Biases and errors in the temporal sampling of random movements

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

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

New sources of data available thanks to Information and Communication Technologies allow to track the trajectories of humans and animals at an unprecedented scale [1]. In general, the continuous spatio-temporal record of the followed individual can be described as a continuous-time random walk [2], where a *rest* time is associated to the endpoint of each *move*. Identifying these different states is an important statistical challenge [3], in particular because these new sources of information and their exploitation have new limits and biases [4] that need to be assessed.

One of these limits is introduced by the temporal sampling of the trajectory. To reconstruct the real movement patterns, one needs a time Δ between sampled points significantly smaller than the characteristic duration of rests and moves in analysis. This is often not the case. Here, we discuss the effect of sampling on the measured statistical properties of random movements. We describe trajectories as an alternating renewal process [5], a generalization of Poisson processes to arbitrary holding times and to two alternating kinds of events, moves and rests, whose durations t and τ are regarded as independent random variables. The sampling time interval Δ depends on the particular experiment and can be either constant or randomly distributed.

We first consider the case of exponential distributions for $P(t) = \overline{t}^{-1} \exp(t/\overline{t})$ and $P(\tau) = \overline{\tau}^{-1} \exp(\tau/\overline{\tau})$, constant sampling time interval $\overline{\Delta}$, and constant speed v. In this case we can obtain explicitly the distribution $P(\ell^*)$ of sampled displacements and its first two moments, that also allow us to quantify difference between the real $\ell = vt$ and the sampled ℓ^* displacement lengths. The observed distribution $P_{\ell^*>0}(\ell^*)$ can have a maximum, even if the original distribution $P(\ell)$ is a monotonically decreasing function. When $\overline{\Delta} > \overline{\tau}$, the result of the sampling is manifestly different from the original exponential distribution.

We can also calculate the fraction $F_{\rm good}(\overline{\Delta})$ of moves that are correctly sampled with a sampling time $\overline{\Delta}$. This quantity is independent of the spatial embedding and of our assumption on the speed v, and represents an excellent measure of the impact of the sampling. We note in particular that there is an optimal sampling time of the same order as $\overline{t}$ and $\overline{\tau}$: $\hat{\Delta} \approx 2\sqrt{\overline{t}\overline{\tau}}$.

We then extend these results numerically, and show that sampling human trajectories in more realistic settings is necessarily worse than the peaked scenario we solved, which therefore allows us to define an upper bound to sampling quality. Finally, we use high-resolution (spatially and temporally) GPS trajectories [6] to verify our predictions on real data. We find that for real cases, characterized by long-tailed rest durations [7], the fraction of correctly sampled movements is dramatically reduced. Constant sampling allows to recover at best 18% of movements, while even idealized methods cannot recover more than 16% of moves from sampling intervals extracted from human communication data [8].

These figures suggest that, in the sampling of a trajectory alternating rests and movements of animals or humans, the assumptions often made that each measure correspond to a rest and that an observed displacement correspond to a move are intrinsically flawed. Further studies and rigorous analysis of the empirical methods used in many studies are thus necessary in order to construct solid foundations for our understanding of human mobility and animal foraging.

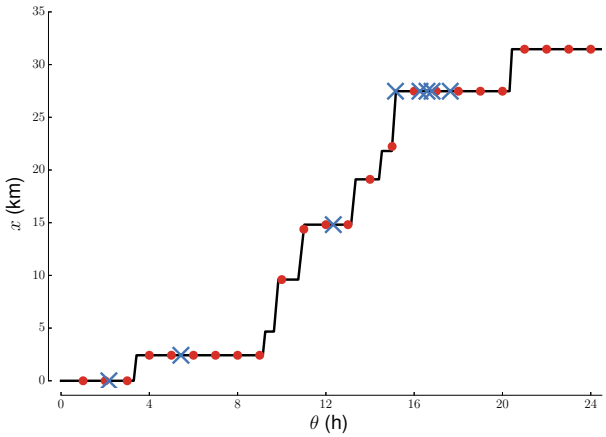

Figure 1: Examples of trajectory sampling with exponentially distributed rest and move durations, we show the case of constant sampling interval (red circles) and the case of random sampling interval (blue crosses) .

[1] Zheng, Y. Trajectory data mining: An overview. *ACM Transactions on Intelligent Systems and Technology* **6**, 29–41 (2015).

[2] Codling, E.A., Plank, M.J. & Benhamou, S. Random walk models in biology. *J R Soc Interface* **5**, 813–834 (2008).

[3] Fryxell, J.M. et al. Multiple movement modes by large herbivores at multiple spatiotemporal scales. *Proc Natl Acad Sci USA* **105**, 19114–19119 (2008).

[4] Cagnacci, F. et al. Animal ecology meets GPS-based radiotelemetry: a perfect storm of opportunities and challenges. *Philos T R Soc B* **365**, 2157–2162 (2010).

[5] Godrèche, C. & Luck, J.M. Statistics of the occupation time of renewal processes. *J Stat Phys* **104**, 489–524 (2001).

[6] Zheng, Y., Xie, X. & Ma, W.-Y. GeoLife: A collaborative social networking service among user, location and trajectory. *IEEE Data Engineering Bulletin* **33** (2), 32–40 (2010).

[7] Proekt, A., Banavar, J.R., Maritan, A. & Pfaff, D.W. Scale invariance in the dynamics of spontaneous behavior. *Proc Natl Acad Sci USA* **109** (26), 10564–10569 (2012).

[8] de Montjoye, Y.A et al. D4D-Senegal: the second mobile phone data for development challenge. *arXiv*:1407.4885 (2014).