An alternative approach to the limits of predictability in human mobility

Ikanovic, Edin Lind; Mollgaard, Anders

Published in:
EPJ Data Science

DOI:
10.1140/epjds/s13688-017-0107-7

Publication date:
2017

Document version
Publisher's PDF, also known as Version of record

Ikanovic and Mollgaard *EPJ Data Science* (2017) 6:12
DOI 10.1140/epjds/s13688-017-0107-7

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

An alternative approach to the limits of predictability in human mobility

Edin Lind Ikanovic and Anders Mollgaard*

*Correspondence: amoellga@nbi.dk
Niels Bohr Institute, University of Copenhagen, Copenhagen, 2100, Denmark

Abstract

Next place prediction algorithms are invaluable tools, capable of increasing the efficiency of a wide variety of tasks, ranging from reducing the spreading of diseases to better resource management in areas such as urban planning. In this work we estimate upper and lower limits on the predictability of human mobility to help assess the performance of competing algorithms. We do this using GPS traces from 604 individuals participating in a multi year long experiment, The Copenhagen Networks study. Earlier works, focusing on the prediction of a participant's whereabouts in the next time bin, have found very high upper limits (>90%). We show that these upper limits are highly dependent on the choice of a spatiotemporal scales and mostly reflect stationarity, i.e. the fact that people tend to *not* move during small changes in time. This leads us to propose an alternative approach, which aims to predict the next location, rather than the location in the next bin. Our approach is independent of the temporal scale and introduces a natural length scale. By removing the effects of stationarity we show that the predictability of the next location is significantly lower (71%) than the predictability of the location in the next bin.

Keywords: human mobility; predictability; limits

1 Introduction

The understanding of human mobility patterns has changed greatly in the last couple of decades. This has mainly been due to new technologies enabling human displacements to be studied with higher accuracy over a longer period of time. Starting with the tracking of bank notes [1] as a proxy for human movement, studies quickly evolved towards the current use of hand held devices for tracking, using either GSM data [2, 3], connections to wifi hotspots [4] or GPS receivers [5] to determine location. The main results from these studies have been the discoveries of power laws governing step size and wait time distributions [1], a universal probability density governing human mobility [6], and simple models capturing many statistical features of human mobility [5–8]. It has furthermore been explored how mobility is affected by recency [9], exploration [10], and return to previously visited places [6] and friends [11]. Such discoveries and models can help predict the spread of diseases [12] and cellphone viruses [13], and also enhance socio-economic forecasting [14–16], city planning [17] and many other fields [5, 18, 19]. Further contribution to progress in these areas can be made if geolocation data can be used to accurately predict an individual's future whereabouts. A crucial part of this work is the construction

of viable evaluation mechanisms, thereby raising the question: what are the upper and lower limits, $\Pi^{\max}$ and $\Pi^{\min}$, on the predictability of human mobility?

This question was initially investigated using call detail records from 45,000 cellphones [3]. Each call corresponded to a known location represented by a Voronoi cell, around the closest cell tower, with an average area of 3 km^2. The known locations were grouped into 1 hour bins, giving a history of locations T_i, for each user i. The work focused on determining how well the best possible algorithm can predict the location of an individual in the next time bin, given T_i. They reported an upper limit narrowly peaked at $\Pi^{\max} = 93\%$ and a lower limit of $\Pi^{\min} = 70\%$.

This work led to questions being raised about possible biases introduced when using call detail records [20] and about the influence of spatiotemporal scales [21]. The temporal resolution [22, 23] and spatial resolution [4, 23, 24] were investigated with GSM and GPS data for smaller populations. Overall, it was found that the predictability increases with temporal resolution and decreases with spatial resolution. The limits of predictability, as defined in [3], therefore depend on the choice of temporal resolution Δt and spatial resolution Δs.

