Communications in Transportation Research 2 (2022) 100081

Contents lists available at ScienceDirect

Communications in Transportation Research

journal homepage: www.journals.elsevier.com/communications-in-transportation-research

Full Length Article

Investigating social media spatiotemporal transferability for transport

Emmanouil Chaniotakis [a], Mohamed Abouelela [b,*], Constantinos Antoniou [b], Konstadinos Goulias [c]

[a] MaaSLab, Energy Institute, University College London, London, WC1E 6BT, UK
[b] School of Engineering and Design, Department of Mobility Systems Engineering, Technical University of Munich, Munich, 80333, Germany
[c] University of California Santa Barbara, Santa Barbara, CA, 93106-4060, USA

ARTICLE INFO

Keywords:
Social media
Data transferability
Transport modelling
International comparisons

ABSTRACT

Social Media have increasingly provided data about the movement of people in cities making them useful in understanding the daily life of people in different geographies. Particularly useful for travel analysis is when Social Media users allow (voluntarily or not) tracing their movement using geotagged information of their communication with these online platforms. In this paper we use geotagged tweets from 10 cities in the European Union and United States of America to extract spatiotemporal patterns, study differences and commonalities among these cities, and explore the nature of user location recurrence. The analysis here shows the distinction between residents and tourists is fundamental for the development of city-wide models. Identification of repeated rates of location (recurrence) can be used to define activity spaces. Differences and similarities across different geographies emerge from this analysis in terms of local distributions but also in terms of the worldwide reach among the cities explored here. The comparison of the temporal signature between geotagged and non-geotagged tweets also shows similar temporal distributions that capture in essence city rhythms of tweets and activity spaces.

1. Introduction

Information and Communication Technologies (ICT) have changed the course of everyday life. New channels of communication and information exchange have emerged and are being heavily used ever since. This new lifestyle that seems to be widely adopted among individuals brings new opportunities in various scientific and business fields that are mainly attributed to the intensity of data production from the capture of the information exchanged, essentially, creating new data sources (Georgiadis et al., 2020; Chaniotakis et al., 2020). These can be categorized from actively generating (by sensors deployed to periodically measure a particular phenomenon, such as weather data) to passively collecting (by sensors that record specific phenomena, such as social media data). Efforts in working with this yet growing amount of data have been directed towards all aspects of the Big Data Life Cycle (data acquisition, information extraction and cleaning, data integration, aggregation and representation, modelling analysis and interpretation; see Sadiq et al., 2018; Kourik and Wang, 2017) constituting a rather multidisciplinary research topic. In transportation, these efforts have been mainly focusing on the aspects of data acquisition – mostly in terms of data collection, information extraction and cleaning and modelling analysis. The analyses most commonly performed are based – to name but a few – on Floating Car Data (Li et al., 2021; Chen et al., 2021b; Astarita et al., 2019, 2020), mobile phone data (Franco et al., 2020; Zhao et al., 2020; Huang et al., 2018; Wang et al., 2018; Zhou et al., 2018), payment and transit card data (Arbex and Cunha, 2020; Tavassoli et al., 2020; Sulis et al., 2018; Yap et al., 2018; Utsunomiya et al., 2006), GPS enabled mobile phone data (Bachir et al., 2019; Bwambale et al., 2017) and social media (Liao et al., 2021; Yao and Qian, 2021; Lock and Pettit, 2020; Hu et al., 2020; Chaniotakis and Antoniou, 2015; Zheng et al., 2016). Of particular interest in regards to the increased data availability is the evolution of pervasive systems (e.g., GPS handsets, cellular networks) and especially the connectivity that has been available to a growing number of individuals, that allow the sharing of different information types such as spatial, temporal, and textual information.

Specifically, Social Media has received attention from the scientific community mainly due to the unprecedented user-generated content that is (in many cases) publicly available. The statistics of Social Media use are astonishing: where social media websites, Facebook, Instagram, and Twitter, are ranked among the top most visited 50 websites globally.[1] In

* Corresponding author. Technical University of Munich, Munich, 80333, Germany.
E-mail address: Mohamed.abouelela@tum.de (M. Abouelela).

[1] www.semrush.com, visited 09/09/2022.

https://doi.org/10.1016/j.commtr.2022.100081
Received 7 March 2022; Received in revised form 4 August 2022; Accepted 4 August 2022
Available online xxxx

2018, there are 187 million daily active users on Twitter sending 500 million tweets every day; while there are around 2.74 billion monthly active users on Facebook, 80% of them access the site via mobile phones.[2] A growing amount of related work has been published in the last few years, showcasing the potential of using Social Media in transportation. Chaniotakis et al. (2016) have provided a comprehensive review of the directions that transportation-related Social Media research is positioned. In short, the directions that the literature takes are either the use of Social Media for modeling and forecasting purposes, including an aspect of the use of Social Media data for OD Estimation (Liao et al., 2021; Osorio-Arjona and García-Palomares, 2019), Attraction Models (Lee et al., 2019; Yang et al., 2018; Hu and Jin, 2018), activity modelling (Cui et al., 2018; Chaniotakis et al., 2017; Hasan and Ukkusuri, 2018; Lee et al., 2016), extraction of mobility-related and spatial characteristics (Ebrahimpour et al., 2020; Hu et al., 2020; Kim et al., 2018; Yao et al., 2018; Jiang et al., 2015; Yang et al., 2019) transportation-related sentiment analysis (Rahman et al., 2021; Bakalos et al., 2020; Sari et al., 2019; Ali et al., 2018, 2019), prediction and event detection (Chaturvedi et al., 2021; Yao and Qian, 2021; Alomari et al., 2019, 2021; Zulfikar et al., 2019; Zhang et al., 2018; Xu et al., 2018; Pereira et al., 2015), and accessibility analysis with the complementary use of Twitter data (Kim and Lee, 2021; Qian et al., 2020; Moyano et al., 2018). On another perspective, social media have also been used mainly from transport providers, for the direct communication that their platform allow with the end users (National Academies of Sciences, Engineering, and Medicine, 2021). Such usages are oriented towards public engagement (Gu et al., 2020; Williamson and Ruming, 2020; Haro-de Rosario et al., 2018; DePaula et al., 2018; Bonsón et al., 2019) and for information sharing (Bokings et al., 2020; Purnomo et al., 2021; Georgiadis et al., 2020; Manetti et al., 2017; Gal-Tzur et al., 2014).

As the literature on Social Media exploitation for transportation studies continues to grow, the questions of transferability of the results and sample specification are becoming central. Its importance is further highlighted by the fact that, due to global availability, Social Media studies are commonly focusing on areas of high Social Media usage, neglecting in a sense the question of how possible would it be to deploy the defined methods in a different context. To the best of the authors' knowledge, little has been done to showcase the potential similarities and differences of deploying Social Media data in transportation research in different cities with the few exceptions to be found on the analysis of the deploying of the natural cities concept by (Jiang and Miao, 2015), the exploratory investigation of millions of Twitter footprints with the extraction of radius gyration for users in USA cities (Cheng et al., 2021), the identification of tourist hot spots in European cities (García-Palomares et al., 2015), the study of how people experience the city on local and global scale through geotagged photos (Paldino et al., 2015), and the use of Social Media as a global mobility proximity (Hawelka et al., 2013). However, in all of the above cases, the methodological approach for the (in some cases indirect) comparison of different city comparison is based on the general tweets data collection (from the Twitter Streaming Application Programming Interface, API) that returns a fraction of the total tweets posted, without focusing on the posting characteristics of individual users. This omission is believed to be of high importance in the comparison of different data analysis settings and the corresponding data uses in transportation modeling, and forecasting. The exploration of the collective entity of posts, users, would allow for a better understanding of the factors that shape the decisions related to post, and the relative characteristics of different origins, related to social media use.

In this paper, we analyze the data collected from Twitter in 10 cities in Europe and USA. Descriptive statistics of Social Media use are explored for the analysis of the different patterns met among different cities. Additionally, we perform a user-centric analysis of the posting activity and the connection with other Social Media Platforms. The latter is performed for users with above-average twitter posting activity for each city, by collecting the posts from their timeline (Chaniotakis and Antoniou, 2015) to identify the capabilities to examine the use of geo-reference in posts, the temporal characteristics of posting, the activity space of individuals, habitual patterns of posting geo-referenced tweets, and the spatial dimensions of this posting activity. The temporal characteristics of time-of-day posting together with the spatial footprint in a city provide unique information about the movements in different cities and the use of transportation and other facilities. We envision the methods here becoming a source of data to validate activity-based models and the identification of hot spots in each city, where crowding happens. This can be used either as a historical record for planning or for emergencies in real-time monitoring and operations. Moreover, correlating the urban form of cities to social media spatio-temporal signatures can lead to different ways in identifying land use, and visualizing changes in land use (e.g., Chen et al., 2021a; Ye et al., 2020; Thakur et al., 2018; McKenzie et al., 2015; Frias-Martinez et al., 2012) and natural changes in time use behavior. The findings from the posting activity of individuals are inferred with socio-demographic characteristics aiming at generalizations concerning the sample differences among cities concerning data availability and the potential hidden latent variables, such as privacy concerns and technology aversion.

2. Dataset construction

The data collection for the comparison of the different cities was performed by first deploying the Random Data Collection (Section 2.1) process that collects random tweets from Twitter (essentially forming a users' dataset), and then based on some filtering criteria, proceeded with the Users-based Data Collection (Section 2.2) that collects a number of the latest tweets from each user.

