International Labour Organization

国际劳工局工作论文 165

三月 / 2026

X

格陵达遥感与随机森林技术在加纳的网格化劳动力市场数据分析

作者 / 金炎，查普·马蒂厄，梅杨，李泽溯

© 国际劳工组织 2026

署名4.0国际（CC BY 4.0）

本作品受Creative Commons Attribution 4.0 International许可。详见： https:// creativecommons.org/licenses/by/4.0/ 用户可以根据许可协议的规定，重复使用、分享（复制和再分发）、改编（混搭、改编和在此基础上进行创作）这些内容。用户必须明确注明国际劳工组织（ILO）是该材料的来源，并说明是否对原始内容进行了修改。禁止在翻译、改编或其他衍生作品中使用国际劳工组织的徽章、名称和标志。

归属 用户必须指明是否进行了修改，并按照以下方式引用作品：Yan, J.，Matthieu, C.，Yang, M.，Zeshuo, L. *格陵达遥感与随机森林技术在加纳的网格化劳动力市场数据分析* 国际劳工组织工作论文165。日内瓦：国际劳工局，2026年。© 国际劳工组织。

翻译。
- 如果对本作品进行翻译，必须在署名信息之外，同时添加以下免责声明： *本译文是国际劳工组织（ILO）受版权保护的作品的翻译版本。本翻译未经国际劳工组织准备、审查或批准，不应被视为国际劳工组织的官方翻译。国际劳工组织对本译文的内容和准确性不承担任何责任。本译文的责任完全由译者承担。*

适应 如对本作品进行改编，必须添加以下免责声明，并附上归属权声明： *这是一份国际劳工组织（ILO）版权作品的改编版本。此改编版本未经ILO准备、审核或认可，不应被视为官方ILO改编。ILO对其内容的准确性和完整性概不负责，责任完全由改编版本作者承担。*

第三方资料 此Creative Commons许可证不适用于本出版物中包含的非国际劳工组织版权材料。如果材料归因于第三方，使用此类材料的人仅对与版权所有者协商权利并对其侵权索赔负责。

任何根据本许可证产生的、无法友好解决的争议，应提交联合国国际贸易法委员会（UNCITRAL）仲裁规则进行仲裁。各方应受此类仲裁产生的任何仲裁裁决的约束，作为此类争议的最终裁决。

关于版权和许可的详细信息，请联系 rights@ilo.org 。有关国际劳工组织出版物和数字产品的详细信息，请访问： www.ilo.org/publns 。

ISBN 9789220432976（印刷版），ISBN 9789220432983（网络PDF版），ISBN 9789220432990（epub版），ISBN 9789220433003（html版）。ISSN 2708-3438（印刷版），ISSN 2708-3446（数字版）。

https://doi.org/10.54394/00033744

ILO出版物中使用的名称，这些名称符合联合国惯例，以及其中材料的呈现，并不意味着ILO对任何国家、地区或领土的法律地位有任何意见表达。

或者关于其权威，或者关于其边境或边界的界定。参见： www.ilo.org/免责声明 .

本出版物中表达的意见和观点是作者的观点，不一定反映国际劳工组织（ILO）的意见、观点或政策。

对企业和商业产品及流程名称的提及并不意味着ILO的认可，未提及特定企业、商业产品或流程也不代表不赞同。

信息关于国际劳工组织出版物和数字产品可以在以下网址找到： www.ilo.org/研究-与-出版物

ILO工作论文总结了正在进行的ILO研究的结果，旨在激发关于劳动世界各领域问题的讨论。欢迎对这篇ILO工作论文提出评论，评论可以发送至 charpe@ilo.org .

授权发表：施密特，多罗西亚

国际劳工组织工作论文可在以下网址找到：www.ilo.org/research-and-publications/working-papers

建议引用：
严，J.，马蒂厄，C.，杨，M.，泽硕，L. 2026. *格网化加纳劳动力市场数据：利用遥感技术和随机森林* ，国际劳工局工作报告 165 (日内瓦，国际劳工局)。 https://doi. org/10.54394/00033744

摘要

这项研究展示了高分辨率（0.005）网格化劳动力市场数据，通过使用随机森林算法和遥感技术对加纳地区层面的普查数据进行降尺度处理而生成。它通过绘制17个就业类别（包括年龄、性别、技能、状态、部门、失业和NEET）来弥补空间分解劳动力市场数据的不足。将辅助数据（64个变量）如土地覆盖、夜间灯光、基础设施和兴趣点整合，以捕捉人口、经济和参与因素。该模型实现了高精度（大多数类别R2 > 90%），并揭示了显著的空间异质性，就业率在像素间从10%到98%不等。结果突出了城乡和南北差异，以及部门集中。变量重要性分析强调了建成区、夜间灯光、道路密度和植被健康状况在预测就业模式中的作用，并在不同就业类别中表现出特异性。该方法通过纳入劳动力市场复杂性，超越了传统的GDP或人口网格化。研究结果展示了机器学习和地理空间数据在数据匮乏环境中增强社会经济地图制作的潜力。

关于作者

颜金，副教授，南京邮电大学物联网学院 & 江苏省智慧健康大数据分析与定位服务工程技术研究中心，中国南京 210023

马蒂厄·夏尔佩，高级经济学家，就业政策部，国际劳工组织，瑞士日内瓦1211，莫里永路4号

杨美 物联网学院，南京邮电大学，中国南京210023

李泽述 物联网学院，南京邮电大学，中国南京210023

目录

图表清单

目录表

X 引言

网格化人口和GDP数据是有效政府政策规划和国际援助分配的重要工具。网格化人口数据对于健康监控和城市发展至关重要，能够实现精确和有针对性的干预（Stevens等人，2015；Tatem，2017；Leyk等人，2019；Chen等人，2019；Lebakula等人，2025）。同时，网格化GDP数据使政策制定者能够监控次国家级别的经济发展，促进更明智和本地化的决策（Kummu等人，2018；Chen等人，2022；Jin等人，2023；Wu等人，2024）。

人口统计数据或社会经济属性存在显著差异。社会经济属性需要将经济活动度与人口联系起来，反映物质和人力资本的积累、聚集效应以及政治和经济制度的发展和质量。

网格化社会经济数据需要超越GDP，鉴于其众所周知的局限性以及通过劳动力市场指标、人类发展指数或贫困测量来衡量个人福祉的必要性。然而，很少有人尝试生产高分辨率的贫困测量（Jean等人，2016；Huber和Mayoral，2024）或性别发展指标（Bosco等人，2017），并且目前尚无现有的劳动力市场网格化数据。

在这篇论文中，我们提出了第一套高分辨率劳动力市场数据地图，该地图基于对加纳2021年住房与人口普查 district-level 统计数据的0.005度分解。降低劳动力市场统计数据的比例背后的理由是较低行政级别的巨大异质性。虽然国家层面的就业率为48%，但区级标准差为11.2个百分点，Nabdam地区区级的最低就业率为16.5%，Ashaiman地区区级的最高就业率为64.7%（见第2.1节和图1）。

高分辨率劳动力市场地图的一个困难是，没有单一的指标能够最好地描述它。虽然就业数量，即总就业人数，是最常讨论的变量，但劳动力市场挑战往往与特定子群体（青年、老年工人、女性）有关，与不充分就业（失业、NEET）相关。[1]），就业质量（技能、就业状况），或行业（农业、制造业、服务业）。就业情况也常以比率表示，需要估算工作年龄人口。[2]

第二难度在于劳资关系依赖于多个因素，而这些因素很难通过GIS和遥感数据捕捉。劳动力供应，即工作的或可工作的人员数量，与人口增长直接相关。然而，还必须考虑其他因素。经济活动水平是劳动力市场的一个重要决定因素，因为它确定了劳动力需求水平。特别是，将GDP的高分辨率地图进行部门分解（金等，2023年；吴等，2024年）是一个挑战。因此，缩放劳动力市场数据需要结合与人口和GDP分散化相关的方法。

[1] NEET代表未就业、未受教育、未接受培训。
[2] 就业率定义为总就业人数除以15岁及以上的劳动年龄人口。这项工作还包括在像素级别对人口类别进行细分（见第4.1节）。

一个关键于劳动力市场的第三个要素是基于个人决定的参与决策。个人参与劳动力市场的决策受到各种因素的影响，如教育水平、性别、家庭结构、工作经验年限、贫困状况或活动领域（参见Aaronson等人（2020年）；Klasen等人（2021年）关于女性劳动力市场参与；Baah-Boateng（2013年）；Amankwah等人（2022年）关于影响加纳劳动力市场参与的具体因素）。

该网格化就业数据集包含17个与就业相关的类别：总就业及其按年龄（青年、成年、老年）分解，按性别（男性、女性），按技能（高技能、低技能），按状态（自营、薪金），按行业（农业、矿业、制造业、工业和服务业），失业和NEET。收集的辅助数据包括64个类别，旨在捕捉人口统计学因素、经济活动水平以及参与劳动市场的决策。这些数据包括GIS数据、遥感数据和兴趣点。劳动力市场统计数据来自加纳2021年住房和人口普查的就业模块，并遵循2019年国际劳工统计学家会议的定义。该模型使用基于区级劳动力统计和辅助数据的随机森林算法进行训练。像素分解在0.005度的分辨率下进行。

方法产生栅格人口数据的速度迅速发展，从简单的面积加权方法——假设目标区域内部分布均匀——转变为利用GIS和遥感技术的丰富资源来估计网格级别人口的大地测量方法（Dobson等，2000；Bhaduri等，2007）。反过来，大地测量方法采用了一系列技术，从统计模型（Briggs等，2007；Chen等，2019）到机器学习（Azar等，2010；Stevens等，2015；Wang等，2020）。全球数据集，如LandScan（Lebakula等，2025）和WorldPop（Tatem，2017），[3] 阐述这些进步。除了降尺度方法之外，基于微观人口普查的从下而上的方法也可以有助于生产全国范围内的网格化人口数据（Wardrop等人，2018；Leyk等人，2019）。

试图超越人口统计数据，纳入社会经济指标的研究往往集中在GDP和贫困问题上。多年的全球GDP栅格数据依赖于测量人口和城市化的统计技术和辅助数据。Murakami等人（2021年）；Murakami和Yamagata（2019年）提出了一种以10年为频率的全球GDP降尺度方法，覆盖了1850年至2100年（以0.5°和1/12°的分辨率），利用梯度提升技术估计的权重。Wang和Sun（2022年）利用夜间灯光和每像素人口数据，将国家和地区的GDP数据投影到像素级别，覆盖了185个国家，同样以2005年起步，以10年为频率。