Here we make the following conjecture:

$$\Pi^{(\max,\min)}(\Delta t, \Delta s) \to 1 \quad \text{when } \Delta t \to 0 \text{ or } \Delta s \to \infty. \tag{1}$$

The rationale behind this expression is that the location of the next time bin will almost certainly *not* change in both limits. At small time scales and at large spatial scales you always know where an individual is going to be in the next time bin: he/she will be in the same spatial bin. We therefore argue that the current limits on the predictability of mobility to a large extent reflect stationarity. Previous results therefore mix two different questions, namely

– How long will an individual stay in his/her current location?
– Where will he/she go next?

Here we propose an analysis that is able to separate out the first question such that we can concentrate on the second. This is achieved by focusing on the *next location*, rather than the location in the next bin. This approach is independent of Δt, provided a small sampling rate. By introducing a natural length scale, we are able to get a single number for the predictability of human mobility, rather than a function of spatiotemporal resolution. Our new approach shows that the upper limit on the predictability of this type of mobility is around $\sim$71%, rather than the >90% found in earlier works. We thereby show that the high upper limits of previous works mostly reflect stationarity, rather than movement.

2 Data and methods

The Copenhagen networks study. Our dataset comes from a large scale study involving approximately 1,000 students over multiple years [25]. Each participant was issued a smartphone capable of recording across multiple channels, including calls, text, bluetooth, and GPS coordinates. In addition to this, the participants answered a questionnaire that, among others, allowed a psychological profile to be inferred. In this paper we mainly use the location data, determined using a combination of the GPS sensors and the network that the phone is connected to. Location data was only available for 849 participants and

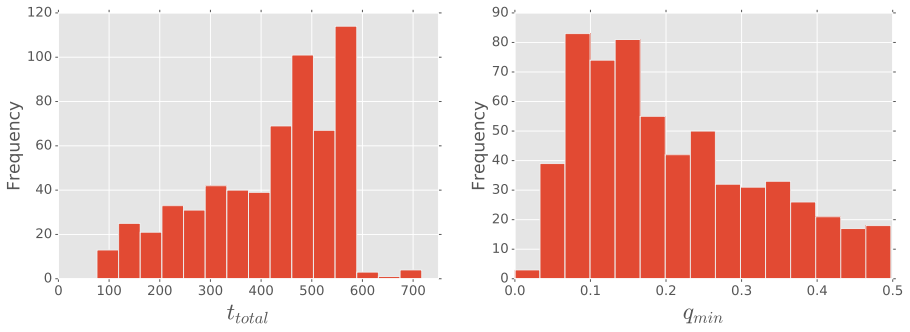


Figure 1 Population statistics. Left panel: the distribution of the number of days that the participants took part in the experiment. Right panel: the distribution of the fraction of missing data for $\Delta t = 15$ min.

consists of $\approx 2.4 \cdot 10^8$ data points. The data was collected from February 2012 up to March of 2015, thus covering a multi year span with a substantial fraction participating for more than a year (see left panel of Figure 1). Each data point consists of latitude and longitude coordinates, together with a timestamp and the accuracy associated with the measurements. These are converted into appropriate time series (see Mobility sequences and predictability for details), and the fraction of bins with unknown locations is denoted $q_{\min}$. For our analysis we need $q_{\min} \leq 50\%$ (see Methods). This reduces the number of participants with sufficient data to 604. The right panel of Figure 1 shows the distribution of $q_{\min}$ at the lowest temporal scale (15 minutes).

Mobility sequences and predictability. The raw GPS data needs to be filtered and converted into a history of discrete locations, T_i, before the limits of predictability can be determined. This can in principle be done in an infinite number of ways, meaning that the GPS trace from a participant can give rise to many different time series T_i depending on the filtering and mapping chosen. In this work we convert the raw data into two different time series:

– T_i^{bins}: Series of time bins.
– T_i^{loc}: Series of locations.

A detailed description of the filters and mappings are given in the Methods section.

