Location relevance and diversity in symbolic trajectories with application to telco data

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

Fatima Hachem
Dept. Computer Science, University of Milan (I)
fatme.hachem@unimi.it

Christian Quadri
Dept. Computer Science, University of Milan (I)
christian.quadri@unimi.it

Sabrina Gaito
Dept. Computer Science, University of Milan (I)
sabrina.gaito@unimi.it

ABSTRACT

We present an approach to the discovery and characterization of *relevant* locations and related mobility patterns in symbolic trajectories built on call detail records - CDRs - of mobile phones (telco trajectories). While the discovery of relevant locations has been widely investigated for continuous spatial trajectories (e.g., stay points detection methods), it is not clear how to deal with the problem when the movement is defined over a discrete space and the locations are symbolic, noisy and irregularly sampled, such as in telco trajectories. In this paper, we propose a methodological approach structured in two steps, called *trajectory summarization* and *summary trajectories analysis*, respectively, the former for removing noise and irrelevant locations; the latter to synthesize key mobility features in a few novel indicators. We evaluate the methodology over a dataset of approx 17,000 trajectories with 55 million points and spanning a period of 67 days. We find that trajectory summarization does not compromise data utility, while significantly reducing data size. Moreover, the mobility indicators provide novel insights into human mobility behavior.

CCS CONCEPTS

• **Information systems → Data management systems**; **Spatial-temporal systems**; **Clustering**.

KEYWORDS

Symbolic trajectories, cellular data, spatio-temporal clustering

1 INTRODUCTION

Trajectories are key resources for the fine-grained analysis of moving objects behavior. Whenever data are collected at very fine temporal scale over a continuous space, the missing points in between consecutive samples can be estimated by interpolation, and the trajectory be qualified as *continuous*. Opposed to continuous trajectories, *symbolic* trajectories are defined over a discrete space consisting of a finite set of points $P = \{p_1, ..p_k\}$ in the embedding space. Given a set of symbols L, and a bijection $m : L \rightarrow P$, a trajectory is a sequence of timestamped symbolic locations

$$T = (l_1, t_1), .., (l_n, t_n), \text{ with } l_i \in L, m(l_i) \in P$$

Commonly, locations are spatially sparse and irregularly sampled, therefore the movement cannot, arguably, be represented by a continuous trajectory or, alternatively, a symbolic time series. An example is the trajectories based on check-in data from geo-social networks.

In this paper, we are concerned with the class of symbolic trajectories built on call detail records (CDR) of mobile phones. CDRs report the communication activities of the subscribers as series of *events*, i.e., voice call start/end, text message, data upload/download, collected by mobile operators for billing purposes. Abstractly, a CDR can be represented by the tuple *(u, e, t, l)*, reporting, in order: the user identifier, the kind of event (e.g. start phone call), the timestamp and the position of the event in the space determined by the position of the cellular network components that are pinged when a phone call is made, e.g. base stations. If we omit the information on the kind of event - marginal for our study - the series of CDRs for a user u, over the observation period $[t_1, t_n]$, can be rewritten as a symbolic trajectory. We refer to this kind of trajectory as *telco trajectory*.

Telco trajectories are complex, noisy and irregularly sampled data, yet of fundamental importance for the study of human mobility [3]. One of the analytical tasks of major interest is to infer the locations that are *relevant* for an individual or community. In the area of telco data analysis, the notion of location relevance is given a statistical meaning, i.e., the relevant locations are those that are frequently visited. For example, in the seminal work [11], locations are ranked by the number of times their position is recorded in the vicinity of the cell tower covering that location. Therefore, for example, the most visited location (likely home) would have rank 1, the second (likely, work place) would have rank 2. A slightly different approach equates location relevance to regularity: trajectories

are split in temporal units, e.g. days, and the relevance of every location l is computed as fraction of units containing l [16].

We argue that methods grounded on location frequency analysis can provide partial and limited information on the locations of interest. Firstly, such approaches tend to ignore the locations that are only visited for a relatively short period of time, for example in occasion of an event, whilst, can classify as interesting locations that, in reality, are transient, e.g., a railway station for a commuter. Secondly, it may be difficult to respond to queries other than top-k relevant locations, for example:

Q_1 : *How many relevant regions do users visit?*
Q_2 : *How popular are those locations in the community?*

A more appealing viewpoint equates relevance to attractiveness, in particular, a location is relevant if the individual intentionally spends some significant time in it. This view is at the basis of the techniques for stop and POIs detection in continuous trajectories, for example reporting the movement of pedestrians and tourists, e.g., [14, 15]. Stop-detection methods, however, call for frequent location sampling. Unfortunately, telco trajectories do not describe the movement at the level of detail requested by those techniques. Therefore, a different approach is needed.

In this paper, we propose a methodology for the characterization and extraction of relevant locations in telco trajectories. The approach is grounded on the idea that the relevance of a location is a time-varying property holding over one or multiple time periods, i.e., the property can be recurrent. Intuitively, a location is relevant when it is assiduously visited for some time, or, put differently, the location is *dominant* in a time period. The methodology consists of two steps: (i) the first step is to reduce the impact of possible noise and irrelevant locations by summarizing telco trajectories. The summarization method is rooted in the conceptual framework we proposed for the density-based segmentation of low-sampling rate spatial trajectories [8]. The outcome of this phase is a set of *summary trajectories*, each reporting the series of locations that are relevant with respect to the input parameters. (ii) In the second step, the extracted locations are further characterized, through the specification of novel mobility indicators enabling the quantification of movement features. In summary, the novel contributions of the paper are:

- We present a novel methodology for the analysis of relevant locations in telco trajectories. The methodology integrates methods from data mining, in particular a variant of a recent density-based trajectory segmentation method, tailored to the discrete space, with novel research on mobility indicators.
- We introduce three novel classes of mobility indicators to measure various aspects regarding both individual and collective mobility. In particular, *summarization rate* and *location diversity* are related to the variety of user's relevant locations; *user diversity* measures the variety of visitors in locations. Indicators are defined in terms of two *diversity* metrics: Richness, and True Diversity associated with the Shannon-Weiner diversity index.
- We experiment with the methodology on a dataset of telco trajectories reporting the movement of ≈ 17,000 individuals in Milan and suburbs over approx 2 months. Moreover we contrast our approach with frequency-based location ranking. We find that trajectory summarization does not compromise data utility, while significantly reducing data size. Moreover, the mobility indicators provide novel insights into human mobility behavior.

