INTERFACE

rsif.royalsocietypublishing.org

Research

Cite this article: Gallotti R, Louf R, Luck J-M, Barthelemy M. 2018 Tracking random walks. *J. R. Soc. Interface* **15**: 20170776.
http://dx.doi.org/10.1098/rsif.2017.0776

Received: 19 October 2017
Accepted: 11 January 2018

Subject Category:
Life Sciences–Physics interface

Subject Areas:
biomathematics

Keywords:
statistical physics, renewal theory, human mobility, animal movement

Author for correspondence:
Marc Barthelemy
e-mail: marc.barthelemy@ipht.fr

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.3982608.v2.

THE ROYAL SOCIETY PUBLISHING

Tracking random walks

Riccardo Gallotti[1], Rémi Louf[2], Jean-Marc Luck[3] and Marc Barthelemy[3,4]

[1]Instituto de Física Interdisciplinar y Sistemas Complejos (IFISC), CSIC-UIB, Campus UIB, ES-07122 Palma de Mallorca, Spain
[2]Centre for Advanced Spatial Analysis (CASA), University College London, London W1T 4TJ, UK
[3]Institut de Physique Théorique, Université Paris-Saclay, CEA and CNRS, 91191 Gif-sur-Yvette, France
[4]CAMS (CNRS/EHESS), 190-198, avenue de France, 75244 Paris Cedex 13, France

MB, 0000-0003-0890-9713

In empirical studies, trajectories of animals or individuals are sampled in space and time. Yet, it is unclear how sampling procedures bias the recorded data. Here, we consider the important case of movements that consist of alternating rests and moves of random durations and study how the estimate of their statistical properties is affected by the way we measure them. We first discuss the ideal case of a constant sampling interval and short-tailed distributions of rest and move durations, and provide an exact analytical calculation of the fraction of correctly sampled trajectories. Further insights are obtained with simulations using more realistic long-tailed rest duration distributions showing that this fraction is dramatically reduced for real cases. We test our results for real human mobility with high-resolution GPS trajectories, where a constant sampling interval allows one to recover at best 18% of the movements, while over-evaluating the average trip length by a factor of 2. Using a sampling interval extracted from real communication data, we recover only 11% of the moves, a value that cannot be increased above 16% even with ideal algorithms. These figures call for a more cautious use of data in quantitative studies of individuals' movements.

1. Introduction

Recent years have witnessed a dramatic increase in the use of large amounts of available data thanks to information and communication technologies. These new sources allow one to monitor and to map the dynamical properties of many complex systems on an unprecedented scale [1] and we now have access to a vast number of spatial trajectories representing movements of objects in geographical space [2]. In particular, such datasets have opened the opportunity to better understand human movements [3–7] and the impact of mobility on important processes such as epidemic spreading [8]. These recent works extend previous studies of movements and foraging patterns of animals [9,10] and rely on tracking man-made inanimate objects [11,12]. However, as is the case for any dataset, these new sources of information have limits and biases [13–16] that need to be assessed.

It is common to approximate the continuous spatio-temporal record of the followed individual (or animal) by a series of straight lines, thus describing the movements of an organism as a sequence of behavioural events called *moves* for animals [17] and *trips* for humans. This empirical approach allows a natural implementation of the theoretical framework of continuous-time random walks [11,18], where a *rest* time is associated with the endpoint of each move. However, this leads to the first major problem due to the lack of behavioural information in the empirical data [19]. Real trajectories always exhibit a large variety of *intertwined* static and dynamic behaviours [20]: slow versus fast movement for animals [19], fixation versus saccade in eye-tracking [21] or activities versus trips in human mobility [22]. Isolating and identifying these behaviours from a series of chronologically ordered points is an important statistical challenge [23] and a growing array of methods based on spatio-temporal characteristics of the trajectories have been developed to perform this task automatically [2,19,21]. These methods are often tailored for the specific dataset in

question [20]. Therefore, even the working definition of a 'move' might vary significantly between studies, depending on the method and the technology used [24].

A second complication comes from the limits of the technology used for collecting the empirical data. In the case of spatial movements, a crucial aspect is the temporal sampling of the trajectory. The simplest and most common method used is *periodic* sampling, where spatial coordinates are recorded at regular time intervals. Alternatively, other data sources are characterized by an *event-based* sampling where locations are recorded at certain (random) events. This is the case, for instance, for the most common sources of human mobility data such as call detail records (CDR) of mobile phone data [25] and geo-located social media accesses [26]. In both cases, the discrete displacements recorded are associated with continuous moves [17], but this is a strong oversimplification, and all derived quantities will depend on the sampling process itself [20,27–29]. The sampling of random processes might even be the principal cause of the emergence of long tails in several statistical distributions [30,31]. For example, in the case of periodic sampling, it has been shown that non-Lévy movements can be erroneously interpreted as Lévy flights when sampling time intervals are larger than the natural timescale of animals' movements [32,33]. The sampling rate is thus a crucial element that has to be taken into account when analysing empirical trajectories [20,34].

For both periodic and event-based sampling, the nature of data forces researchers to make the following naive assumptions:

(i) an individual is always at rest at the location where its position is recorded; and
(ii) every change of position is associated with a single move.

This point of view has been adopted, for instance, in the first important papers where human mobility has been studied with mobile phone data [3,4] and often replicated, even in recent studies [35–37]. However, the use of these new sources of data exacerbates the challenges associated with temporal sampling. Indeed, in these data, trajectories are represented as sequences of positions recorded at the moment of a communication event (which can be a call, a text message or an application access). The trajectory sampling is therefore coupled with the random and bursty nature of human communications [38]. The probability distribution of the time interval between calls [3,4], e-mails [38] and tweets [39] has a long tail which can be fitted by a power law with an exponent value close to -1 (and with a cut-off on the order of days). Only in a few cases, a small set of trajectories sampled every $\Delta = 1$ or 2 h is available [3,4,40]. Even when individuals with a very high call frequency are selected [40], they are still inactive most of the time [41]. In order to identify human mobility patterns, it thus become necessary to introduce ad hoc methods based on reasonable assumptions and almost arbitrary parameters [16,42].

In this paper, we discuss the effect of sampling and assumptions (i) and (ii) on the measured properties of random movements. We will consider one of the simplest and realistic cases where the trajectory consists of two alternating phases, moves and rests, whose durations t and τ are regarded as independent random variables. Trajectories can then be seen as an alternating renewal process, i.e. a generalization of Poisson processes to arbitrary holding times and to two alternating kinds of events. The sampling time interval Δ depends on the particular experiment and can be either constant or randomly distributed. Using methods of renewal theory along the lines of [43], we provide a theoretical estimate for the fraction of correctly sampled trips with periodic sampling, and show the existence of an optimal sampling time interval. We then extend our results numerically to the case of event-based sampling, and with more realistic rest times and speeds. This allows us to show that sampling human trajectories in more realistic settings is necessarily worse than predicted by our analytical model. Finally, we use high-resolution (spatially and temporally) GPS trajectories to verify our predictions on real data.

2. Results

2.1. Theoretical analysis

We study the effect of the periodic sampling rate on the apparent distribution of measured move lengths. We focus on the case of an alternating sequence of rests and moves and we further assume that the movement is one dimensional with a constant velocity v (see electronic supplementary material, section 'Numerical analysis', for other cases). Simplifying the problem to one dimension is here sufficient to point out when the sampling is inadequate. We show below that the diagnostics we use to identify the optimal sampling times are independent of the dimensionality of the space in which a trajectory is embedded. We will not discuss here other issues associated with temporal sampling, like the apparent speed and turning angles in a general two-dimensional case [20,27,29], the possible fits of the displacement distribution [32,33,44,45], or interpolation methods to reconstruct the movements between samplings [46]. The quantities entering this problem are therefore: the move duration t, the move length $\ell = vt$, the resting time τ, and the time interval Δ between two consecutive measures. The distributions $P(t)$ and $P(\tau)$ are characteristics of the specific subject in motion, while the distribution of the sampling interval $P(\Delta)$ is associated with the technology used for tracking the motion. Sampling the trajectory gives us a displacement distribution $P(\ell^*)$ where ℓ^* is the apparent length of a move, and the problem is thus to compute this distribution $P(\ell^*)$ for any given distributions $P(t)$, $P(\tau)$ and $P(\Delta)$.

