

基于大数据的城市人口社会 经济特征精细时空尺度估计 ——数据、方法与应用

刘云霞^{1,2}, 田甜³, 顾嘉钰¹, 刘军民³

(1. 南开大学 周恩来政府管理学院, 天津 300350; 2. 南开大学 计算社会科学
实验室, 天津 300350; 3. 西安交通大学 数学与统计学院, 陕西 西安 710049)

摘要:随着时空大数据的出现和数据挖掘技术的发展,学界在精细时空尺度上对城市人口社会经济特征分布的研究中取得了长足进展,有效弥补了传统数据时效性差、空间精细度低等不足。但这些研究集中在城市信息学、计算社会科学、地理学等领域,其成果尚未广泛应用于以城市人口时空分布为决策依据的资源配置研究中,以解决制约城市精细化管理实践深入的科学问题。为促进信息沟通和学科间对话,就城市人口社会经济特征精细时空尺度估计的数据源、估计方法与模型、结果验证与模型评估、较成熟的人口特征分布数据集、精细尺度人口特征数据的应用背景和主题等展开分析,并展望了相关研究的前景与面临的挑战。分析发现:城市人口特征的精细时空尺度估计所运用的方法包括空间回归、机器学习以及神经网络等。在诸多社会经济特征中,人口数量的时空分布研究最为成熟,经历了从单一数据源到多种数据源、从一般格网化到带有空间信息的格网化、从低时空分辨率到高时空分辨率的发展。城市人口的性别、年龄、种族和教育程度等特征可基于社交媒体数据预测,但尚不成熟。人口特征的精细尺度数据应用主题日趋丰富,包括疾病

收稿日期: 2021-06-21; 修订日期: 2021-10-12

基金项目:国家自然科学基金项目“面向需求的城市社区生活服务圈建设评价与优化发展对策研究”(72074127); 中央高校基本科研业务费专项资金“可持续发展理念下人口老龄化与生态环境治理研究”(63212070); 中国博士后基金项目“居民异质性视角下老旧小区宜居性评价与提升机制研究”(2019M651023); 国家社会科学基金面上项目“模型自适应合成检验方法及其应用研究”(2021ND03050862)。

作者简介:刘云霞,南开大学周恩来政府管理学院,南开大学计算社会科学实验室助理研究员;田甜,西安交通大学数学与统计学院硕士研究生;顾嘉钰,南开大学周恩来政府管理学院本科生;刘军民(通讯作者),西安交通大学数学与统计学院副教授。

防控、贫困识别、气候变化应对、城市风险识别、城市发展历程回溯、城市社会空间结构分析等。

关键词: 城市人口; 时空分布; 社会空间结构; 机器学习; 神经网络; 计算社会科学

中图分类号: C922 **文献标识码:** A **文章编号:** 1000-4149 (2022) 01-0042-16

DOI: 10.3969/j.issn.1000-4149.2021.00.056

一、引言

2017年3月5日,习近平总书记在参加十二届全国人大五次会上海代表团审议时指出,“城市管理应该像绣花一样精细”;《中共中央关于制定国民经济和社会发展第十四个五年规划和二〇三五年远景目标的建议》中提出,要“推进以人为核心的新型城镇化”,在以人为本的城市发展理念下,精细的城市管理离不开精细的人口时空分布数据的支撑。但在目前,人口学研究沿用的数据仍以全国人口普查数据、全国人口抽样调查数据、各地统计年鉴数据等传统人口数据为主,在揭示人口时空分布规律方面存在一定局限性:第一,时效性差。传统统计调查的间隔周期较长,人口普查数据以10年为周期,1%人口抽样调查数据和各地统计年鉴数据分别以5年和1年为周期,时间尺度较大,难以反映更小时间尺度(比如月、天、时)的人口动态变化。第二,空间尺度的精细度低。公开渠道可获得的人口统计数据,最小空间尺度一般为行政区或选区,数据一般基于行政区而非自然地理环境进行划分^[1]。传统人口数据的局限性,制约了城市精细化管理的深入;更精细时空尺度的人口数据,能够更好地助力城市街道和社区层面公共服务与商业配套等与人口密切相关的资源配置,从而加强基层治理体系,统筹推进乡镇(街道)和城乡社区治理。

基于时空大数据的人口社会经济特征的精细时空尺度估计/预测,使精细化的人口社会经济特征分布数据的获得成为可能。人口社会经济特征的时空分布,亦即城市人口特征的时间化、空间化,是将人口的数量、性别、年龄、收入、种族、就业等数据映射到时间轴上、地图上,建立与时空坐标的对应关系^[2],为原本不具有明确时空属性的人口特征数据赋予时间和空间属性,或者将原本时间间隔较长、空间尺度较大的统计数据进行时间标签层面的数据密集化、空间层面的精细化分解,从而得到更精细时空尺度上的人口特征数据。其中,时间维度的精细化取决于数据采集的时间间隔,而空间尺度的精细化则依赖于科学的估计/预测方法。

早期的人口数据空间化主要是指利用一定的方法,对基于较大空间尺度统计的人口特征汇总数据进行空间维度的离散化处理,得到更小空间尺度上的人口特征^[3],对人口统计数据在空间维度进行有益补充。这种由大(尺度)向小(尺度)、自上而下的人口数据精细尺度分解方式,伴随多源大数据的出现和信息技术的发展出现转折。近年来,随着智能终端设备、电子商务平台、社交媒体的兴起以及遥感、人工智能等技术的发展,产生了海量时效性高、空间尺度更精细并且隐含着人口社会经济特征的数据^[4-7],极大丰富了人口特征时空分布研究的数据源^[8]。此外,数据挖掘技术、计算社会科学的发展,为更科学、更精确地研

究人口特征的时空分布提供了有力的支持，使得通过图像等非结构化数据预测人口特征时空分布成为现实。

本研究旨在回答以下问题：①从何而来？——作为估计/预测数据，城市人口社会经济特征精细尺度时空分布估计的数据源有哪些，模型或方法如何演进，估计、预测的精准度如何，面临的挑战和机遇有哪些？②有哪些成果？——既有研究为我们提供了哪些人口社会经济特征数据，这些数据的时空精细程度如何？③如何应用？——这些成果已有哪些应用场景，将有哪些应用前景，如何更好地支撑城市管理的相关研究？为回答以上问题，本文系统分析了城市人口社会经济特征精细时空尺度估计相关研究的数据源、估计模型或方法、产出的数据库成果及其应用现状与前景。

二、城市人口特征精细时空尺度估计：过程与方法

1. 人口特征精细尺度时空估计的过程

多年来，致力于该相关领域研究的学者使用不同的数据源和估计方法，得到高时空分辨率的人口社会经济特征数据。尽管不同的研究使用的数据、方法与模型不同，但通常遵循图 1 所总结的流程。