The rest of the paper is organized as follows. Section 2 describes the characteristics of telco trajectories and overviews the proposed methodology. Section 3 details the trajectory summarization method, Section 4 the mobility indicators. The experimental evaluation is reported in Section 5, while the last two sections report a brief synthesis of the state-of-the-art and concluding remarks, respectively.

2 REQUIREMENTS AND METHODOLOGY

A natural starting point is to describe the nature of empirical data used for this study.

2.1 Telco data

Dataset. The CDR dataset is provided by a major mobile operator in Italy. The dataset covers the city of Milan plus a few surrounding districts, over a period of 67 days, from March to May 2012. The trajectories are given at the spatial granularity of *Location Area*. A Location Area is a set of one or more base stations, grouped together by the mobile operator, and univocally identified by a label. Figure 1 illustrates a few records about phone calls, text messages and Internet data communication. The last sample reports the trajectory combining the records associated to a random user.

CALL RECORDS

source	destination	date	time	start location_area	end location_area	dir	duration
2071153	1553655	26/03/2012	00:45:10	VALLAZZE	VALLAZZE	I	17
1553655	2733807	27/03/2012	18:19:02	BUENOS AIRES PONCHIELLI	MARIA ADELAIDE	O	245
1553655	3568749	29/03/2012	18:24:23	TUNISIA	TUNISIA	O	20
1553655	2577975	30/03/2012	20:27:17	MM2 CENTRALE	MM2 CENTRALE	O	0

SMS RECORDS

source	destination	date	time	location_area	dir
2071153	1553655	26/03/2012	15:32:22	BUENOS AIRES PONCHIELLI	I
1553655	2071153	26/03/2012	15:38:19	BUENOS AIRES PONCHIELLI	O

INTERNET RECORDS

source	date	time	location_area	upload	dowloand
1553655	26/03/2012	00:02:46	PIOLA	0	141
1553655	26/03/2012	00:02:46	PIOLA	4698	7838
1553655	26/03/2012	00:40:27	VALLAZZE	3327	4357

MOBILITY TRACE

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	00:40:27	VALLAZZE
1553655	26/03/2012	00:45:10	VALLAZZE
1553655	26/03/2012	15:32:22	BUENOS AIRES PONCHIELLI
1553655	26/03/2012	15:38:19	BUENOS AIRES PONCHIELLI
1553655	27/03/2012	18:19:02	BUENOS AIRES PONCHIELLI
1553655	29/03/2012	18:24:23	TUNISIA
1553655	30/03/2012	20:27:17	MM2 CENTRALE

Figure 1: A fragment of CDR data

Cells and Location Areas coordinates are not available. However, in previous work, it has been estimated that 75% of the Location Areas in Milan are smaller than 1 square kilometer and concentrated downtown, whilst the largest regions, over 4 square kilometers, are in the suburbs. Figure 2 shows a fragment of the Voronoi polygons used to approximate Location Areas. The set of representative points for the Location Areas forms the telco space. We refer the reader to [16] for further details on the dataset.

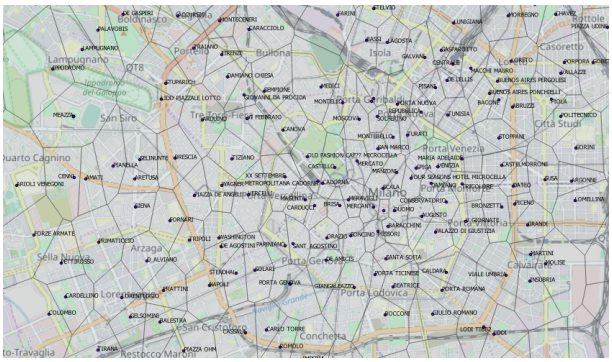

Figure 2: A fragment of the Voronoi diagram obtained from the estimated centroids of Location Areas in Milan

Data characteristics. Abstracting from the specific dataset, telco trajectories have peculiar characteristics:

- Sequences of identical locations. Locations are regions of space. Therefore, as the user's position is matched against the closest base station, it may readily 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 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 [3].
- The locations reported in CDRs can be noisy because of signal fluctuation in the network coverage [5].

2.2 Methodology

The goal is to extract from every trajectory of the reference dataset, the locations relevant for the specific user, and then analyze the characteristics of those locations at population scale. The problem is challenging because telco data are complex. The proposed methodology comprises two steps: trajectory summarization and summary trajectories analysis, the former for removing irrelevant locations; the latter for analyzing supplementary features of relevant locations, and synthesize the individual movement in a number of indicators. The idea behind trajectory summarization is discussed next.

Trajectory summarization. The summarization method is built on the density-based trajectory segmentation technique [7, 8], developed for the detection of stops in noisy spatial trajectories. Density-based trajectory segmentation partitions a trajectory of coordinated points 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*. These concepts can be better understood through the example of trajectory segmentation in Figure 3.

