Data Min Knowl Disc (2018) 32:787–829
https://doi.org/10.1007/s10618-017-0548-4

Data-driven generation of spatio-temporal routines in human mobility

Luca Pappalardo[1,2] · **Filippo Simini**[3]

Received: 16 July 2016 / Accepted: 5 December 2017 / Published online: 27 December 2017

Abstract The generation of realistic spatio-temporal trajectories of human mobility is of fundamental importance in a wide range of applications, such as the developing of protocols for mobile ad-hoc networks or what-if analysis in urban ecosystems. Current generative algorithms fail in accurately reproducing the individuals' recurrent schedules and at the same time in accounting for the possibility that individuals may break the routine during periods of variable duration. In this article we present DITRAS (DIary-based TRAjectory Simulator), a framework to simulate the spatio-temporal patterns of human mobility. DITRAS operates in two steps: the generation of a mobility diary and the translation of the mobility diary into a mobility trajectory. We propose a data-driven algorithm which constructs a diary generator from real data, capturing the tendency of individuals to follow or break their routine. We also propose a trajectory generator based on the concept of preferential exploration and preferential return. We instantiate DITRAS with the proposed diary and trajectory generators and compare the resulting algorithm with real data and synthetic data produced by other generative algorithms, built by instantiating DITRAS with several combinations of diary and trajectory generators. We show that the proposed algorithm reproduces the statistical

Responsible editor: Johannes Fürnkranz.

✉ Luca Pappalardo
lpappalardo@di.unipi.it; luca.pappalardo@isti.cnr.it

Filippo Simini
f.simini@bristol.ac.uk

1 Institute of Information Sciences and Technologies, National Research Council, Pisa, Italy

2 Department of Computer Science, University of Pisa, Pisa, Italy

3 Department of Engineering Mathematics, University of Bristol, Bristol, UK

properties of real trajectories in the most accurate way, making a step forward the understanding of the origin of the spatio-temporal patterns of human mobility.

Keywords Data science · Human mobility · Complex systems · Mathematical modelling · Big data · Spatiotemporal data · Human dynamics · Urban dynamics · Mobile phone data · GPS data · Smart cities

1 Introduction

Understanding the complex mechanisms governing human mobility is of fundamental importance in different contexts, from public health (Colizza et al. 2007; Lenormand et al. 2015) to official statistics (Marchetti et al. 2015; Pappalardo et al. 2016b), urban planning (Wang et al. 2012; De Nadai et al. 2016) and transportation engineering (Janssens 2013). In particular, human mobility modelling has attracted a lot of interest in recent years for two main reasons. On one side, it is crucial in the performance analysis of networking protocols such as mobile ad hoc networks, where the displacements of network users are exploited to route and deliver the messages (Karamshuk et al. 2011; Hess et al. 2015). On the other side human mobility modelling is crucial for urban simulation and what-if analysis (Meloni et al. 2011; Kopp et al. 2014), e.g., simulating changes in urban mobility after the construction of a new infrastructure or when traumatic events occur like epidemic diffusion, terrorist attacks or international events. In both scenarios the developing of generative algorithms that reproduce human mobility patterns in an accurate way is fundamental to design more efficient and suitable protocols, as well as to design smarter and more sustainable infrastructures, economies, services and cities (Batty et al. 2012; Kitchin 2013).

Clearly, the first step in human mobility modelling is to understand how people move. The availability of big mobility data, such as massive traces from GPS devices (Pappalardo et al. 2013b), mobile phone networks (González et al. 2008) and social media records (Spinsanti et al. 2013), offers nowadays the possibility to observe human movements at large scales and in great detail (Barbosa-Filho et al. 2017). Many studies relied on this opportunity to provide a series of novel insights on the quantitative spatio-temporal patterns characterizing human mobility. These studies observe that human mobility is characterized by a stunning heterogeneity of travel patterns, i.e., a heavy tail distribution in trip distances (Brockmann et al. 2006; González et al. 2008) and the characteristic distance traveled by individuals, the so-called radius of gyration (González et al. 2008; Pappalardo et al. 2015b). Moreover human mobility is characterized by a high degree of predictability (Eagle and Pentland 2009; Song et al. 2010b), a strong tendency to spend most of the time in a few locations (Song et al. 2010a), and a propensity to visit specific locations at specific times (Jiang et al. 2012; Rinzivillo et al. 2014).

Building upon the above findings, many generative algorithms of human mobility have been proposed which try to reproduce the characteristic properties of human mobility trajectories (Karamshuk et al. 2011; Barbosa-Filho et al. 2017). The goal of generative algorithms of human mobility is to create a population of agents whose mobility patterns are statistically indistinguishable from those of real individuals. Typ-

ically each generative algorithm focuses on just a few properties of human mobility. A class of algorithms aims to realistically represent spatial properties: they are mainly concerned with reproducing the trip distance distribution (Brockmann et al. 2006; González et al. 2008) or the visitation frequency to a set of preferred locations (Song et al. 2010a; Pappalardo et al. 2015b). Another class of algorithms focus on the accurate representation of the time-varying behavior of individuals, relying on detailed schedules of human activities (Jiang et al. 2012; Rinzivillo et al. 2014). However, the major challenge for generative algorithms lies in the creation of realistic temporal patterns, in which various temporal statistics observed empirically are simultaneously reproduced, including the number and sequence of visited locations together with the time and duration of the visits. In particular, the biggest hurdle consists in the simultaneous description of an individual's routine and sporadic mobility patterns. Currently there is no algorithm able to reproduce the individuals' recurrent or quasi-periodic daily schedules, and at the same time to allow for the possibility that individuals may break the routine and modify their habits during periods of unpredictability of variable duration.

In this work we present DITRAS (DIary-based TRAjectory Simulator), a framework to simulate the spatio-temporal patterns of human mobility. The key idea of DITRAS is to separate the temporal characteristics of human mobility from its spatial characteristics. In order to do that, DITRAS operates in two steps. First, it generates a mobility diary using a diary generator. A mobility diary captures the temporal patterns of human mobility by specifying the arrival time and the time spent in each location visited by the individual. A diary generator is an algorithm which generates a mobility diary for an individual given a diary length. In this paper we propose a data-driven algorithm called Mobility Diary Learner (MDL) which is able to infer from real mobility data a diary generator, MD, represented as a Markov model. The Markov model captures the propensity of individuals to follow quasi-periodic daily schedules as well as to break the routine and modify their mobility habits.

Second, DITRAS transforms the mobility diary into a mobility trajectory by using proper mechanisms for the exploration of locations on the mobility space, so capturing the spatial patterns of human movements. The trajectory generator we propose, d-EPR, is based on previous research by the authors (Pappalardo et al. 2015b, 2016a) and embeds mechanisms to explore new locations and return to already visited locations. The exploration phase takes into account both the distance between locations and their relevance on the mobility space, though taking into account the underlying urban structure and the distribution of population density.

We instantiate DITRAS with the proposed diary and trajectory generators and compare it with nation-wide mobile phone data, region-wide GPS vehicular data and synthetic trajectories produced by other generative algorithms on a set of nine different standard mobility measures. We show that $d\text{-EPR}_{\text{MD}}$, a generative algorithm created by combining diary generator MD with trajectory generator d-EPR, simulates the spatio-temporal properties of human mobility in a realistic manner, typically reproducing the mobility patterns of real individuals better than the other considered algorithms. Moreover, we show that the distribution of standard mobility measures can be accurately reproduced only by modelling both the spatial and the temporal aspects of human mobility. In other words, the spatial mechanisms and the temporal

mechanisms have to be modeled together by proper diary and trajectory generators in order to reproduce the observed human mobility patterns in an accurate way. The generative algorithm we propose, d-EPR $_{\text{MD}}$, captures both the spatial and the temporal dimensions of human mobility and is a useful tool to develop more reliable protocols for ad hoc networks as well as to perform realistic simulation and what-if scenarios in urban contexts. In summary this paper provides the following novel contributions:

- the modeling framework DITRAS which allows for the combinations of different spatial and temporal mechanisms of human mobility and whose code is freely available (https://github.com/jonpappalord/DITRAS);
- the data-driven algorithm MDL to construct from real mobility data a diary generator (MD) which is realistic in reproducing the temporal patterns of human mobility;
- a comparison of existing algorithms as well as algorithms resulting from novel combinations of temporal and spatial mechanisms, on a set of nine mobility measures and two large-scale mobility datasets.

Our modeling framework goes towards a comprehensive approach which combines a network science perspective and a data mining perspective to improve the accuracy and the realism of human mobility models.

This paper is organized as follows. Section 2 revises the relevant literature on human mobility modelling. In Sect. 3 we present the structure of the DITRAS framework. Section 4 describes the first step of DITRAS, the generation of the mobility diary, and in Sect. 4.1 we describe the mobility diary learner MDL and the Markov model. Section 5 describes the second step of DITRAS, the generation of the mobility trajectory, and in Sect. 5.1 we propose a trajectory generator called d-EPR. Section 6 shows the comparison between an instantiation of DITRAS with the proposed diary and trajectory generators with real trajectory data and the trajectories produced by other generative algorithms. In Sect. 6.4 we discuss the obtained results and, finally, Sect. 7 concludes the paper.

2 Related work

All the main studies in human mobility document a stunning heterogeneity of human travel patterns that coexists with a high degree of predictability: individuals exhibit a broad spectrum of mobility ranges while repeating daily schedules dictated by routine (Giannotti et al. 2013). Brockmann et al. study the scaling laws of human mobility by observing the circulation of bank notes in United States, finding that travel distances of bank notes follow a power-law behavior (Brockmann et al. 2006). González et al. analyze a nation-wide mobile phone dataset and find a large heterogeneity in human mobility ranges (González et al. 2008): (i) travel distances of individuals follow a power-law behavior, confirming the results by Brockmann et al.; (ii) the radius of gyration of individuals, i.e., their characteristic traveled distance, follows a power-law behavior with an exponential cutoff. Song et al. observe on mobile phone data that individuals are characterized by a power-law behavior in waiting times, i.e., the time between a displacement and the next displacement by an individual (Song et al.

2010a). Pappalardo et al. find the same mobility patterns on a dataset storing the GPS traces of 150,000 private vehicles traveling during one month in Tuscany, Italy (Pappalardo et al. 2013b). Song et al. study the entropy of individuals' movements and find a high predictability in human mobility, with a distribution of users' predictability peaked at approximately 93% and having a lower cutoff at 80% (Song et al. 2010b). Pappalardo et al. analyze mobile phone data and GPS tracks from private vehicles and discover that individuals split into two profiles, returners and explorers, with distinct mobility and geographical patterns (Pappalardo et al. 2015b). Several studies focus on the prediction of the kind of activity associated to individuals' trips on the only basis of the observed displacements (Liao et al. 2007; Jiang et al. 2012; Rinzivillo et al. 2014), and to discover geographic borders according to recurrent trips of private vehicles (Rinzivillo et al. 2012; Thiemann et al. 2010), or to predict the formation of social ties (Cho et al. 2011; Wang et al. 2011). Other works demonstrate the connection between human mobility and social networks, highlighting that friendships and other types of social relations are significant drivers of human movements (Brown et al. 2013b; Hristova et al. 2016; Wang et al. 2011; Volkovich et al. 2012; Brown et al. 2013a; Hossmann et al. 2011a, b).

How to combine the discovered patterns to create a generative algorithm that reproduces the salient aspects of human mobility is an open task. This task is particularly challenging because generative algorithms should be as simple, scalable and flexible as possible, since they are generally purposed to large-scale simulation and what-if analysis. In the literature many generative algorithms have been proposed so far to model individual human mobility patterns (Karamshuk et al. 2011; Barbosa-Filho et al. 2017).

Some algorithms try to reproduce the heterogeneity of individual human mobility and simulate how individuals visits locations. ORBIT (Ghosh et al. 2005) is an example of such algorithms. It splits into two phases: (i) at the beginning of the simulation it generates a predefined set of locations on a bi-dimensional space; (ii) then every synthetic individual selects a subset of these locations and moves between them according to a Markov chain. In the Markov chain every state represents a specific location in the scenario and proper probability of transitions guarantee a realistic distribution of location frequencies. SLAW (Self-similar Least-Action Walk) produces mobility traces having specific statistical features observed on human mobility data, namely power-law waiting times and travel distances with a heavy-tail distribution (Lee et al. 2012, 2009). In a first step SLAW generates a set of locations on a bi-dimensional space so that the distance among them features a heavy-tailed distribution. Then, a synthetic individual starts a trip by randomly choosing a location as starting point and making movement decisions based on the LATP (Least-Action Trip Planning) algorithm. In LATP every location has a probability to be chosen as next location that decreases with the power-law of the distance to the synthetic individual's current location. SLAW is used in several studies of networking and human mobility modelling and is the base for other generative algorithms for human mobility, such as SMOOTH (Munjal et al. 2011), MSLAW (Schwamborn and Aschenbruck 2013) and TP (Solmaz et al. 2015, 2012).

Small World In Motion (SWIM) is based on the concept of location preference (Kosta et al. 2010). First, each synthetic individual is assigned to a home location,

which is chosen uniformly at random on a bi-dimensional space. Then the synthetic individual selects a destination for the next move depending of the weight of each location, which grows with the popularity of the location and decreases with the distance from the home location. The popularity of a location depends on a collective preference calculated as the number of other people encountered the last time the synthetic individual visited the location. Another category of generative algorithms combine notions about the sociality of individuals with mobility patterns to define socio-mobility models, demonstrating how they can be exploited to design more realistic protocols for ad hoc and opportunistic networks (Borrel et al. 2009; Yang et al. 2010; Fischer et al. 2010; Boldrini and Passarella 2010; Musolesi and Mascolo 2007).

In contrast with many generative algorithms of human mobility, the Exploration and Preferential Return (EPR) model does not fix in advance the number of visited locations on a bi-dimensional space but let them emerge spontaneously (Song et al. 2010a). The model exploits two basic mechanisms that together describe human mobility: exploration and preferential return. Exploration is a random walk process with a truncated power-law jump size distribution (Song et al. 2010a). Preferential return reproduces the propensity of humans to return to the locations they visited frequently before (González et al. 2008). A synthetic individual in the model selects between these two mechanisms: with a given probability the synthetic individual returns to one of the previously visited places, with the preference for a location proportional to the frequency of the individual's previous visits. With complementary probability the synthetic individual moves to a new location, whose distance from the current one is chosen from the truncated power-law distribution of travel distances as measured on empirical data (González et al. 2008). The probability to explore decreases as the number of visited locations increases and, as a result, the model has a warmup period of greedy exploration, while in the long run individuals mainly move around a set of previously visited places. Recently the EPR model has been improved in different directions, such as by adding information about the recency of location visits during the preferential return step (Barbosa et al. 2015), or adding a preferential exploration step to account for the collective preference for locations and the returners and explorers dichotomy, as the authors of this paper have done in previous research by defining the d-EPR model (Pappalardo et al. 2015b, 2016a). It is worth noting that although the algorithms described above are able to reproduce accurately the heterogeneity of mobility patterns, none of them can reproduce realistic temporal patterns of human movements.

Recent research on human mobility show that individuals are characterized by a high regularity and the tendency to come back to the same few locations over and over at specific times (González et al. 2008; Pappalardo et al. 2013b). Temporal models focus on these temporal patterns and try to reproduce accurately human daily activities, schedules and regularities. Zheng et al. (Zheng et al. 2010) use data from a national survey in the US to extract realistic distribution of address type, activity type, visiting time and population heterogeneity in terms of occupation. They first describe streets and avenues on a bi-dimensional space as horizontal and vertical lines with random length, and then use the Dijkstra's algorithm to find the shortest path between two activities taking into account different speed limits assigned to each street. WDM (Working Day Movement) distinguishes between inter-building and intra-building

movements (Ekman et al. 2008). It consists of several submodels to describe mobility in home, office, evening and different transportation means. For example a home model reproduces a sojourn in a particular point of a home location while an office model reproduces a star-like trajectory pattern around the desk of an individual at specific coordinates inside an office building. Although Zheng et al.'s algorithm and WDM provide an extremely thorough representations of human movements in particular scenarios, they suffer two main drawbacks: (i) they represent specific scenarios and their applicability to other scenarios is not guaranteed; (ii) they are too complex for analytical tractability; (iii) they generally fail in capturing some global mobility patterns observed in individual human mobility, e.g., the distribution of radius of gyration. A recent study (McInerney et al. 2013) proposes methods to identify and predict departures from routine in individual mobility using information-theoretic metrics, such as the instantaneous entropy, and developing a Bayesian framework that explicitly models the tendency of individuals to break from routine.

Position of our work. From the literature it clearly emerges that existing generative algorithms for human mobility are not able to accurately capture at the same time the heterogeneity of human travel patterns and the temporal regularity of human movements. On the one hand exploration models accurately reproduce the heterogeneity of human mobility but do not account for regularities in human temporal patterns. On the other hand temporal models accurately reproduce human mobility schedules paying the price in complexity, but fail in capturing some important global mobility patterns observed in human mobility. In this paper we try to fill this gap and propose d-EPR $_{\mathrm{MD}}$, a scalable generative algorithm that creates synthetic individual trajectories able to capture both the heterogeneity of human mobility and the regularity of human movements. Despite its great flexibility, d-EPR $_{\mathrm{MD}}$ is to a large extent analytically tractable and several statistics about the visits to routine and non-routine locations can be derived mathematically. In fact, since the temporal mechanism of d-EPR $_{\mathrm{MD}}$ is based on a Markov chain, using standard results in probability theory one can compute various quantities, including the probability to go between any two states in a given number of steps, the average number of visits to a state before visiting another state, the average time to go from one state to another and the probability to visit one state before another. Moreover the spatial mechanism of d-EPR $_{\mathrm{MD}}$ is based on the EPR model for which various analytical results, such as the distributions of the radii of gyration and of the location frequencies, have been derived (Song et al. 2010a). The data-driven algorithm MDL (Mobility Diary Learner), is another novel contribution of this paper. MDL infers from real mobility data a diary generator for realistic mobility diaries. It is highly adaptive and can be applied to different geographic areas and different types of mobility data.

