

How representative are the geotagged tweets as a proxy for human mobility? Comparing social media data with travel survey and GPS log data

Yuan Liao^{*,a}, Sonia Yeh^a

^aDepartment of Space, Earth and Environment, Division of Physical Resource Theory, Chalmers University of Technology, Gothenburg 41296, Sweden

Abstract

Using geotagged social media data to model travel behaviour (human mobility) has received increasing attention. The low cost of retrieving geotagged tweets makes it especially appealing compared to other data sources. As a proxy for human mobility, such a data source needs careful inspection regarding its representativeness. To date, the works comparing geotagged tweets with the other data sources are not systematic; they either focus on individual trajectory or places network, however, the links between these two perspectives have rarely been constructed. Moreover, the involved data sources are usually limited to one to two categories without careful consideration of different data forms that support different research purposes. To fill the gaps in the literature, this paper attempts to conduct a systematic comparison, concerning longitudinal and lateral data forms where we combine individual trajectory and places network perspectives. This study is based on five data sets; they come from the household travel survey, geotagged tweets, and mobile phone GPS log. The results of this paper contribute a deeper understanding of what can be revealed from geotagged tweets compared with the other data sources to represent the actual travel behaviour.

Keywords: human mobility, data mining, data source comparison, individual trajectory, places network

1. Introduction

Human mobility refers to the movement of human beings, seen as individuals or groups, in space and time (Barbosa et al., 2018). From an individual's perspective, the mobility trajectory is a time series of visits to various locations. On the population level, individuals' mobility trajectories can be aggregated to study the flows of population travelling between different locations/regions. Depending on the spatiotemporal scale for the aggregation, the Origin-Destination matrix (OD matrix) can be constructed with all the combinations of origins and destinations for trips. Therefore, the human mobility study contains two major focuses, **individual trajectory** and **places network**. Empirical data play an important role in understanding human mobility (Gonzalez et al., 2008; Song et al., 2010; Barbosa et al., 2018).

Traditionally, travel behaviour studies depend on household travel surveys that have good accuracy and statistically valid demographic information, however, data from these sources suffer from many disadvantages including high cost which increases over time (Yue et al., 2014), short period (one day for travel diary), and a low response rate which decreases over time (Stopher and Greaves, 2007). To overcome these drawbacks, many efforts have been devoted to the emerging data sources by which, a large-scale collection of human mobility traces becomes feasible.

Along with the development of Information and Communication Technologies (ICT), online social media services, e.g.,

Twitter, has received increasing attention from the transportation research community (Rashidi et al., 2017) focusing on the individual mobility patterns (e.g. Lee et al., 2016), the aggregate-level mobility (typically geographically bounded to city or national level) (e.g. Jin et al., 2014), and the travel demand modelling (e.g. Lee et al., 2015), etc. A tweet typically contains multiple components that can be useful in transport research, including text, hashtag, location, and timestamp. When users choose to report their locations when sending out tweets, these are called **geotagged tweets**. Despite the low proportion of geotagged tweets (1-3%) (Morstatter et al., 2013), these check-ins provide the precise location information and have been increasingly used for understanding human mobility (Lenormand et al., 2014; Jurdak et al., 2015). The low cost of retrieving geotagged tweets makes it especially appealing compared to other data sources (Rashidi et al., 2017).

Geotagged tweets can be obtained in multiple ways. Geotagged tweets collected from the Streaming API are often limited in a geographical bounding box yielding a lateral data set that covers a large number of Twitter users but it takes time to accumulate and individuals' movements that occur outside the bounding box is not captured (Liao et al., 2019). By accessing the user timeline, all the publicly available historical tweets of a specified user can be collected resulting a longitudinal record of individual trajectory without any geographical boundaries. Longitudinal geotagged tweets serve as the only data source without being constrained in a specific area. 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; places network gains much atten-

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

tion because it directly connects to travel demand modelling and has more potentials to support real-world applications (e.g. Jin et al., 2014).

Geotagged tweets as a proxy for human mobility, is especially controversial. The data source is easy and free to access, provides precise location information (Jurdak et al., 2015) and opportunities for long-term tracking of movements free from geographical boundaries (Liao and Yeh, 2018). Geotagged tweets have also been heavily criticised of having population bias (Mislove et al., 2011) and behaviour distortion (Tasse et al., 2017). A recent study shows that people geotag consciously and intentionally in uncommon places (Tasse et al., 2017).

One recent literature review shows a positive view of the experts about the usefulness of such data sources for modelling travel behaviour (Rashidi et al., 2017), despite these identified disadvantages of geotagged tweets. The consensus also points out the necessity of careful inspection of using geotagged social media data to approximate the actual travel behaviour, because data sources have their pros and cons. Despite a few drawbacks, travel surveys contain socio-demographic information and detailed activity records making them not easily replaceable by other emerging data sources (Janzen et al., 2017). The carefully designed sampling to derive statistically representative population-level estimates makes traditional travel survey a vital source for validation/calibration the emerging data sources. **Mobile phone Call Detail Record (CDR)**, i.e., mobile phone data in many studies (e.g. Yue et al., 2014), is the most widely applied among those emerging data sources. CDR can be collected long-term with very large numbers of tracked individuals, however, this data source is often not easy to access, and has the shortcomings of spatiotemporal sparsity and incomplete trajectory (Chen et al., 2018), and is 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 providing a relatively complete and accurate picture of individual mobility trajectory, which makes it close to the “ground truth.” To summarise the literature review, the main characteristics of the four data sources are captured illustratively in Table 1. Compared to the other data sources, geotagged social media data have strengths in long collection duration, a large number of studied individuals, large spatial coverage, easy to access, low cost, and accurate location information. The main weaknesses are incomplete sampling of individual trajectory and lack of socio-demographic information and trip information.

