SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
STATISTICAL METHODS
COMPUTATIONAL BIOLOGY AND BIOINFORMATICS
INFORMATION TECHNOLOGY
EPIDEMIOLOGY

Received
21 May 2013

Accepted
19 September 2013

Published
11 October 2013

Correspondence and requests for materials should be addressed to X.L. (xin.lu@flowminder.org)

Approaching the Limit of Predictability in Human Mobility

Xin Lu[1,2,3,4], Erik Wetter[2,5], Nita Bharti[6], Andrew J. Tatem[7,8] & Linus Bengtsson[2,3]

[1]College of Information System and Management, National University of Defense Technology, 410073 Changsha, China, [2]Flowminder Foundation, 17177 Stockholm, Sweden, [3]Department of Public Health Sciences, Karolinska Institutet, 17177 Stockholm, Sweden, [4]Department of Sociology, Stockholm University, 17177 Stockholm, Sweden, [5]Department of Management and Organization, Stockholm School of Economics, 11383 Stockholm, Sweden, [6]Department of Biology, Center for Infectious Disease Dynamics and Huck Institute of Life Sciences, Penn State University, University Park, PA 16801, USA, [7]Department of Geography and Environment, University of Southampton, Highfield, Southampton, United Kingdom, [8]Fogarty International Center, National Institutes of Health, Bethesda, MD 20892, USA.

In this study we analyze the travel patterns of 500,000 individuals in Cote d'Ivoire using mobile phone call data records. By measuring the uncertainties of movements using entropy, considering both the frequencies and temporal correlations of individual trajectories, we find that the theoretical maximum predictability is as high as 88%. To verify whether such a theoretical limit can be approached, we implement a series of Markov chain (MC) based models to predict the actual locations visited by each user. Results show that MC models can produce a prediction accuracy of 87% for stationary trajectories and 95% for non-stationary trajectories. Our findings indicate that human mobility is highly dependent on historical behaviors, and that the maximum predictability is not only a fundamental theoretical limit for potential predictive power, but also an approachable target for actual prediction accuracy.

Studies of mobility patterns and predictions of individual mobility trajectories are important in many research fields, such as mobile computing, epidemic modeling, traffic planning and disaster response[1–3]. Real-time locations visited by individuals are typically collected through mobile devices equipped with global-positioning system (GPS) capability, mobile phone cell towers, or wireless local area network (WLAN) access points.

Various methods have been proposed to forecast individual trajectories, including Markov chain (MC) models[4,5], neural networks[6], Bayesian networks[7], and finite automaton[8]. Prediction accuracy has been shown to vary according to the algorithm used and the context from which the location data come. For example, in an evaluation of next cell prediction based on more than 6000 users on Dartmouth's campus-wide Wi-Fi wireless network, it was found that the best predictor (the $\mathcal{O}(2)$-MC model) had an accuracy of about 65–72%[9]. In another study where mobility traces of six researchers and GPS-locations of 175 individuals were used, the prediction accuracy was shown to be in the range of 70% to 95% with an $\mathcal{O}(2)$-MC model[10–12]. On the other hand, in an evaluation of MC models for pedestrian-movement prediction, the prediction accuracy was as low as 2%, 45% and 74.4% for the $\mathcal{O}(1)$-MC model, hidden-Markov model, and the mixed MC model, respectively[13].

The above studies investigated small numbers of individuals or special populations, and the results and practical feasibility of the proposed new predictive algorithms were therefore difficult to generalize to the general population. In addition, it was not clear how well these algorithms performed versus the best possible algorithm that could theoretically be constructed; i.e., what is, for the given data type, the best possible accuracy achievable and how well do the predictive algorithms perform versus such a benchmark? The highest potential accuracy of predictability, termed "maximum predictability" (Π^{max}), is defined by the entropy of information of a person's trajectory (frequency, sequence of location visits, etc.). Π^{max} can be calculated by solving a limiting case of Fano's inequality (a relation derived from calculation of the decrease in information in a noisy information channel)[14–16].

By measuring Π^{max}, Song et al showed, using a mobile phone dataset of 50,000 users in a high-income country, that there is a 93% potential predictability in user mobility, despite very large differences in travel distances[17]. Under much more extreme conditions and in a low-income setting, Lu et al analyzed a complete mobile phone dataset of 2.9 million anonymous subscribers after the earthquake in Haiti in 2010, and found that despite massive population movements and increased travel distances following the earthquake, the predictability of people's movements remained as high as 85%, indicating a fundamental regularity in human mobility[1]. These findings are

promising for the design and improvement of predictive algorithms. However, these studies did not show how close to the maximum potential predictability the accuracy of actual algorithms can come in practice.

In this study we aim to fill this gap in knowledge by measuring the maximum predictability and performance of actual prediction algorithms on a mobile phone data set of 500,000 users from Cote d'Ivoire (CIV), West Africa. We also give an overview of population mobility patterns during the data collection period, which took place after the 2011 civil war. We find that the maximum predictability and regularity in mobility in CIV is high, similar to what was found in studies in Haiti and Europe[1,17]. The evaluation of practical predictive algorithms on this dataset reveals that the maximum predictability can be approached with MC-based models. Interestingly, we also show that higher order MC models do not generate improved prediction accuracy when compared to a first order MC model.

Results

The mobile phone dataset. Mobility data was provided by the telecom company Orange and derived from call detail records (CDR) from a random sample of 500,000 anonymous Orange mobile phone subscribers, active during December 1, 2011 to April 28, 2012 in CIV. The user's location was provided as the location of the subprefecture (sous-préfecture in French) of the mobile phone tower through which the call was routed. CIV is composed of 19 regions, which are further divided into 255 subprefectures (237 of these subprefectures have at least one tower, see Fig. 1). The original CDR contains approximately five million users (1/4 of the total

