International Journal of Geo-Information

MDPI

Article

A Mechanistic Data-Driven Approach to Synthesize Human Mobility Considering the Spatial, Temporal, and Social Dimensions Together

Giuliano Cornacchia [1,2,*] and Luca Pappalardo [2]

[1] Department of Computer Science, University of Pisa, 56127 Pisa, Italy
[2] Institute of Information Science and Technologies (ISTI), National Research Council of Italy (CNR), 56124 Pisa, Italy; luca.pappalardo@isti.cnr.it
* Correspondence: giuliano.cornacchia@phd.unipi.it

Abstract: Modelling human mobility is crucial in several areas, from urban planning to epidemic modelling, traffic forecasting, and what-if analysis. Existing generative models focus mainly on reproducing the spatial and temporal dimensions of human mobility, while the social aspect, though it influences human movements significantly, is often neglected. Those models that capture some social perspectives of human mobility utilize trivial and unrealistic spatial and temporal mechanisms. In this paper, we propose the Spatial, Temporal and Social Exploration and Preferential Return model (STS-EPR), which embeds mechanisms to capture the spatial, temporal, and social aspects together. We compare the trajectories produced by STS-EPR with respect to real-world trajectories and synthetic trajectories generated by two state-of-the-art generative models on a set of standard mobility measures. Our experiments conducted on an open dataset show that STS-EPR, overall, outperforms existing spatial-temporal or social models demonstrating the importance of modelling adequately the sociality to capture precisely all the other dimensions of human mobility. We further investigate the impact of the tile shape of the spatial tessellation on the performance of our model. STS-EPR, which is open-source and tested on open data, represents a step towards the design of a mechanistic data-driven model that captures all the aspects of human mobility comprehensively.

Keywords: human mobility; generative models; synthetic trajectories; social network; data science; mechanistic models; mathematical modelling

Citation: Cornacchia, G.; Pappalardo, L. A Mechanistic Data-Driven Approach to Synthesize Human Mobility Considering the Spatial, Temporal, and Social Dimensions Together. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 599. https://doi.org/10.3390/ijgi10090599

Academic Editor: Wolfgang Kainz

Received: 26 July 2021
Accepted: 9 September 2021
Published: 11 September 2021

1. Introduction

Human mobility affects crucial aspects of people lives such as the spreading of viral diseases (e.g., the COVID-19 pandemic) [1–5], public transportation and traffic volumes [4,6,7], urban population and migration [8–10], air pollution [11,12], and well-being [13,14]. Human mobility also plays a fundamental role in the COVID-19 pandemic, as human movements may accelerate the diffusion forcing governments to impose travel restrictions, bans of public gatherings, closures of non-essential businesses, and transitions to homeworking [15].

Mobility data resulting from the rise of ubiquitous computing (e.g., mobile phones, the Internet of Things, social media platforms) provides a precise way to sense human movements and face these societal challenges. Unfortunately, access to individual mobility data is restricted because they contain sensitive information about the individuals whose movements are described, and due to the EU General Data Protection Regulation (GDPR). Even when personal identifiers are removed to anonymize the dataset, there is no guarantee about the protection of the geo-privacy of individuals because they can be re-identified with a small amount of information [16–21].

A solution to deal with geo-privacy issues consists of design generative models of individual mobility, i.e., algorithms able to generate a collection of synthetic trajectories

that are realistic in reproducing fundamental human mobility patterns [22–26]. While disclosing real data requires a hard-to-control trade-off between uncertainty and utility, synthetic trajectories that preserve statistical properties may achieve in multiple tasks performance comparable to real data.

Beyond geo-privacy protection, synthetic trajectories are useful to the performance analysis of networking protocols such as mobile ad hoc networks, where the displacements of network users are exploited to route and deliver the messages [24,26,27]. Moreover, synthetic trajectories are fundamental for urban planning, what-if analysis, e.g., simulating changes in urban mobility in the presence of new infrastructures, epidemic diffusion, terrorist attacks, or international events.

Several models in the literature focus on capturing the patterns of human mobility, such as the existence of a power-law distribution in jump lengths [28–30] and in the characteristic spatial spread of an individual [29], a strong tendency to return to locations they visited before [29,30] and a propensity of move at specific times of the day following a circadian rhythm [24,31]. There are two main approaches to human mobility modelling: mechanistic approaches [22] and deep learning [23] approaches. On the one hand, mechanistic approaches embed directly the fundamental mechanisms of human mobility: they have the advantage of being interpretable, independent on the geography, and non-data demanding, but their realism is limited because of the simplicity of the implemented mechanisms. On the other hand, deep learning approaches let artificial neural networks discover the mechanisms from data: they achieve a greater realism but they are hard to interpret/explain, they are dependent on the training data and hence not geographically transferable, i.e., one model trained on a specific region cannot be directly used on a distinct, non-overlapping region [23].

In this paper, we focus on mechanistic approaches, starting from the observation that the social dimension is often dismissed in mobility modelling, even though about 10–30% of human movements are made due to social purposes [32]. As an exception, the GeoSim model [33] considers the social dimension, thanks to the inclusion of a mechanism related to individual preference and social influence. Unfortunately, the lack of sophisticated spatial and temporal mechanisms limits the realism of GeoSim's trajectories, making this model incomplete and hardly usable in practice.

To this end, we propose the Spatial, Temporal, and Social Exploration and Preferential Return model (STS-EPR), a modelling framework that includes mechanisms to capture the spatial, temporal, and social aspects of human mobility. Namely, STS-EPR includes a mechanism that takes into account the spatial distance between locations as well as the relevance of a location [34,35]; a temporal mechanism able to capture the tendency of individuals to follow a circadian rhythm [31]; and a social mechanism that models the influence of social ties to human displacements [33]. As a further novel contribution, STS-EPR also includes an action correction mechanism, aimed at overcoming borderline cases during the model execution.

The results of our experiments on several discretizations of the geographic space and social media data describing the checkins of thousands of users in several cities around the planet show that the synthetic trajectories generated by STS-EPR are realistic with respect to several social, temporal, and spatial aspects of human mobility. Specifically, we show that the lack of realism in one of the three mechanisms affects the realism of the others.

Our work is a further step towards the design of three-dimensional generative models for human mobility that are at the same time interpretable, geographically transferable, and realistic.

Open Source

The code of STS-EPR is included in the `Scikit-mobility` (https://github.com/scikit-mobility, accessed on 10 September 2021) Python library. The code allows to reproduce the experiments in our paper on open data about checkins from Foursquare.

2. Related Work

Among the many mechanistic generative models proposed for human mobility, the Exploration and Preferential return (EPR) model [36] has turned into a modelling platform given its robustness and modularity, allowing researchers to test their hypotheses by easily replacing or adding specific mechanisms to it. Specifically, EPR relies on two complementary mobility mechanisms: exploration and preferential return. During the exploration mechanism, an agent chooses a new location never visited before, based on a random walk process with truncated power-law jump size distribution. In the preferential return mechanism, an agent returns to a previously visited location based on the number of visits to that specific location, it reproduces the propensity of humans to return to locations they visited before. An agent in the model selects to explore a new location with probability P_{exp}, and with complementary probability $P_{ret} = 1 - P_{exp}$, the agent returns to a previously visited location.

Several studies consequently widened the EPR model by adding increasingly complex mechanisms to reproduce statistical laws more realistically. In the d-EPR model [37], an agent visits a new location depending on both its distance from the current position and collective relevance. In the recency-EPR model [38], the preferential return phase includes information about the recency of location visits. In the memory-EPR model [39], during the exploration mechanism, the agent selects a location with probability proportional to the number of times it visited that location during the previous M days. EPR and its extension, focus only on the spatial aspect of human mobility, neglecting to reproduce realistic temporal patterns. For example, the displacements of individuals are not uniformly distributed during the day but follow the circadian rhythm, a property that is not captured by EPR-like models. Two refined models, namely TimeGeo [40] and DITRAS [31], overcome this problem by including a more refined temporal mechanism.

TimeGeo [40] is a mechanistic modelling framework to produce individual mobility trajectories with realistic spatiotemporal properties. TimeGeo models the temporal dimension through a time-inhomogeneous Markov chain that captures the circadian propensity to travel and the likelihood of arranging short and consecutive activities [40]. It integrates the temporal mechanism with a rank-based version of the EPR model (r-EPR), which assigns a rank to each unvisited location during the selection of a new location to visit, depending on its distance from the trip origin [40].

DITRAS (DIary-based TRAjectory Simulator) [31] generates the synthetic trajectories exploiting two probabilistic models: a diary generator and a trajectory generator. The diary generator consists of a Markov model trained on mobility trajectory data of real individuals, able to capture the probability of individuals to follow or break their routine at specific times of the day [31]. The trajectory generator is an algorithm that, given a weighted spatial tessellation, translates the abstract locations in physical locations using the d-EPR model [34].

Notwithstanding the definite correlation between human mobility and sociality, one of the few mechanistic models that attempts to replicate the socio-mobility patterns is GeoSim [33]. GeoSim takes into account both the mobility and the social dimension, although incorporating a trivial temporal mechanism. GeoSim proposes two mechanisms beyond the explore and preferential return ones: individual preference and social influence. The agent has to decide if its next displacement will be influenced or not by its social contact, respectively, with probability α and $1 - \alpha$.

Position of Our Work

An overview of the literature cannot avoid noticing the lack of mechanistic generative models able to reproduce realistically the spatial, temporal, and social dimensions at the same time. On the one hand, GeoSim can capture meaningful patterns representing the link between mobility and sociality, but cannot reproduce realistic spatiotemporal patterns. On the other hand, TimeGeo and DITRAS reproduce spatial and temporal patterns well but ignore the social dimension. In this paper, we propose the STS-EPR model in which we

combine the mechanisms of existing mechanistic models attempting to reproduce the three dimensions of human mobility.

3. Definitions

3.1. Trajectory

The trajectory of an individual is a sequence of records that allows for reconstructing their movements during the period of observation [41,42]. Formally, a spatiotemporal trajectory is defined as 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 $r_i = (x_i, y_i)$ is a pair of coordinates on a bi-dimensional space in a given reference system (CRS), e.g., latitude and longitude. A pair (r_n, t_n) denotes a visit at location r_n at timestamp t_n.

3.2. Spatial Tessellation

For modelling purposes, the geographic space is discretized through a weighted spatial tessellation. The tiling of the geographic space aims at creating the covering of the area of interest using regular tiles, such as squared or hexagonal tiles, or irregular tiles that may define the shape of buildings, census cells, or administrative units.