The modelling framework we propose, DITRAS, can generate synthetic mobility trajectories and can be easily integrated in transportation forecast models to infer trip demand. Our approach has some similarity with activity-based models (Bellemans et al. 2010), as they both aim to estimate trip demand by reproducing realistic individual temporal patterns, however there are important differences between the two approaches. In fact, while the goal of activity-based models is to produce detailed agendas filled with activities performed by the agents and are calibrated on surveys

with a limited number of participants, our framework produces mobility diaries containing the time and duration of the visits in the various locations without explicitly specifying the type of activity performed there, and is calibrated on a large population of mobile phone users.

A recent paper introduces TimeGeo, a modelling framework to generate a population of synthetic agents with realistic spatio-temporal trajectories (Yang et al. 2016). Similarly to the modelling framework presented here, TimeGeo combines a Markov model to generate temporal patterns with the correct periodicity and duration of visits, with a model to reproduce spatial patterns with the characteristic number of visits and distribution of distances. Albeit having similar aims, there are important differences between our modelling approach and TimeGeo's. In fact, while TimeGeo proposes a parsimonious model which is based on few tunable parameters and is to some extent analytically tractable, the approach proposed in this paper is markedly data driven and parameter-free, with a greater level of complexity which ensures the necessary flexibility to reproduce realistic temporal patterns.

3 The DITRAS modelling framework

DITRAS is a modelling framework to simulate the spatio-temporal patterns of human mobility in a realistic way.[1] The key idea of DITRAS is to separate the temporal characteristics of human mobility from its spatial characteristics. For this reason, DITRAS consists of two main phases (Fig. 1): first, it generates a mobility diary which captures the temporal patterns of human mobility; second it transforms the mobility diary into a sampled mobility trajectory which captures the spatial patterns of human movements. In this section we define the main concepts which constitute the mechanism of DITRAS.

The output of a DITRAS simulation is a sampled mobility trajectory for a synthetic individual. A mobility trajectory describes the movement of an object as a sequence of time-stamped locations. The location is described by two coordinates, usually a latitude-longitude pair or ordinary Cartesian coordinates, as formally stated by the following definition:

Definition 1 *(Mobility trajectory)* A mobility trajectory is a sequence of triples $T = \langle (x_1, y_1, t_1), \ldots, (x_n, y_n, t_n) \rangle$, where t_i $(i = 1, \ldots, n)$ is a timestamp, $\forall_{1 \leq i < n}\ t_i < t_{i+1}$ and x_i, y_i are coordinates on a bi-dimensional space.

For modelling purposes it is convenient to define a sampled mobility trajectory, $S^{(t)}$, which can be obtained by sampling the mobility trajectory at regular time intervals of length t seconds:

Definition 2 *(Sampled mobility trajectory)* A sampled mobility trajectory is a sequence $S^{(t)} = \langle l_1, \ldots, l_N \rangle$, where l_i $(i = 1, \ldots, N)$ is the geographic location where the individual spent the majority of time during time slot i, i.e., between $(i-1)t$ and ti seconds from the first observation. N is the total number of time slots considered. A location l_i is described by coordinates on a bi-dimensional space.

[1] The Python code of DITRAS is freely available for download on a public GitHub repository: https://github.com/jonpappalord/DITRAS

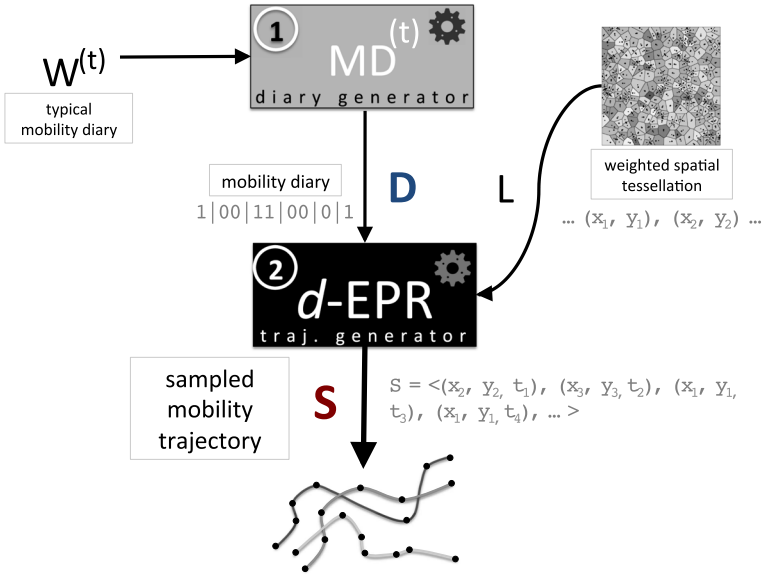

Fig. 1 Outline of the DITRAS framework. DITRAS combines two probabilistic models: a diary generator (e.g., $\text{MD}^{(t)}$) and trajectory generator (e.g., d-EPR). The diary generator uses a typical diary $W^{(t)}$ to produce a mobility diary D. The mobility diary D is the input of the trajectory generator together with a weighted spatial tessellation of the territory L. From D and L the trajectory generator produces a sampled mobility trajectory S

To generate a sampled mobility trajectory DITRAS exploits two probabilistic models: a diary generator and a trajectory generator (see Fig. 1). In this paper we propose as diary generator $\text{MD}^{(t)}$, a Markov model responsible for reproducing realistic temporal mobility patterns, such as the distribution of the number of trips per day and the tendency of individuals to change location at specific hours of the day (González et al. 2008; Jiang et al. 2012). Essentially, $\text{MD}^{(t)}$ captures the tendency of individuals to follow or break a temporal routine at specific times. As trajectory generator we propose the d-EPR generative model (Pappalardo et al. 2015b, 2016a), which is able to reproduce realistic spatial mobility patterns, such as the heavy-tail distributions of trip distances (Brockmann et al. 2006; González et al. 2008; Pappalardo et al. 2013b) and radii of gyration (González et al. 2008; Pappalardo et al. 2013b, 2015b), as well as the characteristic visitation patterns, such as the uneven distribution of time spent in the various locations (Song et al. 2010a; Pappalardo et al. 2013b). d-EPR embeds a mechanism to choose a location to visit on a bi-dimensional space given the current location, the spatial distances between locations and the relevance of each location.

Figure 1 provides an outline of DITRAS and Algorithm 1 describes its pseudocode. DITRAS is composed of two main steps. During the first step, the diary generator builds a mobility diary D of N time slots, each of duration t. The operation of this step is described in detail in Sect. 4. During the second step, DITRAS uses the trajectory generator and a given spatial tessellation L to transform the mobility diary into a sampled mobility trajectory. We describe in detail the second step of DITRAS in Sect. 5. Note

The DITRAS framework

input : $L = \{(l_1, r_1), \ldots, (l_n, r_n)\}$, weighted spatial tessellation
G, diary generator
N, length of trajectory to generate
W, typical diary
output: $S = \langle (x_1, y_1, t_1), \ldots, (x_n, y_n, t_n) \rangle$, sampled mobility trajectory of length N

```
D = generateMobilityDiary(G, N) // use the diary generator DG to create a
    mobility diary D of length N
S = generateMobilityTrajectory(D, L, W) // scan the mobility diary D and create
    a sample mobility trajectory S of length N
return S

Function generateMobilityTrajectory(D, L, W)
    S = newList()
    t = 1
    W_m = assignLocationsTo(W) // assign a physical location to every abstract
        location in typical diary W
    while d < length(D) do
        // scan the mobility diary D
        if D[d] = | then
            // when it sees a separator '|'
            d = d + 1
            continue
        end
        if D[d] = 0 then
            // the individual follows the routine (i.e., she visits a typical
                location)
            S.append((W_m[t], t))
            t = t + 1
        end
        else
            // the individual breaks the routine
            l = TG(S, P) // call the trajectory generator TG to obtain the next
                location to visit
            S.append((l, t))
            t = t + 1
            j = d + 1
            while D[d] = D[j] do
                // stay in location l until the next separator appears
                S.append((l, t))
                t = t + 1
                j = j + 1
            end
            d = j - 1
        end
        d = d + 1
    end
return S
```

Algorithm 1: The algorithm describing the DITRAS framework. Python code is freely available at https://github.com/jonpappalord/DITRAS.

that the two-step process described above is a general framework common to many generative models of human mobility, which are often composed by two sequential parts, the first generating temporal patterns and the second determining the spatial trajectory. However, in some models the division between the temporal and the spatial mechanisms is present but not explicitly acknowledged.

In Sect. 6 we will instantiate DITRAS by using $\text{MD}^{(t)}$ and d-EPR and compare it with other generative models obtained combining diary generators (first step) with trajectory generators (second step).

4 Step 1: Generation of mobility diary

A diary generator G produces a mobility diary, $D^{(t)}$, containing the sequence of trips made by a synthetic individual during a time period divided in time slots of t seconds. For example, $\text{G}^{(3600)}$ and $\text{G}^{(60)}$ produce mobility diaries with temporal resolutions of one hour and one minute, respectively. In Sect. 4.1 we illustrate a data-driven algorithm to construct a diary generator, $\text{MD}^{(t)}$, using real mobility trajectory data such as mobile phone data.

To separate the temporal patterns from the spatial ones, we define the abstract mobility trajectory, $A^{(t)}$, which contains the time ordered list of the "abstract locations" visited by a synthetic individual during a period divided in time slots of t seconds. An abstract location uniquely identifies a place where the individual is stationary, like home or the workplace, but it does not contain any information on the specific geographic position of the location (i.e., its coordinates). The abstract mobility trajectory is thus equivalent to the sampled mobility trajectory where the geographic locations, l_k, are substituted by placeholders, a_k, called abstract locations:

Definition 3 *(Abstract mobility trajectory)* An abstract mobility trajectory is a sequence $A^{(t)} = \langle a_1, \ldots, a_N \rangle$, where a_i $(i = 1, \ldots, N)$ is the abstract location where the individual spent the majority of time during time slot i, i.e., between $(i-1)t$ and it seconds from the first observation.

The mobility diary, $D^{(t)}$, is generated with respect to a typical mobility diary, $W^{(t)}$, which represents the individual's routine. $W^{(t)}$ is a sequence of time slots of duration t seconds and specifies the typical and most likely abstract location the individual visits in every time slot. Here we consider the simplest choice of typical mobility diary, in which the most likely location where a synthetic individual can be found at any time is her home location. It is possible to relax this simplifying assumption and estimate an individual's typical mobility diary from the data by computing her mobility *regularity*, which is the time series of the most visited location in each time slot (Song et al. 2010b). Computing the weekly mobility regularity of individuals on real large-scale mobile phone data and GPS vehicular data and performing a clustering of their typical diaries we find that there is one dominant cluster containing $\approx 90\%$ of the individuals and whose representative typical diary has a single location (see "Appendix A"). This result supports the validity of the simplifying assumption to consider one typical diary with a single location for all agents. The proposed generative model does not change if there are two or more typical mobility diaries which have more than one typical location.

When a synthetic individual is generated it can be randomly assigned to one of the typical diaries in proportion to the overall frequency of the various diaries among real users. Then, the rest of the algorithm remains the same.

Definition 4 *(Typical mobility diary)* A typical mobility diary is a sequence $W^{(t)} = \langle w_1, \ldots, w_N \rangle$ where $w_k = w \ \ \forall k = 1, \ldots, N$ denotes the home location of the synthetic individual. N is the total number of time slots considered.

The mobility diary, $D^{(t)}$, specifies whether an individual's abstract mobility trajectory, $A^{(t)}$, follows her typical mobility diary, $W^{(t)}$, or not. In particular, for every time slot i, $D^{(t)}(i)$ can assume two values:

- $D^{(t)}(i) = 1$ if $A^{(t)}(i) = W^{(t)}(i)$, meaning that the individual visits the abstract location $W^{(t)}(i)$ following her routine, i.e., she is at home;
- $D^{(t)}(i) = 0$ if $A^{(t)}(i) \neq W^{(t)}(i)$, meaning that the individual visits a location other than the abstract location $W^{(t)}(i)$ being out of her routine.

Definition 5 *(Mobility Diary)* A mobility diary is a sequence $D^{(t)}$ of time slots of duration t seconds generated by the regular language $\mathbb{L} = (1^+|(0^+|)^*)^*$, where 1 at time slot i indicates that the individual visits the abstract location in her typical diary at time i, $W^{(t)}(i)$, and 0 indicates a visit to a location different from the abstract location $W^{(t)}(i)$. The symbol "|" indicates a transition or trip between two different abstract locations.

An example of mobility diary generated by language $\mathbb{L}$ is $D^{(t)} = \langle 11|00|0|1 \rangle$. The first two entries indicate that $A^{(t)}(1) = W^{(t)}(1)$ and $A^{(t)}(2) = W^{(t)}(2)$, i.e., the individual follows her routine and she is at home. Next, the third, fourth and fifth entries indicate that $A^{(t)}(3) \neq W^{(t)}(3)$, $A^{(t)}(4) \neq W^{(t)}(4)$ and $A^{(t)}(5) \neq W^{(t)}(5)$, i.e., the individual breaks the routine and visits a non-typical location for two consecutive time slots, then she visits a different non-typical location for one time slot. Finally, the last time slot indicates that $A^{(t)}(6) = W^{(t)}(6)$, the individual follows the routine and returns home. We assume that the travel time between any two locations is of negligible duration.

4.1 Mobility diary learner (MDL)

In this section we propose diary generator $\text{MD}^{(t)}$ and illustrate MDL (Mobility Diary Learner), a data-driven algorithm to compute MD from the abstract mobility trajectories of a set of real individuals (Algorithm 2). We use a Markov model to describe the probability that an individual follows her routine and visits a typical location at the usual time, or she breaks the routine and visits another location. First, MDL translates mobility trajectory data of real individuals into abstract mobility trajectories (Sect. 4.1.1). Second, it uses the obtained abstract trajectory data to compute the transition probabilities of the Markov model $\text{MD}^{(t)}$ (Sect. 4.1.2).

4.1.1 Mobility trajectory data

The construction of $\text{MD}^{(t)}$ is based on mobility trajectory data of real individuals. We assume that raw mobility trajectory data describing the movements of a set of

MDL (Mobility Diary Learner)

input : $D = \{T_1, \ldots, T_n\}$, dataset of real trajectories of n agents
t, time slot length
output: G, a Markov chain

```
G = emptyMarkovChain()
forall the i ∈ {1, ..., n} do
    A_i = createTimeSeries(T_i) // create abstract trajectory of i
    G = updateMarkovChain(A_i) // update the Markov chain using A_i
end
return G

Function updateMarkovChain(A, G)
    slot = 0
    while slot < len(A) − 1 do
        h = slot%24 // hour of the day
        next_h = (h + 1)%24 // next hour of the day
        loc_h = A[slot] // abstract location at the slot
        loc_{h+1} = A[slot + 1] // abstract location at next slot
        if loc_h == 1 then
            if loc_{h+1} == 1 then
                // Case 1: loc_h is typical and loc_{h+1} is typical
                G[(h, 1), (next_h, 1)] = G[(h, 1), (next_h, 1)] + 1
            end
            else
                // Case 2: loc_h is typical and loc_{h+1} is not typical
                τ = 1
                for j = slot + 2 to len(A) do
                    loc2_h = A[j]
                    if loc2_h == loc2_{h+1} then
                        τ = τ + 1
                    end
                    else
                        break
                    end
                end
                h_τ = (h + τ)%24
                G[(h, 1), (h_τ, 0)] = G[(h, 1), (h_τ, 0)] + 1
                slot = j − 1
            end
        end
        else
            if loc_{h+1} == 1 then
                // Case 3: loc_h is not typical and loc_{h+1} is typical
                G[(h, 0), (next_h, 1)] = G[(h, 0), (next_h, 1)] + 1
            end
            else
                // Case 4: both loc_h and loc_{h+1} are not typical
                τ = 1
                for j = slot + 2 to len(A) do
                    loc2_h = A[j]
                    if loc2_h == loc2_{h+1} then
                        τ = τ + 1
                    end
                    else
                        break
                    end
                    h_τ = (h + τ)%24
                    G[(h, 0), (h_τ, 0)] = G[(h, 0), (h_τ, 0)] + 1
                    slot = j − 1
                end
            end
        end
        slot = slot + 1
    end
    G = normalizeMarkovChain(G)
    return G
```

Algorithm 2: Algorithm for the construction of the MD generator.

individuals are in the form $\langle (u_1, x_1, y_1, t_1), \ldots, (u_n, x_n, y_n, t_n) \rangle$ where u_i indicates the individual who visits location (x_i, y_i) at time t_i, $\forall_{1 \leq i < n}\ t_i < t_{i+1}$.

Mobility trajectory data can be obtained from various sources (e.g., mobile phones, GPS devices, geosocial networks) and describe the movements of individuals on a territory. Since the purpose of $\mathrm{MD}^{(t)}$ is to capture the temporal patterns regardless the geographic position of locations, we translate raw mobility trajectory data into abstract mobility trajectories (see definition in Section 3).

