



大数据技术与城市计算

手机信令大数据和用户画像及社会隔离

姚尧 博士，副教授，高级工程师

地理与信息工程学院，地图制图学与地理信息工程

阿里巴巴集团，达摩院，访问学者

Email: yaoy@cug.edu.cn

办公地点：未来城校区地信楼522办公室





主要内容



- 1 手机信令大数据简介
- 2 手机信令大数据处理技术
- 3 手机信令大数据与用户画像
- 4 手机信令大数据与用户位置预测
- 5 手机信令大数据与城市活力
- 6 手机信令大数据与社会隔离



主要内容

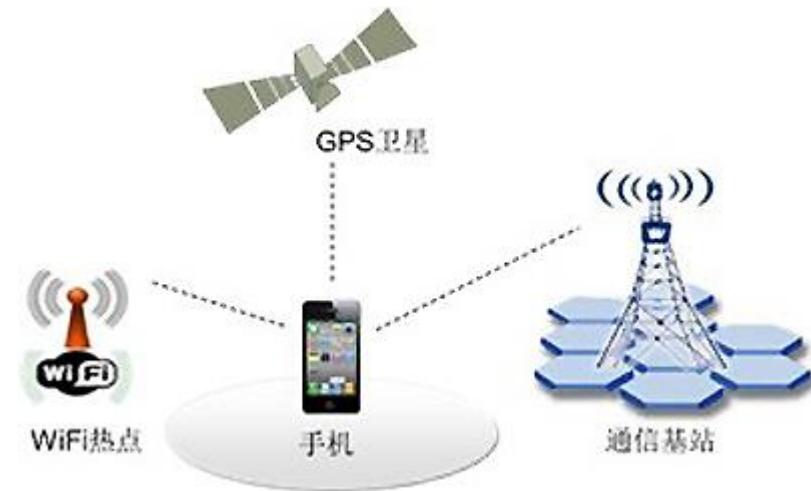


- 1 手机信令大数据简介
- 2 手机信令大数据处理技术
- 3 手机信令大数据与用户画像
- 4 手机信令大数据与用户位置预测
- 5 手机信令大数据与城市活力
- 6 手机信令大数据与社会隔离

01 | 手机信令大数据简介



为什么有时，没有开手机GPS定位，部分APP还是可以获取我们大体的位置信息呢？



01 | 手机信令大数据简介



手机信令数据：是由手机用户在**发生通话、发短信或移动位置等事件**时，被运营商的通信**基站捕获**并记录同一用户信令轨迹所产生的数据。



01 | 手机信令大数据简介



和之前讲过的数据相比，手机信令数据在城市计算中有什么优势？



VS



01 | 手机信令大数据简介



怎样去定量的模拟/分析细尺度空间，城市内部的人群活动模式和城市发展状况？



费时费力，存在主观误差，难以收集整个城市的全部数据。



仅仅记录线路覆盖的城市区域和人群，时间只记录白天，无法记录夜生活。



APP鄙视链，什么年龄段用什么APP,数据存在有偏性。

01 | 手机信令大数据简介



手机信令数据的优势：



VS



- ✓ 收集方便、速度较快、成本较低；
- ✓ 可以无缝的记载全时段（白天+夜晚）；
- ✓ 人群覆盖面广、有偏性弱。

城市规划的“大杀器”

手机信令数据存在的问题：

- ✓ 数据的空间精度依赖于城市中基站的覆盖面积，一般城市中心区域空间分辨率可达200m,但在偏远地区>1km (**分辨率精度不统一**)；
- ✓ 因为基站负载平衡原因，若某一基站负载过多，其负责区域的手机信令将转移给其他基站 (**信号位置跳跃**)；
- ✓ 夜晚为了节约成本，服务商会主动关闭部分基站 (**位置记录存在误差**)。



主要内容



- 1 手机信令大数据简介
- 2 手机信令大数据处理技术
- 3 手机信令大数据与用户画像
- 4 手机信令大数据与用户位置预测
- 5 手机信令大数据与城市活力
- 6 手机信令大数据与社会隔离

02 | 手机信令大数据处理



手机信令数据研究一般流程：

手机信令数据获取

1. 根据需求确定所需数据的时间和地点范围；
2. 明确数据库中各分表的关系和要读取字段信息；
3. 数据库对应表格数据读取和处理；
4. 设计数据库内外读取接口。

数据清洗

1. 数据筛选：筛除非相关数据信息；
2. 数据清洗：去除冗余数据和无法读取的数据；
3. 数据去重：对于同名多号码等数据进行去重处理。

数据处理

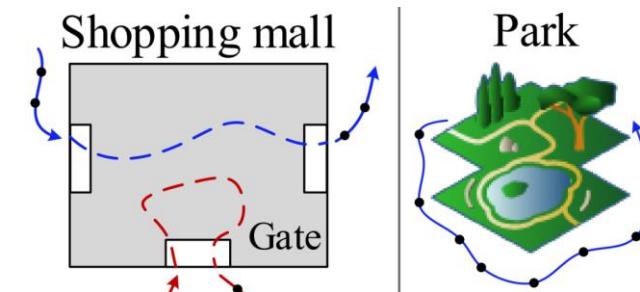
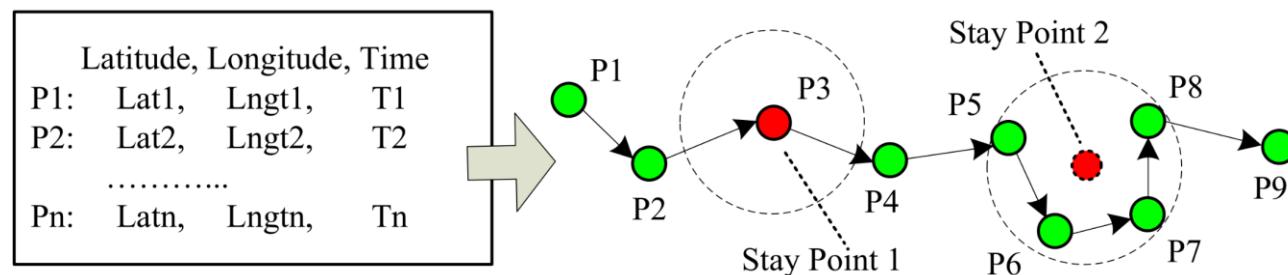
1. 停住点提取：提取轨迹中用户停留时间较长，具有实际意义的活动点；
2. 地图匹配：将活动点匹配至路网地图中；
3. 轨迹分割：将轨迹划分为片段，进行更有目的的研究。

研究应用

1. 用户活动模式挖掘；
2. 城市空间分析；
3. 轨迹预测；
...

手机信令驻点提取：

出行链中轨迹点，并不是同等重要的，有些轨迹点表示人们已经停留的活动点，如购物中心和旅游景点（驻点）。通过驻点提取算法，我们可以提取到手机用户的活动点位置以及驻留时间，并且在一定程度上可以，**减小基站信号跳跃所导致的系统误差**。



迭代遍历手机信令数据出行链中得每个点，寻找用户停留超过一定时间阈值和空间阈值得点。

- 从第一个点开始遍历，计算当前点和后续点的距离；
- 如果距离大于了空间阈值，则计算此段距离得时间跨度；
- 若时间跨度大于给定时间阈值，则进行聚类等算法获取停驻点，如果没有大于给定时间阈值，则放弃，并以该点为下一次计算得起点；
- 不断迭代，直到找到所有停驻点。

Algorithm StayPoint_Detection(P , $distThreh$, $timeThreh$)

Input: A GPS log P , a distance threshold $distThreh$
and time span threshold $timeThreh$

Output: A set of stay points $SP=\{S\}$

```

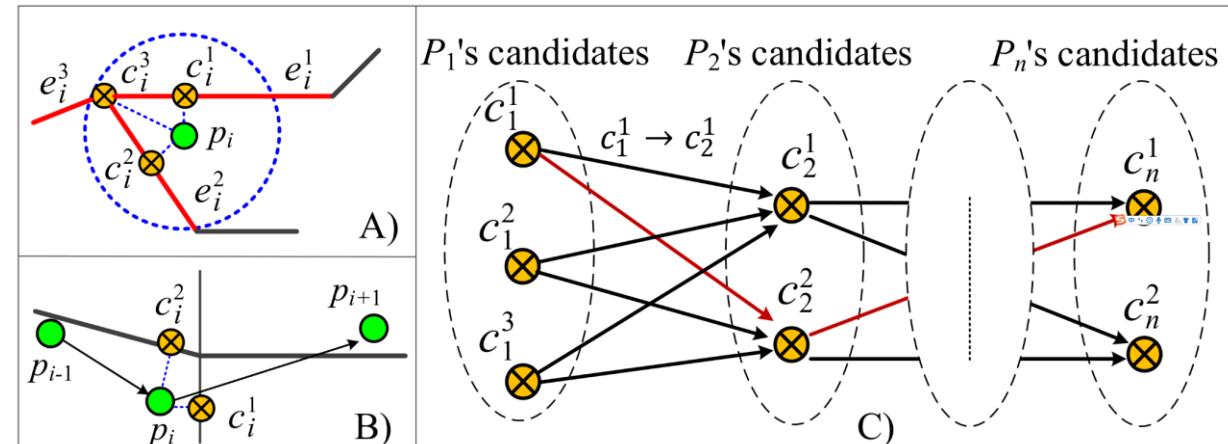
1.  $i=0$ ,  $pointNum = |P|$ ; //the number of GPS points in a GPS logs
2. while  $i < pointNum$  do,
3.    $j:=i+1$ ;
4.   while  $j < pointNum$  do,
5.      $dist=Distance(p_i, p_j)$ ; //calculate the distance between two points
6.     if  $dist > distThreh$  then
7.        $\Delta T=p_j.T-p_i.T$ ; //calculate the time span between two points
8.       if  $\Delta T>timeThreh$  then
9.          $S.coord=ComputMeanCoord(\{p_k \mid i \leq k \leq j\})$ 
10.         $S.arrT=p_i.T$ ;  $S.levT=p_j.T$ ;
11.         $SP.insert(S)$ ;
12.         $i:=j$ ; break;
13.         $j:=j+1$ ;
14.    return  $SP$ .

```

手机信令地图匹配：

地图匹配是指在进行通勤等移动距离计算时，将获取的停驻点或移动点的原始经纬度，转换为连续的和道路网匹配的位置数据。和直接使用原始停驻点计算欧氏距离相比，地图匹配可以获取更为准确的用户移动点。当引入用户活动的上下文语意时，可以提高手机信令数据的定位精度。

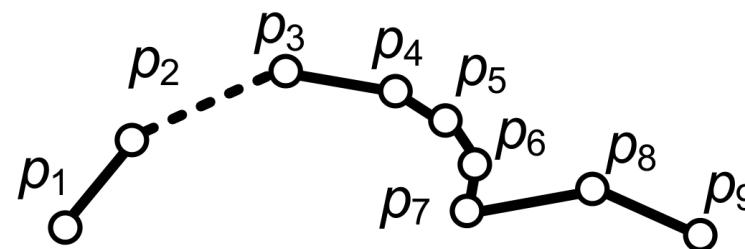
- 计算停驻点和相邻路网边的距离，选取满足距离阈值的候选点；
- 基于用户出行链的上下文信息（如确定方向速度）对每个候选点进行概率权重分配；
- 将停驻点匹配至最概率最大的候选点。



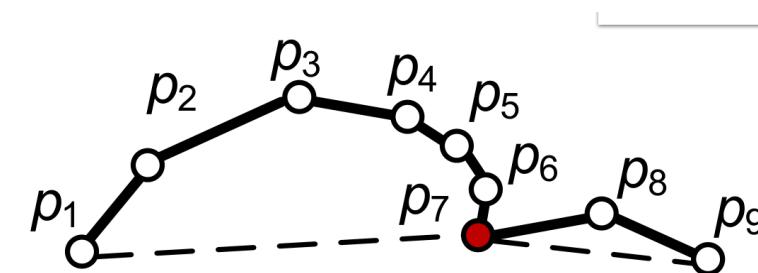
手机信令轨迹分割：

理论上手机信令数据的时间分辨率**可以达到秒级别**，可以记录用户一天详细的出行信息。但是在许多情况下，我们需要对**长时序的**手机信令轨迹进行分割，**分割为多个轨迹段**，使得后续研究更加精细。

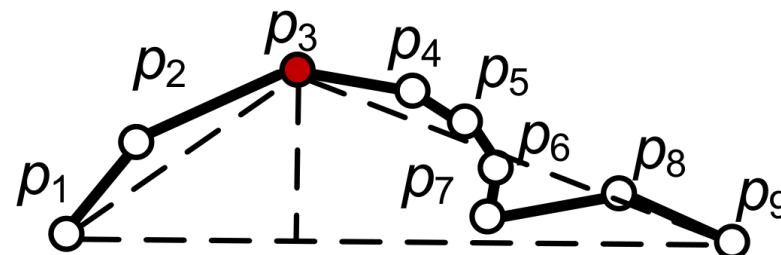
➤ 基于时间间隔：



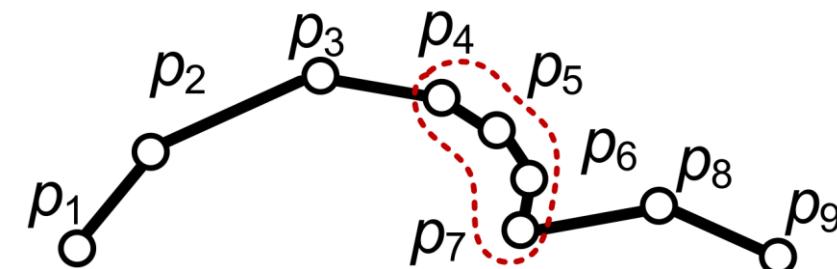
➤ 基于轨迹形状：



➤ 基于形状关键点：



➤ 基于停驻点：



02 | 手机信令大数据处理



手机信令数据研究一般流程：

手机信令数据获取

1. 根据需求确定所需数据的时间和地点范围；
2. 明确数据库中各分表的关系和要读取字段信息；
3. 数据库对应表格数据读取和处理；
4. 设计数据库内外读取接口。

数据清洗

1. 数据筛选：筛除非相关数据信息；
2. 数据清洗：去除冗余数据和无法读取的数据；
3. 数据去重：对于同名多号码等数据进行去重处理。

数据处理

1. 停住点提取：提取轨迹中用户停留时间较长，具有实际意义的活动点；
2. 地图匹配：将活动点匹配至路网地图中；
3. 轨迹分割：将轨迹划分为片段，进行更有目的的研究。

研究应用

1. 用户活动模式挖掘；
2. 城市空间分析；
3. 轨迹预测；
...

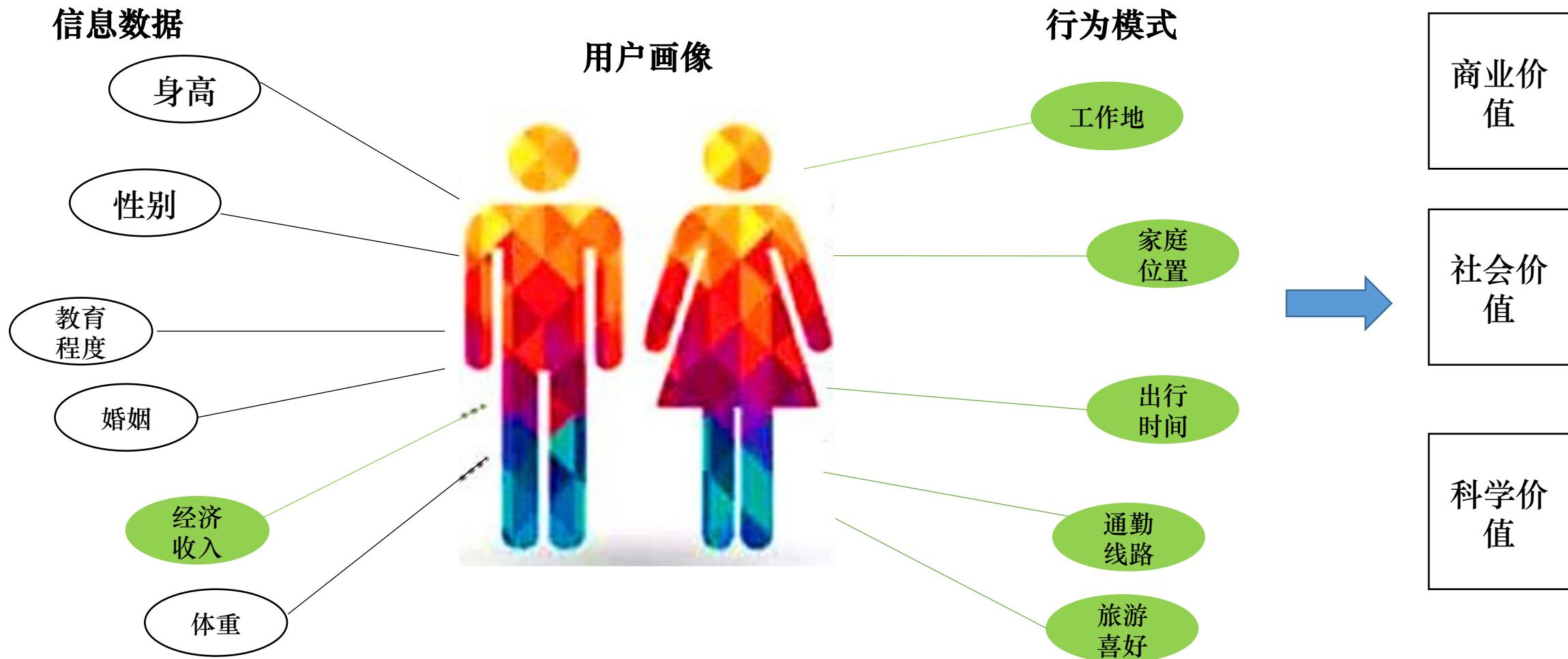


主要内容



- 1 手机信令大数据简介
- 2 手机信令大数据处理技术
- 3 手机信令大数据与用户画像
- 4 手机信令大数据与用户位置预测
- 5 手机信令大数据与城市活力
- 6 手机信令大数据与社会隔离

