An empirical study on human mobility and its agent-based modeling

J. Stat. Mech. (2012) P11024

(http://iopscience.iop.org/1742-5468/2012/11/P11024)

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

An empirical study on human mobility and its agent-based modeling

Tao Jia[1,2], **Bin Jiang**[2], **Kenneth Carling**[3], **Magnus Bolin**[3] **and Yifang Ban**[1]

[1] Division of Geoinformatics, Royal Institute of Technology (KTH), SE-100 44 Stockholm, Sweden
[2] Division of Geomatics, University of Gävle, SE-801 76 Gävle, Sweden
[3] School of Technology and Business Studies, Dalarna University, SE-781 88 Borlänge, Sweden
E-mail: jiatao83@hotmail.com, bin.jiang@hig.se, kca@du.se, mbo@du.se and Yifang@kth.se

Received 5 June 2012
Accepted 12 November 2012
Published 28 November 2012

Online at stacks.iop.org/JSTAT/2012/P11024
doi:10.1088/1742-5468/2012/11/P11024

Abstract. This paper aims to analyze the GPS traces of 258 volunteers in order to obtain a better understanding of both the human mobility patterns and the mechanism. We report the regular and scaling properties of human mobility for several aspects, and importantly we identify its Levy flight characteristic, which is consistent with those from previous studies. We further assume two factors that may govern the Levy flight property: (1) the scaling and hierarchical properties of the purpose clusters which serve as the underlying spatial structure, and (2) the individual preferential behaviors. To verify the assumptions, we implement an agent-based model with the two factors, and the simulated results do indeed capture the same Levy flight pattern as is observed. In order to enable the model to reproduce more mobility patterns, we add to the model a third factor: the jumping factor, which is the probability that one person may cancel their regular mobility schedule and explore a random place. With this factor, our model can cover a relatively wide range of human mobility patterns with scaling exponent values from 1.55 to 2.05.

Keywords: interacting agent models, scaling in socio-economic systems, stochastic processes

1742-5468/12/P11024+20$33.00

Contents

J. Stat. Mech. (2012) P11024

1. Introduction

Studies on human mobility have attracted extensive attention from broad research communities in the past few years. This is because, on one hand, understanding the behavior and the subsequent pattern of human mobility is very helpful in resolving many hot issues of our society, such as urban planning [1], traffic design and forecasting [2], performance evaluation of wireless mobile networks [3], and control and management of infectious disease spreading [4]. On the other hand, with the increasing advancement of location tracking technology, say GPS, we can collect mountains of mobility data. These enormous datasets not only cover the human mobility trajectory for walking with a hand-held GPS or mobile phone [5]–[7], driving with a car equipped with a GPS receiver [2], or even taking a flight with land monitoring system [8], but also stretch to the movements of users of location-based social network applications, such as Flickr [9] and Foursquare [1], which record detailed information wherever and whenever the users log onto the system.

Findings from the literature mainly suggested that the human mobility pattern exhibits a scaling property and, in particular, a Levy flight characteristic. For instance, Brockmann *et al* [10] reported the superdiffusive process of human travel, which was known as the Levy flight characteristic, through the tracking data for bank notes in the US. Gonzalez *et al* [5] found that the human displacement distribution could be well fitted with a truncated power law distribution using cell phone data and further interpreted this finding as the convolution of both population-based heterogeneity and the individual-based Levy flight characteristic. Using the taxicab GPS trajectory data, Jiang *et al* [2] observed the truncated power law distribution of human travel length and attributed it to the topological structure of the underlying urban road network. Similar findings were also reported by [7]: that Levy walks are statistically similar to human walks based on the GPS traces of 101 volunteers, and the authors further evaluated the performance of mobile networks using a truncated Levy walk model with the observed patterns. The above findings on human mobility also conform to the previous studies on the foraging trajectories of animals, such as albatrosses, spider monkeys and sea turtles [11]–[14].

However, there were other findings which indicated that the human mobility pattern could not be characterized as a Levy flight. Azevedo *et al* [15] analyzed, at the city level, several motion components of pedestrian movement in real scenarios and reported that the pause time followed a log-normal distribution. Analyzed by the AIC-based model selection method, the displacements of taxi trajectories were reported to follow an exponential distribution [16]. Moreover, Jiang and Jia [8] investigated human mobility at a country level, using US flight location data, and they found that the flight length could be better fitted with an exponential distribution from the underlying five potential models. From a pure statistical perspective, Scafetta [17] suggested an N-piece-fit Pareto distribution with increasing integer exponents instead of the power law with an exponential cutoff distribution for the human displacement using the datasets from [5] and [7]. With this new distribution, they were able to interpret the human mobility from the multi-scale cost model, but there were still not enough tests to support their argument.

Besides this, sophisticated mobility models were proposed for interpreting and reproducing the observed pattern. Lee *et al* [3] introduced the self-similar least action walk model (SLAW) to emulate human walk behavior, and their model adopted the geometric distance as a determinant factor in choosing the next location of the human walk. This is a relatively partial approach since some people may prefer to choose the destination according to its priority. Han *et al* [18] proposed a hierarchical model based on a square lattice to examine human mobility. Their model considered the priority of choosing the next location and captured well the scaling law of human displacement, and importantly they attributed the scaling property to the hierarchical organization of the traffic system. Unlike the above researchers, whose models which did not take into account the geographic space constraints, Jiang *et al* [2] proposed an agent-based model to simulate human movement on a large street network. Their model reproduced the observed scaling property, but they included a variable travel time which was generated from a power law distribution with the same exponent value as for the observed trail length distribution. Liu *et al* [19] also performed Monte Carlo simulations to interpret the observed trip pattern based on the LandScan population density map, and they suggested that the human mobility pattern could be influenced by geographical heterogeneity and distance decay in human travel.

Inspired from the animal foraging strategy [20], this research assumes that the underlying spatial points of interest (POI) might have a strong influence on the pattern of human mobility. On the other hand, animals in nature have little knowledge of the spatial distribution of available food [14], but humans have a mental map of the available resources in the geographical space. In this respect, we come up with another assumption: that the human preferential selection of available resources might also have a non-negligible influence on the mobility pattern. Our assumptions seem to be more similar to the hypotheses given by [2] and [18] than the others [3, 19]. This paper thereby sets up two steps for verifying these assumptions. We firstly extract the purposive locations (a substitute for the POI) from the GPS logger dataset which contains the mobility locations of 258 volunteers for around one month in Borlänge, Sweden. On the basis of the purposive locations, we subsequently carry out the statistical analysis of human mobility, and the result agrees well with the literature ones. Secondly, we propose a simple agent-based model for implementing the two assumptions. Simulation results from the model suggest that it conforms fairly well to the observed human displacement patterns. Moreover, by assigning each agent a jumping factor, we can tune the simulated scaling exponent of human mobility, which covers the range of most empirical observations [1, 5, 7, 10].

We structure the rest of this paper as follows. In section 2, the dataset and the procedure used to extract purposive locations are elaborated. In section 3, we carry out the statistical analysis of human mobility, and further construct three levels of purpose clusters to allow a better understanding of the underlying spatial structure. Then, we propose an agent-based model to mimic the characteristics of human mobility in section 4. Extension of the model and the related issues are discussed in section 5. Finally, we draw a conclusion in section 6.

