Highlights

- Five data sets from three sources integrating both longitudinal & lateral data forms.
- A systematic comparison framework combining individual trajectory and places network.
- Geotagged tweets capture reasonably well the fundamental laws of individual mobility.
- The effects of social media's population biases and behaviour distortions are mixed.
- Geotagged tweets represent poorly mobility at the population level e.g. trip distance

Using geotagged tweets to assess human mobility: a comparison with travel survey and GPS log data

Yuan Liao*,[a], Sonia Yeh[a]

[a]*Department of Space, Earth and Environment, Division of Physical Resource Theory, Chalmers University of Technology, Gothenburg, Sweden*

Abstract

Understanding human mobility has broad relevance. The drawbacks of current survey methods have spurred interest in emerging data sources, such as Twitter. However, these sources do not provide an unbiased sample. This study attempts to comprehensively examine the validity of using Twitter data to characterize various dimensions and indicators of travel behaviors by carefully comparing individual trajectories and places networks with travel survey data and GPS log data. We find that geotagged tweets from top geotag users capture the fundamental laws of individual mobility reasonably well, including returning effect, diffusive nature, and trip distance distribution, consistent with previous studies. However, the estimated home and work locations of those users may not be reliable due to temporal behavior distortion. Moreover, due to the low geotweeting frequency per user, bursty nature of most users' tweeting, and tendency to report social activities, geotagged tweets poorly represent the mobility indicators at the population level, such as trip distance. We also find that despite good agreement on the network structure of visits, higher preferences of movements are driven by social activities in geotagged tweets . Compared to the other two data sources, geotagged tweets are the only data source that allows us to examine both individual trajectories (domestic and international) and places networks, but the validity of these results vary depending on how they are used.

Keywords: travel behavior, social media data, longitudinal and lateral, comparison, data mining, network science

1. Introduction

Traditionally, studies of travel behavior rely on household travel surveys, featuring accuracy with respect to reported locations, well-resolved temporal data, and statistically valid demographic information. However, the costs of these surveys are increasing (Yue et al., 2014), and the surveys only cover a short period (one day for a travel diary) and suffer from a low and decreasing response rate (Stopher and Greaves, 2007). Emerging data sources associated with mobile/smart phones are increasingly being leveraged to overcome these drawbacks. Global Positioning System (GPS) log data have attracted attention due to high sampling frequency and continuous and passive collection (Toch et al., 2018).). For a detailed review and history, see studies by Chen et al. (2016) and Barbosa et al. (2018).

Alongside the development of Information and Communication Technologies (ICT), interest in online social media services, e.g., Twitter, has grown among the transportation research community (Rashidi et al., 2017) focusing on, for instance, individual mobility (e.g. Lee et al., 2016), population-level mobility (typically geographically bounded to city or national level) (e.g. Jin et al., 2014), and travel-demand modelling (e.g. Lee et al., 2015). A tweet typically contains multiple components that can be useful for transport research, including text, hashtag, location, and timestamp. When users choose to have their location reported when sending out tweets, these are called **geotagged tweets**. Geotagged tweets account for a small proportion (1-3%) (Morstatter et al., 2013). That number varies between regions, 7.4% (George, South Africa), 1.9% (Barcelona, Spain), 1.1% (Kuwait), and 0.3% (Sweden) (Stolf Jeuken, 2017). Despite the low proportion of geotagged tweets, these check-ins provide precise location information and have increasingly been used for understanding mobility (Lenormand et al., 2014; Jurdak et al., 2015).

*Corresponding author. E-mail: yuan.liao@chalmers.se.

The low cost of retrieving geotagged tweets makes them especially appealing compared to other data sources (Rashidi et al., 2017). The data source is free to access, and it provides precise location information with a spatial resolution of around 10 meters compared with 100-200 meters for call detail records (CDR) (Jurdak et al., 2015). Moreover, it allows for long-term tracking of movements that are free of geographical boundaries (Liao and Yeh, 2018). The main criticism pertains to two aspects, a biased population representation and low and irregular sampling. There have been studies comparing multiple data sources to identify/adjust the biases (e.g. Wesolowski et al., 2013; Tasse et al., 2017) and to validate against "ground truth" (e.g. Lenormand et al., 2014). Despite the known disadvantages of geotagged tweets, one recent literature review shows that experts are positive about the usefulness of such data sources for modeling travel behavior (Rashidi et al., 2017). There is also a consensus on the need for careful inspection of using geotagged social media data to approximate the actual travel behavior of the general population.

Geotagged tweets are especially controversial as a proxy for mobility. Efforts have been devoted to understanding the motivations behind geotagged tweets and also the validity of such a data source by comparing it with some other data sources. Previous studies typically treat one data source, often the travel survey data, as the ground truth for validation, and the validation focuses on limited aspects of travel behavior. However, different data sources have their pros and cons. For example, the one-day travel diary and the long-distance travel survey module are typically conducted separately in the household travel survey . There is a tendency to underestimate the total travel demand from the one-day travel diary, and the details, and frequencies of long-distance travels are often poorly characterized (Janzen et al., 2017). GPS log data provide high-frequency location data over multiple days to months, but the information is limited within the service areas of the operator. They are also difficult and expensive to obtain from private service providers (Laurila et al., 2012). Geotagged tweets can capture movements over multiple years and include overseas visits; but the data are "sparse," and thus the picture of actual movements is incomplete (Liao et al., 2019).

By carefully comparing individual trajectories and places networks with travel surveys and GPS log data in this study, we attempt to comprehensively examine the validity of using Twitter data to characterize various dimensions and indicators of travel behaviors.

1.1. Related work

Mobility refers to the movement of individuals or groups of people in space and time (Barbosa et al., 2018). For individuals, the mobility trajectory is a time series of visits to various locations. Individuals mobility trajectories can be aggregated to study the flows of people traveling between different locations/regions. Depending on the spatio-temporal scale of the aggregation, an origin-destination matrix (OD matrix) can be constructed with the origins and the destinations of all trips. Therefore, this mobility study has two major focuses, **individual trajectories** and **networks of places**, here, a "places network." In the last decade, the emerging data sources have significantly improved our understanding of mobility (Gonzalez et al., 2008; Song et al., 2010b; Barbosa et al., 2018). Common emerging data sources are call detail records (CDR), tracking apps on smart phones, GPS-enabled devices, and geotagged social media.

Two forms of data are often used: longitudinal and lateral. A **longitudinal** data set is characterized by the long-term (more than 24 hours) and continuous observations focusing on a group of participants, such as GPS log (e.g. Laurila et al., 2012), call detail records (CDR) (Wesolowski et al., 2012), and Twitter users' geotagged activity trajectories (e.g. Liao and Yeh, 2018). Longitudinal data sets are often applied to reveal the patterns of individual mobility, e.g., the socio-geography of mobility (Phithakkitnukoon et al., 2012) and the activity space estimation (Lee et al., 2016). Because it is possible to observe the individual trajectory over a long period of time, more attention has been paid to the routine mobility (Pianese et al., 2013) and next-location prediction (Do et al., 2015).). A **lateral** data set is often collected based on a particular area, such as a city or a country, during a short-to-medium period, and it usually covers a larger population. It is commonly used to study the travel demand (Jin et al., 2014) and behavior patterns at the aggregate level (Alessandretti et al., 2018). The difference between the aforementioned two data forms is the trade-off between the number of individuals and data collection duration.

The data sources reviewed in this study include household travel surveys, CDR, GPS log data, and geotagged social media data, among which three are non-conventional data sources. The main characteristics of the four data sources are summarized briefly in Table 1. Compared with the other data sources, geotagged social media data have strengths in long collection duration, a large number of studied individuals, large spatial coverage, ease of access, low

cost, and accurate location information. The main weaknesses are incomplete sampling of individual trajectories and lack of socio-demographic information and trip information.

Table 1: Characteristics of the four data sources. [a]Geotagged social media data. [b]Traditional household travel survey. [c]Time length of tracking the same individual. [d]Low cost = +++. Medium cost = ++. High cost = +.

	Check-ins[a]	Travel survey[b]	CDR	GPS log
Time duration[c]	+++	+	+++	++
Number of individuals	++	+++	+++	+
Spatial coverage	+++	++	++	+
Trajectory completion	+	+++	++	+++
Accessibility	+++	++	+	+
Cost[d]	+++	+	++	++
Spatial resolution	+++	++	++	+++
Temporal resolution	+	+++	++	+++
Socio-demographic info.	✗	✓	✗	✓
Trip info.	✗	✓	✗	✗/✓
Passive collection	✓	✗	✓	✗

In Table 2, we further review studies using the four data sources based on their research focus (individual trajectory vs places network) and data form (longitudinal vs lateral). And we also summarize the pros and cons of using them.

Due to the lack of longitudinal data, most previous studies used lateral data (Chen et al., 2016) among which **household travel surveys** were the most prevalent. Travel surveys contain socio-demographic information and detailed activity records making them not easily replaceable by other emerging data sources (Janzen et al., 2017). Because samplings are carefully designed to derive statistically representative population-level estimates, traditional travel surveys remain a vital source for validation/calibration of the emerging data sources. But they also have many shortcomings such as being costly to collect and having low sampling rates, short survey durations, under-reporting of trips, and quickly being out-of-date (Wang et al., 2018). Travel surveys also fail to capture most of the long-distance trips (Janzen et al., 2017).

Mobile phone CDR are the most widely applied among these emerging data sources (Yue et al., 2014). A record in a CDR data set represents a phone call or a text message with the phone activity information (start time, duration, and end time, etc.) and the GPS coordinates of the tower that first channelled the activity. This implies that the spatial accuracy of an individual location depends on the cell tower network's spatial resolution, typically 200-300 meters. CDR can be collected long-term with very large numbers of tracked individuals. For example, a study uses one-year-long CDR series with nearly 15 million tracked individuals to study the impact of mobility on malaria (Wesolowski et al., 2012). Nevertheless, this data source is often not easy to access, and, compared with travel surveys, has the shortcomings of spatiotemporal sparsity and incomplete trajectories (Chen et al., 2018). It is also often not available for follow-up tracking and continuous update.

GPS log data contain the records of GPS coordinates sampled in regular and high frequency. Applied GPS log data can be divided into two main categories: human-carried GPS logger and vehicle-attached logger. The latter is beyond the scope of this paper. As shown in Table 2, most previous studies apply GPS log data from a rather small group of individuals (20-500). Most of these studies come from the computer science community focusing on the individual-based prediction of future locations (e.g. Etter et al., 2012), so do three out of the four GPS log data studies in Table 2. Compared with CDR and household travel surveys, such a data source is used less frequently by the transport research community due to small sample size, high cost, and lack of modal travel information (even though some research efforts specialize in deriving modal estimates from the logged data). Overall, GPS log data provide a relatively complete and accurate picture of individual mobility trajectory, making it close to the "ground truth." Therefore, it is included in this study as a data source with which geotagged social media data are compared.

Geotagged tweets can be obtained in three main ways: 1) Purchase the complete set of public tweets from Twitter Firehose; 2) Access the Streaming API to get a maximum of 1% of the public tweets; 3) Access the user timeline by user name/ID to get a maximum of 3200 historical tweets that are set by the user as publicly accessible. Geotagged tweets collected from the Streaming API are often limited to a geographical bounding box yielding a lateral data set. It covers a large number of Twitter users but takes time to accumulate enough samples, and individuals' movements outside the bounding box are not captured (Liao et al., 2019). By accessing the user timeline, all the publicly available historical tweets by a specified user can be collected resulting in a longitudinal record of the individual trajectory

without any geographical boundaries. As summarized in Table 2, longitudinal geotagged tweets are the only data source that is not constrained to a specific area. This type of longitudinal data has been scaled up to large numbers of Twitter users to study the influence of global cities on human diffusion (Lenormand et al., 2015). Most studies use geotagged tweets in the lateral form, i.e., focusing on a specified area that is often in line with the spatial scale of policy-making and urban planning. For lateral data, the individual trajectory of geotagged tweets is often aimed at validation and understanding of fundamental laws of human mobility, such as the power law distribution of trip distance (Jurdak et al., 2015). Compared to individual trajectories, places networks gain more attention because they connect directly to travel-demand modeling and have greater potential to support applications such as modifying the classic Gravity model by integrating locations posted on Foursquare (Jin et al., 2014).

