国际微观样本整合共享数据库中经过整合的中国、印度、越南和其他十个亚太国家样本数据中初等教育的统计一致性[1]

Robert McCaa, Lara Cleveland, Patricia Kelly-Hall, Steven Ruggles, and Matthew Sobek[2]

[1] 本研究受美国健康研究院的欧亚国家人口普查微观数据一体化项目(IPUMS-EurAsia)的部分资助（项目号为 HD047283)。本文作者感谢那些将原始微观数据委托国际微观样本整合共享数据库予以整合并同意共享数据库向不受国籍、出生地和居住地限制的全球研究人员免费散发的各国国家统计机构。亚太地区的各国国家统计局有：孟加拉国家统计局、柬埔寨国家统计研究院、中国国家统计局、斐济国家统计局、印度统计和项目执行部、印度尼西亚国家中央统计局、吉尔吉斯共和国国家统计委员会、马来西亚国家统计部、蒙古国家统计办公室、巴基斯坦国家统计司、菲律宾国家统计办公室、泰国国家统计办公室和越南统计总办公室。我们感谢评阅人的许多有益的意见和建议。分析和解释中的差错只由本文作者负责。本文早先的一个版本曾于 2014 年 11 月 5-7 日在日本东京召开的第 27 届人口普查大会(ANCSDAAP)上宣读。

[2]美国明尼苏达州，明尼阿波利斯，明尼苏达人口中心

通讯作者
Robert McCaa, Minnesota Population Center, 50 Willey Hall, 226 19th Ave. S., Minneapolis, MN 55455, USA. Email: rmccaa@umn.edu

摘要

国际微观样本整合共享数据库 www.ipums.org/international 向全球免费散发经统一整合过的八十多个国家或地区的人口普查微观样本数据，但服务群体仅限于那些同意严守数据使用条款要求的研究人员和学生。目前数据库收纳了 270 多套样本数据，合计达 6 亿多条个人记录。随着更多国家与微观样本整合共享数据库的合作以及 2010 年普查周期样本整合的完成，每年将有 15-20 套新样本数据颁布。鉴于如此多的微观数据能如此容易获取，对数据质量的关注便是很自然的事了。本文主要探讨不同时间上初等教育完成状况间的一致性。通过对 13 个亚太国家各自多次普查微观数据间不同出生队列初等教育完成百分比的分析，我们发现中国、蒙古、越南和印度尼西亚四个国家的一致性很高，平均差异不到 0.5 个百分点，回归系数(b)在 0.93-1.07 之间，R^2 达 0.99。若将 13 个国家视为一个群体，它们的一致性差异较大；有些国家的一致性较差，与平均值的绝对差异可高达 16 个百分点。这 13 个国家的回归系数的变化范围从 0.62 到 1.44，R^2 在 0.65 到 0.99 之间。总体而言，普查初等教育的统计一致性是较高的。然而，为能专业化地使用统一整合的微观数据，研究人员需要谨慎地细读数据库中整合数据的文献以及原始数据的文献。那些目前尚未与我们合作的国家统计局以及那些尚未将 2010 年周期普查的微观数据委托给我们的国家统计局也请细读这些文献。

关键词

初等教育，统计一致性，国际微观样本整合共享数据库，人口普查样本，整合的微观数据，微观数据获取，中国，印度，越南，亚洲，太平洋地区，孟加拉国，柬埔寨，斐济，印度尼西亚，吉尔吉斯共和国，马来西亚，蒙古，巴基斯坦，菲律宾，泰国

引言

国际微观样本整合共享数据库迄今已有十五年的历史。它向亚太地区以及全球的研究人员散发来自80多个国家超过270套经整合过的普查微观样本数据。中国国家统计局（NBS）是主创成员之一，于2002年签署了合作项目备忘录。1982年的人口普查作为中国的第一个微观样本数据于2003年被整合到数据库中。几年后，1990年人口普查的百分之一的微观样本数据也被收纳。 整合过的高精度的2000年人口普查的微观样本数据计划于2016年散发。但截至发稿，2010年人口普查的样本数据尚未被数据库收入。

国际微观样本整合共享数据库中所有国家的微观数据都可免费获取。但它们不是“开放数据”或“所有人均可获取”的，而是仅限于同意严守数据用户使用条款许可的研究人员和决策者。目前，已批准的用户来自130多个国家，他们可获取6.15亿多条个人记录。这些记录所代表的人口总数约占世界人口的五分之四。随着2010年周期普查数据及原有积压但已授权明尼苏达州人口中心的其他国家的微观数据样本的整合工作的完毕，2020年时，数据库的容量很可能会翻倍。此外，若目前尚未做出决定与本项目合作的其他国家的国家统计局（NSO）参与本项目，这种扩展将会更大（表1）。附录1提供了国际微观样本整合共享数据库项目的更多信息。

由于国际微观样本整合共享数据库管理如此多国家如此大量的微观数据，数据质量问题就变得极为重要。通过对所选配对样本间的统计比较分析，本文拟着重讨论数据质量问题。我们对十三个亚太国家各自两个相邻普查样本间受教育程度状况，特别是接受过初等教育的状况，考查其统计一致性（明尼苏达人口中心 2014 年）。因为中国人口普查微观数据相当重要，我们将中国的分析向后扩展至其第三个样本，即 1982 年普查，向前拓展至 2010 年，并期望不远的将来，中国国家统计局可将其最新的 2010 年人口普查的样本授权于本数据库。本文分析中所用的 2010 年人口普查数据来自中国国家统计局网站上现存的汇总表格 (中国国家统计局 2012 年)。

概念问题

国际微观样本整合共享数据库散发的样本数据的质量问题备受研究人员和各国国家统计局的关注。 Baffour和Valente在最近的综述中将普查的质量界定为“适用性”，并认为数据质量应包含以下六要素或维度：相关性、准确性、及时性、可获得性、可解释性和一致性（2012：122）。虽然准确性和一致性也明显与本文相关，但我们所关心的是一致性。

本文仅分析质量的一个方面，即单个维度不同时间上的一致性[3]和单个指标--初等教育。我们的问题是，同一个国家，最近的人口普查样本中的初等教育完成比例（预测变量）与前一期普查中的同一比例（应变量）相比如何？换言之，对每两次普查而言，我们根据较近一次普查样本中的初等教育完成比例向后推算出前一期普查样本中的初等教育完成比例。问题不是最近的人口普查是否最为准确，而是两次普查中同一队列一致到什么程度？我们对初等教育进行考察不仅是因为普及初等教育是一个千年发展目标，也是因为大多数亚太国家的普查都搜集这一数据，而且在国际微观样本整合共享数据库的样本中，该指标也被广泛使用。