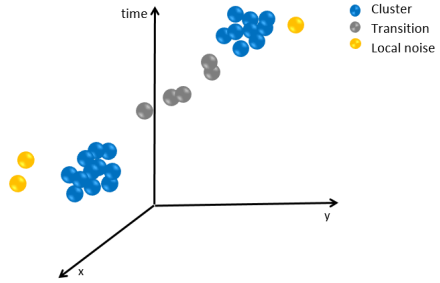

Figure 3: Density-based segmentation of a spatial trajectory [8]. The spatio-temporal points are classified as: cluster point, noise, transition point

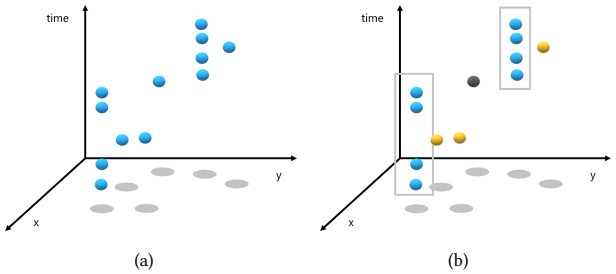

Figure 4: (a) Telco trajectory with 6 different locations (on space) and 12 occurrences (in space-time); (b) The rectangles contains two groups of occurrences for the same symbol, representing clusters along the temporal line. There are also 3 noisy points and one transition point

As this method is shown to be robust against noise and low sampling rates, at first we attempted to apply the technique over the set of coordinated points in the telco space. The resulting clusters, however, did not reveal any pattern of interest. We thus devised a slightly different strategy, to replace the spatial dimension of the locations with the symbolic dimension and therefore treat trajectories as *timed strings*. To convey the intuition of what that means,

Table 1: Notation

T, $\mathcal{T}$	Trajectory, set of trajectories
$\widehat{T}$, $\widehat{\mathcal{T}}$	Summary trajectory, set of
L, l	Symbolic space, location
U, u	Users, user
$w_t(l,j)$, $W_T(l)$	Weight of a symbol
δ	Weight threshold
N	Occurrences threshold
H	Shannon-Weiner entropy
hness (R) R, R_l, R_u	Richness based indexes
TD, TD_l, TD_u	True diversity based indexes
J	Jaccard index
S_{rate}	Summarization rate
ρ	Weighted Spearman Index

consider the example in Figure 4. This example shows a telco space of 6 locations (as points in the plane) and the trajectory of a user. It can be seen that the trajectory 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 happen: the user returns back to the previous location; or the user start frequenting some other location. This suggests a cluster-based segmentation performed over the temporal line with clusters only grouping occurrences of a unique location. Figure 4 shows two clusters, few points of local noise and one point classified as transition. A cluster is characterized by one symbol and by a temporal extent. Our hypothesis is that a cluster identifies the occurrence of a *relevant location.*

3 SUMMARIZATION PHASE

We turn to detail the summarization model and the algorithm.

3.1 Cluster model

To begin, we introduce some basic notation and convention. A trajectory is a sequence of timed symbols from dictionary L. Note that the terms "location" and "symbol" are used interchangeably. Symbol occurrences have a timestamp.

Cluster model. A cluster designates the *dominating* symbol in a time period.

The number of occurrences is a natural indicator of the significance of a symbol. For example, if we have the sequence: (a, t_1) (a, t_2) (a, t_3) (b, t_4), the symbol a appears 3 times in the period $I = [t_1, t_4]$, thus a is dominant in I, irrespective of the presence of b. However, sequences normally contain multiple locations, therefore, different kinds of symbols may compete for the role of dominant symbol. We quantify the significance of a symbol in a time period through two measures, the number of occurrences, and the *weight* of the symbol, respectively. The purpose of the weight function is to award temporally correlated occurrences of the same symbol.

Consider a trajectory $T = (l_1, t_1)..(l_n, t_n)$, defined in the interval $[t_1, t_n]$.

Definition 3.1 (**Weight function**). *Let $w(T, l, j)$ be the function computing the weight of symbol $l \in L$ at position $j \in [2, n]$ in T, defined as:*

$$w(T,l,j) = \begin{cases} |t_j - t_{j-1}|, & \text{if } l_j = l_{j-1} = l \\ 0, & \text{otherwise} \end{cases}$$

The weight $W(T, l)$ over the whole trajectory is given by the sum of weights at the different positions:

$$W(T,l) = \Sigma_{j=2}^{k} w(T,l,j)$$

Example. Consider the following trajectory from t_1 to t_9 containing 9 occurrences of three different symbols, a, b, c. The trajectory is:

$$T = (a,t_1)(a,t_2)(b,t_3)(c,t_4)(b,t_5)(b,t_6)(c,t_7)(a,t_8)(a,t_9) \quad (1)$$

The symbols in T have the following weight:

$$W(T,a) = |t_2 - t_1| + |t_9 - t_8|, W(T,b) = |t_6 - t_5|, W(T,c) = 0$$

Thus the symbol a has the highest weight.

Definition 3.2 (**Dominance**). *Consider a sequence $S = (l_i, t_i)..(l_k, t_k) \subseteq T$. Given $N > 2$ and $\delta \geq 0$, we say that l is dominant in the period $[t_i, t_k]$ iff the following conditions are satisfied:*

i. $l_i = l_k = l$
ii. $W(S, l) \geq \delta$
iii. $|\{l_j \in S | l_j = l\}| \geq N$
iv. *no other symbol satisfies the above conditions in the period.*

Condition (i) states that the sequence S is bounded by l; (ii) and (iii) specify threshold values for the number of occurrences and the weight, respectively; (iv) the dominant symbol is unique in the time frame of S.

Example. Consider again trajectory (1). Without loosing in generality, assume that all symbols are equally spaced in time with $|(t_{i-1} - t_i)| = \Delta$, and let $N = 3$, $\delta = 2\Delta$. It can be seen that a is dominant in the period $[t_1, t_9]$.

Definition 3.3 (**Summary trajectory**). *A summary trajectory is denoted $(I_1, l_1)..(I_k, l_k)$, with (I_j, l_j) meaning:*

- *the symbol l_j is dominant in the period I_j. We say that l_j forms a cluster in the period.*
- *The temporal extent I_j is maximal*

Example. Consider the following sequence of symbols evenly spaced in time, as above, with $|(t_{i-1} - t_i)| = \Delta$, from t_1 to t_{17} (we omit time for brevity) :

$$T = a, b, a, a, a, a, b, a, b, c, d, d, c, d, d, a, d$$

Let $N = 4$ and $\delta = 2\Delta$. The trajectory is summarized in 2 units as follows:

$$\widehat{T} = ([1,8], a)([11,17], d) \quad (2)$$

Note that summary trajectories can be straightforwardly represented using the symbolic trajectories data model in [12].