2.1. *Tweets data collection*

For the extraction of information concerning Social Media usage, data has been collected for 4 major cities in USA (Los Angeles, New York, Orlando, Seattle) and 6 major cities in Europe (Amsterdam, Athens, Copenhagen, London, Munich, Paris). The selection of the particular cities was based on (a) the ability to extract information from textual characteristics (based on known languages), (b) the indicated Social Media usage, (c) the relatively large size of the cities and (d) the diverse characteristics (in terms of size, demographics and Internet penetration). In order to make the data collection as homogeneous as possible at the time of this study, a rather small Random Data Collection period was specified (approximately 2 months); resulting in mostly collecting a user sample. The data collection was performed using the Twitter REST Application Programming Interface (API) and by utilizing the Twitter4J library within a Java program that automatically collects data based on the latitude and longitude of a central point and a radius (Chaniotakis and Antoniou, 2015). It should be noted that the Twitter API returns both geo-referenced (geotagged) tweets as well as tweets without geo-reference (not-geotagged). Additionally, the Twitter REST API returns a limited amount of tweets per query (200 tweets) and has a time quota of 180 queries per 15 min (1 query per 5 s). In Table 1 the results from the initial data collection are presented. As it is clearly evidenced, the use of Social Media in USA (at least within the examined period of time) is much higher than the use in Europe, something that agrees with the statistics (pewinternet.com).

2.2. *User's timeline data collection*

Based on the collected dataset, a random sample of at least 1000 users was selected for each city, to collect their Twitter history. The selection of the particular user sample was solely based on one criterion: the users should have posted at least two geotagged tweets within the examined data collection period. The selection of this minimum number of

[2] blog.hootsuite.com.

Table 1
Tweet data collection.

City	Avg. geotagged tweets (day)	Avg. not geotagged tweets (day)	% of avg. geotagged tweets (day)
Amsterdam, NL	1172	117598	1
Athens, GR	929	NA	NA
Copenhagen, DK	460	54329	0.8
Orlando, USA	454	NA	NA
Seattle, USA	774	61783	1.2
Munich, DE	3551	447179	0.8
New York, USA	16959	NA	NA
London, UK	202	30417	0.7
Los Angeles, USA	6386	1208561	0.5
Paris, FR	4	233086	0.0

NA: Refers to random data collection processes collecting only geotagged data.

geotagged tweets was based on the general data collection characteristics, allowing to collect, for all cities, at least the required users to examine, while confirming that they are using their account for – among others – posting geotagged tweets. The small number of users and the random sample selection aim at the exploration of the potential of using Social Media in transportation studies. The data collection was performed by extracting the latest tweets from the timeline of each user (Chaniotakis and Antoniou, 2015). For each user, the last 600 posted tweets were collected. The data collection process was again performed using the Twitter REST Application Programming Interface (API) and by utilizing the Twitter4J library within a Java program for the collection of tweets using Twitter pagination. It should be noted that the user-based data collection does not include the tweets collected from the random data collection process. This might result in users which have not posted a geotagged tweet in the last 600 tweets, but on the other hand allow us to better compare users and cities.

3. User characteristic analysis

3.1. Social media use

For the analysis of the characteristics, several indicators were selected to be used in order to allow for an adequate comparison of the users. Descriptive statistics were explored for the generic understanding of Twitter use in the examined cities (Table 2). As it is clearly observed, the users that were collected are in general active Twitter users with a large number of tweets posted. The percentage of the number of geotagged tweets posted in each case differs with cities in USA to have a range of geotagged tweets that is higher than that observed in European cities (32.9%–48.4% in USA vs. 11.7%–29.2% in Europe). When comparing the mean percentage of tweets posted in each city respectively to examine the number of users that are using Twitter to post geotagged information in the city of residence, it is rather clearly evidenced that again, there is a clear difference between USA and Europe. Specifically, the highest percentage of geotagged tweets performed in the examined city (to the total geotagged tweets) is found in New York (73.9%), while the lowest is found in Copenhagen (24.3%). Another apparent difference between the collected data in Europe and USA is the percentage of the users who did not post any geotagged tweets. The maximum percentage of the no-geo-taged tweet in European cities is 21.1%, London, while in USA, it is in Orlando with only 2.1%.

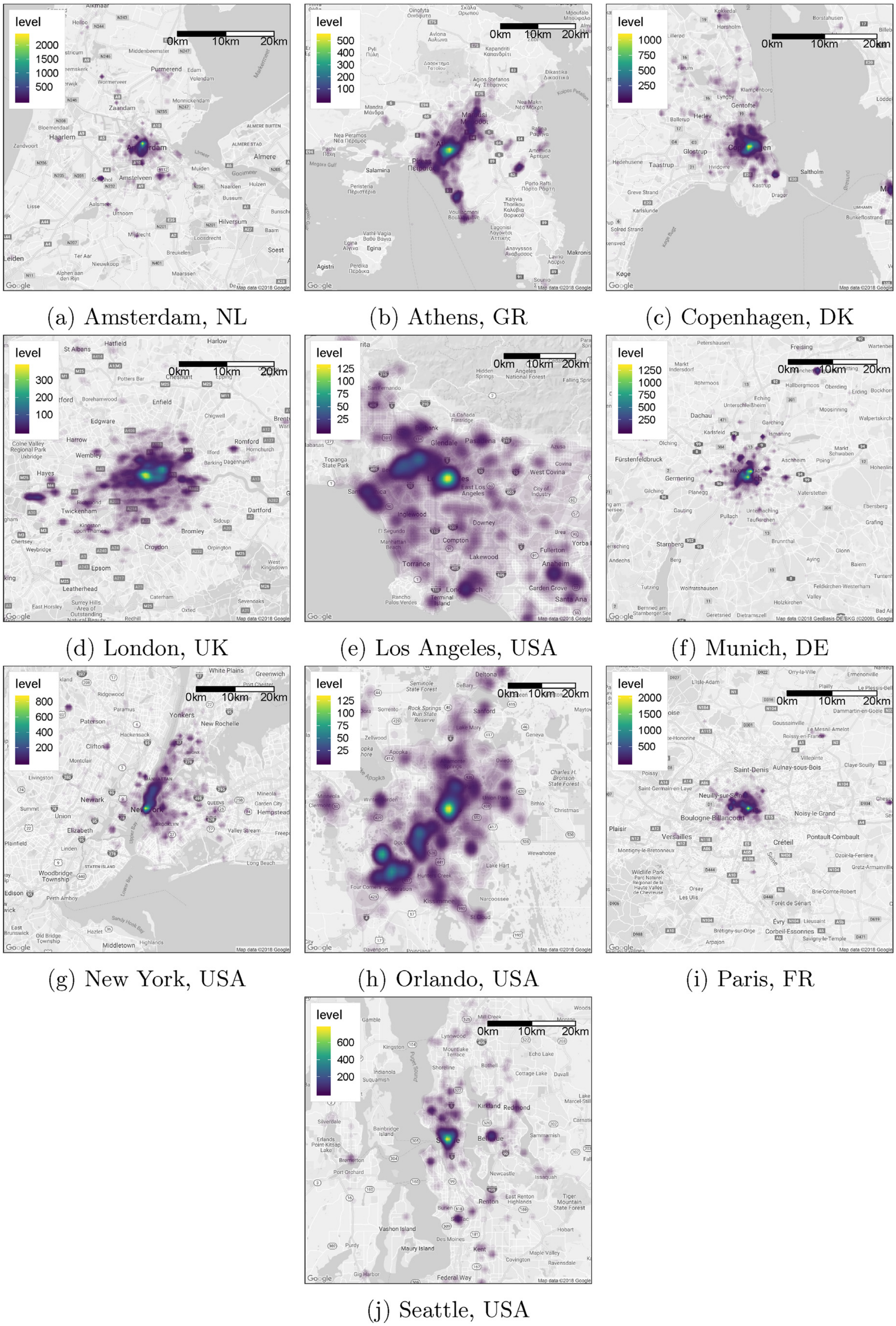
3.2. Spatial analysis

The geotagged tweets seem to widely cover the cities examined, as presented by the spatial density plots in Fig. 1. The observed coverage confirms the fact that even with a small number of users, Social Media (and particularly Twitter) could be used to extract information with a rather large concentration in the central areas of cities, but also on suburban areas. Interestingly, the collected tweets illustrate the differences in the way that cities in USA and Europe are formed (Huang et al., 2007). In particular, by exploring the frequency geotagged tweets performed within the cities examined, it is evidenced that cities in Europe are structured within densely used center, while cities in USA are spread in space, creating multiple centers where individuals dwell and perform activities. This is particularly evidenced in Los Angeles, Orlando, and Seattle, while New York illustrates a concentration of tweets in the Manhattan area and it is clearly a reflection of urban form and business establishments structures in the different regions examined. Such a difference in the urban structure and its analysis (mono/polycentric urban structure observed) is particularly interesting for transportation, as there is strong evidence that urban form is related to travel patterns (Stead and Marshall, 2001). In particular, the mono/polycentric structure has been found to relate to mode choice and distance traveled (Lin et al., 2015; Schwanen et al., 2001). Frequently revisiting the urban form through Social Media, can allow for the definition of proxies for mode and destination choice as well as for the examination of the evolution of cities.

Another interesting stream of research is the global mobility patterns and the potential of extracting them from Social Media. The seminal work of Hawelka et al. (2013) concluded that Twitter can be used as a proxy for human mobility, especially at the country-to-country level. In our case, the User-based Data Collection is not spatially restricted, allowing for the collection of the places visited by individuals around the world. Fig. 2 presents the density plots of areas that users in each examined city have visited. It should be noted that we do not distinguish between tourists and residents of each city. As it is evidenced, the locations visited differ in Europe and USA. More specifically, users from Amsterdam illustrate a concentration of posts in Europe, while also showing posts in other areas, mainly the east coast of USA and regions in Asia. Users from Athens illustrate a higher concentration of posts in Europe; however, there is a much lower dispersion of tweets in Europe in

Table 2
User based data collection.