Formally, given an area A, a set of geographical polygons called tessellation, G, is defined with the following properties: (1) G contains a finite number of polygons, l_i called tiles, $G = \{l_i : i = 1, \ldots n\}$; (2) the locations are non-overlapping, $l_i \cap l_j = \varnothing$, $\forall i \neq j$; and (3) the union of all locations completely covers A, $\bigcup_{i=1}^{n} l_i = A$.

In a weighted tessellation L, at each tile is associated its relevance, namely the popularity of a location among real individuals [31]. The overall number of visits to a location is usually used as an estimation of its relevance.

$L = \langle (r_1, w_1), \ldots, (r_n, w_n) \rangle$, where w_j represents the relevance of the tile j and r_j is the representative point of the tile j.

3.3. Mobility Diary

A Mobility Diary (MD) is an abstract trajectory that describes the locations (in terms of placeholders called abstract locations) and the timestamp at which the user visits that specific abstract location [31]. Generally, in generative models for human mobility, the abstract locations are mapped into physical ones through diverse spatial mechanisms (e.g., EPR [36], or d-EPR [37]).

A mobility diary is generated by a Mobility Diary Generator (MDG) and is defined as follows:

$$\text{MD} = \langle (ab_0, t_1), (ab_1, t_2), \ldots (ab_j, t_{j+1}), (ab_0, t_{j+2}), (ab_1, t_{j+3}) \ldots) \rangle$$

where ab_j denotes an abstract location and t_j the timestamp at which the individual visits ab_j. The abstract location ab_0 refers to the home location of an individual.

3.4. Visitation Pattern

The visitation pattern of an individual a is represented as a vector lv_a of $|L|$ elements, called location vector, where $|L|$ is the total number of locations in L. The j-th element of the location vector, $lv_a[j]$, contains the number of times a visited the location r_j. The visitation frequency of a to a location r_i is: $f_a(r_i) = \frac{lv_a[i]}{\sum_{j=1}^{|L|} lv_a[j]}$.

3.5. Contact Graph

An individual's network of contacts G may influence their movements. We define $G = (V, E)$ as a graph in which V indicates the set of individuals and E the social ties between individuals.

3.6. Problem Definition

A generative mobility model M is any algorithm able to generate a set of n synthetic trajectories $\mathcal{T}_M = \{T_{a_1}, \ldots, T_{a_n}\}$, which describe the movements, during a certain period of time, of n independent agents $a_1, \ldots, a_n$ on a spatial tessellation L [23]. The realism of M is evaluated with respect to:

1. A set of spatial $(s_1, \ldots, s_{m_s})$, temporal $(t_1, \ldots, t_{m_t})$, and social patterns $(o_1, \ldots, o_{m_t})$. The patterns refer to the distributions of mobility measures that quantify aspects related to the spatial, temporal, or social aspects of an individual's mobility (e.g., radius of gyration, mobility entropy, mobility similarity). A realistic $\mathcal{T}_M$ is expected to reproduce as many mobility patterns as possible.
2. A set $\mathcal{X} = \{T_{u_1}, \ldots, T_{u_m}\}$ of real mobility trajectories corresponding to m real individuals $u_1 \ldots u_m$ that move on the same region as the one on which synthetic trajectories are generated. $\mathcal{X}$ is used to compute the set $\mathcal{K}$ of patterns, which are compared with the patterns computed on $\mathcal{T}_M$.
3. A function D that computes the dissimilarity between two distributions. Specifically, for each measure in $f \in \mathcal{K}$, $D(P_{(f,\mathcal{T}_M)}||P_{(f,\mathcal{X})})$ indicates the dissimilarity between $P_{(f,\mathcal{T}_M)}$, the distribution of the measures computed on the synthetic trajectories in $\mathcal{T}_M$, and $P_{(f,\mathcal{X})}$, the distribution of the measures computed on the real trajectories in $\mathcal{X}$. The lower $D(P_{(f,\mathcal{T}_M)}||P_{(f,\mathcal{X})})$, the more realistic model M is with respect to f and $\mathcal{X}$.

4. The STS-EPR Model

The Spatial, Temporal and Social Exploration and Preferential Return model (STS-EPR) extends the Exploration and Preferential Return model (EPR) [34] by considering the social dimension together with the spatial and temporal ones. It also includes a temporal mechanism that reproduces realistically the distribution of the number of movements during the day.

STS-EPR takes as input a spatial tessellation L, a mobility diary generator MDG, the time interval of the simulation $[t_{start}, t_{end}]$, and an undirected graph G describing the social relationships between the N agents. The model outputs N synthetic trajectories describing the displacements of N agents on L during the period $[t_{start}, t_{end}]$.

STS-EPR consists of four phases: initialization, action selection, location selection, and action-correction (see Figure 1). After the initialization phase, the agents perform the action selection, the location selection, and eventually the action-correction phases until a stopping criterion is satisfied (e.g., the number of hours to simulate is reached).

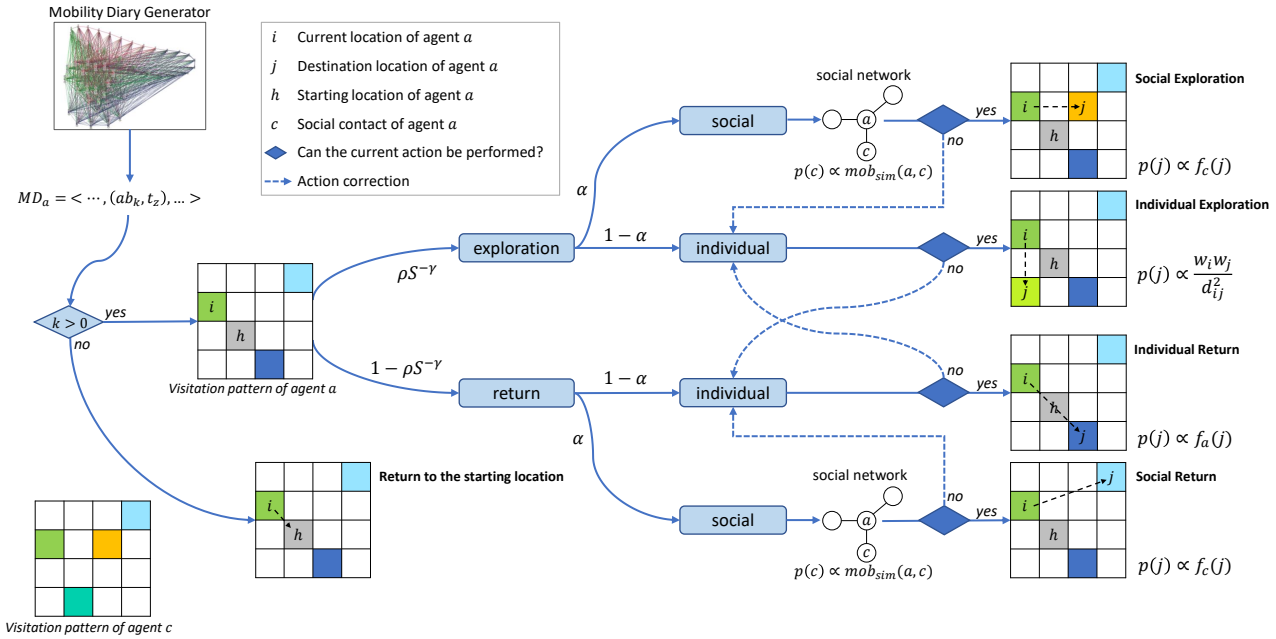


Figure 1. A schematic description of STS-EPR. When an agent a moves according to the entry in its mobility diary MD_a, if the abstract location is ab_0 the individual returns to its starting location, otherwise it decides whether to explore a new location or return to a previously visited one. At that point, a determines whether or not its social contacts affect its choice for the location to visit next. If the selected action cannot be performed, it is corrected with an executable one (dashed arrows indicate action corrections).

4.1. Initialization

The edge's weights in G are initialized to zero and updated during the simulation. The weight of an edge indicates the mobility similarity of the linked agents, i.e., the cosine similarity of their location vectors. The model assigns to each agent a Mobility Diary produced by a Mobility Diary Generator (MDG); in STS-EPR, the MDG considered is a Markov Model that captures the individuals' probability to follow or break their routine at specific times of the day, exploiting the conditional probability of real trajectory data [31]. The mobility diary MD of an agent a is defined as:

$$\text{MD}_a = \langle (ab_0, t_1), (ab_1, t_2), \ldots (ab_j, t_{j+1}), (ab_0, t_{j+2}), (ab_1, t_{j+3}) \ldots) \rangle$$

where ab is an abstract location, ab_0 denotes a's starting location, and t_i is a timestamp. Two distinct consecutive abstract locations must be mapped to two distinct physical locations. After assigning the mobility diaries, the model assigns each agent to a starting location. The probability $p(r_i)$ for an agent of being assigned to a starting location $r_i \in L$ is $\propto w_i$, where w_i is the location's relevance. Each agent will move according to the entries in its mobility diary at the time specified. If the current abstract location is ab_0, the agent returns to its starting location; otherwise, ab_i is converted into a physical location through the next steps.

4.2. Action Selection

When moving, an agent can decide whether to explore a new location or return to a previously visited one by selecting one of two competing mechanisms: exploration and preferential return. Exploration models the decreasing tendency to explore new locations over time [34]. Preferential return reproduces individuals' significant propensity to return to locations they explored before [31,34,37]. An agent explores a new location with probability $P_{exp} = \rho S^{-\gamma}$, or returns to a previously visited one with a complementary probability $P_{ret} = 1 - \rho S^{-\gamma}$, where S is the agent's number of unique visited locations and $\rho = 0.6$, $\gamma = 0.21$ are constants (for these two parameters, we use the values estimated in the literature on mobile phone records [34]). The parameters ρ and γ influence the user's

tendency to explore a new location versus returning to a previously visited location [34]. When the agent returns, it selects a location with a probability proportional to its visitation frequency. At that point, independently of the spatial mechanism selected, the agent determines whether or not the choice of the next location to visit is affected by the other agents, selecting between the individual and the social influence mechanisms. With a probability $P_{soc} = \alpha$, the agent's social contacts will influence its movement [33]. With a complementary probability of $1 - \alpha$, the agent's choice is not influenced by the other agents. The social factor α is equal to 0.2 as in the GeoSim model [33]. Indeed, Toole et al. [33] find that an exponential distribution with a mean value of 0.2 produces a close fit to the distribution of mobility similarity observed in the population. Moreover, this value is consistent with the results of Cho et al. [32], who find that 10–30% of trips are motivated by social reasons. Table 1 summarizes the default value and the role of each parameter.

Table 1. A summary of the parameters of STS-EPR.