Definition 3.4. *The set of symbols appearing in a summary trajectory $\widehat{T}$ are the* relevant locations *of the trajectory*

Note that summarization causes a loss of information because not all the symbols at all times are reported. However, the same symbol can appear multiple times in the summary trajectory (though not consecutively [8]), meaning that the user can return multiple times to the same location.

3.2 Summarization algorithm

The algorithm extracts a series of temporally separated clusters, from the input trajectory, based on the parameters N and δ. The symbols of the sequence are processed one at a time. As a dominant symbol is found, a cluster is created and becomes the active cluster. The algorithm proceeds trying to expand the active cluster, while monitoring at the same time the emergence of other clusters. If the active cluster is no longer expanded, and a new symbol becomes dominant, the active cluster is closed and appended to the output clusters, while a new cluster is created. The process terminates when the scan is complete.

Details. The algorithm is detailed in Figure 1. The information relevant for the processing of symbols is kept in a hash table for the symbols of the telco space. For every distinct symbol of the trajectory, the tuple $(n, w, l)_s$ reports the number of occurrences, the weight and the index of the first occurrence in the portion of trajectory being processed. As a symbol s becomes dominant, the hash table, except for the dominant symbol entry, is reset and the phase of cluster expansion starts. Upon the reading of a symbol s', two cases may occur: if s' is an occurrence of the dominant symbol, the entry is updated while the hash table is reset again, as above. Note that the reset operation is necessary to ensure that clusters are temporally disjoint. If s' is not an occurrence of the dominant symbol, the corresponding entry in the hash table is updated and the input constraints are checked. Hence, if the symbol gets dominant, the pair (I, s), with I denoting the time interval between the first and the last occurrence of s is stored as *unit* of the summary trajectory. The output of the algorithm is a list of units defined over temporally separated time intervals. The run-time complexity is linear with respect to the number of symbol occurrences in the trajectory.

Algorithm

Input: $T = [(l_1, t_1), (l_2, t_2), \ldots]$ *//trajectory* ;
N, δ *//input parameters*;
Result: $\widehat{T}$ *//summary trajectory*

```
C=∅ //Active cluster ;
H //Hash table of |L| entries ;
for (l, t) in T do
    H.UpdateEntry( l, t);
    if C=∅ then
        if getsDominant(l) then
            C ← Cluster(l);
            H.ResetNonDominantSymbols(l);
        end
    else
        if l==dominant(C) then
            H.ResetNonDominantSymbols(l);
        else
            if getsDominant(l) then
                T̂.Add(close(C));
                C ← Cluster(l);
                H.ResetNonDominantSymbols(l);
            end
        end
    end
end
```

Algorithm 1: Summarization algorithm

4 SUMMARY TRAJECTORIES ANALYSIS

After summarizing trajectories, we turn to consider the second phase of the methodology, how to characterize the locations of the summarized space, through the specification of mobility indicators. We recall the basic questions we want to solve:
Q_1 : *How many relevant regions do users visit?*
Q_2 : *How many locations are irrelevant?*
Q_3 : *What is the popularity of those locations?*

We approach the problem by introducing a few variables or *mobility indicators*, on top of the notion of population *diversity metric*.

4.1 Diversity metrics

Diversity is a key concept in innumerable fields, including biology, economy, demography, information theory. For example, diversity quantifies the biodiversity of a geographical area, i.e. diversity of species, the economic diversity of a region, i.e. diversity of companies with respect to their products. In general, diversity is used to characterize populations consisting of objects of different type.

Populations of concern. We are concerned with two kinds of populations: the set of locations (symbols) appearing in every trajectory; the set of users visiting the locations of interest. We refer to those populations as location and user population, respectively. An orthogonal distinction is between populations drawn from the original trajectories and those drawn from summary trajectories.

Location and user diversity. We refer to the diversity of locations population as *location diversity*. Different from the notion of trajectory similarity, which confronts two sequences, location diversity characterizes a single trajectory. In this sense, location diversity can be seen as an individual mobility index. Similarly, we call *user diversity* the diversity of user population.

Diversity metrics. Many different metrics are utilized to measure the diversity of a generic population. A simple measure is given by the count of types. This measure is called *Richness* (R) [13]. In particular, the Richness of a trajectory, either summarized or not, is given by the number of symbol types. For example, the Richness of summary trajectory (2) is R=2. The Richness metric is simple and intuitive, yet it does not take into account the numerosity of occurrences. Therefore this index can result too coarse. To illustrate the informative value of occurrences, consider the following trajectory T:

$$(I_1, a)(I_2, b)(I_3, a)(I_4, c)(I_5, a)(I_6, d)(I_7, a)(I_8, e)(I_9, a) \qquad (3)$$

T has richness $R = 5$. It can be seen, however, that symbol a has 5 occurrences, while b, c, f, e, appear just once. Thus, if we rank those locations by frequency, a results the most frequented by user

u. Alternative rankings can be envisaged, for example based on the cumulative temporal extent of locations (i.e., *dominance time*) or even combining frequency and dominance time. The experiments, however, show that interesting results can also be obtained by simply ranking relevant locations by frequency. This information can provide useful insights into the dynamics of the individual movement.

4.2 Entropy and True Diversity

Probably, the most common diversity metric, sensitive to the numerosity of occurrences, is the Shannon-Wiener index.

The Shannon-Weiner diversity index is based on Shannon entropy. Given a population consisting of n types of elements, the Shannon-Weiner index is defined as:

$$H = -\sum_{i=1}^{n} p_i \ln p_i$$

where p_i the probability that a population sample belongs to type i.

Example: consider the trajectory (3). The population consists of 9 elements of type a, b, c, d, e. The probabilities are respectively $5/9$ (for symbol a) and $1/9$ (for b,c,d,e). Thus, the entropy is $H = 1.303$.

True diversity. During the past decade, a remarkable effort has been conducted, in particular in ecology (see Josh [13]), to clarify the concept of diversity as opposed to that of the "diversity index". This concern is motivated by the lack of a unifying ground for the concept of diversity. In particular, different diversity indexes (e.g., Gini-Simpson, Renyi entropy) result in different measures of diversity. In addition, the values of those indexes do not increase linearly with the number of types, therefore comparing the diversity of different populations is hard. Further, the diversity measures are of difficult interpretation. By contrast, diversity is conceptually straightforward and simply indicates the number of types.