During rests, the displacement is assumed to be zero, and so the succession of rests and moves is associated with a continuous increasing function $x(\theta)$, where θ is the time parameter (figure 1). We sample the position x_k^* for every instant $\theta_k^* = \sum_{j=1}^{k} \Delta_j$, where Δ_j is the value of the jth sampling interval (in the case of constant sampling, $\Delta_j = \bar{\Delta}$, and so $\theta_k^* = k\bar{\Delta}$). The succession of space–time coordinates (θ_k^*, x_k^*) (shown in figure 1 and in the two-dimensional example of electronic supplementary material, figure S1) thus represents all the knowledge we have about the trajectory after sampling. For two consecutive measures at times θ_k^* and θ_{k+1}^*, there is an observed displacement $\ell_k^* = x_{k+1}^* - x_k^*$. Our goal is then to estimate the differences between the distribution of real displacement lengths ℓ and of the observed displacements ℓ_k^*. In particular, we want to understand the biases induced by different choices for $P(\Delta)$.

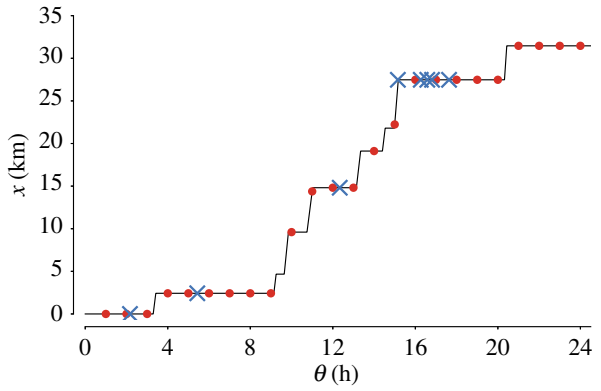

Figure 1. Examples of trajectory sampling. On a trajectory with exponentially distributed rest and move durations, we show the case of constant sampling interval (red circles) and the case of random sampling interval (blue crosses) with $P(\Delta) \propto \Delta^{-1}$ ($\Delta_{\min} = 5$ min, $\Delta_{\max} = 12$ h). See electronic supplementary material, figure S1 for a two-dimensional example. (Online version in colour.)

If we make the naive assumption (ii), discussed in the introduction, that every observed displacement is associated with a single move, the necessary condition for this to be correct is that two subsequent sampling times θ_k^* and θ_{k+1}^* fall in two consecutive rests. We can also easily identify the cases where the sampling times fall in the same rest, because this is the only situation where we exactly have $\ell_k^* = 0$ and which does not lead to a wrong estimate of the individual's movement. Conversely, we must consider as errors all remaining configurations, because at least one of two things necessarily happens: (i) we have a sampling point at movement or (ii) a rest is missed by the temporal sampling. Either of these events leads to a misinterpretation of the individual mobility and to an under- or overestimate of the move lengths [32] and of the number of trips observed [15]. In order to go beyond this simple hand-waving argument, we will consider the case of exponential distributions for $P(t)$ and $P(\tau)$, constant sampling time interval $\bar{\Delta}$, and constant speed v. In this case, we obtain explicitly the distribution $P(\ell^*)$ of sampled displacements. This will allow us to discuss the impact of the sampling, and to show, in particular, that there is an optimal value for $\bar{\Delta}$.

2.1.1. Constant sampling rate and exponential distributions

We will consider the case of exponential distributions for the move and rest durations:

$$P(t) = \left(\frac{1}{\bar{t}}\right) \exp\left(-\frac{t}{\bar{t}}\right) \quad \text{and} \quad P(\tau) = \left(\frac{1}{\bar{\tau}}\right) \exp\left(-\frac{\tau}{\bar{\tau}}\right), \tag{2.1}$$

and a constant sampling interval:

$$P(\Delta) = \delta(\Delta - \bar{\Delta}), \tag{2.2}$$

($\delta(x)$ is Dirac's delta function). In the constant velocity case, the real displacements are also exponentially distributed:

$$P(\ell) = \left(\frac{1}{\bar{\ell}}\right) \exp\left(-\frac{\ell}{\bar{\ell}}\right), \tag{2.3}$$

with $\bar{\ell} = v\bar{t}$.

Using methods of renewal theory [47–49], along the lines of [43], we obtain an explicit expression for the distribution $P(\ell^*)$ of apparent displacements ℓ^* after sampling (see electronic supplementary material, section 'Analytical calculations', and in particular equations (S15), (S33)):

$$P(\ell^*) = \frac{e^{-\bar{\Delta}/\bar{\tau}}}{1 + \bar{t}/\bar{\tau}} \delta(\ell^*) + \frac{e^{-\bar{\Delta}/\bar{t}}}{1 + \bar{\tau}/\bar{t}} \delta(\ell^* - v\bar{\Delta}) + P_{\text{cont}}(\ell^*), \tag{2.4}$$

where the continuous part of this distribution reads

$$P_{\text{cont}}(\ell^*) = \frac{2e^{-(\ell^*/v\bar{t} + (v\bar{\Delta} - \ell^*)/v\bar{\tau})}}{v(\bar{t} + \bar{\tau})} \left[I_0(y) + \left(\frac{\ell^*}{v\bar{\tau}} + \frac{v\bar{\Delta} - \ell^*}{v\bar{t}} \right) \frac{I_1(y)}{y} \right], \tag{2.5}$$

with $y = 2\sqrt{\frac{\ell^*(v\bar{\Delta} - \ell^*)}{v^2 \bar{t}\bar{\tau}}}$, and where $I_0(y)$ and $I_1(y)$ are modified Bessel functions of the first kind.

In the following, we will not consider the discrete part associated with the Dirac's delta function $\delta(\ell^*)$ of the distribution $P(\ell^*)$, as the value $\ell^* = 0$ can be easily recognized and excluded in any practical scenario. The fraction of sampling intervals associated with null movements ($\ell^* = 0$), denoted by $C_0(\bar{\Delta})$, can be significantly large. In the stationary regime [50], we can compute $C_0(\bar{\Delta})$ for any distributions $P(t)$ and $P(\tau)$, and a constant sampling time $\bar{\Delta}$ (see electronic supplementary material, equation (S17)). We can show that it is a decreasing function, varying between $C_0(0) = \bar{\tau}/(\bar{t} + \bar{\tau})$ (i.e. the fraction of time spent at rest, in the continuous sampling limit) and $C_0(\infty) = 0$. In the particular case of exponential distributions (equation (2.1)), C_0 is the prefactor of the $\delta(\ell^*)$ peak in equation (2.4), and can be very large. For instance, $C_0 \approx 60\%$ in the case of car mobility ($\bar{t} = 0.30$ h and $\bar{\tau} = 2.49$ h, see Methods) and $\bar{\Delta} = 1$ h. For this reason, we compare the original data to a rescaled probability distribution which does not include the $\delta(\ell^*)$ peak and is given by (see electronic supplementary material, figure S2)

$$P_{\ell^*>0}(\ell^*) = \frac{1}{1 - C_0(\bar{\Delta})} \left[\frac{e^{-\bar{\Delta}/\bar{t}}}{1 + \bar{\tau}/\bar{t}} \delta(\ell^* - \bar{\Delta}) + P_{\text{cont}}(\ell^*) \right]. \tag{2.6}$$

We show in figure 2*a* the dependence of the continuous part of $P_{\ell^*>0}(\ell^*)$ on $\bar{\tau}$, keeping the average travel time $\bar{t}$ fixed to the experimental value of 0.30 h for car mobility [7].