An illustration of the conversion from GPS-trace into T_i^{bins} is shown schematically in Figure 2. The two dimensional space is covered by a grid with a grid length given by Δs. Each square in the grid is represented by a symbol, such that a human trajectory may look like this

$$T_i^{\text{bins}} = [A, B, B, A, A, A, C, \ldots]. \tag{2}$$

Each symbol corresponds to the grid cell position of a time bin of length Δt. The construction of this trajectory is equivalent to that of earlier works [3, 4, 21–23]. As noted earlier, it depends on the spatiotemporal resolution and includes stationarity.

Next we introduce the new mobility encoding T_i^{loc}, which aims to describe trajectories by a sequence of unique locations. Details can be found in the Methods section. We start by filtering all the GPS information such that travel between locations is removed. This leaves us with a set of stationary GPS points that are distributed around the preferred places

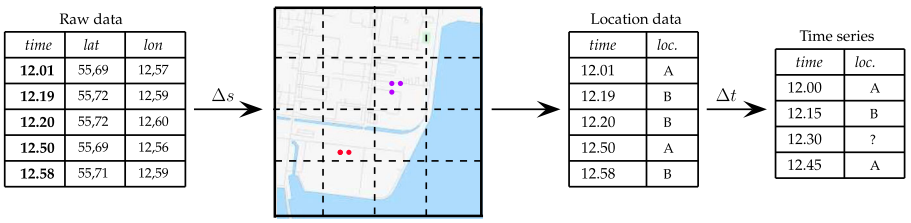


Figure 2 Converting the raw data for participant *i* into a suitable time series, T_i^{bins}. After filtering we plot the data points onto a world map overlaid with square grid cells with side lengths Δs. This converts each data point into a *location* represented by a square grid cell and encoded by a symbol in T_i^{bins}. The location data is then resampled such that each bin in T_i^{bins} corresponds to a time interval Δt. This mobility encoding is similar to that of earlier works and corresponds to the location in a sequence of time bins. We propose to look at a the sequence of locations instead.

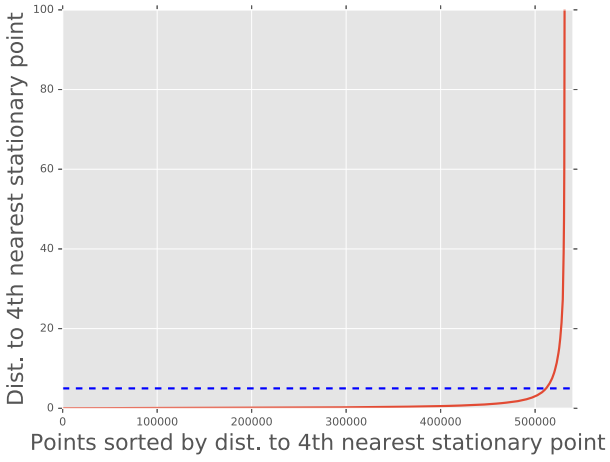


Figure 3 Choosing a length scale. We show the sorted distance to the fourth nearest stationary point for a subset of the entire data set. Standard methods [27] propose setting the parameter $\epsilon_{\text{vicinity}}$ equal to the distance where there is a sharp increase in the distance to the fourth nearest stationary point. Here we choose $\epsilon = 5$ meters as indicated by the dashed line.

of the individual. We then use a clustering algorithm (DBSCAN [26]) on the stationary data points to determine the different locations automatically. This approach results in locations, which better represent the places where individuals spend their time, than the more commonly used Voronoi or square grid cells.

The clustering algorithm takes a length scale as input, which determines whether or not a stationary data point belongs to a location cluster. Here we use $\epsilon_{\text{vicinity}} = 5$ meters meaning that if a stationary data points is more than five meters from all points in a location cluster, then it is considered as not belonging to that location. This length scale is based on an analysis of "the fourth nearest point"-distribution as proposed in [27] (see Figure 3). For the second parameter of the DBSCAN-algorithm, min_pts, we also follow the standards given in the reference, which says to use min_pts = 4. This parameter value defines a location cluster as a minimum of four stationary points, i.e. at least 1 hour must be spent in a five meter vicinity during the full sampling period for a cluster to be considered a location.