Parameter	Default Value	Is the Parameter Fitted from a Dataset?	It Models
ρ	0.6	no	explore or return choice
γ	0.21	no	explore or return choice
α	0.2	no	social factor
$w_1 \dots w_{\lvert L \rvert}$	-	yes	relevances of the $\lvert L \rvert$ locations
Mobility Diary Generator (MDG)	-	yes	mobility diary of agents

4.3. Location Selection

At this point, the agent a decides which location is the destination of its displacement. The sets of locations a can visit or return to are $exp_a = \{i \mid lv_a[i] = 0\}$ and $ret_a = \{i \mid lv_a[i] > 0 \wedge i \notin \{s_a, c_a\}\}$, respectively, where s_a and c_a denote the indices of the starting and current location of agent a. The set of the location visited, without the constraints of the current and starting location, is $vis_a = \{i \mid lv_a[i] > 0\}$. During the location selection step, a can choose among the following actions:

- **Individual Exploration (IE):** a chooses a new location to explore from exp_a. Individuals are more likely to move at small rather than long distances but also take into account the location's collective relevance [31]. We use the gravity law to couple distance and relevance [37]. If a is currently at location r_j, it 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 locations r_i and r_j with relevances w_i, w_j.
- **Social Exploration (SE):** a selects an agent c among its social contacts in the social graph G, i.e., $c \in \{v \in V | (a, v) \in E\}$. The probability $p(c)$ for c to be selected is proportional to the mobility-similarity between them: $p(c) \propto mob_{sim}(a, c)$. After the contact c is chosen, the candidate location to explore is an unvisited location for a that was visited by c, i.e., the location is selected from set $A = exp_a \cap vis_c$. The probability $p(r_i)$ for a location r_i, with $i \in A$, to be selected is proportional to the visitation pattern of c, namely $p(r_i) \propto f_c(r_i)$.
- **Individual Return (IR):** a chooses the return location from the set ret_a with a probability proportional to its visitation pattern. The probability for a location r_i with $i \in ret_a$ to be chosen is: $p(r_i) \propto f_a(r_i)$.
- **Social Return (SR):** c is selected as in SE, and the location a returns to is picked from the set $A = ret_a \cap vis_c$. The probability $p(r_i)$ for a location r_i to be selected is proportional to the visitation pattern of the agent c, namely $p(r_i) \propto f_c(r_i)$.

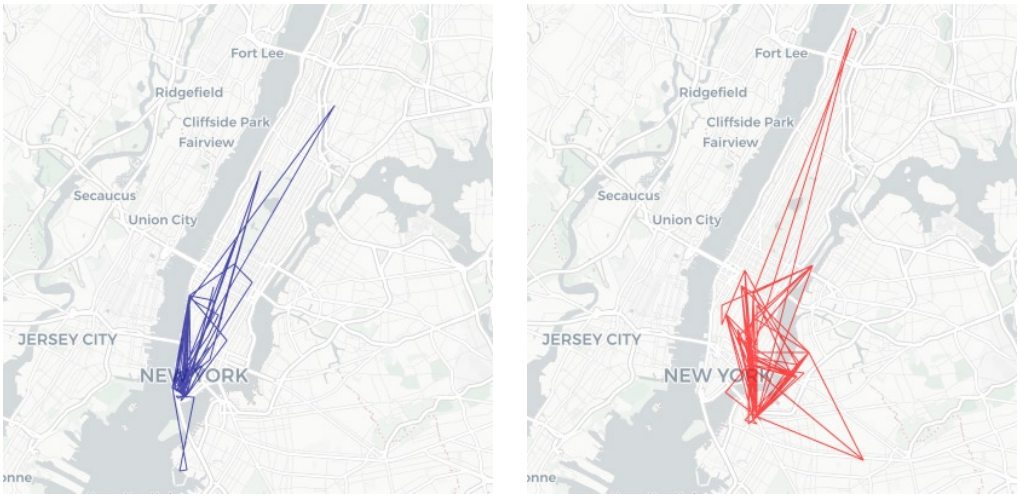
4.4. Action Correction

The set of possible locations an agent can reach is limited. For example, it may happen that the agent visited all locations at least once and there are no new locations to explore. To comply with these kinds of constraints, we introduce an action correction phase, executed if the location selection phase does not allow movements in any location.

- **No location in social choices**: If an agent a decides to move with the influence of a social contact c, but $ret_a \cap vis_c = \varnothing$ or $exp_a \cap vis_c = \varnothing$ (no locations visited by both c and a or no locations visited by c and unvisited by a), we execute an individual action preserving a's choice to explore or return.
- **No new location to explore**: When an agent a decides to explore but it visited all the locations at least once ($exp_a = \varnothing$), we force the agent to make an IR action.
- **No return location**: If an agent a, currently at location r_i, decides to perform an IR, and r_i is the only location visited so far (besides the starting location), it cannot return to any location ($ret_a = \varnothing$). We force a to make an IE.

5. Experiments

In this section, we present the experiments to evaluate the performance of STS-EPR. We simulate the mobility of individuals in eight cities around the globe using STS-EPR and two state-of-the-art models: DITRAS [31] and GeoSim [33]. We evaluate the realism of synthetic trajectories generated by the mentioned models in terms of their statistical similarity with real ones extracted from Foursquare checkins [43] (Figure 2). Furthermore, we also conducted studies to examine the effect of different tile shapes on the trajectories' generation.



(**a**) (**b**)

Figure 2. An example of a real trajectory (**a**) and a synthetic trajectory generated by STS-EPR (**b**) in New York City. Although the two trajectories at a first inspection are quite different, the synthetic trajectory (**b**) preserves some important characteristics of real trajectories. Both trajectories are concentrated in the most relevant locations (close to Manhattan), they have mostly small jump lengths and a small number of trips to distant places (power law behaviour of jump lengths [28]). Plots generated with `scikit-mobility` [44].

5.1. Datasets

5.1.1. Trajectories Dataset

We use a public dataset D_{FS}, collected by Yang et al. [43], which includes a set of global-scale checkins gathered from the social network platform Foursquare over 22 months (from April 2012 to January 2014). We use a lookup dataset D_{loc} to associate the location's identifier with the corresponding geographic coordinates. Table 2 shows some examples of records in the two datasets.

A checkin describes a user's real-time position with its social contacts. A user's time-ordered sequence of checkins can be used to reconstruct their movement considering each checkin as a point in their trajectory. Note that the reconstructed trajectory represents a portion of the user's mobility, and it is biased towards the most captivating places worth sharing on social media (e.g., points of interest).

The authors of the dataset collected the Foursquare checkins from Twitter by searching the Foursquare hashtag [43]. The dataset is associated with a snapshot of the social network obtained from Twitter, antecedent at the collection period.

From the D_{FS} dataset, we extracted, for each city, a validation and a calibration dataset. The validation dataset contains the checkins of a set of users connected through the social graph for three months, from the 10 April 2012 to the 10 July 2012 (Figure 3); it represents a benchmark of genuine trajectories to be used during the validation phase. The calibration dataset contains for the same period the checkins of a set of users not included in the validation dataset. The calibration dataset will be used to calibrate the model, i.e., to compute the location relevance and fit the Mobility Diary Generator. We report the characteristics of these datasets in Table 3.

Table 2. 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.

(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
⋮	⋮	⋮	⋮	⋮

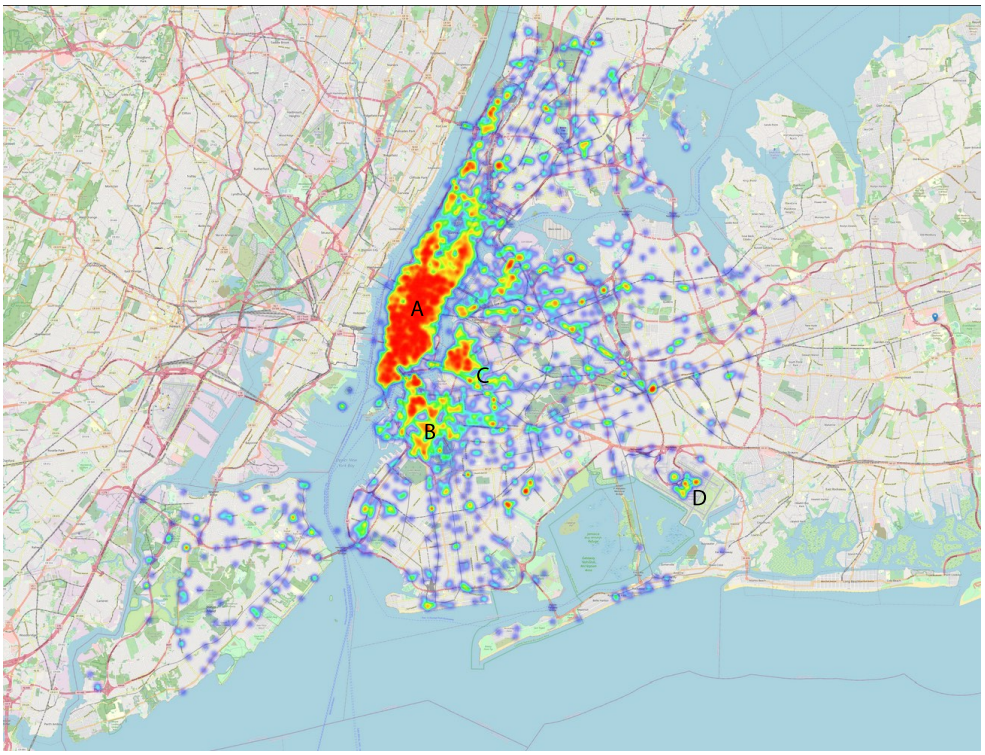


Figure 3. Heatmap of the positions of the 37,489 checkins made by 1001 individuals during three months (April 2012 to July 2012) in New York City. There is a high density of check-ins in Manhattan (A) and its surroundings (upper part of Brooklyn (B) and Queens (C)). The high concentration of check-ins in these areas can be explained because Manhattan is the most densely populated borough and the touristic centre of New York City, containing for example Times Square, Central Park, the Empire State Building, the Statue of Liberty, and Wall Street. As one may expect, another area of dense checkins is the JFK airport (D).

5.1.2. Social Graph

Each of the validation datasets is associated with a social graph that describes the social relationships (i.e., mutual follow on Twitter) between those users that made at least two checkins between the 10 April 2012 and the 10 July 2012. Table 3 reports the characteristics of the contact graphs.

Table 3. A summary of the properties of the eight validation datasets and the corresponding contact graphs extracted from the public Foursquare dataset D_{FS} and of the corresponding calibration datasets.

