Data for Leisure Travel Demand from Social Networking Services

Emmanouil Chaniotakis[*,**], Constantinos Antoniou[*], and Evangelos Mitsakis[**]

[*]National Technical University of Athens
[**]Centre for Research and Technology Hellas – Hellenic Institute of Transport

Abstract

Leisure activities travel demand plays an important role in transportation, as it accounts for a significant amount of trips performed. Due to data limitations, the examination of this trip purpose has received moderate attention so far. The evolution of pervasive systems in conjunction to the increased use of Social Media have provided data that can be used for the development of leisure activities travel demand models. This study provides an initial investigation of the data collected from Social Media and the potential of using it for leisure activities demand modelling. A data analysis framework is provided that distinguish user classes of residents and tourists and investigates the temporal and spatial patterns as well as the probability of deriving activity locations.

1 Introduction

A subject that has lately received attention in transportation modelling is the linkage between leisure activities and travel demand. Trips related to leisure activities and social interactions differ from the widely studied trip purposes (i.e. work and shopping) (Carrasco et al., 2008) and also constitute an important part of the trips performed (Bhat and Gossen, 2004). One of the reasons that this linkage has received little attention so far is the non–existence of detailed data that could allow for a better understanding of the linkage between leisure activities and travel demand (Axhausen, 2008; Carrasco et al., 2008).

Lately, some of the limitation of the conventional types of data and data collection methods can be overcome with the evolution of pervasive systems, such as GPS handsets and cellular networks –that allow sharing of geo-referenced information. Going a step further, the use of Social Media (SM) originated data (such as Facebook, Twitter, Flickr and Google+) allow for sharing of information for activities and users interactions (Grant-Muller et al., 2014; Pei et al., 2014). Data originating both directly from pervasive systems and from SM give the opportunity to the scientific community to steer the focus from overcoming limitations of available data, towards utilizing the use of the increasingly available data, namely the Big Data (Buckley and Lightman, 2015). The validity and the limitations of a variety of this type of data is being investigated, for the identification of travel patterns and investigation of travel behaviour (see Witlox, 2015; Reades et al., 2007).

This study focuses on Social Media and the potential of using it in transportation research. Starting from the question what is Social Media; several definition exist, due

to their dynamic character and the spectrum of services offered (i.e. Bregman, 2012; Kaplan and Haenlein, 2010; Leonardi et al., 2013; Bradley, 2010). The definition by Boyd and Ellison (2007) is considered as mostly representative, in the context of this research; as such adopted here:"Social Media are services which allow users to: a) maintain a public or semi-public personal profile b) articulate a social network by making connections to other users, and c) browse and react to connections". In addition to the definition, Kietzmann et al. (2011) developed a research context by proposing a honeycomb framework that composes of 7 different functionalities of Social Media: a) presence b) sharing c) conversations d) groups e) reputation f) identity g) relationships. Each Social Media site aims at a combination of the above-mentioned, with a tendency to concentrate at three or four functionalities.