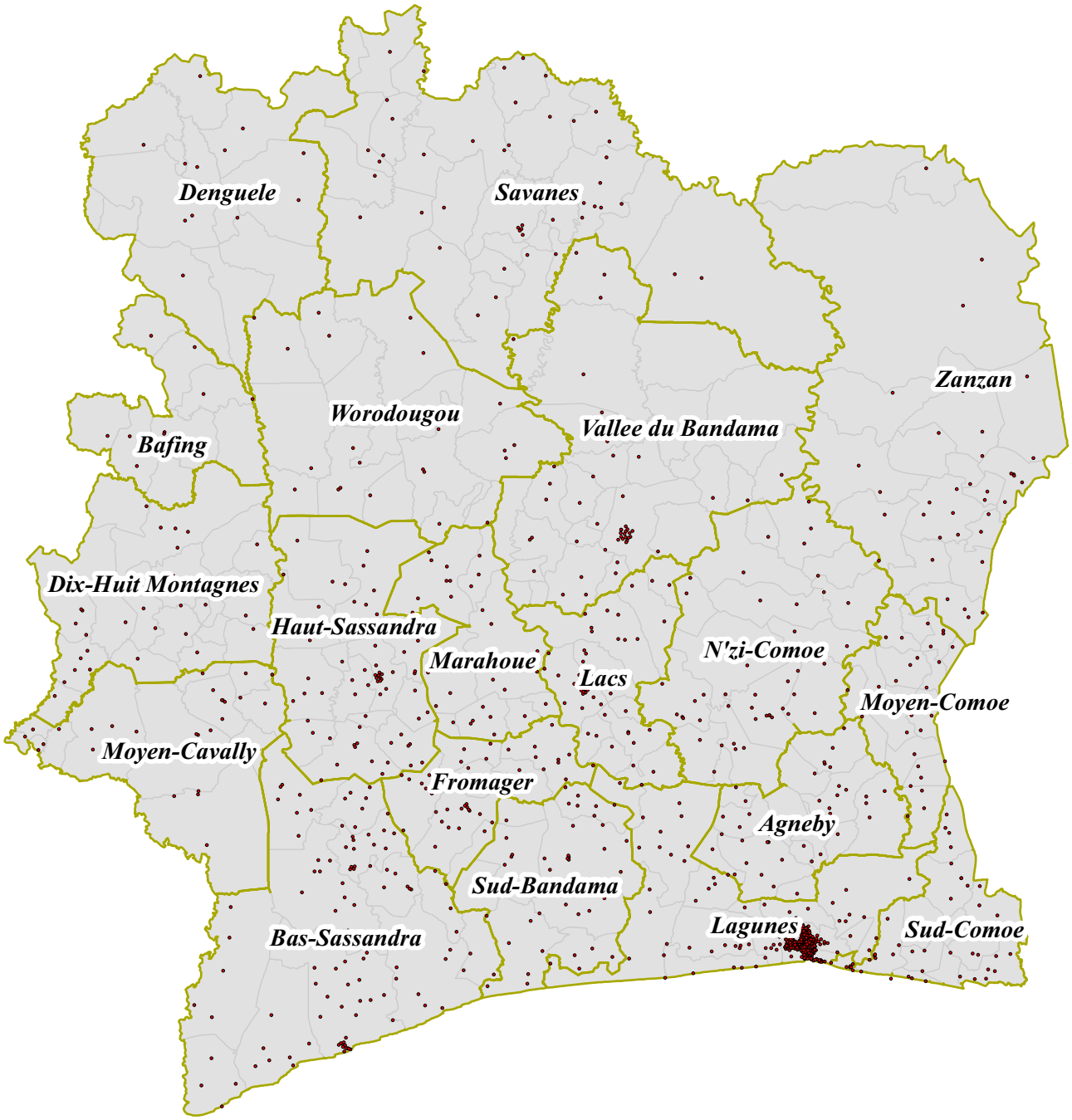

Figure 1 | Administrative map of Cote d'Ivoire and distribution of cell phone towers. (http://sodexo.orange-labs.fr/GEOM_SUB_PREFECTURE.zip).

population of CIV)[18,19]. Detailed description of the data can be found in[20].

The number of Orange subscriptions per person varies considerably throughout the country, as does the overall population density in CIV. For example, the region of Lagunes, which includes the economic capital Abidjan, is home to 25% of the CIV population and is the most frequently visited location for 43% percent of the mobile phone subscribers in this dataset (see Fig. 2A). The distribution of the number of location updates (calls and SMS) follows a log-normal distribution (Fig. 2B), with 81.3% having between 100 and 2000 location updates. Seventy-seven percent of users had at least one location update per day during two thirds of the data collection period (Fig. 2C). Heterogeneity in visitation patterns was high. While sixteen percent of subscribers were only found in one subprefecture during the period, a few users were registered in more than 50 subprefectures (Fig. 2D).

Overview of mobility and aggregated flows. The absolute change in the number of subscribers in each region is dominated by the changes in the region of Lagunes, where Abidjan is situated (Fig. 3A). Seven-day cyclical patterns (workday-weekend cycles) are clearly visible for several regions, e.g. Lagunes and Sud-Comoe, but other more complex trends are also evident. An irregular change in population flow was observed near the end of March and early April when the numbers of users rapidly increased in Abidjan then decreased a few days later (potentially partly related to Easter). In addition, Bas-Sassandra, in the southwest experienced a decrease of users during large parts of the period studied here.

In relative terms (Fig. 3B), several regions showed considerable changes over the period of time analyzed, dominated by Denguele, which showed a small change in absolute terms (see Fig. 3A). As we see from Fig. 3C, both the distribution of average travel distances, $\bar{D}$, and the radius of gyration, r_g, (see Methods for definition) illustrate a skewed decay over increasing traveling distances. While the movements of the vast majority of users were confined within an area of 10 km, a few users traveled on average as far as 100 to 300 km per day (Fig. 3D). Note that the radius of gyration is calculated from location data on the level of the sub-prefecture and thus excludes short movements.

Regularity and potential predictability. We now focus on the regularity of the daily observed trajectories of the users by allocating the last observed location (subprefecture) to each user's trajectory. To avoid the illusion of high predictability stemming from users with many unknown locations, or from users who never traveled to other locations, we include users who visited at least two subprefectures, and were observable for more than 120 days in the period (208,288 users).

The distributions of users' random entropy (S^{rand}), temporal-uncorrelated entropy (S^{unc}) and true-entropy (S^{real}) are presented in Fig. 4A (see Methods). We can see that, consistent with findings from previous studies, the entropy of visited locations is greatly reduced if

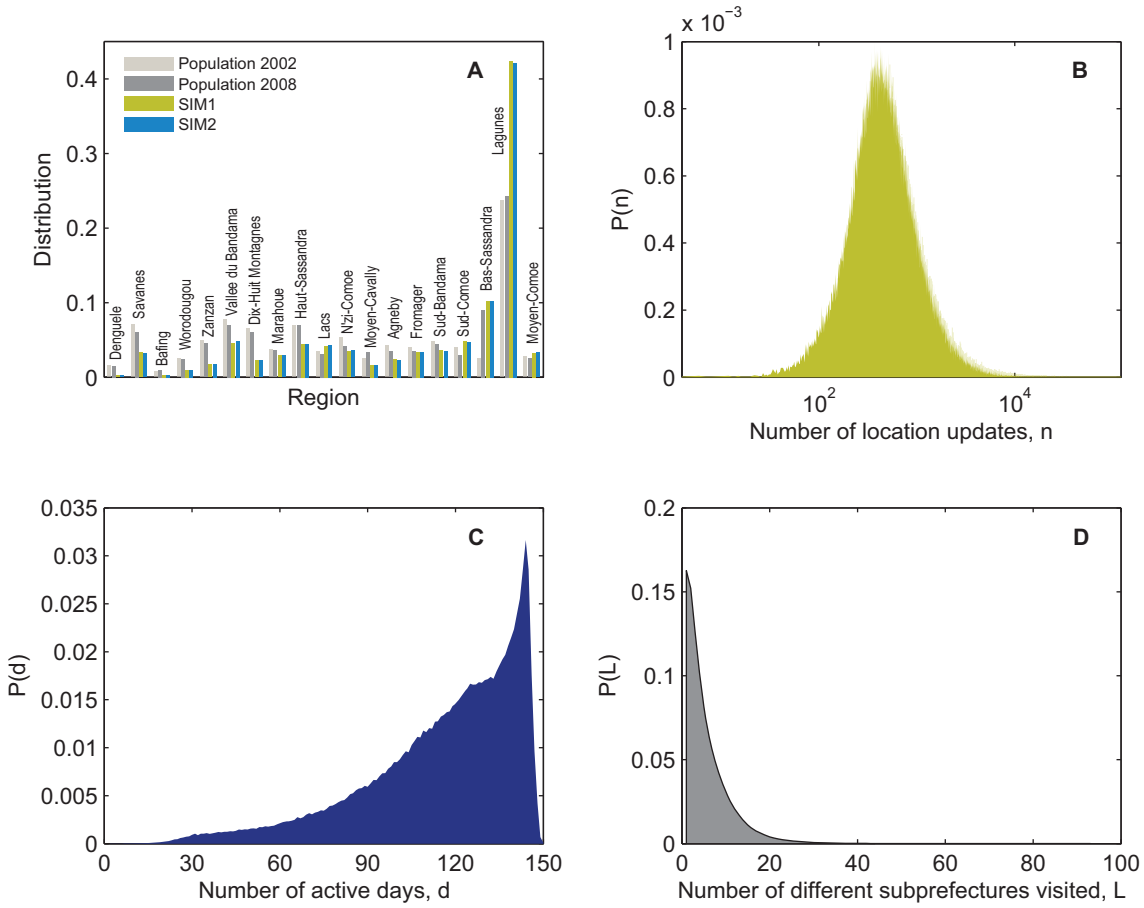

Figure 2 | Characteristics of the mobile phone users. (A) Proportion of users in each region compared to the population. Modelled population data based on 2002 official estimates were obtained from the AfriPop project[27], and 2008 estimates were made by UN OCHA and CNTIG[18]. SIM1: the number of users who made their first calls in this region; SIM2: the number of users who appeared for the majority of their time in this region. We use SIM1 and SIM2 to approximate the number of residential mobile phone users in each region. (B) Distribution of the number of observations for each user during the data collection period. Note that the x-axis is logged. (C) Number of active days on which each user made at least one call. (D) Distribution of the number of different subprefectures visited by each user.

