UNIVERSITÁ DI PISA

DEPARTMENT OF COMPUTER SCIENCE

Master Degree in Computer Science
Data and Knowledge: Science and Technologies

Modeling Human Mobility considering Spatial, Temporal and Social Dimensions

Candidate
Giuliano Cornacchia

Supervisors
Dr. Luca Pappalardo
Dr. Giulio Rossetti

Academic Year 2018 / 2019

Contents

List of Figures

List of Tables

Abstract

The analysis of human mobility is crucial in several areas, from urban planning to epidemic modeling, estimation of migratory flows and traffic forecasting. However, mobility data (e.g., Call Detail Records and GPS traces from vehicles or smartphones) are sensitive since it is possible to infer personal information even from anonymized datasets. A solution to dealing with this privacy issue is to use synthetic and realistic trajectories generated by proper generative models. Existing mechanistic generative models usually consider the spatial and temporal dimensions only. In this thesis, we select as a baseline model GeoSim, which considers the social dimension together with spatial and temporal dimensions during the generation of the synthetic trajectories. Our contribution in the field of the human mobility consists of including, incrementally, three mobility mechanisms, specifically the introduction of the distance and the use of a gravity-model in the location selection phase, finally, we include a diary generator, an algorithm capable to capture the tendency of humans to follow or break their routine, improving the modeling capability of the GeoSim model. We show that the three implemented models, obtained from GeoSim with the introduction of the mobility mechanisms, can reproduce the statistical proprieties of real trajectories, in all the three dimensions, more accurately than GeoSim.

Abstract

L'analisi della mobilità umana è fondamentale per diverse discipline, dalla pianificazione urbana all'epidemiologia computazionale, stima dei flussi migratori e previsioni del traffico. Tuttavia, i dati riguardanti la mobilità (ad esempio Call Detail Records e tracce GPS da automobili o smartphones) sono sensibili perchè è possibile inferire informazioni personali anche da dataset anonimizzati. Una soluzione per affrontare questo problema di privacy è utilizzare traiettorie sintetiche e realistiche generate da adeguati modelli generativi. I modelli generativi meccanicistici esistenti di solito considerano solo le dimensioni spaziali e temporali. In questa tesi, selezioniamo come modello di base GeoSim, che considera la dimensione sociale insieme alle dimensioni spaziali e temporali durante la generazione delle traiettorie sintetiche. Il nostro contributo nel campo della mobilità umana consiste nell'includere, in modo incrementale, tre meccanismi di mobilità, in particolare l'introduzione del concetto di distanza e l'uso di un gravity-model nella fase di selezione della posizione da visitare, infine, includiamo un diary generator, un algoritmo capace di catturare la tendenza degli individui si seguire o interrompere la propria routine, migliorando la capacità di modellazione del modello GeoSim. Mostriamo che i tre modelli implementati, ottenuti da GeoSim con l'introduzione dei meccanismi di mobilità, possono riprodurre le proprietà statistiche delle traiettorie reali, in tutte e tre le dimensioni, in modo più accurato di GeoSim.

Chapter 1

Introduction

Understanding and modeling the mechanisms that govern human mobility is of fundamental importance in different disciplines, such as computational epidemiology, traffic forecasting, urban planning, and what-if analysis [1, 2, 3, 4]. The first step in human mobility modeling is to understand how people move. In recent years, the availability of large mobility datasets such as Call Detail Records (CDR) [4, 3, 2], traces from GPS devices embedded in smartphones and cars [2], and geo-tagged posts on Location Based Social Networks (LBSN) [5], allows to observe and study human movements in great detail.

Many works in the literature exploit mobility data to quantify the patterns that characterize the movements of individuals. These studies show the heterogeneous nature of travel patterns, the existence of a power-law distribution in jump lengths, namely the distance between the source and destination location of a displacement [6, 4], and in the characteristic spatial spread of an individual, referred to as radius of gyration [4]. Humans exhibit a strong tendency to return to locations they visited before [4] and a propensity to be stationary during the night hours while they preferentially move at specific times of the day, following a circadian rhythm [2, 7]. The time spent by an individual in a location is not uniformly distributed, but follows a power-law distribution [3]. Moreover, the sociality shapes the displacements of the individuals, about the 10-30% of the human movements can be explained by social purposes [8].

Mobility datasets are difficult to obtain beacause the companies that own them cannot make the data publicly available because it is possible to infer personal information of the individuals even from anonymized datasets [9, 10]. A solution to dealing with this privacy issue is to use the above descriptions of human mobility patterns to create generative algorithms able to produce a set of synthetic trajectories with the same statistical properties of the real ones. The synthetic trajectories, can be used in what-if analysis or for training a machine learning model without the need of a real dataset.

A significant advantage of using generative models concerns the cost and the time spent in the data collection, which is negligible with respect to the acquisition of a real dataset. Moreover, using generative algorithms allows the simulation of the mobility for a set of agents in an unseen scenario.

Most generative models focus only on the spatial and temporal dimensions of human mobility. In this thesis, we take into account the social dimension too. We propose an extension of GeoSim [11], a state-of-the-art model which takes into account the social dimension introducing two mechanisms in addition to the explore and preferential return ones: individual preference and social influence. We include incrementally three mobility mechanisms to improve its modeling capability. In the first extension, we include a mechanism that takes into account the spatial distance between locations. Then, we include a spatial mechanism for considering both the relevance of a location together with the spatial distance. In the last extension, we include a Mobility Diary Generator, a data-driven algorithm able to capture the tendency of individuals to follow a circadian rhythm [2]. We also propose several additional features that can be included in the proposed generative algorithm to model human mobility, also considering other aspects such as the popularity of a node or include a constrain in the location an agent can reach in a given time.

We compare the synthetic trajectories generated by the proposed extensions with a set of real mobility trajectories relative to the area of New York City, obtained from an LBSN dataset. The similarity between the two sets of trajectories is evaluated between their probability density functions with respect to a set of standard mobility metrics used in human mobility. We show that all the proposed models increase the modeling ability of GeoSim, in particular, the algorithm that uses the Mobility Diary Generator produces the more realistic trajectories. We further validate the modeling ability of the best extension, GS$_{diary}$, simulating the mobility of individuals in different scenarios. We simulate the mobility of individuals moving in London, including in the model different levels of knowledge concerning their mobility behaviors. We compare the synthetic trajectories with a set of real trajectories relative to the urban area of London, obtained as for the New York City dataset, to demonstrate that our model can be applied on different regions without any significant loss of realism.

This thesis is structured as follows. In Chapter 2, we review the state-of-the-art human mobility patterns and generative models. In Chapter 3, we present the baseline model GeoSim (Section 3.2). In Sections 3.3, 3.4 and 3.5 we describe the proposed extensions, respectively GS$_d$, GS$_{gravity}$, and GS$_{diary}$. In Section 4.1, we present the mobility measures used in the evaluation of the generated trajectories, in Section 4.2 we present the LBSN dataset and the filtering process we apply to obtain a mobility dataset relative to the trajectories of 1,001 individuals in New York City, and we present the associated social graph e weighted spatial tessellation. In Chapter 4, we show the experimental setting (Section 4.3.1), then we introduce the scores used to quantify the statistical similarity between the generated and the real trajectories. In Section 4.3.3, we discuss the obtained results and we show some table relative to significant experiments (4.3.4). In Section 4.3.5 we present the London mobility dataset, and we show and discuss the results obtained from the experiments concerning mobility of individuals in London, instantiating the model with different levels of knowledge relative at the scenario to simulate. We then conclude with a discussion on the main contributions of this thesis and further improvements that can be performed (Chapter 5).

Chapter 2

Related Work

2.1 Human mobility patterns

All the main studies in the field of human mobility focus on discovering and explaining the hidden mechanisms that rule the displacements of individuals, showing how human mobility is not a random process and can be explained through simple concepts [1].

Brokmann et al. [6], analyzing the trajectories of 464,760 dollar bills, collected from an online bill-tracking website[1], conclude that the jump length Δr, the distance between two instances of the appearance of a banknote, follows a truncated power-law, independently of the size of the studied population. Gonzalez et al. [4] study the trajectories of 100,000 anonymized mobile-phone users and find that human trajectories show a high degree of temporal and spatial regularity [4]; they confirm the observation made by Brokmann et al., the distribution of the jump length can be approximated by a power-law distribution:

$$P(\Delta r) \approx (\Delta r + \Delta r_0)^{-\beta} e^{-\Delta r/k} \tag{2.1}$$

with the exponent $\beta = 1.75 \pm 0.15$, $\Delta r_0 = 1.5$ km and the cut-off value k depends on the extension of the studied area; this fits the intuition that individuals usually move at short distances and occasionally take long trips. They also find that the distributions of the radius of gyration r_g, the characteristic distance traveled by an individual, can be approximated with a truncated power-law:

$$P(r_g) \approx (r_g + r_g^0)^{-\beta} e^{-r_g/k} \tag{2.2}$$

with the exponent $\beta = 1.65 \pm 0.15$, $r_g^0 = 5.8$ km and the cut-off value $k = 350$ km.

[1] https://www.wheresgeorge.com

Humans exhibit a strong tendency to return to locations they visited before, with a probability proportional to the individual visitation pattern, as demostrated by Gonzalez et al. [4].

Song et al. [12] observed in CDR data that the distribution of the waiting time Δt, namely the time spent by an invidual in location, can be described by a truncated power-law:

$$P(\Delta t) \approx (\Delta t)^{-\beta} e^{-\Delta t/\tau} \tag{2.3}$$

where $\beta = 1.8 \pm 0.1$, $\tau = 17$ h, the latter corresponds to the typical awake time of an individual [12]. In the same work, Song et al., studying the entropy of individuals' displacements, demonstrate that if we use only the heterogeneity of the spatial distribution the predictability of the individuals is insignificant; instead, if we consider a historical record of the visited locations the potential predictability is 93%. Pappalardo et al. analyzing CDR and GPS data discover that individuals can be split into two profiles: returners and explorers, with distinct spatial patterns. [13].

Several studies demonstrate the correlation between human mobility and sociality [11, 14, 8, 15, 5]; the movements of friends are more similar than those of strangers, mainly because we are more likely to visit a location if a social contact explored that location before; another reason is that friends have an higher probability of living or working together, or having the same hobbies, increasing their mobility similarity compared with strangers [15]. The similarity of the movements of a pair of individuals can be seen as an overlap of their trajectories; Cho et al. [8] find that users are more likely to check-in right after a friend has checked-in to the same location, and the probability drop offs following a power-law function. Making a check-in in the same place results in an overlapped point in the trajectory of the users, consequently their mobility similarity increases. The sociality can explain also long trips; according to Cho et al. [8] the probability for an individual of visiting a friend remains constant as a function of the distance. Furthermore, individuals with a similar visitation pattern are more likely to establish a social link; this can be explained due to the fact that if two individuals go often in the same places they are more likely to establish a friendship; the probability of being connected in a social link can be computed through the mobility patterns similarity, as performed by Wang et al. [14] in the link-prediction classification task. According to Cho et al. [8] social relationships can explain about 10-30% of all human movements.

2.2 Generative models

In the literature several mechanistic generative models use the statistical principles and the mechanisms that govern human mobility to generate a set of synthetic and realistic trajectories that reproduces the patterns of human mobility. The exploration and preferential return (EPR) model described by Song et al. [12] is based on two basic mechanisms: exploration and preferential return. The exploration mechanism consists in a random walk process with truncated power-law jump size distribution. The preferential return mechanism reproduces the propensity of humans to return to locations they visited before [4]. If an agent returns to a previously visited location, it selects a location r_i with probability proportional to the number of times the agent visited that location. An agent in the model selects to explore a new location with probability $P_{exp} = \rho S^{-\gamma}$ where S is the total number of locations visited by the agent and ρ and γ are constants; with complementary probability $P_{ret} = 1 - \rho S^{-\gamma}$ the agent returns to a previously visited location. This model does not fix in advance the number of locations but allows them to emerge during the simulation [12].

Pappalardo et al. [16] propose the d-EPR model, an extension of the EPR model. The extension is based on the same mechanisms of exploration and preferential return, but in the exploration phase the individual is driven by a collective force. An individual selects a new location to visit depending on both its distance from the current position, as well as its relevance measured as the total number of visits of all users [16].

Alessandretti et al. [17] propose the limited memory EPR model. In this version of the EPR model, the exploration mechanism is the same as the EPR model, but the agent disposes of a limited memory M [17]. When an agent returns to a previously visited location, it selects a location r_i with probability proportional to the number of times the agent visited that location in the previous M days.

DITRAS (DIary-based TRAjectory Simulator) [2] is a mechanistic modeling framework for generating trajectories that reproduces in a realistic way the spatio-temporal patterns of human mobility. DITRAS generates the trajectories using two probabilistic models: a diary generator and a trajectory generator (Figure 2.1). The diary generator is Markov model trained on mobility trajectory data of real individuals, which captures the probability for individuals to follow or break their routine at specific times [2]; the diary generator builds a mobility diary with abstract locations for each agent in the simulation. The trajectory generator is an algorithm that, given a weighted spatial tessellation, translates the abstract locations in physical locations using the d-EPR model [3]. DITRAS has been validated generating the trajectories of 10,000 agents and comparing statistically the generated trajectories with CDR and GPS data, obtaining very accurate results.

GeoSim, the model proposed by Toole et al [11], is an extension of the EPR model which takes into account also the social dimension together with the spatial and temporal dimensions. GeoSim introduces two mechanisms in addition to the explore and preferential return ones: individual preference and social influence. The agent has to decide if its next displacements will be influenced or not by its social contact, respectively with probability α

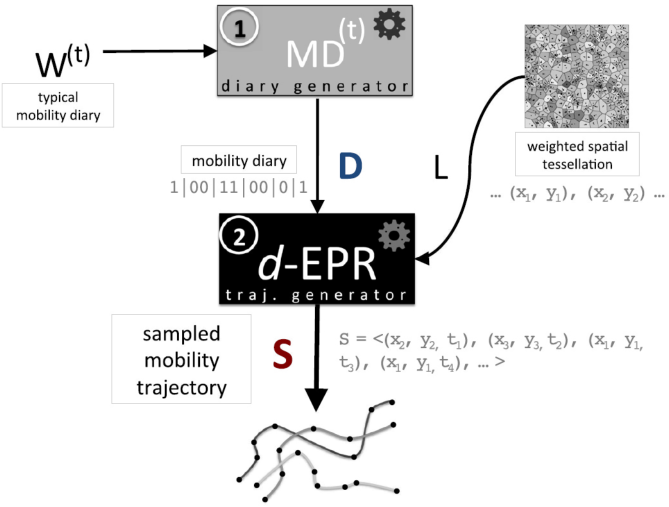


Figure 2.1: A schematic representation of the DITRAS frameworks. Ditras uses two probabilistic models: a diary generator (1) and a trajectory generator (2). The diary generator produces a mobility trajectory that is mapped to a weighed spatial tessellation from the trajectory generator. Figure from [2].

and $1 - \alpha$.

From the literature emerges the lack of generative mobility models able to produce realistic trajectories considering at the same time the spatial, temporal and social dimensions of human mobility. Our work consists in the proposal of three mechanistic generative models, extension of a baseline model GeoSim, able to generate realistic mobility trajectories taking into account also the social dimension together with the spatial and temporal dimensions. Starting from GeoSim, we implement the extensions including, incrementally, three mobility mechanisms: the distance and the use of a gravity-model in the location selection phase and a diary generator able to reproduce the circadian rhythm of the individuals.

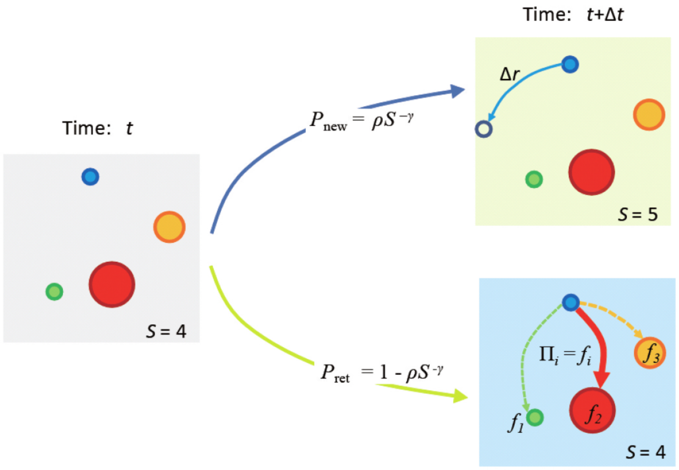


Figure 2.2: A schematic representation of the EPR model. In the EPR model an agent selects whether to return to a previously visited location or explore a new location. Figure from [3].