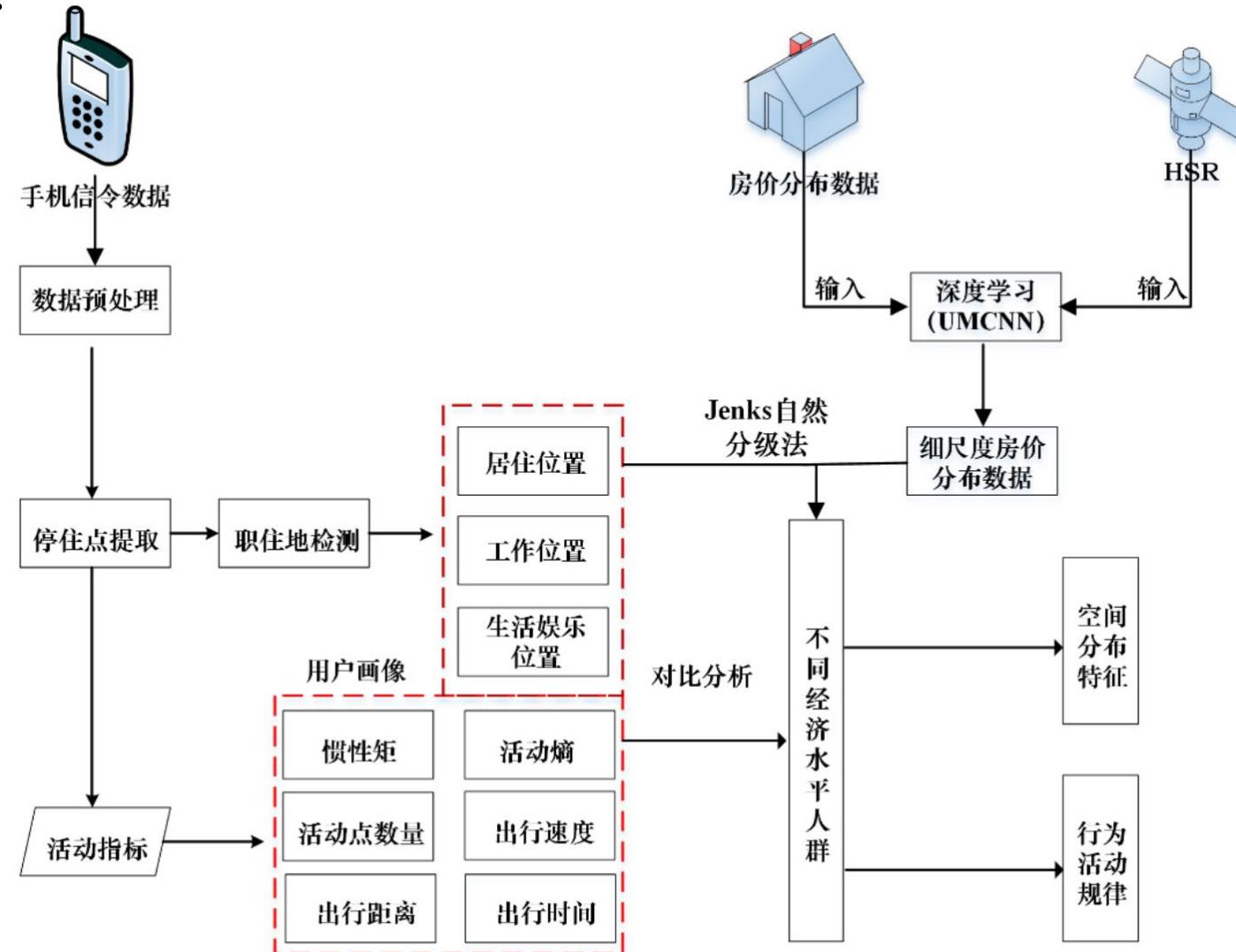
03 | 手机信令大数据与用户画像



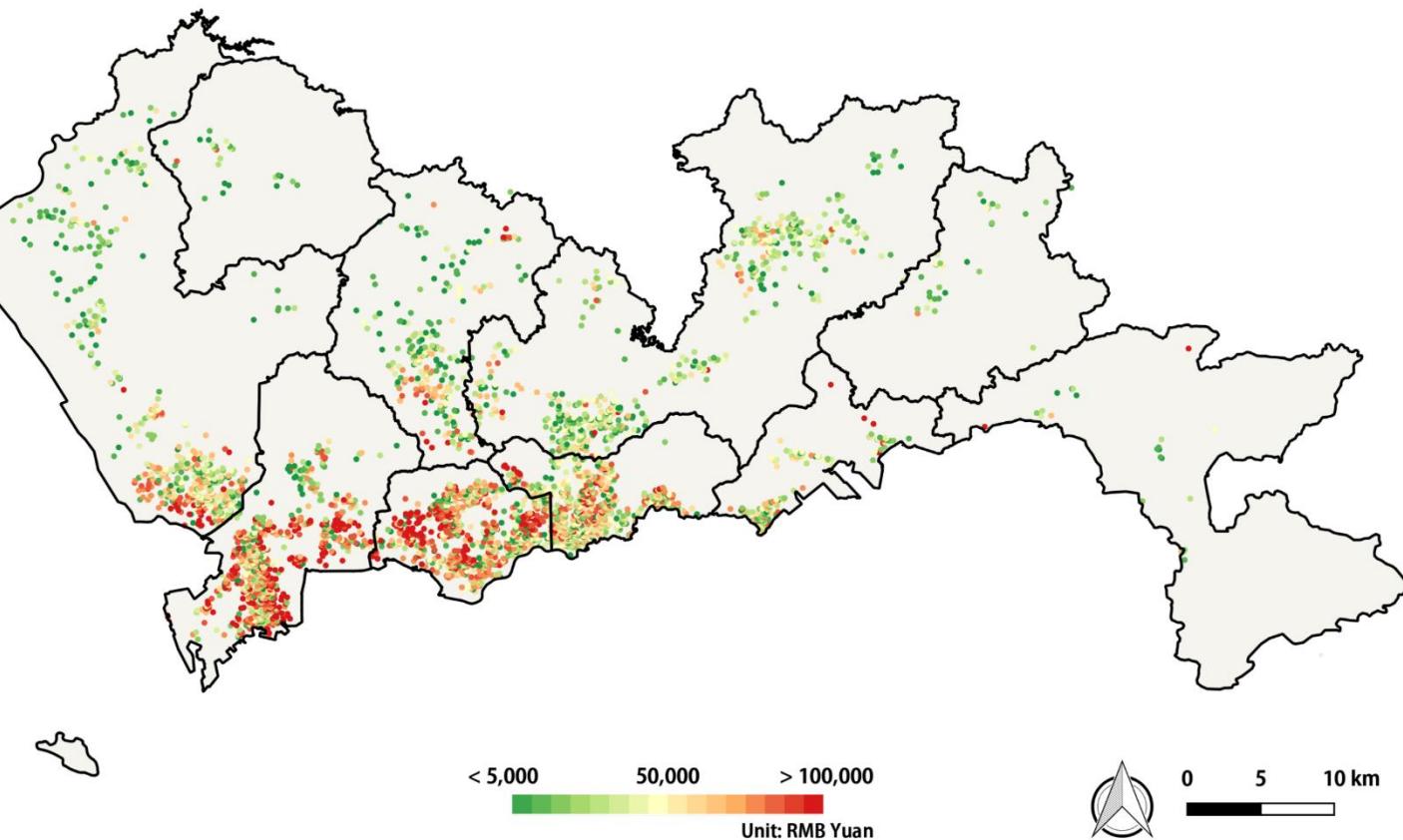
03 | 手机信令大数据与用户画像

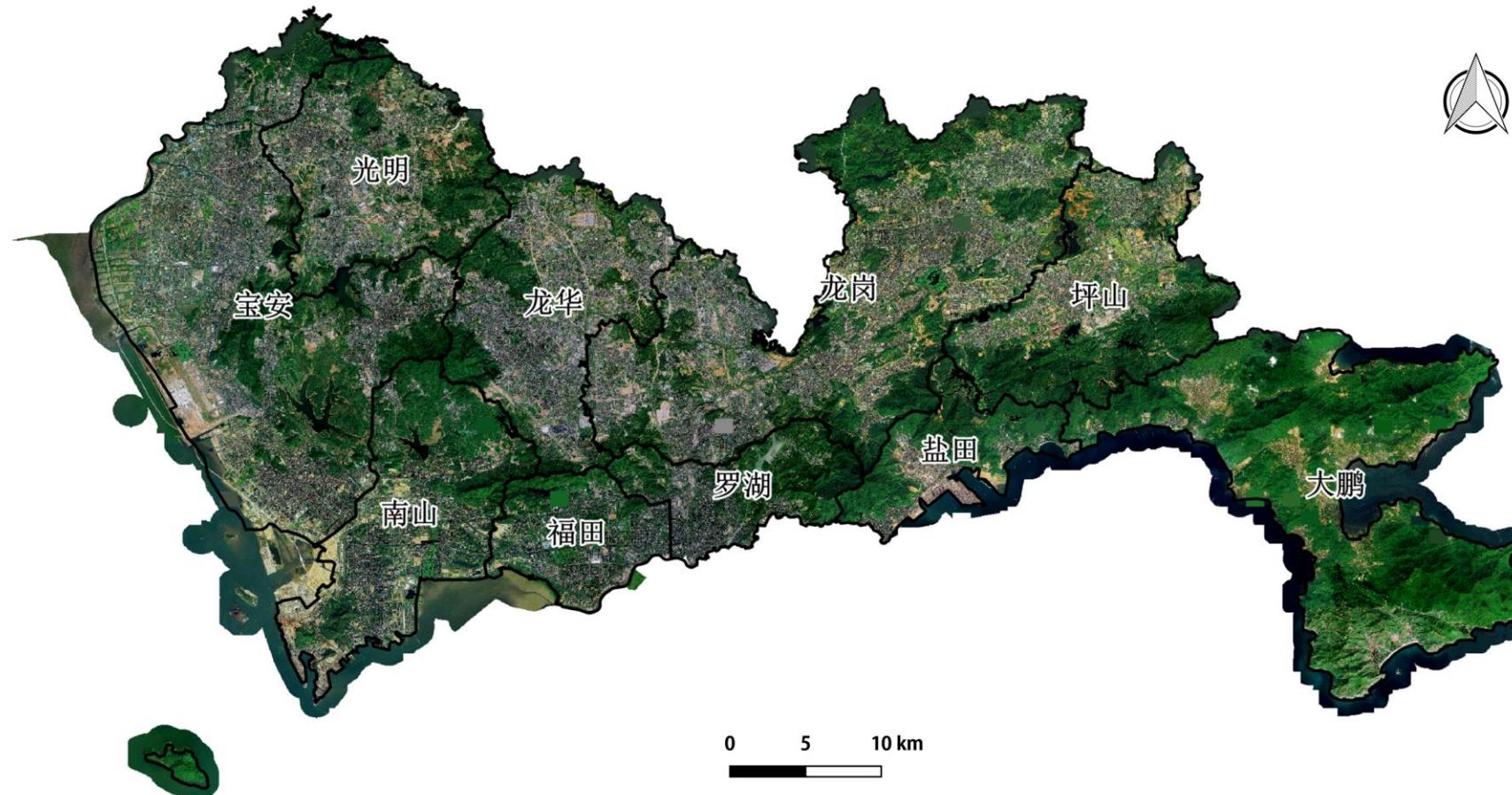


如何构建用户活动模式画像，并定量的分析城市内部不同经济水平人群的活动模式？

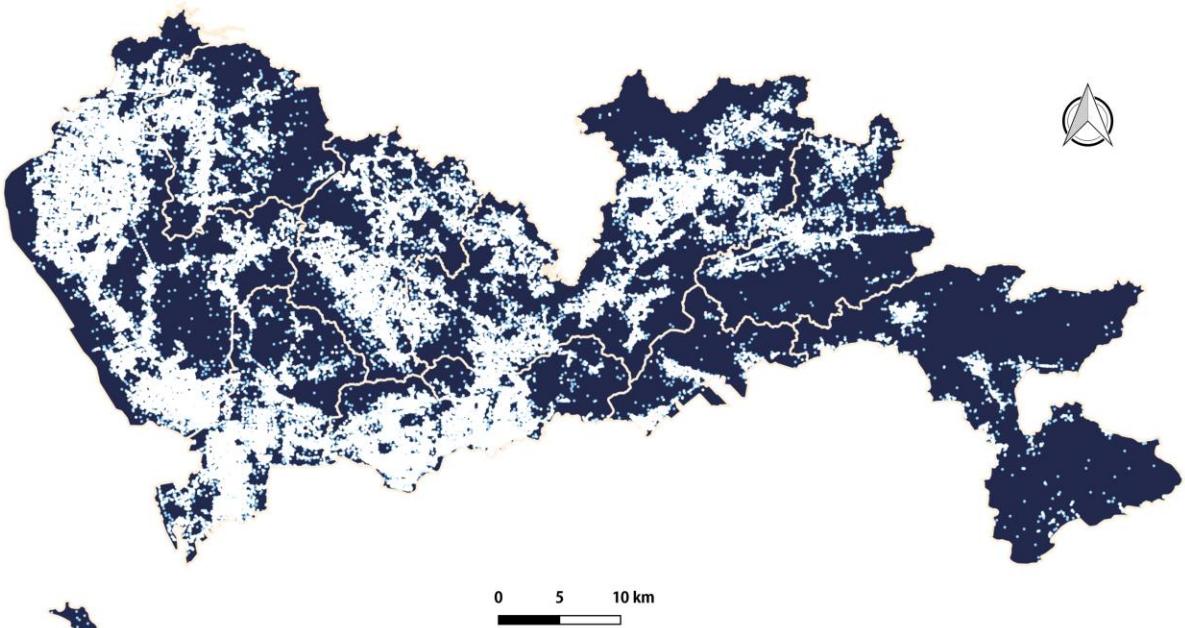


FID	lat_84	lng_84	NAME	PRICE	ADDRESS
0	21.221644	110.406079	假日花园	11089	龙珠大道与龙珠五路交汇处
1	22.258245	114.139387	同富雅苑	4492	向兴路
2	22.268627	114.186761	翠湖茗苑	2524	龙华梅龙路与大和路交汇处
3	22.283032	114.154900	锦鲤小区	3157	观澜大和路
4	22.291601	114.196829	天御豪庭	48795	新湖路与兴华一路交汇
5	22.324931	114.171074	百花小区	10000	布吉镇百花路东一巷4号
6	22.338718	114.152964	骏凯豪庭	8898	南环和宝安大道交汇处
7	22.340854	114.207683	鸿景花园	3571	洪湖路
8	22.395768	113.966597	富豪山庄	17168	坂田大道南路
9	22.469256	114.570667	中兴佳苑	5000	大亚湾区中兴五路
10	22.473449	114.223368	南山世纪	39325	南山工业园南山南路
11	22.480220	113.883673	海天楼	13714	赤湾
12	22.481668	113.908464	顺发大厦	4000	工业一路3号
:	:	:	:	:	:

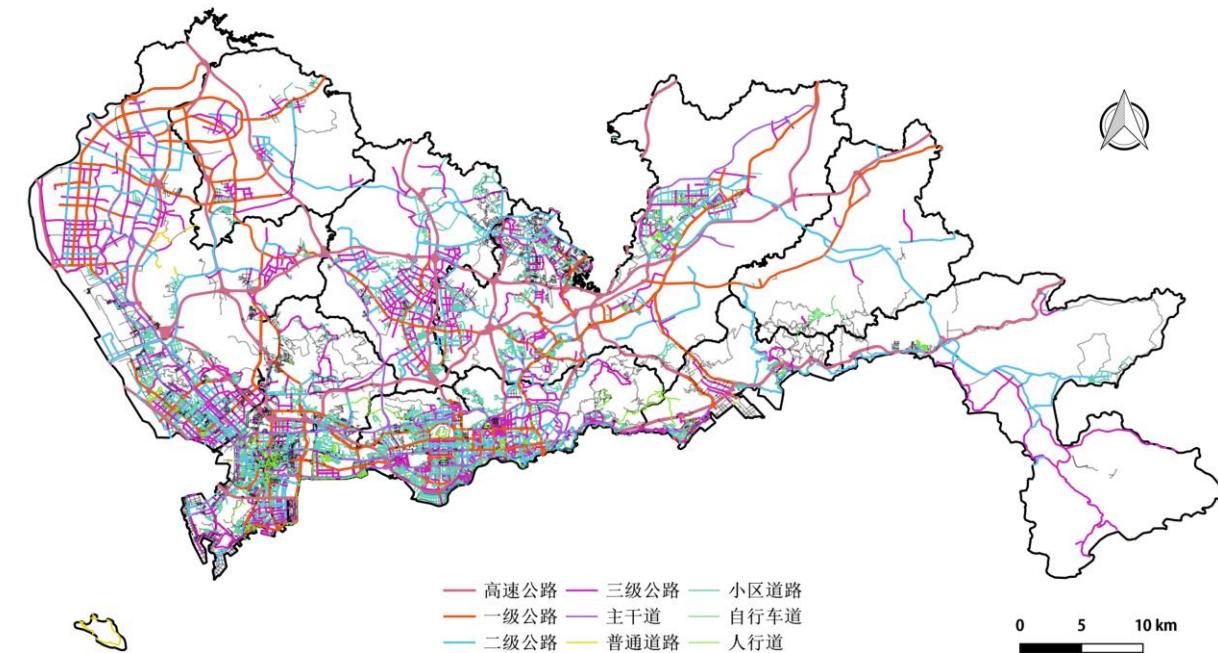




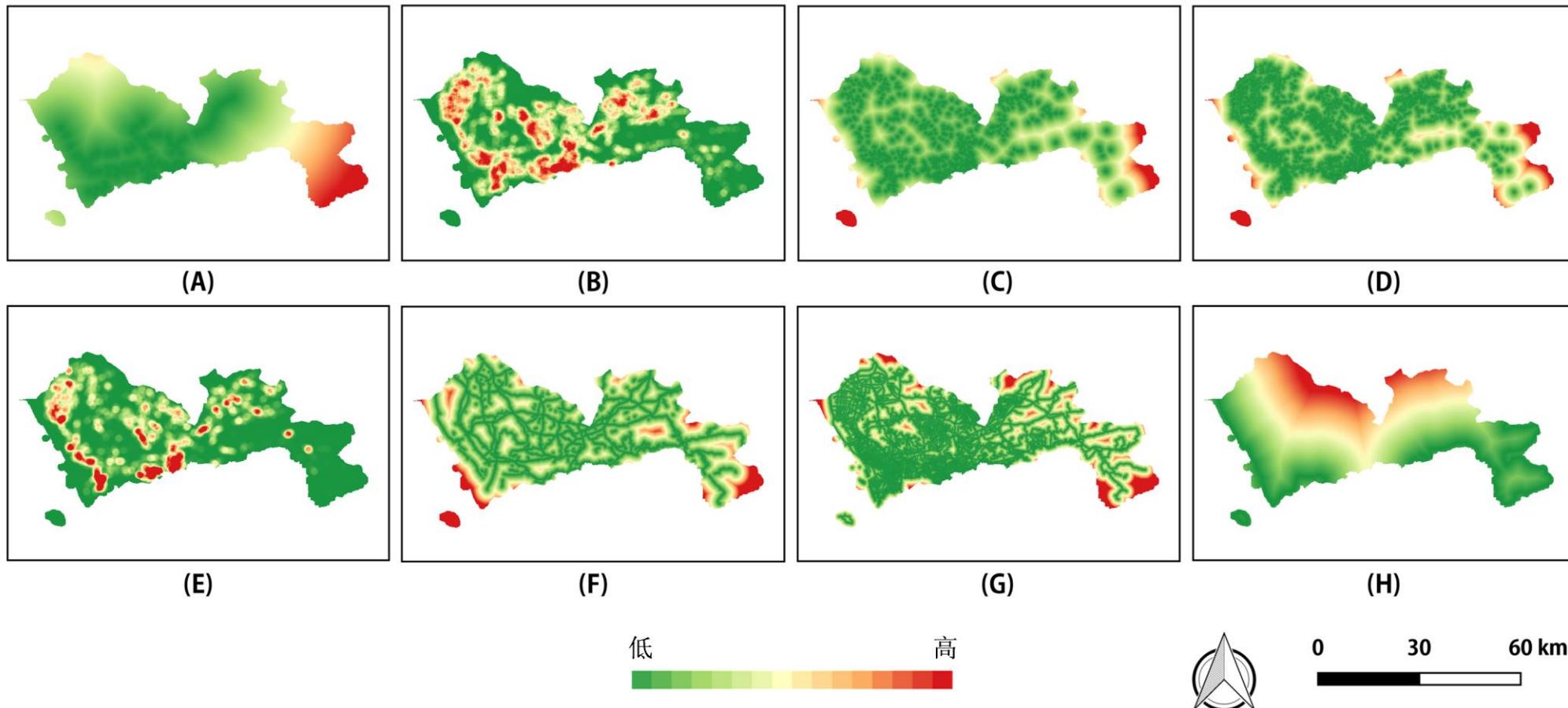
深圳市的**高分辨率遥感影像**
(High Spatial Resolution, HSR)
从**天地图**(tianditu.cn)获取，**分辨率**为5米，大小为13,976 * 22,514，
包含RGB三个波段。



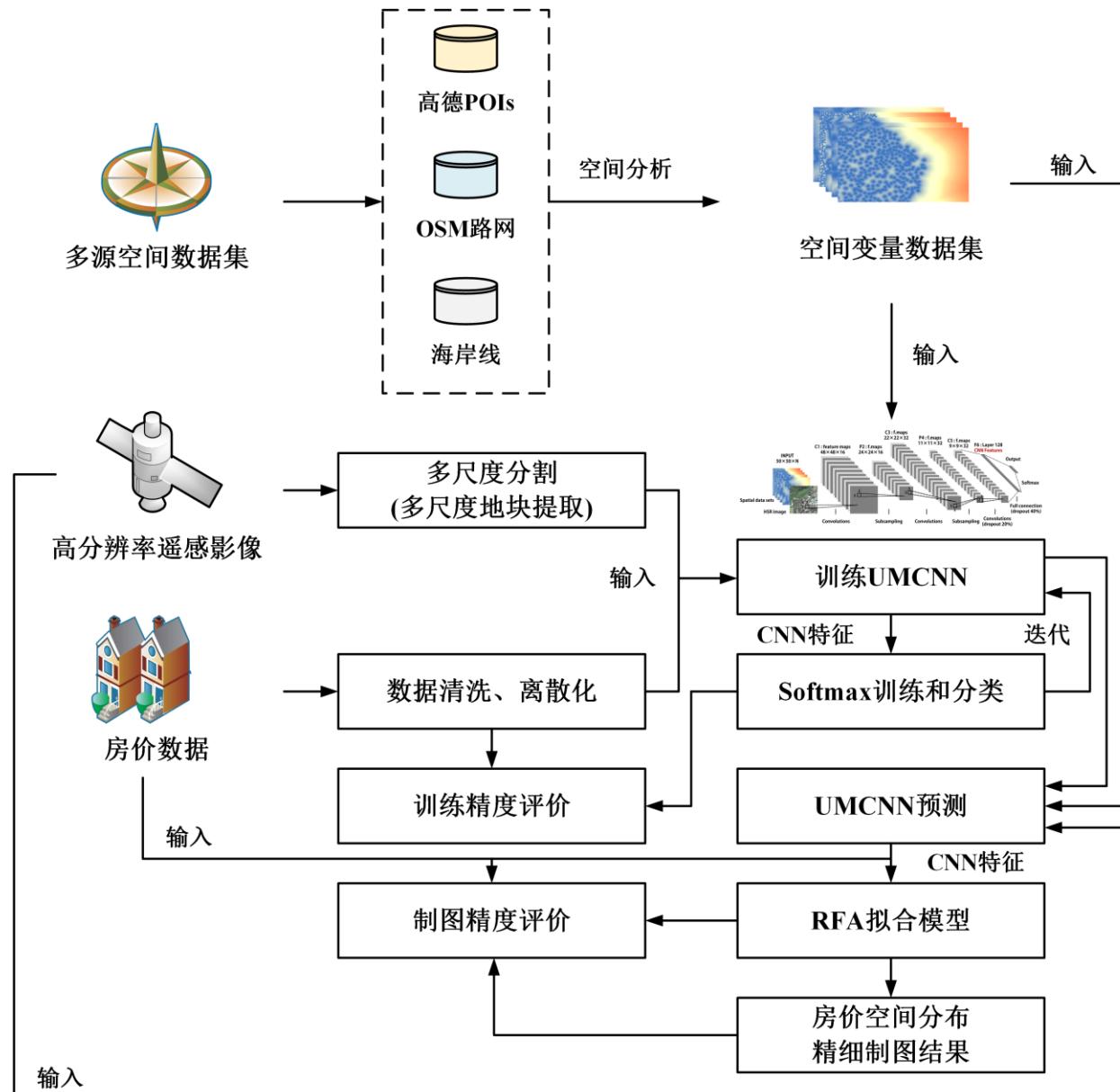
本研究所使用的**感兴趣点数据(Point-of-interest, POIs)**从**高德地图(<http://lbs.amap.com/>)**通过网络爬虫获取，共获得包含20个类别共**211,076条记录**，包括商业设施、教育机构、医疗设施等类别。



