

# 从纵向研究中提高调查估计的质量

## 在 LSMS 面板测量中的应用

*Piero Falorsi P  
aolo Righi Giuli  
a Ponzini*



**WORLD BANK GROUP**

发展经济学发展数据  
组 2025 年 1 月



**Reproducible Research Repository**

一个经过验证的可重复性包可在以下网址获取：<http://reproducibility.worldbank.org>，点击以获得直接访问。  
**here**

## Abstract

纵向调查是分析人类群体随时间变化状态和变动的一个极其宝贵的资源。然而，随着面板调查时间长度的增加，保持估计值的准确性变得更加困难。主要关注的是由于样本损耗（由死亡和搬迁引起）以及新生人口和迁移流的影响导致的样本代表性不足问题。此外，样本疲劳引入了不断增加的测量误差。正确的设计、实施和使用面板调查考虑了一系列方法在统计过程的不同阶段解决这些问题：抽样设计、数据收集和估计。本文专注于具有旋转样本设计的面板案例。这种案例代表了一种强大的混合解决方案，用于应对面板效应的影响。

样本代表性动态。从实证角度出发，本文着重探讨了估计方法。通过乌干达国家面板调查和最长生活标准衡量研究的数据，对提议的技术进行了实验性评估。总结来说，研究发现校正后的加权共享方法为基础的估计器能够提供个体层面的统计数据，这些数据似乎比当前乌干达国家面板调查的估计器更为精确。此外，基于转换矩阵的校正后的加权共享方法的基础横截面估计显示，在改变样本时，其稳定程度通常高于当前乌干达国家面板调查的估计值。

这篇论文是由发展数据组（Development Data Group）和经济发展（Development Economics）部门共同完成的。它是世界银行为提供研究开放访问、并为全球发展政策讨论做出贡献的一项更大努力的一部分。政策研究工作论文也已在网上发布于 <<http://www.worldbank.org/prwp>>。作者可以通过 [gponzini@worldbank.org](mailto:gponzini@worldbank.org) 联系。此论文的经过验证的可再现性包可在 <<http://reproducibility.worldbank.org>> 获取，请点击相应链接。 [here](#) 直接进入。



该政策研究工作论文系列发布正在进行中的研究成果以促进关于发展方面的思想交流。问题。该系列的一个目标是尽快发布研究发现，即使展示的内容尚未完全打磨。论文保留了作者的姓名应按照相应的方式引用。本论文所呈现的研究结果、解释和结论完全代表作者自身的意见。由作者提供。它们不一定代表国际复兴开发银行/世界银行的观点。其附属组织，或世界银行执行董事或他们所代表的政府的组织。

# 从纵向研究中提高调查估计的质量：在 LSMS 面板调查中的应用

Piero Falorsi <sup>1</sup> , Paolo Righi <sup>2</sup> , Giulia Ponzini <sup>3</sup>

JEL 代码: C81, C83, C87, Y80 关键词: 面板调查, 减员, 跟踪, 广义权重份额法, 调查方法

---

<sup>1</sup> 世界银行发展经济学数据集团国际顾问, <sup>1,2</sup> 意大利统计研究所 (ISTAT), <sup>3</sup> 世界银行发展经济学数据组; 本文得到了“到2030年关闭农业数据缺口倡议”(50x2030倡议)的财政支持, 该倡议是一个多边项目, 旨在到2030年通过改造非洲、亚洲、中东和拉丁美洲50个国家的数据系统来弥合全球农业数据差距。如需了解更多信息, 请访问[50x2030.org](http://50x2030.org)。

## 1. INTRODUCTION

纵向调查，基于随时间对同一统计单位进行重复观测，是分析人类群体当前状态及其随时间变化的宝贵资源。以“生活标准测量研究——农业综合调查 (LSMS-ISA)”项目为例，通过传播涉及健康、农业、基本服务获取、营养、贫困状况等议题的家庭面板数据，为撒哈拉以南非洲国家展示了纵向调查的价值。欧洲联盟的收入与生活条件统计 (EU SILC) 和美国统计局的收入与计划参与调查 (SIPP) 也是纵向研究价值的其他关键实例。

在面板调查中，准确性取决于与横截面调查相同的许多因素。然而，随着面板调查时间长度的增加，保持估计值的准确性变得更加困难。一个关键问题是受访者中途退出调查（即面板损耗），这在长期运行的面板调查中可能会变得越来越棘手。此外，移居者——即那些在纵向研究的第一个观察点被访问后搬家的人——代表了一个动态子群体，且难以且成本高昂进行访谈。最后，新加入人口的成员，如移民和新生儿，在传统的面板调查中尤其难以捕捉。

正确的面板调查设计、实施和使用考虑了在统计过程的不同阶段（样本设计、数据收集和估计阶段）解决这些问题的各种方法。尽可能地，样本设计应通过使用刷新样本（例如，旋转面板或分段面板）和纳入之前未参与研究的个体来最小化由于流失、离开者和新加入者导致的数据代表性损失。基于跟踪规则以追踪迁移者，数据收集阶段应设计成能够获取那些已退出样本的个体或家庭的关键信息。采用灵活的数据收集方式（如电话访问）可以接触那些可能不响应传统技术的人群。此外，收集退出者的最少变量数据（例如，通过代理人或上门访问）可以提高调查估计的质量。在估计阶段，需要考虑退出者、迁移者、新加入者以及目标人群随时间的变化动态。这可以通过定义加权程序来实现，该程序更新直接抽样权重以反映面板家庭中新加入的个体，并使用与最新已知人口总数相校准的估计器。

。

这项研究提议通过考虑上述三种调查设计方面来提高面板调查的质量。我们考虑当前时间估计目标人群的横截面参数和纵向参数。这些估计值是从当前时间收集的样本数据中计算得出的，样本数据来自于之前进入样本并在后续根据调查中所考虑的纵向观察规则进行跟踪的个体。这种方法澄清了许多关于面板研究中收集的数据代表性的方面。它适用于当前大规模家庭面板调查中的许多案例（例如，欧盟 SILC、LSMS 等），并且可以轻松扩展到实际调查情况中的更复杂问题。

该论文结构如下：第二部分给出了纵向和横截面人口的正式定义以及纵向研究中需要估计的目标参数。第三部分介绍了基于多源 (Singh 和 Mecatti 2011) 和间接抽样 (Lavallée 2007; Falorsi, Righi 和 Lavallée 2019) 的抽样框架和估计方法。

各种抽样子群体代表性差异以及家庭构成随时间的变化。第4节介绍了场操改进和数据收集的  
提升，以促进提议方法论的实施。第5节简要介绍了生活标准衡量研究 (LSMS)，定义  
了其目标和估计质量的关键方面。第6节通过将提议方法论应用于LSMS-ISA数据，并以乌  
干达国家面板调查2009、2013和2015年的波次作为案例研究，展示了该方法论的实证应用  
。第7节总结了主要发现并得出结论。

## 2. 横截面和纵向种群的正式定义

### 2.1. 在不同时间点观察到的人口

让我们成为当前特定国家的个人人口  $t$ 。被划分为家庭的亚群，表示为  $\{h_1, \dots, h_{H(t)}\}$ 。

我们将家庭集合表示为  $\Omega = \{h \mid h \text{ 是一个家庭}\}$

家庭  $h$  有  $n_h$  个人，表示为  $\{i_1, \dots, i_{n_h}\}$ 。我们将个人集合表示为  $\Omega_p = \{i \mid i \text{ 是一个个人}\}$ 。我们将个人  $i$  在时间  $t$  的集合表示为  $\Omega_{i,t} = \{t \mid i \text{ 在时间 } t \text{ 存在}\}$ 。我们将个人  $i$  在时间  $t$  的集合表示为  $\Omega_{i,t} = \{t \mid i \text{ 在时间 } t \text{ 存在}\}$ 。

$\sum_h \mathbb{1}_{i \in \Omega_h} = 1$  可以参考  $\Omega$  and as **横截面人口** 个人和家庭，因为他们指的是一个特定的时间点。

然而，在时间上的动态变化，以当前时间  $t$  从最初的纵向设置中的截面种群，我们引入了 **纵向种群**。为了简化，\* 我们首先介绍一个没有按家庭划分的人口。因此，表示的纵向人口  $\Omega_{i,t}$  -

let  $\Omega_{i,t}$  个人。人口

从研究范围和纵向观察的角度来看，识别是复杂的

$\Omega_{i,t} \leftarrow \Omega_{i,t} \cap \Omega_{i,t+1}$  \* operability over time (Helliot et al. 2009). The most common definitions are the  $\Omega_{i,t}$   $\Omega_{i,t} \leftarrow \Omega_{i,t} \cap \Omega_{i,t+1}$  交叉点 - 人口，=, and the  $\Omega_{i,t}$  - 人口 =.

十字路口  $\Omega_{i,t} \leftarrow \Omega_{i,t} \cap \Omega_{i,t+1}$  - 人口是两者的适当子集  $\Omega_{i,t} \cap \Omega_{i,t+1}$   $\Omega_{i,t} \leftarrow \Omega_{i,t} \cap \Omega_{i,t+1}$  and  $\Omega_{i,t} \cap \Omega_{i,t+1}$  包括个人  $\Omega_{i,t} \cap \Omega_{i,t+1}$  of 当时仍居住和居住在该国的人  $t$

。纵向测量

\* 随着时间的推移观察同一个人，不包括出生和死亡  $\Omega_{i,t} \cap \Omega_{i,t+1}$  期间

\* study.  $\Omega_{i,t}$  然而，基于交集  $(\cup_t \Omega_{i,t})$ -人口的纵向人口定义在时间上的代表性存在一定缺点。 $t$  并且针对家庭构成动态变化研究的局限性。虽然联合人口能够最好地描述人口动态，但它需要一个动态样本设计来引入出生和新进入者以避免偏差。对于本研究，我们提出一种定义，结合了两种方法的优点。

个人，包括：基于此，

-  $\Omega_{i,t} \leftarrow \Omega_{i,t} \cap \Omega_{i,t+1}$  \* 十字路口  $\Omega_{i,t} \leftarrow \Omega_{i,t} \cap \Omega_{i,t+1}$  - population  $\Omega_{i,t} \leftarrow \Omega_{i,t} \cap \Omega_{i,t+1}$  \* 所有的人  $\Omega_{i,t}$  家庭

所有新成员的家庭的个人在十字路口  $\Omega_{i,t} \cap \Omega_{i,t+1}$ ；

$\Omega_{i,t}$  人口，即使他们不是其中的一部分

根据这个定义，新生儿  $\Omega_{i,t}$  并且那些移民成为交集人口中个体家庭的新成员时，是对人口的新一轮进入。如果该国的移民数量和新生儿数量在数值上不具有显著性。\*，那么纵向人口近似于时间的横截面人口。此外，纵向观察在概念上