Geotagged tweets have been heavily criticized for population bias and behavioral distortion. A study focusing on the U.S. found that Twitter users tend to over-represent dense population regions and are predominantly male (Mislove et al., 2011). There are two possible types of behavioral distortion for Twitter users who geotag: platform-driven behavior, i.e., they only tweet at specified locations or times. On top of that, the decision to geotag only certain or all of the tweets makes it even more complicated. Untangling psycho-social behavior from platform-driven behavior is rarely done, and only a few studies have attempted to do so (Ruths and Pfeffer, 2014). A recent study on the state of the geotags shows that people geotag consciously and intentionally in uncommon places to communicate and show where theyve been, and they geotag soon after being at the place (Tasse et al., 2017).

So far, many of the studies using geotagged tweets focus on identifying the universal laws of human mobility at the aggregate level (e.g. Jurdak et al., 2015), such as the truncated power law of trip distance distribution (Gonzalez et al., 2008), and Zipf's law of the visitation frequency which describes people's tendency to return to a couple of locations they frequently visit (Song et al., 2010a).). One of our studies illustrates that geotagged tweets can be used to identify the different travel behavior patterns (e.g. geographical characteristics and network properties) among four sub-population types (Liao et al., 2019).

When social media data are cross-validated against the higher temporal resolution data such as CDR (Lenormand et al., 2014) and travel surveys (Liao et al., 2019), good agreement is generally found regarding trip distance distribution, etc. When validating geotagged tweets against travel surveys, one study shows that geotagged social media data capture the displacement distribution, length, duration, and start time of trips reasonably well for inferring individual travel behavior (Zhang et al., 2017). Validations using CDR need to be interpreted carefully as CDR and geotagged tweets have similar passive data collection manners that might share some shortcomings. Some studies have compared geotagged tweets with traffic data (Ribeiro et al., 2014) and travel-demand data (Lee et al., 2015), generally achieving good results. To date, the work comparing geotagged tweets with other data sources lacks systematic rigor. On the one hand, the comparisons are typically limited to one to two data sources without comprehensive considerations of the trade-offs and limitations of different data forms. On the other hand, studies either focus solely on individual trajectories or places networks, but not both. The links between these two perspectives have rarely been carefully constructed and validated at the same time.

To fill the gaps in the literature, this paper explores the representativeness of geotagged tweets as a proxy for mobility by conducting a systematic comparison, using both longitudinal and lateral data forms to examine individual trajectories and places networks. This paper is organized as follows. Section 2 describes the data sources and the data sets tailored for comparison as well as their limitations. Section 3 describes the methodological comparison framework. The results are shown in Sections 4 (individual trajectories) and 5 (places networks). Section 6 discusses the findings. Section 7 concludes and discusses future research needs.

2. Data description and preprocessing

2.1. GPS log

The GPS data set is from Mobile Data Challenge (MDC) from October 2009 to the end of March 2011 (Laurila et al., 2012). It contains regularly recorded GPS data from the participants living in Lausanne, Switzerland, during the MDC campaign. The campaign population reached 185 participants (38% female, 62% male), and was concentrated on young individuals, with 22- to 33-year-olds accounting for roughly two-thirds of the population (Laurila et al., 2012). The locations are collected by combining GPS and WiFi readings every 10 s when the phone is detected to be moving. The GPS records have been anonymised using k-anonymity by truncating the location GPS data (longitude,

Table 2: Representative studies based on different data sources. [a] LT = lateral, LD = longitudinal. [b] IT = individual trajectory, PN = places network. [c] Contains GPS log data that are collected using devices carried by individuals. Vehicle-attached GPS loggers are not included under this category. [d] The definition of longitudinal data form for geotagged tweets is stricter than the other sources. Besides long-term data collection period, it also should be free from geographical boundaries; otherwise, it is called lateral geotagged tweets.

Main data source	Data form[a]	Focus[b]	Authors	Main data set	Topic
Household travel survey	LT	IT	Pucher et al. (2011)	The 2001 and 2009 Nationa Household Travel Surveys	Population-level daily activity behavior change (walking and biking)
		PN	Liang et al. (2013)	One year of 46, 000 trips between 2017 zones within a county	Exponential law of intra-urban mobility
Mobile phone CDR	LD	IT	Phithakkitnukoon et al. (2012)	One year of anonymised call detail records of over one million mobile phone users in a country	Geo-social radius
			De Montjoye et al. (2013)	One year of 1.5 million users of a mobile phone operator in a country	The privacy bounds of human mobility
		PN	Gao et al. (2013)	A week of over 74 million phone call records of nearly 1 million mobile subscribers in a city	The clustering structures of spatial-interaction communities
			Iqbal et al. (2014)	One month of 6.9 million users of a mobile phone operator in a city	Development of origin-destination matrices
GPS log[c]	LD	IT	Rhee et al. (2011)	A period of 226 days of 101 individuals' GPS traces in five outdoor sites	Levy-walk nature of human mobility
			Sadilek and Krumm (2012)	Over one month of 307 individuals' GPS traces in a city's metropolitan area	Future location prediction
			De Domenico et al. (2013)	One year of 25 individuals' GPS traces in a country	Human mobility predictability and social interactions
			Zheng et al. (2008)	Ten months of 65 people's GPS traces in a country	Travel mode inference
Geotagged tweets [d]	LD	IT	Liao et al. (2019)	Over three years of 2,933 users' 652, 945 geotagged tweets globally	Sub-groups of individuals identified based on their mobility trajectory metrics
		PN	Hasnat and Hasan (2018a)	One month of 67,000 users' geotagged tweets in a state (raw data)*	Tourists identification and spatial patterns of their destinations
			Lenormand et al. (2015)	Nearly 3 years of 571,893 users' 21 million geotagged tweets globally	Quantification of city influence
	LT	IT	Jurdak et al. (2015)	Eight months of 0.16 million users' 7.8 million geotagged tweets	Universal laws of aggregate mobility behavior
			Zhang et al. (2017)	One year of 9,738 twitter users' geotagged tweets in a few districts	Longitudinal travel behavior features
			Hasan and Ukkusuri (2018)	Near one year of 3,256 users' Foursquare check-ins (via Twitter) in a city	Individuals' next activity prediction given the incomplete trajectory data
		PN	Gao et al. (2014)	One month of 110,868 users' 6.8 million geotagged tweets	Validation of OD trips mined from the geotagged tweets against the large-scale studies'
			Jin et al. (2014)	Near one month of check-ins that are observed in 19,170 venues within a city	Integration of check-ins into Gravity model

latitude) so that the resulting location rectangle, or anonymity-rectangle, contains enough inhabitants (Sweeney, 2002; Laurila et al., 2012). Therefore, the spatial resolution varies between locations depending on the population density.

Some participants have incomplete records due to technical issues. To guarantee data quality, three criteria are applied to the individual trajectories of GPS locations: (a) at least 30% of the trajectory days must have recorded locations (Do et al., 2015); (b) recording time must be at least 90 days (Do et al., 2015); (c) a fraction of missing location data must be less than 80% for each hourly time interval (Song et al., 2010b). After preprocessing, the applied MDC data set includes 61 participants satisfying the aforementioned criteria.

To extract the history of place visits, the location data are further processed. To identify stays (i.e., location data logged when users are engaging in activities), we consider locations at which the participant remains within a radius of 300 meters for at least 10 min as **stays** (Schulz et al., 2012; Jiang et al., 2013). The centroid of stay locations is set as the stay point. There are 37,626 identified stays in the GPS log data set. A recorded day is, therefore, defined as a day when there is at least one stay found.

After identifying stay points, the next step is to further identify stay-regions individually; the DBSCAN algorithm is applied as a density-based clustering method on the stay points (Ester et al., 1996; Schulz et al., 2012). The advantage of DBSCAN is that it can identify clusters of arbitrary shape (Ester et al., 1996). The distance threshold (eps), for merging locations into a stay, is set as 300 meters and a minimum number of stay-points as 1 (Schulz et al., 2012). With this process, the raw GPS log data can be divided into two parts, "stay" and "pass-by," where stay is defined as a location at which the user remains within a radius of 300 meters for at least 10 min, and the rest of the recorded data count as "pass-by." Due to signal loss and varying data quality of pass-bys, this study focuses its analysis on "stays." The stay records of the participants are shown in Figures 1-A, 1-C.

2.2. Geotagged tweets

2.2.1. Lateral data set (Twitter LT)

We collected tweets generated during a one-year period (5 December 2017 - 5 December 2018) within the geographical bounding box of Switzerland (The Tweepy project developers , 2017).

To construct the OD matrix, a valid trip is defined as the connection between two consecutive geotagged tweets generated by the same user satisfying the criteria: (a) the distance between these two geotagged locations is greater than 0, (b) the time interval is between 10 min and 4 h (Lee et al., 2015), and (c) the derived speed of travel is less than 885 kmh (domestic flight speed). Job-posting bots constitute a growing proportion of public geotags (Tasse et al., 2017). We further validate the collected users using BOT score (Davis et al., 2016), to make sure that all users for analysis have the BOT score < 1 implying that they are unlikely to be bot accounts. The same process has also been applied to the longitudinal data set described in the following section (2.2.2). After screening, the final subset includes 37,048 trips produced by 3,249 Twitter users over one year.

2.2.2. Longitudinal data set (Twitter LD)

To collect Twitter LD with sufficiently many geotagged tweets, we need to identify top geotag users in Switzerland. Therefore, we first analyzed tweets generated during a five-month period (5 December 2017 - 16 April 2018) within the geographical bounding box of Switzerland (subset of Twitter LT). Using this data set as a starting point, we further identified 462 non-commercial geotag users who geotagged their tweets most frequently. For this study, we extract those top users' historical tweets from their user timelines, without applying a spatial boundary limit. This method has a maximum number of tweets that can be collected from a specified user, producing varied time spans and varied tweet numbers, as not all users reached the 3200-tweet maximum. Besides time span and tweet number, the geotweeting frequency in general also varies greatly among users.

In order to compare geotagged tweets with GPS log data, we further apply the following rules to pre-process the data to ensure that the studied individuals reside within the Lausanne area in Switzerland and that they have a sufficient number of domestic geotagged tweets to reasonably capture their local activity trajectories: (a) the covered time span is greater than 1 year, and b) the most frequently visited region is in the geographical bounding box of Lausanne area. We further merged the locations that are generated within 10 minutes into one record. Unlike the high sampling frequency of GPS logs, users choose to geotag their tweets, so we only capture some of their visited locations. **Therefore, each generated location record is approximated as a stay**. After screening, we identify 61 users with 31,190 visited locations from their geotagged tweets. Their spatial and temporal distributions are shown in Figure 1-B, 1-D.