		Validation			Calibration	
City	**#checkins**	**#users**	**#edges**	**avg. degree**	**#checkins**	**#users**
New York City (US)	37,489	1001	1755	3.506	247,058	19,416
Osaka (JP)	46,755	823	1734	4.214	48,832	3534
Kuala Lumpur (MY)	78,037	2582	5715	4.427	159,514	13,453
Sao Paulo (BR)	86,654	1651	3725	4.512	266,235	15,733
Jakarta (ID)	99,460	2162	3781	3.498	391,576	35,576
Bangkok (TH)	109,585	2044	5228	5.115	265,670	15,632
Istanbul (TR)	228,755	5089	9918	3.898	555,913	33,454
Tokyo (JP)	229,283	4043	16,137	7.983	185,809	12,825

5.2. Measures

We quantify the models' realism (Section 3.6) with respect to several mobility measures using the six datasets of Section 5.1 and the Kullback–Leibler divergence (KL), defined as:

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

where p is the ground truth distribution and q is a synthetic distribution. We consider the following spatial, temporal, and social mobility measures, which capture well-known statistical patterns of individual human mobility [22,23]:

- Jump Length Δr, the distance between two consecutive locations visited by an individual [28–30]. Formally, $\Delta r = d(r_i, r_{i+1})$ is the geographical distance between two points r_i and r_{i+1} in a trajectory. A truncated power-law well approximates the empirical distribution $P(\Delta r)$ within a population of individuals, with the value of the exponent slightly varying based on the type of data and the spatial scale [28,29].
- Radius of Gyration r_g, the typical distance travelled by an individual u during the period of observation [29,35]. The r_g of individual u defined as $r_g(u) = \sqrt{\frac{1}{n_u}\sum_{i=1}^{n_u} d(r_i, r_{cm})^2}$, where n_u is the number of points in T_u, $r_i \in T_u$ and $r_{cm} = \frac{1}{n_u}\sum_{i=1}^{n_u} r_i$ is the position vector of the centre of mass of the set of points in T_u. A truncated power-law well approximates the distribution of r_g [29,30].
- Visits per Location V_l, the relevance of a location described as its attractiveness at a collective level, indicating the popularity of locations according to how people visit them on the geographic space [31,45].
- Location Frequency $f(r_i)$, the probability of an individual to visit a location r_i [29], identifying the importance of a location to an individual's mobility: the most visited location (likely home or work) has rank 1, the second most visited location (e.g., school or local shop) has rank 2, and so on. The probability of finding an individual at a location of rank L is well approximated by $P(L) \sim 1/L$ [29,30].
- Waiting Time Δt: the elapsed time between two consecutive visits of an individual u: $\Delta t = t_i - t_{i-1}$. Empirically the distribution of waiting times is well approximated by a truncated power-law [34].
- Uncorrelated Entropy S_{unc}: the predictability of the movements of an individual u [36], defined as $S_{unc}(u) = -\sum_{i=1}^{n_u} p_u(i) \log_2 p_u(i)$, where n_u is the number of distinct locations visited by u and $p_u(i)$ is the probability that u visits location i [36,46].
- Activity per Hour $t(h)$: the number of movements made by the individuals at every hour of the day [31,40]. The movements of individuals are not distributed uniformly during the hours of the day but follow a circadian rhythm: people tend to be stationary during the night hours while prefer moving at specific times of the day, for example, to reach the workplace or return home.
- Mobility Similarity mob_{sim}: the cosine-similarity of two individuals' location vectors [32,33,43].
  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.

$$mob_{sim}(u_i, u_j) = \frac{lv_i \cdot lv_j}{\|lv_i\| \|lv_j\|} \tag{2}$$

  Several studies demonstrate the correlation between human mobility and sociality [32,33,43,47,48]: 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.

5.3. Experimental Settings

We synthesize the trajectories of individuals moving for three months in each city using STS-EPR, GeoSim, and DITRAS.

6. Results

For each combination of city and model, we run a trajectory generation for five times and take, for each measure, the average and standard deviation of the resulting KL divergence. Table 4 shows the results for all cities, obtained using a squared tessellation of 300 m. In Sections 6.1–6.3, we present the results obtained using the squared tessellation with tile size of 300 m. We discuss the role of the type and size of tiles in Section 6.4. While the distributions and tables for Kuala Lumpur, Sao Paulo, Jakarta, Bangkok, and Istanbul, together with the distributions and table for the hexagonal tessellation, are reported in the Supplementary Material.

Table 4. Results of STS-EPR, DITRAS, and GeoSim for each city. The results refer to the squared spatial tessellation with tiles of 300 m. For each mobility measure, we show the average and standard deviation of the KL divergence of five generation experiments.

		Δr	r_g	L_i	Vl	Δ_t	$t(h)$	S^{unc}	mob_{sim}
New York City	STS-EPR	**0.017 ±0.0012**	0.2399 ±0.0797	0.0225 ±0.0006	**0.0229 ±0.0044**	**0.0827 ±0.0006**	**0.0227 ±0.0012**	**1.7856 ±0.0688**	**0.1874 ±0.006**
	DITRAS	0.0199 ±0.0021	**0.0848 ±0.0162**	0.1505 ±0.0052	0.077 ±0.0059	0.0848 ±0.0015	0.0233 ±0.0008	3.3201 ±0.3192	0.7903 ±0.1218
	GeoSim	0.7906 ±0.0034	5.3613 ±0.0193	**0.0049 ±0.0004**	4.4898 ±0.0154	0.9752 ±0.0464	0.1801 ±0.0006	7.997 ±0.0771	0.5558 ±0.0116
Tokyo	STS-EPR	**0.0485 ±0.0024**	**0.1517 ±0.0133**	0.0103 ±0.0003	**0.0105 ±0.0006**	0.2406 ±0.0012	**0.0285 ±0.0004**	**1.4906 ±0.0166**	**0.0284 ±0.0025**
	DITRAS	0.066 ±0.0013	0.3905 ±0.0183	0.1132 ±0.0011	0.1201 ±0.009	**0.2398 ±0.0024**	0.0286 ±0.0006	2.1747 ±0.2988	1.0454 ±0.0369
	GeoSim	0.7402 ±0.0032	4.8877 ±0.0058	**0.0001 ±0.0**	2.9007 ±0.0067	1.0047 ±0.0004	0.2874 ±0.0001	6.658 ±0.0238	0.098 ±0.0023
Bangkok	STS-EPR	0.046 ±0.0016	0.2195 ±0.1916	0.0059 ±0.0059	**0.0097 ±0.001**	0.3094 ±0.1848	**0.0145 ±0.0003**	**1.5182 ±0.0574**	**0.0099 ±0.0023**
	DITRAS	**0.0336 ±0.0027**	**0.1992 ±0.0089**	0.0893 ±0.0021	0.0663 ±0.0042	**0.1578 ±0.0007**	0.0147 ±0.0004	2.1252 ±0.0341	1.5523 ±0.105
	GeoSim	0.7391 ±0.003	5.1465 ±0.005	**0.0023 ±0.0002**	3.6032 ±0.0031	0.9575 ±0.0002	0.2928 ±0.0004	4.8474 ±0.0982	0.1742 ±0.003
Osaka	STS-EPR	**0.0346 ±0.0041**	**0.0793 ±0.0096**	0.0049 ±0.0001	**0.0108 ±0.0019**	**0.2447 ±0.0033**	**0.0357 ±0.0007**	**1.8577 ±0.0439**	**0.0449 ±0.0045**
	DITRAS	0.0625 ±0.009	0.125 ±0.0154	- -	0.0566 ±0.0032	0.2474 ±0.0032	0.0362 ±0.0003	2.3668 ±0.0676	1.2222 ±0.0659
	GeoSim	0.7792 ±0.0061	5.0454 ±0.0048	**0.0037 ±0.0002**	4.079 ±0.0112	1.0225 ±0.0009	0.3062 ±0.0004	7.1387 ±0.0007	0.3126 ±0.0142
Istanbul	STS-EPR	**0.0118 ±0.0009**	**0.1051 ±0.0243**	0.0169 ±0.0002	**0.0082 ±0.0007**	0.2924 ±0.1858	0.0059 ±0.0002	**1.8559 ±0.0488**	**0.0828 ±0.0009**
	DITRAS	0.0242 ±0.0016	0.8151 ±0.0133	0.1381 ±0.0009	0.1096 ±0.0049	**0.2164 ±0.186**	**0.0057 ±0.0001**	3.5716 ±0.1668	1.0926 ±0.0422
	GeoSim	0.5296 ±0.0021	4.9211 ±0.0005	**0.0012 ±0.0001**	2.8892 ±0.0021	0.9735 ±0.047	0.2228 ±0.0002	6.0196 ±0.0104	0.3583 ±0.0056

Table 4. *Cont.*

		Δr	r_g	L_i	Vl	Δ_t	$t(h)$	S^{unc}	mob_{sim}
Jakarta	STS-EPR	0.0341 ±0.0023	**0.0329 ±0.006**	0.0077 ±0.0002	**0.017 ±0.0023**	0.1466 ±0.0037	**0.0153 ±0.0003**	1.2139 ±0.027	**0.0269 ±0.0044**
	DITRAS	**0.0203 ±0.0012**	0.2938 ±0.0221	0.0683 ±0.0	0.1003 ±0.0028	**0.1443 ±0.0014**	0.0157 ±0.0005	**1.1002 ±0.054**	1.8143 ±0.1379
	GeoSim	0.7069 ±0.0031	5.3784 ±0.0046	**0.0038 ±0.0004**	3.4635 ±0.0031	0.9631 ±0.0002	0.2198 ±0.0003	5.1984 ±0.1167	0.3275 ±0.0065
Sao Paulo	STS-EPR	0.0314 ±0.0015	**0.0669 ±0.0169**	**0.005 ±0.0002**	**0.0067 ±0.0006**	0.2394 ±0.1604	0.0169 ±0.0007	2.123 ±0.0837	**0.0582 ±0.0026**
	DITRAS	**0.0212 ±0.001**	0.1949 ±0.0149	0.0741 ±0.0007	0.0448 ±0.0035	**0.2386 ±0.1601**	**0.0165 ±0.0004**	**2.1078 ±0.0572**	0.7835 ±0.049
	GeoSim	0.6968 ±0.0032	5.8012 ±0.0387	0.0072 ±0.0004	3.8986 ±0.0289	0.9886 ±0.0002	0.1692 ±0.0003	5.2005 ±0.1224	0.3532 ±0.0062
Kuala Lumpur	STS-EPR	0.0513 ±0.0015	**0.1616 ±0.0068**	0.0087 ±0.0002	**0.0141 ±0.0048**	**0.1448 ±0.0017**	**0.0096 ±0.0001**	**1.4955 ±0.027**	**0.0217 ±0.0072**
	DITRAS	**0.0442 ±0.0047**	0.6756 ±0.0265	- -	0.2834 ±0.0203	0.1452 ±0.0021	0.0097 ±0.0002	1.723 ±0.0369	1.0451 ±0.0626
	GeoSim	0.6494 ±0.0006	4.6041 ±0.0153	**0.0025 ±0.0001**	2.188 ±0.0045	0.978 ±0.0002	0.192 ±0.0005	6.7618 ±0.0996	0.2336 ±0.0044

6.1. Spatial

Our results highlight the importance of coupling distance and relevance in the location selection phase. The models that use the gravity law during the Individual Exploration, i.e., STS-EPR and DITRAS, capture the distribution of the jump length realistically, obtaining similar KL scores for all the cities (best score obtained with STS-EPR in Istanbul with a KL = 0.011, Table 4). In contrast, since GeoSim does not include any mechanism to consider the geographical distance or the relevance, it fails to reproduce the power-law behaviour of the Δr distribution.

