Analyzing metrics to understand human mobility phenomena: challenges and solutions

Luís Rosa, Fábio Silva, and Cesar Analide

Abstract Defining basic metrics that analyze human motion is important for urban planning and population mobility forecasting. These metrics are applied to understand extensive human mobility data generated from multiple sources. This means that our understanding of the basic metrics governing human motion is conditioned by integrating different data sources available. To the best of our knowledge, this article is a comprehensive study of the characteristics and metrics of human mobility patterns. Initially, it focuses on understanding common metrics in human mobility research. Then, it compares metrics such as resilience, displacement, interval and duration in different data types such as Wi-Fi, Call Detail Records (CDRs), Global Positioning System (GPS) and Social Media collected from two individuals. Comparing the results, a variation in movement patterns in both individuals is found in our study. Finally, we uncover a few interesting phenomena that lay a solid foundation for future research.

1 Introduction

The growing rate of urbanization presents many challenges and opportunities. One of the main challenges is how to provide infrastructure solutions that can cope with the stress caused by this massive expansion of populations in concentrated spaces. On the other hand, this population growth and interdependence on infrastructure

Luís Rosa
University of Minho, ALGORITMI Center, Dep. of Informatics, Braga, Portugal e-mail: `id8123@alunos.uminho.pt`

Fábio Silva
CIICESI, ESTG, Politécnico do Porto, Felgueiras, Portugale-mail: `fas@estg.ipp.pt`

Cesar Analide
University of Minho, ALGORITMI Center, Dep. of Informatics, Braga, Portugal e-mail: `analide@di.uminho.pt`

systems have increased availability of mobile phone records, GPS data and other datasets capturing traces of human movements.

Human mobility analysis has drawn much attention in many research fields ranging from the individual level to the commuting fluxes at the collective level [4, 21]. For instance, the Internal Displacement Monitoring Centre's (IDMC) 2020 Global Report on Internal Displacement (GRID) found that new displacements as a result of disasters in 2019 were three times the number of new displacements caused by conflict[3]. Currently, in response to the coronavirus 2019 (COVID-19) pandemic, several national governments have applied lockdown restrictions to reduce the infection rate. For this reason, [11, 1] used walking data to investigate the impact of government control measures on the decrease of human mobility. The results revealed that practicing social and physical distancing and timely decisions on mobility interventions are crucial to slow the spread of the pandemic.

A wide range of studies have established that there is a diversity of metrics in modelling human mobility at different temporal and spatial scales (e.g, temporal resolution of categorisation and aggregation, or spatial scale) [10, 14, 15, 7]. Another metric to deal with human mobility phenomena is the shift from reactive to pro-active policies focusing on resilience [13]. Commonly, human mobility is a critical role to assess resilience [17, 19]. But based on other human mobility metrics we may also understand social and economic activities associated with mobility. Understanding human movements allows us to evaluate not only resilience but also displacement of individuals, the time interval in displacements between two points, as well as the trip duration of the individuals, developing plans for improving living standards and so on.

This paper will focus on the challenges and solutions on human mobility phenomena. For example, people are less likely to move the same way in emergency situations, such as a hurricane, typhoon, earthquake and other natural or man-made extreme events, as they do in normal conditions. Similarly this perturbation generates limitation in the capture of data, once regularly busy area had a break in the data collected on mobile phones or other types of devices and, consequently, the absence of data that enables describing long-term human behaviour. Therefore, this study mainly provides a set of metrics that not only bring useful insights for a variety of applications but also its challenges, without comparing the metrics measurements using records of two individuals from Wi-Fi, CDRs, GPS and Social Media datasets.

The rest of the paper is organized as follows. In Section 2, it defines the concept of five common metrics in human mobility surveys. It analizes human mobility of two individuals, using set metrics, from different data types in Section 3. Then, in Section 4, it discusses the methodologies and techniques that can improve the analysis of movements considering the set of proposed data. Finally, this paper shows that metrics can be effectively used for future human mobility works.

