Multi-scale framework for attraction place identification across human and animal mobility

Maria Luisa Damiani
Dept. Computer Science
University of Milan, Italy
maria.damiani@unimi.it

Fatima Hachem
Dept. Computer Science
University of Milan, Italy
fatme.hachem@unimi.it

Sabrina Gaito
Dept. Computer Science
University of Milan, Italy
sabrina.gaito@unimi.it

ABSTRACT

A large volume of research has been devoted to the concept of mobility, a theme that is transversal to multiple fields of study and applications, from biology and ecology, to computer science and economy. Despite the considerable efforts made by a few scientific communities and the relevant results obtained so far, only very recently the issue of adopting a unifying and comprehensive approach across has been faced. In this paper, we present an overview of the quantitative framework we have recently developed to analyse and model human mobility, based on symbolic trajectories built on CDR data (Call Detail Record) provided by a telco operator [5]. Driven by a location-centric perspective, the framework includes a novel trajectory summarization technique for the extraction of the locations of interest from symbolic trajectories, a *relevance analysis* providing a novel location taxonomy and, inspired by ecological studies, a *diversity analysis* to characterize the movement through a *location diversity profile*. The ultimate goal of this contribution is to stress the need of a methodological integration between human and animal mobility analytical methods.

CCS CONCEPTS

• **Information systems** → **Spatial-temporal systems**; **Data mining**.

KEYWORDS

Human mobility, location diversity, trajectory segmentation

ACM Reference Format:
Maria Luisa Damiani, Fatima Hachem, and Sabrina Gaito. 2021. Multi-scale framework for attraction place identification across human and animal mobility. In *1st ACM SIGSPATIAL International Workshop on Animal Movement Ecology and Human Mobility (HANIMOB'21), November 2, 2021, Beijing, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3486637.3489488

1 INTRODUCTION

Understanding how living beings move on earth at different spatial and temporal scales has long been a challenge from both a theoretical and an application point of view. In fact, the description of human mobility is the basis of many industrial and commercial sectors such as next generation cellular networks, transport, urban planning, migration control, epidemic control and much more. On the other hand, the knowledge of animal mobility is proving to be more and more important and it has been evident to everyone, especially due to the current COVID-19 pandemic, how human and animal mobility are interconnected. There has been a great deal of research conducted by many different scientific communities that have undoubtedly shed light on many relevant patterns. However, just a few, very recent works are dedicated to unifying and comprehensive approaches on mobility [7]. In addition to the inherent complexity of the topic, other factors make it difficult: publicly available datasets are few and of very different sources; several scientific communities have developed specific methodologies and the hybridization between them is just starting. Here, we propose a quantitative framework to model and analyse mobility. Results on human mobility as described by telco data are reported to show the efficacy of the framework in summarizing trajectories and providing useful metrics to characterize mobility behaviors.

HANIMOB'21, November 2, 2021, Beijing, China

ACM ISBN 978-1-4503-9122-1/21/11…$15.00
https://doi.org/10.1145/3486637.3489488

2 TELCO DATA

Mobility data commonly take the form of trajectories reporting the individual location history. From a data modeling perspective, an important class of trajectories for the study of human mobility are the *symbolic trajectories*. Unlike *spatial* trajectories, leveraged both in human and animal mobility studies, where location data is collected at very fine temporal scale over a continuous space and thus the missing points in between consecutive samples can be estimated by interpolation, symbolic trajectories are defined over a discrete space. Locations are spatially sparse and temporally irregular, therefore the sequence cannot be modeled as a continuous trajectory or symbolic time series. A class of trajectories of major importance for the study of human mobility, which can be readily modeled as symbolic trajectories, are those built on CDRs of mobile phones. CDRs report the communication activities of mobile communication subscribers as series of geo-referenced *events*, i.e., voice call start/end, text message, data upload/download, collected by mobile operators for billing purposes, as shown in Figure 1. The importance of CDR data (*game-changing data in the last decade* according to [1]), is substantially due to the significance of the user base, i.e., very large sets of individuals monitored in their daily life. Telco trajectories have some important characteristics:

- Sequences of identical locations. Locations denote regions of space. Therefore, as the user's position is matched against the closest base station, it may happen that consecutive locations are identical. For example, a phone call started and ended at home or in its proximity will generate two records reporting

the same location. Notably, that does not happen in other kinds of trajectories, such as GPS and trajectories of check-in data, where consecutive locations are very unlikely identical, either for technological reasons (signal characteristics) or for the nature of movement (e.g. a check-in is typically performed once).