We obtain similar results for the radius of gyration: GeoSim achieves the worst performance for all the cities examined (KL $\in [4.604, 5.361]$) and fails to capture the shape of the real distribution. The trajectories generated by STS-EPR and DITRAS capture correctly the power-law behaviour of the radius of gyration; they achieve similar scores, nevertheless STS-EPR (KL $\in [0.032, 0.239]$) outscores DITRAS (KL $\in [0.084, 0.815]$) in six cities out of eight.

GeoSim is the best model in terms of KL regarding the location frequency measure, although it underestimates the real distribution (Figures 4d and 5d). STS-EPR captures this measure accurately, achieving a KL score that is on average 90.80% better than the one obtained by DITRAS (Table 4). DITRAS, when applied in some cities, cannot generate trajectories whose users have visited a sufficient number of locations to compute the distribution.

STS-EPR generates trajectories that preserve the distribution of location relevance, too: it is the best model for all the cities, with a KL score that is on average 99.64% and 87.59% better than that of GeoSim and DITRAS, respectively, (Figure 6d); STS-EPR outscores DITRAS even if both the models use the concept of relevance in their spatial mechanisms.

Finally, none of the models approximate the distribution of the entropy measure; however, STS-EPR results as the best model for this measure in six cities (Figures 4h and 5h).

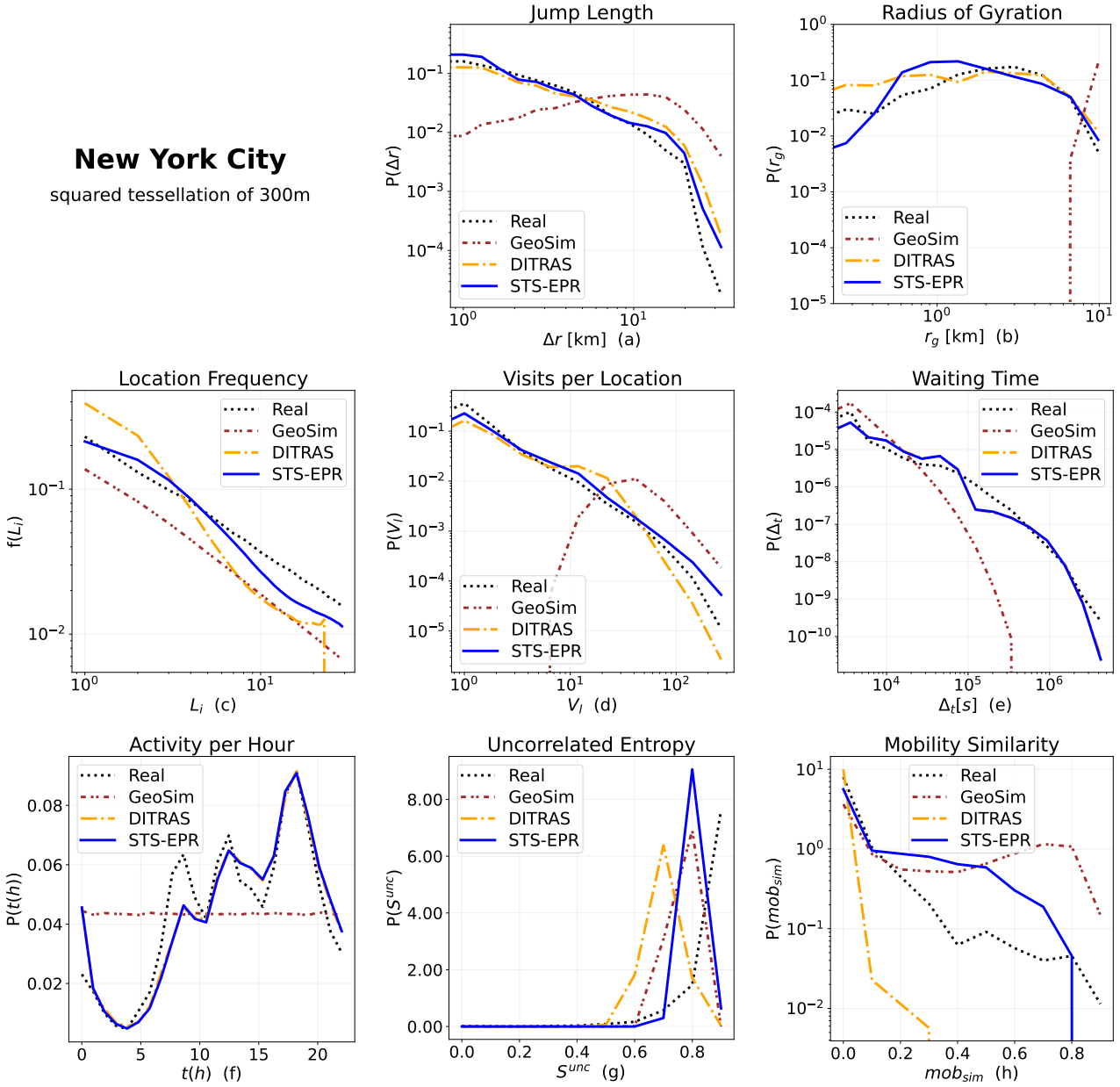


Figure 4. Comparison of the distribution of the mobility measures jump length (**a**), radius of gyration (**b**), location frequency (**c**), visits per location (**d**), waiting time (**e**), activity per hour (**f**), uncorrelated entropy (**g**), and mobility similarity (**h**) of real data (black dotted line) and data produced by GeoSim (red dash-dotted line), DITRAS (orange dash-dotted line), and STS-EPR (blue line), for New York City and the squared tessellation with tiles of 300 m.

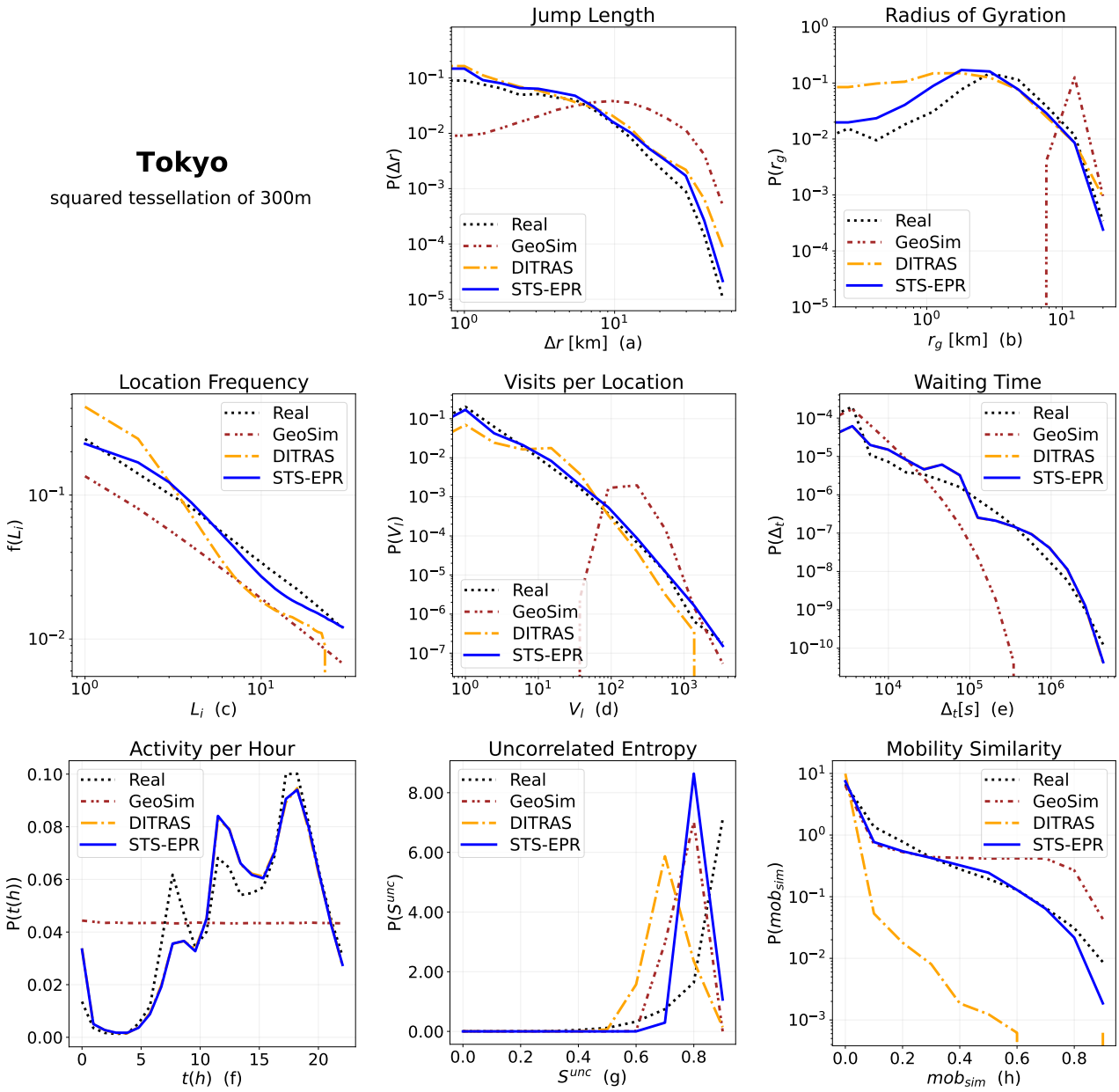


Figure 5. Comparison of the distribution of the mobility measures jump length (**a**), radius of gyration (**b**), location frequency (**c**), visits per location (**d**), waiting time (**e**), activity per hour (**f**), uncorrelated entropy (**g**), and mobility similarity (**h**) of real data (black dotted line) and data produced by GeoSim (red dash-dotted line), DITRAS (orange dash-dotted line), and STS-EPR (blue line), for Tokyo and the squared tessellation with tiles of 300m.

