大数据支持城市人口规模估算和监测的技术途径

钮心毅

同济大学建筑与城市规划学院　教授　博士生导师

2019年5月5日

0 大数据测算城市人口规模的必要性

全国人口大规模流动

- 十年一次人口普查数据更新时间长。
- 快速城市化阶段，常规统计方法容易失真。上海2009年统计数据、2010年六普数据差了288.96万
- 超大城市、特大城市都在困惑：当前城市人口规模、实际服务人口到底是多少？北京、上海、广州、深圳、杭州、苏州...

城市人口规模是城市规划的基础数据

- 城市各种资源、各类设施的规划都与所服务人口相关

建立国土空间规划体系并监督实施的部署要求

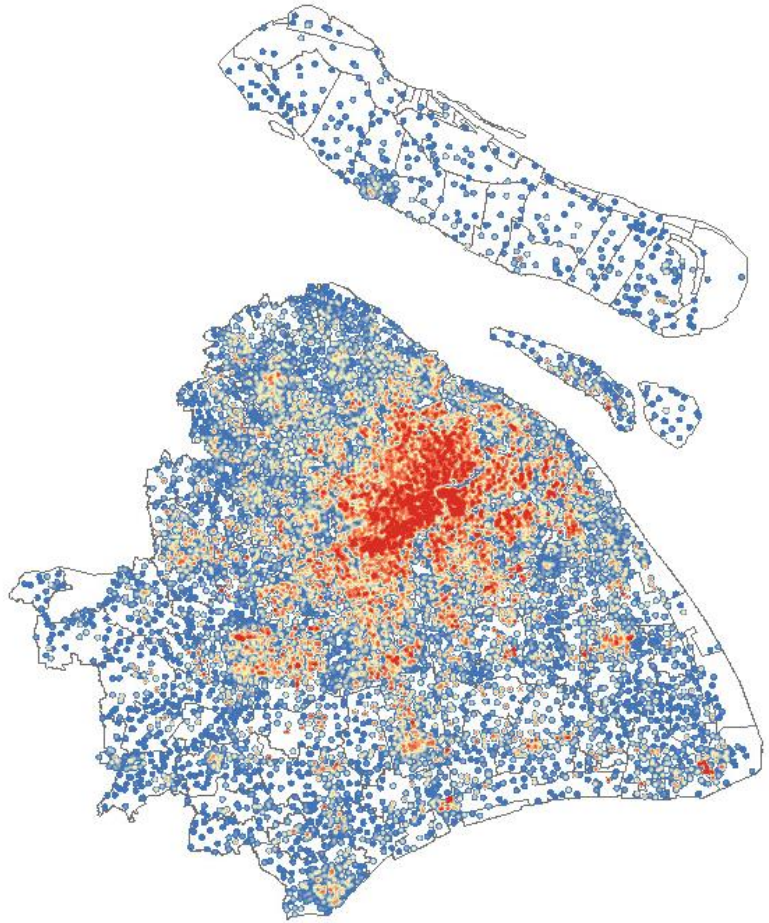
- 以“数字化”推动空间规划体系构建，支持国土空间规划监测、评估、预警
- “智慧”辅助国土空间规划，构建“智慧规划”。

0 大数据测算城市人口规模的必要性

- **移动通信和移动互联网普及**
  - 感知个体活动
- **移动定位大数据源：**
  - 手机信令数据：来自三家移动通信运营商。
  - 移动互联网LBS数据：来自多家互联网公司。
- **当前大数据测算人口的存在的“乱象”**
  - 非传统的“统计机构”纷纷报出各大城市“惊人数字“
  - 各种城市人口的“新”概念层出不穷
  - 学术界“一声不响“
- **根源：数据源、定义、扩样、检验，缺一不可**

1 定义：移动定位大数据反映的“人口”与 “城市人口规模”是一回事吗?

- **“常住人口规模” 概念：**
  - 在某城市连续居住6个月以上。
  - 一种原生概念：**一种城市生活逻辑所定义的概念**
- **移动定位大数据能直接测算的人口：**
  - 即时人口——某一时刻（天、小时）在城市内停留过的人数
  - 过境人口——某日在某城市内短暂停留、过境
  - 短驻人口——短期留驻人口，在某城市内居住不足6个月的人口
  - .....
  - 衍生概念：**一种基于设备用户行为、数据采集逻辑所定义的概念**
- **不能直接对应“城市人口规模（常住人口）“**
  - 能观测到很多“常住人口” 不能反映的人口概念；原因在于两种概念的基础逻辑不同



