From individual to collective behaviours: exploring population heterogeneity of human mobility based on social media data

Liao et al. *EPJ Data Science* (2019) 8:34
https://doi.org/10.1140/epjds/s13688-019-0212-x

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

From individual to collective behaviours: exploring population heterogeneity of human mobility based on social media data

Yuan Liao[1*], Sonia Yeh[1] and Gustavo S. Jeuken[2]

*Correspondence:
yuan.liao@chalmers.se
[1]Department of Space, Earth and Environment, Division of Physical Resource Theory, Chalmers University of Technology, Gothenburg, Sweden
Full list of author information is available at the end of the article

Abstract

This paper examines the population heterogeneity of travel behaviours from a combined perspective of individual actors and collective behaviours. We use a social media dataset of 652,945 geotagged tweets generated by 2,933 Swedish Twitter users covering an average time span of 3.6 years. No explicit geographical boundaries, such as national borders or administrative boundaries, are applied to the data. We use spatial features, such as geographical characteristics and network properties, and apply a clustering technique to reveal the heterogeneity of geotagged activity patterns. We find four distinct groups of travellers: local explorers (78.0%), local returners (14.4%), global explorers (7.3%), and global returners (0.3%). These groups exhibit distinct mobility characteristics, such as trip distance, diffusion process, percentage of domestic trips, visiting frequency of the most-visited locations, and total number of geotagged locations. Geotagged social media data are gradually being incorporated into travel behaviour studies as user-contributed data sources. While such data have many advantages, including easy access and the flexibility to capture movements across multiple scales (individual, city, country, and globe), more attention is still needed on data validation and identifying potential biases associated with these data. We validate against the data from a household travel survey and find that despite good agreement of trip distances (one-day and long-distance trips), we also find some differences in home location and the frequency of international trips, possibly due to population bias and behaviour distortion in Twitter data. Future work includes identifying and removing additional biases so that results from geotagged activity patterns may be generalised to human mobility patterns. This study explores the heterogeneity of behavioural groups and their spatial mobility including travel and day-to-day displacement. The findings of this paper could be relevant for disease prediction, transport modelling, and the broader social sciences.

Keywords: Geotagged activity patterns; Individual mobility; Data mining; Hierarchical clustering

1 Introduction

Understanding travel behaviour can provide insights for a wide range of disciplines, including urban planning [1], transport management [2], epidemiology [3], ecology, and social science [4]. Previous travel behaviour research has used cross-sectional data [5], such as from household travel surveys. Although it is one of the most prevalent data

sources, surveys are costly to collect and therefore typically suffer from small sampling rates, short survey duration, under-reporting of long-distance trips [6], and long lag times between data collection and data availability [7]. Despite some drawbacks, travel surveys contain socio-demographic information and detailed activity records that make them difficult to replace by emerging data sources [6]. Those characteristics enable researchers to examine population-level mobility determinants and large-scale changes in daily mobility. Traditional travel surveys also contain rich explanatory variables that enable the validation/calibration that is essential for utilising emerging data sources.

The rapid development of information and communication technology (ICT) has the potential to address some of the shortcomings mentioned above and broaden the types of questions that can be explored in travel behaviour studies [8]. Emerging data sources, such as records from Global Positioning System (GPS) devices, smart cards, mobile phones, and other online systems, have deepened the understanding of human mobility [9, 10]. Among the emerging data sources, social media data are being gradually accepted as user-contributed data sources in travel behaviour studies, such as activity pattern classification [11], large-scale urban activity [12], and mobility patterns [13].

Geotagged tweets from the Twitter platform represent one type of social media data. A tweet is a short social media text message associated with a unique user on the Twitter platform, and a geotagged tweet also contains the GPS coordinates if the user allows this information to be attached to the tweet. The number of geotagged tweets is low compared to the total number of tweets, with one study finding around 1-3% in Syria [14]. Similarly in our previous study, we also found that geotagged tweets accounted for a limited proportion of overall Twitter users, e.g., 7.4% (George, South Africa), 1.9% (Barcelona, Spain), 1.1% (Kuwait), and 0.3% (Sweden) [15]. The number of geotagged tweets per user also varies among countries. Median (and the 5%th - 95%th percentile in parenthesis) values over a six-month sampling period are 9 (1-190) (Kuwait), 2 (1-50) (Australia), 2 (1-41) (Sweden), and 2 (1-20) (Barcelona, Spain) etc [15]. Despite that, geotagged tweets have proved a useful proxy for tracking and predicting human movement [10]. Such a data source provides precise location information [10], easy and free access [16], and opportunities for continuous tracking activities without a predefined geographic boundaries such as national borders or administrative boundaries [17]. The main criticisms are biased population representation [18] and behaviour distortion [19, 20] regarding when and where locations are reported via geotagged tweets. Some studies have compared multiple data sources to identify or adjust the biases [20, 21] and to validate against "ground truth" [22]. Despite some disadvantages of geotagged tweets, one recent review highlights the usefulness of such data sources for modelling travel behaviour [16] and understanding social behaviors such as urban neighbourhood isolation [23].

1.1 Related work

Geotagged tweets can be obtained by purchasing the complete set of public tweets from Twitter Firehose, using the Streaming API for up to a maximum of 1% of public tweets, or retrieving user timelines by user name/ID for up to 3200 historical tweets that are set publicly accessible by the user [24]. Geotagged tweets are often limited to a geographical bounding box such as national borders or adminstrative boundaries when collected from the Streaming API, yielding a lateral dataset that covers a large number of Twitter users for a snapshot of time. If the movement of a user occurs across or outside the bounding

Table 1 Representative studies of travel behaviour using social media data. T is the time span covered by the dataset. S represents data source. On the Angle column, A indicates aggregate/lateral level and I indicates individual/longitudinal level. On the S column, F indicates Foursquare and T indicates Twitter

Study	Topic	Angle	User	Sample	Geo-scale	T	S
[35]	Travel demand	A	–	19,710	City	21 days	F
[36]	Travel demand	A	54,272	355,059	City	2 days	T
[10]	Representativeness	A	156,607	7,811,004	Country	8 months	T
[26]	City influence	A	571,893	21,017,892	Cities	1000 days	T
[15]	Travel distance	A	791,542	15,719,535	Cities	6 months	T
[32]	User routines	I	825	157,806	City	1 year	F
[11]	Urban activity	I	3256	504,000	City	1 month	T
[37]	Activity space	I	116	63,114	County	5 months	T
[8]	Travel monitoring	I	9738	6,000,000	Districts	1 year	T
This study	Travel patterns	I + A	2926	652,945	Globe	3.6 years	T

box, it is not captured with this method. Geotagged tweets collected from user timelines do not have this geographical boundary limitation, and the historical tweets of a specified user can be collected in a few seconds. These tweets can cover multiple years, creating a longitudinal record of an individual's locations without any geographical boundaries [24]. Non-recurrent mobility that is often under-reported in a one-day travel diary (e.g., tourists' mobility [25]) can be studied using this type of data. It is also feasible to scale up the number of Twitter users to study the influence of global cities [26].

Geotagged social media data have been criticised for non-representativeness due to population bias and behaviour distortion. It is found that Twitter users in the U.S. overrepresent dense population regions and are predominantly male [27]. Behaviour distortion involves both tweeting behaviour and reasons for geotagging, both of which can lead to non-representativeness. The time sparsity of tweets causes the trajectory of geotagged tweets collected from users to be incomplete compared with the actual mobility trajectory of those users. One recent study shows that people geotag consciously and intentionally in uncommon places, and they often geotag soon after arriving at the place [20]. These biases need to be considered before drawing any conclusions from these data sources.

Social media data have been used to study both aggregate mobility behaviour and individual-based activity behaviour [16]. Representative studies are summarised in Table 1. At the aggregate level, studies have shown that social media data can be a reasonable proxy for population mobility. Studies have used social media data to demonstrate the truncated power law of trip distance distribution [28] and Zipf's law of the visitation frequency, which describes people's tendency to return to a couple of locations they frequently visit [29]. Studies generally found good agreements cross validating social media data against other data of higher time resolution, such as mobile phone call detail record (CDR) [22]. At the individual level, studies have used geotagged social media data to infer activity purpose. Such studies usually have specified application, e.g., bike sharing behaviour [30], prediction of next location [31], and lifestyle behaviour [32, 33]. Social media data have been used to identify activity choice patterns [11, 32] and to recognize specific activities [34], combining spatial and temporal information with semantic information in the data.

Previous travel behaviour studies using social media data have been limited by data collection and a lack of understanding of population heterogeneity. Most studies are based on data collected within a specified geographical bounding box over a short time range

(Table 1). Use of a geographical bounding box, such as national borders or administrative boundaries, precludes capture of trips outside or across the boundary of the box, biasing the resulting data toward short-distance travel. On population heterogeneity, most studies of aggregate population behaviours neglect individual differences, while studies of individual mobility usually neglect common features that drive similar behaviours across groups of individuals. Although the travel behaviours of individuals in any population are neither identical to nor independent of each other's, there has been little work on combining aggregate and individual perspectives to gain new insights about travel behaviours of a heterogeneous population. Understanding travel patterns across scales from individuals to a population is the next step in understanding urban mobility and social behaviour [4], especially when comparing mobility in different cities [38].