City	Number of users	Users with no-geotagged tweets (%)	Mean number of tweets	St.Dev. number of tweets	Mean geotagged (%)	St.Dev. geotagged (%)	Mean in examined city (%)
Amsterdam, NL	1127	0.2	556.9	124.1	29.2	27.9	33.9
Athens, GR	2092	12.9	576.7	87.8	22.6	24.8	31.9
Copenhagen, DK	1739	4.1	575.1	90.3	28.3	26.7	24.3
London, UK	2153	21.1	591.4	58.2	11.7	17.6	42.6
Los Angeles, USA	2313	0.0	532.1	160.3	44.3	34.1	64.4
Munich, DE	1389	1.9	545.2	140.3	26.8	27.5	28.5
New York, USA	1997	0.1	566.2	119.3	48.4	32.3	73.9
Orlando, USA	2748	2.1	545.4	142.1	34.3	29.9	35.6
Paris, FR	3856	5.7	583.8	71.3	25.04	25.5	35.3
Seattle, USA	1852	0.1	532.3	155.9	32.9	30.7	47.8



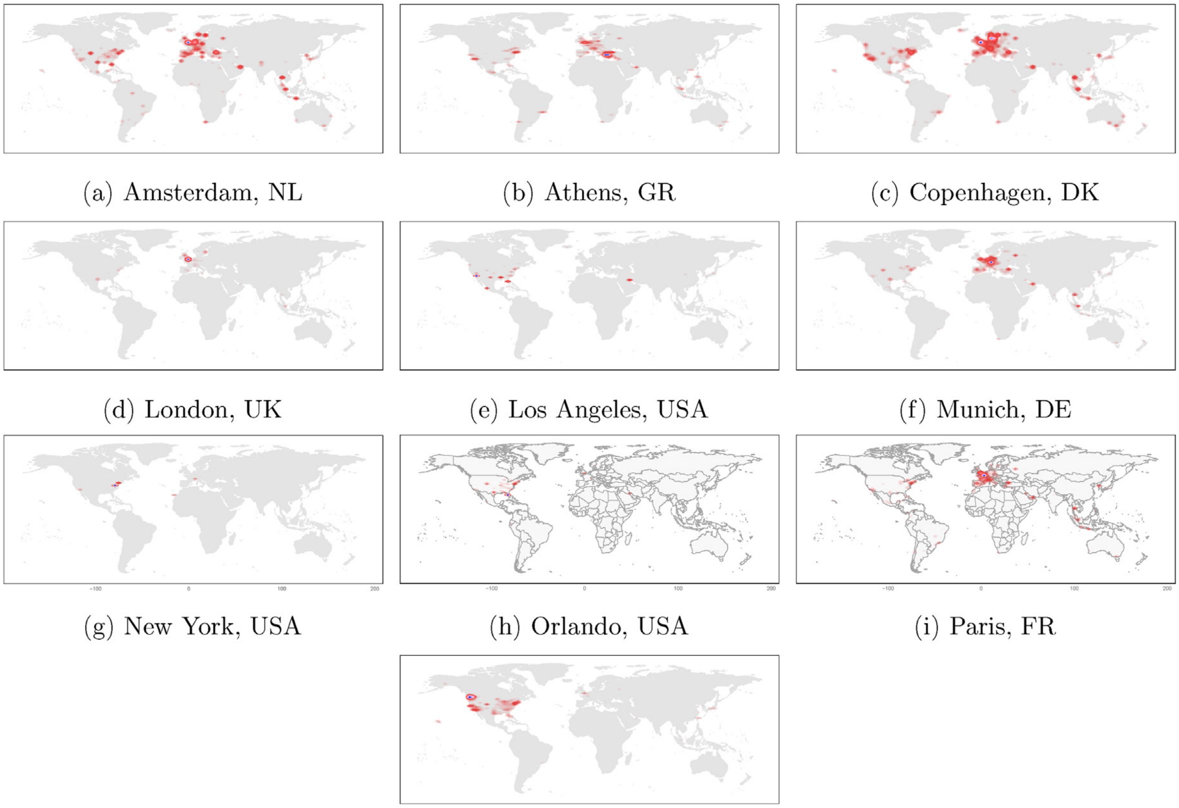
(a) Amsterdam, NL (b) Athens, GR (c) Copenhagen, DK

(d) London, UK (e) Los Angeles, USA (f) Munich, DE

(g) New York, USA (h) Orlando, USA (i) Paris, FR

(j) Seattle, USA

Fig. 1. Density Plots of the tweets performed by the collected users, near the examined city.



(a) Amsterdam, NL (b) Athens, GR (c) Copenhagen, DK

(d) London, UK (e) Los Angeles, USA (f) Munich, DE

(g) New York, USA (h) Orlando, USA (i) Paris, FR

(j) Seattle, USA

Fig. 2. Density Plots of the tweets performed by the collected users on a world scale, as collected from each city.

comparison to users from the rest of the European Cities. Another compelling case is the case of London, where there is a very high concentration of posts in UK, illustrating a much lower number of tweets posted in other areas. The most extensive spread of posts is observed by users originally collected in Copenhagen. On the other hand, in USA there is an apparent concentration of users in the areas collected (Orlando, Los Angeles, New York), while only Seattle illustrates a rather wider spread of posts mainly in USA.

3.3. Temporal analysis

The analysis of the temporal dimension of Twitter posts has been performed in the form of a direct comparison of the temporal distributions for both the geotagged and the not-geotagged tweets. For matters of clarity, it should be noted that the hours in the distribution were adjusted for the different time zones, taking into account the summer time difference when necessary. Fig. 3 presents temporal distributions in different days and hours of the week and Fig. 4 shows the percentage of the in-city geotagged tweets per user.

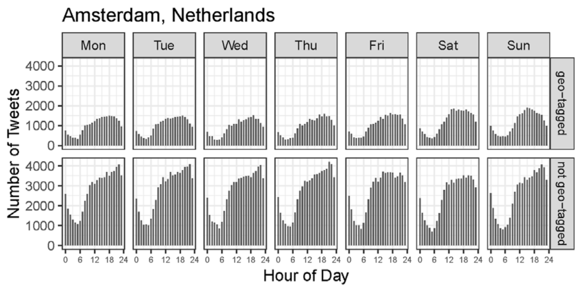
With regards to regularity in time, it is observed that it is much more pronounced than what we can observe for space. This illustrates an almost habitual use of Social Media, which makes it very interesting for the exploration of mobility patterns. Additionally, it is evidenced that there is a rather increased posting activity during weekends, especially for geotagged tweets, and also during evening hours with the peak to be usually around 17:00–20:00. The lowest points for all examined cities is during night hours. Another interesting characteristic of the data collected from New York is the peaks that are observed in most cases 2 h in the day (around 8:00–9:00 and 17:00–18:00). This type of peaks is also observed in the case of Los Angeles.

3.4. Social media activity space characteristics

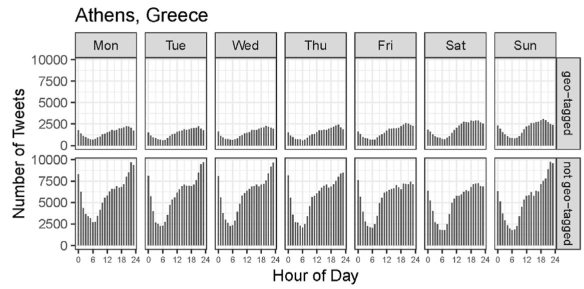
Further explored in the remainder of this paper is the activity space of individuals in the examined cities. Conventionally, activity spaces are estimated as the convex hull of the locations visited by individuals (Golledge, 1997). This definition, however, does not seem to apply for Social Media. The main reason is that activity spaces have been developed for rather small periods of data collection and mainly for habitual travel patters. On the contrary, from Social Media we can get data from the same individuals for a course of years; thus, each user sub-dataset could include traveling in different countries or visiting places once in a few years or just being at a particular location once in a lifetime, which defeats the purpose of defining activity spaces. For this reason, we introduce a methodology that defines activity spaces through a two steps process: (1) investigates the characteristics of frequently visited places by individuals (location recurrence) using clustering techniques, and (2) performs the same clustering analysis on a local scale for the definition of the total activity space. These two different analyses combined provide a much better understanding of the activity space and the resulting habitual patterns of individuals, allowing for a clearer evaluation of the travel patterns that we observe. In both cases, the application of the Density-based Spatial Clustering of Applications with Noise (DBSCAN) algorithm is used (Ester et al., 1996).

3.4.1. Distinction between different user groups

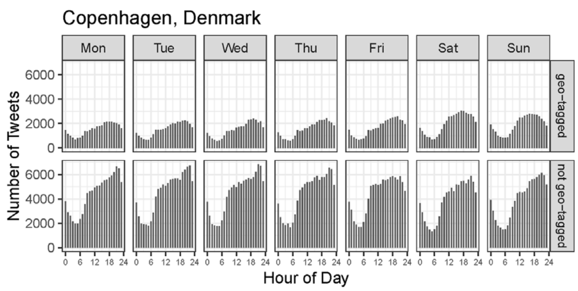
The analysis of the location recurrence and activity space is performed on the total dataset, city residents, and tourist user groups. The collected data does not contain any information concerning the residency location of the different users. Therefore, the identification of the various users' group residency is decided based on the percentage of the