Starting from the raw trajectory data, we assign an abstract location to every time slot in an individual's abstract mobility trajectory $A^{(t)}$ according to the following method. If the individual visits just one location during time slot i, we assign that location to i. If the individual visits multiple locations during slot i, we choose the most frequent location in i, i.e., the location where the individual spends most of the time during the time slot. If there are multiple locations with the same visitation frequency in time slot i, we choose the location with the highest overall frequency. If there is no information in the abstract trajectory data about the location visited in time slot i (e.g., no calls during the time slot in the case of mobile phone data), we assume no movement and choose the location assigned to time slot $i - 1$.

To clarify the method let us consider the following example. A mobile phone user has the following hourly time series of calls: $[A, A, \bullet, \bullet, B, (C, C, B, B)]$, where A, B, C are placeholders for different cell phone towers (i.e., abstract locations). Here the symbol $\bullet$ indicates that there is no information in the data about the location visited during the 1-hour time slot, while all the locations in round brackets are visited during the same time slot. Using the method described above, the abstract mobility trajectory of the individual becomes $A^{(3600)} = \langle A, A, A, A, B, B \rangle$ because: (i) the two $\bullet$ symbols in the third and fourth time slots are substituted by location A assuming no movement with respect to the second time slot; (ii) the location assigned to the last time slot is B since C and B have the same visitation frequency in (C, C, B, B) but $f(B) > f(C)$, i.e., B has the highest overall visitation frequency.

It is worth noting that the choice of the duration of the time slot, t, is crucial and depends on the specific kind of mobility trajectory data used. GPS data from private vehicles, for example, generally provide accurate information about the location of the vehicle every few seconds. In this scenario, a time slot duration of one minute can be a reasonable choice. In contrast when dealing with mobile phone data a time slot duration of an hour or half an hour is a more reliable choice, since the majority of individuals have a low call frequency during the day (Pappalardo et al. 2015b).

4.1.2 Markov model transition probabilities

Let $A_u = \langle a_0^{(u)}, \ldots, a_{n-1}^{(u)} \rangle$ and $W_u = \langle w_0^{(u)}, \ldots, w_{n-1}^{(u)} \rangle$ be the abstract mobility trajectory and the typical mobility diary of individual $u \in U$, where U is the set of all individuals in the data – we omit the superscript (t) for clarity. Elements $a_h^{(u)} \in A_u$ and $w_h^{(u)} \in W_u$ denote the abstract and the typical locations visited by individual u at time slot h with $h = 0, \ldots, N-1$.

A state in the Markov model MD is a tuple of two elements $s = (h, R)$. The state's first element, h, is the time slot of the time series denoted by an integer between 0

Table 1 Formulae to compute the transition probabilities of the Markov chain MD from abstract mobility trajectories

Transition, $s \to s'$	Frequency, $\mathrm{MD}_{ss'}$
$(h,1) \to (h+1,1)$	$\frac{\sum_{u\in U}\sum_{a\in A_u}\delta_h^u(a)\delta_{h+1}^u(a)}{\sum_{u\in U}\sum_{a\in A_u}\delta_h^u(a)}$
$(h,1) \to (h+\tau,0)$	$\frac{\sum_{u\in U}\sum_{a\in A_u}\delta_h^u(a)[1-\delta_{h+1}^u(a)]\prod_{i=1}^{\tau-1}\hat{\delta}_{h+i}^u(a)[1-\hat{\delta}_{h+\tau}^u(a)]}{\sum_{u\in U}\sum_{a\in A_u}\delta_h^u(a)}$
$(h,0) \to (h+1,1)$	$\frac{\sum_{u\in U}\sum_{a\in A_u}[1-\delta_h^u(a)]\delta_{h+1}^u(a)}{\sum_{u\in U}\sum_{a\in A_u}[1-\delta_h^u(a)]}$
$(h,0) \to (h+\tau,0)$	$\frac{\sum_{d\in D}[1-\delta_h^u(a)][1-\delta_{h+1}^u(a)][1-\hat{\delta}_h^u(a)]\prod_{i=1}^{\tau-1}\hat{\delta}_{h+i}^u(a)[1-\hat{\delta}_{h+\tau}^u(a)]}{\sum_{u\in U}\sum_{a\in A_u}[1-\delta_h^u(a)]}$

and $N-1$. The state's second element, R, is a boolean variable that is 1 (True) if at time slot h the individual is in her typical location, $w_h^{(u)}$, and 0 (False) otherwise – just like in the mobility diary. In total there are $N \times 2 = 2N$ possible states in the model. The transition matrix, MD, is a $2N \times 2N$ stochastic matrix whose element $\mathrm{MD}_{ss'}$ corresponds to the conditional probability of a transition from state s to state s', $\mathrm{MD}_{ss'} \equiv p(s'|s)$. The normalization condition imposes that the sum over all elements of any row s is equal to 1, $\sum_{s'} \mathrm{MD}_{ss'} = 1, \forall s$. We consider two types of transitions, $s \to s'$, depending on whether in state s the individual is in typical location or not:

- if the individual is in the typical location at time slot h, i.e., $s = (h,1)$, then she can either go to the next typical location at time slot $h+1$, $s = (h,1) \to s' = (h+1,1)$, or go to a non-typical location and stay there for τ time slots, $s = (h,1) \to s' = (h+\tau,0)$;
- if instead the individual is not in the typical location at time slot h, i.e., $s = (h,0)$, then she can either go to the typical location at time slot $h+1$, $s = (h,0) \to s' = (h+1,1)$, or go to a different non-typical location and stay there for τ time slots, $s = (h,0) \to s' = (h+\tau,0)$.

The formulae to compute the empirical frequencies for the four types of transitions are shown in Table 1. In the table, $\delta_x^u(a) = \delta(a_x^{(u)}, w_x^{(u)})$, $\hat{\delta}_x^u(a) = \delta(a_x^{(u)}, a_{x+1}^{(u)})$, where $\delta(i,j) = 1$ if $i = j$ and 0 otherwise, is the Kronecker delta. By convention, the product $\prod_{i=1}^{\tau-1} \ldots$ is equal to 1 if $\tau = 1$.

5 Step 2: Generation of sampled mobility trajectory

Starting from the mobility diary $D^{(t)}$, the sampled mobility trajectory $S^{(t)}$ is generated to describe the movement of a synthetic individual between a set of discrete locations called weighted spatial tessellation. A weighted spatial tessellation is a partition of a bi-dimensional space into locations each having a weight corresponding to its relevance.

Definition 6 *(Weighted spatial tessellation)* A weighted spatial tessellation is a set of tuples $L = \{(l_1, r_1), \ldots, (l_m, r_m)\}$, where $r_j \in \mathbb{N}$ $(j = 1, \ldots, m)$ is the relevance of a location and the l_j are a set of non-overlapping polygons that cover the bi-dimensional

space where individuals can move. The location of each polygon is identified by the coordinates of its centroid, (x_j, y_j).

The weighted spatial tessellation indicates the possible physical locations on a finite bi-dimensional space a synthetic individual can visit during the simulation. The relevance of a location measures its popularity among real individuals: locations of high relevance are the ones most frequently visited by the individuals (Pappalardo et al. 2015b, 2016a). The relevance is introduced to generate synthetic trajectories that take into account the underlying urban structure. An example of weighted spatial tessellation is the one defined by a set of mobile phone towers, where the relevance of a tower can be estimated as the number of calls performed by mobile phone users during a period of observation, and the polygons correspond to the regions obtained from the Voronoi partition induced by the towers. If information about location relevance is not available to the user of the simulator, the distribution of population can be used to estimate the relevance of the locations. For example, the websites http://sedac.ciesin.columbia.edu/ and http://www.worldpop.org.uk/ provide a fine-grained spatial tessellation for the entire globe, together with an estimate of population density in every location.

First, DITRAS assigns to every abstract location in the typical mobility diary $W^{(t)}$ a physical location on the weighted spatial tessellation L, creating $W_m^{(t)}$, a typical mobility diary where each abstract location has a specific geographic position (Algorithm 1, line 4, procedure `assignLocationsTo`). The geographic position of an abstract location is chosen according to the distribution of location relevance specified in the spatial tessellation, i.e., the more relevant a location is the more likely it is chosen as a geographic position of an abstract location. This choice ensures the generation of synthetic data with a realistic distribution of locations across the territory (Pappalardo et al. 2016a). Next, DITRAS scans $D^{(t)}$ to assign a physical location to every entry. For every entry $D^{(t)}(i) \in D^{(t)}$ we have two possible scenarios:

- $D^{(t)}(i) = 1$, the entry indicates a visit to a typical location, i.e., the abstract location in $W^{(t)}(i)$ (Algorithm 1, line 12). In this scenario the synthetic individual visits location $l = W_m^{(t)}(i)$ which is added to the sampled trajectory at time slot i, i.e. $S^{(t)}(i) = W_m^{(t)}(i)$ (Algorithm 1, lines 14);
- $D^{(t)}(i) = 0$, the entry indicates a visit to a non-typical location (Algorithm 1, line 17). In this second scenario DITRAS calls the trajectory generator to choose a location l to visit, where $l \neq W_m^{(t)}(i)$ (Algorithm 1, lines 19). The chosen location l is added to the sampled mobility trajectory k times, where k is the number of consecutive 0 characters before the next separator character '|' appears in $D^{(t)}$, i.e., the total number of time slots spent in location l (Algorithm 1, lines 23-27).

Example of trajectory generation To clarify how the second step of DITRAS works let us consider the following example. A synthetic individual is assigned a mobility diary $D^{(t)} = \langle 1|00|1\rangle$ and the chosen typical diary is $W^{(t)} = \langle w, w, w, w\rangle$, where w denotes the individual's home. To generate a synthetic sampled mobility trajectory S, DITRAS operates as follows. First, DITRAS assigns a physical location to the individual's home w, generating $W_m^{(t)} = \langle (x_1, y_1), (x_1, y_1), (x_1, y_1), (x_1, y_1)\rangle$. Next, DITRAS starts

from the first entry $D^{(t)}(1)$. Since $D^{(t)}(1) = 1$ the synthetic individual is at home. Therefore, tuple $(x_1, y_1, 1)$ is added to trajectory S. Next, DITRAS processes the second entry $D^{(t)}(2)$, sees a separator and then proceeds to entry $D^{(t)}(3)$. Since $D^{(t)}(3) = 0$, the synthetic individual is not at home in the third time slot. Hence, DITRAS calls a trajectory generator (e.g., d-EPR) which chooses to visit physical location (x_2, y_2). DITRAS hence adds the tuples $(x_2, y_2, 2)$ and $(x_2, y_2, 3)$ to trajectory S, since there two 0 characters until the next separator in $D^{(t)}$. The last entry $D^{(t)}(6) = 1$ indicates that the synthetic individual returns home in the fourth time slot. So, DITRAS adds tuple $(x_1, y_1, 4)$ to trajectory S. At the end of the execution, the sampled mobility trajectory generated by DITRAS is $S = \langle (x_1, y_1, 1), (x_2, y_2, 2), (x_2, y_2, 3), (x_1, y_1, 4) \rangle$.

5.1 The d-EPR model

As trajectory generator we propose the d-EPR individual mobility model (Pappalardo et al. 2015b, 2016a) that assigns a location on the bi-dimensional space to an entry in mobility diary $D^{(t)}$. The d-EPR (density-Exploration and Preferential Return) is based on the evidence that an individual is more likely to visit relevant locations than non-relevant locations (Pappalardo et al. 2015b, 2016a). For this reason d-EPR incorporates two competing mechanisms, one driven by an individual force (preferential return) and the other driven by a collective force (preferential exploration). The intuition underlying the model can be easily understood: when an individual returns, she is attracted to previously visited places with a force that depends on the relevance of such places at an individual level. In contrast, when an individual explores she is attracted to new places with a force that depends on the relevance of such places at a collective level. In the preferential exploration phase a synthetic individual selects a new location to visit depending on both its distance from the current location, as well as its relevance measured as the collective location's relevance in the bi-dimensional space. In the model, hence, the synthetic individual follows a personal preference when returning and a collective preference when exploring. The d-EPR uses the gravity model (Zipf 1946; Jung et al. 2008; Lenormand et al. 2016) to assign the probability of a trip between any two locations in L, which automatically constrains individuals within a territory's boundaries. The usage of the gravity model is justified by the accuracy of the gravity model to estimate origin-destination matrices even at the country level (Erlander and Stewart 1990; Wilson 1969; Simini et al. 2012; Balcan et al. 2009; Lenormand et al. 2016).

Algorithm 3 describes how d-EPR assigns a location on the bi-dimensional space defined by a spatial tessellation L for an entry in mobility diary $D^{(t)}$. The d-EPR takes in input two variables: (i) the current sampled mobility trajectory of the synthetic individual $S = \langle (x_1, y_1, t_1), \ldots, (x_n, y_n, t_n) \rangle$; (ii) a probability matrix P indicating, for every pair of locations $i, j \in L, i \neq j$ the probability of moving from i to j. Every probability p_{ij} is computed as:

$$p_{ij} = \frac{1}{Z} \frac{r_i r_j}{d_{ij}^2},$$

The d-EPR model

input : $S = \langle (x_1, y_1, t_1), \ldots, (x_n, y_n, t_n) \rangle$, the current sample mobility trajectory of the synthetic individual
P, the gravity-probability matrix
output: j, the next location to visit

$\rho = 0.6, \gamma = 0.21$ // distributions' constants (Pappalardo et al. 2015b, 2016a; Song et al. 2010a)

1 $N = |set(S)|$ // number of distinct visited locations
2 $i = last(S)$// the current location of the synthetic individual

3 $p_{new} =$ getReturnProbability() // generate a probability to return or explore
4 **if** $p_{new} \leq \rho N^{-\gamma}$ **then**
5 $j =$ PreferentialExploration(i, P) // explore a new location
7 **return** j
8 **end**
9 **else**
10 $j =$ PreferentialReturn(S) // return to a previously visited location
12 **return** j
13 **end**

1 **Function** PreferentialExploration(i)
2 $j =$ weightedRandom($P[i]$) // choose j according to prob.s in $P[i]$
4 **return** j

1 **Function** PreferentialReturn(S)
2 $j =$ weightedRandom(S) // choose j according to visitation frequency of locations in S
4 **return** j

Algorithm 3: The psuedo-code of the d-EPR trajectory generator. The function `weightedRandom` randomly chooses an element in a vector according to its probability.

where $r_{i(j)}$ is the relevance of location $i(j)$ as specified in the weighted spatial tessellation L, d_{ij} is the geographic distance between i and j, and $Z = \sum_{i,j \neq i} p_{ij}$ is a normalization constant. The matrix P is computed before the execution of the DITRAS model by using the spatial tessellation L.

With probability $p_{new} = \rho N^{-\gamma}$ where N is the number of distinct locations in S and $\rho = 0.6$, $\gamma = 0.21$ are constants (Pappalardo et al. 2015b, 2016a; Song et al. 2010a), the individual chooses to explore a new location (Algorithm 3, line 5), otherwise she returns to a previously visited location (Algorithm 3, line 10). If the individual explores and is in location i, the new location $j \neq i$ is selected according to the probability $p_{ij} \in P$ (Algorithm 3, function `PreferentialExploration`). If the individual returns to a previously visited location, it is chosen with probability proportional to the number of her previous visits to that location (Algorithm 3, function `preferentialReturn`). The d-EPR model hence returns the chosen location j.

It is worth highlighting the difference between typical locations and preferred locations. Typical locations indicate places where individuals repeatedly return as part of their mobility routine. Examples of typical locations are home and work locations, where individuals regularly return in their everyday routine. Besides typical

locations, individuals can also return to preferred locations, i.e., places which are not part of a schematic routine but where people return occasionally, such as cinemas or restaurants. The preferential return mechanism of d-EPR models the existence of such preferred locations, allowing the agents to return to previously visited locations with a probability depending of the past visitation frequency.

6 Results

In this section we show the results of simulation experiments where we instantiate DITRAS by using d-EPR as trajectory generator and $\text{MD}^{(t)}$ as diary generator. We construct $\text{MD}^{(t)}$ from nation-wide mobile phone data covering a period of three month using MDL. We refer to the spatio-temporal model as $d\text{-EPR}_{\text{MD}}^{(\text{CDR})}$ and use it to generate sampled mobility trajectories of 10,000 agents. We compare the resulting sampled mobility trajectories with:

- the trajectories of 10,000 mobile phone users whose mobility is tracked during 3 months in a European country;
- the sampled mobility trajectories produced by other 8 spatio-temporal mobility models created through the DITRAS framework by combining different diary and trajectory generators, whose parameters are fitted on the mobile phone data.

Similarly we instantiate DITRAS by using d-EPR and $\text{MD}^{(t)}$ constructed on GPS vehicular tracks covering a period of one month. We refer to the spatio-temporal model as $d\text{-EPR}_{\text{MD}}^{(\text{GPS})}$. We use this model to generate sample mobility trajectories of 10,000 agents and compare the resulting sample mobility trajectories with:

- the trajectories of 10,000 private vehicles whose mobility is tracked through on-board GPS devices during 4 weeks in Tuscany;
- the sampled mobility trajectories produced by other 8 spatio-temporal mobility models created through the DITRAS framework by combining different diary and trajectory generators, whose parameters are fitted on the GPS vehicular data.