This paper reveals the population heterogeneity of geotagged activity patterns using a long-term dataset without any geographical boundaries, such as national borders or administrative boundaries. Specifically, this study attempts to answer the following three questions:

- Are there any distinct patterns that characterise the observed individual geotagged activities?
- What are the spatial and temporal characteristics derived from different geotagged activity patterns?
- Can geotagged tweets be used as a proxy to approximate the mobility patterns of different behavioural groups?

The dataset includes 652,945 geotagged tweets generated by 2,933 Swedish Twitter users covering time spans of more than one year (3.6 years on average). We first describe the geotagged tweets dataset and validate it against a household travel survey. To identify the population heterogeneity of geotagged activity patterns, we combine aggregate and individual analysis techniques: we first analyse the geotagged trajectories of each user to classify them regarding their activity patterns, and then we conduct an aggregate analysis for each group. We characterise the features of individual trajectories of geotagged tweets using both geographical and network properties. The features describing users' activity patterns are based on those found in the literature. Hierarchical clustering is a descriptive data mining method that can produce new, non-trivial classifications of users based on the available dataset [39]. Patterns of individual trajectories are thus grouped into four categories: local returners, global returners, local explorers, and global explorers. The spatial and temporal characteristics of these four groups of individuals are explored.

2 Methods

We use two datasets in this study: geotagged tweets from Twitter and individual trip information from the Swedish National Travel Survey. We use the household travel survey data to investigate the representativeness of geotagged tweets via a descriptive analysis, comparing spatio-temporal characteristics (behaviour distortion) and the population distribution (population biases).

The rest of this section introduces the methodology that identifies the population heterogeneity of human mobility, as shown in Fig. 1. Six spatial features are proposed to describe the individual geotagged activity patterns in the feature construction. Based on the geotagged activity trajectories, the features are calculated per user, and hierarchical clustering is applied. We identify four groups of users with distinct geotagged activity patterns.

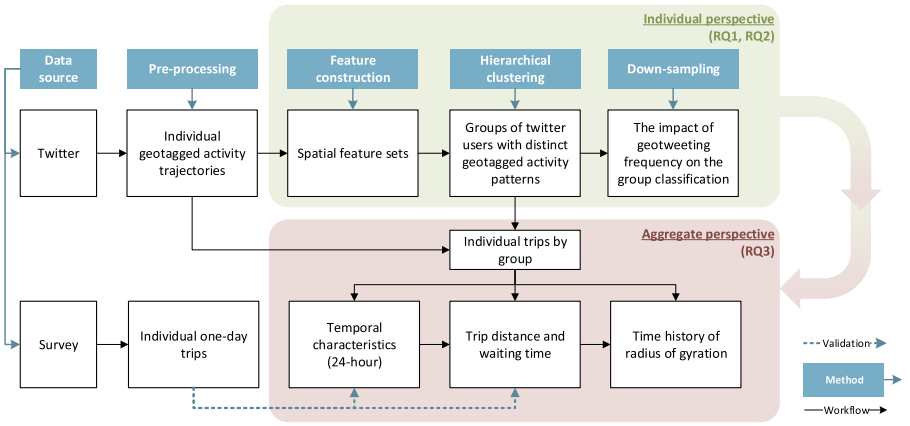

Figure 1 Methods structure. RQ1—Are there any distinct patterns that characterise the observed individual geotagged activities? RQ2—What are the spatial and temporal characteristics derived from different geotagged activity patterns? RQ3—Can geotagged tweets be used as a proxy to approximate the mobility patterns of different behavioural groups?

We further apply down-sampling to test the impact of geotweeting frequency on the group identification.

2.1 Data collection and pre-processing

2.1.1 Twitter data

In a previous study, we used the Gnip database to identify 5000 non-commercial Twitter users who geotagged their tweets most frequently during a six-month period (20 December 2015–20 June 2016) within the geographical bounding box of Sweden [15]. Gnip is a Twitter subsidiary which sells historical tweets in bulk and provides access to the Firehose API. We extract these top users' historical tweets (without applying a spatial boundary) from their user timelines [40]. The data are limited to 3200 tweets per user. This method produces a varied time span and varied tweet number, since not all users reached the 3200-tweet maximum. Because the tweeting frequency varies among users, the time span collected per user also varies: the higher the tweeting frequency, the shorter the time span collected from a user.

We further apply the following rules to pre-process the data to ensure that the individuals included in the study live in Sweden and have a substantial number of geotagged tweets so we can reasonably capture their activity trajectories: (1) the covered time span is above 1 year, (2) the geotweeting frequency (geotagged tweets/day) is above 0.1, or the total amount of geotagged tweets is above 50, and (3) the most frequently visited locations is in Sweden. After screening, we identify 2926 users and 652,945 geotagged tweets.

Using Twitter data, a "trip" is defined as the trajectory between two consecutive geotagged tweets generated by the same user. A trip in this study is equivalent to displacement in some previous studies. Waiting time is defined here as the time interval between two consecutive actions (geotagged tweets in this context) by the same individual [41]. A trip should also have a distance larger than 10 m given the precision of GPS coordinates generated by Twitter.

2.1.2 Swedish travel survey

The survey data come from the Swedish National Travel Survey for the years of 2011–2014 [42]. The survey data are used to compare the trip length distribution with those derived from geotagged tweets. It consists of a total of 31,457 travel diaries spanning the period of a day, with detailed information on individual trip distance, travel time, mode of transportation, and trip purpose [15]. The travel survey also includes a separate dataset containing a total of 9024 trips during 60 days from the same group of participants as in the travel diary data. These include trips that are either longer than 100 km or to neighbouring countries with a distance shorter than 100 km. To be consistent when comparing Twitter data with the survey data, Twitter data are filtered to either only include domestic trips beginning and ending on the same day, with distances longer than one kilometre (minimum distance in the survey), or international trips.

2.2 Geotagged activity pattern: feature construction

The locations that user i visited are first captured using all geotagged tweets by user i with the time stamps: $(X, Y, t)_{i,k}$, $k = 1, 2, \ldots, N_i$ where X is the decimal degree of Latitude, Y is the decimal degree of Longitude, t the time stamp (UTC) of the kth location. We define dom as the indicator to show whether the location is within Sweden: dom = 0 is outside Sweden and dom = 1 is in Sweden. N_i is the total number of locations visited by the user i through his/her geotagged tweets, and T_i is the total captured time span of user i. With t_l as the local time of the tweets, we further calculate the month variable $m \in [1, 12]$, the weekday variable w (weekday = 1 and weekend = 0), and the hour of the day, $h \in [1, 24]$. The time sequence of user's locations (user trajectory) is therefore:

$$S_i = (X, Y, t, t_l, m, w, h, \text{dom})_{i,k}, \quad k = 1, 2, \ldots, N_i. \tag{1}$$

For user i, the number of distinct locations is smaller than or equal to the total number of locations user i visited. Let n_i be the number of distinct locations, $f_{i,j}$ be the visiting frequency of location j, and $T_{i,j}$ be the time interval of two visits of the location j. The vector of visited distinct locations is therefore:

$$\mathbf{L}_i = (X, Y, f, \mathbf{T}, \mathbf{m}, \mathbf{w}, \mathbf{h})_{i,j}, \quad j = 1, 2, \ldots, n_i. \tag{2}$$

A trip, the connection between two consecutive geotagged tweets generated by the same user, is represented by the arc connecting two consecutive geotagged tweets with locations $j-1$ and j. (If $j-1$ and j are within 10 metres of each other or the tweets are within 10 minutes, these are considered to be the same location and not a distinct trip.) The arc connecting these locations has a Haversine distance (distance along the curved surface of the earth), $d > 10$ m, and time interval $\triangle T > 10$ min between the tweets. For each trip, if location $j-1$ and j are located within Sweden, that trip is defined as a domestic trip, dom = 1, and if location $j-1$ and j are located outside Sweden, dom = 2, otherwise dom = 0. The origin-destination matrix that is based on the trajectory of geotagged tweets of user i ($\mathbf{ODM}_i$) is a directed graph with the trip attributes shown below.

$$\mathbf{ODM}_i = (f, d, \text{dom})_{p,q}, \quad p, q = 1, 2, \ldots, n_i. \tag{3}$$

Based on the literature review, we propose two essential aspects: how far one travels and how actively one explores new locations. To do pattern mining, we need to find proper summary statistics as the features to characterise the geotagged activity patterns. Therefore, we first examine the underlying distribution of the trip distance $(d_{i,j}, j = 1, 2, \ldots, n_i)$ and the network node degree $(f_{i,j}, j = 1, 2, \ldots, n_i)$ for all the individuals' geotagged activity trajectories. Specifically, we compare the theoretical distribution that best fits the empirical distribution to see whether the empirical distribution is heavy-tailed [43]. It turns out most users' trip distance and network node degree follow a heavy-tailed distribution, such as the distribution of Cauchy, Lévy, Burr, and Pareto. To deal with the highly skewed data, log transformation is applied to variables $f_{i,j}, d_{i,j}, j = 1, 2, \ldots, n_i$ to calculate the log-mean and the log-variance. The summary statistics below are proposed to quantify the key characteristics of Twitter users' geotagged activity patterns.