$\Omega_{i,t}$  \* 简单。我们在初始时间选择一个样本，然后在下面的调查中我们观察到

\*  $\Omega_{i,t}$

$\Omega_{i,t} \leftarrow \Omega_{i,t} \cap \Omega_{i,t+1}$

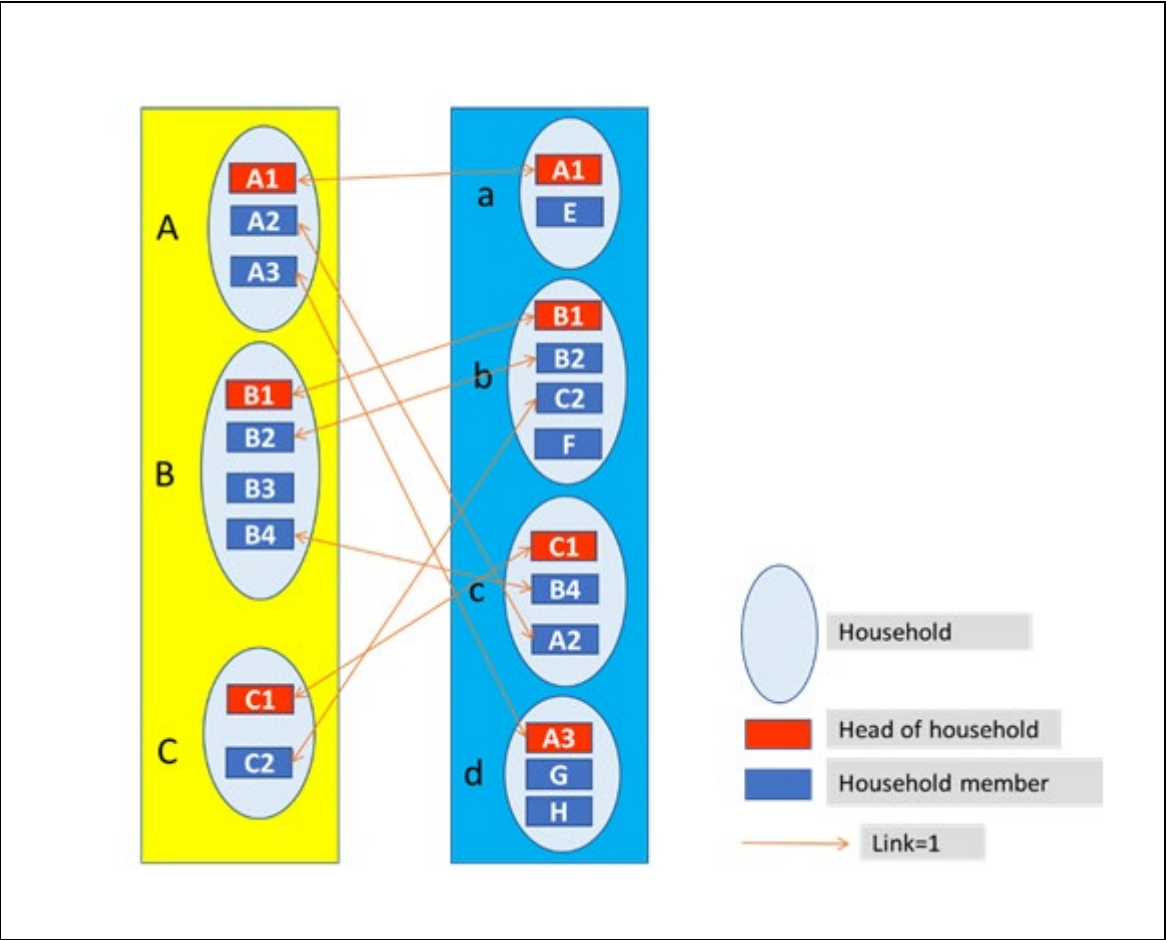
\* 所有家庭成员的初始个体样本。  $\Omega_{i,t}$

$\Omega_{i,t} \leftarrow \Omega_{i,t} \cap \Omega_{i,t+1}$

给定个体纵向人口的定义，我们引入了家庭纵向人口的定义，记作  
随着时间的推移，家庭可能会经历三种类型的变化：消失，融合或分裂。

这些变化直接影响横截面和纵向分析和影响样本代表性 (FAO, 2015)。以图2.1中时间1时的三个家庭A、B和C为例进行说明。在时间2时，家庭A解散，其成员分别迁移到了三个不同的家庭中：A2加入家庭C，A3组建了一个新的家庭D，而A1则与新成员E一同加入了家庭A。此外，在时间2时，家庭B中的成员B3已经消失。

Figure 2.1. Longitudinal housing over time (yellow refers to, blue refers to)  $t$



上述图片展示了纵向家庭定义随时间变化所面临的问题。我们可能采用两种广泛的方法。传统上定义纵向家庭的方式是将其一对一地映射到时间上：一个家庭从时间点生成仅对应一个家庭在时间点。这种做法的主要缺点是，我们可能会面临排除某些家庭的风险，从而导致定义上的不完整或不准确。  
纵向分析一些家庭中的个人 live.

在这一点上，介绍的概念是有帮助的连续性规则。要理解连续性规则的概念，请考虑在时间  $t$  和两个孩子观察到的。两年后，跟随同一个人，测量员面对两个  $t+1$  家庭  $t$  由一个父亲和母亲组成的家庭，其中父亲与母亲分居，父亲独自生活在另一个地方与其中一个孩子一起生活。连续性规则明确规定了哪一户在两年后的观察家庭是最初观察家庭的延续，哪一户不是。

新的家庭单元。因此，连续性被用来定义操作规则，以确定当前时间哪个家庭是之前观察到的家庭的延续。识别纵向家庭的连续性规则可能有所不同 (Falorsi等, 2009)。如图2.1所示，如果连续性规则将具有相同户主的家庭视为纵向家庭，我们将排除家庭  $\setminus(d_t)$  从分析中。\*

为了克服这一问题，我们提出了一种多对多的方法：一个时间点的 households 可能会产生多个时间点的 households；相反，一个时间点的 households 也可能源自另一个时间点的 households。\*

来自时间的几个家庭。人口的家庭 那些包括

。在图 2.1 中，我们看到家庭 A \*

$\square\square\square\square$ ， and d.  $\square\square\square\square \leftarrow \leftarrow * \square\square\square\square$

正式地，许多继续存在的时间点，并且不失一般性，家庭  $\setminus(\mathcal{H}_i)$  是来自时间  $\setminus(t)$  家庭  $\setminus(\mathcal{H})$  的纵向家庭之一，如果  $\setminus(k \in \mathcal{U})$ 。  $\square\square\square\square$

家庭 a、b  $\square\square\square\square \square\square\square\square$

(2.1)  $\square\square\square\square \square\square\square\square \square\square\square\square = \square\square\square\square = t \leftarrow t^* > 0$ ,

}

Where  $\diamond$ ，  $\square\square\square\square = 1$   $\diamond$   $\square\square\square\square \text{ link} = 1$   $\square\square\square\square \square\square\square\square \{ \square\square\square\square 0, 1 \}$   $\square\square\square\square$  变量，如果个体，则取值为 1  $\square\square\square\square \ell$  的家庭在

\* 时间是同一个人  $\square\square\square\square \leftarrow \square\square\square\square$  对于家庭而言，在这一方法下，人口的纵向统计与家庭的统计之间存在完美的对应关系。对于时间点，通过这种方法，可以确保个人层面和家庭层面的数据统计在时间维度上完全一致。\*  $\square\square\square\square \square\square\square\square\square\square\square\square\square\square\square$  每个人的

，我们可以性最初的时候定义他们的家庭，  $\ell$   $U$   $t^*$

$\square\square\square\square \square\square\square\square \square\square\square\square$ ，并且在当前

时间。定义 2.1 包括作为特定情况的一对一连续性规则。在此

## 2.2 感兴趣的参数

### 2.2.1. 横截面参数

case,  $\square\square\square\square$ ， is a 0  $\square\square\square\square$ , 1  $\leftarrow$  二分变量，如果是家庭，则等于 1 是唯一的一个令  $\setminus(x)$  和  $\setminus(y)$  分别为在个体和家庭中观测到的两个定量变量。感兴趣的参数是总和  $\setminus(y \text{ among } \setminus(x))$ 。  $\square\square\square\square$  指的是人口这是家庭的纵向延续。  $t$   $U$   $t$

(2.2) 
$$= \frac{M_t}{N_{t,k}} = \frac{M_t}{N_{t,k}}$$

其中  $\square\square\square\square \diamond$ ， 是的值  $\square\square\square\square = 1$   $\diamond$   $\square\square\square\square = 1$   $\square\square\square\square \square\square$   $y$   $\square\square \square\square\square\square \square\square\square\square$  的个人  $\diamond$   $\square\square\square\square = 1$   $\square\square\square\square \square\square\square\square \square\square\square\square$  家

(2.3) 
$$= \frac{N_{t,k}}{N_{t,k}}$$

是的总数  $\square\square\square\square \square\square\square\square\square\square$

$\square\square \diamond$   $\square\square\square\square$  为家庭  $= 1$   $\square\square$

类别  $\square\square$  其中  $\square\square\square\square\square\square\square\square$ 。

现在我们假设家庭变量的 if person 是一个分类变量，属于 p- th 类 (being = 1, ..., )

。例如，如果

变量描述了三个类别的就业状况 (就业、失业、非就业  $\square\square\square\square \square\square\square\square\square\square\square\square$

{ }  $y$   $p$  在劳动力中)。让  $k$

= 1, if, 表示一个二分变量，用于  $\square\square\square\square$

$\square\square\square\square$  否则.)，等于 0 1 个家庭的个人  $\square\square\square\square \square\square\square\square$

感兴趣的参数是  $\square\square\square\square \square\square\square\square\square\square\square\square$

:  $\square\square\square\square \square\square\square\square$

$\square\square\square\square \diamond$  不同的频率总计 P 变量的类别

$$(2.4) \quad p = \frac{M_t}{N_{t,k}} \quad p = \frac{M_t}{p} \quad ( = 1, \dots, ),$$

在哪里？ $\square\square\square\square$ ，表示变量的总数  $\diamond \square\square\square\square=1 \diamond \square\square\square\square=1 \square\square\square\square \square\square\square\square\square\square\square\square$   
 $\square \diamond \square\square\square\square=1 \square\square\square\square \square\square\square\square \square\square\square\square$ ，即 家庭中的个人数量  $\square\square\square\square k$  以变量类别  
 为特征  $\square\square\square\square$ 。  
 最后，我们可能对通常在家庭层面定义、描述特定家庭状况（例如，贫困或非特定贫困、  
 $\square$  营养不良的状况等）的  $\square$  量感  $\square$  趣。我  $\square$  用  $\square\square \square\square\square\square$  的分  $\square\square$  量表示  $\square$  一情况。  $\cdot =$   
 $(= 1, \dots, )$  如果家庭属于  $p$ - 第一类，我们表示为  $(= 1, \dots, )$  二分  $(0, 1)$  变量的值  $\square$   
 $\square\square\square \square\square\square\square \square\square\square\square \square\square\square\square \square\square\square\square \square\square\square\square \square\square\square\square \square\square\square\square \in \square\square\square\square \square\square\square$   
 $\square \square\square\square$

$$1, = \text{if } \square\square\square\square \square\square\square\square$$

在这种情况下  $\square\square\square\square \square\square\square\square$ ，参数  $\square\square\square\square$   
 $\square\square\square\square \diamond$  th 条件。

2.4 统计中的家庭数量  $p$ -

### 2.2. 2. 个体的纵向参数

净变化参数由

$$(2.5) \quad =$$

参数  $\Delta \square\square\square\square \square\square\square\square \leftarrow 2 * .5$  表示人口交叉构成之间的差异。  $\square\square\square\square - \square\square\square\square \square\square\square$   
 $\square \square\square\square\square$   
 $\square\square\square\square \square\square\square\square * -$  当时的截面总数  $t$  与  $\square$ 。净变化反映了特征和  
 相应的总时间

\*使用横截面汇总估计对净变化进行分析可能会掩盖重要状态

总变动 (Steel 等人，2009 年)。

个体层面的变化，从现在起称为“总变化”，指的是在同一时间序列中以相同的单位测量该现象的变化。  
 大规模的变化可能通过纵向调查得以揭示，即通过对同一人群进行多次访谈来追踪变化。  
 这样我们不仅可以估计每年的趋势，还可以评估持续性 (Lohor 2009)。

当时处于变量类别和类别中的个人的总变化  $p$  相同变量的 (与  $\cdot = 1, \dots *$ ) 在当前时  
 间  $t$  可以表示为：

$$(2.6) \quad p_{t \leftarrow t^*}^{p,g} = \sum_{\ell=1}^{M_{t'}} \sum_{j=1}^{N_{t^*,\ell}} \sum_{k=1}^{M_t} \sum_{i=1}^{N_{t,k}} l_{ki,\ell j}^{t \leftarrow t^*} y_{t,ki}^p y_{t^*,\ell j}^g = \sum_{k=1}^{M_t} \sum_{i=1}^{N_{t,k}} \tilde{y}_{t \leftarrow t^*,ki}^{p,g},$$

在哪里

$$(2.7) \quad p_{*,g}^{p,g} = \frac{M_{t'}}{N_{t^*,\ell}} \quad p_{*,g}^{p,g}$$

是等于 1 的流量变量，如果个人  $\square\square\square\square \square \square\square\square \leftarrow \square\square\square\square \square\square\square\square\square\square\square\square \diamond \ell=1 \diamond \square\square$   
 $\square\square=1 \square\square\square\square \square\square\square\square\square\square\square\square\square\square\square \square\square\square\square \square\square\square\square \square\square\square\square\square\square\square\square \square\square\square\square \square\square\square\square \square\square\square\square$   
 $\square$  更改为 `in time interval()`，否则为 0。因此，总变化可以表示为横截面总数 (在时间  $t^*$ ) 的流量变量

交叉路口人口， $\square\square\square\square\square\square\square\square$ 。此变量只能直接测量  $\square\square\square\square\square\square\square\square$  的单位。  $-$   
 $\square\square\square\square \leftarrow \square\square\square\square * \square\square\square\square \square\square\square\square$

### 2.2.3. 家庭随时间的变化 $\square\square\square\square \cap \square$

$\square\square\square$

The 净变化 的家庭可以表示为 2.5，但总数和参考上面介绍的变量的事实除外。  $\square\square\square$

$$\square \square\square\square y_{t^*}^p$$

$p$

The **总变化** 衡量条件变化的数量，即多少户家庭从不利状态转变为有利状态（例如，从贫困变为非贫困、从营养不良变为非营养不良），从有利状态转变为不利状态（例如，从非贫困变为贫困、从非营养不良变为营养不良），或者保持在相同的有利或不利状态。对于个体而言，总体变化体现在分类变量上。

不同类别的家庭总变化  $p$  变量的(, = 1 , ... , ) 可以表示为

$$g \quad p \quad P$$

$$(2.8) \quad p, g_{t \leftarrow t^*} = \sum_{k=1}^{M_t} \sum_{\ell=1}^{M_{t^*}} \frac{L_{k,\ell}^{t \leftarrow t^*}}{L_k^{t \leftarrow t^*}} Y_{t^*,\ell}^g Y_{t,k}^p$$