We note that $P_{\text{cont}}(\ell^*)$ can have a maximum, even if the original distribution $P(\ell)$ is a decreasing function. The measurements allow us to recover the exponential tail of travel times only if the resting time $\bar{\tau}$ is sufficiently long. Conversely, when the sampling time $\bar{\Delta}$ is larger than the average duration of a rest, the result of the sampling is manifestly different from the original exponential distribution. In figure 2*b*, we take $\bar{t} = 0.30$ h and $\bar{\tau} = 2.49$ h (which are the values observed for vehicular mobility, see Methods) and study the outcome for different sampling times $\bar{\Delta} < \bar{\tau}$. Naturally, $\bar{\Delta}$ acts as a cut-off because all moves longer than this value are necessarily interrupted by the sampling. By contrast, for large values of $\bar{\Delta}$, the number of short travels is underestimated, as subsequent short moves may be joined together and thus appear as an effective long one.

We also computed exactly the first two moments of the distribution equation (2.4) and found for the average

$$\langle \ell^* \rangle = \frac{v\bar{\Delta}}{1 + \bar{\tau}/\bar{t}}, \tag{2.7}$$

(see electronic supplementary material, equation (S19) and equation (S26) for the second moment). Naturally, the exclusion of the null displacements influences the value of the distribution's moments. In particular, the average value of

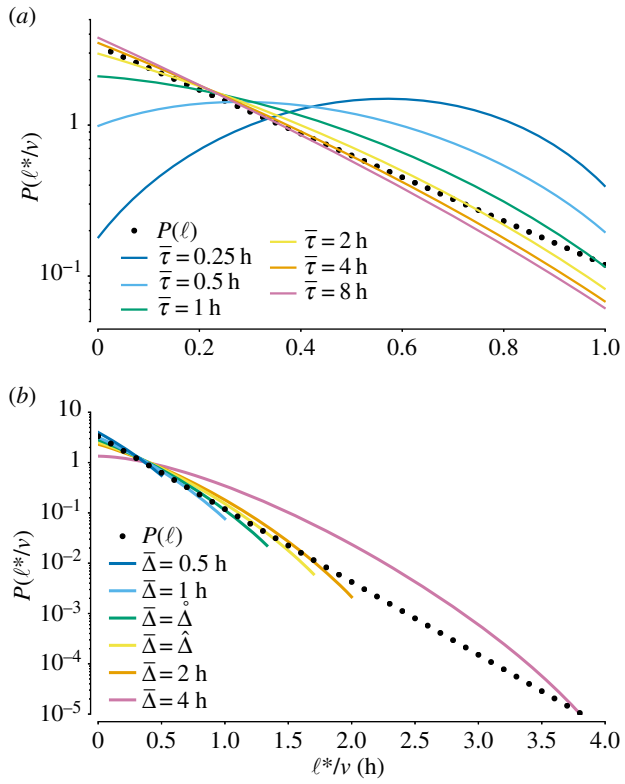

Figure 2. Distributions $P(\ell^*/v)$ obtained from periodic sampling with exponential distribution of rest and move times. (*a*) Dependence of equation (2.6) on $\bar{\tau}$ fixing $\bar{t} = 0.30\,\mathrm{h}$ and $\bar{\Delta} = 1\,\mathrm{h}$. The distribution has a maximum when the average rest times exceed the sampling time, and its value is strictly zero for $\ell^* > v\bar{\Delta}$. (*b*) Dependence of equation (2.6) on $\bar{\Delta}$ fixing $\bar{t} = 0.30\,\mathrm{h}$, $\bar{\tau} = 2.49\,\mathrm{h}$. Short sampling times introduce a cut-off in the distribution. Large deviations can be observed when sampling time intervals are long. (Online version in colour.)

equation (2.6) can be computed by a simple rescaling and reads

$$\langle \ell^* \rangle_{\ell^*>0} = \frac{\langle \ell^* \rangle}{1 - C_0(\bar{\Delta})}. \tag{2.8}$$

This rescaling yields notable changes in the numerical values of the moments. For instance, with realistic values for car mobility ($\bar{t} = 0.30\,\mathrm{h}$ and $\bar{\tau} = 2.49\,\mathrm{h}$), a sampling time of 1 h gives $\langle \ell^* \rangle / v \approx 0.11\,\mathrm{h}$, while excluding the zero-displacement part, we obtain $\langle \ell^* \rangle_{\ell^*>0}/v \approx 0.27\,\mathrm{h}$.

2.1.2. Optimal sampling times

We first note that high-frequency sampling ($\Delta \to 0$) does not automatically allow one to understand the whole trajectories under the naive assumptions (i) and (ii). Indeed, it is only with additional data that we can correctly reconstruct a whole trajectory. It is then necessary to implement a 'segmentation' algorithm that goes beyond the assumption (ii) that an observed displacement corresponds to one single move, as $\Delta \to 0$ implies that any move is cut into a very large number of segments [17]. In addition, high-frequency recordings are known to present uncertainties and systematic errors that need to be taken into account for extracting meaningful information [17,20,51–53]. A good segmentation algorithm should take into account the noise, the spatial scale and characteristic speeds of the tracked subjects. Here, it is not our intent to develop detailed segmentation methods, but to show the quality, and the limits, of the simpler assumption that one observed displacement is equal to one move. In this framework, having $\Delta \to 0$ means that we measure moves over a very short time, obtaining thus a distribution of measured displacement peaked at very small values and indicating that very high-frequency rates are not good under assumption (ii).

We can define an 'optimal constant sampling time' in two different ways: either as the time interval $\mathring{\Delta}$ that correctly estimates the average length of moves, or as the time interval $\hat{\Delta}$ that maximizes the fraction of correctly sampled moves. The second approach offers a more general perspective, introducing a dimensionless measure for the quality of the sampling but which is unfortunately not a natural and common observable in experimental ecology or human mobility. For this reason, we consider in parallel the first approach that is based on a more natural quantity, the average displacement, which also has the merit of focusing on the character of the displacement distribution and therefore on what is perhaps the most controversial topic associated with individual trajectories: the mis-identification of a Lévy walk from empirical data. In the following, we obtain exact formulas for both $\mathring{\Delta}$ and $\hat{\Delta}$ in the exponential–exponential case (i.e. with conditions described by equation (2.1)).

2.1.3. Average move duration and total number of moves

The optimal sampling time $\mathring{\Delta}$ can be obtained by solving for $\bar{\Delta}$ the equation $\langle \ell^* \rangle_{\ell^*>0} = v\bar{t}$. The solution can be written in the form

$$\mathring{\Delta} = \bar{\tau}\, W(-\mathrm{e}^{-\bar{t}/\bar{\tau}-1}) + \bar{t} + \bar{\tau}, \tag{2.9}$$

where $W(x)$ is the Lambert function, such that $W(x)\mathrm{e}^{W(x)} = x$. This function is defined for $x \geq -\mathrm{e}^{-1}$, which always holds in our case because $\bar{t}, \bar{\tau} > 0$. Using the empirical values $\bar{t} = 0.30\,\mathrm{h}$, $\bar{\tau} = 2.49\,\mathrm{h}$, we obtain $\mathring{\Delta} \approx 80$ min. This result is confirmed by Monte Carlo simulations (figure 3), where red circles represent the values for $\mathring{\Delta}$. With this 'optimal' sampling time based on the first moment, the second moment is slightly underestimated. Note that matching the average travel time is equivalent to correctly estimating the number n of trips, i.e. of moves and stops (see inset in figure 3*a*), which is computed by counting the number of consecutive sampled points k and $k+1$ with $\ell_k^* = x_{k+1}^* - x_k^* > 0$. For $\bar{\Delta} > \mathring{\Delta}$, the trajectory is under-sampled ($n^* < n$) and trip lengths are overestimated, while for $\bar{\Delta} < \mathring{\Delta}$ it is over-sampled ($n^* > n$) and trip lengths are underestimated.

