Computers, Environment and Urban Systems 100 (2023) 101934

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Understanding temporal and spatial patterns of urban activities across demographic groups through geotagged social media data

Haifeng Niu *, Elisabete A. Silva

Laboratory of Interdisciplinary Spatial Analysis (LISA), Department of Land Economy, 16-21 Silver Street, Cambridge CB3 9EP, England, United Kingdom

ARTICLE INFO

Keywords:
Urban activities
Human pattern
Location-based social media
Social media enrichment
Deep learning
Demographic inference

ABSTRACT

Large-scale geotagged social media data have been increasingly used for exploring human movement patterns in cities. Challenges of this new data type, such as non-representative users and the lack of activity purposes, remain unsolved and limit its applications in exploring activity-based human patterns in cities. To deal with the above challenges, this paper proposed an analytical framework of social media data enrichment — by revealing the demographic composition of non-representative social media data users and inferring activity purposes of geotagged posts — for better exploring spatial-temporal patterns of human activity in cities. A deep learning model is employed to reveal social media users' age and gender groups from user names, profile images, biographies, and language settings. Eight types of activity purposes are inferred from embedded geo-location by spatially joining with fine-scale building and land use data. Using Greater London as the case study, this paper explores the temporal dynamics of activity purposes with heatmaps of hourly frequency of tweets and identifies spatial differences across age and gender groups using hotspots analysis (Getis–Ord Gi* statistics). This paper demonstrates the application of geotagged social media data in identifying spatial, temporal and demographic patterns of urban activities, which potentially helps shape better place-based and age/gender-sensitive urban policies and planning decisions.

1. Introduction

Geotagged social media data such as Twitter data have been increasingly used in urban analytics as this type of new data with precise timestamps and fine-scale geographic locations records 'whereabouts' of individuals in cities. It provides opportunities for researchers to explore collective patterns in many domains such as mobility patterns (Huang, Yao, Krisp, & Jiang, 2021; Osorio-Arjona & Garcıa-Palomares, 2019; Wu, Zhi, Sui, & Liu, 2014; Yuan, Zheng, & Xie, 2012), urban activities (Ouyang, Fan, Wang, Zhu, & Yang, 2022; Wu et al., 2014; Wu, Ye, Ren, & Du, 2018) and event detection (Xu et al., 2016, 2020) by following a data-driven approach. Despite the potential advantages, some general challenges appear when using geotagged social media data for understanding human patterns in cities. Firstly, the lack of activity purposes embedded in geotagged social media posts is a challenge for exploring activity-based patterns. Although the specific coordinate (latitude and longitude) is embedded in geotagged social media posts, this geographic information is not directly linked to activity locations and needs to be interpreted for activity purposes (Cui, Meng, He, & Gao, 2018).

Another challenge is that social media data are non-representative samples as the users mostly under-represent certain groups such as the elderly and some ethnic minorities. For instance, according to the study of Longley, Adnan, and Lansley (2015), young men and white British people are overrepresented among Twitter users in London, while middle-aged/older women, South Asian, and Chinese users are underrepresented. Thus, the spatio-temporal patterns identified from social media data in London may be biased and not representative of the general population. Moreover, the limited access to data sources can also lead to getting biased samples during data collection. Taking Twitter data as an example, only about 1% of Twitter users are willing to generate geotagged tweets and only 1% of tweets are accessible via the free Twitter Streaming API (Huang & Wong, 2016; Morstatter, Pfeffer, Liu, & Carley, 2013). Malik, Lamba, Nakos, and Pfeffer (2015) examined the population (i.e., gender) bias of geotagged tweets by conducting statistical tests based on tweets and census data in the United States, arguing that geotagged tweets are not representative of the US population. This problem is also noted for other social media data sources, such as Weibo in China (Yuan, Wei, & Lu, 2018). This challenge generally

* Corresponding author.
E-mail address: hn303@cam.ac.uk (H. Niu).

https://doi.org/10.1016/j.compenvurbsys.2022.101934
Received 21 July 2022; Received in revised form 28 November 2022; Accepted 19 December 2022
Available online 17 January 2023

applies to overall urban research employing social media data and has been extensively cited (Kitchin, 2013; Li et al., 2016; Tufekci, 2014). However, it still remains unsolved according to recent reviews and studies (Martı, Serrano-Estrada, & Nolasco-Cirugeda, 2019; Niu & Silva, 2020). With this challenge in mind, there is an urgent need to properly employ this type of geographic big data and unlock its true value for understanding cities better.

When zooming into the context of identifying activity patterns, one question must be answered before exploring the spatio-temporal patterns from the geotagged social media data: who generate the geotagged posts? Specifically, researchers need to reveal the composition of users, e.g., their sociodemographic attributes. This is important because users generating those footprints in cities are only a sample of the general population. Exploring activity patterns from social media data without understanding the demographic composition of data generators might underrepresent certain groups, for instance, senior and female groups in the context of Twitter. Moreover, individuals' demographic backgrounds and socioeconomic status have a significant impact on their activity patterns in cities (Huang & Wong, 2016; Lenormand et al., 2015). Exploring activity patterns of people with different sociodemographic characteristics can help to better understand socio-spatial segregation, mobility inequality and travel demand. Thus, this study aims to answer the following research questions: who are the data contributors of geotagged social media? Seeing through these data, how does it help to reveal the difference in activity patterns across demographic groups? Using Greater London as the case study, this paper firstly proposes an analytical framework to enrich social media data for exploring activity patterns: 1) By employing a deep learning model, certain demographic characteristics (i.e., age and gender) of non-representative Twitter users are inferred from user names, profile images, biographies and language settings; 2) Activities purposes (i.e., commercial, recreation, residential, work, transport, medical, education and sports) of geotagged posts are inferred from their geographic locations that are spatially joined with fine-scale building and land use data. Then, this paper explores the temporal and spatial patterns of urban activities from enriched Twitter data. The proposed hourly frequency heatmap method measures standardised hourly frequencies of tweets over the daily number of tweets for each type of activity. This allows comparing temporal dynamics across paired groups (e.g., weekdays and weekends, and daytime and nighttime). In exploring spatial patterns, standardised differences of tweets in hexagon grids are measured and examined for the local spatial association (Getis-Ord Gi statistic). The spatial association results help to identify the spatial clusters of high or low standardised differences of tweets between groups, revealing the specific locations where age and gender groups have different patterns of urban activities.

The paper is organised as follows. Section 2 reviews the applications of geotagged social media data in exploring human activity patterns and also the demographic inference of geotagged social media data. Section 3 introduces the study area and data used in this study. Section 4 explains the methodology employed herein, including data cleaning, data preparation, data enrichment, and metrics and methods for spatial, temporal and demographic pattern recognition. Section 5 presents the results. Section 6 further discusses results, contributions, limitations and further directions.

2. Related work

2.1. Exploring human activity patterns using geotagged social media data

Among studies exploring activity patterns from geotagged social media, a group of researchers have focused on activity spaces, i.e., the spatial extent of individuals' daily activities, by aggregating the spatio-temporal information at the individual level. Swier, Komarniczky, and Clapperton (2015) linked clusters of geotagged tweets from the UK with building-level functions to identify users' residential zones. Shelton, Poorthuis, and Zook (2015) used social media data to examine the actual activity spaces of segregated neighbourhoods and explored the potential of social media data in examining socio-spatial mobility. With activity patterns extracted from geotagged social media data, studies such as Huang and Wong (2016) and Hu, Li, and Ye (2020) further explored the relationship between activity spaces and socioeconomic statistics from census data. To quantify activity patterns, the aforementioned studies mainly used spatial metrics such as distance of displacement, the radius of gyration, numbers of distinct locations, core activity locations and Shannon entropy of users' trajectories. Another group of researchers used geotagged social media data to infer activity purposes and explore the patterns of activity behaviours, including activity choice patterns, activity forecasts and lifestyles (Rashidi, Abbasi, Maghrebi, Hasan, & Waller, 2017). To infer activity purposes, previous studies have either used check-ins (e.g., Foursquare and Yelp review), which record individual activity with specific locations, or general geotagged social media data by spatially joining land use data or points of interest (POI) data. Hasan and Ukkusuri (2014) integrated Foursquare check-in data with topic modelling to model the activity choice patterns of eight activity categories –home, work, eating, entertainment, recreation, shopping, social services and education. Huang and Li (2016) identified the purposes of activities tagged in social media by spatially joining social media data with land use data to explore the semantic patterns from activity labels formed by time and location information. Combined with fine-scale POI data from Google Places, Cui, Xie, and Liu (2018) used geotagged tweets to infer and predict trip purposes in the categories of eating out, personal, recreation, education, shopping and transportation.

2.2. Sociodemographic inference from social media data

Beyond the spatial and temporal patterns of activities, researchers are also interested in exploring the sociodemographic characteristics of social media users (see Table 1). Some of them assigned sociodemographic attributes to social media users by linking users' visited locations (i.e., activity zones or home locations) with either local knowledge of sociodemographics or census data. Shelton et al. (2015) identified the 'home' neighbourhoods of social media users according to the frequency of their geotagged check-ins in two locally-identified segregated neighbourhoods and then examined the activity patterns of the two groups in terms of socio-spatial inequality. Huang and Wong (2016) identified home and work sites for social media users and inferred their socioeconomic status by matching with census data. Exploring the consuming activities of different social groups, Davis, Dingel, Monras, and Morales (2019) inferred individual home or work locations by mining location-related text from Yelp reviews and locating them in census tracts with demographic and income information. Although this data matching method enriches the individual data with collective attributes, it may be subject to the ecological fallacy, i.e., incorrect inferences about individuals based on aggregate statistics, when drawing conclusions about activity patterns (Li, Ban, Wechsler, & Xu, 2018). To extract individual-level demographics, several studies have introduced name analysis to social media data inference using names from users' profiles. Mislove, Lehmann, Ahn, Onnela, and Rosenquist (2011) found that 64.2% of Twitter users use first names in their profile, although 71.8% of these users are male.

Longley et al. (2015) also used name analysis to infer the ethnicity, gender and age makeup of users from Twitter data. By matching the Twitter data with a name database containing over 300 million names, the study extracted forenames and surnames, which were each assigned a predicted ethnicity, age and gender. Of the sample, 63.6% of the data had probable forenames or surnames, and 49.1% of the data were coded to age and gender. Similarly, Luo, Cao, Mulligan, and Li (2016) performed surname analysis with the Bayesian Improved Surname Geocoding method to identify the ethnicity of users and forename analysis for gender and age by linking to the US Social Security database. In such studies, forenames are generally used for age and gender prediction,

Table 1
Studies of sociodemographic inference from geotagged social media data.

Studies	Data	Information used	Method	Algorithm	Attributes	Supplementary data
Shelton et al. (2015)	Twitter	Check-ins	Data matching with inferred activity zone	Frequent visits	Sociodemographics (segregated neighbourhoods)	–
Huang and Wong (2016)	Twitter	Check-ins	Data matching with inferred home location	Spatial clustering	Socioeconomic status (median house value)	American Community Survey (ACS) data
Davis et al. (2019)	Yelp reviews	Check-ins	Data matching with inferred home location	Location detection	Socioeconomic status (consumption)	–
Jiang, Li, and Ye (2019)	Twitter	Check-ins	Data matching with inferred home location	Frequent visits	Demographics (age, gender, ethnicity, education, household income)	American Community Survey (ACS) data; 2010 Decennial Census
Mislove et al. (2011)	Twitter	User names	Name analysis	–	Demographics (age, gender)	Social Security Administration 2010
Longley et al. (2015)	Twitter	User names	Name analysis	OnoMap software[1]	Demographics (age, gender, ethnicity)	2011 United Kingdom census; OnoMap taxonomy
Luo et al. (2016)	Twitter	User names	Name analysis	Bayesian Improved Surname Geocoding	Demographics (age, gender, ethnicity)	Social Security Administration 2014
Yuan et al. (2018)	Weibo	Biography	Data detection	–	Demographics (gender)	–
Zagheni, Garimella, and Weber (2014)	Twitter	Profile images	Face recognition	Face++ API[2]	Demographics (age, gender)	–
An and Weber (2016)	Twitter	Profile images	Face recognition	Face++ API	Demographics (age, gender, ethnicity)	–
Alowibdi, Buy, and Yu (2013)	Twitter	Profile colour	Supervised machine learning	Classifiers	Demographics (gender)	–
Liu and Ruths (2013)	Twitter	User names	Supervised machine learning	Support vector machine (SVM)	Demographics (gender)	–
Fang, Sang, Xu, and Hossain (2015)	Google+	User names, biography, posts	Supervised machine learning	Relational latent SVM	Demographics (age, gender, ethnicity)	–
Chen, Wang, Agichtein, and Wang (2015)	Twitter	User names, biography, posts, social networks, profile images	Supervised machine learning	SVM	Demographics (age, gender, ethnicity)	–
Zhong, Yuan, Zhong, Zhang, and Xie (2015)	Weibo	Check-ins	Machine learning	Classifiers/ Regression	Sociodemographics (gender, age, education, marital status)	–
Guimaraes, Rosa, Gaetano, Rodrıguez and Bressan (2017)	Twitter	Posts	Deep learning	Deep convolutional neural network	Demographics (age)	–
Wang et al. (2019)	Twitter	User names, biography, posts, profile images	Deep learning	Multimodal deep neural architecture	Demographics (age, gender)	–

[1] https://www.onomap.org/
[2] https://www.faceplusplus.com/

while surnames are used for ethnicity prediction (Cesare, Grant, Nguyen, Lee, & Nsoesie, 2018). However, those matching methods only consider names from profiles, neglecting other profile information such as photos, biography and posts.

