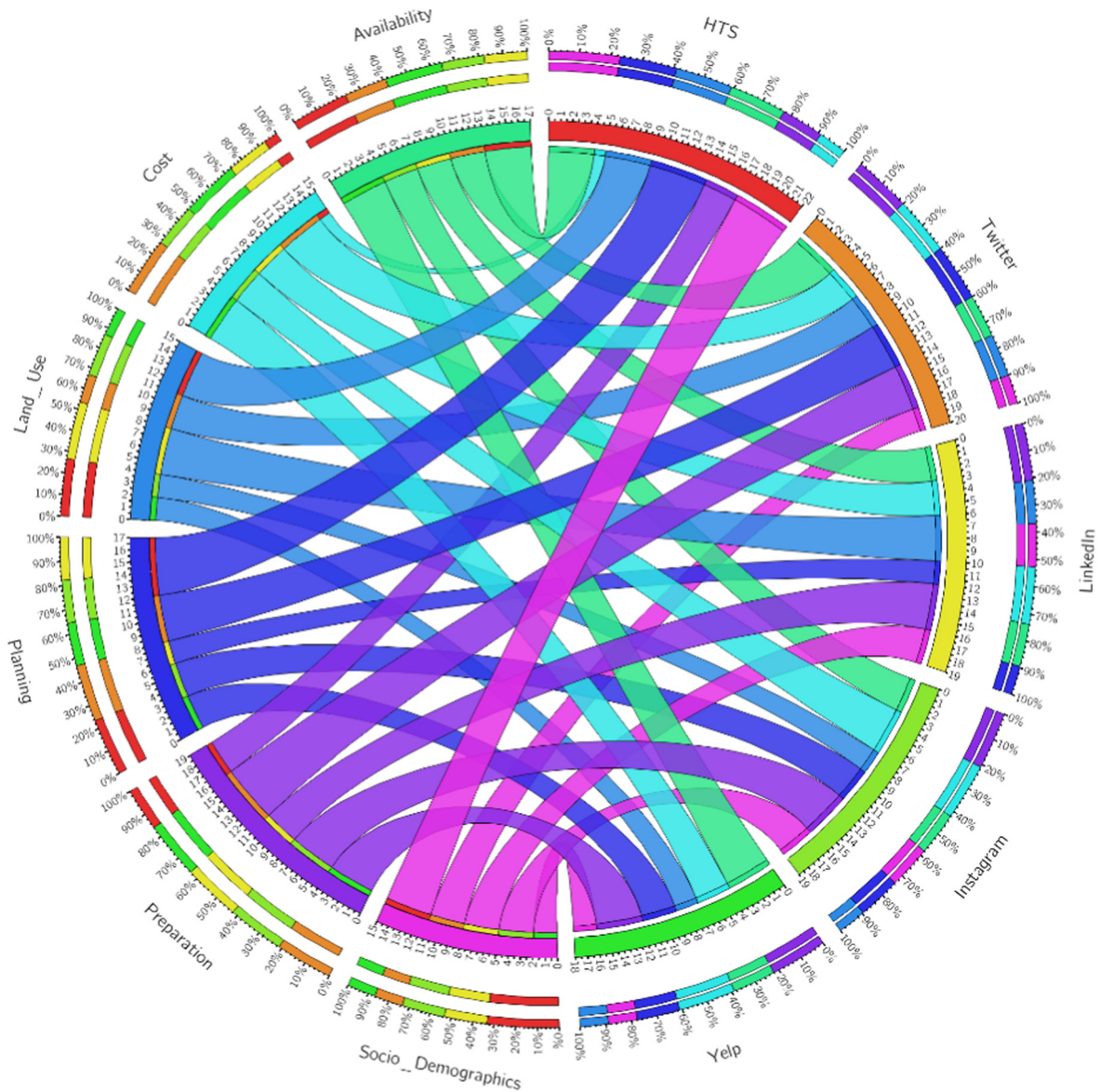


Fig. 5. Relationship between data sources and implementation techniques and potential related context.

without requiring the traveller to complete the travel diary. This effort can significantly decrease the data collection burden and cost (Byon et al., 2009).

Nonetheless the literature is quite slim (Cranshaw et al., 2012; Coffey and Pozdnoukhov, 2013; Steiger et al., 2014; Shang et al., 2016) when it comes to application of social media data courses to extract travel and socio-demographic information.

