基于移动通信数据的居民居住地识别方法研究

宋少飞，李玮峰，杨东援

（同济大学 道路与交通工程教育部重点实验室，上海 201804）

【摘要】从大量手机数据中提取居民活动的时空特征是大数据趋势下的新兴热点。然而现有的研究主要使用的是手机通话数据或者手机定位数据，并且许多研究缺乏相关验证，使得研究可信度不高。本文基于手机信令数据，通过 3 种方法对居民职住地进行识别分析，并对比不同方法之间的差异，研究探讨通过手机大数据对于居民职住地识别的可信性。

【关键词】手机信令数据 ；时间阈值法 ；信息熵 ；时间相对值 ；居住地识别

【中图分类号】U12，F62　　【文献标识码】A　　【文章编号】1000-713X（2015）12-0072-05

Research on the Methods of Home Identification Based on Mobile Phone Data

SONG Shaofei, LI Weifeng, YANG Dongyuan

(Key Laboratory of Road and Traffic Engineering, Ministry of Education, Tongji University, Shanghai 201804, China)

【**Abstract**】Extracting the temporal-spatial characteristics from mobile phone data is a hot topic under the tendency of big data. However, the data used by recent research is mainly the Call Detailed Record data(CDR), and many studies lack the validation which decreases the confidence level. This paper used the Cellular Signaling Data, discussed three methods to identify the individuals' home. Then we compared the differences between these methods.

【**Keywords**】cellular signaling data; time threshold method; information entropy; relative time; home identification

通勤交通与职住区位分布一直是城市规划与交通规划研究的重点，在以往的研究中，获取居民居住地的主要途径是问卷，但大规模的问卷调查需要耗费巨大的人力财力，且调查周期长、数据更新不及时，调查数量一般在几千人，较之人口在几百万到千万的大中型城市，问卷调查样本可能不具有代表性。相比之下，手机数据拥有覆盖面积广、采样及时、更新及时的优势，并且能够长时间跟踪研究居民长期活动的时空特征。

从手机数据中获取居民活动信息，已经成为城市规划、交通规划等相关领域的热点。但现有成果多集中于手机通话数据 [1]（打电话或发短信），而此类数

收稿日期 ：2015-10-21　　录用日期 ：2015-12-03

作者简介 ：宋少飞（1989-），男，硕士研究生，主要研究方向为交通运输规划与管理。

通讯作者 ：ssf0307@126.com

据往往具有很大的随机性，对于居民活动的时空特征的描述准确性很低。本文使用的手机信令数据，是一种包含信息更丰富的手机数据，能够更好地反应用户的活动特征。此外，一些研究将手机数据与其它数据相结合来提取居民活动特征，比如与土地使用类型[2]、城市主要路口视频数据[3]，但这无疑很大程度的增大了数据处理分析的难度，实际可行性不高。

1 研究数据与预处理

1.1 原始数据概况

本研究使用的是上海市 2011 年 9 月的手机信令数据，通过采样分析的方法，从上海市范围内随机提取了 1496 人作为研究对象（共 6441389 条）。与国内外的其它手机数据相比，本文使用的数据有着明显的优势，信令事件类型比较全面，可以有效提高用户位置识别的精确度。原始数据（见表 1）包含 CellularID（用户唯一识别号，经过单项加密）、DateTime（时间戳，信令发生的时间）、LAC（位置区编号）、CI（基站小区编号），LAC 与 CI 唯一标识基站小区。根据 LAC 和 CI，我们可以识别出该条数据的位置（即经纬度坐标），进而，根据时间的连续变化，我们能够还原出用户的出行链信息。

表 1 原始数据示例

CelluarID	DateTime	LAC	CI
912805633C598B3F2C9D03D8A4C4D93F	20110901000302	6150	12914
6478865FA8C6BAABAFADB8A5B31865AD	20110901000253	6204	12626
51E38D9D77BA1979C260994F9CF701F2	20110901000253	6324	62290