我们用人口学中出生队列的概念来估算每一个样本数据中15岁至89岁单个年龄上的一系列人口数，并对比各次普查样本间的数据。当两次普查数据之间存在一致性时，每一出生队列的初等教育完成比例应该相同或非常相似。除了均值和中位数的绝对差异外，我们还使用了相关系数R的平方（R^2）和最小二乘回归系数（b）来测量样本数据间的一致性程度。考虑到不同国家间的可比性，我们的分析仅限于55个出生队列。

---

[3] 经合组织提出的一致性的四个维度包括同一套数据中的一致性、不同数据间的一致性、不同时间上的一致性和不同国家间的一致性 (2011/1:10)。

表 1. 国际微观样本整合共享数据库**: 176 个国家/地区 2010 年普查周期微观数据 分年份的收录状况**

普查年份	国际微观样本整合共享数据库合作伙伴		C. 尚未成为合作伙伴 (人口在 250,000 人及以上的国家或地区)
	A. 2010 年普查周期微观数据已授权于数据库 **(粗体为已经散发)**	B. 2000 年或更早周期的数据已授权，但 2010 年周期数据尚未授权	
2005-9	**2005 喀麦隆,哥伦比亚, 尼加拉瓜,尼日利亚 (NSSO); 2006 布基纳法索, 埃及, 法国, 伊朗, 爱尔兰**, 莱索托; **2007 萨尔瓦多, *斐济群岛*, 巴勒斯坦, 秘鲁, 埃塞俄比亚, 莫桑比克; 2008 柬埔寨,** 以色列, **利比亚, 马拉维, 南苏丹, 苏丹; 2009** 白俄罗斯, **肯尼亚, 吉尔吉斯共和国, Mali**	**2009** 几内亚比索	**2005** 不丹, 科威特, 老挝, 阿联酋; **2006** 香港特别行政区, 利比亚, 澳门特别行政区, 马尔代夫, 尼日利亚 (NPC); **2007** 刚果共和国, 法属波利尼西亚, 斯威士兰; **2008** 阿尔及利亚, 布隆迪, 朝鲜; **2009** 阿塞拜疆, 乍得, 吉布提, 哈萨克斯坦, 新喀里多尼亚, 所罗门群岛
2010	**阿根廷, 巴西, 多米尼加共和国**, **厄瓜多尔, 加纳, 印度 (NSSO), 印度尼西亚, 墨西哥, 巴拿马, 波多黎各**, 特立尼達和多巴哥, **美国, 赞比亚**	佛得角, 中国, 韩国, 马来西亚, 蒙古, 菲律宾, 圣卢西亚, 瑞士, 泰国	巴哈马, 巴巴多斯, 伯利兹, 芬兰, 日本, 卡塔尔, 俄国, 沙特阿拉伯, 新加坡, 台湾, 塔吉克斯坦, 东帝汶, 多哥
2011	**奥地利, 爱美尼亚, 孟加拉**, 博茨瓦纳, 捷克共和国, **法国,** 希腊, 匈牙利, 伊朗, **爱尔兰,** 纳米比亚, **尼日利亚 (国家统计局)**, 波兰, **葡萄牙,** 罗马尼亚, **南非, 西班牙, 乌拉圭**	保加利亚, 加拿大, 哥斯达黎加, 德国, 意大利, 牙买加, 毛里求斯, 尼泊尔, 荷兰, 巴巴亚新几内亚, 斯洛伐克共和国, 斯洛文尼亚, 土耳其, 英国, 委内瑞拉	阿尔巴尼亚, 澳大利亚, 巴林, 比利时, 文莱, 克罗地亚, 塞浦路斯, 丹麦, 厄立特里亚, 爱沙尼亚, 冰岛, 印度 (ORG), 拉脱维亚, 立陶宛, 卢森堡, 马耳他, 黑山, 挪威, 瑞典
2012+		**2012** 玻利维亚, 智利, 古巴, 巴拉圭, 卢旺达, 坦桑尼亚, 土库曼斯坦; **2013** 贝宁, 几内亚-科纳克里, 洪都拉斯, 尼日尔, 塞内加尔 **2014+** 中非共和国, 象牙海岸, 瓜地马拉, 海地, 约旦, 马达加斯加, 摩洛哥, 巴基斯坦, 塞拉里昂, 突尼斯, 乌干达	**2012** 格鲁吉亚, 圭亚那, 马其顿共和国, 诺鲁, 新西兰, 斯里兰卡, 苏里南, 图瓦卢, 津巴布韦; **2013** 波黑, 喀麦隆, 加蓬, 冈比亚, 毛里塔尼亚, 圣多美和普林西比; **2014+** 安哥拉, 刚果民主共和国, 赤道几内亚, 摩尔多瓦共和国, 缅甸, 索马里

来源: http://www.hist.umn.edu/~rmccaa/IPUMSI/census_microdata_inventory.htm
普查日期: http://unstats.un.org/unsd/demographic/sources/census/censusdates.htm

分析的结果相当令人鼓舞。这些结果显示一致性程度较高，甚至非常高。我们将看到中国（R^2=.99, b=.93）、越南和印度尼西亚（R^2=.99, b=.95）2000年和2010年普查之间均值和中位数的差异均低于0.5个百分点。对中国（1990年、1982年与2010年比较）、蒙古和泰国更早先的普查而言，其一致性程度也类似。然而，所有13个国家各自相邻普查间一致性的变化幅度较大(R^2=.65-.99, b=.62-1.44)，表明各国样本数据之间的一致性存在较显著的差异。我们对非洲国家人口普查微观数据的研究显示，其分歧还要大，R^2从0.38到0.99，b从0.46到1.37（McCaa等，2015年）。

人口普查数据是各国花费巨大财力和物力搜集的，是制定公共政策的强有力依据。在社会科学中，它们是应用最广泛的数据源之一，并被政策制定者、研究人员、新闻记者、教师、学生和其他人群广泛使用。鉴于人口普查的社会投资和其广泛认可的效用，以使其潜能最大化的方式散发数据尤为重要。

联合国欧洲经济委员会专家组第十六次会议对人口和住房普查的一致性定义如下（联合国欧洲经济委员会 2014年：4，B.4.f节）：

> 一致性是指在不同时间上以及在一个宽泛的分析框架下人口普查信息可以与其他统计信息能成功地衔接的程度。使用可能已经被国际接受的标准概念、定义和分类有利于促进一致性。

