arXiv:2304.03901v1 [stat.ME] 8 Apr 2023

复合指标的小区域估计：多维贫困发生率案例研究

Alejandra Arias-Salazar[1],Andr´és Guti´érrez[2],Xavier Mancero[2], Stalyn Guerrero-G´ómez [2]，Natalia Rojas-Perilla [3], 张汉文[4]

[1] 哥斯达黎加大学，[2] 拉丁美洲和加勒比经济委员会，[3] 阿联酋大学，[4] 智利自治大学

摘要

本文提出了一种方法，用于在目标为复合指标时获得小区域估计。这些指标对于研究多维现象至关重要，但关于如何在小区域背景下获得此类指标估计的研究却很少。复合指标尤其复杂，因为其构建需要不同的数据来源、聚合程序和加权，这不仅使小区域估计面临挑战，也使不确定性度量的获取变得困难。作为对本文方法的案例研究，我们估计了哥伦比亚市镇层面的多维贫困发生率。此外，我们还基于参数自助法算法提供了不确定性度量。

关键词：复合指标；小区域估计量；广义线性混合模型；多维贫困。

1 引言

官方统计数据是决策者的有用工具，因为它们提供有关一个国家人口特征的信息，并使决策者能够实施和监测针对特定人口群体的公共政策。复合指标通常用于概括涉及两个或更多维度的复杂现象。这类指标应用于不同领域，例如社会科学：人类发展指数（HDI）（联合国开发计划署，2023年b）、性别发展指数（联合国开发计划署，2023年a）、多维贫困指数（联合国开发计划署，2023年c）、腐败感知指数（国际透明组织，2023年）；环境领域：环境绩效指数（Wolf 等，2022年）；经济领域：领先指标综合指数（经济合作与发展组织，2023年）。

复合指标有助于对相关方面进行联合分析，从而更全面地理解复杂现象，更好地传达结果，并促进决策（CEPAL，2013年）。

除了对公共政策的益处外，复合指标的构建和使用也应谨慎处理。例如，不同的加权和汇总方法可能会导致误导性结果，而且应评估用于构建指标的不同数据来源的质量（Freudenberg，2003年）。

复合指标构建及其适用性面临的挑战之一是细分。对细分信息的需求正在迅速增长，因为要制定并实施策略以改善一个国家人口的生活质量，必须掌握这些信息，涉及就业、贫困、教育和健康等方面。近期一个恰当体现这一需求的案例是可持续发展目标，其旨在提供按人口不同相关特征细分的信息（联合国大会，2015年）。

小区域估计（SAE）被提出作为解决细分问题的一种方法。小区域估计方法旨在通过结合两个或更多信息来源，在小区域内生成具有足够精度的准确估计。大多数此类方法通常基于单位层面或区域层面的回归模型；如果现有辅助信息与调查数据之间的相关性足够强，这些方法便能提高估计效率（Pfeffermann，2013年；Rao 和 Molina，2015年；Pratesi，2016年；Tzavidis 等，2018年）。然而，据我们所知，目前尚无文献探讨如何在小区域背景下获得高质量的复合指标估计。与此主题相关的文献主要关注潜在指标的降维，例如经济福祉（参见 Moretti 等，2020年和 Moretti 等，2021年）。

尽管构建复合指标需要遵循一些标准步骤（欧盟委员会联合研究中心和经济合作与发展组织，2008年），但我们仅关注对已明确定义的指标及其权重进行估计。这意味着我们不涉及降维、标准化、汇总和加权等方面。

复合指标可以采用不同方式汇总信息。一种方法是先在总体层面进行汇总，然后再合并各个维度。人类发展指数就是一个例子，它基于国家层面的汇总指标。另一种方法是先跨维度汇总，再跨个人汇总。未满足基本需求指数或 Alkire 和 Foster（2007）提出的多维贫困指数等指数便属于这种情况。本文提出一种后一类型的复合指标方案，首先在个体层面对定义为二分变量的指标进行汇总。

为说明我们提出的方法，我们以多维贫困发生率为例。该指数基于 Alkire 和 Foster（2007）方法，先识别福祉各相关维度上的剥夺，再在家庭层面进行汇总。主要区别在于，本文提出的指数仅对应发生率，而发生率是多维贫困指数的一个组成部分。本文所定义的多维贫困发生率仅用于说明。

多维贫困指数的各组成部分通常使用住户调查数据进行估计。这些调查工具通常收集家庭层面多个不同福祉维度的信息，这是 Alkire 和 Santos（2010）方法以及其他考虑同时存在的匮乏重叠情况的方法所必需的。拉丁美洲调查还具有更新频繁（多数国家每年更新）且样本具有全国代表性的优点。

住户调查的缺点之一是难以针对特定人口群体和地理区域进行细分分析。除国家层面的贫困率外，我们还希望了解哪些人口群体受贫困影响更严重，以及每种具体剥夺对其贫困程度的相对贡献，从而为公共政策的实施提供有用信息。

行政记录和人口普查等数据来源更适合实现更高程度的细分，但它们也有各自的局限。在拉丁美洲国家，行政记录通常无法在个体层面获取，质量不足以生成可靠统计数据，或者无法提供同一个人不同剥夺情况的信息。另一方面，人口普查通常每10年才开展一次，且收集的变量范围有限，无法用于计算完整的多维贫困指数。

本文旨在采用小区域方法，以哥伦比亚为案例研究，对多维贫困指数发生率进行市级分解估计。为此，我们应用基于模型的估计方法，将调查数据与人口普查数据整合起来。该方法力求保留每个匮乏指标的信息，使最终指数（多维贫困发生率）能够按维度分解。作为不确定性度量，我们通过参数自助法推导出均方误差（MSE）。

本文结构如下：第2节阐述了如何获得复合指标的小区域估计，以及如何计算相应点估计的均方误差估计。第3节通过一个案例研究展示该方法的应用。第4节介绍一项用于验证该方法的模拟实验。第5节总结全文并讨论进一步的研究方向。

2 多维贫困发生率：小区域估计方法

与小区域估计中的其他问题不同，在这种情况下，获得域级最终指标（例如总量或均值）并无用处。由于多维贫困发生率的性质，计算时需要掌握人口普查中每个个体在各项指标上的信息。这意味着，首先要对人口普查中的所有人逐一估计其在每项指标上的剥夺状态（是否受到剥夺）。

通常，各类复合指标所需的数据类型可能有所不同。有些指标由数值变量构成，例如预期寿命、收入或受教育年限。本文重点介绍一种方案，用于获取仅由二分变量构成的复合指标的小区域估计。

本节首先介绍案例研究中采用的示例：多维贫困发生率。接着，我们介绍针对二元响应变量（指标）的小区域估计方法。然后，我们提出在人口普查中缺失多个指标时所采用的方案。最后，我们还将说明如何为这一情形确定不确定性度量。

2.1 多维贫困发生率

由于成本高昂，国家统计调查往往无法在小样本量条件下提供可靠的域级社会人口统计估计。小区域估计（SAE）程序是指在小样本量条件下获取此类高度细分目标信息的估计方法。其基本原理是通过构建适当的模型，将调查数据与登记数据相结合，从而改进经典程序。

本文聚焦于基于二分变量构建的复合指标。因此，我们选取多维贫困发生率作为示例，因为它是一个广为人知的指标。Alkire 和 Foster（2007）方法提出了全球多维贫困指数，其一项特点是，每个国家都可以根据自身需求定义自己的指标和维度集合。无论具体定义如何，多维贫困指数都具有相同的特征：用于解释剥夺的指标被归入各个维度。维度对应于与贫困概念相关的福祉要素。对每个维度的