- CDRs are only generated when phones are actively involved in a voice call, text message or internet access. Therefore large temporal gaps exist between consecutive locations. Moreover, trajectories can contain bursts of events, often related to user's activity on the internet (data upload and download), possibly interleaved by long periods of inactivity. The result is a highly inconsistent temporal frequency, which may confound the mobility analysis [1].
- The locations reported in CDRs can be noisy because of signal fluctuation in the network coverage [3] which can lead to ping-pong handovers between neighboring cells [8]. In addition, users can experience brief absences from the locations where they regularly stay, e.g. home. [4].

Source	Date	Time	Location area
1553655	26/03/2012	00:02:46	PIOLA
1553655	26/03/2012	00:02:46	PIOLA
1553655	26/03/2012	00:45:27	VALLAZZE
1553655	26/03/2012	15:38:19	BUENOS AIRES PONCHIELLI

Figure 1: A fragment of CDR data, combining call, SMS and internet records for a single user.

3 THEORETICAL AND ANALYTICAL FRAMEWORK

An important feature of human mobility is related to the number of locations a person visits, namely, stationary people tend to frequent few locations, while people with high propensity to mobility likely visit a higher number of locations. However, locations are not equally significant,e.g., some locations are accidental, frequented by chance, or represent noise, and thus negligible. Therefore, simply counting the number of locations may result in a coarse measure. A different direction is to quantify the locations that appear of major relevance for the user. To address the above challenges, we develop an analytical framework supporting location *relevance* analysis and *diversity* analysis. *Relevance analysis* targets the discovery and classification of attractive locations from telco trajectories. Key component is a cluster-based trajectory segmentation technique, conceptually rooted in [4] and tailored to symbolic trajectories. The outcome is a set of *summary trajectories*. The term 'summary' is used to emphasize that the purpose of the method is to extract key symbols from symbolic trajectories, in analogy with text summarization methods in information retrieval. Locations are then classified based on the two criteria of attractiveness and frequency. *Diversity analysis* targets the quantification of location diversity in summary trajectories. The approach relies on the use of different entropy-based metrics, enabling the homogeneous quantification of location diversity in terms of 'number of types'. These metrics are used to characterize the individual mobility behavior from a location perspective.

Figure 2 displays the functional architecture of the framework. It consists of two building blocks supporting relevance and diversity analysis, respectively. In particular, given a dataset of telco trajectories, relevance analysis accomplishes two main tasks: (a) identification of key locations through trajectories summarization. The quality of clustering is assessed using internal indicators. (b) Classification of relevant locations based on frequency and attractiveness. The second building block is to characterize every single trajectory as a whole using the set of entropy-based indicators, specifically the location diversity profile and entropy rate estimation. The metrics are used for the classification of trajectories at population scale.

3.1 Trajectory summarization algorithm

The approach takes inspiration from SeqScan [4]. SeqScan partitions a spatial trajectory in a series of temporally ordered clusters of arbitrary shape interleaved by sequences of unstructured points called *transitions*. The points that do not belong to any cluster or transition are classified as *local noise*. SeqScan is, however, tailored to spatial trajectories, while proven ineffective when applied to telco trajectories, therefore we have investigated a different strategy. We note that a trajectory typically contains sequences of identical locations, meaning that the user is located in the same region at different times while making a phone call or accessing the Internet. As the user moves elsewhere, two cases can occur: the user returns back to the previous location; or the user starts frequenting some other location. This suggests a cluster-based segmentation performed over the temporal line with clusters only grouping occurrences of a unique location. This method is called *SeqScan-d* [1]. An example is shown in Figure 3. The quality of summarization is assessed using the internal indicator proposed in [6], and the summarization rate S_{rate} expressing the percentage of irrelevant locations in the trajectory; a high S_{rate} means a high level of summarization.