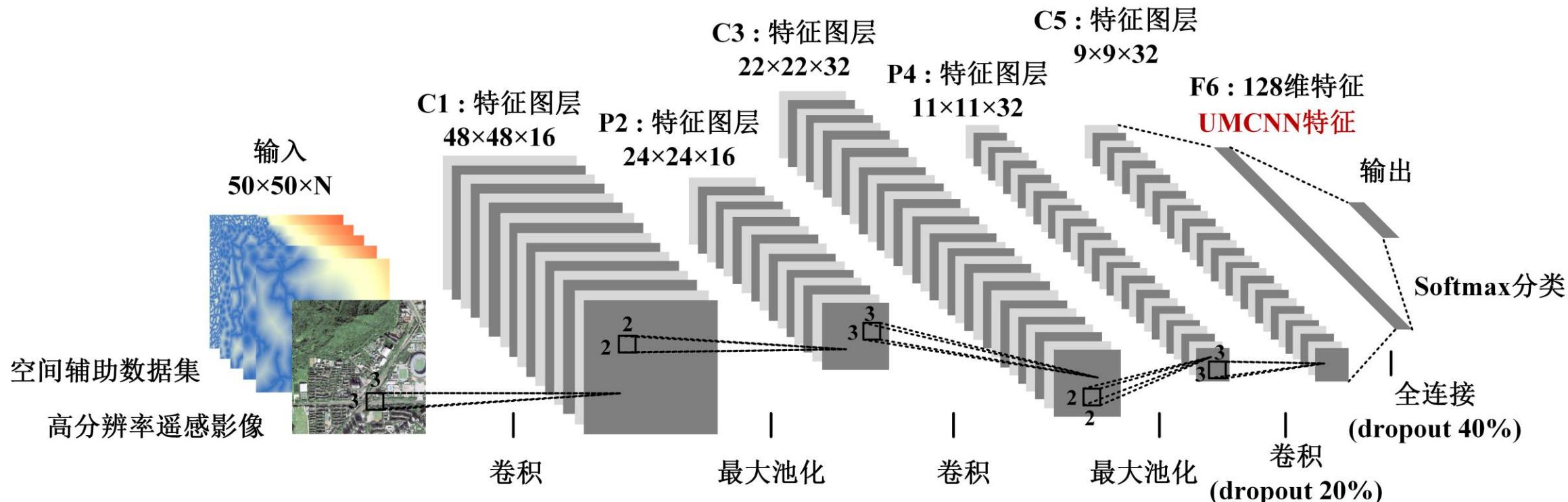
本研究所使用的路网数据从**OpenStreetMap(OSM, <http://www.openstreetmap.org/>)**获取，是由大众共同打造的免费开发和可编辑的地图服务，路网数据中包含了高速公路、主干道、人行道等50多种类型。



(A)到地铁站的距离 (B)公交车站点密度 (C)到幼儿园和小学的距离 (D)到医疗设施的距离
(E)生活服务设施密度 (F)到主要道路的距离 (G)到其他道路的距离 (H)到海岸线的距离



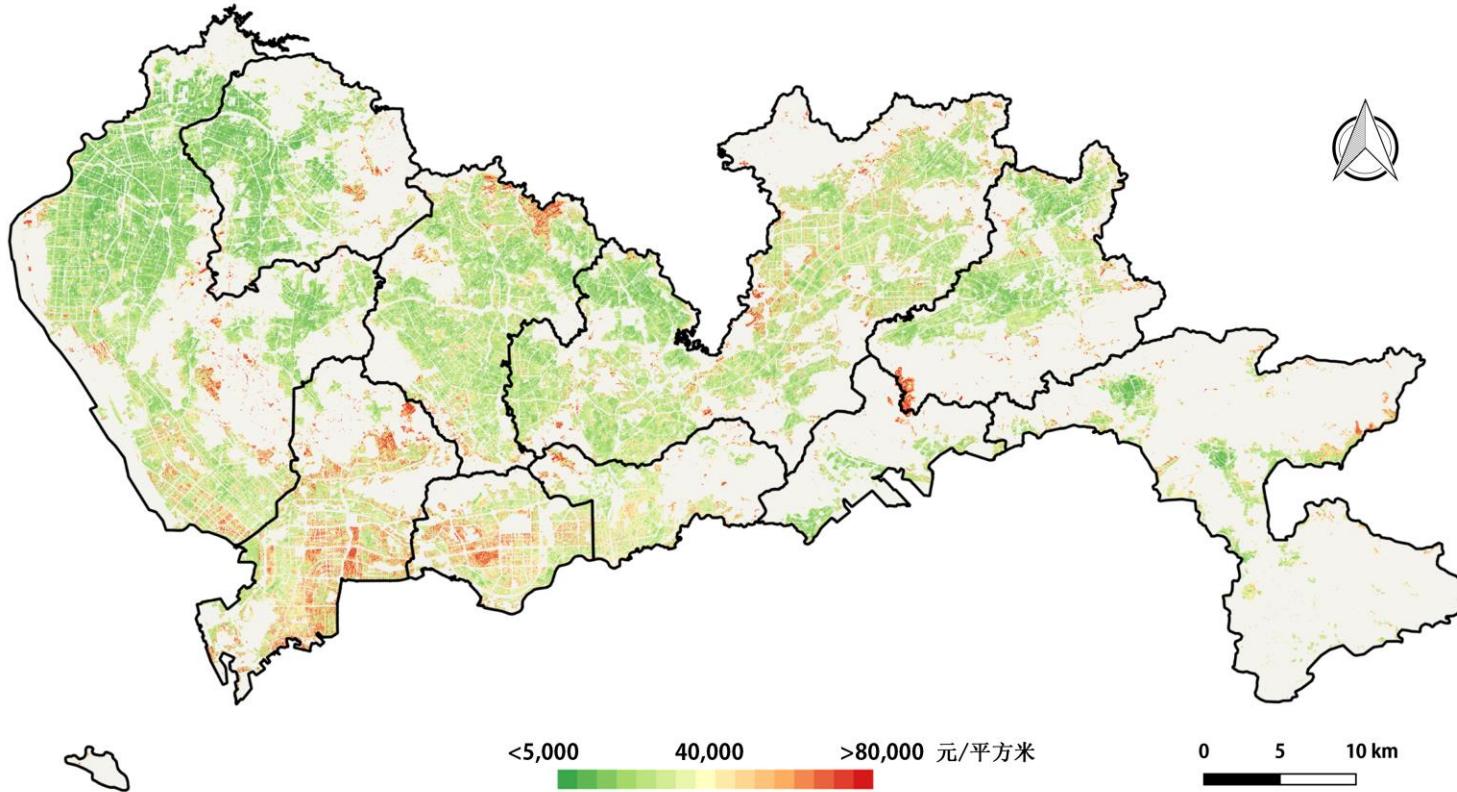
- a) **数据预处理以及多尺度采样，利用八个空间变量以及遥感影像构建多源多尺度的城市房价影响因素数据集**
- b) **利用上述构建的数据集和离散化后的房价数据训练联合挖掘卷积神经网络(Convolution Neural Network for United Mining, UMCNN)，获得最优的模型**
- c) **根据获得的UMCNN模型，去除最后一层分类层，将倒数第二层的向量作为特征输入随机森林拟合模型进行训练**
- d) **根据得到的UMCNN模型和随机森林拟合模型，进行逐像元预测得到深圳市的精细房价，最后进行精度评估和不确定性分析**



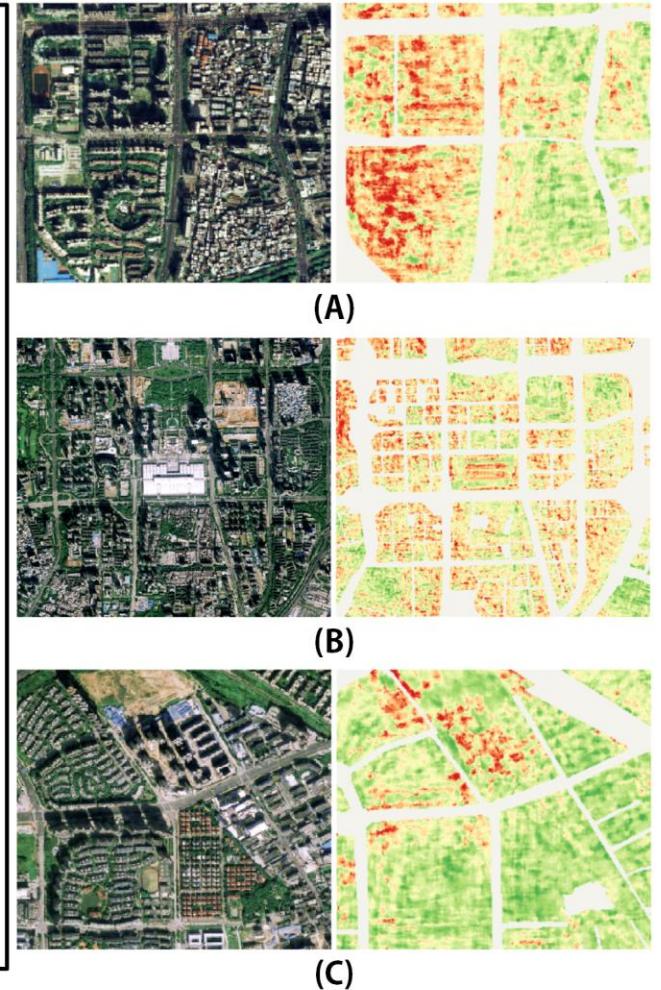
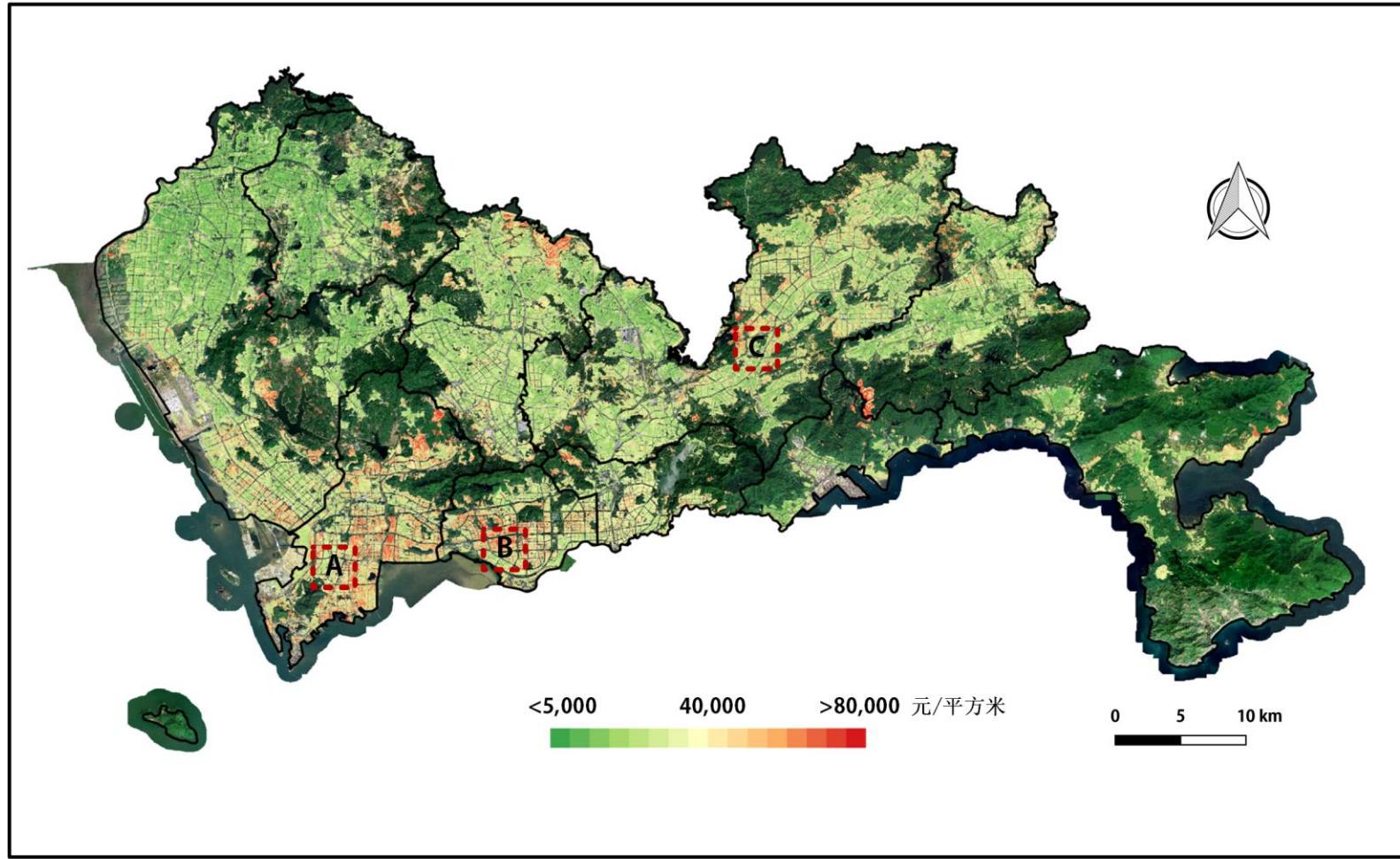
- 出于计算资源以及训练样本的考虑,本研究所使用的联合挖掘神经网络(Convolution Neural Network for United Mining, UMCNN) 是一个相对较为简单的模型
- 该模型有**3层卷积层, 2层最大池化层, 以及一个全连接层和一个Softmax分类层**
- 卷积层采用**3*3 的卷积子, 步长为1个像元**
- 最大池化层采用**2 * 2的窗口, 步长为2**
- 在最后的卷积层以及全连接层中采用**Dropout机制**防止过拟合

窗口大小		25	50	75	100	125	150
训练精度	训练	73.24%	90.63%	90.60%	91.97%	90.91%	90.91%
	验证	74.61%	80.29%	80.70%	83.31%	79.01%	82.30%
	时间 (s)	3125	3777	5067	7188	10336	12898
预测精度	Pearson R	0.764	0.922	0.920	0.919	0.907	0.911
	Standard R2	0.530	0.745	0.743	0.742	0.738	0.740
	RMSE	26.524	18.456	19.887	19.606	20.066	19.141
	%RMSE	21.75%	15.13%	16.31%	16.08%	16.45%	15.69%
	MAE	13.534	9.757	10.578	9.839	9.959	9.817
	%MAE	36.15%	26.20%	28.93%	26.32%	27.31%	26.18%
	P-Value	0.269	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001

- 随着采样窗口的增大，训练精度和验证精度逐渐增大最后趋于平稳；
- 当采样窗口大于 50*50 时，房价模拟结果的精度趋于平稳；
- 另外，本研究的模型采用CPU运行，随着采样窗口的增大，时间的消耗呈指数级增长。



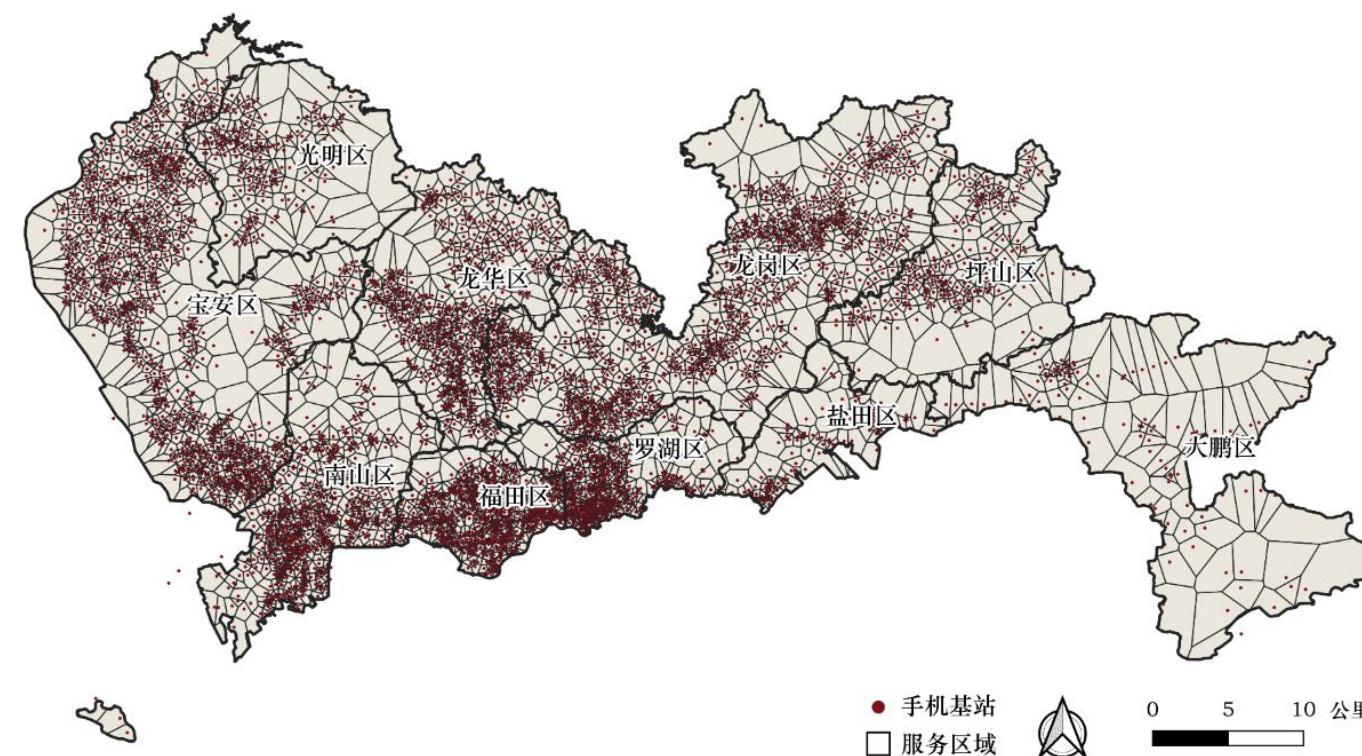
利用UMCNN耦合高分辨率遥感影像和多源空间数据提取特征，结合随机森林拟合模型逐像元进行模拟，得到深圳市**5米分辨率房价精细空间分布**。



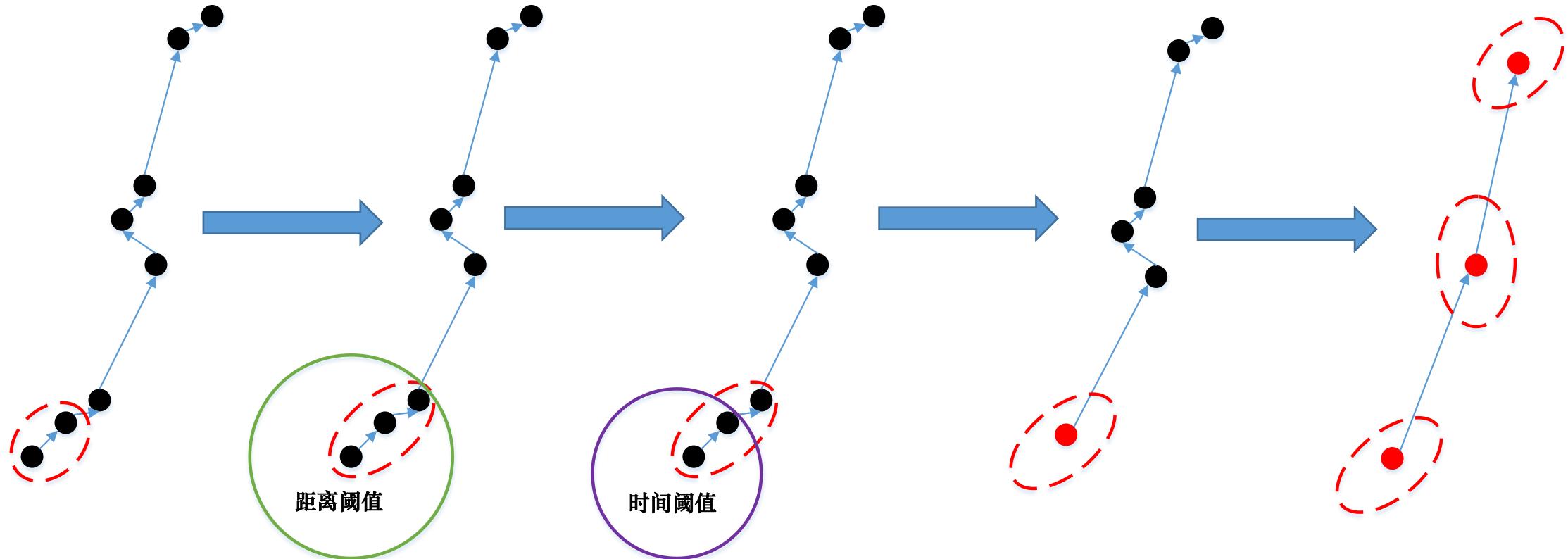
(A)南山区南头 (B)福田区公共行政中心 (C) 龙岗区中海地产的新建商品房

手机信令数据是本实验最重要的数据，本研究共记录深圳某1工作日的1,046.33万人的手机信令，占深圳市总人口80%以上。

用户id	记录次数	记录时刻	记录位置	记录时刻	...
f5d4a*****0205	22	20120323 00:01:32	114.18** 22.64**	20120323 01:28:39	...
0bdf1*****91cb	24	20120322 23:30:13	114.21** 22.60**	20120323 00:30:15	...
1db81*****adf3	23	20120322 23:25:37	114.21** 22.60**	20120323 00:09:29	...
4cdd3*****49a3	9	20120323 12:53:30	114.09** 22.73**	20120323 02:27:50	...
556df*****439c	22	20120322 23:23:27	114.21** 22.60**	20120323 00:26:04	...
...
5790f*****c970	14	20120323 10:55:40	114.35** 22.70**	20120323 11:26:35	...