This discussion has brought to the forefront the concept of *True Diversity* [13, 19]. True Diversity is not another index, but rather a theoretical framework practically enabling the conversion of the most common diversity indexes into a unique measure of diversity, expressed in terms of number of types. In particular, the True Diversity associated with a diversity index X, indicates the number of equally common types determining the value of X. The diversity value can be drawn by calculating the diversity index for equally-common species (each species therefore with a frequency of 1/X) and solving that equation for X [13]. *Richness* is the coarsest form of True Diversity, insensitive to type frequences (True Diversity of level 0). The True Diversity associated with the Shannon-Weiner index (True Diversity of level 1) is defined as follows:

$$D_{SW} = e^{H} = e^{-\sum_{i=1}^{n} p_i \ln p_i}$$

Example: consider trajectory (3). The True Diversity associated with Shannon-Weiner index is 3.7 (types). We can notice the difference from the Richness measure, R=5 (types).

4.3 Mobility indicators

Finally, armed with these concepts, we turn to analyze a dataset of telco trajectories. Let us consider the following components of the dataset: the set of users $\mathcal{U} = \{u_1, .., u_m\}$, the telco space L, the set of non summarized trajectories $\mathcal{T} = \{T_1, .., T_m\}$, the set of summary trajectories $\widehat{T} = \{\widehat{T}_1, .., \widehat{T}_m\}$. We introduce three classes of mobility indicators:

Table 2: Mobility indicators

Class	
Location Diversity	R_l, TD_l
Summarization rate	S_{rate}
User Diversity	R_u, TD_u

Location diversity. Defined at individual level, location diversity specifies the number of different locations in a summary trajectory. Depending on the metric used, the indicator is called location Richness ($R_l(\widehat{T}_i)$) or location True Diversity ($TD_l(\widehat{T}_i) = e^H$ with H entropy of $\widehat{T}_i$).

Summarization rate. Defined at individual level, it specifies the percentage of transient locations in the original trajectory. Given the trajectories $T_i \in \mathcal{T}$ and $\widehat{T}_i \in \widehat{T}$, the summarization rate $S_{rate}(T_i)$ is

$$S_{rate}(T_i) = 1 - \frac{R_l(\widehat{T}_i)}{R_l(T_i)} \quad (4)$$

$S_{rate} \approx 0$ means limited or even no summarization; $S_{rate} \approx 1$ means high summarization level.

User diversity. Defined at community level, it specifies the number of users for which a given location of a summarized space is relevant. We consider user Richness and user True Diversity based on Shannon-Weiner, denoted: $R_u(S), TD_u(S)$, respectively.

5 EXPERIMENTAL EVALUATION

5.1 Experimental setting

We have developed a proof-of-concept sw system to test and evaluate the methodology on a real dataset. The summarization algorithm is written in Python while the dataset is stored in a Postgres database.

Dataset. The dataset consists of 17,168 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 55 million points. The *telco space* consists of 685 locations at the granularity of Location Area, and identified by a label. Table 3 reports the summary statistics on the number of trajectories (i.e. users), number of records, average and standard deviation of trajectory length.

Table 3: Summary statistics of the dataset

# Traj	# Records	# Loc	Avg(trj_len)	Std(trj_len)
17168	54,193,257	685	3151	1650

Methodology. The analytical process consists of three steps:

(i) Calibration of the summarization parameters. This operation is performed iteratively over a random subset of 100 trajectories extracted from the input dataset.
(ii) Dataset summarization. The summarization algorithm is run over the input dataset/s using the parameters specified at the previous stage. The result is 3 summarized datasets.
(iii) Computation of the mobility indicators over one of the summarized datasets: summarization rate, location diversity and user diversity.

Hw/sw platform Data summarization is performed on a Linux-Ubuntu Server DELL T620, 362 GB Ram, data analysis on a standard PC Windows.

5.2 Trajectory summarization rate

The goal of this first experiment is to analyze the impact of the clustering parameters N, δ over the summarization rate. We choose different sets of input parameters (Table 4) focusing in particular on the temporal parameter δ, which is a peculiarity of this technique.

Table 4: Input parameters for data summarization

Summarized Dataset	N	$\delta(day)$
D1	4	$0.0014 \approx 2'$
D2	4	$0.01 \approx 15'$
D3	4	$0.04 \approx 60'$

It can be seen that the value of N is fixed. To convey the intuition of the practical meaning of the parameter setting $N = 4$, consider that an individual engaged in two phone calls from the same region satisfies the constraint $N \geq 4$. Thus the requirement is not excessively strict. As regards the temporal parameter, we recall that δ indicates the minimum weight for a location to be relevant. For this parameter, we have chosen three possible values. These values are expressed as fractions of a day.

Summarization rate. For every summarized dataset and for every user, we compute the summarization rate of the associated compressed trajectory. We obtain three statistical distributions for the S_{rate} indicator. Summary statistics are reported in Table 5, while the Complementary Cumulative Distribution Function (CCDF) for every dataset is reported in Figure 5.

Table 5: Summary statistics for S_{rate}, and data size

Sum. Data	Mean S_{rate} %	Std S_{rate} %	Size(MB)
D1	65	7	65
D2	74	6	51
D3	84	5	44

These statistics highlight a strong correlation between summarization rate and the δ parameter. There is no surprise in this, as the stricter the temporal constraint, the less the number of relevant locations satisfying the constraint. As we can see, the summarization rate is very high. It means that the number of *different* locations appearing in a trajectory is drastically reduced (irrespective of the number of occurrences).

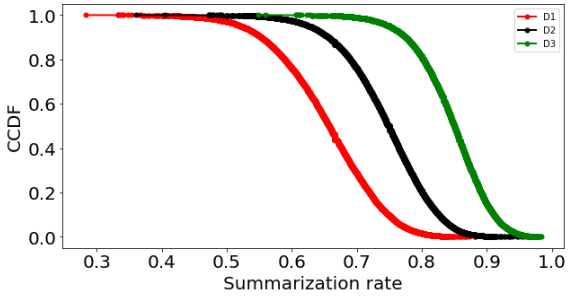

Figure 5: Summarization rate for D1, D2, D3

5.3 Location diversity

In this experiment, we evaluate location diversity for the summarized set D2 using both metrics R_l and TD_l. Further, for comparison, we report the location Richness for the original dataset. Summary statistics are reported in Table 6.