When social media data are cross-validated against the higher temporal resolution data such as CDR (Lenormand et al., 2014), good agreements are generally found regarding trip distance distribution etc., however, CDR and geotagged tweets have similar passive data collection manner that might share some of the same shortcomings. When validated against travel survey, one study shows promising results of geotagged social media having displacement distribution, length, duration, and start time to infer individual travel behaviour (Zhang et al., 2017). Some studies go a step further beyond the fundamental human mobility patterns to compare geotagged tweets with traffic data (Ribeiro et al., 2014) and travel demand data (Lee

Table 1: Characteristics of the four data sources. ^aGeotagged social media data. ^bTraditional household travel survey. ^cTime length of tracking the same individual. ^dLow 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	✓	×	✓	×

et al., 2015) and generally found good results.

To date, the works comparing geotagged tweets with the other data sources lack systematic rigour. On the one hand, the comparisons are typically limited to one to two categories without comprehensive considerations of the trade-offs and limitations of different data forms. On the other hand, those studies either focus solely on individual trajectory or places network, 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 attempts to explore the representativeness of geotagged tweets as a proxy for human mobility by conducting a systematic comparison, using both longitudinal and lateral data forms to examine individual trajectory and places network perspectives.

2. Methodology

To systematically compare geotagged tweets with other data sources, we propose the comparison framework as shown in Figure 1. There are five data sets used in the present study as described in **Supplementary Information (SI)**, Section 1; Twitter LD that is described in SI Section 1.2.1, GPS log (SI Section 1.1), Survey-Lausanne (SI Section 1.3.1), Survey (SI Section 1.3.2), and Twitter LT (SI Section 1.2.2). The comparison is based on two essential dimensions of human mobility, individual trajectory (see results in Section 3.1) and places network (see results in Section 3.2). We apply mixed methods from various disciplines in the present study. More detailed description of methods can be found in SI.

The individual trajectory perspective highlights two aspects: recurrent visit and non-recurrent visit. The more one has recurrent visits, the higher predictability of one’s mobility is. Therefore, we conduct the mobility regularity comparison based on different data sources quantifying the degree to which the revealed individual mobility is regular (SI Methods Section 2.2.1). After showing the mobility regularity degree, we explore the temporal profiles of visits. Bridging spatial and temporal dimensions, we further illustrate the diffusion process comparing Twitter LD with GPS log (SI Methods Section 2.2.2). Trip distance (SI Methods Section 2.2.3), as one key indicator of travel behaviour, is compared across all the data sets;

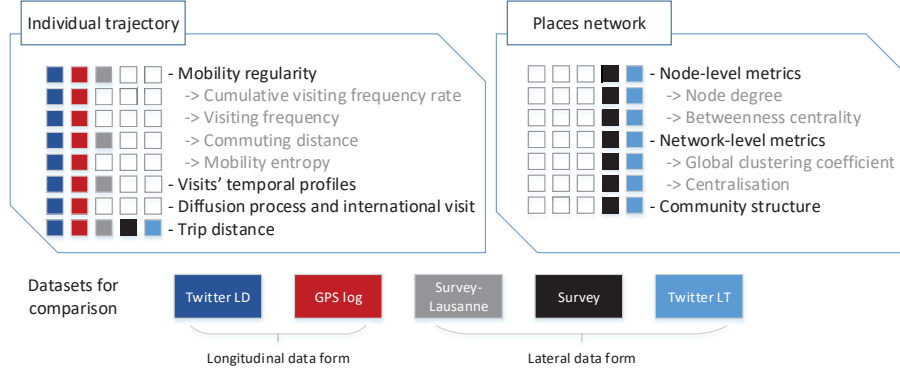


Figure 1: Comparison structure.

we also discuss the difference between two data forms: longitudinal and lateral.

Multiple individuals' geotagged trajectories form places network. In this part of analysis, we present a network-based comparison using node-level metrics (SI Methods Section 2.3.1) and network-level metrics (SI Methods Section 2.3.2) that originate from network science. In addition, we present the detected community structure of Twitter LT in comparison with Survey (SI Methods Section 2.3.3).

3. Results and discussion

3.1. Individual trajectory

3.1.1. Mobility regularity

Figure 2 shows the **cumulative frequency rate** vs the most visited location sorted by their visit frequency. The faster the curve stabilises, the higher number of the visits concentrating on a small number of locations. Twitter data tend to represent more non-recurrent visit compared to mobile phone GPS data.

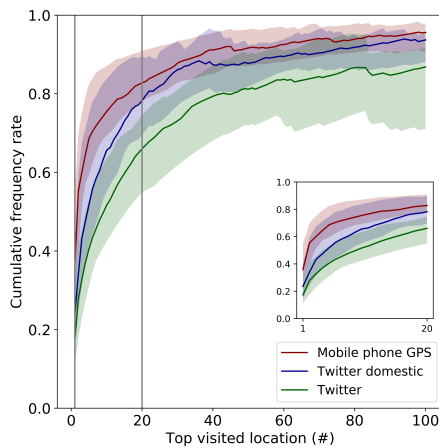


Figure 2: **Cumulative visiting frequency by the most visited locations ordered by their visiting frequency.** The shaded range indicates the upper bound (75%) and lower bound (25%) of the cumulative frequency rate of visits.

A significant difference is observed between three data sets regarding the visits on home and workplace (see Figure 3). Mobile phone GPS data set has the highest **visiting frequency rate**

for both home and workplace (Figure 3A-B) followed by Twitter domestic data set and Twitter data set. One might question that 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 estimated home and workplace **commuting distance** (C). Here, the commuting distance observed from the travel survey can be regarded as the ground truth. Twitter data set yields a similar median value of commuting distance, however, it has a substantial variance between the studied users. Mobile phone GPS data set has a larger median value compared to the travel survey while its variance is smaller than the Twitter data set. The difference suggests that the individual difference plays an important role in the Twitter data set; the representativeness of geotags in Tweets vary between individuals when we want to reveal their daily routine.