■ 停驻点提取



$$l_1 \rightarrow l_2 \rightarrow \dots \rightarrow l_n \Rightarrow S = s_1 \xrightarrow{\Delta t_1} s_2 \xrightarrow{\Delta t_2} \dots \xrightarrow{\Delta t_{n-1}} s_n$$

■ 用户活动分类

- 若用户当日只有一个停留点，将该停留点记为用户居住点；



- 若用户当日停留点大于2个，则按照3:00-5:00的居住时间窗口，以及14:00-16:00的工作时间窗口进行居住点和工作的标记；

- 最后将没有标记的停留点全部归为生活娱乐中，形成完整的时空活动链。

手机位置点

停住点

1-19时

居住点

(a) 停住点个数为1

工作点

12-18时

1-6时

居住点

(b) 停住点个数为2

生活娱乐点

18-19时

12-17时

工作点

1-6时

居住点

(c) 停住点个数大于2

■ 用户活动指标

- 惯性矩:

$$R_g = \sqrt{\frac{\sum_{i=1}^n (\vec{l}_i - \vec{l}_c)^2}{n}}$$

n为用户信令位置的采集次数, i 为第几次的采集, \vec{l}_i 代表的是矢量位置, $\vec{l}_c = \sum \vec{l}_i / n$ 为用户轨迹位置的质心。 R_g 用于量化用户日常活动的空间的大小, 较大的 R_g 值通常表示较大的活动空间, 较小则代表用户的日常活动主要集中在较小的地理区域中。

- 活动位置数

$$A = |\text{set}(s_1, s_2, \dots, s_n)|$$

A较大表示用户当日活动较频繁, 出行频率较高。相反A越小则代表用户的活动地点较少, 可选择活动较为单一。

- 活动熵

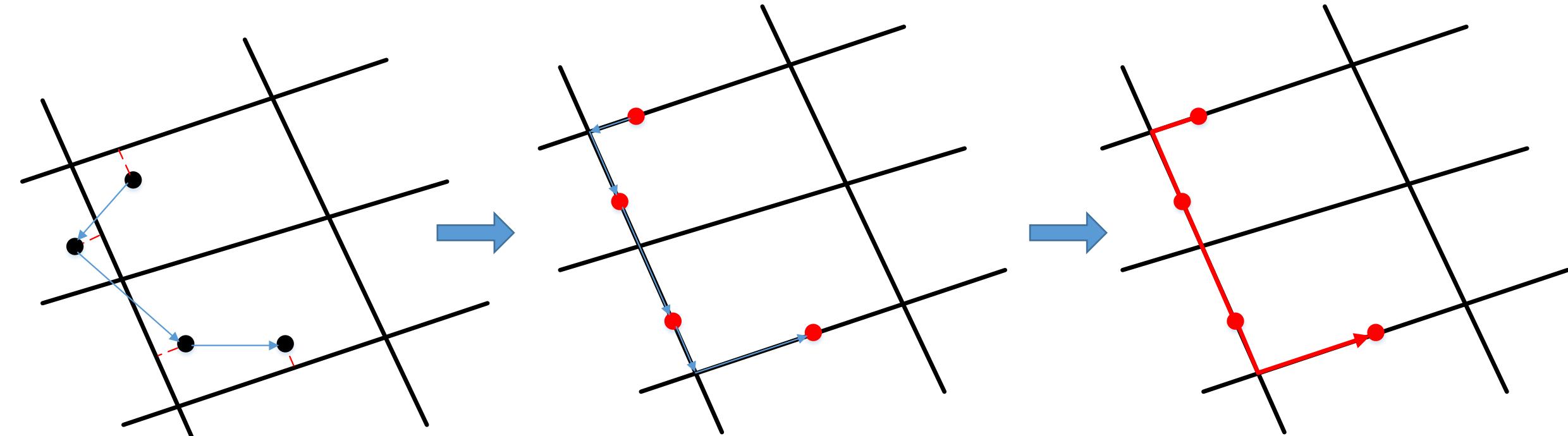
$$p_i = \frac{\sum_{s_j=s_i} \overline{dur_j}}{\sum_{j=1}^n \overline{dur_j}} \quad H_1 = - \sum_{i=1}^A p_i \log(p_i)$$

其中, $\sum p_i = 1$, p_i 表示用户在第*i*个停住点停留时间占总体时间的比例, 而不是停留次数, 此方法可有效降低个别用户突发情况所带来的误差。 H 越大表示用户的活动点的离散程度越高, 活动多样性越大。

03 | 手机信令大数据与用户画像



- 通勤距离、时间和速度



$$S = s_1 \xrightarrow{t_1} s_2 \xrightarrow{t_2} \dots \xrightarrow{t_{n-1}} s_n \Rightarrow M = m_1 \xrightarrow{d_1} m_2 \xrightarrow{d_2} \dots \xrightarrow{d_{n-1}} m_n$$

$$T_t = T_l - T_s$$

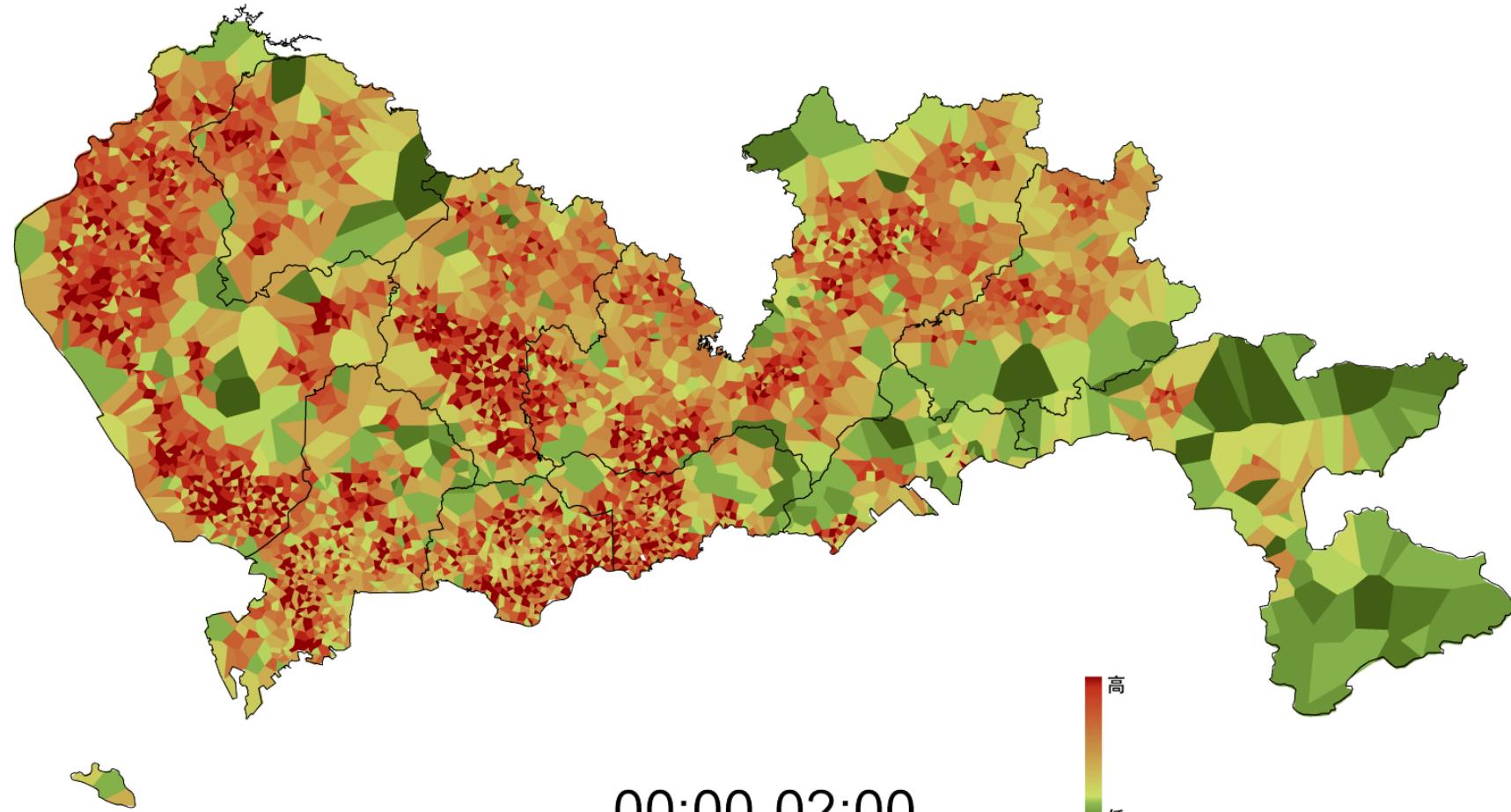
$$D = \sum_{i=1}^n d_i$$
$$v = \frac{D}{T_t}$$

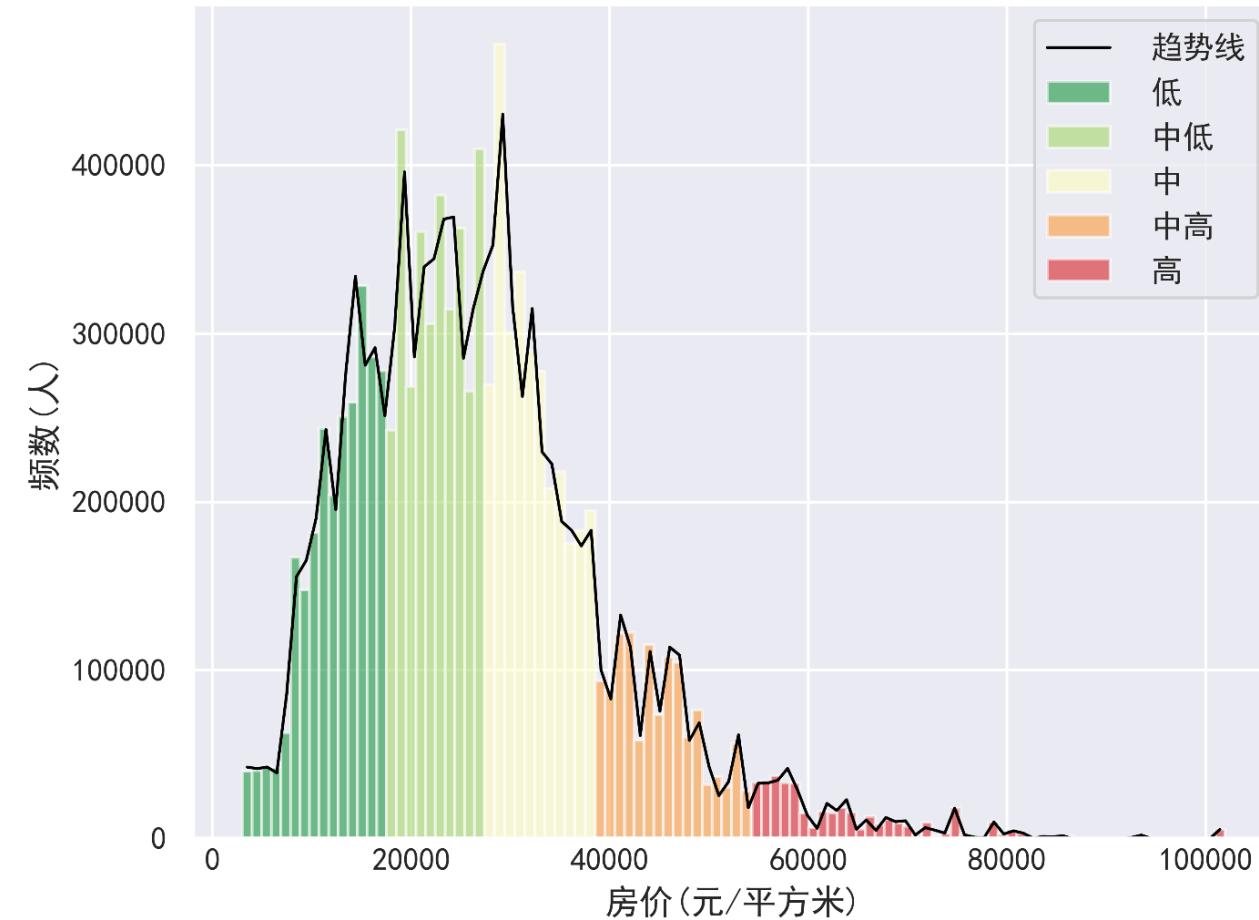
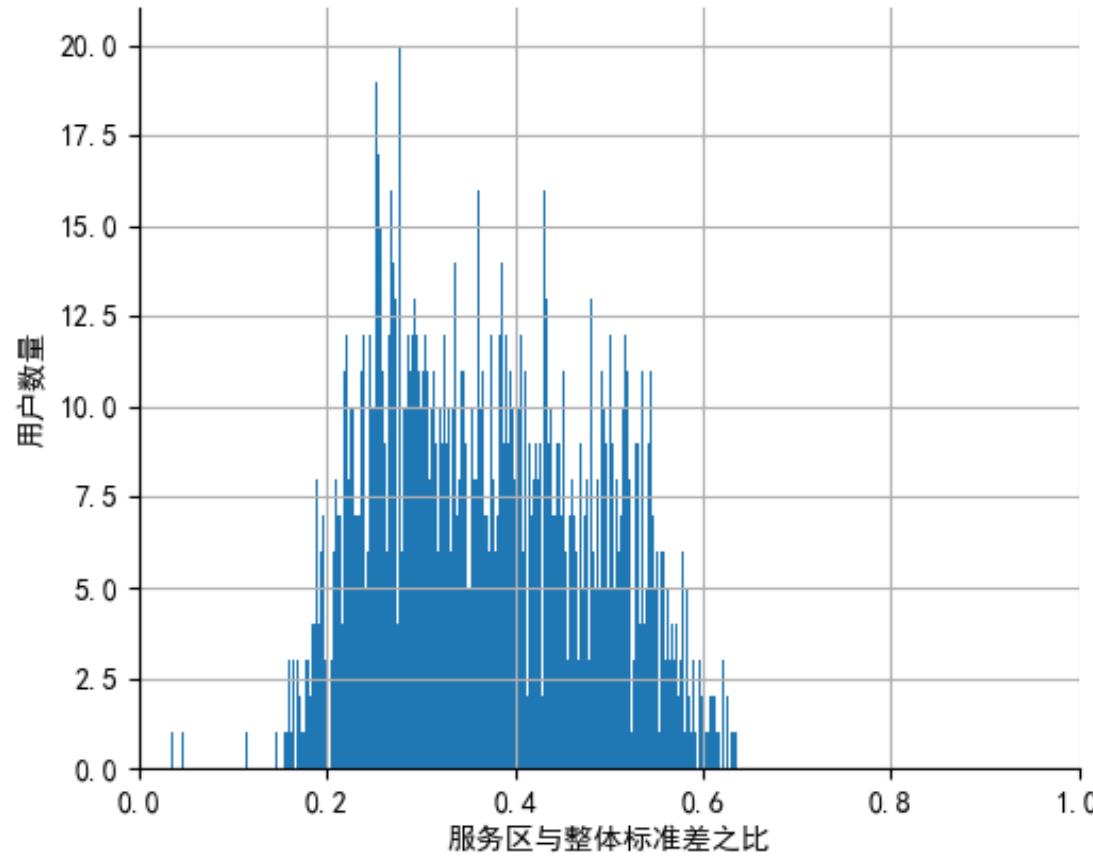
其中 T_l 表示用户原始轨迹序列时间之和， T_s 为停住点时间总和。 m_i 表示道路网匹配后的用户路网位置。 d_i 表示用户相邻匹配点 m_i 和 m_{i+1} 的实际路网距离。 D 为用户的通勤距离， V 表示用户通勤速度。

03 | 手机信令大数据与用户画像



■ 深圳手机用户人群变化





低: <17524.809 元/平方米 中低: 17524.809- 27380.861 元/平方米 中: 27380.861- 38556.434 元/平方米

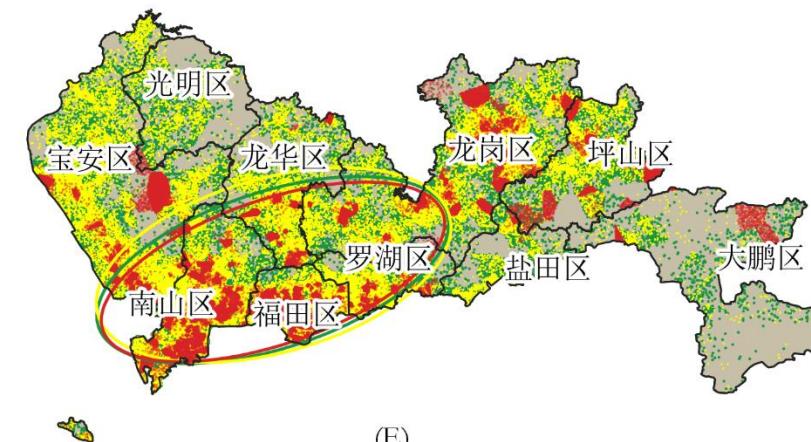
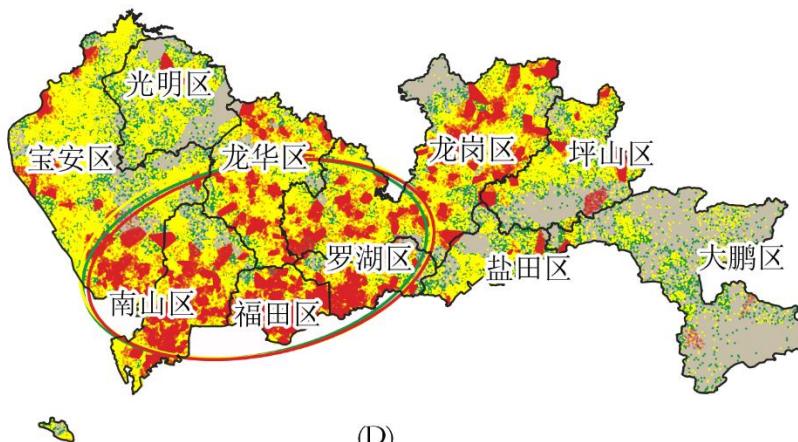
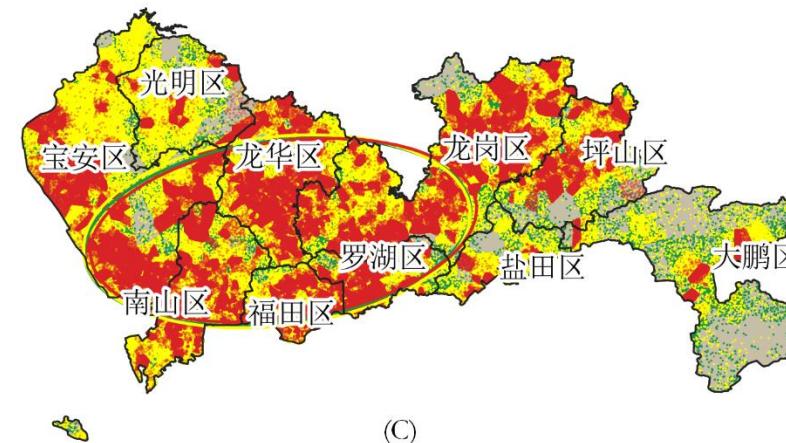
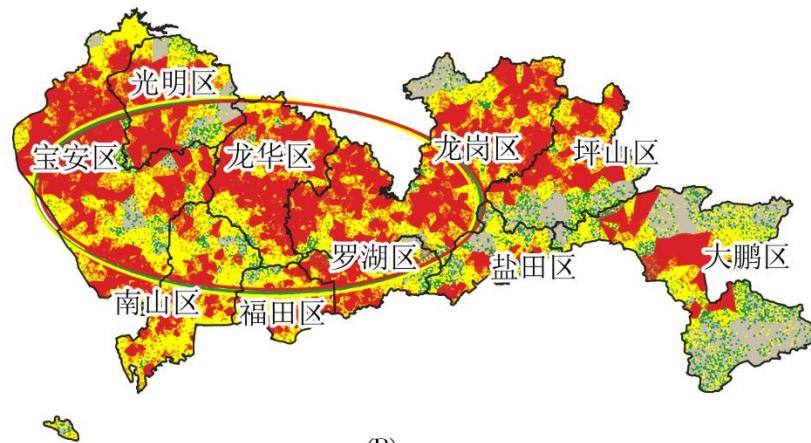
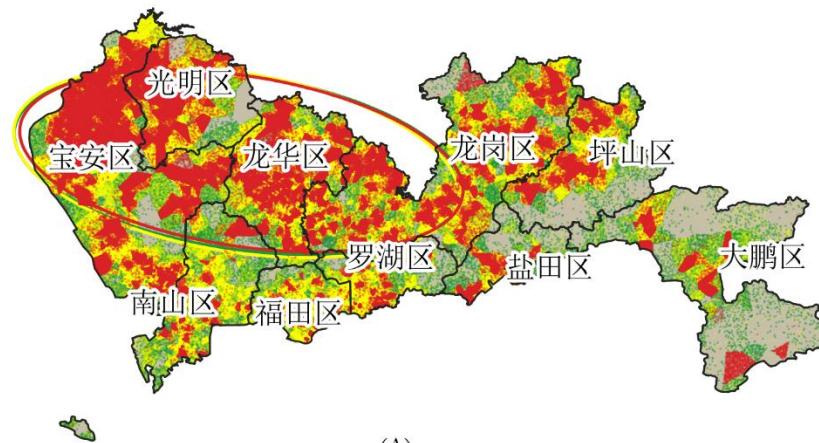
中高: 38556.434- 54232.699 元/平方米 高: > 54232.699元/平方米



nFID,user_id,SES,house_price,reside_location,reside_time,work_location,work_time,num_of_other_activities
 0, f5d4a2f2374a0205,L,13606.433,22.6432 114.1816,0,0,54702,1,2,10217,2508.136,0.43530090571
 4, 4454a3a855641621,M-H,43903.727,22.61 114.3564,10831,22.6158 114.4052,37722,3,5,24653,18
 13, 7391228469afbe5f,L,13778.667968799999,22.6406 114.06705,0,0,68456,1,2,5066,6008.41,0.25
 18, 43ff48c685f2bdef,L,15988.449218799999,22.659 114.0195,16733,22.61 114.3564,51218,1,3,95
 20, 6d1d75f74e4c6187,L,10980.521,22.6406 114.067,0,0,67216,0,1,0,2424.2745,0,0,0,0,0,0.3670,0,0
 25, 87303ccc3c3ee853,M,28756.49,22.5512 114.1427,21836,22.5625 114.0957,36317,1,3,12358,26
 27, 4d1daa72edf89f9a,L,8632.598,22.63 114.1034,34266,22.66935 114.0488,37200,1,3,13525,2849.
 29, 194e8229ab93130a,M,36133.707,22.4842 113.8851,0,0,24454,0,1,0,12250.4195,0,0,0,0,0,0.2185
 38, 233f624adf4c9142,M,31764.736328099996,22.5759 114.2547,28165,22.6795 114.1544000000
 39, 9ba2222c0d386bff,M-L,21364.939,22.6983 113.9578,51265,22.8391 113.8687,4852,3,5,24260,1

用户id	经济等级	房价 (元/平米)	家庭位置	活动点数量	娱乐点	回转半径 (米)	...
f5d4a*****0205	L	13606.433	114.18** 22.64**	2	0	10217	...
4454a*****1621	M-H	43903.727	114.35** 22.61**	4	2	24653	...
43ff42*****bdef	L	15988.449	114.01** 22.65**	3	1	5066	...
...