本文随机抽取了上海市内 1496 个用户作为研究样本。因为不同人群使用手机的频率不同，此外，由于数据为 2011 年，一部分人群（如老年人）对手机的使用率及依赖程度较低，导致不同人群的信令数据数量密度差异度较高，反应在数据上，便是在所研究的 2011 年 9 月共 30 天中，不同用户出现的天数不同（见图 1）。

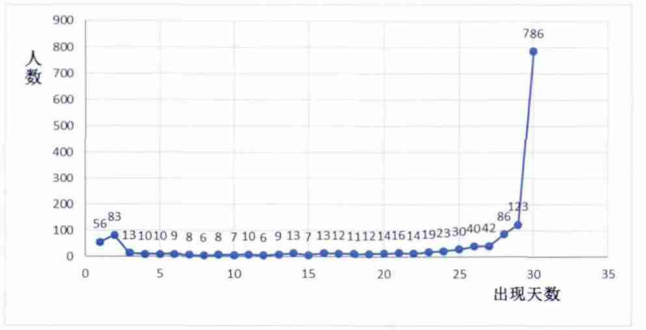

图 1 出现不同天数用户的数量分布

1.2 数据预处理

一部分手机信令数据缺少相应的经纬度坐标，导致一部分数据无效（删除无效数据后，剩余人数为 1451 人）。为了规避乒乓切换现象（手机在服务小区与相邻小区间会来回进行 handover），采取栅格化和分箱法[4]来解决此问题，时间间隔选取为 10 分钟。然后通过逐行相减的方法得到每一个用户在每一个位置的停留时间（单位 秒）以及停留点的经纬度坐标（见表 2，Lat 为纬度，Lon 为经度）。

2 时间阈值法识别居民居住地

时间阈值法是此类研究中经常用到的一种方法，通常将居民每天夜间停留时间超过阈值时间且一个月内出现次数大于阈值次数的停留点，作为居民的居住

表 2 用户每个停留位置的停留时间及其经纬度坐标

CelluarID	DateTime	Lat	Lon	pt
35397788CF36D9BC7A042B99B034C811	20110917084308	31.28045331	121.3743654	1166
35397788CF36D9BC7A042B99B034C811	20110917090234	31.28045331	121.3743654	253
35397788CF36D9BC7A042B99B034C811	20110917090647	31.28045331	121.3743654	2032

表 3　　不同识别规则的识别结果，括号内为识别率

	X ≥ 5h	X ≥ 5.5h	X ≥ 6h	X ≥ 6.5h	X ≥ 7h	X ≥ 7.5h
Y ≥ 10d	1117(74.7%)	1108(74.1%)	1099(73.5%)	1077(72%)	1054(70.5%)	1016(67.9%)
Y ≥ 11d	1099(73.5%)	1089(72.8%)	1069(71.5%)	1040(69.5%)	1011(67.6%)	968(64.7%)
Y ≥ 12d	1065(71.2%)	1046(69.9%)	1030(68.9%)	1005(67.2%)	976(65.2%)	914(61.1%)
Y ≥ 13d	1032(69%)	1016(67.9%)	1001(66.9%)	963(64.4%)	927(62%)	858(57.4%)
Y ≥ 14d	1008(67.4%)	993(66.4%)	969(64.8%)	922(61.6%)	876(58.6%)	811(54.2%)
Y ≥ 15d	958(64%)	947(63.3%)	922(61.6%)	873(58.4%)	808(54%)	740(49.5%)
Y ≥ 16d	930(62.2%)	906(60.6%)	876(58.6%)	824(55.1%)	770(51.5%)	690(46.1%)
Y ≥ 17d	897(60%)	876(58.6%)	843(56.4%)	773(51.7%)	705(47.1%)	635(42.4%)
Y ≥ 18d	862(57.6%)	835(55.8%)	791(52.9%)	715(47.8%)	640(42.8%)	577(38.6%)
Y ≥ 19d	819(54.7%)	786(52.5%)	737(49.3%)	662(44.3%)	587(39.2%)	520(34.8%)
Y ≥ 20d	777(51.9%)	732(48.9%)	681(45.5%)	601(40.2%)	534(35.7%)	454(30.3%)
Y ≥ 21d	731(48.9%)	688(46%)	636(42.5%)	549(36.7%)	473(31.6%)	408(27.3%)
Y ≥ 22d	680(45.5%)	632(42.2%)	572(38.2%)	474(31.7%)	411(27.5%)	342(22.9%)
Y ≥ 23d	631(42.2%)	591(39.5%)	517(34.6%)	419(28%)	353(23.6%)	292(19.5%)
Y ≥ 24d	572(38.2%)	519(34.7%)	441(29.5%)	360(24.1%)	299(20%)	242(16.2%)
Y ≥ 25d	514(34.4%)	447(29.9%)	365(24.4%)	288(19.3%)	235(15.7%)	176(11.8%)