The **mobility entropy** characterises the visit heterogeneity providing a complementary view. The higher the entropy, the harder to predict one's whereabouts. The mobility entropy of the three data sources are ordered from lowest to highest: GPS log, Twitter LD domestic, and Twitter LD. The visits covered by Twitter LD domestic are more decentralised than the Mobile phone GPS data set. When the international visits come into play (the Twitter LD), the entropy increases significantly, not only because the distinct locations increase but also due to the non-recurrent visits, e.g., those locations that are only visited once.

3.1.2. Visits' temporal profile

The individual geotagged tweets and mobile phone GPS records are different on both spatial and temporal dimension (see SI Figure 2 & 1). The mobile phone GPS data have movements which peak during morning and afternoon; this implies the mobility reflects more routine activities. Twitter users can geotag their tweets wherever and whenever they are; the geotags peak during noon and evening. Does that essential difference between those two data sources have an impact on the recorded visits and their corresponding activities? As described in SI Section 2.2.1, the workplace and home are estimated based on the temporal rules. It is important to explore whether those rules output the similar temporal profiles of workplace/home from the above two data sources. It turns

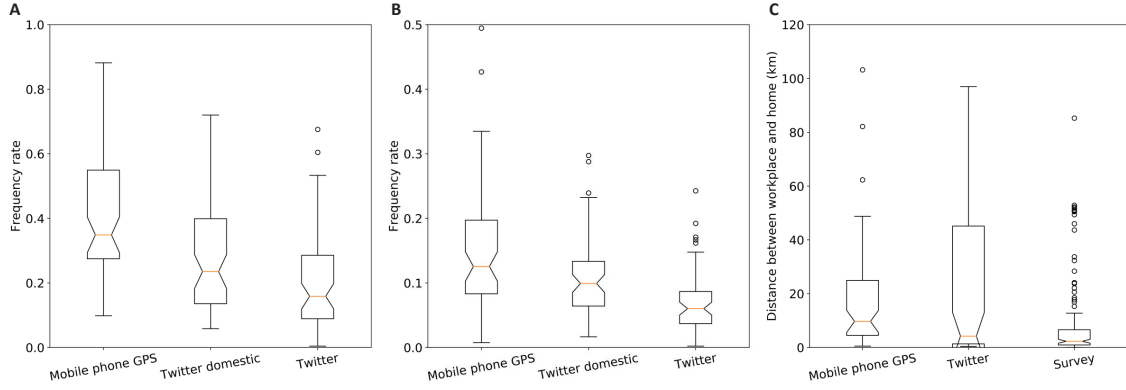


Figure 3: **Estimated home and workplace metrics.** (A) Visiting frequency of estimated home. (B) Visiting frequency of estimated workplace. (C) Commuting distance.

out the probability of being at home based on the mobile phone GPS data resembles the temporal distribution extracted from the travel survey (see the red curve in Figure 4), however, the Twitter users' probability of being at home is below 0.5 across weekdays and weekends (see the blue curve in Figure 4) which reflects the temporal distribution of their geotagging behaviour (see SI Figure 2). In other words, the home location that is derived from the geotags in tweets is less reliable than the one obtained from the mobile phone GPS and travel survey. Having such an observation, it is not surprising to see the commuting distance based on the Twitter data set has a much larger variance than the one based on the travel survey (Figure 3).

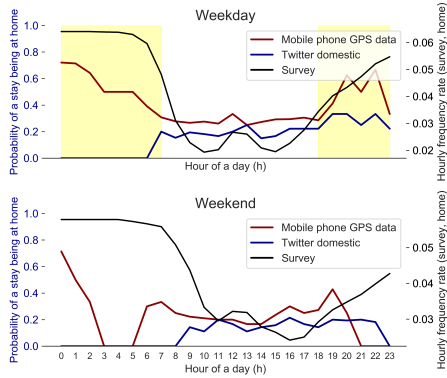


Figure 4: **Hourly probability of being at home in the weekday and weekend.** For the mobile phone GPS data set and the Twitter domestic data set, a data point on the curve represents average value across all the participants about the relative probability of being at home taking consideration of all the other recorded locations that have been visited during that hourly time interval. For the survey data set, the curve represents the home location's visiting frequency rate distribution across an average weekday or weekend.

3.1.3. Diffusion process and international visit

Diffusion process represents how the distance one regularly covers changes over time. The diffusion process is shown in Figure 5. Considering domestic visits only, the geotags in tweets yield similar diffusion process to the GPS log; they stabilise rapidly and remain around a small value of r_g ; when the international visits are included, the geotags produce a signifi-

cantly higher r_g . It suggests that the international visits account for a substantial proportion in the mobility revealed by the geotags. Such an observation is consistent with the number of distinct locations (N) reported in SI Table 1 where the international visited locations account for 44% of total visited locations.

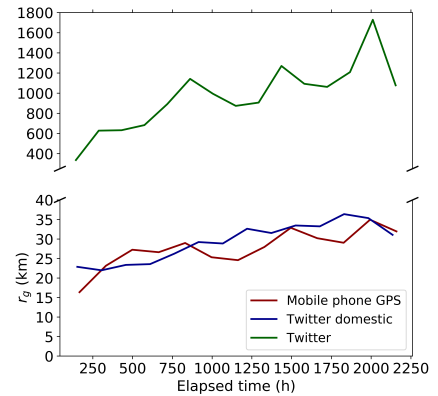


Figure 5: **Diffusion process that is represented by the time history of the radius of gyration.** The origin of the time history is set to the first time that the most visited location is observed. The time frame is 90 days and the number of data points is set to 50. The presented curves represent the 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).

3.1.4. Trip distance

Figure 6 indicates the discrepancy between the trip distance revealed by the geotagged tweets versus the one by the travel survey; on both longitudinal and lateral data form, geotagged tweets yield larger trip distance than the travel survey. In addition, regarding the Haversine trip distance, longitudinal geotagged tweets resemble the travel survey more compared to the lateral data set.