2. The dataset and purposive location extraction

2.1. The dataset

In this research, we deployed 89 BT-338X, a Bluetooth GPS data logger that is a combination of a GPS receiver and a data logger with a Bluetooth interface, to volunteers for recording their daily movement during four periods in three sites of Borlänge, Sweden, including Domnarvet, Kvarnsveden and StoraTuna. These volunteers were recruited from four large sports associations, with high compliance and participation rates. BT-338X was usually attached to the private car by the volunteer for around one week, and the whole data collecting period lasted from 29 March to 15 May in 2011. In total, we obtained a dataset that includes 258 GPS logger files corresponding to 262 021 movement recordings of all volunteers with the removal of 5402 invalid records due to the loss of the GPS signal. It should be noted that each GPS logger file contains the movement information of one volunteer, and also that each record in the GPS logger file includes the information when the GPS signal is received every 5 or 30 s, such as the location in terms of longitude (x) and latitude (y), the time (t) and the velocity (v). The longitude and latitude are referenced using the World Geodetic System 84 (WGS 84) and measured with the accuracy of five meters according to the BT-338X user manual. Although most of the volunteers were residents of Borlänge County, the spatial extent of their movement covered more than

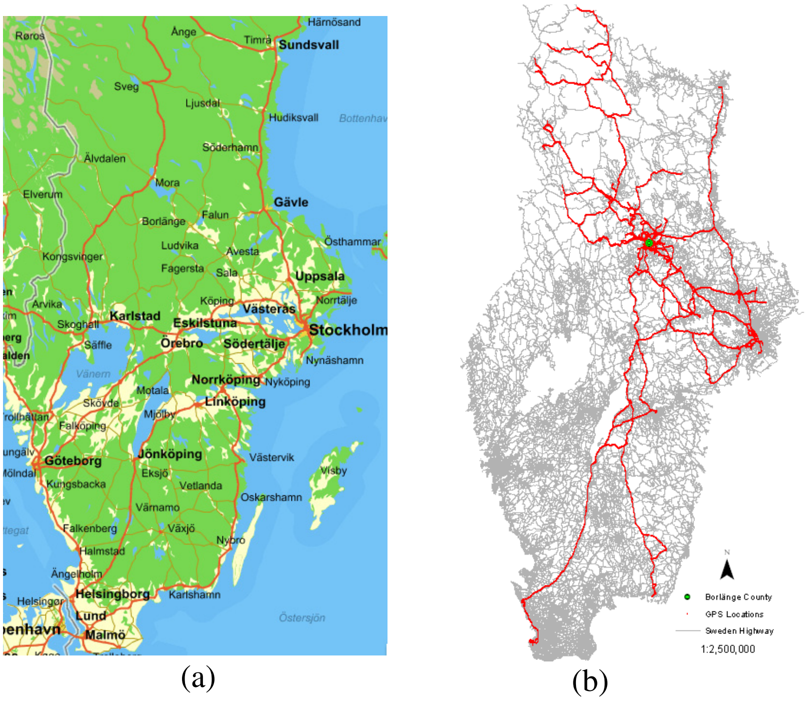


Figure 1. Maps for (a) the region covered by the mobility (source: enirio.se) and (b) the entire GPS location set overlaid on the Sweden highway system (source: openstreetmap.org).

half of the entire territory of Sweden (cf figure 1). In this respect, our dataset reflects a picture of human mobility with high spatial resolution in a relatively large geographical region.

2.2. Purposive location extraction

It is our assumption that purposive locations could be a good proxy for the spatial points of interest (POI), and that they might play a vital role in shaping the human mobility pattern. Moreover, it is our belief that purposive locations are hidden in the mobility data, and thus they can be extracted from the GPS logger dataset. From the perspective of computer vision, a purposive location resembles the interest point in a video sequence where a significant local variation occurs in both space and time [21]. Similarly, from the perspective of the human trajectory, it is characterized as a location with drastic change in time, distance or angle. This change can be identified from two features: large time interval and tortuous behavior. The former relates to situations where the time interval (t) between two consecutive locations exceeds the time threshold (ΔT) and also their distance apart (d) is less than the distance threshold (v^*t, where v is velocity), whereas the latter relates to situations where the angle (φ) formed by three successive locations is less than the angle threshold ($\Delta\varphi$) and those locations tend to cluster together. In reality, the two cases mostly occur when people go to the office or a shopping mall where no GPS signal can be received, or when they are outdoors with a GPS signal but wandered around something interesting. With the two rules, we illustrate identification of purposive

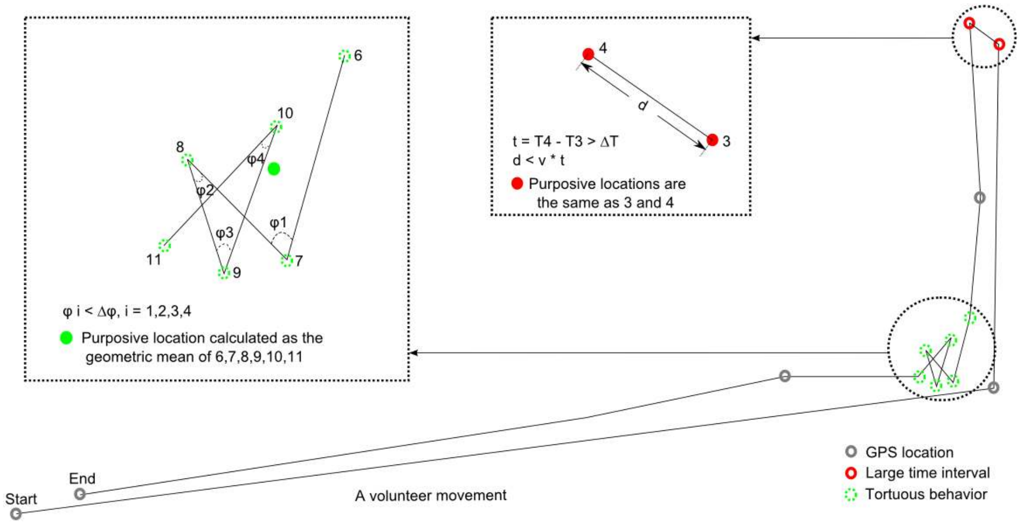


Figure 2. Illustration of identifying purposive locations in the human trajectory.

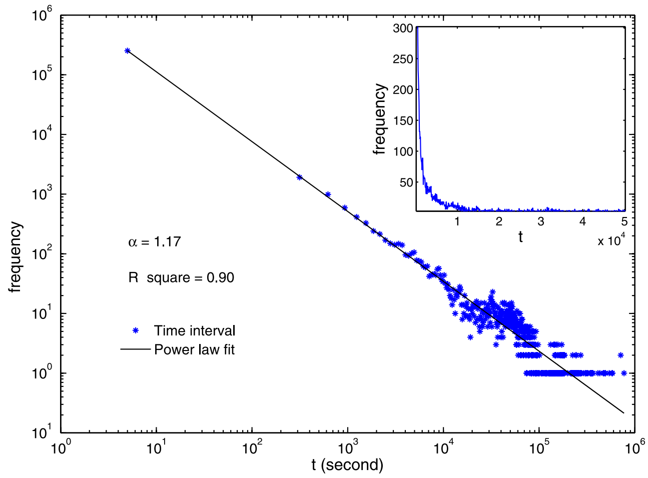


Figure 3. Log–log plot for the time interval (t) of the GPS logger dataset in a histogram with 2500 bins.

locations from the human trajectory in figure 2 and present corresponding pseudo-code in section A.1.

Specifying the values for the two thresholds (ΔT and $\Delta\varphi$) is not a trivial task, because sometimes a small variation in the threshold value might lead to large bias in the result. In this study, for the time threshold (ΔT), we explore the time interval (t) distribution of the whole dataset, and fortunately we find that it can be roughly assumed as a power law distribution (cf figure 3). With this knowledge, we adopt the arithmetic mean value as the time threshold according to the head/tail division rule [22]. The arithmetic mean value equals 550 s, which sounds reasonable in reality. As for the second threshold ($\Delta\varphi$), we are not so lucky: it does not approach a heavy tailed distribution. However, common knowledge from human movement behavior indicates that people tend to walk or drive in a relatively straightforward way rather than a curved backward way, to save both time and energy. This information hints that the angle threshold can be set as 90°, although bias can arise in the case of an extremely curved street. In addition, statistical information tells us that about 90% of total turning angles are greater than 90°. With these settings, we finally extracted 15 423 purposive locations.

3. Empirical analysis of human mobility

In this section, we explore two issues related to the patterns in human mobility. Firstly, we conduct a closely quantitative analysis of the movement path using measurements of factors like the home distance, gyration radius, purpose duration, purpose number and flight length. In particular, we put emphasis on the investigation of the flight length distribution which plays an important role in characterizing human mobility. Secondly, we derive the purpose clusters and examine their properties, which may have a strong influence on forming the pattern of human mobility.