其中

$$(2.9) \quad t \leftarrow t^* = \frac{M_{t^*} \quad N_{t^*,\ell} \quad N_{t,k}}{t \leftarrow t^*}$$

表示的数量 of household. This quantity represents the number of individuals of household  $k$  属于。它对纵向参数的定义都起着核心作用 总变化 并用于产生无偏估计

感兴趣的参数。为了阐明如何在以下架构中计算此数字 2.1，我们为四个家庭中的每个家庭提供 的价值 (  $a$  ,  $b$  ,  $c$ , and  $d$  ) 在图中给出的时间 2

2.1. 架构 2.1。的值 对于图 2.1 中时间 2 的家庭

| * 家庭一 | $L_k^{t \leftarrow t^*}$ |               |
|-------|--------------------------|---------------|
|       |                          | One to - many |
| A     | 1                        | 1             |
| B     | 1                        | 3             |
| C     | 1                        | 3             |
| d     | 0                        | 1             |

If = , 然后家庭  $k$  是当时唯一的家庭  $t$

源自原始

对公式的重新思考 \*\*\*\*\*  
 \*\*\*\*\*  
 家庭 的。 \*\*\*\*\*  
 (2.10) \*\*\*\*\*  
 注意到它可以表示为 \*\*\*\*\*  
 在哪里

$$(2.11) \quad p, g_{t^*} = \frac{M_{t^*} \quad \frac{t \leftarrow t^*}{k, \ell}}{*} \quad \frac{g}{*} \quad \frac{p}{*},$$

是 flow

变量  $\ell=1$  指示时间间隔的变化 ( ) of 家庭  $k$  (按时间标识) 从变量的类别到类别  
 在\*对多连续性规则的情况下 ,  
 间隔。 , 是在 - 中定义的实值变量 [0, 1

在一对一连续性规则的情况下 , , 是 0 ; 1 二分变量。

下 Schema 2.2 示了家庭毛 的 算以及在 中描 的家庭关系。 根据 图 2.1.





□ F+F 是固定不□的家庭子群体，或者根据□□采用的跟踪□□，在适当的□候可以□行跟踪。

\* \*

例如，对于 UNPS，有资格跟踪的家庭是那些至少 □□□□←□□□□ \* 在最初的 □□□□ 中，一个人已经出现在那个时候 □□□□ 和生活 □□□□ 直到时间 □□□□。不包括所有移民的家庭，其中只有新人 \* 在时间流逝之后进入该国，而家庭仅由以下人员组成

□□□□ □□□□ \* 移动后的时间 □□□□ ← 离开他们的教区。 \* □□□□ □□□□ \* 家庭 其成员仅由 移民 or 搬运工 在时间间隔 [ , ] 中。 \* \*\*

\*\*\* 包括 家庭

✓

时间间隔 [ , ]。

\*\* 其成员仅由 移民 or 搬运工 in

分割部分 3.1 和图 3.1。定义 代表性 在两个纵向样本中 (

and 关于子人群。herein 我们称代表性为从样本计算给定人群的直接无偏估计的可能性。

□□□□←□□□□ \*\* □□□□ \* data. □□□□← □□□□ □□□□

For the **Sample**，跟踪规则限制了样本的代表性 只对

亚种群 □□□□←□□□□ \* □□□□ □□□□←□□□□ \*

有资格追踪的时间家庭 □□□□←□□□□ \* 不时 t .

For **Sample**，跟踪规则仅将样本数据的代表性限制为子总体

□□□□←□□□□ \*\* □□□□ □□□□ \* □□□□

有资格追踪的时间家庭 □□□□←□□□□ \*\* 不时。

t

$F_{t \leftarrow t^{**}}$  可以分为两个亚群 和  $F$ ，存在 \*\*  $t \leftarrow t^{*}$   $\mathcal{E}_{t^{**} \leftarrow t^{*}}$  = .

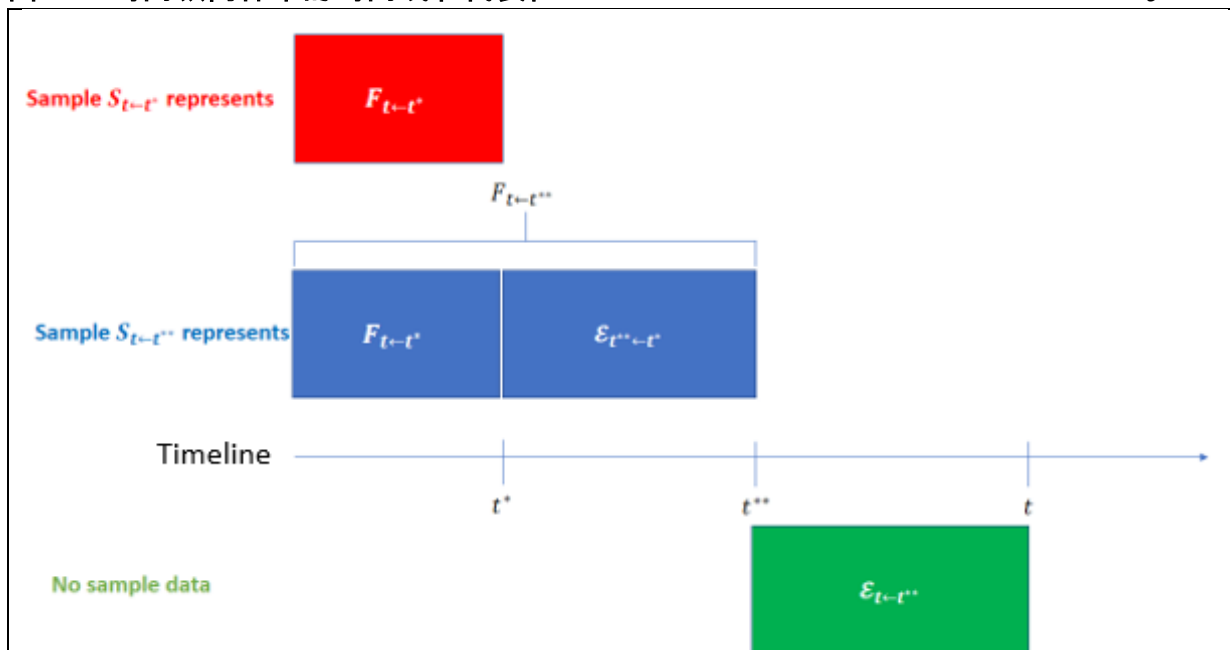
，我们没有属于子种群的人

最后，我们看到在 样本 □□□□←□□□□ \*\* □□□□ □□□□←□□□□ \*  $\cup \mathcal{C}$  □□□□ \*\* ←□□□□ \*，，谁是上次面板刷新后进入人口的人。由于我们没有样本

□□□□←□□□□ \*\*  $\mathcal{C}$  □□□□ \*\* □□□□ \*

关于这个亚群的数据，我们不能计算直接估计。 ←

图 3.1. 时间纵向样本的时间线和代表性



对于的每个家庭，我们可以定义两个二分变量

$$(3.2) \quad F_{t \leftarrow t^*} = \begin{cases} 1 & \text{if } F_{t \leftarrow t^*} = 1 \\ 0 & \text{otherwise} \end{cases}, \quad \varepsilon_{t^{**} \leftarrow t^*} = \begin{cases} 1 & \text{if } \varepsilon_{t^{**} \leftarrow t^*} = 1 \\ 0 & \text{otherwise} \end{cases}.$$

给定于 (0, 1] 否则 0, 否则 0. (3.2) 可以表示为三个附录的总和，每个附录都是 的总和 与三个亚群之一有关， 和  $F_{t \leftarrow t^*}$ ,  $\varepsilon_{t^{**} \leftarrow t^*}$ , and  $\varepsilon_{t \leftarrow t^{**}}$ .

$$(3.3) \quad Y_{t \leftarrow t^*} = F_{t \leftarrow t^*} + Y_{\varepsilon_{t^{**} \leftarrow t^*}} + Y_{\varepsilon_{t \leftarrow t^{**}}}$$

being

$$(3.4a) \quad Y_{F_{t \leftarrow t^*}} = \sum_{k=1}^{M_t} \sum_{i=1}^{N_{t,k}} y_{t,ki} F_{k,t \leftarrow t^*} = \sum_{k=1}^{M_t} Y_{t,k} F_{k,t \leftarrow t^*},$$

$$(3.4b) \quad Y_{\varepsilon_{t^{**} \leftarrow t^*}} = \sum_{k=1}^{M_t} \sum_{i=1}^{N_{t,k}} y_{t,ki} \varepsilon_{k,t^{**} \leftarrow t^*} = \sum_{k=1}^{M_t} Y_{t,k} \varepsilon_{k,t^{**} \leftarrow t^*},$$

$$(3.4c) \quad Y_{\varepsilon_{t \leftarrow t^{**}}} = \sum_{k=1}^{M_t} \sum_{i=1}^{N_{t,k}} y_{t,ki} [(1 - F_{k,t \leftarrow t^*})(1 - \varepsilon_{k,t^{**} \leftarrow t^*})] \\ = \sum_{k=1}^{M_t} Y_{t,k} (1 - F_{k,t \leftarrow t^*})(1 - \varepsilon_{k,t^{**} \leftarrow t^*}).$$

样本的代表性允许仅计算直接无偏估计  $\hat{Y}_{t \leftarrow t^*} = \frac{1}{M_t} \sum_{k=1}^{M_t} Y_{t,k} F_{k,t \leftarrow t^*}$   $\hat{Y}_{\varepsilon_{t^{**} \leftarrow t^*}} = \frac{1}{M_t} \sum_{k=1}^{M_t} Y_{t,k} \varepsilon_{k,t^{**} \leftarrow t^*}$   $\hat{Y}_{\varepsilon_{t \leftarrow t^{**}}} = \frac{1}{M_t} \sum_{k=1}^{M_t} Y_{t,k} (1 - F_{k,t \leftarrow t^*})(1 - \varepsilon_{k,t^{**} \leftarrow t^*})$

The aggregate:  $\hat{Y}_t = \hat{Y}_{t \leftarrow t^*} + \hat{Y}_{\varepsilon_{t^{**} \leftarrow t^*}} + \hat{Y}_{\varepsilon_{t \leftarrow t^{**}}}$

$$(3.5) \quad \hat{Y}_t = \hat{Y}_{t \leftarrow t^*} + \hat{Y}_{\varepsilon_{t^{**} \leftarrow t^*}} + \hat{Y}_{\varepsilon_{t \leftarrow t^{**}}}.$$

代表总数的准确(代理)度量  $\hat{Y}_t$  的准确(代理)度量  $\hat{Y}_t$  只有当聚合 is 可忽略。如果时间点 接近当前时间，则此条件成立。

$\hat{Y}_t$  的准确(代理)度量  $\hat{Y}_t$  只有当聚合 is 可忽略。如果时间点 接近当前时间，则此条件成立。

### 3.3. Estimation process

#### 3.3. 1. Direct estimator

如式(3.5) 所示，对总数的直接估计由下式给出：

$$(3.6) \quad = S_{t \leftarrow t^*} + (1 - \epsilon_{t \leftarrow t^*}) S_{t \leftarrow t^{**}} + \epsilon_{t \leftarrow t^*} S_{t \leftarrow t^{**}},$$

其中  $\epsilon_{t \leftarrow t^*}$  是 GWSM 估计器 (Lavallee 2007) 的估计器，用纵向样本收集的数据获得， $\epsilon_{t \leftarrow t^*}$  和  $S_{t \leftarrow t^{**}}$  是 GWSM 对总数的估计 and 使用的数据计算

$$(3.7) \quad \epsilon_{t \leftarrow t^*} = \frac{1}{N_{t \leftarrow t^*}} \sum_{k=1}^{N_{t \leftarrow t^*}} \epsilon_{t \leftarrow t^*, k}$$

$$(3.8) \quad \epsilon_{t \leftarrow t^*, k} = \frac{1}{N_{t \leftarrow t^*}} \sum_{i=1}^{N_{t \leftarrow t^*}} \epsilon_{t \leftarrow t^*, k, i}$$

在下面的公式(3.9)中提供  $\epsilon_{t \leftarrow t^*, k}$  多少家庭成员  $\epsilon_{t \leftarrow t^*, k}$  在调查期间，收集有关

$$(3.9) \quad \epsilon_{t \leftarrow t^*, k} = \sum_{i=1}^{N_{t \leftarrow t^*}} \sum_{j=1}^{N_{t \leftarrow t^*, k}} \epsilon_{t \leftarrow t^*, k, i, j}$$

$$(3.10) \quad S_{t \leftarrow t^*} = \frac{1}{N_{t \leftarrow t^*}} \sum_{k=1}^{N_{t \leftarrow t^*}} S_{t \leftarrow t^*, k}$$

$$(3.11) \quad \epsilon_{t \leftarrow t^*, k} = \frac{1}{N_{t \leftarrow t^*}} \sum_{i=1}^{N_{t \leftarrow t^*}} \epsilon_{t \leftarrow t^*, k, i}$$