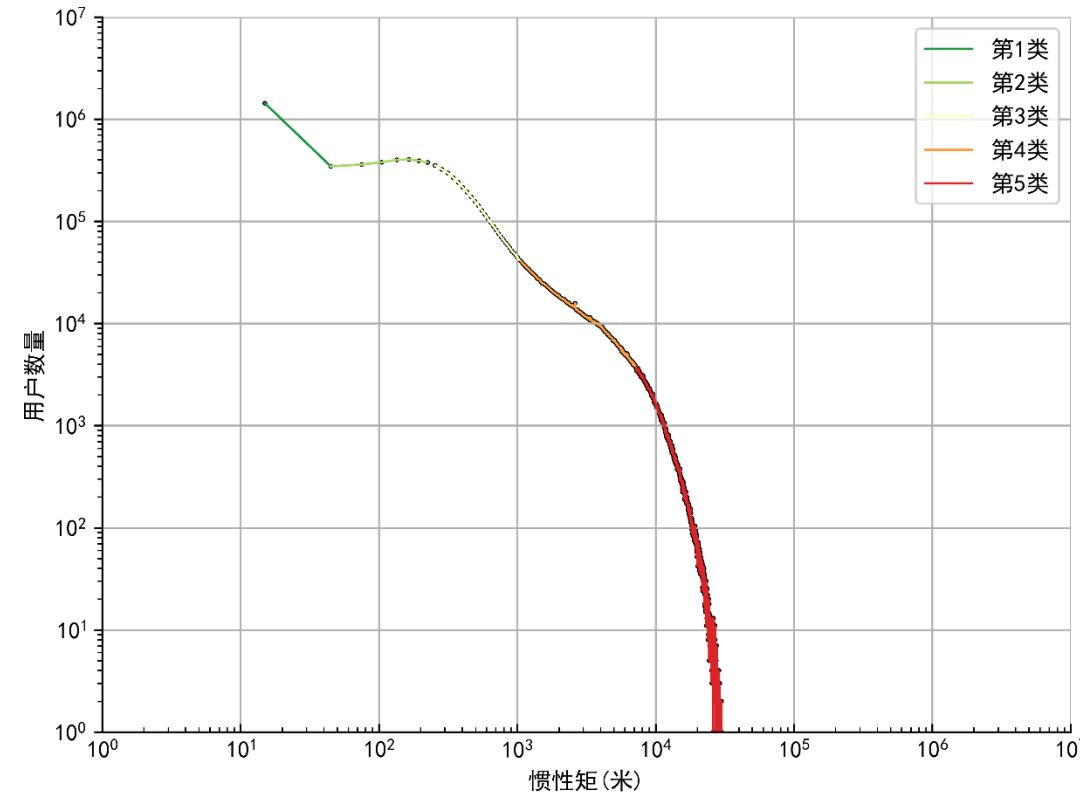
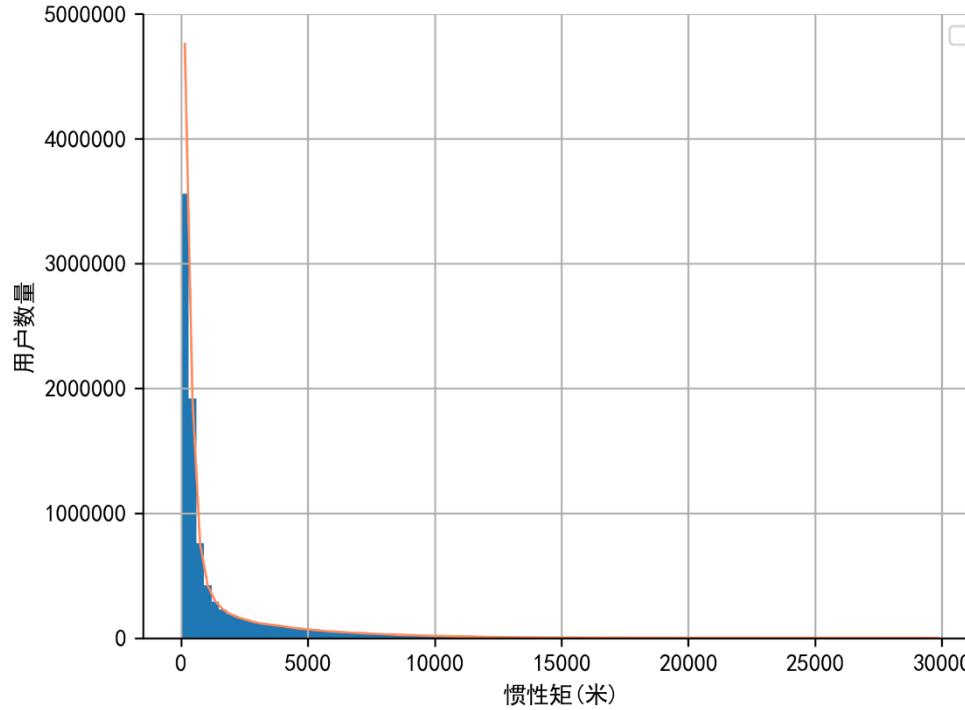
03 | 手机信令大数据与用户画像



■ 居住点
■ 工作点
■ 生活娱乐

0 5 10 公里

深圳市不同经济水平人群活动分布与深圳市各行政区域经济发展水平相似，整体呈现“南高北低，西高东低”的格局。且低经济水平人群和高经济水平人群分布更集中，趋势更明显。

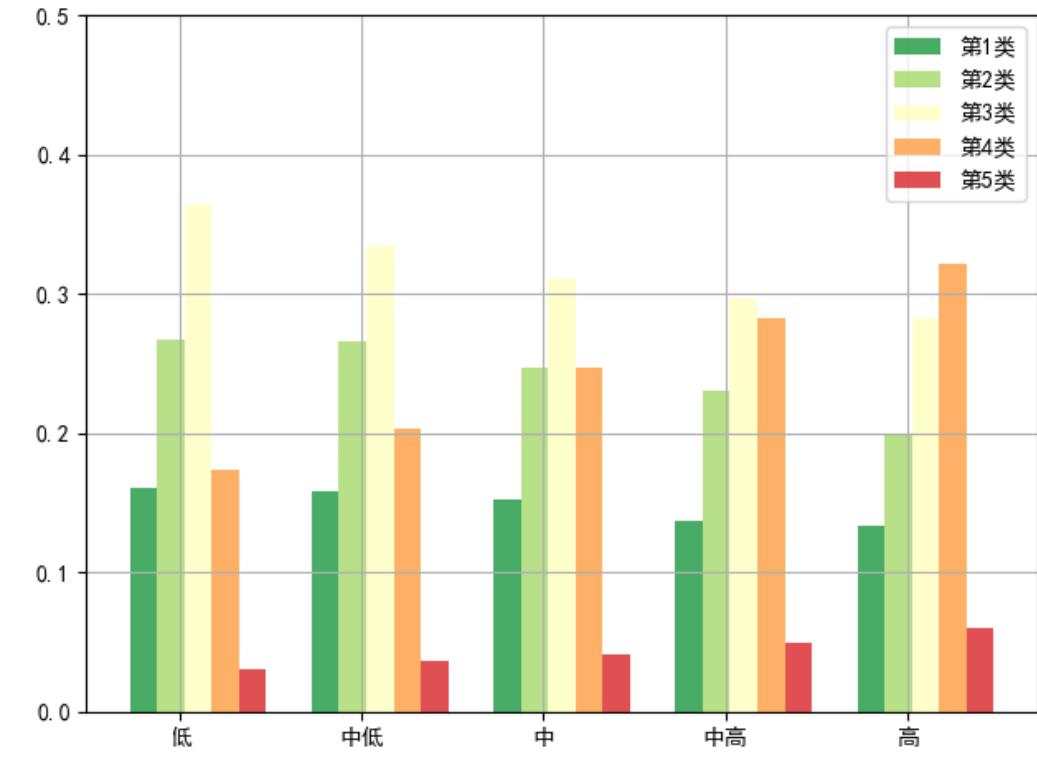
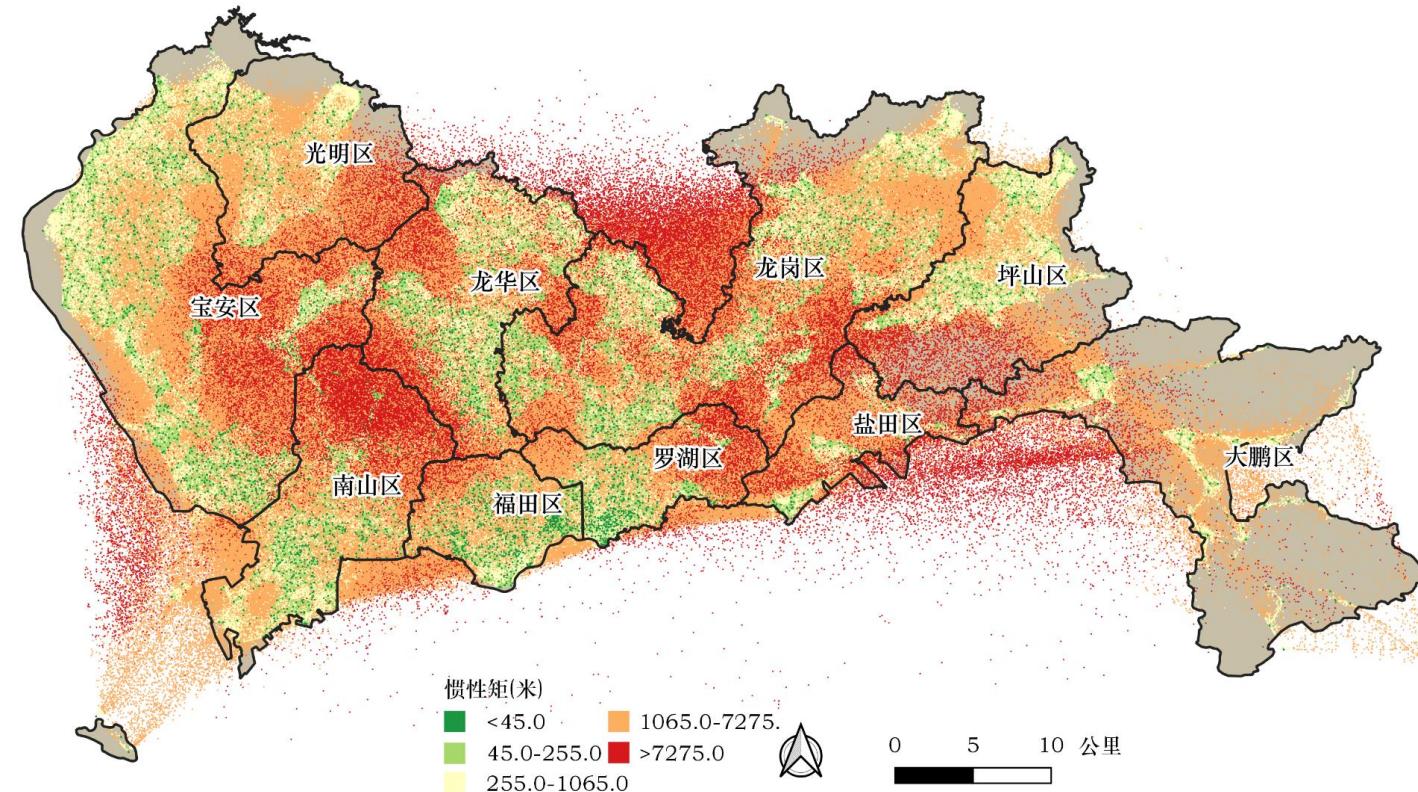


惯性矩密度呈现明显的**幂律分布**。

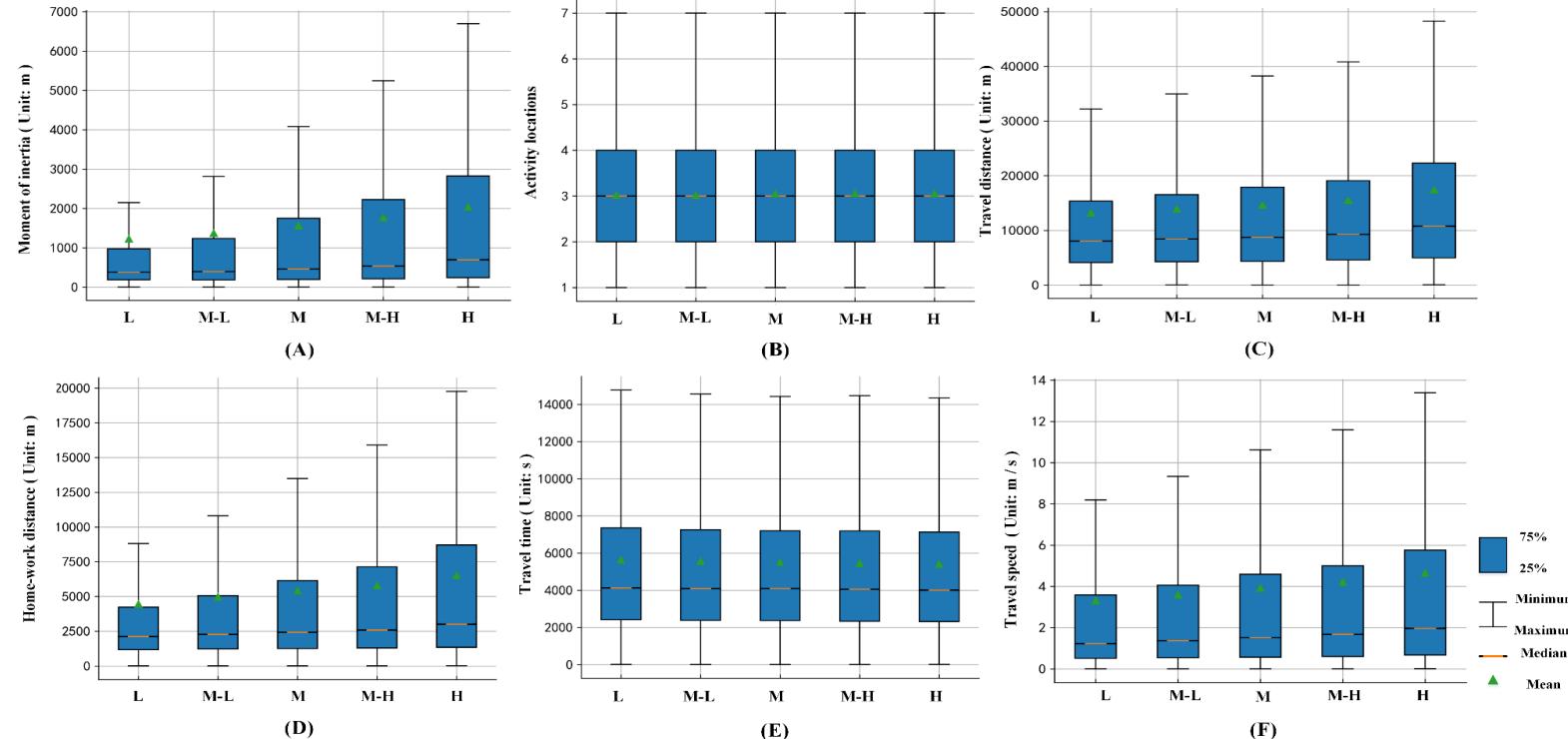
本研究根据用户惯性矩对数分布中的拐点将其分为5类：

第一类：0-45米；第二类：46-255米；第三类：255-1,065米；第四类：1,066-7,275米；第五类：大于7,276米。

03 | 手机信令大数据与用户画像



随着用户经济水平的提高，用户日常活动的范围不断扩大，跨区活动占比越来越高。第三类和第四类惯性矩随经济水平的上升，人群比例有明显的变化。



经济水平	居家时间	工作时间	生活娱乐时间	活动点数量	惯性矩	活动熵	出行时间	出行距离	职住距离	出行速度
低	1.000	1.000	0.942	0.987	0.625	0.982	0.934	0.831	0.674	0.581
中低	0.996	0.994	0.939	0.986	0.692	0.982	0.957	0.863	0.720	0.839
中	0.988	0.986	0.966	0.999	0.780	0.999	0.984	0.898	0.884	0.896
中高	0.965	0.955	0.989	1.000	0.778	1.000	1.000	0.924	0.943	0.934
高	0.943	0.945	1.000	0.999	1.000	0.998	0.997	1.000	1.000	1.000

经济水平的高低并不会影响人们工作日的日常活动点数量以及活动时间安排。但与活动位置数和活动熵不同，不同经济水平人群之间惯性矩存在较大的差异。经济水平状况和通勤距离、职住距离、通勤速度存在正向相关性。经济水平较高人群通勤速度呈现“南高北低”的分布模式。而较低经济水平人群则相反，通勤速度分布为“南低北高”。经济水平和人群活动模式存在较强的相关性，不同经济水平人群遵循一定的活动模式。



主要内容



- 1 手机信令大数据简介
- 2 手机信令大数据处理技术
- 3 手机信令大数据与用户画像
- 4 手机信令大数据与用户位置预测
- 5 手机信令大数据与城市活力
- 6 手机信令大数据与社会隔离