This point of view about the number of moves allows us to extend the validity of this optimal sampling to higher dimensionality (two or three dimensions) and to any distribution $P(v)$. The dimensionality of space indeed does not influence the moves' number counting. To illustrate this, we extend this analysis in the electronic supplementary material, section 'Numerical analysis' with a Monte Carlo simulation in the case where speed is a random variable depending on the move duration [7]. In this case, our exact results for $P(\ell)$ do not hold anymore, because moves have different speeds. Nevertheless, the value given by equation (2.9) only underestimates the mean displacement length with varying speeds by some 5%.

More generally, all our analytical results concern the stationary regime of the renewal process. This stationary regime exists only if the mean values $\bar{t}$ and $\bar{\tau}$ are finite (see electronic supplementary material, section 'Analytical calculations'). The distributions $P(t)$ and $P(\tau)$ can thus have

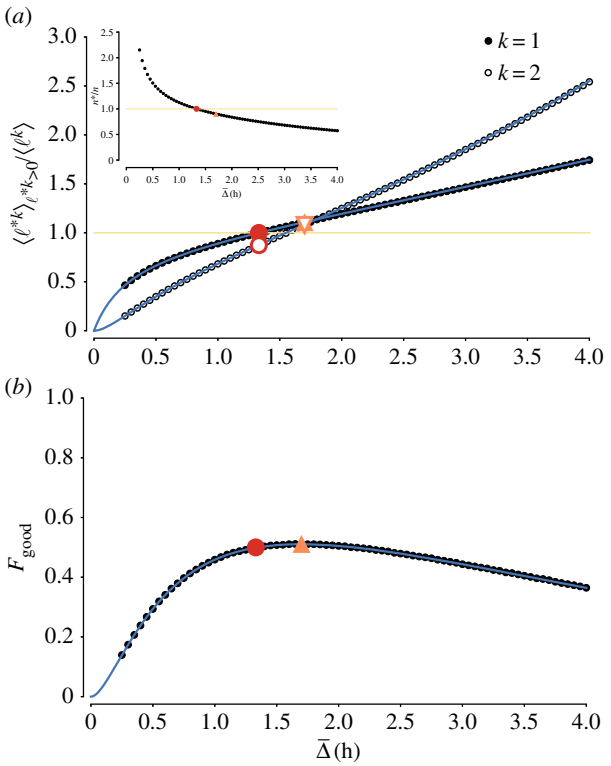

Figure 3. Optimal sampling for exponential distributions and constant sampling. (*a*) We verify numerically (black dots) our analytical results (blue lines) for the first ($k = 1$, equation (2.8)) and second ($k = 2$, electronic supplementary material, equation (S28)) moment of the displacements distribution (normalized by $\langle \ell \rangle$ and $\langle \ell^2 \rangle$, respectively) versus sampling time interval. The original average value $\langle \ell \rangle$ (yellow solid line) is obtained by definition for $\bar{\Delta} = \Delta$ (filled circle), while is overestimated by $\approx 10\%$ for $\bar{\Delta} = \hat{\Delta}$ (up triangle). The second moment ($k = 2$) has a deviation of about 10% for both optimal sampling times (empty circle and down triangle). In the inset, we show the ratio of the estimated number of trips n^* over the actual number of trips n. With $\bar{\Delta} = \Delta$ (circle), we correctly evaluate the number of moves, while $\bar{\Delta} = \hat{\Delta}$ (triangle) yields a slightly underestimated value $n^* \approx 0.90n$. (*b*) The fraction of good moves follows the curve predicted by equation (2.10) (blue line). The maximum value of 51% is reached for $\bar{\Delta} = \hat{\Delta}$ (triangle), but at $\bar{\Delta} = \Delta$ (circle) the value is only 1% lower. We choose here $\bar{t} = 0.30$ h, $\bar{\tau} = 2.49$ h. (Online version in colour.)

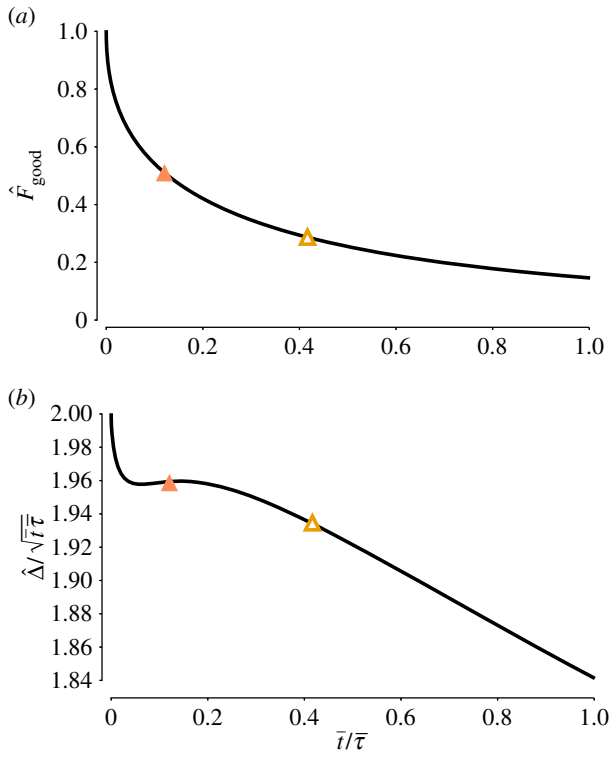

Figure 4. Maximization of F_{good}. (*a*) The maximum $\hat{F}_{\text{good}}$ for exponential distributions. We observe that $\hat{F}_{\text{good}} \to 1$ in the limit for small $\bar{t}$, and decreases as $\bar{t}$ becomes comparable to $\bar{\tau}$. The upper bound to sampling quality is 51% for the car mobility conditions of figure 3 (orange solid triangle) and 29% for GeoLife trajectories of figure 5 (yellow empty triangle). (*b*) The sampling rate $\hat{\Delta}$ optimizing F_{good} has a non-trivial dependence on $\bar{t}$ and $\bar{\tau}$. We identify a relatively weak dependence on $\bar{t}/\bar{\tau}$, of the form $\hat{\Delta} = \alpha\sqrt{\bar{t}\bar{\tau}}$, with α ranging between 1.84 and 2 for all values of $\bar{t} < \bar{\tau}$. In particular, for the characteristic values observed for car mobility (orange solid triangle, $\bar{t} = 0.30$ h, $\bar{\tau} = 2.49$ h), the curve exhibits a plateau, allowing us to approximate $\hat{\Delta} \approx 1.96\sqrt{\bar{t}\bar{\tau}}$. For the GeoLife trajectories (yellow empty triangle), which have significantly shorter rest times ($\bar{t} = 0.33$ h, $\bar{\tau} = 0.80$) the deviation from this approximation is only of about 1.5%. (Online version in colour.)

power-law tails, in principle, for our results to hold, but only with large enough exponents.