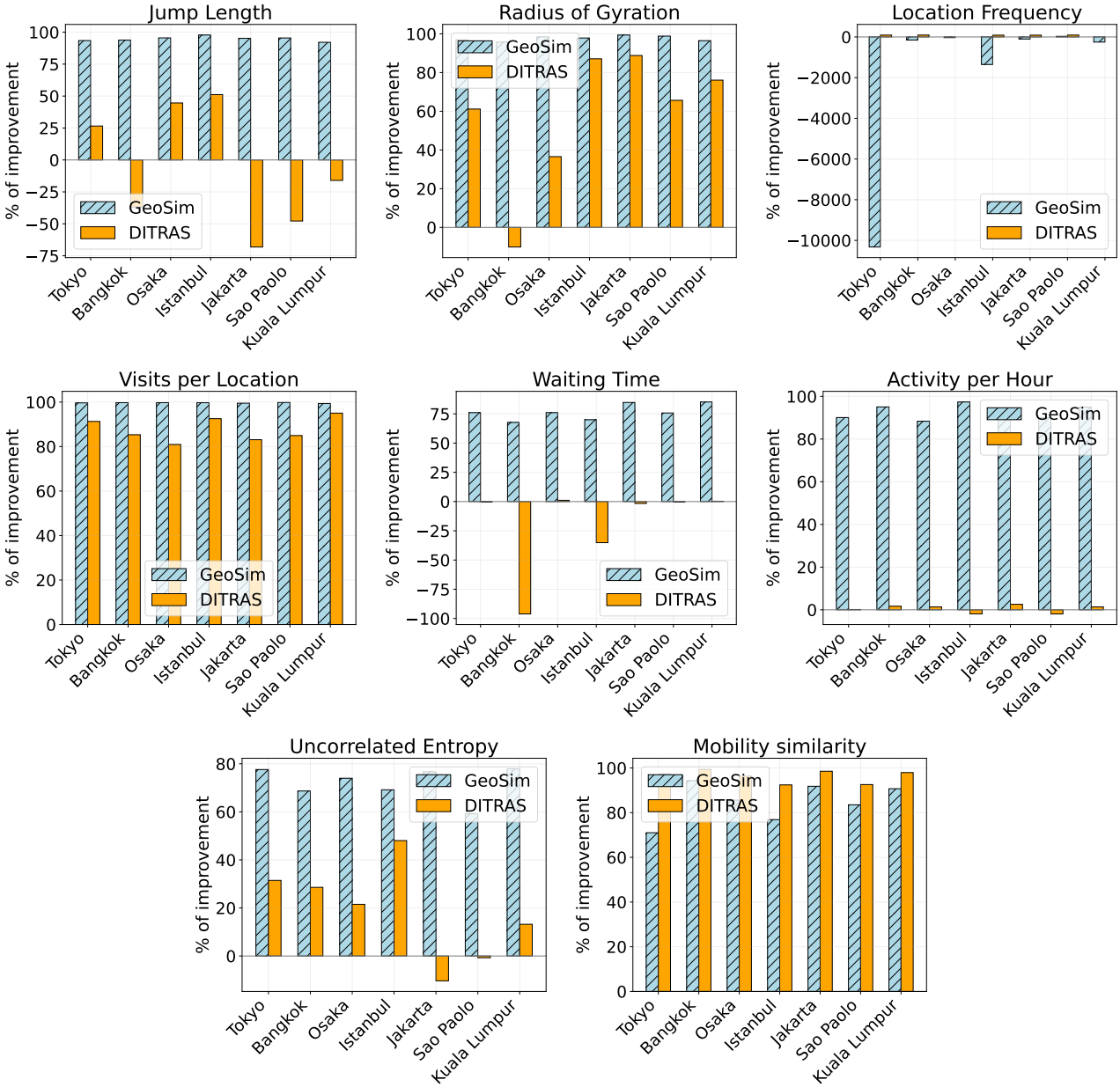


Figure 6. Average improvement in terms of KL divergence (percentage) achieved by using STS-EPR with respect to GeoSim (cyan bars) and DITRAS (orange bars). Note that 100% is an upper bound for the improvement, while there is no lower bound.

6.2. Temporal

The use of the mobility diary generator is essential to generate realistic temporal patterns. GeoSim fails to capture the distribution of the waiting times (KL close to 1 in all the cities, Table 4) because it extracts the waiting times from a pre-defined statistical distribution that does not consider the characteristic waiting time associated with a particular group. The two models that use the Mobility Diary Generator (MDG), STS-EPR, and DITRAS, can better capture this distribution (Figures 4f and 5f). In particular, STS-EPR achieves an average KL improvement of 76.50% and −18.87% with respect to GeoSim and DITRAS, respectively, (Figure 6e).

As for the imitation of people's circadian rhythm, GeoSim produces an unrealistic flat distribution in which the probability of moving is distributed uniformly across the day (Figures 4f and 5f). The two models that use the MDG, namely STS-EPR and DITRAS,

achieve similar KL scores and capture accurately the shape of the distribution of the activity per hour, as well as the peaks of activity during the day (Figures 4g and 5g).

6.3. Social

Our proposal, STS-EPR, is the model that best captures the distribution of mob_{sim}: it achieves KL $\in [0.0099, 0.0828]$ for all the cities but New York City, in which the KL score associated is 0.1874 (Table 4). Since DITRAS does not employ any social mechanism during the generation of the trajectories, as expected, it cannot capture accurately the shape of the distribution (KL $\in [0.783, 1.814]$). GeoSim fails in reproducing values close to 1, achieving KL scores in the range $[0.098, 0.555]$ resulting in the second-best model after STS-EPR. STS-EPR achieves by far the best scores, with a striking KL = 0.0099 for Bangkok (Table 4).

Overall, using STS-EPR guarantees an average KL improvement, computed for the eight cities, of 84.84% and 96.34% (Figure 6h) concerning GeoSim and DITRAS, respectively.

6.4. Impact of Spatial Tessellation

We repeat our experiments for two tessellations that differ in tile shape and surface:

1. Squared tessellation with tiles of size 300 m and area 0.09 km^2;
2. Hexagonal tessellation with tiles of H3 resolution 9 and area 0.10 km^2;

The size of tiles in the hexagonal tessellation has an area close to the squared tessellation. We compute the squared tessellations using the `scikit-mobility` [44] library and the hexagonal tessellations using Uber H3 geospatial indexing (https://eng.uber.com/h3, accessed on 10 September 2021). Table 5 reports the number of squared and hexagonal tiles that cover each of the eight cities.

Table 5. A summary of the properties of the 16 weighted spatial tessellations that cover the eight cities considered during the experiments.

	# Relevant Tiles	
City	**sq. 300 m**	**hex. H3 res 9**
New York City (US)	6734	4498
Osaka (JP)	3690	3270
Kuala Lumpur (MY)	2221	1691
Sao Paulo (BR)	6962	5588
Jakarta (ID)	6160	5215
Bangkok (TH)	7016	5518
Istanbul (TR)	8976	5265
Tokyo (JP)	7417	5844

The choice of the spatial tessellation impacts the results. For the spatial measure jump length, in which distances are fundamental, instantiating STS-EPR with the squared tessellations produces an average KL decrease of 0.02 to the hexagonal one (Figures 7a). For the spatial measure radius of gyration in Bangkok and Kuala Lumpur, the hexagonal tessellation produces a lower KL score (Figure 7b). In general, STS-EPR instantiated on a squared tessellation produces trajectories whose radius of gyration is more similar to the real distribution lowering on average the KL score of 0.012. For other non-distance-based spatial measures, such as location frequency, a hexagonal tessellation produces the best results (Figure 7c). Regarding the number of visits per location (Figure 7d), they are reproduced the best when STS-EPR is instantiated with a squared tessellation. The temporal measures are not affected by the choice of the spatial tessellation, since they do not depend on the space (Figure 7e,f). Similarly, the predictability of the trajectories (entropy) is not affected by the tessellation choice. In contrast, the mobility similarity is reproduced the best with the use of a hexagonal tessellation (Figure 7h), with which STS-EPR produces trajectories with an average KL score decrease of 0.007.