Why is it the case that compared to the lateral level, longitudinal geotagged tweets agree more with travel survey? As shown in SI Table 1, despite infrequent geotweeting of the users included in Twitter LT, once they geotag their tweets, they tend

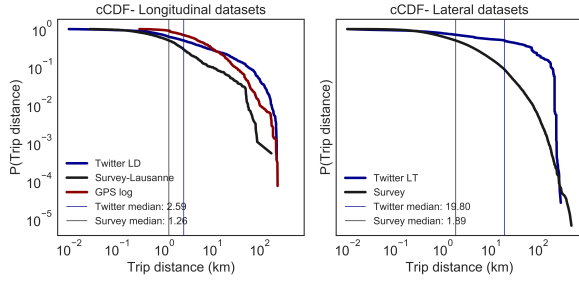


Figure 6: **Haversion distance comparison: domestic trips.** The travel survey and geotagged tweets have been processed to have similar resolution which is around 10 m. However, the minimum distance in the GPS log data set is 300m which has been determined by the filtering algorithm described in Section 1.1.

to produce more geotagged tweets on those covered days (2.8 vs 2.3). Therefore, the non-recurrent trips are more likely to be presented in the lateral data set given its low number of days covered and the burst effect of geotagging behaviour. That explains why the trip distance indicated by Twitter LT is significantly larger than the survey.

3.2. Places network

Trips form a directed network of physical locations that can be represented by an OD matrix. The network properties account for a key facet of individual mobility that emerges at the aggregate level. In this section, the network properties are to be discussed from the node level and the network level (Morstatter et al., 2013) in a comparison between the Twitter LT and the Survey. And we further explore the community structure in the network that two data sets formulated.

3.2.1. Node-level metrics

Constructed OD matrix reveals differences between the two data sets (Figure 6). A larger number of between-cell trips from Twitter LT than Survey is consistent with the observation of longer trip distances reported by the Twitter users. Compared to geotagged tweets, travel survey covers more OD pairs. In other words, the reported trips in the Twitter LD are more concentrated.

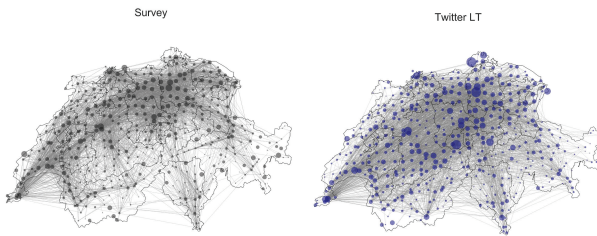


Figure 7: **OD matrix.** The node size is proportional to the node degree (normalised by the sum of total degrees for each data set).

The node degree comparison between two data sets is illustrated in Figure 8. The correlation between the two data sets is strong for both in- and out-degree. Despite the discrepancy in the trip distance between two data sets, the Twitter LT's degree

to which how frequent the trips are generated between or within the cells agrees with the results from the Survey.

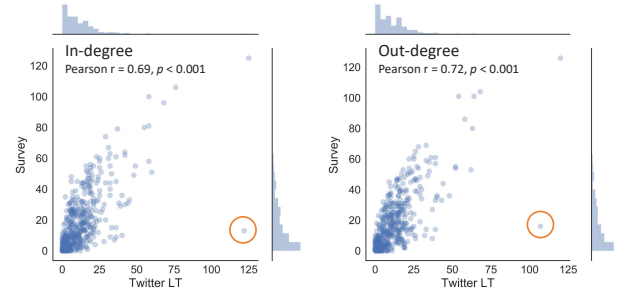


Figure 8: **Geotagged tweets vs travel survey: in-degree and out-degree.** Each data point represents one grid cell's degrees from the two data sets. The orange circle highlights an outlier which is a tourist attraction area located at the geographical centre of Switzerland.

The comparison of betweenness centrality between two data sets is shown in Figure 9. Looking at those nodes with non-zero betweenness centrality, the overall correlation between two data sets is moderate (Pearson correlation coefficient = 0.5, $p < 0.001$). The locations of high betweenness centrality often occur on the shortest paths between other two areas. Those bridge locations identified by the travel survey are likely to be the ones identified by the geotagged tweets, 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 (highlighted on the right graph in Figure 9). The important hubs that are only highlighted by the travel survey are close to transport facilities locations, however, the hubs that are solely identified by the Twitter LT are related to the natural places, such as mountains or parks.

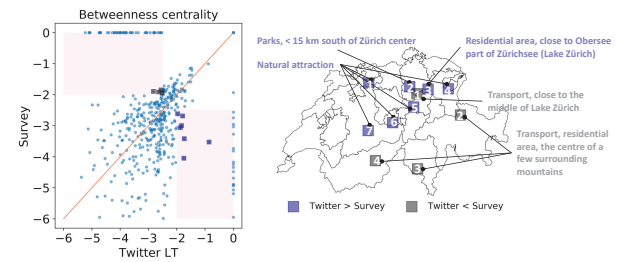


Figure 9: **Betweenness centrality comparison.** The scatter plot on the left hand shows the log value of 463 cells' betweenness centrality for different data sets. The closer the point to the orange diagonal line, the closer two data sets regarding that grid cell. The cells that have large discrepancy between two data sets are highlighted with rectangles on the right map with a short description of that 10km × 10km cell.

3.2.2. Network-level metrics

At the network level, the four centralisation metrics' values stabilise when the observation time period increases (see Figure 10 left graph). With one-year collected data, the travel survey yields larger centralisation of global clustering coefficient than the geotagged tweets (see Figure 10 right graph). Such an observation is visually confirmed by Figure 7. Regarding the centralisation of in- and out-degree, the difference is small. As

for the centralisation of betweenness centrality, the Twitter LT is larger than the Survey indicating the larger variation between the grid cells regarding the hub importance.