3.2 Insights into the nature of locations: the proposal of a location taxonomy

Trajectory summarization is applied to the extraction of attractive locations from telco trajectories. We now turn to discuss the nature of the relationship between location frequency and attractiveness.

The frequency of visiting a location in fact accounts for how many times a person is there but is unable to give any information on the attractiveness that the location has on it. There may be locations, such as the bus stop, which are often visited by a person who has no specific interest in it, but just transits through it. On the contrary, we find examples such as a museum, a place that is rarely visited but that expresses a strong cultural attraction, a place therefore detected as attractive by SeqScan-d, even if visited only sporadically. The synergy between the two metrics reinforces the role that a location plays in a person's life: locations that are both attractive and frequent are clearly very significant, while non attractive and infrequent ones are non-significant. Thus, by combining these two dimensions, we can gain further insights into the nature of the visited locations. In particular, we distinguish four classes of locations that we label: *significant (SL), transit (TL), sporadic (PL), insignificant (IL)*, respectively (in Figure 4).

[1] https://github.com/SeqScan/SeqScan-D

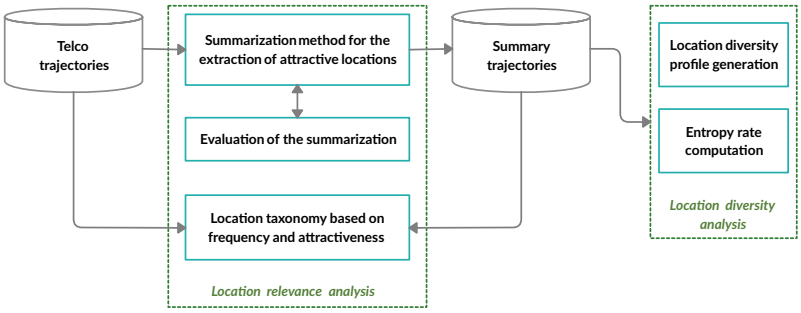

Figure 2: General architecture of the analytical framework: the datasets and the two building blocks for relevance and diversity analysis

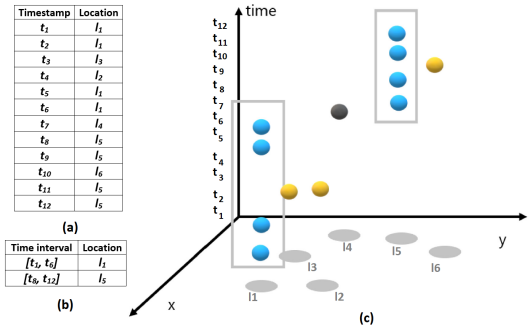

Figure 3: (a) A trajectory with 6 different locations (on space) and 12 occurrences (in space-time); (b) the summary trajectory, output of SeqScan-d; (c) Trajectory representation in space-time: the rectangles contain the clusters along the temporal line (output of SeqScan-d); noisy points (yellow), a transition point (black).

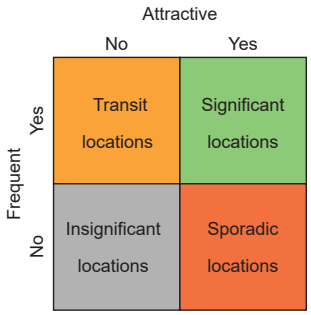

Figure 4: Location taxonomy

The semantics of these classes can be straightforwardly defined in set-based terms. Consider a trajectory T and its summarization $\widehat{T}$ denote with T_u and $\widehat{T}_u$ the two sets of distinct locations in T and $\widehat{T}$, respectively. Moreover, denote with N_u the top-n frequented locations in T, where $n = |\widehat{T}_u| = |N_u|$.