In recent years, researchers have taken advantage of machine learning and deep learning techniques to identify and predict demographics (see Table 1). Several studies have trained supervised classifiers to predict the sociodemographics of social media users based on features from profile information, such as profile images (An & Weber, 2016; Zagheni et al., 2014) and screen name (Liu & Ruths, 2013; McCormick, Lee, Cesare, Shojaie, & Spiro, 2017), to extended information such as profile colour (Alowibdi et al., 2013), posts (Guimaraes, Rosa, Gaetano, Rodrıguez and Bressan, 2017), network (Fang et al., 2015), check-ins (Zhong et al., 2015) and self-reported biography (Chen et al., 2015; Wang et al., 2019). In general, features such as posts, the profile description, profile colour and network are used for gender prediction, while profile photos and forenames are used for age prediction. For ethnicity prediction, previous studies have included features such as profile photos, surnames and posts. In summary, the machine learning methods, along with some deep learning advances in natural language processing and face recognition, have offered new opportunities for inferring sociodemographic characteristics of social media users with increasing accuracy (Cesare et al., 2018). However, few studies have leveraged the advance in demographic detection to tackle the representativeness issue with social media data. To fill this gap, this study will utilise deep learning algorithms to infer the

sociodemographics of social media users before exploring the spatio-temporal patterns of urban activities among different age and gender groups.

3. Study area and data

The study area is Greater London which includes 32 London Boroughs (i.e., local authority districts). According to the census output area population estimates (mid-2019) by the Office of National Statistics (ONS), there are 8.96 million people living in the region of Greater London, an area of 1572 km^2. Several data sets are used in this study:

Twitter data The initial Twitter data in Greater London were collected through the Twitter Streaming API, which provides a 1% sample of tweets from the database. Although the retrieved tweets make up only 1% of the database, they essentially cover 90% of all geotagged tweets (Morstatter et al., 2013). We used the Streaming API to collect tweets by setting a bounding box of Greater London.[1] Although Twitter data show periodical patterns across times of day and days of the week, short-term observations from social media data are not sufficient to capture activity patterns at the individual level (Lee, Davis, Yoon, & Goulias, 2016). Thus, we chose a period of over one year (starting from 2019) to include sufficient data as observations for exploring individuals' activities. Considering that the COVID-19 pandemic has dramatically impacted the movement of citizens in the city, we only included the tweets before the first lockdown in London, resulting in a period of 1 January 2019 to 1 March 2020.

The data collection was collected via the python package Tweepy (Version 3.5.0) and stored in a PostgreSQL database. Each entry in the database included attributes related to the tweet and the user, such as created time of the post, text, language, timestamp, user name, location of the user, profile image link, the profile description, etc. Most importantly, it contained attributes regarding the geo-locations of tweets.

Land and building use data Building use data were acquired from Geomni UKBuildings,[2] a spatial property database that provides individual building characteristics such as use, age and residential types. There are 25 classes of detailed building use, such as retail, recreation, transport, educational use and residential. There are, in total, 2,385,214 building features with specific uses in Greater London. Another dataset used was the Ordnance Survey MasterMap Greenspace layer,[3] which includes open space features (e.g., public parks, garden, and sports grounds) in London.

4. Methodology

4.1. Data cleaning and preparation

Given the focus of this study on spatio-temporal patterns of geotagged social media in London, we first removed bot and inactive accounts according to the following standards. Table 2 shows the number of tweets and users for the above data cleaning process.

- Created time of tweets are within the time range from *2019 to 01-01 00:00:00 + 00* to *2020-03-01 00:00:00 + 00*.
- Geotags referring to bounding boxes of administrations or cities are removed. Only tweets geotagged by coordinates objects and place objects (place type = poi) are selected. (See Fig. 1 for the distribution of distinct point-based geotags in Greater London).
- Geotagged tweets outside of the Greater London boundary are removed.

Table 2
Filtering process with numbers of tweets and users.

Cleaning process	Number of tweets	Users
Tweets collected between Jan 2019 and March 2020	18,169,124	611,348
Tweets with valid geolocation	15,665,004	532,959
Tweets in Greater London Bounding Box	15,196,971	503,847
User filtering process (separately)		
Users whose total number of tweets is <3500[1]	13,349,368	503,665
Users sending at least two geotagged tweets in one day	13,710,968	248,557
Users with spatial mobility[2]	13,319,851	265,913
Tweets after user filtering[3]	11,278,041	205,026

[1] By manually checking high-frequency accounts, we found that users who generated over 3500 geotagged tweets during the whole time period are almost bots.
[2] Users whose movement distance is over 50 m.
[3] Around 62% of tweets are left after the filtering process.

The next step included filtering users such as inactive accounts, bot accounts and spatially inactive account.

- Users who send over 3500 geotagged tweets are identified as bot accounts.
- By summarising daily tweets sent by users, we identify users as active accounts when they send at least two geotagged tweet at least in one day.
- Since this study focuses on the activity of Twitter users, we exclude users with low spatial mobility, referring to users with tags in the same location. To do this, we calculate the centroid of geo-locations for each user and measure the movement distance between the centroids of all geotagged tweets. Users with a movement distance of fewer than 50 m are tagged as spatially inactive accounts and removed from the dataset.

The timestamp recorded in Twitter metadata provides accurate temporal information, allowing for exploring the dynamics of urban activities. Considering the hourly and daily periodicity in human mobility, we extracted the time of day (24 h) and the day of the week (7 days) for each tweet (Song, Koren, Wang, & Barabasi, 2010). We first labelled the tweets sent from 7 am to 7 pm as daytime tweets and those sent from 7 pm to 7 am as nighttime tweets. We also labelled tweets as weekday tweets (from Monday to 5 pm Friday) and weekend tweets (from 5 pm Friday to Sunday).

4.2. Data enrichment with inferred attributes

To explore the activity patterns of different demographic groups from tweets, two types of attributes needed to be extended via data enrichment: demographic characteristics of users and the activity purposes of their geotagged tweets (see Fig. 2). In inferring demographics, we utilised user metadata (i.e., profile image, username, screen name and biography) for training a deep learning model that predicts demographic characteristics, including age and gender.[4] Another data enrichment method intended to reveal the activity purposes related to the locations of tweets. However, only a small number of geotagged tweets were specifically linked with POIs. In this study, we inferred the activity purpose by following two steps: first, we classified the building and land features in London into several activity types (e.g., commercial, recreation, residential, and work) according to their specific uses; then,

[1] Coordinates of the bounding box in WGS84 (EPSG:4326): [−0.510375, 51.28676, 0.334015, 51.691874].
[2] https://www.geomni.co.uk/ukbuildings
[3] https://www.ordnancesurvey.co.uk/business-government/products/mastermap-greenspace

[4] We follow the recommendation from General Data Protection Regulation 2016 (https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679) and Twitter Developer Agreement and Policy (https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases) to collect, store and process data.

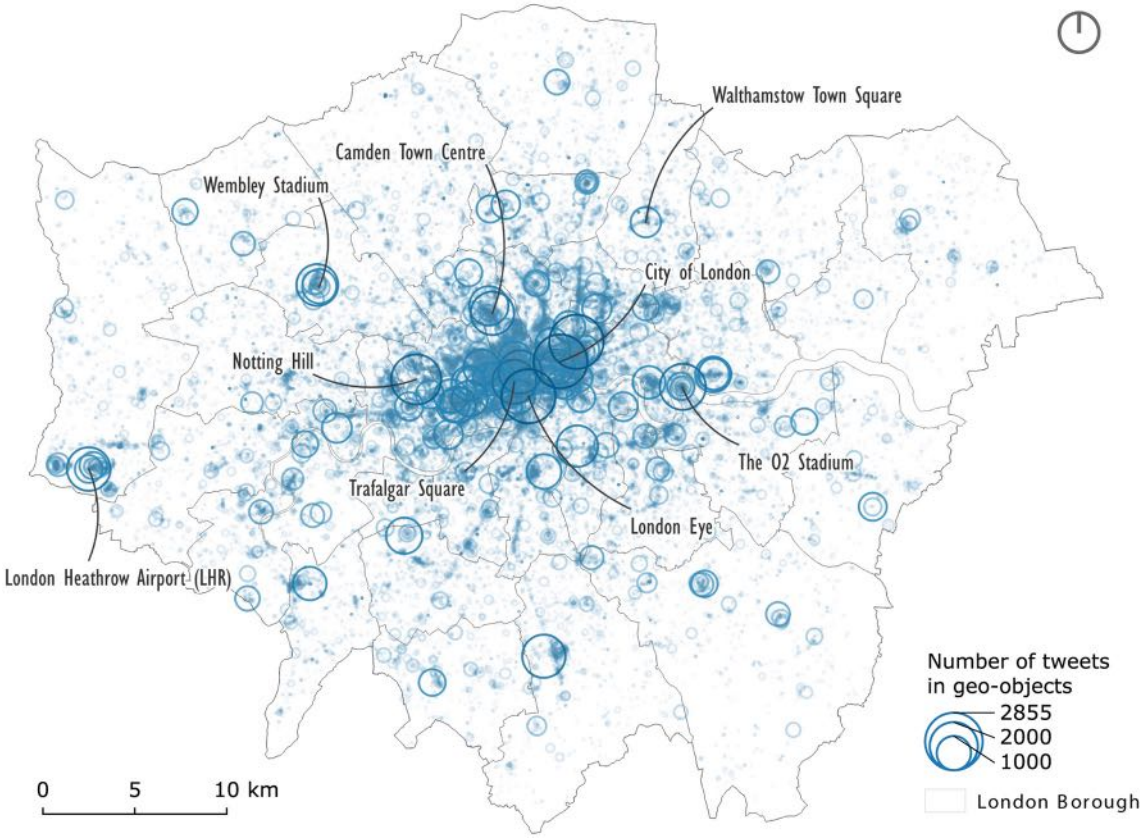


Fig. 1. Distribution of distinct point-based geotags in Greater London. Distinct locations tagged in tweets are represented as circles whose size refers to the total number of tweets at the location. Only tweets geotagged by coordinates objects and place objects (place type = *poi*) are selected in this study. See Appendix A for visualisations of different geographic objects tagged in Greater London.

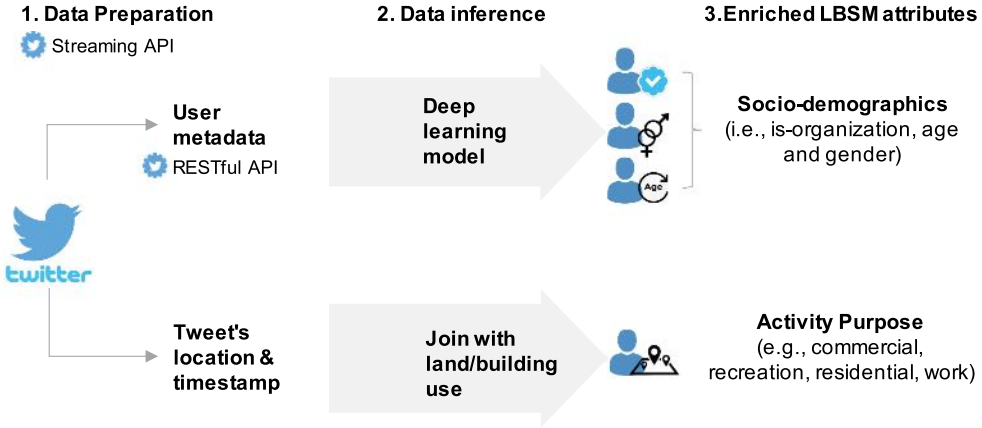


Fig. 2. Data enrichment method for Twitter data.

the geotagged tweets were spatially joined with the above spatial features assigned by activity purposes.