一国或某年GDP数据通常包含大量辅助数据并依赖机器学习。在生产函数方法中，GDP取决于人口统计、资本存量和科技进步。这需要衡量经济活动的辅助数据，比如夜间灯光，以及衡量规模经济或聚集效应的辅助数据——通过基础设施/机构距离来捕捉。陈等（2022年）开发了一个深度学习框架，通过整合多个地理空间大数据集，包括夜间灯光、土地利用、道路网络和兴趣点，对中国的GDP进行高空间分辨率映射。一个主要目标是模拟部门GDP（农业、制造业和服务业）来反映驱使部门经济活动的一系列辅助数据。部门的

[3] https://landscan.ornl.gov/ ； https://www.worldpop.org/

研究方法源于夜间光照对充分捕捉农业GDP的限制，因为农业生产增加主要是通过更高的光照强度间接反映的。Wu等人（2024）提出了一种多尺度融合残差网络模型，该模型结合了中国UMARYR地区的广泛遥感数据和信息点（POI）数据。类似地，Jin等人（2023）基于遥感和POI，提出泰国部门GDP的高分辨率地图，测量自然特征、人口统计、经济活动和部门维度。该模型使用随机森林区域到区域回归克里金进行估算。

与GDP并重，建模社会经济属性包括网格化贫困数据。Jean等人（2016）训练了一个卷积神经网络模型，通过结合夜间灯光数据和白天卫星图像来估计支出数据，这些图像可以作为贫困指标，如屋顶材料。

类似地，Huber和Mayoral（2024）估计了一个在DHS资产指数之间的随机森林模型 [4] 并为11个非洲国家的辅助数据进行了分析。然后，他们预测了42个非洲国家的电网层面的每人大致消费量。Bosco等人（2017年）产生了性别细分的发展指标，如女性识字率、儿童发育不良和避孕使用情况。他们依赖地理定位的DHS调查，使用贝叶斯地理统计和机器学习模型对尼日利亚、肯尼亚、坦桑尼亚和孟加拉国这些发展指标与辅助数据之间的关系进行建模。

第二节介绍材料和方法。第三节展示模型结果、模型评估、高分辨率地图和重要性分析。第四节介绍网格化劳动力市场数据的运用，探讨就业率和城市就业情况。第五节讨论数据、方法和结果。第六节得出结论。

[4] 人口与健康调查

X 1 材料与方法

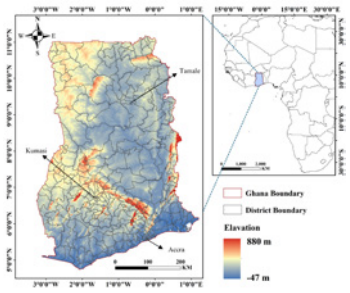
1.1 研究区域及主要劳动力市场统计数据

研究区域为加纳，由16个地区和261个区组成。2021年，加纳人口为3080万，劳动年龄人口（15岁以上）为1990万。根据2019年国际劳工统计学家大会提出的就业定义，总就业人口为960万，意味着就业率为48%（就业率与劳动年龄人口比）。就业率在各年龄组和性别之间存在很大差异。15-24岁年轻人的就业率为22.9%，而25-54岁核心年龄工人的就业率为64.8%，55岁以上老年工人的就业率为39.6%。男女就业差距为8个百分点。失业率低，因为没有失业保险（3.6%）。受教育程度低（基础教育及以下）的工人占比高达63.4%，这在自雇工人（占总就业的63.9%）所占的份额中得到了反映。加纳的产业结构特征是由农业就业（26.5%）向服务业（57%）的转变。

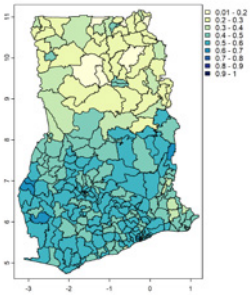
采矿业在就业方面规模较小（1.1%），但特别是金矿业对加纳的经济具有重要意义。制造业和工业占就业总量的近15%，与其他撒哈拉以南国家相比，这一份额相对较高。为了避免在降级时产生边境效应，收集了以下辅助数据，既包括加纳，也包括邻国地区。

图1：选择的劳动力市场指标

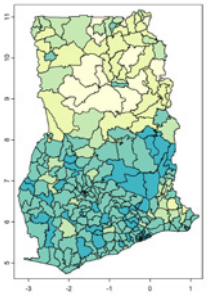
(a) 研究区域位置



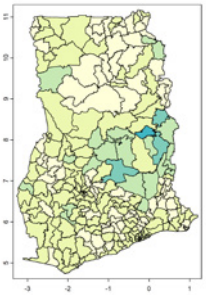
(b)
就业
评级 - 区域



(c)
就业
评价年轻一代
区域



(d)
就业
评定女性
区域



1.2 数据描述与处理

本论文中的方法论包括数据收集、数据准备、模型估计以及生成网格化劳动力市场数据的图2所示。本研究利用两种类型的数据库。第一种类型包括来自加纳统计局的基于普查的区级人口和劳动力市场数据，这些数据被视为因变量。

第二类包括分辨率为0.005度的网格辅助数据，被视为独立变量。加纳统计服务提供了2021年住房和人口普查的10%代表性样本，以及相应的GIS划定的行政边界，包括国家、地区和区级。我们使用2019年国际劳工统计会议对就业类别的定义，并在区级计算了17个就业类别，并应用了普查数据提供的人口权重。这些就业类别包括总就业及其按年龄（青年、成年、老年）、性别（男性、女性）、技能（高技能、低技能）、状态（自雇、领薪）、部门（农业、采矿、制造业、工业和服务）、失业和NEET（未受教育、就业或培训）。年龄类别如下：总人数15岁及以上，青年15-24岁，成年24-54岁，老年55岁及以上。高

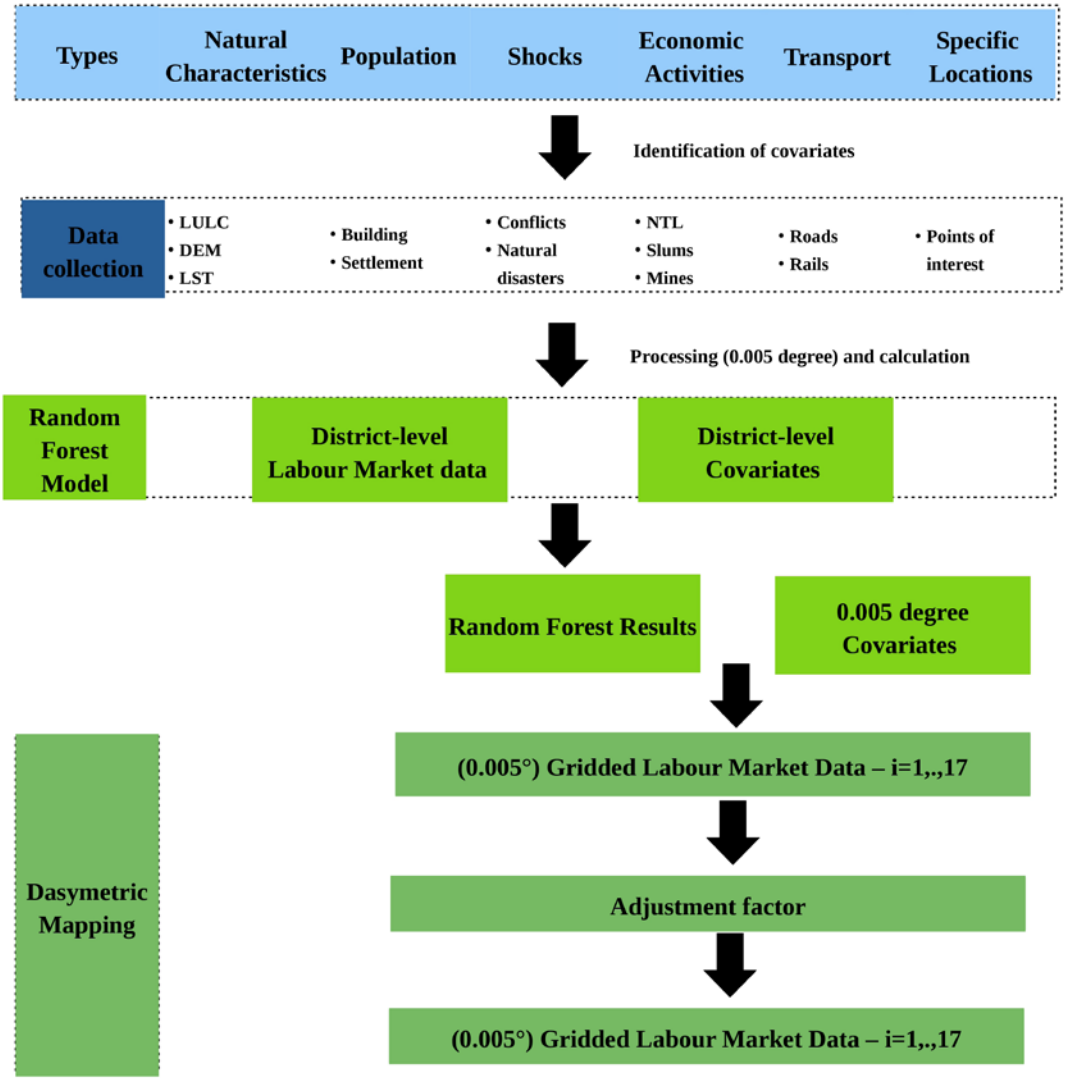
熟练指代中等或高等的教育水平，低熟练指代低于基础水平和基础教育水平。

自雇就业包括自营劳动者、生产合作社成员、家庭贡献劳动者和自雇者。农业包括林业和渔业；工业包括电力、燃气、供水和建筑。服务业包括公共和私营服务。NEET指未就业、未受教育也未接受培训的青年。虽然降级是在就业类别层级上进行的，但一个国家的劳动力市场状况最好用比率来描述。因此，我们计算了7个人口类别，以便能够生成高分辨率的就业率和相关类别地图。 [5] .