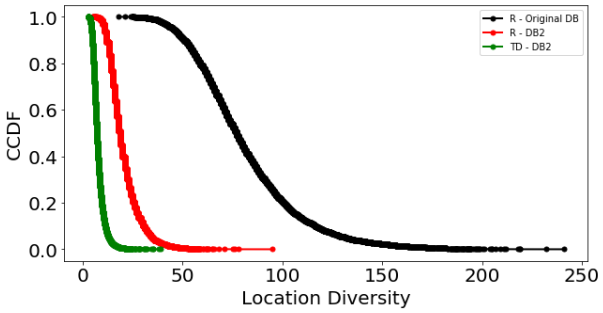

Figure 6: Location diversity for D2

Table 6: Summary statistics for location diversity in D2 (vs. Original DB)

Dataset	Metric	Mean	Std
DB orig	R_l	81	27
D2	R_l	20	7
D2	TD_L	8	3

The distributions for the three location diversity measures are shown in Figure 6. It can be seen that location Richness in the original dataset (black plot) is significantly higher than both Richness and True Diversity in the summarized dataset. For example, the probability of randomly selecting a trajectory with, e.g., more 50 locations (types) is very high in the original dataset, while it is extremely low in the summarized dataset. That is, the locations that are relevant' are few and significantly less in number than irrelevant locations. We can also see, that, compared with R_l, the True Diversity measure is lower. This means that summary trajectories consist of locations of significantly different frequency.

Further details are provided by the histogram in Figure 7, where users are classified in three classes, based on the value of R_l. The

partitioning is obtained by applying the Jenks natural break classification, a method for the clustering of 1-D data. The Jenks method is widely used in GIS platforms for the clustering of features based on the value of a quantitative attribute [1]. Our histogram shows that 58% of users frequents a number of relevant locations ranging in the interval [3, 8), while for 33%, the number of locations varies [8, 12); the third class ranges in [12, 39]. Finally, Figure 8 shows an example trajectory from the original dataset. The representative points of the Location areas are plotted in a spatio-temporal coordinate system, while consecutive points are connected by segments. This trajectory of 39 different locations ($R_l = 39$), once summarized, contains only 5 locations ($R_l = 5$). The mobility is thus concentrated in few locations, as shown in the figure.

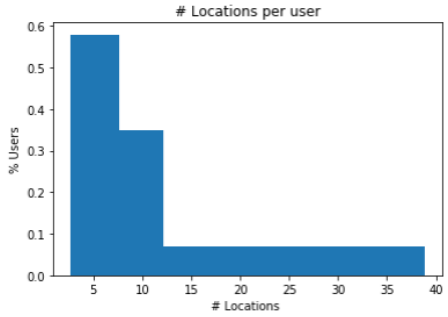

Figure 7: Natural-break classification of location diversity based on true diversity associated to Shannon-Weiner

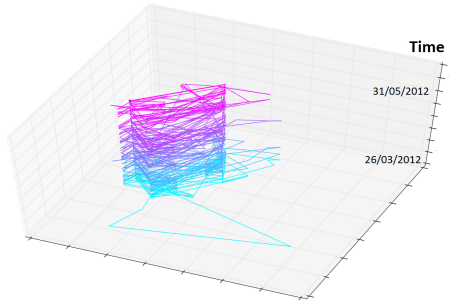

Figure 8: Spatio-temporal representation of a trajectory

5.4 User diversity

We turn to analyze the popularity of the relevant locations obtained from summary trajectories.

User diversity. In general, user diversity is computed with respect to a set of locations L': given L', for every location $l \in L'$, we compute the number of different users passing by l. Since users can pass multiple times by the same locations, it makes sense to utilize both metrics, R_u (user Richness) and TD_u (user True Diversity), the former because it is more intuitive, the latter more detailed.

[1] https://support.esri.com/en/technical-article/000006743

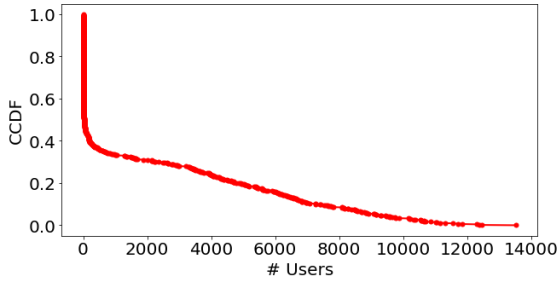

Figure 9: User Richness R_u in the original dataset

Figure 9 shows the user diversity (based on metric R_u) in the telco space L, where $|L| = 685$ locations. It can seen that a large number of locations are visited by very few people. We have found that these locations are in reality districts surrounding Milan, where, plausibly, users are only occasionally located. We can also see that very few locations are highly frequented, by more than 10K users. Coherently, the variance of the variable is extremely high (Table 7).

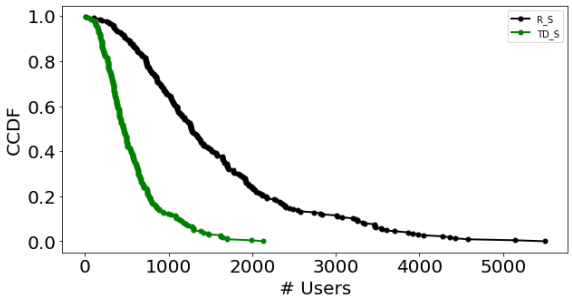

Figure 10: User diversity in the summarized dataset D2

Table 7: Summary statistics for user diversity in D2 (vs. Original DB)

Dataset	Metric	Mean	Std
DB orig	R_u	2033	3220
D2	R_u	1542	1023
D2	TD_u	557	375

In comparison, the distribution of users over the set of relevant locations, obtained from the union set of the summary trajectories, looks quite different. Let $L' \subseteq L$ the set of relevant locations in the summarized dataset $D2$. Figure 10 illustrates the statistical distribution of user diversity over L' where $|L'| = 226$ locations.