Six features of geographical characteristics and network properties are proposed to represent an individual geotagged trajectory. Geographical characteristics are described by features r_g, D_o, and d. Radius of gyration, r_g (km), refers to the travel distance range weighted by the visiting frequency. The total radius of gyration r_g is defined as:

$$r_g = \sqrt{\frac{1}{n_i} \sum_{q=1}^{n_i} f_q \cdot (\mathbf{r}_q - \mathbf{r}_{\mathrm{cm}})^2}, \tag{4}$$

where $\mathbf{r}_q = [X1, X2]_q$ and the mass center of the visited locations:

$$\mathbf{r}_{\mathrm{cm}} = \left[\frac{\sum_{q=1}^{n_i} (X_q \cdot f_q)}{\sum_{q=1}^{n_i} X_q}, \frac{\sum_{q=1}^{n_i} (Y_q \cdot f_q)}{\sum_{q=1}^{n_i} Y_q} \right]. \tag{5}$$

Location distance variance, D_o (km), refers to the geographical dispersion degree of visited locations. $\mathbf{ODM}_d = (d_{p,q})$ represents the linear-scale trip distance matrix where $d_{p,q} = d_{q,p}$, $d_{p,q} = 0, p = q$. The log-transformed trip distance matrix is indicated by $\mathbf{ODM}'_d = (\log(d_{p,q}))$. The only zero elements of the linear-scale $\mathbf{ODM}_d$ entries are on the diagonal for which, instead of using additive smoothing, we retain them on the diagonal of the log-transformed $\mathbf{ODM}'_d$. $\mathbf{ODM}_d^{\mathrm{norm}} = d_{p,q}^{\mathrm{norm}}$ is defined as:

$$\mathbf{ODM}_d^{\mathrm{norm}} = \mathbf{ODM}'_d - \left(\sum_{p=1}^{n_i} \sum_{q=1}^{n_i} \log(d_{p,q}) \cdot f_{p,q} \right) * \frac{\mathbf{J}_{n_i}}{n_i^2}, \tag{6}$$

where J_{n_i} the unit matrix. So the location distance variance D_o is defined by:

$$D_o = \sqrt{\frac{\sum_{p=1}^{n_i} \sum_{q=1}^{n_i} d_{p,q}^{\mathrm{norm}}}{n_i^2}}. \tag{7}$$

Mean value of log-transformed trip distance, d (km), refers to the average log-transformed distance between two consecutive geotagged tweets, defined as:

$$d = \frac{\sum_{k=2}^{N_i} \log(d_{k-1,k})}{N_i - 1}. \tag{8}$$

Table 2 Spatial features characterising individual's activity patterns. The node degree ([a]) is equivalent to the location visiting frequency

Type	Feature		Description
Geographical characteristics	r_g (km)	Radius of gyration	Travel distance range weighted by the visiting frequency.
	D_o (km)	Location distance variance	Geographical dispersion degree of visited locations.
	d (km)	Mean trip distance	Average distance between two consecutive geotagged tweets.
Network properties	$\overline{C}$ (–)	Clustering coefficient	To which degree the visited locations are connected together.
	z (–)	Mean node degree [a]	Overall visiting frequency.
	z_m (–)	Max node degree divided by the sum of total degrees	Degree of how centralised the overall visited locations are by visiting frequency.

Network properties are described by feature clustering coefficient ($\overline{C}$), average node degree (z), and normalised node degree (z_m). Clustering coefficient (average), $\overline{C}$ (–), refers to the degree to which the neighbours of a given node link to each other [44, p. 63]. For a node (location) j with degree (visiting frequency) $f_{i,j}$, its local clustering coefficient is defined as:

$$C_j = \frac{2L_j}{f_j(f_j - 1)}, \tag{9}$$

where L_j indicates the number of links between the k_j neighbours of node j. The average clustering coefficient of the whole network is calculated by:

$$\overline{C} = \frac{1}{n_i} \sum_{j=1}^{n_i} C_j. \tag{10}$$

The mean value of the log-transformed node degree, z (–), represents the overall visiting frequency. Each visited location is seen as one node in the network, and the visiting frequency is equivalent to the node degree; therefore, the average value of the node degree z is one important indicator of the network properties. It is defined as:

$$z = \frac{\sum_{j=1}^{n_i} \log(f_j)}{n_i}. \tag{11}$$

z_m (–) is the max node degree divided by the sum of total degrees, which indicates the how centralised the overall visited locations are. The normalised max node degree z_m is defined as:

$$z_m = \frac{\max[f_j]}{\sum_{j=1}^{n_i} f_j}. \tag{12}$$

The proposed features are summarised in Table 2. The geographical features reflect how far people travel and all have the same units, km. The radius of gyration, which combines the locations' geographical distribution and their visiting frequency, has been widely applied to characterise human mobility patterns [28, 45]. D_o and d describe how the visited

locations are distributed geographically. d indicates the average distance between trips and D_o quantifies the variation of trip distance. Not all the visited locations are fully connected with each other. Hence in the calculation of d and D_o, distance is only counted when there exists a connection between two consecutive geotagged tweets.

The network features describe properties [26] which characterise how actively people explore new locations. Individual trajectories create a complex network, where a node represents one visited location. The clustering coefficient is a measure of the degree to which locations in such a complex network tend to cluster together. For each visited location, one important property is how frequently the users return to a visited location on average (z (–)) as not every location is evenly visited. The frequency of the most visited location reflects how centralised the complex network is (z_m (–)).

2.3 Potential user groups: hierarchical clustering and down-sampling

We apply hierarchical clustering to the six spatial features. The clustering procedure involves: (1) data standardisation, (2) distance calculation, (3) linkage establishment, and (4) splitting the linkage into clusters. For data standardisation, the min-max method is applied to each feature. For the distance calculation, the squared Euclidean distance is applied [46]. For cluster method of linkage establishment, Ward's method is used [47]. Sensible clustering is measured by the small sum of squares of deviations within the same cluster. By limiting the cluster distance larger than a certain threshold, the final clusters are formulated. The average silhouette width provides an evaluation of clustering validity [48]. As a result of hierarchical clustering, each user is categorised into a group with certain characteristics.

To test the impact of geotweeting frequency on the group identity, random down-sampling is applied to raw individual trajectories of geotagged activities. The features are re-calculated based on the down-sampled trajectories. The same procedure of hierarchical clustering is applied to the updated feature sets to re-identify the behaviour group of individuals.

3 Results

In this section, we first briefly summarise the geotagged activity dataset regarding their spatial and temporal characteristics. Two clustering analyses are applied, resulting in four combinations of clusters that categorise users by two independent aspects of geotagged activities: geographical characteristics and network properties. We present the features of these four categories and visualise the typical network structures of the four categories. We further present the statistical characteristics of the users from each category. Based on these four groups of Twitter users, we present their trip distances and diffusion processes in space and time.

3.1 Descriptive analysis of geotagged activity dataset and its comparison with travel survey

Geotagged tweets of the Swedish users are collected without applying any geographical boundaries (Fig. 2(A)). A large proportion of geotagged locations are in Sweden (Fig. 2(B)). The ratio of distinct locations quantifies the variation level of geotagged locations for each user (Fig. 2(D)). The more geotagged locations that are outside the habitually visited locations, the larger the variation level. At the extreme, if the ratio is 1, the geotagged locations are purely random and we have no information on frequently visited locations such

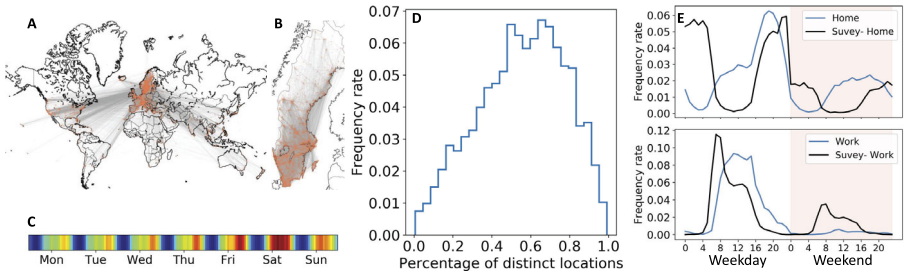

Figure 2 Characteristics of geotagged activity of Swedish users. (**A**) Geotagged activity Origin-Destination Matrix (ODM) on the map. Each point represents a region formulated by DBSCAN clustering with the distance threshold for merging as 1 km and the minimum number of location for a region is set as 1 [10, 49]. (**B**) Geotagged activity ODM on the map where the trip has both origin and destination located within Sweden. (**C**) A week-long geotagging activity pattern that captures the time-dependent characteristic of geotagged locations (average of all the users). The warmer the colour (e.g. red and orange), the higher number of geotagged locations. (**D**) The distribution of the ratio of distinct geotagged locations over total geotagged locations (individually calculated). (**E**) Daily distributions of visiting frequency of the top two most visited locations, weekday vs. weekend (adjusted by the overall distribution of geotagged tweeting frequency over seven days across a week)

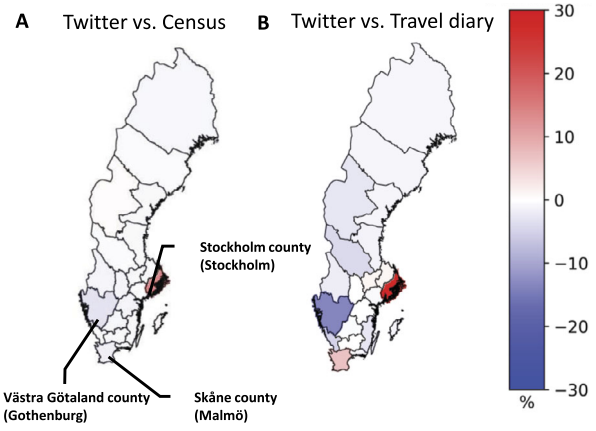

Figure 3 County-level geographical representativeness of estimated home locations from Twitter data: percentage value difference. (**A**) Twitter users vs. residents (Twitter minus Census population) [50]. (**B**) Twitter users vs. Swedish travel survey participants (Twitter minus survey)

as workplace or home. The spread of the distribution, shown in Fig. 2(D), suggests that the proportions of distinct locations are evenly spread out between 0 and 1 among users.