测量通过特定指标来具体实现，这些指标依据可获取的信息选取，通常来自住户调查。对于每个指标，都会使用一个剥夺截断值来判定某人是否应被视为遭受剥夺。

假设有一个包含 K 个指标的指数，这些指标以剥夺状态衡量：$y_{dj}^k = 1$ 表示个体遭受该项剥夺，$y_{dj}^k = 0$ 表示个体未遭受该项剥夺。该指数需要每个个体在 $j = 1$、……、N_d个领域中的 $d = 1$、……、D 项信息，其中 N_d 表示领域 d的人口规模。Alkire 和 Foster（2007）提出的领域 d 全球多维贫困指数计算如下：

$$\text{MDI}_d = \text{H} \cdot \text{A},$$

其中，H 为多维剥夺的人数比率（即发生率），A 为强度。本文仅关注发生率，即人数比率 H，其定义如下：

$$\text{H}_d = \frac{1}{N_d} \sum_{j=1}^{N_d} I(q_{dj} > z). \tag{1}$$

指标函数 $I(\cdot)$在满足条件 $q_{dj} > z$ 时等于 1，其中z 是用于判定个体是否贫困的阈值。

如上所示，要计算多维贫困发生率（H），需要确定每个个体是否遭受剥夺（1）或未遭受剥夺（0）。因此，挑战并非像大多数小区域估计问题那样，获得“最终”的指标估计值（例如按领域划分的特定剥夺人口比例）。在本例中，挑战在于确定人口普查中每个个体在缺失指标上是否遭受剥夺。

2.2 二元变量的小区域估计

在许多应用中，小区域内关注的变量可能是二元变量，例如，y_{dj} =0 或 1，分别表示某种特定特征是否存在。对于二元情形，每个领域 $d = 1$、……、D 的目标估计可以是$\bar{Y}_d = \pi_d = \frac{1}{N_d} \sum_{j=1}^{N_d} y_{dj}$ 人口中具有该特征的比例π_{dj}，即 j 该领域内某个特定单位 d 取值为 1 的概率。

尽管已有其他针对二元结果的方法，例如基于 M-分位数建模的方法（Chambers 等人，2016年），本应用采用基于广义线性混合模型的传统方法。在这种情形下，采用 logit 链接函数的π_{dj} 定义如下：

$$\text{logit}(\pi_{dj}) = \log\Big(\frac{\pi_{dj}}{1-\pi_{dj}}\Big) = \eta_{dj} = \mathbf{x}_{dj}^T \beta + u_d \tag{2}$$

其中 $j = 1, \ldots, N_d$, $d = 1, \ldots, D$, β为固定效应参数向量，u_d为领域 d的区域特定随机效应，其分布为 $u_d \sim N(0, \sigma^2)$。 u_d

假设相互独立且服从 $y_{dj}|u_d \sim$ 伯努利分布(π_{dj})，其中 $E(y_{dj}|u_d) = \pi_{dj}$ 和$Var(y_{dj}|u_d) = \sigma^2_{dj} = \pi_{dj}(1-\pi_{dj})$。此外，$\mathbf{x}_{dj}$ 表示单位层面辅助变量 p 的 $p \times 1$ 取值向量。

由于我们的具体问题是针对多个指标识别剥夺（0,1），因此以这一单位层面的伯努利逻辑混合模型作为起点。关于拟合单位层面逻辑混合模型的不同算法推导，可参见 Morales 等（2021），具体包括：模拟矩法（MSM）、期望最大化（EM）算法、惩罚拟似然（PQL）算法（Gonz´alez z-Manteiga 等，2007），或 Morales 等（2021）所述的最大似然拉普拉斯（ML - Laplace）近似算法。为方便起见，我们采用最新的算法，因为 R 的 lme4 软件包提供了该算法。

当若干小区域未纳入样本，或样本量不足以得出可靠结果时，获得高质量小区域估计的目标可能难以实现。为此，可以定义经验最佳预测值（EBP）（Jiang 和 Lahiri，2001年；Jiang，2003年）。

对于领域内的概率、概率之和和比例等感兴趣的量，可使用蒙特卡罗模拟近似计算 EBP（Hobza and Morales, 2016）。在实践中，通常会避免采用这种方法，因为它没有闭式解，计算时需要进行数值近似（Chambers 等人，2016）。作为一种解决方案，π_{dj} 的 plug-in 预测量定义如下：

$$\hat{\pi}^{in}_{dj} = \frac{\exp(\mathbf{x}^T_{dj}\hat{\beta} + \hat{u}_d)}{1 + \exp(\mathbf{x}^T_{dj}\hat{\beta} + \hat{u}_d)}, \tag{3}$$

由此可以得到 Y_d的 plug-in 预测量：

$$\hat{\bar{Y}}^{in}_d = \frac{1}{N_d}\Big(\sum_{j\in s_d} y_{dj} + \sum_{j\in r_d} \hat{\pi}^{in}_{dj}\Big), \tag{4}$$

其中，s 和 r 分别表示样本内和样本外观测值。

2.3 H 预测量的点估计

设 K 为人口普查中每个个体 j 缺失的指标。所提出的 H 估计程序如下：

1. 使用样本数据，为每个指标拟合个体层面的伯努利 logit 混合模型，并估计 $\hat{}\beta^k$、$\hat{}u^k_d$，最后估计 $\hat{}\pi^{in,\ k}_{dj}$，其中 $k = 1, \cdots, K$如方程 3 所示。

2. 对于 $l = 1, \ldots, L$ 蒙特卡罗模拟：

   • 对人口普查中的每个个体，预测其第 k个指标取值为 1 的概率，即 $\hat{}\pi^{in,\ k,\ (l)}_{dj} \quad \forall\, j \in U_d$。

• 使用 $y_{dj}^{k} \sim$ 伯努利(ˆπ^{in}, $_{dj}$ ^{k})进行蒙特卡罗估计 $\tilde{y}^{k}$, $_{(l)\ dj}$。

• 使用人口普查中已有的指标以及新指标 $\tilde{y}^{k}$, $_{(l)\ dj}$，按照公式1计算$\mathrm{H}_{(l)}$。

3. 每个小区域的最终点估计值 d 通过对每次 L模拟取均值计算得出：

$$\widehat{\mathrm{H}}_d = \frac{1}{L}\sum_{l=1}^{L} \mathrm{H}_d^{(l)}.$$

请注意，根据这一方案，即使有若干指标缺失，也可以估计多维贫困发生率（H） $K \geq 1$。

仅缺失一个指标时，这种方法的证明见附录A；人口普查中缺失两个指标时，证明见附录B。

2.4 均方误差的估计

使用小区域估计量时，估计均方误差（MSE）以衡量精度，是估计社会人口信息的关键步骤。当关注变量为二元变量时，有一些近似方法可用于求得MSE的解析形式。Gonz´alez-Manteiga等人（2007）针对经验预测量的不确定性估计，提出了一种小区域稳健自助法（SAWB）。基于这一自助法方案，我们提出了一项修改，使其能够适用于目标估计量（即多维贫困发生率）包含多个组成部分或指标，且其中一个或多个指标通过小区域估计方法进行估计的情形。

所提出的参数自助法步骤如下：对于每个缺失指标 Y_k，其中 $k = 1, ...,$ K，以及对于 $b = 1, ..., B$， B表示自助抽样的次数：

1. 使用第2.2节中所述的已估计值 ˆβ^{k},ˆσ^{2},$_{u}$ ^{k}，生成u^{*},$_{d}$ $^{k} \sim N(0,$ ˆσ^{2},$_{u}$ $^{k})$ iid。

2. 为每个指标模拟一个自助法超总体 y^{k},$_{dj}^{(b)} \sim$ 伯努利(π^{in}, $_{dj}$ ^{*k},)，其中 π^{in}, $_{dj}$ ^{*k} $=$,$\frac{\exp(\mathbf{x}_{dj}^{T}\ \hat{}\beta^{k}\pm u^{*}\text{,}_{d}^{\ k})}{1+\exp(\mathbf{x}_{dj}^{T}\ \hat{}\beta^{k}+u^{*}\text{,}_{d}^{\ k})}$.