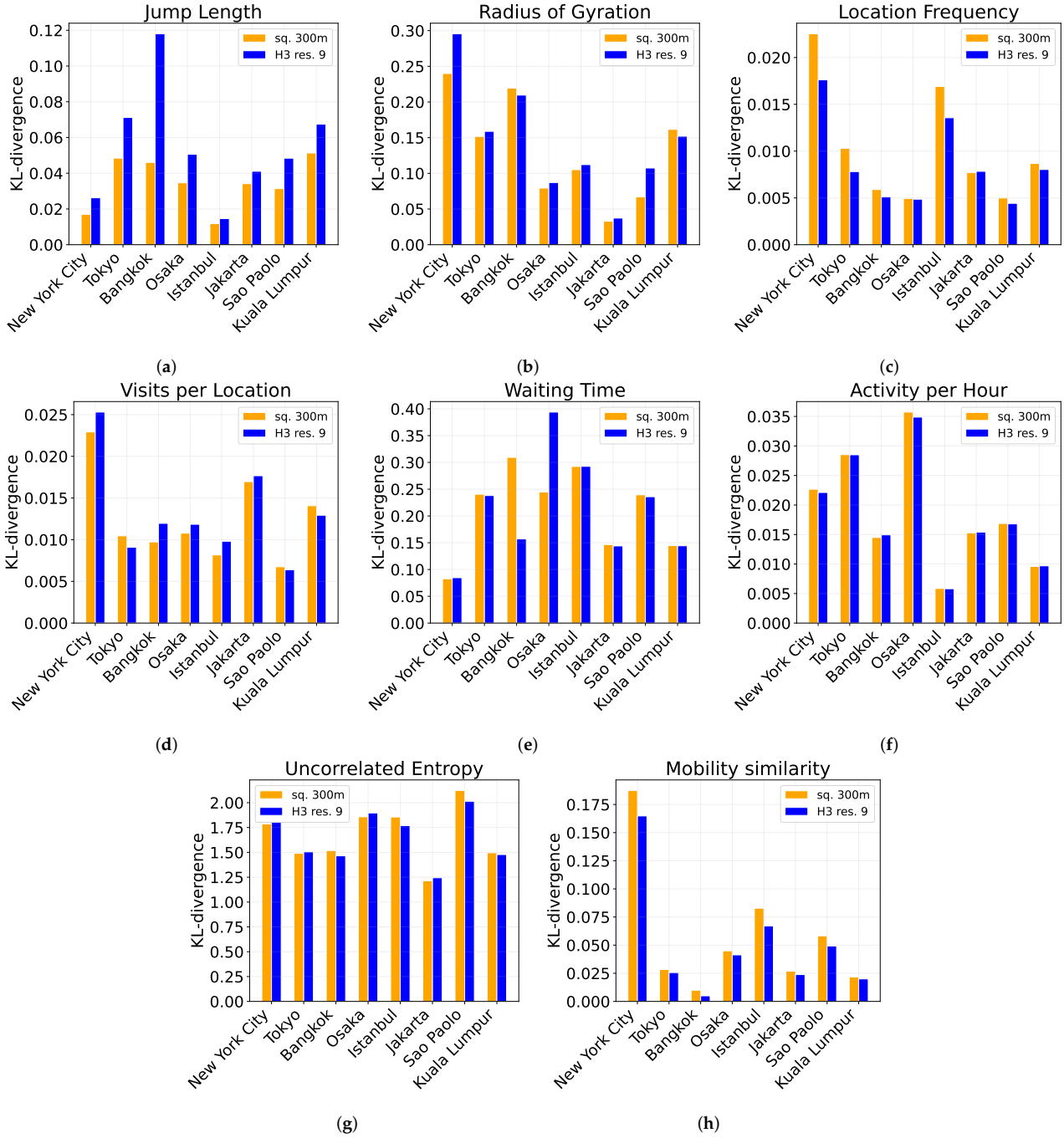


Figure 7. Average KL divergence of STS-EPR for jump length (**a**), radius of gyration (**b**), location frequency (**c**), visits per location (**d**), waiting time (**e**), activity per hour (**f**), uncorrelated entropy (**g**), and mobility similarity (**h**). Each group of bars indicate a city (New York City, Tokyo, London, Chicago, Los Angeles, and Madrid), each coloured bar within refers to the spatial tessellation used during the experiments.

6.5. Discussion

The inclusion of a mechanism that couples spatial distance and location relevance allows significantly improving the realism to spatial measures such as jump length, the radius of gyration, and location relevance. Indeed GeoSim, which does not use the gravity law, cannot capture at all the shape of these distributions.

Although DITRAS and STS-EPR both use the EPR and gravity law as spatial mechanisms, STS-EPR significantly captures better the number of visits per location and the location frequency. This result occurs due to the inclusion in STS-EPR of the social mecha-

nism, which allows an agent to visit locations far from its current position if a social contact visited such a location. Concerning the number of visits per location, the social mechanism rewards relevant locations. A relevant location for an individual may become relevant also for its social contacts when they choose to perform a social action.

The use of the mobility diary generator is crucial to capture the temporal patterns of human mobility. While the usage of the waiting time distribution in GeoSim cannot model the temporal characteristic of the population, using the diary generator allows reproducing both the waiting time and the propensity to move at specific times during the day.

STS-EPR does not depend on the specific characteristics of the geographic area since it produces good quality trajectories in all the examined cities, i.e., it is geographically transferable.

Finally, the combination of realistic social and temporal mechanisms allows STS-EPR to reproduce the mobility similarity realistically. Although GeoSim embeds a social mechanism, its results are comparable to those of DITRAS, which does not embed any social mechanism. This result highlights the importance of sociality: though often neglected in generative mobility models, it is essential to model properly individual human mobility and capture mobility patterns accurately. Indeed, the inclusion of the social mechanism allows capturing the spatial aspects better, as showed by the results achieved by the models for the visits per location measure.

7. Example of Execution

The code of STS-EPR is included in the `Scikit-mobility` (https://github.com/scikit-mobility, accessed on 10 September 2021) Python library. Listing 1 shows how to generate a set of synthetic mobility trajectories using STS-EPR.

In lines 1, 2, and 3 we perform the basic imports to guarantee the correct execution of STS-EPR. In lines 5 and 6 we specify the time interval of the simulation. In lines 12, 15, and 18, we load the weighted spatial tessellation, the social graph, and the mobility diary generator, respectively. Finally lines 21 and 24 instantiate STS-EPR and generate the synthetic trajectories.

The full example of the instantiation and generation of synthetic mobility trajectories using the STS-EPR model and the input files used are available at https://jovian.ai/giuliano-cornacchia/example-sts-epr (accessed on 10 September 2021).

Listing 1: STS-EPR Python example.

```
from skmob.models.sts_epr import STS_epr
import cloudpickle as cp
from urllib.request import urlopen

start = pandas.to_datetime('2012/04/10 00:00:00')
end = pandas.to_datetime('2012/07/10 00:00:00')

#Load the pre-computed data to simulate mobility in New York City
#the full urls are included in the Jovian notebook

#weighted spatial tessellation
nyc_weighted_tex = cp.load(urlopen('squared_tex_300m_nyc.pickle'))

#social graph
nyc_graph = cp.load(urlopen('social_graph_nyc.pickle'))

#mobility diary generator
nyc_diary_generator = cp.load(urlopen('mdg_nyc.pickle'))

#instantiate the model
sts_epr = STS_epr()

#generate the trajectories
syn_trajectories = sts_epr.generate(start, end, social_graph=nyc_graph,
                                    spatial_tessellation=nyc_weighted_tex,
                                    diary_generator=nyc_diary_generator,
                                    random_state=2021,
                                    relevance_column='relevance')
```

8. Conclusions

STS-EPR is a mechanistic data-driven, generative mobility model that embeds the spatial, temporal, and social dimensions together. Our results show that, overall, the modelling of the three dimensions together brings several advantages, making STS-EPR better than existing models that lack either the social, the spatial, or the temporal mechanism.

STS-EPR is particularly suitable in the field of computational epidemiology, in which sociality and mobility are the key factors in the spreading process of a disease. Simulating epidemics may help policymakers make crucial decisions about non-pharmaceutical interventions (e.g., imposing mobility reduction or social distancing).

STS-EPR also has some weaknesses. First, the mechanisms embedded in the model can capture a limited set of mobility measures, and the realism in the distribution of some measures must be improved. Future works should consider adding features in STS-EPR to capture out of the routine trips and the environmental spatial constraints imposed by buildings and road infrastructures. Another opportunity for future improvements consists of embedding deep learning techniques (e.g., Generative Adversarial Networks and Variational Autoencoders) to model aspects of mobility that are not captured by the current mechanisms, to improve the realism of the model.

In the meantime, our model is a step towards the design of a mechanistic data-driven model that can capture all the aspects of human mobility comprehensively.

Supplementary Materials: The following are available online at https://www.mdpi.com/2220-9964/10/9/599/s1, Table S1: Results H3 res 9; Figure S1: New York City sq. 300m; Figure S2: Tokyo sq. 300m; Figure S3: Bangkok sq. 300m; Figure S4: Osaka sq. 300m; Figure S5: Istanbul sq. 300m; Figure S6: Jakarta sq. 300m; Figure S7: Sao Paulo; Figure S8: Kuala Lumpur sq. 300m; Figure S9: New York City hex H3 res. 9; Figure S10: Tokyo hex H3 res. 9, Figure S11: Bangkok hex H3 res. 9; Figure S12: Osaka hex H3 res. 9; Figure S13: Istanbul hex H3 res. 9; Figure S14: Jakarta hex H3 res. 9; Figure S15: Sao Paulo hex H3 res. 9; Figure S16: Kuala Lumpur hex H3 res. 9.

Author Contributions: Conceptualization, Giuliano Cornacchia and Luca Pappalardo; Data curation, Giuliano Cornacchia; Funding acquisition, Luca Pappalardo Methodology, Giuliano Cornacchia and Luca Pappalardo; Project administration, Luca Pappalardo; Resources, Luca Pappalardo; Supervision,

Luca Pappalardo; Validation, Giuliano Cornacchia; Visualization, Giuliano Cornacchia; Writing—original draft, Giuliano Cornacchia and Luca Pappalardo; Writing—review & editing, Giuliano Cornacchia and Luca Pappalardo All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by H2020 SoBigData++ grant number 871042.

Data Availability Statement: The dataset used can be found at: https://sites.google.com/site/yangdingqi/home/foursquare-dataset. The STS-EPR code can be found at: https://github.com/scikit-mobility/scikit-mobility/blob/master/skmob/models/sts_epr.py. An example on how to use the code to generate trajectories using STS-EPR can be found at: https://jovian.ai/giuliano-cornacchia/example-sts-epr. Resources accessed on 10 September 2021.