4.2.1. Inferring demographics from user profile data with deep learning

The demographic inference was based on an open-source model – M3Inference – developed by Wang et al. (2019). The M3Inference model can infer gender, age and organisation identity from multilingual social media data. The model is also multimodal since it can integrate both text and image models, which has been neglected in previous studies. Compared to previous sociodemographic inference methods, this model takes full advantage of Twitter metadata such as a user's name, profile image, biography and language setting that can be accessed via the RESTful API. The full model includes two pipelines for learning images (i.e., profile images) and text (i.e., username, screen name, and biography). Specifically, the image model uses DenseNet (Huang, Liu, van der Maaten, & Weinberger, 2017), a state-of-the-art neural network for visual object recognition, while the text model trains 2-layer long short-term memory networks with character-based embedding (i.e., the character embedding is constructed separately for different languages). Both the image and text models are joined to the full model with a modality drop-out layer, two fully connected layers of size 2048 and a rectified linear unit layer (see Fig. 3). The outputs are separate layers for each task, i.e., gender, age and isorganisation identification, with gender and is-organisation being binary classifications. The categories for age

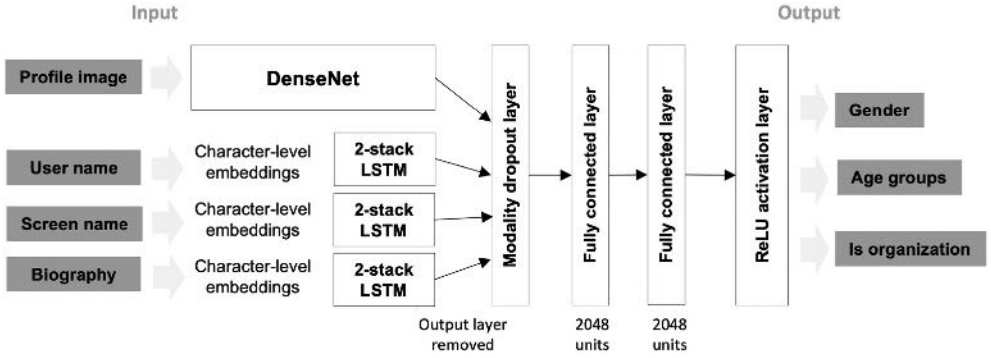


Fig. 3. M3Inference model (adapted from Wang et al. (2019)).

have four levels: ≤ 18, (18, 30), [30, 40) and [40, 99). It is worth noting that the model has been evaluated with the gender and age distribution of the European population dataset (including the UK) provided by the European Statistical Office. M3 model achieved macro F1-scores (i.e., the weighted average of precision and recall) of 0.918, 0.522 and 0.898 in gender, age are organisation status, which outperforms all tested model on Twitter data. Readers are referred to Wang et al. (2019) for more details regarding the m3inference model.

4.2.2. *Inferring activity purposes by joining with building/land use data*

To infer activity purposes from geotagged social media, we followed a data matching method by spatially joining the geotagged tweets with building use data that indicates potential human activity Huang and Li (2016). However, public data for land use classification in the UK, such as the Generalised Land Use Database, only provide simplified classifications (e.g., domestic buildings, non-domestic buildings, roads, green space, and water), which is insufficient as a reference for activity inference, especially in Greater London. To fill this gap, we combined the building use data from Geomni UKBuildings (GUKB) and land features (e.g., public park, garden, sports grounds) from the Ordnance Survey MasterMap Greenspace (OSMG) layer as the reference to infer activity purposes (see the visualisation in Appendix B). We assigned building and land uses based on eight activity types: commercial, recreation, residential, work, transport, medical, education and sports (see Table 3). For each geotagged tweet, we conducted a nearby search and spatially joined the coordinates of the tweet with the nearest building/land features with related activity types. Tweets whose distance to the nearest building/land was over 300 m were eliminated from the dataset. Four categories of attributes were gathered from data enrichment: tweet metadata, user profile, user sociodemographics and activity purpose. The detailed attributes and their acquisition methods are listed in Table 4.

4.3. *Temporal pattern of activity-based tweets with hourly frequency heatmap*

To understand the temporal rhythms of different types of activity, we measured the standardised hourly frequencies of tweets (*z*-score) over the daily number of tweets in each type of activity. We calculated the hourly frequency of tweets separately for eight types of activities using the Eqs. (1) and (2) as:

$$Z_{a,h} = \frac{N_{a,h} - \overline{X}_a}{S_a} \tag{1}$$

$$\overline{X}_a = \frac{\sum_{h=1}^{H} N_{a,h}}{H}, S_a = \sqrt{\frac{\sum_{h=1}^{H} N_{a,h}^2}{H} - (\overline{X}_a)^2} \tag{2}$$

where $N_{a,h}$ refers to the number of tweets related to activity a sent during the hour h, X_a and S_a are the mean and standard deviation of the hourly frequency of tweets related to activity a. The matrix of hourly frequency of activity P is given by using activity types as rows and hours as columns:

$$Z = \begin{bmatrix} Z_{1,1} & Z_{1,2} & \dots & Z_{1,h} & \dots & Z_{1,H} \\ Z_{2,1} & Z_{2,2} & \dots & Z_{2,h} & \dots & Z_{2,H} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ Z_{i,1} & Z_{i,2} & \dots & Z_{i,h} & \dots & Z_{i,H} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ Z_{A,1} & Z_{A,2} & \dots & Z_{A,h} & \dots & Z_{A,H} \end{bmatrix} \tag{3}$$

where A refers to the total number of activity types and H refers to the hour of the day (in 24-h time). The sum of $[Z_{1,1}, Z_{1,2} \dots Z_{1,j} \dots Z_{1,H}]$ is equal to 1. For better visualisation, we converted the hourly frequency matrices of activity into heatmaps.

To compare the temporal patterns of activity between different age or gender groups, we first calculated the matrix for each subset and measured the difference using matrix subtraction. For example, we calculate the difference in temporal patterns between male and female

Table 3
Activity identification based on building and land use in London.

Activity types	Building/land use classifications	Data source[1]
	General commercial - mixed use	GUKB
	General commercial - mixed use - derelict	GUKB
Commercial	Retail - petrol station	GUKB
	Retail - with more recent extensions of different type construction/age	GUKB
	Retail only	GUKB
	Retail with office/residential above	GUKB
	Recreation and leisure	GUKB
Recreation	Public park or garden	OSMG
	Religious grounds	OSMG
	Community - religious	GUKB
	Community - institutional and communal accommodation	GUKB
Residential	Residential only	GUKB
	Residential with retail on ground floor	GUKB
	Industry - manufacturing/processing	GUKB
	Office only	GUKB
	Office with retail on ground floor	GUKB
Work	Community - governmental (central and local)	GUKB
	Storage/warehousing	GUKB
	Storage/warehousing with linked office block	GUKB
	Utilities	GUKB
Transport	Transport	GUKB
Medical	Community - health	GUKB
Education	School grounds	OSMG
	Community - educational	GUKB
Sports	Tennis court	OSMG
	Other sports facility	OSMG

[1] GUKB: Geomni UKBuildings dataset; OSMG: Ordnance Survey MasterMap Greenspace layer.

Table 4
Enriched attributes of geotagged social media data.

Attribute type	Attribute name	Acquisition method
Tweet metadata	tweet_id	Metadata
	user_id	Metadata
	created_at	Metadata
	geom	Metadata
	hour	Extracted from metadata
	day	Extracted from metadata
	weekdays/ weekends	Extracted from metadata
	daytime/nighttime	Extracted from metadata
User profile	user_id	Metadata
	user_name	Metadata
	user_bio	Metadata
	user_profile_image	Metadata
User sociodemographics	age	Inferred by deep learning
	gender	Inferred by deep learning
	is/non-organisation	Inferred by deep learning
	home borough	Inferred by spatio-temporal clustering
Activity purpose	activity type	Inferred from building/land use data

users in London by subtracting the z-score of male users, $Z_{(gender=male)}$, from the z-score of female users, $Z_{(gender=female)}$. When measuring the standardised difference in temporal patterns across demographic groups, we chose the matrix of overall hourly frequency of activities, $Z_{(Subset)}$, as the subtrahend and the equivalent for specific gender and age subsets $Z_{(all)}$, as the minuend.

4.4. Spatial pattern of activity-based tweets with hotspot identification

To explore the spatial pattern of activities, we created a hexagon grid with a perimeter of 200 m to aggregate tweets in the city. Considering the uneven sample sizes, we calculated z-scores, i.e., standardised frequency of tweets, for all hexagonal cells. Following the method used in Lansley and Longley (2016), we then compared the z-scores in the same cell between different subgroups to delineate their spatial disparities. For instance, when identifying the spatial patterns of gender groups, we split the entire dataset of tweets into two subsets according to the gender attribute of Twitter users. For each cell, we calculated the standardised frequencies of tweets separately for the male and female groups. By using the standardised frequencies (z-scores) of tweets sent by men to subtract the equivalent for women, we achieved the standardised differences (z-score difference) of tweets between men and women in the hexagon cell. A positive standardised difference in a cell means there were more tweets sent from female users than male users at that specific location. This analysis was also applied to the exploration of the spatial differences between tweets sent from the four age groups of Twitter users at different times of day (i.e., daytime and nighttime) and days of the week (i.e., weekdays and weekends).

However, hexagon cells with high or low values (i.e., z-score difference) may not be statistically significant hotspots. To be a statistically significant hotspot, the cell should have a high value while being surrounded by other cells with high values. The same applies to cold spots. Thus, we utilised **hotspot analysis (Getis–Ord Gi* statistic)** to measure the local spatial association of high or low standardised differences between different groups. The Getis–Ord Gi* statistic is given as:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \overline{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{\left[\sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j}\right)^2\right]}{n-1}}} \quad (4)$$

$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - (\overline{X})^2}, \overline{X} = \frac{\sum_{j=1}^{n} x_j}{n} \quad (5)$$

where x_j is the z-score difference in hexagon cell j, $w_{i,j}$ is the spatial weight between cells i and j, and n is equal to the total number of hexagons. The results of the Getis–Ord Gi* statistical analysis include a G_i^* score for each spatial feature and a p-value as its statistical significance (95%, 99%, 99.9%). To understand the activity pattern within the hotspots, we further calculated the hourly frequencies of tweets only located in the hotspot cells by type of activity. We visualised the statistics as a heatmap by using the method explained in Section 4.3, but summarised by boroughs.

The methodology was implemented on a machine running the Windows 10(x64) operating system. PostgreSQL with the PostGIS extension was used to manage the geospatial data and perform geoprocessing. We calculated the Getis–Ord Gi* statistics with GeoDa version 1.18. The results were visualised with QGIS version 3.10 and Python version 3.7 with the Matplotlib (https://matplotlib.org/) and Seaborn (https://seaborn.pydata.org/) packages.

5. Results

5.1. Statistics of social media users with inferred demographic characteristics

The results suggest that there is an overrepresentation of men among the Twitter users in Greater London (see Table 5). The ratio of male Twitter users to all Twitter users in London is 15% higher than that of official population estimates of mid–2019 (Park, 2020). Compared to the results of a previous study by Sloan (2017), this ratio is also 7.9% higher than the one found in the Twitter dataset of users in Great Britain. In terms of age, underrepresentation was noted both in the groups of young (aged <18) and senior citizens (aged over 40). On the contrary, the ratio of Twitter users aged 19–29 to users of all ages is 17.8% higher than that of the official population estimates (see Table 6). Fig. 4 shows the difference in gender ratios for each age group between Twitter data and government statistics in London. The comparison highlights that the overrepresentation of men in the sample of Twitter users exists in each age group, but especially for the 30–39 and over 40 age groups.