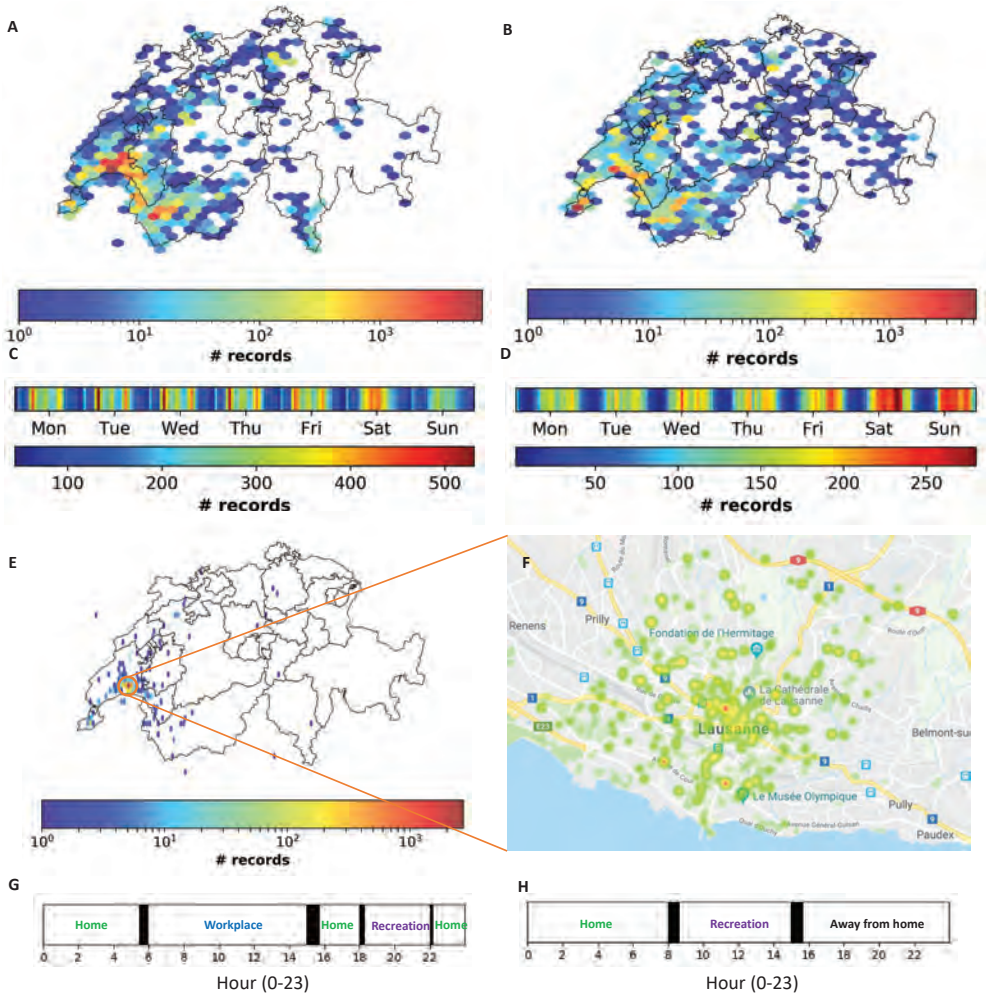

Figure 1: **Spatial and temporal distribution of three data sources**. (A) Spatial distribution of GPS log stays. (B) Spatial distribution of domestic geotagged tweets (individuals). (C) Temporal (hourly intervals) distribution of GPS log stays. (D) Temporal distribution of domestic geotagged tweets (individuals). (E) Locations in Switzerland in travel diaries. (F) Locations in Lausanne area in travel diaries. (G) Temporal distribution of a typical weekday (Thursday) in travel diaries. (H) Temporal distribution of a typical weekend day (Sunday) in travel diaries. The black lines in temporal distribution graphs, G and H, indicate the duration of moving from two adjacent stay locations.

2.3. Travel survey data

The travel survey data are extracted based on the results of the Mobility and transport microcensus 2015 from the Federal Statistical Office of Switzerland. It involves 57,090 participants who together reported 193,880 trips in their travel diaries (one day each). We further formulate two subsets to compare with the GPS log and the Twitter LD (Survey-Lausanne), and the Twitter LT (Survey) respectively. The GPS coordinates pairs are obtained by feeding the address (at the street level) of origin and destination into Google Places API (Google).

2.3.1. Lausanne residents (Survey-Lausanne)

We extract 573 Lausanne residents who have one-day detailed records of their trips. A stay in the Survey-Lausanne is defined as either the origin or the destination of a trip. There are in total 3,566 residents and 4,482 stays within Lausanne.

Compared to GPS log and geotagged tweets, only one day is captured for each participant, as shown by the spatial distribution (Figure 1E-F) and temporal distribution of two typical days (Figure 1G-H). Most stay locations are in Lausanne where the participants live. They represent more routine daily mobility patterns compared to GPS log or

geotagged tweets. Figures 1G-H show two distinct patterns for weekdays and weekends, with relatively accurate start time, end time, and activity purpose for each stay.

To explore international visits, we also extracted from the travel diary 112 residents in Lausanne who responded to the module of "traveling with overnight stays" in the travel survey. In that module, the participants were asked how many times they have taken a one-night-or-longer private or business trip in the last 4 months. The reported trips in this module are at least one night long, less frequent than once a week, not a daily routine, and they do not serve the daily living. This overnight trip module's results are used in Section 4.3.2.

2.3.2. All participants (Survey)

To compare the travel survey with Twitter LT, all the trips that both originate and end in Switzerland are extracted; the number of participants is 42,806 with 147,936 valid trips reported. A valid trip has complete information about trip distances (> 10 m) and purposes, etc.

2.4. Characteristics and limitation of data sets for comparison

Both Twitter LD and GPS log data are unevenly distributed in time (Figure 1D and 1C). Compared to GPS log data, geotagged tweets are more spread out over 24 hours and peak at noon and night. For GPS log data, the records peak during mornings and afternoons, implying that commuting is the main driver of mobility.

The statistics of five data sets are summarized in Table 3. GPS log, Twitter LD, and Survey-Lausanne are longitudinal data sets. Survey and Twitter LT are lateral data sets. Despite there being two forms of data sets, all the statistics are produced following two steps: (1) calculate each indicator based on the data points of each individual and (2) based on the indicators of covered individuals in the data set, calculate the mean value, 25% value, and 75% value.

Table 3: Statistics of the five data sets. LD indicates longitudinal, i.e. the data set is collected based on the individuals. LT indicates lateral, i.e. the data set is collected based on a certain area/geographical bounding box. N: the number of distinct regions (equivalent to stays for mobile phone GPS). n: the total number of visited locations. F_g: stays/visited regions per recorded day. q: the completeness of mobile phone GPS data, defined as the fraction of missing location data for hourly time intervals. [a] Only includes domestic locations.

Source	GPS log		Twitter LD		Twitter LD domestic[a]		Survey-Lausanne		Survey		Twitter LT	
Data form	LD		LD		LD		LT		LT		LT	
# individuals	61		61		61		573		42,806		3249	
Statistics	median	IQR	median	IQR	median	IQR	median	IQR	median	IQR	median	IQR
Time span	356	(267, 512)	960	(681, 1444)	-	-	388	-	388	-	365	-
Days covered	213	(105, 302)	241	(128, 396)	-	-	1	-	1	-	1	(1, 2)
N	160	(102, 223)	235	(140, 338)	96	(75, 176)	3	(2, 4)	2	(2, 3)	6	(4, 10)
n	565	(281, 716)	469	(200, 688)	236	(132, 452)	5	(3, 6)	5	(3, 6)	4	(2, 8)
F_g	2.5	(2.2, 2.8)	1.6	(1.3, 2.0)	1.4	(1.4, 1.8)	5	(3, 6)	5	(3, 6)	2	(2, 4)
q	0.7	(0.7, 0.8)	-	-	-	-	-	-	-	-	-	-

The GPS log data set has a one-year time span while Twitter LD covers roughly a three-year time period. Despite the difference in time span, the data sets cover a similar number of days because the sampling frequency is lower for geotagged tweets. The sets also show a comparable number of visited regions, as seen from both N and n in Table 3. Comparing the number of distinct stay regions with the number of all recorded regions, Twitter LD tends to present less regularly visited regions compared to GPS log (N/n = 14.5% vs 9.5%). GPS log data have much higher sampling frequency once a movement is detected compared to Twitter data; however, Twitter LD has F_g at the same level as GPS log, implying that participants stay in a limited number of regions every day. The fraction of hour-long intervals when a user's location is unknown to us is labeled q (Song et al., 2010b). The distribution of q is consistent with the previous findings (Song et al., 2010b) that people are stationary rather than moving around during most of the day, especially on weekdays. Survey-Lausanne, covering one day for each participant, has an average of 5.1 stays per person per day. It provides more complete information on stays than the other two sources: 2.7 for GPS log and 2.3 for Twitter LD.

Both lateral data sets, Survey and Twitter LT, cover a time span of one year. Twitter LT has more visited locations than Survey does; however, Survey has more visited locations per recorded day (F_g) compared to Twitter LT. Although Twitter LD comes from the top geotag users, it has a smaller F_g than Twitter LT. However, Twitter LT has a much lower proportion of days covered during its time span (32/365 = 8.8%) compared to Twitter LD (24.9%).

The main limitation is that the applied data sets do not cover the same period or the same group of individuals. Having the data covering the same period with the same group of individuals would be ideal and valuable, but we are constrained by limited access to and availability of multiple data sources. However, GDP, population growth, and travel demand are relatively stable in Switzerland 2010-2018 (The World Bank, 2019). According to the latest statistics in 2015, there was little change in the daily distance per person compared with 2010 (Federal Office for Spatial Development ARE, 2017). Therefore, we expect that the travel patterns in Switzerland are very similar between the late 2010 and 2018.

3. Methodology

3.1. Definitions

Depending on the research community and the data source, mobility terminology is often not consistent. In this paper, we define the following terms to better present the comparison methods and results.

- **Geotag** is the location's GPS coordinates pair attached to a tweet. The text part of a tweet is beyond the scope of this study.
- **Individual trajectory** refers to a series of visits during a certain period by an individual.
- **Places network** refers to a set of locations visited by multiple individuals.
- **Stay**, for GPS log, is defined as being at a location where a participant remains within a radius of 300 meters for at least 10 min. For travel survey, a stay is defined as either the origin or the destination, where various activities occur, of a reported trip.
- **Visit** is defined as one geotagged tweet or one stay as identified in the GPS log data or travel survey.
- **Trip** is defined as the connection between two consecutive visitsstays generated by the same individual. For geotagged tweets, we include only trips from two consecutive visits generated within a time interval smaller than 4 hours. "Trip" is also equivalent to "displacement" commonly defined in many studies.

3.2. Comparison framework

To systematically compare geotagged tweets with other data sources, we propose the comparison framework shown in Figure 2. There are five data sets in total: Twitter LD, described in Section 2.2.2; GPS log (Section 2.1); Survey-Lausanne (Section 2.3.1); Survey (Section 2.3.2); and Twitter LT (Section 2.2.1). The comparison is based on two dimensions of mobility, the individual trajectory and the places network.

The individual trajectory highlights two aspects: recurrent visits and non-recurrent visits. The more recurrent visits an individual makes, the more predictable that persons mobility. Therefore, we compare the mobility regularity based on different data sources by quantifying the degree to which the revealed individual mobility is regular. After showing the degree of mobility regularity, we explore the temporal profiles of visits. Bridging spatial and temporal dimensions, we further illustrate the diffusion process by comparing Twitter LD with GPS log. Trip distance, a key indicator of travel behavior, is compared across all the data sets. We also discuss the difference between two data forms: longitudinal and lateral.

Combining multiple individuals' geotagged trajectories yields a places network. We present a network-based comparison using node-level metrics and network-level metrics that originate in network science. In addition, we present the detected community structure of Twitter LT in comparison with Survey.

We apply mixed methods from various disciplines. To better present the comparison results, we use two main comparison sections, Individual trajectory (Section 4) and Places network (Section 5).). In each section, we first describe the methods and then present the results.