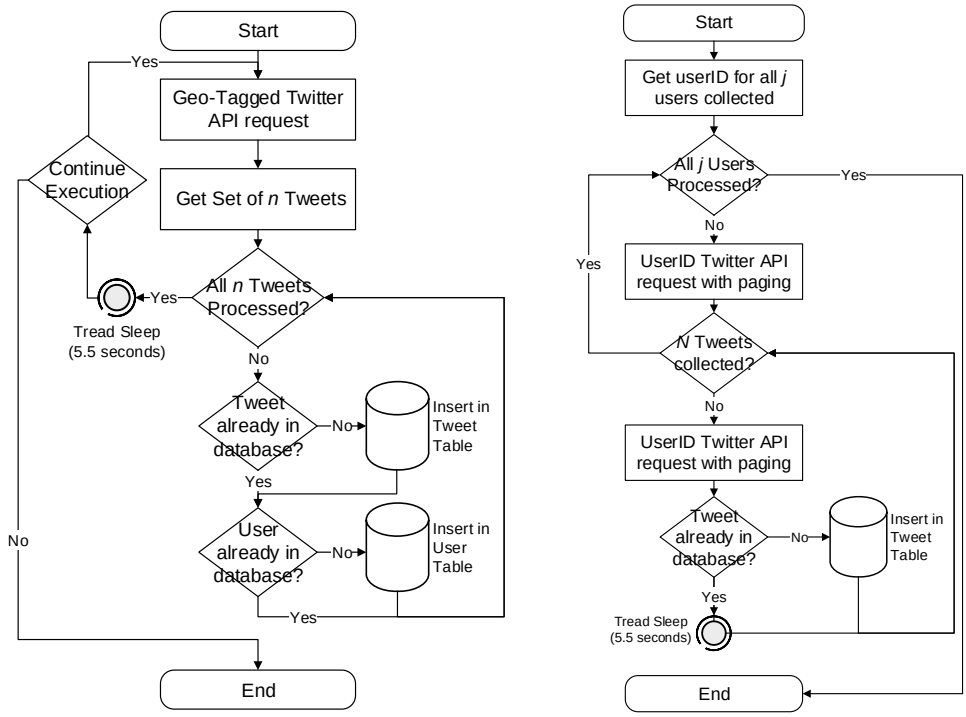
The use of Social Media in transportation studies can be summarised in the following areas: (1) information sharing and communication with costumers and (2) data collection and mining. Previous studies presented the potential of using this data source. To name but a few, Wanichayapong et al. (2011) used synthetic analysis to classify traffic related incidents in spatial dimensions. Schulz et al. (2013) presented the identificaiton of road incidents based on semantic enrichment and text classification. Abel et al. (2012) presented Twitcident for automated collection and filtering of emergency related information, Pereira et al. (2013) focused on the identification of non–recurrent events from web–pages in order to estimate the resulted traffic. Finally, Kumar et al. (2014) used Twitter to detect poor road conditions in order to be used as a "road hazard alert system". On a different context, Cheng et al. (2011) presented some first statistics on mobility patterns from social media data. Later, (Liu et al., 2014) presented some similar findings on human mobility patterns in China using check–in data. Hawelka et al. (2013) used Tweet for the exploration of global mobility patterns, focusing on tourism from and to different countries. On a more transport modelling oriented approach, Yang et al. (2014) presented a framework for the dynamic estimation of Origin Destination matrices using social media data. To the best of the authors knowledge there is no research on the exploration of leisure activities using Social Media, while it is believed that the characteristics of sharing are ideal for such an investigation.

In this paper, we examine the potential of using SNS originated data for the estimation of travel demand for leisure activities. We focus on the inference of transportation–related information from Twitter together with important initial results from the analysis of empirical data. We base our work and conclusions on the analysis of data collected from three European cities of varying size (London, Athens, and Thessaloniki), during a period of 4 months. Given the characteristics of the data sample, we focus on the Athens case study. The paper is structured as follows: first the social media data collection and sample selection is presented (Section 2). Afterwards, a short data analysis is included in order to get a better understanding the nature of the tweeting process (Section 3), followed by the evaluation of the leisure activities in the concept of naturally derived areas on interest (Section 4). Finally the findings are discussed and the future worked is outlined (Section 5).

2 Social Media Data & Sample Specification

Data was collected from three cities of different sizes, with two of them located in Greece (Athens and Thessaloniki) and one located in the United Kingdom (London). The data collection methodology used includes two components (Figure 1): first data is collected randomly from the Twitter API for a specific area (Real–Time Data Collection (RTDC) component – Figure 1a) and second, data is collected from those users to collect a number of Tweets per user (Historic Data Collection (HDC) component – Figure 1b) Chaniotakis and Antoniou (2015). This methodology allows the collection of data from

an area which essentially form a user sample and later, based on that user sample, the derivation of a set of locations visited by each individual. Consequently, locations of geo–tagged Tweets and trips made are not restricted in the geographical area imposed when collecting data using the RTDC component.



(a) Real–Time Tweeter Mining

(b) Historical Data Mining

Figure 1: Tweeter Data Collection adopted by (Chaniotakis and Antoniou, 2015)

The data collection methodology was implemented in Java Runtime Environment using the Twitter4J library developed by Yamamoto (2007). Data was stored in a two–table Structured Query Language (SQL) (*Tweets* and *Users*) that included the following fields: a) *Tweets* Table: *tweetID, usedID, tweetDate, tweetText, latitude, longitude, place, dateAdded* b) *Users* Table: *userID, userName, dateAdded, description*. After 4 months of continuous data collection, and given the number of users and Tweets collected using the RTDC component from each area (Table 1), it was chosen to use data originating from Athens, as it provided a suitable amount of data, taking into account the required computational time. This choice was also made upon the existence of corresponding data, for the city of Athens and familiarity with the region.