经济水平和人群活动模式存在较强的相关性，不同经济水平人群遵循一定的活动模式。

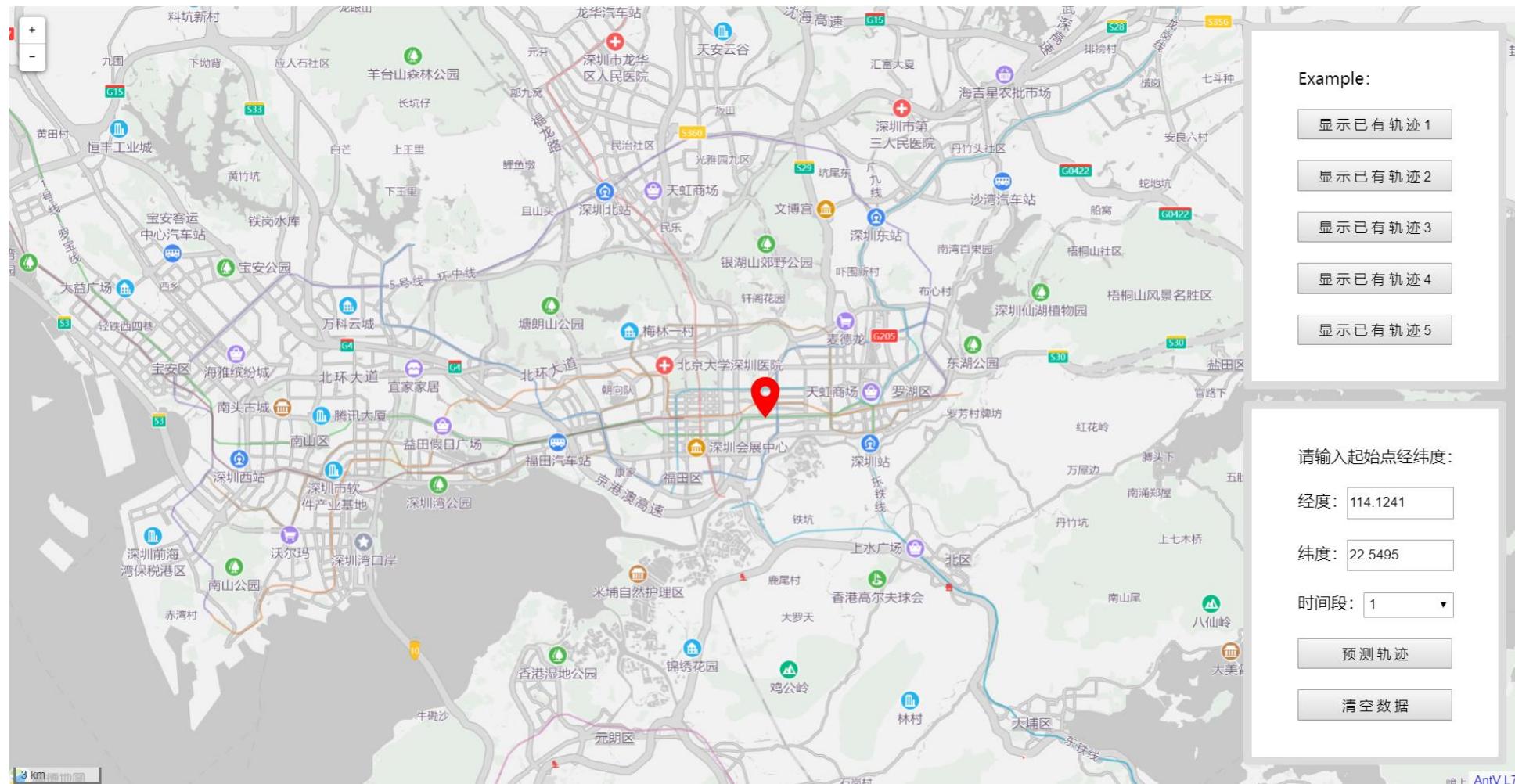


如何进一步挖掘个人的时空**移动模式**，并基于个人移动模式实现用户**位置的预测**？

04 手机信令大数据与用户位置预测



■ 可视化系统界面



04 | 手机信令大数据与用户位置预测



■ 预测功能

在左下角的信息框内，输入起始点的经纬度信息和开始时间段，点击预测轨迹按钮即可进行预测。

以输入 经度：114.048585，纬度：22.530502，时间：4（即12:00-16:00）为例。

左图为网页输入内容，右图为后台计算结果：

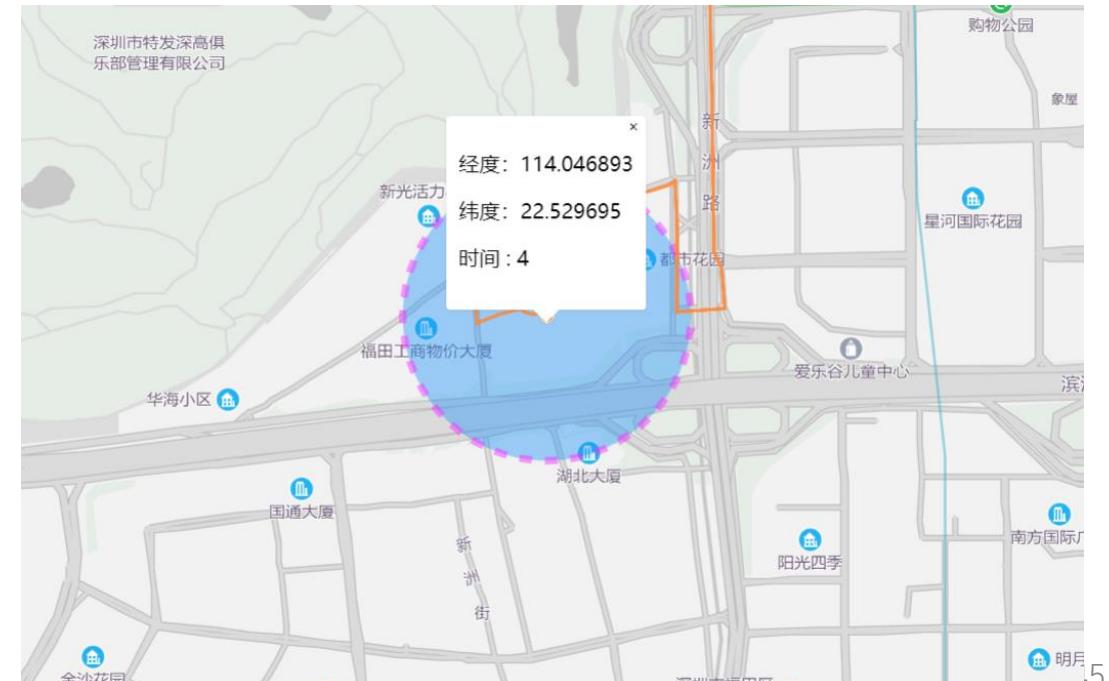
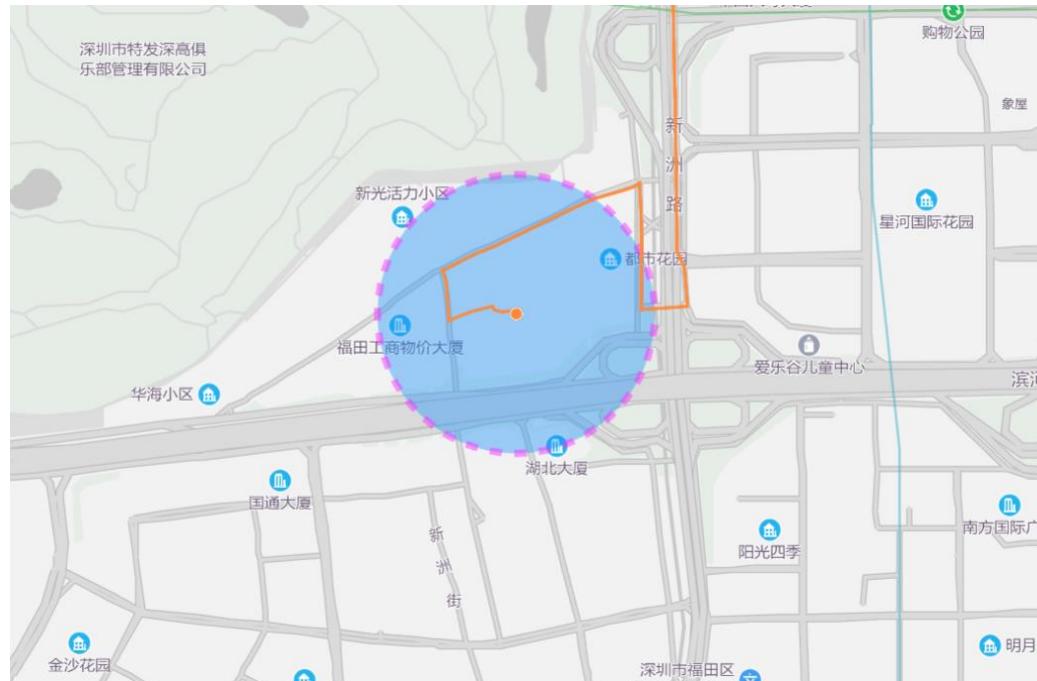
```
114.048585  
22.530502  
4  
3612  
map has finished...  
import the model successfully !  
result:  
features: 3612 4 predict loc: 2353  
features: 2353 5 predict loc: 248  
features: 248 6 predict loc: 248  
features: 248 1 predict loc: 114  
features: 114 2 predict loc: 114
```

04 | 手机信令大数据与用户位置预测



■ 预测结果展示

在下图中，中心的橙色原点为输入的起始位置点，淡蓝色、半径为250m的圆形区域为居民在该点处的活动范围，橙色线段为居民的移动轨迹。



04 | 手机信令大数据与用户位置预测



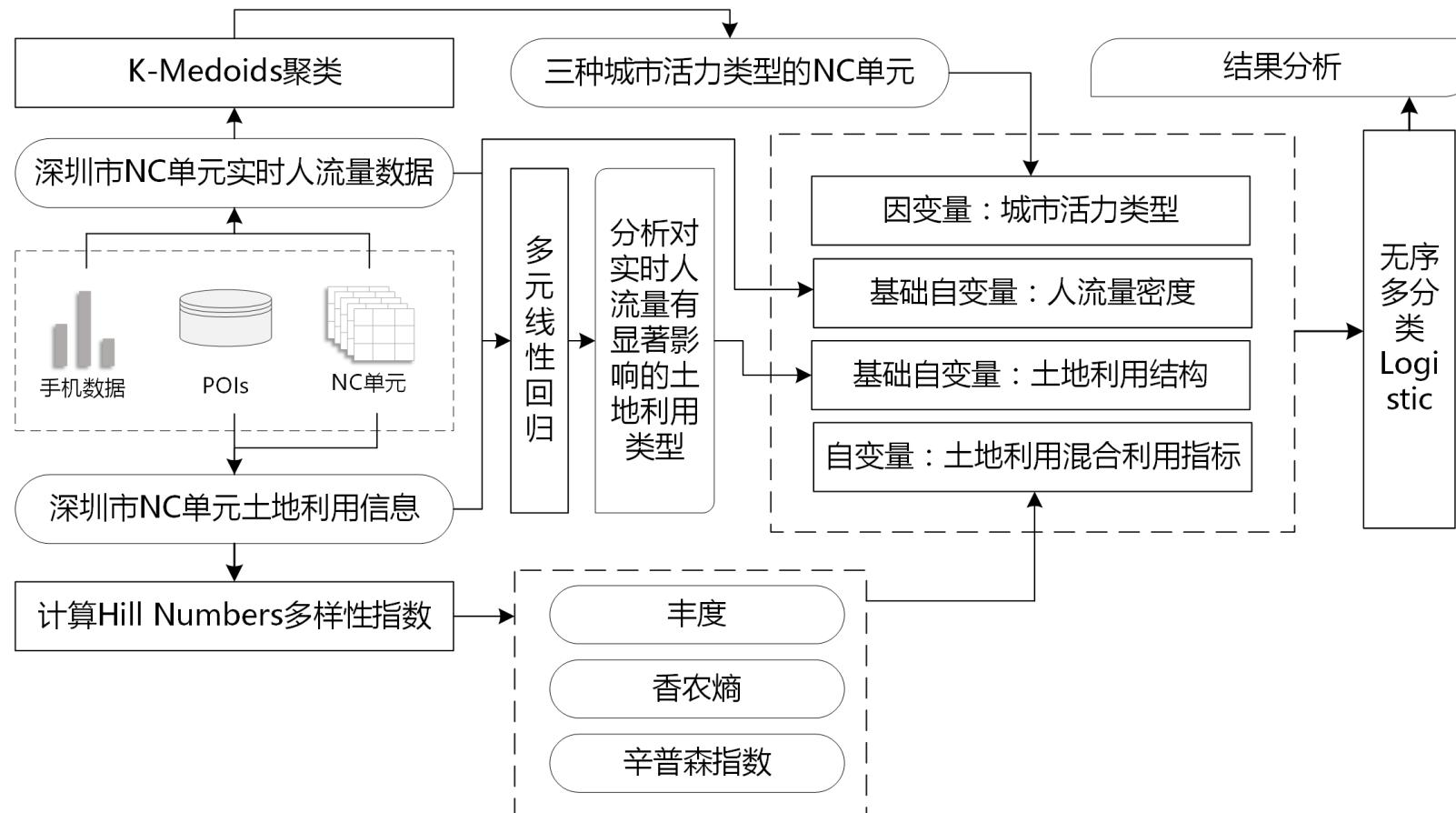


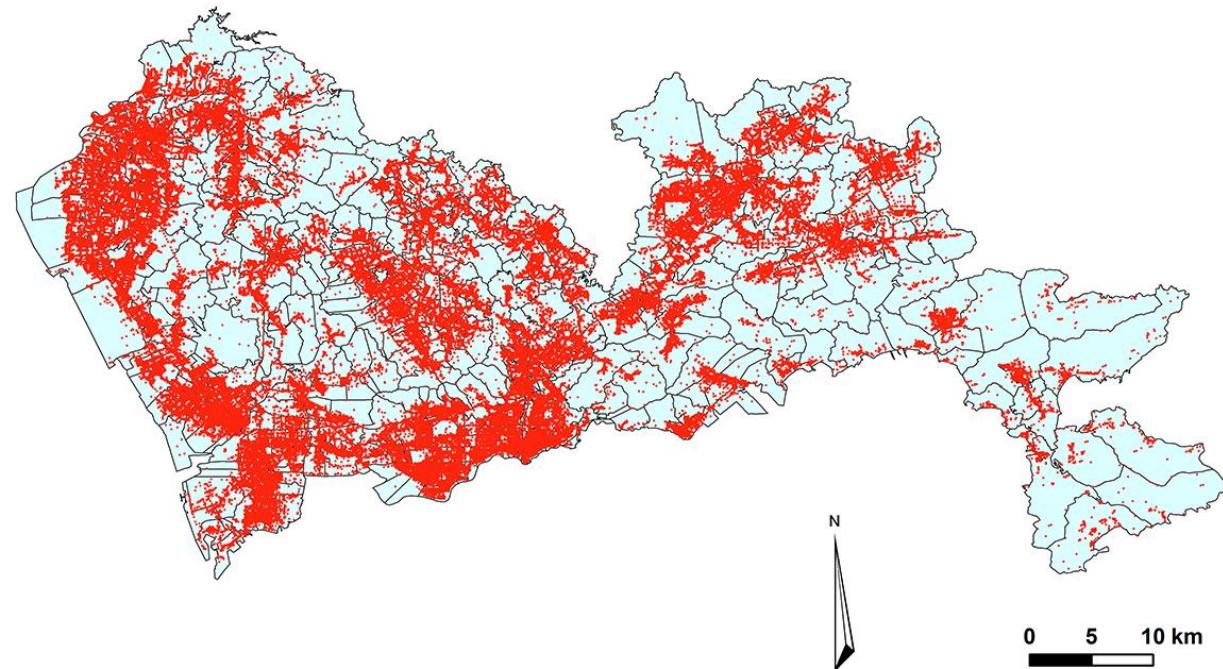
主要内容



- 1 手机信令大数据简介
- 2 手机信令大数据处理技术
- 3 手机信令大数据与用户画像
- 4 手机信令大数据与用户位置预测
- 5 手机信令大数据与城市活力
- 6 手机信令大数据与社会隔离