3.1. Characteristics of human mobility

3.1.1. The home distance pattern and gyration radius. Thanks to having the home addresses of our volunteers, we can explore individual home distance (m) patterns. Home distance measures the distance from the current location to the home at time t, and its change with time can reflect the regularity of daily movement. Here we present the home distance pattern of an anonymous volunteer in figure 4(a), from which we can clearly observe one long trip accompanied by the majority regular movement pattern in terms of away from home (at work) during the day time and staying at home at night. On the other hand, we derive the gyration radius for each volunteer with the formula $r = \sqrt{(\sum(x_t - x_{\mathrm{mean}})^2 + (y_t - y_{\mathrm{mean}})^2)/N}$, where (x_t, y_t) is the position at time t, $(x_{\mathrm{mean}}, y_{\mathrm{mean}})$ is the average position of the movement and N is the total number of positions. It is well known that the gyration radius can measure the extent to which the volunteer has traveled in geographical space, and thus we present the gyration radius distribution for all volunteers in figure 4(b). We find that it can be approximated by a power law distribution with exponent value equal to 1.93 and P value equal to 0.11 ([23]; cf section A.2), which spans almost two decades of magnitudes and explains about 63% of the empirical data. This finding is roughly consistent with the literature [5, 24] although with a slightly high exponent value, and importantly it indicates the heterogeneity of individual movements; for example, most people travel a short distance whereas a few of them take a long trip.

3.1.2. The daily purpose count and purpose duration. It is assumed that human movement could be associated with the purposive locations, and two basic questions about this concern how many purposes one volunteer may have in her/his daily journey and how long it takes for her/him at each purposive location. The two questions are helpful for understanding the interaction between human mobility and the underlying urban structure. To answer the first question, we present the distribution of the daily purpose count for each volunteer in figure 5(a). From this figure, we clearly observe that it can be fitted very well with an exponential distribution with λ value equal to 0.13 and P value equal to 0.45 [25]. In this respect, the number of personal daily purposes can be better modeled as the Poisson process, which states that on average there are around

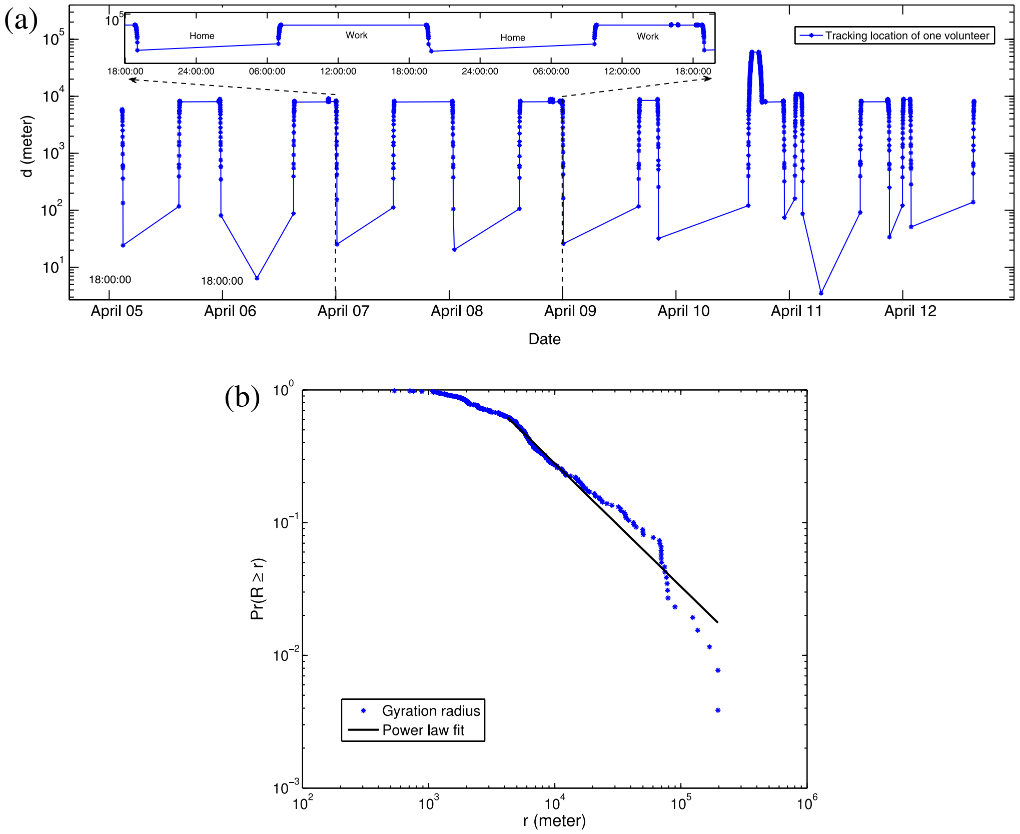
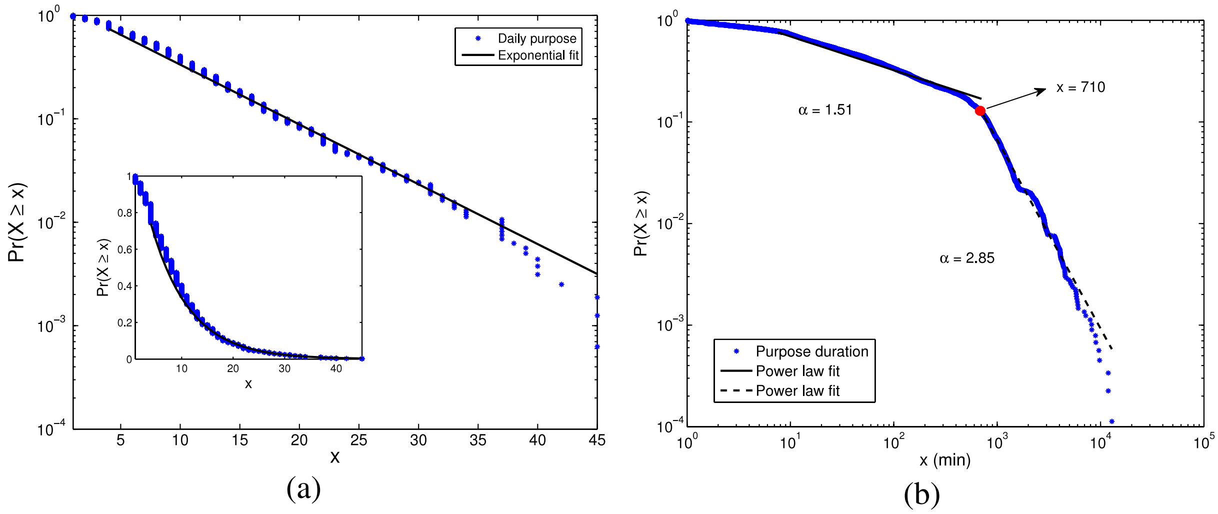


Figure 4. (a) Semi-log plot of an anonymous home distance pattern and (b) log–log plot of the gyration radius distribution.

$1/\lambda \approx 7$ purposes that one may have in one day. This finding is in line with [35] which assumes that individuals arrange their daily agenda independently in a random way.

The second question relates to the personal duration at each purposive location, and it has also been investigated in other literature with different contexts [7, 15, 16]. This measurement has a great influence on the diffusion [6, 7] of human mobility, and hence it plays an important role in infectious disease control and urban planning. We show the purpose duration distribution in figure 5(b), where a double-power-law distribution is identified with exponent values equal to 1.51 and 2.85 respectively. This observation is similar to the distribution of pause time in [7] and interevent time in [16], but is different from the log-normal distribution of pause time in [15] and the exponential distribution of elapsed time in [16]. Besides, we notice that the time duration at around 12 h partitions the whole distribution into two power law components. The first component with small exponent value is likely to reflect the scaling pattern of common human mobility, such as the daily shopping time, work time or sleep time. On the other hand, the second component with large exponent value is more likely to indicate the scaling pattern of uncommon human mobility, such as a long time of travel by train or air with the device turning off. Furthermore, the double-power-law behavior of the purpose duration may be explained by a mechanism similar to the multi-scale cost model proposed in [17].



(a) (b)

Figure 5. Distribution of (a) daily count of purposes in a semi-log plot and (b) purpose duration in a log–log plot.

3.1.3. Flight length. In this part, we examine the distribution of flight length, which characterizes the mobility pattern and has shown many advantages. For instance, previous research [11, 12, 14] on animal foraging has found the Levy flight characteristic of animal mobility and suggested that it could drastically increase the chance of finding food or resources, whereas recent study on human mobility has also reported this property [7] and found that it could significantly improve the routing performance of mobile networks. However, it is not a trivial thing to obtain the flight from the movement path, although it belongs to a part of the movement path. Thanks to the purposive locations that we elaborated above, we can define the flight as the path between any two successive purposive locations of one volunteer. Note that the path can be defined by two forms: the straight line and the shortest street segments. The former may be coined as 'air fly' which is the straight line connected by two consecutive purposive locations and has been involved in many studies [5, 7, 10], whereas the latter can be named as 'street fly' which is the shortest path along the street network from the starting purposive location to the ending purposive location and has recently been investigated by [2]. To make a comparison and understand the role of the street network on human mobility, we hereby examine both of the definitions (cf figure 6(a)).