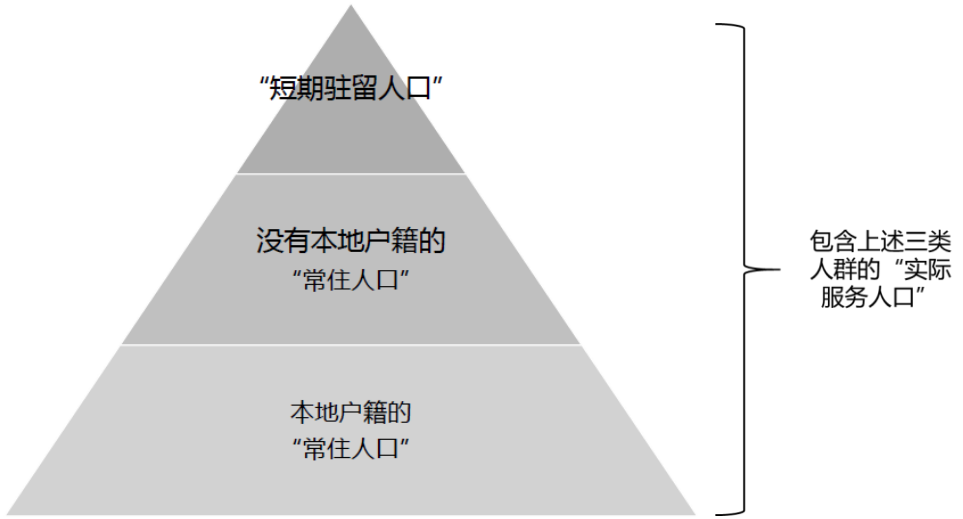
移动定位大数据直接测算的是设备（用户）行为

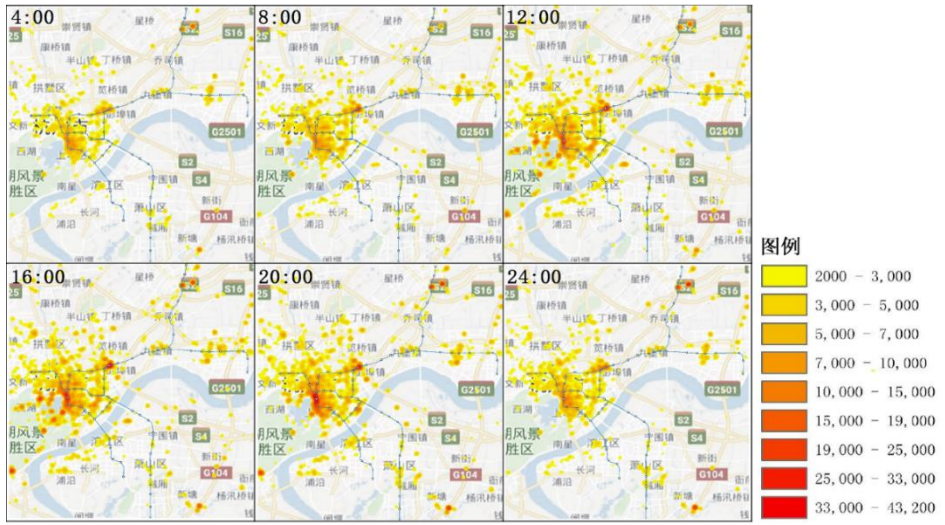
1 定义：移动定位大数据反映的“人口”与 “城市人口规模”是一回事吗?

- **移动定位大数据测算“常住人口”必须符合“城市人口规模”的定义**
  - 基本原理是对手机用户**长时间时空轨迹规律**进行测算，以**多日夜间的时空轨迹**推算居住地。
  - 每日的即时人口，计算当晚居住地
  - **6个月时长数据，计算50%以上（重复率）日期**均居住在该城市内
- **不能以移动定位大数据某种特点，替代“城市人口规模”定义**
  - （1）不能以数据源难以获取的原因，以短时间周期计算常住人口（比如1-2周）
  - （2）不能以大数据的人口数据采集逻辑替代城市人口生活逻辑
  - 不同数据源（信令数据、移动互联网LBS数据）均应遵循

1 定义：移动定位大数据反映的“人口”与 “城市人口规模”是一回事吗?

- **移动定位大数据提供了传统的统计概念中很难定义的人口测算途径**
  - 北京、上海、广州等城市总体规划中强调“**实际服务人口**”。
  - 短期留驻人口：短期在某城市居住停留，但不满6个月
  - 已有杭州等特大城市案例，使用移动定位大数据估算出短期驻留人口的规模能达到同期常住人口总量的10%
  - 移动定位大数据使得测算城市实际服务人口有了可能。





施澄 陈晨 钮心毅. 面向 “实际服务人口”的特大城市空间规划响应——以杭州市为例. 城市规划学刊，2018（4）：41-48

1 定义：移动定位大数据测算一个“ID” 对应一个“人” 吗?

- **多种大数据产生和采集原理差异，形成了不同的特征**
  - 物联网设备：即便是手机卡，一卡未必等于一个人
    - 未来5G时代，物联网是移动通信主要新增点
  - 一人多设备（一机多卡）、一个设备多个APP。
    - 手机普及率，全国移动电话用户普及率达到112.2部/百人（2018年）

- **应对方法**
  - 手机信令数据存在剔除物联网用户的技术关节。
    - 专用号段
  - 移动互联网LBS数据
    - 融合统一的ID——一设备多个APP、一人多设备，

