# Limits of Predictability in Human Mobility

2016年06月27日-07月2日 一周研究报告

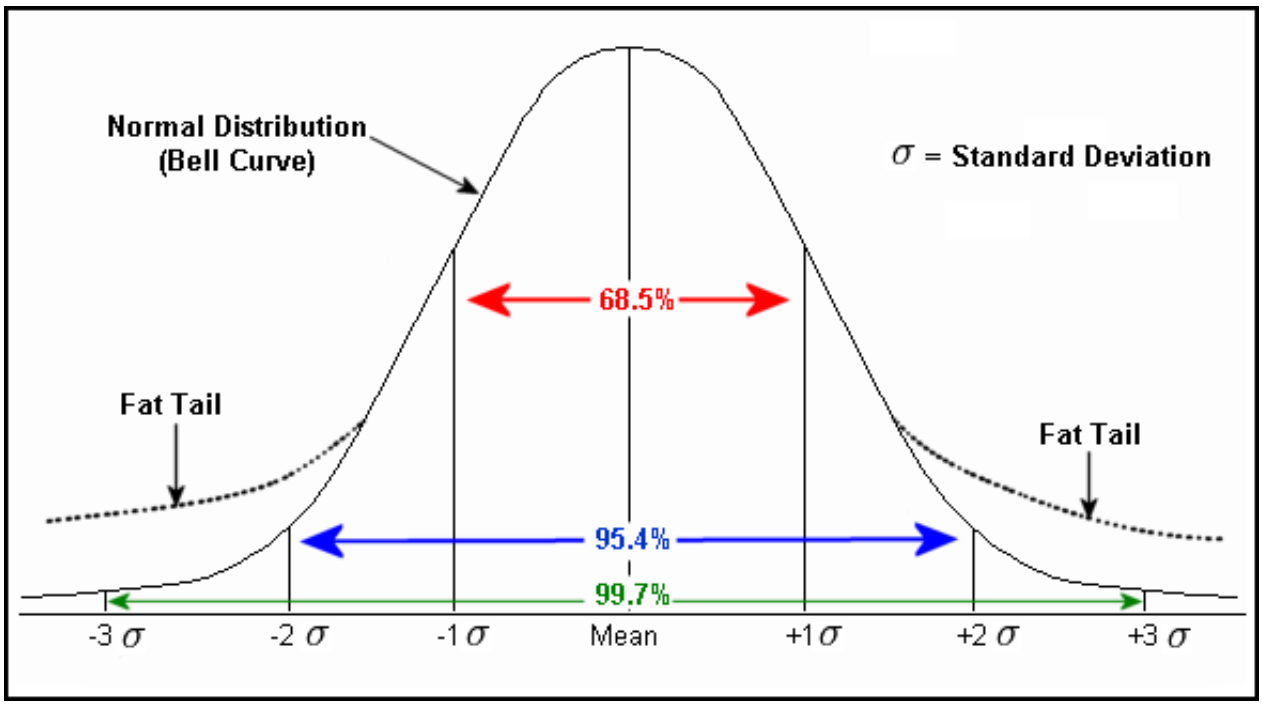
## 1. 概述

本周的主要工作就是对人类移动性的行为模式进行研究与学习，并思考如何将其应用于群智感知中去。

## 2. 术语定义

回转半径(radius of gyration ): 在文中意思为以某基站为圆心，个体用户的出行半径；

重尾分布（fat-tailed distribution）: 一种概率分布模型，在许多情况下，右边尾部的部分比较受到重视，但左边尾部比较厚，或是两边尾部都比较厚的状况，也被认为是一种重尾分布。




## 3. 文献介绍

### 3.1 引言与数据集介绍

文中同样现对目前学者专家们对人类移动模式的研究进展进行介绍，以及前人做的工作，研究的程度、常用的研究方法等等。

本文作者研究所使用数据集都是匿名用户的移动电话的数据，一部分是记录移动手机每次进行呼叫时记录下的接入基站等相关位置信息，时长跨度较大（**14**周），用户数量大（**50000**用户）；另一部分是用户主动参与该实验，每个小时不断的上报自己的位置相关信息，时间、附近的基站信息，时长为八天。