对于样本中收集的家庭在调查期间，收集器可以收集的值，可以等于 0 或 1。

如果所有家庭成员对问题 1 或问题 2 给出肯定的答案，则等于 1。

Let  $\epsilon_{t \leftarrow t^*}$



重量  $w_i$ ，用于估计  $\hat{Y}$ ，因为这些尽可能接近  $Y$ ，直接赋予权重，并允许重新生成某些辅助变量已知的总和。总量的校准估计量  $\hat{Y}$  定义为：

$$(3.16) \quad \hat{Y} = \frac{\sum_{i \in S} w_i y_i}{\sum_{i \in S} w_i} \quad \text{其中 } w_i = \frac{N_{t,k}}{N_{t,k} - (y_i - \bar{y}_k)}$$

或者如果变量是指  $y_i$

As Särndal and Lundström (2005, 第88页) 指出，就像 (3.16) 对家庭来说，就像 (3.17) 对个体来说，

As Särndal and Lundström (2005, 第88页) 指出，就像 (3.16) 对家庭来说，就像 (3.17) 对个体来说，

关于非响应的丰富信息，Särndal 和 Lundström (2005, 第88页) 指出，校准方法可以轻松利用响应数据；如果辅助变量能够解释非响应情况，校准估计量可以减少非响应偏差。通过这种方式，估计值可以共同平滑非响应偏差并校准已知的人口总量。这种方法对于通常用于生成调查估计的方法具有三个优势，该方法基于对非响应倾向得分进行初步调整，并生成个体级别的权重，这意味着同一户家庭中的人可能具有不同的抽样权重。

The first 优势在于，抽样方差估计更为坚实且透明。校准估计器表现出渐进性质，即它们以概率收敛至回归估计器。它们为抽样设计和统计模型均生成稳健的推断，具有良好的统计特性。

The second 好处是所需的调整步骤较少，从而使估计流程更为简便。Särndal 和 Lundström (2005, 第88页) 指出，如果非响应信息仅在样本级别可用，校准方法可以在两步中完成，第一步在校选样本级别对已知总量进行校准，这包括受访者和非受访者。两步校准方法在退样原因与抽样覆盖率不足的原因不同时，可以提高估计的准确性。因此，实施了两次校准：第一次校准调整了面板退样的基础权重，第二次校准使用调整后的权重来生成最终权重以解决覆盖率不足的问题。

最后，第三优势在于，校准方法产生的估计值是在 Household (户) 层面定义的。这意味着同一个家庭中的所有个体具有相同的最终权重，这显著地促进了估计值与个体层面定义的变量以及基于家庭层面变量构建的变量之间的一致性。

为了介绍这个估计器，让  $X_1, \dots, X_k$  是行政数据中人口的辅助已知总数的列向量，最新的人口普查，或人口统计。让  $X_1, \dots, X_k$  是辅助的向量

$k$  这样的  $X_k$  的变量

家庭  $h$

家庭  $h$

Let  $X_{hk}$  是为  $X_k$  家庭  $h$  的个体  $i$  的变量

。关于辅助变量 (在家庭层面定义) 之间的关系，可能存在两种情况。假设变量  $X_k$  在家庭层面定义 (例如，家庭的贫困状况)。

假设变量  $X_k$  在个体层面定义 (例如，个人的教育水平)。

是  $X_{i..} = X_{i...}$ ，在这种情况下，变量  $X_{i..}$  作为  $X$  值的平均值。另一方面，如果变量  $X$  是一个计数变量（例如，特定年龄组的女性），然后  $X_{i..}$  它可以作为  $X$  变量在  $i$  家庭中的第一个人有给定的个人水平在哪里  $X_{i..}$ 。再次考虑上述示例， $X = 1$  如果人是女性， $X = 0$  否则。如果我们假设向量  $X_{i..}$  对于整个人口来说是已知的，并且对于样本家庭来说，向量是已知的。已校准的权重  $w_i$  得到以下最小约束问题的解

$$(3.17) \quad \min \sum_{i \in \mathcal{I}} w_i D(d_i)$$

其中： $D(d_i) = \sum_{j \in \mathcal{J}} (d_{ij} - d_j)^2$  是截断距离函数 (Singh 和 Mohl 1996) 问题 (3.17) 可以用开源软件 Regenesees 解决，以及其他解决方案。这允许  $w_i$  带权重  $w_i$  作为 (3.17) 的解获得，具有以下积极特征：截断的线性和截断的逻辑系统 (Gardero 2015) 定义权重  $w_i$  水平，我们可以用它们来估计个人和家庭的参数。因此，它们确保了家庭和个人估计的一致性 (The estimates of the auxiliary variables are accurate)。在国家一级定义。因此，权重，以已知总数为基准 (或由大型和  $X_{i..}$ )，保证样本估计值与国家层面的统计系统的一致性。使用截断距离函数意味着  $\mathbf{X}$  重量  $w_i$  总是积极的，其中  $w_i$  权重的异常值对最终估计的影响有限。

## 总 2.5 的校准估计器由

校准后的毛额估计值  $\hat{Y}_{i..}$  对个人的改变，如公式 2.6 所定义，可以简单地获得流量变量的加权和  $\hat{Y}_{i..}$  在表达式中定义家庭总变化的校准估计，如公式 2.10 所定义的，可以简单地获得为流量变量表达式 2.11 的加权和。定义于  $\hat{Y}_{i..} = \sum_{j \in \mathcal{J}} w_{ij} Y_{ij}$

<sup>2</sup> <https://www.istat.it/en/methods-and-tools/methods-and-it-tools/process/processsing-tools/regenesees>, 2019 年 9 月访问。

## 4. 数据收集

### 4.1. Tracking rules

统计办公室定义跟踪规则以应对迁徙者造成的潜在偏差和样本大小减少。这些规则的定义必须考虑提高估计准确性和在全国范围内追踪迁徙者的成本之间的权衡。通常，这种权衡会导致一个妥协方案，即只在有限区域内追踪迁徙者。在这种情况下，当个体在首次观察区域内的移动时被追踪。

要进行跟踪，调查需要根据以下条件定义跟踪协议：

- The **跟踪目标** 为了具体描述这一过程，让我们回顾第4.1节中已经介绍过的三次浪潮、和。在当前时间段的研究工作开始时，\*\*\*，我们有最初在 中选择的个人和家庭样本，或者这些个人纳入 □□□□ □□□□ □□□□ d 在连续浪潮中的家庭中。用于考试 \* pl e，作为跟踪规则，我们可以指定 \*\* □□□□ 调查计划应用跟踪协议仅适用于在 □□□□ 初始样品和。
- The □□□□ **该地区的划界** 在其中进行跟踪(即地区，教区等) □□□□ \* □□□□ □□ □□ \*\*，**枚举区**，。
- The **问卷定义**。例如，在某些个体已经决定搬迁的情况下，收集有助于下次调查重新联系这些人的信息可能会很有用(例如，受访者是否计划在面板的下一波次搬迁，搬迁的原因，以及潜在搬迁者的电话号码)。
- 的协议 **代理面试** 用于非响应者通过挨家挨户访谈收集的最小变量集，以及用于迁移者通过代理访谈收集的最小变量集。
- 的定义 **搬运工的数据收集程序**，这考虑了访谈的区域(例如，如果搬迁者留在样本区域内，则数据收集可以通过面对面访谈进行；如果搬迁者移至样本区域外，则可通过电话访谈进行)。

### 4.2. 用于计算基本权重的链接和其他辅助变量

我们称基权重为根据通用加权份额方法(Lavallée, 2007)计算的权重除以重复因子(第3.3.1节)。这些权重在家庭层面定义，每个个体拥有所属家庭的基权重。

如3.3.1节所述，GWSM(加权渐进样本均值)需要以下两个变量：(i) 将个体与其所属的初始样本链接起来的变量(在之前的例子中我们在  $\backslash(t_t\backslash)$  和  $\backslash(t_t\backslash)$  有两个原始样本)，以及(ii) 将个体与目标人群链接起来的变量，即他们的年份。\* 首次出现在目标人群中(对于新生儿和移民)。\*\*

计算基础权重的过程包括倍数因子，该因子表示家庭在初始样本中被选中的机会数量。每个更新后的样本为单位提供了一次潜在的被选中机会。如第4.15节所述，如果

， 我们根据以下两个条件将多重性因子分配给当时观察到的家庭：

$\square\square\square\square \leftarrow \square\square\square\square * \cong \square\square\square\square \square\square\square\square \leftarrow \square\square\square\square ** \square\square\square\square$

**条件 1** 如果家庭至少有一个成分属于波浪中的目标人口且不是迁移者，则多重因子为2。

**条件 2** 。如果没有家庭的组成部分满足 \* 条件 1 ， 则多重性因子为 1 。

请注意， 条件 1) 和 2) 引入了 固定的个人在时间 (时间上不可追踪的移动者) 和固定的个人 (移动者或移民 \*

关于时间和固定的时间 \* )。 A fixed individual at \*\* 有个人

多重性因子等于 2 ； 固定个体在 \*  $\square\square\square\square$  个体多重性因子等于 \*\*  $\square\square\square\square$

$\square\square\square\square *$

1.  $\square\square\square\square\square\square ** \square\square\square\square \square\square\square\square$

因此， 在波浪中， 可能会出现以下家庭多重性值：

1. 若在第 . 波次的原始样本中至少有一个组件被选中，则多重性因子  $\square\square\square\square$  等于 2。

2. 在时间点  $\square\square\square\square$ ， 多重性因子等于 2。 \* 如果至少选择了家庭的一个组成部分， 家庭等于 1 。

\*\* \*

3. 如果家庭的至少一部分是  $\square\square\square\square$ ，  $\square\square\square\square \leftarrow \square\square\square\square$  和变量 家庭的数量等于 0 。

在波浪中选择

\*\*  $\square\square\square\square \square\square\square\square \leftarrow *$

最后 个体在当前波和第一波中的反应指标变量  $\square\square\square\square \square\square\square\square$  波 (对于在原始样本中选择的个体) 完成信息以构建基本权重。

为了便于操作理解在当时构建基重所涉及的步骤 t ， 我们在表 4.1 中列出了此过程中涉及的变量以及它们可以采用的值。

表4.1. 连接和其他辅助变量 $t$ 用于计算基权重的时间点示例。例如：两个原始样本（第一样本，和更新样本，）

| 变量                            | 符号 | Description | 价值观                                 | $t^*$                                       | $t^{**}$                                        |
|-------------------------------|----|-------------|-------------------------------------|---------------------------------------------|-------------------------------------------------|
| Personal 识别码                  |    |             | $v_1$                               | 代码。一段时间内的唯一代码。                              |                                                 |
| 原始样本选择                        |    |             | $v_2$                               | 原始样品的时间，包括个人。失踪时个人被纳入样本中家庭。                 | 时间 $t$ 丢失。                                      |
| 家用识别码在原始样品                    |    |             | $v_3$                               | 个人家庭代码原始示例。唯一代码时间。失踪时，个人是纳入样本家庭。            | 代码；丢失。                                          |
| 包含概率等于（                       |    |             | $v_4 = \pi_{t^0,k}^*$<br>or $v_4^*$ | 个体的包容概率原始样本。当个人被纳入样本家庭。                     | Numeric: constant in the                        |
| 当前的家庭代码识别码在当前的波               |    |             | $v_5$                               | wave.                                       | Code.                                           |
| 人口成员资格*                       |    |             | $v_6$                               | 目标的个人成员资格与样本相关的人口（在与变量的关系）。 $v_2$           | 1 (当个人来自的目标人口原始样品)；<br>0 (当个人不在目标人群中原始样品)。      |
| 个人多重性因子当前的波()                 |    |             | $v_7$                               | 个体的多重性因素在原始样本中选择尊重人口固定个人与尊重或(建成考虑变量)。 $v_2$ | 2 (当个人是固定的个人与尊重人口)；<br>1 (当个人是人口)；<br>samples). |
| 家用多重性因子 的值的最大值(在中定义公式 4.15)   |    |             | $v_8 = \frac{1}{v_7}$               | 家庭多重性因素。考虑 2 (当列出个人家庭中的个人。 $v_7$            | $v_7 = 2$ )。家庭有 1 (否则)。                         |
| 响应指示器在原始示例 (仅适用于估计总收入 - 更改参数) |    |             | $v_9$                               | 个人的反应指标原始样本。                                | 1 (答辩人)；<br>0 (非答复者或不在原始样品)                     |
| 响应指示器在当前的波                    |    |             |                                     | 1 (受访者) 中个人的反应指标；当前的波浪。                     | 0 (非答复者)。                                       |