首先，对模型自变量和因变量数据进行收集与预处理。在这一阶段研究人员需要清洗数据，处理异常值以及缺失值，保证数据可用性。人口分布的空间估计所基于的数据源趋向多元化，从早期的手机信令数据^[9-10]，发展到夜间灯光数据^[6]、土地利用数据^[11-13]，并越来越多地运用 POI 数据^[14]、轨迹数据^[15-17]、公交乘车数据、社交媒体数据（用户上传的图像、在软件/平台上以各种形式发布的文本信息）、街景图片、房屋交易数据等^[18-21]。在数据源多元化的同时，数据的质量也在不断提高。此外，多源数据的精准匹配与有效融合，为人口空间估计工作在更精准、精细尺度上的推进提供了新的思路^[22]。

其次，初步提取源数据特征，再根据其与因变量的相关性、与自变量之间的共线性或者研究人员的经验，对提取的特征进行筛选，确定最后用于建模的自变量。自变量的形式与源数据类型有关：如果源数据是空间数据，自变量可以是土地覆盖等；如果源数据是夜间灯光图像/遥感图像，自变量可以是灯光辐射总量、灯光辐射面积、图像灰度值等；如果源数据是移动互联网/通信相关的新型数据，自变量可以是与手机或者社交网络相关的数据。因变量即研究者预测的人口社会经济特征，例如人口数量、人口密度、收入水平等。此外，关于因变量，需要将研究年份的人口/经济统计调查数据与研究地区的行政边界进行匹配，再根据每个网格单元所对应的自变量取值，建模计算划分的每个网格单元里人口社会经济特征的具体数值（例如不同收入、性别、教育水平对应的人数等）。

再次，研究者需根据数据的实际情况进行传统的统计建模或者构建机器/深度学习模型，将每个网格对应的人口社会经济特征数值作为因变量，筛选所得特征作为自变量，将自变量取值输入对应的模型从而得到人口社会经济特征的空间分布。

最后，进行结果验证与模型评估。如果是回归型的预测任务（例如计算具体的人口数量等），可以选择均方误差（MSE）、平均绝对误差（MAE）、均方根误差（RMSE）、决定

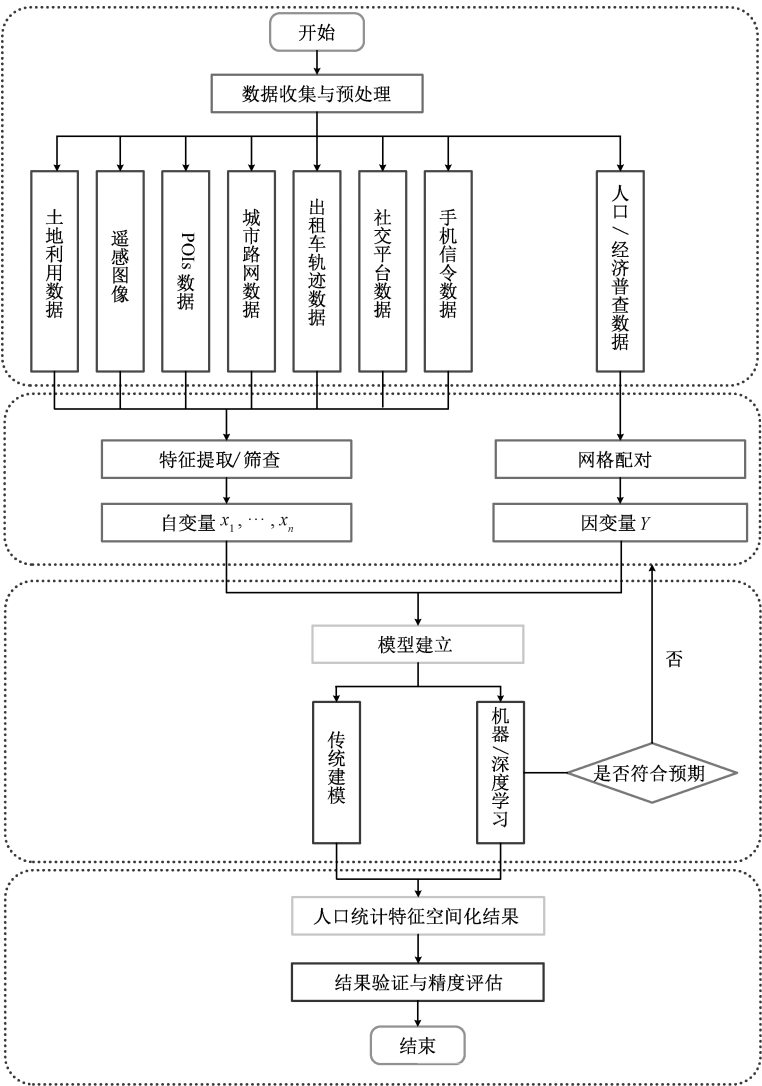


图 1 城市人口特征精细时空尺度估计的过程

系数 (R^2) 等指标进行结果验证以及模型评估。如果是分类型的预测任务 (例如预测对应网格的贫困等级等) 可以选择准确率 (Accuracy)、召回率 (Recall)、F1-Score 等指标进行结果验证以及模型评估。其中, 均方误差 (MSE)、平均绝对误差 (MAE)、均方根误差 (RMSE) 的值越小, 表示模型越好, 预测的效果越佳; 而决定系数、准确率 (Accuracy)、精确率 (Precision)、召回率 (Recall)、F1-Score 等指标的取值范围为 0—1, 值越接近于 1, 表示预测的效果越佳。

2. 人口特征精细尺度时空估计的方法与模型

城市人口特征的精细时空尺度估计所运用的方法包括空间回归、机器学习、神经网络等 (见表 1)。空间回归分析常用于拟合或推断人口社会经济特征的空间分布, 因为城市人口的空间分布特征与文化资源、经济资源、地理环境等因素紧密相关, 且这些因素间具有空间相