From what has been discussed so far, a pattern can be recognized toward using social media data in determining general mobility of travellers and aggregate travel models. Nonetheless, usefulness of such data source for disaggregate level modelling is yet to be explored. Therefore, it is useful to understand, if social media data is envisioned to be used for extracting information about individuals what level of accuracy can be expected. The following part discusses how social media can be used for extracting individual level information about travellers which can supplement household travel survey data.

## 5. Advanced travel demand models and social media data

Using social media data to extract information to be used in advanced travel demand models requires interpreting detailed information about travel diary of people. Advanced travel demand models like activity-based models include several components about day-to-day travel related decisions of individuals (Auld and Mohammadian, 2009). This section discusses some possible directions for using social media data as input to some elements of activity-based models.



### 5.1. Trip purpose

Trip purpose is one of the most essential travel attributes in travel demand modelling. Extracting information about the purpose of the activity from the text of tweets is a challenging task. This requires linguistic mining techniques which are still at their infancy although some techniques such as Latent Dirichlet Allocation (LDA) method is used widely in the literature (Kosala and Adi, 2012; Steiger et al., 2014). The complication is related to the fact that one person may tweet about a restaurant but not necessarily meaning to have an outdoor recreational eating activity. When the tweet is combined with the geolocation of the correspondent, it can facilitate extracting the purpose of the trip. The first step for developing understanding about how tweets can be interpreted to determine the purpose of a potential trip associated with is to build a data dictionary. Based on words used in a tweet and their correlation with different activities, activity purpose can be determined. Other than looking at words used in a tweet, combination of word in the sentences should be looked at which results in conceptually mining the text. This requires further advanced linguistic mining techniques which are not yet used in search engine. Data sources such as Yelp and Foursquare makes determining trip purpose easier as posts are categorised based on industry classifications of the visited place.

### 5.2. Departure time and activity location

Determining departure time, given the fact that tweets have a time tag makes the process of finding the departure time easier. The main complication relates to identifying tweets that are related to an activity. Further, the time tag might be for an activity happening after or before the time the tweet is posted. Nonetheless, it should be noted here that all tweets are related to an activity but the activity might be an indoor activity which is of secondary importance. Using Twitter data to develop understanding about in-home activities is a topic that will be further discussed later.

If an activity is identified to be associated with a tweet, based on the text (i.e., content) of tweet, it can be determined whether the activity has happened, is happening or is scheduled to happen in future. Preceding and succeeding a tweet can also help to determine the time of an activity and the departure time.

Similarly, activity location can be determined by looking at the text of tweet and the associated geocode, if provided. Unlike check-in (or in some cases, hash-tag) data for which the location of the activity can be easily determined, the geocode coming with tweets does not necessarily imply the location of the activity, because the respondent might have tweeted before the activity happened or even after the fact. As a result, a combination of GIS methods (to link land use data to the location of tweets) and data mining is required to determine the location of an activity that is related to a tweet.

### 5.3. Travel route, activity duration and traffic condition

If nothing is noted in the tweet text about the route or duration of the activity, then, extracting information for these two travel attributes require multiple tweets to be jointly considered to determine a chain of tweets. This is possible specifically for people who frequently tweet as it makes the record of their tweet like a GPS tracker which is also associated with notes about each location or the preceding and succeeding point.

Similarly, traffic condition and travel time on different links can be determined by looking at travelers tweeting while travelling, the same way GPS information is used. The significant advantage of Twitter data to GPS data is that each point is accompanied by a note which may contain more information about the traffic condition.

### 5.4. Mode of transport and party composition

Similar to trip purpose, mode of transport and party composition should be determined using text mining and natural language processing approaches. Nonetheless, constructing a dictionary for this purpose is not as complicated.

### 5.5. Socio-demographic attributes

#### 5.5.1. Twitter

Using data mining techniques, it is possible to determine the location of home, job and school the same way activity locations are determined. Further, it is more likely that more tweets are posted from home, school and work. Keywords related to home, work and school are expected to be present more in tweets posted from these locations. Therefore by looking at a history of tweets by one respondent it becomes possible to figure out these three important locations.

#### 5.5.2. LinkedIn

A panel data for job relocation history of people is a costly type of data requiring a long term project to follow people and observe their relocation pattern. Alternatively a retrospective data can be collected for this purpose. The freely available data from LinkedIn can be used for this purpose in which people retrospectively report their job/school relocation pattern. The significance of using LinkedIn data pertains to the accuracy of information about job/school location, relocation timing and type. By mining the Twitter and LinkedIn data some demographic information such as gender, estimated income, social network type and age can be determined.