3. 按照方程1计算 $\mathrm{H}_d^{(b)}$。

4. 提取自助样本，并按照第2.2节所述的点估计—蒙特卡罗方法计算 $\widehat{\mathrm{H}}_d^{(b)}$。

5. $\widehat{\mathrm{MSE}}\left[\widehat{\mathrm{H}}_d\right] = 1/B\sum_{b=1}^{B}\left[\mathrm{H}_d^{(b)} - \widehat{\mathrm{H}}_d^{(b)}\right]^2$.

3 案例研究：哥伦比亚成年人口的多维贫困发生率

本文的案例研究以多维贫困指数发生率为例（见公式 1）。我们使用该指数仅作示例。本文所呈现的指数构成（维度和指标）并非哥伦比亚国家统计局（西班牙语：Departamento Administrativo Nacional de Estad´ ıstica，DANE）或其他国际组织所定义的构成。

在本例中，我们使用哥伦比亚的数据，因为该国拥有近期的人口和住房普查（from2018年）以及同年的住户调查。

表1描述了本例所使用指数的构成。该指数基于联合国拉丁美洲和加勒比经济委员会此前关于多维贫困指数可能结构的研究；该指数根据国家住户调查信息的可获得性制定，可用于拉丁美洲国家之间的比较（CEPAL，2014年；Santos和Villatoro，2018年）。该指数包括5个维度（住房；供水与卫生设施；能源与连接；教育；以及就业与社会保障）和 $K = 8$ 个指标。除两种情况外，成年人与老年人所适用的指标和界限值相同。对于教育维度，教育不足的界限值为：18至29岁的成年人12年，30至59岁的成年人9年，以及60岁及以上人群4年。对于就业与社会保障维度，成年人的“失业或就业收入不足”指标，在老年人中替换为“没有养老金或养老金收入不足”。

人口普查通常包含计算该指数所需的信息，但就业与社会保障维度除外，因为这些维度需要个人收入数据。本文假设人口普查中也没有教育信息，因此必须通过小区域估计方法进行估计。

表 1：指数构成、哥伦比亚人口普查中指标的可用性及目标人口。

维度	指标	权重	在人口普查中	目标
住房	劣质住房材料	1/10	Yes	成年人、老年人
	过度拥挤	1/10	Yes	成年人、老年人
供水和卫生设施	饮用水短缺	1/10	Yes	成年人、老年人
	缺乏卫生设施	1/10	Yes	成年人，老年人
能源和互联互通	缺乏互联网服务	1/10	Yes	成年人、老年人
	缺乏电力	1/10	Yes	成年人、老年人
教育	未完成的学业	2/10	**No**	成年人，老年人
就业和社会保障	无养老金或养老金不足	2/10	**No**	老年人
	失业或不足就业相关收入			成年人

我们关注方程中所述的多维贫困发生率

1. 此处，我们需要 $K = 8$ 个以剥夺为衡量标准的指标：如果个人存在该项剥夺，则记为 y_{dj}^{k} =1；如果个人不存在该项剥夺，则记为 $y_{dj}^{k} = 0$ 。

该指数需要每个个体在 $j = 1, \ldots\ldots, N_d$ 个领域中的信息，其中 N_d 表示该领域 d的人口规模。

当条件 $q_{dj} > z$ 满足时，指标函数 $I(\cdot)$等于 1。本文采用 z 的值为 0.4，即当 $q_{dj} > 0.4$时， $I(\cdot)$等于 1。q_{dj} 是一个加权量，考虑了构成该指数的 $K = 8$ 个指标（见表 1）：

$$q_{dj} = 0.1 \sum_{k=1}^{6} y_{dj}^{k} + 0.2 \sum_{k=7}^{8} y_{dj}^{k}.$$

求和的第一部分考虑住房、供水与卫生、能源与互联互通维度的指标。第二部分考虑教育、就业和社会保障维度的指标。后两者实际上是人口普查中缺失的两个指标，将采用第2.3节介绍的方法进行估计。

图 1 展示了在构成 H 的八个指标中，有六项存在剥夺的人口比例，即多维贫困发生率。这些地图依据人口普查信息，在市级层面绘制。由于人口普查未包含所需信息，因此未绘制两个缺失指标的地图。此外，无法计算 H。

3.1 数据来源

3.1.1 全国人口与住房普查（CNPV）2018年

全国人口与住房普查（西班牙语：Censo Nacional de Poblaci´on y Vivienda，CNPV）由DANE开展。虽然计划每10年进行一次，但由于行政和经济原因，最近一次人口普查比原计划晚了三年。首次在传统面对面访谈之外，采用了电子自填式调查收集信息。数据收集阶段于2018年持续了10个月（DANE，2019年）。