2 扩样：扩样的逻辑

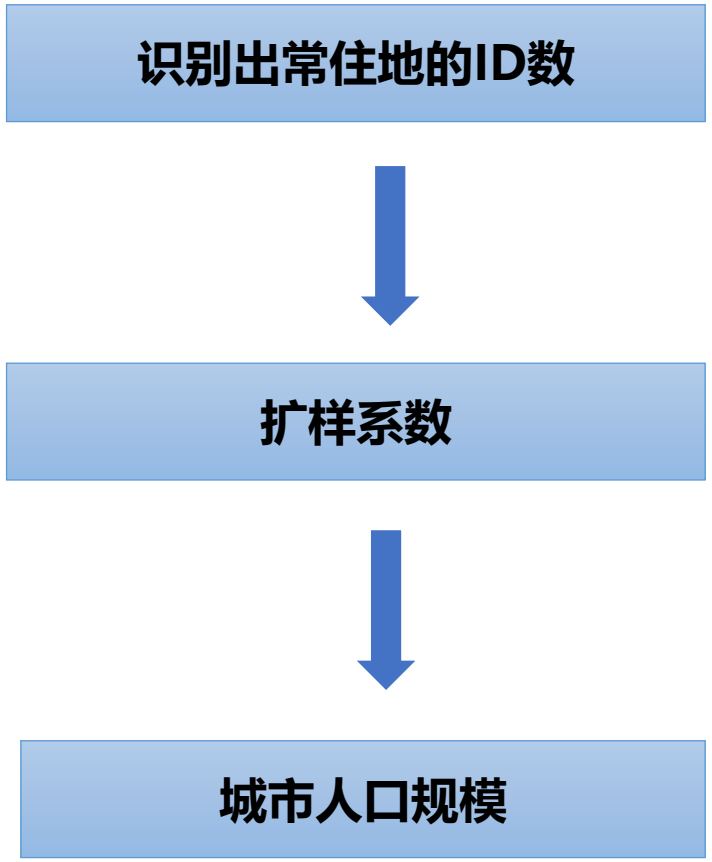
没有一家数据源覆盖了100%的人群

- 一家运营商肯定需要扩样。
- 三家运营商合计在一起也需要另一种“扩样”。因为仍存在没有手机的人、不常用手机的人。
  - 使用多家运营商来源数据，其主要意义在于相互对照、比较，而不是合并扩样。
- 移动互联网LBS数据也需要扩样，设备识别数肯定不能覆盖100%人群。