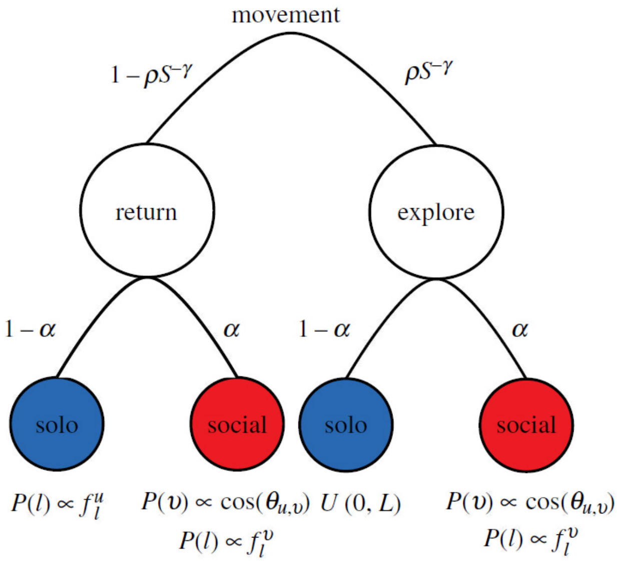


Figure 2.3: A schematic representation of the GeoSim model. In the GeoSim model an agent first selects whether to return to a previously visited location or explore a new location, and after determine if the choice of the location will be influenced or not by a social contact. Figure from [11]

Chapter 3

Modeling Individual Mobility

Generative models of human mobility have the purpose of generating realistic trajectories for a set of agents [2]. All the generative models considered in this work are mechanistic, they assumes that *a complex system can be understood by examining the workings of its individual parts and the manner in which they are coupled.*

The models considered exploit two competing mechanisms to describe human mobility: exploration and preferential return [3, 2, 16, 13]. The exploration mechanism models the scaling law presented by Song et al. [3]: the tendency to explore new locations decreases over time. Preferential return reproduces the significant propensity of individuals to return to locations they explored before [3, 2, 16]. An agent explores a new location with probability $P_{exp} = \rho S^{-\gamma}$, or returns to a previously visited location with a complementary probability $P_{ret} = 1 - \rho S^{-\gamma}$, where S is the number of unique locations visited by the individual and $\rho = 0.6$, $\gamma = 0.21$ are constants [3]. When the agent returns, the location is selected with a probability proportional to its location frequency (Equation 4.3). 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 [2, 16, 7].

In Section 3.1 we introduce some of the basic concepts for modeling human mobility. Then we present GeoSim (Section 3.2) and the three extensions we propose. Starting from GeoSim, we incrementally include some mechanisms obtaining different and more detailed models. In GS$_d$ (Section 3.3), the first extension of GeoSim, we include a mechanism for modeling the role of the spatial distance from the current location and the location to explore. Next in Section 3.4 we describe GS$_{gravity}$, which includes a gravity law mechanism for taking into account, during the choice of the location to explore, both the spatial distance from the current position and its relevance. In GS$_{diary}$ (Section 3.5), the last extension, we use a diary generator able to capture the tendency of individuals to follow or break their routine. In Section 3.6 we describe the proposed additional features and finally in

3.7 we present the implementative choices.

3.1 Key Mobility Concepts

The term human mobility refers to the movements of individuals in space and time [1]. For modeling purposes, human mobility can be defined as *the movements of a set of individuals within a discrete number of locations during an observation period.* The displacements of an individual over time can be described by a mobility trajectory.

Definition 3.1. A mobility trajectory is a sequence $T = \langle (r_1, t_1), \ldots, (r_n, t_n) \rangle$ where t_i is a timestamp such that $\forall i \in [1, n)$ $t_i < t_{i+1}$ and r_i is defined as (x_i, y_i) where the components are coordinates on a bi-dimensional space.

The individuals move on a weighted spatial tessellation defined as follows.

Definition 3.2. A weighted spatial tessellation L represents the tiling of a bi-dimensional space, resulting in a *non-overlapped* set of locations.
Every location has a weight corresponding to its relevance at a global level [2]; furthermore, every location is associated with a representative point, generally the centroid of the tile expressed as a pair of coordinates.
$L = \langle (r_1, w_1), \ldots, (r_n, w_n) \rangle$ where w_j is the weight of the tile j and r_j is the representative point of the tile j.

The number of times a user visits each location during a period of observation represents its visitation pattern.

Definition 3.3. The visitation pattern of an individual a can be represented as a vector lv_a of $|L|$ elements, called location vector, where $|L|$ is the total number of locations. The *j-th* element of the location vector, $lv_a[j]$, contains the number of times a visited the location r_j.

3.2 GeoSim

GeoSim, proposed by Toole et al. [11], extends the EPR model presented by Song et al. [3]. The EPR model considers only the spatial and temporal dimensions during the generation of the synthetic trajectories. GeoSim, instead, takes into account also the social dimension to model the fact that about 10-30% of the human movements can be explained by social purposes [8]. GeoSim introduces two mechanisms in addition to the presented ones: individual preference and social influence. When an agent decides to explore or return, it has to determine if the choice will be affected or not by the other agents, with a probability, respectively, of α and $1 - \alpha$, where $\alpha = 0.2$ is a constant [11]. In the individual preference mechanism, the agent behaves selecting the location according to its visitation pattern, without any influences of the other agents involved. Otherwise, in the social influence mechanism, an agent selects the location to explore or return with the influence of the visitation patterns of its social contacts (i.e., a user i is a social contact for u if exists an edge (u, i) in the social graph).

The model takes in input:

- N: the number of synthetic individuals;
- L: a tessellation of the space where the agents are allowed to move;
- G: an undirected graph modeling the sociality of the N agents;
- min_{wt}: the minimum amount of time an agent spend in a location, set to 1 hour;
- $[start, end]$: the time interval of the simulation;

and outputs the generated synthetic trajectories.

GeoSim can be decomposed into three phases: initialization, action selection and location selection.

In the initialization phase, a set of N agents are distributed randomly across $|L|$ locations and connected in an undirected graph, called social graph, describing the social links between agents. The weight corresponding to each edge represents the mobility similarity between the linked agents. The location vector lv of each agent is initialized to reflect its single visit, and the number S of visited locations is set to one. At the end of this phase, each agent selects a waiting time $\Delta t \geq min_{wt}$ from the empirical probability density function $P(\Delta_t) \approx \Delta_t^{-\beta} e^{-\Delta_t/\tau}$ [3]; each agent is allowed to move after the picked Δt.

After the initialization phase, every agent decides whether or not to move according to its own waiting time. When an agent moves, performs the action and location selection steps, then chooses a new waiting time, updates the location vector and, if the agent explored a new location, S is incremented by one. The agents iterates these actions until a stopping criterion is satisfied (e.g., the number of hours to simulate is reached).

During the action selection phase (Figure 3.1), the agent first chooses to either visit a new location with probability $P_{exp} = \rho S^{-\gamma}$ or to return to a previously visited location with probability $P_{ret} = 1 - \rho S^{-\gamma}$. At that point, independently of the mechanisms selected, the agent chooses to perform the location selection phase individually with probability $1 - \alpha$ or with the influence of its contacts with probability α.

After the agent picked the action among the four combinations, in the location selection phase the agent selects the destination of its move according to the chosen mechanisms. For an agent a, we define the sets containing the indices of the locations a can explore or return respectively, as follows:

$$exp_a = \{i \mid lv_a[i] = 0\} \tag{3.1}$$

$$ret_a = \{i \mid lv_a[i] > 0\} \tag{3.2}$$

The frequency of visits of an individual a relative to a location r_i is referred as $f_a(r_i)$.

- **Exploration-Individual:** During the individual exploration, an agent a chooses a new location to explore from the set exp_a uniformly at random. The probabiliy $p(r_i)$ for a location r_i with $i \in exp_a$ to be selected is $\frac{1}{|exp_a|}$.
- **Exploration-Social:** In the social exploration action, an agent a selects a social contact c. The probability $p(c)$ for a social contact c to be selected is directly proportional to the mobility-similarity between them: $p(c) \propto mob_{sim}(a, c)$. After the contact c is chosen, the location to explore is selected from the set $A = exp_a \cap ret_c$; the probability $p(r_i)$ for a location r_i, with $i \in A$, to be selected is proportional to the visitation pattern of the agent c, namely $p(r_i) \propto f_c(r_i)$.
- **Return-Individual:** In the individual return action, an agent a picks the return location from the set ret_a with a probability directly proportional to its visitation pattern. The probability for a location r_i with $i \in ret_a$ to be chosen is: $p(r_i) \propto f_a(r_i)$.
- **Return-Social:** The contact c is selected as in the Exploration-Social action, while the set where the location is selected from is defined as $A = ret_a \cap ret_c$; the probability $p(r_i)$ for a location r_i to be selected is proportional to the visitation pattern of the agent c, namely $p(r_i) \propto f_c(r_i)$.

3.3 GeoSim-distance

A modeling limit of GeoSim concerns the spatial patterns of the generated synthetic trajectories. GeoSim does not take into account the distance from the current location and the location to explore [11], since the destination of the next move is chosen uniformly at random. Consequently, the probability density functions for both jump size and radius of gyration of the generated trajectories do not follow the proper empirical distribution [4].

The power-law behavior of the probability density function of the jump length suggests that individuals are more likely to move at small rather than long distances. To take into account this observation in the first extension of GeoSim, namely GS$_d$, in the Exploration-Individual action an agent a currently at location r_j, selects an unvisited location r_i, with $i \in exp_a$, with probability $p(r_i) \propto \frac{1}{d_{ij}}$ where d_{ij} is the geographic distance between location r_i and r_j.

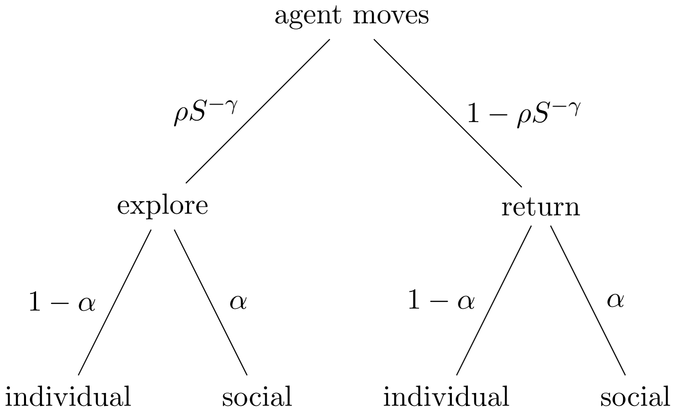


Figure 3.1: A schematic description of the action selection phase of the considered models. The individual first decides whether to explore a new location or return to a previously visited one. Then the agent determines if its social contacts will affect or not its choice for the location to visit next.

3.4 GeoSim-gravity

Individuals do not consider the distance from a place as the only discriminant factor while selecting the next location to explore. They are driven by a preferential-exploration force in the selection of the new location to explore [16, 2]. The individuals take into account also the relevance of a location at a collective level together with the distance from their current location. The method used for coupling both the distance and the relevance is the same used in the *d-EPR* model [16]: the use of a gravity law. 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 [16].

In GS$_{gravity}$, the second proposed extension of GeoSim, an agent a currently at location r_j, during the Exploration-Individual action selects an unvisited location r_i, with $i \in exp_a$, with probability $p(r_i) \propto \frac{w_i w_j}{d_{ij}^2}$ where d_{ij} is the geographic distance between location r_i and r_j and w_i, w_j represent their relevance.

The relevance of a location can be estimated through the measure visits per location (Sect. 4.1), using a real mobility dataset. In case the real information is not available, the relevance of a location can be estimated using the population density [2].

3.5 GeoSim-diary

Both the proposed extensions focus on the improvement of the spatial patterns of the generated trajectories. As discussed before, the displacements of human beings follow a circadian rhythm since individuals are mainly stationary during the night and move at a specific time of the day [2, 7]. In the discussed models is not included any mechanism able to capture the individual's recurrent and periodic schedules. in fact the periodicity of the displacement of an individual is given from its waiting times.

To model the individuals' routines, we propose GS$_{diary}$, an extension of GeoSim; in GS$_{diary}$, we use a mobility diary generator, an algorithm able to capture the tendency of individuals to follow or break their routine [2]. During the first step, the model assigns at each agent a mobility diary produced by the mobility generator. A mobility diary MD for an agent a is defined as:

$$\mathrm{MD}_a = \langle (ab_0, t_1), (ab_1, t_2), \dots (ab_j, t_{j+1}), (ab_0, t_{j+2}), \dots) \rangle \tag{3.3}$$

Where ab is an *abstract location*, ab_0 denotes the home location of the agent a and the visits between two home locations are called *run*. The abstract location ab_0 is assigned randomly to a physical location in the initialization step.

The agent will move according to the entries in its mobility diary: if the current abstract location is ab_0 the agent visits the home location, otherwise converts the abstract location into a physical one selecting the location to visit with the same approach as in GS$_{gravity}$. The physical locations visited by an individual during a run must be distinct from each other, but the physical location resulting from the mapping of ab_i can be different in different runs. In GS$_{diary}$ there is no need to select a waiting time from the empirical distribution [12], since it is intrinsically specified in the mobility diary.

3.6 Additional features

For each of the presented models, we introduce some additional features that may capture in a realistic way some crucial aspects of human mobility.

Relevance-based Starting Locations

Given a weighted spatial tessellation L, during the initialization phase, the agents are assigned to a starting location r_i with a probability $p(r_i) \propto \frac{1}{|L|}$. With the introduction of the concept of relevance at a collective level for a location, we decide to propose a new strategy for assigning the agents at the starting location, the RSL (Relevance-based Starting Locations). In this strategy the agents aren't assigned uniformly at random, the probability $p(r_i)$ for an agent of being assigned to a starting location r_i is $\propto w_i$, where w_i is the relevance of the location at a collective level.

Reachable locations

When a synthetic individual is allowed to move, it is associated with a waiting time, extracted from the empirical distribution $P(\Delta_t) \approx \Delta_t^{-\beta} e^{-\Delta_t/\tau}$ [3], or specified in the mobility diary of the agent, depending on which extension of GeoSim in considered. The agent associated with a waiting time Δ_t can not physically visit every location, but realistically only the locations the individual can reach moving at a certain speed for the picked amount of time.

We define $speed_{agent}$ as the typical speed of an individual and I as the set of all the locations the agent can visit, the set R of the reachable locations for an agent starting from the location r_j is computed as follows:

$$R = \{i \in I \mid dist(r_j, r_i) \leq dist_{max}\} \tag{3.4}$$

where $dist_{max} = \Delta t \cdot speed_{agent}$

Social Choice by Degree

When an agent a performs a social action, selects a contact c with a probability $p(c) \propto mob_{sim}(a, c)$. The choice of the social contact is personal for the agent; in fact, it is determined using only individual information, and no collective information is considered. We propose a new contact selection method. During a social return, the selection process is the same as the presented models. Instead, in the social exploration, the contact is determined using its popularity at the collective level. The popularity *pop* of an individual u within a social graph G is defined as:

$$pop(u, G) = deg(u, G) \tag{3.5}$$

where the degree of a node n in the graph G is denoted as $deg(n, G)$. In the social exploration, an agent a selects a social contact c with probability $p(c) \propto pop(c, G)$.

The intuition underlying this proposal is very simple. When an individual decides to explore a new location with the influence of a contact, a popular contact will be more likely to influence its decision than an unpopular one, even if their mobility patterns are very different. For example, an event promoter (generally a popular node within a social graph) has a high probability of influencing the choice of one of its contacts, during the selection of the next location to explore, even though they can have different mobility behaviors. In contrast, when an individual decides to return at an already visited location with the influence of its social contacts, it is reasonable to think that the choice of the contact is conducted using individual information. During the return action, individuals follow their routine, consequently they are more likely to select a contact with a similar mobility pattern.

Action-correction phase

The location an agent can select as the destination of its next displacement is constrained. In several cases, the set of locations where the individual is allowed to move is empty; in the literature, it is unspecified how to deal with these situations. We extend the models including a new phase: the action correction phase. The action correction phase is executed after the location selection, if the latter is too restrictive and do not allow movements in any location.

- **No new location to explore**: When an individual a performs the selection action phase (Figure 3.1) and decides to explore individually an unvisited location, it selects the location from the set exp_a (Equation 3.1). In case the individual visited all the locations at least one time, no choice can be made since $exp_a = \varnothing$. We deal with this case correcting the action (Figure 3.2) of the individual from Exploration-Individual to Return-Individual, preserving in this way the choice of performing the location selection without any influence of its social contacts.

- **No location in social choices**: If the agent a decide to move with the influence of a social contact c, and the set A computed for the relative action $A = ret_a \cap ret_c$ or $A = exp_a \cap ret_c$ is empty, we correct the action from the current to Return-Individual (Figure 3.2).

- **No reachable locations**: When a synthetic individual is allowed to move, it is associated with a waiting time Δt, and it can reach the locations in the set R (Eq. 3.6). The set R may be empty even if $I \neq \varnothing$, meaning there is no location the agent can visit within the radius $dist_{max}$ but the set of possible choices is not empty. If the agent was performing an exploration, a new $\Delta t_1 > \Delta t$ is selected to expand the area the agent can cover during its displacement. After picking the new waiting time the set R is computed, if $R = \varnothing$ then a new $\Delta t_2 > \Delta t_1$ is picked, (this procedure is repeated for a maximum of n_{max} time) and an Exploration-Individual action is performed with the new waiting time. If the agent was performing a Return-Social action or in the case the incrementing of the waiting time was performed n_{max} times, then the action is corrected with a Return-Individual; note that Return-Individual can not fail, since the agent can always return in its current location r_j from the moment that $dist(r_j, r_j) = 0$ (Figure 3.3).