### 5.6. Other aspects of an advanced travel demand model

Other than the above-mentioned advantages of using social media data three other aspects of transport planning and travel demand modelling can be discussed that are more related to the activity-based modelling scheme.

- a. In-home activity data: There is a significant challenge in the area of travel demand modelling to obtain data about in-home activities of people. This is important to travel demand modellers, specifically activity-base modellers, because there is a trade-off 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 modellers and planners.
- b. 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.
- c. Future activities: 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.

## 6. Summary

This paper presented a bibliometric analysis of the transport literature with a focus on applications of social media data. The rapidly growing part of the literature of transport where social media plays a significant role has experienced a sharp jump in the last few years creating expectations that more is expected to be seen on application of social media data in different domains of transport engineering including travel demand estimation, operation, traffic management, network design and planning. A comprehensive review of the literature was discussed in this paper revealing that studying human mobility has been the most dominant topic related to transport where social media has been used.

A special focus was given to explore the cons and pros of using social media data in travel demand modelling. A qualitative survey reflecting the opinion of several experts in the area of travel demand modelling was discussed.

Based on the in-depth and critical review of the literature and the results of the survey the advantages and disadvantages of the using social media data can be summarized as follows.

### 6.1. Advantages

The survey results showed that there seems to be significant potentials for using social media data to develop models for estimating travel demand, managing operation and long term planning purposes, while special caution is required in doing so due to the biases associated with social media data. This bias is becoming less severe as social media users are growing making the sample a close representative of the population. Nonetheless, sampling bias correction methods can be employed to adjust this bias especially with the use of supplementary databases in which demographics of users, for example, of the big media data of interest are reported (Wesolowski et al., 2013). Further, sampling bias can be adjusted as it has been done by travel demand modellers (Lee et al., 2015) by using the output of a model developed based on social media movements and observed travel patterns.

Acquisition cost of obtaining the data is the most important aspect making these data sources appealing. Nonetheless, the total cost of having useful data for modelling purposes is significantly larger than the acquisition cost.

Social media data encompasses information that is revealed by users in realistic situations, especially if the data comes with text content, then such data is free from sampling, surveying or laboratory biases.

Social media usage is growing due to the inflation in smart phones and tablets usage. As a result, such data is quite commonly associated with geo-location information which is valuable information for transport planning, management and operation purposes.

### 6.2. Disadvantages

The most challenging issue in front of using social media data pertains to complications associated with extracting useful information from the content of the data which requires employment of advanced text and data mining techniques. This is even more challenging if such data sources are considered for estimating demand at the individual level versus aggregate and zone-based level for which acceptable modelling structures have already been proposed by travel demand modellers. When used for demand estimation purposes, social media data should be adjusted for over-representation of such system



users. Further, such data is over-represented by discretionary and leisure activities. As a result, the tour of activities should be inferred which leaves a methodological challenge ahead of travel demand modellers. In short, although the acquisition cost of social media data is quite low (almost free), the processing cost is still quite high before it becomes useful for planning purposes. Nonetheless, the literature shows a vibrant environment where researchers are trying to extract useful information from the big data has become available to them through social media resources.

Further, an important concern regarding social media data is individual-specific information that cannot be shared publicly unless the “access to public” is released by the individual or the owner of the data (Smith et al., 2012). Access to such privately identifiable information (PII) is restricted by stringent confidentiality clauses, as a result any analysis conducted on personalized big data such social media data requires careful attention to aggregate the geotagged information of people so that it is not identifiable (Henne et al., 2013; Chen and Zhang, 2014). Nonetheless, travel demand models essentially reflect only the overall behaviour of individuals in the form of a set of parameters determining the impact of used explanatory variables. As a result, no information can be tracked back to individuals if modelling results are reported which is the case in the majority of travel demand models including activity-based models.