简单扩样

- **设备数 / k = 扩样后人口值**。

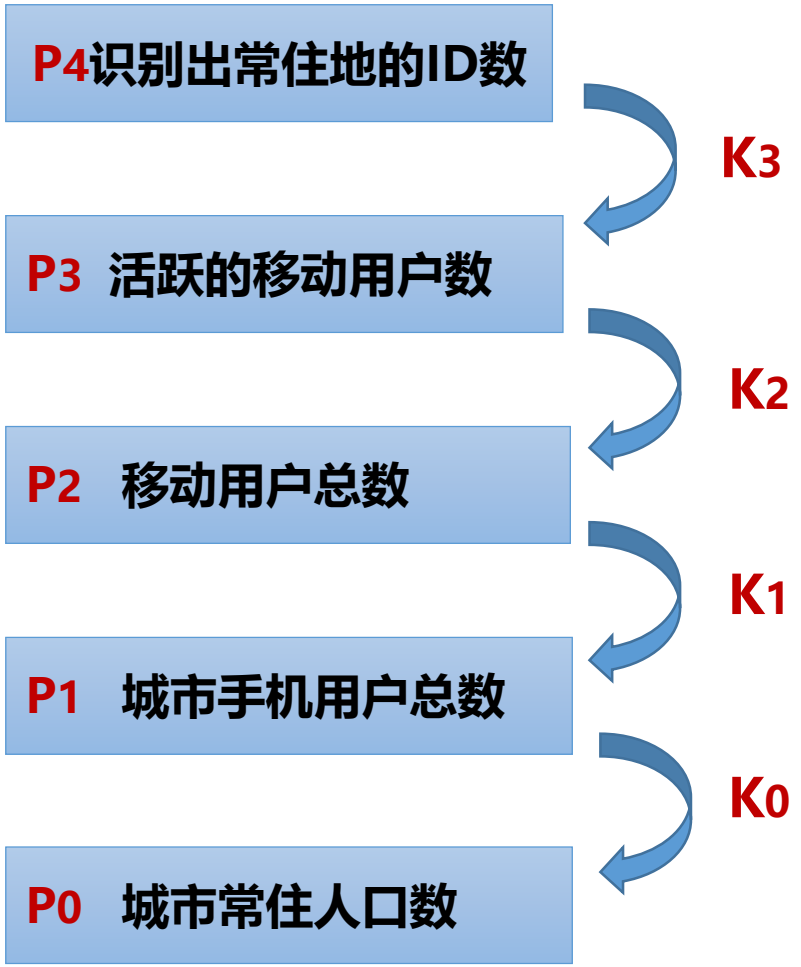
K： 扩样系数， 例如70%



简单扩样一般流程

2 扩样：简单扩样不可行

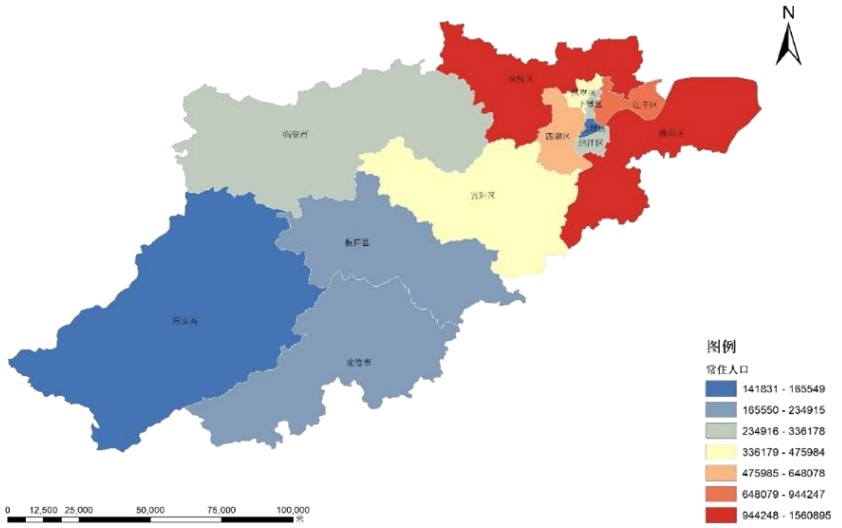
- **以移动公司的手机信令数据为例**
  - 常住人口算法只能从信令数据得出 **P4** 值
  - 假如某地移动公司提供信息：当地市场份额60%，只是告诉 **$K_1 \approx 60\%$**
  - **K_3 K_2 K_0** 取值无法准确预估。
    - **城市常住人口中多少比例没有手机？—— K_0**
    - **多少移动手机是备用卡（发了卡、不常用）—— K_2**
    - **多少用户ID行为不规律导致常住地无法计算（夜间不开机）—— K_3**
  - 假如4个K值取值都比真实系数大 5%，最终P0会比真实城市人口规模大21.5%
- **移动互联网LBS数据，同理。**



扩样系数 $K = K_0 * K_1 * K_2 * K_3$

2 扩样：扩样的逻辑

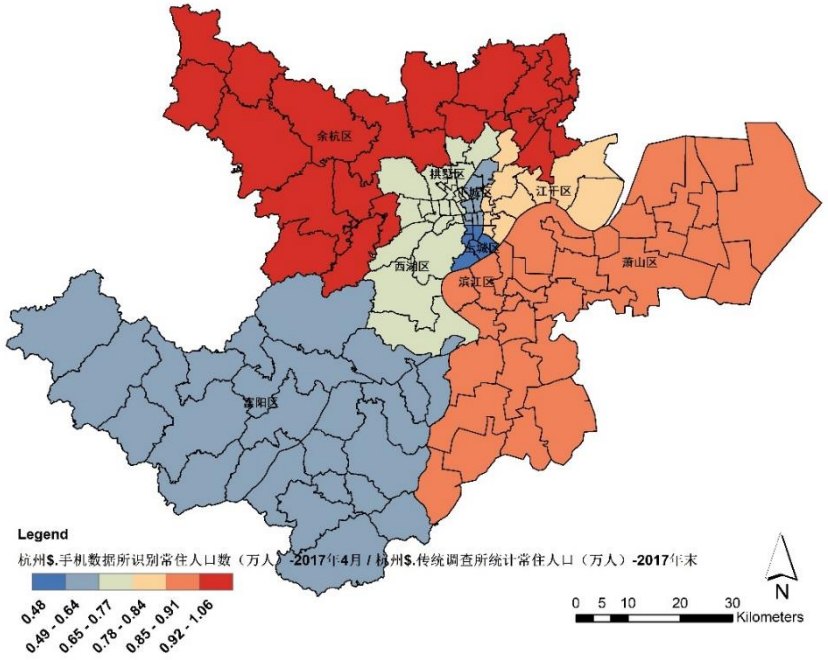