- **Run in the Mobility Diary**: In the mobility diary the locations visited by an individual during a *run* (defined as the visits between two home return) must be distinct from each other. Given a run $d = \langle ab_1, ab_2, \ldots, ab_n \rangle$ of length n, all the abstract locations $ab_i \in d$

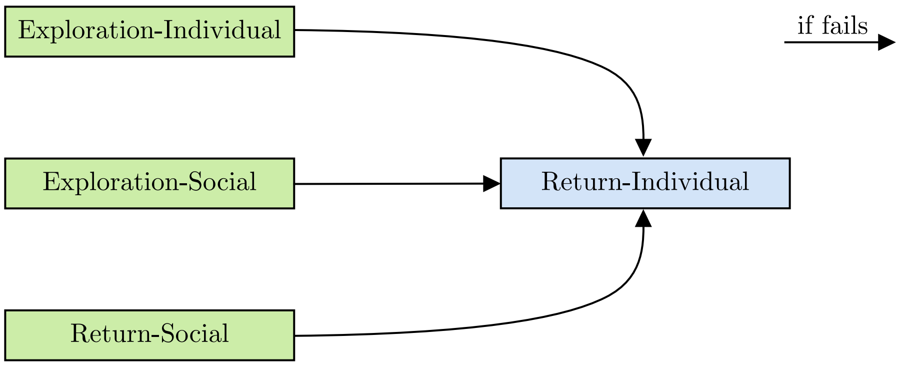


Figure 3.2: A description of the action correction routine in the case the agent can neither explore nor perform a social choice. The green rectangle denotes a starting state while the light blue a final state. When one of the already mentioned action fails, a Return-Individual is executed; the individual return at an already visited location can always be performed since the set ret_a contains always at least one element, the starting location of the individual.

must be assigned to distinct physical locations, the mapping between the abstract locations in d and the real locations in L must be injective. The injectivity of the mapping can not always be guaranteed: the location selection can fail due to the three cases presented above. With the use of the mobility diary, also the Individual-Return action can fail, since the agent can not even visit its current location.
In the action correction phase (Figure 3.4), if a social choice can not be completed, the next action executed is the action with the same mechanism performed without the influence of social contacts. In the case an individual action can not be performed, the complementary individual action is performed. If even the complementary individual action fails, the agent returns to the home location.. When an agent returns to the home location due to action failure during the assignment of the abstract location $ab_j \in d$, the run d is splitted in $d1 = \langle ab_1, \dots, ab_j \rangle$, $d2 = \langle ab_{j+1}, \dots, ab_n \rangle$ and the agent start the mapping of the new run $d2$.

3.7 Implementation

We implement GeoSim and the presented extension in `GeoSim`, a `Python` open source algorithm for generating realistic mobility trajectories [1]. `GeoSim` is build using the mobility structure of `scikit-mobility` [18].

[1] The Python source code of `GeoSim` is freely available for download at https://github.com/kdd-lab/2019_Cornacchia

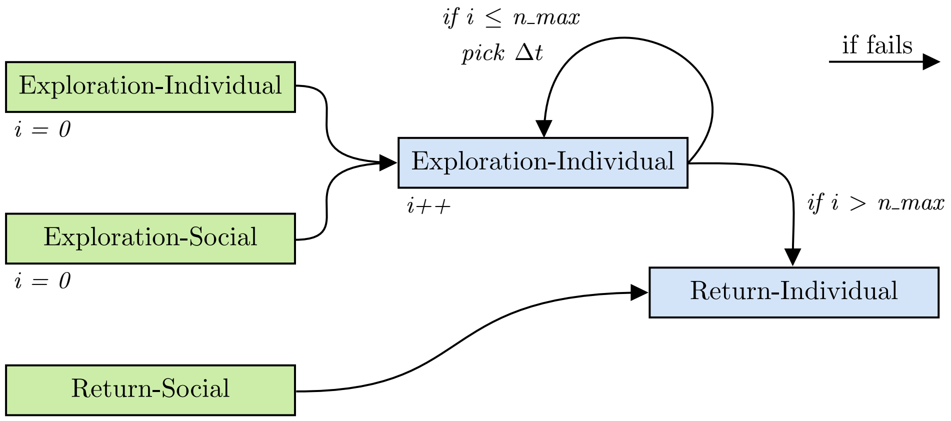


Figure 3.3: In the action correction routine designed for handling cases where the agent can not reach any of the possible locations, if the individual was performing an exploration action a new waiting time, greater than the current one, is picked; if it still can not reach any location the latter procedure is repeated at most n_{max} times. If the agent was performing a Return-Social action or after n_{max} increments, the action is corrected with a Return-Individual; Return-Individual can always be performed since the agent can return in its current location r_j from the moment that $dist(r_j, r_j) = 0$.

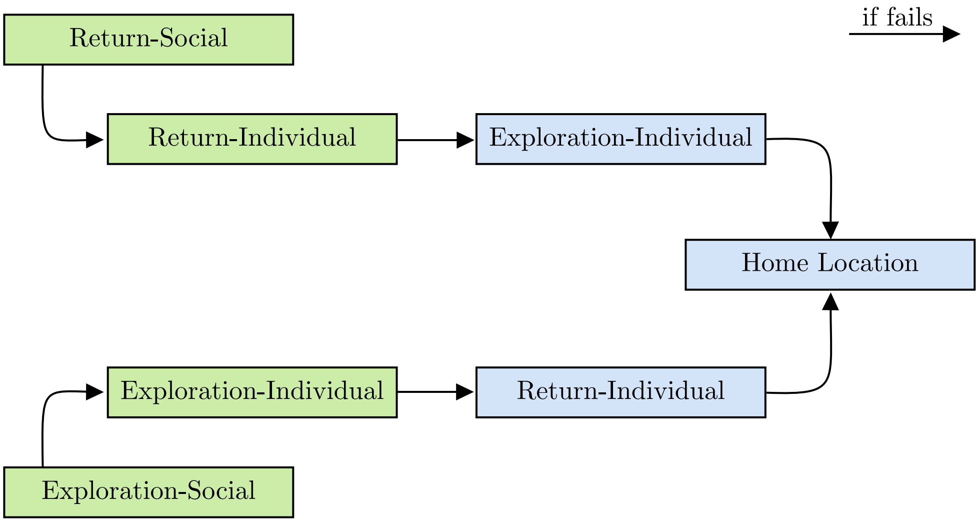


Figure 3.4: In the action correction routine, in order to guarantee the injective mapping of the abstract location visited during a run, if the agent was executing a social choice the action is corrected in an individual one, preserving the mechanism selected. If an individual action can not be performed the complementary one is executed, if it fails then the agent returns home.

parameter	`rho`	`gamma`	`alpha`	`beta`	`tau`	`min_wt_hours`
value	0.6	0.21	0.2	0.8	17	1

Table 3.1: The default values of the distributions' parameters.

Parameters

The main class, `GeoSim`, takes several parameters in input; they can be classified into four categories:

- **distributions**: The parameters in this category defines the constants used in the model. The constants ρ and γ affect the probability for an individual to return or explore, α corresponds at the social factor, namely the probability of being influenced by a social contact in the choice of the next displacement, τ and min_wt_hours are the constants used in the waiting time distribution. The default values for the constants are reported in Table 3.1.

- **simulation**: The parameters in this category specifies the number `n_agents` of agents to simulate, the extremes of the time interval to simulate, respectively `start_date` and `end_date`; the parameter `social_graph` if specified is interpreted as the edge list of the social graph and the weighted spatial tessellation is specified in the parameter `spatial_tessellation`.

- **model**: The combinations of the values of the parameters in this group define which one of the presented generative models will be executed (Table 3.2). The boolean parameter `distance` defines if the models will use the distance as a factor in the location selection, `gravity` indicates if the model will use the gravity law in the selection of the destination of the next displacement of an individual, and the parameter `diary_generator`, if specified, denotes a `MarkovDiaryGenerator` object which represents the mobility diary generator, the model will use it generate the mobility diary for each agent.

- **additional features**: The parameters in this category control the additional features used in the execution; the boolean parameter `rsl` defines whether or not agent are assigned to a starting location with probability proportional to the relevance of a location or uniformly at random, the parameter `max_speed_km_h`, if specified, indicates the typical speed $speed_{agent}$ of the agents, the boolean parameter `degree_exp_social` defines if during the social exploration the is chosen proportionally to its degree or not.

models	`distance`	`gravity`	`diary_generator`
GeoSim	False	False	None
GS$_d$	True	False	None
GS$_{gravity}$	True	True	None
GS$_{diary}$	True	True	mobility_diary

Table 3.2: The different combinations of the models' parameter for each presented model.

Data Structures

The most important data structures used for modelling efficiently the different aspects of the human mobility are the following:

- **Mobility Pattern**: The visitation pattern lv_a of an agent a is modeled as a vector of $|L|$ elements, where $|L|$ it the total number of locations.

- **Distance Matrix**: The distance matrix D is a square matrix of dimension $|L| \times |L|$ where $|L|$ is the total number of locations. Each entry $d_{ij} \in D$ contains the spatial distance in km between location r_i and location r_j. For optimizing the space used by the process, the matrix is represented in memory as a sparse matrix, and it is populated only when necessary.

- **Trajectories**: the trajectories generated by the model are described by a `TrajDataFrame`, a data structure of the library scikit-mobility [18]; it contains the user identifier, the coordinates of the location and a timestamp.

- **Spatial Tessellation**: the spatial tessellation is represented through a `DataFrame`, containing for each location its identifier, the coordinates and the relevance at a collective level.

Time complexity

In the initialization phase, the agents are assigned randomly to a home location, this procedure costs $\mathcal{O}(n)$, where n is the number of agents in the simulation. In the worst case scenario, each of the n agents move every min_{wt} hours, performing $m = \frac{total_h}{min_{wt}}$ movements, where $total_h$ corresponds at the number of hours simulated; consequently the number of actions performed in the simulation is $O(n \cdot m)$, in each action selection phase the agent generates at most four random numbers, each generation costs $\mathcal{O}(1)$ and during the location selection phase the model iterates through the location vector $\mathcal{O}(L)$ in order to build the set of the locations the agent can visit. In the case all the locations are visited at least one time, all the entries in the distance matrix M are computed, this operation costs $\mathcal{O}(L^2)$. The time complexity of the proposed implementation is $\approx \mathcal{O}(n + 4nm + Lnm + L^2)$.

Example of execution

The following code shows how to simulate the mobility of 50 agents connected in a social graph given a weighted spatial tessellation during an observation period of three months. The choice of the parameters instantiate the model $\text{GS}_{gravity}$, as shown in Table 3.2.

```
import GeoSim as gs

#load a weighted spatial tessellation
tex = pickle.load(open("tessellation_250m.pickle","rb"))

#load a social graph of 50 individuals (edge list)
graph = pickle.load(open("social_graph_50_agents.pickle","rb"))

# define the period of the simulation
start_date = pandas.to_datetime("2012/04/10␣00:00:00")
end_date = pandas.to_datetime("2012/07/10␣00:00:00")

#create a new istance of GeoSim
geosim = gs.GeoSim()

#generate the synthetic trajectories with the specified parameters
tdf = geosim.generate(start_date, end_date, n_agents = 50,
                    spatial_tessellation = tex, rsl = True,
                    distance = True, gravity = True,
                    social_graph = graph)
```

The output of the method `geosim.generate` is a `TrajDataFrame` that contains the simulated displacements made by the agents. Each record in the `TrajDataFrame` contains a user identifier, a timestamp corresponding at the time of the displacement and the coordinates of the visited location; an example of the output generated by the model is shown in Table 3.3. The output in sorted in ascending order by `uid` and `timestamp`.

	uid	lat	lng	timestamp
0	19	40.734004	-73.988473	2012-04-10 00:00:00.000000
1	19	40.734004	-73.988473	2012-04-10 03:29:57.112640
2	19	40.734004	-73.988473	2012-04-10 08:34:19.532325
3	19	40.713580	-73.950294	2012-04-10 21:38:10.585138
4	19	40.734004	-73.988473	2012-04-11 01:21:49.630914
⋮	⋮	⋮	⋮	⋮

Table 3.3: An example of a `TrajDataFrame` structure containing the generated trajectories. Each record contains a user identifier `uid`, the timestamp corresponding at the synthetic displacement `timestamp` and the coordinates of the visited location `lat` and `lng`.

The class `GeoSim` uses the structure of the library *scikit-mobility* and could be easily merged in that framework. The constructor method `GeoSim()` instantiate a new object; it takes as parameters the distribution's parameters (the default values are shown in Table 3.1). The method `generate()` takes in input the parameters in the categories simulation, model and additional features; it outputs a `TrajDataFrame` containing the generated trajectories of the agents. Through a `GeoSim` object is possible to access at the attribute `social_graph` containing a representation of the social graph used in the simulation, and at the attribute `agents` containing the information of the agents (e.g., location vector, starting position and number of distinct locations visited).

Chapter 4

Results

In this chapter, we show the results of the experiments that simulate the mobility of 1,001 agents in the urban area of New York City, for an observation period of three months. We compare the synthetic trajectories obtained with the trajectories of the real individuals moving in the same city for the same number of months. The similarity between the two sets of trajectories concerns a set of measures that characterize the patterns of human mobility.

The measures we consider are defined formally in Section 4.1 and the scores used for estimating their similarity are presented in Section 4.3.2. In Section 4.2, we present the dataset and the filtering process applied to obtain the trajectories of 1,001 real individuals in New York City; then we describe the distributions of the resulting dataset, the obtained social graph and the weighted spatial tessellation used in the experiments. In Section 4.3.1, we describe the experimental settings, in Section 4.3.3 we discuss the obtained results and in Section 4.3.4 we report the scores of the most significant experiment for each measure.

4.1 Measures

The socio-mobility measures considered in this thesis can be classified along the spatial, temporal and social dimensions [7]. All the measures, except the social one, are computed through the *scikit-mobility*[1] library [18].

Jump Length

A key factor in modeling human mobility is the distance an individual travels in an amount of time. Given a trajectory, the jump length Δr is defined as the geographical distance between two consecutive locations visited by an individual u, more formally:

$$\Delta r = dist(r_i, r_{i+1}) \tag{4.1}$$

[1] https://github.com/scikit-mobility

where r_i and r_{i+1} are two consecutive spatial points in the trajectory of u and $dist$ is the distance on the spherical earth between two points.

Radius of Gyration

The radius of gyration $r_g(u)$ describes the typical distance traveled by an individual during the period of observation. It characterizes the spatial spread of the locations visited by the individual u from the locations' center of mass r_{cm}.

$$r_g(u) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} dist(r_i(u), r_{cm}(u))^2} \tag{4.2}$$

where N is the number of locations in the trajectory of the individual u and the center of mass $r_{cm} = \frac{1}{N}\sum_{i=1}^{N} r_i$.

Visits per Location

A useful measure to understand how individuals move in a physical space is the number of visits per location. This quantity describes the relevance of a location, namely the attractiveness at a collective level.

Location Frequency

Humans exhibit a strong tendency to return to locations they visited before [4]. The location frequency $f(r_i)$ measures the probability of visiting a location r_i.

$$f(r_i) = \frac{n(r_i)}{n_u} \tag{4.3}$$

where $n(r_i)$ is the number of visits to location r_i and n_u is the total number of points in the trajectory of the individual u.
One method to describe the importance of a location for an individual u is the concept of location's rank; a location r_i has rank k if it is the *k-th* most visited location by an individual u.

Waiting Time

The waiting time Δ_t is defined as the elapsed time between two consecutive points in the mobility trajectory of an individual u, or equivalently as the time spent in a location.

$$\Delta_t = t_{i+1} - t_i \tag{4.4}$$

Uncorrelated Entropy

The uncorrelated entropy gives an estimation of the *predictability* of the movements of an individual u [12].

$$E_{unc}(u) = -\sum_{i=1}^{N_u} p_u(i)\, log_2(p_u(i)) \tag{4.5}$$

where N_u is the number of distinct locations visited by u and $p_u(i)$ is the historical probability that location i was visited by user u.

Activity per Hour

The movements of individuals aren't distributed uniformly during the hours of a day. Humans' actions follow a circadian rhythm [2, 7]; people tend to be stationary during the night hours while they preferentially move at specific times of the day, for example, to reach the workplace or return home.
To measure this distinctiveness of human mobility, we compute the number of movements made by the individuals at every hour of the day.