In Sect. 6.1 and in Sect. 6.2 we describe respectively the mobile phone data and the GPS vehicular data we use in our experiments to describe the mobility of real individuals and the pre-processing operations we carry out on the data. In Sect. 6.3 we provide a comparison on a set of spatio-temporal mobility patterns of $d\text{-EPR}_{\text{MD}}^{(\text{CDR})}$'s trajectories, mobile phone data's trajectories, and the trajectories produced by the other models. These simulations are performed by using a weighted spatial tessellation induced by the mobile phone towers. Analogously, we provide a comparison on a set of spatio-temporal mobility patterns of $d\text{-EPR}_{\text{MD}}^{(\text{GPS})}$'s trajectories, GPS data's trajectories, and the trajectories produced by the other models. These simulations are performed by using a weighted spatial tessellation induced by the census cells in Tuscany. All the simulations are performed using a time slot duration $t = 3600\text{s} = 1\text{h}$.

6.1 CDR data

We have access to a set of Call Detail Records (CDRs) gathered by a European carrier for billing and operational purposes. The dataset records all the calls made during 11 weeks by ≈ 1 million anonymized mobile phone users. CDRs collect geographical, temporal and interaction information on mobile phone use and show an enormous potential to empirically investigate the structure and dynamics of human mobility on a society wide scale (Reades et al. 2007; Hidalgo and Rodriguez-Sickert 2008; González et al. 2008; Jiang et al. 2012; Calabrese et al. 2011; Pappalardo et al. 2015b, a). Each time an individual makes a call the mobile phone operator registers the connection between the caller and the callee, the duration of the call and the coordinates of the phone tower communicating with the phone, allowing to reconstruct the user's approximate position. Table 2 illustrates an example of the structure of CDRs.

CDRs have been extensively used in literature to study different aspects of human mobility, due to several advantages: they provide a means of sampling user locations at large population scales; they can be retrieved for different countries and geographic scales given their worldwide diffusion; they provide an objective concept of location, i.e., the phone tower. Nevertheless, CDR data suffer different types of bias (Ranjan et al. 2012; Iovan et al. 2013), such as: (i) the position of an individual is known at the granularity level of phone towers; (ii) the position of an individual is known only when she makes a phone call; (iii) phone calls are sparse in time, i.e., the time between consecutive calls follows a heavy tail distribution (González et al. 2008; Barabási 2005). In other words, since individuals are inactive most of their time, CDRs allow to reconstruct only a subset of an individual's mobility. Several works in literature study the bias in CDRs by comparing the mobility patterns observed on CDRs to the same patterns observed on GPS data (Pappalardo et al. 2013b, 2015b, 2013a, c) or handover data (data capturing the location of mobile phone users recorded every hour

Table 2 Example of call detail records (CDRs)

Timestamp	Tower	Caller	Callee
(a)			
2007/09/10 23:34	36	4F80460	4F80331
2007/10/10 01:12	36	2B01359	9H80125
2007/10/10 01:43	38	2B19935	6W1199
⋮	⋮	⋮	⋮

Tower	Latitude	Longitude
(b)		
36	49.54	3.64
37	48.28	1.258
38	48.22	− 1.52
⋮	⋮	⋮

Every time a user makes a call, a record is created with timestamp, the phone tower serving the call, the caller identifier and the callee identifier (a). For each tower, the latitude and longitude coordinates are available to map the tower on the territory (b)

or so) (González et al. 2008). The studies agree that the bias in CDRs does not affect significantly the study of human mobility patterns.

Data preprocessing In order to cope with sparsity in time of CDRs and focus on individuals with reliable call statistics, we carry out some preprocessing steps. Firstly, for each individual u we discard all the locations with a visitation frequency $f = n_i/N \leq 0.005$, where n_i is the number of calls performed by u in location i and N the total number of calls performed by u during the period of observation (Schneider et al. 2013; Pappalardo et al. 2015b). This condition checks whether the location is relevant with respect to the specific call volume of the individual. Since it is meaningless to analyze the mobility of individuals who do not move, all the individuals with only one location after the previous filter are discarded. We select only active individuals with a call frequency threshold of $f = N/(h * d) \geq 0.5$ calls per hour, where N is the total number of calls made by u, $h = 24$ is the hours in a day and $d = 77$ the days in our period of observation. Starting from ≈ 1 millions users, the filtering results in 50, 000 active mobile phone users.

Weighted spatial tessellation The weighted spatial tessellation L we use in the experiments is defined by the mobile phone towers in the CDR data. The relevance of a phone tower is estimated as the total number of calls served by that tower by the 50,000 active mobile phone users during the 3 months. Every location's position on the space is identified by the latitude and longitude coordinates of a phone tower.

6.2 GPS data

The GPS dataset stores information of approximately 9.8 Million different trips from 159,000 private vehicles tracked during one month (May 2011) which passed through Tuscany (central Italy). The GPS traces are provided by Octo Telematics Italia Srl,[2] a company that provides a data collection service for insurance companies. The GPS device is embedded in the private vehicles' engine and automatically turns on when the vehicle starts. The sequence of GPS points that the device transmits every 30 seconds to the server via a GPRS connection forms the global trajectory of a vehicle. When the vehicle stops no points are logged nor sent.

We exploit these stops to split the global trajectory into several sub-trajectories, corresponding to the trips performed by the vehicle. Clearly, the vehicle may have stops of different duration, corresponding to different activities. To ignore small stops like gas stations, traffic lights, bring and get activities and so on, we choose a stop duration threshold of at least 20 minutes: if the time interval between two consecutive observations of the vehicle is larger than 20 minutes, the first observation is considered as the end of a trip and the second observation is considered as the start of another trip. We also performed the extraction of the trips by using different stop duration thresholds (5, 10, 15, 20, 30, 40 minutes), without finding significant differences in the sample of short trips and in the statistical analysis we present in the paper. Since GPS data do not provide explicit information about visited locations, we assign each origin and destination point of the obtained sub-trajectories to the corresponding census cell,

[2] http://www.octotelematics.com/.

according to the information provided by the Italian National Institute of Statistics (ISTAT).[3] We hence obtain a data format similar to CDR data, where we describe the movements of a vehicle by the time-ordered list of census cells where the vehicle stopped. We filter the data by discarding all the vehicles with only one visited location or with less than one trip per day on average during the period of observation. This filtering results in a dataset of 46,121 vehicles.

Weighted spatial tessellation The weighted spatial tessellation L we use in the experiments is defined by the census cells in Tuscany. The relevance of a location is estimated as the total number of stops in the corresponding cell by the 159,000 private vehicles during the month of observation. Every location's position on the space is identified by the latitude and longitude coordinates of the census cell.

6.3 Models comparison and validation

We use the DITRAS framework to build 18 models (9 models fitted on CDRs and 9 models fitted on GPS data) which use different combinations for the diary generator and the trajectory generator. In particular, we consider three diary generators – MD, RD and WT – and three trajectory generators – d-EPR, SWIM and LATP. For every model we simulate the mobility of 10,000 agents for a period of $N = 1,848$ hours (3 months) and $N = 744$ hours (1 month) for models fitted on CDRs and GPS data respectively. Tables 3 and 4 show the ability of every model in reproducing a set of characteristic statistical distributions derived from the CDR and the GPS data respectively, quantified by two measures: *(i)* the Root Mean Square Error, $\text{RMSE}(\mathbf{y}, \hat{\mathbf{y}}) = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$ where $\hat{y}_i \in \hat{\mathbf{y}}$ indicates a point of the synthetic distribution $\hat{\mathbf{y}}$, $y_i \in \mathbf{y}$ the corresponding point in the empirical distribution $\mathbf{y}$ and n the number of observations; *(ii)* the Kullback-Leibler divergence, $\text{KL}(\mathbf{y}||\hat{\mathbf{y}}) = H(\mathbf{y}, \hat{\mathbf{y}}) - H(\mathbf{y})$, where $H(\mathbf{y}, \hat{\mathbf{y}})$ is the cross entropy between the real distribution and the empirical distribution and $H(\mathbf{y})$ is the entropy of the real distribution. Here we use the notation TG_{DG} to specify that trajectory generator TG is used in combination with diary generator DG. For example, $d\text{-EPR}_{\text{MD}}$ indicates the model using diary generator MD in combination with trajectory generator d-EPR. Notation $\text{TG}_{\{\text{DG}_1,\ldots,\text{DG}_k\}}$ indicates the set of models $\{\text{TG}_{\text{DG}_1}, \ldots, \text{TG}_{\text{DG}_k}\}$. Similarly, notation $\{\text{TG}^1, \ldots, \text{TG}^k\}_{\text{DG}}$ indicates the set of models $\{\text{TG}^1_{\text{DG}}, \ldots, \text{TG}^k_{\text{DG}}\}$.

Diary generators In the Random Diary (RD) generator a synthetic individual is in perpetuum motion: in every time slot of the simulation she chooses a new location to visit. We use RD to highlight the difference between the diary generator we propose, MD (Sect. 4.1), and the temporal patterns of a non-realistic diary generator.

In the Waiting Time (WT) diary generator a synthetic individual chooses a waiting time Δt between a trip and the next one from the empirical distribution $P(\Delta t) \sim \Delta t^{-1-\beta} \exp^{-\Delta t/\tau}$, with $\beta = 0.8$ and $\tau = 17$ hours as measured on CDR data (Song et al. 2010a). WT is the temporal mechanism usually used in combination with mobility models like EPR (Song et al. 2010a) and SWIM (Kosta et al. 2010). It reproduces in a realistic way the distribution of the time between two consecutive trips (Song et al.

[3] www.istat.it.

Table 3 Error of fit between CDR data and synthetic data

CDR	Δr	r_g	S^{unc}	T	D	Δt	V	N	$f(L)$
MD									
***d*-EPR**	*.0001*	.0026	*.9643*	*.0061*	*.0659*	*.0014*	$2.6E^{-5}$	*.0218*	*.0122*
	.0006	.0247	*29.34*	*.0101*	*.0682*	*.1915*	.0016	*.5449*	*.1200*
SWIM	.0005	–	3.6069	.0062	.0683	.0029	$5.6E^{-5}$	–	.0669
	.0067		60.97	.0101	.0808	.4996	.0451		1.2892
LATP	.0001	.0061	3.2236	.0062	.0684	.0027	$6.3E^{-5}$	–	.0625
	.0008	.3223	258.46	.0101	.0802	.3282	.0600		.9353
RD									
d-EPR	.0004	.0027	1.1745	.0232	.2098	.0024	$4.1E^{-5}$	.0235	.0521
	.0029	.0161	20.8015	.197	4.3558	.2048	.0191	1.1773	.3876
SWIM	.0041	–	–	.0232	–	.0033	$7.2E^{-5}$	–	.0947
	.1501			.1974		.3773	.0460		4.4057
LATP	.0002	–	–	.0232	–	.0033	$4.6E^{-5}$	–	.0874
	.0014			.1974		.6967	.0321		2.2051
WT									
d-EPR	.0003	*.0024*	1.1666	.0232	.1790	.0023	$4.0E^{-5}$	.0224	.0502
	.0019	*.0130*	20.00	.1970	3.9769	.1946	.0189	1.0395	.3537
SWIM	.0033	–	–	.0232	.2036	.0033	$1.9E^{-5}$	–	.0943
	.0601			.1975	4.3806	.1146	*.0070*		3.9605
LATP	.0001	–	–	.0232	.2037	.0033	$7.2E^{-5}$	–	.0866
	.0010			.1975	4.5672	.6322	.0309		2.1015
Best model	*d*-EPR	*d*-EPR	*d*-EPR	*d*-EPR	*d*-EPR	*d*-EPR	SWIM	*d*-EPR	*d*-EPR
	MD	WT	MD	MD	MD	MD	WT	MD	MD

Every row i is a model and every column j a mobility measure. A cell (i, j) indicates the RMSE (first row) and the KL divergence (second row) of a synthetic distribution w.r.t. the real distribution. The best RMSE values are in italic. Symbol—indicates that the synthetic distribution is not comparable with the real distribution. We highlight in bold the combination of temporal and spatial model leading to the highest number of Italic cells

2010a; Pappalardo et al. 2013b) but does not model the circadian rhythm and the tendency of individuals to be in certain places and specific times.

We construct two diary generators, $\text{MD}_{\text{(CDR)}}$ and $\text{MD}_{\text{(GPS)}}$, by applying algorithm MDL (Sect. 4.1) on CDR data and GPS data respectively. These diary generators are based on Markov models and can reproduce the circadian rhythm of individuals and their tendency to follow or break the routine.

Trajectory generators The trajectory generator SWIM (Kosta et al. 2010) is a modelling approach based on location preference. The model initially assigns to each synthetic individual a home location L_h chosen randomly from the spatial tessellation. The synthetic individual then selects a destination for the next movements depending on the weight of each location (Kosta et al. 2010):

Table 4 Error of fit between GPS data and synthetic data

GPS	Δr	r_g	S^{unc}	T	D	Δt	V	N	$f(L)$
MD									
***d*-EPR**	.0254	*.0148*	*1.9855*	*.0053*	.1334	.0738	.0123	*.0113*	*.0323*
	.5346	*.2850*	*156.92*	*.0156*	.2992	.7567	.1415	*.0411*	*.2429*
SWIM	.0229	–	3.8403	.0054	.1232	.0589	.0123	.0319	.0358
	.8970		210.87	.0156	.2634	.7321	.1522	1.6923	.4914
LATP	.0258	.0225	3.7636	.0054	.1233	.0655	.0178	.0315	.0324
	.5968	.9508	151.35	.0157	.2636	.7148	.4639	1.9085	.3811
RD									
d-EPR	*.0031*	.0237	–	.0231	.0923	.0349	*.0042*	.0271	.0560
	.0420	.9939		.1906	1.2493	.4221	*.0360*	3.3216	.5258
SWIM	.0274	–	–	.0231	–	.2647	.0102	–	.0915
	1.6628			.1912		1.4443	.0919		3.6641
LATP	.0169	–	–	.0231	–	.1599	.0168	–	.0899
	.1381			.1912		1.1524	.3609		2.9663
WT									
d-EPR	.0069	.0223	–	.0231	.0923	*.0291*	.0045	.0270	.0530
	.0518	.8217		.1906	1.0593	*.4369*	.0394	2.132	.4623
SWIM	.0180	–	–	.0231	*.0923*	.1608	.0095	–	.0908
	.7278			.1912	*.9510*	1.0941	.0823		3.2346
LATP	.0190	–	–	.0231	.0923	.1027	.0166	–	.0890
	.1840			.1913	1.0398	.9187	.4282		2.6838
Best model	*d*-EPR RD	*d*-EPR MD	*d*-EPR MD	*d*-EPR MD	SWIM WT	*d*-EPR WT	SWIM WT	*d*-EPR MD	*d*-EPR MD

Every row i is a model and every column j a mobility measure. A cell (i, j) indicates the RMSE (first row) and the KL divergence (second row) of a synthetic distribution w.r.t. the real distribution. The best RMSE values are in italic. Symbol—indicates that the synthetic distribution is not comparable with the real distribution. We highlight in bold the combination of temporal and spatial model leading to the highest number of italic cells

$$w(L)_{\text{swim}} = \alpha * d(L_h, L) + (1 - \alpha) * r(L), \quad \alpha = 0.75 \tag{1}$$

which grows with the relevance $r(L)$ of the location and decreases with the distance from the home (Kosta et al. 2010):

$$d(L_h, L) = \frac{1}{(1 + distance(L_h, L))^2}.$$

SWIM tries to model both the preference for short trips and the preference for relevant locations, though it does not model the preferential return mechanism.

The trajectory generator LATP (Least Action Trip Planning) (Lee et al. 2012, 2009) is a trip planning algorithm used as exploration mechanism in several mobility models, such as SLAW (Lee et al. 2012, 2009), SMOOTH (Munjal et al. 2011), MSLAW (Schwamborn and Aschenbruck 2013) and TP (Solmaz et al. 2015, 2012). In LATP

a synthetic individual selects the next location to visit according to a weight function (Lee et al. 2012, 2009):

$$w(L)_{\text{latp}} = \frac{1}{distance(c, L)^{1.5}}. \tag{2}$$

LATP only models the preference for short distances and does not consider the relevance of a location nor model the preferential return mechanism.

We compare the synthetic mobility trajectories of the nine models with CDR trajectories and GPS trajectories on the distributions of several measures capturing salient characteristics of human mobility. Tables 3 and 4 display the mobility measures we consider, which are: trip distance Δr (González et al. 2008; Pappalardo et al. 2013b), radius of gyration r_g (González et al. 2008; Pappalardo et al. 2013b, 2015b), mobility entropy S^{unc} (Song et al. 2010b; Eagle and Pentland 2009; Pappalardo et al. 2016b), location frequency $f(L)$ (Song et al. 2010a; Hasan et al. 2013; Pappalardo et al. 2013b), visits per location V (Pappalardo et al. 2016a), locations per user N (Pappalardo et al. 2016a), trips per hour T (González et al. 2008; Pappalardo et al. 2013b), time of stays Δt (Song et al. 2010a; Hasan et al. 2013) and trips per day D.