Baffour和Valente（2012年：126）区别了两种类型的一致性：内部一致性（单个普查中不同结果间的一致性）和外部一致性（两次或以上人口普查之间比较的一致性）。为了实现统计一致性，定义、概念、框架和分类必须在国家和国际层面上做到明确和统一。若这些条目发生变化时，描述新旧条目之间相似性和差异的文字说明是必不可少的。 Baffour和Valente的结论是“理想地，(普查)问项应保持长期不变，以便进行纵向比较”，而且应该对数据中任何不寻常的趋势或不一致性加以解释。

对于2010年周期的人口普查，联合国统计司推荐“受教育程度”为一个核心主题，并在普查后的处理中使用1997年修订的《国际教育标准分类》(ISCED)，以方便国际比较(联合国统计司2008年：149-150)。《国际教育标准分类》涵盖初等教育，通常为4-7年完成，最常见的为六年完成(联合国教科文组织 2012年：17)。

数据和方法

人口普查反映了一个国家和其人民的人口发展史。各次高质量的全国性人口普查所揭示的结果应该相似或一致。历史人口学家（本文五位作者中的四位作者为历史人口学家）的分析工具包含队列内部比较法，即各次普查间可通过比较出生队列来测量某一统计量。

关于外部一致性，我们提出的一个简单问题是：对于每一个出生队列，最新样本中申报完成初等教育的比例与大约十年前样本中的比例类似吗？以越南为例，我们要问：比如出生于1970年的队列中，小学毕业的比例在2009年人口普查样本和1999年的人口普查样本中相同吗？事实上，答案是肯定的，几乎一模一样：根据2009年人口普查样本，越南出生于1970年的队列中，73.8％的人完成了初等或以上教育，与1999年样本中的72.5％相比，仅相差一个多百分点。

为进一步探讨这一问题，我们涵盖所有的出生队列：从人口普查开始前的15年（只有极少数人在15岁以后才小学毕业）一直向后延伸至队列的绝对频数小到不可靠为止，比如说89岁。如下面图1所示，越南1999年普查与2009年普查相比，我们发现b=0.93，尽管不是完美的1.0，但表明两者高度一致。R^2为0.99。

在评估外部一致性时，至少需要注意三个事项：普查机构的组织实施方式，国际微观样本整合共享数据库的规范统一性和偏差。首先，各次普查中设计的问项、定义和类别及现场

普查员的培训必须予以考虑，还要考虑国家权威普查机构对数据处理和编辑的方式。第二，鉴于我们分析所用的数据是经国际微观样本整合共享数据库整合过的，所以还必须考虑数据库团队对微观数据进行规范化的做法，以及对各次普查统一编码方案的决定正确与否。第三，本方法假定不同受教育程度人的死亡率、迁移率以及普查申报登记状况相同。本方法还假定没有成人教育运动项目而使超过常规年龄后的小学毕业比例增加。当受教育程度低的人具有较高的死亡风险时，将会出现系统性高估。同样，当一个国家迁入或迁出的可能性与受教育程度有关时，与普查质量无关的国际迁移将夸大普查间的不一致性。另外，还有源自受访者，特别是受教育程度较低人群的申报误差以及年龄尾数申报的偏好，比如0和5等。本方法的其它详细信息参见Feeney（2014年）。

整合受教育程度 - 国际微观样本整合共享数据库方法

研究人员和各国类似国家统计局机构的从国际微观样本整合共享数据库获得的最主要好处是，经统一规范化后的各国历时几十年的微观数据样本--通常是从各国现存的或经恢复过的最早的人口普查微观样本至2010年周期或更新的数据。当项目刚开始时，只有少数国家的国家统计局散发普查样本。而今，则是大多数国家都散发普查样本。然而，即使在今天，各国国家统计局鲜有发布文献以促进对两次或多次普查样本进行比较分析的情况；将早期普查重新整理生成交叉表以帮助研究人员统一规范化各次普查中的变量的情况甚至更少。大多数统计部门，人员编制严重不足，财力和人力资源也严重匮乏。各国国家统计局的一般做法是简单地随机抽取一个样本，匿名散发。通常情况下，很少出版如何对各次普查间的微观数据进行比较的指导性文献。

国际微观样本整合共享数据库系统地收集、归档和散发原始的源文件，包括普查表、现场调查指南、编码本、技术手册和官方出版物。我们使用这些所有一整套的文件，对这些高精度的微观数据中的每一个样本，每一个变量和每一编码一一进行整合。原来数据中的序列码被重新编为层次码或复合码以便于比较，但仍保留其在原始数据中的所含信息（Esteve and Sobek，2003年）。通过对综合性的原始源文件的细致研究，重新编写整合后的元数据的文档。数据库团队将新形成的数据库中的每个整合后的变量在元数据中归为六类（参见下面的图2中“标签”）：

1. 编码
2. 总体概述
3. 可比性讨论
4. 适用性问题
5. 概念的可用性
6. 原始文本的详细措辞（“问卷文本”链接到各国官方语言版和翻译的英文版的原始问卷），和
7. 链接到用于创建各个整合变量的“源变量”。

整合的基本目标是在不遗漏任何有用信息的原则下便利微观数据的使用。这是一项富于挑战性的任务，因为为了简化数据以利于不同时空上的比较分析，就有必要开发可比较的适用于所有普查样本的编码方案。微观数据整合后，每个样本中相同的概念（变量、类别）就有了相同的代码。为了避免遗漏那些更详细样本中的重要信息，我们使用了一个复合编码来保留所有的原始信息，并同时在不同样本间标明可比较的代码。通过复合码，研究人员可以很容易地进行时空对比，并甄别它们的差异。

第一个位数，被称为“总代码”，反映所有样本中共有的信息（最小公分母）。接下来的一位或两位码，表明样本中的额外信息。数据库中，相当多的样本数据均具有这一代码。尾部位数表明其他信息，但数据库中只有很少的样本数据才有这一代码。某一位数上为0，表示该样本数据中没有这项信息。

对于本文的一致性分析，我们关注数据库中使用最广泛的变量，即受教育程度（EDATTAN）。大多数搜集这一变量信息的人口普查微观数据均按照联合国教科文组织公布的《国际教育标准分类》方案（2012年）为依据，分为四个水平或阶段：被访人是否已经完成（a）从未上过学,（b）小学,（c）中学, 或（d）更高水平的教育。因此，国际微观样本整合共享数据库中的复合码包括四个类别（代码为1-4），再加数据缺失码（代码为9）和“不适用”码（代码为0-孩子太小而无法上学，或普查中那些没有被要求申报此问项的人）。