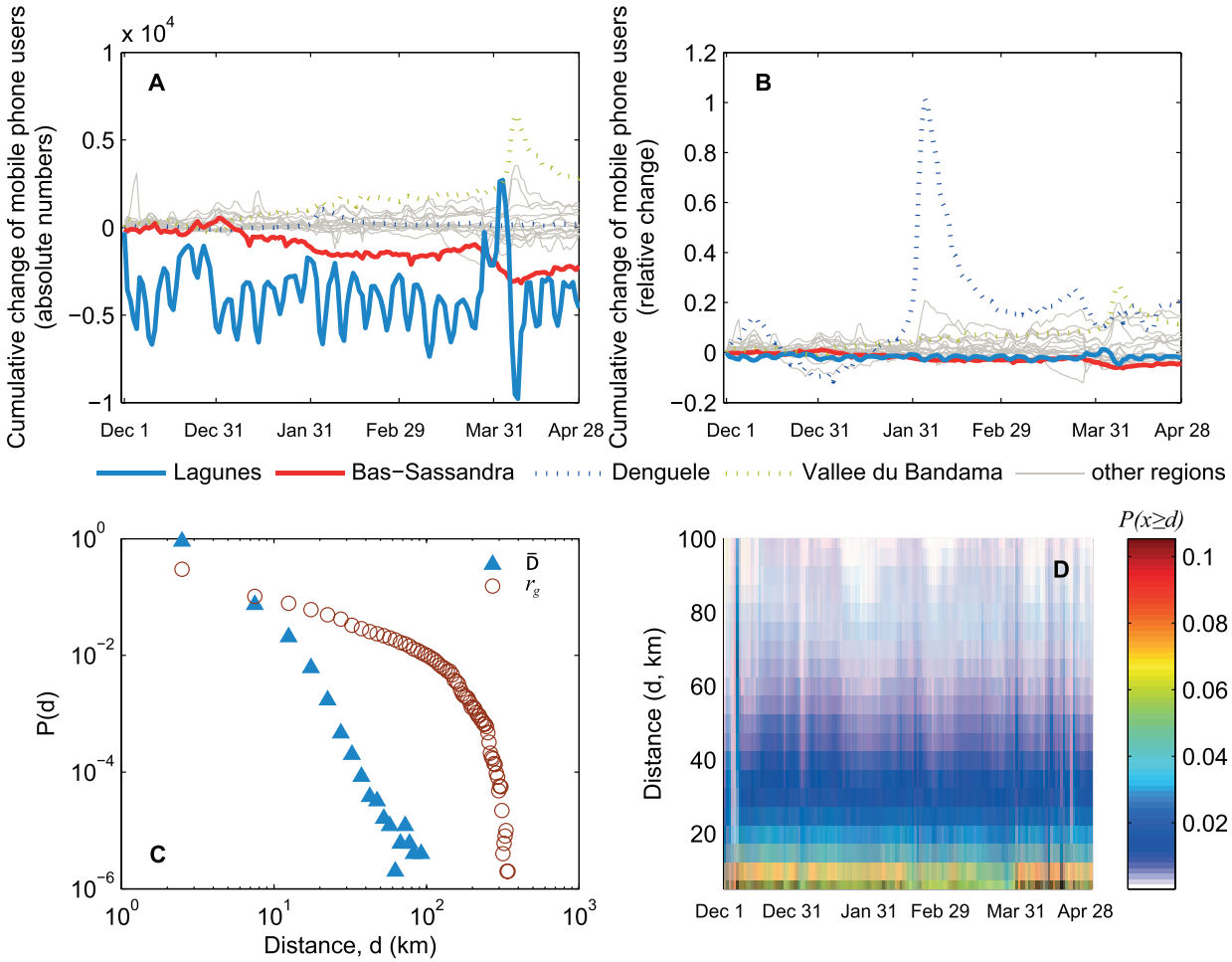

Figure 3 | Overview of population movements: (A) Cumulative change in number of users in each region. (B) Same data as in panel (A) but changes are shown in proportion to the number of users in each region at the beginning of the period. (C) Distribution of average travel distances $\bar{D}$ and the radius of gyration, r_g. (D) Cumulative probability distribution of average daily travel distance over the 150 study days.

we consider both the spatial and temporal correlation of the visit sequences. The median value of S^{rand} is 2.0, indicating that if we assume that individuals randomly choose a location to visit the next day, a typical individual could be found in any of $2^{2.0} \approx 4$ locations. On the other hand, if we use information contained in the frequency and sequence order of the trajectory of individuals, the uncertainty in a typical individual's whereabouts reduces to $2^{S^{unc}} = 2^{0.91} \approx 1.88$ and $2^{S^{real}} = 2^{0.71} \approx 1.64$, in less than two locations.

Not surprisingly, the reduced uncertainty leads to increased maximum predictability, as shown in Fig. 4B. If information is available only on the number of unique locations visited, L_i, the accuracy of any predictive algorithm cannot exceed 0.35. With the additional information on frequency and temporal correlation, the average predictability increases to $<\Pi^{unc}> \approx 0.84$, and $<\Pi^{max}> \approx 0.88$, respectively. Additionally, we evaluated the sensitivity of our entropy and predictability measures to the sampling rate of the data, without finding any important biases (see Supporting Information S1).

In Fig. 4C, we investigated the correlation between the radius of gyration and the average predictability, $<\Pi>$. There is a steady decrease of $<\Pi^{rand}>$ and $<\Pi^{unc}>$ when r_g increases (measured based on the centroid of each subprefecture). On the other hand, $<\Pi^{max}>$ stays around 0.85 for a wide range of $r_g \in [20, 300]$. This finding is consistent with previous studies, revealing the independence of predictability on travel distance in human mobility[1,17]. However, we have also examined other travel distance measurements. Increases in average travel distance ($\bar{D}$) cause a slight decrease in predictability. $<\Pi^{max}>$ ranging from 0.9 to 0.7 when $\bar{D}$ increases from 1 to 20 km, and stays around 0.63 0.68 for $\bar{D} \in [20,70]$. However, interestingly, predictability decreases considerably with an increasing number of unique locations visited. From Fig. 4E, we can see that the average predictability $<\Pi^{unc}>$ and $<\Pi^{max}>$ decays almost linearly with the number of unique locations visited.

Prediction accuracy based on Markov-chain models. The predictability analysis in the previous section reveals that, by combining information on frequency with temporal correlation of the trajectory, the theoretical upper bound of prediction accuracy can get as high as 0.88. However, the largest prediction accuracy that can be achieved with properly designed predictive algorithms is not given directly by this measure. In this section, we use MC(n) models to predict the location of users on each day, by considering all previous data points in the trajectory (see Methods). The accuracy of these models is presented in Fig. 5A and shows accuracies of more than 0.8 for almost all days. The accuracy of MC-based models ($<\gamma> \approx 0.91$), MC(1) toMC(7), produce substantially higher accuracies than the estimation method based only on frequency information, i.e., MC(0) ($<\gamma> \approx 0.85$).

There is however little difference between the performance of MC-based models of different orders. At the beginning of the period when historical information is limited, the accuracy of MC(1) is slightly higher than the other MC models, however this difference becomes very small when the historical trajectory is over 100 data points.