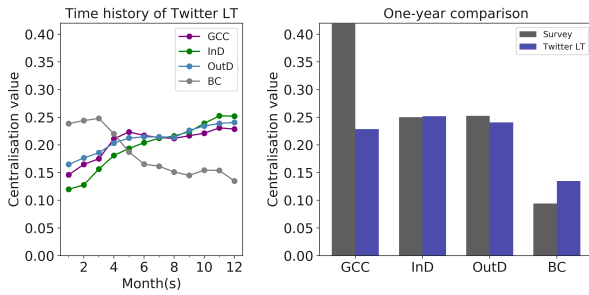


Figure 10: **Network-level comparison: centralisation of global clustering coefficient (GCC), in-degree (InD), out-degree (OutD), and betweenness centrality (BC).** The centralisation value is between 0 to 1; the larger the centralisation, the larger the variation between nodes in the given network. Time history of Twitter LT shows how the centralisation metrics change their values when more months' data have been involved in the calculation.

3.2.3. Community structure

Figure 11 shows the community structure of OD matrix that is produced from Survey and Twitter LT. Their global modularity differs; one-day travel diary has higher modularity than the geotagged tweets. And the detected communities are more constrained by the geographical distances between the nodes for the Survey compared to the Twitter LT. This is because the one-day travel diary mainly captures the routine mobility, however, the formulated communities in geotagged tweets tend to represent the non-recurrent travelling preferences.

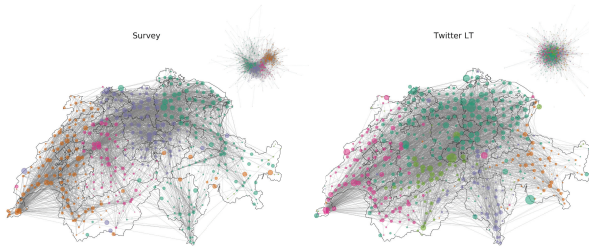


Figure 11: **Community structure.** Global modularity: Twitter LT = 0.33, Survey = 0.48.

4. Remarks

In this study, we explore the representativeness of geotagged tweets, as a proxy for human mobility; we compare them with household travel survey and GPS log data. Below is a brief summary of the main findings. A more thorough discussion is in plan.

From the perspective of individual trajectory, geotagged tweets represent more non-recurrent (less routine) visits than the other two data sources. Moreover, there is a large individual difference regarding the mobility regularity observed by their geotagged activities. Because of the behaviour distortion, i.e.,

Twitter users tend to tweet during noon and evening, home and workplace, derived using temporal distribution method, are less reliable than the other data sources. International visits account for a substantial proportion in geotagged tweets; they tend to represent those destinations that are more distant than the ones in the travel survey. Geotagged tweets reveal longer trip distance than the other two data sources in general, however, such difference is amplified in the lateral data form given the burst effect of geotweeting behaviour.

From the perspective of places network (Twitter LT vs Survey), Twitter-based OD matrix spreads more towards mountain areas which have low population density but high natural attraction density. Geotagged tweets generate lower modularity community structure than the one-day travel diary; geotagged communities represent more non-recurrent travelling preferences rather than traffic-network constraints. Despite some differences, geotagged tweets and the one-day travel diary have correlated zonal demand and the degree to which an area is important connecting other areas.

References

- 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*.
- 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.
- Freedman, D., Diaconis, P., 1981. On the histogram as a density estimator: L 2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 57, 453–476.
- Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L., 2008. Understanding individual human mobility patterns. *Nature* 453, 779–782.
- Janzen, M., Müller, K., Axhausen, K.W., 2017. Population synthesis for long-distance travel demand simulations using mobile phone data, in: 6th Symposium of the European Association for Research in Transportation (hEART 2017).
- Jin, P., Cebalac, 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.
- 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., 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.
- 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*.
- 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*.
- 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.
- 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.
- Song, C., Qu, Z., Blumm, N., Barabási, A.L., 2010. Limits of predictability in human mobility. *Science* 327, 1018–1021.
- Stopher, P.R., Greaves, S.P., 2007. Household travel surveys: Where are we going? *Transportation Research Part A: Policy and Practice* 41, 367–381.
- Tasse, D., Liu, Z., Sciuto, A., Hong, J.I., 2017. State of the geotags: Motivations and recent changes., in: *ICWSM*, pp. 250–259.
- 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.

How representative are the geotagged tweets as a proxy for human mobility?

Comparing social media data with travel survey and GPS log data

Supplementary Information

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

February 8, 2019

Contents

1	Data description and preprocessing	1
1.1	GPS log	1
1.2	Geotagged tweets	1
1.2.1	Longitudinal dataset (Twitter LD)	1
1.2.2	Lateral dataset (Twitter LT)	2
1.3	Travel survey data	2
1.3.1	Lausanne residents (Survey-Lausanne)	2
1.3.2	All participants (Survey)	2
1.3.3	Characteristics of datasets for comparison	3
2	Methods	4
2.1	Definitions	4
2.2	Individual trajectory	4
2.2.1	Mobility regularity	4
2.2.2	Diffusion process and international visit	4
2.2.3	Trip distance	5
2.3	Places network	5
2.3.1	Node-level metrics	5
2.3.2	Network-level metrics	6
2.3.3	Community structure	6

*Corresponding author, yuan.liao@chalmers.se

1 Data description and preprocessing

1.1 GPS log

The GPS dataset in this study is selected from a Mobile Data Challenge (MDC) which is run from October 2009 to the end of March 2011 (Laurila et al., 2012). This dataset contains regularly recorded GPS from the participants living in Lausanne, Switzerland during the MDC campaign. The raw locations are collected by combining GPS and WiFi readings per 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, latitude) so that the resulting location rectangle, or anonymity-rectangle, contains enough inhabitants (Sweeney, 2002; Laurila et al., 2012). Therefore, the spacial resolution varies between locations depending on the population density.