人口普查旨在收集哥伦比亚人口及其生活条件的人口统计信息，包括住房和家庭特征。这些信息对于该国的区域规划和决策至关重要（DANE，2019年）。

3.1.2 综合家庭调查（GEIH）2018年

大型综合住户调查（西班牙语名称为 Gran Encuesta Integrada de Hogares，简称 GEIH）由 DANE 每年开展。该调查提供有关

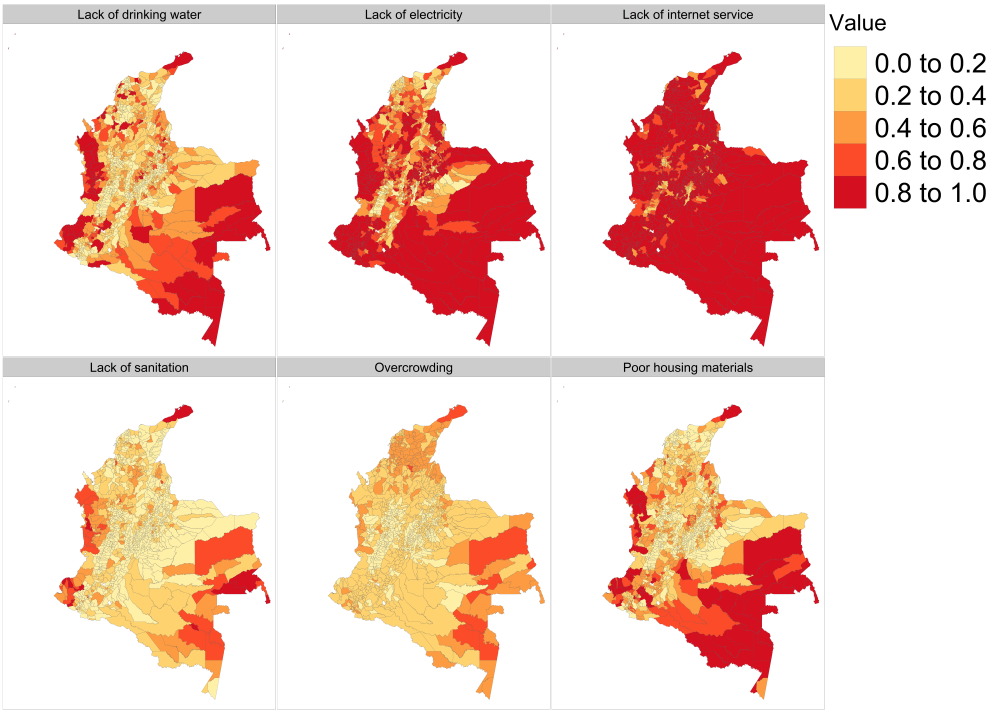


图 1：市级人口普查中可用的 H 指标

劳动力规模和结构，以及人口和家庭的社会人口学特征、住房、教育程度、社会保障参保情况、收入等方面的信息（DANE，2021年）。该调查是就业和收入贫困指标的官方信息来源，并包含计算联合国拉丁美洲和加勒比经济委员会多维剥夺指数所需的变量。在行政区划上，哥伦比亚分为 5 个大区、33 个省和 1122 个市镇；GEIH 提供具有代表性的国家层面、城市和农村地区、大区以及 33 个省中 24 个省的信息。2018年，有效调查样本包括 231,128 个家庭和 762,753 名个人。

本文使用的 GEIH 版本来自住户调查数据银行（BADEHOG）。该数据库由联合国拉丁美洲和加勒比经济委员会统计司维护，收录了 18 个拉丁美洲国家的住户调查数据。在该数据库中，各项变量均经过协调处理，以便构建各种指标并进行跨国比较。这一特点有助于简化对该地区各国多维贫困发生率的估计。

3.1.3 卫星影像

地理空间数据已成为提升地方区域估计质量的有力资源。尽管大数据总体上已因这一用途而广受欢迎，

地理空间数据有两项独特优势：易于获取，而且不存在偏差选择问题（Masaki 等，2020年）。

卫星影像作为辅助信息来源，已应用于多个小区域估计问题，例如福祉（Engstrom 等，2022年）、人口密度（Harvey，2002年；Deng 和 Wu，2013年；Steinnocher 等，2019年）、贫困制图（Babenko 等，2017年；Chandra 等，2018年）以及多维贫困（Pokhriyaland Jacques，2017年；Koebeetal.，2022年）等领域。

在本案例研究中，我们使用 Earth Engine Data Catalog（Gorelick 等，2017年）提供的资源。在众多可用产品中，我们将夜间灯光强度、城市覆盖比例和农作物覆盖比例等区域层面信息纳入模型。

3.2 结果

本节介绍将所提方法应用于哥伦比亚各省和市镇，以估计第 2.1 节所述多维贫困发生率所得的主要结果。

表2展示了各目标领域的样本规模和人口规模分布。人口普查规模为 $N = 34,180,812$，样本规模为 $n = 549,077$，覆盖33个省中的24个，以及1122个市镇中的438个。

表2：各省和市镇的样本规模与人口普查规模分布

领域	来源	Min	第一四分位数	中位数	Mean	第三季度	Max
省							
（样本内：）73%	调查	6613	21601	22344	22878	24692	35264
	人口普查	21037	256244	752416	1035781	1036702	5847519
市镇							
（样本内：）40%	调查	14	127	220	1254	369	24594
	人口普查	103	4433	8440	30464	17601.25	5847519

如表2所示，在本应用中，只有部分市镇的样本量较小，因此，提高估计准确度并非本例的主要问题，而这在许多小区域估计问题中却是主要问题。尽管如此，我们仍根据变异系数分析估计准确度，其定义如下：

$$CV(\widehat{\mathrm{H}}_d) = \frac{\sqrt{MSE(\widehat{\mathrm{H}}_d)}}{\widehat{\mathrm{H}}_d} \times 100.$$

直接估计和基于模型的估计所得变异系数的汇总统计量见表3.2。此处还按域级（省和市镇）对变异系数进行了细分

，并针对每个缺失指标（教育和就业）分别进行了分析。在省级层面，两项指标的变异系数均相对较小，直接估计和基于模型的估计也都是如此。样本外省份带来的不确定性可以解释基于模型的估计为何数值较高。采用小区域估计方法的优势在市镇层面最为明显，因为平均变异系数与直接估计相近，但第三四分位数和最大值均低于直接估计。教育和就业这两项指标都呈现出这一特征。

对于样本内市镇，图2清楚地显示，基于模型的估计具有稳定且较低的变异系数。

表3：各省和市镇变异系数的描述性统计（百分比）

领域/指标	估计	Min	第一四分位数	中位数	Mean	第三季度	Max
省							
教育	直接	0.64	1.24	1.43	1.47	1.67	2.09
	基于模型的	0.36	0.44	0.49	1.56	2.85	9.22
就业	直接	0.87	1.39	1.73	1.80	2.04	3.40
	基于模型的	0.26	0.40	0.53	1.66	2.90	9.24
市镇							
教育	直接	0.00	2.60	5.37	6.27	8.75	39.81
	基于模型的	0.45	3.81	5.37	5.23	6.51	10.87
就业	直接	0.00	2.33	5.19	5.91	8.46	27.60
	基于模型的	0.46	3.69	5.58	5.53	7.03	18.78

由于我们观察到，两个指标的直接估计所得变异系数均可接受，因此，应用我们提出的小区域估计方法的主要益处，是为样本外域获得可靠的估计：包括9个省级行政区和684个市镇，从而提供所有相关领域中目标人口处于多维贫困状态的比例。结果见图3至图5。

图3展示了在人口普查数据中无法获得、因而需要使用小区域估计方法来获取估计值的市级两个指标。

总体而言，哥伦比亚是一个拥有多个大区的国家，每个大区都面临着各自独特的经济和社会挑战。如图3所示，亚马孙地区是需要特别关注的大区之一，该地区包括瓜伊尼亚、沃佩斯和瓜维亚雷等省，以及乔科、瓜希拉和苏克雷等其他省。这些省的特点是贫困程度较高，教育和就业水平较低，且获得住房、供水和卫生设施等基本服务的机会有限。尤其是，教育是决定个体福祉和生活水平的关键因素。尽管这些省近年来在扩大教育机会方面取得了进展，但其入学率和受教育程度往往较低，导致许多儿童无法接受优质教育，并面临更高的贫困风险以及被正规经济排斥的风险。就业