Table 1: Number of Users and Tweets per City using the RTDC methodology

City	Number of Tweets	Number of Users
Athens	233243	13848
London	6745852	334521
Thessaloniki	64827	9420

Based on the above–mentioned, the HDC methodology component was applied for

the collection of 1000 latest Tweets per users. The implementation of the HDC component resulted in a 1.9 GB database that contained 9,085,152 Tweets (both geo–tagged and not geo–tagged) from the 13,848 users. From those Tweets 1,351,165 were geo–tagged (approx. 15%).

The sample specification focused on the distinction between inhabitants and tourists as well as the choice of users who can illustrate a high degree of information. This characterization is not always straightforward as information on the place of residence is not provided by Twitter. For that reason and in order to avoid misconceptions due to users' posting only a very small number of Tweets; users with above average posting activity (total number of geotagged Tweets posted larger than 97 Tweets) were selected. For the subset of users selected (that included 3917 users) the frequencies of the number of Tweets in the Metropolitan Athens' area to the total number of Tweets per user were estimated (Figure 2). As it is illustrated in Figure 2, there is a number of users of whom Tweets are mainly posted outside the boundaries of the Metropolitan Geographical Area, that can be characterised as tourists (percentage of in examined area Tweets <0.25), a set of users of whom Tweets are mainly posted inside the Metropolitan Geographical Area that can characterised as residents (percentage of in examined area Tweets <0.75) and a set of users of whom the residence location is unclear (percentage of in examined area Tweets >0.25 and <0.75). This classification in 3 user classes lead to 3 subsets – namely *tourists*, *residents*, *unclear*– that contained 2232 users, 1169 users and 516 users respectively. Note that in this study we only focus on the following two user classes: *residents* and *tourists*.

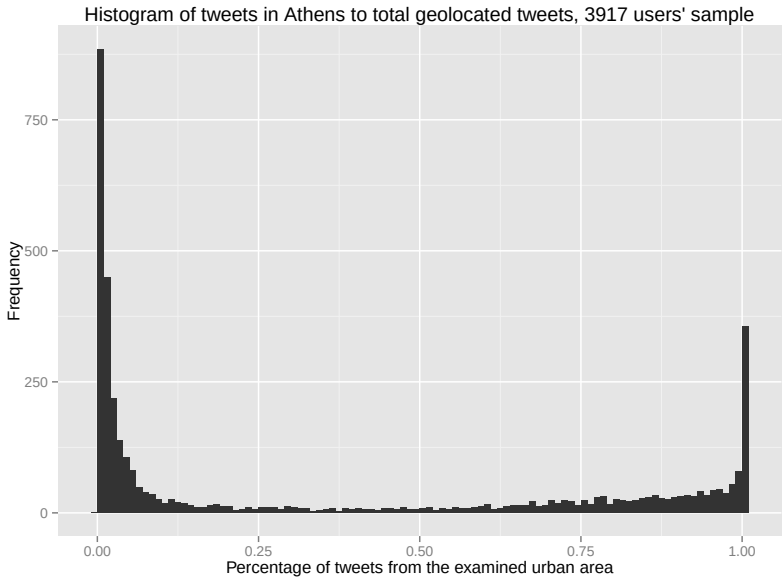


Figure 2: Frequencies of the number of Tweets in the Metropolitan Athens' area to the total number of geo–tagged Tweets per user (bins width = 0.01) for the above average posting activity sample