We assume that the first and the second most visited locations by users are either work or home. These two locations have distinct temporal distributions in a day. We apply a hierarchical clustering to the instances of users' daily time distribution of visiting frequency for these two locations. We find two significantly different patterns that fit work and home respectively (Fig. 2(E)). Individual geotagged activity is unevenly distributed in time (Fig. 2(C)). People's weekend activity is more dispersed and they spend less time at the two most visited locations; therefore, the frequencies of visits are lower compared to weekdays.

Figure 3 shows how representative those Twitter users are regarding their estimated home locations compared with the Swedish travel survey and with the Census population.

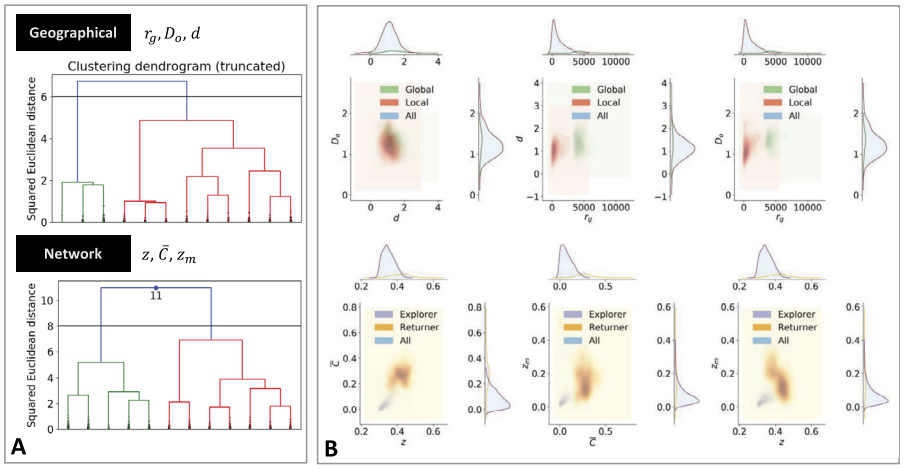

Figure 4 Clustering results. (**A**) Clustering dendrogram. The clusters are formulated by splitting groups into clusters when the cluster distance is above five. (**B**) Kernel density estimation of cluster feature distributions of geographical characteristics and network properties. The height of the one-dimensional distribution is proportional to the fraction of the number of individuals in each cluster over the total individuals. The Silhouette width for both clustering results is above 0.5, indicating that the identified categories are sensible

Not surprisingly, compared with the general population, the top Twitter users in Sweden seem to over-represent the residents in Stockholm county, while the rest of the top Twitter users seem to be distributed similarly to the population distribution (Fig. 3(A)). Compared with the travel survey (Fig. 3(B)), the top Twitter users are more concentrated in Stockholm and Malmö, the third biggest city but under-represent the residents in Västra Götaland county where the second biggest city Gothenburg is located. It is worth noting that the design of travel survey can over- or under- sample certain population segments depending on the expected response rate, usage patterns etc., in order to get representative samples.

3.2 Behavioural categories

There are four categories identified through two clustering analyses (see Fig. 4), one for geographical characteristics (namely global vs. local) and one for network properties (returner vs. explorer).

- Global returner. Geotagged locations are geographically remote and diverse. These individuals generate high proportions of international trips (the destination is outside of Sweden). They also exhibit a centralised network structure. We call this group of individuals *global returner*.
- Global explorer. Geotagged locations are geographically remote and diverse. Like global returners, these individuals frequently travel internationally; however, their visited locations are distributed in a more decentralised way, i.e., the locations are more evenly geotagged. We call this group of individuals *global explorer*.
- Local returner. Geotagged locations are not geographically remote or diverse. These individuals usually visit locations near and connected to a frequently visited centre. The more clustered sub-structures in their location network reflect their occasional explorations around a centralised location. We call this group of individuals *local returner*.

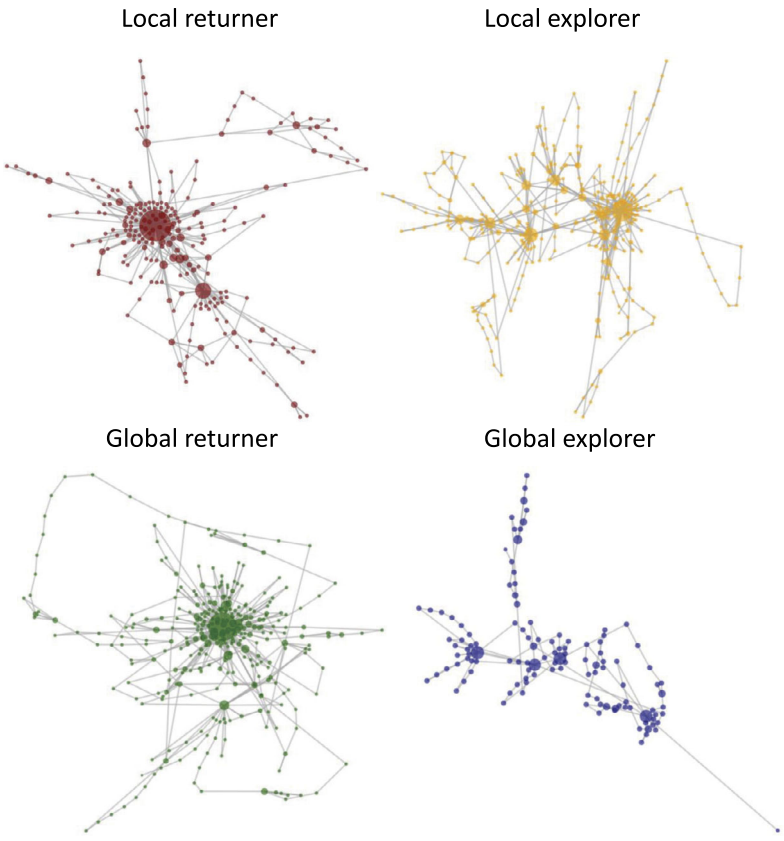

Figure 5 Network visualisation of four representative individuals from each behavioural group. Each node represents one visited location. The diameter of the node is proportional to the node degree

- Local explorer. Geotagged locations are not geographically remote or diverse. There are multiple locations that these individuals visited more frequently than the other locations, and those locations are relatively distant from each other, so the trip distances between them is large. Nevertheless, overall the visited locations are less centralised. Most users are in this category, which we call *local explorer*.

Figure 5 shows the network visualisation of four typical users' trajectories. To better illustrate the network structure, the location position is displayed optimally rather than according to its geographical position. The returners visit different places, centring on a large-degree node (frequently visited location). The chain structure of the explorers is characterised by the lack of a recognisable centre, implying a low returning rate. It is worth noting that the returners have more clustered sub-structures that correspond to daily mobility, i.e., people move near home locations for regular activities (e.g., commuting and shopping), and move around the locations of those regular activities [51].

A statistical summary of the four categories is shown in Table 3. It shows an imbalanced distribution of Twitter users across four groups: most of them are local explorers (78.0%), followed by local returners (14.4%), while the rest are global explorers (7.3%) and global returners (0.3%). The ratio of domestic trips (dom), the returning rate of the most frequently visited location (R), and the geotweeting frequency (F_g) are different between categories (Kruskal–Wallis test, $p < 0.001$). The Mann–Whitney U test is applied to test the variable difference between each pair of categories. Regarding dom, a significant dif-

Table 3 Statistics of four behaviour groups. dom represents the percentage of trips where both the origin and destination are in Sweden (0), among the destination and the origin, there is one location outside Sweden (1), and both the origin and destination are outside of Sweden (2). *R* denotes the ratio of visiting frequency of the most frequently visited location over the total number of geotagged locations. F_g denotes the geotweeting frequency

Name	User (%)	dom (%)			*R*	F_g (/day)	Characteristics
		0	1	2			
Local returner	14.4	81.3	7.0	11.7	0.38	0.55	Small r_g, d, and D_o. Large z, $\overline{C}$, and z_m.
Local explorer	78.0	88.4	5.0	6.6	0.21	0.28	Small r_g, d, and D_o. Small z, $\overline{C}$, and z_m.
Global returner	0.3	45.9	10.0	44.1	0.41	1.57	Large r_g, d, and D_o. Large z, $\overline{C}$, and z_m.
Global explorer	7.3	39.6	12.1	48.3	0.22	0.29	Large r_g, d, and D_o. Small z, $\overline{C}$, and z_m.