## References

- Auld, J., Mohammadian, A., 2009. Framework for the development of the agent-based dynamic activity planning and travel scheduling (ADAPTS) model. *Transp. Lett.* 1 (3), 245–255.
- Beirão, G., Cabral, J.S., 2007. Understanding attitudes towards public transport and private car: a qualitative study. *Transp. Policy* 14 (6), 478–489.
- Byon, Y.-J., Abdulhai, B., Shalaby, A., 2009. Real-time transportation mode detection via tracking global positioning system mobile devices. *J. Intelligent Transp. Sys* 13 (4), 161–170.
- Cebelak, M.K., 2013. Location-based social networking data: doubly-constrained gravity model origin-destination estimation of the urban travel demand for Austin, TX. Master's Thesis. The University of Texas at Austin.
- Chang, J., Sun, E., 2011. Location3: How users share and respond to location-based data on social networking sites. In: Proceedings of the International Conference on the Weblogs and Social Media (ICWSM).
- Chen, C., 2006. CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature. *J. Am. Soc. Inform. Sci. Technol.* 57 (3), 359–377.
- Chen, C.P., Zhang, C.Y., 2014. Data-intensive applications, challenges, techniques and technologies: a survey on Big Data. *Inform. Sci.* 275, 314–347.
- Cheng, Z., Caverlee, J., Lee, K., Sui, D.Z., 2011. Exploring millions of footprints in location sharing services. *ICWSM*, 81–88.
- Cho, E., Myers, S.A., Leskovec, J., 2011. Friendship and mobility: user movement in location-based social networks. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1082–1090.
- Coffey, C., Pozdnoukhov, A., 2013. Temporal decomposition and semantic enrichment of mobility flows. In: Proceedings of the 6th ACM SIGSPATIAL International Workshop on Location-Based Social Networks. ACM.
- Collins, C., Hasan, S., Ukkusuri, S.V., 2013. A novel transit rider satisfaction metric: rider sentiments measured from online social media data. *J. Public Transp.* 16 (2), 21–45.
- Cramer, H., Rost, M., Holmquist, L.E., 2011. Performing a check-in: emerging practices, norms and conflicts' in location-sharing using foursquare. In: Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services.
- Cranshaw, J., Schwartz, R., Hong, J.L., Sadeh, N.M., 2012. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. In: *ICWSM*.
- Davis, A., Goulias, K.G., 2015. Building better activity-travel behavior models with harvested situational awareness information from social media and place surveys. *IATBR 2015, WINDSOR*.
- Edler, D., Rosvall, M., 2015. The MapEquation software package, available online at <<http://www.mapequation.org>>.
- Francis, R.C., McGee, J.P., Sainati, R.A., Sheehan, Jr., R.L., Tong, S.K.K., 2003. U.S. Patent No. 6,600,418. Washington, DC: U.S. Patent and Trademark Office.
- Fu, K., Nune R., Tao, J.X., 2015. Social media data analysis for traffic incident detection and management. In: Transportation Research Board 94th Annual Meeting, Washington D.C., 14–4022.
- Gao, H., Tang, J., Liu, H., 2012. Exploring social-historical ties on location-based social networks. In: *ICWSM*.
- Gao, H., Tang, J., Hu, X., Liu, H., 2013. Exploring temporal effects for location recommendation on location-based social networks. In: Proceedings of the 7th ACM conference on Recommender systems, ACM, pp. 93–100.
- Golder, S.A., Macy, M.W., 2014. Digital footprints: opportunities and challenges for online social research. *Sociology* 40 (1), 129.
- Hasan, S., 2013. Modeling urban mobility dynamics using geo-location data. PhD Dissertation, Purdue University.
- Hasan, S., Ukkusuri, S.V., 2013. Social contagion process in informal warning networks to understand evacuation timing behavior. *J. Public Health Manage. Pract.* 19, S68–S69.
- Hasan, S., Ukkusuri, S.V., 2014. Urban activity pattern classification using topic models from online geo-location data. *Transp. Res. Part C: Emerging Technol.* 44, 363–381.
- Hasan, S., Ukkusuri, S.V., 2015. Location contexts of user check-ins to model urban geo life-style patterns. *PLoS ONE* 10 (5), e0124819.
- Hasan, S., Zhan, X., Ukkusuri, S.V., 2013. 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. ACM.
- Hasan, S., Ukkusuri, S.V., Zhan, X., 2016. Understanding social influence in activity-location choice and life-style patterns using geo-location data from social media. *Front. ICT* 3, 10.
- Henne, B., Szongott, C., Smith, M., 2013. SnapMe if you can: privacy threats of other peoples' geo-tagged media and what we can do about it. In: Proceedings of the sixth ACM conference on Security and Privacy in Wireless and Mobile Networks, pp. 95–106.
- Jin, P.J., Cebelak, M., Yang, F., Ran, B., Walton, C.M., Zhang, J., 2014. Location-based social networking data: exploration of use of doubly constrained gravity model for origin-destination estimation. In: The 93rd Annual Meeting of Transportation Research Board, Washington DC, 14–5314.
- Jurdak, R., Zhao, K., Liu, J., Aboujaoude, M., Cameron, M., Newth, D., 2015. Understanding Human Mobility from Twitter. *PLoS ONE* 10 (7), e0131469.
- Kaigo, M., 2012. Social media usage during disasters and social capital: Twitter and the Great East Japan earthquake. *Keio Commun. Rev.* 34, 19–35.
- Kosala, R., Adi, E., 2012. Harvesting real time traffic information from Twitter. *Proc. Eng.* 50, 1–11.
- Katakis, I., Tsoumakas, G., Vlahavas, I., 2008. Multilabel text classification for automated tag suggestion. *ECML PKDD discovery challenge*, 75.
- Lee, J.H., Davis, A., Yoon, S.Y., Goulias, K.G., 2016. Activity Space Estimation with Longitudinal Observations of Social Media Data. In: Transportation Research Board 95th Annual Meeting, 16–0070.
- Lee, J.H., Gao, S., Janowicz, K., Goulias, K.G., 2015. Can Twitter data be used to validate travel demand models?, In: *IATBR, WINDSOR*.
- Lian, D., Xie, X., 2011. Collaborative activity recognition via check-in history. In: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks. ACM.
- Lindsay, B.R., 2011. Social media and disasters: current uses, future options, and policy considerations. In: CRS Report for Congress, R41987.
- Luong, T.T., Houston, D., 2015. Public opinions of light rail service in Los Angeles, an analysis using Twitter data. In: Conference 2015 Proceedings Philadelphia USW.