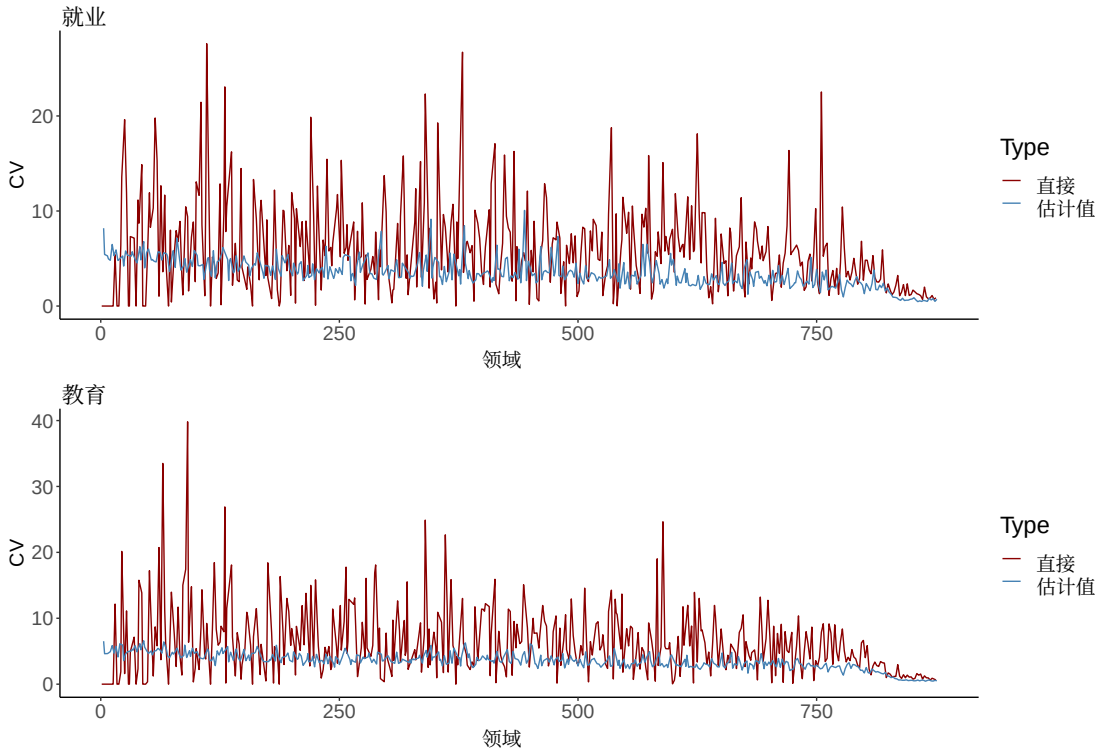


图2：就业和教育指标在省级和市级层面的直接估计值与基于模型的估计值的变异系数（百分比）。

与社会保障是估计的第二项贫困与福祉关键指标。尽管哥伦比亚的失业率近年来有所下降，但全国各地的就业状况仍存在差异。上述各省的失业率往往较高，非正规部门规模也较大，导致许多人无法获得正规就业，并面临更高的贫困风险以及被正规经济排斥的风险。在波哥大等一些省，正规就业和受教育程度较高，且相当一部分人口在正规部门工作并接受过高等教育。

利用这两个指标，现在可以计算最终的 H 值，结果如图4所示，涵盖哥伦比亚的各个市镇。最终 H 值的变异系数低于15%（图5），表明这些估计值在精度方面是“可接受的”。

通过所提出的方法，可以从两个角度进行更深入的分析：哪些剥夺最为严重？以及总体而言哪些地区受影响最严重？我们以考卡山谷省为例，该省 H 值处于最低水平之列（图4：该省位于国家中西部，多维剥夺指数为38.5%）。

考卡山谷省成年人口在各项剥夺中的占比较低

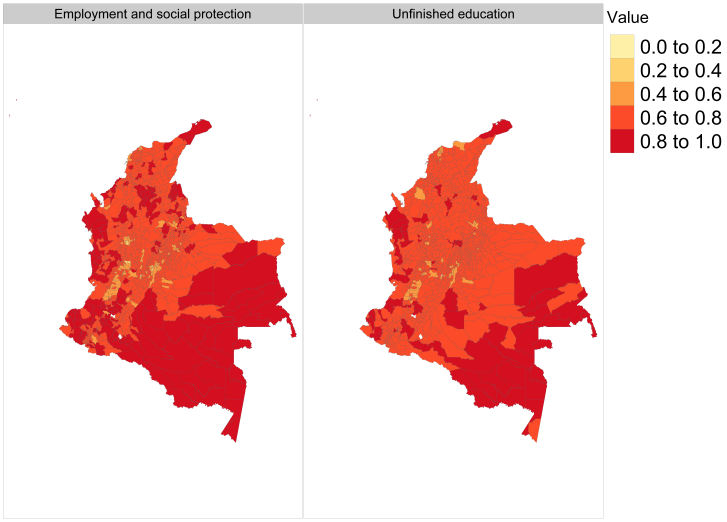


图3：市级就业与社会保障指标以及未完成教育指标的基于模型的估计。

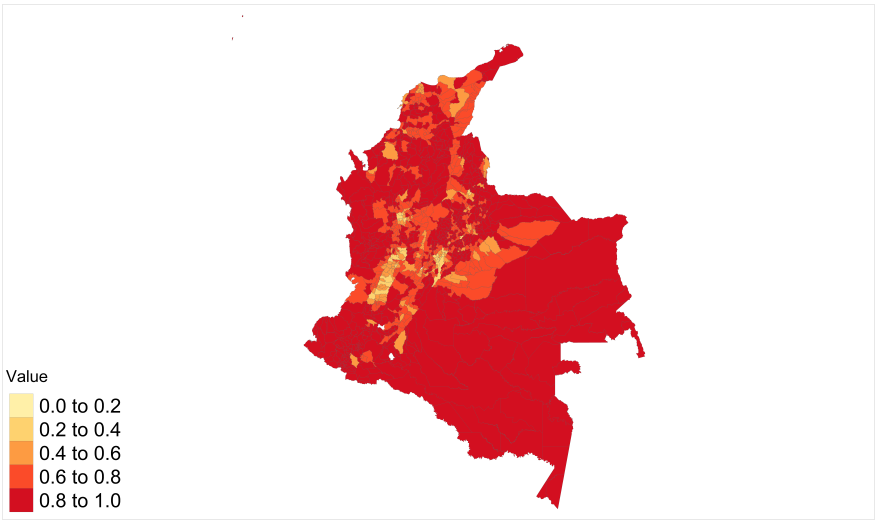


图4：市级最终 H。

在住房、过度拥挤、饮用水、卫生设施和电力方面（介于0%至20%），而在互联网服务、教育、就业和医疗保险方面则处于中等水平（40%至60%）（见图6）。此外，观察总体 H 时，可以看出该省的˙各市镇之间存在显著差异。图7显示，El Aguila和Dagua的H值较高（分别为92%和82%），而卡利和图卢阿等其他地区的数值相对较低（30%和39%）。值得