Some participants have incomplete records due to technical issues. To guarantee the data quality, three criteria are applied to the individual trajectory of GPS locations: (a) at least 30% of their days have recorded locations (Do et al., 2015); (b) recording time of at least 90 days (Do et al., 2015); (c) a hour-long-interval-based fraction of missing location data is less than 80% (Song et al., 2010). After preprocessing, the applied MDC dataset includes 61 participants satisfying aforementioned criteria. The raw records of the participants are shown in Figure 1.

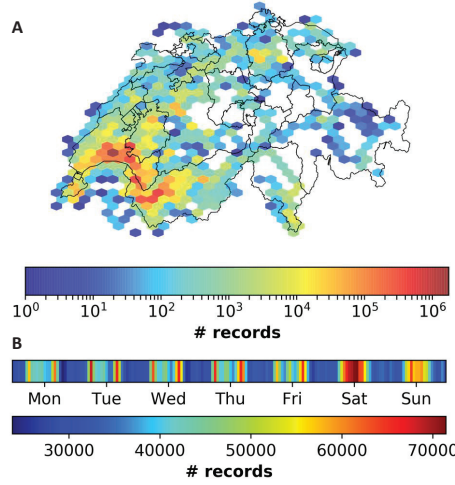


Figure 1: **Spatial (A) and temporal (in hourly interval)(B) distribution of GPS log data.**

To extract the history of place visits, the raw location data are further processed. To identify stays (i.e., location data logged when users are engaging in activities), we consider a location which the participant remains within a radius of 300 meters for at least 10 min as a **stay** (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 from the GPS log dataset. 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).

1.2 Geotagged tweets

1.2.1 Longitudinal dataset (Twitter LD)

We collected Tweets generated during a five-month period (5 December 2017 - 16 April 2018) within the geographical bounding box of Switzerland. Using this dataset, 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 user timeline, without applying a spatial boundary limit (The Tweepy project developers, 2017). This method has a limit on the maximum number of tweets that can be collected from a specified user, producing varied time span and varied tweet number 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 Lausanne area, Switzerland, and have a substantial number of domestic geotagged tweets to reasonably capture their local activity trajectories: (a) the covered time span is above 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 that we capture partly their visited locations. **Therefore, each generated location by users 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 2. Job-posting bots constitute a growing proportion of public geotags (Tasse et al., 2017). We further validate the 61 users using BOT score (Davis et al., 2016). All users have the BOT score < 1 implying that they are unlikely to be bot accounts. The same process has also been applied to the lateral dataset described in the following section (1.2.2).

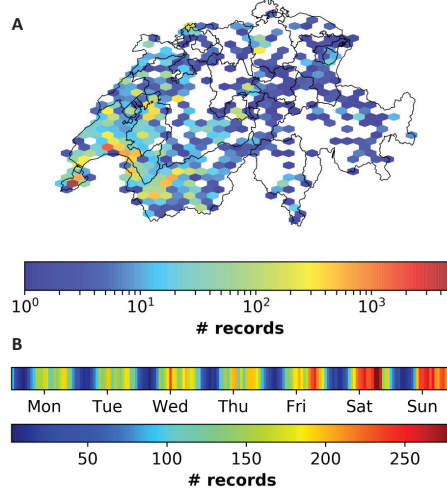


Figure 2: **Spacial (A) and temporal (in hourly interval)(B) distribution of domestic geotagged tweets (individuals).**

1.2.2 Lateral dataset (Twitter LT)

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

To construct 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 larger 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). After screening, the final subset includes 37,048 trips produced by 3,249 Twitter users over one year.

1.3 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 reported 193,880 trips during their travel diary day. We further formulate two subsets to compare with the GPS log & the Twitter LD (Survey-Lausanne), and the Twitter LT (Survey) respectively. The GPS coordinates pair is obtained by feeding the address (at the street level) of origin and destination into Google Places API (Google).

1.3.1 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 their 4,482 stays within Lausanne.

Compared to the GPS log and geotagged tweets, only one day is captured for each participant, as shown by the spatial distribution (Figure 3a-b) and temporal distribution of two typical days (Figure 3c-d). Most stay locations are in Lausanne where the participants live. They represent a more routine daily mobility patterns compared to GPS log and geotagged tweets. Figures 3c-d show two distinct patterns between weekday and weekend with relatively accurate start time, end time, and activity purpose of each stay.

1.3.2 All participants (Survey)

To compare the travel survey with the Twitter LT, all the trips that have both the origin and destination 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.

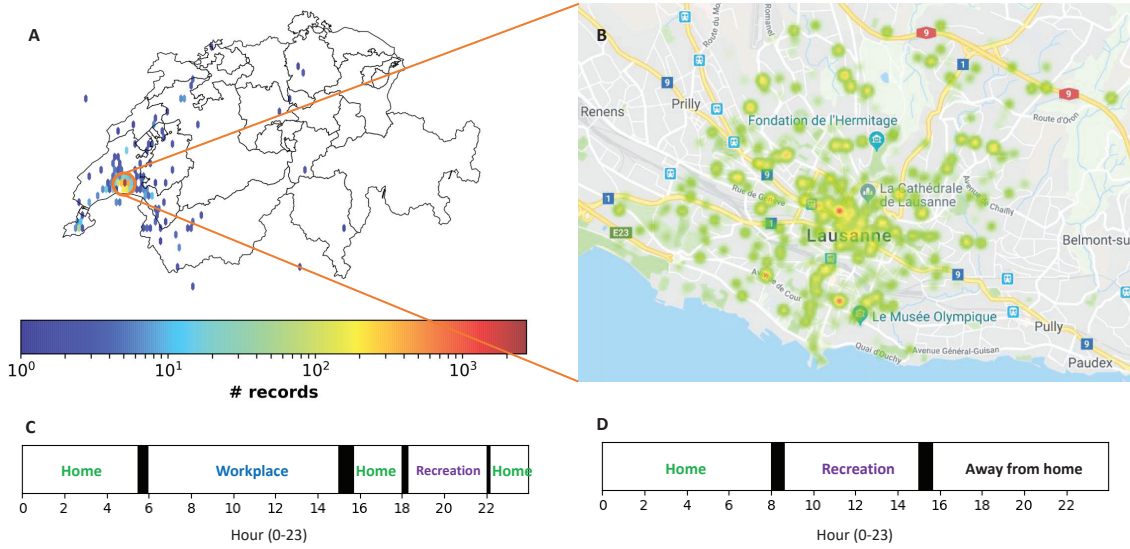


Figure 3: **Spatial and temporal (in minute) distribution of domestic reported locations.** (A) Locations in Switzerland. (B) Locations in Lausanne area. (C) Temporal distribution of a typical weekday (Thursday). (D) Temporal distribution of a typical weekend day (Sunday). The black line in temporal distribution graphs, C and D, indicates the duration of moving from two adjacent stay locations.