### 3.2 研究过程

首先，在数据集中选出两个典型用户，将他们的移动模式做了抽样，简单刻画其行为模式。

第一个用户中记录的基站数量为**22**，表示其大约在**22**个基站附近活动，其活动半径大约为***30km***，第二个用户中记录的基站数量为**76**，同理其大约在**76**个基站附近进行活动，活动半径大约为**90km**。将这些基站在地图中一一进行标注，并将其连接起来，如下图**1**所示。

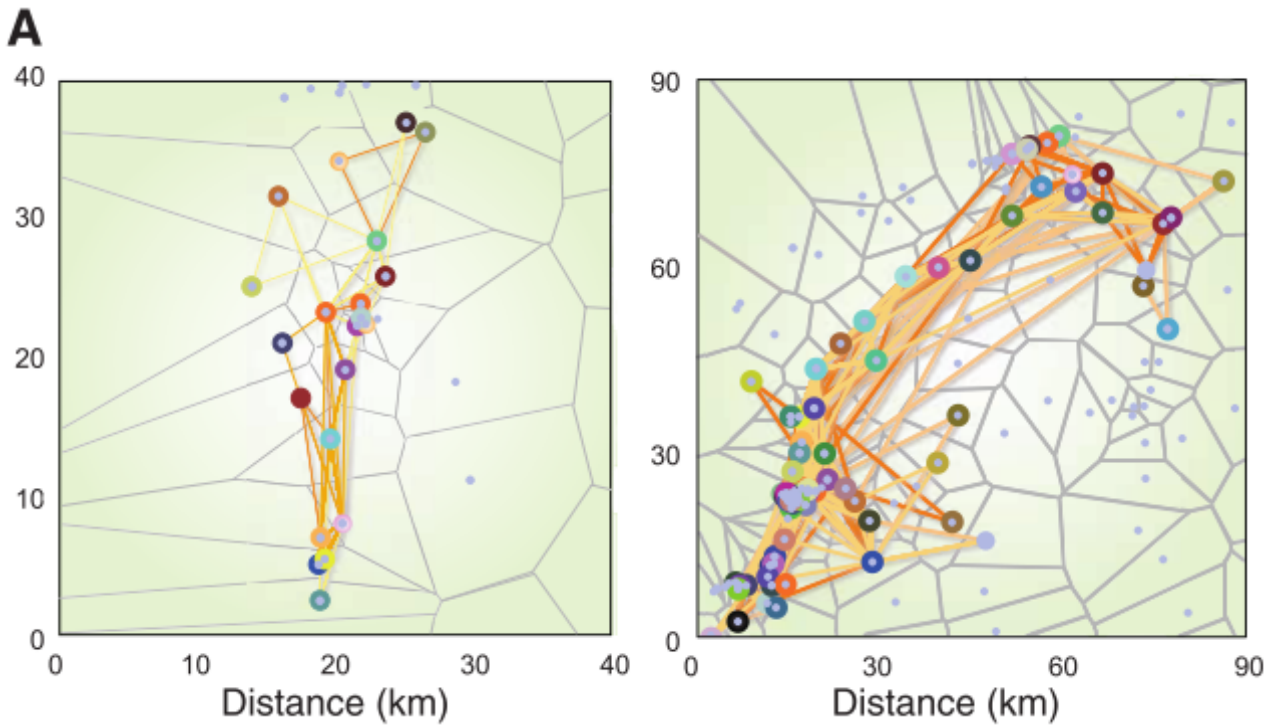



图 **1**　两个匿名用户出行距离图

然后分别将两个用户在每个地点停留的时间长短进行统计，可得下图。图**2**中点形状的大小代表了用户在这个地点停留的时长占总时长的比例。

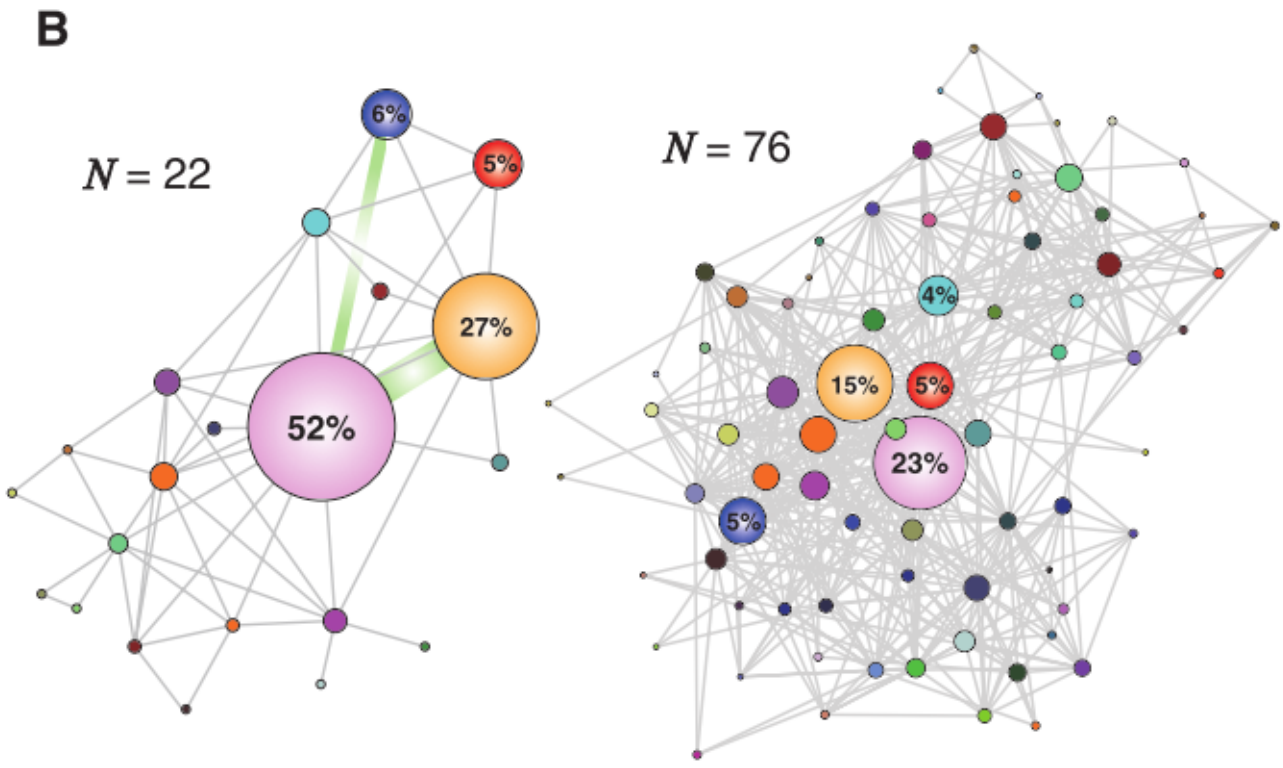



图 **2**两个匿名用户的出行网点图

经过简单的行为模式刻画之后，分别定义了三个参数-熵值，进行定量预测。

- 随机熵(*Random Entropy*))：

$$S_i^{rand} \equiv \log_2 N_i$$

其中*Ni*是用户i访问过的不同地方的数量，随机熵值表示了如果每个地方以相同概率被访问的话，用户下一个地点的可预测程度。

- 时空不相关熵(*Temporal-uncorrelated Entropy*):