2 Understanding Common Metrics in Human Mobility Research

Based on the literature, we have chosen to focus on four common metrics. These metrics can be applied to analyze the human tracking, looking at a large population of free-will and autonomous decision-making individuals, or at any event that implies a restriction in mobility. In addition, they can have a wide range of applications on human mobility research.

2.1 Spatial resolution of categorisation and aggregation

With spatial resolution of categorisation, we obtain data about communities and they are used to compute thematic maps of the territory. For example, unique locations that are visited by a user might be defined by tile grids, tower catchment areas, or GPS from points of interest for internal analysis. For each community researchers retrieve all the cells associated to its nodes and join them in a cluster, i.e. a geometric representation of the area covered by the community [9]. Areas corresponding to different communities are rendered with different colours. Normally, there are holes in this grid, since these cells don't contain any human movement. The phenomenon is more evident for smaller resolutions, as many small cells don't contain roads, streets, or any pedestrian walking way.

The spatial resolution of aggregation is necessary to define a finite set of places visited by pedestrians, focused in description of the size of the regions of interest for which data are aggregated before being shared. The recorded pedestrian positions can hereby be distinguished in two different scenarios:

1. limited in a finite set of predefined positions (e.g, cells of a mobile phone network or positions of sensors);
2. arbitrary positions received from mobile devices and capable of measuring absolute spatial positions (e.g, GPS devices).

Regardless of the size of the dataset, the different positions should be aggregated. In scenarios with large data, the spatial aggregation should across all users in an area of interest (e.g, city or neighbourhood). Moreover, spatial data may give the required set of points for aggregation. Often they are aggregated using arbitrary divisions (e.g, administrative districts and regular grids), which do not respect the spatial distribution of the data. Another method divides the spatial data into convex polygons of approximately equal size based on the places [22].

2.2 Temporal resolution of categorisation and aggregation

In temporal resolution of categorisation the generated demand data is time-dependent [6]. It aims to compute for each user the regularity, i.e., observe the data volume generated by the user during an interval time basis (e.g, hourly, daily, monthly, etc) [5]. Moreover, it provides an opportunity to categorise every user's location. For example, the modal location that a user logs data in every hour. Similarly, in order to take full advantage of this resolution, we should quantify the data observed to estimate human movements demand. This method must balance trade offs between small, but complete data for a short period of time and large, but incomplete data over a longer period of time. In both cases noise and biases must be carefully dealt with to produce valid measurements.

In temporal resolution of aggregation, it is important to address issues about privacy challenges. In practice, it is often difficult to get access to this kind of data for research purposes, privacy concerns being a major hindrance. Use of aggregation by temporal resolution protects peoples' identities [8]. Simultaneously, this kind of aggregation describes the time window for which data is aggregated for all users based on a specific region. For example, we might be interested in the average number of locations visited by users in a specific region over the course of an interval time window. However, human mobility data is scarce for developing countries. Understanding the limitations of using time-averaged human mobility data and model approximations to human movement is essential to guide the choice of mobility measures in models and further development in this field.

3 Detailed Analysis of Human Mobility Patterns

In this section, we outline considerations for analysing aggregated data from several devices, including representativeness, situational context, and modelling approaches to capture human mobility. These devices gathered four main types of data: Wi-Fi, GPS, CDR and Social Media. Then we define new detailed metrics that can be applied in human motion analysis. Through analysis of these metrics and data fitness, we obtain important conclusions about human mobility phenomena.

3.1 Data Collection

Wi-Fi, GPS, CDR and Social Media datasets collect traces of the different movements of people, thus producing a network of people who are moving in different places, trajectories, etc. In the case of Wi-Fi, we use the historical listing of LinkNYC Kiosks that save their location, and the status of the Link's wifi, tablet, latitude, longitude, geometry point, type of kiosk, registry date and phone [16]. These devices are connected daily for thousands of users within New York City. Although, the available data was collected between January 2012 and December 2013, we consider only the user interactions from January 2013 to October 2013.