1.3.3 Characteristics of datasets for comparison

Both the Twitter LD and GPS log are unevenly distributed in time (Figure 1B and 2B). Compared to GPS log, geotagged tweets distribute less concentrated across 24 hours and peak at noon and night. For GPS log, the records peak during morning and afternoon implying that commuting is the main driver of mobility.

The five datasets' statistics are summarised in Table 1. GPS log, Twitter LD, and Survey-Lausanne are longitudinal datasets. Survey and Twitter LT are lateral datasets. Despite two forms of datasets, 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 indicator of covered individuals in the dataset, calculate the mean value, 25% value, and 75% value.

Table 1: Statistics of the five datasets. LD indicates longitudinal, i.e. the dataset is collected based on the individuals. LT indicates lateral, i.e. the dataset 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 regions. F_g : stays/visited regions per recorded day. q : the completeness of mobile phone GPS data, defined as a one-hour-interval-based fraction of missing location data. ^a Only includes domestic locations.

Source	GPS log		Twitter LD		Twitter LD domestic ^a		Survey-Lausanne ^b		Survey ^b		Twitter LT	
Statistics	mean	IQR	mean	IQR	mean	IQR	mean	IQR	mean	IQR	mean	IQR
Time span	369.5	(267, 514)	1152.4	(681, 1444)	-	-	388	-	388	-	365	-
Days covered	221.1	(105, 302)	287.1	(130, 396)	-	-	1	-	1	-	32.0	(1, 6)
N	89.5	(51, 113)	160.6	(70, 172)	71.5	(39, 88)	3.2	(2, 4)	2.9	(2, 3)	7.4	(2, 5)
n	616.8	(281, 716)	609.2	(233, 905)	346.2	(161, 467)	5.1	(3, 6)	4.8	(3, 6)	22.8	(2, 8)
F_g	2.7	(2.2, 2.9)	2.31	(1.4, 2.7)	1.4	(0.9, 1.5)	5.1	(3, 6)	4.8	(3, 6)	2.8	(1.7, 4.0)
q	0.7	(0.7, 0.8)	-	-	-	-	-	-	-	-	-	-

The GPS log dataset has a one-year time span while the Twitter LD covers roughly a three-year time period. Despite the difference in time span, both datasets have a similar number of days covered. It is obvious that sampling frequency is lower for geotagged tweets than GPS log, however, they show a comparable number of visited regions, as seen from both N and n in Table 1. Comparing the number of distinct stay regions with the number of all recorded regions, the Twitter LD tends to present less regularly visited regions compared to the GPS log ($N/n = 14.5\%$ vs 9.5%). The GPS log has much higher sampling frequency once a movement is detected compared to Twitter data, however, the Twitter LD has F_g at the same level as the GPS log, implying that participants stay in a limited number of regions every day. q quantifies the fraction of hour-long intervals when the user's location is unknown to us (Song et al., 2010). The distribution of q is consistent with the previous findings (Song et al., 2010) that humans stay rather than move during most of the time in a given day, especially for weekdays. Survey-Lausanne, covering one day for each participant, has an average of 5.1 stays per person per day which provides more complete information than the other two sources, 2.7 for GPS log and 2.3 for Twitter LD.

Both two lateral datasets, Survey and Twitter LT, cover a time span of one year. Twitter LT has more visited locations than Survey does, however, Survey has a larger number of visited locations per record day (F_g) compared to Twitter LT. As geotagged social media data, Twitter LD and Twitter LT are different in many aspects. Despite Twitter LD comes from

the top geotag users, it has 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%).

2 Methods

2.1 Definitions

Depending on the research community and the data source, human mobility’s 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 that is 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 the GPS log, is defined as being at a location where the participant remains within a radius of 300 meters for at least 10 min. For the 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 and travel survey.
- **Trip** is defined as the connection between two consecutive visits/stays generated by the same individual. For geotagged tweets, it is also required to have the time interval smaller than 4 hours. “Trip” is also equivalent to “displacement” commonly defined in many studies.

2.2 Individual trajectory

2.2.1 Mobility regularity

In mobility studies, trips have purposes such as home-based-work and home-based-other. Unlike travel surveys, which record trip purposes, (Schneider et al., 2013; Çolak et al., 2015), such information must be derived for other data sources including the GPS log and the Geotagged tweets. 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 a large 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 human mobility can differ between mobile phone GPS, geotagged tweets, and travel survey. To compare those data sources, we propose a set of indicators to quantify the expressed regularity. Based on the identified three types of locations, the regularity is measured by their **cumulative frequency rate**, **visiting frequency rate distribution**, **commuting distance**. The **mobility entropy** can be expressed as $S = -\sum_{j=1}^{n'} p(j) \log p(j)$ characterising the heterogeneity of visitation patterns (Song et al., 2010). The larger the mobility entropy, the higher the heterogeneity.