Mobility Similarity

Several studies demonstrate the correlation between human mobility and sociality [11, 14, 8, 15, 5]; the movements of friends are more similar than those of strangers, mainly because we are more likely to visit a location if a social contact explored that location before. Furthermore, individuals with a similar visitation pattern are more likely to establish a social link.
We define the mobility similarity mob_{sim} between two individuals u_i, u_j as the cosine-similarity of their location vectors lv_i, lv_j.

$$mob_{sim}(u_i, u_j) = \frac{lv_i \cdot lv_j}{||lv_i||\,||lv_j||} \tag{4.6}$$

4.2 Dataset

The mobility dataset used for the evaluation of the generated trajectories, is obtained from an LBSN (Location Based Social Network) dataset collected by Yang et al. [5]; it contains a set of global-scale check-ins gathered from the social network Foursquare over 22 months (from April 2012 to January 2014). A check-in made by a user describes the individual's real-time position with its social contacts. In Foursquare, the check-ins made by a user are not publicly available; despite this, many users share their check-ins on Twitter to make them public. The authors of the dataset, collected the Foursquare check-ins from Twitter by searching the Foursquare hashtag [5]. The dataset is associated with a lookup dataset for the locations, and with a snapshot of the social network obtained from Twitter, antecedent at the collection period. In the next subsection, we explain the filtering process performed to obtain the set of trajectories of 1,001 individuals regarding the urban area of New York City, during an observation period of three months, then we describe the social graph and the weighted spatial tessellations used during the experiments.

4.2.1 Mobility Dataset of New York City

The LBSN dataset D_{FS} [5], contains 90,048,627 check-ins made by 2,733,324 users all around the globe. The attributes of D_{FS} are an anonymized user identifier, an identifier of the location where the user made the check-in, the UTC (Coordinated Universal Time) when the check-in occurred and the location's timezone offset (Table 4.1). It is also included a lookup dataset D_{loc} that associates the location's identifier with the respective coordinates and other information. LBSN datasets allow to reconstruct the mobility of an individual considering the check-ins as points in the individual's trajectories.

(a)

user_id	location_id	UTC time	timezone
⋮	⋮	⋮	⋮
268846	42872fd9b60caeb	Tue Apr 03 18:27:37 2012	-240
377500	3c38c65be1b8c04	Tue Apr 03 18:27:38 2012	-240
248657	1855f964a520be3	Tue Apr 03 18:27:38 2012	-240
⋮	⋮	⋮	⋮

(b)

location_id	latitude	longitude	category	cc
⋮	⋮	⋮	⋮	⋮
42872fd9b60caeb	41.660393	-83.615227	College Cafeteria	US
6200f964a520ee3	40.722206	-73.981720	Theater	US
9cadf964a521fe3	44.972814	-93.235313	Student Center	US
⋮	⋮	⋮	⋮	⋮

Table 4.1: An example of records for the dataset D_{FS} (a) and the lookup dataset D_{loc} (b), In D_{loc} the `location_id` is associated with the coordinates, the category and the country code.

Dataset creation

To extract the mobility dataset $D_{NYC} \subset D_{FS}$ relative at the area of New York City, for comparing the synthetic trajectories during the evaluations of the models, we performed a join between D_{FS} and D_{loc} on the attribute `location_id`, obtaining all the displacements made by the individuals in New York City, associated with the relative coordinates. Before performing the join, we apply some filters to obtain only the check-ins in NYC and to reduce the number of records in D_{FS} and D_{loc}, avoiding a computationally expensive operation. The resulting dataset D_{FS_loc} is composed of 925,289 check-ins relative to 80,146 users. At this point, after converting the UTC time in the time of New York City, we remove all the users that are not included in the snapshot of the social graph G scraped from Twitter. Of the 80,146 users only 8,452 appear in G (10.5%). In the next filter operation, we take only the check-ins of the 8,452 users performed during a period of three months, from April 2012 to July 2012, in this period of observation the check-ins made by the filtered users are 80,032. Then, we substitute the *fast check-ins*, defined as a set of check-ins performed by an individual such that the time difference between them is less or equal than $t = 7s$, with a single check-in where the coordinates and timestamp are computed as the average of the respective attributes for the *fast check-ins*. Then, we select only the users with mobility (at least two check-ins) and the users who appear in at least one edge with another of the filtered users. After the latter filtering operations, the users left are 1,780. We removed the users that are not in the main component of the social graph G (considering only the edges between the 1,780 users). The final dataset, D_{NYC}, contains 37,489 check-ins made by 1,001 connected users during an observation period of three months (April 2012 to July 2012).

Data distributions

We analyzed the probability density functions (PDF) of the measures presented in Section 4.1, to check whether or not the obtained trajectories $\in D_{NYC}$ present the significant analytical proprieties of individuals' displacement.

The distribution of the number of check-ins per user is heavy-tailed (Figure 4.2) This behavior is typical of the LBSN datasets and was also observed by the authors of the original dataset [5]. The movements of the individuals in New York City are not highly predictable, as attested by the uncorrelated entropy measure (Figure 4.2).

The jump length confirms the tendency of individuals to move at small rather than long distances [6, 4], as we can see from Figure 4.3 individuals in the area of NYC move rarely at distances greater than ≈ 22km. The distribution of the radius of gyration (Figure 4.3) shows that the typical spatial spread of individuals' displacements is likely to be included between 1 and 7 km.

The probability for an individual to visit location of rank i (Figure 4.4), namely the location frequency, follows a distinctive distribution of this measure, the Zipf law [4]. The number of visits per each location, which correspond to the relevance of a tessellation, results in a power law distribution (Figure 4.4); most locations have a few visits while only rare locations receive a significant number of visits [2]. Figure 4.1 shows an heatmap of the check-ins $\in D_{NYC}$, that can be considered as a continuous and non-aggregated form of relevance.

The distribution of the time spent in a location, for the 1,001 individuals, during the observation period of three months follows a power law [2, 7] (Figure 4.5). Humans' actions follow a circadian rhythm [2, 7]: the activity per hour measures, depicts the non-uniform distribution of the movements of individuals during the hours of a day (Figure 4.5).

We compute the mobility similarity for the users connected in the real social graph G and in a random one with the same number of nodes and edges. As we can see from Figure 4.6, the mobility similarity within users connected in the social graph is generally higher than the ones of random pairs of users. This result confirms the correlation between human mobility and sociality, the movements of friends are more similar than those of strangers [11, 14, 8, 15].

4.2.2 Social Graph

The social graph G, a snapshot relative to March 2012 of the social connections among a subset of the users in D_{FS}, obtained from Twitter, is composed of 114,324 nodes and 363,704 edges. During the data preprocessing, G was filtered several times, obtaining a new graph G_{NYC} (Figure 4.7); it is an undirected and connected graph, with 1,001 nodes which represent the users and 1,755 edges which represent the social connections between users. The statistical properties of the graph follow the well-know significant properties of social graphs; the node degree distribution follows a power-law and the average path length is 5.154 (≈ 6 in social graphs, according to Karamshuk et al. [7]). The density and average node degree are respectively $4 \cdot 10^{-3}$ and 3.506.

4.2.3 Weighted Spatial Tessellation

To partition the area of New York City into a discrete number of relevant and *non-overlapped* locations, we used a weighted spatial *squared* tessellation L. A first consideration is that some of the locations are completely included in the water area of New York City and, consequently, they are unreachable for the agents in our models. We exclude these locations obtaining a new tessellation $L_{\text{LAND}} \subseteq L$, since we consider only displacements within land locations. The relevance w_i of each location $r_i \in L_{\text{LAND}}$ is computed as the total number of check-ins in D_{NYC} made in that location by the individuals; we assigned a default relevance of 0.1 at the locations without any associated check-in. We compute another subset of locations, $L_{\text{REL}} \subseteq L_{\text{LAND}}$ defined as follows:

$$L_{\text{REL}} = \{r_i \in L_{\text{LAND}} \mid w_i \geq 1\} \tag{4.7}$$

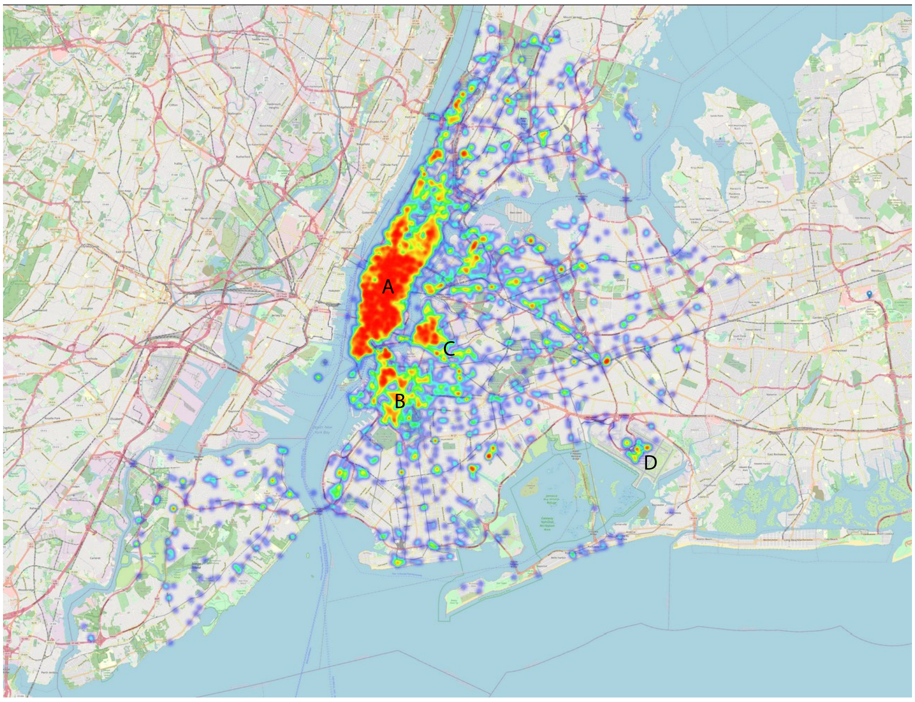
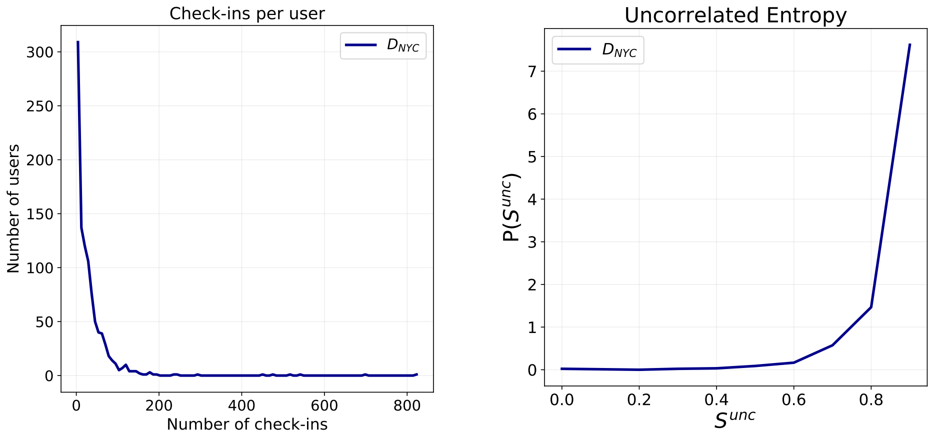
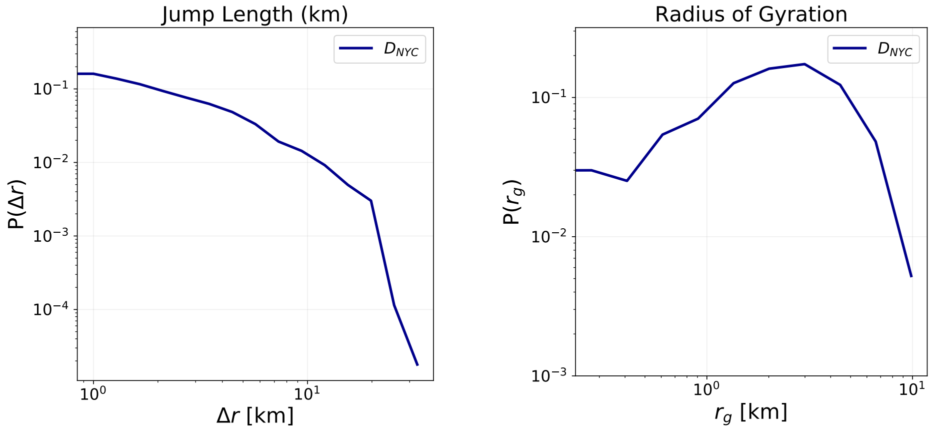


Figure 4.1: The heatmap relative at the 37,489 check-ins made by 1,001 individuals during an observation period of three months (April 2012 to July 2012) in New York City. There is a high concentration of check-ins in the borough of Manhattan (A) and in its surroundings (upper part of Brooklyn (B) and of Queens (C)); this high concentration of check-ins in that area can be explained mainly because Manhattan is the most densely populated of the five boroughs of New York City; another reason is that Manhattan is the touristic center of New York City, it contains attractive locations such as Times Square, Central Park, the Empire State Build, Statue of Liberty, Wall Street, One World Trade Center, and many others. Another area of dense check-ins is the one that is associated to the JFK airport (D). The distribution of the check-ins in the physical space can be considered as a continuous and non-aggregated form of relevance, from the moment that the relevance of a location is computed as the number of check-ins made in that location.



(a) (b)

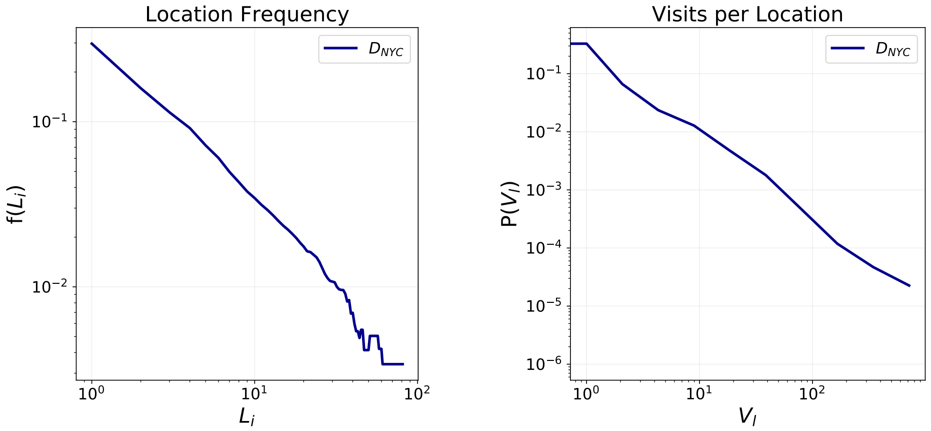
Figure 4.2: The distributions of the check-ins per user (a) and uncorrelated entropy (b) for the filtered dataset D_{NYC}.



(a) (b)

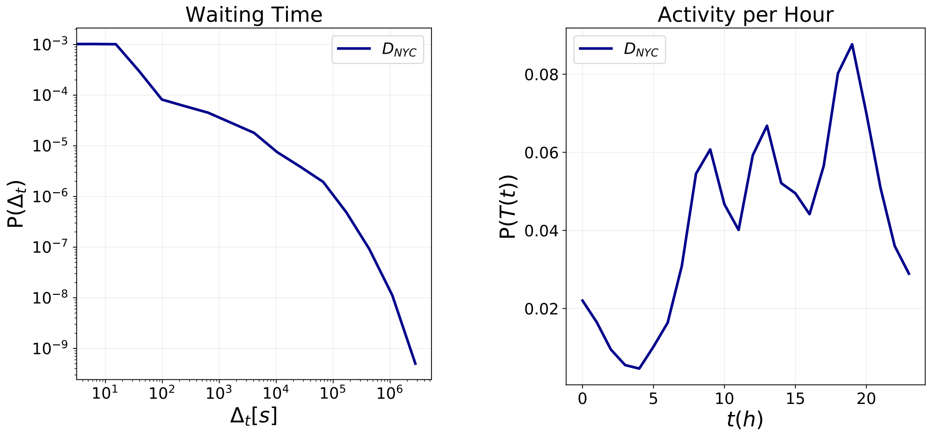
Figure 4.3: The distributions of the spatial measures jump length (a) and radius of gyration (b) for the dataset D_{NYC}.

We build the tessellations L_{LAND} and L_{REL} for different levels of granularity, we select the side s, in meters, of the squared tiles from the set $\{250, 500, 750, 1000, 2000\}$; a tessellation with locations of side s is referred as $L_{\text{LAND}}(s)$ or $L_{\text{REL}}(s)$. The number of locations in each tessellation is shown in Table 4.2.