5.2. Temporal and spatial patterns of activity-based tweets

5.2.1. Daily dynamics

The results of the hourly patterns of tweets for the eight types of activities show that residential-related tweets dominate the whole dataset, followed by tweets inferred as recreation and work activities (see Fig. 5a). In total, around 52.3% of 11,278,041 tweets are identified as residential-related, and 21.9% of all tweets are identified as recreation-related. The proportion of work-related tweets is 11.34%. Fig. 5b depicts the heatmap of standardised hourly frequencies of tweets. It shows that most of the geotagged tweets in London were sent between 9 am and 10 pm. In each row, the continuous cells in the red colour reveal the peak times of tweets related to the specific type of activity. Focusing on the peak hours of activities inferred from geotagged tweets, residential and recreation activities share a similar pattern in that most of the related tweets were sent during the nighttime from 6 pm to 10 pm. Work- and medical-related tweets were primarily

Table 5
Population of Twitter users by gender.

	Male	Female
London Twitter users	66.2%	33.8%
London Twitter users (refined)[1]	64.9%	35.1%
London Population estimates 2019[2]	49.9%	50.1%
GB Twitter users[3]	57.0%	43.0%

[1] Refined by excluding tweets from organisation accounts.
[2] Population estimates by Office for National Statistics (ONS).
[3] Source: Sloan (2017).

Table 6
Population of Twitter users by age.

	≤18	19–29	30–39	>40
London Twitter users	10.0%	31.3%	22.0%	36.6%
London Twitter users (refined)[1]	10.2%	33.7%	22.8%	33.3%
London Population estimates 2019[2]	23.7%	15.9%	18.1%	42.3%

[1] Refined by excluding tweets from organisation account.
[2] Source: Population estimates by Office for National Statistics (ONS).

sent during working hours (9 am to 7 pm) and peaked around noon. The active time for transport-related tweets started earlier, at 8 am, and ended around 7 pm, while sports-related tweets had a short active time between 11 am and 9 pm. Compared to tweets related to all other activity purposes, sports-related tweets had a shorter active time but had two surprising peak times, around 2 pm and 7 pm. It should be noted that the hourly frequency of tweets only reflects the general trends of activity purposes, as these check-in locations are not necessarily activity destinations.

5.2.2. Weekly dynamics

To explore the temporal rhythms of activities in a week, we first calculated the standardised hourly frequency of tweets on weekdays and weekends and then compared the weekday and weekend frequencies of the tweets. The standardised differences shown in Fig. 6 were calculated by subtracting the standardised hourly frequency of tweets on weekends from those on weekdays. A red-violet cell indicates that there were more tweets related to that specific activity sent during the weekdays than those sent during the weekends at the same time. The results show that, generally, tweets sent on weekdays had higher frequencies during the early morning (i.e., 7–8 am) than those sent on weekends. It is also evident that tweets related to commercial, recreation, residential, education and sports were more active in the evenings on weekdays, presumably during after-work or after-class time. Especially for tweets inferred as sports activities, there were two peaks identified in the last row of Fig. 6: one around 2–4 pm on weekends and the other around 7 pm on weekdays. Transport-related tweets during the weekdays were more concentrated in the mornings and evening rush hours compared to the tweets sent on the weekends. Furthermore, there was an increase of tweets inferred as commercial, recreation and education activities during the daytime (i.e., 9 am–5 pm) on weekends, which is reasonable as people usually have more free time on the weekends. It was surprising to see that work-related tweets were more active during the nighttime on weekends instead of on weekdays.

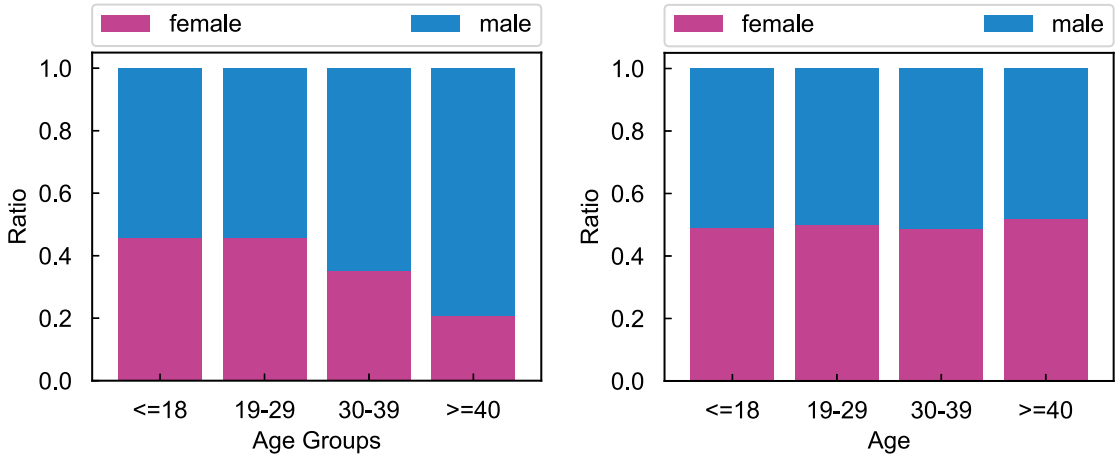
5.2.3. Spatial differences between daytime and nighttime tweets

To examine the spatial patterns of tweets from different times of day, we calculated the standardised frequencies of tweets (z-scores) during the daytime and nighttime. For each hexagon cell, the attribute value refers to the standardised difference between z-scores for daytime and nighttime, as Z_{Diff}. By applying the hotspot analysis based on the Getis–Ord Gi* statistic, we identified those statistically clustered cells with high Z_{Diff} values as hotspots and those with low Z_{Diff} values as cold spots (see Fig. 7). The colour shade indicates different levels of significance, where the darkest shades indicate that the spatial cluster of high values or low values is significant at a 99.9% confidence level. To the right of the hexagon map, the heatmap further illustrates the hourly frequencies of tweets by activity types only for hotspots or cold spots within each borough. Considering that hotspots or cold spots are normally concentrated in some areas, we only included the heatmaps for certain boroughs. Fig. 7a focuses on the hotspots (i.e., red cells) where more tweets were sent during the daytime. Daytime activities were mainly distributed in boroughs such as the City of London, Camden and Islington, but just around the city centre. Most of the daytime tweets in hotspots within the City of London were work-related. Within the Borough of Camden, hotspots of daytime activities included tweets related to commercial, work, transport and education activities, which is reasonable considering that places such as Holborn, King's Cross and University College London are included there. Hotspots in the London Borough of Islington covered areas such as Bunhill and Clerkenwell, with more tweets inferred as work, commercial, recreation and sports. Moreover, areas within Southwark and Lambeth along the River Thames had more clusters for daytime activities.

Fig. 7b focuses on the cold spots (i.e., blue cells) where there were more tweets sent during the night, revealing areas that may have more nighttime activities. A major and contiguous cluster of nighttime activities was found in the City of Westminster, including the West End, Soho, Mayfair, Head Park Estate and Trafalgar Square, in which commercial-, work- and recreation-related activities dominated because of the high-density and mixed-use commercial, leisure, retail and office spaces. The other centres for nightlife shown in the hexagon map are not obvious, except for the O2 stadium in the Borough of Greenwich, where most tweets were inferred as recreation activities. This is reasonable as the stadium hosts many matches, concerts and other events during the night.

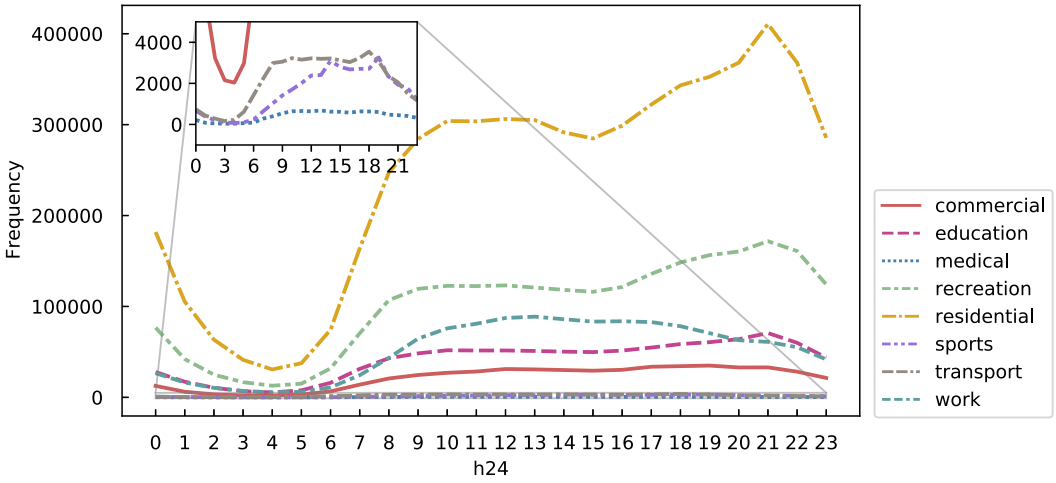
5.2.4. Spatial differences between weekdays and weekends tweets

Fig. 8 shows the spatial disparity between weekday and weekend tweets. Similarly, hotspots (i.e., red cells) denote places where more

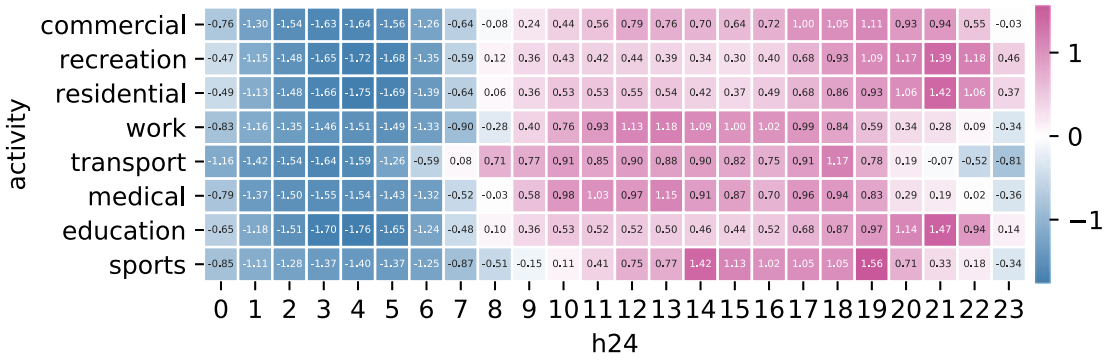


(a) Twitter users in Greater London (b) Population statistics from ONS

Fig. 4. Comparison of gender ratios in age groups between Twitter data and population estimates from ONS.



(a) Hourly trends of tweets



(b) Standardised hourly frequency of tweets

Fig. 5. Daily rhythms of geotagged tweets with inferred activity purposes. (a) Line plot of the hourly frequency of geotagged tweets by eight types of activity. (b) Heatmap of hourly frequency of geotagged tweets by eight types of activity. Standardised differences in the heatmap are within ±1.5 standard deviation.

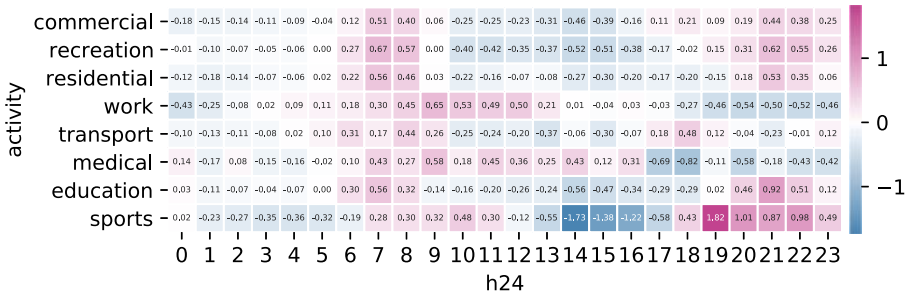


Fig. 6. Heatmap of standardised differences in temporal patterns of geotagged tweets between weekdays and weekends. The value in each cell refers to the standardised difference calculated by subtracting the standardised hourly frequency of tweets sent during weekdays $Z_{(weekdays)}$ from the number for weekends $Z_{(weekends)}$. Standardised differences in the heatmap are within ±1.5 standard deviation.

tweets were sent during weekdays than on weekends, while cold spots (i.e., blue cells) highlight places with more tweets during the weekends. Londoners sent more tweets during weekdays from central London, including most areas of the City of London, southern parts of Camden and Islington and certain areas distributed over the City of Westminster and South Bank. Specifically, tweets from hotspots in the City of London were almost all work-related, while those from Camden, Westminster, and Kensington and Chelsea were tagged as commercial-, recreation-, residential- and work-related activities. It is noted that hotspots in the City of Westminster only existed around Westminster, Warwick, Marylebone High Street, Hyde Park and Bayswater, not including West End, Soho and Covent Garden, which had more tweets during weekends than during weekdays.