First, there is no longer evidence of locations frequented by few people (Figure 9). That is coherent with the interpretation that those locations are occasional, and thus irrelevant for our model. We can also see a significant gap in the maximum number of visitors. Plausibly, many locations, although frequently visited, are not really relevant for a large number of people, i.e., are *transient*.

Further details are provided by the histogram in Figure 11 reporting the Jenks-based classification of locations based on the number of visitors. More than 50% of location are visited by less than 1327 users, while the most popular locations are visited by a number of individuals ranging between 2740 and 5500. The percentage of locations classified as top frequented amounts to 13%.

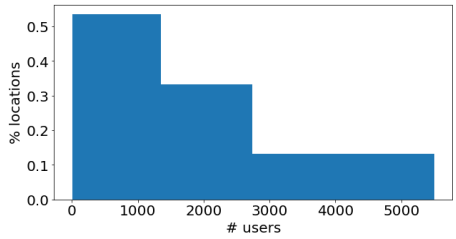

Figure 11: Natural-break classification of user diversity based on richness

5.5 Relevance vs. regularity: comparison

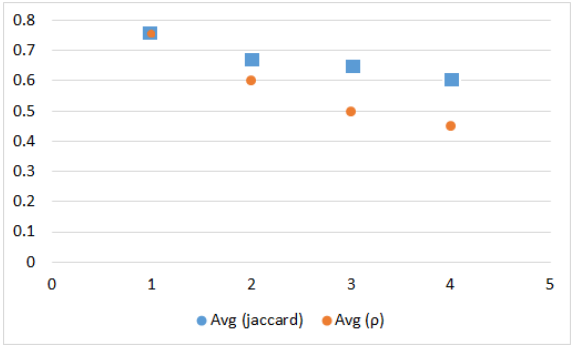

Figure 12: Location ranking comparison; Jaccard and ρ metrics

In this experiment we compare the most relevant locations with the most frequent locations.

In particular, for every user, we compare the top-k frequent locations in the summarized dataset D2 with the top-k frequent locations in the original, uncompressed trajectories. We denote the two rankings as τ_1 and τ_2, respectively. In the former case, locations are ranked based on the number of times they appear in the trajectory. For example, in the summary trajectory $(I_1, a)(I_2, b)(I_3, a)$, location a appears twice and b once. In the latter case, the ranking is based on the fraction of days every location is visited.

Metrics. Given τ_1, τ_2, we want to measure how similar the two rankings are for every trajectory. We use two different metrics. The first is the Jaccard index: $J = \frac{\tau_1 \cap \tau_2}{\tau_1 \cup \tau_2}$. This index is simple, yet the comparison is only performed on the sets of values, irrespective of the ranking.

Popular measures sensitive to ranking are the Spearman correlation coefficient and Kendall τ distance that measure the total number of pairwise inversions in the rankings. Both measures, however, apply to permutations of a unique domain, while in our case, the top-k rankings contain different elements. More recent metrics try to capture element weights, position weights, and pairwise distances between permutations, e.g.,[9]. For the problem at hand, we rely on a recent metric called *Weighted Spearman Rank Distance* [6].

Let $q = |\tau_1 \cup \tau_2|$. The penalty weight w_i for an element i in the lists τ_1 or τ_2 is computed as follows:

$$w_i = \begin{cases} 1 - \frac{1}{|x_i - y_i| + 1}, & i \in \tau_1 \text{ and } i \in \tau_2 \\ 1, & \text{otherwise} \end{cases}$$

where x_i, y_i indicate the position of the element in τ_1 and τ_2, respectively. The Weighted Spearman rank coefficient value is then computed as:

$$\rho_w = \frac{\sum_{i=1}^{q} w_i}{q}$$

In the experiment we consider the complement: $\rho = 1 - \rho_w$

High correlation is found when $\rho_w \approx 1$ (i.e., very few penalties are assigned); low correlation when $\rho_w \approx 0$ (i.e., many penalties are assigned). If the two lists have no common element, $\rho = 0$.

Experiment. For every trajectory, we compute the indices J and ρ for K=1,2,3,4. The average values over the dataset are reported in Figure 12. It can be seen that the Jaccard index is high for the first two locations (approx 0.7). It means that the top-2 locations are both "relevant" and "regularly visited". That makes sense. For example, *home* and *work* are both relevant and frequent locations. The value of ρ gives supplementary information, that the distance between the rankings increases rapidly for K>2. This suggests that location relevance does not equate to location frequency. This is what we wanted to demonstrate.

5.6 Discussion

Main findings. Back to the research questions presented in the introductory section:

- The majority of people exhibit limited mobility across regions. More than 90% of the population frequent at most 12 (relevant) locations. These results are qualitatively in line with human mobility studies (see next section).
- More than 60% of the locations reported in CDRs are irrelevant, with respect to the location relevance model.
- Approx 13% of the locations in the summarized dataset are highly frequented, meaning that those locations are also relevant for the community, not only for the single individual. We have not found comparable results in literature.

Remarks.

- The summarization technique finds concentrations of symbols in timed strings. In this sense the technique can be generalized beyond the telco domain.
- The summary trajectories obtained from the experiments are highly compressed, more than 70% of the location types are removed. The summarization rate, however, is arguably related to the spatial granularity of locations. It remains to

analyze the summarization rate with locations at varying resolution, e.g., at cell level.

- We find that the combined use of two diversity metrics, Richness and True Diversity associated with the Shannon-Weiner diversity index, allows a better understanding of the data characteristics. The former is simple and intuitive, the latter provides further details on data distribution. The indicators based on these diversity metrics have been shown to be expressive and effective.
- We have found that the summarized trajectories preserve, at least to some extent, important properties of uncompressed trajectories. In particular, the top-2 frequented locations (the so-called home and work) can be identified with good accuracy. This indirect approach to validation presents interesting challenges.
- Performance. The core of the approach is trajectory summarization. Summarization is, however, not scalable unless relying on parallel and big data architectures. We leave the architectural issue for future work.