许多样本中含有诸如接受过初等、中等、甚至高等教育但并没有毕业等的更多信息。本码的第二位数反映了这一信息。第三位数区分了技校、正规教育和其他教育。成功的国际层面上的数据整合必须记录这种区别，以使研究人员可以很容易地了解这些以及其他成千上万的细节。

表2 列举了13个国家受教育程度变量的一般和详细的编码（由两位标准的国家或地区代码表示，ISO3166）。该表的上半部分显示，所有样本都具有四个一般水平：小学未毕业、小学毕业、中学毕业和大学毕业。表的下半部分则表明，普查中搜集的受教育程度的水平信息在不同国家和不同样本间千差万别。每个单元中的频数指的是不同样本数据中相应代码上未加权的样本人数。这些频数纯粹是描述性的。正如我们将在下一节中看到的，一致性可以通过用加权后的代码和出生年份生成的交叉表中的百分比来进行评估。

国际微观样本整合共享数据库整合元数据的目标是在条件允许前提下尽可能多地提供重要信息，便于人们点击网站获取数据，以推动微观数据的知情分析。需要注意的是元数据是免费获取。但微观数据的获取是受到限制的，以尊重所有合作的各国国家统计局同意的使用条件。整合过的微观数据是经过深度的测试和强化的。数据库团队花费了数千小时来分析、讨论、辩论、测试和再测试，直到整合的微观数据被验证可以向研究人员散发为止。因有新的样本被整合到数据库中，该过程每年都加以重复。

主要结果

越南：图 1 描绘了基于国际微观样本整合共享数据库中整合的越南 1989 年、1999 年和 2009 年普查微观数据得到的不同出生年份接受过初等教育的状况。这些入学率曲线揭示了各次普查间惊人的一致性。1999 年与 2009 年以及 1989 年与 2009 年的相关系数分别高达 0.93 和 0.92，接近于相关系数等于 1.0 时完全吻合的状态。兴许这些几乎完全一致的结果不值得奇怪，因为这些数据都是来自同一个统计部门。然而，这三次普查在分别间隔十年的三个不同时点上每次都动员了成千上万的普查登记员而搜集的。三次普查数据的处理和编码都使用了不断完善的高技术，从而避免了很多可能的错误。而且，这些数据是基于由国际微观样本整合共享数据库团队整合过的变量而得到的。这种整合是前所未有的。不管怎么说，图 1 中数据的吻合度很高。研究人员应该对越南相邻普查的微观样本数据间的高度吻合度感到满意。

值得一提的是外部一致性，这是因为尽管统计总办公室在三次普查中使用了统一的面对面实地访谈，但因样本的地区个数发生了变化从而导致了抽样方式的不同。样本的地区个数从1989年的80个上升到1999年的122个和2009年的几百个。抽样的密度也有很大变化，从1989年的5%缩小到1999年的3%，再重新扩大到2009年的15%[4]。

[4]https://international.ipums.org/international/sample_designs/sample_designs_vn.shtml

表 2. EDATTAN (受教育程度): 国际微观样本整合共享数据库中 13 个国家经过统一化总码和细目码

各单元值是指相应代码在经统一化后的最新微观数据中的未加权频数

	国家 代码 (ISO 3166)	BD	KH	CN	FJ	IN	ID	IR	KG	MY	MN	PH	TH	VN
	普查年份	2011	2008	1990	2007	2004-5	2010	2006	2009	2000	2000	2000	2000	2009
代码	标签													
总码														
0	NIU (不适用)	1,117,354	136,274	1,418,185	·	·	2,253,453	131,235	72,044	·	35,396	935,577	43,640	1,517,591
1	小学未毕业	3,216,705	766,314	4,383,067	24,403	316,386	6,117,917	195,404	124,184	223,334	84,105	2,132,120	284,685	4,675,806
2	小学毕业	2,065,976	376,009	5,069,640	40,755	172,721	10,135,303	467,961	66,697	166,637	54,743	1,967,457	179,347	6,140,145
3	中学毕业	639,020	47,837	915,562	17,684	85,023	4,394,068	195,055	251,330	10,486	55,050	1,689,518	69,705	1,316,274
4	大学毕业	166,665	13,010	49,493	1,468	28,290	702,308	59,970	48,606	25,456	14,431	305,054	20,933	527,774
9	未知	·	677	·	13	413	·	250,200	2,125	9,387	·	388,084	6,209	·
细目码														
0	NIU (不适用)	1,117,354	136,274	1,418,185	·	·	2,253,453	131,235	72,044	·	35,396	935,577	43,640	1,517,591
100	小学未毕业		·	·	·	·	·	·	56,168	·	40,263	·	·	·
110	未上过学	1,961,034	297,550	2,145,035	10,890	220,227	1,986,754	41,776	·	100,909	·	466,783	55,479	892,633
120	上过一些学	1,255,671	468,764	2,238,032	13,513	96,159	4,131,163	153,628	·	122,425	·	1,665,337	229,206	3,783,173
130	小学(4 年)	·	·	·	·	·	·	·	68,016	·	43,842	·	·	·
	小学毕业但中学没有毕业													
	小学毕业													
211	小学(5 年)	1,256,266	·	·	·	88,352	·	233,865	·	·	·	·	·	·
212	小学 (6 年)	·	256,570	2,822,479	24,932	·	6,539,863	·	·	80,005	·	1,967,457	116,450	2,671,203
	中学未毕业													
221	正规或非特殊教育	809,710	119,439	2,247,161	15,823	84,369	3,595,440	234,096	46,187	86,632	47,742	·	62,897	3,468,942
222	技校	·	·	·	·	·	·	·	20,510	·	7,001	·	·	·
	中学毕业													
	正规或非特殊教育													
311	正规教育毕业	639,020	41,385	640,916	10,489	49,669	3,592,138	127,008	208,581	8,878	40,677	814,182	24,371	1,074,774
312	接受过一点大学教育	·	·	43,450	380	29,237	·	29,805	15,106	·	·	715,722	17,237	151,141
320	技校教育	·	·	·	·	·	·	·	·	·	14,373	·	·	·
321	中专		2,228	148,554	·	·	400,543	38,242	27,643	·	·	·	13,106	90,359
322	大专		4,224	82,642	6,815	6,117	401,387	·	·	1,608	·	159,614	14,991	·
400	大学毕业（本科毕业）	166,665	13,010	49,493	1,468	28,290	702,308	59,970	48,606	25,456	14,431	305,054	20,933	527,774
999	未知/缺失	·	677	·	13	413	·	250,200	2,125	9,387	·	388,084	6,209	·