$$S_i^{unc} \equiv -\sum_{j=1}^{N_i} p_i(j) log_2 p_i(j)$$

其中*pi(j)*是用户*i*以往访问位置*j*的概率，该熵表示了用户访问的每个地方之间是存在不同的。

- 真实熵(*Actual Entropy*):

$$S \equiv -\sum_{T_i' \subset T_i} P(T_i') \log_2 \left[ P(T_i') \right]$$

该熵值得的大小不仅仅取决于用户*i*访问某个地点的频率，而且取决有用户方位该地点的时间顺序以及用户在该地点停留的是将的长短，具有时空相关性，准确的刻画了用户一个用户的移动模式的时空相关性。具体来说，假设

$$T_i = \{X_1, X_2, X_3, \cdots, X_L\}$$

表示用户i以小时为间隔单位所被观察到的出行规律，那么真实熵就是上面的***Si***计算方法。

为了计算出真实熵，需要一个连续时间间隔内的用户位置信息，由于数据集中用户位置信息是每次进行手机通话的时候进行记录的，然后用户的通话行为呈现出一定突发性，短时间内多次通话，然后很长一段时间都是沉默的（如下图3中D所示），所以数据集中的数据具有很大的不完整性。这种数据的不完整性我们用***q***来描述，代表着我们每个小时内不知道用户信息的比例，并得到了数据集中***q***的概率密度分布图，如图3中*E*所示.