- *Significant* locations are those that are both frequent and attractive, i.e., $SL = N_u \cap \widehat{T}_u$;
- *Transit* locations are those that are frequent but not attractive, for example the locations where the user passes quickly daily, i.e., $TL = N_u \setminus \widehat{T}_u$;
- *Sporadic* locations are those that are attractive but not frequent, for example, the locations hosting some event of interest for the user, e.g., $PL = \widehat{T}_u \setminus N_u$;
- *Insignificant* locations are those that are neither attractive or frequent, i.e., $IL = T_u \setminus (\widehat{T}_u \cup N_u)$.

As a result, given an input trajectory T and its summarization $\widehat{T}$, we can label each distinct location in T, based on the proposed taxonomy.

3.3 Location Diversity Analysis

A natural property of locations is their diversity. *Diversity metrics* quantify the heterogeneity of a set consisting of elements of different type (*population*). While a simple measure is the count of types, more sophisticated metrics are used in practice. The notion of diversity is key in innumerable fields, including biology, economy, demography, information theory. For example, diversity can be used in ecology to measure the biodiversity of a geographical area, i.e. diversity of species; in economy, the economic variety of a region, i.e. diversity of companies with respect to their products; in demography, the racial heterogeneity in various regions. In this work, we employ the concept of diversity to characterize the heterogeneity of relevant locations in a trajectory. We refer to that measure as *location diversity*. As the number of locations as well as the number of visits vary significantly among individuals, we hypothesize that location diversity can act as discriminant for the categorization of mobility behaviors. Diversity measures provide complementary information on mobility. Since all of them are relevant, we follow the approach presented by ecologists in [2], and introduce the notion of *location diversity profile*. The location diversity profile provides a multi-level characterization of location heterogeneity through a multiplicity of diversity indicators, globally enhancing the interpretation of the user behavior. The set of locations in a summary trajectory represents the population of concern, where locations have a type and a number of occurrences. We consider the three indices: R, indicating the location richness; the true location diversity of order 1, based on Shannon Entropy, denoted TD_H; and the true location diversity of order 2, the inverse of the Simpson index, denoted TD_S.
The location diversity profile of trajectory i is defined by the triple:

$$Pr(i) = (R^i, TD_H^i, TD_S^i) \quad (1)$$

The diversity indices presented in the location diversity profile are all based on either the number of unique visited locations or their relative frequency, so that – for example – the two trajectories $\widehat{T}_1 = a, a, a, b, b, c$ and $\widehat{T}_2 = a, b, a, c, a, b$ have the same location diversity profile. Clearly these two trajectories convey different information about the mobility of the user they refer to; in other words, intuition suggests that, in addition to the frequency of visitation, the *order* in which locations are visited may also be a relevant property in the assessment of the mobility behavior of an individual.

To this intent, we consider the *entropy rate* as another useful indicator of users' mobility. Entropy rate is a well-known indicator and has already been used in [9] to estimate an upper bound of predictability in human mobility. For further details, we refer the reader to [5].

4 RESULTS

Dataset. The dataset consists of 100,000+ telco trajectories in the area of Milan, of various length and duration, over a period of 67 days. The total number of samples in the dataset amounts to about 54 million points. The *telco space* consists of 685 locations, identified by a label. Table 1 reports the summary statistics on the number of trajectories (i.e. users), number of records, average and standard deviation of trajectory length.

Table 1: Summary statistics of the dataset

# Traj	# Records	# Loc	Avg(trj_len)	Std(trj_len)
104,413	54,193,257	685	3151	1650

Data summarization. SeqScan-d is applied on the dataset for the extraction of attractive locations. SeqScan-d is capable of identifying key locations for each user and dispensing with the irrelevant ones, despite the high noise and uncertainty of the data. In general, the summarization rate S_{rate} is quite high, with the percentage of irrelevant locations (types) varying approx from 40% to 90%.