- **显著影响扩样的结果其他因素**
  - 空间单元 (地级市、区县、街道)：
  - 地域差异（城市 / 乡村、东部 / 中西部 .....)
- **简单扩样不可行**
- **1、不扩样常住人口照样有价值**
  - 要习惯看不扩样的常住人口值
- **2、不追求过小的集计空间单元**
  - 现阶段区县单元统计是适宜单元
- **有效的扩样解决方案要依赖机器学习等人工智能方法**

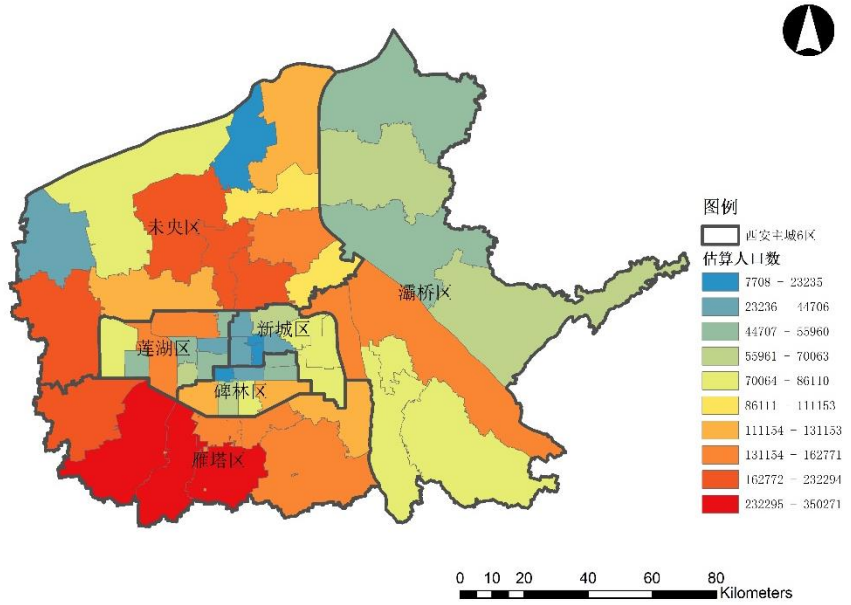


区县	识别常住用户/万人	统计常住人口/万人	比值
上城区	16.55	35.32	47%
下城区	30.32	53.60	57%
拱墅区	43.48	58.20	75%
西湖区	64.81	84.42	77%
江干区	94.42	106.15	89%
滨江区	31.88	33.56	95%
萧山区	146.86	157.20	93%
余杭区	156.09	135.90	115%
富阳区	47.60	73.64	65%
临安区	33.62	58.85	57%
淳安县	14.18	34.96	41%
建德市	22.56	44.70	50%
桐庐县	23.49	42.30	56%
总计	725.87	918.80	79%