Fig. 8b focuses on cold spots scattered outside of central London. Local town centres in Greater London such as Camden Town, Hampstead Town, Romford Town, Twickenham Riverside and Blackheath saw more tweets sent during the weekends than on weekdays. A similar pattern can be seen in recreation and sports facilities such as Twickenham Stadium, Wembley Stadium, Copthall Stadium, Olympic Park, Victoria Park and Tottenham Hotspur.

Stadium, which people visit more during the week.

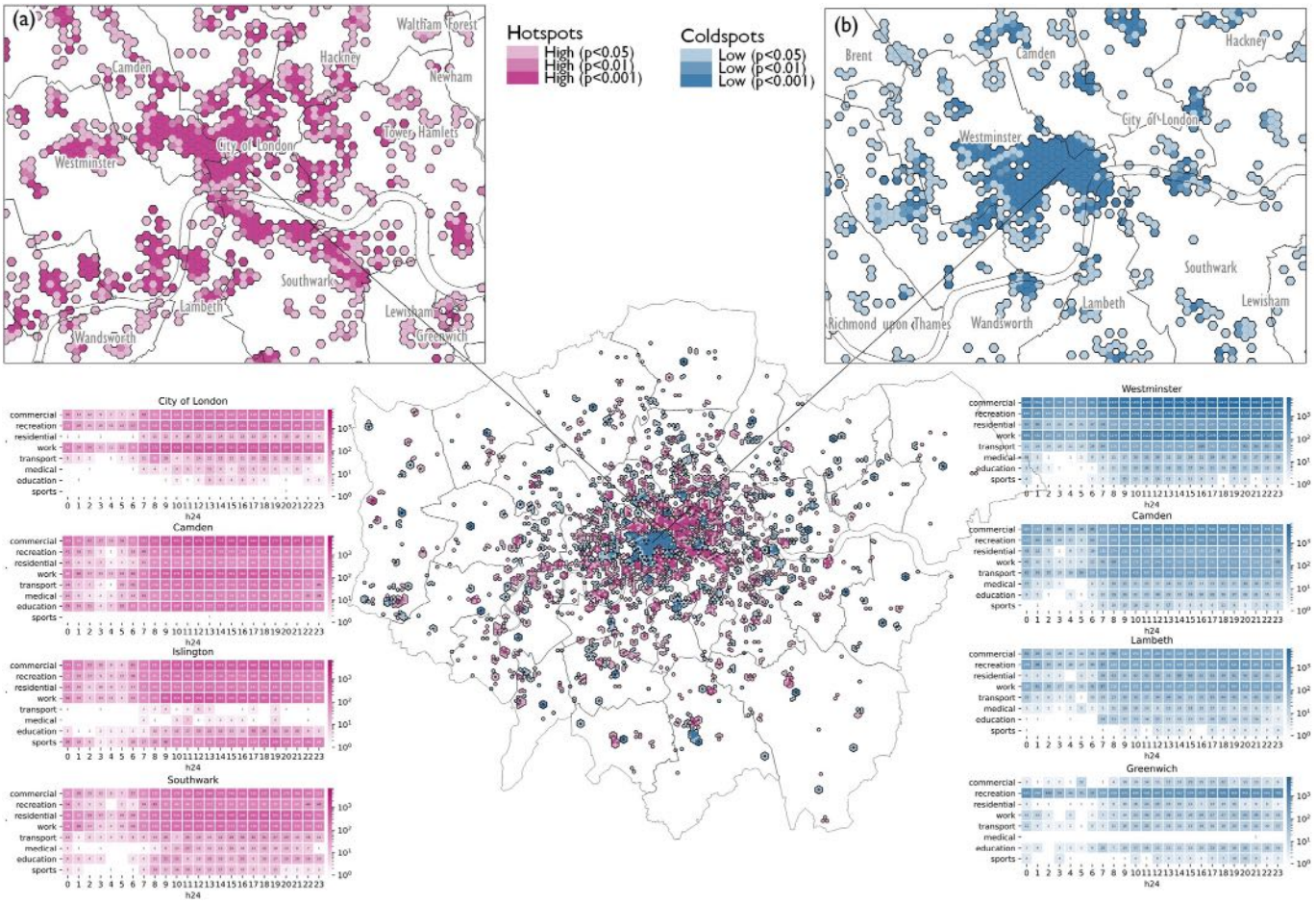


Fig. 7. Spatial clusters of standardised differences between frequencies of tweets during daytime and nighttime. (a) Map zoomed in on hotspots where red indicates there were more tweets sent during the daytime than the nighttime. (b) Map zoomed in on cold spots where blue indicates there were more tweets sent during the nighttime than the daytime. The colour shade indicates the confidence level: 99.9%, 99% and 95%, where the darkest shades indicate the highest confidence interval. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5.3. Temporal pattern of activity-based tweets across demographic groups

5.3.1. Temporal differences between gender groups

The difference in temporal patterns of activities between genders is not evident. For example, according to the rows for recreation and work activity shown in Fig. 9, the absolute values of the z-score differences are <0.2 standard deviations, indicating a similar temporal pattern between the male and female groups. Minor differences can be found during the night time after 7 pm when female users were involved in more commercial, work, transport and education activities, and also during the morning when female users were involved in more commercial, medical, education and sports activities than male users. Especially for sports activities, there was a significant pattern that tweets from female users were mainly sent during the morning and around noon.

5.3.2. Temporal differences across age groups

Fig. 10 illustrates the difference in temporal patterns of activities between tweets sent from each age group and all users. By comparing the red-violet cells across the four visualisations in Fig. 10, we see that the younger groups under 29 years old sent more tweets during the night or after midnight, specifically involving recreation, residential and education activities. People between 19 and 29 tweeted about more commercial- and work-related activities after 8 pm than users in other age groups. They also sent more sports-related tweets from 5 to 10 pm. The Fig. 10d for users over 40 years old shows that users in this group tended to send more tweets related to almost all types of activity during the daytime, although the difference was not significant. Also, there were more tweets inferred as recreation, residential and education activities in this group than others, especially during the morning from 7 to 9 am.

5.4. Spatial pattern of activity-based tweets across demographic groups

5.4.1. Spatial differences between gender groups

Within the central areas of London around the West End, there was a contiguous hotspot of commercial-, recreation- and work-related tweets sent by women covering Soho, Mayfair, Marylebone and Bloomsbury, Covent Garden and St James's, likely due to the densely concentrated high streets (see Fig. 11). In some especially popular destinations such as Buckingham Palace, Tower Bridge, London Bridge, North Greenwich, King's Cross station and Victoria Station, women were more active than men in tagging tweets. Similar patterns were also found in some residential areas, such as Notting Hill and Brook Green in Hammersmith. On the south bank of the River Thames, cold spot clusters were found in northern Lambeth, including Waterloo, Vauxhall and Oval, indicating more tweets sent from men than women in this area. Another main cold spot area was around Paddington, including Bayswater and Westbourne. Within the City of London, there was not much spatial disparity in geotagged tweets between the gender groups. Near this district, there were two cold spot areas, Old Street from the Borough of Islington and Whitechapel from the Borough of Tower Hamlets, and a hotspot area, Spitalfields and Shoreditch Streets, between the aforementioned areas.

Fig. 8. Spatial clusters of standardised differences between frequencies of tweets during weekdays and weekends. (a) Map zoomed in on hotspots where red indicates there were more tweets sent during the weekdays than on weekends. (b) Map zoomed in on cold spots where blue indicates there were more tweets sent during the weekends than on weekdays. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

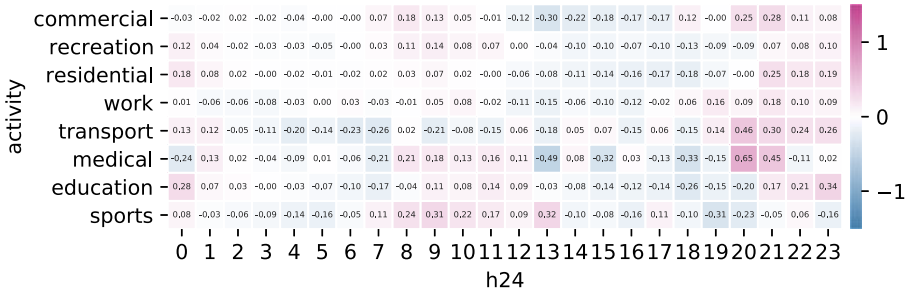
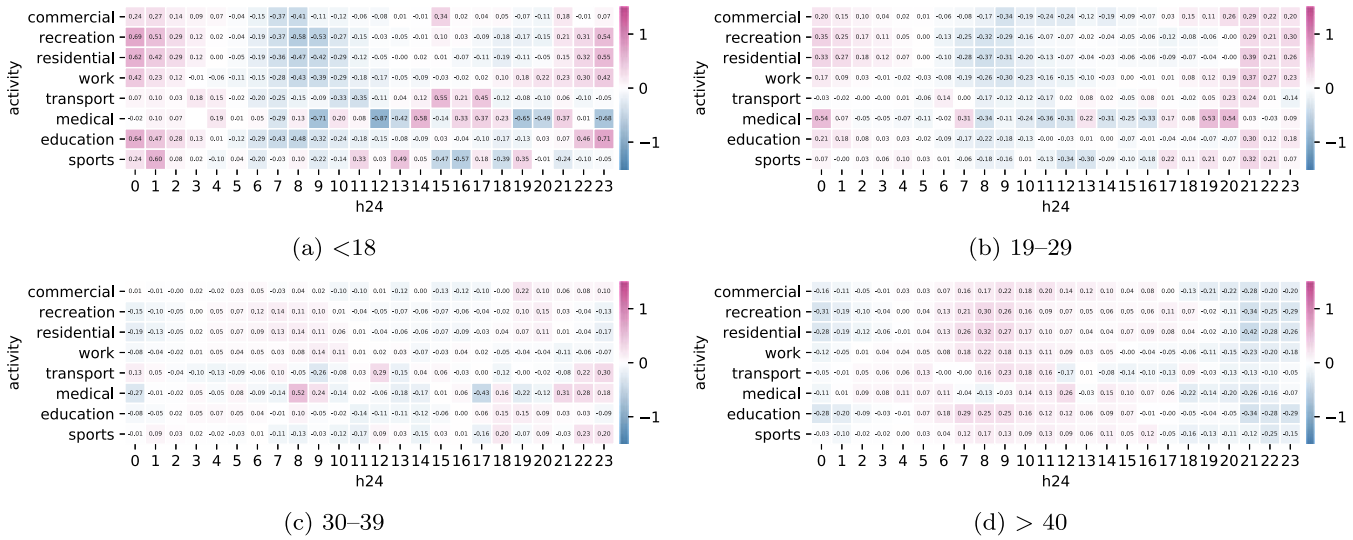


Fig. 9. Heatmap of standardised differences in temporal patterns of geotagged tweets between female and male groups. The value in each cell refers to the standardised difference calculated by subtracting the standardised hourly frequency of tweets sent by men Z(female) from the number for women Z(male). Standardised differences in the heatmap are within ±1.5 standard deviation.