Trip distance The distance of a trip Δr is the geographical distance between the trip's origin and destination locations. We compute the trip distances for every individual and then plot the distribution $P(\Delta r)$ of trip distances in Fig. 2a–c (CDR data) and Fig. 3a–c (GPS data). Figure 2a compares the distribution of trip distance of CDR data with the distributions produced by d-EPR$^{(\text{CDR})}_{\text{MD}}$, SWIM$^{(\text{CDR})}_{\text{MD}}$ and LATP$^{(\text{CDR})}_{\text{MD}}$. We observe that d-EPR$^{(\text{CDR})}_{\text{MD}}$ and LATP$^{(\text{CDR})}_{\text{MD}}$ are able to reproduce the distribution of $P(\Delta r)$ although slightly overestimating long-distance trips. In contrast SWIM$^{(\text{CDR})}_{\text{MD}}$ cannot reproduce the shape of the empirical distribution resulting in a RMSE(SWIM$^{(\text{CDR})}_{\text{MD}}$) and KL(SWIM$^{(\text{CDR})}_{\text{MD}}$) higher than the other two models (see Table 3). The shape of the synthetic distributions do not vary significantly by changing the diary generator (Fig. 2b–c). In other words, the choice of the diary generator does not affect the ability of the model to capture the distribution $P(\Delta r)$. This is also evident from Table 3 where the RMSEs and the KLs in the first column vary a little by changing the diary generator. Model d-EPR$^{(\text{CDR})}_{\text{MD}}$ produces the best fit with CDR data, as we notef in Fig. 2c and Table 3. This suggests that modelling preferential return and location preference is crucial to reproduce $P(\Delta r)$ as well as the preference for short-distance trips. Although SWIM embeds a preference for short-distance trips (Eq. 1) the distance is chosen with respect to the home location L_h leading to an underestimation of short-distance trips (Fig. 2a–c). Figure 3a–c compares the distribution of trip distance of GPS data with the distributions produced by the generative algorithms. Results on GPS data confirm the observations on CDRs: in contrast with SWIM, d-EPR and LATP are able to reproduce the distribution of $P(\Delta r)$, regardless the diary generator. Also in this case, d-EPR$^{(\text{GPS})}_{\text{RD}}$ is the model generating the most realistic synthetic data (Table 4).

Radius of gyration The radius of gyration r_g is the characteristic distance traveled by an individual during the period of observation (González et al. 2008; Pappalardo et al. 2013b, 2015b). In detail, r_g characterizes the spatial spread of the locations visited by an individual u from the trajectories' center of mass (i.e., the weighted mean point of the locations visited by an individual), defined as:

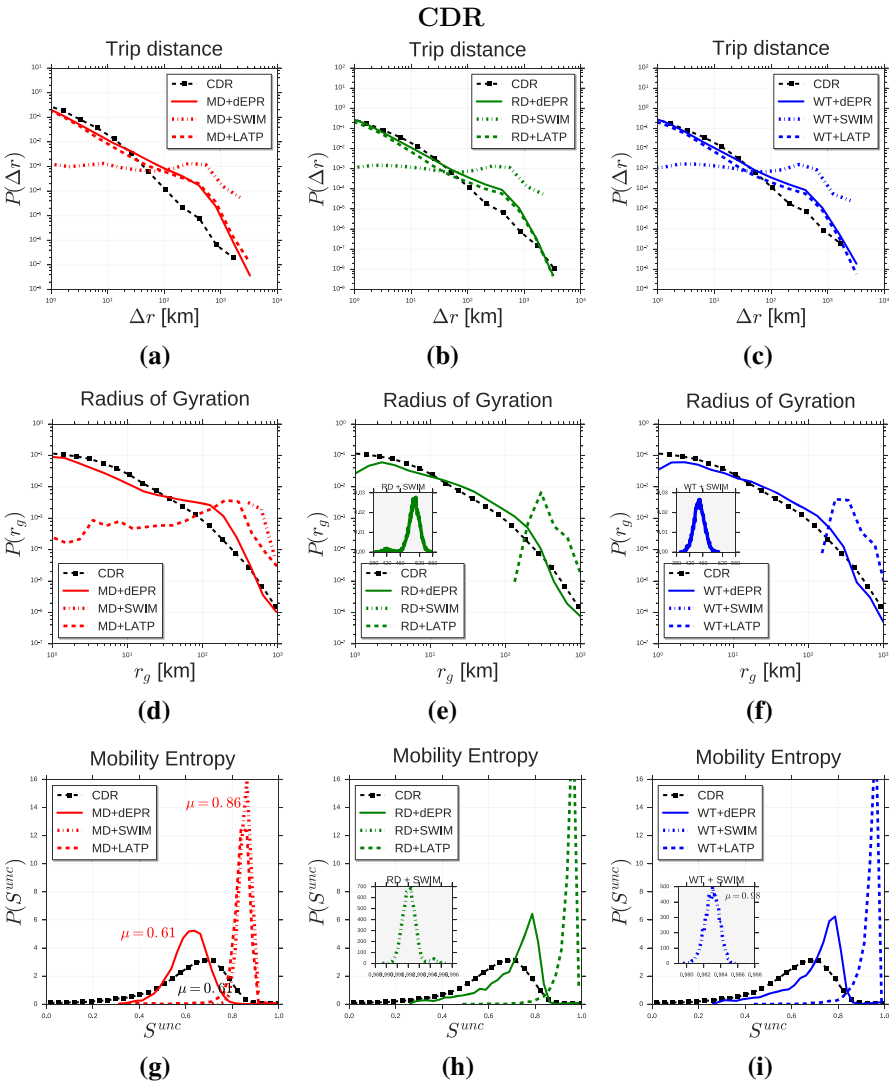

Fig. 2 Distributions of human mobility patterns (CDR). The figure compares the models and CDR data on trip distance, radius of gyration and mobility entropy. Plots in (**a**–**c**) show the distribution of trip distances $P(\Delta r)$ for real data (black squares) and data produced by three trajectory generators (d-EPR, SWIM and LATP) in combination with the MD generator (**a**), the RD generator (**b**) and the WT generator (**c**). Plots in (**d**–**f**) show the distribution of radius of gyration r_g, while plots in (**g**–**i**) show the distribution of mobility entropy S^{unc}

$$r_g = \sqrt{\sum_{i \in L^{(u)}} p_i (l_i - l_{cm})^2}, \tag{3}$$

where l_i and l_{cm} are the vectors of coordinates of location i and center of mass, respectively (González et al. 2008; Pappalardo et al. 2013b, 2015b), $L^{(u)} \subseteq L$ is

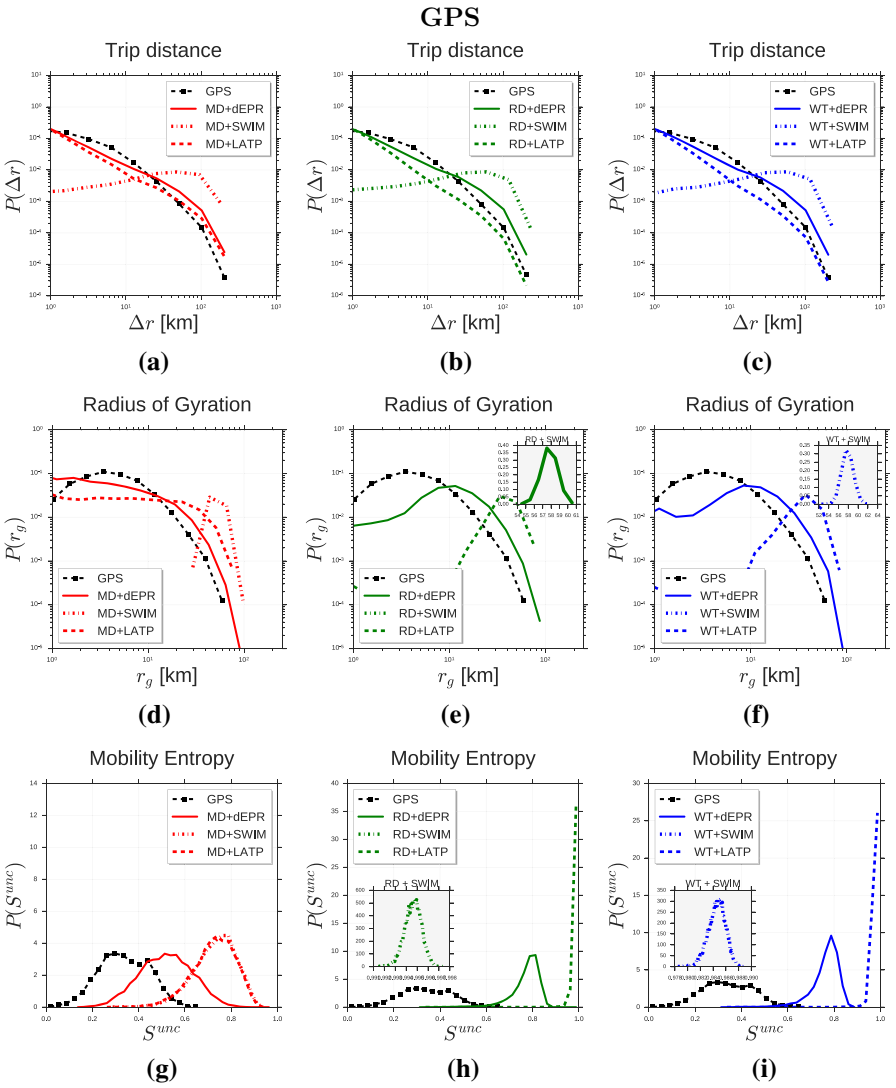

Fig. 3 Distributions of human mobility patterns (GPS). The figure compares the models and GPS data on trip distance, radius of gyration and mobility entropy

the set of locations visited by individual u, $p_i = n_i/|L^{(u)}|$ is the individual's visitation frequency of location l_i, equal to the number of visits to l_i divided by the total number of visits to all locations. In Fig. 2a we observe that d-EPR$_{\text{MD}}^{(\text{CDR})}$ is the only model capable of reproducing the shape of $P(r_g)$ of CDR data, though overestimating the presence of large radii (see Fig. 2d). RMSE(d-EPR$_{\text{MD}}^{(\text{CDR})}$) for r_g is indeed lower than RMSE(SWIM$_{\text{MD}}^{(\text{CDR})}$) and RMSE(LATP$_{\text{MD}}^{(\text{CDR})}$) as shown in Table 3. SWIM$_{\text{MD}}^{(\text{CDR})}$ and LATP$_{\text{MD}}^{(\text{CDR})}$ cannot reproduce the shape of $P(r_g)$ because r_g also depends on the preferential return mechanism (Song et al. 2010a; Pappalardo et al. 2015b) which is not

modeled in SWIM and LATP. In a previous work (Pappalardo et al. 2016a) we also show that $P(r_g)$ depends on the preferential exploration mechanism of d-EPR since a version of d-EPR without preferential exploration – the s-EPR model – is not able to reproduce the shape of $P(r_g)$. We also observe that while $d\text{-EPR}^{(\text{CDR})}_{\{\text{MD,RD,WT}\}}$ produce similar distributions of r_g, SWIM and LATP produce different distributions of r_g with different choices of the diary generator (Fig. 2e, f). The shape of $P(r_g)$ for GPS data is slightly different from the same distribution of CDR data, since short radii are less likely in GPS due to the nature of car travels (Pappalardo et al. 2013c, b, a). Also for GPS we observe that, in contrast with LATP and SWIM, d-EPR is the only model that can reproduce the shape of $P(r_g)$. In particular $d\text{-EPR}^{(\text{GPS})}_{\text{MD}}$ produces the best fitting with GPS data in terms of both RMSE and KL (Table 4).

Mobility entropy The mobility entropy S^{unc} of an individual u is defined as the Shannon entropy of her visited locations (Song et al. 2010b; Eagle and Pentland 2009; Pappalardo et al. 2016b):

$$S^{unc}(u) = \frac{\sum_{i \in L^{(u)}} p_i \log(p_i)}{\log |L^{(u)}|}, \tag{4}$$

where p_i is the probability that individual u visits location i during the period of observation and $\log |L^{(u)}|$ is a normalization factor. The mobility entropy of an individual quantifies the possibility to predict individual's future whereabouts. Individuals having a very regular movement pattern possess a mobility entropy close to zero and their whereabouts are rather predictable. Conversely, individuals with a high mobility entropy are less predictable.

We observe that the average $\overline{S}^{unc}$ produced by $d\text{-EPR}^{(\text{CDR})}_{\text{MD}}$ data equals the average $\overline{S}^{unc}$=0.61 in CDR data, although $d\text{-EPR}^{(\text{CDR})}_{\text{MD}}$ underestimates the variance of distribution $P(S^{unc})$ (Fig. 2g). In contrast, $\text{SWIM}^{(\text{CDR})}_{\text{MD}}$ and $\text{LATP}^{(\text{CDR})}_{\text{MD}}$ largely overestimate $\overline{S}^{unc}$ and underestimate the variance of $P(S^{unc})$, resulting in RMSE and KL much higher than $\text{RMSE}(d\text{-EPR}^{(\text{CDR})}_{\text{MD}})$ and $\text{KL}(d\text{-EPR}^{(\text{CDR})}_{\text{MD}})$, as shown in Table 3. This is because SWIM and LATP do not model the preferential return mechanism, which increases the predictability of individuals since they tend to come back to already visited locations. $P(S^{unc})$ is not robust to the choice of diary generator: diary generator RD and WT make the models to largely overestimate $\overline{S}^{unc}$ (Fig. 2h, i). In particular $\text{SWIM}^{(\text{CDR})}_{\{\text{RD,WT}\}}$ and $\text{LATP}^{(\text{CDR})}_{\{\text{RD,WT}\}}$ produce distributions with $\bar{S}^{unc} \approx 1$, indicating that the typical synthetic individual is much more unpredictable than a typical individual in CDR data. This makes those distributions not comparable with the distribution of MD models. Hence, distribution $P(S^{unc})$ highly depends on both the choice of the trajectory generator and the choice of the diary generator. We observe similar results for GPS data, where only $\{d\text{-EPR, SWIM, LATP}\}^{(\text{GPS})}_{\text{MD}}$ can reproduce $P(S^{unc})$ in reasonable agreement with real data. All the other models produce distributions that are not comparable with the entropies of private vehicles (Fig. 3g–i).

Location frequency Another important characteristic of an individual's mobility is the probability of visiting a location given the location's rank. The rank of a location

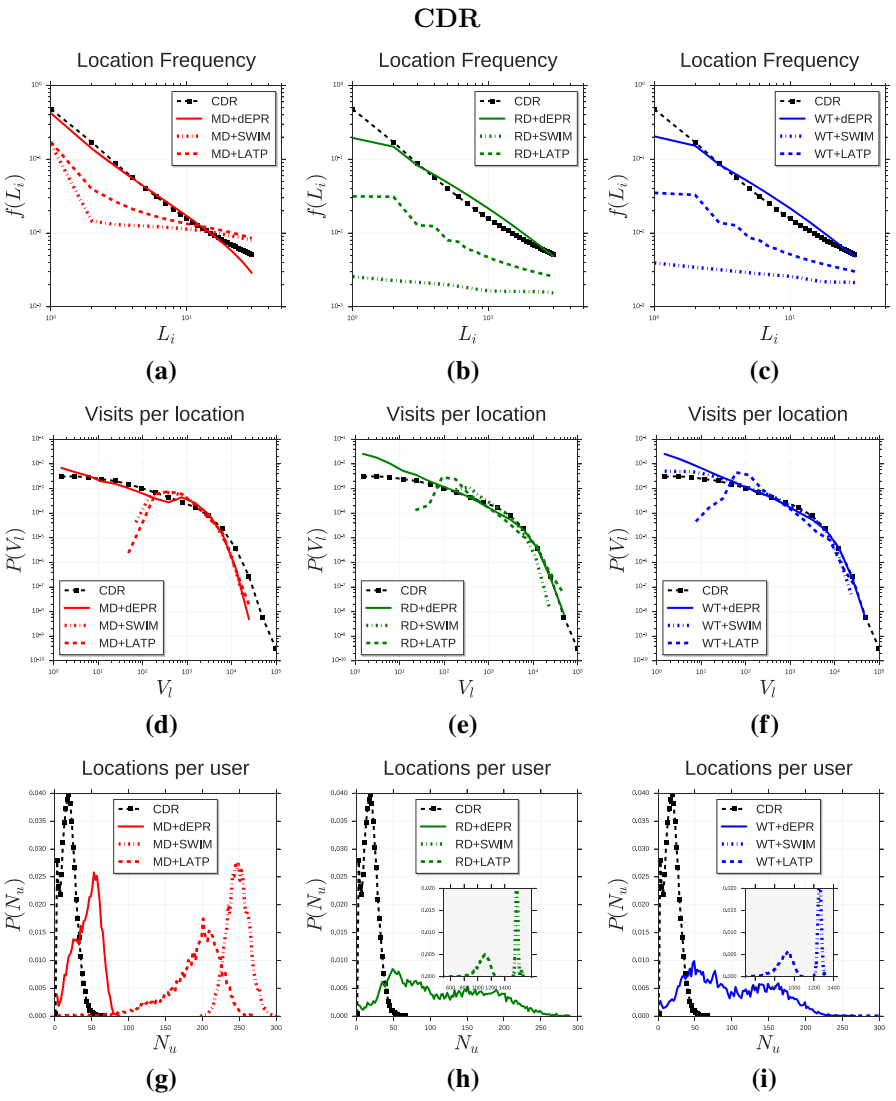

Fig. 4 Distributions of human mobility patterns (CDR). The figure compares the models and CDR data on location frequency, visits per location and locations per users. Plots in (**a**–**c**) show the distribution of location frequency $f(L)$ for d-EPR, SWIM and LATP used in combination with MD, RD and WT respectively. Plots in (**d**–**f**) show the distribution of the number V of visits per location and plots in (**g**–**i**) show the distribution of the number N of distinct visited locations per user

depends on the number of times the individual visits the locations over the period of observation. For instance, rank 1 represents the most visited location (generally home place); rank 2 the second most visited location (e.g., work place) and so on. We compute the frequency of each of these ranked locations for every individual and plot the distribution of frequencies $f(L_i)$ in Figs. 4a–c (CDR) and 5a–c (GPS). For CDR