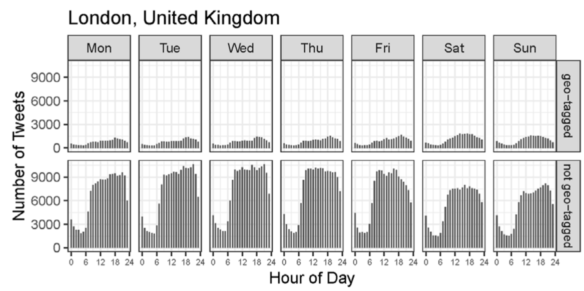
(a) Amsterdam, NL



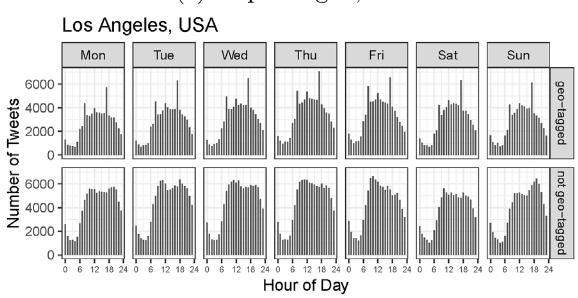
(b) Athens, NL



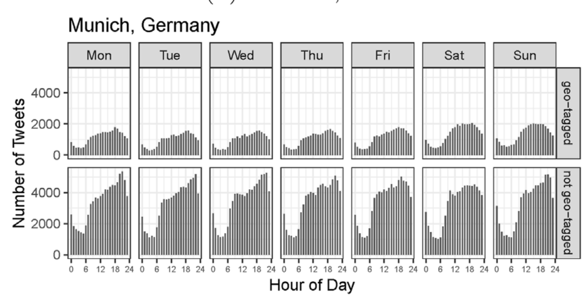
(c) Copenhagen, NL



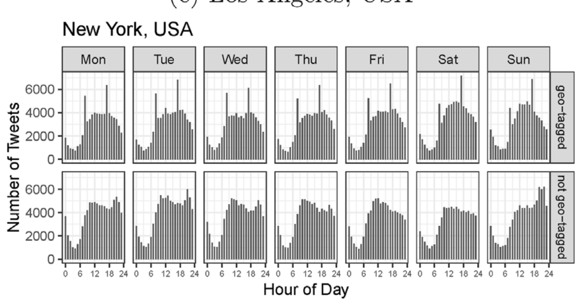
(d) London, UK



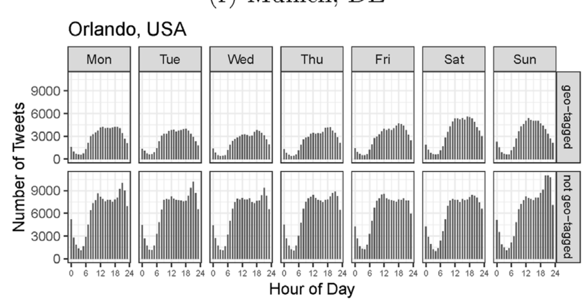
(e) Los Angeles, USA



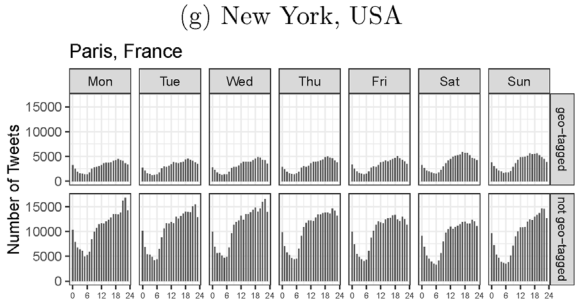
(f) Munich, DE

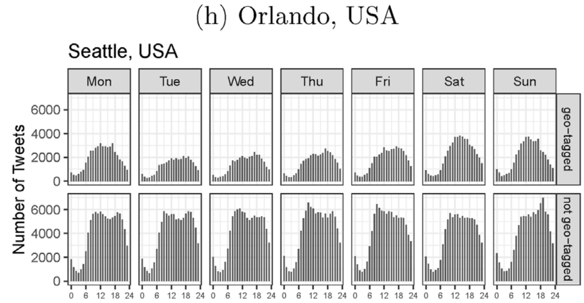


(g) New York, USA



(h) Orlando, USA



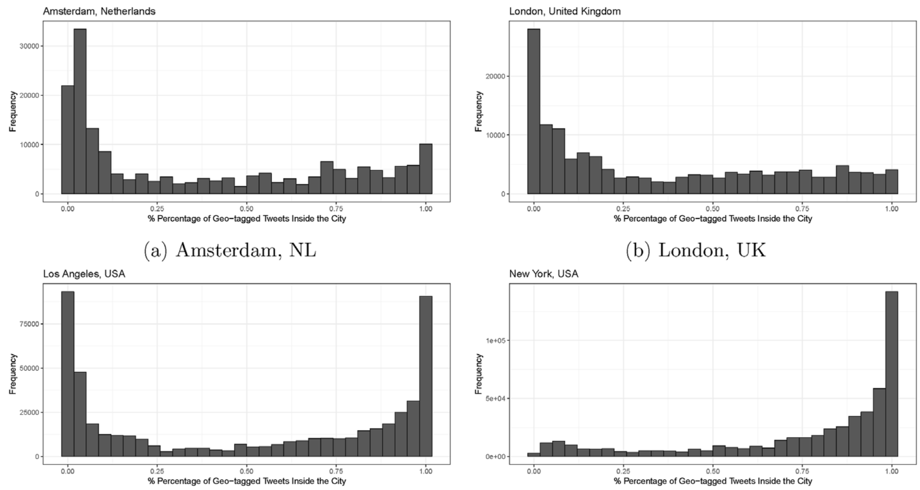


11

(i) Paris, FR

(j) Seattle, USA

Fig. 3. Examples of Temporal Distributions for geotagged and not-geotagged tweets.



(a) Amsterdam, NL (b) London, UK

(c) Los Angeles, USA (d) New York, USA

Fig. 4. Examples of Inside City User geotagged tweets.

Table 3
Percentage of the different user groups per city.

City	% of resident	% of unclear	% of tourist
Amsterdam, NL	58.39	27.60	13.84
Athens, GR	53.35	11.47	22.28
Copenhagen, DK	57.16	14.38	24.38
London, UK	46.68	12.22	20.02
Los Angeles, USA	47.00	28.06	24.90
Munich, DE	66.74	16.49	14.83
New York, USA	49.87	25.54	24.49
Orlando, USA	56.30	20.92	20.67
Paris, FR	55.63	16.88	21.84
Seattle, USA	58.75	28.40	12.74

geotagged tweets inside the city boundaries to the total tagged tweets. If people are having < 0.25 of geotagged tweets inside the cities' limits they are considered tourists, and people having > 0.50 of geotagged tweets within city boundaries are residents. Nevertheless, the status of the third group with geotagged tweets ouside of the chosen range of 0.25–0.50 is unclear; therefore they are out of the interest of this paper. Table 3 shows the percentage of each user group for the collected data for each of the subject cities.

3.4.2. Location recurrence

Starting from the location recurrence, for each user the geotagged tweets are investigated for the formation of spatial clusters of different in-time posts. This sheds some light on aspects of transferability of methods and solutions which use Social Media. When examined, we could identify if there are strong differences or similarities with regards to habitual posting from specific locations that can be associated with specific activities (Chaniotakis et al., 2017). Here, this was performed by specifying the characteristics of the clusters based on the GPS accuracy (Chaniotakis et al., 2017). The analysis was implemented in R using the dbscan library, which applies the density-based algorithm for discovering clusters in large spatial databases with noise originally developed by Ester et al. (1996). The parameters were selected after examination of various settings taking into account the GPS accuracy (Schaefer and Woodyer, 2015) and the number of tweets that each individual posts: the neighborhood of a point parameter (Eps) was defined to be 0.002 and the minimum number of points to be 5. The usability of this analysis is based on the investigation of the way users use Twitter, distinguishing the use of Twitter to post extraordinary locations visited (in terms of the users' "mean" activity space) from the ordinary Social Media use. The analysis of the location recurrence yield interesting results (Table 4).

First, for the aggregated data representing all the users groups, the mean number of clusters varies from 3 (London) to 8 (New York). For the resident group, the mean number of clusters ranges between 2 (London) and 6 (New York and Los Angeles). The tourist group mean number of clusters varies between 4 (London) and 11 (Seattle). For the total aggregated data, there is a clear difference between Europe and USA, with USA cities illustrating a larger mean number of clusters, except for Seattle that illustrates a mean number of clusters close to the European mean. The resident user groups represent the same pattern for the different cities with a lesser mean number of clusters compared to the total data. The tourist user groups show a slightly different pattern with the increase of the mean number of clusters for all the cities compared to the previously described two groups. The larger number of clusters for

Table 4
Analysis of user location recurrence for different groups.

City	Mean number of clusters			Mean point in cluster			Mean noise point			Mean cluster in exam city		
	Total	Res*	Tour**	Total	Res*	Tour**	Total	Res*	Tour**	Total	Res*	Tour**
Amsterdam, NL	5.11	3.92	10.26	82.05	82.21	121.38	77.57	55.15	163.47	1.26	2.00	0.08
Athens, GR	5.32	4.35	5.67	69.46	59.44	72.42	88.03	67.34	89.56	1.09	1.97	0.00
Copenhagen, DK	5.74	4.63	6.09	81.38	69.91	86.61	92.19	72.73	89.22	0.98	1.61	0.01
London, UK	3.04	2.35	3.76	39.42	30.97	53.36	60.15	50.37	59.18	0.81	1.61	0.01
Los Angeles, USA	6.14	5.88	5.40	155.65	185.76	152.10	70.18	54.37	78.16	3.78	5.28	0.85
New York, USA	7.95	5.88	10.95	190.37	226.08	144.63	73.63	38.84	124.04	5.23	5.42	3.15
Munich, De	4.78	3.54	7.86	66.93	49.92	112.88	78.43	56.69	135.43	0.91	1.30	0.04
Orlando, Usa	6.42	5.64	7.23	108.07	99.64	130.75	71.39	56.23	89.99	2.02	3.32	0.04
Paris, FR	5.54	4.62	6.44	73.52	62.00	92.22	86.60	72.07	90.08	1.53	2.55	0.05
Seattle, USA	5.47	4.45	11.36	92.57	92.72	147.73	70.23	45.56	171.24	2.20	3.02	0.21

Res* = Resident, Tour** = Tourist.

Table 5
Power-Law properties of location recurrence clusters.