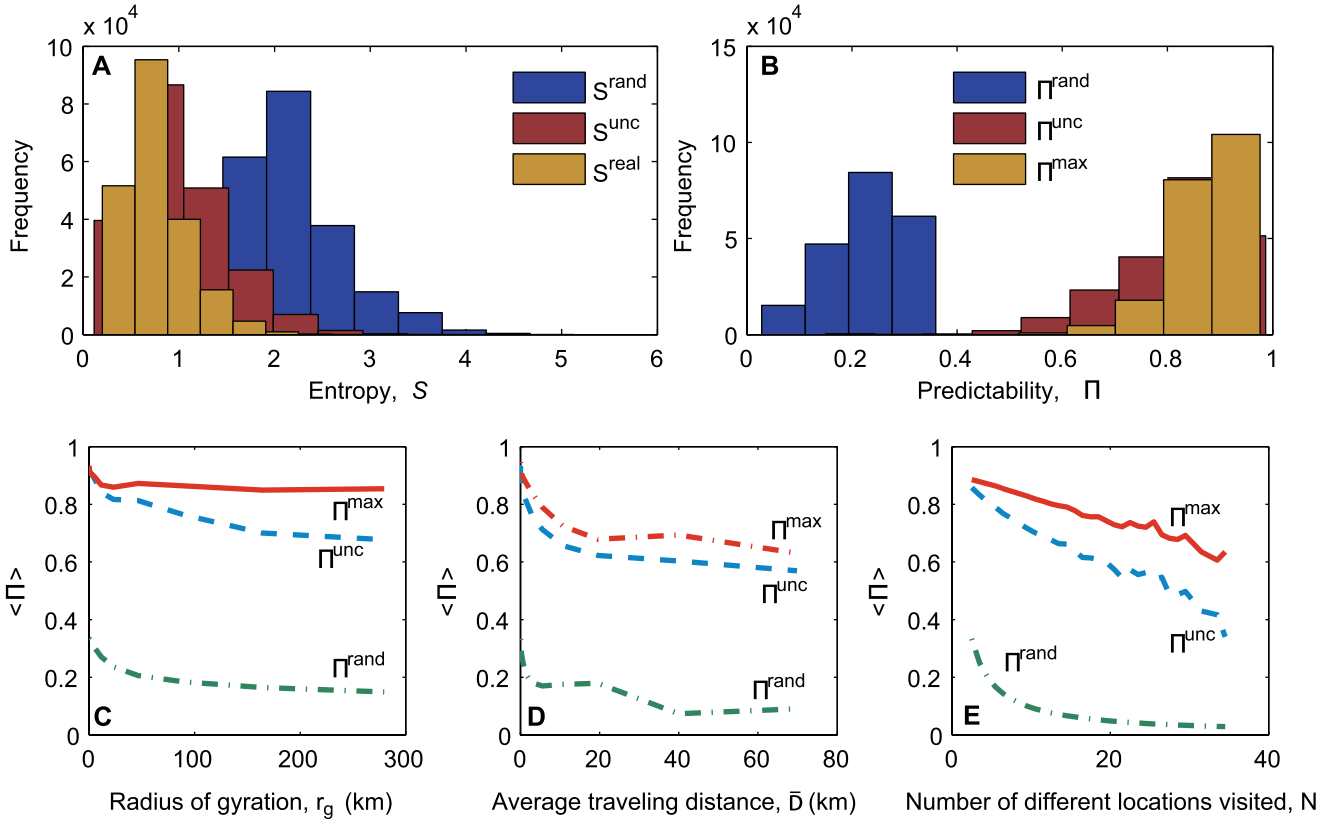

Figure 4 | Entropy and predictability analysis based on trajectory of visited subprefectures. (A) Shows the frequency distribution of S^{rand}, S^{unc} and S^{real}. (B) Shows the frequency distribution of Π^{rand}, Π^{unc} and Π^{max}. (C) Shows the correlation between radius of gyration and Π^{max}. (D) Shows the correlation between average travel distance $\bar{D}$ and Π. (E) Shows the correlation between the number of different locations visited and Π.

Another interesting finding from Fig. 5A, is that the MC-based models perform more robustly than MC(0). For example, during the later period of the data, there is a sharp decrease in the accuracy of MC(0) (from 0.88 to 0.77), while the accuracy of MC-based models shows a much smaller decrease, from 0.92 to 0.87. Irruptions of decreased accuracy from the MC(0) model indicate that people moved abnormally from their regular patterns. The sustainability of MC-based models reveals that such abnormalities can be captured partly by considering the temporal correlation of visiting sequences in the trajectories.

The increase of $<\gamma>$ over the observation period is not very apparent from Fig. 5A, as $<\gamma>$ is calculated based on a combination of users with long and short historical trajectories. To investigate the effect of trajectory length on the performance of algorithms, we removed, for each user, the unknown locations and calculated the average prediction accuracy for users with valid historical trajectories of the same length L_{hist}. The results (Fig. 5B) show that the accuracy of MC(0) approaches relative stability after around 15 historical data points. For a wide range of $L_{hsit} \in [15, 120]$, $<\gamma>$ is steady around 0.85, indicating that the visiting behavior on frequency is relatively stable over time for users with valid historical trajectories of this range. On the other hand, there is a steady increase of $<\gamma>$ for the MC-based models. When the available historical trajectories contain more than 100 data points, the average prediction accuracy climbs to over 0.9.

The performance of MC-based models indicates that, while the predictability of a typical user is $\Pi^{max} = 0.88$, which gives an upper

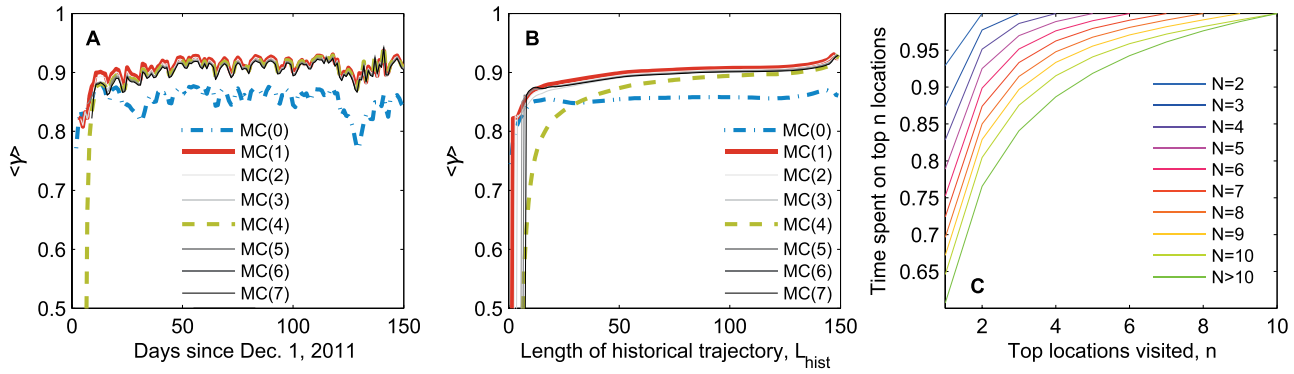

Figure 5 | Visiting behavior and prediction accuracy. (A) Proportion of accurate predictions for each day based on historical data (users who were not active on a day are excluded in the prediction). In (B), the accuracy of predictive algorithms increases with the length of historical trajectories. (C) Fraction of time users spent in their most n visited subprefectures. Subscribers are divided into 10 groups based on the number of distinct locations they visited (N).

bound for the accuracy of any predictive algorithm when the trajectory is stable, the MC-based models are able to produce estimates as high as 0.91, even higher than the theoretical upper limit. Possible reasons for why the practical algorithm can produce accuracies higher than Π^{max} could be that the trajectory data contains only one data point for each day, which means that the maximum length of the trajectory can only be 150 and the movement patterns of individuals may not yet have stabilized. To investigate the effect of stability on the performance of prediction algorithms, we use the Geweke diagnostic[21,22] to classify X_i into stationary and non-stationary trajectories (see Methods). In Fig. 6, we can see that there is a clear difference in the prediction accuracy of MC models between stationary and non-stationary trajectories: after 50 historical observations, the average prediction rate is about 0.95 for non-stationary trajectories and only 0.87 for stationary trajectories. This finding confirms that, given that the trajectory is stationary, the maximum predictability, Π^{max}, provides an upper bound of accuracy for any prediction algorithm. However, for non-stationary trajectories we show that prediction accuracy can greatly surpass the maximum predictability.

MC-models considering higher orders (longer correlations of previous locations) do not necessarily improve prediction accuracy. For example, for trajectories with the same historical length, the MC(4) model always produces less precise predictions compared to other MC-models (Fig. 5B). This finding is consistent with previous studies, in which the MC($n > 2$) model was found to not bring important improvement at the cost of a significant overhead in terms of computation and space for the learning and storing of the mobility model[9,12]. It is worth noting that a large part of the predictive power of the studied prediction algorithm is due to the fact that many individuals spent a substantial time in his/her top visited locations. For example, users who visited four distinct subprefectures, still spent almost 80% of their time in their most visited locations (Fig. 5C).