怎样去定量的模拟/分析细尺度空间，城市内部的人群**活动模式**和**城市发展状况**？



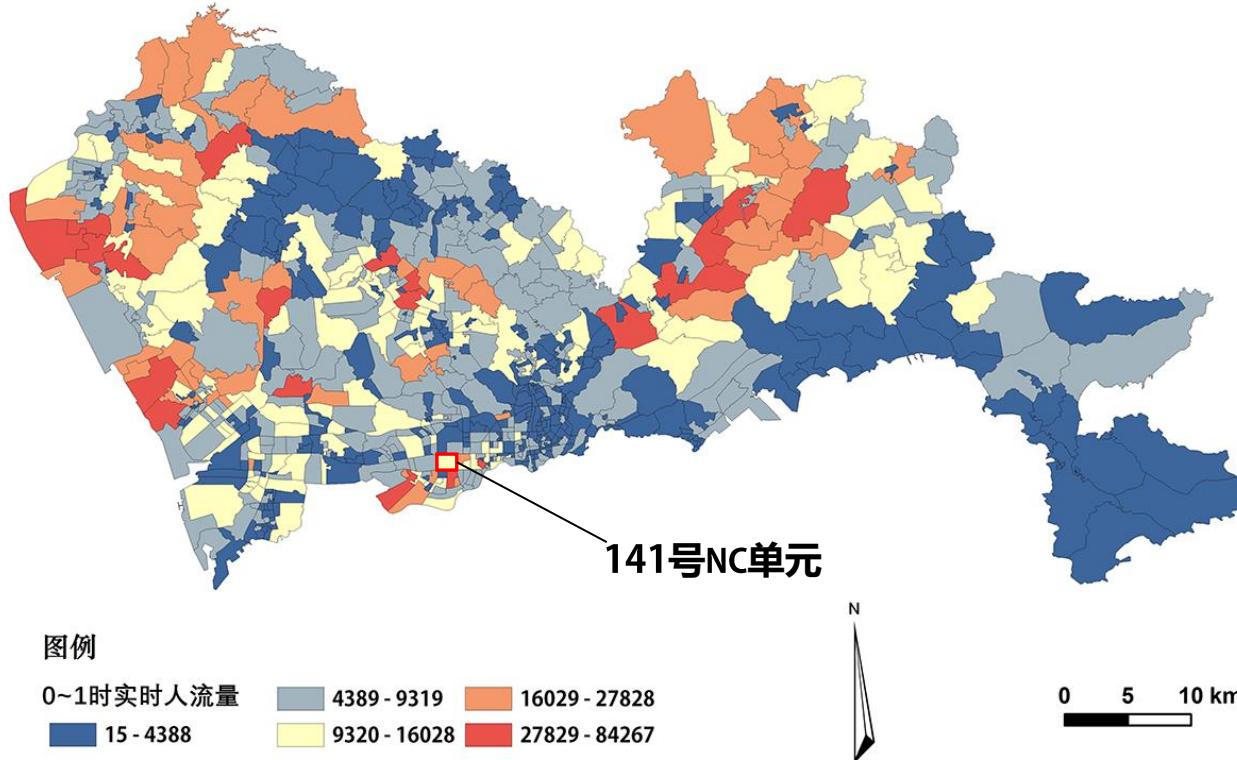


深圳市poi数据

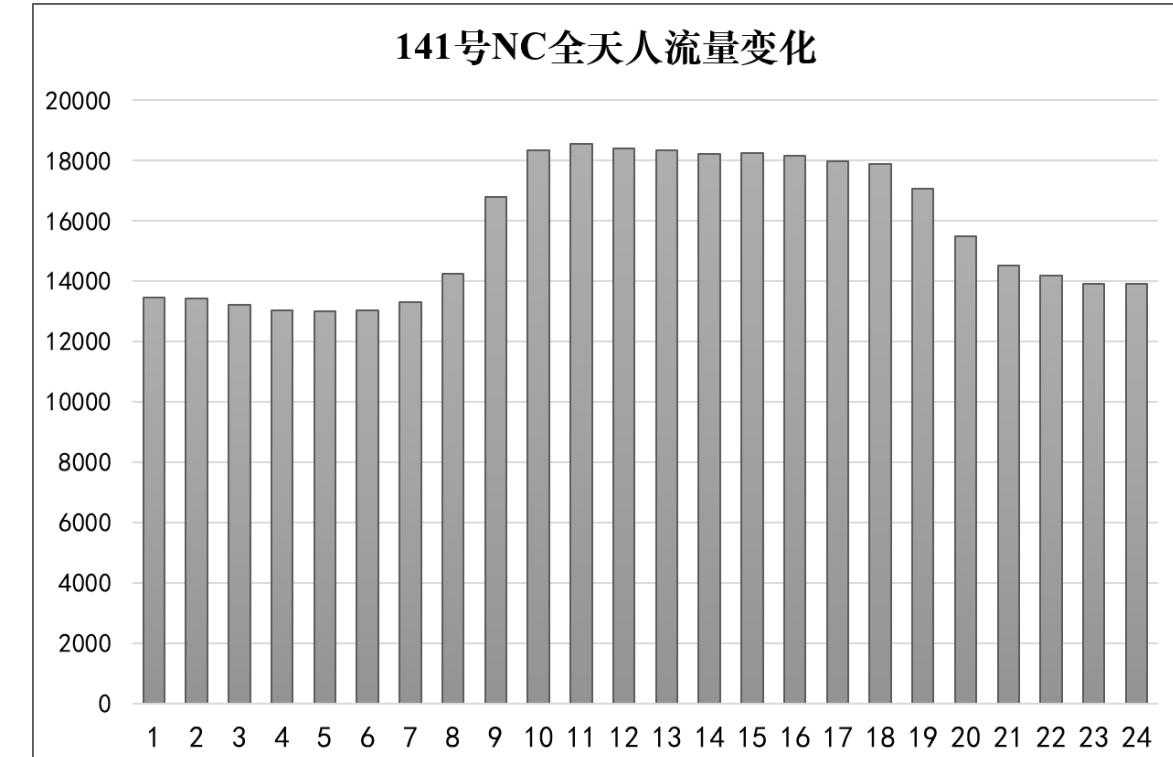


深圳市NC数据

05 | 手机信令大数据与城市活力



2016年3月中国联通信号基站记录的手机定位数据，每间隔一个小时有一条采样记录，取31天平均数据，将其分配到各个NC单元中。

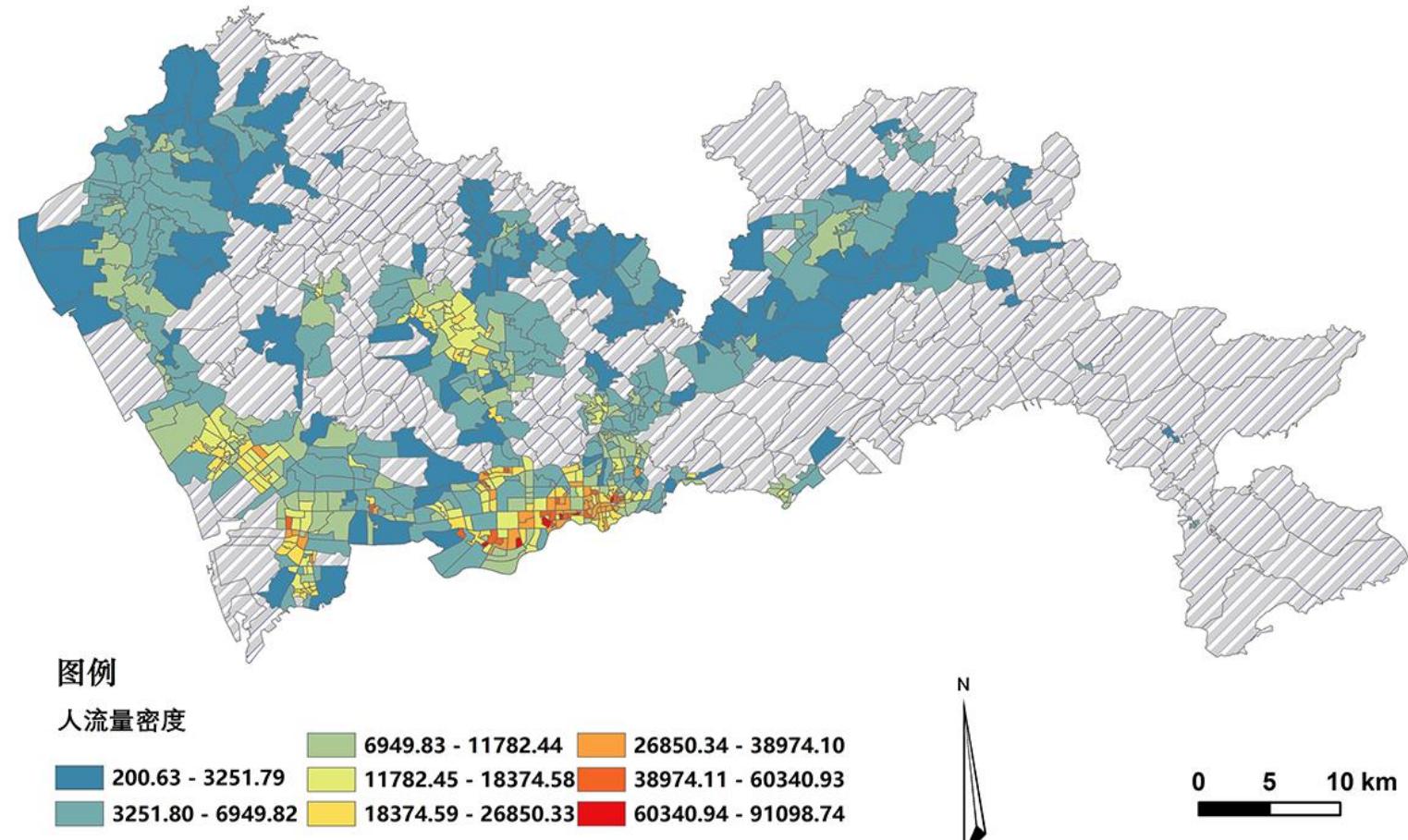


横坐标为时间，纵坐标表示实时人流量，每个NC单元都有12条记录。

05 | 手机信令大数据与城市活力

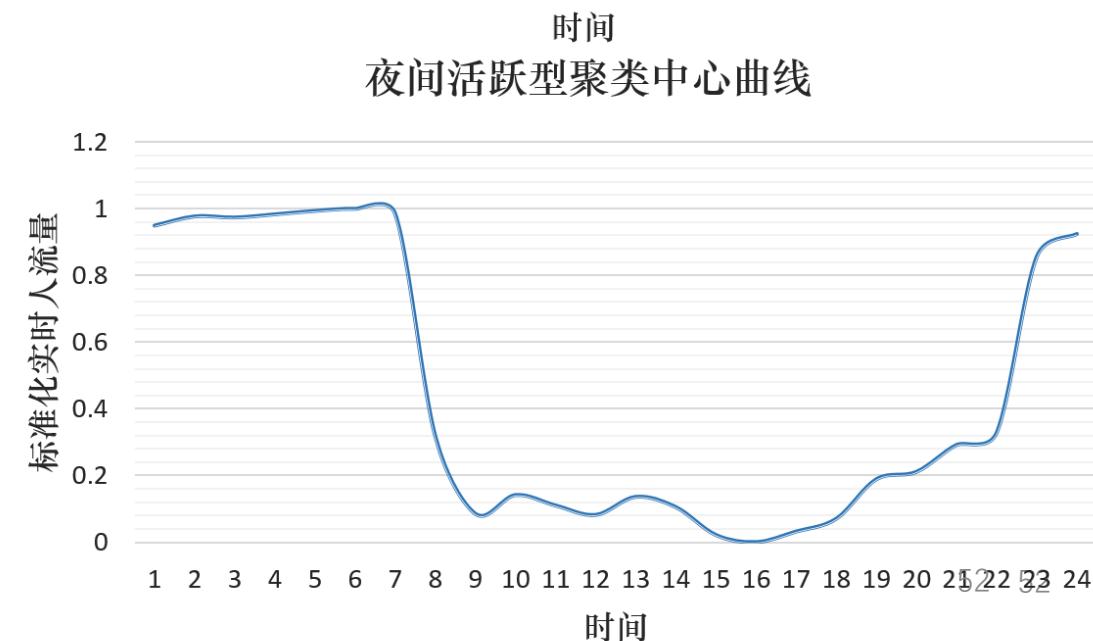
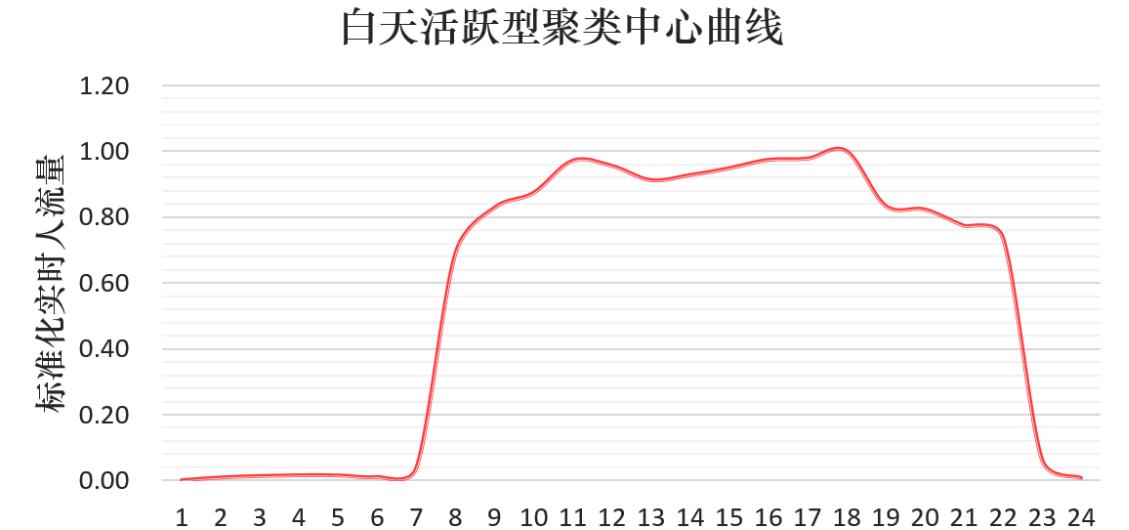


- 工信部手机普及率数据，广东已经达到了为141.2(部/百人)
- 将各NC单元凌晨2: 00~5: 00的平均人流量密度，作为各NC单元的人口密度。
- 采用《世界发展报告2009》中的定义方法，将人口密度小于200人/平方公里的区域作掩膜处理。



2: 00~5: 00平均人流量密度图，单位：人/平方公里

- 不同功能区的人流总是集中在特定的时间段，城市功能分离会对城市发展造成负面影响。
- 使用K-Medoids聚类算法对各个NC单元的全天人流量曲线进行聚类。





- Hill numbers多样性指数：丰度(Richness)、香农熵(Shannon Entropy)、辛普森指数(Simpson index)

$$^qD = \left(\sum_{i=1}^s P_i^q \right)^{1/(1-q)}$$

- 丰度——土地利用类型丰富度： $^0D = \sum_{i=1}^s P_i^0$
- 香农熵——无序性： $^1D = \exp\left(-\sum_{i=1}^s P_i \ln(P_i)\right)$
- 辛普森指数——聚集度： $^2D = 1 / \left(\sum_{i=1}^s P_i^2 \right)$



- 由每个NC单元内不同类别POI的数量表示土地利用结构。
- 建立24个多元线性回归模型（即一个小时建立一次模型），每个NC单元作为一个样本进行多元线性回归。
- 模型变量：
 - 因变量：各NC单元每小时的人流量，以 $y_0, y_1, y_2, y_3, \dots, y_{23}$ 表示。
 - 自变量：各NC单元内12类POI的个数，分别用 $x_1, x_2, x_3, x_4, \dots, x_{12}$ 表示。
- 第*i*个时间段土地利用结构与区域实时人流量的关系模型可以表示为：

$$\begin{aligned}y_i = & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 \\& + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11} + \beta_{12} x_{12}\end{aligned}$$

- 因变量：城市活力类型—全天活跃型（数值：0），夜间活跃型（数值：1），
白天活跃型（数值：2），发生概率分别为 π_0, π_1, π_2 表示。
- 自变量：
 - 手机信令数据反演的NC单元人口密度： $x_{population}$
 - 土地利用结构（线性回归模型的结果）： $x_{finance}, x_{residence}, x_{traffic}$
 - 土地混合利用指标： $x_{richness}, x_{entropy}, x_{simpson}$
- 构建四个模型：

变量	模型1	模型2	模型3	模型4
NC单元人口密度	✓	✓	✓	✓
商务金融用地比例	✓	✓	✓	✓
住宅用地比例	✓	✓	✓	✓
交通用地比例	✓	✓	✓	✓
丰度	✓	N\A	N\A	✓
香农熵	N\A	✓	N\A	✓
辛普森指数	N\A	N\A	✓	✓



- 四个模型均以夜间活跃型为参照对象，以模型4为例，无序多类分类Logistic回归的数学表达式为：

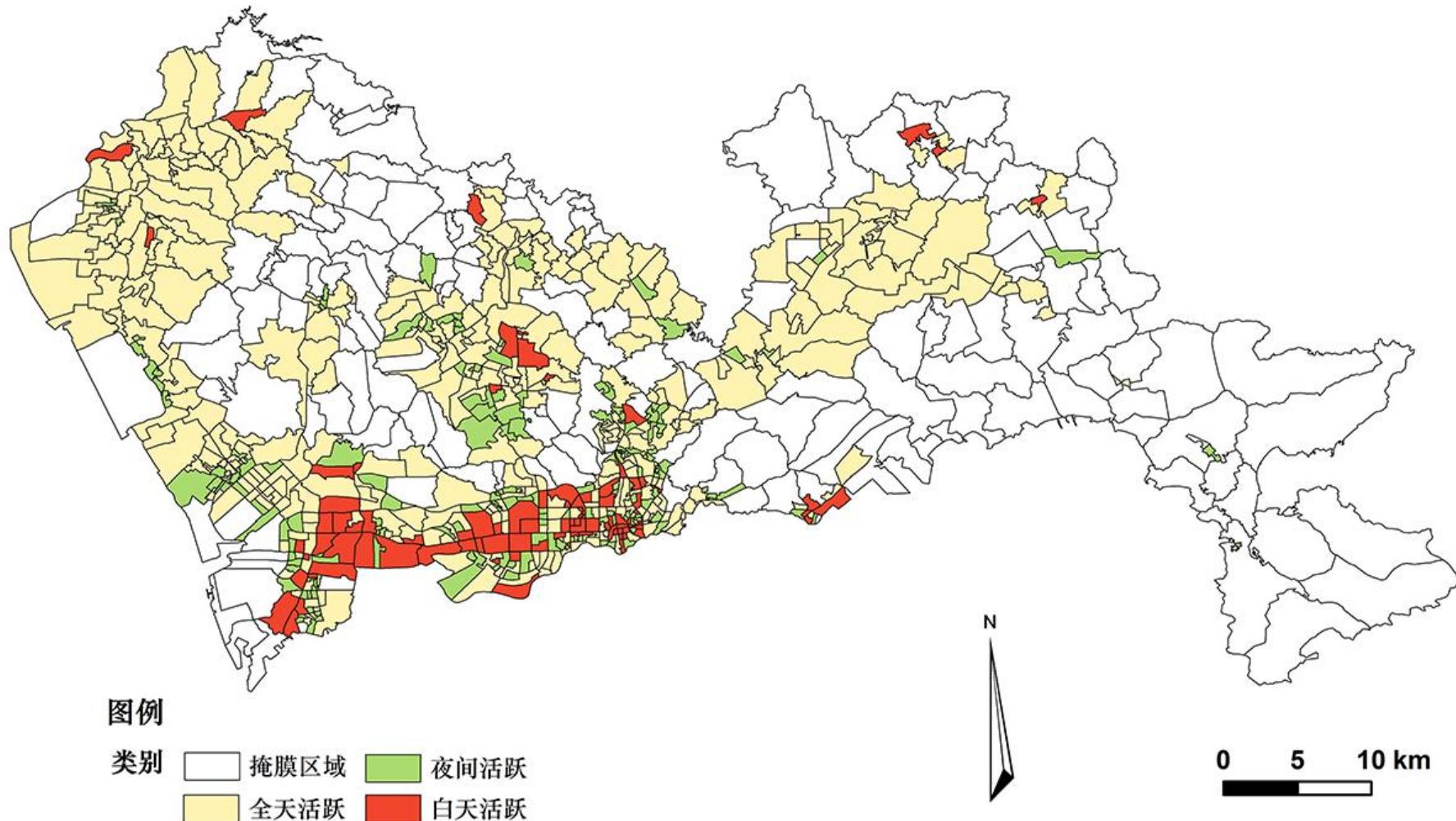
$$\text{Logit} \frac{\pi_0}{\pi_1} = \beta_{00} + \beta_{01}x_{population} + \beta_{02}x_{finance} + \beta_{03}x_{residence} + \beta_{04}x_{traffic} + \beta_{05}x_{richness}$$

$$+ \beta_{06}x_{entropy} + \beta_{07}x_{simpson}$$

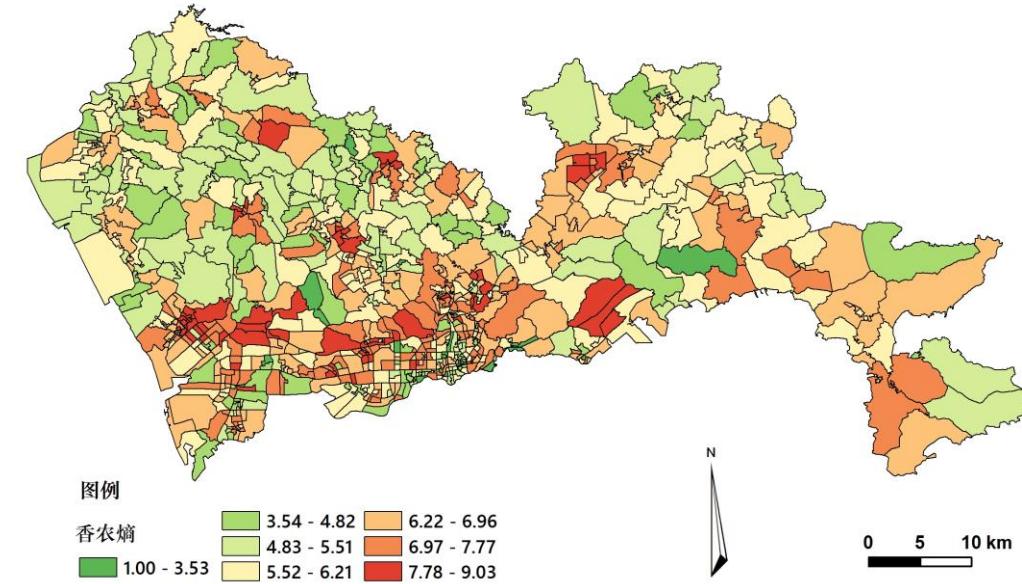
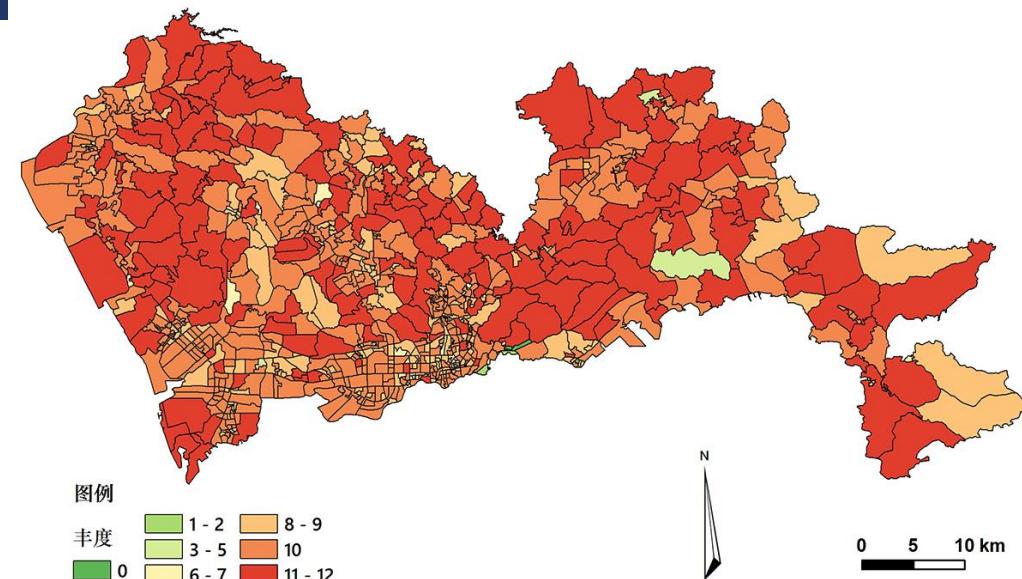
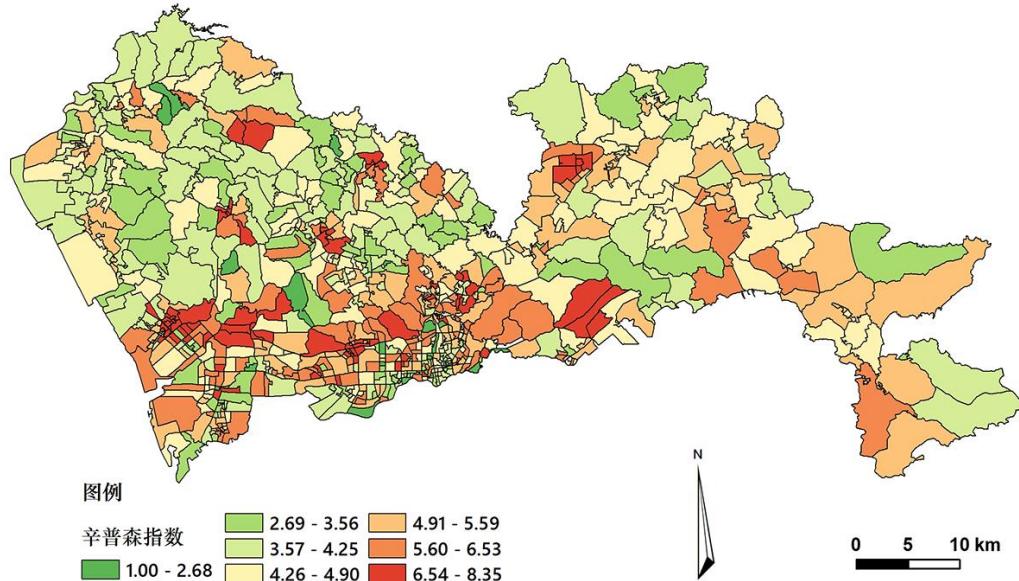
$$\text{Logit} \frac{\pi_2}{\pi_1} = \beta_{10} + \beta_{11}x_{population} + \beta_{12}x_{finance} + \beta_{13}x_{residence} + \beta_{14}x_{traffic} + \beta_{15}x_{richness}$$

$$+ \beta_{06}x_{entropy} + \beta_{07}x_{simpson}$$

- 深圳市各个NC单元被分为，分为白天活跃型、夜间活跃型和全天活跃型，三种类型作为Logistic回归的因变量。



- 深圳市各基本社区单元 (NC) 丰度、香农熵、辛普森指数计算结果图。



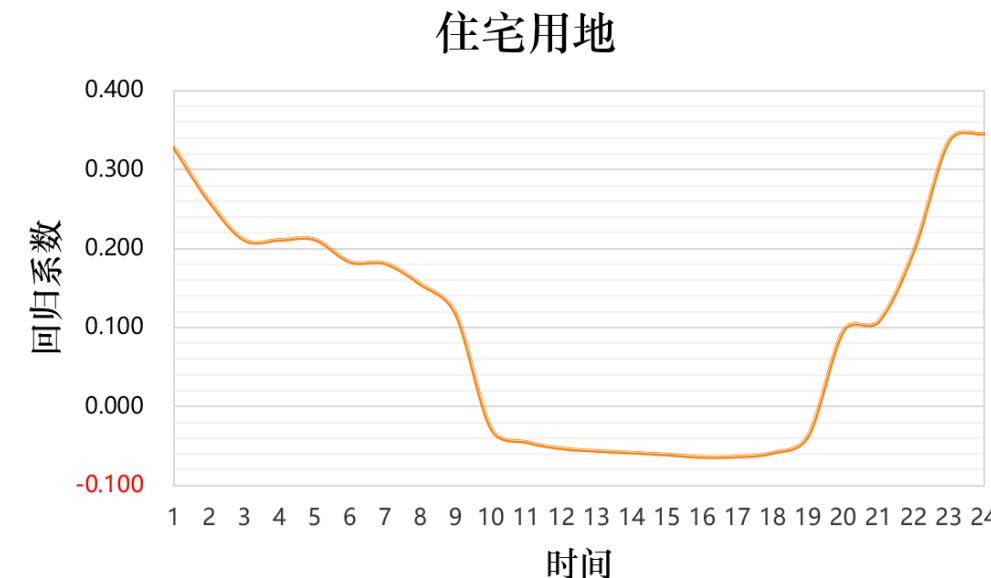
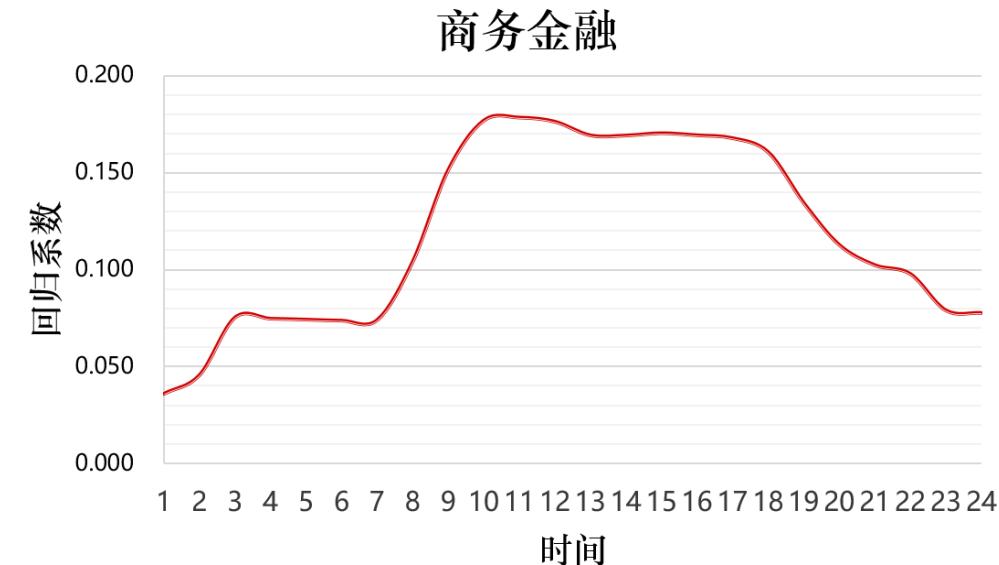