City	Number of points in cluster			Number of noise points			Number of clusters		
	p-value	x_{min}	α	*p*-value	x_{min}	α	*p*-value	x_{min}	α
Total Data									
Amsterdam, NL	0.463	302	1.429	0.07	196	1.265	0.155	18	1.728
Athens, GR	0.287	321	1.445	0.345	257	1.254	0.00	16	1.695
Copenhagen, DK	0.493	364	1.434	0.035	212	1.247	0.93	30	1.677
London, UK	0.37	220	1.541	0.01	178	1.284	0.365	22	1.907
Los Angeles, USA	0.00	164	1.345	0.075	158	1.271	0.27	26	1.7
Munich, DE	0.83	406	1.467	0.28	172	1.264	0.00	18	1.744
New York, USA	0.00	233	1.313	0.945	234	1.26	0.77	33	1.598
Orlando, USA	0.00	317	1.387	0.285	130	1.267	0.00	21	1.66
Paris, FR	0.00	277	1.439	0.86	293	1.252	0.00	11	1.676
Seattle, USA	0.00	201	1.413	0	113	1.277	0.61	26	1.742
Resident data									
City	Number of points in cluster			Number of noise points			Number of clusters		
	p-value	x_{min}	α	*p*-value	x_{min}	α	*p*-value	x_{min}	α
Amsterdam, NL	0.26	282	1.42	0.88	132	1.30	0.00	13	1.82
Athens, GR	0.01	165	1.48	0.01	171	1.27	0.00	6	1.80
Copenhagen, DK	0.01	198	1.47	0.00	150	1.26	0.10	20	1.78
London, UK	0.49	159	1.60	0.02	128	1.29	0.00	10	2.09
Los Angeles, USA	0.00	156	1.33	0.06	154	1.28	0.00	8	1.74
Munich, DE	0.44	273	1.53	0.00	72	1.28	0.00	14	1.90
New York, USA	0.00	178	1.30	0.18	114	1.31	0.00	16	1.75
Orlando, USA	0.00	306	1.40	0.02	125	1.28	0.90	26	1.70
Paris, FR	0.00	138	1.47	0.00	182	1.26	0.00	16	1.75
Seattle, USA	0.00	91	1.42	0.00	119	1.31	0.00	7	1.85
Tourist Data									
City	Number of points in cluster			Number of noise points			Number of clusters		
	p-value	x_{min}	α	*p*-value	x_{min}	α	*p*-value	x_{min}	α
Amsterdam, NL	0.00	126	1.36	0.62	218	1.29	0.64	17	1.51
Athens, GR	0.00	131	1.43	0.07	189	1.32	0.17	18	1.64
Copenhagen, DK	0.33	344	1.41	0.31	310	1.26	0.01	15	1.62
London, UK	0.00	110	1.48	0.06	106	1.30	0.00	6	1.76
Los Angeles, USA	0.00	59	1.35	0.00	106	1.28	0.02	14	1.79
Munich, DE	0.96	420	1.36	0.39	147	1.29	0.28	18	1.53
New York, USA	0.55	386	1.33	0.27	232	1.29	0.95	27	1.47
Orlando, USA	0.11	390	1.36	0.00	102	1.28	0.00	12	1.64
Paris, FR	0.32	281	1.40	0.52	308	1.26	0.01	18	1.60
Seattle, USA	0.04	217	1.34	0.00	127	1.22	0.53	21	1.51

the tourist groups replicates the expected tourist behavior of visiting more locations compared to the city residents. The difference between USA and Europe is still evident for the tourist groups except for Amsterdam having a mean of 10 clusters, which is higher than Los Angeles and Orlando mean number of clusters, this finding supports the criteria used to differentiate between the city's residents and tourists.

The same characteristics are observed in the mean number of points in clusters and the opposite in the mean number of points characterized as noise. Finally, the number of clusters inside the city seem to follow the same pattern (more frequent visits at the same areas) in USA in comparison to Europe. The Users' Location Recurrence analysis illustrate the differences in use of Social Media in the examined areas and particularly between Europe and USA. It is found that in Europe, users mainly post geotagged tweets when visiting locations that can be characterized as not frequently visited (presumably for leisure activities or special events). On the other hand, users from USA tend to post geotagged tweets from locations they visit frequently.

Apart from the generic analysis of the location clusters and in order to extract information concerning the transferability of the findings to other cities, the examination of the classification-related variable distribution properties took place. Particularly, and given the observed resulting distributions, in addition to the findings of the literature, the fitting of Power-Law distribution was found to be the most prominent distribution. The examination of the applicability of the Power-Law properties in the Location Recurrence Clusters data was performed using the Power-Law library in R (Clauset et al., 2009). The results of the fitting are presented in Table 5 with an example of the log–log plots for the number of user location recurrency clusters is presented in Fig. 5.

For the total data-set, it is clearly evident that in half of the cases and only starting from a large number of x_{min}, we were able to detect Power-Law distributions (p-value $\leq$ 0.05 where 0.05 is the significant level examined – using the Kolmogorov Smirnoff test), indicating that the examined variables (number of clusters, number of points in a cluster and number of noise points) cannot be characterized by the Power-Law distribution (Clauset et al., 2009). For the resident group data, the Power-Law distribution is evident in most of the cities (7 cities). For the tourist group data, Power-Law distribution is not evident in half of the cases. This finding further supports the previously examined characteristics that indicate the strong difference in Social Media use across the two examined continents and even across different cities. The findings of the Power-Law distributions are related to the first law of Geography. Essentially, what we observe is a different degree of relatedness in different cities. This is a function of the people that live and visit those places, the infrastructure and spatial distribution of business establishments, predominant cultural traits, and reporting habits using social media.

3.4.3. Cluster-based activity space

The activity space from Social Media was examined on clustered data based on points mutual distances again using the DBSCAN algorithm.

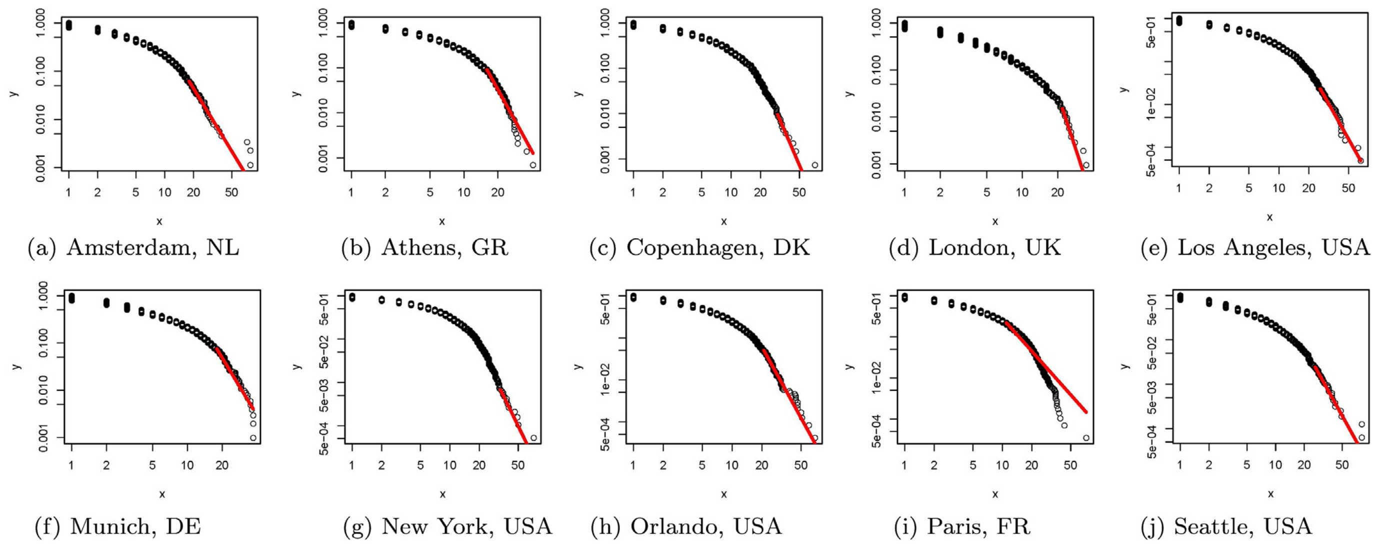


Fig. 5. Power-Law plot for users locations clusters.

This choice was based on the need to identify the areas visited by individuals on a local scale, avoiding the formation of very large activity spaces that could result from long distance traveling for tourism or business purposes, as this is out of the scope of this work. The parameters for this analysis were selected taking into account the upper level of commuting distance for the countries examined: the neighborhood of a point parameter (Eps) was defined to be 1.2 and the minimum number of points to be 3. Additionally, in order to extract characteristics of the city examined, the clusters were characterized as near the city examined (in case the distance between the center of the cluster and the center of the city was smaller 120 km). The cluster analysis was performed on the basis of the cluster in the city examined, as well as including the two additional largest clusters (in terms of points in the clusters). Table 6 presents some aggregated activity space characteristics. This analysis also confirms the differences in the use of Social Media in Europe and in USA, while it is worth noting that all European cases have a significantly larger activity space in comparison to the activity space of the USA cases for the different users groups. Besides, the area of the tourist groups activity space is more extensive than the resident groups' activity space which confirm the rational travel pattern of tourists.

As presented in Table 7, in the majority of the cases, the number of geotagged tweets belong to the examined city's cluster. It should be noted that the fact that a large number of tweets is classified as of being in the city might seem contradictory, when compared with the percentage of in city tweets, however it could illustrates that in many cases, the strict administrative areas of the cities do not necessarily represent the individuals who commonly use the city; while it should also be taken into account that in extreme cases some tweets could even be posted almost 300 km away from the city center and still belong to the classified as in the city classification (as we only consider the distance of the class center to the city center). The largest percentage is observed in New York city, while the lowest in Copenhagen and Munich. Another interesting fact is that apparently only a small percentage of geotagged tweets do not belong to a cluster. This finding is subject to the low minimum number of point specification and the large Eps parameter used. Finally, in most cases, a vast majority of geotagged tweets are included in either one of three examined clusters.