2.1.4. Fraction of correctly sampled moves

In order to estimate $\hat{\Delta}$, we have to compute the fraction F_{good} of movements that are correctly measured. This occurs when two consecutive sampling times fall during the rests immediately before and after a move, say θ_k^* in the rest τ_m and $\theta_{k+1}^* = \theta_k^* + \bar{\Delta}$ in the rest τ_{m+1}. The probability P_{good} of the latter event and the fraction $F_{\text{good}} = P_{\text{good}}/(1 - C_0)$ are calculated in the electronic supplementary material, section 'Analytical calculations'. In the case of exponential distributions, we obtain the explicit expression (see electronic supplementary material, equation (S37))

$$F_{\text{good}}(\bar{\Delta}) = \frac{\bar{t}\bar{\tau}}{(\bar{\tau} - \bar{t})^2} \frac{e^{-\bar{\Delta}/\bar{t}} + ((\bar{\tau} - \bar{t})\bar{\Delta}/\bar{t}\bar{\tau} - 1)e^{-\bar{\Delta}/\bar{\tau}}}{1 + \bar{t}/\bar{\tau} - e^{-\bar{\Delta}/\bar{\tau}}}. \qquad (2.10)$$

In figure 3*b*, we compare the shape of F_{good} for fixed values of $\bar{t}$ and $\bar{\tau}$ with the result of a Monte Carlo simulation. For empirical values valid for car mobility ($\bar{t} = 0.30$ h, $\bar{\tau} = 2.49$ h), the curve has a maximum $\hat{F}_{\text{good}} \approx 51\%$ for a sampling time given by $\hat{\Delta} = 1.70$ h (102 min). Both the value of $\hat{\Delta}$ and the height $\hat{F}_{\text{good}}$ of the maximum of $F_{\text{good}}(\bar{\Delta})$ depend on the ratio $\bar{t}/\bar{\tau}$ (figure 4*a*). They are however independent of the spatial embedding and of the characteristics of $P(v)$. The quantity $\hat{F}_{\text{good}} \approx 51\%$ is associated with the largest value of $\bar{\tau}$ for the data sources we have analysed (mobile data, GPS trajectories and car mobility, see electronic supplementary material, table S1), and thus represents the best possible value associated with human mobility at an urban scale. It is remarkable that the optimal fraction $\hat{F}_{\text{good}}$ of sampled movements in human mobility is so low that essentially one half of the moves are cut or merged during the sampling, limiting the possibility of understanding the individuals' behaviour. We also note that the value $F_{\text{good}}(\Delta)$ is not far from 51% (figure 3*b*). We thus see that, even if the measured and real distributions are similar with comparable first moments, we are often describing different movements. The nature of the process, characterized by $\bar{\tau}$ and $\bar{t}$, limits our knowledge of the system for any value of $\bar{\Delta}$.

The maximal value $\hat{F}_{\text{good}}$ is naturally associated with another optimal sampling time representing the conditions

for which we sample correctly the largest number of moves. This optimal sampling time $\hat{\Delta}$ is of the same order as $\bar{t}$ and $\bar{\tau}$: $\hat{\Delta} = \alpha\sqrt{\bar{t}\bar{\tau}}$. The function $\alpha(\bar{t}/\bar{\tau})$ can be approximated as a constant when studying human mobility at an urban scale, or other datasets sharing similar $\bar{t}/\bar{\tau}$ ratios (figure 4):

$$\hat{\Delta} \approx 1.96\sqrt{\bar{t}\bar{\tau}}. \tag{2.11}$$

This result suggests that the sampling with $\bar{\Delta} \ll \bar{t},\bar{\tau}$ (that is, substantially more frequently than the time frame of an average move or rest) is not optimal and will lead to incorrect results. This is apparently paradoxical, because if the trajectory is very well sampled, then it would be relatively straightforward to build an algorithm that reconstructs correctly moves and rests. However, such a high-frequency sampling is useful only when we have additional information that allows one to reconstruct the trajectory which can be done with more advanced technologies that do not need assumptions (i) and (ii).

2.2. Sampling human movements

The conditions of equations (2.1) and (2.2) define a process where both travel and rest times have a short-tailed distribution and the trajectory sampling is strictly periodic. While this allowed us to find exact analytical expressions and to uncover important effects of sampling on the statistical properties of trajectories, real-world problems are much more involved. Indeed, human travel times are characterized by short-tailed distributions (see [7] and references therein), and resting time can be broadly distributed for both humans [4,54] and animals (see [55] and references therein). In addition, the trajectory can be sampled with a random inter-sampling time.

We expect, in general, to observe the same behaviour as the exponential–exponential case (described by equation (2.1)) studied above for any peaked distribution of rest and move durations (i.e. when both the first two moments converge). We show here that when rests or sampling times are broadly distributed, the outcome of the sampling will be necessarily worse. The exponential–exponential conditions discussed above therefore correspond to the best-case among the typical scenarios observed empirically (although better sampling might be eventually obtained in marginal scenarios such as fixed rest and move times for example). We first confirm with Monte Carlo simulations the validity for more complex cases of the results obtained above for a constant sampling time interval. In particular, we show (see electronic supplementary material, section 'Numerical analysis' and table S1 for details) how the sampling quality F_{good} for cars' mobility progressively decreases from the upper bound of 51% when introducing randomness in sampling times (exponential or power-law) and in rest durations. For instance, introducing a broad $P(\tau)$ yields values of F_{good} lower than 40%, while a broad $P(\Delta)$ yields a F_{good} lower than 30%. We finally predict that, when coupling a broad $P(\Delta)$ and a broad $P(\tau)$ (as observed for mobile phone data), the quality of the sampling decreases significantly, with F_{good} falling to 23%.

We illustrate these different results on a spatio-temporal high-resolution dataset, namely the GeoLife GPS trajectories [56,57]. The data consist of coordinates given every 5 s, thus allowing us to perform a speed-based sequencing (see Methods). We measure the properties of the sequenced

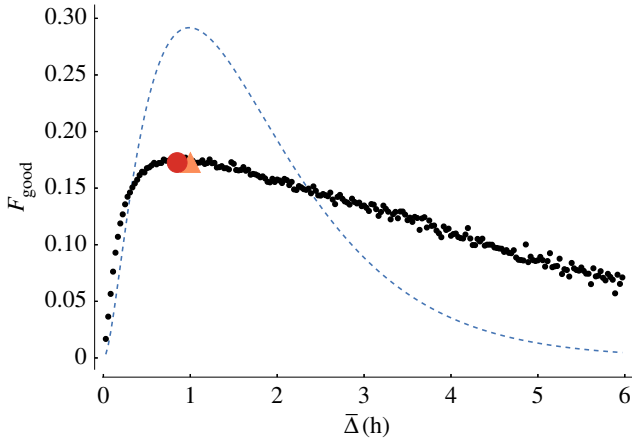

Figure 5. Constant sampling on GPS data. Results are obtained by sampling the GeoLife GPS data with a constant sampling interval Δ. We show (black dots) the fraction of moves correctly sampled as a function of the length of the sampling interval $\bar{\Delta}$. The dashed blue line corresponds to the theoretical curve computed for exponential distributions. The red circle corresponds to $\Delta = 52$ min, while orange triangles correspond to the empirical maximum $\hat{\Delta} = 1$ h of F_{good}. Strikingly, the latter coincides with the theoretical value of $\hat{\Delta}$ for exponential distributions. (Online version in colour.)

trajectories and find again an average trip time $\bar{t} \approx 0.33$ h. The average rest time drops to $\bar{\tau} \approx 0.8$ h, because data allow us to define activities at a finer scale. Using the functional form for F_{good} given by equation (2.10) for the ideal case, we find that the upper bound for the sampling quality declines substantially to $\hat{F}_{\text{good}} = 29\%$. In the following, we use these GeoLife GPS trajectories to study the effect of sampling on real trajectories. In particular, we will validate the previous results by studying the effect of constant sampling and then use mobile phone data to sample the GPS trajectories with a random sampling interval.