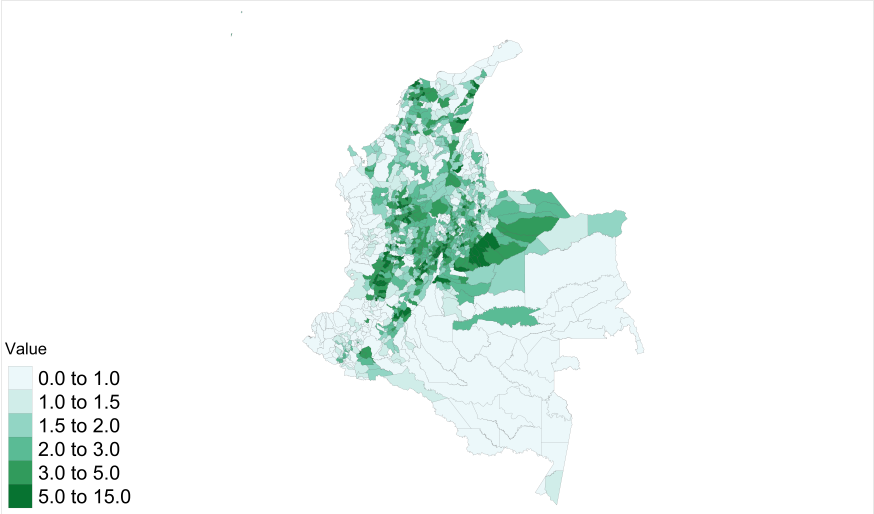


图 5：市级 H 的变异系数。

请注意，该省各市之间的贫困发生率仍存在显著差异。这凸显了采取有针对性的干预措施、应对这些挑战并减少该省多维贫困的必要性。

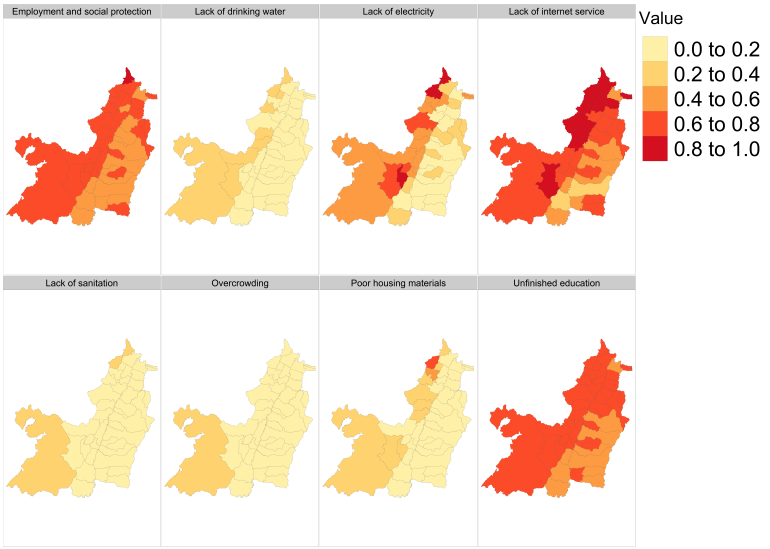


图 6：考卡山谷省：市级 H 指标。

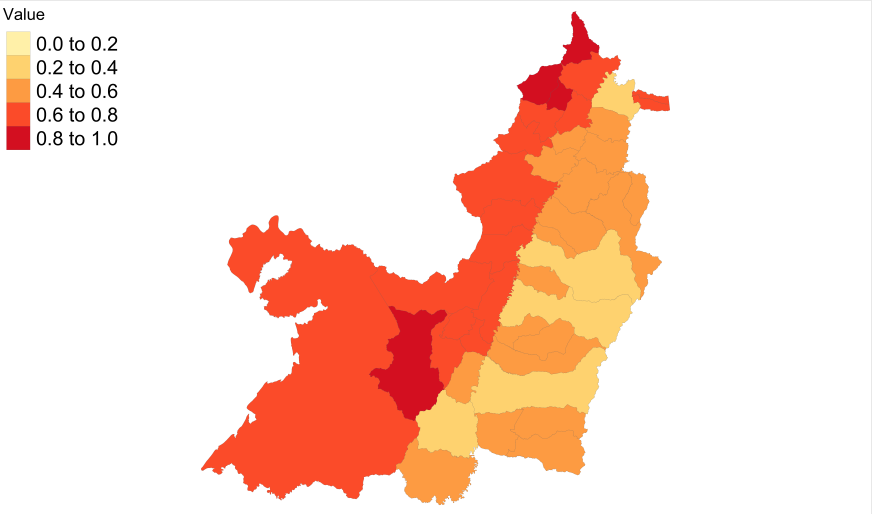


图7：考卡山谷省：市级最终 H。

4 评估

对所提出方法的评估分为两部分。首先，对直接估计和基于模型的估计进行内部比较。其次，介绍一项基于设计的模拟研究，以评估第2.2节所提出估计量的性能。

图8比较了样本内领域的直接估计和基于模型的估计。正如预期，在省级和市级两种情况下，采用所提方法得到的点估计都非常接近直接估计，皮尔逊相关系数分别为0.984和0.880。

采用基于抽样设计的模拟实验评估第2.2节提出的估计量的性能。

为在真实情境下进行评估，我们将哥伦比亚2018年全国人口与住房普查作为固定总体，并在不同样本量下按照简单随机抽样设计从中反复抽取样本：a）500个、b）5000个、c）50000个观测值。最后，还考虑原案例研究中的复杂抽样设计，进行情景d）的评估，样本包含550000个观测值。该复杂抽样设计反映了哥伦比亚2018年大型综合住户调查（GEIH）的特点，相关描述见第3.1小节。

与应用中的情况类似，目标指标是哥伦比亚小区域（即各省和市镇）的多维贫困指数。人口数据中已有八个变量中的六个，这些变量将保持固定。另外两个变量将按照第2.1节中介绍的方法获得。小区域预测量的性能通过变异系数（前文已述）、偏差以及

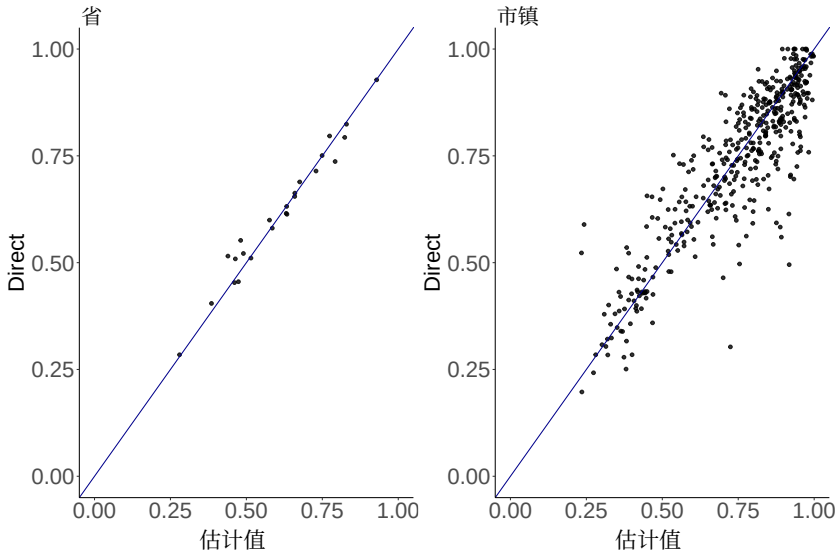


图8：省级和市级直接估计值与基于模型的估计值的 H 对比。

各领域的均方根误差进行评估。后两个度量定义如下：

$$\text{Bias}(\widehat{\text{H}}_d) = \frac{1}{T}\sum_{t=1}^{T}\left(\widehat{\text{H}}_d^{(t)} - \text{H}_d\right);$$

$$\text{RMSE}(\widehat{\text{H}}_d) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{\text{H}}_d^{(t)} - \text{H}_d\right)^2},$$

其中，H_d 在第2.1节中定义，下标 t 表示 T 模拟运行次数；在本练习中，我们设定 $T = 100$。

图 9 展示了所评估估计量在各领域中的领域特定偏差、均方根误差和变异系数的分布。正如预期，样本量增加会降低偏差并提高精度。对于“复杂”情形，这一点尤为明显，因为其样本量大于其他三个情景。

5 结论与进一步研究

复合指标在复杂现象研究中具有重要价值，且广泛应用于公共政策。随着对细分信息的需求日益增长，如何进一步提高这类指标的细分程度也成为一项挑战。小区域估计方法可以解决这一问题，但目前尚无针对复合指标这一特定情形的相关文献。本文旨在通过提出一种方法来弥补这一信息缺口，以获得小区域