3 校核：多源数据校核是必不可少的步骤

- **三类原因会影响估算结果的可靠性**
  - 数据源本身导致（定义）
  - 常住地计算方法导致
  - 扩样方法导致
- **判断估算结果依据**
  - 往往是统计局数据可能已经不能反映实际人口
  - 结合**多方数据源**对估算结果可靠度进行验证
- **两类检验方法：**
  - （1）多源大数据的相互校核
  - （2）与社会经济统计数据、城市建设实际情况对照检验



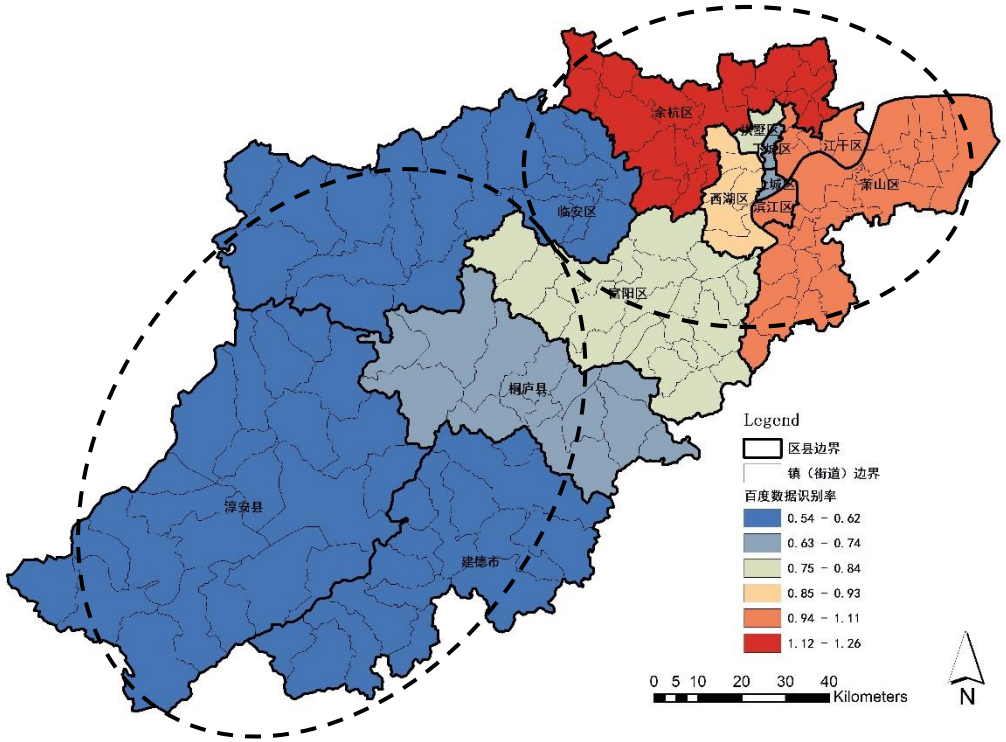


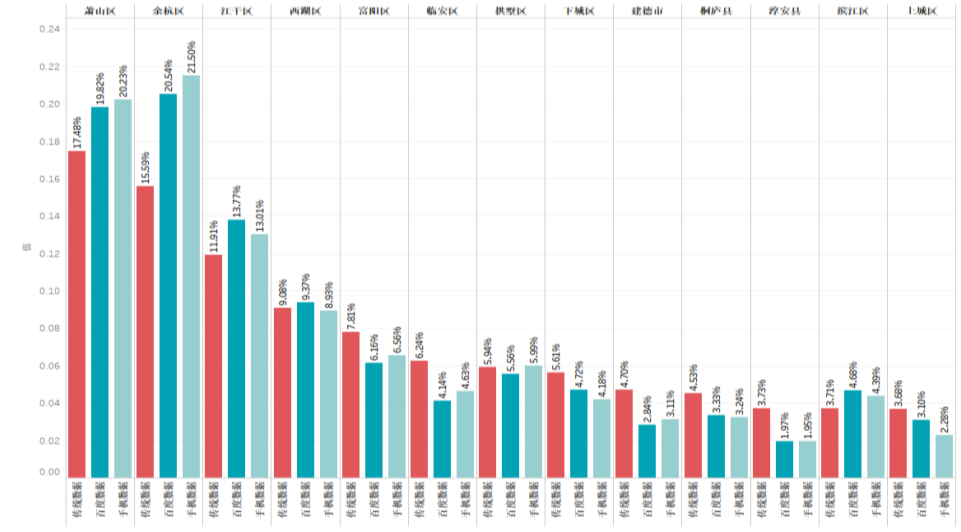
3 校核：多源大数据的相互校核

两种不同来源的移动定位大数据估算结果相互对照

- 两种数据反映的趋势、空间特征是否相同
- 人口规模变动趋势、空间特征是否符合社会经济状况

杭州市区县名称	杭州市统计局发布常住人口-万人（2017 年末）	手机信令扩样后常住人口-万人（以60%全市用户识别率推算，2017 年 4 月）	百度数据扩样后推测常住人口-万人（2018 年上半年）	手机信令数据扩样后常住人口识别率-%	百度慧眼数据扩样后常住人口识别率-%
上城区	34.8	27.6	39.3	79%	113%
下城区	53.1	50.5	60.0	95%	113%
江干区	112.8	157.4	174.8	140%	155%
拱墅区	56.2	72.5	70.6	129%	126%
西湖区	86	108.0	119.0	126%	138%
滨江区	35.1	53.1	59.4	151%	169%
萧山区	165.5	244.8	251.5	148%	152%
余杭区	147.6	260.2	260.7	176%	177%
富阳区	73.9	79.3	78.1	107%	106%
临安区	59.1	56.0	52.5	95%	89%
桐庐县	42.9	39.2	42.3	91%	99%
淳安县	35.3	23.6	25.0	67%	71%
建德市	44.5	37.6	36.1	84%	81%
总人口	946.8	1209.8	1269.2	128%	134%