5.4.2. Spatial differences between age groups

To identify the spatial pattern of different age groups among the Twitter users, we calculated the standardised differences between each age group and all users. Each subfigure in Fig. 12 shows the preferred locations tagged by each age group. According to the concentrated and contiguous hotspots, the youngest group (under 18 years old) appeared distinctly more often in Central London, especially in the areas around the River Thames. The hotspots covered places such as the City of Westminster, Marylebone, Holborn, Finsbury and the riversides of Vauxhall Bridge, Blackfriars Bridge and Tower Bridge. In contrast, the hotspots of geotagged tweets from users aged 19–29 were mostly distributed outside the Central Activities Zone (CAZ) of London. Clusters were also apparent around Hyde Park, Regent's Park, Paddington Station, Camden Town, Bethnal Green and St Katharine's & Wapping. Further from the city core, users from this group also clustered around Shepherd's Bush in the Borough of Hammersmith and South Hamstead in the Borough of Camden. Users in more senior groups in London had more geotagged tweets outside the city core (see cold spots in both Figs. 12c and 12d). For instance, the main hotspots for users aged 30–39 were located around Fulham, especially the riversides, although some small clusters appeared around the City of London. This pattern is more noticeable in the group of users over 40, who were more active in



(a) <18 (b) 19–29

(c) 30–39 (d) > 40

Fig. 10. Heatmap of standardised differences in temporal patterns of geotagged tweets between each age group and all users. The value in each cell refers to the standardised difference calculated by subtracting the standardised hourly frequency of tweets sent by all $Z_{(all)}$ from that of the age group $Z_{(subset)}$. Standardised differences in the heatmap are within ±1.5 standard deviation.

Fig. 11. Spatial clusters of standardised differences of tweets frequencies between male and female user groups. (a) Map zoomed in on hotspots where red indicates that more tweets were sent by women than men. (b) Map zoomed in on cold spots where blue indicates that more tweets were sent by men than women. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

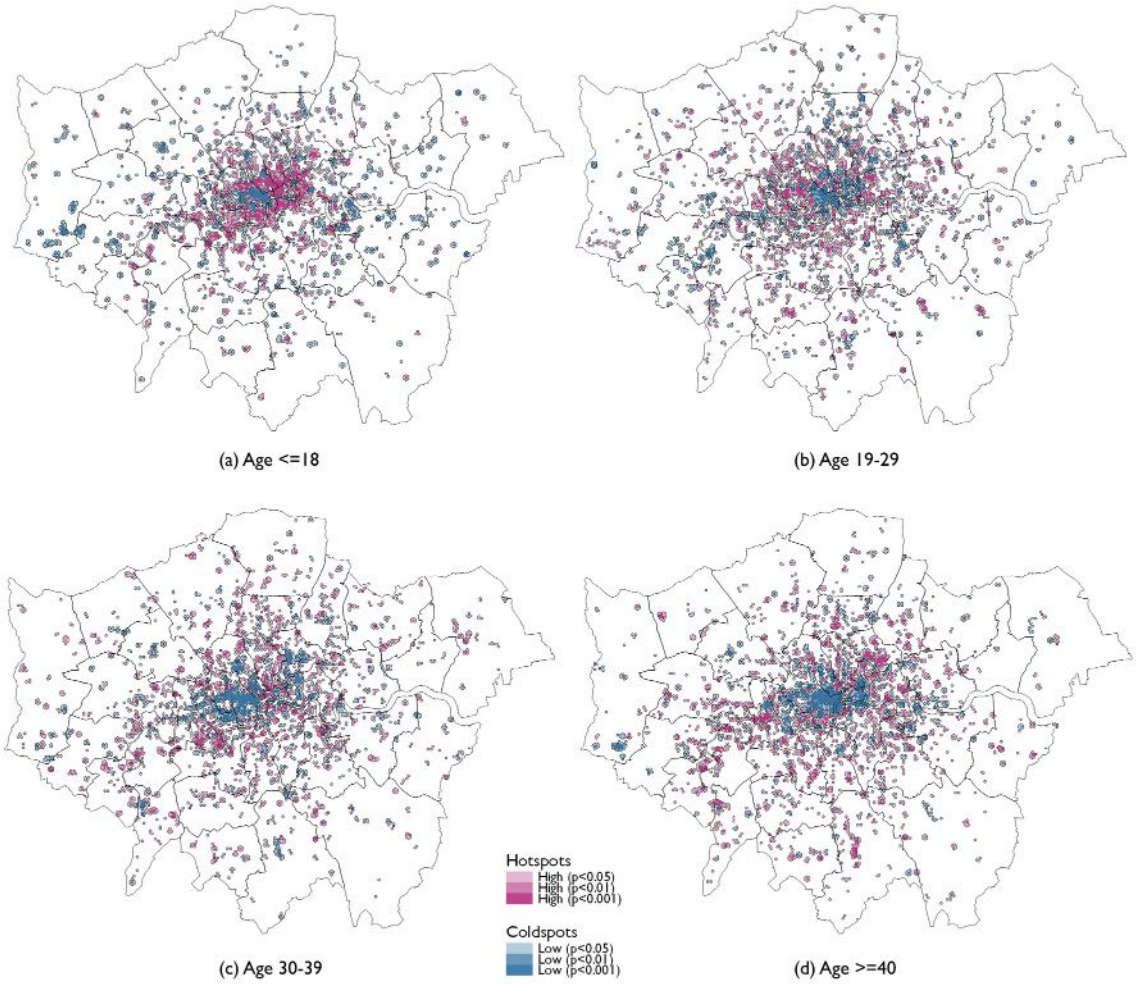


Fig. 12. Spatial clusters of standardised differences of tweets frequencies between each age group and by all users. (a) Hotspots (in red) indicates there are more tweets sent by users aged under 18 than general users. (b) Hotspots (in red) indicates there are more tweets sent by users aged between 19 and 29 than general users. (c) Hotspots (in red) indicates there are more tweets sent by users aged between 29 and 39 than general users. (d) Hotspots (in red) indicates there are more tweets sent by users aged over 40 than general users. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

southwest London along the River Thames (e.g., Hammersmith, Hounslow, Richmond upon Thames and Kingston upon Thames) and northeast London (e.g., Hackney). In outer London, hotspots for this group are also evident in Croydon, Haringey and Ealing, which differed from the younger groups. However, inside the city core, the hotspots only appear in Westminster, Chelsea and St George's Cathedral in Southwark.

6. Discussion

The result of inferring demographic characteristics of Twitter users in London confirmed the overrepresentation of male users noted in previous studies (Longley et al., 2015; Sloan, 2017) and further identified the overrepresentation of Twitter users aged between 19 and 40 in Greater London, compared with data from official population estimates. When using geotagged tweets to infer activity patterns for people in London, the patterns are highly likely to be biased toward the group of male Londoners over 30. This shows the importance of understanding the demographic attributes of users before conducting urban analysis with social media data.

The result also shows that, by spatially joining geotagged social media posts with fine-scale land use data, purposes of activities can be inferred and used for the identification of activity patterns. The heatmap of the hourly frequencies of tweets (Figs. 5 and 6) shows the daily and weekly rhythms of geotagged tweets with eight inferred activity purpose categories. The results of analysing daily tweet frequency (Fig. 5b) show that patterns of commercial, recreation, residential and education activities are similar. This analysis also revealed that there were two peak hours of sports activities in the city. The two peaks are impressively explained by the fact that the 2–4 pm peak appeared only during the weekends while the one around 7 pm appeared during weekdays. In exploring the differences in the temporal patterns of activities within demographic groups, the findings for the gender groups were limited as men and women shared similar daily activity patterns (see Fig. 9). For the age groups, there were more findings in terms of the differences in temporal patterns as shown in Fig. 10. This may suggest that the differences in temporal patterns between gender groups are less evident than those across age groups, which might require further exploration in the future.

In exploring the spatial patterns of activity-based tweets, the results shown in Fig. 7 indicate the spatial differences between daytime and nighttime activities in the central area of London. By looking into the frequency of tweets in the eight activity categories in the nighttime hotspots of the City of Westminster, it was unexpected to find many work-related activities during the night time and throughout the whole day. This appeared presumably because the geotagged tweets were accidentally inferred as work-related activities due to the high-density and mixed-use buildings in these areas that were used for process enrichment. In comparison, a difference between weekday and weekend

tweets was not evident likely because tweets sent during the weekdays mainly dominate the city core. Only certain areas such as West End, Soho and Covent Garden were also identified as clustered areas where there were more tagged activities during the weekends. This suggests that the spatial difference between daytime and nighttime activity patterns is more obvious than that of weekdays and weekends.

The spatial difference of activity-based tweets between gender groups (Fig. 11) shows that female Londoners tend to participate in commercial and recreation activities in those highdensity and mixed-use areas. Interestingly, the distribution of hotspots of women's tweets overlapped with the hotspots for nighttime activities, especially in the City of Westminster, where many commercial and recreation services are provided (see Fig. 7b). The results in Fig. 12 shows that the younger groups generate more location-based tweets within the city core, while senior groups are relatively more active in outer areas. The youngest groups generate more location-based tweets in more often in Central London that are mostly commercial and recreation-related. The hotspots for the group aged 19–29 are more scattered but mainly located within Inner London. Geotagged tweets sent from this group seem to have more commercial and work-related activities. The hotspots for 30–39 and more senior groups are mostly in town centres of outer areas. These patterns identified for each age group provide spatial-explicit evidence for age-sensitive planning and policy-making.

6.1. Contribution to the literature

This study is distinct from many other studies employing social media data in exploring activity patterns in cities. Firstly, the study deal with the unsolved non-representativeness issue of geotagged social media data, which has been extensively criticised in many previous studies (Kitchin, 2013; Li et al., 2016; Martı et al., 2019; Tufekci, 2014). To reveal the demographic composition of non-representative Twitter users, this study utilised a deep learning model with using users' metadata, i.e., profile image, user name, description and language setting, as inputs. Two gender groups and four age groups are inferred. Comparing with previous studies, it have much higher accuracy than other methods such as name analysis (Cesare et al., 2018; Longley et al., 2015; Luo et al., 2016; Mislove et al., 2011) and the home detection method (Davis et al., 2019; Huang & Wong, 2016). This provides opportunities to further explore the spatio-temporal patterns of urban activities across different age and gender groups.

By taking advantage of the demographic inference of Twitter users in London, this paper is the first attempt to explore the spatio-temproal patterns of human activity across different demographic groups (Hu et al., 2020; Huang & Wong, 2016). To better illustrate the differential patterns across age and gender groups temporally and spatially, this paper also proposes new analytical approaches. An hourly frequency heatmap of activity-based tweets was introduced to illustrate the daily and weekly rhythms. This study calculated the difference in standardised frequency of tweets in hexagon cells between different groups (e.g., daytime and nighttime, weekdays and weekends, and male and female groups) and utilised hotspot analysis to visualise the spatial disparities between them.

6.2. Implications for urban planning and policy

By inferring the age and gender from geotagged social media data, this paper reveals the demographic composition of non-representative Twitter users. The following analysis of activity patterns across demographic groups provides urban planners and policymakers with quantitative evidence related to specific groups, such as female and middle-aged people in London. For instance, Fig. 7 reveals the major and contiguous cluster of nighttime activities in the City of Westminster and identifies specific locations such as the West End, Soho, Mayfair, Head Park Estate and Trafalgar Square. This provides nearly real-time evidence to support policymaking regarding the nighttime economy with spatial interventions (GLA, 2020). Moreover, based on the finding that the spatial difference between daytime and nighttime activity patterns is more obvious than those between weekdays and weekends, the government should implement spatial interventions based on daily rhythms instead of weekly rhythms. Understanding the activity patterns for specific demographic groups, such as women, can also help support policies such as the ‘Women's Night Safety Charter’ recently issued to combat the violence against women during the night.

The study could further be used to support the ‘High Streets Strategy’ in the UK with the spatial patterns linked to the high streets in London (Ministry of Housing, Communities, & Local Government, 2021). For instance, most of the high streets in the City of Westminster are more active during the nighttime, and the high streets around Soho, Mayfair, Marylebone and Bloomsbury, Covent Garden and St James's are more active on weekends. These findings will be helpful for understanding the patterns in the high streets of different areas, which can support policymaking for the high streets in the local government.

Although the abovementioned applications of geotagged social media data for planning and policy-making are promising, it should be noted that geotagged social media data is not a silver bullet. Researchers and practitioners need to ask the flip side of the first research question – who has not contributed to the geotagged tweets and who has been missed from the result? The digital divide driven by internet access, digital skills and usage preferences hugely impacts the generation of geotagged social media data. Specific groups such as the elderly group, teenagers, people with lower education and people with limited access to the internet may not be reflected in the geotagged social media data. When applying social media data to answer urban questions related to the above groups, it is necessary to introduce other data sources, such as public surveys and census data, as complements.