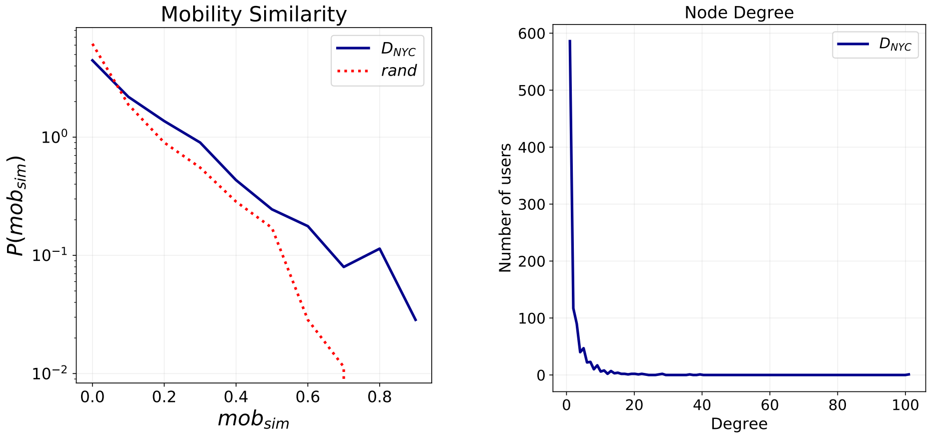
(a) (b)

Figure 4.4: The distributions of the location frequency (a) and visits per location (b) for the dataset D_{NYC} computed with a weighted squared tessellation with size 1000 meters.



(a) (b)

Figure 4.5: The distributions of the waiting time (a) and activity per hour (b) for the dataset D_{NYC}; in the plot (b) we can notice how the circadian rhythm of the individuals presents three peaks at hours corresponding to the following activities: reach the workplace (8-9 am), work lunch break (12am-1pm) and return home (6-7pm).



(a)

(b)

Figure 4.6: The probability density function of the mobility similarity (a), computed among the pair of users connected in the social graph G_{NYC} (blue solid line) and for random pairs of users not connected in the social graph (red dotted line). In the figure (b) is shown the distribution of the node degree for the social graph G_{NYC}.

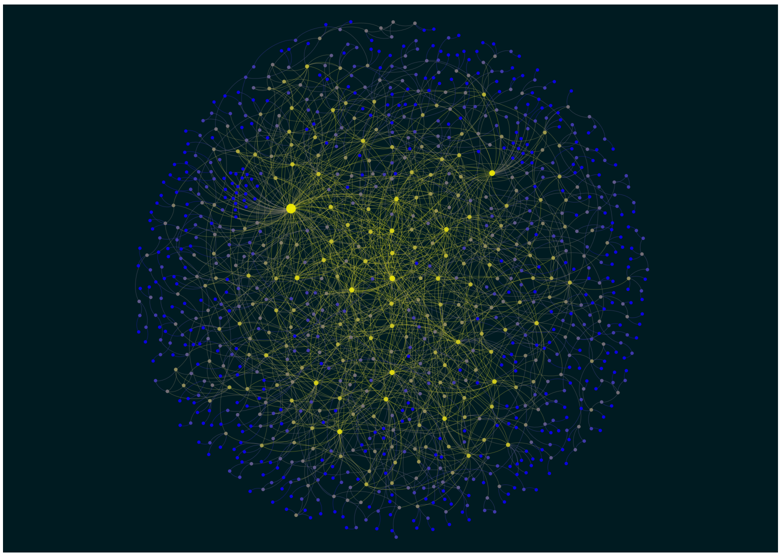


Figure 4.7: A visualization of the social graph G_{NYC}. The size of a node is proportional to the degree, as well as the color that varies from purple to yellow.

4.3 Results

4.3.1 Experiments settings

We generate the synthetic trajectories, simulating for three months the displacements of 1,001 individuals connected in the graph G_{NYC}, moving in

tile size	$\lvert L\rvert$	$\lvert L_{\text{LAND}}\rvert$	$\lvert L_{\text{REL}}\rvert$
250 m	34,408	23,740	2,893
500 m	8,745	6,285	1,604
750 m	3,951	2,918	1,082
1000 m	2,256	1,709	800
2000 m	596	475	333

Table 4.2: The number of locations for each tile size for the tessellation L, L_{LAND} and L_{REL}.

the urban area of New York City, represented through the tessellation L_{REL} presented in Section 4.2.3. We use the baseline model, GeoSim, and the three proposed extensions: GS$_d$, GS$_{gravity}$ and GS$_{diary}$. Each of the proposed extensions are instantiate with the additional features RSL and the action correction phase [2]. The Markov model, relative to the mobility diary generator, is trained on the displacements of the individuals included in the dataset D_{NYC}.

For each model we use the weighted spatial tessellations L_{REL} for different levels of granularity, we select the side s in meters from the set $\{250, 500, 750, 1000, 2000\}$.

In the experimental phase, for each model and for each tessellation we make five executions, to collect the mean and the standard deviation for each mobility measure, resulting from the comparison of the synthetic trajectories with the trajectories of real individuals.

4.3.2 Scores

The statistical similarity between the distributions of the measures which characterizes human mobility, of the generated and the real trajectories, is quantified using five scores [19, 2].

- **RMSE**: The Root Mean Square Error (RMSE) between a ground truth distribution p and a synthetic distribution q is defined as:

$$\text{RMSE}(p,q) = \sqrt{\frac{\sum_{i=1}^{n}(p_i - q_i)^2}{n}} \tag{4.8}$$

where $q_i \in q$, $p_i \in p$ and the number of observations in both the distributions is n.

- **Kullback–Leibler divergence**: The Kullback–Leibler divergence (KL) between a ground truth distribution p and a synthetic distribution q quantifies how much information is lost when q is used to

[2] We tried to include also the other additional features during the experiments, namely reachable locations and social choice by degree, but they did not improve the performance of the generative models.

approximate p.

$$\mathrm{KL}(p \parallel q) = \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right) \tag{4.9}$$

- **Hellinger distance**: The Hellinger distance (H) measure the distance between two distributions p and q.

$$\mathrm{H}(p,q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{n}(\sqrt{p_i} - \sqrt{q_i})^2} \tag{4.10}$$

- **Pearson's correlation coefficient**: The Pearson's correlation coefficient (r) is a measure of the linear relationship between two set of observations p and q.

$$r_{pq} = \frac{\sum_{i=1}^{n}(p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\sum_{i=1}^{n}(p_i - \bar{p})^2}\sqrt{\sum_{i=1}^{n}(q_i - \bar{q})^2}} \tag{4.11}$$

  where $\bar{p}$ and $\bar{q}$ are the mean values of p and q respectively.

- **Spearman's Rank correlation coefficient**: The Spearman's Rank correlation coefficient (ρ) measure the monotonic relationships (linear or non-linear) while Pearson's correlation measure only linear relationships.

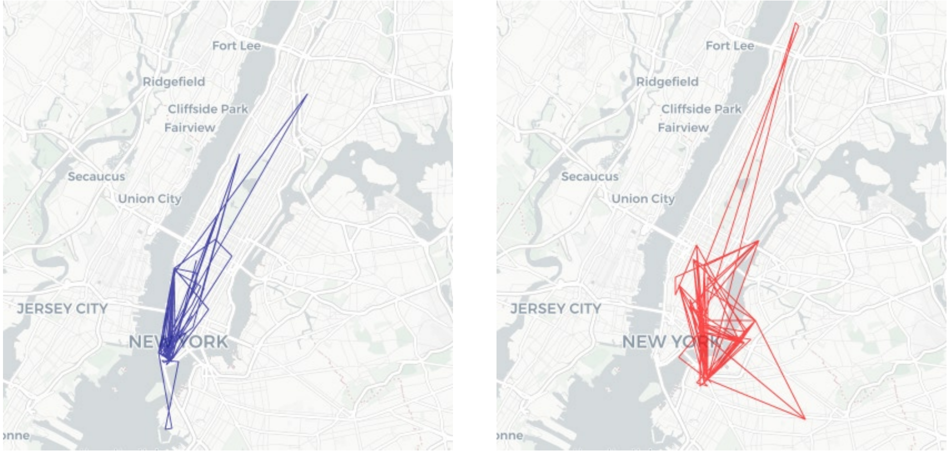
$$\rho_{pq} = 1 - 6\frac{\sum_{i=1}^{n}(r_k(p_i) - r_k(q_i))^2}{n(n^2-1)} \tag{4.12}$$

  where $r_k(p_i)$ is the *rank* of value p_i in the sorted list $(p_1, ..., p_n)$, analogously $r_k(q_i)$.

4.3.3 Discussion of results

In this section we discuss the results obtained through the simulation of the mobility for a set of individuals with the proposed methods.

The experiments show that the role of both the model mechanisms and the tessellation granularity is crucial to produce realistic trajectories. The probability density function of the spatial measures computed over the generated trajectories, shapes according to the mechanism used in the generative model. For what concerns the jump length (Figure 4.9), in GeoSim, no mechanism takes into account the spatial distance between locations, consequently, the model can not even replicate correctly the monotonicity of the distribution. In GS_d, with the introduction of a mechanism able to model the spatial distance between locations, the probability density function of the jump length follows the power-law behavior; the tendency of the individuals to move at small rather than long distances is preserved. However, taking into account only the distance is not sufficient, GS_d underestimates small-distance trips and overestimates long-distance trips. With the introduction of the gravity model, GS$_{gravity}$ generates trajectories that reproduce in a very accurate way the real distribution, although slightly overestimating small displacements. To reproduce the jump length distribution accurately, as shown in Figure 4.9, the granularity of the tessellation plays a crucial role together with the mechanism used in the generative model. With a fine-grained weighted spatial tessellation, the model generates more realistic trajectories (Figure 4.8 shows a real and a synthetic trajectory).

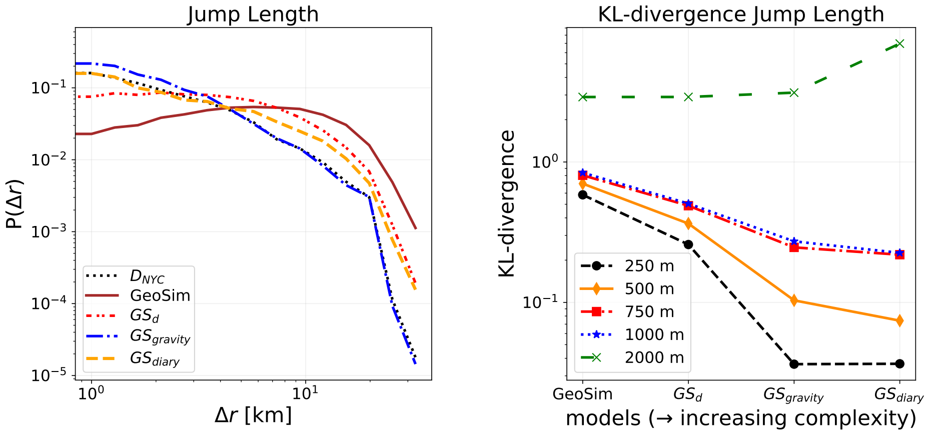


(a) (b)

Figure 4.8: A spatial representation of the trajectory of a real individual (a) and a synthetic individual (b); the latter is generated using the model GS$_{diary}$ with the weighted spatial tessellation $L_{\text{REL}}(250)$. Figures generated with *scikit-mobility* [18].

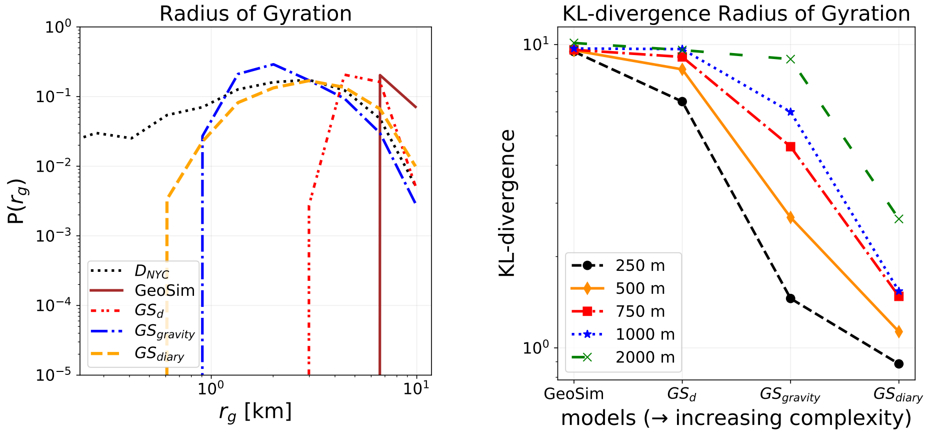
The considerations made for the jump length also hold for the radius of gyration; the typical spatial spread of the agents is mechanism and tessellation dependent (Figure 4.10). With every introduction of a more sophisticated mechanism the models generate more realistic synthetic trajectories. The model GS$_{diary}$ produces the most accurate trajectories, outperforming GS$_{gravity}$ in terms of Kullback-Leibler divergence despite both use the gravity law. This can be due to the fact that the number of check-ins per user using the diary generator is similar to the real ones, while the other models overestimate this measure, as shown in Figure 4.15(a). None of the proposed models are able to reproduce correctly the radius of gyration for values smaller than $1km$.

The location frequency distribution of the real data is better reproduced by the model GS$_{diary}$ (Figure 4.11(a)); GS$_{diary}$ underestimates the location frequency of the top ten locations visited by the individuals. For this measure the use of the fine-grained tessellation does not ensure better results;



(a) (b)

Figure 4.9: The probability density function of the jump length (a) for the real trajectories and for each of the proposed spatial mechanisms computed with the tessellation $L_{\mathrm{REL}}(250)$. The Kullback-Leibler divergence (b) for each model and for each granularity of the weighted spatial tessellation.

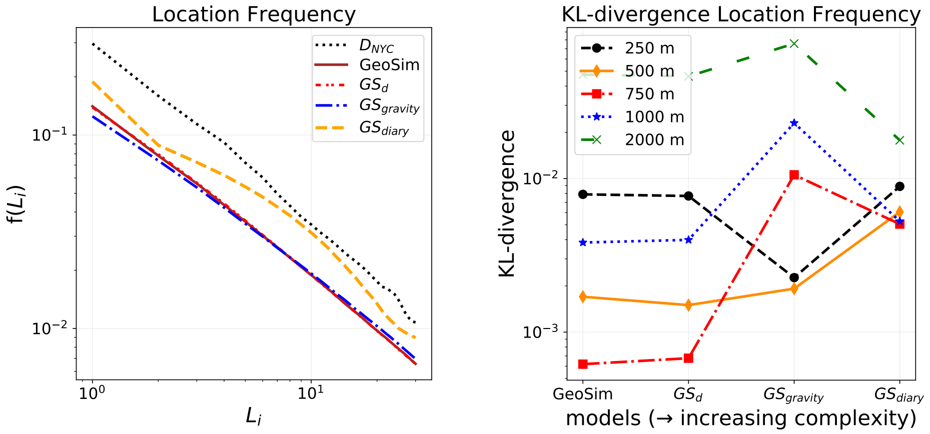


(a) (b)

Figure 4.10: None of the presented models can reproduce the radius of gyration for small values (a). The Kullback-Leibler divergence for each model and for each granularity of the weighted spatial tessellation (b).

all the weighted spatial tessellation with size $\leq 1000\ m$ produce good result in terms of KL-divergence (order of 10^{-3} for different models) as shown by the plot in Figure 4.11(b).

The distribution of the visits per location, equivalent at the relevance of each location, for the generated trajectories changes according to the indi-



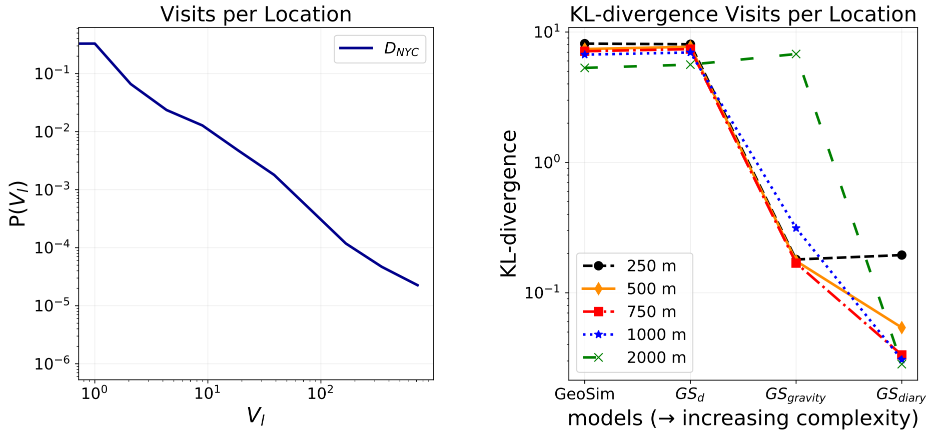
(a)

(b)

Figure 4.11: The probability density function of the top 20 locations visited by an individual (a) and the Kullback-Leibler divergence for each model and tessellation granularity; in this case a fine granularity does not ensure better result.

vidual exploration mechanism used in the model. The introduction of more sophisticated mechanisms produces trajectories with a more realistic number of visits per location. Both GeoSim and GS$_d$ underestimate the number of locations with less than 30 and 50 visits respectively (Figure 4.12(a)). With the introduction of the gravity law and the concept of relevance, the models GS$_{gravity}$ and GS$_{diary}$ replicates accurately the power-law behavior of the number of visits per location. The choice of the spatial partition is crucial for the number of visits; for the baseline model as well as GS$_d$ and GS$_{gravity}$ the tessellation used does not play a crucial role: all the tessellation with granularity $\leq 1000m$ produce similar results. In GS$_{diary}$ the results are better with a tessellation $> 250m$ because the agents perform a small number of displacement. Consequently, an aggregated number of visits considering larger locations are more similar to the real one.