关性。但当空间回归分析假设较多时，模拟结果的优劣取决于研究者对自变量和因变量的选择与处理。

表 1 人口特征估计的主要建模方法

建模方法	具体模型	代表性研究及主要贡献
空间回归	地理加权回归	使用四个土地利用区域变量（高密度城市利用、低密度城市利用、农田和森林）来估计亚特兰大的人口 ^[27]
	空间自回归	建筑体积作为人口预测指标。证明在人口普查区这一空间尺度下，基于体积的模型始终优于基于建筑面积和长度的模型 ^[28]
机器学习	随机森林	基于人口普查和辅助数据，使用随机森林模型预测 100 米×100 米空间分辨率下的人口密度 ^[29]
	弹性网回归	证明了通过手机数据模拟的数百万个人特征，可以准确地重建整个国家的财富分布 ^[9]
	LASSO 回归	利用大众点评餐厅数据对四项人口社会经济特征（夜间人口分布、日间人口分布、公司数量和消费水平）进行模拟 ^[30]
深度学习	支持向量机等	使用 DMSP/OLS 夜间灯光图像和机器学习方法预测某地区的贫困程度 ^[31]
	CNN 和岭回归	利用 CNN 从高分辨率卫星图像提取统计特征，结合岭回归和统计特征估计五个非洲国家消费支出和资产财富 ^[32]
	CNN 分类算法	将目标检测和分类算法结合，证明了社会经济属性可以从含有汽车的街景图像中推断出来 ^[33]
	VGG-16 网络	基于街道图像和 VGG-16 分类算法，测量收入、教育、失业、住房、生活环境、健康和犯罪的空间分布 ^[34]

大数据的出现为机器学习提供了海量数据，在机器学习的助力下，计算社会科学的发展有了质的提升^[23]，机器学习算法越来越多地应用于人口社会经济特征的空间分布推测。目前，该领域最常用的机器学习算法有：贝叶斯模型、多元线性回归、支持向量机、随机森林与神经网络等。其中，与其他算法相比，随机森林算法无须考虑特征之间多重共线性的问题，因此可以有效避免过度拟合。此外，随机森林算法对噪声数据的敏感度低，能很好地适应高维数据，所以该算法在城市人口特征的精细时空尺度估计中应用较广泛。

神经网络利用计算机模拟动物神经网络的行为特征，并通过该网络从大数据中学习一般规律^[24]。过去十年，神经网络在计算机视觉、自然语言处理领域及其应用方面取得了显著的成就，这也给人口社会经济特征估计带来了新的研究范式，使得通过图像以及移动应用平台文本数据来推测人口社会经济特征成为可能^[25]。但由于神经网络是一种数据驱动的计算法，目前仍无法确定它对于不同场景数据的适应性和泛化性，因此结果的可解释性有待商榷。此外，神经网络需要大量有标注的数据集进行训练，这也是它目前难以广泛应用的原因^[26]。

三、城市人口数量的精细尺度空间估计

城市人口的社会经济特征众多。但在诸多特征中，人口数量的时空分布研究最为成熟。该领域研究经历了从单一数据源到多种数据源、从一般格网化到带有空间信息的格网化、从低空间分辨率到高空间分辨率、从静态的空间数据到含时间属性的动态空间数据的转变。

1. 静态的人口空间分布

早期研究者们基于城市地理学相关理论，引入场论，建立了人口密度模型。利用人口密度模型实现人口空间化的典型研究有：刘爱华等基于 2000 年和 2010 年天津市街镇人口数

据,利用克拉克(Clark)模型和基于中心地理论建立的回归方程对天津市主城区人口空间结构变化进行了研究^[35];康停军等对广州市2000年的城市人口进行了单中心和多中心的模型拟合,解决了因行政单元面积差异大而导致的郊区人口密度被低估的问题^[36];陈浩等以兰州为例,对乡镇街道尺度的常住人口数据进行空间化处理,提出变形椭圆圈层法,并结合GIS空间分析和Smeed模型拟合等方法,多角度揭示2000—2010年兰州市人口空间分布格局及其演变特征^[37]。

在早期的人口分布研究中,除了人口密度模型之外,常用的方法还有空间插值法,其中又细分为点插值法和面插值法。面插值法中常用的面积权重法,需基于人口在基本行政单元中均匀分布的前提假设,仅能完成数据形式的转换,未能考虑人口分布中的空间信息;而空间插值法产生的人口空间分布数据空间分辨率较低,一般只能达到千米级的精度,难以满足更小细粒度的城市人口研究。

随着遥感图像和地理信息技术的快速发展,许多研究开始利用遥感图像提取出的人口相关信息,建立统计模型估计人口数量。统计模型方法需首先提取与人口空间分布相关的特征,再建立统计模型将所提取的特征与人口数量之间的关系表征化^[38],从而实现人口数量估计。

统计建模方法和人口密度模型方法、空间插值方法有所不同,后两者一般是利用人口统计数据进行空间化,因此人口总量数据与统计数据一致,但是时空分辨率较低。而统计建模方法中,由于统计调查数据会参与建模过程,最终的人口分布估计结果可能与实际人口分布情况存在偏差,但却能有效解决利用多源数据进行人口建模的难题,通过融合多种不同空间细粒度的数据,在一定程度上提高空间分辨率。目前,以统计建模的方法可实现的人口空间分辨率在25米×25米到1千米×1千米不等^[17]。

伴随大数据技术的迅猛发展,新的数据源不断涌现,例如社交网络的用户签到数据、POI数据、出租车轨迹数据等,人口空间化的研究方法出现了新的范式。一些研究者利用单一数据源,结合机器/深度学习算法研究人口分布;另一些研究者将社交媒体数据与遥感图像数据等结合,预测人口的时空分布。例如,基于大众点评的餐厅数据、刷卡消费数据与手机信令数据等,运用LASSO回归算法在不同的空间尺度上(边长范围为0.5千米—5千米的正方形网格)预测了北京市的昼夜人口分布、公司数量以及消费水平,并在成都、沈阳等其他8个不同的城市实现了跨城市迁移预测^[30];基于15个法国城市的OpenStreetMap数据,使用卷积神经网络(CNN)计算了一组城市特征,并结合1千米×1千米的人口数据,估算出200米×200米的人口空间数据^[39];利用遥感图像以及POI数据基于随机森林回归模型绘制北京市25米×25米的人口空间分布^[40];基于夜间灯光、兴趣点及土地利用等多源地理信息数据分区建模,运用机器学习算法开展了以广州市中心六区为研究区域的150米×150米的人口分布格网模拟研究,模拟结果与街(镇)实际人口的相关系数为0.834,决定系数 R^2 为0.695^[41]。

2. 动态的人口空间分布

人口的空间分布在时间维度上具有动态性,人口空间分布的时间动态规律对于城市公共

安全与危机管理、流动人口分布、城市活力研究具有重要意义^[42]。其中，对城市人口精细时间尺度上的昼夜分布动态分析较为成熟^[43]。带有时间信息的轨迹数据^[44]、社交媒体数据、手机信令数据^[45-47] 等使得人口空间分布的时间分辨率精细化成为可能。表 2 列举了含有人口时空分布估计的典型研究。