We also study the GPS dataset using a dataset that was obtained through a Freedom of Information Law (FOIL) request from the New York City Taxi & Limousine

Commission (NYCT&L). The NYCT&L display trip data of Uber pickups in New York City [20]. This dataset consists of over 4.5 million Uber pickups from April 2012 to September 2013, and 14.3 million more Uber pickups from January to June 2013. For this article, we use the data from January to June 2013.

Another dataset that we use is CDR dataset. It is associated with a phone that made a voice call with the locations (latitude and longitude) of the cellular towers and reconstructs the spatio-temporal data of a mobile phone. Our research use 311 calls or service requests in New York City from January 2011 to December 2014 [18].

Finally, Gowalla is a location-based social networking website where users share their locations by checking-in. The friendship network is undirected and was collected using their public API [23], and consists of 196,591 nodes and 950,327 edges. We have collected a total of 6,442,890 check-ins of these users over the period of February 2009 - October 2010. However, this dataset stores a huge amount of data and we only consider data from March 2010 to October 2013.

3.2 Exploring Human Mobility Metrics

As we mentioned, we can illustrate the differences in the human mobility behaviors of individuals based on datasets represented in Subsection 3.1. However, we need to refine our metrics. In other words, we can use data types and suitable methods, so that we know the influence between data type and metrics in human movements aspects. Therefore, we leverage four metrics to analyze human mobility: Displacement, Resilience, Duration and Interval.

3.2.1 Displacement

To the best of our knowledge, Haversine formula, shown in Eq. 1, is suitable for human mobility analysis because given a pair of points (latitude and longitude), this formula calculates the great-circle distance between two points within a time interval.

$$d_{ij} = 2r \times sin^{-1}\left(\sqrt{sin^2\left(\frac{\Phi_2 - \Phi_1}{2}\right) + cos(\Phi_1)cos(\Phi_2)sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \qquad (1)$$

where r is the radius of earth (6371 km), d is the distance between two points, $\phi_1 = \phi_{i+1,j}$, $\phi_2 = \phi_{i,j}$ is latitude of the two points and $\lambda_1 = \lambda_{i+1,j}$, $\lambda_2 = \lambda_{ij}$ is longitude of the two points respectively.

3.2.2 Resilience

Previously, several concepts of resilience have been proposed. [12] have reviewed the methods of defining and quantifying resilience in various fields. In its turn, [2] developed a framework for measuring resilience considering four dimensions: (i) robustness reflecting the strength or ability of the system to reduce the damage; (ii) rapidity representing the rate or speed of recovery; (iii) resourcefulness reflecting the ability to apply materials and human resources by prioritizing goals when an event occurs; and (iv) redundancy representing the capacity to achieve goals by prioritizing objective to restrain loss and future disruptions. Here, using Bruneau's approach, we calculate the loss of resilience which is equivalent to the size of dysfunction/degradation of human mobility.

$$R_L = \int_{t_0}^{t_1} [100 - Q(t)]dt \tag{2}$$

where, R_L denotes Resilience Loss, $Q(t)$ denotes a Quality function for infrastructure service at time t and $(t_1 - t_0)$ is the recovery time. However, this formula needs many variables which are difficult to collect in the context of human mobility because it involves a large geographical area. In addition, measuring resilience, in a mobility context, has been difficult due to the lack of appropriate metrics over longer time periods.

3.2.3 Interval

The interval is the time spanning the two reference points. An interval of a year, for example, means that we are interested in whether an individual's location at time t is different than their location at time $t + Xyear$. In order to avoid double counting observations, we can constrain the duration so that it is always less than or equal to the interval. So in the case where we want to know the human mobility rate given an interval of X and a duration of Y (X and Y are measured by time units), we infer each individual location at time t using the time frame $[(t\breve{}Ydays), (t+Ydays)]$ and at time $t + Xyear$ using the time frame $[(t + Xyear\breve{}Ydays), (t + Xyear + Ydays)]$ and then we can see if the two locations are the same.