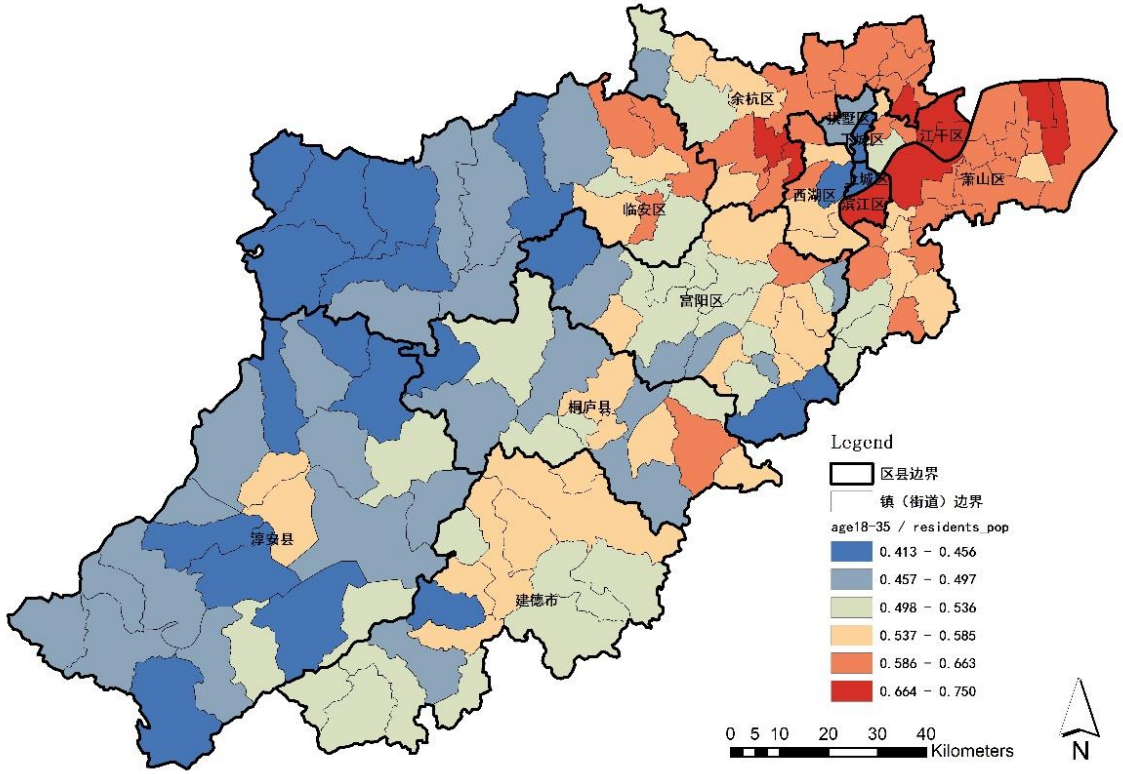




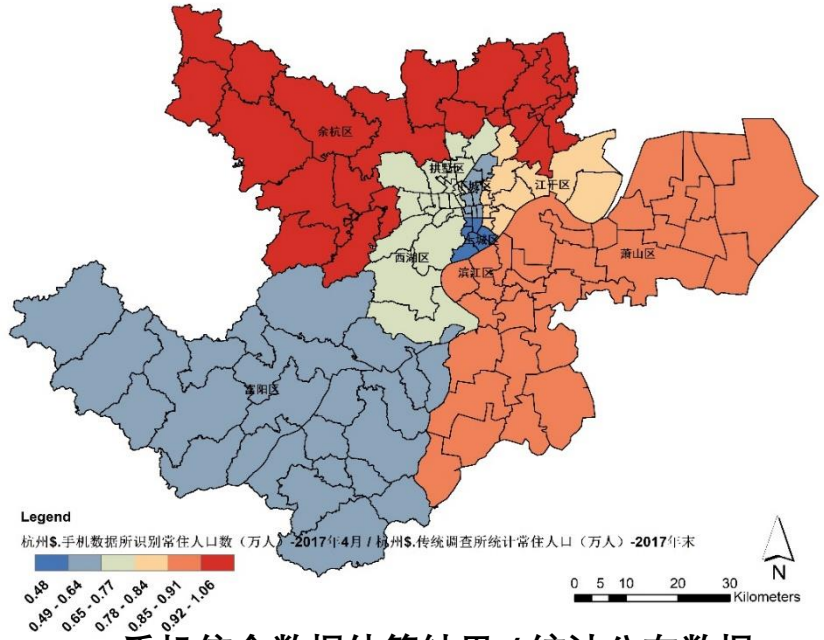
3 校核：多源大数据的相互校核

人口规模变动趋势、空间特征是否符合社会经济状况

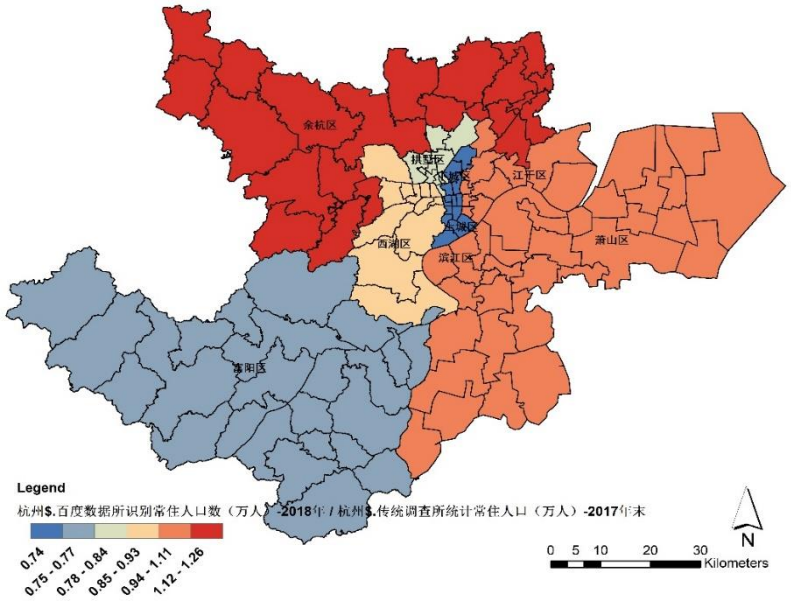
- 例如：用中青年就业年龄段的常住人口占比解释估算结果。快速城市化中的外来就业人口在近郊区大规模增加



杭州市域范围内18至35周岁（中青年就业年龄段）居民在各镇（街道）常人口中的占比（%）（百度慧眼数据测算）



手机信令数据估算结果 / 统计公布数据

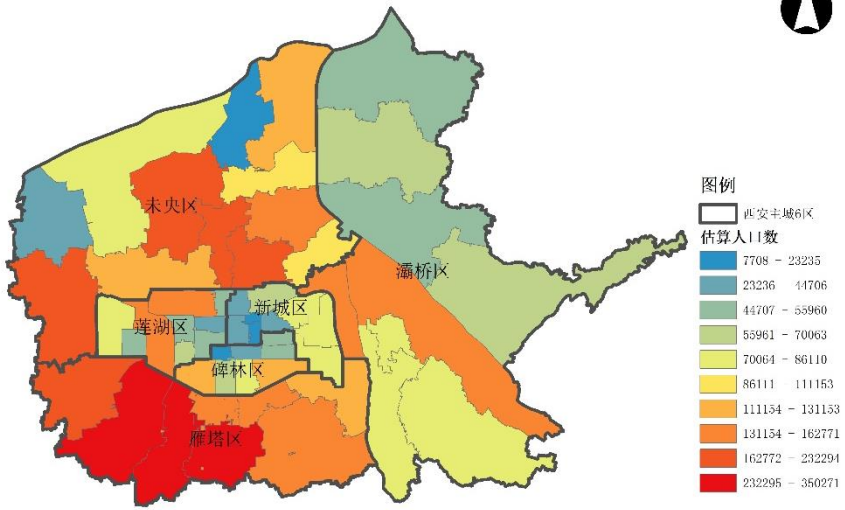


百度数据估算结果 / 统计公布数据

3 校核：与社会经济统计数据、城市建设实际情况对照检验

与社会经济统计数据、城市建设实际情况对照

- 移动定位大数据估算人口的空间分布特征是否能从其他社会经济数据中得到解释
- 土地使用、住房建设、经济产值、产业结构……



区县名	2010-2017总销售面积（平方米）	2010-2017总竣工面积（平方米）	估算人口区间中间值与"六普"统计人口差值
未央区	53482195	28327760	810123
雁塔区	69392645	25540304	638425
灞桥区	21565383	7384411	204500
莲湖区	10540581	7606055	26366
碑林区	13230605	14561092	-34411
新城区	5755131	2995653	-112663