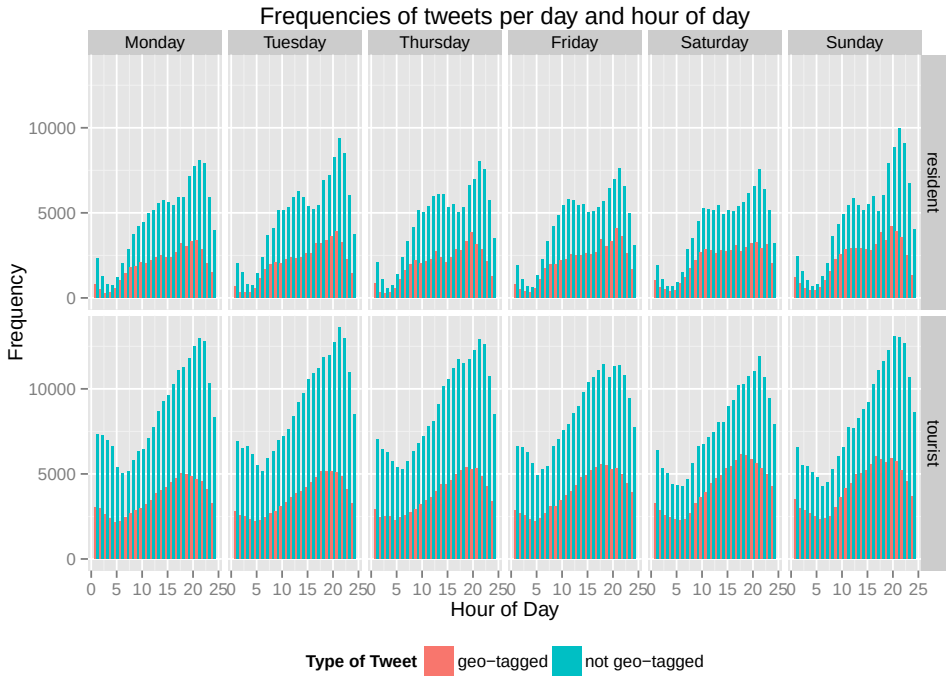


Figure 3: Temporal Distribution of Tweets in the Athens Metropolitan area based on the user class, the day of week and the hour of day

3 Data Analysis

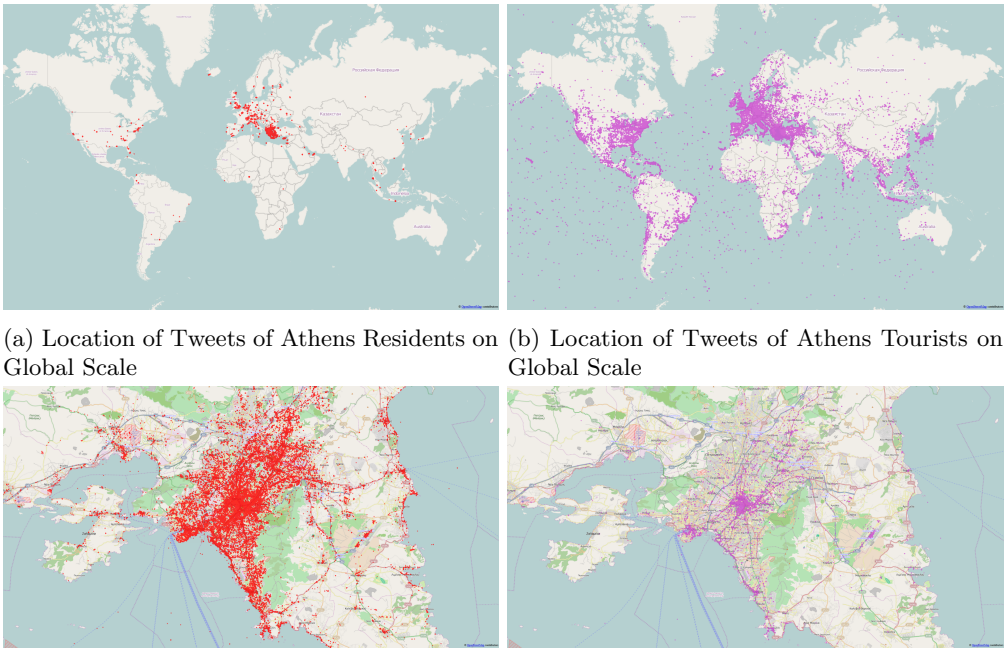
3.1 Temporal Data Analysis

A temporal statistical evaluation was performed for the exploration of the temporal patterns that defined Tweeting, in the three different classes per day of week and hour of day (Figure 3). Concerning the hourly posting activity, there is a almost similar pattern during all days and for each user class. Users of the *residents* user class, are more active during non–working hours with the peak to be found at around 21:00 o'clock. The posting pattern of the *tourists* user class illustrate a rather different behaviour. There is an increasing posting activity during the day, that peaks at around 21:00 o'clock. Temporal data analysis provides a starting point on the examination of activities related to posting at social media. It is indicated that most Tweets are posted during evening hours that allow for forming the hypothesis that posting at social media can be related to leisure activities. Note that geo–tagged and not geo–tagged Tweets illustrate almost similar temporal patterns, among the different user classes.