来源: https://international.ipums.org/international-action/variables/EDATTAN#codes_section

注: “IN 2004-5”是指印度全国抽样调查机构 调查表 10 的样本数; 其他全部为普查样本。

下文讨论前面提及的三个方面：普查机构的组织实施方式、微观样本整合共享数据库的规范统一性和偏差。

第一，正如表3所示，总体而言，越南统计总办公室设计的三次普查中的受教育程度问项是一致的。除1989年只要求填写接受教育的年数外，其余两次普查均要求同时提供接受教育的年数和水平。关于是否上过学，1989年和1999年的普查表提供了三个选项：是否现在在上学、是否曾经上过学和未上过学。这种区分在2009年的普查表中删去了，因为我们只考察14岁以上人口中接受初等教育的情况，这一删除对我们的分析几乎没有影响。

第二，国际微观样本整合共享数据库对受教育程度的统一规范化有两个方案，分别为国际层面(EDATTAN)和国家层面(越南为EDUCVN)。对于国际层面记录，只需9个码，而越南的EDUCVN需要95个码。在微观样本整合共享数据库的元数据可比性讨论中，需要用500多字来对EDUCVN加以说明。但下面两句话足以概括这两个变量之间的不同。微观样本整合共享数据库元数据中这样写道：

> 受教育程度国际规范统一化的划分反映在EDATTAN上，它附加了一些条件以对各个国家的数据进行规范化处理。相反，EDUCVN保留了原有越南数据中所有有关受教育程度的详细信息。[5]

在国际层面的重新编码中，只要允许，我们对数据库中的所有样本数据都予以了规定，需要接受过六年的教育才算完成初等教育。值得注意的是，国际微观样本整合共享数据库系统为研究人员利用自己的标准进行重新编码，或者将国际微观样本整合共享数据库整合过的变量复原，检查数据库在规范化过程中各种结果之间的一致性和准确性提供了机会。图2中的 “Source Variable”(原始变量)键就可起到这个作用。针对本研究，图2解释了初等教育完成的比例是基于每次普查的受教育年数问项中至少接受过6年或以上的编码而估算的。请注意，我们将这一标准应用于数据库中的全部样本数据，即使是越南的国家教育体系规定只需要接受五年教育即可达到小学毕业。

若仅对单个国家感兴趣，研究人员可能倾向于使用(下载)诸如 EDUCVN、 EDUCBD、EDUCCN 等国家层面的编码变量。但对各国之间的差异感兴趣时，研究者人员可选择国际层面的变量 EDATTAN。

在评估外在一致性时，第三个注意事项是源于假定的迁移、死亡、成人教育或申报等引起的偏差。若1989年和1999年两次普查中各个年龄上的比例存在非常小的系统性低估，可能说明存在成人教育运动，或者说明教育程度较低的人比教育程度较高的人具有较高的死亡率或迁出率从而导致小学毕业的比例在相邻普查间升高。在申报过去较久远的事件时，也可能存在高估的偏差，尽管在现在的情况下，这种事件的申报偏差看起来非常小。

[5] https://international.ipums.org/international-action/variables/EDUCVN#comparability_section

表**3. 受教育程度问项内容在越南1989年、1999年 和2009年三次 普查中的差异较大，但总体上仍具可比性**

2009

13. What is the highest grade of education/training [NAME] is attending or has attained?

ABBREVIATION:

TRADE VOC. SCHOOL - TRADE VOCATIONAL SCHOOL

VOC. SCHOOL - VOCATIONAL SCHOOL

PRE-SCHOOL .....00 ☐ Q16

PRIMARY.....01 ☐

LOWER SECONDARY.....02 ☐

SHORT TERM TRAINING....03 ☐

HIGHER SECONDARY.....04 ☐

TRADE VOC. SCHOOL.....05 ☐

VOC. SCHOOL .....06 ☐

TRADE COLLEGE.....07 ☐

COLLEGE .....08 ☐

UNIVERSITY.....09 ☐

MASTER .....10 ☐

DOCTOR.....11 ☐

14. What is the highest grade/year of education/training [NAME] is attending or has completed at the above-mentioned grade?

GRADE/YEAR ☐☐

1999

10. Is (Name) attending now, stopped or never attending school?

• *School:* Only record persons currently attending/attended general schools (or equivalent) and higher educational schools.

GENERAL SCHOOL.....1

ATTENDED IN THE PAST.....2

NOT STATED NEVER ATTENDED.....3 → (Q.12)

11*. What is the highest educational level that (NAME) is attending or completed?

• Persons who completed/ attending secondary vocational schools or lower, record his/her general education grades.

GENERAL SCHOOL 1

GRADE SYSTEM _____ ☐☐

UNDER GRADUATE 2

GRADUATE 3

POST GRADUATE 4

CHECK Q.11:

UNDER GRADE 5 ____

GRADE 5 OR HIGHER _

1989

8-a/School attendance or equivalent

Attending now — 1

Attended in the past — 2

Never attended — 3

b/Highest grade completed

Grade . . . . . — —

FOR PERSONS BORN ON OR BEFORE 1-4-1984 (AGED 13 AND OVER) ANSWER F

9- a/ Highest qualification or trade

None — 1

Technical worker with certificate — 2

Technical worker no certificate — 3

Middle vocational education — 4

College / university degree — 5

Post-graduate — 6

b/ Field of study

. . . . . — — —

来源: https://international.ipums.org/international/enum_materials.shtml

注: 在上述来源网页中的显示的所有问项均包括越南官方语言版和由国际微观样本整合共享数据库翻译的非官方英文版。

图 1 越南 1989 年、1999 年和 2009 年三次人口普查样本中的初等教育完成(EDATTAN)高度的统计一致性

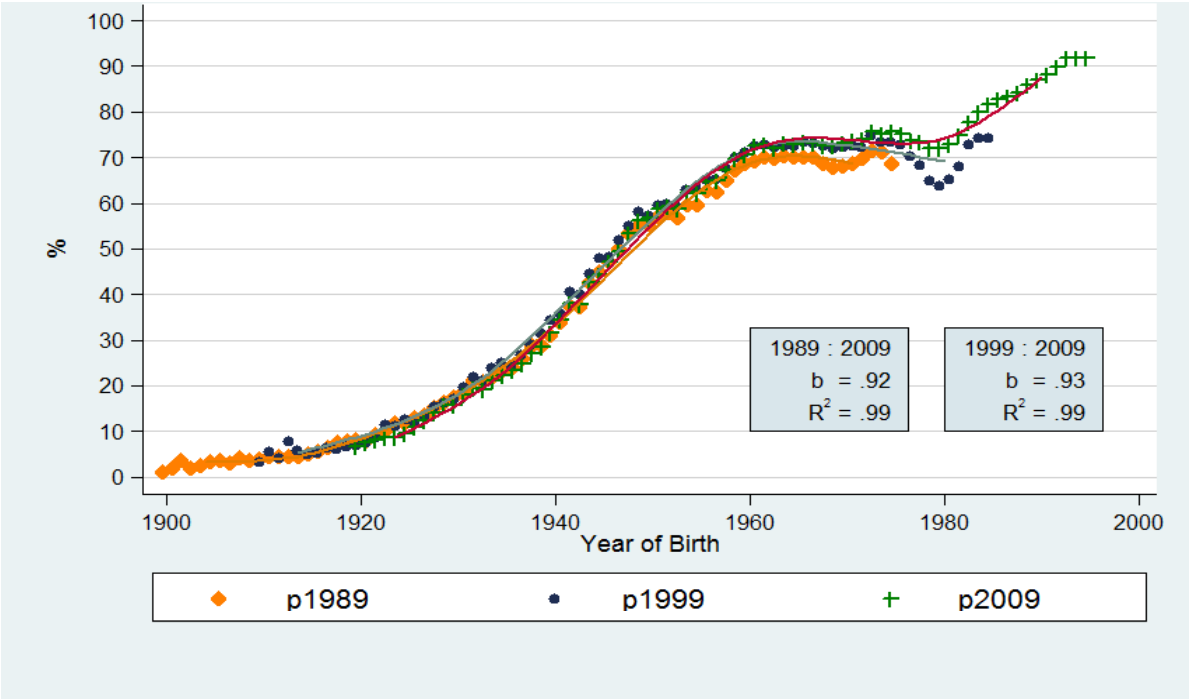
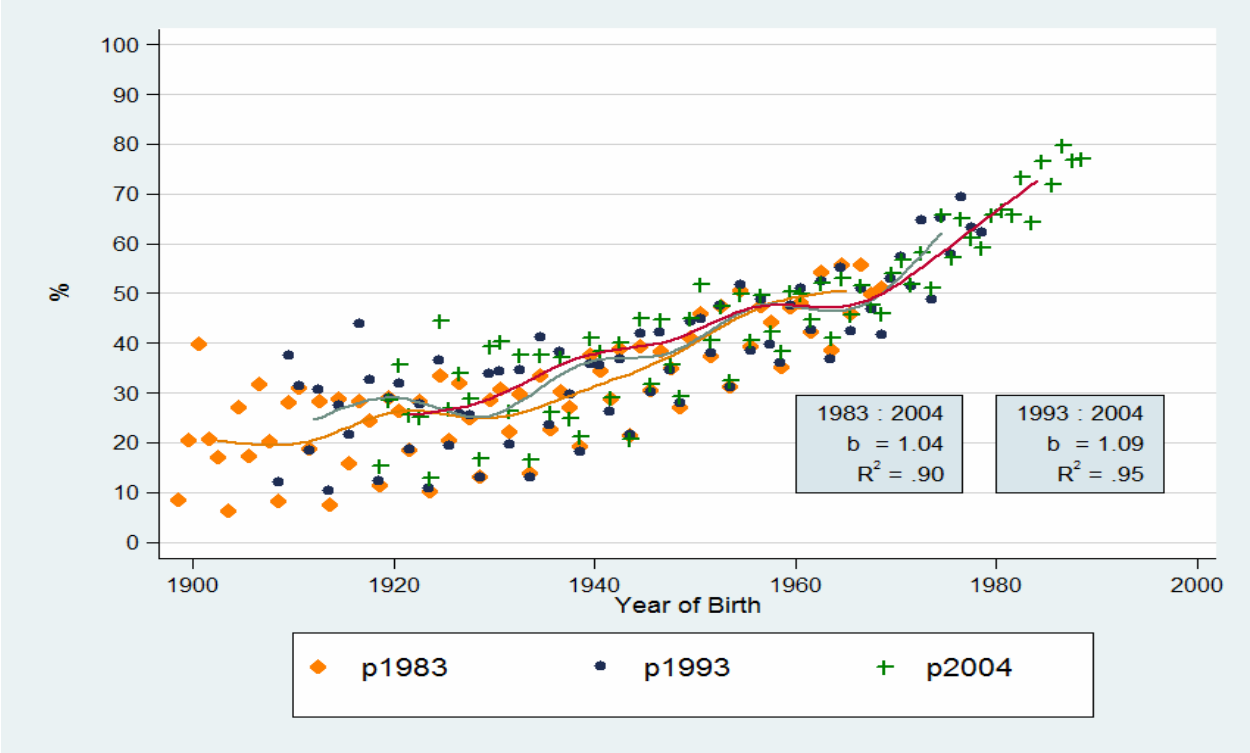


图 2 国际微观样本整合共享数据库元数据屏幕截图：受教育程度--国际编码可比性讨论，越南篇

印度: 我们用印度作为第二个检验。印度的微观数据不是来自人口普查而来自由统计和规划实施部出资，全国抽样调查机构负责组织的全国性的抽样调查样本(“社会经济调查，户调查表10”)。虽然调查表10的调查每五年进行一次，且所有各次的数据均由国际微观样本整合共享数据库散发，但本文只检查1983年(整个完整日历年)、1993年（1993年7月至1994年 6月）和2004年（2004年7月至2005年6月）三套数据(参见图3)。

图 3. 印度全国抽样调查机构组织的三个周期的调查（2004-5 年、1993-4 年和 1983 年）显示，尽管年龄堆积严重但统计一致性良好



印度全国抽样调查机构在其二十多年来的问卷中始终对初等教育完成使用统一的定义，但调查问卷中没有受教育年限的这一问项 。因此EDATTAN变量的编码与印度国家的惯例保持一致：小学5年完成，初中8年完成，高中10年完成。总体而言，尽管有强烈的年龄尾数申报偏好，汇总统计中显示，初等教育在各次调查间的一致性较高（R^2=0.95, b=1.09，平均离差为0.7个百分点）。