Computation of the street fly distance is harder than that of the air fly distance, because we have to calculate the shortest path. Moreover, a conventional shortest path algorithm, like the Dijkstra algorithm, is extremely time-consuming for such a large street network with 780 512 streets, and we have to resort to the more efficient A^* search algorithm which was first proposed by Hart *et al* [26]. In total, we obtain 15 107 pairs of air fly and street fly distances from the purposive locations, and we present their distributions in figure 6(b). From the following table (cf table 1), we notice that both measurements can be better approximated by a power law with an exponential cutoff distribution, say $p(x) \propto x^{-1.93} * \mathrm{e}^{-0.000\,004x}$ for the air fly distance distribution and $p(x) \propto x^{-1.94} * \mathrm{e}^{-0.000\,005x}$ for the street fly distance distribution (cf figure 7(b)). Importantly, in this study, the two

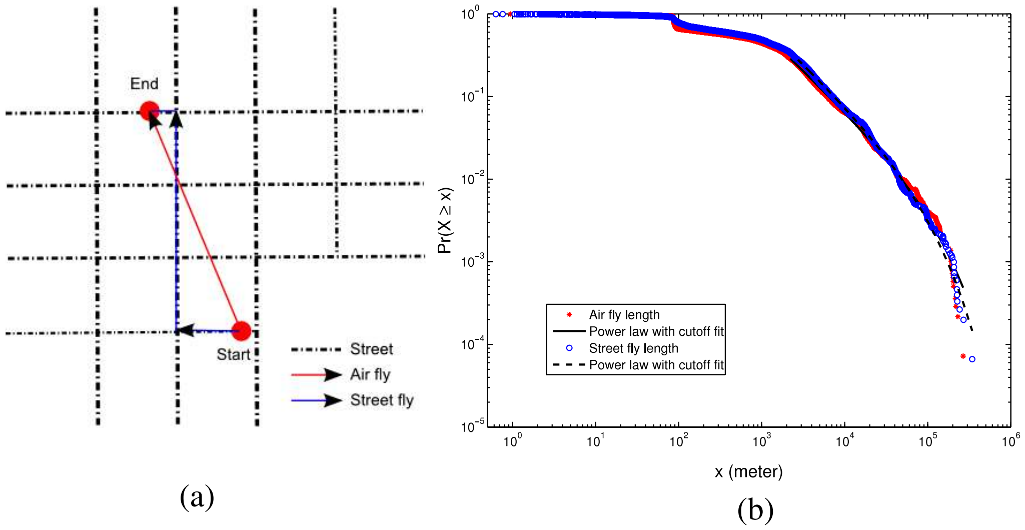


Figure 6. (a) Illustration of air fly and street fly and (b) the log–log plot of their length distribution.

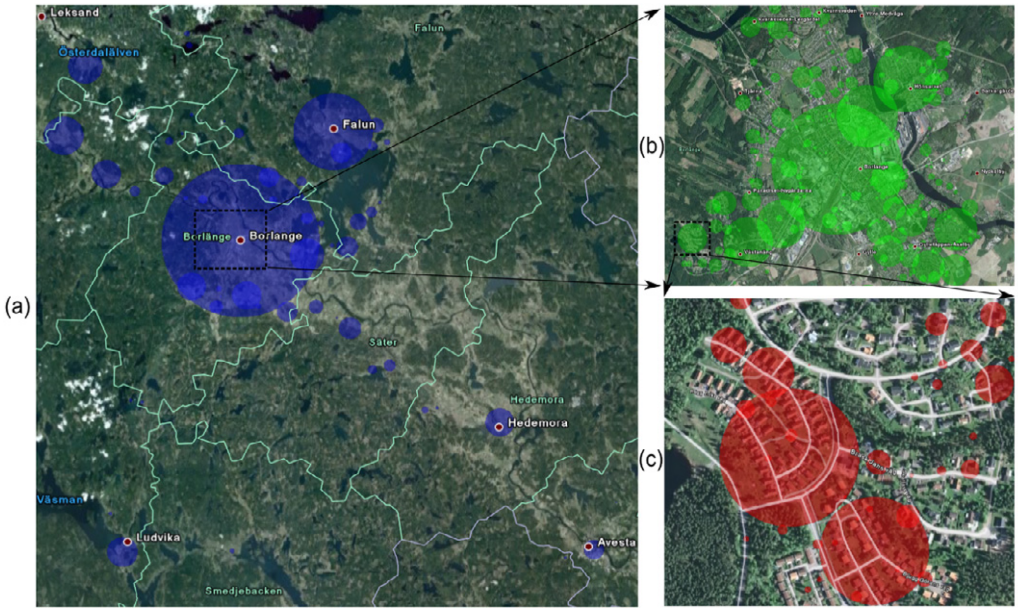


Figure 7. Enlarged maps for three levels of purpose clusters on Google Earth$^{\mathrm{TM}}$: (a) high level, (b) middle level and (c) low level.

measurements seem to play a similar role in characterizing the human mobility pattern, and hence we adopt the air fly distance as the flight length due to its simplicity and constant direction. Consequently, our finding supports the Levy flight property of human mobility, and it is generally consistent with most of the previous studies [2, 5, 7, 10], although the exponent value is larger than the ones in a large scale space [5, 10] and is smaller than the one in a small scale space [2].

Table 1. Flight length in terms of street fly and air fly. (Note: VTS stands for Vuong's test statistic [23, 34]—the smaller the VTS, the more plausible the competing model—and p is the significance level.)

Flight length	Exponential		Log-normal		Stretched exponential		Power law with cutoff	
	VTS	p	VTS	p	VTS	p	VTS	p
Street fly	−9.0	0	−137.4	0	−124.4	0	−10 500.1	0
Air fly	−3.0	0	−149.6	0	−120.3	0	−8 708.7	0

3.2. Purpose clusters

Purposive locations are assumed to be the product of human interaction with the underlying spatial structure, and they should be associated with the underlying spatial points of interest (POI), such as work place, gas station, parking lot, etc. In this respect, purposive locations may serve as the underlying spatial structure on which human mobility relies, and hence it is important to investigate the property of the purposive location in an aggregated way to obtain a better understanding of the human mobility. Here we adopt the entropy-based hierarchical clustering method [27], which selects the best level of clustering with the maximum entropy change through an iterative decomposition of the triangular irregular network (TIN) model of the spatial locations. An advantage of this method is that no parameter is needed to derive the clusters, and with this method we firstly classify the entire purposive locations into three levels of clusters (1 = high, 2 = middle and 3 = low). It is supposed that clusters in different levels might represent different levels of geographical entities—for instance, high level clusters corresponding to entities of city or town, middle level clusters corresponding to the city districts and low level clusters corresponding to the city blocks. For a better demonstration, we visualize the three levels of clusters in Google Earth$^{\mathrm{TM}}$ (cf figure 7), where each cluster is symbolized as a circle whose radius is proportional to the number of purposes.

In addition, a general glance at figure 7 gives an impression of the heterogeneity in terms of the radii of the circles. This image strengthens our intuition that the clusters in each level might have a scaling property. As expected, the size of clusters in each level is indeed found to obey a power law distribution (cf figure 8), and here the size refers to either the number of volunteers or purposes (note that one volunteer may have multiple purposes in one cluster). We present the detailed parameters of the power law model in table 2. This finding suggests that the scaling property of purpose clusters might play a significant role in shaping the observed pattern of human mobility. Our explanation is roughly similar to the ones given by [2] and [18], and it also agrees well with the arguments from [11]. However, to make our point more convincing, we further verify it in the agent-based simulation of section 4.