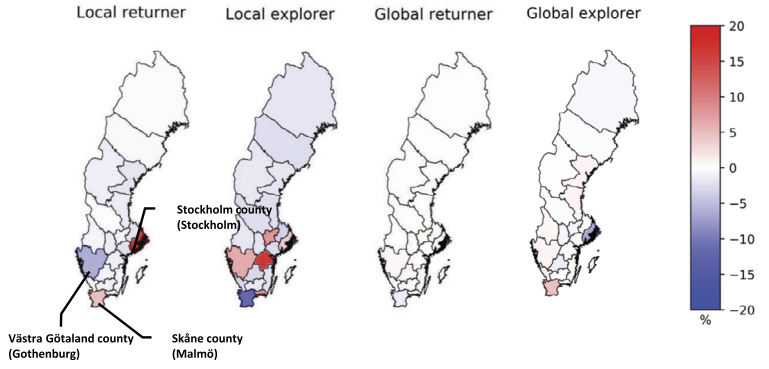

Figure 6 Geographical distribution of estimated home locations of four behavioural groups (county level). The displayed percentage by county is the value of each group minus overall users' value. The warmer the county is shaded, the more its residents are represented in a certain group

ference is found across all category pairs ($p < 0.001$). As for *R* and F_g, there is no significant difference between global returner vs. local returner, and global explorer vs. local explorer. That finding indicates that a high returning rate and frequent geotweeting behaviour are associated with the centralised network structure of geotagged locations. The estimated home locations of different behaviour groups show an interesting spatial pattern (Fig. 6). Compared to overall Twitter users, local returners concentrate more in Stockholm and Malmö. Local explorers concentrate more in the middle of Sweden. Global returners only account for a small proportion of total users, and their geographic distribution is close to the overall studied Twitter users.

3.3 The impact of geotweeting frequency on group identification

Table 3 shows a significant difference in geotweeting frequency between returners and explorers. It is possible that this difference affects the network properties of these users, and thus their group identity, i.e., if the returners' tweeting frequency is reduced to the same rate as explorers, there is a chance that they will be categorised as explorers without changing their actual travel behaviours.

To test the above assumption, we randomly remove 50% of the geotagged tweets from the individuals' original trajectories. Then we calculate the features based on their new

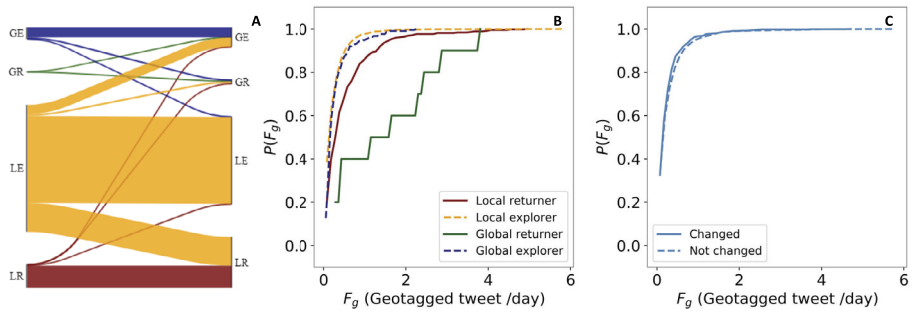

Figure 7 Reclassification results. (**A**) Group changes by the original groups. GE = Global explorer, GR = Global returner, LE = Local explorer, LR = Local returner. (**B**) Geotweeting frequency of four behavioural groups. (**C**) Geotweeting frequency of the users who are re-identified as the original group vs. the ones who are not

geotagged activity trajectories and apply hierarchical clustering to get the new behavioural group. The results are shown in Fig. 7(A). The down-sampling has changed the group identity of a small proportion (around 25%) of users (Fig. 7(A)). The most frequent group change is from local explorer, the largest identity group, to local returner, and from local explorer to global explorer. Figure 7(B) shows the distribution of geotweeting frequency across four behavioural groups. Returners have higher geotweeting frequency in general, however, the group changes are not related to their geotweeting frequency (Fig. 7(C)). Hence, the assumption that the distinct patterns of four user groups are solely due to their difference in geotweeting frequency does not hold. We conclude that the group identities of the users are robust regardless of the users' geotweeting frequency.

3.4 Collective behaviours: trip distance and diffusion in space

In this section, we aggregate all trips within each user category to explore the collective behaviours of the four behavioural groups.

3.4.1 Trip distance

The definition of a trip in the context of geotagged activity, as defined in Sects. 2.1.1 and 2.2, is different from the one in the Swedish travel survey (one-day diary). A trip in Twitter data is the connection between two consecutive geotagged tweets of the same user. It provides incomplete mobility information of individuals because of the spatiotemporal sparsity of tweets. Despite that, at an aggregate level and over large samples, studies generally find good agreements of trip distance comparing Twitter data vs. other sources of data including Call Detail Records (CDR) and censuses [22].

The minimum trip distance for the travel survey data is 1 km [42]. To be comparable with the survey, the Twitter data is reanalysed with a minimum trip distance also set to 1 km and a time frame of 24 hours, which excludes 24.8% of previously-analysed Twitter trips. Only 0.4% trips in the Swedish travel survey are international, while geotagged tweets show 3.6% international trips on a comparable basis.

The geotagged tweets approximate the 1-day travel survey data well for over 90% of the observed one-day domestic trip distances; however, the geotagged data have relatively more long-distance trips than the survey data (Fig. 8(A)). For international trips, despite the similarity between all users' distribution and the survey data, a large population variance exists between different user categories (Fig. 8(B)).

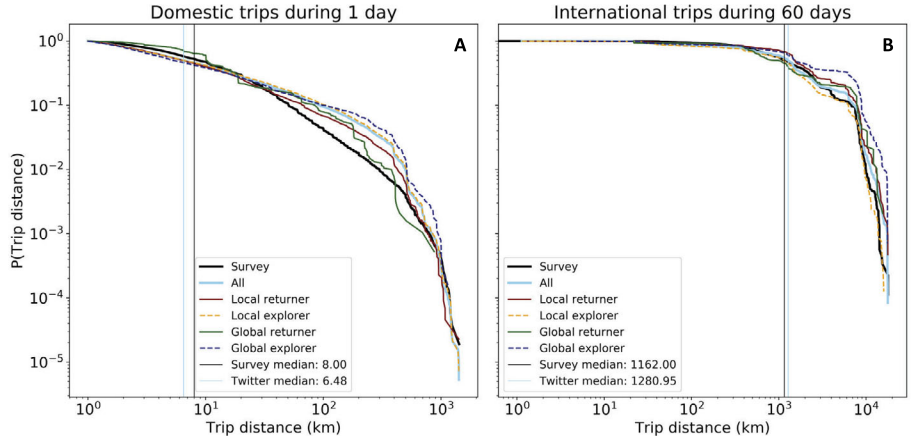

Figure 8 The complementary cumulative distribution function (cCDF) of trip distance. (**A**) Domestic trips validated against the one-day travel survey. (**B**) International trips validated against the 60-day long-distance/international travel survey

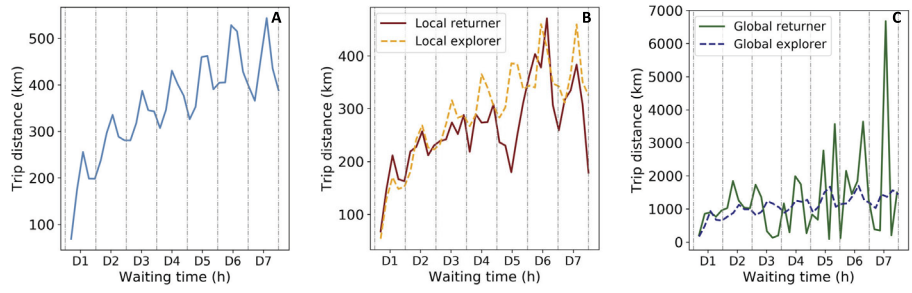

Figure 9 Trip distance vs. waiting time during 7 days. (**A**) All travellers. (**B**) Local travellers. (**C**) Global travellers. Waiting time is defined as the time interval between two consecutive geotagged tweets generated by the same Twitter user

The trip distance as a function of waiting time (the time interval between two consecutive geotagged tweets by the same individual) is shown in Fig. 9. The trip distance generally increases with the waiting time over a multiple-day period at a decreasing rate to up to 7 days (Fig. 9). The correlation between trip distance and waiting time suggests that the observed trip distance increases with waiting time. The diffusive nature of human mobility and the returning effect (e.g., return to home or return to work) create two distinct mechanisms that interact with each other: the diffusion effect causes the observed trip distance to increase with increasing waiting time derived, and the returning effect causes some of the distances to decrease to zero periodically, i.e., every 24 hours (Fig. 9).