地，具体识别规则如下。

（1）夜间时间规定为 8:00pm~ 次日 8:00am 共 12 个小时，在此时间段内用户在停留点 A 停留时间超过 X 个小时。

（2）在一个月的观察周期内，停留点 A 符合规则（1）的天数超过 Y 天。

根据具体情况，我们可以通过确定不同组合的 X 小时与 Y 天数，来识别研究样本的居住地（见表 3、图 2）。

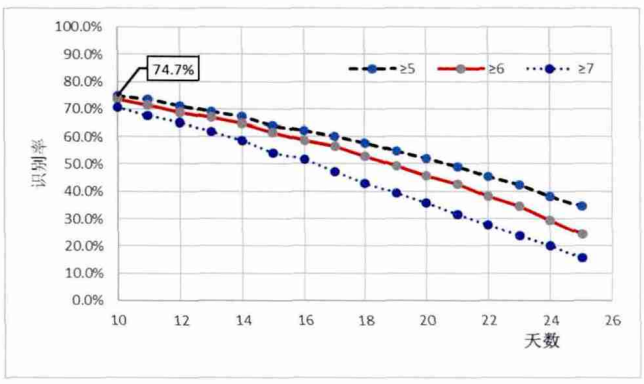

图 2　不同识别规则识别率比较

我们看到，对于不同的 X 和 Y 的值，识别结果差异较大，因此，此种方法并不是非常准确。但是，在要求不是非常严格的情况下，此种方法有一定的实用性。比如，每天停留时间超过 5 小时，重复天数 10 天以上，是能够作为居住地的识别标准的（即 X=5，Y=10）。此时对于研究样本的识别率为 74.7%。此外，1496 人中，出现天数不足 10 天的共有 210 人（见图 1），这些用户对手机的使用程度以及依赖程度较低，考虑到数据为 2011 年，一部分老人可能对手机依赖程度不高，固其使用天数虽然很少，但如果以停留时间相对值作为识别居住地的标准，这 210 人中将有很大一部分用户的居住地将能够识别出来。但是由于这 210 人中，还包含有外地来沪旅游、出差等人群，用停留时间相对值很难将其与常住老人区别开来，故此文未进行讨论。

3　根据信息熵识别居民居住地

香农（C.E.Shannon）在 1948 借鉴热力学的概念，把信息熵定义为信息中排出了冗余后的平均信息量。我们将所研究的每个手机用户作为信息源来考虑，则其在每一个位置停留的时间长短以及位置之间的变化

频率变可以作为一种信息量来研究。由此，我们便可以定义每一个用户在研究时间段内（30 天）的信息熵如下。

将每个用户个体作为信号源：X_i（本研究中 i=1~1496），假设信号源 X_i 在研究的时段内共在 n 个位置停留过，则将其每一个停留位置作为一个信源输出符号 U_j（j=1~n），相对应的概率为 $P_{ij}=T_{ij}/T$，其中 T_{ij} 为用户 X_i 在停留点 j 的停留时间，T 为研究时间段总时间。则信息熵为（公式 1）：

$$H(X_i)=H(U_i)=E(-\log_2 P_{ij})=-\sum_{j=1}^{n} P_{ij}*\log_2 P_{ij} \quad (1)$$