Entropy, predictability and prediction accuracy. The evaluation of predictive algorithms above reveal that, for this dataset, which is a combination of stationary and non-stationary trajectories, the maximum predictability Π^{max} can be achieved with a first-order Markov chain model (MC(1)). In this section, we investigate whether the individual predictability, Π_i^{max}, is correlated with the accuracy in predicting all the locations when the trajectory increases from 1 to T. We measure the individual prediction accuracy ($<\gamma_i>$) by the proportion of accurate predictions over all days for each individual (days without any location data are excluded).

First, we check the correlation between prediction accuracy and the disorder in the trajectory, i.e., S^{real}. As we can see from Fig. 7A, $<\gamma_i>$ is highly correlated with the trajectory's entropy; the larger the entropy, the lower the prediction accuracy. The correlation coefficient between S^{real} and $<\gamma_i>$ is as high as -0.849, with $p < 0.000$. Second, we investigate the correlation between prediction accuracy and the maximum predictability, i.e., Π_i^{max}. Not surprisingly, Π_i^{max} also correlates highly with $<\gamma_i>$, with a correlation coefficient of 0.802, $p < 0.000$ (Fig. 7B).

The high correlation between predictability and prediction accuracy of the MC(1) model reveals that, as a measurement for disorder and potential predictability, S^{real} and Π_i^{max} capture the theoretical limits for the predictive analysis of human movement behaviors,

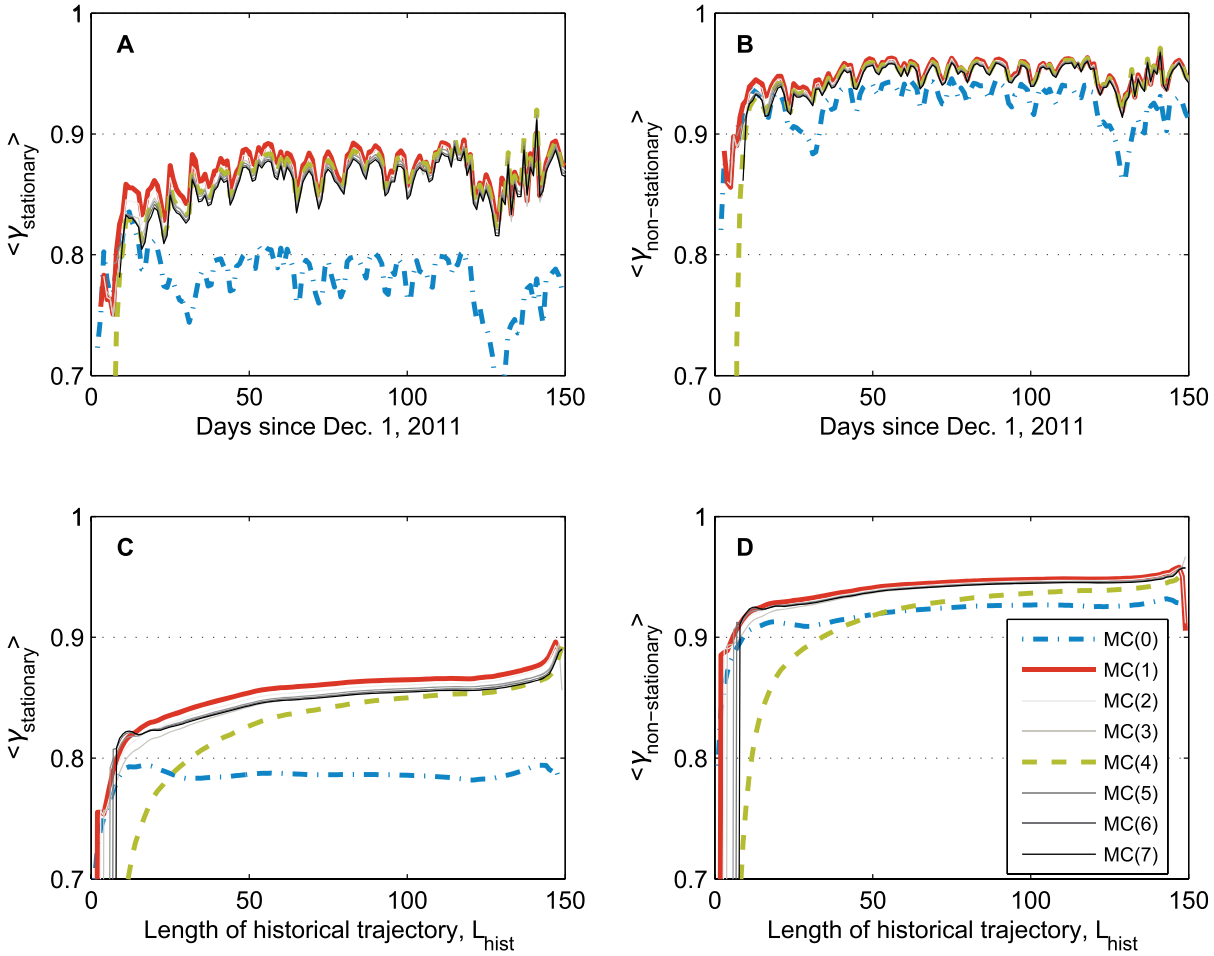

Figure 6 | Prediction accuracy on stationary (A, C) and non-stationary (B, D) trajectories.

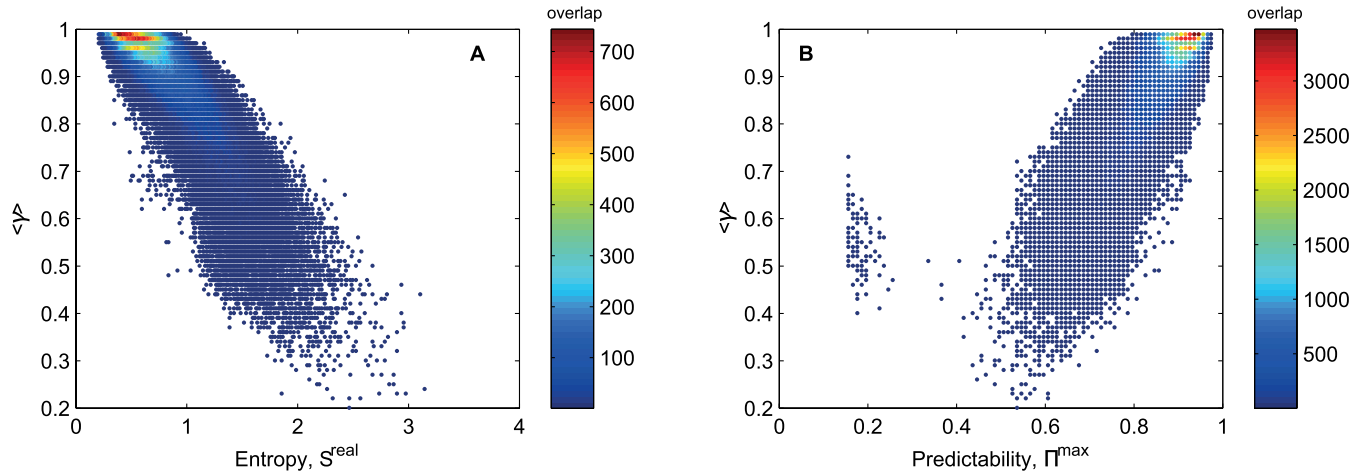

Figure 7 | Correlation between entropy, predictability and prediction accuracy. Data points are aggregated into intervals of equal lengths. (A) Correlation between entropy and prediction accuracy. (B) Correlation between predictability and prediction accuracy.

and provide an *approachable* upper bound of predictive power for this type of mobility data. More broadly, the approach used here provides an important strategy to evaluate and guide the design and improvement of mobility prediction algorithms.