The time spent in a location, namely the waiting time, has a minimum temporal resolution of one hour for the proposed model, since $min_{wt} = 1$. Since in the real distribution of the waiting time, there are values $< 1h$, to compare the synthetic and the real distributions we consider three cases: (i) we compare the distributions as we do with the other measures (Figure 4.13(a)); (ii) we cut from the real distribution the waiting times $< 1h$ (Figure 4.13(b)); and (iii) we map all the values $< 1h$ in $1h$, preserving the number of points in the distribution (Figure 4.13(c)). All the models that assign the waiting time using the empirical distribution computed by Song et al. [3] behave in the same way, while the model with use the diary generator is able to reproduce more accurately the characteristic waiting time of the set



(a)

(b)

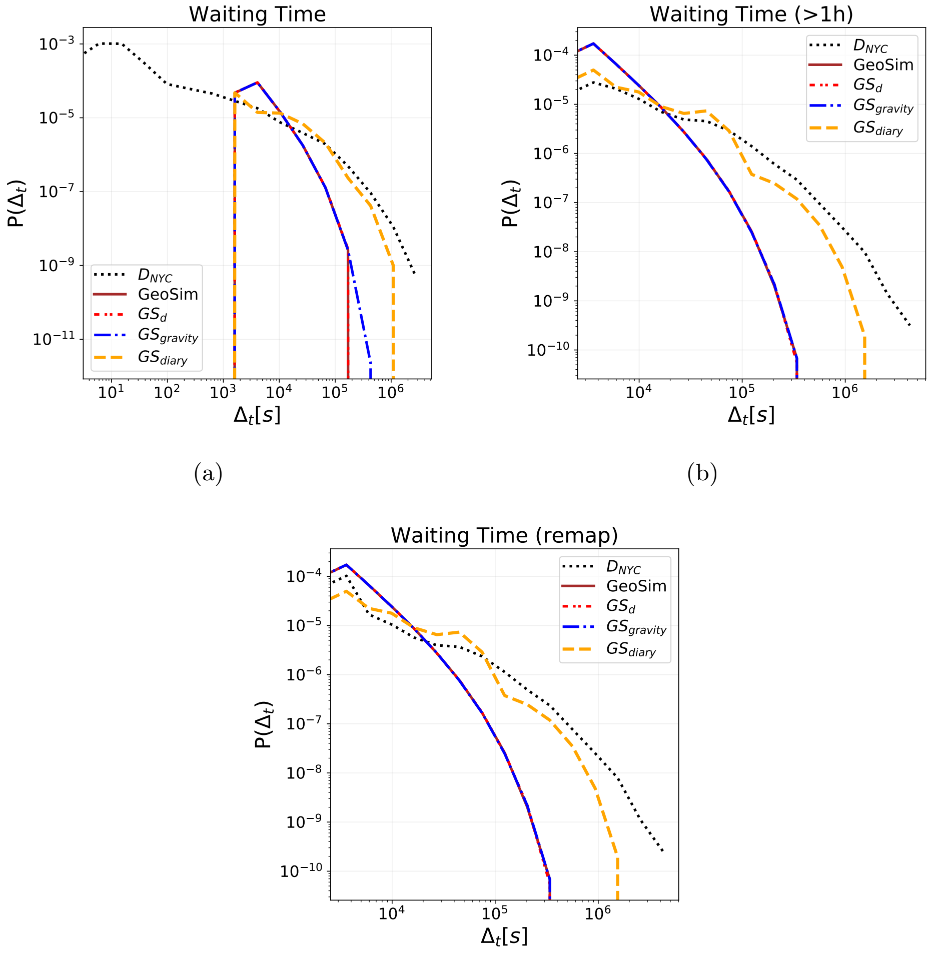
Figure 4.12: The models without the gravity law mechanism overestimate the number of visits per location (a). The tessellation does not play a crucial role except for the GS_{diary} model.

of individuals in New York City.

The number of trips made at each hour of the day, namely the activity per hour, depicts the tendency of individuals to move at a certain hour of the day. This measure is affected only by the time of the movements of individuals and neither by the mechanism on which they choose the next location to explore nor by the spatial tessellation used during the experiments. As we can see in Figure 4.14(b), the models produce the same results with every tessellation, and the only model able to reproduce the circadian rhythm of the individuals is GS_{diary}. This is caused by the fact that GS_{diary} takes into account the preference of individuals to move at specific times, while GeoSim, GS_d and $\mathrm{GS}_{gravity}$ takes into account only the waiting times, without considering the hour of the day nor the preference of an individual.

The behavior of the synthetic agents is slightly more predictable than the real counterparts; this can be caused by the fact that in the presented models, a fraction α (the social factor) of the displacements of an agent are based on the previous movements of its social contacts; this can increase the predictability of the movements of the agents. The model that reproduces in a more similar way this characteristic is GS_{diary} diary.

The probability density function of the mobility similarity changes according to the used model; as we can see from Figure 4.16, the baseline model and GS_d produces similar shapes: the users connected in the graph G_{NYC} have higher mobility similarity than random pairs of non connected users. The correlation between mobility and sociality also holds for the trajectories produced by $\mathrm{GS}_{gravity}$ and GS_{diary}; the model that reproduces the



(a)

(b)

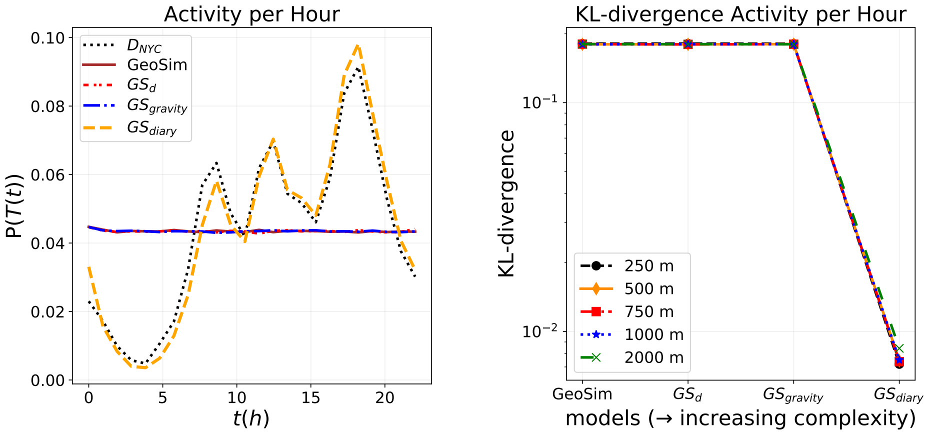
(c)

Figure 4.13: The three options considered for dealing with waiting time values under one hour; in the first case we compare the distributions with all the values (a), in the second options we cut the real distribution (b) and finally we re-map all the values $< 1h$ in the value $1h$.

mobility similarity more accurately is GS$_{diary}$ (Figure 4.16(d)).

Except for the models that does not use the gravity law, the fine-grained tessellations produce better results (Figure 4.17(b)).

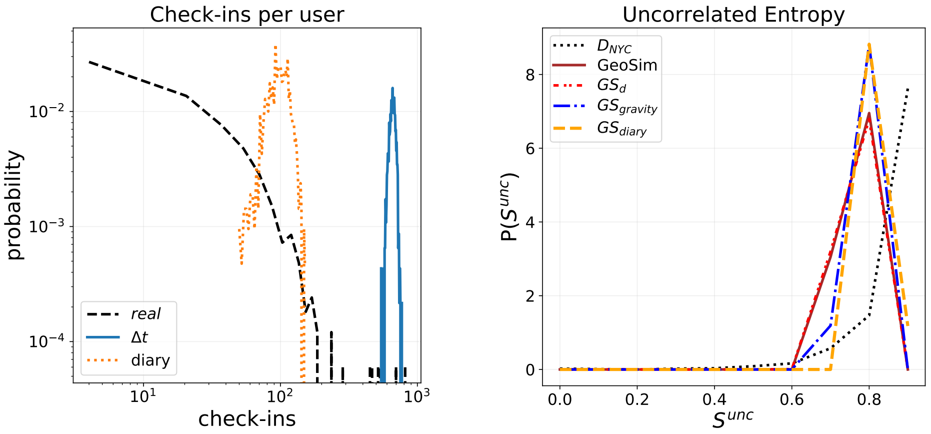
Our experiments reveal three main results. First, the proposed extension GS_{diary} produces synthetic trajectories having in general the best fit to the trajectories in the dataset D_{NYC}. Second result is that the choice of the spatial mechanism and the temporal mechanism (empirical distribution or diary generator) used for picking the waiting time is important in order to reproduce accurately some properties of the mobility trajectories; in general



(a)

(b)

Figure 4.14: The only model able to reproduce the circadian rhythm of the individuals in the dataset D_{NYC} is $\text{GS}_{gravity}$ (a); in the other models the probability of moving at a certain hour is uniform. The choice of the tessellation does not influence this temporal measure (b).

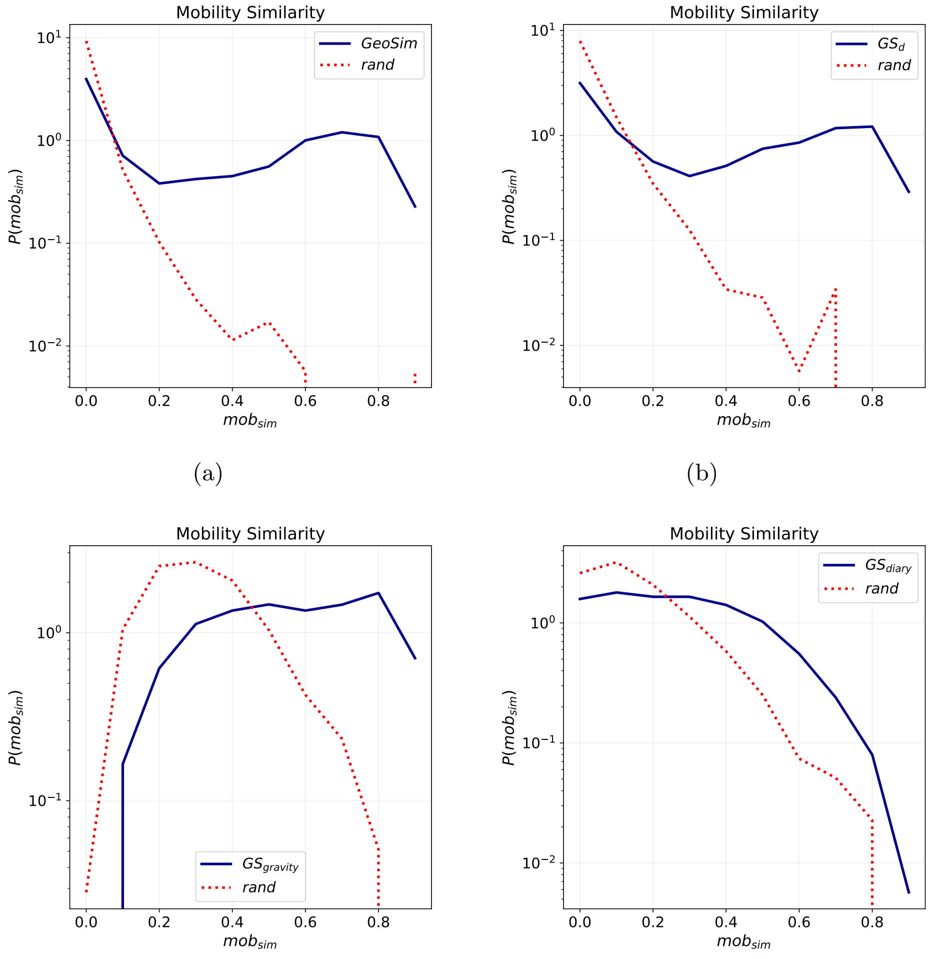


(a)

(b)

Figure 4.15: The probability density function of the number of check-ins made by the real individuals and for models that pick the waiting time from the empirical distribution of Song et al. [12] and for the model that use the diary generator (a). The probability density function of the uncorrelated entropy (b).

the gravity-law is the crucial mechanism for reproducing correctly almost all the measure. Last, the choice of the granularity of the weighted spatial tessellation depends on which measure we consider: for spatial measures (jump length and radius of gyration) as well as for the location frequency and



(a)

(b)

(c)

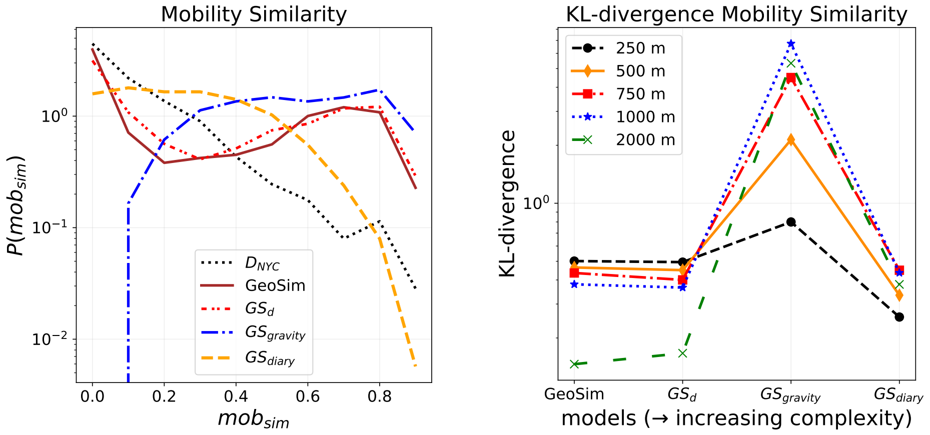
(d)

Figure 4.16: The shape of the generated mobility similarity changes according to the model used; all the generative models preserve the correlation between human mobility and sociality, the movements of friends are more similar than those of strangers

mobility similarity a fine tessellation produces better results. In contrast, for the other measures a larger tessellation produces better trajectories.

4.3.4 Tables of results

For each mobility measure we report the table of the scores obtained through the experiments, referred at the tessellation which gives the best overall result for that specific measure in terms of the five scores used to quantify the similarity between the real and the synthetic distributions. For each



(a)

(b)

Figure 4.17: The probability density function of the mobility similarity (a); generally a fine grained tessellation produces better results, in GeoSim the good score with the tessellation $L_{\mathrm{REL}}(2000)$ can be caused by the smaller dimension of the location vectors assigned at an agent, and consequently more likely to be similar at the one of its social contacts.

score, we report mean and standard deviation, the best values for each score are reported in bold.

Jump Length

The following table summarizes the scores for the jump length, using a tessellation $L_{\mathrm{REL}}(250)$. The best models are the ones that use the gravity-law mechanism in the choice of the next location to explore.

	GeoSim	GS_d	$\boldsymbol{GS_{gravity}}$	$\boldsymbol{GS_{diary}}$
RMSE	0.1044±0.0002	0.0821±0.0004	**0.0364±0.0006**	0.0371±0.001
KL	0.5838±0.0052	0.2583±0.004	**0.0363±0.0014**	0.0366±0.0016
Hellinger	0.4298±0.0009	0.2783±0.0024	0.1149±0.0032	**0.1114±0.004**
spearman	0.3412±0.0213	0.9594±0.0183	**1.0±0.0**	**1.0±0.0**
pearson	0.2934±0.0132	0.7201±0.0048	0.9366±0.0018	**0.9787±0.0031**

Table 4.3: Scores of the jump length measure.

Radius of Gyration

The scores reflect the fact that all the proposed models are not able to reproduce small radii correctly (Figure 4.10). The most accurate model is

GS_{diary}, the latter is the only model able to reduce in a significant way the Kullback-Leibler divergence and the Hellinger-distance score.

	GeoSim	GS_d	$GS_{gravity}$	$\boldsymbol{GS_{diary}}$
RMSE	0.1064±0.0004	0.0957±0.0003	0.0627±0.0065	**0.0345±0.0044**
KL	9.481±0.411	6.4871±0.0229	1.4549±0.0442	**0.8855±0.2103**
Hellinger	0.6507±0.0031	0.5645±0.0011	0.2977±0.0136	**0.2446±0.0112**
spearman	-0.3256±0.05	0.0842±0.0	0.8281±0.0	**0.8299±0.0208**
pearson	-0.2666±0.0057	0.1402±0.0122	0.8613±0.017	**0.8826±0.0361**

Table 4.4: Scores of the radius of gyration measure; the considered spatial tessellation is $L_{\text{REL}}(250)$.

Location Frequency

For this measure the tessellation considered is $L_{\text{REL}}(250)$; all the models are good in terms of Kullback-Leibler divergence, the only model able to reduce the Hellinger-distance is GS_{diary}.