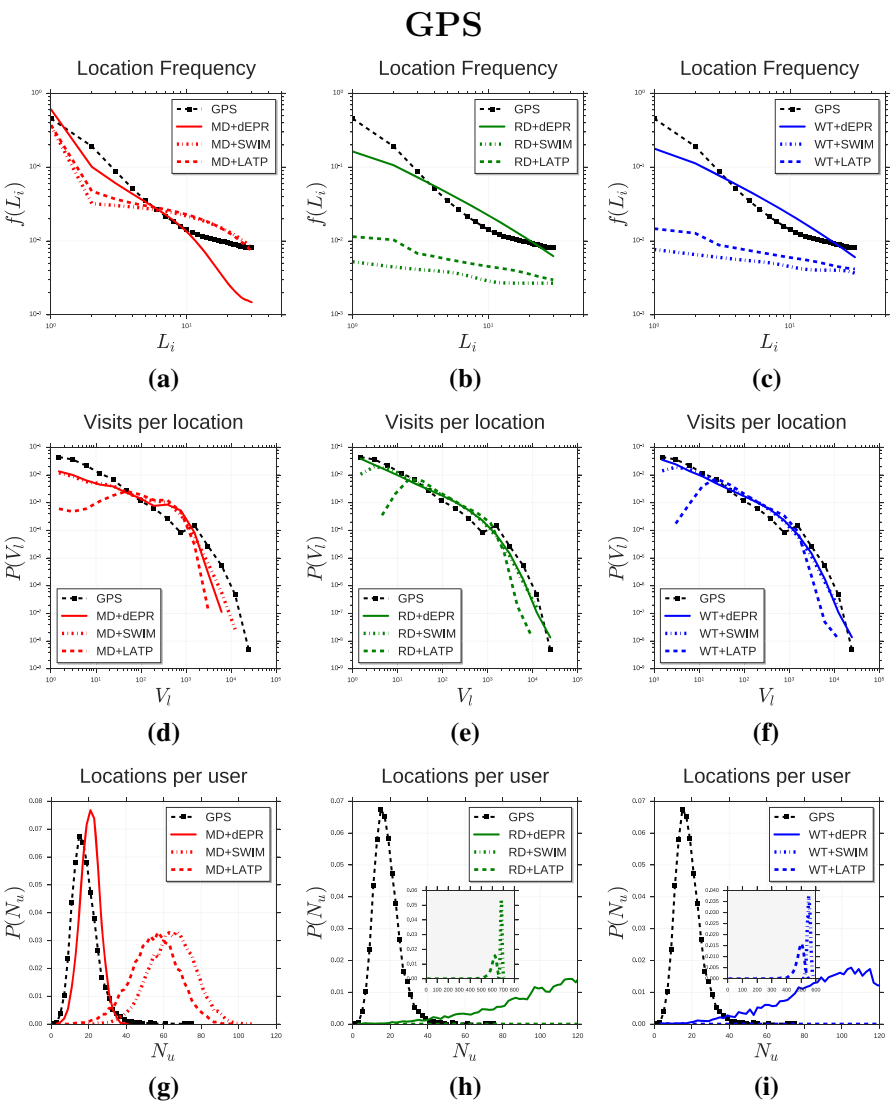

Fig. 5 Distributions of human mobility patterns (GPS). The figure compares the models and GPS data on location frequency, visits per location and locations per users

data, we observe that d-EPR$^{(CDR)}_{MD}$ reproduces the shape of $f(L_i)$ (with RMSE=0.0122 and KL = 0.12) better than SWIM$^{(CDR)}_{MD}$ and LATP$^{(CDR)}_{MD}$ (which have RMSE = 0.0669, KL = 1.2892 and RMSE = 0.0626, KL = 0.9353 respectively). If we change the diary generator in the model, d-EPR$^{(CDR)}_{\{RD,WT\}}$ underestimate the frequency of the top-ranked location and slightly overestimate the frequency of the less visited locations with respect to CDR data (Fig. 4b, c). A reason for this discrepancy is that RD and WT do not take into account the circadian rhythm of individuals, hence underestimating the

number of returns to the most frequent location (usually the home place). In $\text{SWIM}^{(\text{CDR})}_{\text{MD}}$ and $\text{LATP}^{(\text{CDR})}_{\text{MD}}$, the absence of a preferential return mechanism produce a more uniform distribution of location frequencies (Fig. 4b, c), which is further exacerbated for $\text{SWIM}^{(\text{CDR})}_{\{\text{RD,WT}\}}$ and $\text{LATP}^{(\text{CDR})}_{\{\text{RD,WT}\}}$. Location frequency $f(L_i)$ is another case where the choice of the diary generator and the choice of the trajectory generator are both crucial to reproduce the shape of the distribution in an accurate way. Experiments on GPS data confirm results observed on CDRs (Fig. 5a–c): model $d\text{-EPR}^{(\text{GPS})}_{\text{MD}}$ produces the best fit with real data, while changing either the diary or the trajectory generators produces worse fits.

Visits per location A useful measure to understand how a set of individuals exploit the mobility space is the number V of overall visits per location, i.e., the total number of visits by all the individuals in every location during the period of observation. For every dataset, we compute the number of visits for every location of the weighted spatial tessellation and plot the distribution $P(V)$ in Fig. 6d–f (CDR) and Fig. 7d–f (GPS). As for CDR data, $d\text{-EPR}^{(\text{CDR})}_{\text{MD}}$ produces a $P(V)$ which follows a heavy tail distribution: the majority of locations have just one visit while a minority of locations have up to several thousands visits during the 11 weeks. The value of V of a location depends on two factors: (i) its relevance in the weighted spatial tessellation; (ii) its position in the weighted spatial tessellation. The higher the relevance of a location in the weighted spatial tessellation, the higher is the probability for the location to be visited in the exploration mechanisms of d-EPR and SWIM. Indeed, from Fig. 6e, f we observe that d-EPR and SWIM are the models which better fit $P(V)$. In contrast LATP does not take into account the relevance of a location during the exploration being unable to capture the shape of $P(V)$. Experiments on GPS data substantially confirm these results (Fig. 7d–f): d-EPR and SWIM generates the most realistic distributions of $P(V)$.

Locations per user The number N_u of distinct locations visited by an individual during the period of observation describes the degree of exploration of an individual, i.e., how the single individuals exploit the mobility space. In Fig. 4g we observe that the MD models do not capture the shape of $P(N_u)$ in CDR data: the average number of distinct locations $\overline{N}$ according to $d\text{-EPR}^{(\text{CDR})}_{\text{MD}}$ is about twice $\overline{N}$ in CDR data, while $\text{SWIM}^{(\text{CDR})}_{\text{MD}}$ and $\text{LATP}^{(\text{CDR})}_{\text{MD}}$ produce distributions whose $\overline{N}$ is more than ten times $\overline{N}$ in CDR data. By changing diary generator (Fig. 4h, i) the difference with CDR data becomes even larger: $d\text{-EPR}^{(\text{CDR})}_{\{\text{RD,WT}\}}$ produce a much broader variance of $P(N_u)$, $\text{SWIM}^{(\text{CDR})}_{\{\text{RD,WT}\}}$ and $\text{LATP}^{(\text{CDR})}_{\{\text{RD,WT}\}}$ predict a number of distinct visited locations very far from CDR data. These results suggest that the considered models overestimate the degree of exploration of individuals. In the case of $d\text{-EPR}^{(\text{CDR})}_{\text{MD}}$ the overestimation may depend on the distribution of time of stays, as the distribution of time stays $P(\Delta t)$ produced by $d\text{-EPR}^{(\text{CDR})}_{\text{MD}}$ overestimates the number of short stay times, leading to a larger total number of visited locations (Fig. 6g). For GPS data, model $d\text{-EPR}^{(\text{GPS})}_{\text{MD}}$ produces a $P(N)$ that is more realistic than the other models, as it is evident from Fig. 7g and from Table 4.

Trips per hour Human movements follow the circadian rhythm, i.e., they are prevalently stationary during the night and move preferably at specific times of the day

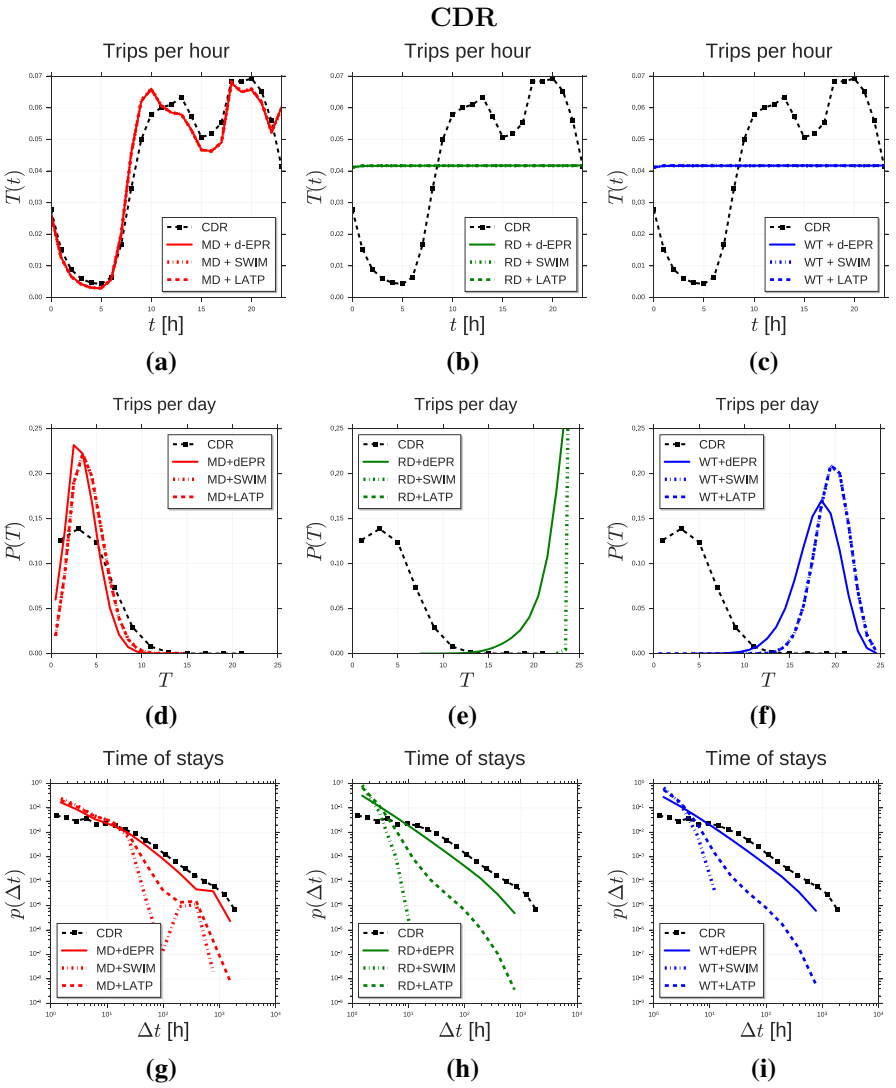

Fig. 6 Distributions of human mobility patterns (CDR). The figure compares the models and CDR data on trips per hour, trips per day and time of stays. Plots in (**a**–**c**) show the distribution of the number T of trips per hour of the day for d-EPR, SWIM and LATP used in combination with MD, RD and WT respectively. Plots in (**d**–**f**) show the distribution of the number D of trips per day, plots in (**g**–**i**) show the distribution of time of stays Δt

(González et al. 2008; Pappalardo et al. 2013b). To verify whether the considered models are able to capture this characteristic of human mobility, we compute the number of trips T made by the individuals at every hour of the period of observation. Figures 6a–c and 7a–c show how T distribute across the 24 hours of the day, for CDRs and GPS data respectively. We observe that, regardless the trajectory generator used,

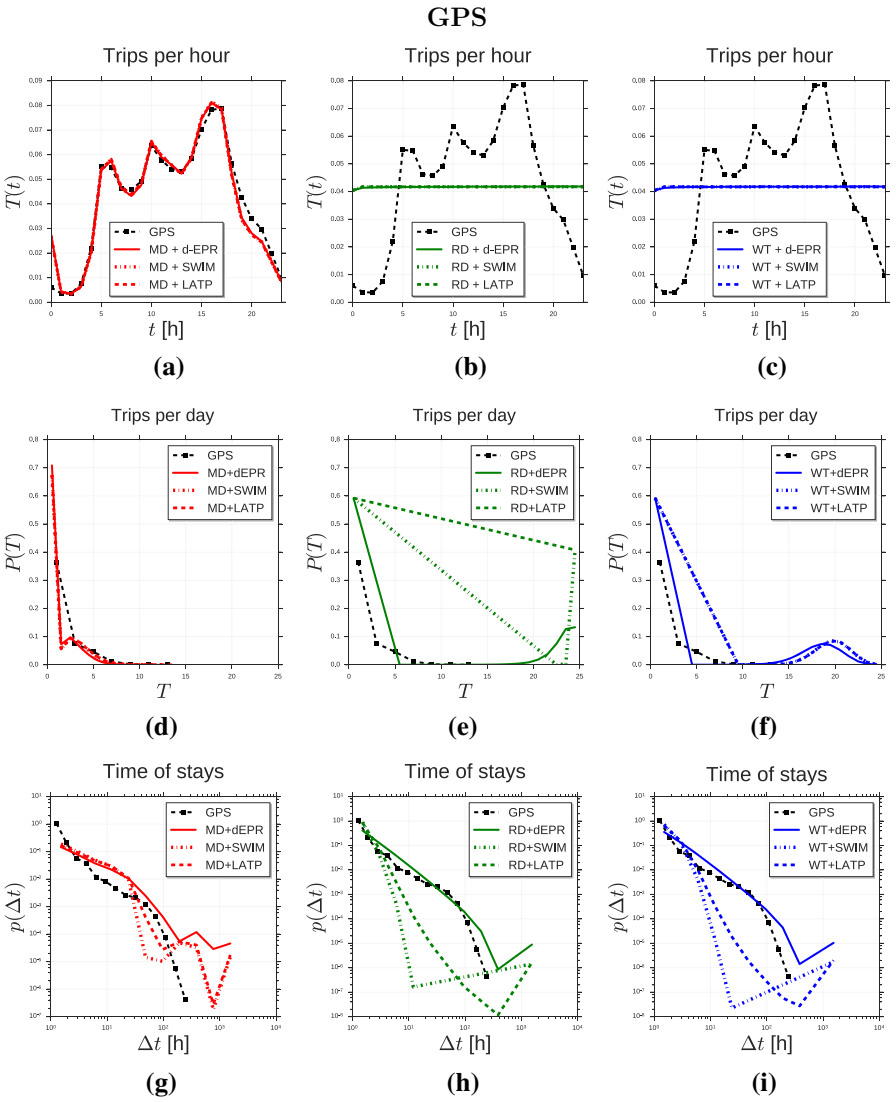

Fig. 7 Distributions of human mobility patterns (GPS). The figure compares the models and GPS data on trips per hour, trips per day and time of stays

diary generator MD produces a distribution of trips per hour very similar to real data (Figs. 6a and 7a). The mobility diary generator MD proposed in Sect. 4 is hence able to create mobility diaries which reproduce the circadian rhythm of individuals in an accurate way. In contrast, diary generators RD and WT are not able to capture this distribution, regardless the trajectory generator used (Figs. 6b, c and 7b, c). This is because: (i) in RD individuals are always in motion; (ii) WT takes into account the

waiting times but not the preference of individuals to move at specific times of the day.

Trips per day The number of trips per day D indicates the tendency of individuals to travel in their every-day life. For every dataset, we compute the number of trips per day made by each individual during the period of observation and plot the distribution $P(D)$ in Fig. 6d–f (CDR) and Fig. 7d–f (GPS). We observe that $d\text{-EPR}_{\text{MD}}^{(\text{CDR, GPS})}$, $\text{SWIM}_{\text{MD}}^{(\text{CDR, GPS})}$ and $\text{LATP}_{\text{MD}}^{(\text{CDR, GPS})}$ are able to capture the shape of $P(D)$ but overestimate the variance of the distribution (Fig. 6d). The other diary generators, RD and WT, are not able to reproduce the CDR distribution since the average number $\overline{D}$ of trips per day is much higher than CDR data (Fig. 6e, f). Again, this is because in RD individuals are always in motion and because WT does not take into account the circadian rhythm of individuals.

Time of stays The distribution of stay times Δt is another important temporal features observed in human mobility. Stay time is the amount of time an individual spends at a particular location. In our experiments we compute the stay time as the number of hours every individual spends in her visited locations and plot the distribution $P(\Delta t)$ in Fig. 4g–i (CDR) and Fig. 5g–i (GPS). We observe that, for both CDRs and GPS data, $d\text{-EPR}_{\{\text{MD,RD,WT}\}}^{(\text{CDR, GPS})}$ capture the shape of the distribution while the other models do not, though overestimating the presence of short time stays.