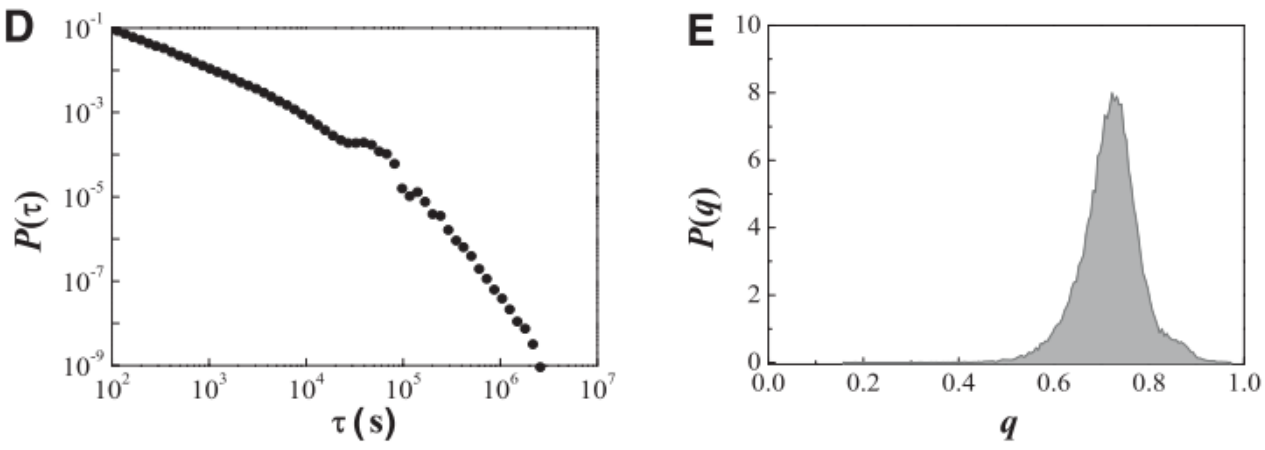



图 **3 D**通话概率与时间间隔该路密度分布图，**E**数据不完整性**q**分布图

上图中***q***的峰值出现在0.7左右，通过实验我们得到当***q***的值大于0.8时不能提供一定量的位置信息，所以我们保留了***q***值小于*0.8*的部分进行熵值计算。

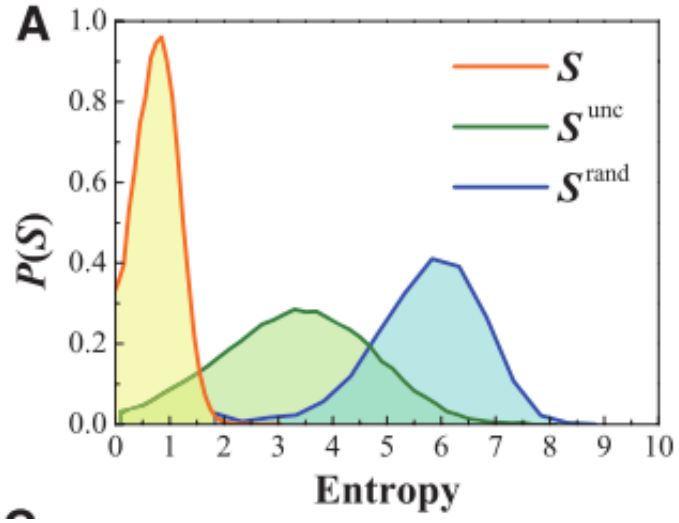



图 **4**三种熵值概率密度分布

计算完三个熵值，分别得到其分布图。图中比较有价值的分别是 和 ，前者的峰值出现在6左右，后者的峰值在*0.8*左右，通过熵值的逆运算我们可以得到随机熵值中位置数为64，表示如果使用随机熵来预测用户下一刻出现的位置，有*64*种可能。用真实熵0.8逆运算约为*1.7*，少于两个地方。