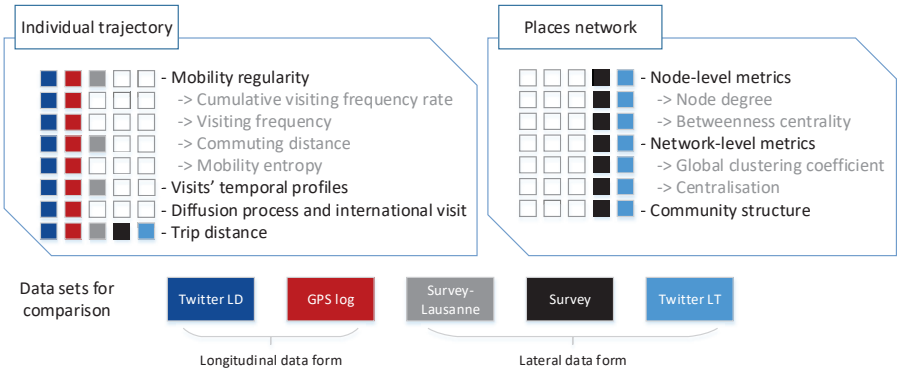

Figure 2: **Comparison framework for individual- and population-level mobility (individual trajectory and places network, respectively) from three different data sources**. Each box shows a list of the analyses and the corresponding data sets used.

4. Individual trajectory

4.1. Mobility regularity

4.1.1. Methods

In mobility studies, trips have purposes such as home-based-work trips and home-based-other trips. Unlike travel surveys, which record trip purposes, (Schneider et al., 2013; Çolak et al., 2015), GPS log and geotagged tweets can only derive this information indirectly. To extract visit (the destination of a trip) purposes, each stay region is assigned a tag as home (H), workplace (W), or other (O) (Schneider et al., 2013) based on the following method. Given that home and workplace represent the largest proportion of visited locations, we assume that the most visited location during weekends and 7pm-8am on weekdays is the home location. Once the home location is identified, it is eliminated and the most visited location during 8am-8pm on weekdays is identified as one's workplace.

The captured regularity of mobility can differ between GPS log, Twitter LD, and Survey-Lausanne. For purposes of comparison, we propose a set of indicators. Based on the assigned three types of locations (H, W, and O), the regularity is measured by their **cumulative frequency rate**, **visiting frequency rate distribution**, and **commuting distance**. The **mobility entropy** can be expressed as $S = -\sum_{j=1}^{n'} p(j) \log p(j)$ characterizing the heterogeneity of visitation patterns (Song et al., 2010b). The larger the mobility entropy, the greater the heterogeneity.

4.1.2. Results

Both the **cumulative frequency rate** and **mobility entropy** quantify the variation of visiting frequency across locations. Figure 3A shows the cumulative frequency rate vs. the most visited locations sorted by their visiting frequency. The faster the curve approaches 1, the higher the share of the visits concentrating on a small number of locations. Focusing on the top 20 locations, the univariate analysis of variance shows that the data source has a significant impact on the cumulative frequency rate ($F = 709$, $p < 0.001$). Post hoc tests show significant differences among three data sources (LSD corrected). Twitter LD tends to represent more non-recurrent visits compared to GPS log data. Mobility entropy characterizes the visit heterogeneity and provides a complementary view (see Figure 3B). By definition, mobility entropy is determined by two factors, the number of distinct visited locations and the frequency rate of each visited location. The higher the entropy, the harder it is to predict a subject's whereabouts. The mobility entropy of the three data sources are ordered from lowest to highest: GPS log, Twitter LD domestic, and Twitter LD. Twitter LD domestic has a smaller number of distinct visited locations compared with GPS log data (median value 96 vs. 160). However, Twitter LD domestic has higher entropy, suggesting the recorded visits are more decentralized compared with the GPS log data. When international visits are included (Twitter LD), the entropy increases significantly, not only because the distinct locations increase but also due to the non-recurrent visits being included, e.g., those locations that are only visited once.

There are significant differences between the three data sets regarding the visiting frequency for home and workplace (see Figure 4). GPS log has the highest visiting frequency for both home and workplace (Figure 4A-B), followed

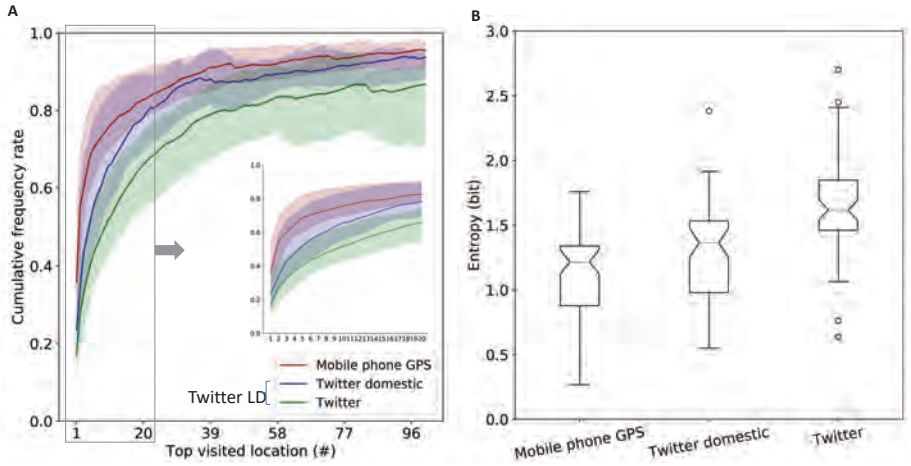

Figure 3: **Variation of visiting frequency between locations**. (A) Cumulative visiting frequency by the most visited locations ordered by their visiting frequency. The shaded area indicates the upper bound (75%) and lower bound (25%) of the cumulative frequency rate of visits. Inlay graph shows zoom on the first twenty top visited locations. (B) Mobility entropy of three data sets: GPS log, Twitter LD domestic and Twitter LD (all trips). The higher the entropy, the more randomness and the harder it is to predict one's whereabouts.

by Twitter LD domestic data and Twitter LD (include all international trips). One might question whether the estimated workplace and home are reliable, especially for Twitter data sets. We further explore the question by looking into the commuting distance, i.e., the direct distance between the estimated home and workplace (Figure 4C). The commuting distance observed from Survey-Lausanne can be regarded as the ground truth since such information is stated explicitly in the survey. Twitter LD yields a similar median value of commuting distance, but the variance is substantial. GPS log has a larger median value compared to Survey-Lausanne, while its variance is smaller than for Twitter LD. We conclude that the reliability of derived home and work locations from geotagged tweets is probably not as good as for the other data sources due to the lower observed visiting frequency (Figures 4A and 4B) and the much larger variations in commuting distance compared with the Survey (Figure 4C). The Mobile phone GPS participants are younger compared to the general population, and this could potentially explain the longer and wider range of commuting distances compared with Survey.

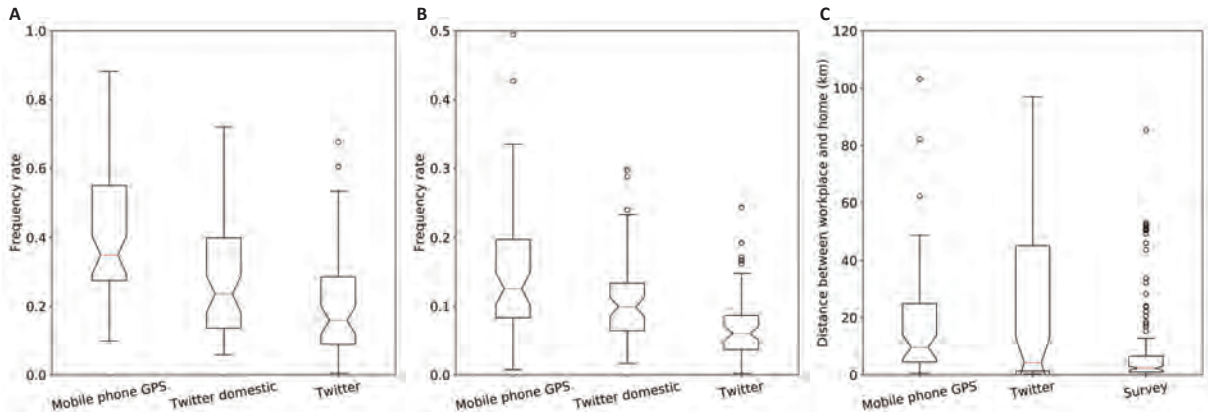

Figure 4: **Estimated home and workplace metrics**. (A) Visiting frequency of estimated home. (B) Visiting frequency of estimated workplace. (C) Commuting distance.

4.2. Temporal profile of visits

Twitter LD and GPS log are different in both spatial and temporal dimensions (see Figures 1A-D). GPS log has movements that peak during the morning and afternoon, implying that the captured mobility reflects more routine

activities. Top geotag users' geotags peak during noon and evening. Does the difference have an impact on the recorded visits and their corresponding activities? As described in Section 4.1.1, workplace and home locations are estimated based on the temporal rules. It is therefore important to explore whether those rules produce similar temporal profiles of workplace/home locations for these two data sets.

Figure 5 shows the temporal profile of home visits [1]. It turns out that the probability of being at home based on GPS log resembles the temporal distribution of Survey-Lausanne (see the red curve in Figure 5). Twitter LD users' probability of being at home is below 0.5 across weekdays and weekends (see the blue curve in Figure 5), reflecting the temporal distribution of their geotagging behavior (see Figure 1B, D). In other words, the home location derived from geotagged tweets is less reliable than those estimated with GPS log and Survey-Lausanne. Therefore, it is not surprising to see the commuting distances based on Twitter LD have a much greater variance than with Survey-Lausanne (Figure 4).

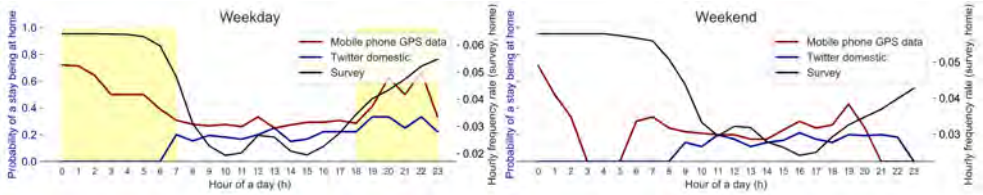

Figure 5: **Average probability of being at home on weekdays and weekends per hour**. For GPS log and Twitter LD domestic, a data point on the curve represents the average value across all participants the relative probability of being at home. For Survey-Lausanne (Survey in the legends), the curve represents the visiting frequency of home location distributing over the course of an average weekday or weekend. The yellow time interval is 7 pm - 8 am during weekdays, highlighting a presumably high probability of people being at home.

4.3. Diffusion process and international visit

4.3.1. Methods

The total radius of gyration r_g is defined as:

$$r_g = \sqrt{\frac{1}{N}\sum_{q=1}^{N} p_q \cdot (\mathbf{r}_q - \mathbf{r}_{cm})^2} \tag{1}$$

where $\mathbf{r}_q = [X1, X2]_q$ and the mass center of the visited locations $\mathbf{r}_{cm}$ is defined as

$$\mathbf{r}_{cm} = \left[\sum_{q=1}^{q=N}(X1 \cdot p_q)/\sum_{q=1}^{q=N} X1, \sum_{q=1}^{q=N}(X2 \cdot p_q)/\sum_{q=1}^{q=N} X2\right] \tag{2}$$

The **radius of gyration** (r_g) is widely used to indicate the distance that one covers on a regular basis. It combines the geographical distribution of the locations and their visiting frequency and has been widely applied to characterize mobility patterns (Gonzalez et al., 2008; Song et al., 2010b; Lu et al., 2013; Jurdak et al., 2015; Liao and Yeh, 2018). The time history of r_g, starting from one location, indicates how people diffuse during a particular time frame. The length of the time interval between two consecutive visits varies within and between individuals. For the sake of aggregation and comparison across data sources, data must be pre-processed to produce the time history of r_g during a given time period. A detailed description of the pre-processing method can be found in our previous study (Liao and Yeh, 2018).

r_g is less affected by those visits that happen infrequently. Therefore, movements that happen in a comparatively confined space will have a small r_g, even if a longer movement happens occasionally (Lu et al., 2013). Besides

[1] Zero means that for certain hours during an average weekday or weekend, there are no observed stays for any of the participants in that data set. For missing data in Mobile phone GPS data, the reason could be lost signals or turned-off GPS, etc. For Twitter LD domestic, it means that no users geotag their tweets during that hour.

recurrent visits, **international travel** captures less frequent trips that cover longer distances. How do the international visits affect the diffusion process? This can only be examined by Twitter LD that includes all trips. To make Twitter LD comparable with the international visits reported by the World Tourism Organization (UNWTO) (Federation, Swiss Tourism, 2018), continuous geotagged tweets within a foreign country are merged to count as one visit to that particular country.