Location taxonomy. We conduct an experiment to classify the location based on the taxonomy presented in Section 3.2. It can be seen that:

- The vast majority of locations in native trajectories are classified as *insignificant*. The percentage varies from approx. 67% to 84%. Hence, this result is in agreement that around 70% of the locations in the native trajectory are not important for the user.Therefore it is substantially in line with the literature.
- The percentage of *significant* locations, i.e. frequent and attractive, computed for the whole set of locations visited by an individual varies between 9% and 19%.
- Transit (frequent but not attractive) and sporadic (infrequent but attractive) locations, together, account for 7-14%, on average, of the locations.

Diversity profile. The summary statistics for the variables of the profile are reported in Table 2. It can be seen that, on average, a trajectory contains about 15 relevant locations, while the true diversity of order 1 (TD_H) amounts to about 7 locations, and the true diversity of order 2 (TD_S) to about 4 locations. There is thus a significant gap between the values of R and TD_x, indicating that locations are not evenly frequented. Moreover, the diversity of core locations is quite small.

Table 2: Summary statistics on the variables of the location diversity profiles.

Metric	Max	Min	Avg	Std
R	108	1	15.61	8.87
TD_H	91	1	6.95	3.84
TD_S	74.6	1	4.49	2.41

Entropy rate. The mean value of the entropy rate of the summary trajectories is 1.97 bits; the most complex trajectory has an entropy rate of 5.11 bits.

5 CONCLUSIONS

This work focuses on a possible characterization of human mobility, complementary to the prevailing models grounded on statistical mechanics. In particular, inspired by ecological concepts, we try to characterize mobility in terms of location diversity. Moreover, since locations are not equally relevant, we apply the analysis to those locations that are more intensively frequented. An interesting issue we are working on is whether our notion of mobility profile can be effective in partitioning the user population in homogeneous groups.

REFERENCES

[1] Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R. James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J. Ramasco, Filippo Simini, and Marcello Tomasini. 2018. Human mobility: Models and applications. *Physics Reports* 734 (2018), 1–74.

[2] Anne Chao, Nicholas J. Gotelli, T C. Hsieh, Elizabeth L. Sander, K H. Ma, Robert Colwell, and Aaron M. Ellison. 2014. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs* 84, 1 (2014), 45–67.

[3] Balázs Csáji, Arnaud Browet, Vincent Traag, Jean-Charles Delvenne, Etienne Huens, Paul Van Dooren, Zbigniew Smoreda, and Vincent Blondel. 2013. Exploring the mobility of mobile phone users. *Physica A: statistical mechanics and its applications* 392, 6 (2013), 1459–1473.

[4] Maria Luisa Damiani, Fatima Hachem, Hamza Issa, Nathan Ranc, Paul Moorcroft, and Francesca Cagnacci. 2018. Cluster-based trajectory segmentation with local noise. *Data Mining and Knowledge Discovery* 32, 4 (2018), 1017–1055.

[5] Maria Luisa Damiani, Fatima Hachem, Christian Quadri, Matteo Rossini, and Sabrina Gaito. 2020. On Location Relevance and Diversity in Human Mobility Data. *ACM Trans. Spatial Algorithms and Systems* 7, 2, Article 7 (2020), 38 pages.

[6] Maria Luisa Damiani, Hamza Issa, Guiseppe Fotino, Marco Heurich, and Francesca Cagnacci. 2016. Introducing 'presence'and 'stationarity index'to study partial migration patterns: an application of a spatio-temporal clustering technique. *International Journal of Geographical Information Science* 30, 5 (2016), 907–928.

[7] Urška Demšar, Jed A. Long, Fernando Benitez-Paez, Vanessa Brum Bastos, Solène Marion, Gina Martin, Sebastijan Sekulić, Kamil Smolak, Beate Zein, and Katarzyna Siła-Nowicka. 2021. Establishing the integrated science of movement: bringing together concepts and methods from animal and human movement analysis. *International Journal of Geographical Information Science* (2021), 1–36.

[8] Lars K. Rasmussen and Ian Oppermann. 2003. Ping-pong effects in linear parallel interference cancellation for CDMA. *IEEE Transactions on Wireless Communications* 2, 2 (2003), 357–363.

[9] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010. Limits of Predictability in Human Mobility. *Science* 327, 5968 (2010), 1018–1021.