We first sample the trajectories with a constant time interval $\bar{\Delta}$ that varies between 1 min and 6 h. For each value of $\bar{\Delta}$, we compute the fraction of the trips that are correctly identified. The results are represented in figure 5. They confirm our analytical predictions. Indeed, we find that there exists an optimum value of the sampling time $\bar{\Delta}_{\text{opt}} \approx 60$ min. Even though this was not expected, because of a non-exponential $P(\tau)$, this value coincides with the predicted maximum $\hat{\Delta} \approx 1.96\sqrt{\bar{t}\,\bar{\tau}} \approx 60$ min (the theoretical curve is represented as a dashed line in figure 5). The fraction of correctly sampled moves is lower than in the idealized case with at best 18% of the trips that are recovered ($F_{\text{good}} \approx 0.18$) with a constant sampling interval.

We also estimate the average length of the sampled trips for every value of the sampling interval and compare it with the average trip length in the original sequenced trajectory (results are represented in electronic supplementary material, figure S3). The optimal value of the sampling time $\bar{\Delta} \approx 15$ min is much smaller than the one maximizing the number of correctly sampled trips. Furthermore, we find that, at the optimal sampling interval $\bar{\Delta} \approx 60$ min, the average sampled trip two-dimensional displacement is about two times larger than the average trip length of the original trajectories.

In the case of geo-localized data obtained from devices such as mobile phones, position and time are recorded at random times corresponding to a call or another event. The sampling time intervals are thus random variables. In

general, they are distributed according to a broad law such as a power-law with exponent close to -1 [3,4]. Here, we use CDR mobile phone data from Senegal [58] and, as commonly done [35,40], extract the duration between calls of the users with extremely high average call frequency, in the same spirit as in [3]. We then sample the sequenced GPS trajectory using these durations. The result is staggering: only 11% of the trips are correctly sampled. One may argue that calls and rests are correlated, or that calls done during moves can be filtered out. We thus computed the proportion of correctly sampled trips at different levels of correlation (see electronic supplementary material, section 'Correlations between calls and rests in empirical sampling' for details), and find that, at best—when we only have calls during rests—only 16% of trips are recovered. The use of CDR mobile phone data or of any dataset presenting a long-tailed inter-event time distribution to study mobility is thus very questionable. We note that forcing a perfect correlation between calls and rests amounts to forcing assumption (i) presented above. Yet, the trajectory is still poorly sampled, meaning that assumption (ii) is flawed.

3. Discussion

A key aspect of every experimental science is to be aware of the limits of the experiment's set-up and of the measuring apparatus. Unfortunately, this point has often been neglected in the recent trend of data-driven studies. The desire for novel, large-impact results is leading to studies where many corners are cut. As a consequence, a large number of quantitative results are sustained almost exclusively by the sheer amount of data gathered, even when those data are not adequate for the problem at hand: not all biases do average themselves out. This is particularly true for the study of trajectories from sampling movements in space. The choices taken for trajectory segmentation, together with the temporal and spatial granularity of the measures, influence all quantities associated with these trajectories [20,34].

In this paper, we have shown that for any sampling of a trajectory alternating rests and movements (of animals, human or artefacts) the assumptions that each measure corresponds to a rest and that an observed displacement correspond to a move are intrinsically flawed. We solved analytically an idealized case which shows that the fraction of trips that are correctly identified with a constant sampling time interval is intrinsically limited, and that this limit is *at best* 51% for humans moving at the urban scale. We also showed that this fraction is significantly lower in any other realistic scenario, especially when mobility is being studied through the lens of mobile phone communications: using phone calls in order to track mobility gives correct predictions for 23% of the trips made with a car. Result gets even worse if one wants to investigate mobility at a finer scale: using high-resolution GPS data the value drops down to 11%, and we estimate that no more of 16% of movements can be recovered, even if a perfect stay-point identification algorithm is applied. These figures (summarized in electronic supplementary material, table S1) cast a shadow on the possibility of understanding [3] and modelling [4] human mobility from CDR data. Our ability to predict individuals' movements [40] is limited not only by the temporal and spatial scales of analysis [59,60], but also and highly predominantly by limitations inherent to the data sources. We provided new analytical tools to evaluate the quality of a sampled trajectory for the study of both animal and human movements. Positions must be collected (or, when necessary for historical comparisons, down-sampled [34]) at least with a frequency commensurate with the underlying moving and resting dynamics ($\bar{\Delta} \approx 1.96\sqrt{\bar{t}\bar{\tau}}$). Alternatively, stay points can be reconstructed from high-frequency sampling ($\langle\Delta\rangle \ll \bar{\tau}$), but not when one has bursty inter-event times, because during the numerous extreme events constituting the long tail of the distribution $P(\Delta)$ the information on the movements is simply absent. Further studies and rigorous analysis of the empirical methods used in many studies are thus necessary in order to construct solid foundations for our knowledge.

4. Methods

4.1. GPS data

In order to prove the validity of our claims, we test the above predictions on high-resolution data, the GeoLife GPS trajectories [56]. This dataset consists of the trajectories of 182 subjects registered by a GPS device over the course of 3 years. The database contains 17 621 trajectories for a cumulated travel length of more than 1 000 000 km. Most trajectories are logged with a temporal precision of the order of the second.

Because the term 'rest' has a behavioural connotation, we will talk in the following about stay points [42]. These are locations where an individual stays for a certain period of time and from which he or she does not depart too much. Of course, the identification of stay points depends on the spatial and temporal granularity of the data [20].

As mentioned in the Introduction, the absence of contextual information forces us to make more or less realistic assumptions in order to identify travelling times and rests. We begin by filtering out the trajectories that are less than 1 km long, as they are not representative. We then proceed to identify stay points as follows:

— we consider all points around the point p_t in a moving time window of duration $\tau = 10$ s around t;
— in this window, we compute the average movement speed between successive trajectory points;
— if the average speed is lower than 2 m s^{-1} (fast walker), we identify p_t as a stay point;
— we iterate the procedure for all points in the trajectory; and
— we aggregate consecutive stay points if the move in-between is less than 100 m and aggregate consecutive moves if the intermediate rest is shorter than 5 min.

The last passage is introduced in order to minimize the impact of fluctuations in the GPS reading. After this procedure, we obtain individual trajectories where stay points are identified.

We find the average travel and rest times $\bar{t} = 20$ min and $\bar{\tau} = 48$ min. The average travel time is identical to that observed for vehicular mobility. The average duration $\bar{\tau}$ of a rest is, however, significantly shorter.

4.2. Call detail records data

We use the dataset 2 'fine-grained mobility' of the Orange data made available for the D4D challenge [58] that provides anonymized individual CDR records. For privacy reasons, the caller IDs are reshuffled every 15 days. The dataset spans 25 such 15-day periods. The selection procedure that is most often used is the one proposed in [40], i.e. selecting only the individuals whose average call frequency is greater than 0.5 calls/hour.

Here, we allowed for a more conservative margin by selecting only the 1.1% of individuals who had more than 1 call/hour in a period of 15 days. Furthermore, the data provide call time stamps with a 10-minutes granularity. We apply a smoothing procedure which consists of picking a time uniformly at random between $M-5$ and $M+5$, where M is the value in minute indicated by the time stamp. One should bear in mind that the mobile phone CDR and GPS trajectories come from two independent datasets describing two different populations and times of the year. For this reason, we did not enforce calendar synchronization between the datasets, but used the CDR data to randomly extract real inter-event times with the appropriate minimal frequency. More accurate numbers would thus be obtained in a situation where information on calls and trajectories would be available for the same user.

4.3. Characteristic times for car mobility

We need to identify the values of $\bar{t}$ and $\bar{\tau}$ in conditions that realistically describe human mobility. We do this by using the results of the analysis of urban and inter-urban traffic of private vehicles in Italy [7].