3.4.2 Diffusion process

The individual diffusion process is described by the time history of the radius of gyration r_g. We first sort the distinct locations of each individual based on their visiting frequency. The r_g time history begins when the top location has been visited for the first time in one's trajectory of geotagged tweets, and it continues for 90 days thereafter. Previous studies have shown that r_g tends to stabilise within 2000 hours (around 3 months), e.g., [28]. The value of r_g is updated each time a geotagged tweet appears during the 90 days.

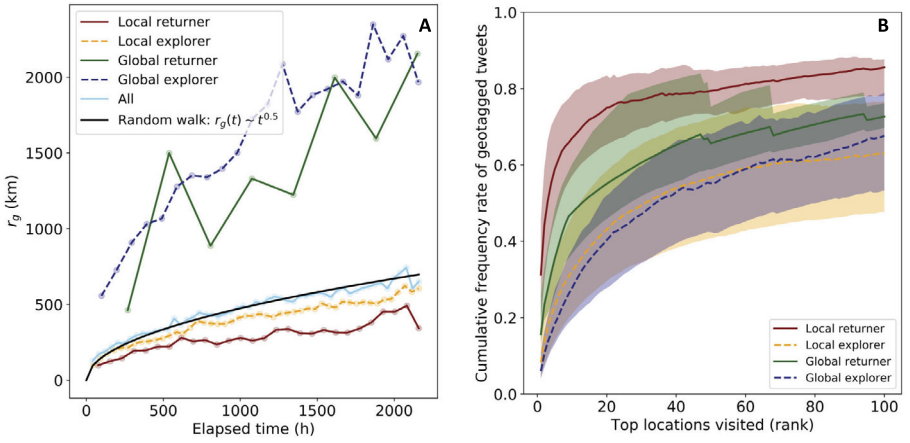

Figure 10 Diffusion process. (**A**) Time history of radius of gyration within 90 days. The time history starts from the first time observing the most visited location; each data point indicates the mean value of r_g across the same group of users. (**B**) Visiting frequency by the ranking order of the most visited locations. The shaded range indicates the upper bound (75%) and lower bound (25%) of the cumulative frequency rate of visits

Each time history is required to contain at least 10 instances of r_g. We normalise the time sequences to the same data length (50 data points) by using nearest-neighbour interpolation to sequences shorter than 50 and randomly down-sampling the sequences longer than 50. Hence, we get a normalised 90-day sequence of r_g for each user who satisfies the conditions above (2303 valid users in total) [17].

Figure 10(A) shows the r_g of the returners compared with the explorers and the time history from the random walk process. The global travellers have a larger mobility range than the local travellers throughout the 90 days. Their mobility range also increases continuously throughout the time period, whereas the returners' mobility range tends to saturate earlier. If individual trajectories followed a random walk [52], then the radius of gyration should follow the solid black line $r_g(t) \sim t^{1/2}$ in Fig. 10(A).

Figure 10(C) shows the cumulative distribution function of the visiting frequency rate vs. the most visited locations ordered by their visiting frequency. The cumulative frequency rate reflects the regularity of users' visiting behaviour. Returners have more concentrated visits to a fewer number of locations than the explorers do. Not only does the cumulative frequency rate start higher and rise faster for returners than for explorers, but the cumulative frequency saturates around a mean value below 80% for returners compared to a mean value below 60% for explorers. The variations for explorers are higher as well (Fig. 10(B)).

4 Discussion

This study presents a picture of population heterogeneity of geotagged activity patterns through a novel combination of individual and aggregate perspectives in the analysis framework. In addition, we collect and apply the geotagged social media dataset spanning a long period and without any geographical boundaries.

4.1 Four distinct geotagged activity patterns: population heterogeneity and collective behaviours

In this study, we propose two essential dimensions of individual mobility, how far one travels and how actively one explores new locations. Based on the correspondingly constructed feature set, most users are identified as local explorers followed by local returners, global explorers, and global returners. Local returners are characterised by the relatively short-range trip distance compared to the global travellers. Returners' trajectories form complex networks that have more concentrated structure than explorers do. Daily mobility makes most people local travellers: they move between and around home, work, and locations of regular activities most of the time, with occasional long distance travel or travel abroad. This explains why most Twitter users are categorised as either local explorers or local returners.

Those two dimensions have been explored separately in some previous studies. Using geotagged tweets, one study found two distinct types of Twitter users with low randomness and high randomness, respectively [10]. In their study, randomness represents the visiting frequency distribution across distinct locations: the more the visiting frequency spreads, the higher the randomness. But that study did not capture the other dimension, how far one travels, which can also differ among sub-populations. Another study used a high-resolution dataset from a mobile navigation app, Sygic in Australia, where two distinct groups of users were found; "travellers" who visit different areas with distinct, salient characteristics, and "locals" who cover shorter distances and revisit many of their locations [53]. But due to high-dimensional indicators, that study did not show the essential differences in human mobility which make the results less intuitive to interpret. In our study, we capture the randomness by using the network structure's properties to quantify how actively one explores new locations. The names of the four groups are inspired by previous studies suggesting a returner-explorer dichotomy in human mobility using GPS log and mobile phone data [29, 54]. Those studies showed two distinct network structures based on individual mobility trajectories: one user type recurrently travelled between many different locations (explorer) and the other had a smaller number of different locations (returner). The network structure was found to be invariant across the distances that one regularly covers (r_g.) Based on that study, we further created geographical characteristics to quantify "how far people travel" with the attempt to achieve a more complete description of human mobility patterns.

The aforementioned studies also have a similar drawback: they apply data sources that only capture individuals' mobility within a country. This incomplete tracking fails to capture international trips and narrows their contributions to domestic mobility only.

The present study has no restrictions from national or adminstrative spatial boundary. We found that even with high time sparsity, social media data can still capture differences in mobility patterns across sub-populations. We illustrate the diffusion process of four groups as one aspect of the collective human mobility patterns in Fig. 10. On the one hand, trip distance increases with waiting time yet decreases at each 24-hour cycle, indicating both the returning effect and the increased probability of exploring new locations (the diffusive nature of mobility). The correlation between trip distance and waiting time agrees with the previous findings from mobile phone data [28] and Twitter data [10]. On the other hand, the time history of r_g highlights the differences between four types of travellers with global explorers' mobility range increasing continuously whereas the returners' mobility

range tending to saturate earlier. The mechanism of stabilising r_g (Fig. 10(A)) has also been described in another study [28]. Randomness (the degree of location predictability) plays an important role in explorers' identity as observed in Fig. 10(B). Explorers have a lesser tendency towards stabilisation in the cumulative frequency rate of visited locations than returners, which describes the essential difference between returners and explorers.

4.2 Implications of human mobility and population heterogeneity

Individual mobility is defined by a person's capabilities, social network and opportunities, i.e. an individual's ability to move, social needs and desire; and the availability of transportation resources such as infrastructure [4]. The understanding of population heterogeneity will benefit a broad range of disciplines from travel behaviour modelling to social sciences. For example, heterogeneity can be applied to generate more accurate agents in transport demand modelling [55]. The continuous tracking and long-term observation of individuals as illustrated in this article can benefit disease prediction by providing a more dynamic and temporal perspective of how people diffuse in space [56] and the importance of adding population heterogeneity to improve the predictions and develop effective mitigation strategies [57, 58]. Putting individuals into different groups or places of residence according to their travel behaviour can also enable new research related to the adoption of new technology [59], etc. The population heterogeneity identified in this study can be combined with sociodemographic information of individuals or groups, e.g., race and income level, in future studies to further understand factors such as the effect of neighbourhoods on travel behaviours of individuals or groups [23], and the relationships between short-term mobility and long-term migration [60]. A study on location-based social networks shows that the shared visited locations are informative in predicting the social connections between individuals [61]. The distinct behavioural groups identified in this study can provide additional insights that contribute to the inference of friendship, such as the relationship between people's mobility and their social network, where a large proportion of places visited are within a small distance of their nearest (geographical) social ties' locations [62]. The relationship between social ties and mobility can be further explored to form a more complete picture [4]. Questions such as whether individuals' social network shapes their mobility behaviour or the other way around can be further studied using data we presented here. Would explorers have a different social network structure compared with returners? Does such a difference contribute to their distinct travel behaviours?

4.3 Representativeness of geotagged social media data as a proxy for human mobility

Compared to the one-day travel diary, the Twitter dataset in this study has strengths and weaknesses as a proxy for human mobility. Based on another ongoing study where we compare geotagged tweets with different data sources, the main strengths of geotagged social media data are in long collection duration, a large number of involved individuals, boundary-free spatial coverage, ease of access, low cost, and accurate location information. The main weaknesses are incomplete individual trajectories caused by high sparsity in the time dimension (plus behaviour bias), lack of socio-demographic information (plus population bias), and lack of trip information [24].