辅助数据分为5个类别。第一个类别主要衡量自然特征，紧密遵循文献中用于降尺度人口数据的方 法，旨在识别人口密集地区与自然地区。ESA-WorldCover提供10米分辨率的9个土地覆盖分类类别，包括树木覆盖、灌木丛、草地、耕地、建成区、稀疏植被、永久性水体、草本湿地和红树林。 [6] 我们通过计算不同土地覆盖类型在0.005度分辨率下的百分比来生成一组网格化的辅助变量。数字高程模型测量地形的坡度和崎岖度。Copernicus数字高程模型的分辨率为30米，据此提取了4个指标：DEM、坡度、崎岖度和坡度≤5度的比例。这些数据被重采样到0.005度网格。植被通过三个指数的中值来测量：增强植被指数（EVI）、归一化植被指数（NDVI）和植被近红外反射率（NIRv）。这些指数来自MODIS MOD13A1.061产品，频率为16天，分辨率为500米。MODIS MYD11A1.061产品提供了中值地表温度。频率为每日，分辨率为1公里×1公里。植被健康和生产力另一个衡量指标是净初级生产力（NPP），它测量光合作用固定的碳量。它是MODIS产品（MOD17A3HGF.061）之一，以年度频率和0.05度分辨率提供。年度净初级生产力（NPP）通过将给定年份MOD17A2H产品中所有8天的净光合作用（PSN）值相加来计算。PSN值代表总初级生产力（GPP）和维护呼吸（MR）之间的差异。降雨估计来自CHIRPS全球数据。分辨率为0.05度。通过最小值、最大值、平均值和总降水量，将每日观测数据汇总为年度数据。归一化差异水指数（NDWI）评估植被和土壤中的水分含量。相应的MODIS产品是MODIS/006/MOD09GA。每日数据以最小值、最大值和中值汇总为年度数据。分辨率为0.005度。河流和湖泊通过HydroRIVERS和HydroLakes数据库识别。根据HydroRIVERS使用了到最长上游路径和最长下游路径的距离的网格化数据。此外，使用数据库提供的河流网络计算了河流网络密度。使用湖泊多边形数据计算了从栅格中心点到合并后的湖泊数据的最近距离。此外，利用OpenStreeMap，从像素中心到最近的水类（湖泊、运河、排水沟、河流、溪流）的距离测量补充了HydroSheds数据。

[5] 总人口数，劳动年龄人口总数，按年龄段（青年、壮年、老年）及性别（男、女）分解。
[6] 11个类别减去“雪和冰”、“苔藓”类别。

X 图2：网格化劳动力市场估算方法



该第二类数据集收集了识别人口位置的辅助数据。Open Buildings 2.5D 时间数据集以0.5米的分辨率识别建筑存在、建筑高度和建筑比例。GRID3 居住地数据提供了以0.005度分辨率的年度多边形。居住地多边形与区域栅格相匹配，每个单元格的居住地比例被测量。

大型冲击，可能对当地经济产生影响，通过冲突数据和自然灾害数据被识别。乌普萨拉冲突数据计划识别致命暴力的个别事件，并提供伤亡人数以及冲突性质的估计。考虑到较小的空间分辨率，单位是10公里半径内的伤亡人数以及对应年份最近冲突的距离。自然灾害通过干旱、洪水、地震和滑坡来衡量。《地理编码灾害数据集》（GDIS）为各种自然灾害提供了多边形和点估计。干旱从1999年至2018年进行测量。过去洪水通过叠加全球洪水数据被识别。

洪水数据库和地理编码灾害数据集（2009-2018年）。基于全球洪水数据库，通过识别最新年份可用的受灾区域后续像素，构建了第二次洪水测量。地震数据由美国地质调查局全球地震数据集提供。[7] 滑坡数据由全球滑坡目录——NASA戈达德（1970-2019）提供，我们提取了2000年至2019年的滑坡记录。[8] 对于上述四种自然灾害数据，每个网格单元中心到最近的灾害类别距离是以0.005度网格构建的。

地方经济活动通过使用具有15弧秒空间分辨率的年度夜间灯光卫星图像进行估计。我们利用来自黑玛瑙的数据，特别是关注各个角度的无雪辐射亮度。这些年度数据经过调整以考虑阳光、眩光、月光、云层和极光相关的照明效果。数据以辐射单位测量，值为65535的像素以及标记为低质量的像素被排除在分析之外。

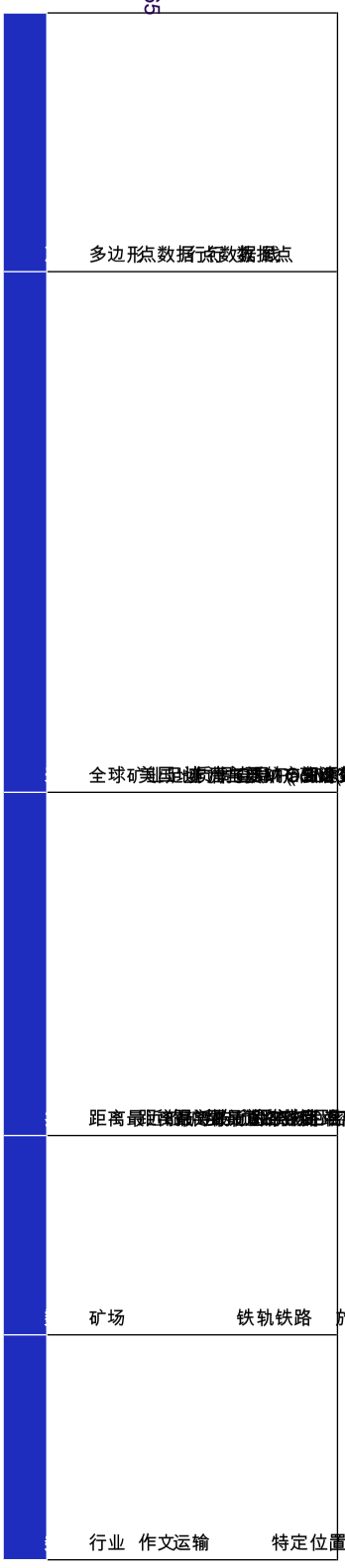
在经济活动和就业水平之外，收入分配和就业质量是本工作中预测的一些就业类别决定因素。由于通过地理空间数据测量收入分配相当复杂，贫民窟是一个合适的候选对象。没有全球贫民窟数据。然而，Li等（2023）通过假设密集排列、小型建筑区域往往暗示贫民窟高发，利用城市形态学估计了四个撒哈拉以南国家的贫民窟。

最终类别包括辅助数据，这些数据有助于识别行业维度以及基础设施（如道路和学校）密度，这可能捕捉到某些推动劳动力市场参与的因素。矿井数据来源于三个来源：i）全球采矿足迹（唐和韦尔纳，2023），ii）美国地质调查局非洲矿业及相关基础设施（帕迪拉等，2021），以及iii）区域资源开发测绘中心非洲主要矿产数据集。第一个数据库由多边形组成，而其他两个由点数据组成。对于每个像素，计算到最近多边形或最近点数据的距离。交通基础设施数据来自对应年份的OpenStreetMap。道路数据分为六个类别：高速公路、主要道路、次要道路、不适于汽车的道路、非常小的道路和未知。对于每个类别，在像素级别构建密度。对于高速公路和主要道路，生成到最近道路段距离。对于铁路，也计算了类似的措施。兴趣点（POI）来自OpenStreetMap。这些POI分为15种主要类型，包括公共设施、教育、健康、休闲、运动、餐饮、室内住宿、室外住宿、购物、货币、旅游、燃料停车、交通、工作场所和其他。这15个类别的密度是缩减劳动力市场统计数据的重要因素，因为它们的集中度可以反映就业的行业构成，同时捕捉到诸如健康、教育和宗教等可能影响劳动力市场参与决策的特定因素的重要性。密度通过带宽为3000的核密度估计来衡量。

[7] https://developers.google.com/earth-engine/datasets/catalog/GLOBAL_FLOOD_DB_MODIS_EVENTS_V1
[8] https://gee-community-catalog.org/projects/landslide/

表1 据来源

自然 特点	LULCDEM 植被地表温度/水/降水/火灾/湖泊	9门课程 [illegible]	ESA WorldCover MODIS [illegible]	10米 30米 0.05度 0.01度 0.005度 0.05度 0.005度 0.005度 [illegible]
人口	Nuclear Power Plants 建筑高度	[illegible]	[illegible]	[illegible]
大规模震动	冲突 解决 自然灾害	[illegible] 比例 [illegible] 滑坡 [illegible] 地震	OSM 全球洪水 [illegible] 全球滑坡	0.005度 数据点 数据点 [illegible]
经济活动 收入分配	夜间灯光图	NTL, [illegible] VCNLI	[illegible]	[illegible] 2019）



本表展示了用于人口和就业数据降尺度的辅助数据来源。

总共共有64个辅助数据。每个栅格数据已重采样至0.5公里×0.5公里的分辨率。此外，每个栅格正在根据...
x
miǎn

，与 *miǎn* 并且 *最大*
mx min
数据x的最小值和最大值。归一化使得不同量级的数据可比较，并有助于机器学习算法更快收敛。此外，像素群仅限于人口密集的像素。如果一个像素的夜间灯光、建筑高度、建筑比例、定居点或贫民窟的值不为零，则该像素被视为人口密集。

POIs被转换成分辨率为0.005度的密度层，每个类别都使用王等人（2020年）；金等人（2023年）的核密度估计（KDE）。KDE使用以下带宽进行估计：500米、600米、1000米、1200米、1500米、3000米、5000米和8000米。通过实验不同的带宽，我们注意到增加带宽会导致核密度结果的峰值减少。由于核密度分布的趋势是一致的，为了避免带宽过大——从而掩盖局部特征，通常建议将带宽设置为栅格分辨率的1到2倍。这种方法在平滑和保留细节之间取得平衡。因此，我们可以选择1000米作为带宽进行进一步建模。

劳动力市场统计的网格化是通过人口密度的降尺度来完成的。鉴于与像素大小随纬度变化而变化的潜在问题，人口密度被测量为人口除以像素数。在0.005度分辨率下，只有一个像素。然而，机器学习算法在区县级别训练数据，对于这个行政级别，像素数量很重要。我们变量的密度公式也与涉及更高分辨率（如0.01度）的鲁棒性检查相关。

1.3 缩放方法

我们使用随机森林模型对人口和劳动力市场数据进行降尺度。该过程分为三个步骤。第一步是收集和处理数据。如前文所述，区级人口和劳动力市场数据作为因变量，而五个类别的辅助变量被用作自变量。在这里，区级辅助变量是从0.005度网格辅助变量中汇总的。第二步是模型构建和训练。区级人口、劳动力市场数据和辅助变量被用作训练随机森林模型的输入。训练好的模型随后被应用于使用0.005度网格辅助变量作为输入，预测0.005度网格级别的人口和劳动力市场数据。为了实现高精度映射，引入了区级校准因子以匹配官方普查统计。最后，使用区级人口和劳动力市场数据来评估0.005度网格预测数据的准确性。