The average travel time observed for Italian cars is $\bar{t} = 0.30$ h. Moreover, as discussed in [7] and references therein, in private as in public transportation, the distribution of trip durations $P(t)$ in a city is short-tailed. A similar result has been found in taxi rides, in survey data (where also $\bar{t} \approx 0.30$ h) and on the GPS data [57] we use in this work (for separated modes of transport). For this reason, we can safely limit our numerical analysis to the case of exponential $P(t)$.

Concerning rest times, two different functional forms have been proposed for the distribution $P(\tau)$. Car parking durations have been fitted with a stretched exponential:

$$P(\tau) = \frac{\exp\left(-(\tau/\tau_0)^{\beta}\right)}{\tau_0 \Gamma(1+1/\beta)}, \tag{4.1}$$

with $\tau_0 \approx 10^{-4}$ h and $\beta \approx 0.19$ [7]. For mobile phone data, a truncated power-law fit has been proposed:

$$P(\tau) \propto \tau^{-\gamma} \exp\left(-\frac{\tau}{\tau_e}\right), \tag{4.2}$$

with $\gamma \approx 1.8$ and $\tau_e \approx 17$ h [4]. This fit is made on movements sampled at best with $\bar{\Delta} = 1$ h (it is thus expected to be influenced by the sampling issues described above), and does not allow one to identify rests shorter than 1 h. Note that in estimating the distribution's average below, we are extending the distribution (4.2) below this experimental range.

In our analytical study, we assume the distribution $P(\tau)$ to be exponential, while it is not in general. We therefore estimate the parameter $\bar{\tau}$ averaging the distributions (4.1) and (4.2) between 5 min and 24 h, which corresponds to selecting only individuals moving every day, we obtain average rest times of 2.49 h and 0.55 h, respectively. To have a consistent description of car mobility, we choose to use the value $\bar{\tau} = 2.49$ h. As our results suggest that the larger $\bar{\tau}$ the better the sampling, our choice also defines a best-case scenario. In our numerical study, we will instead use the whole distributions given above.

Data accessibility. The GeoLife GPS Trajectories used here for empirical validation data are publicly available [56] and available at https://research.microsoft.com/en-us/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/ (last accessed 11 October 2017).

Authors' contributions. R.G., R.L., J.M.L. and M.B. designed the research and wrote the text. R.G. performed the numerical analysis. R.L. performed the data analysis. J.M.L. performed the analytical calculations. R.G. and R.L. prepared the figures.

Competing interests. We declare we have no competing interests.

Funding. R.G. has received funding from the SESAR Joint Undertaking under grant agreement no. 699260 included in the European Union's Horizon 2020 research and innovation programme. R.L. acknowledges support from the James S. McDonnell Foundation through a Postdoctoral Fellowship.

Acknowledgements. R.G. thanks M. Lenormand and T. Louail for useful discussions.

References

1. Vespignani A. 2012 Modelling dynamical processes in complex socio-technical systems. *Nat. Phys.* **8**, 32–39. (doi:10.1038/nphys2160)
2. Zheng Y. 2015 Trajectory data mining: an overview. *ACM. Trans. Intell. Syst. Technol.* **6**, 29–41.
3. González MC, Hidalgo CA, Barabási A-L. 2008 Understanding individual human mobility patterns. *Nature* **453**, 779–782. (doi:10.1038/nature06958)
4. Song C, Koren T, Wang P, Barabási A-L. 2010 Modelling the scaling properties of human mobility. *Nat. Phys.* **6**, 818–823. (doi:10.1038/nphys1760)
5. Raichlen DA, Wood BM, Gordon AD, Mabulla AZ, Marlowe FW, Pontzer H. 2014 Evidence of Lévy walk foraging patterns in human hunter–gatherers. *Proc. Natl Acad. Sci. USA* **111**, 728–733. (doi:10.1073/pnas.1318616111)
6. Gallotti R, Bazzani A, Rambaldi S. 2015 Understanding the variability of daily travel-time expenditures using GPS trajectory data. *EPJ Data Sci.* **4**, 1–14. (doi:10.1140/epjds/s13688-015-0055-z)
7. Gallotti R, Bazzani A, Rambaldi S, Barthelemy M. 2016 A stochastic model of randomly accelerated walkers for human mobility. *Nat. Commun.* **7**, 12600. (doi:10.1038/ncomms12600)
8. Balcan D, Colizza V, Gonçalves B, Hu H, Ramasco JJ, Vespignani A. 2009 Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc. Natl Acad. Sci. USA* **106**, 21 484–21 489. (doi:10.1073/pnas.0906910106)
9. Viswanathan GM, Da Luz MGE, Raposo EP, Stanley HE. 2011 *The physics of foraging: an introduction to random searches and biological encounters*. Cambridge, UK: Cambridge University Press.
10. Reynolds A. 2015 Liberating Lévy walk research from the shackles of optimal foraging. *Phys. Life Rev.* **14**, 59–83. (doi:10.1016/j.plrev.2015.03.002)
11. Brockmann D, Hufnagel L, Geisel T. 2006 The scaling laws of human travel. *Nature* **439**, 462–465. (doi:10.1038/nature04292)
12. Phithakkitnukoon S, Wolf MI, Offenhuber D, Lee D, Biderman A, Ratti C. 2013 Tracking trash. *IEEE Pervas Comput.* **12**, 38–48. (doi:10.1109/MPRV.2013.37)
13. Cagnacci F, Boitani L, Powell RA, Boyce MS. 2010 Animal ecology meets GPS-based radiotelemetry: a perfect storm of opportunities and challenges. *Phil. Trans. R. Soc. B* **365**, 2157–2162. (doi:10.1098/rstb.2010.0107)
14. Sagarra O, Szell M, Santi P, Díaz-Guilera A, Ratti C. 2015 Supersampling and network reconstruction of urban mobility. *PLoS ONE* **10**, e0134508. (doi:10.1371/journal.pone.0134508)
15. Williams NE, Thomas TA, Dunbar M, Eagle N, Dobra A. 2015 Measures of human mobility using mobile phone records enhanced with GIS data. *PLoS ONE* **10**, e0133630. (doi:10.1371/journal.pone.0133630)
16. çolak S, Alexander LP, Alvim BG, Mehndiratta SR, González MC. 2016 Analyzing cell phone location data for urban travel. *Transp. Res. Record* **2526**, 126–135. (doi:10.3141/2526-14)
17. Turchin P. 1998 *Quantitative analysis of movement: measuring and modeling population redistribution in animals and plants*. Sunderland, MA: Sinauer Associates.
18. Codling EA, Plank MJ, Benhamou S. 2008 Random walk models in biology. *J. R. Soc. Interface* **5**, 813–834. (doi:10.1098/rsif.2008.0014)
19. Hebblewhite M, Haydon DT. 2010 Distinguishing technology from biology: a critical review of the use

of GPS telemetry data in ecology. *Phil. Trans. R. Soc. B* **365**, 2303–2312. (doi:10.1098/rstb.2010.0087)