4. Agent-based modeling of human mobility

Agent-based modeling has been widely used in many applications [28] and considered as an effective tool for capturing the emergent collective behavior due to the simple individual

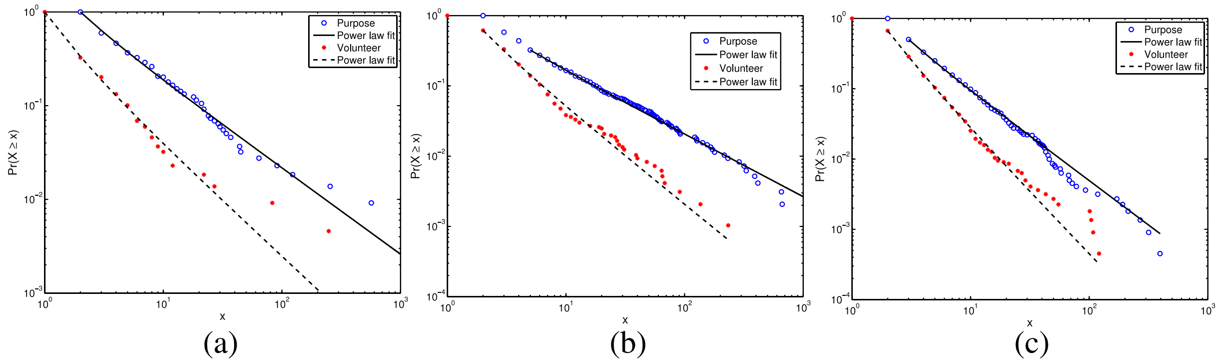


Figure 8. Log–log plots for the three levels of clusters in terms of both number of events and number of volunteers ((a) first level, (b) second level and (c) third level).

Table 2. Table of the power law model for the three levels of clusters. (Note: for details about the P value, please refer to[23]; cf section A.2.)

Cluster	Number of purposes		Number of volunteers	
level	Alpha	P	Alpha	P
High	1.92	0.19	2.18	0.9
Middle	1.89	0.89	2.37	0.05
Low	2.26	0.40	2.76	0.32

behavioral rules [29]. To mimic the observed Levy flight characteristic of human mobility, we illustrate agent-based modeling in this section. Generally, our model is composed of two components: the hierarchical purpose cluster graph and the agent mobility behavior. The first component is constructed on the basis of the three levels (high, middle and low) of purpose clusters and is aimed at setting the spatial structure of the mobility of the agents. In this graph, two clusters are connected if they have the sibling (belonging to the same parent) or parent–child relationship, or if they both belong to the high layer clusters. Specifically, this hierarchical graph resembles our conventional thinking on traveling in a road network [30] or between different cities [18]. For instance, as shown in figure 9(a), people traveling from node a to node b have to go through the intermediate node A and B in the middle level. On the other hand, the agent has two behaviors bestowed on them: uniform and preferential, to choose the next destination. A uniform agent considers its neighbors with equal probability of being visited (cf figure 9(b)), while the preferential agent has a high probability of visiting the neighbors with large numbers of purposes (cf figure 9(c)). Therefore, it is our belief that two different images will emerge with respect to the two different behaviors.

Before delving into the simulation to uncover the two different images, it is necessary to discuss an important characteristic of this model from the perspective of the random walk [31], namely the probability of the edge being visited. This is because the main concern of this study is the mobility displacements (flights) between consecutive nodes visited by the agent. In this respect, the proposed model can be regarded as a random

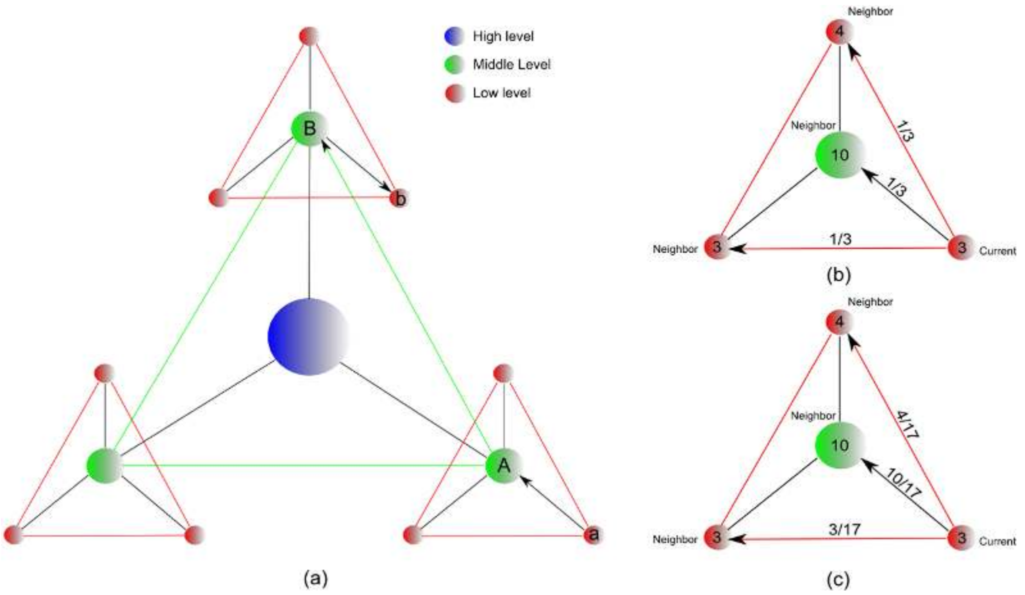


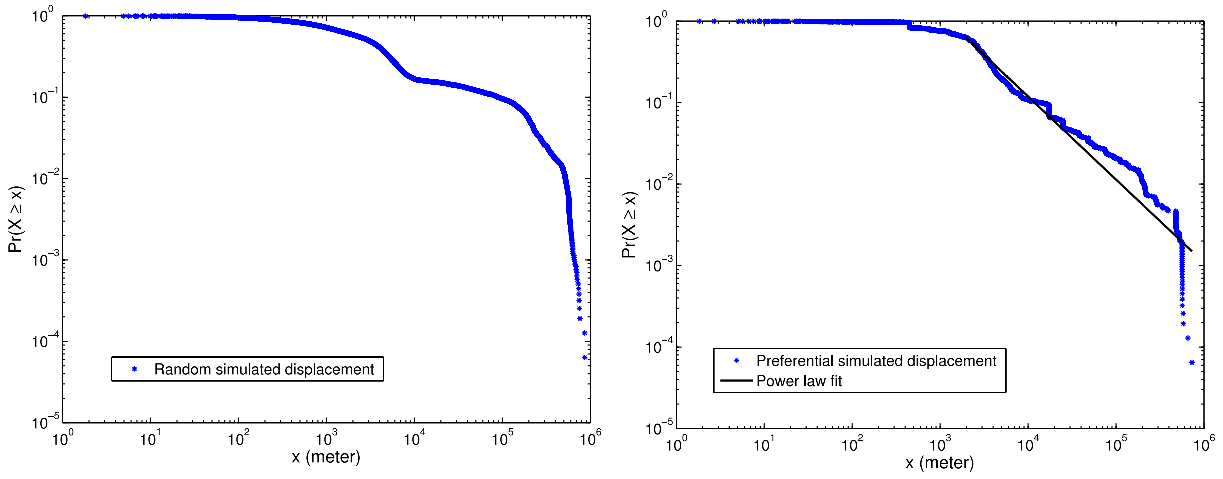
Figure 9. Demonstration of a synthetic three-level hierarchical human mobility model.

walk model in a hierarchical graph [31]. Moreover, it reflects a stochastic process in which an agent continuously visits the neighboring node based on the current node uniformly or preferentially. Here, the probability of the graph edge ($p(e_{ij})$) being visited by the uniform agent or the preferential agent can be determined roughly, and it is equal to the probability of the node v_i being visited ($p(v_i)$) multiplied by the probability of the node v_j being visited subsequently ($p(v_j|v_i)$). For the uniform agent, we can derive the equation (1) where m is the number of edges in the graph, and hence each edge of the graph has the same probability of being visited [31], say $p(e) = 1/2m$:

$$p(e_{ij}) = p(v_i) * p(v_j|v_i) = \frac{d(v_i)}{\sum_{v\in V} d(v)} * \frac{1}{d(v_i)} = \frac{d(v_i)}{2m} * \frac{1}{d(v_i)} = \frac{1}{2m}. \tag{1}$$

On the other hand, for the preferential agent, we can express this as equation (2) below, where $N(v)$ represents the adjacent neighbors of node v and s_v denotes the corresponding number of purposes. Therefore, we can conclude that each edge has a probability of being visited proportional to the product value of the number of purposes of two ending nodes, say $p(e_{ij}) \propto s_i * s_j$. Through the above analysis, two facts can be obtained. One is that the random walk of the uniform agent can be considered as a special case of the random walk of the preferential agent with the number of purposes of each node set as 1, and the other one is that the edge connecting two nodes with a high number of purposes is much more often visited by the preferential agent than the uniform one, which conforms well to the real situation. Consequently, we conjecture that the preferential agent would outperform the uniform one.

$$\begin{aligned} p(e_{ij}) = p(v_i) * p(v_j|v_i) &= \frac{s_i * \sum_{k\in N(i)} s_k}{\sum_{v\in V} s_v * \sum_{k\in N(v)} s_k} * \frac{s_j}{\sum_{k\in N(i)} s_k} \\ &= \frac{s_i * s_j}{\sum_{v\in V} s_v * \sum_{k\in N(v)} s_k} = \frac{s_i * s_j}{C}. \end{aligned} \tag{2}$$



(a)

(b)

Figure 10. Simulated displacement length distribution (SN = 258 and ST = 60) for the (a) random and (b) preferential agents in log–log plots.