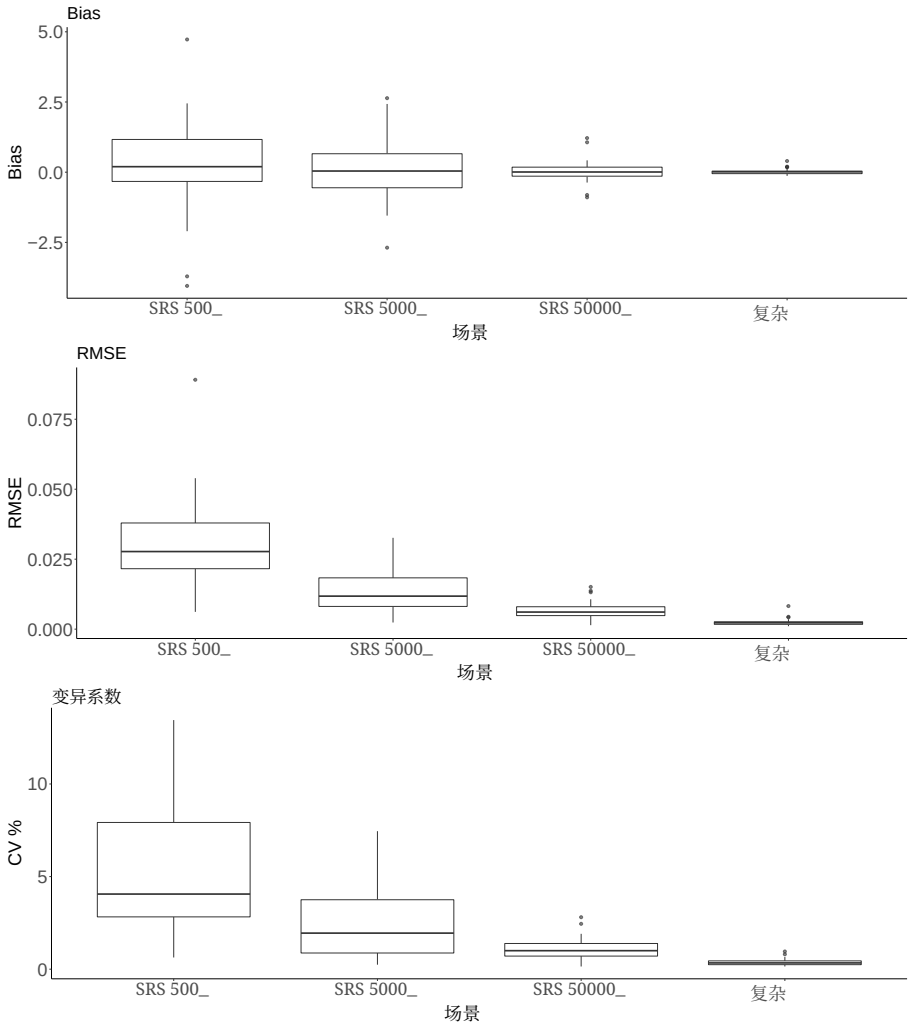


图 9：不同样本量和抽样设计下，多维贫困发生率区域特定点估计的性能度量。

估计值，适用于研究指标由二分变量构成的情形。我们以多维贫困发生率（H）为例说明这一方法。

在处理复合指标（如 H）时，如何生成小区域估计这一挑战引出了许多有待进一步研究的问题。首先，可以考虑一种更为通用的方法，将对指标和维度之间协方差结构及依赖关系的分析纳入其中。在本案例研究中，恰好估计了两个指标，并假定它们彼此之间（以及与其他指标之间）不存在依赖关系。其次，信息来源（如人口普查数据和调查数据）之间的时间间隔也是一个值得研究的问题。尤其是当某些指标可从人口普查数据中获得，而其他指标需要使用最新的调查数据进行估计时。第三，关于点

估计和均方误差估计，可以通过基于模型的模拟来验证所提出的方法。第四，可进一步研究基准程序，以纠正不同估计值之间可能存在的不一致（例如由抽样误差和非抽样误差造成的影响）。最后但同样重要的是，采用小区域估计方法得到的最终 H，旨在捕捉贫困在福祉多个维度上的复杂性——住房、供水与卫生设施、能源与互联互通、教育、就业与社会保障。然而，家庭内部不平等以及贫困人口中的不平等仍有待进一步研究加以捕捉。此类衡量方法能够在最需要的细分区域层面呈现一个国家贫困的全貌，已成为政策相关应用的重要基础。

当构成复合指标的变量并非二分变量时，还需要进一步的方法来获取小区域估计。人类发展指数（HDI）就是一个明显的例子，它由预期寿命、受教育年限和基尼系数组成。本文重点估计全球多维贫困指数的一个组成部分。还需要进一步研究，以便计算完整的指数，并考虑贫困强度。

致谢作者谨此感谢阿联酋大学阿联酋大学初创研究基金所提供的支持。

参考文献

Alkire, S. 和 Foster, J.（2007年）。计数与多维贫困测量。牛津贫困与人类发展倡议工作论文第7号。网址：https://ophi.org.uk/working-paper-number-07/（访问日期：2021年6月）。Alkire, S. 和 Santos, M. E.（2010年）。急性多维贫困：一项面向发展中国家的新指数。牛津贫困与人类发展倡议（*OPHI*）工作论文。Babenko, B.、Hersh, J.、Newhouse, D.、Ramakrishnan, A. 和 Swartz, T.（2017年）。利用经过高分辨率和中分辨率卫星图像训练的卷积神经网络进行贫困制图，并将其应用于墨西哥。CEPAL（2013年）。拉丁美洲社会概览´。智利圣地亚哥。CEPAL（2014年）。拉丁美洲社会概览´。智利圣地亚哥。Chambers, R.、Salvati, N. 和 Tzavidis, N.（2016年）。采用半参数小区域估计对二元结果进行估计，并将其应用于英国地方当局的失业率估计。《皇家统计学会学报：*A*辑（社会统计）》，179(2)：453–479。Chandra, H.、Aditya, K. 和 Sud, U. C.（2018年）。印度比哈尔邦贫困发生率的局部化估计与空间制图——小区域估计技术的应用。*PLOS ONE*，13(6)：1–14。DANE（2019年）。2018年全国人口与住房普查方法说明。技术报告。Deng, C. 和 Wu, C.（2013年）。改进小区域人口估计：一种整合地理与人口学的方法。美国地理学家协会年刊，103(5)：1123–1141。Engstrom, R.、Hersh, J. 和 Newhouse, D.（2022年）。太空视角下的贫困：利用高分辨率卫星影像估计经济福祉。《世界银行经济评论》，36(2)：382–412。Freudenberg, M.（2003年）。国家表现的复合指标。González-Manteiga, W.、Lombardía, M. J.、Molina, I.、Morales, D. 和 Santamaría, L.（2007年）。逻辑混合模型下小区域线性参数预测量均方误差的估计。《计算统计与数据分析》，51(5)：2720–2733。Gorelick, N.、Hancher, M.、Dixon, M.、Ilyushchenko, S.、Thau, D. 和 Moore, R.（2017年）。Google Earth Engine：面向所有人的行星尺度地理空间分析。《环境遥感》。