3.2 Spatial Analysis

The *residents* and *tourists* classes derived were examined upon the spatial distribution of their posting activity. As the HDC data collection methodology was applied Tweets were collected based on users (collected from the first methodology); as such geo–tagged Tweets were collected from places all over the world. Figure 4a and 4b illustrate the different spatial distribution of posting activity on a world scale. Users categorised as residents are found to have a higher density of Tweets in the area of Greece.

On the other hand, users categorised as tourists for the area of Athens are found to post Tweets from various places around the world with higher density of Tweets in Europe and the USA. This difference is also evidenced when examining the spatial patterns of posting activity, in the metropolitan area of Athens. Tweets from users characterised as residents are distributed in the city area, while Tweets from users characterised as tourists are mainly gathered at places that are common tourist attractions, or ports and airports.



(a) Location of Tweets of Athens Residents on Global Scale

(b) Location of Tweets of Athens Tourists on Global Scale

(c) Location of Tweets of Athens Residents in Athens Metropolitan Area

(d) Location of Tweets of Athens Tourists in Athens Metropolitan Area

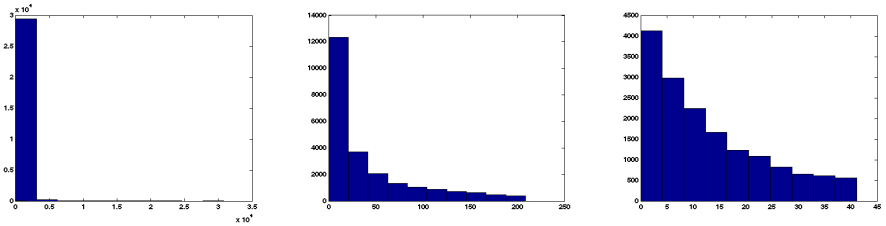
Figure 4: Location of Tweets collected for *residents* and *tourists* user classes

4 Areas of Interest

The inference of activities derived from social media was built upon the premises of the Natural Cities (Jiang et al., 2015) in order to identify natural Areas of Interest (AoI) using Social Media spatial data. The natural definition of cities is based on the fact that in many cases, data illustrate an imbalance that can be described by a heavy–tailed distribution (L-shaped) (Jiang and Miao, 2014). The data is divided in classes iteratively, until the data from classes do no illustrate a heavy–tailed distribution. This methodology is named head/tail methodology. In this study we focus on the part of the data characterized as the head of the distribution and we use them to identify locations which have a high density of Tweets for the *Residents* and *Tourists* user class.

The implementation requires at first, the definition of Triangulated Irregular Network (TIN) that connects Tweet points with lines and creates triangles. Then the length of those lines is used for the identification of areas with defined by lines of small length indicating places of high density posting activity. Figure 5 illustrates the implementation of the head/tail methodology for the *tourists* class. It should be noted that especially for the *residents* class the inference of Areas of Interest should take into account the

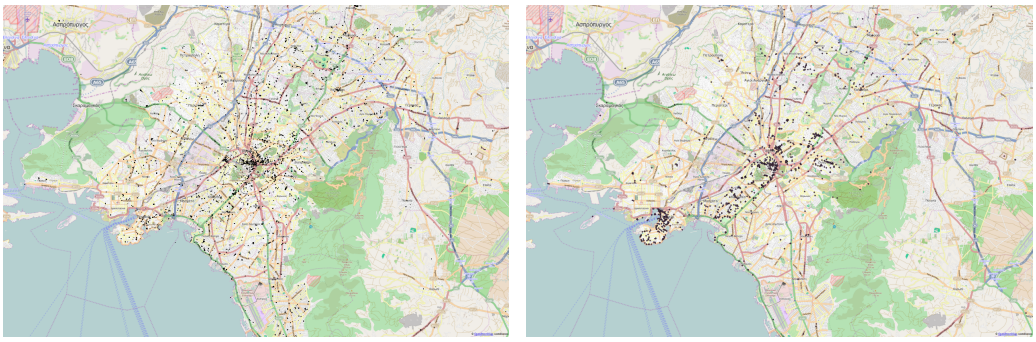
fact that those users might post repeatedly from their homes or places of work, however not included here. The finally selected sample included lines with lengths under 12.83 and 41.09 meters for *residents* and *tourists* user classes respectively.