\* 示例：在 wave 中，家庭在 的原始样本中至少选择了一个人  $10$  (目标人群为目标人群)：(个人属于目标人群)；当个体在价值为0时被选中，该个体在家庭中的纳入个体在时间 $t$ 不属于目标人群，则值为1；否则为0。  $v_4^* = v_4^* \cdot v_4^*$  当家庭中的个人在目标人群中时，值 = 1)。

time  $v_4^* = v_4^* \cdot v_4^*$ ；  
 $v_4^* = v_4^* \cdot v_4^*$  (即新生儿或移民后

\*\*  $v_4^* = v_4^* \cdot v_4^*$

根据上表中列出的变量， $v_4^* = v_4^* \cdot v_4^*$  base 第 3.3. 1 节介绍的家庭重量可以计算为：

$k$

$$k = \frac{1}{v_4^*} \frac{1}{v_4^*} \frac{1}{v_4^*} \dots \frac{1}{v_4^*} \dots$$

每个人在哪里 is  $v_4^* = v_4^* \cdot v_4^*$

$v_4^* = v_4^* \cdot v_4^*$

$v_4^* = v_4^* \cdot v_4^*$

$v_4^* = v_4^* \cdot v_4^*$ ，being  $v_4^* = v_4^* \cdot v_4^*$   $\sum$

$v_4^* = v_4^* \cdot v_4^*$  家庭代码。最后，我们注意到，随着家庭动态的变化，家庭代码可能会随时间发生变化，除非某些特定的  $(i \in k_k)$  家庭  $(v_v)$  除外。  $v_4^* = v_4^* \cdot v_4^*$  for  $v_4^* = v_4^* \cdot v_4^*$   $v_4^* = v_4^* \cdot v_4^*$  代码不变。  $v_4^* = v_4^* \cdot v_4^*$   $v_4^* = v_4^* \cdot v_4^*$

#### 4.3. 确定通过入户访问收集非响应者数据的最小变量集以及通过代理访问收集搬迁者数据的最小变量集

样本规模的减少必须提高横截面和纵向估计的准确性。流失（人们退出以及人们迁移）可能会使估计值失真，如果退出或迁移的人群系统性地与留在调查中的人群不同。例如，假设当人们开始就业时他们就会退出，因为不再有时间参与调查。在这种情况下，面板将产生低估就业人数的估计值。面板更新、追踪和校准估计方法可以解决样本单位的损失问题。

为了减少样本流失偏差，在第一阶段可以收集非响应者的额外信息，这些信息可以用作两步校准方法中的辅助变量。在此方法中，已知选定样本级别的总值包括来自响应者的信息以及从非响应者收集到的信息衍生出的变量。为此，需要进行两次校准：第一次校准调整基本权重以应对面板样本流失问题，第二次校准处理覆盖率不足的问题。如前所述，如果样本流失的原因与覆盖率不足的原因不同，则两步方法可以提高估计值的准确性。然而，在将此类信息的收集引入当前统计流程之前，必须仔细设计和实施，并通过实证研究评估其效用和准确性。

额外的变量可以通过一个简短的问卷模块收集，该模块涵盖了人口统计变量（如年龄、性别等）、社会变量（就业状况、literacy水平等），以及对于搬迁者而言，还包括他们离开调查区域的原因（如经济原因、个人原因等）。最后，应记录搬迁者的电话号码，以便与调查中未能找到的人进行电话访谈。

**门口的采访。** 上门访谈是一种简化的访谈方式，用于获取非响应者特征的关键信息，这些非响应者往往与所关注的目标现象有关。例如，在官方调查语言中，低识字率的移民群体可能在识字相关的非响应方面集中度较高，或许在一般识字率方面也是如此。然而，由于缺乏对这些群体的进一步信息，这种集中程度目前尚不清楚。上门访谈是一种简短而简单的访谈，针对不愿意参与主要调查的家庭（或个人）进行，旨在获取这些非响应者的基本信息。如上所述，人口统计、识字和就业问题是一些可以向非响应者提出的基本问题。

**代理面试。** 代理访谈可以在两种不同的背景下使用：i) 如果家庭中有一人是搬迁者，则另一家庭成员可以回答简短的问卷调查；ii) 如果整个家庭搬迁，则邻居（或社区）可以回答简短的问卷调查。在两种情况下，获取搬迁者的电话号码可能允许进行电话访谈，前提是跟踪规则不允许进行跟踪。

如家访一样，代理访问的目的是确定样本中特定特征的集中程度。可以对代理受访者进行模块测试。然后添加两个额外的问题：搬家公司电话号码以及搬离家庭的原因。

#### 4.4. 规划多模式数据收集

调查的运营成本限制了跟踪规则；因此，在所有情况下，离开采样区域（即普查区、教区、区等）的搬迁者无法被追踪。为了获得关于搬迁者的合成信息并减少由于搬迁导致的面板样本流失，一种经济有效的策略是进行电话访谈。在这种情况下，使用简化的问卷更为优选，或者在更复杂的情况下，调查设计可以包括计算机辅助电话访谈（CATI），并且面板调查将依赖多模式数据收集。

最终，标准问卷需要询问电话号码以避免在未来波次中出现搬迁者的不回应。

#### 5. LSMS - ISA 小组调查的简短说明

我们通过使用Living Standards Measurement Study – Integrated Surveys on Agriculture (LSMS-ISA) 的实际数据进行模拟，对这一方法论框架进行了测试。这项研究得到了世界银行的支持。模拟的目的是验证所提出的估计器的属性和可用性，并将其估计值与特定应用案例中当前产生的估计值进行比较。LSMS-ISA调查在整个估计过程中具有相当大的异质性，包括抽样计划、跟踪协议、目标参数、辅助变量集以及使用的估计器。为了简化起见，所提出的估计器不包含模拟所用调查的所有具体问题和细节。相反，我们专注于所有纵向调查共有的主要步骤，以得出不针对特定调查的一般结论。

LSMS-ISA调查是全国代表性的纵向家庭调查，旨在理解生活水平与农业之间的关系。自2008年启动以来，LSMS-ISA项目已在撒哈拉以南非洲八个合作伙伴国家与统计机构合作，提供技术援助以设计和实施多用途家庭面板调查，并重点关注农业，同时推动数据收集方法的创新和效率。表5.1总结了合作伙伴国家、支持的面板调查以及调查实施的年份。

表 5.1 LSMS - ISA 调查

| Country | Survey                   | Year                                                                                   |
|---------|--------------------------|----------------------------------------------------------------------------------------|
| 布基纳法索   | 布基纳法索环境和谐生活条件 (EHCVM)    | 2018/19, 2022                                                                          |
| 埃塞俄比亚   | 埃塞俄比亚社会经济调查 (ERSS)       | 2011/2012, 2013/2014, 2015/2016, 2018/2019, 2021/2022                                  |
| 马拉维     | 第三次综合住户调查 (IHS3)         | 2010                                                                                   |
|         | 综合住户调查 (IHPS)            | 2013, 2016/2017, 2019/2020                                                             |
| Mali    | 农业科学 (EAC-I)             | 2014/2017                                                                              |
| 尼日尔     | 全国家庭生活调查条件与农业 (ECVM / A) | 2011/2014                                                                              |
| 尼日利亚    | 综合住户统计调查 (GHS)           | 2010/2011, 2012/2013, 2015/2016, 2018/2019                                             |
| 坦桑尼亚    | 坦桑尼亚国家小组调查 (TZNPS)       | 2008/2009, 2010/2011, 2012/2013, 2014/2015, 2020/2021                                  |
| 乌干达     | 乌干达国家小组调查                | 2009/2010, 2010/2011, 2011/2012, 2013/2014, 2015/2016, 2018/2019, 2017/2018, 2019/2020 |

总体而言，LSMS-ISA的时间序列样本完全或部分（即作为子样本）与现有的农业或家庭样本调查相一致。例如，埃塞俄比亚农业样本调查（AgSS）社会经济调查（ESS）对现有农业家庭的一部分进行了访谈，将其纯粹的农村样本扩展到一个城镇EAs的样本。坦桑尼亚国家面板调查（NPS）和乌干达国家面板调查（UNPS）由来自家庭预算调查的子样本组成，具体而言分别是坦桑尼亚家庭预算调查（THBS）和乌干达国家家庭调查（UNHS）。在马拉维，综合家庭面板调查（IHPS）跟踪并重新访谈了第三次综合家庭调查（IHS3）的一部分家庭样本。在尼日利亚，普通家庭调查-面板（GHS-Panel）是从核心横截面调查中抽取的一个子样本。最后，在尼日尔，纵向研究追踪了全国住户生活条件和农业调查（ECVM/A）的整个样本。

LSMS-ISA 调查包括两阶段的概率样本，使用通用人口普查作为抽样框架。在大多数 LSM S-ISA 样本中，以大小比例概率选择 enumeration areas（EA）作为主要抽样单位。然后从选定的 EAs 中所有家庭的完整列表中随机选择样本家庭。因此，LSMS-ISA 样本构成 EA、家庭和个人的随机样本。

LSMS-ISA样本旨在在全国范围内代表家庭和个体。理想情况下，纵向研究能够保持代表性随时间的变化，表明样本应代表每次调查时的当前人口，并反映初始人口随着时间动态变化的情况。为了同时保持这两种类型的代表性，纵向调查会追踪之前调查轮次中被访谈的人，并添加新的个体，以确保包括新加入的人口群体，如移民和新生儿（Glewwe和Jacoby 2000）。为此，面板调查制定了规则来定义后续轮次中的访谈目标，并创建了特定的程序来跟踪从原始地点迁移的家庭和个人。此外，他们采用更新后的样本（即旋转面板或分层面板）以及纳入未参与先前研究的个体的程序，以确保这些个体也被包含在调查中。

在讨论面板调查中的追踪问题时，Witoelar (2011) 引入了访谈目标和追踪目标的概念。访谈目标指的是需要进行访谈的对象；追踪目标是在他们搬迁后如果能找到的话需要追踪的对象，并且如果找到，应当对其进行访谈。同样地，作者在定义根据对象是否在同一地点被发现或为搬迁者而制定的受访标准时，区分了跟进规则与追踪规则。在一定程度上，LSMS-ISA面板调查可以被视为包含家庭和个人的面板样本。在基线阶段被发现并进行访谈的所有家庭和个人在整个面板样本历史中都是访谈目标。

LSMS-ISA 跟踪方法更为多样化，主要在以下方面存在差异：（i）跟踪单位的目标对象，（ii）跟踪目标的特性，以及（iii）跟踪的地理范围。这三个要素的选择将影响代表性和成本。

LSMS-ISA的目标单位和重新访谈对象各不相同。通常情况下，这些面板调查会追踪房屋、家庭或个人，每种选择都有其独特的含义。2022年，布基纳法索EHCVM回访了2018/19年访问过的房屋。尼日利亚和埃塞俄比亚追踪并重新访谈参与首次数据收集的家庭，但不包括从家庭分离的个人。马拉维IHPS、坦桑尼亚NPS和乌干达UNPS则追踪个人。

跟踪住房而非家庭进行追踪是一种更简单的选项。随着时间的推移对家庭进行跟进需要定义哪个家庭是原家庭的继承者，这并非一项简单的任务，因为家庭结构和边界往往不够明确，并且家庭可能会分裂或重组。正如Witoelar ( 2011 ) 所指出的那样，家庭成员身份和构成会迅速变化，家庭结构和边界往往模糊不清，而家庭的定义也常常根据调查目的的不同而有所不同。Falorsi ( 2009 ) 提出了并讨论了连续性规则，用以界定哪些家庭被视为延续了上次调查中观察到的家庭。

LSMS-ISA面板调查已经确立了实用规则来定义跟踪的家庭。例如，埃塞俄比亚ESS将仍在原始居所生活的家庭成员组视为原家庭。如果整个家庭搬迁并仍居住在埃塞俄比亚境内，则继续追踪并重新访谈；如果所有家庭成员离开该国，则该单位将从样本中移除。如果家庭分隔，与先前波次中识别的家庭户主一起搬迁的家庭成员部分将被追踪并进行访谈。同样，尼日利亚GHS-Panel对仍在原始居所的家庭成员进行访谈，并追踪那些虽然搬迁出居所但仍位于尼日利亚境内的家庭。如果家庭分裂，原家庭被定义为与上一轮家庭户主共同居住的家庭成员组。如果找不到原家庭户主，则原家庭被定义为包含最多原家庭成员数量的一组。

住户和家庭面板的一个缺点是，那些搬出基线家庭的成员不再包含在样本中，尽管他们仍然属于人口的一部分。因此，剩余成员的样本在当前以及随着时间演变的人口中代表性会减弱。然而，如我们在论文后文所述，这种代表性不足得到了缓解，因为所有LSMS-ISA调查都会纳入新加入被访谈（并追踪）的家庭的个体，例如新生儿或移民。

随着时间的推移，代表性损失可以通过追踪原始个体来解决。个体追踪在多大程度上能减轻代表性问题，取决于对追踪目标的特征以及追踪地理范围的决策。理论上，决定追踪哪些个体依赖于需要长期代表的人口特性。专注于特定子群体的研究目标与旨在描述全国人口的LSMS-ISA面板调查有所不同。