Harvey, J. T.（2002）。利用卫星影像估计普查区人口：一些方法与局限。《国际遥感杂志》，23(10):2071–2095。Hobza, T. 和 Morales, D.（2016）。单位层面 Logit 混合模型下的经验最佳预测。《官方统计学杂志》，32(3):661。Jiang, J.（2003）。基于广义线性混合模型的小区域推断经验最佳预测。《统计规划与推断杂志》，111(1-2):117–127。Jiang, J. 和 Lahiri, P.（2001）。基于二元数据的小区域推断经验最佳预测。《统计数学研究所年刊》，53(2):217–243。欧盟委员会联合研究中心和经济合作与发展组织（2008）。《复合指标构建手册：方法与用户指南》。经济合作与发展组织。Koebe, T., Arias-Salazar, A., Rojas-Perilla, N. 和 Schmid, T.（2022）。利用结构保持方法和卫星影像进行普查间更新。《皇家统计学会学报：*A*辑（社会统计）》。Masaki, T., Newhouse, D., Silwal, A. R., Bedada, A. 和 Engstrom, R.（2020）。利用地理空间数据进行非货币贫困小区域估计。Morales, D., Esteban, M. D., P´erez, A. 和 Hobza, T.（2021）。《小区域估计与混合模型教程》。施普林格。Moretti, A., Shlomo, N. 和 Sakshaug, J. W.（2020）。多维潜在经济福祉指标的多变量小区域估计。《国际统计评论》，88(1):1–28。Moretti, A., Shlomo, N. 和 Sakshaug, J. W.（2021）。潜在经济福祉的小区域估计。《社会学方法与研究》，50(4):1660–1693。经济合作与发展组织（2023）。综合领先指标。https://www .
oecd-ilibrary.org/economics/data/main-economic -indicators/composite-leading-indicators data-00042 -en。访问日期：2023年3月12日。Pfeffermann, D.（2013）。小区域估计领域的重要新进展。《统计科学》，28:40–68。Pokhriyal, N. 和 Jacques, D. C.（2017）。整合异质数据源以改进贫困预测与制图。《美国国家科学院院刊》，114(46):E9783–E9792。

Pratesi, M.（2016）。《利用小区域估计分析贫困数据》。John Wiley & Sons。Rao, J. N. K. 和 Molina, I.（2015）。《小区域估计》。Wiley，纽约，第2版。Santos, M. E. 和 Villatoro, P.（2018）。拉丁美洲多维贫困指数。《收入与财富评论》，64(1):52–82。Steinnocher, K., Bono, A. D., Chatenoux, B., Tiede, D. 和 Wendt, L.（2019）。利用立体卫星影像估计城市人口分布模式。《欧洲遥感学报》，52(sup2):12–25。国际透明组织（2023）。腐败感知指数。https://www.transparency.org/en/cpi/2021。访问日期：2023年3月12日。Tzavidis, N., Zhang, L.-C., Luna Hernandez, A., Schmid, T. 和 Rojas-Perilla, N.（2018）。从头到尾：生产小区域官方统计数据的框架。《皇家统计学会学报：*A*辑（社会统计）》，181(4):927–979。联合国开发计划署（2023a）。性别发展指数。https://hdr.undp.org/gender-development-index#/indicies/GDI。访问日期：2023年3月12日。联合国开发计划署（2023b）。人类发展指数。https://hdr.undp.org/data-center/human-development-index#/indicies/HDI。访问日期：2023年3月12日。联合国开发计划署（2023c）。多维贫困指数。https://hdr.undp.org/content/2022-global-multidimensional-poverty-index -mpi#/indicies/MPI。访问日期：2023年3月12日。联合国大会（2015）。第70/1号决议。《改变我们的世界：2030年可持续发展议程》。技术报告，联合国大会。Wolf, M. J., Emerson, J. W., C, E. D., de Sherbin, A. 和 A, W. Z.（2022）。*2022*年环境绩效指数。康涅狄格州：耶鲁大学环境法律与政策中心。

缺少一个指标时的证明

证明。 我们定义 $W=\alpha Y$，不难求得 W 的概率质量函数如下:

$$f_W(w)=\begin{cases}1-\pi & \text{if } w=0,\\ \pi & \text{if } w=\alpha,\\ 0 & \text{otherwise.}\end{cases}$$

另一方面:

$$\begin{aligned}E(Z)&=1-P(X<\delta),\\ &=1-P(\alpha Y+k<\delta),\\ &=1-P(\alpha Y<\delta-k),\\ &=1-P(W<\delta-k).\end{aligned}$$

上述表达式取决于 $\delta-k$ 的值，具体如下:

Case 1 When $\delta-k\leq 0$, $P(W<\delta-k)=0$.

情形2 当 $0<\delta-k\leq\alpha$,

$$P(W<\delta-k)=P(W=0)=1-\pi.$$

情形3 当 $\alpha<\delta-k$，因此

$$P(W<\delta-k)=P(W=0)+P(W=\alpha)=1.$$

由此，我们得出:

$$E(Z)=\begin{cases}1 & \text{if } \delta-k\leq 0,\\ \pi & \text{if } 0<\delta-k\leq\alpha,\\ 0 & \text{if } \alpha<\delta-k.\end{cases}$$

□

B 缺失两个指标时的证明

证明。 我们将随机变量Y_1 y Y_2的线性组合定义为 $W=\alpha(Y_1+Y_2)$。注意，该组合 Y_1+Y_2 只能取值0、1或2。因此， W 可以取值0、 α 和 2α，其对应的概率如下:

$$
\begin{aligned}
P(W=0) &= P(Y_1=0, Y_2=0) = (1-\pi_1)(1-\pi_2) \\
P(W=\alpha) &= P(Y_1=0, Y_2=1) + P(Y_1=0, Y_2=1) \\
&= \pi_2(1-\pi_1) + \pi_1(1-\pi_2) \\
P(W=2\alpha) &= P(Y_1=1, Y_2=1) = \pi_1\pi_2.
\end{aligned}
$$

W 的概率质量函数定义如下:

$$
f_W(w) = \begin{cases}
(1-\pi_1)(1-\pi_2) & \text{if } w=0, \\
\pi_2(1-\pi_1) + \pi_1(1-\pi_2) & \text{if } w=\alpha, \\
\pi_1\pi_2 & \text{if } w=2\alpha, \\
0 & \text{otherwise.}
\end{cases}
$$

另一方面:

$$
\begin{aligned}
E(Z) &= 1 - P(X < \delta), \\
&= 1 - P(\alpha(Y_1+Y_2) + k < \delta), \\
&= 1 - P(\alpha(Y_1+Y_2) < \delta - k), \\
&= 1 - P(W < \delta - k).
\end{aligned}
$$

上述表达式取决于 $\delta - k$ 的值，具体如下:

情形1 当 $\delta - k \leq 0$， $P\left(W < \delta - k\right) = 0$。

情形2 当 $0 < \delta - k \leq \alpha$,

$$
P(W < \delta - k) = P(W=0) = (1-\pi_1)(1-\pi_2).
$$

情形3 当 $\alpha < \delta - k \leq 2\alpha$，因此

$$
\begin{aligned}
P(W < \delta - k) &= P(W=0) + P(W=\alpha) \\
&= (1-\pi_1)(1-\pi_2) + \pi_2(1-\pi_1) + \pi_1(1-\pi_2).
\end{aligned}
$$

情形4 当 $2\alpha < \delta - k$，因此

$$
P(W < \delta - k) = P(W=0) + P(W=\alpha) + P(W=2\alpha) = 1.
$$

根据上述情况，我们得出结论:

$$
E(Z) = \begin{cases}
1 & \text{if } \delta - k \leq 0, \\
\pi_2(1-\pi_1) + \pi_1(1-\pi_2) + \pi_1\pi_2 & \text{if } 0 < \delta - k \leq \alpha, \\
\pi_1\pi_2 & \text{if } \alpha < \delta - k \leq 2\alpha, \\
0 & \text{if } 2\alpha < \delta - k.
\end{cases}
$$

□