- 观察分析各种土地利用类型的回归系数及其时序分布特征。
- 选出对实时人流量有显著影响的类型作为后续Logistic回归的变量。

时间	截距	批发零售	住宿餐饮	商务金融	其他商服
0 ~ 1	1426.22	0.13	0.21	0.04	0.07
1 ~ 2	1437.03	0.13	0.21	0.05	0.07
2 ~ 3	1442.33	0.13	0.20	0.08	0.07
3 ~ 4	1446.35	0.13	0.21	0.07	0.07
4 ~ 5	1444.73	0.13	0.21	0.07	0.07
5 ~ 6	1437.21	0.13	0.21	0.07	0.07
6 ~ 7	1424.31	0.12	0.22	0.07	0.07
7 ~ 8	1186.59	0.11	0.20	0.10	0.08
8 ~ 9	1004.37	0.09	0.18	0.15	0.08
9 ~ 10	852.49	0.08	0.26	0.18	0.08
10 ~ 11	755.57	0.08	0.26	0.18	0.08
11 ~ 12	711.84	0.08	0.26	0.18	0.08
12 ~ 13	700.82	0.09	0.26	0.17	0.08
13 ~ 14	695.25	0.08	0.26	0.17	0.09
14 ~ 15	671.52	0.08	0.26	0.17	0.09
15 ~ 16	662.64	0.08	0.26	0.17	0.09
16 ~ 17	686.01	0.08	0.26	0.17	0.09
17 ~ 18	715.73	0.09	0.26	0.16	0.09
18 ~ 19	842.53	0.10	0.26	0.13	0.09
19 ~ 20	946.54	0.11	0.25	0.11	0.08
20 ~ 21	1033.70	0.12	0.25	0.10	0.08
21 ~ 22	1102.73	0.12	0.24	0.10	0.08
22 ~ 23	1371.47	0.13	0.23	0.08	0.07
23 ~ 0	1415.49	0.13	0.22	0.08	0.07

模型	adjust R^2	F值	显著性(sig)
0 ~ 1	0.6287	111.0696	<0.0000
1 ~ 2	0.6269	110.2173	<0.0000
2 ~ 3	0.6259	109.7527	<0.0000
3 ~ 4	0.6251	109.367	<0.0000
4 ~ 5	0.6248	109.2399	<0.0000
5 ~ 6	0.6255	109.5451	<0.0000
6 ~ 7	0.6272	110.3788	<0.0000
7 ~ 8	0.6393	116.1942	<0.0000
8 ~ 9	0.6381	115.6167	<0.0000
9 ~ 10	0.6314	112.3408	<0.0000
10 ~ 11	0.6277	110.6125	<0.0000
11 ~ 12	0.624	108.8524	<0.0000
12 ~ 13	0.6233	108.5359	<0.0000
13 ~ 14	0.6185	106.3703	<0.0000
14 ~ 15	0.6154	104.9932	<0.0000
15 ~ 16	0.6138	104.3214	<0.0000
16 ~ 17	0.6131	104.0165	<0.0000
17 ~ 18	0.6201	107.1023	<0.0000
18 ~ 19	0.639	116.048	<0.0000
19 ~ 20	0.6541	123.9161	<0.0000
20 ~ 21	0.6546	124.1959	<0.0000
21 ~ 22	0.6538	123.7321	<0.0000
22 ~ 23	0.6395	116.2941	<0.0000
23 ~ 0	0.6312	112.2355	<0.0000

- 商务金融用地、住宅用地、交通用地三种土地利用类型对区域实时人流量有显著的影响且在全天内存在明显的时序变化。



05 | 手机信令大数据与城市活力



- **丰度**: 增加一种土地利用类型, **全天活跃型**的发生概率是夜间活跃型的1.967倍, 白天活跃型的1.099倍。
- 增加一单位无序性(**香农熵**), **全天活跃型**的发生概率是夜间活跃型的1.795倍, 是白天活跃型的1.129倍。
- 增加一单位聚集度(以**辛普森指数**衡量), **全天活跃型**的发生概率是夜间活跃型的1.515倍, 白天活跃型的**0.984**倍。
- 合理的土地混合利用能够提升城市活力。

	变量	B	显著性	Exp(B)
全天活跃型	截距	-2.388	0.004	
	人口密度	0.223	<0.000	0.701
	商务金融	0.008	<0.000	1.218
	住宅用地	-0.019	0.016	0.881
	交通用地	0.031	<0.000	1.231
	辛普森指数	0.335	0.012	1.515
	丰度	0.329	0.009	1.967
	香农熵	0.333	0.039	1.795
白天活跃型	截距	-0.29	0.730	
	人口密度	0.239	0.910	0.712
	商务金融	0.006	0.007	1.216
	住宅用地	-0.021	0.030	0.879
	交通用地	0.083	<0.000	1.297
	辛普森指数	0.351	0.005	1.539
	丰度	0.234	0.018	1.789
	香农熵	0.212	0.043	1.59

注:参考类别为夜间活跃型



主要内容



- 1 手机信令大数据简介
- 2 手机信令大数据处理技术
- 3 手机信令大数据与用户画像
- 4 手机信令大数据与用户位置预测
- 5 手机信令大数据与城市活力
- 6 手机信令大数据与社会隔离

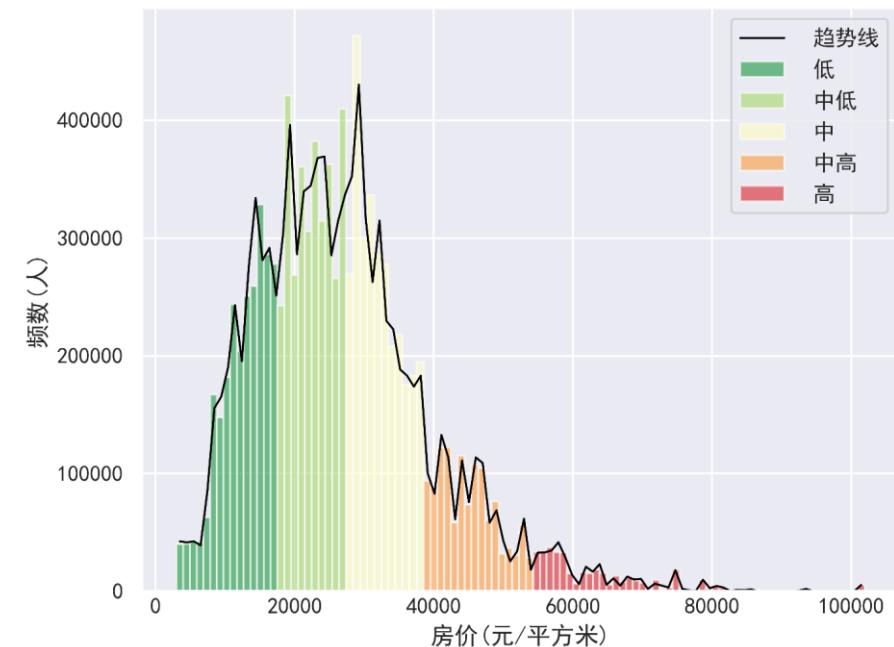
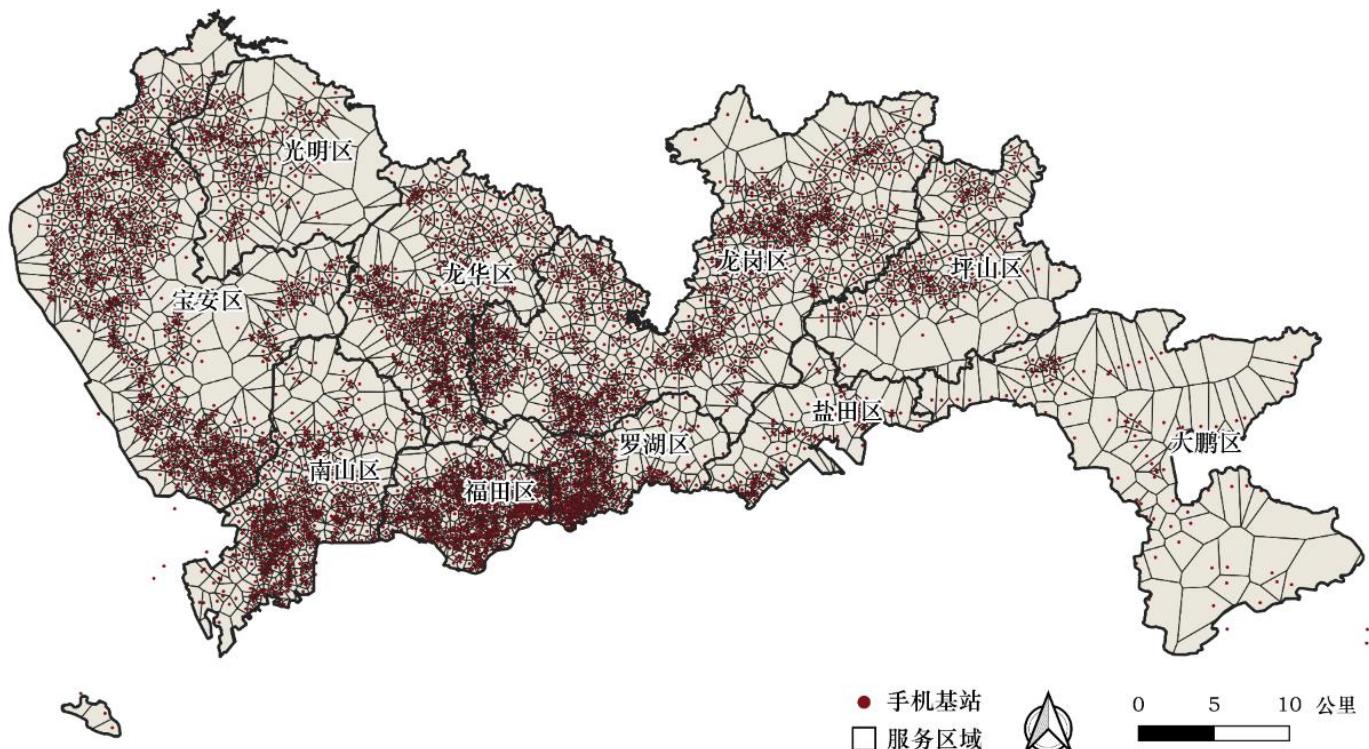
- 据联合国预测，2050年，世界上超过三分之二的人口将居住在城市地区。快速的城市化会导致社会隔离问题的加剧，分析并量化城市人群隔离对社会资源合理分配、社会公平、解决贫困至关重要。

城市中**不同区域**，对于人群活动存在**不同的影响**。



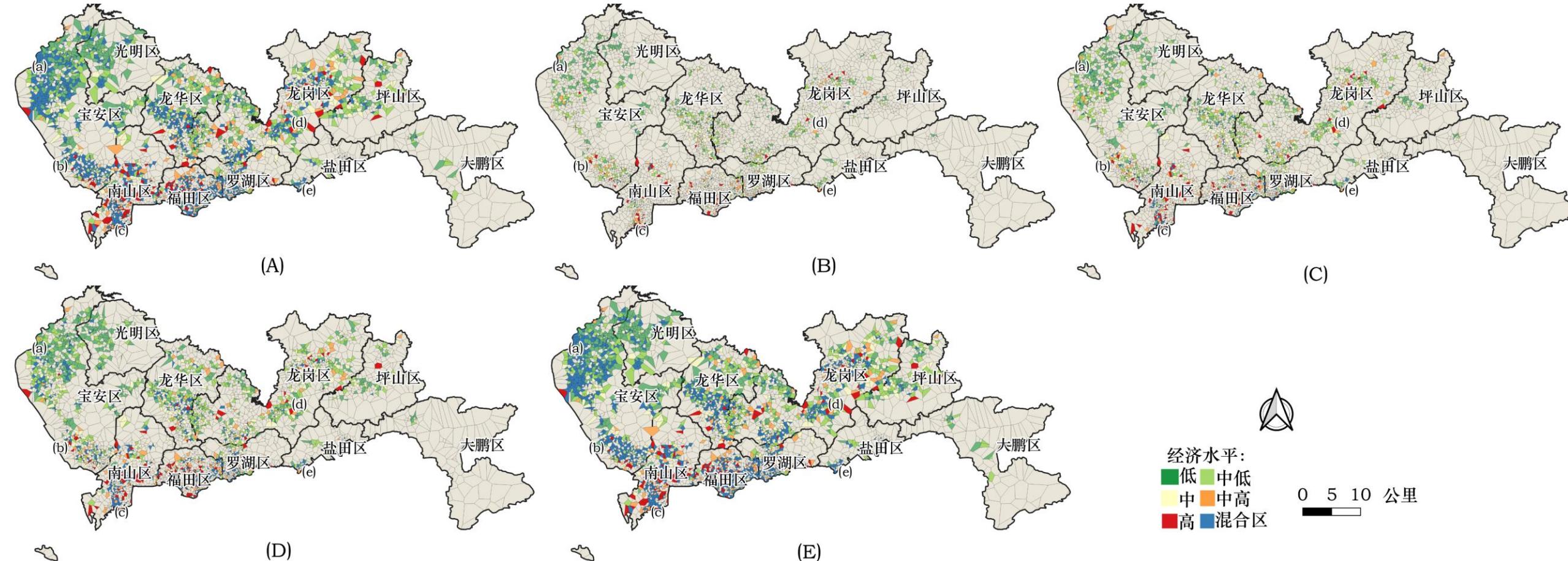
如何基于手机信令数据提取不同人群的活动区域，并对比分析不同经济水平人群是否存在对社会空间利用不平衡的**社会空间隔离现象**？

用户id	记录次数	记录时刻	记录位置	记录时刻	...
f5d4a*****0205	22	20120323 00:01:32	114.18** 22.64**	20120323 01:28:39	...
0bdf1*****91cb	24	20120322 23:30:13	114.21** 22.60**	20120323 00:30:15	...
1db81*****adf3	23	20120322 23:25:37	114.21** 22.60**	20120323 00:09:29	...
4cdd3*****49a3	9	20120323 12:53:30	114.09** 22.73**	20120323 02:27:50	...
556df*****439c	22	20120322 23:23:27	114.21** 22.60**	20120323 00:26:04	...
...
5790f*****c970	14	20120323 10:55:40	114.35** 22.70**	20120323 11:26:35	...

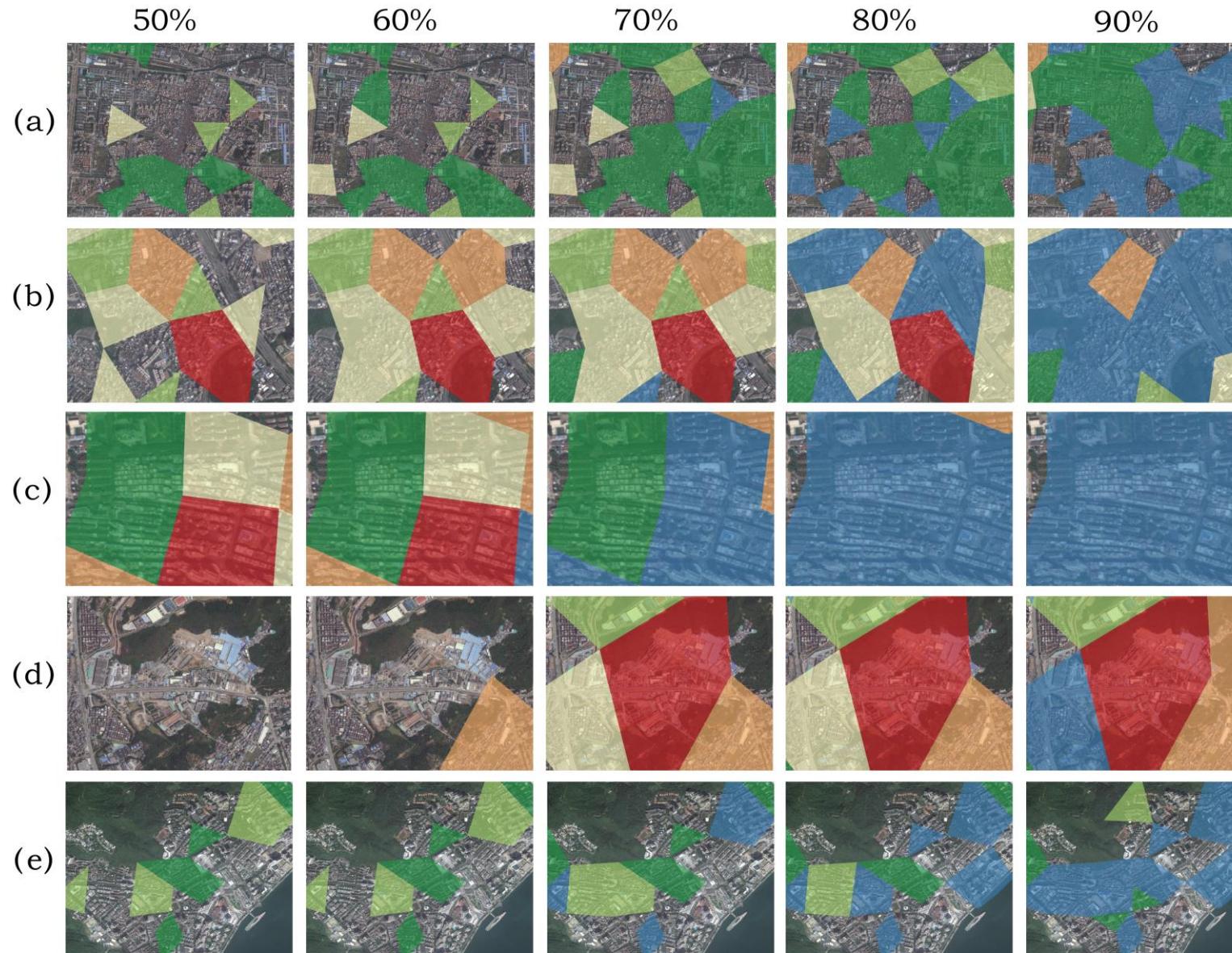


经济水平 \ 活动占比	50.00%	55.00%	60.00%	65.00%	70.00%	75.00%	80.00%	85.00%	90.00%	95.00%
L	99.01%	99.41%	99.24%	97.61%	92.48%	83.28%	64.32%	45.30%	25.38%	10.85%
M-L	99.23%	98.69%	98.13%	94.75%	89.49%	75.11%	55.64%	35.16%	18.49%	6.76%
M	98.74%	98.41%	96.90%	93.01%	85.52%	68.25%	49.26%	31.78%	18.84%	8.95%
M-H	97.40%	96.81%	94.89%	85.02%	71.43%	55.61%	40.55%	30.45%	22.66%	14.00%
H	98.70%	97.94%	90.15%	80.65%	62.41%	52.41%	42.14%	32.20%	22.21%	13.95%

深圳市不同经济水平人群在日常活动区域选择时，彼此之间存在较大的差异，不同经济水平人群之间存在社会空间隔离现象。



低经济水平人群活动网格有明显的**聚集**现象。而**高**经济水平人群活动区域开始呈现**全市多点分布**。



经济水平**较高**和**较低**人群之间的空间隔离更**明显**，且**较低**经济水平人群为**被动隔离**，而**高**经济水平则为主动隔离。

本章总结



本章介绍了手机信令大数据的**优劣势、数据处理过程、以及其在居民活动模式分析、城市功能、社会公平等方面的应用。**

手机信令大数据主要具有**收集快速、方便，时间空间覆盖高，人群有偏性小**的特点。

手机信令大数据在**居民活动分析、居民出行预测、城市内部功能分析、城市社会公平评价**等研究领域中都有应用。

手机信令大数据和其他多源数据耦合可以实现“ $1+1>2$ ”的效果，在分析**城市居民出行模式和城市内部细尺度的社会经济情况**等方面有广泛的应用前景。

手机信令大数据可以从最基本的单位-城市居民-入手，为了解社会经济情况提供帮助，为**城市精细尺度规划、城市规划**等提供决策支持。