在本次研究中，随机森林算法的具体表示如下。首先，一个ran-
多目标森林模型
d
d

d

成立于区县级别的，基于该等式 yxx
dd
x
d
k

，其中表示地区级别的种群变量或劳动力市场变量。
1 2 I'm sorry, but the text you've provided, \"r,\" is incomplete and does not contain sufficient information to provide a meaningful translation. Please provide a full sentence or paragraph for translation. I'm sorry, but the text you've provided, \"r,\" is incomplete and does not contain sufficient information to provide a meaningful translation. Please provide a full sentence or paragraph for translation.
x
d
y
d

k

*k*表示该辅助变量的总数量，与

表2：最小化均方根误差的调整参数

名字	流行	wap	wap_{y}（根据给定的文字，“y”直接对应汉字“y”，没有其他上下文时，只能将其直译。）	wap_{w}	wap_{m}	Et	Ety	Etpa	Etold	Etw.
mtry	16	8	10	9	5	8	7	10	12	10	5	8
ntree	300	300	800	300	500	50	500	500	50	500	50	500
名字	Etm	Ut.	NEET	Eths	ETLs.	Etse	Etees	Etag1	Etmi1	Etma1	Etind1	Etserv1
mtry	13	4	19	10	13	7	9	11	20	18	13	13
ntree	500	800	100	300	300	800	300	500	100	800	500	500

此表格展示了在调整各人口和就业类别时，使RMSE最小化的变量组合和树模型的组合。

然后，随机森林模型

d

在区县级用于预测目标变量 g

I'm sorry, but the text you've provided, \"r,\" is incomplete and does not contain sufficient information to provide a meaningful translation. Please provide a full sentence or paragraph for translation.

y

（根据给定的文字，“y”直接对应汉字“y”，没有其他上下文时，只能将其直译。）

g 在0.005度网格分辨率下，如方程所示

d

y x x

gg g

2

何处 g de- 表示在0.005度网格单元格中的辅助变量。最后，为了匹配区级数据，

I'm sorry, but the text you've provided, \"r,\" is incomplete and does not contain sufficient information to provide a meaningful translation. Please provide a full sentence or paragraph for translation.

x

1 k

x

在建立随机森林模型之后，我们采用重复的k折交叉验证方案以确保预测的有效性，并通过参数调整进一步优化模型。k折交叉验证涉及将数据随机划分成10个大小相等的子集。对于每一次的10次迭代，使用一个子集进行测试，而剩余的子集则用于训练。这个过程重复了10次，确保所有子集都被用作测试集一次。

在使用随机森林模型进行预测之前，我们通过调整模型参数来优化其性能。调整的参数是在随机森林中生长的树的数量。树的数量的取值为5、10、20、50、100、300、500和800。第二个参数是每次分裂时随机选择的变量数，即节点划分为两个或更多子节点的过程。此值介于1到30之间。在调整过程中，使用均方根误差、平均绝对误差和R平方来评估每个参数组合的准确性。在这三个指标中，我们依赖于均方根误差。

（根据给定的文字，“y”直接对应汉字“y”，没有其他上下文时，只能将其直译。）

y

（根据给定的文字，“y”直接对应汉字“y”，没有其他上下文时，只能将其直译。） g y

g

p

i

y

何处 y

g

n i

是修正后的预测变量 g 是 y

预测的网格分辨率目标变量

[1] p

表2展示了最终确定的结果。mtry参数决定在每个决策树节点处随机选择多少个特征作为分割的候选。引入这种随机性有助于减少单个树之间的相关性，从而提高模型的鲁棒性和降低过拟合的风险。使RMSE最小的mtry参数值范围从4（失业）到20（采矿），平均在每个类别中选择10个特征。使RMSE最小化的树的数量，与mtry参数（ntree）结合，范围为50到800，平均为410。通过mtry参数选出的变量不能直接与出现在变量重要性中的那些变量进行比较（请参见下面的讨论），因为后者有助于识别哪些特征对做出准确预测最重要。然而，它们之间存在间接联系：算法倾向于青睐mtry中的变量，这些变量产生有效的分割，从而降低节点不纯度。

是的，该变量是区县级目标变量，代表 y

The district

n区表示区域的数量，并通过重复

/

三步，首先在粗分辨率下构建随机森林模型，其次在更高分辨率下应用随机森林模型，最后对预测结果进行区域级校正，可以创建一个0.005度分辨率的指标变量图。

[9] 我们使用R包，具体来说是caret和randomForest库。

在人口类别中抽取的最佳变量数量与相应就业类别的变量数量相似，尽管有时略小。[10] 差异在于树木数量，对于人口类别而言比就业类别更多。样本变量的最优数量以及树木的数量，往往随着剩余的就业类别和行业就业类别的增加而增加。对于后者来说，这一点尤为明显。一种解释是，这些类别衡量的是非充分就业和就业质量，而就业水平与低收入国家的_population_和人口统计密切相关。这一点在农业中尤为明显，因为农业在加纳仍然是生计活动，因此与_population_密切相关。相比之下，采矿、工业和制造业是那些需要特定技能和能力 的部门。适应这些类别的辅助变量和树木的数量可能随着被估计类别的复杂性的增加而增加。

该模型将树木数量估计和变量选择纳入随机森林估计中。该模型生成了不同的指标来评估估计的质量，包括均方残差和解释方差百分比。此外，该模型还生成了变量重要性和节点纯度指标，以讨论不同人口和就业类别预测背后的主要辅助变量。得到的模型随后用于预测不同类别的像素值。通过平均所有单独树木的预测结果，以生成更准确和稳健的预测。

[10] 15岁及以上，青年，盛年，老年人，女性，男性

2 结果

2.1 模型评估

我们的两个估计质量指标是均方残差（MSR）和解释的变异百分比。[11] 在不同就业类别中，MSR的范围从0.07到0.2。这些低值表明模型具有良好的准确性。模型的预测值更接近实际观察到的值，这表明模型与数据的拟合度更好。模型的良好性能得到了超过90%的变异解释百分比的证实，除了NEET（89%）和一些行业就业（农业68%和矿业67%）。在17个与就业相关的类别中，有7个达到了94%的变异百分比。模型在农业方面的相对较弱的表现与现有尝试缩小行业GDP数据的规模的工作一致（Wu等人，2024年）。这可能与需要收集更多辅助数据以更精确地衡量作物产量有关。Wu等人（2024年）在中国UAMRYR省测量行业GDP数据时，使用随机森林在农业方面达到了0.63的R2，在制造业方面达到了0.65，在服务业方面达到了0.74。[12] 我们同样发现，衡量就业水平和衡量就业质量的类别之间没有显著差异。请注意，由于篇幅限制，关于人口类别的指标已复制在附录中。

表3：剩余平方和及解释的变异百分比

	(1)	(2)		(3)	(4)
名字	MSR	百分比差异	名字	MSR	百分比差异
Et	0.16	0.94	Eths	0.18	0.94
Ety	0.13	0.92	ETLs.	0.15	0.92
Etpa	0.16	0.94	Etse	0.17	0.92
Etold	0.12	0.92	Etees	0.17	0.94
Etw.	0.15	0.93	Etag	0.10	0.68
Etm	0.14	0.94	Etmi.	0.07	0.67
Ut.	0.10	0.92	Etma	0.16	0.91
NEET	0.15	0.89	Etind.	0.10	0.94
			Etserv	0.20	0.94

本表展示了每个就业相关类别的平均平方残差以及解释的方差百分比。

2.2 网格化劳动力市场地图

The gridded employment data are mapped in Figure 3 at a resolution of 0.005 degrees. Due to space limitations, we present only a selection of the 17 categories. The remaining categories can

网格化的就业数据在图3中以0.005度分辨率进行了绘制。由于篇幅限制，我们仅展示了17个类别中的部分。其余类别可以

$$\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}$$

[11] $MSR = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}$

$R^2 = 1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$

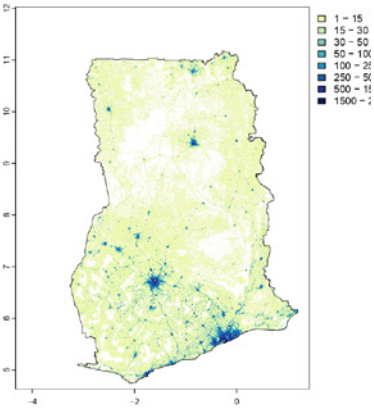
他们取得更高的 R^2。

[12] 采用Res-FuseNet模型而非随机森林。

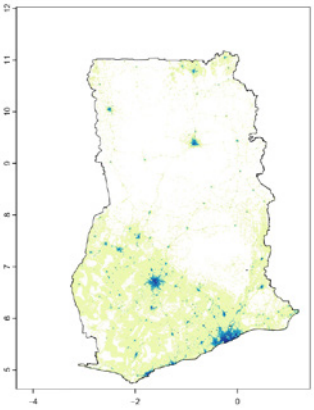
可在附录中找到，图8至11。以下各就业类别以绝对值表示：0、15、30、50、100、250、500、1500、2500。数据有效地反映了加纳主要城市聚集区的人口密度较高。城市主要位于加纳南部，介于阿克拉和库马西两座城镇之间。较高的人口密度也转化为较高的就业值（面板A）。就业也集中在加纳沿海和连接城市的交通基础设施沿线。在国家北部，就业较为稀疏，主要集中在四个主要聚集区：塔马莱、瓦、博尔冈塔加、巴库。后两个聚集区靠近布基纳法索边界。

图3：（所选）就业类别空间分布

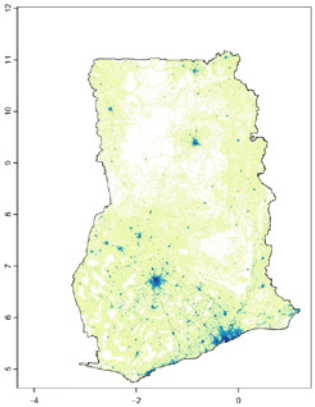
Employment – 就业 – 0.005°



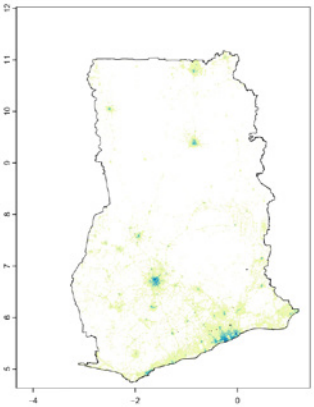
(b) 员工 – 0.005°



（c）低技能 —— 0.005°



(丁) 制造 - 0.005°



除了就业数量外，就业质量也是描绘地方劳动力市场的一个重要维度。我们通过在图（b）中绘制工资就业来展示就业质量。工资就业在像素级别的分布与总就业分布相似，反映了人口密度。然而，工资就业主要分布在