Discussion

By investigating the movement of 500,000 mobile phone users during the post-civil war period in Cote d'Ivoire, we have found a potential predictability in user mobility as high as 88% in this West-African, lower-middle income setting. The finding of high predictability is consistent with two previous studies which investigated the mobility patterns of mobile phone users in very different settings, one in a high-income country with stable social conditions[17] and one in a low-income country following an extreme natural disaster[1]. By applying MC-based estimate algorithms, we found that the first order MC model (MC(1)) is able to produce an average predictive accuracy of 91%, with stationary and non-stationary trajectories having a predictive accuracy of 87%, and 95% respectively.

One would perhaps assume that Markov chains of second or seventh order would produce next-location estimates with higher accuracy, as aggregated flows based on mobile phone data frequently show weekly cycles, see e.g., Fig. 3A, and[1,2,17]. However, our evaluations on the MC(n) models show that this information is not necessary in this setting. This could be due to the fact that many people in Cote d'Ivoire do not have a two-day weekend, and that unplanned journeys are less common in resource-limited settings, which may be true of travel in general[23,24]. The trajectories used for prediction contain only the last observed location on each day, which makes it difficult for the time series to reach stability. As we can see, there is a big difference in the predictive power between stationary and non-stationary trajectories, implying that the diagnostics for convergence are critical for drawing conclusions from predictability analysis on human mobility. Nevertheless, we believe that the evaluation of predictive performance on a daily basis is most practical for the long-term investigation of population movements. For the purpose of this study, we have only included the Markov chain based models in the evaluation of predictive performance of algorithms. Future studies may want to compare other predictive algorithms, such as dynamic Bayesian networks (DBN)[25], neural networks[6] and finite automaton[8], and to evaluate the feasibility of predicting aggregated population movements with individual-based travel behavior models.

In summary, this paper is, to the best of our knowledge, the first attempt to investigate both the maximum predictability and how close to this value practical algorithms can come when applied on a large mobile phone data set. Our results not only show that the predictability of human mobility is high, but also show that this high predictability is achievable for daily population movement predictions. These findings indicate that human mobility behavior is far from random, and that individuals' movements are highly influenced by their historical behavior. With a good understanding of individuals' travel patterns, mobility modeling and public policy decision making, such as epidemic modeling, urban planning, and traffic design, may be significantly improved.

Methods

Measures of mobility. We use the average travel distance, $\bar{D}$, and the radius of gyration of the trajectory, r_g, to measure the mobility property of individuals. Specifically, let $M_i = \{m_1, m_2, \ldots, m_n\}$ be the sequence of observed location updates for person i during the period of data collection. Then $\bar{D}$ and r_g are defined by:

$$\bar{D}(i) = \sum_{j=2}^{n} |m_j - m_{j-1}| \quad (1)$$

and $r_g(i) = \sqrt{\frac{1}{n}\sum_{j=1}^{n} |m_j - \bar{m}|^2}$, where $|m_j - m_{j-1}|$ is the distance between location m_j and m_{j-1}, and $\bar{m} = \frac{1}{n}\sum_j m_j$ is the center of mass of the trajectory[26].

The radius of gyration is different from the average travel distance. Someone who moves in a comparatively confined space will have a small radius of gyration even though he or she covers a large distance. On the other hand, r_g can be larger than $\bar{D}$ if someone travels with small steps but in a fixed direction or in a large circle. Note that in the dataset used here we only know the location of each individual by subprefecture, consequently, the centroid of each subprefecture is used to approximate the coordinates of individuals. Such an approximation can introduce imprecision for the measure of travel distances, but still provides useful information when comparing mobility between users, as those who traveled across many subprefectures will have larger r_g and $\bar{D}$ than those who spent most of their time in one or two subprefectures.

Measures of entropy and predictability. We are primarily interested in the stable, long-term patterns of population mobility behavior as opposed to short-term movements. Here we focus on entropy and predictability analysis of day-to-day movements of individuals. Let $X_i = \{x_1, x_2, \ldots, x_T\}$ be the sequence of daily locations for person i during the data collection period of T days. x_j is the last observed location ID of person i on day j, otherwise we mark x_j "unknown". The uncertainty (or disorder) of the trajectories can be measured by entropy. Larger entropy indicates greater disorder, and consequently reduces the predictability of an individual's movements.

Entropy. Following notation in[17] we measure: (i) the random entropy, $S_i^{rand} = \log_2 L_i$, capturing the predictability of each user by assuming that the person's whereabouts are uniformly distributed among L_i distinct locations in X_i; (ii) the temporal-uncorrelated entropy, $S_i^{unc} = -\sum_{k=1}^{L_i} p_k \log_2 p_k$, where p_k is the frequency at which the person visited the k^{th} location among the L_i distinct locations. S_i^{unc} takes into account the number of different locations visited as well as the proportion of time i spent at each location, decreasing the uncertainty of the trajectory, and; (iii) the true-entropy, $S_i^{real} = -\sum_{X'_i \subset X_i} P(X'_i) \log_2[P(X'_i)]$, where $P(X'_i)$ is the probability of finding a subsequence X'_i in X_i, considering both spatial and temporal patterns.

Predictability. Given the entropy E for an individual i, Fano's inequality gives an upper limit for the predictability of i, i.e., the level of accuracy the best possible predictive algorithm can achieve:

$$\Pi_i \le \Pi_i^{\mathrm{Fano}}(E, L_i) \tag{2}$$

where Π_i^{Fano} is given by

$$E = H\left(\Pi_i^{\mathrm{Fano}}\right) + \left(1-\Pi_i^{\mathrm{Fano}}\right)\log_2(L_i - 1) \tag{3}$$

and

$$H\left(\Pi_i^{\mathrm{Fano}}\right) = -\Pi_i^{\mathrm{Fano}}\log_2\left(\Pi_i^{\mathrm{Fano}}\right) - \left(1-\Pi_i^{\mathrm{Fano}}\right)\log_2\left(1-\Pi_i^{\mathrm{Fano}}\right) \tag{4}$$

Let $\Pi_i^{\mathrm{rand}} = \Pi_i^{\mathrm{Fano}}\left(S_i^{\mathrm{rand}}, L_i\right)$, $\Pi_i^{\mathrm{unc}} = \Pi_i^{\mathrm{Fano}}\left(S_i^{\mathrm{unc}}, L_i\right)$ and $\Pi_i^{\mathrm{max}} = \Pi_i^{\mathrm{Fano}}\left(S_i^{\mathrm{real}}, L_i\right)$, since $S_i^{\mathrm{rand}} \ge S_i^{\mathrm{unc}} \ge S_i^{\mathrm{real}}$, it is true that $\Pi_i^{\mathrm{max}} \ge \Pi_i^{\mathrm{unc}} \ge \Pi_i^{\mathrm{rand}}$. Comparing between these three predictability measurements provides us with the ability to investigate how the spatial distribution and temporal correlations in an individual's trajectory improve potential predictive power. Since Π_i^{max} provides the best possible predictive power (because it uses the maximum information from S_i^{real}) we refer to it in this paper as the "maximum predictability".

Prediction algorithms. *Predicting a user's next location using Markov chain models.* To investigate how close we can get to achieving Π with actual prediction algorithms we implement several variants of Markov chain (MC) based models.

In an MC-based model, the trajectory of each individual is modeled as a Markov chain of order n, which assumes that the movement of individuals between the L_i locations is a process with limited memory in the sense that the future location is visited depending only on the previous n visited location, i.e., $P\left(X_i^{t+1} = x^{t+1} \middle| X_i^t = x^t, X_i^{t-1} = x^{t-1}, \ldots, X_i^1 = x^1.\right) = P\left(X_i^{t+1} = x^{t+1} \middle| X_i^t = x^t, X_i^{t-1} = x^{t-1}, \ldots, X_i^{t-n+1} = x^{t-n+1}.\right)$, where X_i^t is a random variable representing the location for individual i at time t.