4. Conclusions and discussion

With the development of disruptive technological concepts, heterogeneous data sources will continue to emerge. Although not strictly defined as transportation data, they might illustrate properties that would potentially enrich our understanding of the transportation system. Here, we examine aspects of Social Media use and particularly Twitter for the examination of the potential of using Social Media data in various settings. An exploratory, empirical analysis is presented on commonly used spaces in transportation: spatial, temporal and activity spaces. In order to extract aggregated characteristics, classification techniques are implemented and Power-Law distributions are examined, without afterall identifying clear Power-Law properties.

As illustrated in the pertinent literature and supported by evidences in this study, SM spatio-temporal patterns could be used for augmenting transport-related activities, commonly underrepresented in transport surveys (e.g., leisure activities) and capture aspects such as city development dynamics (e.g., polycentric versus monocentric) as well their evolution (e.g., changes due to gentrification or market conditions), areas of interest and destination choices. At the same time, the definition of activity spaces using data from SM-data that spans across larger timescales can provide a more informative view of the boundaries of areas that are commonly visited by individuals and provide valuable information for travel behaviour research. In addition to the above, the analysis performed can shed light into the differential use of Social Media in different areas around the world, providing a basis for the evaluation of the suitability of the use of different (or re-calibrated) techniques that combine transport and social media data for different areas. Also, the user-based approach followed provide a better understanding of the social-media related behaviour and has the potential for exploration of transferability of methods across different contexts.

As analysis suggests, the distinction between residents and tourists is fundamental for the development of models, as there are major differences in terms of behavior, for all variables examined. Additionally, the exploration of activity spaces is believed to benefit from a clustering based examination, especially when used for social media data analysis. With the proposed methodology, the definition of activity spaces can be formalized to represent different contexts of activities such as areas that are frequently visited in a time frame or in a particular area, avoiding points which act as noise. This is especially suitable in case of wider in time-span datasets (such as data originating from Social Media), due to possibility of using activities performed only once are in different countries. Additionally, activity space definition also benefits from the distinction between tourists and residents, who illustrated different behaviours. The exploration of the Power-Law distribution yielded also interesting results concerning the various activity space characteristics. Although Power-Law distribution could not be safely defined, it is believed that the exploration of the parameters of the distribution fit

Table 6
Activity space characteristics, for different groups.

City	Mean number of clusters			Mean points in exam city cluster			Mean points in Second Larger Cluster			Mean points in First Larger Cluster		
	Total	Res*	Tour**	Total	Res*	Tour**	Total	Res*	Tour**	Total	Res*	Tour**
Amsterdam, NL	5.24	2.96	13.95	77.43	112.36	12.77	48.81	14.67	119.95	19.7	15.07	25.90
Athens, GR	5.59	3.78	6.19	63.11	76.19	8.06	58.02	33.56	75.79	24.55	16.39	26.13
Copenhagen, DK	6.26	4.88	5.98	58.98	74.13	13.06	63.47	36.78	87.07	25.83	15.99	38.3
London, UK	3.82	3.07	3.98	54.04	60.27	37.87	35.45	10.46	65.78	16.31	9.77	24.31
Los Angeles, USA	3.86	2.23	5.55	180.27	223.78	172.18	52.99	12.32	98.87	20.89	10.5	33.8
New York, USA	3.54	1.46	7.34	212.06	258.7	110.05	36.18	8.79	73.64	19.48	7.53	28.78
Munich, DE	5.73	4.06	10.63	42.37	52.11	14.59	53.69	30.07	99.22	22.66	16.9	32.35
Orlando, USA	4.23	2.88	6.93	105.48	123.71	97.93	61	23.55	100	21.2	12.64	30.85
Paris, FR	5.62	4.64	5.81	61.32	80.24	10.27	54.53	22.27	95.51	23.83	15.52	36.38
Seattle, USA	4.92	2.32	16.00	96.15	123.6	30.25	34.83	12.55	85.46	19.18	9.66	38.48
City	Mean No. of noise point			Activity space in exam. city (1E+07 km^2)			Activity space in First Larger Cluster (1E+07 km^2)			Activity space in First Larger Cluster (1E+07 km^2)		
	Total	Res*	Tour**	Total	Res*	Tour**	Total	Res*	Tour**	Total	Res*	Tour**
Amsterdam, NL	4.62	2.35	14.53	2.1	1.32	4.95	2.56	1.87	4.88	3.21	2.81	5.04
Athens, GR	4.29	3.05	4.99	2.05	1.03	6.44	2.46	1.33	2.72	3.07	1.76	3.35
Copenhagen, DK	5.13	4.44	4.89	2.84	1.9	5.32	2.84	2.19	2.33	3.28	2.55	2.85
London, UK	3.54	3.24	2.98	1.89	1.65	2.14	2.24	2.39	1.44	2.93	3.18	2.08
Los Angeles, USA	4.31	2.39	6.44	1.79	1.31	2.22	2.18	2.03	1.59	3.05	2.97	2.41
New York, USA	2.97	1.18	6.34	1.57	0.60	2.98	2.25	1.18	3.22	2.78	1.42	3.55
Munich, DE	5.95	3.81	14.99	1.87	1.42	4.44	2.15	1.67	3.4	2.53	2.07	4.07
Orlando, USA	3.84	2.38	7.60	1.06	0.84	1.44	1.34	1.16	1.6	1.77	1.59	2.21
Paris, FR	4.42	3.81	4.66	2.53	1.68	5.45	2.62	2.03	2.76	3.08	2.45	3.44
Seattle, USA	4.77	1.86	18.35	1.14	0.57	2.89	1.68	1.01	3.26	2.14	1.45	3.38

Res* = Resident, Tour** = Tourist.

Table 7
Clustering characteristics for different groups.

City	Mean % in examined to geotagged			Mean % in First Larger Cluster to geotagged			Mean % in Second Larger Cluster to geotagged			Mean % of noise to geotagged		
	Total	Res*	Tour**	Total	Res*	Tour**	Total	Res*	Tour**	Total	Res*	Tour**
Amsterdam, NL	0.34	0.53	0.02	0.31	0.16	0.46	0.13	0.09	0.11	0.05	0.05	0.05
Athens, GR	0.32	0.51	0.00	0.36	0.25	0.55	0.14	0.12	0.19	0.05	0.05	0.07
Copenhagen, DK	0.24	0.40	0.00	0.36	0.26	0.55	0.14	0.11	0.22	0.05	0.06	0.05
London, UK	0.43	0.68	0.01	0.33	0.14	0.62	0.13	0.09	0.21	0.07	0.07	0.07
Los Angeles, USA	0.64	0.92	0.28	0.31	0.07	0.62	0.10	0.04	0.15	0.03	0.02	0.04
Munich, DE	0.28	0.41	0.01	0.35	0.28	0.52	0.14	0.14	0.15	0.07	0.07	0.06
New York, USA	0.74	0.94	0.30	0.15	0.03	0.30	0.08	0.02	0.12	0.01	0.01	0.03
Orlando, USA	0.36	0.59	0.00	0.34	0.19	0.51	0.11	0.08	0.13	0.03	0.03	0.03
Paris, FR	0.35	0.57	0.01	0.34	0.19	0.60	0.14	0.11	0.20	0.05	0.05	0.05
Seattle, USA	0.48	0.69	0.02	0.22	0.09	0.30	0.11	0.05	0.13	0.04	0.02	0.07

Res* = Resident, Tour** = Tourist.

illustrates that there are similarities with regards to activity spaces for different areas.

This study does not come without shortcomings. To begin with, the choice of the social media platform examined bounds the analysis to the specific-online-population, imposing also biases related to the usage of the SM platform by the users. These biases relate both to the primary SM functionality (e.g., Twitter is perceived to be used primarily for staying informed, while Facebook is perceived as a platform to stay connected) and also to the different use of the platform in different countries and by users of different personality Traits (Gil de Zúñiga et al., 2017). Notwithstanding that the perception of privacy and consequently the use of geotagging functionalities could differ between different countries (e.g., Germany versus USA). Additionally, although the sampling method is believed to be suitable for the analysis performed, the number of users should be increased for more conclusive analysis. Additionally, the methods used, could be improved. In this paper, we take a simple approach in defining the residents and tourist groups, resorting to percentages of tweets performed. However, future research could focus on the use of different methods (such as text analysis and language detection) to further the accuracy of the groups distinction.

As presented in Chaniotakis et al. (2019), Latent Class clustering could be used for the investigation of differences and similarities between different groups of individuals; thus could be applied for the case of social media users as well. Finally, the investigation of distributions and the definition of models would further enhance the understanding of the differences and similarities.

Replication and data sharing

The codes for reproducing the results reported in this paper are available at https://github.com/besheer110/Transferability-and-Sample-Specification-for-Social-Media-data-a-Comparative-Analysis-for-Transport/blob/main/Code. Twitter data used in the analysis can be downloaded using a token obtained for scientific research from https://developer.twitter.com/en/docs/authentication/oauth-1-0a/obtaining-user-access-tokens. Data collected for this research cannot be shared publicly based on the data collection agreement of Twitter.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study was partially funded by the DAAD Project (No. 57474280) Verkehr-SuTra: Technologies for Sustainable Transportation, within the Programme: A New Passage to India — Deutsch-Indische Hochschulkooperationen ab 2019, the German Federal Ministry of Education and Research, Bundesministerium für Bildung und Forschung (BMBF), project FuturTrans: Indo-German Collaborative Research Center on Intelligent Transportation Systems, and by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 815069 (project MOMENTUM (Modelling Emerging Transport Solutions for Urban Mobility)).

References

Ali, F., Shaker, E.S., Ali, A., Kwak, K., Kwak, D., 2018. Sentiment Analysis of Transportation Using Word Embedding and Lda Approaches. Korea Telecommunications Society, pp. 1111–1112.