We can now construct the trajectory of an individual among his/her locations, using the clusters found by the DBSCAN algorithm. In this encoding we do not include the time spent at the different locations, but represent each location by just a single symbol, e.g.:

$$T_i^{\text{loc}} = [A, B, A, C, \ldots]. \tag{3}$$

Compare this with the sequence in (2) and note that the stationarity has been removed, i.e. no similar symbols in a row.

We expect the sequence of locations to be less predictable than the sequence of time bins, since it encompasses the more complicated spatial dynamics. In order to quantify this intuition, we need a measure of predictability. Here we use a slightly modified version of the scheme developed by [3] (see Methods for details). First, the entropy rate of the mobility sequence is determined using an estimator based on the Lempel-Ziv compression algorithm. Since all the sequences are affected by missing data, one must extrapolate the entropy rate from missing data to full data. By testing our extrapolation on periods with complete data, we find that we can predict the true entropy within 10%, even when 50% of the sequence is missing. Having estimated the entropy rate H_{est} we are in a position to determine the upper limit of predictability $\Pi^{\max}$. This is done by solving [3]

$$H_{\text{est}} = -\Pi^{\max}\log_2\left(\Pi^{\max}\right) - \left(1-\Pi^{\max}\right)\log_2\left(1-\Pi^{\max}\right) + \left(1-\Pi^{\max}\right)\log_2(N-1), \tag{4}$$

where N is the number of unique locations in the time series. The upper limit found represents a tight upper bound attainable by an appropriate, but for now unknown, algorithm.

We also examine the lower limit of predictability. For the location sequence T_i^{loc}, we use a first order Markov chain to predict the next location [28], i.e. we expect the location that most often follows the current location. If the current location has not been explored before, then we expect the most visited location as the next one. For the time bin sequence T_i^{bins} we use a simple predictor, which expects the current location to continue into the next time bin. This predictor will be referred to as "the trivial predictor" and it measures the amount of stationarity in the mobility sequence.

3 Results

We start by presenting our results for T_i^{bins}, i.e. the mobility encoding that people have been using previously. As noted earlier, the predictability of these sequences depend on the spatiotemporal resolution. In the left panel of Figure 4 we fix $\Delta s = 400$ m and vary Δt to determine how the upper and lower limits depend on the temporal scale. The predictability grows towards 1 as the time scale is decreased, just as expected by our conjecture (1). Note the high performance of the trivial predictor (70%-91%).

Next we fix the temporal scale $\Delta t = 15$ min and vary the spatial scale Δs (Figure 4, right panel). Both the upper limit (squares) and lower limit (discs) increase when Δs is increased, again in agreement with (1). We note that the upper limit is not very sensitive to the spatial scales investigated here ($\Delta s > 100$ m). We furthermore note the impressive performance of the trivial predictor at large spatial scales. For comparison we also compute the limits of predictability at the spatiotemporal scales considered in [3] ($\Delta t = 60$ min and $\Delta s = 1.7$ km). We find that the trivial predictor is successful in $88.3 \pm 3.8\%$ of the cases, while the upper bound is $95.5 \pm 1.8\%$, i.e. almost all of the predictability reflects the fact that people do *not* change location.