表 2 动态的人口空间分布研究

代表性研究以及主要贡献	数据源
以土地利用数据为媒介，构建城市人口、昼夜、空间的关系模型，同时基于多种面积权重构建人口空间化模型 ^[48]	2010 年北京市人口普查数据以及土地利用等空间数据
利用手机数据基于传统的统计建模方法估计动态人口分布 ^[15]	人口普查数据、电话数据集：2007 年 8 月、2007 年 11 月、2008 年 6 月（葡萄牙）；2007 年 5—10 月（法国）
使用了 BP 神经网络从手机通话记录中重建人类轨迹，利用对数线性回归来估计每小时人口密度 ^[16]	2015 年 12 月 6 日北京用户手机通话记录数据以及 2010 年北京市人口普查数据
提出了全时序列动态人口分布，并通过深度神经网络（DeepDPM）实现了动态人口分布映射 ^[17]	2017 年 7 月 1—31 日上海市移动通讯记录数据、对应时间段 POI 数据
提出了一种基于锚点的轨迹重建方法来重建手机轨迹，实现动态人口分布的估计 ^[49]	2010 年上海市人口普查数据、某个工作日的手机信令数据
利用了 POI 数据、高分辨率的卫星图像以及土地覆盖数据建立了高空间分辨率的人口密度模型，并绘制了北京奥运村的夜间人口分布 ^[50]	2015 年的人口统计调查数据、高分辨率立体图像、2015 年年底奥运村房价数据、2010 年土地覆盖数据

四、人口性别、年龄、种族、贫富等特征的精细空间尺度估计

除人口数量外，性别、贫富、年龄、职业等其他人口社会经济特征的空间分布对于城市精细化管理也有重要作用^[51]。例如，年龄和性别是影响疾病负担估算和死亡率的重要人口社会经济特征，也是公共服务和公共设施空间配置的重要决策依据。研究表明，可以利用社交媒体数据（社交软件中的用户数据，博客帖子等）预测研究地区的性别、年龄、种族和教育程度等分布，并在精细的空间分辨率下绘制其分布图^[52]。在人口数量的精细空间尺度估计方法的启发下，一些学者开始使用机器学习或深度学习算法推测精细空间尺度的人口社会经济特征。例如，基于美国人口普查中的两个较小的行政普查单位 block group 和 block，通过使用含有地理标签的博文来推测美国的性别和种族分布。作者在 block group 尺度下得到性别分布的平均相关性为 0.671，种族分布的平均相关性为 0.692，在 block 空间分辨率下的性别分布和种族分布的平均相关性为 0.585 和 0.487^[52]；利用预训练的 CNN 和 Google 街景图像来估计在 Layer Super Output Area（一个 LSOA 平均人口数量为 1614 人）空间分辨率下的伦敦市人类福祉指标分布，人类福祉指标具体包括收入、就业、教育等人口社会经济特征^[34]。

但是，由于性别、年龄、种族和受教育水平等特征没有直接的视觉特征（例如夜间灯光的饱和程度、住房条件和生活环境等），因此这些人口社会经济特征研究中很少直接使用图像作为预测数据源。目前在相关研究中，大多需使用社交媒体中的用户推文和个人资料进行文本分析并提取变量，但是囿于各个国家的语种以及用户语言习惯的差异，利用社交媒体数据估计性别、年龄、种族和受教育水平等人口社会经济特征仍然处于起步阶段。

五、城市人口特征精细时空尺度估计的成果

经过 30 余年的发展,城市人口特征精细尺度时空估计取得了丰硕成果^[53]。较成熟的精细尺度人口特征数据集包括 Gridded Population of the World (GPW)、Global Rural-Urban Mapping Project (GRUMP)、LandScan、WorldPop、中国人口空间分布千米网格数据集等。这些人口数据集都按一定周期在网站更新,为人口学、城市管理、人文地理、资源与环境等领域的学者提供精细时空尺度的人口数据。

GPW 数据集(<https://sedac.ciesin.columbia.edu/data/collection/gpw-v3/sets/browse>):由哥伦比亚大学和 NASA 共同完成,主要提供了 2000—2020 年全球的人口空间数据,目前该数据集能达到的最高空间分辨率为 30 弧秒,相当于赤道处 1 千米。该数据集主要是通过面积权重法将全球人口统计数据进行空间化。

GRUMP 数据集(<https://sedac.ciesin.columbia.edu/data/collection/grump-v1/sets/browse>):同样由哥伦比亚大学完成,该数据集在 GPW 数据集的基础上更加注重城乡差异化。目前 GRUMP 数据集一共包含了人口数量网格数据集、人口密度数据集、城市居住点等 8 个全球数据集。

LandScan 数据集(<https://www.satpalda.com/product/landscan/>):由美国橡树岭国家实验室主持完成,主要提供了 1998—2019 年间 1 千米×1 千米的网格化人口数据集。该数据集主要是通过 Dasymetric 方法对国家行政边界内人口数量进行空间化。

WorldPop 数据集(<https://www.worldpop.org>):由美国佛罗里达大学所主持,主要提供了 2000—2020 年中南美洲、非洲和亚洲人口空间分布数据集。此外,除了人口数量的数据,WorldPop 还囊括了精细空间尺度上部分地区的人口性别、年龄、人口迁移等数据。该数据集所支持的精度有 100 米×100 米和 1 千米×1 千米两种,是基于人口普查数据、夜间灯光数据以及土地利用数据,再结合随机森林模型实现的人口空间化^[54]。

中国人口空间分布千米网格数据集(<http://www.resdc.cn/DOI>):由中国科学院地理科学与资源研究所完成。该数据集的空间分辨率为 1 千米×1 千米,时间覆盖范围为 1990—2015 年。主要依据土地利用数据、夜间灯光数据、居民点密度的人口分布权重数据,基于多因子权重分配法绘制人口空间数据集。

上述 5 种人口数据集,只有 LandScan 为商业数据集无法免费获得,其余数据集均可在官网免费下载。各数据集的对比情况详见表 3。

尽管研究成果丰富,但这些精细时空尺度的人口特征数据集,被“数据生产者”之外的人口学及相关领域研究者广泛应用,还面临一些困境:一方面,这些数据带有空间坐标信息,其数据存储类型(比如.tif)决定了数据的输出和解读有赖于 ArcGIS 等地理信息软件,而这些软件对人口学等学科的学者来说并不常用。因此,数据使用过程中根据研究区域对数据集的进一步清理、再加工,成为数据使用无形的壁垒。另一方面,一些数据集的输出是全球范围内的,数据量大,研究者根据自己的研究需要进行数据提取存在难度。以 WorldPop 的年龄数据为例,数据呈现单元为国家或地区,而非更具体的省份、城市等,数据所占存储空间大(单中国大陆地区的数据近 20G)。另外其数据切割方式,难以适应学者以区域为导