3.2.4 Duration

Duration (the length of time spent in a destination) is another important evaluation metric and closely tied to trip displacement. Studies investigated the spatial distribution of human mobility conditioned on trip duration and found distinct differences between short and long duration trips. In short-trip duration travel networks, trips are skewed towards urban destinations, compared with long-trip duration networks where human mobility is more evenly spread among locations. To perform trip duration calculation we should consider the Eq.3.

$$w'_{ij} = [(w_{ij}|t_1, t_2)/max(w_{ij})] \times 100 \quad (3)$$

Each sub-network contained only trips where duration of travel (t) was within the interval $[t_1, t_2]$ and is scaled by the overall maximum edge weight observed in the full travel network $max(w_{ij})$. However, network topology changes on account of trip duration. To further explore this phenomenon, we can use a subset data into X duration-restricted subnetworks reflecting intervals of trip duration with duration intervals continuing up to Y days, months, years, etc.

4 Results and Discussion

Based on the set of data types represented in Subsection 3.1 we calculate the displacement, resilience and duration metrics. However, firstly, we define some requirements. In this work, we mine the data to avoid privacy issues. All declared personal information of users was anonymized to ensure privacy protection. Then, we remove irrelevant attributes, focusing only on essential attributes in our research such as geo-locations, register data and username (anonymized attribute). Another factor we take into account was to filter data. Although, we have chosen data within a specific time interval, to simplify the analysis, we consider the data regarding 1 day (14 June 2012) and consider only 100 coordinates of two users (User A and User B), discarding the rest of the line informations. Then, we constructed a API in Python, so that to calculate and compare the metrics presented in Subsection 3.2 using the pre-processed data. Table 1 provides a summarized comparison outcome of different metrics in data types.

Table 1: Different metrics in human mobility analysis.

Data Type	User	Displacement (km)	Resilience (%, hours)	Interval (hours)	Duration (hours)
Wi-Fi	A	2.29	Min(15, 2) & Max(70, 4)	0.75	1
	B	1.36	Min(5, 3) & Max(64, 5)	0.4	1
GPS	A	36.10	Min(13, 2) & Max(30, 2)	2.2	8
	B	16.70	Min(24, 2) & Max(15, 3)	1.7	4
CDR	A	10.10	Min(21, 1) & Max(78, 2)	0.4	1.1
	B	5.51	Min(1, 1) & Max(42, 2)	0.1	0.1
Social Media	A	0.7	Min(22, 5) & Max(84, 7)	0.1	0.1
	B	3.1	Min(11, 6) & Max(90, 8)	0.2	0.2

Observing the results, in Wi-Fi dataset the user A presents a major displacement within same interval than user B. Moreover, if we compare the behaviour of these users with other dataset, the Wi-Fi presents the second lowest displacement measurement. This can mean that recording data from wi-fi is more abundant than from other data sources. In fact, LinkNYC platforms are accessible and spread throughout many places in New York City, allowing users to be connected anytime and anywhere. However, Social Media dataset shows the lowest displacement. Abundance of data and close displacement points in a small space are key factors that contribute to these results. Moreover, alongside CDR dataset, it can be used to track and identify migration patterns at the sub-regional level or city. On the other hand, greatest displacement values were recorded in users in the GPS dataset. Based on them, we can conclude that this data type can help to assess human mobility movements in country-level.

We also quantify the loss of resilience using these four datasets. Since resilience changes over time, the analysis of this metric was carried out based on graphs. Due to space restrictions we decided to present the maximum and minimum resilience values for each user in function of the data type. These graphs show that resilience results vary from individual to individual. The *min* and *max* values are not similar in the datasets.

Interval metric is another important human mobility metric. It presents different values (hour) to understand the average time spanning the two reference points of two users. The availability of devices in the capture of user data has an influence on the displayed values. If we compare the interval time to displacement, versus source data location, LinkNYC (Wi-Fi dataset), Uber pickups (GPS dataset) and Gowalla (Social Media dataset) present low average values.