信息熵值的单位为“比特”，大小表示了所研究个体活动强度，其值越小表示该个体越稳定。如一个个体在所观测时间段内，没有任何移动（即一直停留在同一个在位置），那么该个体在所观测时间段内的信息熵为零，而个体移动的越频繁，则其信息熵值越大。为了能够直观的反应出个体的运动强度，表 4 给出了一些信息熵的参考值。

表 4　不同情景下的信息熵参考值

	个体在 6 小时的观测时间段内的运动情况	信息熵值 H/ 比特
情景 1	个体一直停留在 A 点 6 小时	0
情景 2	在 A 点停留 3 小时，在 B 点停留 3 小时	1
情景 3	在 A、B、C、D 分别停留 1.5 小时	2
情景 4	在 A 点停留 5.5 小时，在 B 点停留 0.5 小时	3.71

针对我们所研究的问题，绝大多数居民在 0:00~6:00 之间，应该处于睡眠状态，因此，在这一时间段居民的信息熵值应该较小，而相应的空间位置应该属于该用户的居住地。因此，我们计算了研究样本在该时间段的信息熵值，结果见图 3。

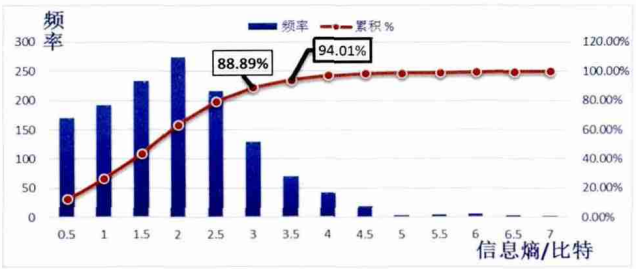

图 3　信息熵分布情况

根据表 4 的情景 4，我们假设在时间段 0:00~6:00 内，一个用户如果提前出门半小时或者晚归半小时，其信息熵也应小于 3.71（如果一直停留在家，则熵值为 0），因此，我们可以将 3.5 作为判断识别用户居住地的标准，即若 $H(i) \le 3.5$，则可以判断该地点为居住地。此时的识别率为 94.1%，如果将该标准严格到 3，则 88.9% 的识别也是非常理想的。

为了进一步直观的描述此方法，图 4 给出了所研究样本 2011 年 9 月 3 日 ~2011 年 9 月 30 日完整的四个星期的信息熵变化情况的热力图。

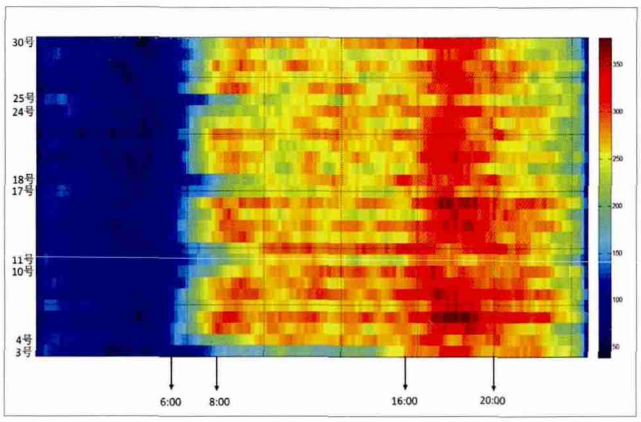

图 4　样本信息熵随时间变化

图 4 中，横坐标为一天的 0~24 小时，纵坐标为 2011 年 9 月 3 号至 30 号（其中该月的 3、4、10、11、17、18、24、25 为周末）。蓝色代表熵值很小，个体处于稳定状态（很少移动，最小值为 0），而红色代表个体处于活跃状态。我们可以看出，在早 6 点之前，颜色几乎为深蓝色，表明用户基本处于静止状态。因此，我们可以将在该时间段内信息熵很小的地点作为用户的居住地。