4.3.2. Results

The diffusion process represents how the distance one regularly covers changes over time. The diffusion process estimated from three different sets is shown in Figure 6. Considering domestic visits only, Twitter LD domestic yields a similar diffusion process to GPS log; they both stabilize quickly and remain around a small value of r_g. When international visits are included, Twitter LD produces a significantly higher r_g. It suggests that international visits account for a substantial proportion of visits revealed by the geotagged tweets. Such an observation is consistent with the number of distinct locations (N) reported in Table 3, where the international visited locations account for 44% of total visited locations.

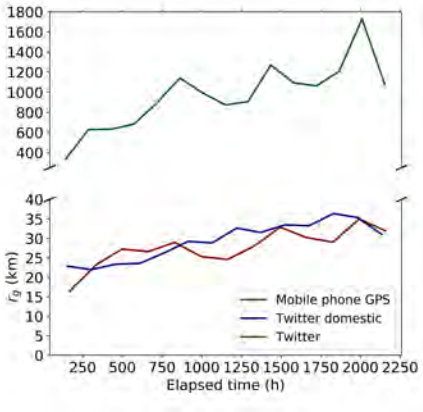

Figure 6: **Diffusion process represented by the time history of the radius of gyration**. The origin of the time history is set to the first time that the most visited location is observed. The time frame is 90 days and the number of data points is set to 50. The curves represent the mean value of all the individuals' data points in each time bin whose width is decided by the Freedman Diaconis Estimator that takes into account data variability and data size (Freedman and Diaconis, 1981).

The top visited countries for international travel are shown in Table 4. Tourist Arrivals 2015 only include tourism trips. The overnight trip module in the Survey-Laussane data set collected a maximum of 3 trips per person over the past 3 months. That overnight trip module is only implemented for one-third of the participants who live in Lausanne. It is worth noting that all three data sources are collected in a different manner with their own incompleteness. Twitter LD and Tourist Arrivals share 7 of their 10 most-visited international destinations. Twitter LD and international visits in Survey-Lausanne share 4 of their 10 most-visited international destinations.

4.4. Trip distance

Trip distance characterizes how far people travel. It can be measured by the **Haversine distance** (shortest distance) or the **actual trip distance**. The actual trip distance can only be obtained when the route and accurate time sequences are known. Among the data sources in this study, only travel survey data provide the actual trip distance estimated by individuals (but keep in mind that the accuracy of that self-estimated trip distance varies (Stopher and Greaves, 2007)). GPS log has detailed mobility trajectories with high temporal resolution, but the data set in this study, like many other GPS log data sets, has various quality issues, such as signal loss and the anonymization process when participants approach their homes. We therefore only estimate the Haversine distances from GPS log and Twitter and compare these with the Haversine distances calculated from the travel survey.

In the travel survey, the reported actual trip distances (y) in the travel survey and the Haversine distances (x) show a strong linear relationship ($y = 1.36 \times x + 0.29, R^2 = 0.96$): the actual travel distance is on average 36% greater than

Table 4: Top visited countries during international travel: Twitter LD, Tourist Arrivals (2015) (countries visited by residents of Switzerland) (Federation, Swiss Tourism, 2018), and Survey-Laussane.[a] International visits among reported overnight trips.

Rank	Twitter LD	Tourist Arrivals 2015	Survey-Laussane[a]
1	France	France	France
2	UK	USA	Italy
3	China	Spain	Spain
4	USA	China	Portugal
5	Turkey	Italy	Germany
6	Italy	Turkey	Netherlands
7	Spain	Germany	USA
8	Argentina	UK	Greece
9	Luxembourg	Mexico	Czech Republic
10	Brazil	Thailand	Thailand

the Haversine distance. The observed relationship is reflective of the regions spatial characteristics, e.g., the transport networks in Switzerland, and may not be generalized to the other regions. Given that only the Haversine distance is known for GPS log and Twitter, we thus use the Haversine trip distance across data sets as a proxy for the actual trip distance.

Figure 7 shows the difference between the trip distance captured by geotagged tweets versus travel survey, for domestic trips. For both the longitudinal and lateral data (Twitter LD and Twitter LT), trip distances are greater with geotagged tweets than with the travel survey (Survey-Lausanne and Survey), and more so for Twitter LT. Despite the infrequent geotweeting by the users included in Twitter LT, once they geotag their tweets, they tend to produce more geotagged tweets on those covered days (F_g = 2 vs 1.6). Overall, we conclude that the low geotweeting frequency, tendency to over-report non-recurrent trips (Figure 3B), and bursty nature of geotagging more prominent in Twitter LT result in significantly longer estimated trip distances than with Survey.

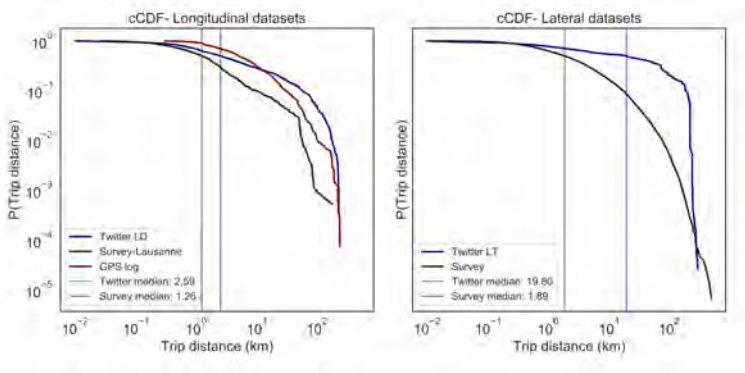

Figure 7: **Comparison of trip distance (Haversine distance, domestic only).** Travel survey and geotagged tweets have been processed to have similar resolution around 10 m. However, the minimum distance in the GPS log data set is 300 m, which has been determined by the filtering algorithm described in Section 2.1.

5. Places network

Trips form a directed network of physical locations that can be represented by an OD matrix. The network properties of trips emerge at the aggregate level of mobility patterns. In this subsection, the network properties are examined based on the node level and the network level (Morstatter et al., 2013), and we explore the community structure in the network, explained below.

5.1. Methods

The origins and the destinations of all trips are converted into a corresponding grid cell with the Military Grid Reference System (MGRS), using a precision level of 10 km (Langley, 1998). Switzerland is thus divided into 463

grid cells, and the center of each cell is determined by the center-most pair of GPS coordinates of all the observed visits in that cell.

The node-level comparison mainly focuses on calculating the measures based on nodes, i.e., grid cells. **Node degree** is a basic concept of centrality that counts the number of neighbors. Neighbors can be divided into two categories; in-neighbors that the node is connected to as their destination, and out-neighbors that the node is connected to as their origin. Two node degrees, in-degree and out-degree, are measured to quantify the degree centrality (see SI Equation **??**). Another centrality measure is **betweenness centrality**(Freeman, 1978), which identifies the degree of a node that bridges different location communities (see SI Equation **??**). A node with higher betweenness centrality would have more control over the network because more trips will pass through that node.

The network-level metrics quantify the OD matrix generated from different data sources. The **global clustering coefficient** (GCC) (see SI Equation **??**) measures the total number of closed triangles in a network to quantify the overall degree to which nodes are clustering (Barabási et al., 2016, p. 69). **Centralization** measures how equal the nodes are in a network regarding any given metric. It is defined as the difference between the value of the maximum-value node and all other node values compared to the theoretically maximum possible difference (Freeman, 1978). At the node level, we introduce three metrics, in-degree, out-degree, and betweenness centrality. At the network level, we select the corresponding centralization of the node's in-degree (InD), out-degree (OutD), and betweenness centrality (BC) to compare different data sources (see SI Equation **??** and **??**).

In network science, a community is a group of nodes that have a higher likelihood of connecting to each other than nodes from other communities (Barabási et al., 2016, p. 322). In other words, a community is a locally dense connected subgraph in a network. Aggregated individual movements in a certain area naturally create a complex network. Modularity quantifies whether a community partition is better than some other one (Barabási et al., 2016, p. 339) for detecting the community structure (see SI Equation **??**). Louvain Method—is used as a greedy optimization method—for community detection (Blondel et al., 2008).

5.2. Results

The OD matrix extracted from Twitter LT and Survey are shown in Figure 8A-B. With 463 grid cells (10 km $\times$ 10 km), a trip has 214,369 potential combinations of origin and destination. Not surprisingly, the OD matrix is sparse with a small proportion of non-zero OD pairs (3.9% for Survey vs. 2.6% for Twitter LT). For all the trips generated from Survey, within-cell trips account for 62% and between-cell trips 38%. By contrast, Twitter LT yields 32% within-cell trips, while between-cell trips account for a large proportion, 68%. The total average the trip number for each OD pair is 17.7 for Survey, the number of within-cell trips is 246.3, and the number of between-cell trips is 7.1. For Twitter LT, the average trip number per OD pair is 7.0, and the within- and between-cell trip numbers are 42.7 and 5.0 respectively. A larger number of between-cell trips from Twitter LT than Survey is consistent with the observation of longer trip distances found for the Twitter users in 7.

The node degree comparison between two data sets is illustrated in Figure 8C-D. Despite the discrepancy in the trip distance between two data sets, the correlation between the two data sets is strong for both in- and out-degree, implying that Twitter LT reveals similar zonal travel demand to Survey.

Looking at the nodes with non-zero betweenness centrality, the overall correlation between two data sets is moderate (Pearson correlation coefficient = 0.5, $p < 0.001$) as shown in Figure 9A. The locations of high betweenness centrality often occur on the shortest paths between two other areas. Those bridge locations identified by Survey are likely to be ones identified by Twitter LT, but not vice versa. It is particularly interesting to look at those cells that are identified as "bridge" (high betweenness centrality) by only one data set but not by the other one (Figure 9B). The important hubs that are only highlighted by Survey are close to transport facilities locations. However, the hubs that are solely identified by Twitter LT are related to places of leisure activities, such as mountains or parks.

The values for the four centralization metrics stabilize when the observation time period increases (Figure 10A), implying that the network structure can change initially as we collect more data but becomes robust after 4-6 months. Survey has greater centralization of the global clustering coefficient than geotagged tweets (Figure 10B), consistent with the observations in Figure 8A-B. As for the centralization of in- and out-degree, the difference is small. Twitter LT is larger than Survey for the centralization of betweenness centrality, indicating the larger variation between the grid cells regarding the hub importance.