- Maghrebi, M., Abbasi, A., Rashidi, T.H., Waller, T., 2015. Complementing travel diary surveys with Twitter data: application of text mining techniques on activity location, type and time. In: *IEEE 17th International Conference on Intelligent Transportation Systems (ITSC)*, 2015. IEEE.
- Mai, E., Hranac, R., 2013. Twitter interactions as a data source for transportation incidents. In: *Proc. Transportation Research Board 92nd Ann. Meeting*, Washington D.C., 16–1636.
- Miller, E., Lee-Gosselin, M., Habib, K.N., Morency, C., Roorda, M., Shalaby, A., 2014. A Framework for Urban Passenger Data Collection. In: *10th International Conference on Transport Survey Methods*, Leura, Australia.
- Nik Bakht, M., Kinawy, S.N., El-Diraby, T.E., 2015. 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*, Washington DC, 15–0117.
- Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., Mascolo, C., 2012. Correction: a tale of many cities: universal patterns in human urban mobility. *PLoS ONE* 7 (9), 10.1371.
- Onnela, J.-P., Arbesman, S., González, M.C., Barabási, A.L., Christakis, N.A., 2011. Geographic constraints on social network groups. *PLoS ONE* 6 (4), e16939.
- Pender, B., Currie, G., Delbosc, A., Shiwakoti, N., 2014. Social media use during unplanned transit network disruptions: a review of literature. *Transport Rev.* 34 (4), 501–521.
- Pianese, F., An, X., Kawsar, F., Ishizuka, H., 2013. Discovering and predicting user routines by differential analysis of social network traces. In: *IEEE 14th International Symposium and Workshops World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, 2013, pp. 1–9.
- Rashidi, T.H., Kanaroglou, P., 2013. The Next Generation of Transportation Demand Models, Toward an Interdisciplinary Science. In: Miller, E.J., Roorda, M.J. (Eds.), *International Association for Travel Behaviour Research Book*. Emerald Group Publishing, Bingley, U.K., pp. 201–229.
- Rashidi, T.H., Mohammadian, A., Koppelman, F., 2012. Modelling interdependencies between vehicle transaction, residential relocation and job change. *Transportation* 38 (6), 909–932.
- Rashidi, T.H., Mohammadian, A., 2011. Household travel attributes transferability analysis: application of a hierarchical rule based approach. *Transportation* 38 (4), 697–714.
- Rashidi, T.H., Mohammadian, A., Zhang, Y., 2010. Effect of variation in household sociodemographics, lifestyles, and built environment on travel behavior. *Transp. Res. Rec.: J. Transp. Res. Board* 2156 (1), 64–72.
- Ribeiro, Jr., S.S., Davis, C.A., Oliveira, D.R.R., Meira, W., Gonçalves, T.S., Pappa, G.L., 2012. Traffic observatory: a system to detect and locate traffic events and conditions using Twitter. In: *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Location-Based Social Networks*. ACM, pp. 5–11.
- Rosvall, M., Bergstrom, C.T., 2010. Mapping change in large networks. *PLoS ONE* 5 (1), e8694.
- Ruths, D., Pfeffer, J., 2014. Social media for large studies of behavior. *Science* 346 (6213), 1063–1064.
- Sakaki, T., Okazaki, M., Matsuo, Y., 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In: *Proceedings of the 19th International Conference on World Wide Web*. ACM.