然而，追踪所有个体迁移者会非常昂贵。追踪目标越雄心勃勃（例如，个体必须被追踪的距离越远离其原来的住所），成本将越高。通常情况下，LSMS-ISA的个体追踪目标是根据个人的年龄及其与家庭户主的关系来定义的。马拉维、坦桑尼亚和乌干达会追踪并跟进与家庭户主有生物学关系的迁移者，排除当时是仆人或客人的个体。此外，所有这三个调查都设定了一个下限，即在家庭成员迁移出家庭的情况下，不会对其进行追踪（马拉维为12岁，坦桑尼亚和乌干达为15岁）。

跟踪的地理范围也可能影响未来轮次中面板调查的代表性，如果那些移动最远的人口或家庭与其他人差异很大。通常，要跟踪多远的距离是基于行政边界、地理边界或距离（Witoelar, 2011）。仅依赖本地跟踪的面板调查只在相同的选举区（EAs）或附近寻找家庭。远程跟踪则允许追踪远离原住所的迁移者。无论是否跟踪家庭，

对于个人而言，所有LSMS-ISA调查均采用远程跟踪方法来追踪并采访原EA（Enumeration Area）中的迁出者，只要他们仍居住在国内。

在马拉维、坦桑尼亚和乌干达的面板调查中，当个人组建新的家庭时，与该个人同住的所有家庭成员会进入样本并成为追踪目标。将联合家庭纳入样本可以考虑可能被其他情况下忽视的新人口，并捕捉到总体人口动态的变化。然而，包括搬迁者的新人户会增加数据收集成本。在纵向研究中提升代表性的一种更经济的方法是采用轮转面板或样本刷新，一些LSMS-ISA调查已使用这种方法。例如，乌干达UNPS 2013/14阶段淘汰了初始样本中三分之一的EAs（行政区域）。坦桑尼亚NPS样本在2014/15年进行了刷新，但随后2020/21年的NPS数据收集仅跟进刷新样本，这导致了从最初阶段开始的纵向代表性出现断层。2022年，埃塞俄比亚ESS样本完全刷新，布基纳法索Enquête Harmonisée sur le Conditions de Vie des Ménages (EHCVM) 则继续跟踪其2018/19年的家庭，通过从较近的抽样框架中选择一组新的EAs来补充样本。最后，尼日利亚在2018/19年也对其调查样本部分刷新。

## 6. 乌干达数据的实证评价

作为案例研究，我们应用了上述讨论的估计方法来分析乌干达国家面板调查（UNPS）。UNPS是由乌干达国家统计局（UBOS）实施的一项多用途家庭面板调查，得到了世界银行LSMS-ISA项目的技术支持。该调查始于2009/10年，作为2005/6年乌干达国家户户调查的后续项目，其主要目的是为预算决策和政策干预提供信息，并监测主要的国家政策和项目。每年进行一次的UNPS还提供了关于收入动态和消费支出估算的代表性信息，以在其他全国性调查努力之间监控贫困情况。此外，UNPS收集了高质量的农业数据，这些数据整合在一个多主题框架内，有助于理解农业、福利与其他社会人口统计特征之间的联系。

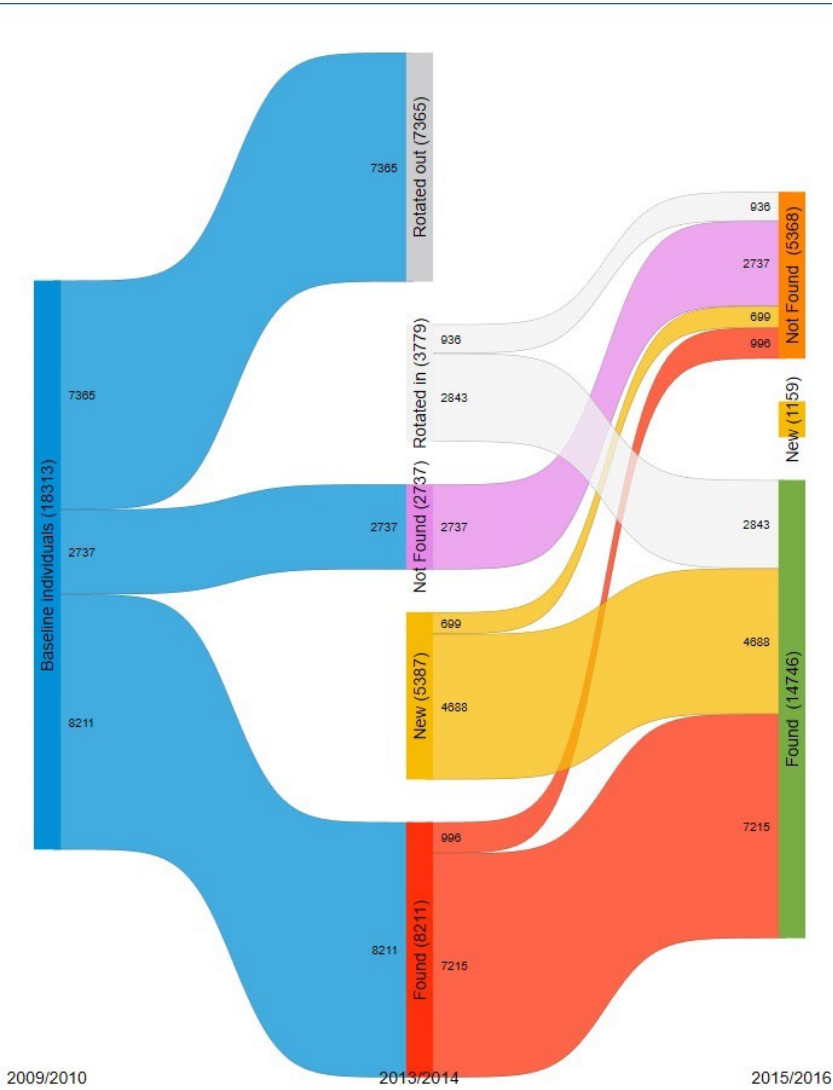
自2009年启动以来，国家统计局开展了八轮约3000户家庭的数据收集工作。UNPS 2009/10年之后分别进行了2010/11年、2011/12年、2013/14年、2015/16年、2017/18年、2018/19年和2019/20年的调查。其样本在国家级、城乡和地区级具有代表性。本研究使用了UNPS 2009/10（第一波）、UNPS 2013/14（第四波）和UNPS 2015/16（第五波）的数据。UNPS 2009/10涵盖了2975户家庭，UNPS 2013/14包括3118户家庭，而UNPS 2015/16则访谈了3304户家庭。经过数据清洗和分析前的数据准备后，我们的最终数据集包含第一波18,313人，第四波17,377人，以及第五波15,905人。

图6.1展示了本研究中三次数据收集波次中受访个体状态的变化。2009/10年的水平代表了第一波数据收集时接受访谈的18,313名个体。在这批个体中，有45%（8,211人）在2013/14年被重新找到并进行了重新访谈，15%的人没有被找到，40%的人被替换掉了。2013/14年的样本更新带来了5,387名新的个体，约占原样本量的30%。在非搬迁户和拆分户的新个体中，共有3,779个新的单位，相当于原样本量的21%。从2013/14年到2015/16年，2013/14年波次样本中有31%的个体丢失，而85%的个体被找到并重新进行了访谈。该图表还展示了旧样本和更新样本随时间的变化动态。例如，我们可以看到，在更新样本中，有87%的个体

在 2015 / 16 轮中接受采访 ， 而在 2013 / 14 浪潮中接受采访的旧样本中有 82% 。

为了本论文的目的，我们专注于2009/10-2015/16和2013/14-2015/16两个面板。在每个面板中，只有在两年都出现的人被纳入分析。由于在第3波离开样本后没有人在第5波重新进入样本，第一个和第二个面板的最终观察次数分别为7,215次和4,688次。

图 6.1. UNPS 2009 / 10 - 2013 / 14 - 2015 / 16 小组样本中的个人动态



除了2013/14年的样本更新外，图6.1的结果是由复杂的协议定义的跟踪规则所决定的。UNPS跟踪方案通过考虑目标人口的流动性来确定跟踪方案，具体包括：i) 整个家庭从初始位置搬迁的情况；ii) 家庭中部分成员迁移到另一个地点而其他成员则迁往其他地方的情况。

并且分化为分户 ( split-offs ) , 以及iii ) 原来家庭中迁出的个体, 形成新的分户家庭。对于上述所有三种追踪目标, 跟踪规则在面板轮次的过程中发生了变化。

在前三个波次中, UNPS 跟踪了所有从 2005/06 年原始位置迁移到其他任何位置的原住户, 尽管初始样本中的并非所有住户都会被单独跟踪, 特别是当其成员离开原住户时。仅对一部分原住户进行个体跟踪, 即每户调查点 ( EA ) 随机选取 2 户 ( 占原住户样本的 20% ) , 这些住户被称为“分家户跟踪目标”。此外, 在这些“分家户跟踪目标”中, 仅追踪 15 岁及以上且与户主有生物学关系的个人。即使他们移出了原始 EA/ 行政村, 这些个人和他们形成的分家户也会被找到并进行访谈。一旦访谈完成, 这些分家户及其新家庭的所有成员将被纳入 UNPS 样本, 并在后续的所有 UNPS 波次中接受访谈和跟踪。

自UNPS 2013/14开始, 分拆目标追踪的范围扩展至包含所有位于乌干达境内的样本家庭, 不论其与原家庭位置的距离、是否为原始家庭或分拆家庭。在这些家庭中, 仅年满15岁且在上一轮调查中被识别为主户头、配偶或主户头子女的家庭成员有资格进行追踪。其他成员仅在与上述核心成员同住的情况下进行访谈。这意味着, 如果在最后已知的位置找不到任何核心成员, 则即使之前的家庭成员仍居住在那里, 该家庭也不会被访谈。

在第四波次中, UNPS的样本得到了更新。原样本中三分之一的家庭被替换, 作为面板更新的一部分, 这些家庭不再被跟踪或访谈。

在当前的UNPS设置下, 追踪个体需要填写一份个体追踪表, 该表包含所有分户人员和/or 单户搬迁者的全部联系方式。所需的新地址信息通常是从其之前的家庭成员或其他知情人士处收集的。对于每个已搬迁的核心成员, 都需要填写一份追踪表。根据表格中填写的信息, 联系并访谈搬迁个体。

尽管追踪目标样本仅包括每个家庭的核心成员, 但所有与这些核心成员共同生活的人均被访谈并成为联合国综合人口统计样本 ( UNPS ) 的一部分。最后, 如果这些个人成为新分户的家庭核心成员, 即使他们移居到不同的地点, 他们也会在后续的UNPS波次中接受访谈。

## 6.1. 横截面估计的实证评价

我们计算了2015/16年的UNPS估计值。分析中包含了先前波次的数据, 以识别家庭动态 ( 例如, 迁居者、移民和新生人口 ) 。

我们校准了基础权重 ( 表示式3.8中给出 ) , 将其与2014年人口普查数据和UBOS官方2015年的官方预测相结合, 以匹配已知的按性别和年龄组划分的人口总数。校准所使用的预测值见表6.1。此后, 我们将这些权重称为校准后的GWSM基础权重。

我们将校准后的GWSM基准权重应用于2015/16年UNPS数据集中的若干变量, 并将其与乌干达官方统计数据进行比较, 以评估这些权重的功能。

使用UNPS 2015/16抽样权重。基线权重是在家庭层面定义的（见第三部分）。UNPS抽样权重是在个人层面定义的，并且目前由UBOS（2011）用于调查估计。这些权重是根据一个复杂的过程获得的，该过程基于非响应倾向得分对非响应进行初步调整。具体而言，最终的个人层权重通过以下六个方法步骤获得：i) 定义基线权重为纳入概率的倒数，ii) 计算公平份额修正因子，iii) 将第三步定义的权重合并，iv) 计算流失率修正因子，v) 剔除至第四步的权重，和vi) 后期分层。

**表 6.1. 按年龄和性别分列的人口预测 (2015 年)**

| 年龄组   | Male       | Female     |
|-------|------------|------------|
| 0-4   | 3,220,300  | 3,028,800  |
| 5-9   | 2,870,800  | 2,730,500  |
| 10-14 | 2,528,000  | 2,500,100  |
| 15-19 | 2,008,700  | 2,097,500  |
| 20-24 | 1,509,200  | 1,790,100  |
| 25-29 | 1,181,200  | 1,403,800  |
| 30-34 | 938,700    | 1,082,800  |
| 35-39 | 743,900    | 836,500    |
| 40-44 | 630,600    | 677,500    |
| 45-49 | 470,800    | 486,800    |
| 50-54 | 382,300    | 443,200    |
| 55-59 | 241,700    | 277,900    |
| 60-64 | 194,600    | 241,600    |
| 65-69 | 140,400    | 171,100    |
| 70-74 | 114,900    | 158,600    |
| 75-79 | 71,500     | 90,700     |
| 80+   | 96,400     | 140,600    |
| Total | 17,344,000 | 18,158,100 |