6.4 Discussion of results

Two main results emerge from our experiments. First, model $d\text{-EPR}_{\text{MD}}$ produces sampled mobility trajectories having in general the best fit to both CDR data and GPS data (i.e., having the lowest RMSE and KL for most of the measures), as evident in Tables 3 and 4. Diary generator MD, indeed, simulates in a realistic way temporal human mobility patterns such as the distribution of location frequency (Fig. 4a) and the distribution of trips per hour (Figs. 6a, 7a). This is mainly because MD reproduces the circadian rhythm of individuals, while RD and WT do not. Moreover, trajectory generator d-EPR embeds two mobility mechanisms: preferential return and preferential exploration. The preferential return mechanism – absent in SWIM and LATP – allows for a realistic simulation of, for example, the distribution of radius of gyration (Figs. 2d, 3d) and the distribution of stay times (Fig. 6g). The preferential exploration mechanism, which is modeled by both d-EPR and SWIM but it is absent in LATP, allows for a realistic description of the territory exploitation by individuals, in terms of the distribution of the number of visits per location (Figs. 4d, 5d). Also, model $d\text{-EPR}_{\text{MD}}$ produces realistic distributions for both CDR and GPS data, suggesting that it can be used in different simulation scenarios where its parameters are fitted on different types of data and different spatio-temporal resolutions.

Second interesting result is that the temporal and the spatial mechanisms have different roles in shaping the distribution of standard mobility measures. Some measures, such as trip distance (Figs. 2a–c, 3a–c), radius of gyration (Figs. 2d–f, 3d–f), visits per location (Figs. 4d–f, 5d–f) and time of stays (Fig. 2g–i) mainly depend on the

choice of the trajectory generator, i.e., on the spatial mechanism of the model. Indeed, by changing the underlying diary generator the shape of these distribution, the RMSE and the KL divergence w.r.t. real data do not change in a significant way. Other measures, such as trips per hour (Figs. 6a–c, 7a–c) and trips per day (Fig. 6d–f) mainly depend on the choice of the diary generator, i.e., on the temporal mechanism of the model. Conversely, both the spatial and the temporal mechanism are determinant in reproducing the distribution of some other measures like mobility entropy (Figs. 2g–i, 3g–i) and locations per user (Figs. 4g–i, 5g–i). Moreover the right combination of diary and trajectory generator, d-EPR $_{\text{MD}}$, leads to more accurate fits w.r.t. both CDR data and GPS data for the majority of measures (Tables 3, 4). Human mobility patterns depend on both where people go and when people move: our results show that to reproduce them in an accurate way we need proper choices for the spatial and the temporal generative models to use in the DITRAS framework.

7 Conclusion and future work

In this paper we propose DITRAS, a framework for the generation of individual human mobility trajectories with realistic spatio-temporal patterns. The framework consists of two steps: (i) the generation of a mobility diary by using a diary generator; (ii) the generation of a mobility trajectory by using a trajectory generator. In the paper we propose a novel diary generator MD together with MDL, a data-driven algorithm to build it from real mobility data.

We instantiate DITRAS by using MD and the state-of-the-art trajectory generator d-EPR and obtain a novel generative algorithm, d-EPR $_{\text{MD}}$. We use it to generate the spatio-temporal trajectories of thousands of agents visiting the locations on a large European country and a region in Italy. The generated sampled mobility trajectories are compared with CDR data, GPS vehicular data, and the trajectories produced by other generative algorithms, each obtained by using a different combination of diary generator and trajectory generator in the DITRAS framework. Among the considered algorithms, d-EPR $_{\text{MD}}$ produces the best fit with respect to both CDR data and GPS data. We also observe that different combinations of diary and trajectory generators show different abilities to reproduce the distribution of standard mobility measures. This result highlights the importance of considering both the spatial and temporal dimensions in human mobility modelling.

The proposed model d-EPR $_{\text{MD}}$ has a limited number of parameters to fit. The generation of the mobility diary is parameter-free as the Markov chain is a non-parametric model where each element of the transition matrix MD is estimated using the empirical frequencies observed in the data. The generation of the mobility trajectory is based on the d-EPR model. The details on how to fit the d-EPR parameters are explained in detail in (Pappalardo et al. 2015b, 2016a). Here, for the two parameters of the exploration probability p_{new}, we choose the values $\rho = 0.6$ and $\gamma = 0.21$ that have been estimated in previous work (Song et al. 2010a). For the gravity model used in the exploration phase, we use a power law deterrence function of the distance with exponent -2, although other types of gravity or intervening opportunities models can be used. Given that the model is non-parametric or depends on a very small number

of parameters, it does not suffer from training/test issues and its calibration is quite robust to changes in the size of the training set.

Applications Given its flexibility, DITRAS can be used in a wide range of applications. Here we provide three examples where DITRAS and d-EPR $_{\text{MD}}$ can be particularly useful and profitably applied.

In urban science, the generation of what-if scenarios to imagine the new mobility that could emerge from the construction of new infrastructures requires the generation of realistic mobility data and hence the presence of an accurate generative algorithm (Barbosa-Filho et al. 2017; Kopp et al. 2014). d-EPR $_{\text{MD}}$ could be used to generate synthetic data given the tessellation of the territory that emerges from the construction of the new infrastructure, allowing urban planners and managers to quantify changes in urban mobility and visualize preferred path that could emerge from the simulation.

Computational epidemiology has attracted particular attention in the last decade, as the arrival of the 2009 flu pandemic prompted scientists to develop realistic mobility models to simulate the spread of viruses on a territory (Merler et al. 2013; Ajelli et al. 2010; Venkatramanan et al. 2017). The possibility to use DITRAS to combine different temporal and spatial mechanisms is particularly valuable for this type of studies, as generative algorithms for individual human mobility are the basic mechanism used in computational epidemiology to generate synthetic population mimicking at an individual level the realistic aspects related to disease propagation.

Opportunistic Networks (OppNets) enable communications in disconnected environments in the absence of an end-to-end path between the sender and the receiver. In OppNets, the mobility of nodes (e.g., mobile devices such as smartphones and tables) help the delivery of messages by connecting, asynchronously in time, otherwise disconnected subnetworks. This means that the network protocols responsible for finding a route between two disconnected devices must embed patterns in human movements and make prediction of future encounters. Realistic generative algorithms for human mobility are fundamental for testing the efficiency of OppNets protocol, as real data about the functioning of the network is obviously not available during the protocol design (Tomasini et al. 2017). DITRAS can be used to instantiate many generative algorithms and then generate realistic mobility routines to test the efficiency of a given network protocol for OppNets. Given its accuracy in reproducing human mobility patterns, d-EPR $_{\text{MD}}$ can be used to uncover the characteristics of the network protocol in real-life, such as the speed of message delivery.

A possible application of DITRAS and d-EPR $_{\text{MD}}$ in data mining is anomaly detection. The proposed model can be used to detect individuals with an anomalous mobility behavior with respect to the typical mobility patterns of the majority of the individuals. In particular, within our framework an individual is anomalous if her trajectory is not a likely outcome of the model, i.e., if the probability that the model would generate such trajectory is below a given threshold. To this end, the log-likelihood of each individual's trajectory can be computed and the individuals can be ranked according to their log-likelihood values: individuals with a low rank and a very high log-likelihood values would be the most typical, whereas individuals with the highest ranks and low log-likelihood values would be the most anomalous.

Improvements The instantiation of DITRAS we propose, d-EPR $_{\mathrm{MD}}$, can be further improved in several directions. First, in this work the construction of the diary generator $\mathrm{MD}^{(t)}$ through the mobility diary learner MDL is based on the simplest possible typical diary $W^{(t)}$, where the most likely location where a synthetic individual can be found at any time is her home location. More complex typical diaries can be used specifying, for example, the typical times where an individual can be found at work, school, friends' home and so on. Such a composition of $W^{(t)}$ can be constructed by using surveys or generative algorithms describing the daily schedule of human activities (Rinzivillo et al. 2014; Jiang et al. 2012; Liao et al. 2007) as a way to enrich an individual's trajectory with information about the type of activity associated to a location.

Second, in d-EPR the preference for short-distance trips is embedded in the preferential exploration phase only. A preference for short-distance trips can be introduced during the preferential return mechanisms as well, in order to eliminate the overestimation of long-distance trips and long-distance radii observed in Figs. 2a and 2d.

Third, in d-EPR $_{\mathrm{MD}}$ we make the simplifying assumption that the travel time is of negligible duration. This may not be a good assumption especially when the duration of the time slot is one hour or less. The proposed algorithm can be modified to explicitly include realistic information on the travel time between locations, which imposes constraints on the locations that are reachable in a given time window and on the time that can be spent in a location given the travel time needed to reach the next location in the mobility diary. Moreover, another interesting improvement can be to map the sampled mobility trajectories to a road network specifying specific road routes with specific velocities. This mapping would be of great help, for example, in what-if analysis where we want to study how human mobility changes with the construction of a new infrastructure in an urban context.

Finally, there is a large number of studies that demonstrate the connection between human mobility and social networks (Brown et al. 2013b; Hristova et al. 2016; Wang et al. 2011; Volkovich et al. 2012; Brown et al. 2013a; Hossmann et al. 2011a, b), as well as several approaches that include information on social connections in human mobility models (Borrel et al. 2009; Yang et al. 2010; Fischer et al. 2010; Boldrini and Passarella 2010; Musolesi and Mascolo 2007). A mechanism to account for the influence of social connections on human mobility can be introduced in DITRAS as a third phase, between the mobility diary generation and the sampled trajectory construction.

We leave these improvements of DITRAS for future work.

Acknowledgements We thank Paolo Cintia, Gianni Barlacchi and Salvatore Rinzivillo for their invaluable suggestions. This work has been partially funded by the EU under the H2020 Program by project Cimplex Grant n. 641191. Filippo Simini has been supported by EPSRC First Grant EP/P012906/1.

Appendix A: Homogeneity of typical mobility diaries

We investigate to what extent the typical mobility diaries of real individuals are homogeneous by performing a clustering experiment. For every individual in the GPS dataset we compute her typical week, i.e. a time series of length 168 hours. Every time slot is the most frequent location of the individual in that hour of the week. We then apply the DBSCAN clustering algorithm (Ester et al. 1996) to group the typical weeks in dense clusters. We use the Levenshtein metric (Navarro 2001) to measure the similarity between two typical weeks. DBSCAN takes two input parameters: $minPts$ and eps (Ester et al. 1996). We set $minPts = 4$ and $eps = 70$. We estimate the value of these parameters using the procedure suggested in (Tan et al. 2005): (i) we fix $minPts = 4$ and compute for every typical week the distance d to its 4th nearest neighbor; (ii) we sort the typical weeks in increasing order with respect to d and set eps to the distance corresponding to an elbow in the curve of Fig. 8a. We observe no significant differences in the clustering results by varying $minPts$ in the range [2, 5].

DBSCAN produces two clusters, one of them consisting of $\approx 90\%$ of the typical weeks (Fig. 8b). The silhouette coefficient of the clustering (Rousseeuw 1987), a measure of how similar a typical diary is to its own cluster compared to other clusters, is $s = 0.50$ (in general, $s \in [-1, 1]$). The typical weeks in the biggest cluster have typically one or two locations, while the representative typical week (i.e., the medoid of the cluster) consists of just one location, the most frequent location of the individual (Fig. 8c, d). This result supports the validity of the simplifying assumption to consider one typical diary with a single location for all agents.

Appendix B: Computational complexity of d-EPR$_{\mathrm{MD}}$

Learning phase. In the learning phase, two main tasks are performed:

(1) the construction of the MD model by the MDL algorithm (Algorithm 2). The procedure `UpdateMarkovChain` has computational complexity $\mathcal{O}(N)$, where N is the number of slots in the period of observation. As we repeat the procedure for all the n individuals in the dataset, the computational complexity of Algorithm 2 is $\mathcal{O}(Nn)$. When $n \gg N$, (e.g., when the period of observation is short and the dataset contains hundreds of thousands of individuals), the factor N is negligible and the computational complexity of Algorithm 2 can be approximated to $\mathcal{O}(n)$.
(2) the construction of the probability matrix P in the d-EPR model, which has complexity $\mathcal{O}(L^2)$ where L is the number of locations in the spatial tessellation.

Generation phase. In the generation phase, the generation of the mobility diary with MD has complexity $\mathcal{O}(N)$. The generation of the trajectory from the mobility diary has complexity $\mathcal{O}(LNn)$ (Algorithm 3): in the worst case, for each individual we assign a spatial location in each time slot, and the assignment of a spatial location requires a call to procedure `weightedRandom` which has complexity $\mathcal{O}(L)$. When $n \gg N$, the computational complexity can be approximated to $\mathcal{O}(Ln)$.

The total complexity of the generation phase is hence $\mathcal{O}(L^2 + Ln)$ when the probability matrix has to be constructed for the first time. In this case, when $L \sim n$ the

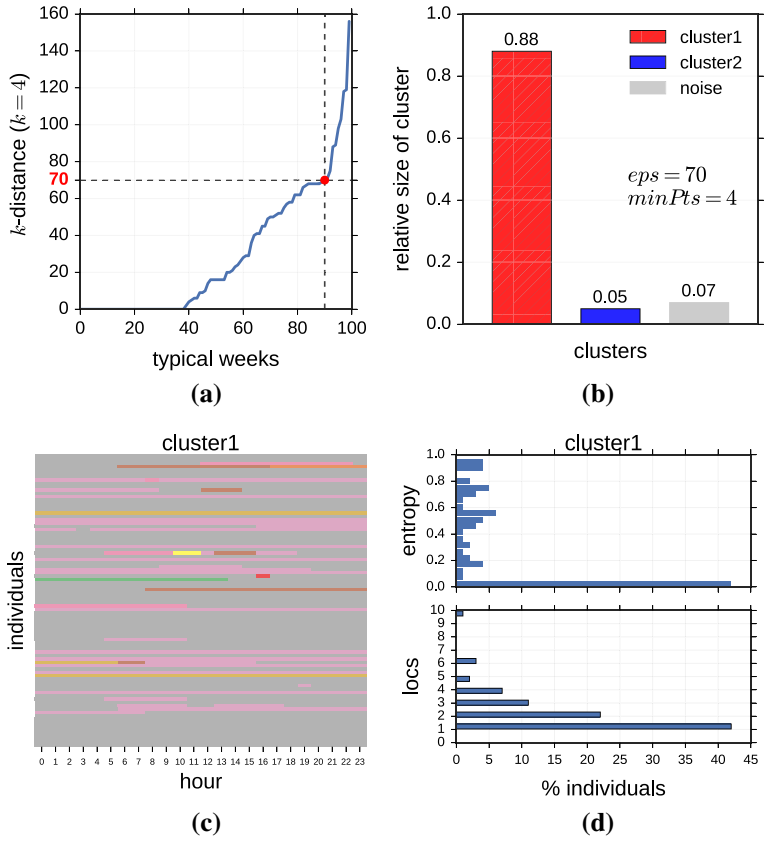

Fig. 8 First row: **a** Typical weeks sorted by distance to the 4th nearest neighbor, the elbow suggests to use $eps = 70$; **b** relative size of the clusters resulting from DBSCAN algorithm with $minPts = 4$ and $eps = 70$ and their relative size. Second row: **c** Visualization of a day of the typical weeks of 100 individuals in the GPS dataset for the first cluster. Every color represents a different abstract location in the typical diary. **d** Distribution of abstract location entropy and number of distinct abstract locations of time series of individuals in cluster 1

computational complexity can be approximated to $\mathcal{O}(L^2)$. If the probability matrix is already available or has been already computed, the computational complexity of the generation phase is $\mathcal{O}(LNn)$, which can be approximated to $\mathcal{O}(Ln)$ when $n \gg N$.