To verify the above conjecture, simulations are conducted for both the uniform agents and the preferential ones. On the basis of the statistical analysis of observed human mobility, we set the values of the two model parameters, namely the number of agents (SN) and simulation steps (ST), as the values of the observed number of volunteers (258) and the average number of purposes per volunteer (total number of purposes/number of volunteers $= 15\,423/258 \cong 60$) respectively. It sounds roughly reasonable to set the value of the second parameter in this way because of every step of agent movement corresponding to a purpose being achieved. With these settings, we present the simulated displacement length distribution in figure 10. The plot shown in figure 10(a) does not display a power law-like property, which might hint that the uniform agent could not mimic the Levy flight characteristic of human mobility. On the other hand, the plot in figure 10(b) displays a power law distribution with alpha equal to 2.02, which might suggest that the preferential agent does indeed capture the Levy flight characteristic of the human mobility. Moreover, we have verified that the simulation with the empirical settings has reached saturation (cf section A.3), which means that the simulation result comes to a stable status with little bias coming from the fluctuation of model parameters, such as ST or SN.

5. Discussion

We have shown that simulated human displacements can be approximated by a power law distribution with exponent value equal to 2.02, which does not deviate very much from the observed one: 1.93. This simulation result is based on two ingredients of the model: the structure of the hierarchical purpose clusters and the individual preferential behavior. Indeed, the underlying geographical structure has a significant influence on the mode and/or distance of human travel; for instance, most people would probably like to travel the short distances to neighboring places by car whereas some people might prefer to travel long distances to far places by flying, and this argument has also been suggested by other studies [2, 18]. On the other hand, the individual preferential behavior reflects

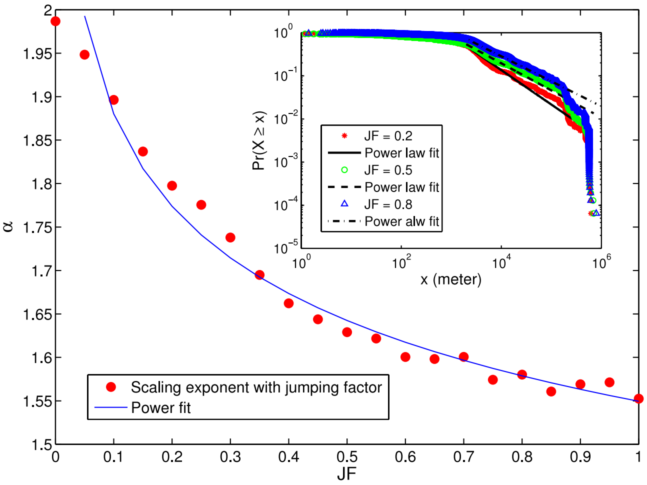


Figure 11. The relationship between the jumping factor (JF) and the exponent value. (Note: the inset is a log–log plot.)

the strategies adopted by people in the process of making decisions. For example, people are more likely to visit popular places because their needs can be better satisfied there. This is also similar to the process of preferential attachment [32] in society, where rich people get richer and poor people get poorer.

However, in reality, human mobility is so complex or diverse that no simple rule or model can fully capture its characteristics. This is why different studies have reported different scaling exponent values based on different mobility datasets, although they have all identified the Levy flight characteristic. Our model can mimic the observed Levy flight characteristic of volunteer mobility, but it lacks the power to generate a wide range of mobility patterns in terms of different exponent values. Given this complexity and shortcoming, we add another ingredient to the model to allow a better understanding of the human mobility pattern: the jumping factor (JF), which is similar to the damping factor in the PageRank algorithm [33]. This factor supposes that in reality one person might have a probability of canceling the regular mobility schedule and immediately make a decision to move to another place, and here it refers to the probability of going to a random low level cluster. To test the influence of the jumping factor on the mobility pattern, we carry out simulations for each specified jumping factor value and then obtain the corresponding exponent value.

In figure 11, we illustrate the relationship, and we find that it can be approximated very well by a power relationship ($\alpha = 1.55 \times \mathrm{JF}^{-0.084}$) with an R square value as high as 0.98. This indicates that the larger the value of the jumping factor, the more heterogeneous the human mobility. An explanation for this relationship can be given as follows: with a larger jumping factor, people will be less likely to be constrained to neighboring places and more likely to explore a random new place; this randomness may increase the probability of taking long journeys and consequently lead to the entire mobility pattern becoming more heterogeneous. In the extreme case, where the jumping factor is set to 1, the movement is a totally random walk on the low level of clusters. In general, our model can cover a relatively wide range of human mobility patterns with exponent values from 1.55 to 2.05.

In these respects, the proposed model strengthens the findings reported in [18] (that the scaling pattern of human mobility is mainly attributable to the hierarchical organization of the traffic systems) by constraining the human movement to a real geographic world instead of an ideal square lattice, and it also consolidates the argument in [2] (that the topological structure of the street network plays a vital role in shaping the patterns of human mobility) by removing the additional travel time variable following a power law distribution with the same exponent value as the observed human trail length distribution. Besides, the introduction of the jumping factor enables it to reproduce the empirical findings reported in a wide range of studies [5, 7, 10]. Therefore, the results in this study will not only help the research community by verifying and reproducing the reported empirical findings, but also be helpful in the fields of urban planning, traffic management and even infectious disease control.

6. Conclusion

In this paper, we analyze a GPS logger dataset including the movement recordings of 258 volunteers for a period of one month. We find both the regular and scaling properties of human mobility from several measurements, and we further report its Levy flight characteristic which is consistent with most previous studies. An explanation for the scaling properties of human mobility is given starting from two assumptions: (1) the scaling and hierarchical properties of the purpose clusters which serve as the underlying spatial structure, and (2) the preferential individual behavior. We subsequently implement the two assumptions in an agent-based model for a convincing confirmation.

We show that the simulated human displacements can be approximated by a power law distribution with exponent value equal to 2.02, which does not deviate very much from the observed one: 1.93. Besides, to enable the model to reproduce more mobility patterns, we add one more ingredient, the jumping factor, to the model. Through several simulations, we report a power relationship between the jumping factor and the simulated scaling exponent value. Importantly, with this factor, our model can cover a relatively wide range of human mobility patterns with different exponent values from 1.55 to 2.05.