注：UBOS 对 2015 年的官方预测和基于 2014 年的人口普查，可在 <https://www.ubos.org/wp-content/uploads/2018/06/Population-Projections-2018.xlsx> 于 2021 年 6 月 28 日下载。

我们主要使用国家人口和住房普查（NPHC）2014年（UBOS 2016；UBOS 2017）和乌干达国家住户调查（UNHS）2016/17的数据和报告作为官方统计数据的主要来源。前者由UBOS每十年左右进行一次，旨在收集乌干达人口的基准人口统计和社会经济数据。后者是自1999/20年起由UBOS实施的第六次追踪调查，是一项横断面调查，旨在收集人口统计和社会经济特征的数据，样本量为15,636户家庭。UNPS是其2005/06年调查的后续调查，并且很大程度上采用了相同的方法论。

The NPHC 2014和UNHS 2016/17是我们在找到的与UNPS 2015/16最接近的数据来源。UNPS 2015/16的参考时期为3月。

2015年至2016年3月期间，而NPHC 2014的数据收集于2014年8月至9月进行，UNHS 2016/17的数据收集则在2016年6月至2017年6月间实施。

表6.2展示了国家级个体和家庭特征的比较估算指标。前三列分别显示了未使用抽样权重（无权重估计器）、使用原始UNPS权重（当前UNPS估计器）以及使用校准GWSM基权重（校准GWSM基估计器）的指标。无权重估计列是有帮助的，因为它表明了各种加权策略如何作用于最终估计值，从而改变无权重估计值。

在过去的两列中，表格呈现了官方统计数据及其来源。总体而言，在个体指标层面，使用校准后的GWSM基线估计器而非当前的UNPS估计器进行估算更为准确。采用校准后的基线GWSM权重计算的女性人口比例和18岁以下儿童的比例，与官方统计数据非常接近。这一结果符合预期，因为权重是基于我们作为基准使用的2014年人口普查数据进行校准的，该数据用于人口总量预测。

无论我们赋予其多大的权重，2015/16年的UNPS 高估了10岁以上识字人口的“真实”官方值。尽管这三种估计是一致的，但由于这些估计基于大约16,000名个体的样本，表中报告的差异在统计上是显著的。对于如此大的样本量，考虑到一个 *deft* 值 (Kish, 1966) 范围从2到3 (意味着 *deff* 值范围从4到9)，比例的采样标准偏差在0.007和0.010之间的值附近变化。

未加权的统计数据显著高于官方统计数据（80.7%比72.2%），这表明非响应可能影响了文盲人口。因此，这一子群体在样本中被低估。然而，观察到的差异可能是由于问卷中询问信息的方式导致的可比性问题。不同的加权系统似乎无法解决这一问题。最后，校准后的GWSM基权重估计值对于两项关键调查指标——劳动力人口和全国范围内的贫困头数——显得更为准确。

在家庭层面，当前的UNPS和校准后的GWSM基估计值在可比性上更为接近，尽管前者能产生更准确的统计数据，尤其是在城市家庭占比和农业家庭占比的变量上。在少数情况下——具体而言是电力接入和住房所有权——无论使用何种权重，UNPS 2019/20均会产生不准确的官方统计数据估计值。

**表 6.2 未加权、当前 UNPS 和校准 GWSM 基本估计的个人和家庭估计**  
已校准

|              | 未加权  | UNPS | GWSM<br>base | 官方统计 |              |
|--------------|------|------|--------------|------|--------------|
|              | Mean | Mean | Mean         | Mean | 来源           |
| <b>个人</b>    |      |      |              |      |              |
| Female       | 51.1 | 50.8 | 51.1         | 51.0 | 人口普查(2014 年) |
| % 18 岁以下儿童   | 51.2 | 52.1 | 55.0         | 55.0 | 人口普查(2014 年) |
| 识字人口(+ 10 年) | 80.7 | 80.8 | 81.7         | 72.2 | 人口普查(2014 年) |
| 工作人口         | 76.8 | 76.1 | 80.1         | 78.8 | UNHS (2016)  |
| 贫困人数         | 20.4 | 19.3 | 19.9         | 21.4 | UNHS (2016)  |
| <b>家庭</b>    |      |      |              |      |              |
| 城市家庭份额       | 25.3 | 25.1 | 22.9         | 25.0 | 人口普查(2014 年) |
| 家庭的大小        | 4.8  | 5.0  | 4.8          | 4.7  | 人口普查(2014 年) |
| 农业家庭份额       | 78.7 | 80.1 | 79.4         | 80.0 | 人口普查(2014 年) |
| 中的房间数量       | 2.2  | 2.3  | 2.1          | 2.4  | UNHS (2016)  |
| 家庭           |      |      |              |      |              |
| 家庭拥有一部手机     | 74.0 | 72.9 | 75.4         | 74.3 | UNHS (2016)  |
| 家庭可以访问       | 15.4 | 15.1 | 15.6         | 21.0 | 人口普查(2014 年) |
| 电力           |      |      |              |      |              |
| 家庭可以获得安全     | 75.2 | 72.2 | 71.7         | 72.0 | 人口普查(2014 年) |
| 水            |      |      |              |      |              |
| 家庭拥有住宅       | 81.9 | 82.9 | 81.1         | 71.8 | UNHS (2016)  |
| 房子有砖墙        | 67.2 | 63.8 | 64.2         | 66.6 | UNHS (2016)  |

表6.3按居住地区划分了劳动力人口的比例。校准后的GWSM基估计值（82.0%）似乎比当前UNPS估计值（78.8%）更接近官方统计数据，特别是在农村地区（82.6%）。如前所述，尽管这些差异很小，但它们在统计上是显著的。在城市地区，估计值更为争议性，我们没有找到最优估计值。

**表 6.3. 按居住地区分列的就业人口**

| 校准 GWSM | 未加权  | UNPS | base | 官方统计 |           |
|---------|------|------|------|------|-----------|
|         | Mean | Mean | Mean | Mean | Year 来源   |
| 农村      | 79.6 | 78.8 | 82.0 | 82.6 | 2016 UNHS |
| Urban   | 67.5 | 67.4 | 72.7 | 69.0 | 2016 UNHS |

在完成国家级统计数据和分析后，我们转向研究居住区域内的贫困情况（见表6.4），该表未显示出各种估计值之间的具体差异。农村地区的贫困发生率远高于城市地区。

**表 6.4. 按居住面积分列的贫困人数比率**

| 已校准 | 未加权 | UNPS | GWSM 底座 | 官方统计 |
|-----|-----|------|---------|------|
|-----|-----|------|---------|------|

|       | Mean | Mean | Mean | Mean | Year | 来源   |
|-------|------|------|------|------|------|------|
| 农村    | 23.8 | 22.2 | 22.3 | 25.0 | 2016 | UNHS |
| Urban | 9.0  | 9.6  | 10.2 | 9.6  | 2016 | UNHS |

## 6.2. 纵向估计的实证评价

面板调查产生净变化和总变化估计（第二章第二节）。总变化提供了关于净变化的重要见解，因此，通过转换矩阵获得的横截面估计值应与整体UNPS 2015/16样本所得的估计值一致。

表6.5展示了2009年至2015年和2013年至2015年间就业状况的变化动态。特别地，前三列报告了2015年的就业状态相对于2009年的状态。仅当个体出现在2009年和2015年的数据中时，它们才会被表列出。最后三列则对2015年那些在2013年也参与调查的个体，显示了相同的统计信息。

聚焦于2009/2015年的估计值，我们注意到所有估算值均高于表6.2中呈现的估算值。这一结果表明，在这些年份中离开的人群——占大多数人员流失——主要是失业者。三个估算器显示，面板中包含了78.6%、77.6%和81.4%的就业人员。重新平衡操作调整了面板数据。关注2013/2015年面板给出的横截面估算值，根据未加权估算器，77.4%的个体是就业状态；根据当前UNPS估算器，这一比例为76.7%；而根据校准后的GW-SM基权重估算器，这一比例为80.8%。

我们可以断言，在考虑不同的子样本时，即2009/2015年和2013/2015年面板以及整个2015年的样本时，最后一个估计器产生的估计值更加稳定且一致。三个估计值之间的相对变异系数（标准误与均值之比）为0.8%。UNPS估计器的相对误差为1.0%，而未加权估计器的相对标准误为1.2%。

**Table 6.5. Employment transition matrix estimates by unweighted, current UNPS, and calibrated GWSM base estimators**  
2009/15 2013/15

|                                        | 未加权 UNPS |      | 已校准<br>GWSM 底座 | 未加权 UNPS |      | 已校准<br>GWSM<br>base |
|----------------------------------------|----------|------|----------------|----------|------|---------------------|
| 失业<br>2009 (or 2013)<br>and 2015       | 5.1      | 5.4  | 3.7            | 11.6     | 12.1 | 9.3                 |
| 失业<br>2009 年(或 2013 年) 和<br>2015 年受雇   | 6.0      | 5.6  | 5.8            | 8.5      | 8.1  | 7.6                 |
| 2009 年就业<br>(或 2013 年) 和 2015 年        | 72.6     | 72.0 | 75.6           | 68.9     | 68.6 | 73.2                |
| 2009 年就业<br>(或 2013 年) 和<br>失业<br>2015 | 16.3     | 16.9 | 14.9           | 11.0     | 11.2 | 9.9                 |

表6.6中显示的2009/2015和2013/2015面板按居住地划分的就业动态分解表明，面板流失主要集中在城市地区的失业者。与表6.3中所见一致，校准后的GWSM在估计2009/2015样本中的农村地区就业情况时比联合国人口统计系统（UNPS）面板权重更准确。对于2013/15面板的结果表明，更新过程提高了准确性。在城市地区，所有方法都给出了上行估计，特别是在2009/15样本中。研究结果表明，离开UNPS的人群主要是城市地区的失业者，这暗示了有必要改进追踪规则以减少估计值潜在偏误的需求。

**表6.6. 按居住地迁移矩阵估计、未加权当前UNPS以及校准后的GWSM基估计值分区域就业情况**  
2009/15 2013/15

| 未加权 UNPS                                  |      |      | 校准未加权 UNPS |      |      | 已校准  |
|-------------------------------------------|------|------|------------|------|------|------|
|                                           |      |      | GWSM       |      |      | GWSM |
|                                           |      |      | base       |      |      | base |
| 2009 年农村失业者<br>(或 2013 年) 和 2015 年城市      | 3.8  | 4.2  | 3.1        | 8.9  | 9.2  | 7.4  |
|                                           | 9.5  | 9.7  | 6.9        | 20.3 | 21.3 | 16.8 |
| 2009 年农村失业<br>(或 2013 年) 和<br>2015 年城市就业  | 4.4  | 4.0  | 4.3        | 8.1  | 7.7  | 7.1  |
| 11.9                                      |      | 11.3 | 13.3       | 10.0 | 9.4  | 9.7  |
| 2009 年农村就业 (或<br>2013) and 2015<br>Urban  | 74.6 | 74.6 | 77.3       | 71.8 | 71.3 | 75.2 |
| 2009 年就业 (或农村<br>2013) and<br>2015 年城市失业者 | 62.5 | 63.1 | 66.8       | 59.5 | 59.9 | 65.5 |
| 16.4                                      | 16.4 | 17.3 | 15.3       | 11.3 | 11.8 | 10.4 |
| 16.1                                      |      | 15.8 | 13.0       | 10.1 | 9.3  | 8.1  |

我们对贫困头计率进行了类似的分析（表6.7）。从2009/2015面板数据得出的估计值大致与官方统计数据相符：分别有21.9%、21.1%和21.4%对应未加权、UNPS和校准后的GWSM基估计器。重新抽样操作（2013/2015样本）将频率降低至20.5%、19.3%和20.0%。当我们把这些估计值与官方统计数据（21.4%）进行比较时，并没有明显看出某个估计器在质量上优于其他估计器。然而，当我们比较贫困头计率三种估计值的百分比相对变异度时（其中两种来自纵向估计，一种来自横截面估计器），我们注意到UNPS估计器的百分比相对变异度为5.2%，而校准后的GWSM估计器为4.1%，未加权估计器为4.0%。

**Table 6.7. Poverty rate transition matrix estimates by the unweighted, current UNPS, and calibrated GWSM base estimators**  
2009/15 2013/15

|                                      | 未加权 UNPS |      | 已校准<br>GWSM 底座 | 未加权<br>d | 校准<br>UNPS d | GWSM<br>base |
|--------------------------------------|----------|------|----------------|----------|--------------|--------------|
| 2009 年非贫困 HH<br>( 或 2013 年) 和 2015 年 | 63.1     | 64.8 | 63.6           | 65.0     | 66.5         | 65.4         |
| 2009 年非贫困 HH<br>( 或 2013 年) 和差的 HH   |          |      |                |          |              |              |
| 2015                                 | 12.3     | 12.1 | 13.0           | 8.3      | 8.1          | 8.3          |
| 在 2009 年糟糕的 HH 中 (或<br>2013) 和差的 HH  |          |      |                |          |              |              |
| 2015                                 | 9.6      | 9.0  | 8.4            | 12.2     | 11.2         | 11.7         |
| 在 2009 年糟糕的 HH 中 (或<br>2013) 和非贫困 HH |          |      |                |          |              |              |
| 2015                                 | 15.0     | 14.2 | 15.1           | 14.4     | 14.1         | 14.5         |