Figure 10C-D shows the community structure of the OD matrix from Survey and Twitter LT. Their global modularities differ: the modularity is greater for the one-day travel diary than for geotagged tweets. The higher modularity

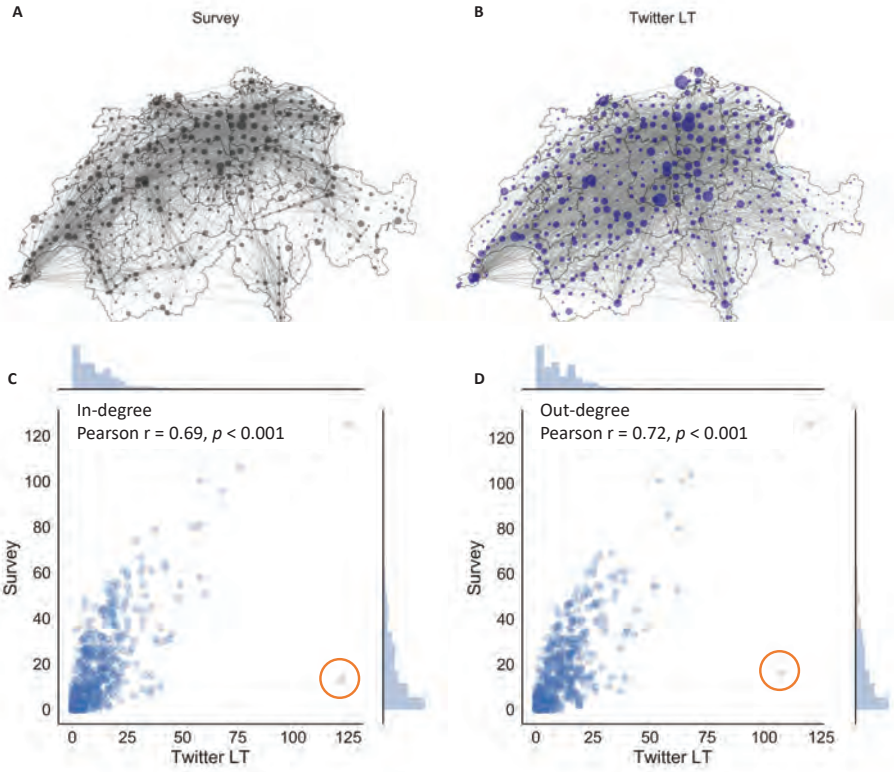

Figure 8: **Node-level comparison**. (A) & (B) OD matrix comparison The node size is proportional to the node degree (normalized by the sum of total degrees for each data set). (c) & (D) In-degree and out-degree comparisons of Twitter LT vs. Survey. Each data point represents one grid cell's degrees from the two data sets. The orange circle highlights an outlier that is a tourist attraction area located at the geographical center of Switzerland.

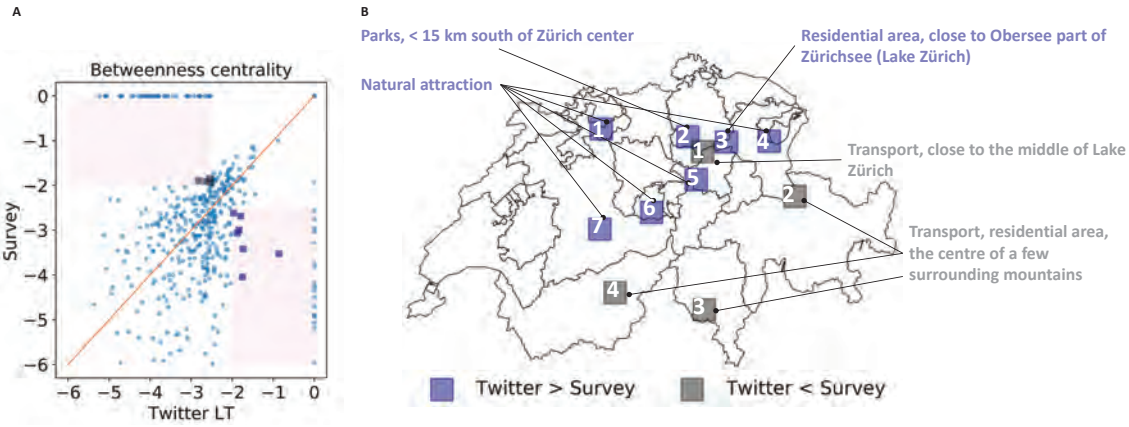

Figure 9: **Betweenness centrality**. (A) Log value of the 463 cells' betweenness centrality for Survey and Twitter LT. The closer the point to the orange diagonal line, the closer the two data sets on grid cells' centrality value. (B) Grid cells with large differences between the two data sets. The cells that have large discrepancy between two data sets (Twitter LT and Survey) are highlighted with rectangles on the map, with a short description of the key features within that 10km × 10km cell.

a data source produces, the more clustered the sub-graph structure in its network. And the detected communities are more constrained by the geographical distances between the nodes for Survey compared with Twitter LT. This is because the one-day travel diary mainly captures routine mobility, whereas the communities in geotagged tweets tend

to represent non-recurrent traveling preferences.

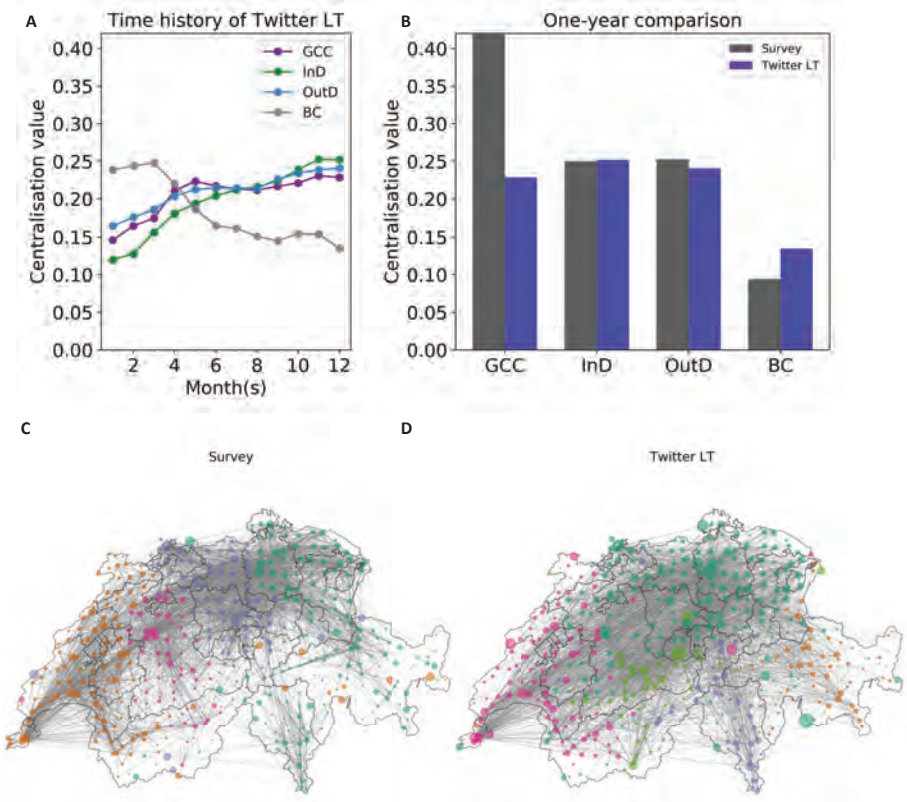

Figure 10: **Network-level comparison**. (A) Centralization values of Twitter LT stabilize over time. Centralization of the global clustering coefficient (GCC), in-degree (InD), out-degree (OutD), and betweenness centrality (BC). The centralization value is between 0 and 1; the greater the centralization, the greater the variation between nodes in the given network. (B) Centralization values based on the two data sets over a one-year period. (C) & (D) Community structure based on the OD matrix from Survey (C) and Twitter LT (D). Global modularity: Twitter LT = 0.33, Survey = 0.48.

6. Discussion

We compare data from a household travel survey, GPS log data, and geotagged tweets as proxies for mobility and propose a framework that considers both longitudinal and lateral data forms. Without perfect measurements, our characterization of mobility is limited by the data sources we have. As a result, mobility measurement has been divided into different categories: domestic trips vs. international trips, day trips vs. trips with overnight stays, etc., as defined in travel surveys. The taxonomy varies and the corresponding categories overlap. For instance, a daily trip is not necessarily domestic if one lives close to the border. Our study adopts a more robust taxonomy of mobility for comparison: individual trajectories (Section 4) and places networks (Section 5) from which we draw the following key insights and compare them with the literature.

6.1. More non-recurrent visits and fewer visits to home/workplace

One of the main criticisms of geotagged social media data is the behavior distortion of geotagged tweets, with users deciding when and where to geotag. We show that geotagged tweets represent more non-recurrent (less routine) visits compared with travel surveys and GPS logs (see Table 3). This is consistent with a study on Foursquare geotagging behavior (Lindqvist et al., 2011). A geotag usage survey based on 400 US residents shows that 70% of their geotags happen in places that people go infrequently (Tasse et al., 2017). We found the average visiting

frequency per location (n/N) is more than once, e.g., 2.0 for Twitter LD, 2.5 for Twitter LD domestic, and 1.5 for Twitter LT. For top geotag Twitter users, the cumulative frequency of visited locations shows the returning nature of mobility (see Figure 3A), i.e., people visit a limited number of locations repeatedly. Our previous study (Liao et al., 2019) also shows that people geotweet from locations that they visit frequently as well as new locations. Moreover, there is a large individual difference regarding the mobility regularity observed by their geotagged activities. That is to say, despite more non-recurrent visits in the geotagged tweets, the top geotag Twitter users' routine activities are preserved reasonably well in the data.

Tasse et al. (2017) suggest that most geotag users geotag their tweets within an hour of arrival (if at all), thus geotagging may be a timely indicator of the start time of the activity. However, our study finds that the temporal distribution of the reported activities is not similar to that shown in travel survey or GPS log data. We find that Twitter users tend to geotweet midday and in the evening. Such a temporal distribution resembles a leisure activity pattern according to the travel survey statistics (Federal Office for Spatial Development ARE, 2017).A study that found Twitter data to be a good proxy for local commuting patterns only used simple heuristics based on frequency counts to estimate home and workplace (McNeill et al., 2017) and did not integrate the temporal dimension into identifying home and workplace. Using a more detailed temporal distribution method, our study finds that the estimated home and work locations are not as reliable as those estimated with the two other data sources. Even with sophisticated methods, e.g., hierarchical classification (Mahmud et al., 2014), the prediction accuracy is low in general (¡ 70%) depending on the spatial resolution. This study identifies two underlying reasons: (1) locations are skewed in the time dimension, (2) home and workplace account for a smaller proportion of recorded mobility compared with GPS log data and travel survey. As a result, the estimated commuting distance using geotagged tweets has a long tail, suggesting the large uncertainty and therefore the unreliability of the estimates.

6.2. Less accurate trip distance with lateral data despite larger sample size than longitudinal data

For geotagged tweets, international visits account for a 43% on average, though the proportions can vary depending on sub-population types (Liao et al., 2019). That is in line with a study showing geotag users tend to geotag locations that are not within their neighborhood; and the geotagged locations concentrate substantially at locations farther away than the daily mobility area (Tasse et al., 2017). We find that these international visits captured through geotagged tweets are consistent with those destinations that are popular with Swiss tourists, suggesting the reliability of using geotagged social media data for modeling international visits (e.g. Chua et al., 2016; Hasnat and Hasan, 2018b). This is a key strength of geotagged tweets, given that international longitudinal mobility, tracing same individuals consistently, is missing from most commonly used mobility data sources, traditional household travel surveys, CDR, and GPS log data.

Trip distance, i.e., the displacement length in many studies, is often used to prove the usefulness of geotagged social media data (e.g. Jurdak et al., 2015; McNeill et al., 2017). We find that geotagged tweets reveal longer trip distances than the other two data sources; such a difference is amplified in the lateral data form given the bursty geotweeting behavior of the majority of Twitter users. Our study illustrates that the long-term observation of longitudinal geotagged tweets by top users compensates for the time sparsity and helps to recreate a more complete image of individual mobility, and, therefore, is more reliable for the travel distance estimation than the lateral data set. More specifically, lateral geotagged tweets are dominated by non-recurrent and long-distance trips empirically explained by the bursty effect of less-frequent geotag users, i.e., they tend to geotag more frequently during a shorter time frame compared to those top geotag users in the longitudinal data set. The derived trip distance becomes unreliable despite a more than 50-fold greater sample size. This has not been discussed in the literature and is one of the important observations of this paper.