最大的城市区域。在次级城市以及连接这些次级城市的地区，这种现象不太常见。在加纳的北部半部，有薪工作仅限于少数地区。

除了就业状态外，我们通过描绘低技能工作者（见面板(c)）来说明就业质量。低技能就业的分布与总就业分布相吻合，证实了工作队伍中有相当一部分人的教育水平较低。一个显著特点是，在城市中心地区，低技能就业不如总就业普遍，反映出高技能工作在最大的城市区域核心地带的集中。

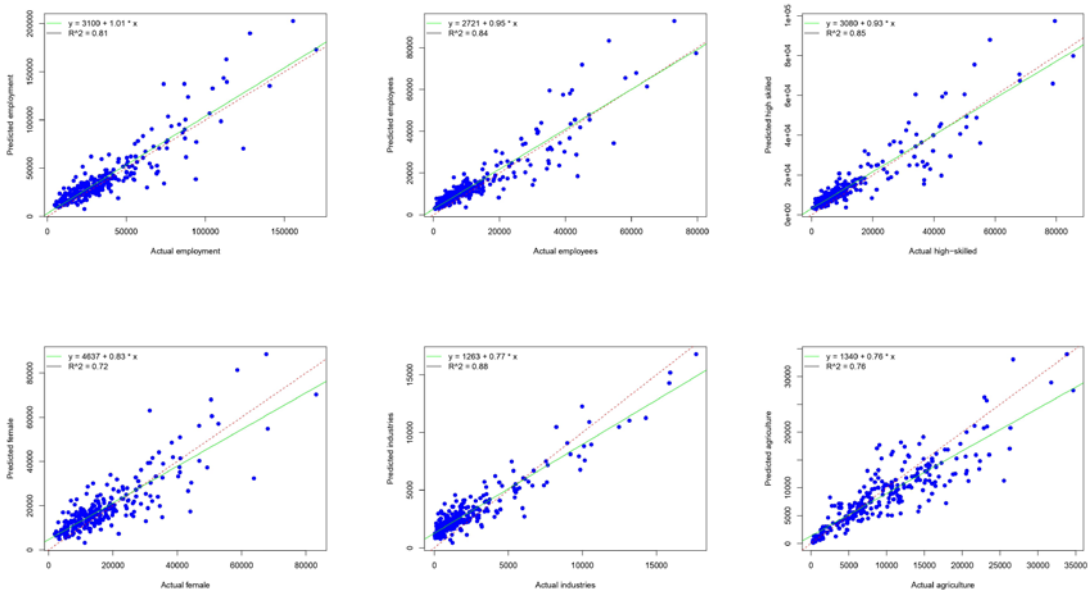
作为最终说明规模缩小的示例，我们展示了制造业就业的分布图。我们以前已经指出，加纳的经济主要依靠服务业和农业，而工业和制造业的就业仍然微乎其微。这一点在图(d)中表现得非常明显，全国范围内的制造业就业几乎可以忽略不计。然而，在某些高度集中的城市中心，制造业并不能被忽视。这表明，尽管从整体来看制造业就业相对较小，但加纳仍有一些地区具有制造业的专业化特征。

2.3 实际与预估的区域统计数据

在没有更细的行政单位在公开可用的普查数据中，我们只能比较实际与预测的劳动力市场变量在区级水平（图4）。预测变量在区调整实施前呈现。实际的区级统计数据基于普查数据，预测值则是对应区层单个像素值的汇总。

尽管各区域之间及0.005度像素之间的规模差异很大，但实际值与预测值之间的相关性仍令人满意。对于下面显示的就业类别，选择如下：总就业的0.84，高技能就业的0.85，以及产业的0.88。回归系数接近于1：分别为1.01、0.84、0.85和0.88。

X 图4：实际与估算统计数据 - 区级 - 选择类别



关于就业、雇员和高技能就业，红色45度线周围的分散程度随着地区人口的增长而增加。对于人口较少的地区，点状数据接近45度线，并分布在其上下。随着人口的增加，数据点逐渐远离红色线。此外，模型低估了中等人口规模地区的地区级就业，然后高估了高人口规模地区的就业。

对于行业而言，模式略有不同。分散度较小。然而，该模型高估了没有工业专业化的地区的行业就业，同时低估了有工业专业化的地区的行业就业。农业就业也出现了类似的模式。随着地区绝对规模的增加，该模型对农业就业的预测不足。

由于空间限制，上面仅绘制了六个类别。对于剩余的类别，行业就业总体表现良好，制造业和服务的R2值分别为0.6和0.85。这些行业层面良好表现的主要原因在于辅助数据直接且与土地利用、土地覆盖和农业植被健康指数相关，而其他行业则与城市相关辅助数据相关，如不透水表面和交通网络。

R关于就业观察到技能高级技能就业和低技能就业（）示模比。这主要是由实际值与预测值之间的差异随着区域规模的增加而增加所解释的。

R最后价值观方面有男女就业点的差异与成年劳动力、老年工人与成年劳动力之间的就业率分别相差0.2和0.3个百分点。有趣的是，……

这些类别最易受到参与劳动力市场决策的影响。在引言中讨论的推动劳动力参与的因素，通常与家庭结构和教育相关，这些因素难以通过辅助数据来捕捉。

2.4 重要性分析

重要性分析提供了一种评估每个特征对模型准确度贡献的方法。这有助于识别在预测中最具影响力的变量，并支持特征选择和模型解释。所使用的指标是“均方误差增加”，它评估每个辅助变量对模型准确度的贡献。[13]

关于总就业，三种类型的相关数据突出（见图5）。建成区相关变量捕捉人口密度，并涉及就业的劳动力供给方面——即有人就有工作。夜间灯光和贫民窟等变量捕捉就业关系的需求侧及其组成（即就业质量）。最后，道路密度突出显现，并捕捉就业关系中至关重要的不同维度。道路密度反映了人口密度，因为道路网络往往会随着人口增加而增加。它还反映了就业的行业维度，因为城市地区有更多专门从事服务业、制造业和工业的道路。道路密度也可能反映公共基础设施的一般密度，这可以作为就业质量（技能和地位）的替代指标。最后，在经济地理学中，经济活动从道路的邻近性中受益，因为这降低了运输成本。这对受到聚集效应影响的行业，如制造业，尤为重要。

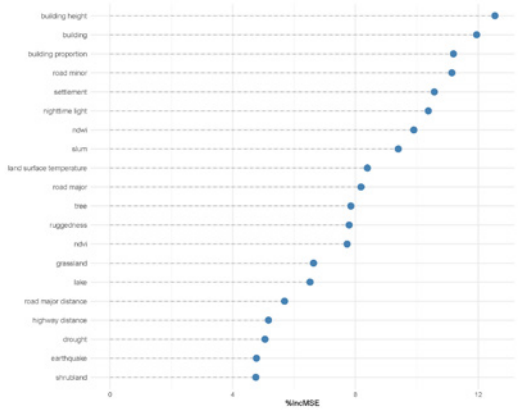
相比之下，农业就业主要受反映植被的辅助数据影响，如土地利用和土地覆盖投入，包括灌木和草本植被。如归一化差水分指数（NDWI）等指示植被健康状况的变量，在模型中也显得至关重要。气候变化变量——尤其是干旱——成为第三大重要因素，因为降水不足直接影响到作物产量。矿业就业受距离矿区和铁路距离等变量的驱动，因为铁路基础设施通常被用来运输矿产开采，以及公路基础设施（Jedwab和Moradi，2016）。

制造业、工业和服务业的就业与不透水表面、道路基础设施和夜间灯光有关。这反映了这些行业在城区的集中。

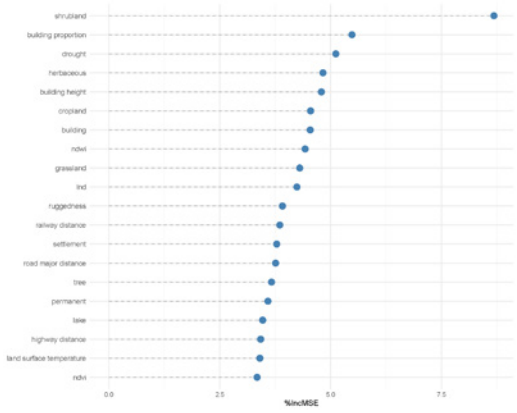
服务业，尤其是在人口密集地区的公共服务位置，受到很大影响。制造业和工业也受到交通运输成本和集聚经济效益的影响，这些可以通过这些不同类别的辅助数据有效地加以捕捉。

[13] 第二项措施，可根据需求获取，是提高节点纯度。它表示每个辅助变量对节点纯度的贡献。这两项措施产生类似的结果。

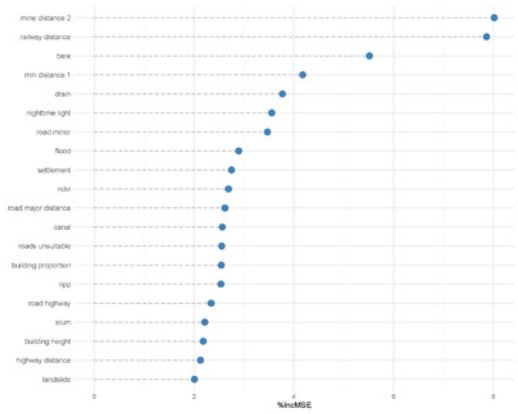
X 图5：重要性分析



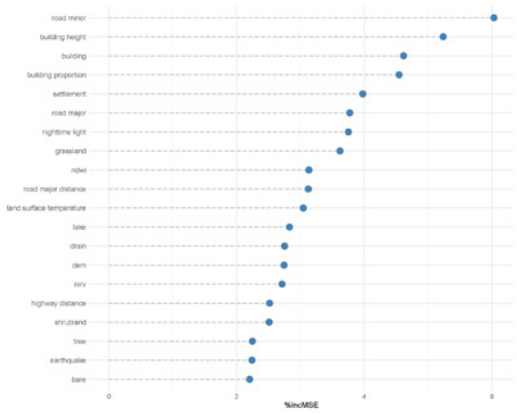
(a) employment



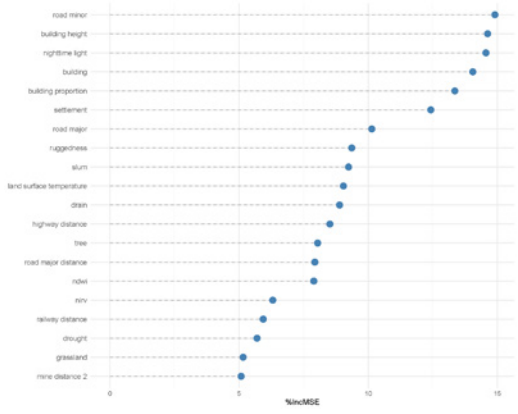
(b) 农业



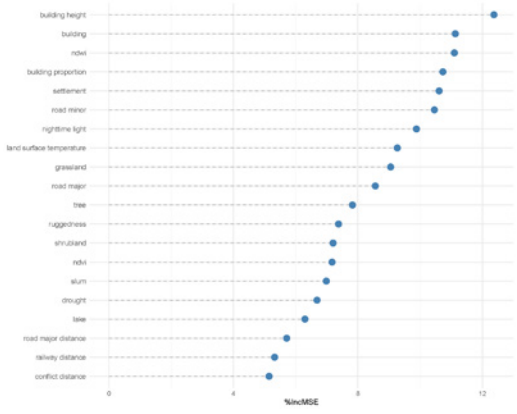
（c）采矿



（d）制造



(e) 高技能



(f) 低技能

关于就业质量，无论考虑何种指标，高技能就业都受其集聚在城区的因素驱动——即建成区域相关类别、夜灯光和道路基础设施——而不论哪种计量。相反，低技能就业在城市内的生存服务业以及农村的生存农业中都非常普遍。