Given the previous n locations $X_i^t = x_i^t, X_i^{t-1} = x_i^{t-1}, \ldots, X_i^{t-n+1} = x_i^{t-n+1}$, the prediction is then determined by the transition matrix, P, choosing the destination location $x^{pre} (1 \le pre \le L_i)$ which maximizes the probability:

$$P\left(X_i^{t+1} = x^{pre} \middle| X_i^t = x_i^t, X_i^{t-1} = x_i^{t-1}, \ldots, X_i^{t-n+1} = x_i^{t-n+1}\right)$$
$$= \max_{k=1}^{L_i}\left\{P\left(X_i^{t+1} = x^k \middle| X_i^t = x^t, X_i^{t-1} = x^{t-1}, \ldots, X_i^{t-n+1} = x^{t-n+1}\right)\right\}$$

Increases of the order n in the Markov chain do not necessarily lead to improvement in the prediction accuracy. However, to investigate the correlation of predictive powers with the length of trips to historical locations considered, we vary n from 1 to 7 (one day to one week). If predictions for a higher ordered MC(n) model did not exist (i.e., the order of the previous n locations is unique in history), the prediction from a lower ordered model, MC($n-1$), was used.

The performance of each model was evaluated by the accuracy, γ, which is the proportion of accurate predictions from all predictions made:

$$\gamma = \frac{\text{number of correct predictions}}{\text{total number of predictions}}. \tag{5}$$

Users who were not active on a specific day were excluded from the prediction.

Next place prediction using historical frequency data. For comparison we implemented a simple algorithm predicting the next location based on the most visited location in the historical trajectory: $P(x^{pre}) = \max_{k=1}^{L_i}\left\{p_k \middle| X_i^t = x_i^t, X_i^{t-1} = x_i^{t-1}, \ldots, X_i^1 = x_i^1\right)\}$, where p_k is defined the same as in S^{unc}. As no temporal correlation is considered in this algorithm, we refer it as MC(0).

Using the MC models, we repeatedly updated the transition matrices and the visiting frequency for each user when new locations were observed in the trajectory. We predicted for each user the most likely location s/he would visit on each day based on all the historical information, i.e., for each day t, the transition matrices and visiting frequency are constructed based on the trajectory from day 1 to day $t-1$.

Geweke diagnostic. The Geweke Diagnostic[21,22] is a test to detect failure of convergence by comparing values in the early part of a Markov chain to those in the latter part of the chain.

Let $X_i^1 = \left\{x_{i,t}^1 : t = 1, \ldots, n_1\right\}$ and $X_i^2 = \left\{x_{i,t}^2 : t = n_a, \ldots, n\right\}$, where $1 < n_1 < n_a < n$. let $n_2 = n - n_a + 1$ and define

$$\bar{\theta}_1 = \frac{1}{n_1}\sum_{t=1}^{n_1} x_{i,t}^1, \tag{6}$$

$$\bar{\theta}_2 = \frac{1}{n_2}\sum_{t=n_a}^{n_2} x_{i,t}^2. \tag{7}$$

Then the statistic

$$Z_n = \frac{\bar{\theta}_1 - \bar{\theta}_2}{\sqrt{\frac{\hat{s}_1(0)}{n_1} + \frac{\hat{s}_2(0)}{n_2}}} \tag{8}$$

converges to a standard normal distribution as $n_1 \to \infty$ given that the chain is stationary and $(n_1 + n_2)/n < 1$.

In the above equation, $\hat{s}_1$ and $\hat{s}_2$ denote consistent spectral density estimates at zero frequency[21,22] for X_i^1 and X_i^2, respectively.

Large Z_n-scores then indicate rejection of the null hypothesis and provide evidence that the chain is non-stationary. For the purpose of this study, we first converted the values of $x_{i,t}$ into unique integers monotonically increasing from 1, and used a significant level of $\alpha = 0.05$ and let $n_2/n = 50\%$. A trajectory is said to be stationary only if it passes all the tests at $n_1/n = 0.2$, $n_1/n = 0.3$, $n_1/n = 0.4$, and $n_1/n = 0.5$. By the end, the proportion of trajectories that passed the Geweke test (stationary trajectories) was 49%, see also supporting figure S2.

1. Lu, X., Bengtsson, L. & Holme, P. Predictability of population displacement after the 2010 haiti earthquake. *Proc Natl Acad Sci U S A* **109**, 11576–81 (2012).
2. Bengtsson, L., Lu, X., Thorson, A., Garfield, R. & von Schreeb, J. Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: A post-earthquake geospatial study in haiti. *Plos Medicine* **8**, (2011).
3. Kenett, D. Y. & Portugali, J. Population movement under extreme events. *Proceedings of the National Academy of Sciences* **109**, 11472–11473 (2012).
4. Ross, S. M. *Introduction to probability models* (Academic press, 2009).
5. Liu, G. & Maguire Jr, G. A class of mobile motion prediction algorithms for wireless mobile computing and communication. *Mobile Networks and Applications* **1**, 113–121 (1996).
6. Liou, S.-C. & Lu, H.-C. Applied neural network for location prediction and resources reservation scheme in wireless networks. In *International Conference on Communication Technology Proceedings, 2003. ICCT 2003*, vol. 2, 958–961 (IEEE, 2003).
7. Akoush, S. & Sameh, A. Mobile user movement prediction using bayesian learning for neural networks. In *Proceedings of the 2007 international conference on Wireless communications and mobile computing* 191–196 (ACM, 2007).
8. Petzold, J., Bagci, F., Trumler, W. & Ungerer, T. Global and local context prediction (2003).
9. Song, L., Kotz, D., Jain, R. & He, X. Evaluating next-cell predictors with extensive wi-fi mobility data. *Mobile Computing, IEEE Transactions on* **5**, 1633–1649 (2006).
10. Killijian, M.-O., Roy, M. & Trédan, G. Beyond san francisco cabs: Building a*-lity mining dataset. In *Proceedings of the Workshop on the Analysis of Mobile Phone Networks,* 75–78 (2010).
11. Zheng, Y., Li, Q., Chen, Y., Xie, X. & Ma, W.-Y. Understanding mobility based on gps data. In *Proceedings of the 10th international conference on Ubiquitous computing,* 312–321 (ACM, 2008).
12. Gambs, S., Killijian, M.-O. & del Prado Cortez, M. N. Next place prediction using mobility markov chains. In *Proceedings of the First Workshop on Measurement, Privacy, and Mobility,* 3 (ACM, 2012).
13. Asahara, A., Maruyama, K., Sato, A. & Seto, K. Pedestrian-movement prediction based on mixed markov-chain model (2011).
14. Kontoyiannis, I., Algoet, P. H., Suhov, Y. M. & Wyner, A. J. Nonparametric entropy estimation for stationary processes and random fields, with applications to english text. *Information Theory, IEEE Transactions on* **44**, 1319–1327 (1998).
15. Fano, R. M. Transmission of information: A statistical theory of communications. *American Journal of Physics* **29**, 793–794 (1961).
16. Brabazon, A. & O'Neill, M. *Natural computing in computational finance*, vol. 1, (Springer, 2008).
17. Song, C. M., Qu, Z. H., Blumm, N. & Barabasi, A. L. Limits of predictability in human mobility. *Science* **327**, 1018–1021 (2010).
18. UN OCHA Côte d'Ivoire and Le Comité National de Télédétection et d'Information Gégraphique (CNTIG). Common and fundamental operational datasets registry (2011). http://cod.humanitarianresponse.info/fr/country-region/c%C3%B4te-divoire [accessed Febuary 13, 2013].
19. Bank, T. W. Population total in cote d'ivoire (2011). http://data.worldbank.org/country/cote-divoire [accessed Febuary 13, 2013].
20. Blondel, V. D. *et al.* Data for development: the d4d challenge on mobile phone data. *arXiv:1210.0137v2* (2013).
21. Cowles, M. K. & Carlin, B. P. Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association* **91**, 883–904 (1996).
22. Geweke, J. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *IN BAYESIAN STATISTICS* 169–193 (Oxford: Oxford University Press, 1991).
23. Rubio, A., Frias-Martinez, V., Frias-Martinez, E. & Oliver, N. Human mobility in advanced and developing economies: A comparative analysis. In *AAAI Spring Symposia Artificial Intelligence for Development, AI-D, Stanford, USA* (2010).
24. Wesolowski, A., Eagle, N., Noor, A. M., Snow, R. W. & Buckee, C. O. The impact of biases in mobile phone ownership on estimates of human mobility. *Journal of the Royal Society Interface* **10**, (2013).
25. Eagle, N., Clauset, A. & Quinn, J. A. Location segmentation, inference and prediction for anticipatory computing. In *AAAI Spring Symposium: Technosocial Predictive Analytics*, 20–25 (2009).
26. Gonzalez, M. C., Hidalgo, C. A. & Barabasi, A. L. Understanding individual human mobility patterns. *Nature* **453**, 779–782 (2008).