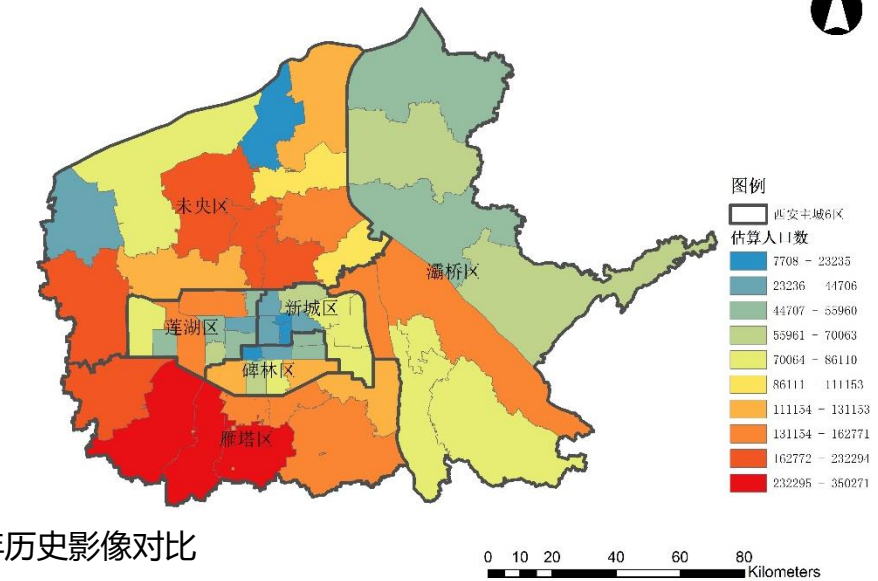
区县名	估算人口与2017年统计人口比值	与"六普"统计数据比值
未央区	1.71	1.91
雁塔区	1.38	1.47
灞桥区	1.19	1.28
莲湖区	0.94	0.99
碑林区	0.85	0.90
新城区	0.73	0.77

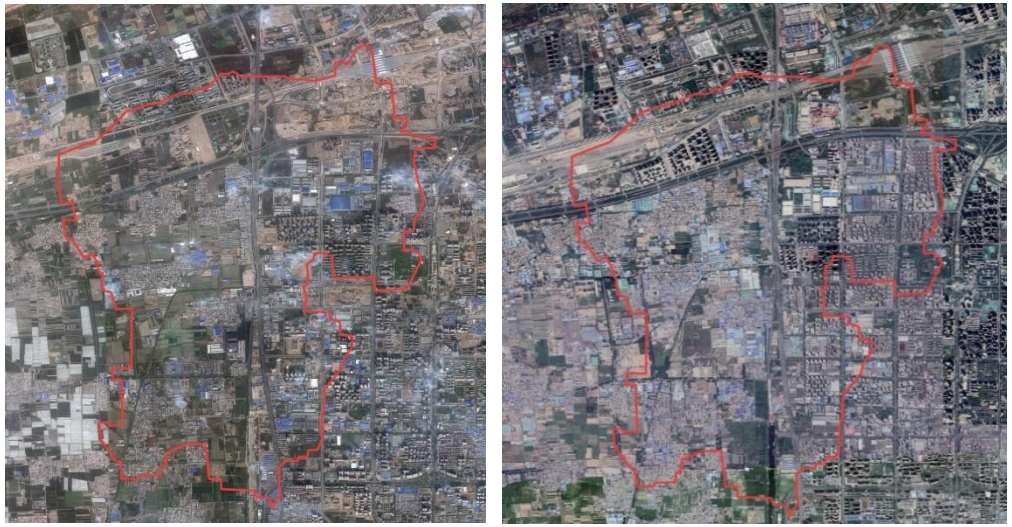
3 校核：与社会经济统计数据、城市建设实际情况对照检验

与社会经济统计数据、城市建设实际情况对照

- 例如：影像数据中反映城市住房变化与估算结果能对应

大明宫街道2010年与2017年历史影像对比





汉城街道2010年与2017年历史影像对比

区县名	估算人口与2017年统计人口比值	与"六普"统计数据比值
未央区	1.71	1.91
雁塔区	1.38	1.47
灞桥区	1.19	1.28
莲湖区	0.94	0.99
碑林区	0.85	0.90
新城区	0.73	0.77

4 应用方向

- **国土空间规划的实施监测、评估、预警**
  - 监测城市人口**规模变化的趋势、空间分布变化趋势**
  - 不追求监测人口规模的精确绝对总数。
- **城乡人口大规模流动、大规模变化的城市**
  - 特大城市、超大城市快速、短周期的人口规模估算
  - 便捷、经济、高效判断常住人口变化
- **与传统人口普查的关系、与统计局人口数据关系**
  - 10年周期的全国人口普查是最详细的人口数据
  - 移动定位大数据优势在两次人口普查之间，有效地估算城市人口
  - 辅助、优化人口统计的重要手段

4 结语

- （1）**迫切性：国土空间规划、国土空间基础信息平台**
  - 移动定位大数据估算“常住人口规模”， 是国土空间规划实施监测、评估的有效工具。
  - 移动定位大数据为测算“实际服务人口规模” 等概念提供了可能。
- （2）**技术途径：定义、扩样、检验、应用**
  - 移动定位大数据优势短周期估算在于城市人口规模变化的趋势、空间分布变化趋势。
  - 要特别关注“定义、扩样、检验“三个技术环节。
- （3）**时机：国土空间规划、第七次人口普查**
  - 在2020年“七普”时间节点是推进大数据估算监测人口规模研究、实践的恰好时机

谢 谢

niuxinyi@tongji.edu.cn