6.3. Limitations and future work

When carrying out the study at the scale of a metropolitan area such as Greater London, the geotagged social media dataset collected is spatially imbalanced since people in the city core generate many more posts than people in the outer areas. When examining spatial patterns for those areas, we see small, randomly scattered clusters with low significance levels. However, for densely populated areas, geotagged social media posts provide a valid source for exploring spatial patterns, as shown in the results. Another limitation is related to the data enrichment for inferring activity purposes. As this study used a ‘spatial join’ method (i.e., linking geotagged social media with building and land use data) to infer the activity purposes related to tweets, there were a few errors generated, especially in highdensity and mixed-use areas. For example, there were a certain number of tweets in the City of Westminster inferred as work-related even for the group under 19 years old during the nighttime. In the future, this limitation can be improved by increasing the accuracy of inference for activity purposes. For instance, the activity purpose of geotagged tweets could be inferred through deep learning with multiple inputs. Furthermore, future studies can further explore individual characteristics beyond age and gender, such as race, income and social status. Exploring urban activity patterns with socioeconomic characteristics will facilitate the understanding of how citizens use urban spaces and how cities provide functions for different groups of people.

7. Conclusion

The study demonstrates the application of exploring urban activity patterns with geotagged social media data while dealing with challenges such as the non-representativeness issue (i.e., the data are not representative of the population) and the lack of activity purposes of social media data. This study utilised a deep learning model to infer the certain sociodemographic groups of users from users' profile information and then explored their activity patterns in Greater London. This study first

answered the question 'who are the data contributors?' to reveal the composition of the sample of users contributing geotagged social media data, which provided a feasible approach to deal with the non-representativeness issue of social media data. The results show that, in terms of geotagged tweets, Twitter users in London underrepresent the female population and age groups of the youngest (aged <18) and senior citizens (aged over 40). The overrepresentation of males in London Twitter users is around 15%, mainly among the 30–39 and over 40 age groups.

This paper also integrates geotagged social media data with fine-scale building and land use data in London to infer the activity purposes of social media posts. Among eight types of activity purposes, residential, recreation, and work ranked the top three. Building on this, the study explores the spatial and temporal patterns of human activity among different age and gender groups. The result shows that female Londoners tend to participate in commercial and recreational activities in those high-density and mixed-use areas. Younger groups generate more location-based tweets within the city core, while senior groups are relatively more active in outer areas. Via the lens of geotagged social media data, we are able to understand the temporal (i.e., daily and weekly) rhythms and spatial patterns of different activities across age and gender groups, which potentially provide data-driven evidence for better planning and policymaking.

Author agreement statement

We the undersigned declare that this manuscript is original, has not been published before and is not currently being considered for publication elsewhere. We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us. We understand that the Corresponding Author is the sole contact for the Editorial process. He is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs.

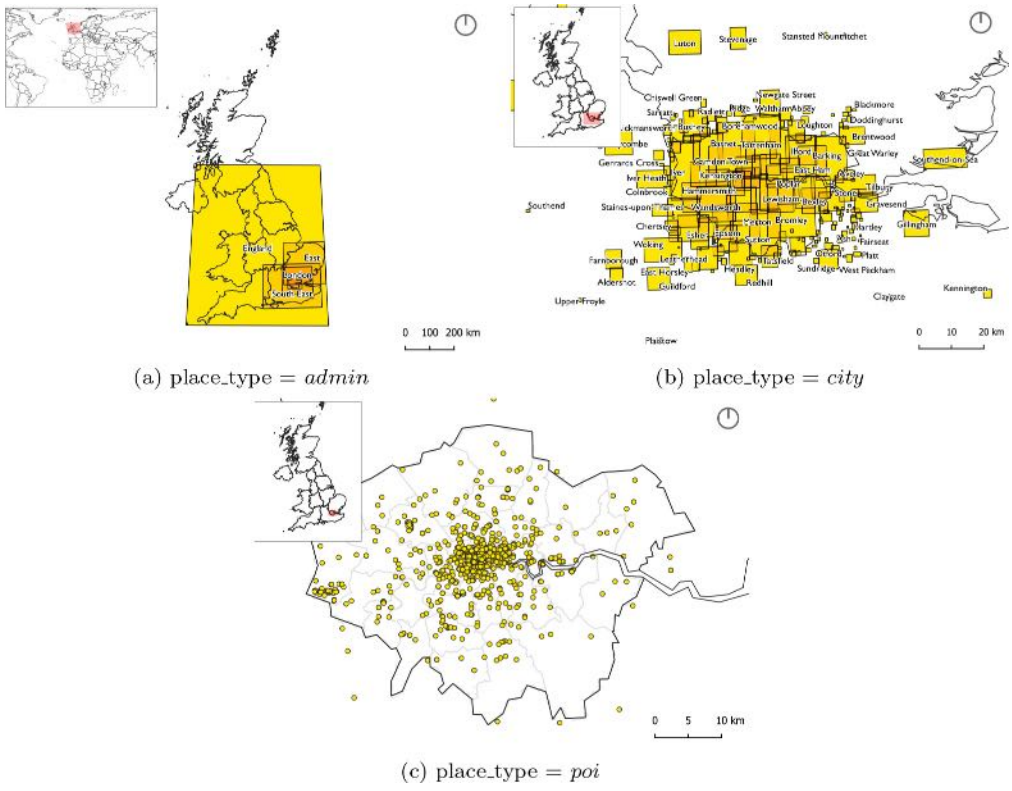
Data availability

The authors do not have permission to share data.

Acknowledgement

This research is supported by a scholarship from the China Scholarship Council (CSC No.201808060346). We thank the anonymous reviewers for their comments and suggestions.

Appendix A. Visualisation of place and coordinates geo-objects tagged in geo-tagged tweets



(a) place_type = *admin*

(b) place_type = *city*

(c) place_type = *poi*

Fig. A.1. *Place* objects tagged in tweets in Greater London.

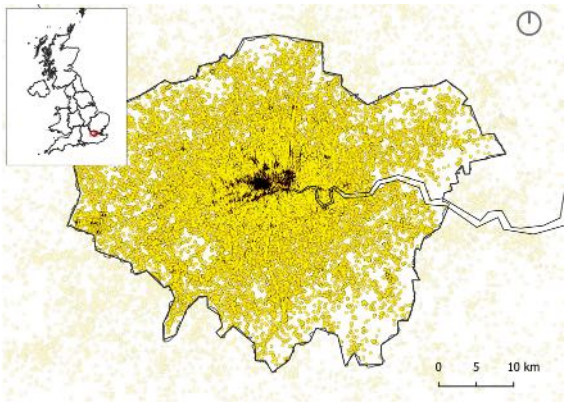


Fig. A.2. *Coordinates* objects tagged in tweets in Greater London.

Appendix B. Visualisation of land and building use data in Greater London

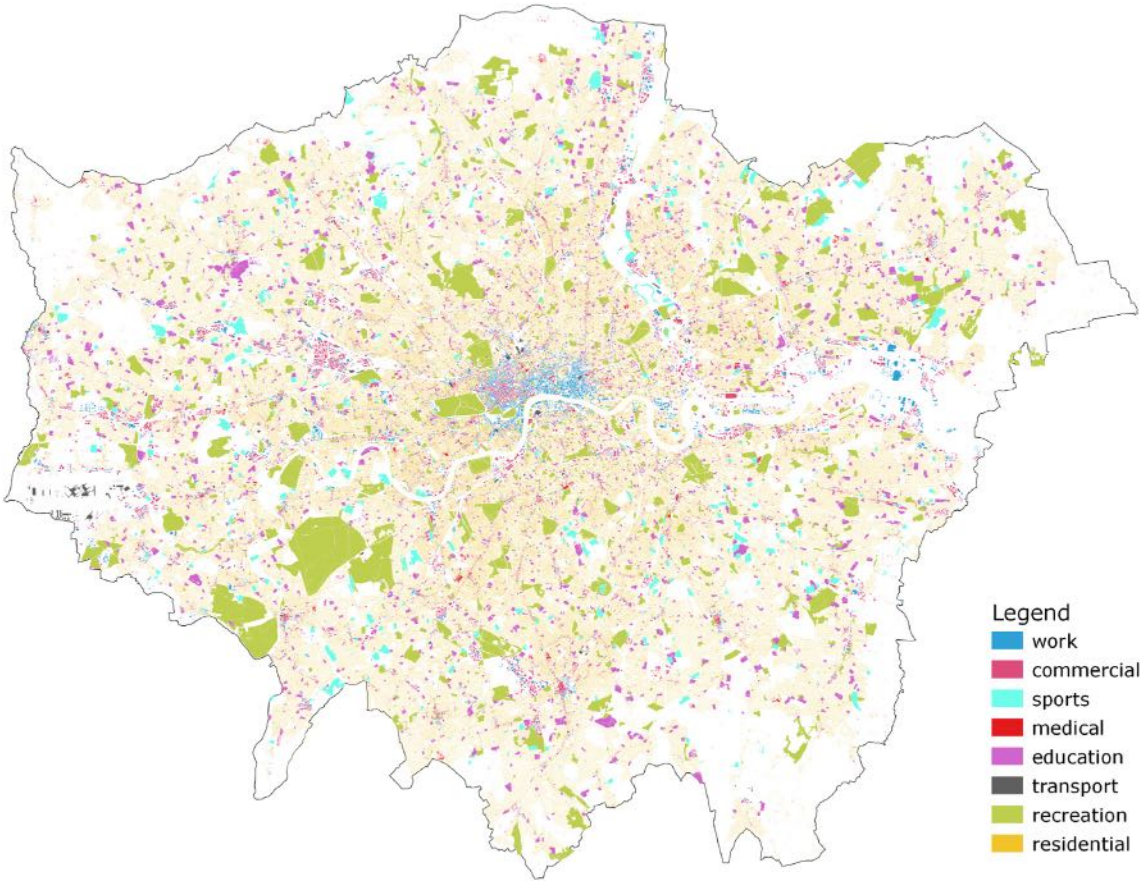


Fig. B.1. Classification of building and land use in Greater London based on activity types. Data are from Geomni UKBuildings dataset and Ordnance Survey MasterMap Greenspace layer.

References

Alowibdi, J. S., Buy, U. A., & Yu, P. (2013). Language independent gender classification on Twitter. In *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining ASONAM '13* (pp. 739–743). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2492517.2492632 00079.

An, J., & Weber, I. (2016). #greysanatomy vs. #yankees: Demographics and hashtag use on Twitter. http://arxiv.org/abs/1603.01973, 00050 arXiv: 1603.01973.

Cesare, N., Grant, C., Nguyen, Q., Lee, H., & Nsoesie, E. O. (2018). How well can machine learning predict demographics of social media users?. URL: http://arxiv.org/abs/1702.01807.

Chen, X., Wang, Y., Agichtein, E., & Wang, F. (2015). A comparative study of demographic attribute inference in Twitter. In *, 9. Proceedings of the international AAAI conference on web and social media*. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14656, 00066 Number: 1.

Cui, Y., Meng, C., He, Q., & Gao, J. (2018). Forecasting current and next trip purpose with social media data and Google Places. *Transp. Res. Part C: Emerg. Technol., 97*, 159–174. URL: https://www.sciencedirect.com/science/article/pii/S0968090X18304455 https://doi.org/10.1016/j.trc.2018.10.017.00025.

Cui, Y., Xie, X., & Liu, Y. (2018). Social media and mobility landscape: Uncovering spatial patterns of urban human mobility with multi source data. *Frontiers of Environmental Science & Engineering, 12*, 7. https://doi.org/10.1007/s11783-018-1068-1, 00010.

Davis, D. R., Dingel, J. I., Monras, J., & Morales, E. (2019). How segregated is urban consumption? *Journal of Political Economy*. https://doi.org/10.1086/701680 (pp. 000–000). 701680. 00074 tex.ids= davis2019segregated publisher: The University of Chicago Press Chicago, IL.

Fang, Q., Sang, J., Xu, C., & Hossain, M. S. (2015). Relational user attribute inference in social media. *IEEE Transactions on Multimedia, 17*, 1031–1044. https://doi.org/10.1109/TMM.2015.2430819. 00054 Conference Name: IEEE Transactions on Multimedia

GLA. (2020). Night time strategy guidance. URL: https://www.london.gov.uk//what-we-do/arts-andculture/24-hour-london/night-time-strategy-guidance.