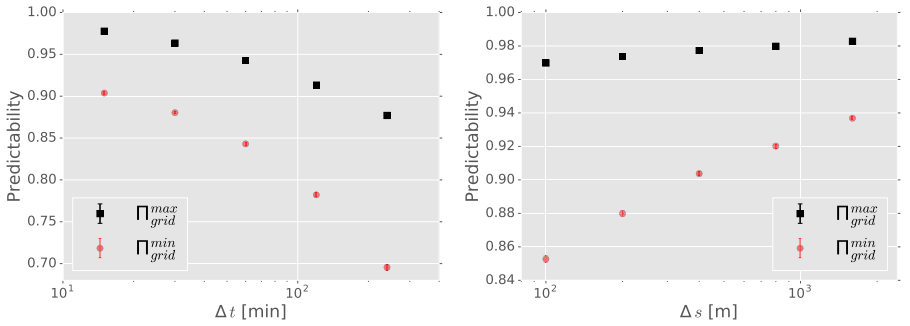
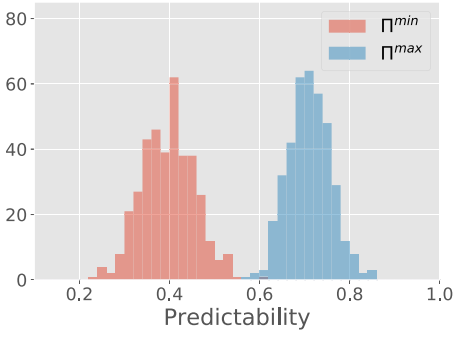


Figure 4 ***Next bin*** **predictability depends on spatiotemporal resolution.** Left panel: The temporal resolution dependency of the upper limits (squares) and lower limits (discs) of predictability for the next bin approach. Each location is represented by a square grid with $\Delta s = 400$ m. Error bars are included but are smaller than the symbols. Right panel: The spatial dependency at a fixed temporal resolution of $\Delta t = 15$ min. The lower limit shows that, depending on resolution, 85% to 94% of the predictability is due to people *not* moving.

Figure 5 The distributions of the upper and lower limits for *next location* predictions. Both results, $\Pi^{\max} = 71.1 \pm 4.7\%$ and $\Pi^{\min} = 39.8 \pm 5.9\%$, are significantly smaller than the limits found for *next bin* predictions (Figure 4). We conclude that previous work overestimates the predictability of dynamic human mobility.



The limits presented in Figure 4 follow our postulate and are in agreement with earlier works with smaller populations. We now test what happens when we remove the stationarity from the spatial dynamics, i.e. when we consider the predictability of the next location instead. In Figure 5 we show the distributions of the upper and lower limits for next location predictability. Both limits are strongly reduced when compared to the results for next bin predictability. For the upper limit we find $\Pi^{\max} = 71.1 \pm 4.7\%$, i.e. a significant reduction from the >90% predictability found in previous works. We find that this value is very robust to increases in the length scale and that it only changes by a few percent as ϵ is increased towards 25 meters. The lower limit is found to be $\Pi^{\min} = 39.8 \pm 5.9\%$, which is at least 30% lower than any of the lower limits found by the trivial predictor for next bin sequences.

We note that another group has simultaneously been working on the same data set with the same methods and they have found $\Pi^{\max} = 0.68$ [29]. Despite the close match in results they have actually been using very different DBSCAN paramters, namely $\epsilon_{\text{vicinity}} = 50$ (we use $\epsilon_{\text{vicinity}} = 5$) and min_pts = 2 (we use min_pts = 4), thereby further underlining the robustness of the results. Our main contribution relative to their work is to derive the length scale from the data, to directly state and investigate conjecture (1), and to relate the predictability of the next location to psychological factors.

Table 1 Examining which factors impact the predictability of human mobility patterns. r_g is the radius of gyration, $\mathbf{eff}_{\mathbf{places}}$ is the effective number of places an individual chooses from when changing to a new location and is defined as $2^{H_{\mathbf{unc}}}$. We also examine the impact of basic personality traits using the Big Five psychological profile [30]. Error bars are determined using the bootstrap method

Measure	Correlation with $\Pi_{\mathbf{max}}$
r_g	-0.05 ± 0.05
$\mathbf{eff}_{\mathbf{places}}$	-0.26 ± 0.05
$\Pi_{\mathbf{min}}$	0.49 ± 0.04
Agreeableness	-0.05 ± 0.06
Conscientiousness	0.04 ± 0.06
Extroversion	-0.13 ± 0.05
Neuroticism	0.06 ± 0.06
Openess	-0.004 ± 0.059

The above results raise the question: what factors impact the predictability of human mobility? Our partial answer to this question can be found in Table 1, where we correlate $\Pi^{\max}$ to a range of variables. We find that radius of gyration, representing typical distances traveled, does not impact next place predictability. A related result has been reported earlier, using next bin predictability [3]. While this result can seem counterintuitive, our next result is able to shed more light on the matter. $\Pi_{\max}$ is anti-correlated with the effective number of places an individual chooses from, when determining where to go next. Therefore, the predictability of an individual does not depend on the reach of his/her travels, but rather on the number of places visited.