4　根据相对停留时间识别居住地

绝大多数居民，在所观测时间段内（0:00~6:00），都处于休息、睡眠状态。在这段时间，居民应该主要停留在居住地，因此，我们可以计算居民在每个停留点的相对停留时间，作为识别居民居住地的指标。样本 i（i=1，2，……，1496）在停留点 j（j=1，2，……，

J）的相对停留时间为 ：

$$P_i^j = \frac{T_i^j}{\sum_{j=1}^{J} T_i^j} \quad (2)$$

试中，T_i^j为样本 i 在停留点 j 总的停留时间。因此，我们有$\sum_{j=1}^{J} P_i^j = 1$。

针对每一个居民，我们选取使得 P 取值最大的 j 作为居民的居住地，即 ：

$$\arg\max_j (P_i^j) \quad (3)$$

表 5 给出了 5 个样本的前 5 个 P 的取值（按照由大到小排列），我们可以发现，通常每个用户的最大的相对停留时间要远大于其它相对停留时间。

表 5　　5 个样本的相对时间值

	样本 1	样本 2	样本 3	样本 4	样本 5
1	0.820	0.469	0.668	0.697	0.564
2	0.070	0.161	0.232	0.126	0.117
3	0.057	0.154	0.062	0.064	0.099
4	0.017	0.147	0.023	0.056	0.078
5	0.017	0.069	0.016	0.049	0.062

按照这种方法，我们能够将所有用户的居住地识别出来。但是，这种方法忽视了上夜班或者作息不规律的一部分人，可能将其的工作地错误的识别为居住地。但考虑到此类情况在居民中所占比例不大，这种方法在一定误差范围内也能够被接受。

5　结论

本文基于手机信令数据，通过 3 种不同的方法，对 1496 个研究样本的居住地进行了识别，并得到了以下结论。

（1）现有文献中最常用的方法为时间阈值法，但经过本文的讨论，发现不同的时间阈值对于识别结果差异很大。因此，时间阈值法并不是一个非常科学、严谨的方法，不建议在此类问题中使用该方法。

（2）信息熵能够体现出一个个体在所观测时间段内的运动情况。由于居民在夜间休息时与在白天时的活动情况迥然不同，因此，通过信息熵来识别居民居住地是一个非常理想的方法，而且识别率较高。

（3）时间相对值判别法能够将所有研究样本的居住地识别出来。但是，由于部分居民的工作时间为夜间，也有些居民的作息时间极不规律，因此，此种方法会有一定的错误率。但介于这一部分的人群数量不是很大，这种方法也有一定的可行性。

大数据背景下的交通问题都有了新的思路与研究方法[5]。对于通勤交通与职住分布来说，能够通过新型数据找到问题突破口，能够避开传统数据的须多弊端。本文研究了手机信令数据在识别居民居住地问题上的表现，下一步的工作将对工作地进行识别，并研究居民相应的通勤与职住分布情况。

【参考文献】

[1] Järv O., Ahas R., Witlox F.. Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records[J]. Transportation Research , 2014, 38(1):122–135.

[2] 许宁 , 尹凌 , 胡金星 . 从大规模短期规则采样的手机定位数据中识别居民职住地 [J]. 武汉大学学报（信息科学版）, 2014(6):750-756.

[3] Iqbal M. S., Choudhury C. F., Wang P., et al. Development of origin–destination matrices using mobile phone call data[J]. Transportation Research , 2014, 40(1):63–74.

[4] Li W., Cheng X., Duan Z., et al. A framework for spatial interaction analysis based on large-scale mobile phone data[J]. Computational Intelligence & Neuroscience, 2014, 2014:363502-363502.

[5] 杨东援 , 段征宇 . 大数据环境下城市交通分析技术 [M]. 上海 ：同济大学出版社 , 2015:51.