Acknowledgments: This work has been supported by project H2020 SoBigData++ grant agreement 871042.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pepe, E.; Bajardi, P.; Gauvin, L.; Privitera, F.; Lake, B.; Cattuto, C.; Tizzoni, M. COVID-19 outbreak response, a dataset to assess mobility changes in Italy following national lockdown. *Sci. Data* **2020**, *7*, 1–7.
2. Kraemer, M.U.; Yang, C.H.; Gutierrez, B.; Wu, C.H.; Klein, B.; Pigott, D.M.; Du Plessis, L.; Faria, N.R.; Li, R.; Hanage, W.P.; et al. The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* **2020**, *368*, 493–497.
3. Oliver, N.; Lepri, B.; Sterly, H.; Lambiotte, R.; Deletaille, S.; De Nadai, M.; Letouzé, E.; Salah, A.A.; Benjamins, R.; Cattuto, C.; et al. Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle. *Sci. Adv.* **2020**, *6*, eabc0764.
4. Andrienko, G.; Andrienko, N.; Boldrini, C.; Caldarelli, G.; Cintia, P.; Cresci, S.; Facchini, A.; Giannotti, F.; Gionis, A.; Guidotti, R.; et al. (So) Big Data and the transformation of the city. *Int. J. Data Sci. Anal.* **2021**, *11*, 311–340, doi:10.1007/s41060-020-00207-3.
5. Huang, X.; Li, Z.; Jiang, Y.; Ye, X.; Deng, C.; Zhang, J.; Li, X. The characteristics of multi-source mobility datasets and how they reveal the luxury nature of social distancing in the U.S. during the COVID-19 pandemic. *Int. J. Digit. Earth* **2021**, *14*, 424–442, doi:10.1080/17538947.2021.1886358.
6. Rossi, A.; Barlacchi, G.; Bianchini, M.; Lepri, B. Modelling Taxi Drivers' Behaviour for the Next Destination Prediction. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 2980–2989.
7. Khaidem, L.; Luca, M.; Yang, F.; Anand, A.; Lepri, B.; Dong, W. Optimizing Transportation Dynamics at a City-Scale Using a Reinforcement Learning Framework. *IEEE Access* **2020**, *8*, 171528–171541.
8. Pappalardo, L.; Ferres, L.; Sacasa, M.; Cattuto, C.; Bravo, L. Evaluation of home detection algorithms on mobile phone data using individual-level ground truth. *EPJ Data Sci.* **2021**, *10*, 29.
9. Deville, P.; Linard, C.; Martin, S.; Gilbert, M.; Stevens, F.R.; Gaughan, A.E.; Blondel, V.D.; Tatem, A.J. Dynamic population mapping using mobile phone data. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 15888–15893.
10. Sîrbu, A.; Andrienko, G.; Andrienko, N.; Boldrini, C.; Conti, M.; Giannotti, F.; Guidotti, R.; Bertoli, S.; Kim, J.; Muntean, C.I.; et al. Human migration: The big data perspective. *Int. J. Data Sci. Anal.* **2020**, *11*, 341–360.
11. Bohm, M.; Nanni, M.; Pappalardo, L. Quantifying the presence of air pollutants over a road network in high spatio-temporal resolution. In Proceedings of the NeurIPS 2021 Workshop—Tackling Climate Change with Machine Learning, Online, 13–14 December 2020.
12. Nyhan, M.; Kloog, I.; Britter, R.; Ratti, C.; Koutrakis, P. Quantifying population exposure to air pollution using individual mobility patterns inferred from mobile phone data. *J. Expo. Sci. Environ. Epidemiol.* **2019**, *29*, 238–247.
13. Pappalardo, L.; Vanhoof, M.; Gabrielli, L.; Smoreda, Z.; Pedreschi, D.; Giannotti, F. An analytical framework to nowcast well-being using mobile phone data. *Int. J. Data Sci. Anal.* **2016**, *2*, 75–92.
14. Voukelatou, V.; Gabrielli, L.; Miliou, I.; Cresci, S.; Sharma, R.; Tesconi, M.; Pappalardo, L. Measuring objective and subjective well-being: Dimensions and data sources. *Int. J. Data Sci. Anal.* **2020**, *11*, 279–309.
15. Newlands, G.; Lutz, C.; Tamò-Larrieux, A.; Villaronga, E.F.; Harasgama, R.; Scheitlin, G. Innovation under pressure: Implications for data privacy during the Covid-19 pandemic. *Big Data Soc.* **2020**, *7*, 2053951720976680, doi:10.1177/2053951720976680.
16. Montjoye, Y.A.; Hidalgo, C.; Verleysen, M.; Blondel, V. Unique in the Crowd: The Privacy Bounds of Human Mobility. *Sci. Rep.* **2013**, *3*, 1376, doi:10.1038/srep01376.
17. Montjoye, Y.A.; Gambs, S.; Blondel, V.; Canright, G.; Cordes, N.; Deletaille, S.; Engø-Monsen, K.; García-Herranz, M.; Kendall, J.; Kerry, C.; et al. On the privacy-conscientious use of mobile phone data. *Sci. Data* **2018**, *5*, 180286, doi:10.1038/sdata.2018.286.
18. Pellungrini, R.; Pappalardo, L.; Pratesi, F.; Monreale, A. A Data Mining Approach to Assess Privacy Risk in Human Mobility Data. *ACM Trans. Intell. Syst. Technol.* **2017**, *9*, 31:1–31:27, doi:10.1145/3106774.
19. Pellungrini, R.; Pappalardo, L.; Simini, F.; Monreale, A. Modeling Adversarial Behavior Against Mobility Data Privacy. *IEEE Trans. Intell. Transp. Syst.* **2020**, doi:10.1109/TITS.2020.3021911.
20. Mir, D.J.; Isaacman, S.; Cáceres, R.; Martonosi, M.; Wright, R.N. DP-WHERE: Differentially private modeling of human mobility. In Proceedings of the 2013 IEEE International Conference on Big Data, Silicon Valley, CA, USA, 6–9 October 2013; pp. 580–588.

21. Fiore, M.; Katsikouli, P.; Zavou, E.; Cunche, M.; Fessant, F.; Hello, D.L.; Aivodji, U.M.; Olivier, B.; Quertier, T.; Stanica, R. Privacy in trajectory micro-data publishing: A survey. *arXiv* **2019**, arXiv:1903.12211.
22. Barbosa-Filho, H.; Barthelemy, M.; Ghoshal, G.; James, C.; Lenormand, M.; Louail, T.; Menezes, R.; Ramasco, J.J.; Simini, F.; Tomasini, M. Human mobility: Models and applications. *Phys. Rep.* **2018**, *734*, 1–74.
23. Luca, M.; Barlacchi, G.; Lepri, B.; Pappalardo, L. A Survey on Deep Learning for Human Mobility. *arXiv* **2021**, arXiv:2012.02825.
24. Karamshuk, D.; Boldrini, C.; Conti, M.; Passarella, A. Human mobility models for opportunistic networks. *IEEE Commun. Mag.* **2011**, *49*, 157–165.
25. Solmaz, G.; Turgut, D. A Survey of Human Mobility Models. *IEEE Access* **2019**, *7*, 125711–125731.
26. Hess, A.; Hummel, K.A.; Gansterer, W.N.; Haring, G. Data-driven human mobility modeling: A survey and engineering guidance for mobile networking. *ACM Comput. Surv. (CSUR)* **2015**, *48*, 1–39.
27. Tomasini, M.; Mahmood, B.; Zambonelli, F.; Brayner, A.; Menezes, R. On the effect of human mobility to the design of metropolitan mobile opportunistic networks of sensors. *Pervasive Mob. Comput.* **2017**, *38*, 215 – 232.
28. Brockmann, D.; Hufnagel, L.; Geisel, T. The Scaling Laws of Human Travel. *Nature* **2006**, *439*, 462–5, doi:10.1038/nature04292.
29. Gonzalez, M.C.; Hidalgo, C.; Barabasi, A.L. Understanding Individual Human Mobility Patterns. *Nature* **2008**, *453*, 779–82, doi:10.1038/nature06958.
30. Pappalardo, L.; Rinzivillo, S.; Qu, Z.; Pedreschi, D.; Giannotti, F. Understanding the patterns of car travel. *Eur. Phys. J. Spec. Top.* **2013**, *215*, 61–73, doi:10.1140/epjst/e2013-01715-5.
31. Pappalardo, L.; Simini, F. Data-driven generation of spatio-temporal routines in human mobility. *Data Min. Knowl. Discov.* **2017**, *32*, doi:10.1007/s10618-017-0548-4.
32. Cho, E.; Myers, S.; Leskovec, J. Friendship and Mobility: User Movement In Location-Based Social Networks. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 21–24 August 2011; pp. 1082–1090, doi:10.1145/2020408.2020579.
33. Toole, J.; Herrera-Yague, C.; Schneider, C.; Gonzalez, M.C. Coupling Human Mobility and Social Ties. *J. R. Soc. Interface/R. Soc.* **2015**, *12*, doi:10.1098/rsif.2014.1128.
34. Song, C.; Koren, T.; Wang, P.; Barabasi, A.L. Modelling the scaling properties of human mobility. *Nat. Phys.* **2010**, *6*, 818–823, doi:10.1038/nphys1760.
35. Pappalardo, L.; Simini, F.; Rinzivillo, S.; Pedreschi, D.; Giannotti, F.; Barabasi, A.L. Returners and explorers dichotomy in human mobility. *Nat. Commun.* **2015**, *6*, 8166, doi:10.1038/ncomms9166.
36. Song, C.; Qu, Z.; Blumm, N.; Barabasi, A.L. Limits of Predictability in Human Mobility. *Sciences* **2010**, *327*, 1018–1021, doi:10.1126/science.1177170.
37. Pappalardo, L.; Rinzivillo, S.; Simini, F. Human Mobility Modelling: Exploration and Preferential Return Meet the Gravity Model. *Procedia Comput. Sci.* **2016**, *83*, 934–939, doi:10.1016/j.procs.2016.04.188.
38. Barbosa, H.; de Lima-Neto, F.B.; Evsukoff, A.; Menezes, R. The effect of recency to human mobility. *EPJ Data Sci.* **2015**, *4*, 21, doi:10.1140/epjds/s13688-015-0059-8.
39. Alessandretti, L.; Sapiezynski, P.; Lehmann, S.; Baronchelli, A. Evidence for a Conserved Quantity in Human Mobility. *Nat. Hum. Behav.* **2018**, *2*, 485–491, doi:10.1038/s41562-018-0364-x.
40. Jiang, S.; Yang, Y.; Gupta, S.; Veneziano, D.; Athavale, S.; Gonzalez, M.C. The TimeGeo modeling framework for urban mobility without travel surveys. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, E5370–E5378, doi:10.1073/pnas.1524261113.
41. Zheng, Y.; Capra, L.; Wolfson, O.; Yang, H. Urban computing: Concepts, methodologies, and applications. *ACM Trans. Intell. Syst. Technol. (TIST)* **2014**, *5*, 1–55.
42. Zheng, Y. Trajectory data mining: An overview. *ACM Trans. Intell. Syst. Technol. (TIST)* **2015**, *6*, 29.
43. Yang, D.; Qu, B.; Yang, J.; Cudre-Mauroux, P. Revisiting User Mobility and Social Relationships in LBSNs: A Hypergraph Embedding Approach. In Proceedings of the 2019World WideWeb Conference (WWW '19), San Francisco, CA, USA, 13–17 May 2019; pp. 2147–2157, doi:10.1145/3308558.3313635.
44. Pappalardo, L.; Barlacchi, G.; Simini, F.; Pellungrini, R. Scikit-mobility: A Python library for the analysis, generation and risk assessment of mobility data. *arXiv* **2019**, arXiv:1907.07062.
45. Ouyang, K.; Shokri, R.; Rosenblum, D.S.; Yang, W. A Non-Parametric Generative Model for Human Trajectories. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), Stockholm, Sweden, 13–19 July 2018; pp. 3812–3817.
46. Eagle, N.; Pentland, A.S. Eigenbehaviors: Identifying structure in routine. *Behav. Ecol. Sociobiol.* **2009**, *63*, 1689–1689.
47. Wang, D.; Pedreschi, D.; Song, C.; Giannotti, F.; Barabasi, A.L. Human mobility, social ties, and link prediction. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 21–24 August 2011; pp. 1100–1108, doi:10.1145/2020408.2020581.
48. Fan, C.; Liu, Y.; Huang, J.; Rong, Z.; Zhou, T. Correlation between social proximity and mobility similarity. *Sci. Rep.* **2016**, *7*, 11975, doi:10.1038/s41598-017-12274-x.