乡村地区。这种双重存在解释了建筑相关变量与植被相关变量的混合。例如，归一化差异水指数（NDWI）在平均均方残差方面排名第3。地表温度和草地也出现在最重要的十个变量之中。

2.5 模型更粗略的分辨率和质量

像素级劳动力市场变量的投影分辨率为0.005度，接近赤道时约为0.5公里乘以0.5公里。这是一个可能被认为过于精细的粒度分辨率，因为地方劳动力市场通常在更粗的分辨率下表现最佳，如1公里乘以1公里或5公里乘以5公里。因此，模型在0.05度（约5公里乘以5公里）的粗分辨率下进行估计。估计质量仍然很高。然而，平均平方残差（MSR）和解释的方差百分比都变差。平均而言，所有就业子类别中的MSR增加了13个百分点，这一增长主要是由制造业、工业和服务业等行业类别驱动的。解释的方差百分比下降更为有限，平均约为1个百分点。一个可能的解释是，用于细化的辅助数据在粗分辨率下平均时会失去解释力。例如，以0.5公里乘以0.5公里分辨率的道路密度比聚合到5公里乘以5公里时更有信息量。由于篇幅限制，不同就业子类别的MSR和方差百分比在附录表4中展示。

X 3个应用

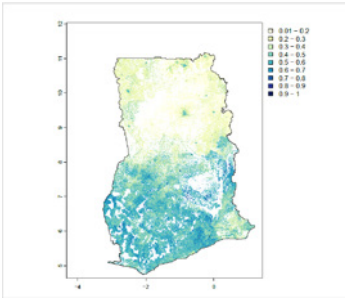
3.1 就业率，行业百分比

这项工作中的降级练习涉及劳动力市场的级别类别。然而，为了更好地描述本地级别的就业情况，测量就业率是很重要的，因为它提供了就业与劳动年龄人口（定义为15岁以上）的比例感。为了构建这些类别，这项工作还在像素级别上对人口类别进行了细分。[14] 这些类别包括总人口、劳动年龄人口总数、青年、适龄人群、老年、男性和女性。人口类别估计的质量与就业类别估计的质量相同，参见附录5表。就业率的数据细度越高，其异质性越强。全国就业率为48%。在地区级别上，标准差为11.2个百分点，最低为16.5%，最高为64.7%。虽然在总体层面上相对较低，但在加纳的一些地区，就业率之高甚至达到高收入国家水平。在像素级别，标准差增加到15个百分点，就业率在0.1至0.98之间波动（图6A面）。

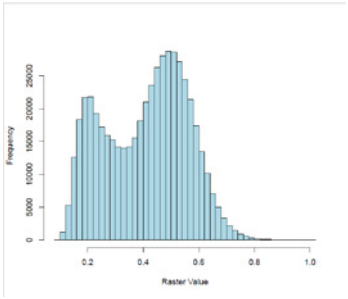
[14] 这些类别包括总人口、劳动年龄总人口、青年、盛年、老年、男性和女性。对于人口类别估计的质量与就业类别估计的质量相当。对于每个类别，我们都在括号中列出MSR和R2：总人口（0.16；0.92），劳动年龄总人口（0.15；0.93），青年（0.15；0.92），盛年（0.14；0.94），老年（0.14；0.92），男性（0.14；0.93），女性（0.13；0.94）。此外，异常值像素的百分比，即大于1的比率，对于所有类别来说都不为零，除了女性就业率（4%），这可能反映了对女性就业的过高估计或对女性劳动年龄人口的过低估计。

X 图6：（选定）就业率及百分比的分布

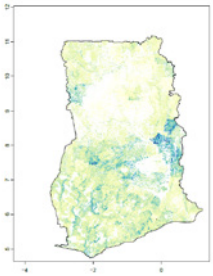
(a)（a）
就业
速度



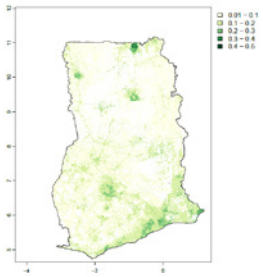
(b)
直方图
就业
速度



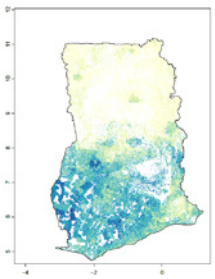
（c）青年就业
就业率



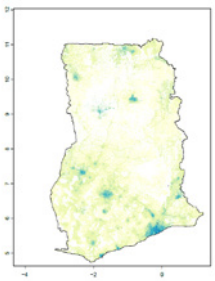
(d) 人类与工业
%



(e) 旧工作
ers employ-
月薪



高级技能
就业
%



关于就业率的地理分布，虽然就业水平反映出城市地区的普遍性，但就业率描述了一种南北差异。在南部加纳，无论是农村还是城市地区，就业率都较高。国家北部，除了之前提到的四个主要城市中心外，就业率普遍较低。特别是东北部的就业率特别低（图6面板A）。在像素级上，就业率的分布呈中心在0.2和0.49的二项分布（图6面板B）。

不同年龄段中，就业率分布呈现出年轻工人和老年工人不同的模式（图C和D）。在城镇地区，青年就业率较低，反映了城市地区对年轻工人就业机会的缺乏。然而，加纳中心地区的青年就业率较高，尤其是中部-东部地区。西部地区一些非常细分的区域也出现了较高的青年就业率。相比之下，南部城市中心老年工人的就业率低于北部城市中心。从地理分布模式来看，老年工人在加纳的西部地区具有较高的就业率；就业率越高，越靠近科特迪瓦边界。

最后两个图显示的是高素质就业人数及制造业与工业就业人数占总就业人数的百分比（E面和F面）。就高素质就业份额而言，尽管从绝对数量上来看高素质就业集中在城市地区，但在次级城市、沿海地区和城市中心之间，高素质工人所占的比例似乎相对较大。类似地，制造业和工业

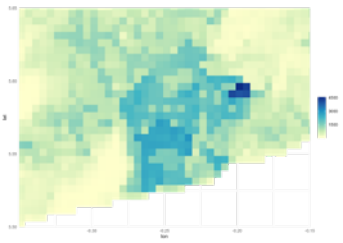
当以总就业人数的百分比来衡量时，就业分布并不集中在少数几个城市中心，而不是在绝对值上。在城市地区，制造业和工业就业可以占总就业人数的相当比例。尽管从整体来看，这一比例仍然很小，但在地方层面，高比例的存在意味着存在专业化中心。

3.2 市级劳动力市场

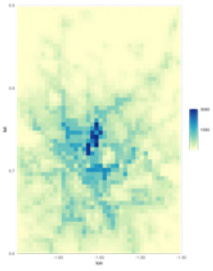
网格化就业数据使我们能够为任何村庄、镇或城市生产一个详细的劳动力市场全景。在本节中，我们提供了在聚集层面的说明。图7聚焦于阿克拉、库马西和塔马莱等城市。该模型区分了就业集中的市中心和就业较少的城市边缘。阿克拉表现出高密度的就业，以及聚集体内的异质分布。在库马西和塔马莱，模型确定了就业集中的核心区域，外围的就业岗位则围绕道路基础设施聚集。

X 图7：阿克拉、库马西和塔马莱的就业分布

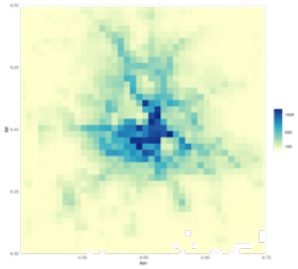
(a) 阿克拉 - 0.005°



（b）库马西 – 0.005°



(c) 热狗 0.005°



为了衡量城乡之间就业分布情况，我们将网格化劳动力市场统计数据与Africapolis生成的集聚层叠加。[15] Africapolis将城市聚集定义为建筑物和建筑之间距离少于200米的建成区域，总人口超过10,000人的连续体。其定义基于空间方法，与现有城市区域的定义不同，后者依赖于行政标准。利用Africapolis的聚集层，我们识别出129个拥有超过10,000人口的聚集区。对于这些城市，我们将格网人口数据汇总，生成劳动力市场统计数据。重点关注五个最大的聚集区——阿克拉、库马西、塔马利、塔科腊迪和科佛里杜亚，我们发现它们占总人口的32%，但占总就业的38%，占高技能工作的53%，占高技能员工的50%，占制造业就业的42%，占服务业就业的53%，以及失业者的41%。这表明对城市区域的偏好，这些区域集中了大多数就业类别。总体来看，拥有超过10,000居民的129个城市聚集区集中了49%的人口，但集中了68%的高技能就业和58%的制造业就业。

[15] https://africapolis.org

4 讨论

虽然人口变量或GDP的缩减通常仅限于少数几个变量——如总人口、年龄细分或行业GDP——这项工作的一个贡献是在17个就业类别中进行缩减。主要动机是就业是一个复杂的现象，结合了数量、未充分利用（失业或非经济活动人口NEET）、质量（技能和地位）以及行业分布等方面。这种复杂性在中等收入国家尤为明显，那里的就业挑战远远超出了整体就业水平。人口（建成区）、总需求（夜间灯光）和基础设施的可及性成为各种就业类别的重要决定因素。

我们观察到，除了模型的良好表现及其预测的准确性外，对劳动力市场参与决策敏感的分类可以进一步改善。在文献中，家庭结构、教育水平和是否工作或学习的决策是影响青年、老年工人和女性就业的重要因素。我们对辅助数据的选择可以进一步优化，以更好地捕捉这些维度。在这方面，兴趣点（POI）在变量重要性方面并未排名靠前。这可能是由于在加纳主要城市和农村地区，兴趣点的覆盖范围存在异质性。