Finally, utilizing the psychological profiles of the participants, we are able to examine the impact of their psychological traits on their predictability. The only significant correlation we find here is with extroversion, meaning that the next location of an extroverted individual is statistically harder to predict.

4 Conclusion

Our results show that it is possible to extract a wide range of upper and lower limits of predictability of human mobility depending on the filtering and discretization scheme chosen. We have shown the strong dependency of "next bin" predictability on spatiotemporal scales. Furthermore, we have shown that the predictability at large spatial scales and small temporal scales mostly reflect stationarity, namely that people stay in the same spatial bin. This raises the need for an alternative approach to estimate the predictability of human mobility patterns.

The task of predicting human mobility is two fold: how long will a person stay in a certain location and where they will go next. Here we determined an upper limit on the predictability of the latter. We found that the upper limit of this task is much lower than the previously stated ones of $\sim$93%. In particular, by using the natural length scale of human locations we found an upper limit on predictability of $71.1 \pm 4.7\%$. A lower limit was likewise found using a first order Markov chain model with a success rate of $39.8 \pm 5.9\%$. Overall, our results indicate that it might not be so trivial to predict human mobility after all.

5 Methods

Converting the raw data into T_i^{bins}. We start by employing an accuracy filter, which removes all the data points with an accuracy below 50 meter. The grid map used is char-

acterized by two parameters: a length scale Δs and the origin of the map. The Technical University of Denmark, where most of the participants were enrolled, was chosen as the origin. This ensured that the grid cells had sides of approximately equal length Δs at the locations where most of the data was collected. The length scales used are $\Delta s \in [100, 200, 400, 800, 1600]$ meters.

Small changes in the origin of the grid map can effect the number of locations detected [24]. To mitigate the possible bias introduced by having a fixed origin of the grid map, we add a random offset for each participant chosen randomly from a uniform distribution on $[0, \Delta s]$.

Our data was not sampled at a fixed rate. A time binning with a fixed temporal resolution Δt allowed us to convert the raw data into a time series. The binning is done such that for each time bin we chose the most visited location. If two or more locations are the most visited locations, then we chose one of them at random. The time scales used are $\Delta t \in [15, 30, 60, 120, 240]$ minutes. Time bins with no recorded locations are denoted using a special ? marker. Thus we end up with a time series T_i^{bins} which depends primarily on Δs and Δt.

Converting the raw data into T_i^{loc}. Again we start by employing the accuracy filter. To reduce the number of data points associated with travel, we employ a second filter inspired by the *pause-based* model used in [5]. It detects all the data points which are 15 ± 1.5 min apart and for which the distance between the two measurements are less than 100 m. These two measurements are then averaged into a single data point representing a place where a participant stood still for roughly a quarter of an hour. This filters out most of the travel information in the dataset, except interruptions such as traffic jams and waiting for public transport.

The list of locations is binned with a fixed temporal resolution $\Delta t = 15$ min as described above. After this we compress every time series such that all instances where a participant stood still for more than one time bin are represented by just a single symbol. This is best explained by an example. A time series obtained by the procedures described above could look like: $T_i = [A, ?, A, B, B, A, A, A, C, \ldots]$. After compression this time series is converted into:

$$T_i^{\text{loc}} = [A, B, A, C, \ldots]. \tag{5}$$

The resulting time series are independent of Δt provided that Δt is small. The smallest sampling rate that we dare use in this study is $\Delta t = 15$, since smaller sampling rates would make it difficult to distinguish stationarity from movement because of the limited accuracy of the GPS.