Ali, F.D., Khan, P., El-Sappagh, S., Ali, A., Ullah, S., Kim, K.H., Kwak, K.S., 2019. Transportation sentiment analysis using word embedding and ontology-based topic modeling. Knowl. Based Syst. 174, 27–42.

Alomari, E., Mehmood, R., Katib, I., 2019. Road traffic event detection using twitter data, machine learning, and Apache spark. In: 2019 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing. Internet of People and Smart City Innovation, pp. 1888–1895.

Alomari, E., Katib, I., Albeshri, A., Yigitcanlar, T., Mehmood, R., 2021. Iktishaf+: a big data tool with automatic labeling for road traffic social sensing and event detection using distributed machine learning. Sensors 21, 2993.

Arbex, R., Cunha, C.B., 2020. Estimating the influence of crowding and travel time variability on accessibility to jobs in a large public transport network using smart card big data. J. Transport Geogr. 85, 102671.

Astarita, V., Giofrè, V.P., Guido, G., Vitale, A., 2019. A single intersection cooperative-competitive paradigm in real time traffic signal settings based on floating car data. Energies 12, 409.

Astarita, V., Giofré, V.P., Festa, D.C., Guido, G., Vitale, A., 2020. Floating car data adaptive traffic signals: a description of the first real-time experiment with "connected" vehicles. Electronics 9, 114.

Bachir, D., Khodabandelou, G., Gauthier, V., El Yacoubi, M., Puchinger, J., 2019. Inferring dynamic origin-destination flows by transport mode using mobile phone data. Transport. Res. C Emerg. Technol. 101, 254–275.

Bakalos, N., Papadakis, N., Litke, A., 2020. Public perception of autonomous mobility using ml-based sentiment analysis over social media data. Logistics 4, 12.

Bokings, S.H., Nurmandi, A., Loilatu, M.J., 2020. How twitter works in public transportation: a case study of bus rapid transit in jakarta and semarang. CommIT J. Commun. Inf. Technol. 14, 53–63.

Bonsón, E., Perea, D., Bednárová, M., 2019. Twitter as a tool for citizen engagement: An empirical study of the Andalusian municipalities. Govern. Inf. Q. 36, 480–489.

Bwambale, A., Choudhury, C.F., Hess, S., 2017. Modelling trip generation using mobile phone data: a latent demographics approach. J. Transport Geogr. 76, 276–286.

Chaniotakis, E., Antoniou, C., 2015. Use of geotagged social media in urban settings: empirical evidence on its potential from twitter. In: IEEE 18th International Conference on Intelligent Transportation Systems, (ITSC), IEEE, pp. 214–219.

Chaniotakis, E., Antoniou, C., Pereira, F., 2016. Mapping social media for transportation studies. IEEE Intell. Syst. 31, 64–70.

Chaniotakis, E., Antoniou, C., Aifadopoulou, G., Dimitriou, L., 2017. Inferring activities from social media data. Transportation Research Record. J. Transport. Res. Board 2666, 29–37.

Chaniotakis, E., Davis, A., Aifadopoulou, G., Antoniou, C., Goulias, K., 2019. A latent class cluster comparison of travel behavior between Thessaloniki in Greece and San Diego in California. In: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing. IEEE, pp. 239–248.

Chaniotakis, E., Efthymiou, D., Antoniou, C., 2020. Data aspects of the evaluation of demand for emerging transportation systems. In: Demand for Emerging Transportation Systems. Elsevier, pp. 77–99.

Chaturvedi, N., Toshniwal, D., Parida, M., 2021. Combinatorial approach of feature generation for traffic event detection using social media data: CFGA. Preprint (Version 1) available at. https://doi.org/10.21203/rs.3.rs-876469/v1.

Chen, B., Tu, Y., Song, Y., Theobald, D.M., Zhang, T., Ren, Z., Li, X., Yang, J., Wang, J., Wang, X., Gong, P., Bai, Y., Xu, B., 2021a. Mapping essential urban land use categories with open big data: results for five metropolitan areas in the United States of America. ISPRS J. Photogrammetry Remote Sens. 178, 203–218.

Chen, Y., Chen, C., Wu, Q., Ma, J., Zhang, G., Milton, J., 2021b. Spatial-temporal traffic congestion identification and correlation extraction using floating car data. J. Intell. Transport. Syst. 25, 263–280.

Cheng, Z., Caverlee, J., Lee, K., Sui, D., 2021. Exploring millions of footprints in location sharing services. ICWSM 25, 81–88.

Clauset, A., Shalizi, C.R., Newman, M.E., 2009. Power-law distributions in empirical data. SIAM Rev. 51, 661–703.

Cui, Y., Meng, C., He, Q., Gao, J., 2018. Forecasting current and next trip purpose with social media data and google places. Transport. Res. C Emerg. Technol. 97, 159–174.

National Academies of Sciences, Engineering, and Medicine, 2022. Uses of social media in public transportation. The National Academies Press, Washington, DC. https://doi.org/10.17226/26451.

DePaula, N., Dincelli, E., Harrison, T.M., 2018. Toward a typology of government social media communication: Democratic goals, symbolic acts and self-presentation. Govern. Inf. Q. 35, 98–108.

Ebrahimpour, Z., Wan, W., Velázquez García, J.L., Cervantes, O., Hou, L., 2020. Analyzing social-geographic human mobility patterns using large-scale social media data. ISPRS Int. J. Geo-Inf. 9, 125.

Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96). AAAI Press, pp. 226–231.

Franco, P., Johnston, R., McCormick, E., 2020. Demand responsive transport: generation of activity patterns from mobile phone network data to support the operation of new mobility services. Transport. Res. A: Policy Pract. 131, 244–266.

Frias-Martinez, V., Soto, V., Hohwald, H., Frias-Martinez, E., 2012. Characterizing urban landscapes using geolocated tweets. In: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing. IEEE, pp. 239–248.

Gal-Tzur, A., Grant-Muller, S.M., Minkov, E., Nocera, S., 2014. The impact of social media usage on transport policy: issues, challenges and recommendations. Procedia - Soc. Behav. Sci. 111, 937–946.

García-Palomares, J.C., Gutiérrez, J., Mínguez, C., 2015. Identification of tourist hot spots based on social networks: a comparative analysis of European metropolises using photo-sharing services and GIS. Appl. Geogr. 63, 408–417.

Georgiadis, G., Nikolaidou, A., Politis, I., Papaioannou, P., 2020. How public transport could benefit from social media? evidence from european agencies. In: Conference on Sustainable Urban Mobility. Springer, pp. 645–653.

Gil de Zúñiga, H., Diehl, T., Huber, B., Liu, J., 2017. Personality traits and social media use in 20 countries: how personality relates to frequency of social media use, social media news use, and social media use for social interaction. Cyberpsychol., Behav. Soc. Netw. 20, 540–552.

Golledge, R.G., 1997. Spatial Behavior: A Geographic Perspective. Guilford Press, New York, USA.

Gu, T., Harrison, T.M., Zhu, Y., 2020. Municipal government use of social media: an analysis of three Chinese cities. In: Proceedings of the 53rd Hawaii International Conference on System Sciences. Maui, Hawaii, USA. University of Hawaii, pp. 1803–1812.

Haro-de Rosario, A., Sáez-Martín, A., del Carmen Caba-Pérez, M., 2018. Using social media to enhance citizen engagement with local government: twitter or facebook? New Media Soc. 20, 29–49.

Hasan, S., Ukkusuri, S.V., 2018. Reconstructing activity location sequences from incomplete check-in data: a semi-markov continuous-time bayesian network model. IEEE Trans. Intell. Transport. Syst. 19, 687–698.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., Ratti, C., 2013. Geo-located Twitter as the proxy for global mobility patterns. http://arxiv.org/abs/1311.0680.

Hu, W., Jin, P., 2018. Dynamic origin-destination estimation based on time delay correlation analysis on location-based social network data. In: Transportation Research Board 97th Annual Meeting. Washington DC, USA.

Hu, L., Li, Z., Ye, X., 2020. Delineating and modeling activity space using geotagged social media data. Cartogr. Geogr. Inf. Sci. 47, 277–288.

Huang, J., Lu, X.X., Sellers, J.M., 2007. A global comparative analysis of urban form: applying spatial metrics and remote sensing. Landsc. Urban Plann. 82, 184–197.

Huang, Z., Ling, X., Wang, P., Zhang, F., Mao, Y., Lin, T., Wang, F.Y., 2018. Modeling real-time human mobility based on mobile phone and transportation data fusion. Transport. Res. C Emerg. Technol. 96, 251–269.

Jiang, B., Miao, Y., 2015. The evolution of natural cities from the perspective of location-based social media. Prof. Geogr. 67, 295–306.

Jiang, S., Alves, A., Rodrigues, F., Ferreira Jr., J., Pereira, F.C., 2015. Mining point-of-interest data from social networks for urban land use classification and disaggregation. Comput. Environ. Urban Syst. 53, 36–46.

Kim, J., Lee, J., 2021. An analysis of spatial accessibility changes according to the attractiveness index of public libraries using social media data. Sustainability 13, 9087.

Kim, J., Rasouli, S., Timmermans, H.J., 2018. Social networks, social influence and activity-travel behaviour: a review of models and empirical evidence. Transport Rev. 38, 499–523.

Kourik, J.L., Wang, J., 2017. The intersection of big data and the data life Cycle: impact on data management. Int. J. Knowl. Eng. 3, 32–36.

Lee, J.H., Davis, A.W., Yoon, S.Y., 2016. Activity space estimation with longitudinal observations of social media data. Transportation 43, 955–977.

Lee, J.H., Davis, A., McBride, E., Goulias, K.G., 2019. Statewide comparison of origin-destination matrices between California travel model and twitter. In: Mobility Patterns, Big Data and Transport Analytics. Elsevier, pp. 201–228.