(a) TIN Length (b) TIN Length under 208.3 meters (c) TIN Length under 41.09 meters

Figure 5: Selected histograms illustrated the head/tail distribution

Those were compared with Points of Interest (POIs) for the city of Athens from Openstreetmap (OSM) based on its *tagging* system. Figure 6 illustrate the inference of Natural Areas of Interest derived for residents and Points of Interest from OSM. It is found that Natural Areas of Interest include a vast majority of POIs from OSM that are characterized as bars, cafe, cinema, fast–food, music–venues, night clubs and restaurants. Furthermore, there are areas, that are not described as POIs, which are located at areas defined as recreational land uses areas. Those AoI can be safely characterized as areas that are visited by users but have not yet been added in OSM.



(a) Natural Areas of Interest in the city of Athens (b) Leisure POIs from OSM for the city of Athens

Figure 6: Natural Areas of Interest and Points of Interest from Social Media Data and OSM for residents of Athens

Concerning tourist the inference of activities resulted on larger Areas of Interest, due to the smaller number of tweets in the city of Athens (approx. 15400). In this case the activities were not examined upon leisure amenities but tourists attractions again defined as POIs from OSM. The results of the inference are presented in Figure 7. As it is clearly presented a large majority of tourists attractions (approx. 71%) are within the boundaries of the Areas of Interest. This finding further supports the fact that Social Media and specifically Twitter is used to share leisure activities in the areas visited.

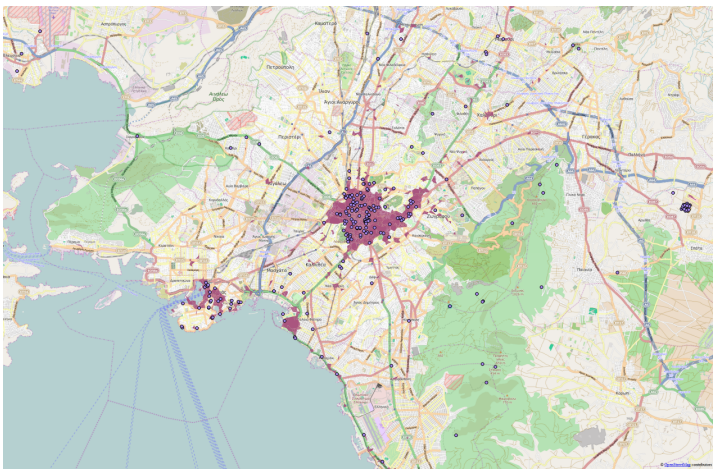


Figure 7: Natural Areas of Interest and Points of Interest from Social Media Data and OSM for tourists of Athens

5 Conclusions and Future Work

This research presented a part of the investigation of leisure activities travel demand based on social media data. User classes of *tourists* and *residents* are defined for which the evaluation of the type of visited areas is based on existing POIs from Open Street Maps. Exploratory work has been performed, in order to identify the types of activities related to Social Media posting activity. It has been found that many of the collected tweets are also leisure–related POIs. This feature sets the ground for further investigation of leisure activities from Social Media data as it confirms the hypothesis that posting geo–tagged tweets are connected to leisure activities.

As this study is exploratory, its finding are the basis for further research. The first domain is the further characterization of dynamic Areas of Interest as a concept that could eventually replace POIs and allow for a further specification of leisure attractions. Another important aspect is the inference of social network that can be derived from Social Media for the exploration of the impact that social network have to travel behaviour.

Acknowledgements

This research has been supported by the Action: ARISTEIA–II (Actions Beneficiary: General Secretariat for Research and Technology), co–financed by the European Union (European Social Fund ESF) and Greek national funds project. Authors wish to thank professor Bin Jiang for his contribution to the work performed for this paper.

References

Abel, F., Hauff, C., Houben, G.-J., Stronkman, R., and Tao, K. (2012). Twitcident: Fighting fire with information from social web streams. In *Proceedings of the 21st International Conference Companion on World Wide Web*, WWW '12 Companion, pages 305–308, New York, NY, USA. ACM.