表 3 典型人口精细尺度空间分布数据集对比

数据集	时间分辨率	空间分辨率	基础数据	绘制方法
GPW	2000—2020 年	1 千米×1 千米	人口普查数据和相应的地理界线	面积权重法
GRUMP	1990—2010 年	1 千米×1 千米	夜间灯光和辅助数据	构建城乡掩膜，配置人口
LandScan	1998 年， 2000—2019 年	1 千米×1 千米	人口普查数据和地理数据	基于权重模型绘制人口分布
WorldPop	2000—2020 年	1 千米×1 千米， 100 米×100 米	人口普查数据、夜间灯光数据以及土地利用数据	随机森林、Dasymetric
中国人口空间分布千米网格数据集	1990—2015 年	1 千米×1 千米	土地利用数据，夜间灯光数据，辅助数据以及人口普查数据	多因子权重分配法

向的研究。例如，中国大陆地区 100 米×100 米网格的年龄数据，按男性和女性分 2 类，又按年龄区间划分了 18 类。性别与年龄段组合后，该数据集包括 36 个子数据集，每个数据集的大小约 0.5G。这样的数据切割方式下，如果学者拟研究中国某城市不同年龄人群的时空分布，需下载约 18G 的数据集，再进行数据合并然后按研究区域裁剪等，工作量较大且对跨专业技能要求较高。因此，在这些精细尺度数据集的生产和终端使用之间，还需要对数据的再加工，输出更加终端友好性的数据集。

近年来，众多国内外学者在上述数据集的基础上，根据自己的需求绘制了特定地区时空分辨率的人口数据集。例如，王明明等利用 WorldPop 数据集、夜间灯光数据、街道/乡镇级人口统计数据和行政边界数据绘制了 2000 年和 2010 年山东省街道/乡镇级格网化人口空间分布数据集^[55]；叶婷婷等利用夜间灯光数据、POI 数据和 WorldPop 人口数据集，结合随机森林模型获取了北京、上海、广州和重庆的 2010 年 100 米×100 米的空间化数据集^[14]；杨晓荣等以多源数据、WorldPop 以及中国公里网格人口分布数据集为基础数据，在县级尺度上采用了空间回归模型构建 2015 年福建省 500 米×500 米人口空间分布数据^[56]；郭山山等采用中国公里网格人口分布数据集以及浙江省的乡镇级行政区域面属性数据，乡镇级的建筑物数据和灯光数据，利用加权计算的方法，得到了浙江省 250 米×250 米的网格人口分布数据^[57]。

六、精细时空尺度城市人口特征数据的应用

精细时空尺度的城市人口社会经济特征数据是城市问题解决、公共危机应对与公共风险管理的决策依据，伴随城市化的发展，城市内部由于医疗、教育等公共服务的均等性造成的社会可持续问题出现，制约着以人为本、改善民生导向下的宜居和谐城市建设。而更精细的人口分布数据有助于发现并解决资源配置的空间公平性方面的问题，促进城市可持续发展。同时，高分辨率的城市人口空间分布地图，也有助于建设韧性城市，提高风险应对能力。比如在灾害发生后，精准的空间定位和受难人数估计，对更准确地进行灾难影响评估、实现灾后救助有重要作用。我国抗击新冠肺炎疫情的经验也说明，掌握精细尺度上的人口活动轨迹数据，实现精准防控，对有效应对突发公共卫生事件起到重要作用。另外，人口社会经济特征及其在空间精细尺度上的分布数据对于制定合理的地方经济发展政策和减少区域之间的不平等也具有重要的决策支持作用。

1. 精细时空尺度人口特征研究的热度及学科分布

为了解精细时空尺度人口特征估计的学科及期刊分布, 对相关学术成果展开文献计量分析。英文文献主要通过 Web of Science 数据库进行检索, 检索词包括 “population prediction/estimation”、“high resolution”、“mapping”、“machine learning/remote sensing” 等关键词及其组合, 中文文献主要是在中国知网 (CNKI) 进行搜索, 检索词包括 “人口”、“空间化/空间分布”、“遥感/多源数据”、“预测” 等关键词以及其组合。发表年份设定为 2010—2020 年 (其中 2020 年数据截至 10 月 21 日前在线发表), 文献类型为 article 或 review。通过关键词搜索初步得到相关学科范围内相应研究主题的 896 篇相关文献, 运用 CiteSpace 软件展开文献计量分析 (见图 2)。

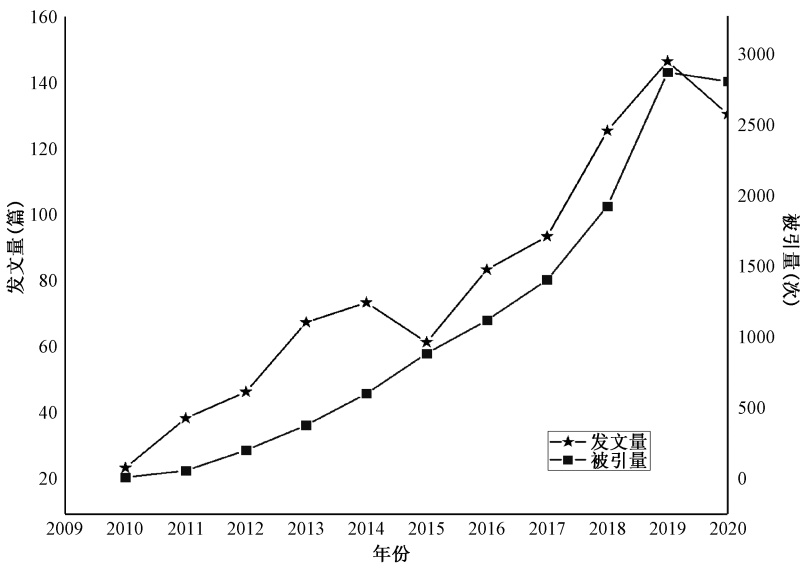


图 2 人口特征精细尺度估计的文献数量及被引量

统计数据显示: 人口社会经济特征的精细尺度估计/预测的研究正在受到越来越多学者的关注。从文章数量来看, 自 2010 年开始, Web of Science 数据库中的发文数量整体呈快速增长态势 (2015 年除外), 从 2010 年的 20 余篇增加至 2020 年的 120 余篇。此外, 这些研究成果的被引量也呈稳步上升态势, 2017 年以后增长迅速。其中, 文章的所发表的期刊主要有 *Nature*、*Science*、*PNAS*、*Scientific Data*、*Landscape and Urban Planning*、*Cities*、*Urban Studies*、*Remote Sensing* 等。涉及的学科较广, 包括环境科学、生态学、测绘学、地质学、地理科学、计算机科学、城市学等不同学科。