在数据集中每个个体日常活动的距离（回转半径 ）大多都在1-10km，很少用户超过10km在几百km，符合重尾分布。因此，作者假设用户行为的可预测都也应该符合重尾分布。

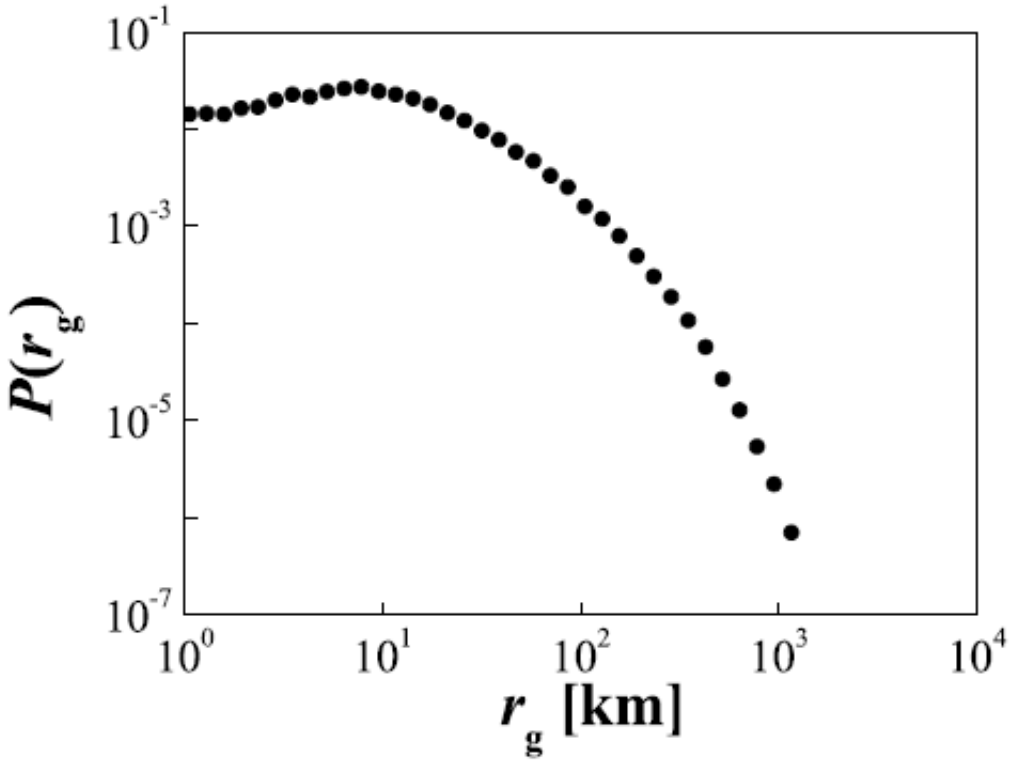



图 **5**回转半径的重尾分布图

用来说明可预测都大小的为可预测概率，并且提出了预测算法（此处不介绍）。可预测概率可以解释为，例，表示一个人至少***80%***的时间的行为模式是随机的，只有剩余***20%***的时间的行为行踪是可以进行预测。同时也可以理解为，即便我们的预测算法是***100%***正确的我们也只有20%的可能预测出他的位置，所以代表了每个用户移动可预测的极限。

有了预测算法，对数据集中的每个用户进行了可预测度计算，并将这些数据做成分布图，但是*p*的峰值出现在了***0.93***的位置，并不符合前面假设符合重尾分布，如下图6中B所示。这种分布规律表明，虽然我们每个人的行为模式看似都是随机的，但是每个人的里使用移动轨迹中却隐藏着高度可预测性。