Axhausen, K. W. (2008). Social networks, mobility biographies, and travel: survey challenges. *Environment and planning. B, Planning & design*, 35(6):981.

Bhat, C. R. and Gossen, R. (2004). A mixed multinomial logit model analysis of weekend recreational episode type choice. *Transportation Research Part B: Methodological*, 38(9):767 – 787.

Boyd, D. M. and Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230.

Bradley, A. J. (2010). A new definition of social media. *Gartner blog network*, 7.

Bregman, S. (2012). *Uses of social media in public transportation*, volume 99. Transportation Research Board.

Buckley, S. and Lightman, D. (2015). Ready or not, big data is coming to a city (transportation agency) near you. In *Transportation Research Board 94th Annual Meeting*, number 15-5156 in TRB2014.

Carrasco, J. A., Hogan, B., Wellman, B., and Miller, E. J. (2008). Collecting social network data to study social activity-travel behavior: an egocentric approach. *Environment and planning. B, Planning & design*, 35(6):961.

Chaniotakis, E. and Antoniou, C. (2015). Use of Geotagged Social Media in Urban Settings: Empirical Evidence on its Potential from Twitter. In *Proceedings of Intelligent Transportation Systems (ITSC) 2015 IEEE (accepted for publication)*.

Cheng, Z., Caverlee, J., Lee, K., and Sui, D. Z. (2011). Exploring millions of footprints in location sharing services. *ICWSM*, 2011:81–88.

Grant-Muller, S. M., Gal-Tzur, A., Minkov, E., Nocera, S., Kuflik, T., and Shoor, I. (2014). Enhancing transport data collection through social media sources: methods, challenges and opportunities for textual data. *IET Intelligent Transport Systems*.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., and Ratti, C. (2013). Geo-located Twitter as the proxy for global mobility patterns.

Jiang, B. and Miao, Y. (2014). The Evolution of Natural Cities from the Perspective of Location-Based Social Media. *ArXiv e-prints*.

Jiang, S., Alves, A., Rodrigues, F., Ferreira Jr., J., and Pereira, F. C. (2015). Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Computers, Environment and Urban Systems*.

Kaplan, A. M. and Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, 53(1):59 – 68.

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., and Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, 54(3):241–251.

Kumar, A., Jiang, M., and Fang, Y. (2014). Where not to go?: detecting road hazards using twitter. *Proceedings of the 37th international ACM ...*, 2609550:1223–1226.

Leonardi, P. M., Huysman, M., and Steinfield, C. (2013). Enterprise social media: Definition, history, and prospects for the study of social technologies in organizations. *Journal of Computer-Mediated Communication*, 19(1):1–19.

Liu, Y., Sui, Z., Kang, C., and Gao, Y. (2014). Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. *PloS one*, 9(1):e86026.

Pei, T., Sobolevsky, S., Ratti, C., Amini, A., and Zhou, C. (2014). Uncovering the Directional Heterogeneity of an Aggregated Mobile Phone Network. *Transactions in GIS*, 18(S1):126–142.

Pereira, F. C., Rodrigues, F., and Ben-Akiva, M. (2013). Using Data From the Web to Predict Public Transport Arrivals Under Special Events Scenarios. *Journal of Intelligent Transportation Systems*, (June 2014):1–16.

Reades, J., Calabrese, F., Sevtsuk, A., and Ratti, C. (2007). Cellular census: Explorations in urban data collection. *Pervasive Computing, IEEE*, 6(3):30–38.

Schulz, A., Ristoski, P., and Paulheim, H. (2013). I see a car crash: Real-time detection of small scale incidents in microblogs. *Lecture Notes in Computer Science (including*

subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7955 LNCS:22–33.

Wanichayapong, N., Pruthipunyaskul, W., Pattara-atikom, W., and Chaovalit, P. (2011). Social-based traffic information extraction and classification. In *ITS Telecommunications (ITST), 2011 11th International Conference on*, pages 107–112.

Witlox, F. (2015). Beyond the data smog? *Transport Reviews*, 35(3):245–249.

Yamamoto, Y. (2007). Java library for the twitter api@ONLINE.

Yang, F., Jin, P. J., Wan, X., Li, R., and Ran, B. (2014). Dynamic origin-destination travel demand estimation using location based social networking data. In *Transportation Research Board 93rd Annual Meeting*, number 14-5509.