表6.8展示了居住区域转移矩阵的情况，其中2009/2015年的面板估计值大于2013/2015年的估计值。贫困人口的过度估计表明，非贫困人口，尤其是农村居民，更为活跃且更频繁地从UNPS样本中迁出。在2013/2015年的面板中，数据趋于稳定。所有统计结果与来自2015年横截面样本的国家层面估计一致。2013/14年的样本更新似乎减轻了非贫困人口的流失。居住的农村和城市地区表现出相同的趋势。比较贫困头均率的三个估计值（两个来自纵向估计值，一个来自横截面估计器）的相对百分比变异，对于农村子人口，UNPS估计器的相对百分比变异为5.3%，而校准后的GWSM估计器为3.5%。对于城市子人口，UNPS估计器的相对百分比变异为0.6%，校准后的GWSM估计器为3.0%，未加权估计器为5.4%。因此，在这种情况下，我们无法确定哪个估计器更加稳定。

表6.8. 居住地转换矩阵估计的贫困率，由未加权、当前UNPS和校准GWSM基估计值给出

| 2009/15                    |      |      |                     | 2013/15  |      |                     |
|----------------------------|------|------|---------------------|----------|------|---------------------|
| 未加权 UNPS                   |      |      | 已校准<br>GWSM<br>base | 未加权 UNPS |      | 已校准<br>GWSM<br>base |
| <hr/>                      |      |      |                     |          |      |                     |
| 2009 年非贫困 HH 农村地区          |      |      | 59.4                | 59.4     | 61.4 | 61.0                |
| ( 或 2013 年) 和 2015 年 Urban | 81.1 | 83.5 | 81.6                | 84.2     | 83.8 | 83.4                |
| 2009 年非贫困 HH 农村            |      |      |                     |          |      |                     |
| ( 或 2013 年) 和穷人            | 13.8 | 13.7 | 14.0                | 9.2      | 8.8  | 9.1                 |
| 2015 年城市 HH                |      |      |                     |          |      |                     |
|                            | 7.0  | 6.5  | 7.7                 | 5.5      | 5.6  | 5.0                 |
| 2009 年贫困 HH 农村 (或          |      |      |                     |          |      |                     |
| 2013) and 2015 Urban       | 11.4 | 10.6 | 9.7                 | 14.7     | 13.4 | 13.2                |
|                            | 3.0  | 3.1  | 2.1                 | 3.8      | 4.1  | 5.4                 |
| 2009 年在贫困的 HH (或农村         |      |      |                     |          |      |                     |
| 2013) 和非贫困 HH              | 16.7 | 16.3 | 16.3                | 16.7     | 16.4 | 16.6                |
| 2015 年城市                   |      |      |                     |          |      |                     |
|                            | 8.8  | 6.9  | 8.6                 | 6.5      | 6.5  | 6.2                 |

### 6.3. 实证结果总结

比较估算器的评估工作在两个层面上进行：统计性质和实际应用。就其统计性质而言，第四部分介绍了新校准估算器的性质，第六部分则提供了在实际数据上的具体结果。鉴于声称一种估算器优于另一种估算器涉及的复杂性，我们使用基准统计数据来进行这种评估。结果总结如下：

- 校准的 GWSM 基估计器似乎可以产生更准确的个体水平统计数据比目前的 UNPS 估计器还要多。
- 这两个估计量在家庭层面产生同样准确的统计数据，这表明 GWSM 基估计器可以通过考虑家庭已知总数。
- 虽然基于转移矩阵的横截面估计通常不会根据整个 UNPS 样本再现精确的横截面估计，示例刷新操作缓解了这个问题。这表明两个估计器都可以仅部分处理人口的动态性质，这对 i) 很重要相干地实现周期性样本刷新 (即旋转面板) 和 ii) 改进跟踪规则，因为面板减员会影响感兴趣的特定亚群。

经过校准的GWSM基线跨截面估计在转移矩阵中的表现似乎在改变样本（例如2009/2015或2013/2015）时比当前的UNPS估计更为稳定。

尽管模拟的范围是为了获得一般性结论，就其实际实施层面而言，考虑到乌干达案例的具体性，我们可以给出以下指示，旨在尽可能少地改变当前过程：

- 根据表5.1，乌干达面板数据存储了变量，和。由于原始样本（第一波）中的非回应者与未包含在原始样本中的新个体无法明确区分，因此变量并未被显式存储。<sup>1</sup> □□□□<sup>3</sup> □□□□<sup>5</sup> 样本<sup>2</sup>
- 我们通过使用其他存储变量并作出一些准确假设，人为创建了人口成员变量。该变量应该直接询问（\ (vv\ ) ）。<sup>6</sup>

在新版问卷中（即，“若家庭至少包含原样本中从某一人口群体中选取的一个成员，该个体是否属于同一人口群体或为新加入者（新生儿或移民）？”）。

- 我们通过使用其他存储变量并做出一些精确假设，人为地创建了个体多样性因子。该变量应在新的问卷调查版本中直接询问（即，“原始样本中的个体是否发生变化”）。 □

□□□<sup>7</sup>  
2009 年至 2013 年的教区（原始样本的年份？）“）。

- 基于的值，我们计算了家庭多重性因子。

- 我们人为地重建了每个波中的响应指标变量，<sup>7</sup> and  
做一些准确的假设

<sup>9</sup> 我们存储了每个被选中个体的变量和概率，即使对于样本中的个体  $\mathbf{v}$ （  
而不是已经纳入的个体）。if 不回应原始的，did  
（即，在当前面板数据中，缺少权重）<sup>4</sup>。

我们在计算GWSM权重时并未依赖于家长家庭的概念，这在实际操作中可能是一项复杂的任务。

- 我们使用标准校准估计器计算最终权重，并将这些权重应用于横截面和纵向估计，从而创建一个独特的抽样权重向量，简化了个体和家庭估计的一致性。

## 7. conclusions

鉴于所探讨主题的复杂性，我们现在回顾主要步骤以提供各种结果的统一图景。在本文中，我们探索了通过结合按先前时间选择并遵循特定规则的样本来追踪面板成员从初始调查访谈地点移动后的横截面和纵向估计质量提升的问题。

我们已经认识到问题的复杂性，并且需要从多个层面采取行动，这些层面涉及相互关联但又不同的科学领域。

第一个方面是定义所需的数量，考虑到总体既具有横截面特征也具有纵向特征。这导致了对统计单位的适当定义，这些单位随时间变化并在关注的不同点上进行考虑，以便构建横截面和纵向估计值。对于简单的统计单位（如个人），解决方案相对无争议；但对于复合单位（如家庭），可能有多种解决方案，每种方案在代表自然总体的真实情况以及观察感兴趣的现象的便利性方面都有利有弊。我们提出了多对多的方法来定义家庭随着时间的延续性，这意味着初始时间的一个家庭可能会在下一个时间点产生多个家庭。相反，当前的家庭可能源自前几次的家庭。通过这种方法，纵向的人口与家庭人口之间存在完美的对应关系。此外，多对多方法包括一对一连续性规则作为一种特殊情况。

接下来，我们考虑了观测过程，使理论上定义的人口量得以观测。我们了解到，在重复观测的时间线上必须加以考虑，以避免由于将相同时间点的数据重复纳入分析而产生的失真现象。

观察到某一群体被多次观察，或者相反地，未能代表某些可能从经济和社会角度看具有特殊利益的亚群体，例如搬迁者、新生儿或某一特定时刻进入人口的个体。

我们现在转向估计方法论。GWSME估计算法提供了一种方法，可以提供无偏估计，并且易于实施。在家庭层面校准估计值并纠正权重具有多种理论和应用上的优势。方差估计具有良好的渐近行为，更为透明且有坚实的理论基础。由于估计基于少量调整阶段，流程更加简单。最后，这种方法显著提高了基于家庭层面变量创建的估计值与基于个体层面定义的变量创建的估计值之间的一致性。

最后，研究的最后一方面关注数据收集以及可以实施的不同策略，以减轻由于非响应和面板样本流失导致的偏差。

在结论中，我们使用了乌干达国家面板调查（UNPS）的数据来实验性地评估建议的技术。总结来说，我们发现校准后的GWSM基估计器产生的个体层面统计量似乎比当前UNPS估计器更为准确。此外，校准后的GWSM基横截面估计值在转换样本时显示出更高的稳定性，相较于当前UNPS估计值更为一致。

我们强调，上述结果是通过使用当前UNPS估计器较少的辅助信息获得的，原因在于部分相关信息并未明确记录，同时我们的目标之一是提出一个相对于现有估计器更为简化、计算步骤更少且基于科学文献中广为人知的稳健采样理论的估计器。

## 参考文献

Deville, J. C. and Särndal C. E. (1992)。调查抽样中的校准估计器。美国统计协会杂志。Vol.87. N° 418.

FAO (2015). 综合调查框架，指南，罗马。[http://www.gsars.org/wp-content/uploads/2015/05/ISF-Guidelines\\_12\\_05\\_2015-WEB.pdf](http://www.gsars.org/wp-content/uploads/2015/05/ISF-Guidelines_12_05_2015-WEB.pdf)。2015年12月16日访问。

FAO (2014). 全球战略以改进农业和农村统计：综合调查框架技术报告，技术报告系列 GO - 02 - 2014，[http://gsars.org/wp-content/uploads/2014/07/Technical\\_report\\_on-ISF-Final.pdf](http://gsars.org/wp-content/uploads/2014/07/Technical_report_on-ISF-Final.pdf)。2014年12月1日访问。

格罗夫斯，R. M.，Don A. Dillman，J. L. Eltinge，R - J. A. Little。(2001)。调查无响应 威利，纽约。

Falorsi P. D., P. Righi, P. Lavallée。(2019)。不同种群综合观测的最优抽样。调查方法，Vol. 45, No. 3, pp. 485 - 511. Statistics Canada, Catalogue No. 12 - 001 - X.

Falorsi, F, Russo A. (2009)。在 Alleva G., Falorsi P. 每个 Valutazione degli squilibri territoriali。米兰，佛朗哥·安吉利。ISBN：9788856812077。

格莱威 , P. 和雅各比 , H. ( 2000年 ) .“收集面板数据的建议”, 在M. Grosh和P. Glewwe 主编的《为发展中国家设计家庭调查问卷：生活标准衡量研究15年的经验教训》一书中。 华盛顿特区，世界银行。

霍维茨 , D. G. 和 D. L. 汤普森。 (1952)。没有从有限宇宙替换的抽样的概括 . J. Amer. 统计学家. Assoc ., 47,663-685.

Kish, L. (1965), Survey Sampling, New York: Wiley. ISBN 978 - 0471109495.

Laval é e, P. (2007). Indirect Sampling. Springer: Ottawa.

林恩 , P. , (2009) , 《小组调查方法》 , 威利。威利父子有限公司 , 出版物。

梅卡蒂 , F. ( 2007) 。用于多帧调查的单帧多重性估计器。 调查方法 , 33(2): 151-157.

纳兰 , R. D. 1951 年。关于没有替换的抽样 , 有不同的概率。 J 、印第安纳州。阿格里尔。统计学家 , 3, 169 - 174 。

S ä rndal C. E. 和 Lundstr ö m S. (2015). 无响应调查中的估计 威利。纽约。

辛格 , A. C. 和梅卡蒂 , F. (2011) 。广义多重性调整后的 Horvitz - Thompson 估计作为多框架调查的统一方法。 官方统计杂志 , 27(4): 633-650.

乌干达统计局 ( 2011 年) , 乌干达国家小组调查(UNPS) 2009 / 10 - 主要报告 , 乌干达坎帕拉。

乌干达统计局 ( 2017) , 2014 年全国人口和住房普查 - 专题报告系列中的教育 , 乌干达坎帕拉。

乌干达统计局 ( 2016 年) , 2014 年全国人口和住房普查 - 主要报告 , 乌干达坎帕拉。

埃塞克斯大学 . (2016), 英国家庭小组调查 ,  
<https://www.iser.essex.ac.uk/bhps>

Witoelar Firman. (2011). 长itudinal 家庭调查追踪. 世界银行生活标准测量研究-整合农业调查 ( LSMS-ISA ) . 世界银行.

世界银行 . (2020), 基本信息文件 - 乌干达国家小组调查 (UNPS) 2015 / 2016 . <http://survey.worldbank.org/lsms/programs/integrated-survey-agriculture-ISA/uganda#bootstrap-panel> 。

Zardetto, D . (2015) 。 ReGenesees : 一种先进的 R 系统 , 用于复杂抽样调查中的校准 , 估计和抽样误差评估。 官方统计杂志 , 31 (2), 177.