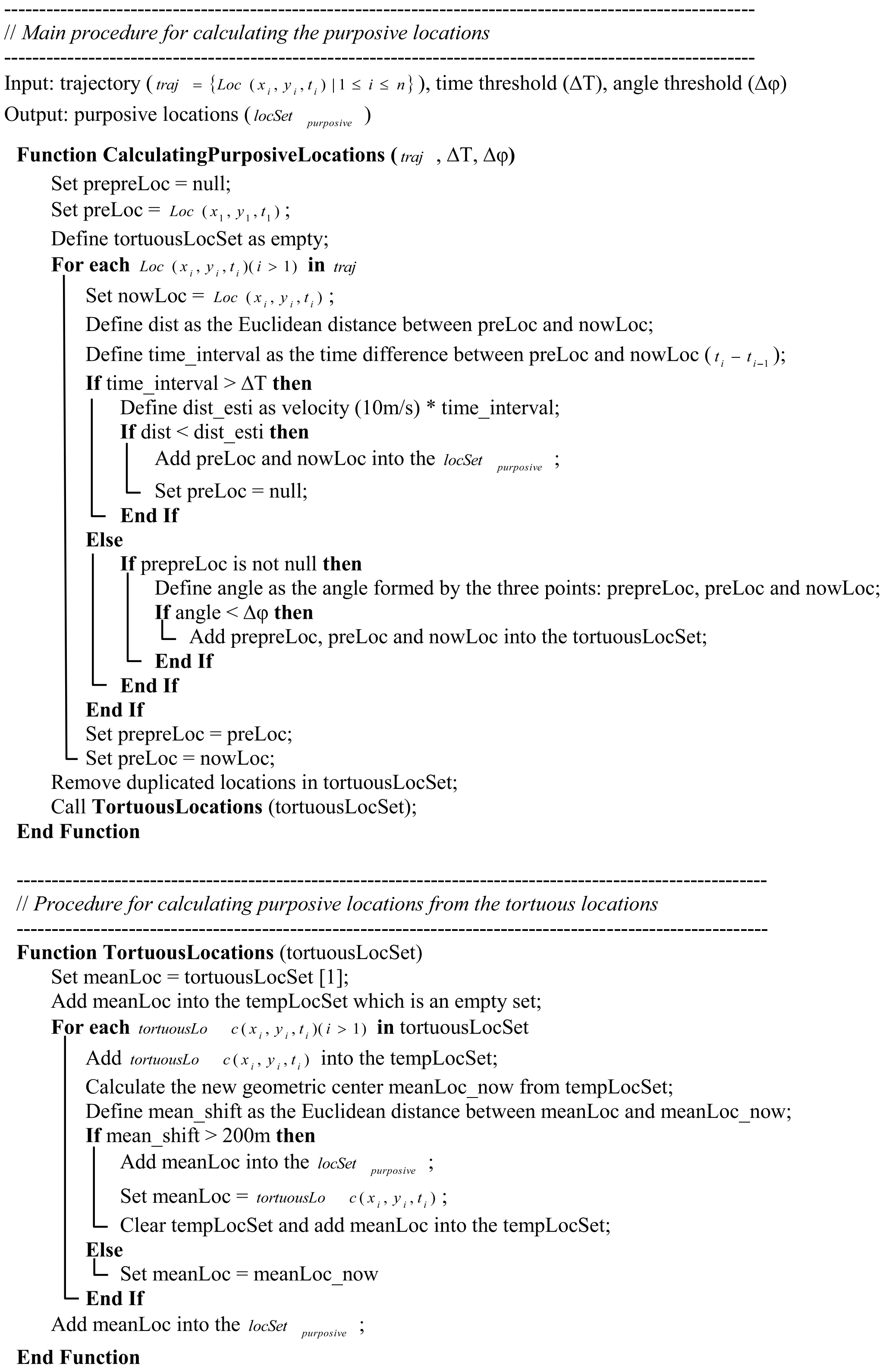
Acknowledgments

Financial support from the Swedish Retail and Wholesale Development Council is gratefully acknowledged. We thank the anonymous volunteers for their contribution to the GPS logger dataset, we thank Dr Aaron Clauset for kind help on how to test the power law distribution, and we also thank the anonymous referees for constructive comments.

Appendix

In this appendix, we supply the pseudo-code used to calculate the purposive locations in section A.1, the procedure for testing the power law model in section A.2, and the agent-based simulation saturation test in section A.3.

A.1. Calculating the purposive locations from a trajectory



```
--------------------------------------------------------------------------------
// Main procedure for calculating the purposive locations
--------------------------------------------------------------------------------
Input: trajectory ($traj = \{Loc(x_i, y_i, t_i) \mid 1 \le i \le n\}$), time threshold (ΔT), angle threshold (Δφ)
Output: purposive locations ($locSet_{purposive}$)
  Function CalculatingPurposiveLocations ($traj$, ΔT, Δφ)
      Set prepreLoc = null;
      Set preLoc = $Loc(x_1, y_1, t_1)$;
      Define tortuousLocSet as empty;
      For each $Loc(x_i, y_i, t_i)(i > 1)$ in $traj$
          Set nowLoc = $Loc(x_i, y_i, t_i)$;
          Define dist as the Euclidean distance between preLoc and nowLoc;
          Define time_interval as the time difference between preLoc and nowLoc ($t_i - t_{i-1}$);
          If time_interval > ΔT then
              Define dist_esti as velocity (10m/s) * time_interval;
              If dist < dist_esti then
                  Add preLoc and nowLoc into the $locSet_{purposive}$;
                  Set preLoc = null;
              End If
          Else
              If prepreLoc is not null then
                  Define angle as the angle formed by the three points: prepreLoc, preLoc and nowLoc;
                  If angle < Δφ then
                      Add prepreLoc, preLoc and nowLoc into the tortuousLocSet;
                  End If
              End If
          End If
          Set prepreLoc = preLoc;
          Set preLoc = nowLoc;
      Remove duplicated locations in tortuousLocSet;
      Call TortuousLocations (tortuousLocSet);
  End Function

  --------------------------------------------------------------------------------
  // Procedure for calculating purposive locations from the tortuous locations
  --------------------------------------------------------------------------------
  Function TortuousLocations (tortuousLocSet)
      Set meanLoc = tortuousLocSet [1];
      Add meanLoc into the tempLocSet which is an empty set;
      For each $tortuousLoc(x_i, y_i, t_i)(i > 1)$ in tortuousLocSet
          Add $tortuousLoc(x_i, y_i, t_i)$ into the tempLocSet;
          Calculate the new geometric center meanLoc_now from tempLocSet;
          Define mean_shift as the Euclidean distance between meanLoc and meanLoc_now;
          If mean_shift > 200m then
              Add meanLoc into the $locSet_{purposive}$;
              Set meanLoc = $tortuousLoc(x_i, y_i, t_i)$;
              Clear tempLocSet and add meanLoc into the tempLocSet;
          Else
              Set meanLoc = meanLoc_now
          End If
      Add meanLoc into the $locSet_{purposive}$;
  End Function
```

A.2. The procedure used to calculate the P value for testing the power law distribution

Given an observed dataset x, it is straightforward to denote its power law model as

$$p(x) = cx^{-\alpha}(x \geq x_{\min}, \alpha > 1) \tag{A.1}$$

where $x_{\min}$ is the smallest value above which the power law model holds and c and α can be calculated through the normalization process and by the maximum likelihood estimation (MLE) method respectively;

$$c = (\alpha - 1)x_{\min}^{\alpha-1} \tag{A.2}$$

$$\alpha = 1 + n\left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}}\right]^{-1}. \tag{A.3}$$

However, it cannot be guaranteed that the hypothesized power law distribution is a plausible fit to the data. In other words, we need to test the power law hypothesis quantitatively. According to the modified Kolmogorov–Smirnov (KS) test proposed by Clauset *et al* [23], a P value is calculated, to measure the plausibility of the power law model. In total, five steps are involved in this procedure.

(1) Obtain the power law model $p(x)$ for the observed dataset and calculate the KS statistic which is the maximum distance (D) between the cumulative distribution functions of the data $(G(x))$ and the fitted power law model $(P(x))$:

$$D = \max_{x \geq x_{\min}} |G(x) - P(x)|. \tag{A.4}$$

(2) Generate 2500 synthetic datasets with a similar distribution to the observed dataset, that is the values above $x_{\min}$ are drawn from a pure power law $p(x)$ and the values below $x_{\min}$ are uniformly selected from the observed dataset with values below $x_{\min}$.

(3) For each synthetic dataset, obtain its power law model $s(x)$ and calculate its KS statistic as D_i.

(4) The P value is calculated as the probability of synthetic datasets having D_i greater than D:

$$P = \frac{\text{the number of } D_i \text{ greater than } D}{2500}. \tag{A.5}$$

(5) Retain the hypothesized power law model if the P value is greater than or equal to the criterion or threshold value 0.05; or reject it if the P value is less than the criterion or threshold value 0.05.