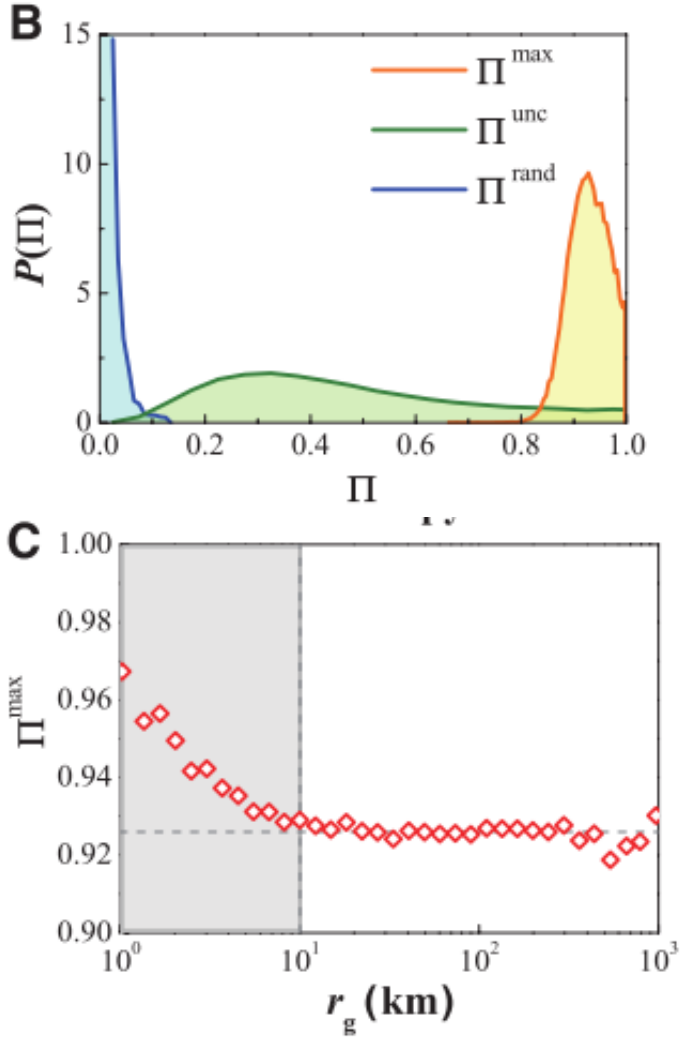



图 **6 B**可预测性分布图，**C**可预测概率与回转半径函数分布图

上面提出的假设是，用户的回转半径符合重尾分布，可预测性也应该符合重尾分布，但是假设用户的可预测性并不符合重尾分布，假设失败上图所示，当用户的回转半径大于***10km***，甚至到了***100km***或者几百km其可预测性的最大值与回转半径为10km的可预测性的值几乎一样。那么失败的原因是什么?

因为在回转半径比较小的时候，随着观察用户的回转半径的增大，访问的数量会增加，所以可预测性会降低（上图6中图C中回转半径***1-10km***时候），但是当观察回转半径达到一定量时候，用户访问的地点数量不再增加，那么可预测性概率的大小保持稳定（上图6中图C中回转半径10-1000km时候）

上面介绍了 为个体行为可预测性的上限，下面介绍了个体行为可预测性的下限。

为了理解观察到的个体行为的高度可预测性，将一周的时间分成了168小时，每个小时我们都对用户最常去的地方做记录。比如，在星期一，一个用户在8-9点的时候记录了十次，其中出现在了A位置1次，B位置2次，C位置7次，说明在这个小时内最可能出现的位置为C，设概率值为R，意义为在某个时间段发现一个人出现最有可能的那个地方的概率。经过对数据集中个体数据的统计，发现R的值大约为0.7，这意味着平均70%的时候用户方位的位置与实际位置相同。这个值得大小依赖于具体的某个时间段，比如在半夜的时候R的峰值为0.9，在中午到下午6/7点的时候，R具有最小值。如图7中A所示。