2. 精细时空尺度人口数据的应用场景

为探究精细时空尺度人口数据的应用场景, 本文对相关文献进行了关键词共现分析, 其中时间切片为 1 年, 显示的关键词频设置为 10 个及以上, 得到该主题下 896 篇文献的关键词共现图谱, 该图谱共包含 356 个节点, 1374 条边 (见图 3)。分析发现: 相关研究所基于的情境不局限于欧美发达国家, 中国、印度等亚洲国家以及部分非洲国家等发展中国家背景

下的相关研究也在不断涌现。人口特征的精细尺度分布作为资源分配的基础工作,应用主题越来越丰富,包括疾病预防及控制、电力消费空间格局分布、空气污染的人因分析、空气污染对人的身心健康和人类活动影响、碳排放空间估计及格局分析、贫困及收入差距的识别、韧性城市导向下基于人口的城市风险及脆弱性识别、城市发展演化过程回溯、城市物理及社会结构分析、城市社区及邻里的重塑等。



图3 人口社会经济特征的空间估计的相关文献关键词共现

具体说来, 在环境领域中, 慕航等利用 1999—2016 年 PM 2.5 数据集, 结合精度为 1 千米×1 千米的人口栅格数据, 分析了“一带一路”沿线 65 个国家 PM 2.5 污染与暴露风险的时空变化特征^[58]; 梁昌一等采用 LandScan 全球人口空间分布栅格数据和 2000—2016 年 PM2.5 空间分布栅格数据, 以及中国城市层面的经济统计面板数据, 采用 OLS 以及工具变量法识别了城市空间发展模式对雾霾污染的影响^[59]; 江伟康等以粤港澳大湾区为例, 运用 InVEST 模型与 CA-Markov 模型研究大湾区 2005—2030 年生态环境质量的时空演变特征, 并结合 1 千米×1 千米的人口分布数据探究各地区城市发展与生态环境质量的权衡关系, 预测了 2030 年大湾区不同情景下生态环境质量状况^[60]; 有学者提出了一个通用框架, 该框架同时考虑了影响二氧化碳排放量的各个因素 (城市单元中的人口和面积) 之间的相关性, 揭示了城市人口、面积与二氧化碳排放量的耦合关系^[61]。

在经济方面,王好等利用 LandScan 人口数据和夜间灯光数据,以全球 132 个国家为研究对象,采用面板固定效应模型和两阶段最小二乘法探讨城市人口规模分布对地区收入差距

的影响^[62];基于中国 306 个城市的网格化人口数据,通过探索性空间数据分析确定其城市(副)中心,构建衡量其多中心度和分散度的指标,并对空间结构关于城市生产力的影响进行建模^[63]。

在城市发展与规划方面,黄一琦等基于 POI 数据对零售业进行空间热点识别,获取热点聚集特征,再基于人口空间分布数据、出行轨迹数据获得人口分布特征以及人群集聚特征,最后对比分析厦门市零售业空间冷热点与人口分布特征的匹配程度,以期优化厦门市零售业布局^[64];基于不同细粒度的人口数据对中国 22 个城市区域的多中心城市发展进行了多个地理尺度的多中心发展模式对比分析^[65];基于夜间灯光图像和人口分布数据,量化了 1990 年至 2010 年中国不同城市的城市扩张^[66]。

七、人口特征精细尺度时空估计:机遇与挑战

城市人口特征的时空尺度精细化无论从深度还是广度来说都积累了丰硕的成果,也将迎来更多的发展机遇:开源数据的更新频率加快,将有助于人口分布数据在时间维度精细化,更加深入地揭示人口分布的动态变化规律。在人口空间分布的时间维度回溯性方面,未来的研究将重点关注历史时期人口分布精度评估,回答现代的方法在多大程度上适用于历史时期人口的精细时空尺度分布格局重建。未来除了继续在更高的时空分辨率上对人口数量、财富水平等人口社会经济特征进行推断之外,也将尝试并完善性别、民族、年龄等其他人口特征的推断。另外,在同一时空分辨率上,利用建模方法和多源数据同时对多个人口社会经济学特征进行预测,使得不同的特征具有匹配性、可比性,也是研究者们目前比较关心的热点话题。

人口特征的精细时空尺度估计也面临着挑战:第一,多源数据的有效融合问题。考虑到单一数据源包含的语义信息匮乏,目前许多研究者尝试通过多源数据融合预测人口社会经济特征的时空分布,例如将遥感数据和兴趣点等数据融合。但是由于多源数据的获取时间、获取方式和相应的测量误差有所不同,简单的数据叠加可能会降低信息的准确性。因此,不同数据源的有效融合方式仍有待完善。第二,网格单元的适宜性问题。在精细尺度的人口空间分布研究中,网格单元的大小是人口社会经济特征分布数据在精细尺度上精确与否的决定性因素,其适宜性的确定是关键问题,但是,目前的许多研究关于其网格大小的选择大多都是根据经验进行选择的,其适宜性以及网格大小选取的合理性有待考证。第三,深度学习方法的泛化性问题。目前很多研究都证明了卷积神经网络等方法可以直接用于提取与人口有关的社会经济特征,但是这种方法需要大量有标注的训练集。这些数据在一些发达国家容易获得,但在一些数据基础较差的国家却较难获得,给这些国家的人口社会经济特征的精细尺度时空推断带来困难。如何提高深度学习等前沿方法的泛化性是未来需解决的难点问题,即训练基础数据完备国家的模型如何应用于基础数据缺乏的国家,通过跨区域的迁移学习,使城市人口特征估计/预测能够通过预训练的网络加少量目标区域中有标注的数据实现。

参考文献:

- [1] WARDROP N A, JOCHEM W C, BIRD T J, et al. Spatially disaggregated population estimates in the absence of national population and housing census data [J]. *Proceedings of the National Academy of Sciences*, 2018, 115 (14): 3529–3537.
- [2] 董南, 杨小唤, 蔡红艳, 等. 人口数据空间化研究进展 [J]. *地球信息科学学报*, 2016 (10): 1295–1304.
- [3] 柏中强, 王卷乐, 杨飞, 等. 人口数据空间化研究综述 [J]. *地理科学进展*, 2013 (11): 1692–1702.
- [4] NIEBERGALL S, LOEW A, MAUSER W. Integrative assessment of informal settlements using VHR remote sensing data: the Delhi case study [J]. *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing*, 2009, 1 (3): 193–205.
- [5] DEHURY B, MOHANTY S K. Regional estimates of multidimensional poverty in India [R], 2015.
- [6] AZAR D, GRAESSER J, ENGSTROM R, et al. Spatial refinement of census population distribution using remotely sensed estimates of impervious surfaces in Haiti [J]. *International Journal of Remote Sensing*, 2010, 31 (21): 5635–5655.
- [7] NOOR A M, ALEGANA V A, GETTING P W, et al. Using remotely sensed night-time light as a proxy for poverty in Africa [EB/OL]. (2008–10–21) [2021–12–07]. <http://doi.org/10.1186/1478-7954-6-5>.
- [8] CHEN X, NORDHAUS W D. Using luminosity data as a proxy for economic statistics [J]. *Proceedings of the National Academy of Sciences*, 2011, 108 (21): 8589–8594.
- [9] BLUMENSTOCK J, CADAMURO G, ON R. Predicting poverty and wealth from mobile phone metadata [J]. *Science*, 2015, 350 (6264): 1073–1076.
- [10] POKHRIYAL N, JACQUES D C. Combining disparate data sources for improved poverty prediction and mapping [J]. *Proceedings of the National Academy of Sciences*, 2017, 114 (46): E9783–E9792.
- [11] ZENG C, ZHOU Y I, WANG S, et al. Population spatialization in China based on night-time imagery and land use data [J]. *International Journal of Remote Sensing*, 2011, 32 (24): 9599–9620.
- [12] QIE Z J, RONG L L. Spatial-temporal human exposure modeling based on land-use at a regional scale in China [J]. *Safety Science*, 2016, 87 (4): 243–255.
- [13] DUQUE J C, PATINO J E, RUIZ L A, et al. Measuring intra-urban poverty using land cover and texture metrics derived from remote sensing data [J]. *Landscape and Urban Planning*, 2015, 135 (11): 11–21.
- [14] YE T, ZHAO N, YANG X, et al. Improved population mapping for China using remotely sensed and points-of-interest data within a random forests model [J]. *Science of The Total Environment*, 2019, 658 (12): 936–946.
- [15] DEVILLE P, LINARD C, MARTIN S, et al. Dynamic population mapping using mobile phone data [J]. *Proceedings of the National Academy of Sciences*, 2014, 111 (45): 15888–15893.
- [16] LIU Z, MA T, DU Y, et al. Mapping hourly dynamics of urban population using trajectories reconstructed from mobile phone records [J]. *Transactions in GIS*, 2018, 22 (2): 494–513.
- [17] ZONG Z, FENG J, LIU K, et al. DeepDPM: dynamic population mapping via deep neural network [C]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.
- [18] NAIK N, KOMINERS S D, RASKAR R, et al. Computer vision uncovers predictors of physical urban change [J]. *Proceedings of the National Academy of Sciences*, 2017, 114 (29): 7571–7576.
- [19] FAN Z, DING Z, YU L, et al. Representing place locales using scene elements [J]. *Computers, Environment Urban Systems*, 2018, 71 (11): 153–164.
- [20] HU L, HE S, HAN Z, et al. Monitoring housing rental prices based on social media: an integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies [J]. *Land Use Policy*, 2019, 82 (12): 657–673.
- [21] CHEN Y, LIU X, LI X, et al. Mapping the fine-scale spatial pattern of housing rent in the metropolitan area by using online rental listings and ensemble learning [J]. *Applied Geography*, 2016, 75 (8): 200–212.

- [22] ZHAO X, YU B, LIU Y, et al. Estimation of poverty using random forest regression with multi-source data: a case study in Bangladesh [J]. *Remote Sensing*, 2019, 11 (4): 375–392.
- [23] ARIMOND A. A distributed system for pattern recognition and machine learning [R], 2012.
- [24] LECUN Y, BENGIO Y, HINTON G. Deep learning [J]. *Nature*, 2015, 521: 436–444.
- [25] REICHSTEIN M, CAMPS-VALLS G, STEVENS B, et al. Deep learning and process understanding for data-driven earth system science [J]. *Nature*, 2019, 566: 195–204.
- [26] SCHMIDHUBER J. Deep learning in neural networks: an overview [J]. *Neural Networks* 2015, 61 (9): 85–117.
- [27] LO C. Population estimation using geographically weighted regression [J]. *GIScience & Remote Sensing*, 2008, 45 (2): 131–148.
- [28] QIU F, SRIDHARAN H, CHUN Y. Spatial autoregressive model for population estimation at the census block level using LIDAR-derived building volume information [J]. *American Cartographer*, 2010, 37 (3): 239–257.
- [29] STEVENS F R, GAUGHAN A E, LINARD C, et al. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data [J]. *PLoS One*, 2015, 10 (2): e0107042.
- [30] DONG L, RATTI C, ZHENG S. Predicting neighborhoods’ socioeconomic attributes using restaurant data [J]. *Proceedings of the National Academy of Sciences*, 2019, 116 (31): 15447–15452.
- [31] LI G, CAI Z, LIU X, et al. A comparison of machine learning approaches for identifying high-poverty counties: robust features of DMSP/OLS night-time light imagery [J]. *International Journal of Remote Sensing*, 2019, 40 (15–16): 5716–5736.
- [32] JEAN N, BURKE M, XIE M, et al. Combining satellite imagery and machine learning to predict poverty [J]. *Science*, 2016, 353 (6301): 790–794.
- [33] GEBRU T, KRAUSE J, WANG Y, et al. Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States [J]. *Proceedings of the National Academy of Sciences*, 2017, 114 (50): 13108–13113.
- [34] SUEL E, POLAK J W, BENNETT J E, et al. Measuring social, environmental and health inequalities using deep learning and street imagery [J]. *Scientific Reports*, 2019, 9 (1): 1–10.
- [35] 刘爱华, 邹哲, 刘森. 基于人口密度模型的大都市空间结构演化——以天津市为例 [J]. *城市发展研究*, 2015 (3): 141–144.
- [36] 康停军, 张新长, 赵元, 等. 基于多智能体的城市人口分布模型 [J]. *地理科学*, 2012 (7): 790–797.
- [37] 陈浩, 权东计, 赵新正, 等. 西部欠发达城市人口空间分布与演变——以兰州市为例 [J]. *世界地理研究*, 2019 (4): 105–114.
- [38] WU S S, QIU X M, WANG L. Population estimation methods in GIS and remote sensing: a review [J]. *Mapping Sciences & Remote Sensing*, 2005, 42 (1): 80–96.
- [39] GERVASONI L, FENET S, PERRIER R, et al. Convolutional neural networks for disaggregated population mapping using open data [C]. 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 2018: 594–603.
- [40] LI K, CHEN Y, LI Y. The random forest-based method of fine-resolution population spatialization by using the international space station nighttime photography and social sensing data [J/OL]. *Remote Sensing*, 2018, 10 (10): 1650; <http://doi.org/10.3390/rs10101650>.
- [41] 成方龙, 赵冠伟. 分区策略与机器学习的人口分布精细化模拟 [J]. *测绘科学*, 2020 (9): 165–173.
- [42] 张善余. 人口地理学概论 [M]. 上海: 华东师范大学出版社, 2013.
- [43] 陈慕琳. 基于互联网挖掘信息及建筑数据估算的精细尺度昼夜人口空间格局分析 [D]. 武汉: 华中师范大学, 2019: 52–56.
- [44] YU B, LIAN T, HUANG Y, et al. Integration of nighttime light remote sensing images and taxi GPS tracking data for