另一个讨论点是区级现有就业数据与0.005分辨率下预测的粒度之间的差异。第3.3节讨论了预测模型在预测区级统计数据方面的良好表现。然而，从更细粒度的行政级别，例如议会，开始可能会提高模型的准确性。

模型改进也可以依赖于替代估计方法。本文中，方法依赖于随机森林。如XGBoost、多元线性回归、支持向量回归或Res-FuseNet等替代估计技术可以帮助提高像素级投影的性能。模型改进的一个来源在于其更好地预测高人口密度地区的能 力。本文讨论了随着区域规模的增加，实际值与估计值之间的分散程度增加。这是高空间分辨率人口和GDP数据中的常见结果。Jin等人（2023）表明，区域到区域的克里金模型往往能更好地估计大人口单位。在其他高分辨率降尺度研究中也发现了类似的挑战。例如，美国州级职业就业数据已降至人口普查区，为超过800个职业产生了区级估计（Wang等人，2024）。深度学习方法也已被应用于绘制发展不均衡地区的精细社会经济变量图，包括人口和就业规模（Ahn等人，2023）。这些例子既展示了精细空间分解的可能性，也强调了其局限性，进一步证实了在像素级投影中方法改进的潜力。

我们的分析在选取建成区时聚焦于住宅区，将通勤和就业地点问题留待未来研究。总体来说，这些结果表明，高分辨率就业数据可以为地方劳动力政策和规划提供宝贵信息。尽管在精细尺度预测方面存在一些局限性，但该模型在大多数就业类别和空间单元上提供可靠的估计。

5 结论

这项研究介绍了一种开创性方法，用于生成高分辨率（0.005）网格状加纳劳动力市场数据，解决了空间分层数据集中一个关键的空缺。通过运用随机森林算法、遥感技术和包括土地利用覆盖、夜间灯光、基础设施和兴趣点在内的64个辅助变量，我们将区级人口普查数据成功下缩小至17个就业类别，涵盖了年龄、性别、技能、状况、部门、失业和青年无业人员（NEET）。

该模型实现了高精度（多数类别）并揭示了加纳劳动力市场存在显著的地理异质性。像素级别的分解提供了新颖、细致的视角，揭示了就业分布的独特模式。就业高度集中在主要城市聚集地，尤其是在加纳南部的阿克拉和库马西之间，以及沿海和主要交通基础设施沿线。相比之下，国家的北部半部分显示出较为稀疏的就业，主要集中在四个关键聚集地。薪金就业主要与城市地区相关联，这些地区要么是公共行政，要么是制造业和工业活动所在地。同样，低技能就业在市中心的总就业中所占比例较低，这反映了高技能工作主要集中在主要城市地区的核心区域。尽管制造业就业在全国范围内仍然微不足道，但在某些局部城市中心却不容忽视，这表明加纳存在专门的制造业中心。对就业率的分析显示出广泛的差异，像素之间的就业率从10%到98%不等，并突显了明显的南北差异。网格化数据也强调了城市就业的突出地位。

变量重要性分析突出了建成区、夜间灯光、道路密度和植被健康在预测就业模式中的作用，并考虑了特定行业的细微差别——例如植被指数对农业就业和基础设施对制造业的影响。尽管模型表现稳健，但在捕捉劳动力市场参与决策方面仍存在挑战，尤其是对于青年、妇女和老年工作者，家庭结构和教育在其中起着关键作用，但在地理空间数据中反映较少。

该方法超越了传统的GDP或人口网格方法，通过引入劳动力市场的复杂性，提供了一个适用于数据稀缺环境的可扩展框架。未来的工作可以对辅助数据的选取进行细化，探索替代的机器学习技术（例如，逐步随机森林、随机森林区域能量相关回归克里金法、XGBoost或Res-FuseNet），以及整合更细的行政区划或额外的辅助数据，进一步提高精确度。

尽管人口普查包括就业模块，但劳动力市场通过专门的调查，如劳动力调查或家庭收入支出调查，更能有效地捕捉。这些调查提供了详细的劳动力市场指标——如非正规就业或工资/收入类别——并且其较高的频率使得每年或每两年有可能生成网格化的劳动力市场数据。然而，它们的样本规模较小，往往限制了在区级层面的代表性。一种解决方案是利用调查家庭的地理坐标（纬度和经度），在公开可用的前提下，生成高分辨率的网格化劳动力市场数据。

网格化劳动力市场地图为政策制定者提供了一个宝贵资源，帮助他们设计有针对性的干预措施，解决当地就业挑战，并在加纳等类似背景下促进更包容的经济增长。

参考文献

Aaronson, D., R. Dehejia, A. Jordan, C. Pop-Eleches, C. Samii, 和 K. Schulze (2020, 08). 过去两百年生育对母亲劳动供给的影响。 *《经济期刊》131* (633), 1–32.

Ahn, D., M. Song, S. Lee, Y. Choi, J. Kim, S. Park, H. Yang, and M. Cha (2023). Using distributional adjustment for fine-grained socio-economic prediction from satellite images. In *会议录：第32届ACM国际信息与知识管理大会* ，CIKM '23，纽约，纽约州，美国，第3717-3721页。美国计算机协会。

Amankwah, A.，P. Castaing，N. Owoo，A. Palacios-Lopez（2022）。加纳劳动力市场参与与就业选择：个人性格特征和性别角色态度是否重要？ *世界银行政策研究工作论文10664*

D. Azar，J. Graesser，R. Engstrom，J. Comenetz，R. M. L. Jr.，N. G. Schechtman，T. Andrews (2010)《利用遥感估算的 Haiti 不透水表面来优化人口分布的空间细分》 *国际遥感杂志 31* (21), 5635–5655.

巴哈博滕，W.（2013）。加纳失业的成因。 *非洲发展评论 25* (4), 385–399.

Bhaduri, B.，E. Bright，P. Coleman，M. L. Urban（2007）。Landscan usa：一种高分辨率地理空间和时间模型方法，用于人口分布和动态变化。 *地理杂志 69* (1/2), 103–117.

Bosco, C., V. Alegana, T. Bird, C. Pezzulo, L. Bengtsson, A. Sorichetta, J. Steele, G. Hornby, C. Ruktanonchai, N. Ruktanonchai, E. Wetter, 和 A. J. Tatem (2017). 探索按性别划分的发展指标的高分辨率制图。 *《皇家学会接口》杂志 14* (129), 20160825.

Briggs, D. J., J. Gulliver, D. Fecht, 和 D. M. Vienneau (2007). 使用土地覆盖和光辐射数据对小区域人口分布进行等距建模。 *环境遥感*
108 (4), 451–466.

陈，Y.，吴，G.，葛，Y.，徐，Z.（2022）。利用深度学习与多源地理大数据绘制中国网格化国内生产总值分布图。 *IEEE应用地球观测与遥感选集期刊* 15, 1791–1802.

陈，Y.，张，R.，金，Y.，夏，Z.（2019年10月）。使用地理加权面积到点回归克里金法降尺度人口普查数据以进行网格化人口地图绘制。 *IEEE 访问* 7, 1–1.

Dobson, J. E., E. A. Bright, P. R. Coleman, R. C. Durfee, 和 B. A. Worley (2000). Landscan：用于估计风险人群的全球人口数据库。 *摄影测量工程与遥感 66* (7), 849–857.

胡贝尔，J. D. 和 L. 玛约尔（2024年3月）。像素中的经济发展：夜灯的局限性以及新的空间分解消费和贫困测量。工作论文1433号，巴塞罗那经济学院。

Jean, N.，Burke, M.，Xie, M.，Davis, W. M.，Lobell, D. B.，Ermon, S. (2016)．结合卫星影像与机器学习预测贫困 *科学* 353 (6301), 790–794.

Jedwab, R. 和 A. Moradi (2016). 贫困国家交通革命永久性影响的证据：来自非洲的研究 *《经济与统计评论》98期* (2), 268–284.

Jin, Y. 等 (2023). 使用随机森林区域至区域回归克里金模型在泰国以1公里分辨率绘制国内生产总值分布图。
ISPRS国际地理信息期刊12（12） .

Klasen, S., L. Tu-Thi-Ngoc, J. Pieters, 和 M. Santos-Silva (2021). 什么推动了女性劳动力参与？来自八个发展中经济体和新兴经济体的可比微观证据 *《发展研究杂志》第57期* (3), 417–442.

Kummu, M.，M. Taka，以及J. H. A. Guillaume（2018年）。关于1990-2015年全球国内生产总值和人类发展指数的栅格化全球数据集 *科学数据 5* (1).

Lebakula, V.，K. Sims，A. Reith，A. Rose，J. McKee，P. Coleman，J. C. Kaufman，M. Urban，C. Jochem，C. Whitlock，M. Ogden，J. Pyle，D. Roddy，J. Epting，E. A. Bright（2025）。2000年至2022年全球30弧秒年全球网格化人口数据集Landscan global。 *科学数据 12* ，495.

Leyk, S.，A. E. Gaughan，S. B. Adamo，A. de Sherbinin，D. Balk，S. Freire，A. Rose，F. R. Stevens，B. Blankespoor，C. Frye，J. Comenetz，A. Sorichetta，K. MacManus，L. Pistolesi，M. Levy，A. J. Tatem，和M. Pesaresi（2019）。人口空间分配：大规模栅格人口数据产品的综述及其适用性。 *地球系统科学数据11* (3), 1385–1409.

李，C.，余，L.，奥洛，F.，奇米姆巴，E. G.，坎博姆贝，O.，阿萨莫阿，M.，奥波库，P. D.，奥格韦诺，V. W.，法韦特，D.，洪，J.，邓，X.， Gong，P.，Wright，J.（2023）。撒哈拉以南非洲紧凑型和郊外社区的贫民窟和城市贫困。 *可持续城市与社会 99* ，104863.

村上，D. 和 山田，Y. (2019). 基于空间显式统计降尺度的网格化人口和GDP情景估计。 *可持续性 11* (7).

Murakami, D.，Yoshida, T.，Yamagata, Y. (2021)。与五个SSPs（共同社会经济途径）相兼容的网格化GDP预测。 *《建筑环境前沿》第7期*

Padilla, A., D. Otarod, S. Deloach-Overton, R. Kemna, P. Freeman, E. Wolfe, L. Bird, A. Gulley, M. Trippi, C. Dicken, J. Hammarstrom, 和 A. Brioche (2021). 编制非洲矿产资源及相关基础设施的地理空间数据（GIS）。美国地质调查局数据发布。

Stevens, F. R., A. E. Gaughan, C. Linard, and A. J. Tatem (2015, 02). Using random forests to disaggregate census data for population mapping based on remote sensing and auxiliary data. *PLOS ONE* 10 (2), 1–22.