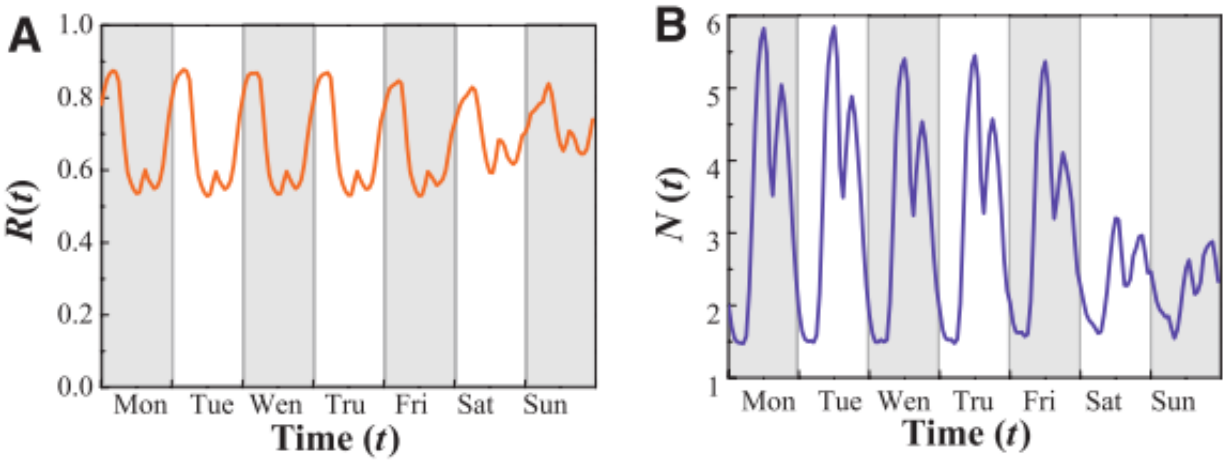



图 **7 A**图为**R**随着时间的分布图，**B**图访问地点数量随着时间变化的分布图

上面R值的规律也可以通过图7中B图得到验证，当A图中R出现最小值的时候，B图响应时刻正好对应的访问地点的数量呈现出最大值，说明当访问地点数量所得时候，可预测的概率小；当A图中R呈现峰值的时候，对应B图中响应时刻访问的地点数量是最少的时候，此时可预测的概率最大，这样在A、B图中使得R为可预测得到了相互的验证。

加入用户在访问的所有的地点中随即移动话，根据上面随机熵值可以到 为0.016，远远小于0.7，所以再一次证明了用户的移动性与随机移动的方式相去甚远。除此之外，文中还将相对预测规律 与回转 半径做了描述，如下图8所示。

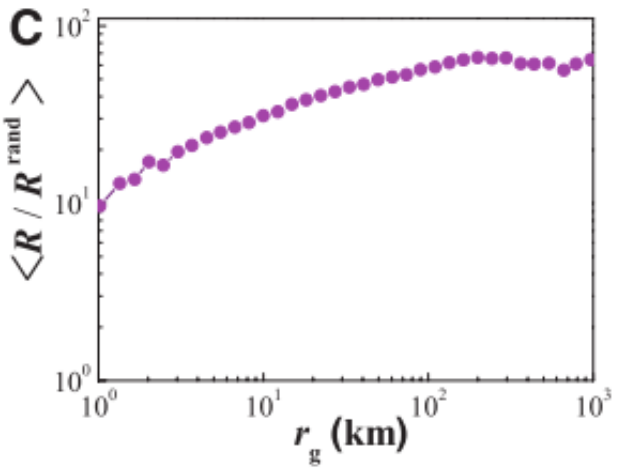



图 **8**相对规律 与回转半径变化图

为了测试上述实验的适用范围，以及其他因素是否会影响个体行为的可预测性。文中在不同地区，不同年龄段的人群做了相同的测试，包括使用不同语言的地区，人口密度不同的地区，贫富程度不同的地区，均没有发现非常大的区别。

### **3.3** 结论

综上所述，我们通过实验所得，个体的行为可预测都高达93%，但是更见成果是，这一结果并不会因为人群的不同而出现，较大偏差，即适用范围很广。除此之外，根据不同地区人群，以及描述的参数不同，个体的行为也不尽相同，经常出行且距离远的用户行为的可预测性要比不经常出行的用户的可预测性要低。

## **4.** 参考文献

[1] *Song C, Koren T, Wang P, et al. Modelling the scaling properties of human mobility[J]. Nature Physics*, 2010, 6(10):818-823.

[2] 陆锋, 刘康, 陈洁. 大数据时代的人类移动性研究[J]. 地球信息科学学报, 2014, 16(5):665-672.

[3] 刘瑜, 肖昱, 高松,等. 基于位置感知设备的人类移动研究综述[J]. 地理与地理信息科学, 2011, 27(4):8-13.