6.3. Spatial distribution, zonal travel demand, and community structure

Places networks are constructed based on the population movements that emerge from individual mobility. Compared with the other data sources, we find that the individual trajectory derived from geotagged tweets has more non-recurrent visits with longer distances, and less regular temporal profiles, suggesting recreational activity. These characteristics are reflected in the OD matrix results: Twitter-based OD matrixes point more towards mountain areas and recreational attractions. Previous studies have used social media data to model zonal demand (Lee et al., 2015; Jin et al., 2014). However, this study highlights the importance of context when geotagged tweets are used for exploring zonal demand, which captures leisure traveling patterns reasonably well but may overestimate the demand level.

We find that geotagged tweets generate a community structure with lower modularity than the one-day travel diary. In addition, the one-day travel diary produces the community structure that visually represents the geographical closeness, i.e., the places that are close to each other sharing the same area tend to cluster together. Geotagged communities, however, represent more non-recurrent traveling preferences rather than geographical constraints. One-day travel diaries formulate a daily mobility network driven by the purposes of daily routines whereas geotagged tweets create a network that is both routine but also socially driven (Tasse et al., 2017), e.g., "showing I was at a cool place" and "keeping family/friends updated."

Despite some differences on community structure and trip spatial distribution, geotagged tweets and the travel survey agree well on most node-level metrics and network-level metrics. Geotagged tweets and the one-day travel diary have strongly correlated zonal demand (see Figure 8). Good agreement is also found on the degree to which an area is important in connecting other areas (see Figure 9). Areas of high betweenness importance as shown by the travel survey are most likely to be recognized as having the same importance with geotagged tweets, but not vice versa. This suggests that geotagged tweets represent a broader zonal betweenness importance. The results of the places network comparison suggest a promising potential for using geotagged tweets to understand places dynamics. Besides directly using the OD matrix for travel demand modeling (Gao et al., 2014; Lee et al., 2015), broader topics have been studied using geotagged places networks, including urban communities dynamics (Lenormand and Ramasco, 2016), urban congestion detection (Wang et al., 2014), and strategies of locating new business (Tasse and Hong, 2014).

6.4. The implications of population bias and behavior distortion are mixed

A main criticism of geotagged social media data is that social media users are not representative of the overall population. Studies typically select top geotag users so that the number of geotagged tweets is large enough for reasonable analysis (e.g. Jurdak et al., 2015; Zhang et al., 2017; Hasan and Ukkusuri, 2018), despite the fact that top geotag users show large differences in socio-demographic characteristics compared to both the overall population and the average Twitter user (Tasse et al., 2017). Despite population biases, many studies have found reasonable agreement on mobility patterns between geotagged tweets and other data sources (e.g. Jurdak et al., 2015; McNeill et al., 2017).

Our findings in this study are mixed. We find population bias and behavior distortion contributing to different mobility estimates from those found with the travel survey and GPS log data. On the other hand, we also find robust results despite these differences.

As discussed in Section 6.2, we find that the observed trip distance derived from top geotag users agrees well with the travel survey and GPS log data, suggesting long-term collection can compensate for the time sparsity. The evidence in this study also implies that geotagged tweets do not come from a limited set of locations or only the new locations that have rarely been visited. Compared with GPS log data collected during one year, domestic geotagged tweets (3.2 years) from user timelines cover a comparable number of distinct visited locations. Despite less reliable routine mobility, the fundamental laws of mobility including returning effect, diffusive nature, and trip distance are almost identical in geotagged tweets and the travel survey and GPS log data. Moreover, heterogeneity in mobility patterns can also be identified in geotagged tweets (Liao and Yeh, 2018; Liao et al., 2019).

Behavior distortion (low and irregular sampling) is influenced by social factors. For example, one study finds that Twitter users tend to geotag uncommon places, almost as a kind of postcard sent to their followers (Tasse et al., 2017). Understanding the underlying motivations is crucial, and any application of geotagged social media data requires more awareness of why people geotag. Some of the findings in this study can be explained by the behavioral motivations identified in the previous studies, such as the more non-recurrent mobility (see Figure 3A-B and Figure 4) and longer trip distance (see Figure 7) than the other data sources. The behavioral distortion also reflects the observed difference in places networks as discussed in Section 6.3.

To better understand the socio-psychological factors of geotagging, text-mining of geotagged tweets can provide rich contextual information, e.g., tourist travel or leisure travel. Location type, activity, and time can be obtained by applying text-mining of Twitter data to complement travel diary surveys (Maghrebi et al., 2015). One study infers the city-level locations dynamics by only using the text content (Cheng et al., 2010). Such an approach can also provide further validation of our findings. For example, we find that the temporal distribution of geotagged tweets shows the influence of leisure activities. With text mining, we can further verify leisure-related geotagged tweets among the overall geotagged tweets.

7. Conclusions

This study presents a systematic comparison involving three data sources that are widely used in mobility research: geotagged tweets, traditional household travel surveys, and GPS log data. The major contributions of this paper are:

1) **A systematic and comprehensive comparison framework**. Our paper departs from previous studies in that we do not assume a single data source as the ground truth; rather, we explore how our understanding of mobility changes when different data sources are used. When aggregated, individual trajectories form places networks, and in theory these two perspectives should be consistent. However, in reality, these two essential perspectives of human mobility have been studied separately using different data sources to answer different research questions. The proposed comparison framework integrating these two perspectives provides a holistic view on data sources, study designs, and research questions. The outcomes of such a framework benefit a broad range of topics in travel behavior studies.

2) **New insights on the strengths and weaknesses of different data sources in capturing mobility**. A travel survey is a lateral data form that measures places networks and population mobility at the aggregate level, whereas GPS log data are a longitudinal data form that measures individual trajectories. Yet both data sources fail to capture the international travels of the same individuals. The key strength of geotagged tweets, especially the ones from top geotag users, is that, due to the long-term collection that compensates for the time sparsity, geotagged tweets are representative of individual trajectories including international travels. Geotagged tweets from top geotag Twitter users capture reasonably well the fundamental laws of individual mobility, including returning effect, diffusive nature, and trip distance distribution. However, the estimated home and work locations from top geotag Twitter users may not be reliable due to data sparsity and behavior distortion. In addition, due to the low number of days covered and the bursty tweeting behavior of most users, and the tendency to report social activities and locations, geotagged tweets poorly represent the mobility indicators at the population level such as trip distance. We also find strong agreement on the network structure of visits and zonal demand between Twitter and the travel survey, despite higher preferences for movements driven by social activities in geotagged tweets. Geotagged tweets are the only data source that allows us to examine individual trajectories (domestic and international) and places networks simultaneously, but the validity of the results and their implications need further validations and careful interpretations.

This study has three major limitations for potential future work. The first limitation is the geographical scale. The GPS log data were collected from a group of residents of Switzerland, and our study therefore focuses on Switzerland, which has a lower rate of Twitter users compared with some other countries (Hawelka et al., 2014). Future work includes a cross-regional study to apply the proposed comparison framework to other areas. The second limitation is the lack of text-based analysis to complement the collected geotagged tweets. As we discussed, semantic information can be a useful resource for understanding the underlying motivations behind the geotagged activities, but such research is outside the scope of the current study. A potential direction is combining textual context and social networks, etc., with geotags to further tailor the geotagged tweets that work for specific research questions. The third limitation is the inconsistent time periods and individuals in the data sets used in this study. More careful design in future studies to increase the number of participants across multiple data sources with consistent time frames and participants can have valuable contributions.

Acknowledgment

The authors acknowledge the financial support of the Swedish Research Council for Sustainable Development (Formas, project number 2016-01326).

References

The Tweepy project developers , 2017. Tweepy: v3.5.0.

Alessandretti, L., Sapiezynski, P., Sekara, V., Lehmann, S., Baronchelli, A., 2018. Evidence for a conserved quantity in human mobility. Nature Human Behaviour , 1.

Barabási, A.L., et al., 2016. Network science. Cambridge university press.

Barbosa, H., Barthelemy, M., Ghoshal, G., James, C.R., Lenormand, M., Louail, T., Menezes, R., Ramasco, J.J., Simini, F., Tomasini, M., 2018. Human mobility: Models and applications. Physics Reports .

Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment 2008, P10008.

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. Transportation research part C: emerging technologies 68, 285–299.
Chen, G., Hoteit, S., Viana, A.C., Fiore, M., Sarraute, C., 2018. Enriching sparse mobility information in call detail records. Computer Communications 122, 44–58.
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, ACM. pp. 759–768.
Chua, A., Servillo, L., Marcheggiani, E., Moere, A.V., 2016. Mapping cilento: Using geotagged social media data to characterize tourist flows in southern italy. Tourism Management 57, 295–310.
Çolak, S., Alexander, L.P., Alvim, B.G., Mehndiratta, S.R., González, M.C., 2015. Analyzing cell phone location data for urban travel: current methods, limitations, and opportunities. Transportation Research Record: Journal of the Transportation Research Board , 126–135.
Davis, C.A., Varol, O., Ferrara, E., Flammini, A., Menczer, F., 2016. Botornot: A system to evaluate social bots, in: Proceedings of the 25th International Conference Companion on World Wide Web, International World Wide Web Conferences Steering Committee. pp. 273–274.
De Domenico, M., Lima, A., Musolesi, M., 2013. Interdependence and predictability of human mobility and social interactions. Pervasive and Mobile Computing 9, 798–807.
De Montjoye, Y.A., Hidalgo, C.A., Verleysen, M., Blondel, V.D., 2013. Unique in the crowd: The privacy bounds of human mobility. Scientific reports 3, 1376.
Do, T.M.T., Dousse, O., Miettinen, M., Gatica-Perez, D., 2015. A probabilistic kernel method for human mobility prediction with smartphones. Pervasive and Mobile Computing 20, 13–28.
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, pp. 226–231.
Etter, V., Kafsi, M., Kazemi, E., 2012. Been there, done that: What your mobility traces reveal about your behavior, in: Mobile Data Challenge by Nokia Workshop, in conjunction with Int. Conf. on Pervasive Computing.
Federal Office for Spatial Development ARE, 2017. Population's transport behaviour 2015. Technical Report.
Federation, Swiss Tourism, 2018. Swiss tourism in figures 2017: Structure and industry data. Technical Report.
Freedman, D., Diaconis, P., 1981. On the histogram as a density estimator: L 2 theory. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 57, 453–476.
Freeman, L.C., 1978. Centrality in social networks conceptual clarification. Social networks 1, 215–239.
Gao, S., Liu, Y., Wang, Y., Ma, X., 2013. Discovering spatial interaction communities from mobile phone d ata. Transactions in GIS 17, 463–481.
Gao, S., Yang, J.A., Yan, B., Hu, Y., Janowicz, K., McKenzie, G., 2014. Detecting origin-destination mobility flows from geotagged tweets in greater los angeles area, in: Eighth International Conference on Geographic Information Science (GIScience'14), Citeseer.
Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L., 2008. Understanding individual human mobility patterns. Nature 453, 779–782.
Google, . Place search — places api — google developers.
Hasan, S., Ukkusuri, S.V., 2018. Reconstructing activity location sequences from incomplete check-in data: A semi-markov continuous-time bayesian network model. IEEE Transactions on Intelligent Transportation Systems 19, 687–698.
Hasnat, M.M., Hasan, S., 2018a. Identifying tourists and analyzing spatial patterns of their destinations from location-based social media data. Transportation Research Part C: Emerging Technologies 96, 38–54.
Hasnat, M.M., Hasan, S., 2018b. Understanding tourist destination choices from geo-tagged tweets, in: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), IEEE. pp. 3391–3396.
Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., Ratti, C., 2014. Geo-located twitter as proxy for global mobility patterns. Cartography and Geographic Information Science 41, 260–271.
Iqbal, M.S., Choudhury, C.F., Wang, P., González, M.C., 2014. Development of origin–destination matrices using mobile phone call data. Transportation Research Part C: Emerging Technologies 40, 63–74.
Janzen, M., Müller, K., Axhausen, K.W., 2017. Population synthesis for long-distance travel de-mand simulations using mobile phone data, in: 6th Symposium of the European Association for Research in Transportation (hEART 2017).
Jiang, S., Fiore, G.A., Yang, Y., Ferreira Jr, J., Frazzoli, E., González, M.C., 2013. A review of urban computing for mobile phone traces: current methods, challenges and opportunities, in: Proceedings of the 2nd ACM SIGKDD international workshop on Urban Computing, ACM. p. 2.
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. Transportation Research Record: Journal of the Transportation Research Board , 72–82.
Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., Newth, D., 2015. Understanding human mobility from twitter. PloS one 10, e0131469.
Langley, R.B., 1998. The utm grid system. GPS world 9, 46–50.
Laurila, J.K., Gatica-Perez, D., Aad, I., Bornet, O., Do, T.M.T., Dousse, O., Eberle, J., Miettinen, M., et al., 2012. The mobile data challenge: Big data for mobile computing research, in: Pervasive Computing.
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.
Lee, J.H., Gao, S., Goulias, K.G., 2015. Can twitter data be used to validate travel demand models, in: 14th International Conference on Travel Behaviour Research.
Lenormand, M., Gonçalves, B., Tugores, A., Ramasco, J.J., 2015. Human diffusion and city influence. Journal of The Royal Society Interface 12, 20150473.
Lenormand, M., Picornell, M., Cantú-Ros, O.G., Tugores, A., Louail, T., Herranz, R., Barthelemy, M., Frias-Martinez, E., Ramasco, J.J., 2014. Cross-checking different sources of mobility information. PLoS One 9, e105184.
Lenormand, M., Ramasco, J.J., 2016. Towards a better understanding of cities using mobility data. Built Environment 42, 356–364.
Liang, X., Zhao, J., Dong, L., Xu, K., 2013. Unraveling the origin of exponential law in intra-urban human mobility. Scientific reports 3, 2983.
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. pp. 2068–2073.