20. Laube P, Purves RS. 2011 How fast is a cow? Cross-scale analysis of movement data. *Trans. GIS* **15**, 401–418. (doi:10.1111/j.1467-9671.2011.01256.x)
21. Salvucci DD, Goldberg JH. 2000 Identifying fixations and saccades in eye-tracking protocols. In *Proc. of the 2000 Symp. on Eye tracking Research & Applications*, pp. 71–78. New York, NY: ACM.
22. Kitamura R, Fujii S, Pas EI. 1997 Time-use data, analysis and modeling: toward the next generation of transportation planning methodologies. *Transp. Policy* **4**, 225–235. (doi:10.1016/S0967-070X(97)00018-8)
23. Fryxell JM, Hazell M, Börger L, Dalziel BD, Haydon DT, Morales JM, McIntosh T, Rosatte RC. 2008 Multiple movement modes by large herbivores at multiple spatiotemporal scales. *Proc. Natl Acad. Sci. USA* **105**, 19 114–19 119. (doi:10.1073/pnas.0801737105)
24. Edwards AM. 2011 Overturning conclusions of Lévy flight movement patterns by fishing boats and foraging animals. *Ecology* **92**, 1247–1257. (doi:10.1890/10-1182.1)
25. Blondel VD, Decuyper A, Krings G. 2015 A survey of results on mobile phone datasets analysis. *EPJ Data Sci.* **4**, 10. (doi:10.1140/epjds/s13688-015-0046-0)
26. Hawelka B, Sitko I, Beinat E, Sobolevsky S, Kazakopoulos P, Ratti C. 2014 Geo-located twitter as proxy for global mobility patterns. *Cartogr. Geogr. Inform. Sci.* **41**, 260–271. (doi:10.1080/15230406.2014.890072)
27. Codling EA, Hill NA. 2005 Sampling rate effects on measurements of correlated and biased random walks. *J. Theor. Biol.* **233**, 573–588. (doi:10.1016/j.jtbi.2004.11.008)
28. Mansilla R. 2010 Stroboscopic observation of a random walker. (http://arxiv.org/abs/1011.5929)
29. Rosser G, Fletcher AG, Maini PK, Baker RE. 2013 The effect of sampling rate on observed statistics in a correlated random walk. *J. R. Soc. Interface* **10**, 20130273. (doi:10.1098/rsif.2013.0273)
30. Reed WJ, Hughes BD. 2002 From gene families and genera to incomes and internet file sizes: why power laws are so common in nature. *Phys. Rev. E* **66**, 067103. (doi:10.1103/PhysRevE.66.067103)
31. Mosetti G, Jug G, Scalas E. 2007 Power laws from randomly sampled continuous-time random walks. *Phys. A.* **375**, 233–238.
32. Plank MJ, Codling EA. 2009 Sampling rate and misidentification of Lévy and non-Lévy movement paths. *Ecology* **90**, 3546–3553. (doi:10.1890/09-0079.1)
33. Codling EA, Plank MJ. 2011 Turn designation, sampling rate and the misidentification of power laws in movement path data using maximum likelihood estimates. *Theor. Ecol.* **4**, 397–406. (doi:10.1007/s12080-010-0086-9)
34. Johnson LR, Boersch-Supan PH, Phillips RA, Ryan SJ. 2017 Changing measurements or changing movements? Sampling scale and movement model identifiability across generations of biologging technology. *Ecol. Evol.* **7**, 9257–9266. (doi:10.1002/ece3.3461)
35. Pappalardo L, Simini F, Rinzivillo S, Pedreschi D, Giannotti F, Barabási A-L. 2015 Returners and explorers dichotomy in human mobility. *Nat. Commun.* **6**, 8166. (doi:10.1038/ncomms9166)
36. Barbosa H, de Lima-Neto FB, Evsukoff A, Menezes R. 2015 The effect of recency to human mobility. *EPJ Data Sci.* **4**, 21. (doi:10.1140/epjds/s13688-015-0059-8)
37. Pappalardo L, Rinzivillo S, Simini F. 2016 Human mobility modelling: exploration and preferential return meet the gravity model. *Procedia Comput. Sci.* **83**, 934–939.
38. Barabási A-L. 2005 The origin of bursts and heavy tails in human dynamics. *Nature* **435**, 207–211. (doi:10.1038/nature03459)
39. Bild DR, Liu Y, Dick RP, Mao ZM, Wallach DS. 2015 Aggregate characterization of user behaviour in Twitter and analysis of the retweet graph. *ACM Trans. Internet Technol.* **15**, 4. (doi:10.1145/2700060)
40. Song C, Qu Z, Blumm N, Barabási A-L. 2010 Limits of predictability in human mobility. *Science* **327**, 1018–1021. (doi:10.1126/science.1177170)
41. Bagrow JP, Lin Y-R. 2012 Mesoscopic structure and social aspects of human mobility. *PLoS ONE* **7**, e37676. (doi:10.1371/journal.pone.0037676)
42. Jiang S, Fiore GA, Yang Y, Ferreira Jr J, Frazzoli E, González MC. 2013 A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In *Proc. of the 2nd ACM SIGKDD Int. Workshop on Urban Computing, UrbComp'13, Chicago, Illinois*, pp. 1–9.
43. Godrèche C, Luck JM. 2001 Statistics of the occupation time of renewal processes. *J. Stat. Phys.* **104**, 489–524. (doi:10.1023/A:1010364003250)
44. Reynolds A. 2008 How many animals really do the Lévy walk? Comment. *Ecology* **89**, 2347–2351. (doi:10.1890/07-1688.1)
45. Benhamou S. 2008 How many animals really do the Lévy walk? Reply. *Ecology* **89**, 2351–2352. (doi:10.1890/08-0313.1)
46. Hoteit S, Secci S, Sobolevsky S, Ratti C, Pujolle G. 2014 Estimating human trajectories and hotspots through mobile phone data. *Comput. Netw.* **64**, 296–307. (doi:10.1016/j.comnet.2014.02.011)
47. Feller W. 1968 *An introduction to probability theory and its applications*. New York, NY: Wiley.
48. Cox DR. 1962 *Renewal theory*. London, UK: Methuen.
49. Cox DR, Miller HD. 1965 *The theory of stochastic processes*. London, UK: Chapman & Hall.
50. Metzler R, Jeon JH, Cherstvy AG, Barkai E. 2014 Anomalous diffusion models and their properties: non-stationarity, non-ergodicity, and ageing at the centenary of single particle tracking. *Phys. Chem. Chem. Phys.* **16**, 24 128–24 164. (doi:10.1039/C4CP03465A)
51. Giannotti F, Nanni M, Pinelli F, Pedreschi D. 2007 Trajectory pattern mining. In *13th ACM SIGKDD Int. Conference, San Jose, California*, pp. 330–339.
52. Wang H, Calabrese F, Di Lorenzo G, Ratti C. 2010 Transportation mode inference from anonymized and aggregated mobile phone call detail records. In *13th Int. IEEE Conference on Intelligent Transportation Systems, Funchal, Portugal*, pp. 318–323.
53. Ranacher P, Brunauer R, Trutschnig W, Van der Spek S, Reich S. 2016 Why GPS makes distances bigger than they are. *Int. J. Geogr. Inf. Sci.* **30**, 316–333. (doi:10.1080/13658816.2015.1086924)
54. Gallotti R, Bazzani A, Rambaldi S. 2012 Towards a statistical physics of human mobility. *Int. J. Mod. Phys. C* **23**, 1250061. (doi:10.1142/S0129183112500611)
55. Proekt A, Banavar JR, Maritan A, Pfaff DW. 2012 Scale invariance in the dynamics of spontaneous behavior. *Proc. Natl Acad. Sci. USA* **109**, 10 564–10 569. (doi:10.1073/pnas.1206894109)
56. Zheng Y, Xie X, Ma W-Y. 2010 GeoLife: a collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.* **33**, 32–40.
57. Zhao K, Musolesi M, Hui P, Rao W, Tarkoma S. 2015 Explaining the power-law distribution of human mobility through transportation modality decomposition. *Sci. Rep.* **5**, 9136. (doi:10.1038/srep09136)
58. de Montjoye YA, Smoreda Z, Trinquart R, Ziemlicki C, Blondel VD. 2014 D4D-Senegal: the second mobile phone data for development challenge. (http://arxiv.org/abs/1407.4885)
59. Gallotti R, Bazzani A, Degli Esposti M, Rambaldi S. 2013 Entropic measures of individual mobility patterns. *J. Stat. Mech.* **2013**, P10022. (doi:10.1088/1742-5468/2013/10/P10022)
60. Cuttone A, Lehmann S, González MC. 2016 Understanding predictability and exploration in human mobility. (http://arxiv.org/abs/1608.01939)