表4提供了这些调查以及其它国家样本数据间有关一致性评估的更多细节。大数法则可能使人们认为样本规模越大，统计一致性越高，但事实并非如此。中国和巴基斯坦同为大样本，但只有中国的样本显示出高度的统计一致性。蒙古和泰国同为微观数据库中的小样本数据。他们的样本规模虽很"微小"，但其统计的一致性却较高。

诸如在印度全国抽样调查机构的调查样本以及孟加拉国和巴基斯坦的人口普查样本中，未接受教育人群的年龄申报存在的强烈数字偏好歪曲了不同时间上的比较。表4中的惠普尔年龄堆积指数表明，印度全国抽样调查机构的抽样样本中的年龄申报质量是“非常低劣的”。尽管如此，我们发现，1970年的出生队列在 2004-5年样本中申报小学毕业的比例为53.9%，与他们在1993-4年样本中申报的58.2%只差4.2个百分点。此外，两次调查中共有的55个队列的平均差异只有0.7个百分点；其中位数的差异更小，只有0.4个百分点。

表 4. 13个亚太国家微观样本中初等教育的统计一致性

国家	数据年份	样本规模 (*10^6)	惠普尔指数及数值	SCHOOL 状态	YRSCHL 年数	EDATTAN 水平	1970 年出生队列小学毕业比例 (%)	所有 55 个出生队列 均数	差异 均数	差异 中位数	R^2	b
				类别				%				
孟加拉	2011- 5%	7.2	262-e	2	14	7	41.4	33.8	-3.1	-2.6	0.93	0.93
	2001-10%	12.4	300-e	3	15	10	42.8	36.9				
柬埔寨	2008-10%	0.3	110-b	3	16	10	44.6	28.1	5.4	5.3	0.97	0.93
	1998-10%	0.2	118-b	3	16	8	40.5	22.7				
中国	2000- 1%	11.8	100-a	5	-	10	93.9	53.9	0.2	0.0	0.99	0.99
	1990- 1%	11.8	101-a	4	-	7	93.6	54.0	-3.5	-3.4	0.99	0.96
	1982- 1%	10.0	102-a	-	-	8*	91.1	48.7				
斐济群岛	2007-10%	0.1	105-b	4	15	9	96.6	80.1	6.9	2.9	0.94	1.44
	1996-10%	0.1	102-a	3	15	9	95.6	73.2				
印度	2004/5-0.1%	0.6	193-e	5	-	9	53.9	42.7	0.7	0.4	0.95	1.09
	1993/4-0.1%	0.6	221-e	3	-	8	58.2	42.0				
印度尼西亚	2010-10%	23,6	114-c	4	-	9	86.7	65.2	0.5	0.1	0.99	0.93
	2000-10%	20.1	152-d	-	-	8	84.0	65.8				
吉尔吉斯共和国	2009-10%	1.1	100-a	2	-	10	99.1	82.9	1.9	0.9	0.98	0.98
	1999-10%	0.5	99-a	4	-	9	99.1	81.1				
马来西亚	2000- 2%	0.4	115-c	3	-	8	76.1	35.4	-16.1	-13.7	0.93	1.02
	1991- 2%	0.3	114-c	4	15	7	89.2	51.5				
蒙古	2000-10%	0.2	99-a	3	-	8	94.8	59.1	0.0	0.1	0.99	0.98
	1989-10%	0.2	100-a	-	-	8	92.3	59.1				
巴基斯坦	1998-10%	13.1	187-e	-	-	9	40.4	22.0	-2.7	-3.6	0.65	0.62
	1981-10%	8.4	264-e	-	-	8	32.6	24.7				
菲律宾	2000-10%	7.4	110-b	3	17	9	85.4	65.4	2.3	1.7	0.99	1.04
	1990-10%	6.0	111-b	3	18	9	84.4	63.0				
泰国	2000- 1%	0.6	110-b	4	24	11	82.4	26.8	0.9	0.7	0.99	1.00
	1990- 1%	0.5	105-b	4	24	12	82.3	25.9				
越南	2009-15%	14.2	101-a	5	23	9	73.8	50.8	0.2	-0.4	0.99	0.93
	1999- 3%	2.4	100-a	5	19	9	72.5	50.6				

来源: www.ipums.org/international *数据没有区分教育水平的开始或完成。

注: 惠普尔指数: a. 数据申报质量很好(<105), b. 质量较好 (105-109.9), c. 一般 (110-124.9), d.质量较差 (125-174.9) , e. 质量很差 (>=175). 作者自己的计算值。表中有关受教育程度的三个变量均由国际微观样本整合共享数据库整合后创建的。

中国：尽管1982年的人口普查并没有严格沿用受教育程度的国际标准定义，1982年、1990年和2000年中国的三次人口普查的微观数据明显地表现出近乎完美的一致性。 1982年的人口普查没有区分毕业者与未毕业者（即那些接受过某一特定教育水平却未能完成的人）。因此，1982年微观数据也未作这种区分。尽管如此，我们可通过按1982年的界定对1990年和2000年微观数据中受教育程度重新分组制表(见表4)。研究结果证实了三个样本中的统计一致性很高。

将1990年的数据与2000年相比，相关系数和回归系数都近乎完美，为0.99，平均差异为0.2个百分点。中位数的差异完全为零。这是本文所有国家数据中统计一致性最好的案例。1982年与2000年数据间的一致性虽不完美但仍较高，回归系数“只有”0.96，平均差值比1990年与2000年数据间的差异大近30倍。但即使是这样，1982年与2000年数据间的平均差异也仅有-3.5个百分点。