2.2.2 Diffusion process and international visit

The total radius of gyration r_g is defined as:

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

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

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

The **radius of gyration** (r_g) is widely used to indicate a distance that one covers on a regular basis; it combines the locations’ geographical distribution and their visiting frequency, has been widely applied to characterising human mobility patterns (Gonzalez et al., 2008; Song et al., 2010; Lu et al., 2013; Jurdak et al., 2015; Liao and Yeh, 2018). Therefore, the movements that happen in a comparatively confined space will have a small r_g even though a large distance movement happens occasionally (Lu et al., 2013).

The time history of r_g , starting from one location, indicates how people diffuse during a particular time frame. The length of time interval between two consecutive visits varies within and between individuals. For the sake of aggregation and comparison across data sources, a necessary pre-process is needed to produce the time history of r_g during a time period of given length. The detailed description of such a method can be found in our previous study (Liao and Yeh, 2018).

Representing a regular mobility range, r_g is less affected by those visits that happen infrequently. Besides the recurrent mobility that serves everyday life, **international travel** is also another important facet of human mobility. International travel is often perceived to happen infrequently and to cover longer distances. How do the international visits impact on the diffusion process? In this study, the answer can only be approached by the Twitter LD that includes the trajectory of individuals' geotags without the geographical constraint as the GPS log has. Despite the inherent differences between datasets, we attempt to reveal how the geotags reveal the Lausanne Twitter users' **top visited countries of international travel**. It is commonly observed that the studied Twitter users tend to geotag their tweets more frequently when they are travelling abroad. To make such a longitudinal Twitter dataset comparable with the international visits reported by the Swiss Statistics Agency, those repeated geotags have been merged to count as one visit to a particular country.

2.2.3 Trip distance

Trip distance is a key metric of human mobility that characterises how far people travel. It can be calculated in different ways; the **Haversine distance** (shortest distance) and the **actual travel distance**. The actual travel distance directly reflects the traffic condition, route selection, and energy consumption. However, such a distance can only be obtained when the route and accurate time sequences are known. Among the available data sources in this study, only the travel survey provides this information. Then comes the question that, from a more generic perspective, is it possible to approach the actual travel distance by looking into the Haversine distance?

Figure 4 shows the comparison between two distances from the reported domestic trips in the travel survey. It turns out there exists a strong linear relationship between the actual travel distance and the Haversine distance. Such a relationship's parameters are determined by the overall spatial constraints, e.g., the transport networks in Switzerland. Given those trips are collected rather evenly during a year, the impact of temporal (e.g., hour of the day) has been averaged out.

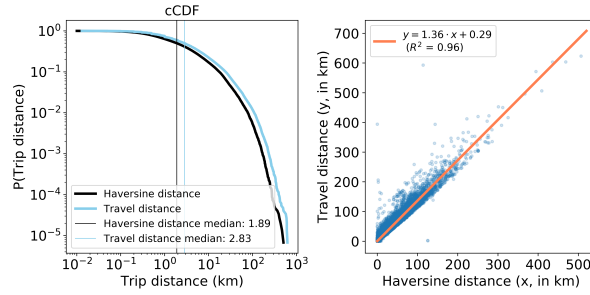


Figure 4: **Distances of domestic trips indicated by the travel survey.** cCDF indicates the complementary cumulative distribution function.

The revealed linear relationship suggests that once the Haversine distance is known, on the aggregate level, it can be converted into the actual travel distance by the established linear model. That observation supports the following comparison between the datasets regarding the Haversine distance.

2.3 Places network

For all the trips, both their origin and destination are converted into a corresponding grid cell using Military Grid Reference System (MGRS) with precision level of 10 km (Langley, 1998). By doing so, Switzerland is divided into 463 grid cells. The centre of each cell is determined by the centre-most coordinates pair of all the observed visits in that cell.

2.3.1 Node-level metrics

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 neighbours. In the current context, the neighbours can be divided into two categories; in-neighbours that the node is connected to as their destination, and out-neighbours that the node is connected to as their origins. Therefore, two node degrees, in-degree and out-degree, are measured to quantify the degree centrality. Besides the degree centrality, another centrality measure has been defined as **betweenness centrality**. Betweenness centrality (Freeman, 1978) identifies the degree of a node that bridges different location communities. A node with higher betweenness centrality would have more control over the network because more trips will pass through that node.

2.3.2 Network-level metrics

The network-level metrics quantify the OD matrix generated, as a whole, from different data sources. The **global clustering coefficient** (GCC) 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). **Centralisation** measures how equal the nodes are in a network regarding any given metric. It is defined as the difference of the value of the maximum-value node to all other node values compared to the theoretically maximum possible difference (Freeman, 1978). At the node level, we introduce three metrics. At the network level, correspondingly, we select centralisation of the node's in-degree (InD), out-degree (OutD), and betweenness centrality (BC) to compare different data sources.

2.3.3 Community structure

In a network science, a community is a group of nodes that have a higher likelihood of connecting to each other than to 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 where the corresponding community is an important facet of mobility patterns. Modularity quantifies whether a community partition is better than some other one (Barabási et al., 2016, p. 339). It can be used to detect the community structure. In this study, we apply the Louvain Method for community detection to extract communities which is a greedy optimisation method (Blondel et al., 2008).

References

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

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.

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.

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.

Freeman, L.C., 1978. Centrality in social networks conceptual clarification. *Social networks* 1, 215–239.

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.

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.

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., Gao, S., Goulias, K.G., 2015. Can twitter data be used to validate travel demand models, in: *14th International Conference on Travel Behaviour Research*.

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.

- Lu, X., Wetter, E., Bharti, N., Tatem, A.J., Bengtsson, L., 2013. Approaching the limit of predictability in human mobility. *Scientific reports* 3, srep02923.
- 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., Qu, Z., Blumm, N., Barabási, A.L., 2010. Limits of predictability in human mobility. *Science* 327, 1018–1021.
- Sweeney, L., 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 557–570.
- Tasse, D., Liu, Z., Sciuto, A., Hong, J.I., 2017. State of the geotags: Motivations and recent changes., in: *ICWSM*, pp. 250–259.