Guimaraes, R. G., Rosa, R. L., Gaetano, D. D., Rodrıguez, D. Z., & Bressan, G. (2017). Age groups classification in social network using deep learning. *IEEE Access, 5*, 10805–10816. https://doi.org/10.1109/ACCESS. 2017.2706674. 00063 Conference Name: IEEE Access

Hasan, S., & Ukkusuri, S. V. (2014). Urban activity pattern classification using topic models from online geo-location data. *Transp. Res. Part C- Emerg. Technol., 44*, 363–381. https://doi.org/10.1016/j. trc.2014.04.003. 00074 publisher: Elsevier

Hu, L., Li, Z., & Ye, X. (2020). Delineating and modeling activity space using geotagged social media data. *Cartography and Geographic Information Science, 47*, 277–288. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85079421705&doi=10.1080%2f15230406.2019.1705187&partnerID=40&md5=8410e71cbfb5b62fb45696ed8e1bdd48 https://doi.org/10.1080/15230406.2019.1705187. 00006.

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). *Densely connected convolutional networks* (pp. 4700–4708). URL: https://openaccess.thecvf.com/content_cvpr_2017/html/Huang_Densely_Connected_Convolutional_CVPR_2017_paper.html14689.

Huang, H., Yao, X. A., Krisp, J. M., & Jiang, B. (2021). Analytics of location-based big data for smart cities: Opportunities, challenges, and future directions. *Computers, Environment and Urban Systems, 90*, Article 101712.

Huang, Q., & Wong, D. W. S. (2016). Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us? *International Journal of Geographical Information Science, 30*, 1873–1898. https://doi.org/10.1080/13658816.2016.1145225, 00105 publisher: Taylor & Francis.

Huang, W., & Li, S. (2016). Understanding human activity patterns based on space-time-semantics. *ISPRS Journal of Photogrammetry and Remote Sensing, 121*, 1–10. URL: http://www.sciencedirect.com/science/article/pii/S0924271616303203 https://doi.org/10.1016/j.isprsjprs.2016.08.008.00031.

Jiang, Y., Li, Z., & Ye, X. (2019). Understanding demographic and socioeconomic biases of geotagged Twitter users at the county level. *Cartography and Geographic Information Science, 46*, 228–242, 00000 Publisher: Taylor & Francis eprint: https://doi.org/10.1080/15230406.2018.1434834.

Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues Hum. Geogr., 3*, 262–267. https://doi.org/10.1177/2043820613513388, 00381.

Lansley, G., & Longley, P. A. (2016). The geography of twitter topics in London. *Computers, Environment and Urban Systems, 58*, 85–96. URL: http://linkinghub.elsevier.com/retrieve/pii/S0198971516300394 https://doi.org/10.1016/j.compenvurbsys.2016.04.002 (00049 publisher: Elsevier tex.ids=lansleyGeographyTwitterTopics2016b).

Lee, J. H., Davis, A. W., Yoon, S. Y., & Goulias, K. G. (2016). Activity space estimation with longitudinal observations of social media data. *Transportation, 43*, 955–977. https://doi.org/10.1007/s11116-016-9719-1. 00058

Lenormand, M., Louail, T., Cantu-Ros, O. G., Picornell, M., Herranz, R., Murillo Arias, J., … Ramasco, J. J. (2015). Influence of sociodemographics on human mobility. *Scientific Reports, 5*, 10075. https://doi.org/10.1038/srep10075. 00002. WOS: 000355391400001 number: 1 publisher: Nature Publishing Group tex.ids= lenormandInfluenceSociodemographicCharacteristics2015.

Li, L., Ban, H., Wechsler, S. P., & Xu, B. (2018). Spatial data uncertainty. In B. Huang (Ed.), *Comprehensive geographic information systems* (pp. 313–340). Oxford: Elsevier. URL: https://www.sciencedirect.com/science/article/pii/B978012409548909610X https://doi.org/10.1016/B978-0-12-409548-9.09610-X.

Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., Pettit, C., Jiang, B., Haworth, J., Stein, A., & Cheng, T. (2016). Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing, 115*, 119–133. URL: http://www.sciencedirect.com/science/article/pii/S0924271615002439 https://doi.org/10.1016/j.isprsjprs.2015.10.012 (00064 publisher: Elsevier tex.ids= liGeospatialBigData2016b).

Liu, W., & Ruths, D. (2013). What's in a name? Using first names as features for gender inference in Twitter. In *2013 AAAI spring symposium series* (p. 00204). Citeseer.

Longley, P. A., Adnan, M., & Lansley, G. (2015). The geotemporal demographics of Twitter usage. *Environment and Planning A, 47*, 465–484. https://doi.org/10.1068/a130122p, 00000 WOS:000352363600014 publisher: SAGE Publications Sage UK: London, England tex.ids= longleyGeotemporalDemographicsTwitter2015a.

Luo, F., Cao, G., Mulligan, K., & Li, X. (2016). Explore spatiotemporal and demographic characteristics of human mobility via Twitter: A case study of Chicago. *Applied Geography, 70*, 11–25. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84962526700&doi=10.1016%2fj.apgeog.2016.03.001&partnerID=40&md5=d7273f82d33dcdcdc34cd5b2ee446881 https://doi.org/10.1016/j.apgeog.2016.03.001, 00145 publisher: Elsevier.

Malik, M., Lamba, H., Nakos, C., & Pfeffer, J. (2015). *Population bias in geotagged Tweets. WS-15-18* pp. 18–27). URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0849645890 60&partnerID=40&md5=6e912871773a8a6aa15da14fec2a733dtex.ids=malikPopulationBiasGeotagged2015a.

Martı, P., Serrano-Estrada, L., & Nolasco-Cirugeda, A. (2019). Social media data: Challenges, opportunities and limitations in urban studies. *Computers, Environment and Urban Systems, 74*, 161–174. URL: http://www.sciencedirect.com/science/article/pii/S0198971518302333 https://doi.org/10.1016/j.compenvurbsys.2018.11.001 (00031 tex.ids= martiSocialMediaData2019a).

McCormick, T. H., Lee, H., Cesare, N., Shojaie, A., & Spiro, E. S. (2017). Using Twitter for demographic and social science research: Tools for data collection and processing. *Sociological Methods & Research, 46*, 390–421. https://doi.org/10.1177/0049124115605339, 00132 Publisher: SAGE publications Inc.

Ministry of Housing, Communities & Local Government. (2021). Government strategy to regenerate high streets. URL: https://www.gov.uk/government/news/government-strategy-to-regenerate-highstreets, 00000.

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. (2011). Understanding the demographics of Twitter users. In *, 5. Proceedings of the international AAAI conference on web and social media*. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14168, 00865 Number: 1.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? Comparing data from Twitter's streaming api with Twitter's firehose. In *arXiv: 1306.5204 [physics]*. http://arxiv.org/abs/1306.5204. 00582. arXiv: 1306.5204 tex.ids= morstatterSampleGoodEnough2013a.

Niu, H., & Silva, E. A. (2020). Crowdsourced data mining for urban activity: Review of data sources, applications, and methods. *Journal of Urban Planning and Development, 146*, 04020007. URL: https://ascelibrary.org/doi/full/10.1061/(ASCE)UP.1943-5444.0000566. doi:10.1061/(ASCE)UP.1943-5444.0000566. 00008 tex.ids= niu2020a publisher: American Society of Civil Engineers.

Osorio-Arjona, J., & Garcıa-Palomares, J. C. (2019). Social media and urban mobility: Using twitter to calculate home-work travel matrices. *Cities, 89*, 268–280. URL: http://www.sciencedirect.com/science/article/pii/S0264275118312976 https://doi.org/10.1016/j.cities.2019.03.006.00011.

Ouyang, J., Fan, H., Wang, L., Zhu, D., & Yang, M. (2022). Revealing urban vibrancy stability based on human activity time-series. *Sustainable Cities and Society, 85*, Article 104053.

Park, N. (2020). *Population estimates for the UK, England and Wales, Scotland and Northern Ireland, provisional: mid-2019*. Hampshire: Office for National Statistics.

Rashidi, T. H., Abbasi, A., Maghrebi, M., Hasan, S., & Waller, T. S. (2017). Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. *Transp. Res. Part C: Emerg. Technol., 75*, 197–211. URL: http://www.sciencedirect.com/science/article/pii/S0968090X16302625 https://doi.org/10.1016/j.trc.2016.12.008 (00154 tex.ids= rashidiExploringCapacitySocial2017a publisher: Elsevier).

Shelton, T., Poorthuis, A., & Zook, M. (2015). Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and Urban Planning, 142*, 198–211. https://doi.org/10.1016/j.landurbplan.2015.02.020, 00289 publisher: Elsevier.

Sloan, L. (2017). Who tweets in the United Kingdom? Profiling the twitter population using the British social attitudes survey 2015. *Soc. Media + Soc., 3*. https://doi.org/10.1177/2056305117698981, 2056305117698981. 00038 Publisher: SAGE Publications Ltd.

Song, C., Koren, T., Wang, P., & Barabasi, A.-L. (2010). Modelling the scaling properties of human mobility. *Nature Physics, 6*, 818–823. URL: https://www.nature.com/articles/nphys1760 https://doi.org/10.1038/nphys1760.00993. Number: 10 Publisher: Nature Publishing Group.

Swier, N., Komarniczky, B., & Clapperton, B. (2015). *Using geolocated twitter traces to infer residence and mobility. 00017*. Technical Report Office for National Statistics.

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. arXiv:1403.7400 [physics], URL: http://arxiv.org/abs/1403.7400, 00587 arXiv: 1403.7400.

Wang, Z., Hale, S., Adelani, D. I., Grabowicz, P., Hartman, T., Flock, F., & Jurgens, D. (2019). Demographic inference and representative population estimates from multilingual social media data. In *The world wide web conference WWW '19* (pp. 2056–2067). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3308558.3313684, 00035.

Wu, C., Ye, X., Ren, F., & Du, Q. (2018). Check-in behaviour and spatio-temporal vibrancy: An exploratory analysis in Shenzhen, China. *Cities, 77*, 104–116. https://doi.org/10.1016/j.cities.2018.01.017, 00019 WOS:000433269800013 tex.ids= wu2018check publisher: Elsevier.

Wu, L., Zhi, Y., Sui, Z., & Liu, Y. (2014). Intra-urban human mobility and activity transition: Dvidence from social media check-in data. *PLoS One, 9*. https://doi.org/10.1371/journal.pone.0097010. e97010. 00077 tex.Ids= wu2014intra publisher: Public library of science San Francisco, USA.

Xu, Z., Liu, Y., Yen, N. Y., Mei, L., Luo, X., Wei, X., & Hu, C. (2020). Crowdsourcing based description of urban emergency events using social media big data. *IEEE Trans. Cloud Comput., 8*, 387–397. https://doi.org/10.1109/TCC.2016.2517638, 00168 Conference Name: IEEE Transactions on Cloud Computing.

Xu, Z., Zhang, H., Hu, C., Mei, L., Xuan, J., Choo, K.-K., Sugumaran, V., & Zhu, Y. (2016). Building knowledge base of urban emergency events based on crowdsourcing of social media. *Concurr. Comput., 28*, 4038–4052. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84957812535&doi=10.1002%2fcpe.3780&partnerID=40&md5=ab60a9098b95d5c78ca2b935787539f5 https://doi.org/10.1002/cpe.3780.00061.

Yuan, J., Zheng, Y., & Xie, X. (2012). Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 186–194). New York, United States: ACM. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84

866045445&doi=10.1145%2f2339530.2339561&partnerID=40&md5=10aedac56fa811e1fce011223d0b4cc6 https://doi.org/10.1145/2339530.2339561, 00611.

Yuan, Y., Wei, G., & Lu, Y. (2018). Evaluating gender representativeness of location-based social media: A case study of Weibo. *Annals of GIS, 24*, 163–176, 00013 Publisher: Taylor & Francis eprint: https://doi.org/10.1080/19475683.2018.1471518.

Zagheni, E., Garimella, V. R. K., & Weber, I. (2014). Inferring international and internal migration patterns from twitter data. In *Proceedings of the 23rd international conference on world wide web* (pp. 439–444). ACM, 00163.

Zhong, Y., Yuan, N. J., Zhong, W., Zhang, F., & Xie, X. (2015). You are where you go: Inferring demographic attributes from location check-ins. In *Proceedings of the Eighth ACM international conference on web search and data mining WSDM '15* (pp. 295–304). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2684822.2685287 00153.