To explore the impact of trip duration on human mobility patterns, we assessed travel by trip duration of four datasets. The short-duration trips can be characterized by a high proportion of trips to nearby and densely populated areas in Wi-Fi, CDR and Social media. These patterns can change incrementally when assessed over a more continuous set of duration-restricted subnetworks reflecting long-duration trips such as we can interpret based on the values presented of GPS dataset.

5 Conclusions

In this work, we propose a set of metrics to analyze the characteristics of human mobility in different data types. In the initial phase, the approach utilizes common metrics in human mobility research to fit the patterns of human movement. Then we reviewed the state of the art in these metrics and illustrated their characteristics and limitations. Furthermore, we explored seven basic types of human mobility metrics and discussed their principles.

We also discussed data types based on different sources that have been used on several human mobility surveys. Then, a range of real data sets (i.e Wi-Fi, CDRs, GPS and Social Media) were analyzed based on a spectrum of mathematical models. We did a summary of different data types in human mobility analysis where proposed datasets present privacy issues. On the other hand, a ubiquitous occurrence of mobile devices and social media has made human behaviors more closely linked to social networks, maybe sparking a new driving force for modeling human mobility. Although this new data source has challenges in extracting useful insights, social media data is the most promising in the literature of human mobility.

Finally, we explore other metric trends to understand the dynamical behaviors of human mobility. Through the integration of displacement, resilience, duration, and interval metrics in data, this data may provide valuable insights for policy decision-makers. We also described the characteristics of these metrics and how they contribute to modeling human mobility. We hope that they enable a deeper understanding of the literature of human mobility and give rise to new research interests in this field.

In future work, we plan to introduce more users/individuals records in our research, investigating urban community patterns. Although, we only use human mobility metrics on the individual level, we should analyze human mobility movements adding a larger number of individuals in order to reach crowd-level mobility data. This kind of data provides a better understanding of human mobility phenomena, helping create policies aimed to build resilient communities. Additionally, other metrics can be used to research human movements with accuracy. For example, using CDR data to confirm individuals' movements that can be detected by analyzing the footprint left by the user in the places they called from. We can choose an algorithm cluster the cell towers points, which were first sorted in a descending order. Datasets with multiple attribute such as Wi-Fi and CDR dataset can use Deep Learning techniques. In its turn, an approach usually applied in Social Media data is cluster techniques. They are formed by connecting neighboring grid points that have a relative duration larger than a threshold percentage. Therefore, appropriate strategies or techniques to manage the Wi-Fi, GPS, CDR and Social Media datasets can be useful on understanding dynamic human mobility from individual-level to community-level.

Acknowledgments

This work has been supported by FCT - Fundacao para a Ciencia e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020. It has also been supported by national funds through FCT – Fundação para a Ciência e Tecnologia through project UIDB/04728/2020.

References

1. Atalan, A.: Is the lockdown important to prevent the COVID-9 pandemic? Effects on psychology, environment and economy-perspective. Annals of Medicine and Surgery 56, 38–42 (aug 2020)
2. Bruneau, M., Chang, S.E., Eguchi, R.T., Lee, G.C., O'Rourke, T.D., Reinhorn, A.M., Shinozuka, M., Tierney, K., Wallace, W.A., Von Winterfeldt, D.: A Framework to Quantitatively

Assess and Enhance the Seismic Resilience of Communities (nov 2003)