References

Ajelli M, Gonçalves B, Balcan D, Colizza V, Hu H, Ramasco JJ, Merler S (2010) Comparing large-scale computational approaches to epidemic modeling: agent-based versus structured metapopulation models. BMC Infect Dis 10(1):190. https://doi.org/10.1186/1471-2334-10-190. ISSN 1471-2334

Balcan D, Colizza V, Gonçalves B, Hu H, Ramasco JJ, Vespignani A (2009) Multiscale mobility networks and the spatial spreading of infectious diseases. Proc Natl Acad Sci 106(51):21484–21489. https://doi.org/10.1073/pnas.0906910106

Barabási A-L (2005) The origin of bursts and heavy tails in human dynamics. Nature 435(7039):207–211. https://doi.org/10.1038/nature03459

Barbosa H, de Lima-Neto FB, Evsukoff A, Menezes R (2015) The effect of recency to human mobility. EPJ Data Sci 4(1):1–14. https://doi.org/10.1140/epjds/s13688-015-0059-8. ISSN 2193-1127

Barbosa-Filho H, Barthelemy M, Ghoshal G, James CR, Lenormand M, Louail T, Menezes R, Ramasco JJ, Simini F, Tomasini M (2017) Human mobility: models and applications. arXiv:1710.00004

Batty M, Axhausen KW, Giannotti F, Pozdnoukhov A, Bazzani A, Wachowicz M, Ouzounis G, Portugali Y (2012) Smart cities of the future. Eur Phys J Spec Top 214(1):481–518. https://doi.org/10.1140/epjst/e2012-01703-3. ISSN 1951-6401

Bellemans T, Kochan B, Janssens D, Wets G, Arentze T, Timmermans H (2010) Implementation framework and development trajectory of feathers activity-based simulation platform. Transp Res Rec J Transp Res Board 2175:111–119

Boldrini C, Passarella A (2010) Hcmm: Modelling spatial and temporal properties of human mobility driven by users' social relationships. Comput Commun 33(9):1056–1074. https://doi.org/10.1016/j.comcom.2010.01.013. ISSN 0140-3664

Borrel V, Legendre F, Dias de Amorim M, Fdida S (2009) Simps: using sociology for personal mobility. IEEE/ACM Trans Netwrking 17(3):831–842. https://doi.org/10.1109/TNET.2008.2003337. ISSN 1063-6692

Brockmann D, Hufnagel L, Geisel T (2006) The scaling laws of human travel. Nature 439(7075):462–465. https://doi.org/10.1038/nature04292

Brown C, Nicosia V, Scellato S, Noulas A, Mascolo C (2013a) Social and place-focused communities in location-based online social networks. Eur Phys J B 86(6):290. https://doi.org/10.1140/epjb/e2013-40253-6. ISSN 1434-6036

Brown C, Noulas A, Mascolo C, Blondel V (2013b) A place-focused model for social networks in cities. In: 2013 International conference on social computing (SocialCom). pp 75–80. https://doi.org/10.1109/SocialCom.2013.18

Calabrese F, Colonna M, Lovisolo P, Parata D, Ratti C (2011) Real-time urban monitoring using cell phones: a case study in rome. IEEE Trans Intell Transp Syst 12(1):141–151. https://doi.org/10.1109/TITS.2010.2074196. ISSN 1524-9050

Cho E, Myers SA, Leskovec J (2011) Friendship and mobility: user movement in location-based social networks. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, KDD'11. ACM. pp 1082–1090

Colizza V, Barrat A, Barthelemy M, Valleron A-J, Vespignani A (2007) Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions. PLoS Med 4(1):1–16. https://doi.org/10.1371/journal.pmed.0040013

De Nadai M, Staiano J, Larcher R, Sebe N, Quercia D, Lepri B (2016) The death and life of great italian cities: a mobile phone data perspective. In: Proceedings of the 25th international conference on world wide web, WWW '16, pp. 413–423, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/2872427.2883084. ISBN 978-1-4503-4143-1

Eagle N, Pentland AS (2009) Eigenbehaviors: identifying structure in routine. Behav Ecol Sociobiol 63(7):1057–1066. https://doi.org/10.1007/s00265-009-0830-6

Ekman F, Keränen A, Karvo J, Ott J (2008) Working day movement model. In: Proceedings of the 1st ACM SIGMOBILE workshop on mobility models, MobilityModels '08, ACM, New York, NY, USA. pp 33–40. https://doi.org/10.1145/1374688.1374695. ISBN 978-1-60558-111-8

Erlander S, Stewart NF (1990) The gravity model in transportation analysis: theory and extensions. Topics in transportation. VSP, Utrecht, The Netherlands. http://opac.inria.fr/record=b1117869. ISBN 90-6764-089-1

Ester M, Kriegel HP, Jorg S, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the second international conference on knowledge discovery and data mining (KDD). pp 226–231

Fischer Daniel, Herrmann Klaus, Rothermel Kurt (2010) Gesomo—a general social mobility model for delay tolerant networks. In: MASS, IEEE Computer Society. pp 99–108. http://dblp.uni-trier.de/db/conf/mass/mass2010.html#FischerHR10. ISBN 978-1-4244-7488-2

Ghosh J, Philip SJ, Qiao C. (2005) Sociological orbit aware location approximation and routing in manet. In: 2nd international conference on broadband networks, 2005, vol 1. pp 641–650 https://doi.org/10.1109/ICBN.2005.1589669

Giannotti F, Pappalardo L, Pedreschi D, Wang D (2013) A complexity science perspective on human mobility. In: Mobility data: modeling, management, and understanding. pp 297–314

González MC, Hidalgo CA, Barabási A-L (2008) Understanding individual human mobility patterns. Nature 453(7196):779–782. https://doi.org/10.1038/nature06958

Hasan S, Schneider CM, Ukkusuri SV, González MC (2013) Spatiotemporal patterns of urban human mobility. J Stat Phys 151(1–2):304–318. https://doi.org/10.1007/s10955-012-0645-0

Hess A, Hummel KA, Gansterer WN, Haring G (2015) Data-driven human mobility modeling: a survey and engineering guidance for mobile networking. ACM Comput Surv 48(3):38:1–38:39 10.1145/2840722

Hidalgo CA, Rodriguez-Sickert C (2008) The dynamics of a mobile phone network. Phys A Stat Mech Its Appl 387(12):3017–3024. https://doi.org/10.1016/j.physa.2008.01.073. ISSN 0378-4371

Hossmann T, Spyropoulos T, Legendre F (2011a) A complex network analysis of human mobility. In: 2011 IEEE conference on computer communications workshops (INFOCOM WKSHPS). pp. 876–881 https://doi.org/10.1109/INFCOMW.2011.5928936

Hossmann T, Spyropoulos T, Legendre F (2011b) Putting contacts into context: mobility modeling beyond inter-contact times. In: Proceedings of the twelfth ACM international symposium on mobile ad hoc networking and computing, MobiHoc '11, vol 11. ACM, New York, NY, USA. pp 18:1–18. https://doi.org/10.1145/2107502.2107526. ISBN 978-1-4503-0722-2

Hristova D, Noulas A, Brown C, Musolesi M, Mascolo C (2016) A multilayer approach to multiplexity and link prediction in online geo-social networks. EPJ Data Sci 5(1):24. https://doi.org/10.1140/epjds/s13688-016-0087-z. ISSN 2193-1127

Iovan C, Olteanu-Raimond A-M, Couronné T, Smoreda Z (2013) Moving and calling: mobile phone data quality measurements and spatiotemporal uncertainty in human mobility studies. In: Springer (ed) 16th international conference on geographic information science (AGILE'13). pp 247–265 https://doi.org/10.1007/978-3-319-00615-4_14

Janssens D (2013) Data science and simulation in transportation research, 1st edn. IGI Global, Hershey. ISBN 1466649208, 9781466649200

Jiang S, Ferreira J Jr, González MC (2012) Clustering daily patterns of human activities in the city. Data Min Knowl Disc 25(3):478–510. https://doi.org/10.1007/s10618-012-0264-z

Jung WS, Wang F, Stanley HE. Gravity model in the korean highway. EPL: Europhys Lett 81(4):48005 http://stacks.iop.org/0295-5075/81/i=4/a=48005

Karamshuk D, Boldrini C, Conti M, Passarella A (2011) Human mobility models for opportunistic networks. IEEE Commun Mag 49(12):157–165. https://doi.org/10.1109/MCOM.2011.6094021

Kitchin R (2013) The real-time city? big data and smart urbanism. GeoJournal 79(1):1–14. https://doi.org/10.1007/s10708-013-9516-8. ISSN 1572-9893

Kopp C, Kochan B, May M, Pappalardo L, Rinzivillo S, Schulz D, Simini F (2014) Evaluation of spatio-temporal microsimulation systems. In: Knapen L, Janssens D, Yasar A (eds) Data on science and simulation in transportation research. IGI Global, Hershey

Kosta S, Mei A, Stefa J (2010) Small world in motion (SWIM): modeling communities in ad-hoc mobile networking. In 2010 7th annual IEEE communications society conference on sensor, mesh and ad hoc communications and networks (SECON). IEEE. pp 1–9. https://doi.org/10.1109/secon.2010.5508278. ISBN 978-1-4244-7150-8

Lee K, Hong S, Kim SJ, Rhee I, Chong S (2009) Slaw: a new mobility model for human walks. In: INFOCOM 2009. IEEE. pp 855–863 https://doi.org/10.1109/INFCOM.2009.5061995

Lee K, Hong S, Kim SJ, Rhee I, Chong S (2012) Slaw: self-similar least-action human walk. IEEE/ACM Trans Netw 20(2):515–529. https://doi.org/10.1109/TNET.2011.2172984. ISSN 1063-6692

Lenormand M, Gonçalves B, Tugores A, Ramasco JJ (2015) Human diffusion and city influence. J R Soc Interface 12(109). https://doi.org/10.1098/rsif.2015.0473. ISSN 1742-5689

Lenormand M, Bassolas A, Ramasco JJ (2016) Systematic comparison of trip distribution laws and models. J Transp Geogr 51:158–169. https://doi.org/10.1016/j.jtrangeo.2015.12.008. ISSN 0966-6923

Liao L, Donald J P, Fox D, Kautz H (2007) Learning and inferring transportation routines. Artif Intell 171(5–6):311–331. https://doi.org/10.1016/j.artint.2007.01.006

Marchetti S, Giusti C, Pratesi M, Salvati N, Giannotti F, Pedreschi D, Rinzivillo S, Pappalardo L, Gabrielli L (2015) Small area model-based estimators using big data source. J Off Stat 31(2):263–281. https://doi.org/10.1515/jos-2015-0017

McInerney J, Stein S, Rogers A, Nicholas R J (2013) Breaking the habit: measuring and predicting departures from routine in individual human mobility. Pervasive Mob Comput 9(6):808–822

Meloni S, Perra N, Arenas A, Gómez S, Moreno Y, Vespignani A (2011) Modeling human mobility responses to the large-scale spreading of infectious diseases. Sci Rep 1(62):08. https://doi.org/10.1038/srep00062

Merler S, Ajelli M, Fumanelli L, Vespignani A (2013) Containing the accidental laboratory escape of potential pandemic influenza viruses. BMC Med 11(1):252. https://doi.org/10.1186/1741-7015-11-252. ISSN 1741-7015

Munjal A, Camp T, Navidi WC (2011) Smooth: a simple way to model human mobility. In: Proceedings of the 14th ACM international conference on modeling, analysis and simulation of wireless and mobile systems, MSWiM '11. ACM, New York, NY, USA. pp 351–360. https://doi.org/10.1145/2068897.2068957. ISBN 978-1-4503-0898-4

Musolesi M, Mascolo C (2007) Designing mobility models based on social network theory. SIGMOBILE Mob Comput Commun Rev 11(3):59–70. https://doi.org/10.1145/1317425.1317433. ISSN 1559-1662

Navarro G (2001) A guided tour to approximate string matching. ACM Comput Surv 33(1):31–88. https://doi.org/10.1145/375360.375365. ISSN 0360-0300

Pappalardo L, Rinzivillo S, Pedreschi D, Giannotti F (2013a) Validating general human mobility patterns on gps data. In: Proceedings of the 21th Italian symposium on advanced database systems (SEBD2013)

Pappalardo L, Rinzivillo S, Qu Z, Pedreschi D, Giannotti F (2013b) Understanding the patterns of car travel. Eur Phys J Spec Top 215(1):61–73. doi:10.1140/epjst%252fe2013-01715-5

Pappalardo L, Simini F, Rinzivillo S, Pedreschi D, Giannotti F (2013c) Comparing general mobility and mobility by car. In: Proceedings of the 2013 BRICS congress on computational intelligence and 11th Brazilian congress on computational intelligence, BRICS-CCI-CBIC '13, IEEE Computer Society, Washington, DC, USA. pp 665–668. https://doi.org/10.1109/BRICS-CCI-CBIC.2013.116. ISBN 978-1-4799-3194-1

Pappalardo L, Pedreschi D, Smoreda Z, Giannotti F (2015a) Using big data to study the link between human mobility and socio-economic development. In: 2015 IEEE international conference on big data, big data 2015, Santa Clara, CA, USA, October 29–November 1, 2015, pp 871–878. https://doi.org/10.1109/BigData.2015.7363835

Pappalardo L, Simini F, Rinzivillo S, Pedreschi D, Giannotti F, Barabasi A-L (2015b) Returners and explorers dichotomy in human mobility. Nat Commun 6. https://doi.org/10.1038/ncomms9166

Pappalardo L, Rinzivillo S, Simini F (2016a) Human mobility modelling: exploration and preferential return meet the gravity model. Proc Comput Sci 83:934–939. https://doi.org/10.1016/j.procs.2016.04.188. ISSN 1877-0509. The 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016) / The 6th International Conference on Sustainable Energy Information Technology (SEIT-2016) / Affiliated Workshops

Pappalardo L, Vanhoof M, Gabrielli L, Smoreda Z, Pedreschi D, Giannotti F (2016b) An analytical framework to nowcast well-being using mobile phone data. Int J Data Sci Anal 2(1–2):75–92. https://doi.org/10.1007/s41060-016-0013-2

Ranjan G, Zang H, Zhang Z-L, Bolot J (2012) Are call detail records biased for sampling human mobility? SIGMOBILE Mob Comput Commun Rev 16(3):33–44. https://doi.org/10.1145/2412096.2412101. ISSN 1559-1662

Reades J, Calabrese F, Sevtsuk A, Ratti C (2007) Cellular census: explorations in urban data collection. IEEE Pervasive Comput 6(3):30–38. https://doi.org/10.1109/MPRV.2007.53. ISSN 1536-1268

Rinzivillo S, Mainardi S, Pezzoni F, Coscia M, Pedreschi D, Giannotti F (2012) Discovering the geographical borders of human mobility. Künstl Intell 26(3):253–260. https://doi.org/10.1007/s13218-012-0181-8

Rinzivillo S, Gabrielli L, Nanni M, Pappalardo L, Pedreschi D, Giannotti F (2014) The purpose of motion: learning activities from individual mobility networks. In: Proceedings of the 2014 international conference on data science and advanced analytics, DSAA'14. pp 312–318. https://doi.org/10.1109/DSAA.2014.7058090

Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7. ISSN 0377-0427

Schneider CM, Belik V, Couronné T, Smoreda Z, González MC (2013) Unravelling daily human mobility motifs. J R Soc Interface 10(84). https://doi.org/10.1098/rsif.2013.0246. ISSN 1742-5689

Schwamborn M, Aschenbruck N (2013) Introducing geographic restrictions to the slaw human mobility model. In: 2013 IEEE 21st international symposium on modelling, analysis and simulation of computer and telecommunication systems. pp 264–272. https://doi.org/10.1109/MASCOTS.2013.34

Simini F, González MC, Maritan A, Barabási AL (2012) A universal model for mobility and migration patterns. Nature 484:96–100. https://doi.org/10.1038/nature10856

Solmaz G, Akbaş Mİ, Turgut D (2012) Modeling visitor movement in theme parks. In: 2012 IEEE 37th conference on local computer networks (LCN). pp 36–43. https://doi.org/10.1109/LCN.2012.6423650

Solmaz G, Akbaş Mİ, Turgut D (2015) A mobility model of theme park visitors. IEEE Trans Mob Comput 14(12):2406–2418. https://doi.org/10.1109/TMC.2015.2400454. ISSN 1536-1233

Song C, Koren T, Wang P, Barabási A-L (2010a) Modelling the scaling properties of human mobility. Nat Phys 6(10):818–823. https://doi.org/10.1038/nphys1760. ISSN 1745-2473

Song C, Qu Z, Blumm N, Barabási A-L (2010b) Limits of predictability in human mobility. Science 327(5968):1018–1021. https://doi.org/10.1126/science.1177170

Spinsanti L, Berlingerio M, Pappalardo L (2013) Mobility and geo-social networks. In: Mobility data: modeling, management, and understanding. pp 315–333

Tan P-N, Steinbach M, Kumar V (2005) Introduction to data mining, 1st edn. Addison-Wesley Longman Publishing Co. Inc., Boston. ISBN 0321321367

Thiemann C, Theis F, Grady D, Brune R, Brockmann D (2010) The structure of borders in a small world. PLoS ONE 5(11):e15422

Tomasini M, Mahmood B, Zambonelli F, Brayner A, Menezes R (2017) On the effect of human mobility to the design of metropolitan mobile opportunistic networks of sensors. Pervasive Mob Comput 38(Part 1):215–232. https://doi.org/10.1016/j.pmcj.2016.12.007. ISSN 1574-1192

Venkatramanan S, Lewis B, Chen J, Higdon D, Vullikanti A, Marathe M (2017) Using data-driven agent-based models for forecasting emerging infectious diseases. Epidemics https://doi.org/10.1016/j.epidem.2017.02.010. ISSN 1755–4365

Volkovich Y, Scellato S, Laniado D, Mascolo C, Kaltenbrunner A (2012) The length of bridge ties: structural and geographic properties of online social interactions. In: Proceedings of the sixth international conference on weblogs and social media, Dublin, Ireland, June 4–7 http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4670

Wang D, Pedreschi D, Song C, Giannotti F, Barabási A (2011) Human mobility, social ties, and link prediction. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '11. ACM, New York, NY, USA. pp 1100–1108. https://doi.org/10.1145/2020408.2020581. ISBN 978-1-4503-0813-7

Wang P, Hunter T, Bayen AM, Schechtner K, González MC (2012) Understanding road usage patterns in urban areas. Sci Rep 2(1001). https://doi.org/10.1038/srep01001

Wilson AG (1969) The use of entropy maximising models, in the theory of trip distribution, mode split and route split. J Transp Econ Policy 111(1):108–126. https://doi.org/10.2307/20052128

Yang S, Yang X, Zhang C, Spyrou E (2010) Using social network theory for modeling human mobility. IEEE Netw 24(5):6–13. https://doi.org/10.1109/MNET.2010.5578912. ISSN 0890-8044

Yang Y, Jiang S, Gupta S, Veneziano D, Athavale S, Gonzalez MC (2016) The TimeGeo modeling framework for urban mobility without travel surveys. PNAS 113(37). https://doi.org/10.1073/pnas.1524261113

Zheng Q, Hong X, Liu J, Cordes D, Huang W (2010) Agenda driven mobility modelling. IJAHUC 5(1):22–36. https://doi.org/10.1504/IJAHUC.2010.03

Zipf GK (1946) The p1p2/d hypothesis: On the intercity movement of persons. Am Sociol Rev 11(6):677–686