	GeoSim	GS_d	$GS_{gravity}$	$\boldsymbol{GS_{diary}}$
RMSE	0.0254±0.0002	0.0256±0.0002	0.0279±0.0001	**0.0139±0.0002**
KL	0.0079±0.0005	0.0077±0.0005	**0.0023±0.0001**	0.0089±0.0001
Hellinger	0.219±0.0005	0.2195±0.0003	0.2232±0.0002	**0.1175±0.0005**
spearman	**1.0±0.0**	**1.0±0.0**	**1.0±0.0**	**1.0±0.0**
pearson	0.9991±0.0002	**0.9993±0.0002**	0.9992±0.0003	0.9943±0.0003

Table 4.5: Scores for the location frequency measure, the spatial tessellation considered is $L_{\text{REL}}(250)$.

Visits per Location

From the scores (Table 4.6) emerges the overestimation of both GeoSim and GS_d (Kullback-Leibler of 7.7095 and 6.9719 respectively) in the number of visits per location. The introduction of the concept of relevance and the mechanism of the gravity-law give the best scores.

	GeoSim	GS_d	$GS_{gravity}$	$\boldsymbol{GS_{diary}}$
RMSE	0.1013±0.0	0.1013±0.0	0.0869±0.0008	**0.0464±0.0025**
KL	6.7095±0.0067	6.9719±0.0137	0.3135±0.024	**0.0307±0.0052**
Hellinger	0.4674±0.0	0.4674±0.0	0.2696±0.0047	**0.1187±0.0055**
spearman	-0.5267±0.0286	-0.5965±0.0196	**1.0±0.0**	0.9982±0.0036
pearson	-0.1955±0.0011	-0.2276±0.0066	0.8856±0.0166	**0.9976±0.0014**

Table 4.6: Scores for the visits per location measure, the spatial tessellation considered is $L_{\text{REL}}(1000)$.

Waiting Time(s)

From the three tables below we can see how the scores changes according to the distribution of the waiting time considered; considering also values $1 < h$ (Table 4.7) no model is able to reproduce correctly the real distribution, cutting from the real distribution all the waiting times $1 < h$ (Table 4.8) or remapping to an hour (Table 4.9) produce better scores; the model that best fits this measure is GS_{diary}.

	GeoSim	GS_d	$GS_{gravity}$	$\boldsymbol{GS_{diary}}$
RMSE	**0.0004±0.0**	**0.0004±0.0**	**0.0004±0.0**	**0.0004±0.0**
KL	3.6461±0.0008	3.6405±0.0008	3.642±0.0025	**2.9627±0.0057**
Hellinger	0.0344±0.0	0.0344±0.0	0.0344±0.0	**0.0341±0.0**
spearman	0.1588±0.0	0.2658±0.0	0.2444±0.0428	**0.322±0.0655**
pearson	-0.1667±0.0001	**-0.1653±0.0001**	-0.1655±0.0006	-0.1779±0.0022

Table 4.7: Scores for the waiting time distribution.

	GeoSim	GS_d	$GS_{gravity}$	$\boldsymbol{GS_{diary}}$
RMSE	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**
KL	0.5619±0.0011	0.5609±0.0003	0.5644±0.0027	**0.0423±0.0031**
Hellinger	0.0065±0.0	0.0065±0.0	0.0065±0.0	**0.0014±0.0001**
spearman	0.5179±0.0	0.5179±0.0	0.5254±0.0149	**0.8851±0.0281**
pearson	0.9154±0.0001	0.9154±0.0002	0.9153±0.0003	**0.9769±0.0009**

Table 4.8: Scores for the waiting time distribution where are considered only values ≥ 1 hour.

	GeoSim	GS_d	$GS_{gravity}$	$\boldsymbol{GS_{diary}}$
RMSE	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**
KL	0.1976±0.0006	**0.1972±0.0004**	0.198±0.0007	0.2045±0.0051
Hellinger	0.0043±0.0	0.0043±0.0	0.0042±0.0	**0.0029±0.0**
spearman	0.5179±0.0	0.5179±0.0	0.5254±0.0149	**0.8851±0.0281**
pearson	0.975±0.0001	0.9749±0.0002	**0.9751±0.0001**	0.9178±0.0015

Table 4.9: Scores for the waiting time distribution where the values ≥ 1 hour are considered as 1 hour.

Activity per Hour

The three models that use the empirical distribution of Song et al. [12] in the waiting time choice behave the same. The only model able to reproduce the circadian rhythm is GS_{diary}.

	GeoSim	GS_d	$GS_{gravity}$	$\boldsymbol{GS_{diary}}$
RMSE	0.0234±0.0001	0.0234±0.0	0.0234±0.0	**0.0047±0.0001**
KL	0.1797±0.0007	0.1801±0.0004	0.18±0.0006	**0.0075±0.0003**
Hellinger	0.2268±0.0004	0.227±0.0002	0.2269±0.0004	**0.0431±0.0009**
spearman	-0.1078±0.2374	-0.2607±0.1109	-0.2214±0.1681	**0.9786±0.0016**
pearson	-0.2008±0.1717	-0.2962±0.0873	-0.2531±0.1352	**0.9834±0.0006**

Table 4.10: Scores for the activity per hour measure.

Uncorrelated Entropy

The best tessellation for this measure is $L_{\text{REL}}(2000)$; all the models are not able to reproduce in an accurate way the distribution of the uncorrelated entropy of the individuals in New York City.

	GeoSim	GS_d	$GS_{gravity}$	$\boldsymbol{GS_{diary}}$
RMSE	2.0942±0.0405	**2.0842±0.0207**	2.4114±0.0141	2.2734±0.0129
KL	7.3433±0.0099	7.3404±0.0048	7.4983±0.0147	**3.4449±0.2474**
Hellinger	1.7999±0.0085	1.7974±0.0042	1.9059±0.0078	**1.7716±0.0242**
spearman	0.5276±0.0	0.5276±0.0	0.5276±0.0	**0.768±0.0**
pearson	0.5544±0.0021	**0.555±0.0011**	0.5357±0.0009	0.5456±0.0014

Table 4.11: Scores for the uncorrelated entropy measure.

Mobility Similarity

The best model for what concern the distribution of the mobility similarity with respect the social graph G is GS_{diary}. In this case the use of the gravity-law with the waiting times chosen from the empirical distribution of Song et al. [12] gives the worst result; using a diary generator and the gravity law like in GS_{diary} is the best choice according to the results presented in Table 4.12.

	GeoSim	GS_d	$GS_{gravity}$	$\boldsymbol{GS_{diary}}$
RMSE	1.3609±0.0236	1.3677±0.021	1.904±0.0478	**0.9007±0.0318**
KL	0.5013±0.0085	0.495±0.007	0.7998±0.0403	**0.2568±0.0238**
Hellinger	1.2843±0.0091	1.2743±0.0092	1.5087±0.0284	**0.7958±0.018**
spearman	0.1345±0.0225	0.1879±0.0153	0.4594±0.0145	**0.9222±0.0102**
pearson	0.9432±0.0041	0.9517±0.0033	0.8638±0.0311	**0.9785±0.0028**

Table 4.12: The scores referred to the measure mobility similarity using a tessellation $L_{\text{REL}}(250)$.

4.3.5 Modeling ability in new scenarios

For assessing the modeling ability of the proposed model in other scenarios beyond New York City, we simulate the mobility for a set of individuals moving in the area of London for a period of three months. We used the model GS$_{diary}$ instantiated with a weighted spatial tessellation with granularity of 250 meters. The dataset D_{LON}, relative at the mobility of the individuals in London, is created using the same *modus operandi* as for the New York City dataset D_{NYC}; D_{LON} contains 14,895 check-ins (Figure 4.18) made by 622 users connected in a social graph G_{LON} which has 1,185 edges. The weighted spatial tessellation $L_{\text{LON}}(250)$ computed over D_{LON} contains 2,800 relevant locations.

First, we verify that the model is not *city-dependent*; we instantiate the model with full knowledge of the London scenario (diary generator MD$_{\text{LON}}$ and weighted spatial tessellation $L_{\text{LON}}(250)$); as shown in Table 4.13, the model is able to generate trajectories (Figure 4.19) with realistic mobility patterns.

Next, we simulate three cases scenario where we have partial or no information about the displacements and the circadian rhythm of the individuals in London. In the first scenario, called None, we assume to know nothing about the mobility of the individuals in London; we use the Mobility Diary Generator MD$_{\text{NYC}}$ computed for New York City, and the weighted spatial tessellation is generated assigning a relevance w_i for a location r_i from a truncated power-law $P(w) \approx (w)^{-\beta}e^{-w/\lambda}$ where $\beta = 1.25$ and $\lambda = 104$; the parameters are fitted over the distribution of the relevance of the locations

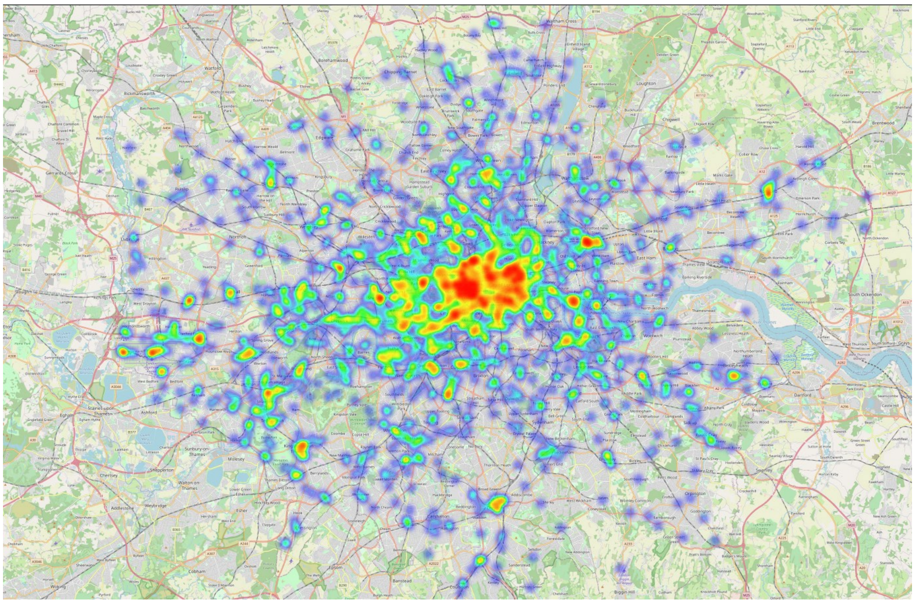
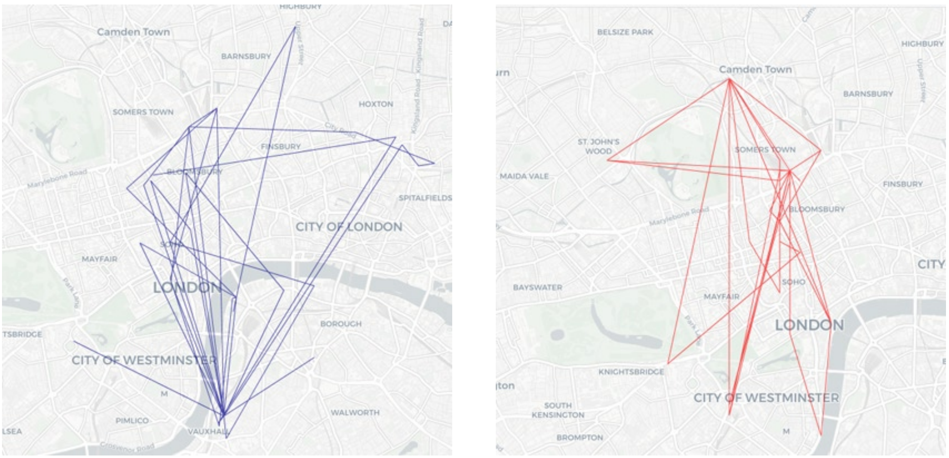


Figure 4.18: The heatmap relative at the 14,895 check-ins made by 622 individual during an observation period of three months (April 2012 to July 2012) in London. From the heatmap emerges an high concentration of check-ins in the borough of Westminster and City of London.

in New York City. In the second scenario (referred as MD), we assume to know only the routine of the individuals in London using the Mobility Diary Generator MD_{LON} and the weighted spatial tessellation used by the model is the one fitted on the New York's locations. In the last scenario (referred as L_{LON}), we assume to know only the real weighted spatial tessellation $L_{\text{LON}}(250)$ using the Mobility Diary Generator MD_{NYC}.

As shown in Table 4.13, GS_{diary} is able to reproduce in an accurate way the standard mobility measures with respect to real trajectories. We obtain results similar to the ones obtained in the experiments concerning the area of New York City. The model is able to generate realistic trajectories even with lack of information; for example, without including neither the diary generator of the individuals in London nor the weighted spatial tessellation computed over D_{LON} (first row Table 4.13) the scores are in general worst than the full-knowledge scenario, but they are good considering no information of the mobility behaviour of the London's individuals is included in the model. Including in the model, in a complementary way, the real weighted spatial tessellation and the real Mobility Diary, results in synthetic trajectories with more accurate spatial and spatio-temporal patterns respectively.

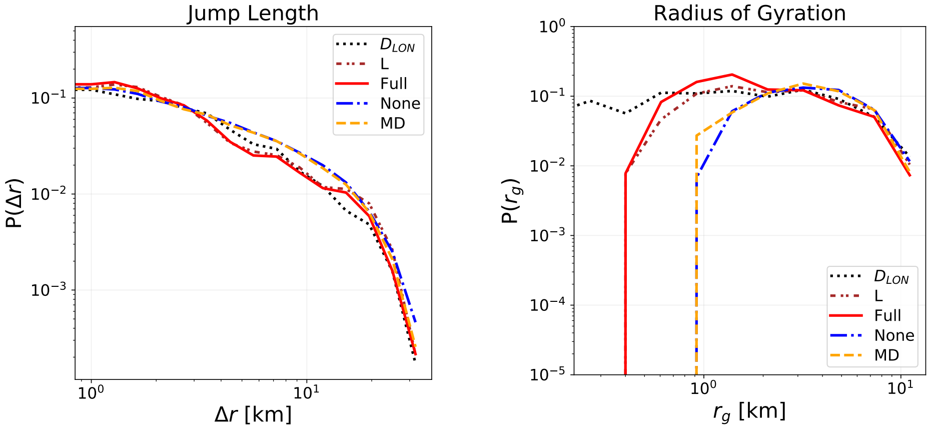
The distribution of the jump length associated with the synthetic trajectories (Figure 4.20(a)) is replicated accurately; instead the models cannot replicate the shape of the distribution of the radius of gyration (Figure



(a) (b)

Figure 4.19: A spatial representation of the trajectory of a real individual (a) and a synthetic individual (b) moving in London; the latter is generated using the model GS$_{diary}$ with the weighted spatial tessellation $L_{\mathrm{REL}}(250)$ computer over D_{LON}. Figures generated with *scikit-mobility* [18].

4.20(b)).



(a) (b)

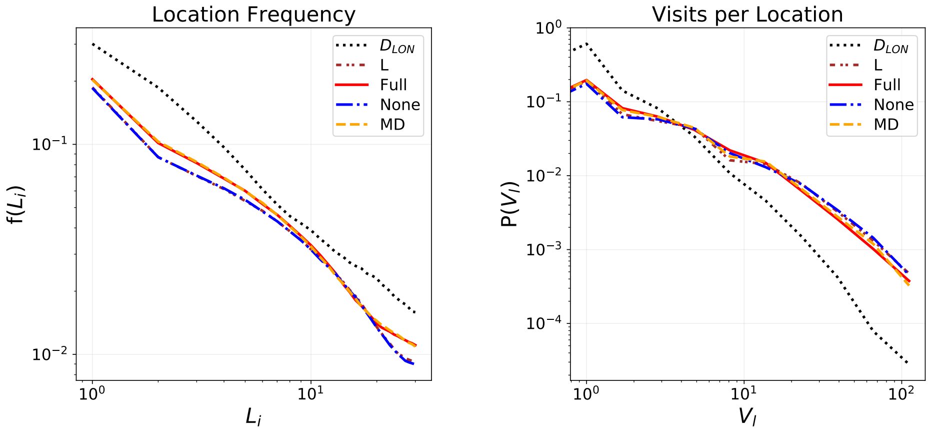
Figure 4.20: The probability density function for the jump length (a) and radius of gyration (b) computed for the real and synthetic trajectories.