Li, J., Boonaert, J., Doniec, A., Lozenguez, G., 2021. Multi-models machine learning methods for traffic flow estimation from floating car data. Transport. Res. C Emerg. Technol. 132, 103389.

Liao, Y., Yeh, S., Gil, J., 2021. Feasibility of estimating travel demand using geolocations of social media data. Transportation 49, 137–161.

Lin, D., Allan, A., Cui, J., 2015. The impacts of urban spatial structure and socio-economic factors on patterns of commuting: a review. Int. J. Urban. Sci. 19, 238–255.

Lock, O., Pettit, C., 2020. Social media as passive geo-participation in transportation planning–how effective are topic modeling & sentiment analysis in comparison with citizen surveys? Geo Spatial Inf. Sci. 23, 275–292.

Manetti, G., Bellucci, M., Bagnoli, L., 2017. Stakeholder engagement and public information through social media: a study of canadian and american public transportation agencies. Am. Rev. Publ. Adm. 47, 991–1009.
McKenzie, G., Janowicz, K., Gao, S., Yang, J.A., Hu, Y., 2015. POI pulse: a multi-granular, semantic signature–based information observatory for the interactive visualization of big geosocial data. Cartogr. The Int. J. Geogr. Inf. Geovisualization 50, 71–85.
Moyano, A., Moya-Gómez, B., Gutiérrez, J., 2018. Access and egress times to high-speed rail stations: a spatiotemporal accessibility analysis. J. Transport Geogr. 73, 84–93.
Osorio-Arjona, J., García-Palomares, J.C., 2019. Social media and urban mobility: using twitter to calculate home-work travel matrices. Cities 89, 268–280.
Paldino, S., Bojic, I., Sobolevsky, S., Ratti, C., González, M.C., 2015. Urban magnetism through the lens of geo-tagged photography. EPJ Data Science 4, 5.
Pereira, F.C., Rodrigues, F., Polisciuc, E., Ben-Akiva, M., 2015. Why so many people? explaining nonhabitual transport overcrowding with internet data. IEEE Trans. Intell. Transport. Syst. 16, 1370–1379.
Purnomo, E.P., Loilatu, M.J., Nurmandi, A., Salahudin, Z.Q., Sihidi, I.T., Lutfi, M., 2021. How public transportation use social media platform during covid-19: study on jakarta public transportations' twitter accounts? Webology 18, 1–19.
Qian, T., Chen, J., Li, A., Wang, J., Shen, D., 2020. Evaluating spatial accessibility to general hospitals with navigation and social media location data: a case study in nanjing. Int. Res. J. Publ. Environ. Health 17, 2752.
Rahman, R., Redwan Shabab, K., Chandra Roy, K., Zaki, M.H., Hasan, S., 2021. Real-Time Twitter data mining approach to infer user perception toward active mobility. Transport. Res. Rec. 2675, 947–960.
Sadiq, S., Dasu, T., Dong, X.L., Freire, J., Ilyas, I.F., Link, S., Miller, M.J., Naumann, F., Zhou, X., Srivastava, D., 2018. Data quality: the role of empiricism. ACM SIGMOD Record 46, 35–43.
Sari, E.Y., Wierfi, A.D., Setyanto, A., 2019. Sentiment analysis of customer satisfaction on transportation network company using naive bayes classifier. In: 2019 International Conference on Computer Engineering. Network, and Intelligent Multimedia (CENIM), pp. 1–6.
Schaefer, M., Woodyer, T., 2015. Assessing absolute and relative accuracy of recreation-grade and mobile phone GNSS devices: a method for informing device choice. Area 47, 185–196.
Schwanen, T., Dieleman, F.M., Dijst, M., 2001. Travel behaviour in Dutch monocentric and policentric urban systems. J. Transport Geogr. 9, 173–186.
Stead, D., Marshall, S., 2001. The relationships between urban form and travel patterns. an international review and evaluation. Eur. J. Transport Infrastruct. Res. 1, 113–141.
Sulis, P., Manley, E., Zhong, C., Batty, M., 2018. Using mobility data as proxy for measuring urban vitality. J. Spatial Inf. Sci. 16, 137–162.
Tavassoli, A., Mesbah, M., Hickman, M., 2020. Calibrating a transit assignment model using smart card data in a large-scale multi-modal transit network. Transportation 47, 2133–2156.
Thakur, G., Sims, K., Mao, H., Piburn, J., Sparks, K., Urban, M., Stewart, R., Weber, E., Bhaduri, B., 2018. Utilizing geo-located sensors and social media for studying population dynamics and land classification. In: Human Dynamics Research in Smart and Connected Communities. Springer, pp. 13–40.
Utsunomiya, M., Attanucci, J., Wilson, N., 2006. Potential uses of transit smart card registration and transaction data to improve transit planning. Transport. Res. Rec. 1971, 118–126.
Wang, Z., He, S.Y., Leung, Y., 2018. Applying mobile phone data to travel behaviour research: a literature review. Travel Behav. Soc. 11, 141–155.
Williamson, W., Ruming, K., 2020. Can social media support large scale public participation in urban planning? The case of the# MySydney digital engagement campaign. Int. Plann. Stud. 25, 355–371.
Xu, S., Li, S., Wen, R., 2018. Sensing and detecting traffic events using geosocial media data: a review. Comput. Environ. Urban Syst. 72, 146–160.
Yang, C., Xiao, M., Ding, X., Tian, W., Zhai, Y., Chen, J., Liu, L., Ye, X., 2019. Exploring human mobility patterns using geo-tagged social media data at the group level. J. Spatial Sci. 64, 221–238.
Yang, F., Jin, P.J., Cebelak, M., Ran, B., Walton, C.M., 2018. The application of venue-side location-based social networking (vs-lbsn) data in dynamic origin-destination estimation. In: Intelligent Transportation and Planning: Breakthroughs in Research and Practice. IGI Global, pp. 355–375.
Yao, W., Qian, S., 2021. From twitter to traffic predictor: next-day morning traffic prediction using social media data. Transport. Res. C Emerg. Technol. 124, 102938.
Yao, H., Xiong, M., Zeng, D., Gong, J., 2018. Mining multiple spatial–temporal paths from social media data. Future Generat. Comput. Syst. 87, 782–791.
Yap, M., Nijënstein, S., Van Oort, N., 2018. Improving predictions of public transport usage during disturbances based on smart card data. Transport Pol. 61, 84–95.
Ye, Y., An, Y., Chen, B., Wang, J., Zhong, Y., 2020. Land use classification from social media data and satellite imagery. J. Supercomput. 76, 777–792.
Zhang, Z., He, Q., Gao, J., Ni, M., 2018. A deep learning approach for detecting traffic accidents from social media data. Transport. Res. C Emerg. Technol. 86, 580–596.
Zhao, P., Liu, D., Yu, Z., Hu, H., 2020. Long commutes and transport inequity in China's growing megacity: new evidence from beijing using mobile phone data. Travel Behav. Soc. 20, 248–263.
Zheng, X., Chen, W., Wang, P., Shen, D., Chen, S., Wang, X., Zhang, Q., Yang, L., 2016. Big data for social transportation. IEEE Trans. Intell. Transport. Syst. 17, 620–630.
Zhou, X., Yeh, A.G., Yue, Y., 2018. Spatial variation of self-containment and jobs-housing balance in shenzhen using cellphone big data. J. Transport Geogr. 68, 102–108.
Zulfikar, M.T., et al., 2019. Detection traffic congestion based on twitter data using machine learning. Procedia Comput. Sci. 157, 118–124.

Emmanouil (Manos) Chaniotakis is a Lecturer (Assistant Professor) in Transport Modelling and Machine Learning, and the deputy head of MaaSLab (Mobility as a Service Lab) at the Energy Institute of University College London (UCL). His research focuses on modelling and simulation of transportation systems, including conventional and emerging transportation systems, demand modelling, and machine learning in transportation. He holds a Ph.D. from the Technical University of Munich, an M.Sc. degree in Transportation Infrastructure and Logistics from the Delft University of Technology, and an engineering diploma from the Aristotle University of Thessaloniki. Dr. Chaniotakis has authored more than 50 scientific publications in peer-reviewed journals, conferences, and books on the above research topics and has co-edited a book on demand for emerging mobility services.

Mohamed Abouelela is a lecturer and associate researcher in the Chair of Transportation Systems Engineering at the Technical University of Munich, Germany. His research focuses on big data analytics, emerging mobility systems, travel behavior, human factors modeling, and transportation in developing countries. He holds an M.Sc. in Transportation Traffic Demand Management from the Technical University of Munich and a Bachelor of Civil Engineering from Ain Shams University, Egypt.

Constantinos Antoniou is a Full Professor in the Chair of Transportation Systems Engineering at the Technical University of Munich (TUM), Germany. He holds a Diploma in Civil Engineering from NTUA (1995), an M.Sc. in Transportation (1997), and a Ph.D. in Transportation Systems (2004), both from MIT. His research focuses on modeling and optimization of transportation systems, data analytics and machine learning for transportation systems, and human factors for future mobility systems. He has authored more than 400 scientific publications, including more than 150 papers in international, peer-reviewed journals.

Konstadinos Goulias has three degrees in Civil Engineering earned in 1986 (B.S. and M.S. in Italy), 1987 (MSCE at University of Michigan), and in 1991 (Ph.D. in University of California at Davis). From 1991 to 2004, he was a Professor of Transportation in the Civil and Environmental Engineering Department of Penn State University. He is now a Professor of Transportation in the Department of Geography at the University of California Santa Barbara where he is teaching courses in Transportation Planning, Modeling, and Simulation as well as Smart Cities. He has more than 340 papers and reports to sponsors and is the co-Editor-in-Chief of the journal Transportation Letters. He worked in the Netherlands, Italy, Japan, Germany, Portugal, Australia, and USA, developing new modeling techniques, simulation frameworks, and expert reviews of technologies and engineering policies. See also www.kostasgoulias.com.