Estimating the entropy rate. The entropy rate is found using an estimator based on the Lempel-Ziv compression algorithm [3]:

$$H_{\text{rate}} = \left(\frac{1}{n} \cdot \sum_{i=1}^{n} \frac{\Lambda_i}{\log(n)} \right)^{-1}, \tag{6}$$

where n is the length of the time series and Λ_i is the length of longest substring in the time series starting from position i and not encountered earlier from position 1 to $i - 1$. This estimator has been shown to converge rapidly towards the entropy rate [31].

The fraction of missing data, q, changes the entropy rate estimate. By artificially removing data in complete records we can study possible extrapolation methods. We have used a subset of 47 individuals with a complete location record spanning at least 2 weeks. For each of these complete records we determined H_{true} using the estimator (6). Removing data from these complete records and comparing the entropy rate determined by our method, H_{est}, with H_{true}, we found that we could estimate H_{true} within $\pm 10\%$ as long as $q \leq 0.5$. Our method is thus able to determine the entropy rate even when we only know half of the locations visited. Earlier this method has been used up to $q \leq 0.7$ [3], but our tests show reliable results only when $q \leq 0.5$.

Our extrapolation works as follows. For each time series we determine the amount of time the participants location was unknown. This fraction of the total time was denoted q_{min}. We then found both $H_{\text{unc}}(q)$ and $H_{\text{rate}}(q)$ for each $q \in [q_{\text{min}}, q_{\text{min}} + 0.05, q_{\text{min}} + 0.1, \ldots, 0.9 - q_{\text{min}}]$. Here H_{unc} is the entropy of the time series, found using $H_{\text{unc}} = -\sum_{i=1}^{N} p_i \log_2(p_i)$, where the sum runs over all the N different locations visited and p_i is the fraction of time spent at i. This enabled us to calculate $\sigma(q) = H_{\text{rate}}(q)/H_{\text{unc}}(q)$. Earlier it has been shown [3] that $\sigma(q)$ depends linearly on q. This linear relation has not been found when using data with a higher sampling rate [22]. Our set of complete records showed that $\sigma(q)$ could be fitted well with an offset exponential function. Using these fits we could extrapolate and determine $\sigma_{\text{est}} = \sigma(q = 0)$. The entropy rate was then found using

$$H_{\text{est}} = \exp^{\sigma_{\text{est}}} \cdot H_{\text{unc}}(q). \tag{7}$$

Funding
The study received funding through the UCPH 2016 Excellence Programme for Interdisciplinary Research.

List of abbreviations
GSM: Global System for Mobile Communications
GPS: Global Positioning System
DBSCAN: Density-based spatial clustering of applications with noise

Availability of data and materials
Data are part of larger study "Social Fabric" involving researchers at the Technical University of Denmark and University of Copenhagen. Due to privacy consideration regarding subjects in our dataset, including European Union regulations and Danish Data Protection Agency rules, we cannot make all data used here publicly available. The data contains detailed information on mobility and daily habits at a high spatio-temporal resolution. We understand and appreciate the need for transparency in research and are ready to make the data available to researchers who meet the criteria for access to confidential data, sign a confidentiality agreement, and agree to work under our supervision in Copenhagen.

Ethics approval and consent to participate
The "Social Fabric" study was reviewed and approved by the appropriate Danish authority, the Danish Data Protection Agency (Reference number: 2012-41-0664). The Data Protection Agency guarantees that the project abides by Danish law and also considers potential ethical implications. All subjects in the study gave written informed consent.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
Conceived and designed the study: EI AM. Analyzed the data: EI. Wrote the paper: EI AM.