Despite high time sparsity, one of the most appealing features of geotagged social media data is the capability of continuous and long-term tracking of individual mobility via

their geotagged activities. We demonstrate how that particular feature helps capture the heterogeneity of mobility. A one-day travel diary only captures trips generated within 24 hours for each survey participant, while the Twitter dataset covers on average 3.6 years for each participant. Although Twitter data is extremely sparse, the long-term and continuous tracking compensates for the time sparsity, allowing us to obtain a realistic picture of user's trips on an average day.

Given the reported biases in the geotagged social media data, the current study carefully conducts a descriptive analysis in comparison with the travel survey. The behaviour of using social media is complex and multidimensional. For example, more than 20 tweeting features have been used to characterise "how you tweet" including various time-related statistics [63]. If users constantly and regularly tweet during a certain daily time frame or only from a few selected locations, then the locations we capture are skewed to the locations that they tend to visit during that time frame. However, as seen in our study Fig. 2(D), it is not the case that people only geotweet from a few fixed locations. Despite peaks during lunch time and night (Fig. 2(C)), geotagged tweets capture many routine activities (Fig. 2(D)), as seen from the temporal profile of the first and second most visited locations that share some similarities with the "ground truth" in the travel survey. We explore population bias by comparing the geographical distribution of the Twitter users' estimated home location in this study with those from the travel survey. It appears that the top Twitter users are over-represented by residents of big cities. This is consistent with the observation by a previous study [27].

Some of the disadvantages mentioned above can potentially be mitigated. Text mining could be applied to derive location information from the contents of the tweets. One study has proposed such an approach to infer city-level location of tweets, partially mitigating the time sparsity of geotagged tweets [64]. Similarly, data fusion could be promising to obtain better application performance, e.g., activity prediction [65].

5 Conclusions

In this study, we develop a novel analysis framework to categorise individuals regarding their geotagged activity patterns to reveal population heterogeneity of mobility patterns. Based on the classification results, trip distance and diffusion process in space are presented by distinct group. The major contributions of this study include:

(1) Datasets and analysis framework. This study involves two data sources; a household travel survey and geotagged social media data. The geotagged tweets dataset covers a long period (3.6 years on average) without any geographical boundaries. The descriptive analysis of geotagged tweets reveals behaviour and population differences between the two data sources. Our analysis framework provides a coherent picture of the geotagged activity patterns by combining the individual perspective with the aggregate perspective.

(2) Four distinct groups of users. We propose two essential aspects to quantify population heterogeneity of human mobility: how actively one explores new locations and how far one travels. A set of features are defined to describe geotagged activity patterns from the perspectives of geographical characteristics and network properties. A hierarchical clustering analysis is applied, and four types of Twitter users are identified: local explorers, local returners, global explorers, and global returners.

(3) Population heterogeneity. On the aggregate level, we present the diffusion process in space based on geotagged social media data. It shows good agreements with previous studies, while the key differences between user groups are quantified.

One limitation of the current study is the small number of individuals categorised as global returners. Therefore, their behaviour captured in this study is less reliable than the rest of the groups. However, for the sake of completeness reflecting the clustering analyses discussed in Sect. 3.2, we keep all the groups' identities to show how geotagged tweets reveal the heterogeneity of travelling behaviour. Future studies can increase the sample size of this subpopulation to explore the robustness of the results presented here.

The other major limitation is that our conclusions from the geotagged activity patterns may not be generalised to the overall population due to the population and behaviour biases introduced by using geotagged tweets [24]. Given the known shortcomings, more systematic research efforts are required to identify and correct for these biases. Our next step is to systematically compare multiple data sources, such as travel surveys, geotagged social media, call detail records, and GPS logs to reach a deeper understanding of the strengths and weakness of each data source [24]. With such understanding, we can further develop mobility models informed by revealed population heterogeneity, leveraging geotagged social media data to estimate more accurate and timely travel patterns and demand.

Acknowledgements
The authors would like to express their sincere gratitude to the two anonymous reviewers whose comments have greatly improved this manuscript.

Funding
This research is funded by the Swedish Research Council Formas (Project Number 2016-1326).

Abbreviations
ICT, Information and Communication Technology; GPS, Global Positioning System; RQ, Research Question; API, Application Programming Interface; UTC, Coordinated Universal Time; ODM, Origin-Destination Matrix; GC, Geographical Cluster; NC, Network Cluster; CDR, Call Detail Records; cCDF, Complementary Cumulative Distribution Function.

Availability of data and materials
The data that support the findings of this study are available from *Twitter* (https://twitter.com). The dataset generated and analysed for this study is not publicly available due to privacy protection of the subjects in our dataset, which can be potentially used to identify individuals.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
YL, SY and GJ designed the study. YL analyzed the data. YL and SY wrote the paper. All authors edited and approved the final version of this manuscript.

Author details
[1]Department of Space, Earth and Environment, Division of Physical Resource Theory, Chalmers University of Technology, Gothenburg, Sweden. [2]School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH Royal Institute of Technology, Stockholm, Sweden.

Publisher's Note

Received: 1 November 2018 Accepted: 29 October 2019 Published online: 14 November 2019

References

1. Noulas A, Scellato S, Lambiotte R, Pontil M, Mascolo C (2012) A tale of many cities: universal patterns in human urban mobility. PLoS ONE 7(5):37027. https://doi.org/10.1371/journal.pone.0037027
2. Treiber M, Kesting A (2013) Traffic flow dynamics. Traffic flow dynamics: data, models and simulation. Springer, Berlin. https://doi.org/10.1007/978-3-642-32460-4
3. Balcan D, Colizza V, Gonçalves B, Hu H, Ramasco JJ, Vespignani A (2009) Multiscale mobility networks and the spatial spreading of infectious diseases. Proc Natl Acad Sci 106(51):21484–21489. https://doi.org/10.1073/pnas.0906910106
4. Kaufmann v, Bergman M, Joye D (2004) Motility: mobility as capital. Int J Urban Regional 28-4:745–756. https://doi.org/10.1111/j.0309-1317.2004.00549.x