27. Tatem, A. & Linard, C. High spatial resolution data on persons per grid square in cote d'ivoire (2010). http://www.afripop.org [accessed October 3, 2013].

Acknowledgements

The authors would like to thank the operator France Telecom-Orange and the "Data for Development" committee for sharing the mobile phone dataset and organizing the D4D challenge. This paper forms part of the output of the Human Mobility Mapping Project (www.thummp.org) and the AfriPop population mapping project (www.afripop.org). AJT acknowledges funding support from the RAPIDD program of the Science and Technology Directorate, Department of Homeland Security, and the Fogarty International Center, National Institutes of Health, and is also supported by grants from NIH/NIAID (U19AI089674) and the Bill and Melinda Gates Foundation (#49446 and #1032350). NB acknowledges funding from the Branco Weiss - Society in Science.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Lu, X., Wetter, E., Bharti, N., Tatem, A.J. & Bengtsson, L. Approaching the Limit of Predictability in Human Mobility. *Sci. Rep.* **3**, 2923; DOI:10.1038/srep02923 (2013).

Supporting Information:

Approaching the Limit of Predictability in Human Mobility

Xin Lu[1,2,3,4,*], Erik Wetter[2,5], Nita Bharti[6], Andy Tatem[7,8], and Linus Bengtsson[2,3]

1, College of Information System and Management, National University of Defense Technology, 410073 Changsha, China

2, Flowminder Foundation, 17177 Stockholm, Sweden

3, Department of Public Health Sciences, Karolinska Institutet, 17177 Stockholm, Sweden

4, Department of Sociology, Stockholm University, 17177 Stockholm, Sweden

5, Department of Management and Organization, Stockholm School of Economics, 11383 Stockholm, Sweden

6, Department of Biology, Center for Infectious Disease Dynamics and Huck Institute of Life Sciences, Penn State University, University Park, PA 16801, USA

7, Department of Geography and Environment, University of Southampton, Highfield, Southampton, United Kingdom

8, Fogarty International Center, National Institutes of Health, Bethesda, MD 20892, USA

S1. Distribution of Entropy and Predictability with Increased Sampling Rates

To examine the effect of sampling rates on the distributions of entropy and predictability, we have increased the sampling rates of trajectory from once per day to two times (S_{12} and Π_{12}) and three times (S_8 and Π_8) per day. Specifically we construct a trajectory for each individual by his/her observed location every 12 hours (12:00, 0:00) and every 8 hours (8:00, 16:00, 0:00), respectively. Distribution of entropy and predictability for trajectories with more than 80% known locations are shown in **Fig. *S1***.

We can see that, despite the different sampling rate, the distributions of S and Π are very similar to those shown in the main text, indicating that the sampling rate has little effect on our conclusion: that considering both the spatial and temporal correlation can reduce the uncertainties and improve predictability significantly.

* Email: xin.lu@flowminder.org

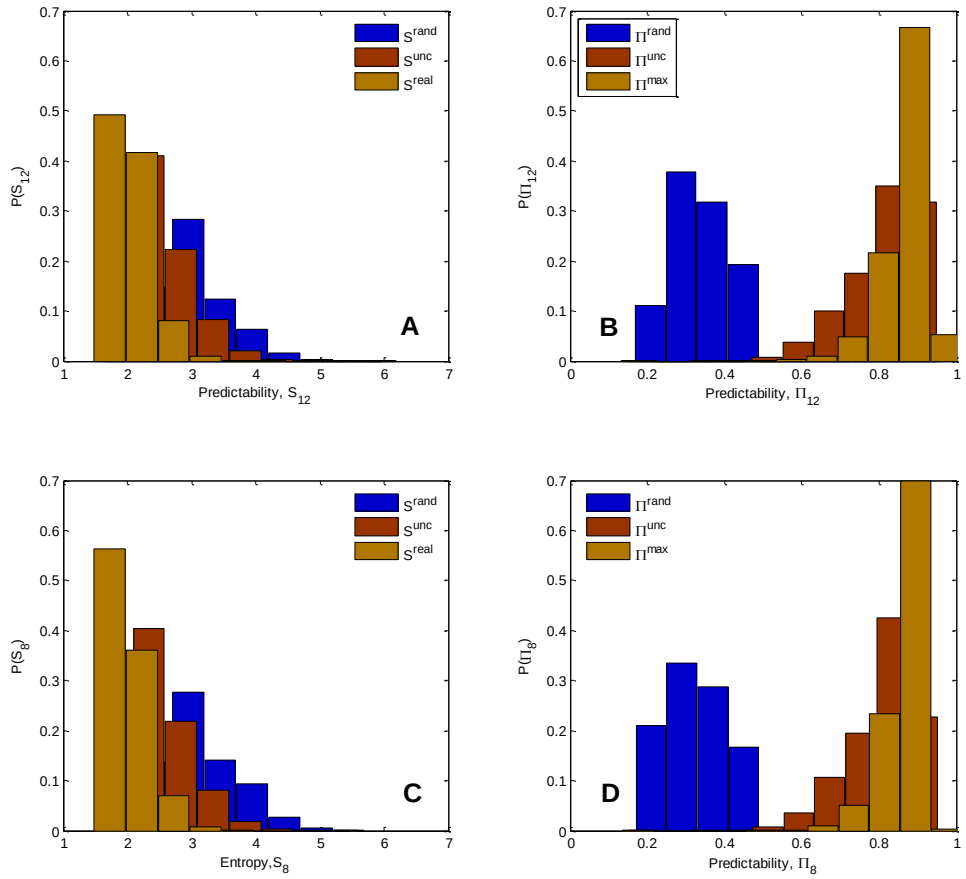

Fig. S1. Distribution of entropy and predictability for trajectories sampled two times a day (*A, B*) and three times a day (*C, D*). In comparison to the daily trajectory from the main text, which has the median entropy (predictability) of 2.0 (0.25), 0.90 (0.84) and 0.71 (0.89), the median entropy (predictability) with increased sampling rate is 2.32 (0.20), 0.97 (0.84), 0.69 (0.90) for trajectories sampled twice a day, and is 2.59 (0.17), 1.04 (0.83), 0.70 (0.90), respectively.

S2. Geweke Diagnostic

Based on the Geweke Diagnostic method presented in the main text, the proportion of trajectories that passed the convergence test is presented in **Fig. *S2***. We can see that when the length of tested chains $X_i^1 = \{x_{i,t}^1 : t = 1, \ldots, n_1\}$ varies from 20% to 50% of the length of the trajectory, the percentage of trajectories that passed the convergence test decreases from 65% to 56%. By the end, the percentage of trajectories passed all the tests (i.e., for n1/n={0.2, 0.3, 0.4, 0.5}) is 49%.

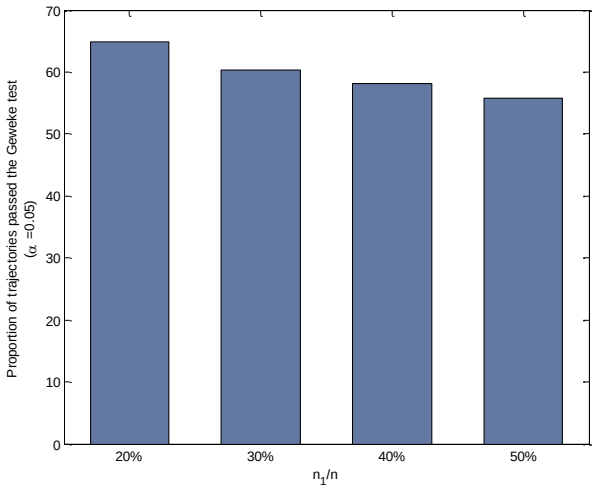

Fig. S2. Proportion of individual trajectories that passed the Geweke test.