Publisher's Note

Received: 19 September 2016 Accepted: 4 June 2017 Published online: 19 June 2017

References

1. Brockmann D, Hufnagel L, Geisel T (2006) Nature 439(7075):462
2. Gonzalez MC, Hidalgo CA, Barabasi AL (2008) Nature 453(7196):779

3. Song C, Qu Z, Blumm N, Barabási AL (2010) Science 327(5968):1018 http://science.sciencemag.org/content/327/5968/1018. doi:10.1126/science.1177170
4. Qian W, Stanley KG, Osgood ND (2013) In: Web and wireless geographical information systems. Springer, Berlin, pp 25-40
5. Rhee I, Shin M, Hong S, Lee K, Kim SJ, Chong S (2011) IEEE/ACM Trans Netw 19(3):630
6. Song C, Koren T, Wang P, Barabási AL (2010) Nat Phys 6(10):818
7. Jiang S, Yang Y, Gupta S, Veneziano D, Athavale S, González MC (2016) Proceedings of the National Academy of Sciences p 201524261
8. Pappalardo L, Simini F (2016) arXiv preprint arXiv:1607.05952
9. Barbosa H, de Lima-Neto FB, Evsukoff A, Menezes R (2015) EPJ Data Sci 4(1):21
10. Pappalardo L, Simini F, Rinzivillo S, Pedreschi D, Giannotti F, Barabási AL (2015) Nature communications 6
11. Toole JL, Herrera-Yaqüe C, Schneider CM, González MC (2015) J R Soc Interface 12(105):20141128
12. Colizza V, Barrat A, Barthelemy M, Valleron AJ, Vespignani A (2007) PLoS Med 4(1):e13
13. Kleinberg J (2007) Nature 449(7160):287
14. Gabaix X, Gopikrishnan P, Plerou V, Stanley HE (2003) Nature 423(6937):267
15. Pappalardo L, Vanhoof M, Gabrielli L, Smoreda Z, Pedreschi D, Giannotti F (2016) Int J Data Sci Anal 2(1-2):75
16. Frias-Martinez V, Virseda J (2012) In: Proceedings of the fifth international conference on information and communication technologies and development. ACM, New York, pp 76-84
17. Makse HA, Andrade JS, Batty M, Havlin S, Stanley HE et al (1998) Phys Rev E 58(6):7054
18. Kitamura R, Chen C, Pendyala RM, Narayanan R (2000) Transportation 27(1):25
19. Krings G, Calabrese F, Ratti C, Blondel VD (2009) J Stat Mech Theory Exp 2009(7):L07003
20. Ranjan G, Zang H, Zhang ZL, Bolot J (2012) Mob Comput Commun Rev 16(3):33
21. Lin M, Hsu WJ (2014) Pervasive Mob Comput 12:1
22. Jensen BS, Larsen JE, Jensen K, Larsen J, Hansen LK (2010) In: Machine learning for signal processing (MLSP), 2010 IEEE international workshop on. IEEE, New York, pp 196-201
23. Smith G, Wieser R, Goulding J, Barrack D (2014) In: Pervasive computing and communications (PerCom), 2014 IEEE international conference on. IEEE, New York, pp 88-94
24. Lin M, Hsu WJ, Lee ZQ (2012) In: Proceedings of the 2012 ACM conference on ubiquitous computing. ACM, New York, pp 381-390
25. Stopczynski A, Sekara V, Sapiezynski P, Cuttone A, Madsen MM, Larsen JE, Lehmann S (2014) PLoS ONE 9(4):e95978
26. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) J Mach Learn Res 12:2825
27. Tan PN, Steinbach M, Kumar V (2005) Introduction to data mining. Pearson, Upper Saddle River
28. Lu X, Wetter E, Bharti N, Tatem AJ, Bengtsson L (2013) Scientific reports 3
29. Cuttone A, Lehmann S, González MC (2016) arXiv preprint arXiv:1608.01939
30. Digman JM (1990) Annu Rev Psychol 41(1):417
31. Kontoyiannis I, Algoet PH, Suhov YM, Wyner AJ (1998) Information theory IEEE Trans Inf Theory 44(3):1319