5. Chen C, Ma J, Susilo Y, Liu Y, Wang M (2016) The promises of big data and small data for travel behavior (aka human mobility) analysis. Transp Res, Part C, Emerg Technol 68:285–299. https://doi.org/10.1016/j.trc.2016.04.005
6. Janzen M, Müller K, Axhausen KW (2017) Population synthesis for long-distance travel demand simulations using mobile phone data. In: 6th symposium of the European association for research in transportation (hEART 2017).
7. Wang Z, He SY, Leung Y (2018) Applying mobile phone data to travel behaviour research: a literature review. Travel Behav Soc 11:141–155. https://doi.org/10.1016/j.tbs.2017.02.005
8. Zhang Z, He Q, Zhu S (2017) Potentials of using social media to infer the longitudinal travel behavior: a sequential model-based clustering method. Transp Res, Part C, Emerg Technol 85:396–414. https://doi.org/10.1016/j.trc.2017.10.005
9. Yue Y, Lan T, Yeh AGO, Li Q-Q (2014) Zooming into individuals to understand the collective: a review of trajectory-based travel behaviour studies. Travel Behav Soc 1(2):69–78. https://doi.org/10.1016/j.tbs.2013.12.002
10. Jurdak R, Zhao K, Liu J, AbouJaoude M, Cameron M, Newth D (2015) Understanding human mobility from Twitter. PLoS ONE 10(7):0131469. https://doi.org/10.1371/journal.pone.0131469
11. Hasan S, Ukkusuri SV (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
12. Gao S, Yang JA, Yan B, Hu Y, Janowicz K, McKenzie G (2014) Detecting origin-destination mobility flows from geotagged tweets in greater Los Angeles area. In: Proceedings of the eighth international conference on geographic information science, pp 1–4
13. Hasan S, Schneider C, Ukkusuri S, González M (2013) Spatiotemporal patterns of urban human mobility. J Stat Phys 151(1–2):304–318. https://doi.org/10.1007/s10955-012-0645-0
14. Morstatter F, Pfeffer J, Liu H, Carley KM (2013) Is the sample good enough? Comparing data from Twitter's streaming api with Twitter's firehose. In: Seventh international AAAI conference on weblogs and social media. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6071/6379
15. Stolf Jeuken G (2017) Using big data for human mobility patterns—examining how Twitter data can be used in the study of human movement across space. Master's thesis. http://studentarbeten.chalmers.se/publication/250155-using-big-data-for-human-mobility-patterns-examining-how-twitter-data-can-be-used-in-the-study-of-hu
16. Rashidi TH, Abbasi A, Maghrebi M, Hasan S, Waller TS (2017) Exploring the capacity of social media data for modelling travel behaviour: opportunities and challenges. Transp Res, Part C, Emerg Technol 75:197–211. https://doi.org/10.1016/j.trc.2016.12.008
17. Liao Y, Yeh S (2018) Predictability in human mobility based on geographical-boundary-free and long-time social media data. In: 2018 21st international conference on intelligent transportation systems (ITSC). IEEE Press, New York, pp 2068–2073. https://doi.org/10.1109/ITSC.2018.8569770
18. Malik MM, Lamba H, Nakos C, Pfeffer J (2015) Population bias in geotagged tweets. In: Ninth international AAAI conference on web and social media, pp 18–27. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/viewPaper/10662
19. Ruths D, Pfeffer J (2014) Social media for large studies of behavior. Science 346(6213):1063–1064. https://doi.org/10.1126/science.346.6213.1063
20. Tasse D, Liu Z, Sciuto A, Hong JI (2017) State of the geotags: motivations and recent changes. In: Eleventh international AAAI conference on weblogs and social media, pp 250–259. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/viewPaper/15588
21. Wesolowski A, Eagle N, Noor AM, Snow RW, Buckee CO (2013) The impact of biases in mobile phone ownership on estimates of human mobility. J R Soc Interface 10(81):20120986. https://doi.org/10.1098/rsif.2007.1218
22. Lenormand M, Picornell M, Cantú-Ros OG, Tugores A, Louail T, Herranz R, Barthelemy M, Frias-Martinez E, Ramasco JJ (2014) Cross-checking different sources of mobility information. PLoS ONE 9(8):105184. https://doi.org/10.1371/journal.pone.0105184
23. Wang Q, Phillips NE, Small ML, Sampson RJ (2018) Urban mobility and neighborhood isolation in America's 50 largest cities. Proc Natl Acad Sci 115(30):7735–7740. https://doi.org/10.1073/pnas.1802537115
24. Liao Y, Yeh S (2020) Using geotagged tweets to assess human mobility: a comparison with travel survey and GPS log data (under review). Transp Res, Part C, Emerg Technol
25. Hasnat MM, Hasan S (2018) Identifying tourists and analyzing spatial patterns of their destinations from location-based social media data. Transp Res, Part C, Emerg Technol 96:38–54. https://doi.org/10.1016/j.trc.2018.09.006
26. Lenormand M, Gonçalves B, Tugores A, Ramasco JJ (2015) Human diffusion and city influence. J R Soc Interface 12(109):20150473. https://doi.org/10.1098/rsif.2015.0473
27. Mislove A, Lehmann S, Ahn Y-Y, Onnela J-P, Rosenquist JN (2011) Understanding the demographics of Twitter users. In: Fifth international AAAI conference on weblogs and social media, pp 554–557. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2816/3234
28. Gonzalez MC, Hidalgo CA, Barabasi A-L (2008) Understanding individual human mobility patterns. Nature 453(7196):779–782. https://doi.org/10.1038/nature07850
29. Song C, Koren T, Wang P, Barabási A-L (2010) Modelling the scaling properties of human mobility. Nat Phys 6(10):818–823. https://doi.org/10.1038/nphys1760
30. Coffey C, Pozdnoukhov A (2013) Temporal decomposition and semantic enrichment of mobility flows. In: Proceedings of the 6th ACM SIGSPATIAL international workshop on location-based social networks. LBSN'13. ACM, New York, pp 34–43. https://doi.org/10.1145/2536689.2536806
31. Chang J, Sun E (2011) Location3: how users share and respond to location-based data on social networking sites. In: Proceedings of the fifth international AAAI conference on weblogs and social media, pp 74–80
32. Pianese F, An X, Kawsar F, Ishizuka H (2013) Discovering and predicting user routines by differential analysis of social network traces. In: 2013 IEEE 14th international symposium and workshops on a World of wireless, mobile and multimedia networks (WoWMoM). IEEE Press, New York, pp 1–9. https://doi.org/10.1109/WoWMoM.2013.6583383
33. Hasan S, Ukkusuri SV (2015) Location contexts of user check-ins to model urban geo life-style patterns. PLoS ONE 10(5):0124819. https://doi.org/10.1371/journal.pone.0124819

34. Yang D, Zhang D, Zheng VW, Yu Z (2015) Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns. IEEE Trans Syst Man Cybern Syst 45(1):129–142. https://doi.org/10.1109/TSMC.2014.2327053
35. Jin P, Cebelak M, Yang F, Zhang J, Walton C, Ran B (2014) Location-based social networking data: exploration into use of doubly constrained gravity model for origin-destination estimation. Transp Res Rec 2430:72–82. https://doi.org/10.3141/2430-08
36. Lee JH, Gao S, Goulias KG (2015) Can Twitter data be used to validate travel demand models. In: 14th international conference on travel behaviour research.
37. Lee JH, Davis AW, Yoon SY, Goulias KG (2016) Activity space estimation with longitudinal observations of social media data. Transportation 43(6):955–977. https://doi.org/10.1007/s11116-016-9719-1
38. Keuschnigg M, Mutgan S, Hedström P (2019) Urban scaling and the regional divide. Sci Adv 5(1):0042. https://doi.org/10.1126/sciadv.aav0042
39. Kantardzic M (2011) Data mining: concepts, models, methods, and algorithms. Wiley, Hoboken. https://doi.org/10.1109/9780470544341
40. The Tweepy project developers: Tweepy: v3.5.0 (2017). http://tweepy.readthedocs.io/en/v3.5.0/
41. Barabási A-L (2005) The origin of bursts and heavy tails in human dynamics. Nature 435(7039):207–211. https://doi.org/10.1038/nature03459
42. Official Statistics of Sweden: Swedish National Travel Survey (RVU Sweden) 2011–2016. (2016). https://www.trafa.se/en/travel-survey/travel-survey/
43. Markovich N (2008) Nonparametric analysis of univariate heavy-tailed data: research and practice, vol 753. Wiley, Chichester
44. Barabási A-L et al (2016) Network science. Cambridge University Press, Cambridge
45. Song C, Qu Z, Blumm N, Barabási A-L (2010) Limits of predictability in human mobility. Science 327(5968):1018–1021. https://doi.org/10.1126/science.1177170
46. Deza MM, Deza E (2009) Encyclopedia of distances. Springer, Berlin. https://doi.org/10.1007/978-3-642-00234-2
47. Ward JH Jr (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58(301):236–244. https://doi.org/10.1080/01621459.1963.10500845
48. Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7
49. 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. In: Kdd, vol 96. AAAI Press, Palo Alto, pp 226–231.
50. Statistics Sweden: Population of Sweden in 2016, by county (2016). https://www.statista.com/statistics/526617/sweden-population-density-by-county/
51. Golledge RG, Stimson RJ (1997) Spatial behavior: a geographic perspective. Guilford Press, New York
52. Brockmann D, Hufnagel L, Geisel T (2006) The scaling laws of human travel. Nature 439(7075):462. https://doi.org/10.1038/nature04292
53. Scherrer L, Tomko M, Ranacher P, Weibel R (2018) Travelers or locals? Identifying meaningful sub-populations from human movement data in the absence of ground truth. EPJ Data Sci 7(1):19. https://doi.org/10.1140/epjds/s13688-018-0147-7
54. Pappalardo L, Simini F, Rinzivillo S, Pedreschi D, Giannotti F, Barabási A-L (2015) Returners and explorers dichotomy in human mobility. Nat Commun 6. https://doi.org/10.1038/ncomms9166
55. Anda C (2018) A time-space model of disaggregated urban mobility from aggregated mobile phone data. In: 15th international conference on travel behavior research (IATBR 2018). Future Cities Laboratory (FCL), Zurich. https://doi.org/10.3929/ethz-b-000300714
56. Xu Z, Glass K, Lau CL, Geard N, Graves P, Clements A (2017) A synthetic population for modelling the dynamics of infectious disease transmission in American Samoa. Sci Rep 7(1):16725. https://doi.org/10.1038/s41598-017-17093-8
57. Merler S, Ajelli M (2010) Human mobility and population heterogeneity in the spread of an epidemic. Proc Comput Sci 1(1):2237–2244
58. Dobra A, Bärnighausen T, Vandormael A, Tanser F (2019) A method for statistical analysis of repeated residential movements to link human mobility and hiv acquisition. PLoS ONE 14(6):0217284
59. Vannoy SA, Palvia P (2010) The social influence model of technology adoption. Commun ACM 53(6):149–153
60. Fiorio L, Abel G, Cai J, Zagheni E, Weber I, Vinué G (2017) Using Twitter data to estimate the relationship between short-term mobility and long-term migration. In: Proceedings of the 2017 ACM on web science conference. ACM, New York, pp 103–110
61. Pelechrinis K, Krishnamurthy P (2016) Socio-spatial affiliation networks. Comput Commun 73:251–262
62. Phithakkitnukoon S, Smoreda Z, Olivier P (2012) Socio-geography of human mobility: a study using longitudinal mobile phone data. PLoS ONE 7(6):39253
63. Pennacchiotti M, Popescu A-M (2011) A machine learning approach to Twitter user classification. In: Fifth international AAAI conference on weblogs and social media. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewPaper/2886
64. Cheng Z, Caverlee J, Lee K (2010) You are where you tweet: a content-based approach to geo-locating Twitter users. In: Proceedings of the 19th ACM international conference on information and knowledge management, CIKM'10. ACM, New York, pp 759–768. https://doi.org/10.1145/1871437.1871535
65. Zhu Z, Blanke U, Tröster G (2014) Inferring travel purpose from crowd-augmented human mobility data. In: Proceedings of the first international conference on IoT in urban space. URB-IOT '14. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, pp 44–49. https://doi.org/10.4108/icst.urb-iot.2014.257173