Liao, Y., Yeh, S., Jeuken, G., 2019. From individual to collective behaviours: Exploring population heterogeneity of human mobility based on social media data (under review). EPJ Data Science .

Lindqvist, J., Cranshaw, J., Wiese, J., Hong, J., Zimmerman, J., 2011. I'm the mayor of my house: examining why people use foursquare-a social-driven location sharing application, in: Proceedings of the SIGCHI conference on human factors in computing systems, ACM. pp. 2409–2418.

Lu, X., Wetter, E., Bharti, N., Tatem, A.J., Bengtsson, L., 2013. Approaching the limit of predictability in human mobility. Scientific reports 3, srep02923.

Maghrebi, M., Abbasi, A., Rashidi, T.H., Waller, S.T., 2015. Complementing travel diary surveys with twitter data: application of text mining techniques on activity location, type and time, in: Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on, IEEE. pp. 208–213.

Mahmud, J., Nichols, J., Drews, C., 2014. Home location identification of twitter users. ACM Transactions on Intelligent Systems and Technology (TIST) 5, 47.

McNeill, G., Bright, J., Hale, S.A., 2017. Estimating local commuting patterns from geolocated twitter data. EPJ Data Science 6, 24.

Mislove, A., Lehmann, S., Ahn, Y.Y., Onnela, J.P., Rosenquist, J.N., 2011. Understanding the demographics of twitter users. ICWSM 11, 25.

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: ICWSM.

Phithakkitnukoon, S., Smoreda, Z., Olivier, P., 2012. Socio-geography of human mobility: A study using longitudinal mobile phone data. PloS one 7, e39253.

Pianese, F., An, X., Kawsar, F., Ishizuka, H., 2013. Discovering and predicting user routines by differential analysis of social network traces, in: World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2013 IEEE 14th International Symposium and Workshops on a, IEEE. pp. 1–9.

Pucher, J., Buehler, R., Merom, D., Bauman, A., 2011. Walking and cycling in the united states, 2001–2009: evidence from the national household travel surveys. American journal of public health 101, S310–S317.

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. Transportation Research Part C: Emerging Technologies 75, 197–211.

Rhee, I., Shin, M., Hong, S., Lee, K., Kim, S.J., Chong, S., 2011. On the levy-walk nature of human mobility. IEEE/ACM transactions on networking (TON) 19, 630–643.

Ribeiro, A.I.J.T., Silva, T.H., Duarte-Figueiredo, F., Loureiro, A.A., 2014. Studying traffic conditions by analyzing foursquare and instagram data, in: Proceedings of the 11th ACM symposium on Performance evaluation of wireless ad hoc, sensor, & ubiquitous networks, ACM. pp. 17–24.

Ruths, D., Pfeffer, J., 2014. Social media for large studies of behavior. Science 346, 1063–1064.

Sadilek, A., Krumm, J., 2012. Far out: Predicting long-term human mobility., in: Twenty-sixth AAAI Conference on Artificial Intelligence.

Schneider, C.M., Belik, V., Couronné, T., Smoreda, Z., González, M.C., 2013. Unravelling daily human mobility motifs. Journal of The Royal Society Interface 10, 20130246.

Schulz, D., Bothe, S., Körner, C., 2012. Human mobility from gsm data-a valid alternative to gps, in: Mobile data challenge 2012 workshop, June, pp. 18–19.

Song, C., Koren, T., Wang, P., Barabási, A.L., 2010a. Modelling the scaling properties of human mobility. Nature Physics 6, 818.

Song, C., Qu, Z., Blumm, N., Barabási, A.L., 2010b. Limits of predictability in human mobility. Science 327, 1018–1021.

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.

Stopher, P.R., Greaves, S.P., 2007. Household travel surveys: Where are we going? Transportation Research Part A: Policy and Practice 41, 367–381.

Sweeney, L., 2002. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10, 557–570.

Tasse, D., Hong, J.I., 2014. Using social media data to understand cities, in: Proceedings of NSF Workshop on Big Data and Urban Informatics.

Tasse, D., Liu, Z., Sciuto, A., Hong, J.I., 2017. State of the geotags: Motivations and recent changes., in: ICWSM, pp. 250–259.

The World Bank, 2019. Switzerland. Data retrieved from World Country View, `http://data.worldbank.org/icountry/switzerland?view=chart`.

Toch, E., Lerner, B., Ben-Zion, E., Ben-Gal, I., 2018. Analyzing large-scale human mobility data: a survey of machine learning methods and applications. Knowledge and Information Systems , 1–23.

Wang, J., Wei, D., He, K., Gong, H., Wang, P., 2014. Encapsulating urban traffic rhythms into road networks. Scientific reports 4, 4141.

Wang, Z., He, S.Y., Leung, Y., 2018. Applying mobile phone data to travel behaviour research: A literature review. Travel Behaviour and Society 11, 141–155.

Wesolowski, A., Eagle, N., Noor, A.M., Snow, R.W., Buckee, C.O., 2013. The impact of biases in mobile phone ownership on estimates of human mobility. Journal of the Royal Society Interface 10, 20120986.

Wesolowski, A., Eagle, N., Tatem, A.J., Smith, D.L., Noor, A.M., Snow, R.W., Buckee, C.O., 2012. Quantifying the impact of human mobility on malaria. Science 338, 267–270.

Yue, Y., Lan, T., Yeh, A.G., Li, Q.Q., 2014. Zooming into individuals to understand the collective: A review of trajectory-based travel behaviour studies. Travel Behaviour and Society 1, 69–78.

Zhang, Z., He, Q., Zhu, S., 2017. Potentials of using social media to infer the longitudinal travel behavior: A sequential model-based clustering method. Transportation Research Part C: Emerging Technologies 85, 396–414.

Zheng, Y., Li, Q., Chen, Y., Xie, X., Ma, W.Y., 2008. Understanding mobility based on gps data, in: Proceedings of the 10th international conference on Ubiquitous computing, ACM. pp. 312–321.

Using geotagged tweets to assess human mobility: a comparison with travel survey data and GPS log data
Supplementary Information

Yuan Liao* and Sonia Yeh
Department of Space, Earth and Environment
Chalmers University of Technology, Gothenburg 41296, Sweden

July 19, 2019

Network metrics for places network comparison

Switzerland is divided into 463 cells (i) using MGRS, therefore, the generated network based on Survey or Twitter LT has 463 nodes ($v_i, i = 1, ..., N, N = 463$). Degree is a key property of a node, representing the number of links/trips it has to other nodes. Degree has the following characteristic:

$$deg(v_i) = deg^+(v_i) + deg^-(v_i) \tag{1}$$

where $deg^+(v_i)$ is the number of trips that are connected to node i regarding it as the origin, and the number of trips that are connected to node i regarding it $deg^-(v_i)$ as the destination.

A shortest path between node j and node k is the path with the fewest number of links/trips. Betweenness centrality (BC) is defined as:

$$BC(v_i) = \sum_{j<k} g_{jk}(v_i)/g_{jk}, (i, j, k = 1, 2, ..., N) \tag{2}$$

where g_{jk} stands for the number of shortest paths connecting node j and node k, and $g_{jk}(v_i)$ stands for the number of those paths that node i is on.

Besides the above node-level metrics, we apply a series of network-level metrics to further quantify the difference between Survey and Twitter LT. Global clustering coefficient (GCC) is defined as:

$$GCC = \frac{\text{(Number of Triangles)} \times 3}{\text{Number of Connected Triples of Nodes}} \tag{3}$$

where a triplet is an ordered set of three nodes such that A connects to B and B connects to C. And a triangle is three nodes that connect with each other accounting for 3 triplets, such as ABC, BCA, and CAB (Barabási et al., 2016, p. 70).

Centralization of a network measures how different the nodesâ properties are compared to each other. It is defined based on $deg^+(v_i)$, $deg^-(v_i)$, and BC, as the difference of the most central node's value (maximum) to all other node values compared to the maximum possible value of such a difference. Therefore, for $deg^+(v_i)$ and $deg^-(v_i)$, the centralization is defined as:

$$C_{deg} = \frac{N \cdot \max(deg(v_i)) - \sum_i deg(v_i)}{(N-1) \cdot (N-2)} \tag{4}$$

And the centralisation of betweenness centrality (BC) is defined as:

*Corresponding author, yuan.liao@chalmers.se

$$C_{BC} = \frac{N \cdot \max(BC(v_i)) - \sum_i BC(v_i)}{N \cdot N \cdot (N-2)} \tag{5}$$

In addition to the aforementioned metrics, we also compare the community structure of places networks based on Survey and Twitter LT. Modularity takes values between -1 and 1 and indicates the density of edges inside communities to edges outside communities. The underlying assumption is that randomly wired networks lack an inherent community structure (Barabási et al., 2016). Therefore, the modularity is defined to measure the network's real wiring diagram matrix (A_{jk}) and the expected number of links between node j and node k if the network is randomly wired:

$$Q = \frac{1}{2L} \sum_{jk} (A_{jk} - p_{jk}) \tag{6}$$

where L stands for the total number of links/trips, G_c stands for a group of nodes formulating a subgraph, and p_{jk} stands for the expected number of links between node j and node k if the network is randomly wired. p_{jk} can be calculated by:

$$p_{jk} = \frac{deg(v_j) \cdot deg(v_k)}{2L} \tag{7}$$

The Louvain method of community detection is to maximise the modularity Q to find the best possible grouping of the nodes of a given network. However, in reality the computational load is too high, so heuristic algorithms are used (Blondel et al., 2008).

References

Barabási, A.L., et al., 2016. Network science. Cambridge university press.

Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment 2008, P10008.