- Schweitzer, L., 2014. Planning and social media: a case study of public transit. *J. Am. Plan. Assoc.* 80 (3), 218–238.
- Shang, S., Guo, D., Liu, J., Zheng, K., Wen, J.-R., 2016. Finding regions of interest using location based social media. *Neurocomputing* 173, 118–123.
- Smith, M., Szongott, C., Henne, B., von Voigt, G., 2012. Big data privacy issues in public social media. In: *2012 6th IEEE International Conference on Digital Ecosystems and Technologies (DEST)*, pp. 1–6.
- Steiger, E., Albuquerque, J.P., Zipf, A., 2015. An advanced systematic literature review on spatiotemporal analyses of twitter data. *Trans. GIS* 19 (6), 809–834.
- Steiger, E., Ellersiek, T., Zipf, A., 2014. Explorative public transport flow analysis from uncertain social media data. In: *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, 2014. ACM.
- Steur, R., 2015. Twitter as a spatio-temporal source for incident management. Faculty of Geosciences Theses, Master thesis.
- Tasse, D. and Hong J.I. 2014. Using social media data to understand cities. In: *Proceedings of NSF Workshop on Big Data and Urban Informatics*, August 2014.
- Tian, Y., Zmud, M., Chiu, Y.-C., Carey, D., Dale, J., Smarda, D., Lehr, R., James, R., 2016. Quality assessment of social media traffic reports – a field study in Austin, Texas. In: *Transportation Research Board 95th Annual Meeting*, 2016, Washington D.C., 16–6852.
- Ukkusuri, S., Zhan, X., Sadri, A., Ye, Q., 2014. Use of social media data to explore crisis informatics: Study of 2013 Oklahoma Tornado. *Transp. Res. Rec.: J. Transp. Res. Board* 2459, 110–118.
- Wanichayapong, N., Pruthipunyaskul, W., Pattara-Atikom, W., Chaovalit, P., 2011. Social-based traffic information extraction and classification. In: *11th International Conference on ITS Telecommunications (ITST)*, 2011. IEEE, pp. 107–112.
- Wesolowski, A., Eagle, N., Noor, A.M., Snow, R.W., Buckee, C.O., 2013. The impact of biases in mobile phone ownership on estimates of human mobility. *J. Royal Soc. Interface* 10 (81), 20120986.
- Wilde, N., Hänsel, K., Haddadi, H., Alomainy, A., 2015. Wearable Computing for Health and Fitness: Exploring the Relationship between Data and Human Behaviour. *arXiv preprint arXiv*, pp. 1509.05238.
- Williams, S., Currid-Halkett, E., 2014. Industry in motion: using smart phones to explore the spatial network of the garment industry in New York City. *PLoS one* 9 (2), e86165.
- Wu, X., Zhu, X., Wu, G.Q., Ding, W., 2014. Data mining with big data. *IEEE Trans. Knowledge Data Eng.* 26 (1), 97–107.
- Yin, Z., Fabbri, D., Rosenbloom, S.T., Malin, B., 2015. A scalable framework to detect personal health mentions on Twitter. *J. Med. Internet Res.* 17 (6).
- Zhang, Y., Mohammadian, A., 2010. Bayesian updating of transferred household travel data. *Transp. Res. Rec.: J. Transp. Res. Board* 2049, 111–118.
- Zhu, Z., Blanke, U., Tröster, G., 2014. Inferring travel purpose from crowd-augmented human mobility data. In: *Proceedings of the First International Conference on IoT in Urban Space*, 2014. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).