6 RELATED WORK

Research on human mobility patterns span many different fields, from statistical physics, to geography, complex networks and pervasive computing [3]. A large body of research targets the discovery of general rules underlying human movement and use principled methodologies grounded on statistical methods. In that respect, CDR datasets are key sources [3, 4]. Foundational work by Gonzalez et al. [11] found that human trajectories show a high degree of temporal and spatial regularity. The regularity is mainly due to the fact that users spend most of their time in a small number of locations. These findings are also supported by Song et al. [18], which show a model mixing the propensity of users to return to previously visited locations and a drift for exploration [4]. Notably, Csaji e al. [5] show how small the number of frequently visited locations is. They define a frequently visited location of a user as a place where more than 5% of phone calls were initiated. The authors found that the average number of frequently visited locations is only 2.14, and that 95% of the users visit frequently less than 4 locations. A related approach by Bagrow et al. [2] is to group frequently visited locations representing recurrent mobility into a habitat. Compared to these approaches, our methodology is different: the locations of interest are those around which the individual gravitates for relatively short periods, not necessarily those that are frequent over long periods. Further we refer to a CDR dataset, also reporting Internet communications which are frequent and bursty, and thus more complex to handle. Qualitatively, the results we obtain are in line with the literature in that a large percentage of people frequent few locations, though locations have a different meaning. From the data management viewpoint, various lines of research are related to the discovery of locations from user's traces, especially revolving around the concepts of semantic trajectories, e.g. [15], trajectory segmentation algorithms, e.g. [1, 14, 17], trajectory data mining, e.g. [10, 20]. The work presented in this paper combines methods from data mining with methods inspired by research on human mobility pattern analysis.

7 CONCLUSIONS

This paper presents a two-step methodology to the discovery of the regions of interest. A major contribution is the summarization technique for the discovery of concentrations of symbols in timed strings based on temporal criteria, which extends the notion of density-based segmentation to the symbolic space. We have also proposed three novel mobility indicators, relying on the concepts of location and user diversity, and discussed possible extensions of the work, especially in the direction of a scalable architecture. For the generality of the concepts presented, the methodology can be of interest beyond the telco domain.

REFERENCES

[1] B. Aronov, A. Driemel, M. V. Kreveld, M. Löffler, and F. Staals. Segmentation of trajectories on nonmonotone criteria. *ACM Trans. Algorithms*, 12(2):26:1–26:28, 2015.

[2] J.P. Bagrow and Y.-R. Lin. Mesoscopic structure and social aspects of human mobility. *PloS one*, 7(5):37676, 2012.

[3] H. Barbosa, M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, and M. Tomasini. Human mobility: Models and applications. *Physics Reports*, 734:1 – 74, 2018.

[4] V. Blondel, A. Decuyper, and G. Krings. A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4(10), 2015.

[5] B. Csaji, A. Browet, V.A. Traag, J. Delvenne, E. Huens, P. Van Dooren, Z. Smoreda, and V. D. Blondel. Exploring the mobility of mobile phone users. *Physica A: Statistical Mechanics and its Applications*, 392(6):1459 – 1473, 2013.

[6] D. Chicco D, E. Ciceri, and M. Masseroli. Extended Spearman and Kendall coefficients for gene annotation list correlation. In *Serio C.D., Lio P., Nonis A., Tagliaferri R. (eds) Computational Intelligence Methods for Bioinformatics and Biostatistics*. Springer,Lecture Notes in Computer Science, 2015.

[7] M. L. Damiani, H. Issa, and F. Cagnacci. Extracting stay regions with uncertain boundaries from GPS trajectories: a case study in animal ecology. In *Proc. ACM SIGSPATIAL*, 2014.

[8] M.L. Damiani, F. Hachem, H. H. Issa, N. Ranc, P. Moorcroft, and F. Cagnacci. Cluster-based trajectory segmentation with local noise. *Data Mining and Knowledge Discovery*, 32(4):1017–1055, 2018.

[9] R. Fagin, R. Kumar, M. Mahdian, D. Sivakumar, and E. Vee. Comparing partial rankings. *SIAM J. DISCRETE MATH.*, 20(3):628–648, 2006.

[10] Z. Feng and Y. Zhu. A survey on trajectory data mining: Techniques and applications. *IEEE Access*, 4:2056–2067, 2016.

[11] M.C. Gonzáles, C.A. Hidalgo, and A.L. Barabási. Understanding individual individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[12] R.H. Güting, F. Valdés, and M. L. Damiani. Symbolic trajectories. *Trans. Spatial Algorithms and Systems*, 1(2):7:1–7:51, 2015.

[13] l. Jost. Entropy and diversity. *Wiley Online Library Oikos*, 113(2):363–375, 2006.

[14] A. T. Palma, V. Bogorny, B. Kuijpers, and L. O. Alvares. A clustering-based approach for discovering interesting places in trajectories. In *Proc. ACM Symposium on Applied Computing*, 2008.

[15] C. Parent, S. Spaccapietra, C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, M. L. Damiani, A. Gkoulalas-Divanis, J. Macedo, N. Pelekis, Y. Theodoridis, and Z. Yan. Semantic trajectories modeling and analysis. *Comput. Surv.*, 45(4):1–32, 2013.

[16] C. Quadri, M. Zignani, S. Gaito, and G. P. Rossi. On non-routine places in urban human mobility. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 2018.

[17] A. Soares Junior, V. Cesario Times, C. Renso, S. Matwin, and L. A. F. Cabral. A semi-supervised approach for the semantic segmentation of trajectories. In *2018 19th IEEE International Conference on Mobile Data Management (MDM)*, 2018.

[18] C. Song, T. Koren, P. Wang, and A. Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6:818–823, 2010.

[19] H. Tuomisto. A consistent terminology for quantifying species diversity? Yes, it does exist. *Oecologia*, 164(4):853–60, 2010.

[20] Y. Zheng. Trajectory data mining: an overview. *ACM Trans. Intelligent Systems and Technology*, 6(3):1–41, 2015.