3. Centre's, I.D.M.: Data on human mobility in disaster contexts: Where are we and what comes next? (2020), https://www.internal-displacement.org/events/data-on-human-mobility-in-disaster-contexts-where-are-we-and-what-comes-next
4. Cornacchia, G., Rossetti, G., Pappalardo, L.: Modeling Human Mobility considering Spatial, Temporal and Social Dimensions (jul 2020)
5. van Duynhoven, A., Dragićević, S.: Analyzing the effects of temporal resolution and classification confidence for modeling land cover change with long short-term memory networks. Remote Sensing 11(23) (dec 2019)
6. Ebrahimpour, Z., Wan, W., García, J.L.V., Cervantes, O., Hou, L.: Analyzing social-geographic human mobility patterns using large-scale social media data. ISPRS International Journal of Geo-Information 9(2), 125 (feb 2020)
7. Esteban Ortiz-Ospina: How do people across the world spend their time and what does this tell us about living conditions? (2020), https://ourworldindata.org/time-use-living-conditions
8. Falconi, T.M., Estrella, B., Sempértegui, F., Naumova, E.N.: Effects of data aggregation on time series analysis of seasonal infections. International Journal of Environmental Research and Public Health 17(16), 1–21 (aug 2020)
9. Feng, X., Li, J.: Evaluation of the Spatial Pattern of the Resolution-Enhanced Thermal Data for Urban Area. Journal of Sensors 2020 (2020)
10. Forde, J., Hopfe, C.J., McLeod, R.S., Evins, R.: Temporal optimization for affordable and resilient Passivhaus dwellings in the social housing sector. Applied Energy 261, 114383 (mar 2020)
11. Hadjidemetriou, G.M., Sasidharan, M., Kouyialis, G., Parlikad, A.K.: The impact of government measures and human mobility trend on COVID-19 related deaths in the UK. Transportation Research Interdisciplinary Perspectives 6, 100167 (jul 2020)
12. Hosseini, S., Barker, K., Ramirez-Marquez, J.E.: A review of definitions and measures of system resilience. Reliability Engineering and System Safety 145, 47–61 (jan 2016)
13. Huang, Q., Huang, R., Hao, W., Tan, J., Fan, R., Huang, Z.: Adaptive Power System Emergency Control Using Deep Reinforcement Learning. IEEE Transactions on Smart Grid 11(2), 1171–1182 (mar 2020)
14. Kévorkian, A., Grenet, T., Gallée, H.: Tracking the Covid-19 pandemic: Simple visualization of the epidemic states and trajectories of select European countries & assessing the effects of delays in official response (2020), https://hal.archives-ouvertes.fr/hal-03065691
15. Kulkarni, V., Mahalunkar, A., Garbinato, B., Kelleher, J.D.: Examining the limits of predictability of human mobility. Entropy 21(4) (2019)
16. LLC, I.C.c.Q.C.S.: LinkNYC Kiosks: Free super fast Wi-Fi and that's just the beginning. (2021), https://www.link.nyc/
17. Mohamed, M.A., Chen, T., Su, W., Jin, T.: Proactive Resilience of Power Systems against Natural Disasters: A Literature Review. IEEE Access 7, 163778–163795 (2019)
18. NYC Open Data: 311 Service Requests from 2010 to Present (2021), https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/7ahn-ypff
19. Schwerdtle, P.N., Bowen, K., McMichael, C., Sauerborn, R.: Human mobility and health in a warming world (jan 2019)
20. Shikun, L.: Uber Pickups in New York City (2016), https://www.kaggle.com/fivethirtyeight/uber-pickups-in-new-york-city
21. Sun, H., Zhen, F., Jiang, Y.: Study on the characteristics of urban residents' commuting behavior and influencing factors from the perspective of resilience theory: Theoretical construction and empirical analysis from Nanjing, China. International Journal of Environmental Research and Public Health 17(5) (mar 2020)
22. Yin, Z., Jin, Z., Ying, S., Li, S., Liu, Q.: A spatial data model for urban spatial–temporal accessibility analysis. Journal of Geographical Systems 22(4), 447–468 (oct 2020), https://doi.org/10.1007/s10109-020-00330-6
23. Zitnik, M., Sosi, R., Maheshwari, S., Leskovec, J.: SNAP: Network datasets: Gowalla (2014), https://snap.stanford.edu/data/loc-gowalla.html