唐，李；韦纳，T. T.（2023）。从高分辨率卫星图像绘制的全球采矿足迹图。 *《通信地球与环境》4* ，134.

Tatem, A. J. (2017). Worldpop，空间人口统计的开放数据. 科学数据 4，170004.

王，S.，Agrawal，S.，Mack，E. A.，Kalani，N.，Cotten，S. R.，Chang，C.-H.，Savolainen，P. T.（2024）。从州级到普查区级对职业就业数据的降尺度。 *应用地理学 170* ，103349.

王，T. 和 孙，F.（2022）。与共享社会经济路径一致的全球网格GDP数据集。 *科学数据 221* ，2052–4463.

王怡，黄晨，赵明，侯嘉，张怡，顾军（2020）。利用NPP/VIIRS和基于随机森林模型的兴趣点数据绘制中国大陆人口密度图。 *遥感 12* (21).

Wardrop, N. A., W. C. Jochem, T. J. Bird, H. R. Chamberlain, D. Clarke, D. Kerr, L. Bengtsson, S. Juran, V. Seaman, A. J. Tatem（2018年）。无国家人口和住房普查数据的区域分解人口估算。 *《美国国家科学院院刊》115卷* (14), 3529–3537.

Wu, N., J. Yan, D. Liang, Z. Sun, R. Ranjan, and J. Li (2024). 利用遥感与POI数据的多尺度特征融合进行高分辨率GDP地图绘制。 *《国际应用地球观测与地理信息期刊》129* ，103812.

附录

表4：模型评估粗分辨率 - 0.05°

名字	(1) MSR	(1) 百分比差异
Et	0.21	0.92
Ety	0.23	0.89
Etpa	0.21	0.92
Etold	0.24	0.91
Etw.	0.21	0.92
Etm	0.20	0.92
Ut.	0.29	0.91
NEET	0.21	0.87
Eths	0.27	0.93
ETLs.	0.20	0.91
Etse	0.22	0.90
Etees	0.29	0.92
Etag	0.19	0.71
Etmi.	0.76	0.75
Etma	0.36	0.89
Etind.	0.29	0.92
Etserv	0.28	0.93

此表展示了每个就业类别在分辨率0.05下的均方残差和解释的变异百分比。

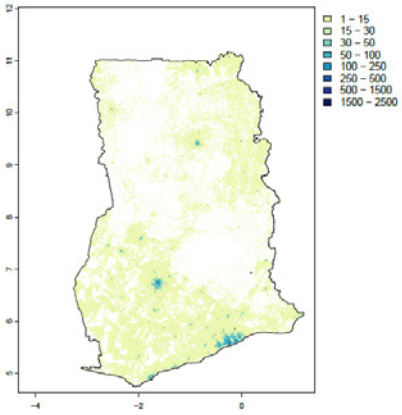
X **表格5：模型评估人口类别**

名字	(1) MSR	(1) 百分比差异
流行	0.16	0.92
wap	0.15	0.93
wap y	0.15	0.92
WAP PA	0.14	0.94
wap旧	0.14	0.92
WAP W	0.14	0.93
wap m	0.13	0.94

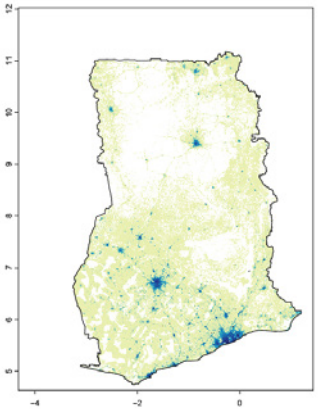
该表格展示了每个种群相关类别均方残差和解释的变差百分比。

图8：（所选）就业类别空间分布

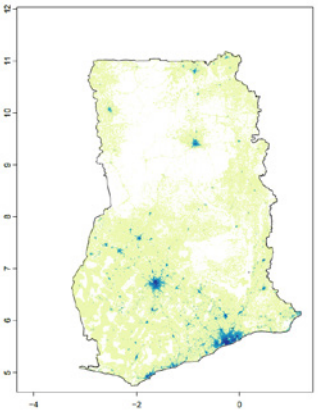
（a）青年 - 0.005°



(b) 最好年龄 - 0.005°



(c) 老工人 - 0.005°



(女性) - 0.005°

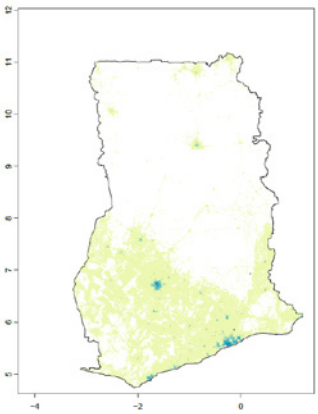
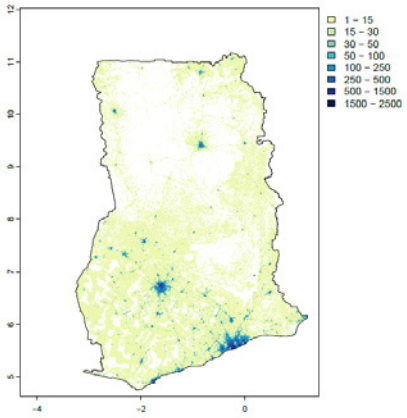
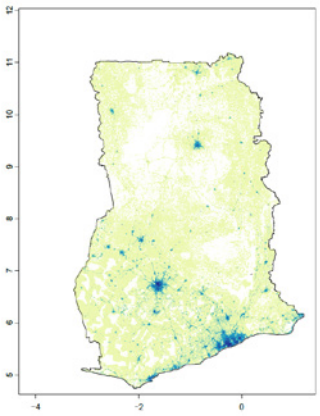


图9：（选定）就业类别空间分布

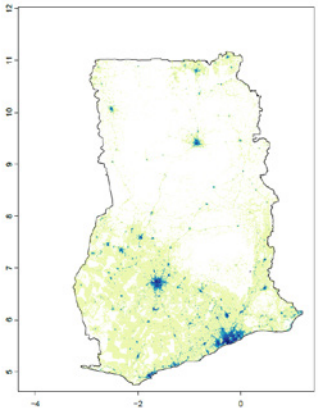
(a) 男性 - 0.005°



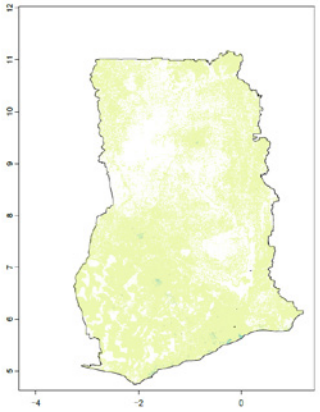
(b) 自雇 - 0.005°



（c）高技能人才——0.005°

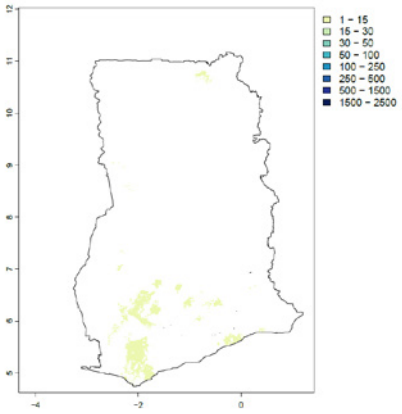


(农业 - 0.005°)

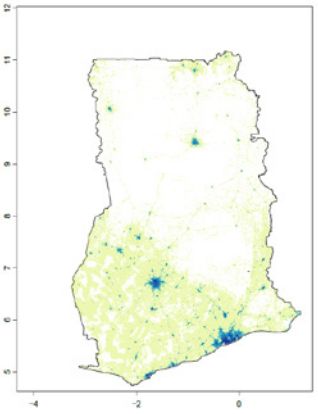


X **图10：（选定）就业类别空间分布**

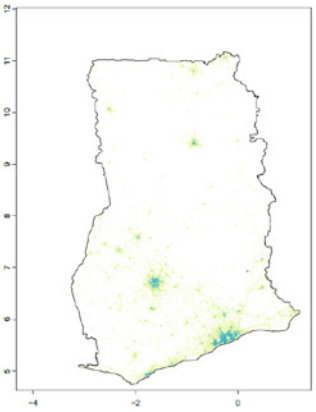
(a) 矿业 - 0.005°



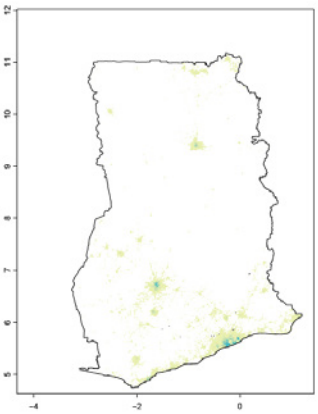
(b) 服务 - 0.005°



(工) 行业 - 0.005°

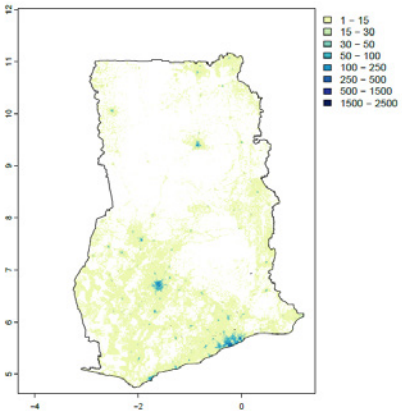


(d) 失业率 - 0.005°



X 图11：额外就业类别空间分布

(a) NEET - 0.005°



致谢

我们感谢Sara Elder、Lorenzo Guarcello、Luca Fedi、Marie Urban、Carter Christopher、Viswadeep Lebakula、Andrew Tatem、Chris Nnanatu和Yacouba Diallo就宝贵意见和建议所进行的讨论。我们感谢Yves Perardel和Vipasana Karkee在住房与人口普查的劳动力市场模块方面的交流。我们对加纳劳工部Sandra Ofosuapea女士，以及加纳统计服务部门的Johnson Kagya和Anthony Oduro-Denkyirah表示感激。我们还要感谢Mathieu Chenut在编辑方面的支持。

推进社会正义，促进体面劳动

国际劳工组织是联合国在劳动领域的机构。我们汇聚政府、雇主和工人，共同提升所有人的工作生活，通过创造就业、工作权利、社会保障和社会对话，推动以人为中心的未来工作方式。

联系信息

就业政策部门（EMPLOYMENT）

国际劳工组织
Route des Morillons 4
1211 日内瓦 22 瑞士
T +41 22 799 7861
employment@ilo.org
www.ilo.org/employment

ISBN 9789220432983