Both the measures concerning the frequency and the number of visits in each location are reproduced accurately by the synthetic trajectories; the model in the experiments which uses the Mobility Diary Generator of New York City underestimates the frequency for the first ten locations (Figure

	Δr	r_g	L_i	Vl	$\Delta_t map$	$t(h)$	S^{unc}	mob_{sim}
None	0.044	3.1222	0.0152	0.1575	0.2455	0.066	2.2479	0.2699
	0.0033	0.726	0.0006	0.0112	0.0016	0.0011	0.0779	0.02
MD	0.0345	2.3151	**0.0102**	0.1503	0.1656	0.0117	1.3514	**0.2278**
	0.0021	0.5801	**0.0005**	0.0123	0.0042	0.0006	0.2133	**0.0167**
L_{LON}	0.017	1.4486	0.0153	0.1534	0.2472	0.0655	2.1396	0.2847
	0.0025	0.3124	0.0005	0.0139	0.0039	0.0008	0.2578	0.0223
Full	**0.0119**	**0.9861**	0.0106	**0.1407**	**0.1643**	**0.0113**	**1.3445**	0.2342
	0.0008	**0.4517**	0.0003	**0.0078**	**0.0045**	**0.001**	**0.2267**	0.016
NYC	0.0366	0.8855	0.0089	0.1947	0.1665	0.0072	2.1176	0.2568
	0.0016	0.2103	0.0001	0.0161	0.0054	0.0003	0.3345	0.0238

Table 4.13: Table of the results for the experiments of individuals' mobility in London. Every of the first four rows corresponds to a different scenario described above; the last row reports the results of the experiments in New York City using the tessellation granularity of 250 meters. Every column refers to a standard mobility measure. Every cell reports the mean Kullback-Leibler score (first row) and the standard deviation (second row).

4.21(a)); the same holds for the visits per location measure where is present a slightly underestimation of the number of location with a small number of visits. (Figure 4.21(b)).

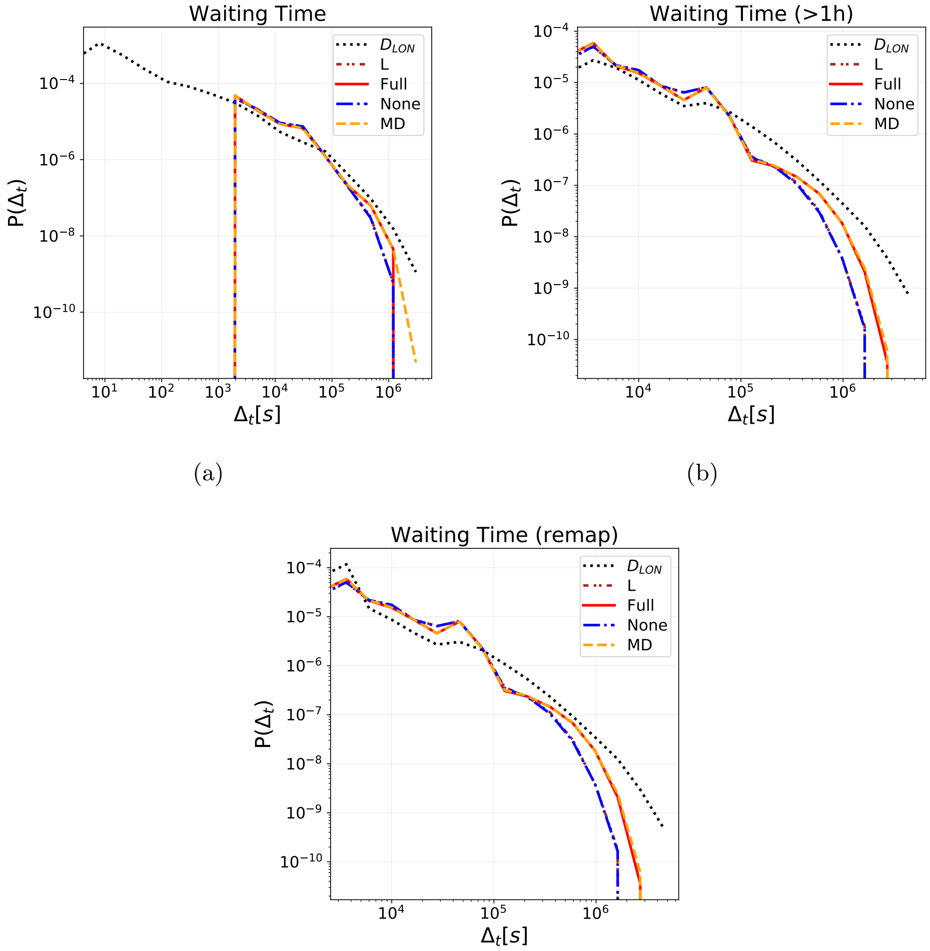


(a)

(b)

Figure 4.21: The probability density function for the location frequency (a) and visits per location (b) computed for the real and synthetic trajectories.

Figure 4.22 shows how the waiting time, in the three variants presented before, is affected by the amount of real information known by the model for values $> 10^5$.



(a)

(b)

(c)

Figure 4.22: The three options considered for dealing with waiting time values under one hour; in the first case we compare the distributions with all the values (a), in the second options we cut the real distribution (b) and finally we re-map all the values $< 1h$ in the value $1h$. The knowledge of the model affects the distribution only for values $> 10^5$.

From Figure 4.23(a) is evident the different circadian rhythm between the individuals in New York City and London. The circadian rhythm of the individuals in New York is characterized by three peaks, while for individuals in London is characterized by two peaks. This can be explained due to the different socio-cultural behaviours of the two studied populations.

The predictability of the synthetic agents is not influenced by the knowledge modeled by GS$_{diary}$ (Figure 4.23(b)).

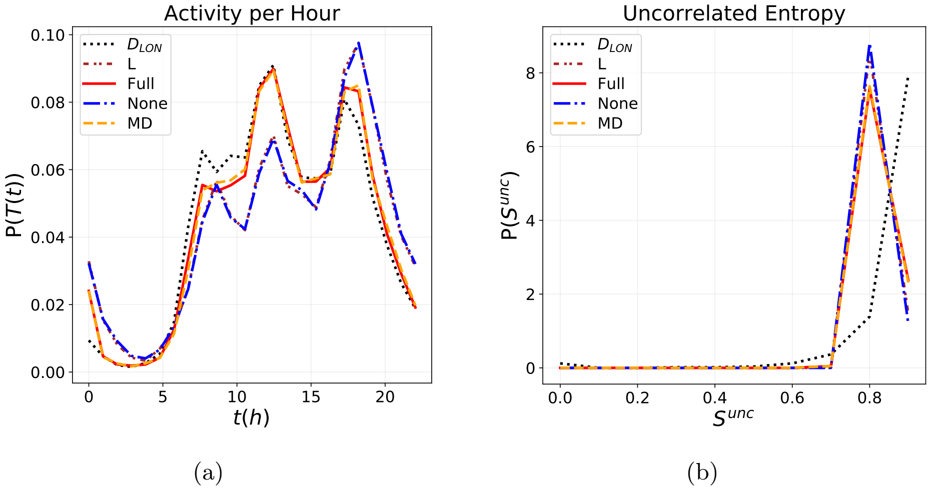
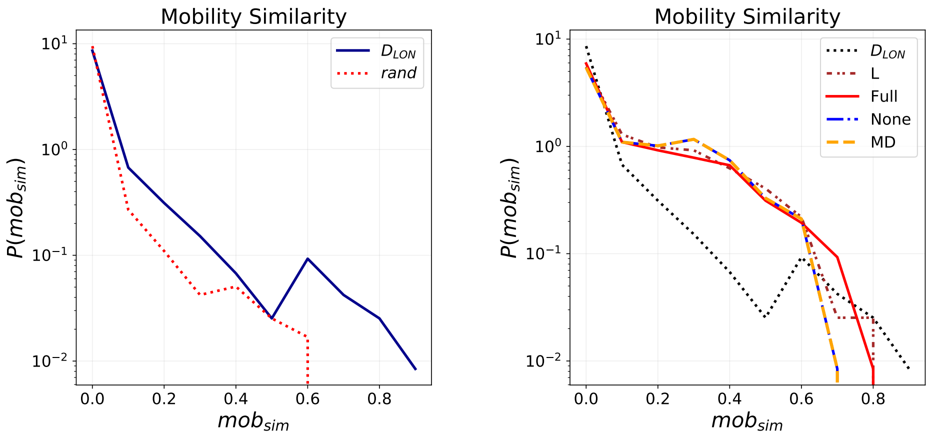


Figure 4.23: The distribution of the activity per hour measure (a) where is evident the different routine from individuals of New York City and London. Figure (b) shows that the predictability of the agents does not change significantly according to the information known.

The mobility similarity distribution between the generated trajectories of the agents, changes according to the real information included in the model (Figure 4.17(a)); without the use of the real weighted spatial tessellation the model underestimate the mobility similarity between connected users.

These experiments demonstrate and validate the applicability of the model GS$_{diary}$ in urban areas beyond New York City. It can be used in an unseen scenario using a pre-computed Mobility Diary Generator[3] and a weighted spatial tessellation with relevance picked from the empirical truncated-power law $P(w) \approx (w)^{-\beta} e^{-w/\lambda}$ where $\beta = 1.25$ and $\lambda = 104$.

[3] A pre-computed Mobility Diary Generator is available at https://github.com/kdd-lab/2019_Cornacchia



(a)

(b)

Figure 4.24: The distribution of the mobility similarity for the individuals in London considering the actual social graph and a random one (a). The distribution of the mobility similarity for the generated trajectories; without the use of the real weighted spatial tessellation the model underestimate the mobility similarity between connected users.

Chapter 5

Conclusions and Future Works

In this thesis, we develop a model that considers the social dimension together with spatial and temporal dimensions during the generation of the synthetic trajectories.

Starting from GeoSim, a state-of-the-art socio-mobility model, we include incrementally three mobility mechanisms to improve its modeling ability. In GS_d, the first proposed extension, we introduce a mechanism that takes into account the distance from the current location and the location to explore. In the second extension $GS_{gravity}$, we take into account the relevance of a location together with the distance from the current location using a gravity-law. In the last model, GS_{diary}, we include a Mobility Diary Generator, an algorithm able to capture the tendency of individuals to follow or break their routine.

For each of the extensions, we propose four additional features: (i) the RSL (Relevance-based Starting Location), which considers the relevance during the assignment of the agents at their starting location; (ii) the reachable location concept, where we model the fact that an agent, associated with a waiting time, can visit only the locations reachable traveling at a certain speed for the associated amount of time; (iii) we introduce the concept of popularity of an agent at a collective level during the contact selection phase, when an agent decides to explore a new location; and (iv) we specify how to deal with borderline cases, proposing an action correction phase.

We evaluate the proposed models simulating the displacements of 1,001 individuals connected in a social graph, moving for three months in the urban area of New York City. The generated trajectories are compared with a real mobility dataset relative to the same urban area, extracted from an LBSN dataset. Among the proposed extensions, GS_{diary} produces the more realistic trajectories. From the experiments, it emerges that the role of both the model mechanisms and the tessellation granularity is crucial to produce realistic trajectories. We further validate the modeling ability of the best extension, GS_{diary}, simulating the mobility of individuals in different scenarios. We simulate the mobility of individuals moving in London, including in the model different levels of knowledge concerning their mobility behaviors. From the results obtained we can conclude that the model is able to generate realistic mobility trajectories independently from the urban area considered, and even with lack of information the model is still able to generate a set of trajectories that presents the well-know statistical patterns of real trajectories.

The proposed models can be further improved in several directions. The concept of spatial distance between the current location and the location to visit in the next displacement is considered only in the individual exploration; it can be also considered in the other cases together with the visitation pattern of the individual. In the proposed models the social graph is static; an interesting improvement can be to consider a dynamic social graph where the agents can create new links between them. Another consideration is that, for example, during the working hours an individual tends to interact mainly with its colleagues, in contrast, evening activities will be influenced mostly by its family or friends [7]. The social relationships of an individual change with time and the social graph can be modeled as a time-varying social graph: a graph where the weight of the connections changes according to the time and the social community of the contact. Currently, in the models, there is no such a representation of the urban infrastructure like urban roads. An improvement can be to consider the road network of a city, and the speed limit associated with each road. In this scenario, the agent reaches its selected location traveling through the road infrastructure, respecting the speed limits associated with the roads on its path.

An interesting improvement can be to use a different weighted spatial tessellation from the squared one used during our experiments. For example, a tessellation where the tiles do not cover the same area, but rather that contains a specific amount of population. In this way, a high population area will be partitioned in many small tiles; in contrast, a low population area will be represented with a small number of large tiles. Two libraries that allow this partition of the space in hierarchical hexagon tessellation are Uber's H3[1] and Google S2 Geometry[2]. From the experiments emerges that the use of a correct Mobility Diary Generator is important in order to generate realistic trajectories; as shown in Section 4.3.5, the circadian rhythm varies between populations according to their socio-cultural tradition. The circadian rhythm of a population can be shaped from different factors; an interesting improvement can be to design a model able to generate a plausible Mobility Diary Generator, if it is not available, starting

[1] https://eng.uber.com/h3

[2] https://s2geometry.io

from information about a population, such as the working hours schedule, opening hour of activities (e.g., schools, restaurants, gyms, pubs, shops and schools) and many others. Also the relevance of the locations plays a crucial role; when information about the relevance is not available a solution besides the use of the population density is to assign a relevance according to the empirical power-law distribution presented in Section 4.3.5. However, the heatmaps relative to the check-ins in New York City and London shows that the relevant locations are clustered in space, and this can not be modeled using only the power-law distribution. A solution can be to create a model in which the relevance of a location is more likely to be high if it surrounded by relevant locations. Artificial Intelligence techniques, such as Generative Adversarial Networks (GANs), are used to generate synthetic trajectories that follow the distribution of a real mobility trajectories used as a train dataset [20]. GANs can be embedded into the mechanistic generative models in order to produce more realistic trajectories, capturing the aspects of human movements that can not be modeled from the mechanisms of such generative models [20]. This could represent a further step forward in modeling and understanding human mobility.

Bibliography

[1] H. Barbosa, M. Barthelemy, G. Ghoshal, C. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, and M. Tomasini, "Human mobility: Models and applications," *Physics Reports*, 09 2017.

[2] L. Pappalardo and F. Simini, "Data-driven generation of spatio-temporal routines in human mobility," *Data Mining and Knowledge Discovery*, vol. 32, 12 2017.

[3] C. Song, T. Koren, P. Wang, and A.-L. Barabasi, "Modelling the scaling properties of human mobility," *Nature Physics*, vol. 6, 10 2010.

[4] M. C. Gonzalez, C. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, pp. 779–82, 07 2008.

[5] D. Yang, B. Qu, J. Yang, and P. Cudre-Mauroux, "Revisiting user mobility and social relationships in lbsns: A hypergraph embedding approach," pp. 2147–2157, 05 2019.

[6] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, pp. 462–5, 02 2006.

[7] D. Karamshuk, C. Boldrini, M. Conti, and A. Passarella, "Human mobility models for opportunistic networks," *IEEE Communications Magazine*, vol. 49, pp. 157–165, 12 2011.

[8] E. Cho, S. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," pp. 1082–1090, 08 2011.

[9] Y.-A. Montjoye, C. Hidalgo, M. Verleysen, and V. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific reports*, vol. 3, p. 1376, 03 2013.

[10] Y.-A. Montjoye, S. Gambs, V. Blondel, G. Canright, N. Cordes, S. Deletaille, K. Engø-Monsen, M. García-Herranz, J. Kendall, C. Kerry, G. Krings, E. Letouzé, M. Luengo-Oroz, N. Oliver, L. Rocher, A. Rutherford, Z. Smoreda, J. Steele, E. Wetter, and L. Bengtsson,

"On the privacy-conscientious use of mobile phone data," *Scientific Data*, vol. 5, p. 180286, 12 2018.

[11] J. Toole, C. Herrera-Yague, C. Schneider, and M. C. Gonzalez, "Coupling human mobility and social ties," *Journal of the Royal Society, Interface / the Royal Society*, vol. 12, 02 2015.

[12] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi, "Limits of predictability in human mobility," *Science (New York, N.Y.)*, vol. 327, pp. 1018–21, 02 2010.

[13] L. Pappalardo, F. Simini, S. Rinzivillo, D. Pedreschi, F. Giannotti, and A.-L. Barabasi, "Returners and explorers dichotomy in human mobility," *Nature Communications*, vol. 6, 09 2015.

[14] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi, "Human mobility, social ties, and link prediction," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1100–1108, 08 2011.

[15] C. Fan, Y. Liu, J. Huang, Z. Rong, and T. Zhou, "Correlation between social proximity and mobility similarity," *Scientific Reports*, vol. 7, 09 2016.

[16] L. Pappalardo, S. Rinzivillo, and F. Simini, "Human mobility modelling: Exploration and preferential return meet the gravity model," *Procedia Computer Science*, vol. 83, 05 2016.

[17] L. Alessandretti, P. Sapiezynski, S. Lehmann, and A. Baronchelli, "Evidence for a conserved quantity in human mobility," *Nature Human Behaviour*, vol. 2, 07 2018.

[18] L. Pappalardo, G. Barlacchi, F. Simini, and R. Pellungrini, "scikit-mobility: a python library for the analysis, generation and risk assessment of mobility data," 10 2019.

[19] S. Jiang, Y. Yang, S. Gupta, D. Veneziano, S. Athavale, and M. C. Gonzalez, "The timegeo modeling framework for urban mobility without travel surveys," *Proceedings of the National Academy of Sciences*, vol. 113, p. 201524261, 08 2016.

[20] X. Liu, H. Chen, and C. Andris, "trajgans : Using generative adversarial networks for geo-privacy protection of trajectory data (vision paper)," 2018.