Note that there is a difference between the P value used in the conventional hypothesis testing and the P value adopted in this study. Generally speaking, the difference is in how we use it. In the conventional hypothesis testing, the 'goal' is to get a small P value so that one can reject the null hypothesis; whereas in our study, the 'goal' is to fail to reject the null hypothesis, which is that some data can be plausibly fitted by the model suggested in the null hypothesis, for example, the power law model. Thus, a high P value is interesting for us, while in the more conventional use of hypothesis testing a low P value is interesting. Besides, as regards the criterion or the threshold value, the authors [23] argued that most researchers would like to set it as either 0.1 or 0.05, although they further stressed that

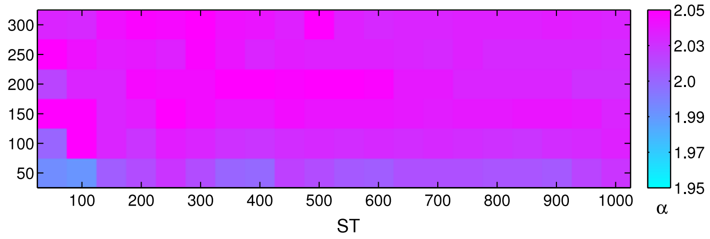


Figure A.1. Saturation status test for the parameters in agent-based simulation.

it depends on the judgment of the researcher in a particular situation at hand. In this study, we set it as 0.05, which is similar to the approach of a previous study [2]. In other words, the threshold value 0.05 means that the null hypothesis can be retained if there is a probability of 5% or more that the synthetic datasets with the same size drawn from the hypothesized distribution would have a KS statistic value larger than that of the observed dataset.

A.3. The agent-based simulation saturation test

In this part, we supply the agent-based simulation saturation test to check whether the simulation under the empirical settings in section 4 reaches saturation status. Here saturation means that the simulation result comes to a stable status with little bias coming from the fluctuation of model parameters, such as ST or SN. Besides this, there are many ways of measuring the degree of saturation [2, 8], but we adopt a naive exhaustive way to visualize the effect of parameter change on the simulation result.

We start with the simulation by setting the parameter SN as 50 and ST as 50, and then the next simulation with an increment of 50 in SN or ST until SN reaches 300 and ST reaches 1000. Thus, the parameter space covers the extent of empirical values and has a total number of 120 combinations (SN $\in \{50, 100, \ldots, 250, 300\}$ and ST $\in \{50, 100, \ldots, 950, 1000\}$). From figure A.1, we can clearly see that the scaling exponent values fall within the range from 1.95 to 2.05 for all the combinations, and consequently, we conclude that the simulation with the empirical settings reaches the saturation status.

References

[1] Noulas A, Scellato S, Lambiotte R, Pontil M and Mascolo C, *A tale of many cities: universal patterns in human urban mobility*, 2012 *PLoS ONE* **7** e37027

[2] Jiang B, Yin J and Zhao S, *Characterizing human mobility patterns in a large street network*, 2009 *Phys. Rev.* E **80** 021136

[3] Lee K, Hong S, Kim S J, Rhee I and Chong S, *SLAW: a new mobility model for human walks*, 2009 *IEEE Proc. INFOCOM* pp 855–63

[4] Bajardi P, Poletto C, Ramasco J J, Tizzoni M, Colizza V and Vespignani A, *Human mobility networks, travel restrictions, and the global spread of 2009 H1N1 pandemic*, 2011 *PLoS ONE* **6** e16591

[5] Gonzalez M, Hidalgo C A and Barabási A L, *Understanding individual human mobility patterns*, 2008 *Nature* **453** 779–82

[6] Zignani M and Gatio S, *Extracting human mobility patterns from GPS-based traces*, 2010 *Wireless Days 2010 IFIP* pp 1–5

[7] Rhee I, Shin M, Hong S, Lee K and Chong S, *On the Levy-walk nature of human mobility*, 2011 *IEEE/ACM Trans. Netw.* **19** 630–43

[8] Jiang B and Jia T, *Exploring human mobility patterns based on location information of US flight*, 2011 arXiv:1104.4578

[9] Kalogerakis E, Vesselova O, Hays J, Efros A A and Hertzmann A, *Image sequence geolocation with human travel priors*, 2009 *IEEE Proc. 12th ICCV* pp 253–60

[10] Brockmann D, Hufnage L and Geisel T, *The scaling laws of human travel*, 2006 *Nature* **439** 462–5

[11] Viswanatha G M, Afanasyev V, Buldyrev S V, Murphy E J, Prince P A and Stanley H E, *Levy flight search patterns of wandering albatrosses*, 1996 *Nature* **381** 413–5

[12] Ramos-Fernandez G, Morales J L, Miramontes O, Cocho G, Larralde H and Ayala-Orozco B, *Levy walk patterns in the foraging movements of spider monkeys*, 2004 *Behav. Ecol. Sociobiol.* **55** 223–30

[13] Edwards A M *et al*, *Revisiting Levy flight search patterns of wandering albatrosses, bumblebees and deer*, 2007 *Nature* **449** 1044–8

[14] Sims D W, Southall E J, Humphries N E, Hays G C, Bradshaw C A, Pitchford J W and Metcalfe J D, *Scaling laws of marine predator search behavior*, 2008 *Nature* **451** 7182

[15] Azevedo T S, Bezerra R L, Campos C A V and Moraes L F M, *An analysis of human mobility using real traces*, 2009 *Proc. 2009 IEEE WCNC* pp 2390–5

[16] Liang X, Zheng X D, Lv W F, Zhu T Y and Xu K, *The scaling of human mobility by taxis is exponential*, 2012 *Physica* A **391** 2135–44

[17] Scafetta N, *Understanding the complexity of the Levy-walk nature of human mobility with a multi-scale cost/benefit model*, 2011 *Chaos* **21** 043106

[18] Han X P, Hao Q, Wang B H and Zhou T, *Origin of the scaling law in human mobility: hierarchical organization of traffic systems*, 2011 *Phys. Rev.* E **83** 036117

[19] Liu Y, Kang C G, Gao S, Xiao Y and Tian Y, *Understanding intra-urban trip patterns from taxi trajectory data*, 2012 *J. Geogr. Syst.* **14** 463–83

[20] Stephens D W and Krebs J R, 1986 *Foraging Theory* (Princeton: Princeton University Press)

[21] Laptev I and Lindeberg T, *Interest point detection and scale selection in space–time*, 2003 *Proc. 4th Int. Conf. on Scale Space Methods in Computer Vision* pp 372–87

[22] Jiang B and Liu X, *Scaling of geographic space from the perspective of city and field blocks and using volunteered geographic information*, 2012 *Int. J. Geogr. Inf. Sci.* **26** 215–29

[23] Clauset A, Shalizi C R and Newman M E J, *Power-law distributions in empirical data*, 2009 *SIAM Rev.* **51** 661–703

[24] Song C M, Koren T, Wang P and Barabasi A L, *Modeling the scaling properties of human mobility*, 2010 *Nature Phys.* **6** 818–23

[25] Jia T and Jiang B, *Building and analyzing the US airport network based on en-route location information*, 2012 *Physica* A **391** 4031–42

[26] Hart P E, Nilsson N J and Raphael B, *A formal basis for the heuristic determination of minimum cost paths*, 1968 *IEEE Trans. Syst. Sci. Cybern.* **4** 100–7

[27] Jia T and Jiang B, *Scaling property of urban systems using an entropy-based hierarchical clustering method*, 2012 *Proc. AGILE' 2012 Int. Conf. on Geographic Information Science* ed J Gensel, D Josselin and D Vandenbroucke

[28] Jiang B and Jia T, *Agent-based simulation of human movement shaped by the underlying street structure*, 2011 *Int. J. Geogr. Inf. Sci.* **25** 51–64

[29] Miller J H and Page S E, 2007 *Complex Adaptive Systems: An Introduction to Computational Models of Social Life* (Princeton, NJ: Princeton University Press)

[30] Kalapala V, Sanwalani V, Clauset A and Moore C, *Scale invariance in road networks*, 2006 *Phys. Rev.* E **73** 026130

[31] Lovasz L, *Random walks on graphs: a survey*, 1996 *Combinatorics: Paul Erdos is Eighty* ed D Miklos *et al* (Budapest: János Bolyai Mathematical Society) pp 353–98

[32] Barabasi A L and Albert R, *Emergence of scaling in random networks*, 1999 *Science* **286** (5439) 509–12

[33] Page L and Brin S, *The anatomy of a large-scale hypertextual web search engine*, 1998 *Proc. 7th Int. Conf. on World Wide Web* vol 7 pp 107–17

[34] Vuong Q H, *Likelihood ratio tests for model selection and non-nested hypotheses*, 1989 *Econometrica* **57** 307–33

[35] Bazzani A, Giorgini B, Rambaldi S, Gallotti R and Giovannini L, *Statistical laws in urban mobility from microscopic GPS data in the area of Florence*, 2010 *J. Stat. Mech.* P05001