- population surface enhancement [J]. *International Journal of Geographical Information Science*, 2019, 33 (4): 687–706.
- [45] BECKER R, C  CERES R, HANSON K, et al. Human mobility characterization from cellular network data [J]. *Communications of the Acm*, 2013, 56 (1): 74–82.
- [46] BENGTTSSON L, LU X, THORSON A, et al. Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti [J]. *Plos Medicine*, 2011, 8 (8): e1001083.
- [47] MOUMNI B, FRIASMARTINEZ V, FRIASMARTINEZ E. Characterizing social response to urban earthquakes using cell-phone network data: the 2012 Oaxaca earthquake [J]. *Enriquefrias*, 2013: 1199–1208.
- [48] 戚伟, 李颖, 刘盛和, 等. 城市昼夜人口空间分布的估算及其特征——以北京市海淀区为例 [J]. *地理学报*, 2013 (10): 1344–1356.
- [49] LI M, ZHANG H, CHEN J. Fine-grained dynamic population mapping method based on large-scale sparse mobile phone data [C]. 2019 20th IEEE International Conference on Mobile Data Management (MDM), 2019: 473–478.
- [50] XU M, CAO C, JIA P. Mapping fine-scale urban spatial population distribution based on high-resolution stereo pair images, points of interest, and land cover data [J/OL]. *Remote Sensing*, 2020, 12 (4): 608; <http://doi.org/10.3390/rs12040608>.
- [51] XIE M, JEAN N, BURKE M, et al. Transfer learning from deep features for remote sensing and poverty mapping [C]. Thirtieth AAAI Conference on Artificial Intelligence, 2016: 3929–3935.
- [52] MONTASSER O, KIFER D. Predicting demographics of high-resolution geographies with geotagged tweets [C]. Thirty first AAAI Conference on Artificial Intelligence, 2017: 1460–1466.
- [53] BALK D L. Determining global population distribution: methods, applications and data [J]. *Advances in Parasitology*, 2006, 62 (1): 119–156.
- [54] GAUGHAN A E, STEVENS F R, HUANG Z, et al. Spatiotemporal patterns of population in mainland China, 1990 to 2010 [J]. *Scientific Data*, 2016, 3 (1): 16005.
- [55] 王明明, 王卷乐. 山东省街道/乡镇级格网化人口空间分布数据集 (2000, 2010 年) [J]. *中国科学数据 (中英文网络版)*, 2019 (3): 182–192.
- [56] 杨晓荣, 陈楠. 基于多源数据的福建省人口数据空间化研究 [J]. *贵州大学学报 (自然科学版)*, 2019 (2): 79–84.
- [57] 郭山山, 龚俊, 尹晶飞. 基于 DMSP/OLS 的人口分布网格精细化研究 [J]. *地震研究*, 2016 (2): 321–326.
- [58] 慕航, 何超, 阮秋明, 等. “一带一路”沿线国家 PM_{2.5} 污染与人口暴露风险的时空分布特征 [J]. *环境科学学报*, 2020 (6): 2229–2240.
- [59] 梁昌一, 刘修岩, 李松林, 等. 城市空间发展模式与雾霾污染——基于人口密度分布的视角 [J]. *经济学动态*, 2021 (2): 80–94.
- [60] 江伟康, 吴隽宇. 基于地区 GDP 和人口空间分布的粤港澳大湾区生境质量时空演变研究 [J]. *生态学报*, 2021 (5): 1747–1757.
- [61] RIBEIRO H V, RYBSKI D, KROPP J P. Effects of changing population or density on urban carbon dioxide emissions [J]. *Nature Communications*, 2019, 10 (1): 1–9.
- [62] 王妤, 孙斌栋. 城市规模分布对地区收入差距的影响——基于 LandScan 全球人口数据库的实证研究 [J]. *城市发展研究*, 2021 (6): 25–32.
- [63] LI Y, LIU X. How did urban polycentricity and dispersion affect economic productivity? a case study of 306 Chinese cities [J]. *Landscape Urban Planning*, 2018, 173 (1): 51–59.
- [64] 黄一琦, 许俊萍. 基于大数据的零售业空间热点识别及人口分布特征分析——以厦门市为例 [J]. *中外建筑*, 2020 (12): 107–111.
- [65] LIU X, DERUDDER B, WANG M. Polycentric urban development in China: a multi-scale analysis [J]. *Environment Planning*, 2018, 45 (5): 953–972.

- [66] XU G, JIAO L, YUAN M, et al. How does urban population density decline over time? an exponential model for Chinese cities with international comparisons [J]. *Landscape and Urban Planning*, 2019, 183 (11): 59-67.

Fine Spatio-temporal Scale Estimation of Urban Population's Socio-economic Characteristics Based on Big Data: Data, Methals and Applications

LIU Yunxia^{1,2}, TIAN Tian³, GU Jiayu¹, LIU Junmin³

(1. Zhou Enlai School of Government, Nankai University, Tianjin 300350, China;

2. Lab of Computational Social Sciences, Nankai University, Tianjin 300350, China;

3. School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: With the emergence of spatio-temporal big data and the development of data mining technology, scholars have made great progress in estimation of urban population's socio-economic characteristics on fine spatio-temporal scale, which makes up for the shortcomings of the traditional data with poor timeliness and low spatial precision effectively. However, related studies are restricted in urban informatics, computational social science, geography and other fields, which have not been widely applied to the resource allocation based on spatio-temporal distriution of urban population. This problem restricts the in-depth practice of urban fine management. In order to promote information communication and interdisciplinary dialogue, this paper prospected the research prospects and challenges according to the analysis of data source, estimation methodology, result verification and model evaluation, mature fine scale population characteristic distribution data set. The results found that spatial regression, machine learning and neural network can be used in the fine spatio-temporal scale estimation of urban population characteristics. Among many socio-economic characteristics, research on fine scale population distribution is the most developed, which has transformed from single data source to multiple data sources, from general grid to grid with spatial information, and from low spatio-temporal resolution to high spatio-temporal resolution. The gender, age, race and education of urban population can also be predicted based on social media data, but it is not mature. The application of fine scale urban population characteristics data is rich, which includes disease prevention and control, poverty identification, climate change mitigation, urban risk management, urban development backtracking, urban social spatial structure analysis, etc.

Keywords: urban population; spatial-temporal distribution; social spatial form; machine learning; neural network; computational social science

[责任编辑 刘爱华]