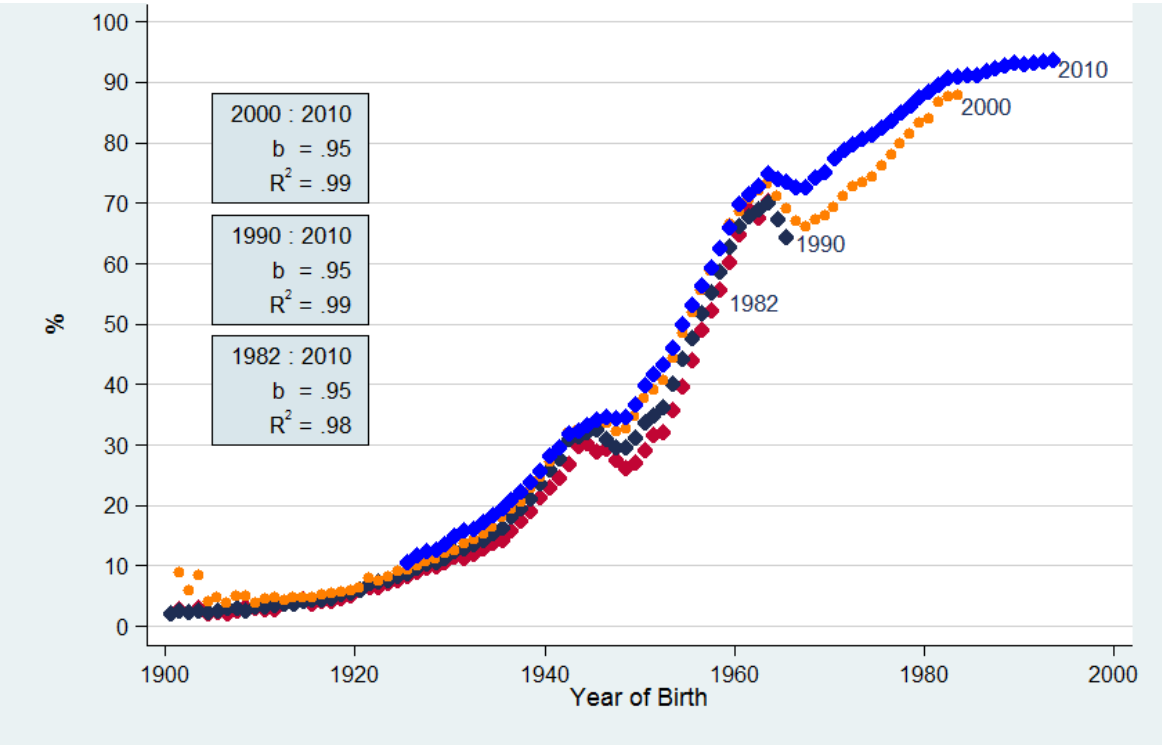
中国国家统计局在所有后续的人口普查中严格沿用国际标准的受教育程度定义的做法值得称道。 2000年和2010年的人口普查中，与教育有关的三个问项（水平、状态和成年人继续教育）是最为详细和最新的国际最佳实践。相比之下，1982年人口普查只包括受教育水平一个问项，未对毕业和未毕业作出区分，从而其提供的信息量很少。

因缺乏2010年人口普查的微观数据，统计一致性可以通过采用一个略微不同的概念，即不用“毕业”，而用“高于毕业”，再与1982年至2000年间的各次微观数据进行比较。换言之，我们用“小学毕业后的其他教育”，而不用 “小学毕业”。这个概念变得不单纯指完成小学教育，而实际是指初等教育毕业后进入更高层次的学校教育。如果有2010年人口普查的微观数据，我们就有可能对各水平的教育程度（包括小学毕业）统一代码。由于表格采用固定的类别和定义，只用表格数据，这种一致性是不容易实现的。对于2010年的人口普查中某些受教育程度的水平（比如高于小学毕业)，可以用从中国国家统计局的网站上下载的受教育程度的表格，通过重新划分使其与早期普查微观数据样本中的定义相一致。

如图4所示，中国1982年至2010年的人口普查间，“高于小学毕业”的统计一致性很高。以2010年的人口普查为准绳，1982年、1990年和2000年的回归系数几乎均为0.95，非常接近表4中的0.99和理想的1.0。相关系数的平方也几近完美，1982年为0.98，1990年和2000年均为0.99。对于百分率的比较，为确保1982年人口普查前有足够多的年份，我们考察1960年出生队列。研究发现，1982年至2010年四次普查的百分率分别为64.8％，66.1％，68.6％和69.8％。1925年至1963年39个出生队列的总体平均比例，从1982年人口普查中的30.0％上升为1990年的32.2％，2000年的34.4％和2010年的35.7％。这一增长与更高学历人群的预期寿命较高相一致，但也与在更大年龄阶段接受更多的教育(如成人教育)相一致。年轻队列中的差异比年长队列中的差异更显著可能归因于成人教育。从1978至1998年，成人高等教育的招生数每年增长3.5％；1999后至2010年，每年的增长率猛升到10.4％（Lai 2014年：62）。我们相信， 2010年人口普查微观数据中的结果比仅基于表格数据的比较更具一致性，因为可通过用有“仪式”象征意义的变量，比如“小学毕业”，而非“高于小学毕业”这样奇怪、不直观的概念进行比较。

其他国家的比较：表4总结了对目前由国际微观样本整合共享数据库散发的十三个亚太国家的样本分析。中国、蒙古和泰国三个国家的一致性几近完美。这三国各自的平均差异小于一个百分点，b与1的差异小于+/-0.2个百分点。孟加拉国、柬埔寨、印度、印度尼西亚、吉尔吉斯共和国、菲律宾和越南为第二组，其平均差异稍大且回归系数略低。第三组国家的平均差异较大，从六个百分点到十六个百分点，但回归系数差异较小。巴基斯坦为最后一组。因为在整体人口结构中存在非常严重的年龄性别堆积现象，巴基斯坦样本数据中的一致性很异常：一方面平均绝对差相对小，为-2.7个百分点；另一方面，回归系数和相关系数都在0.7以下。

图 4. 中国 1982 年、1990 年、2000 年和 2010 年四次普查间较高的统计一致性



讨论

相同队列内部比较是测量不同时间上普查样本数据间一致性的一个较强的统计检验方法。然而，除非有普查个体数据且这些数据已经整合过了，否则这种方法因其应用上的困难而较少被使用。一旦普查个体数据被整合成单个数据库，比如像国际微观样本整合共享数据库，这种方法就很容易被应用于那些有“仪式”象征意义的变量，诸如受教育程度、已婚与否和曾生子女数等。

研究人员应该明白国际微观样本整合共享数据库对数据的整合是普查事后整合。各国国家统计局(微观数据的所有者)并不对国际微观样本整合共享数据库团队的决定以及整合和统一的数据编码设计负责。相反, Eurostat (欧洲统计数据库)的普查数据发布平台是由欧洲各国的统计办公室在2010年普查周期之前建立和运作的，所以整合是在实际普查实施之前完成的。[6]

国际微观样本整合共享数据库团队面临的挑战是当各国的国家统计局不能提供变量界定时如何处理每一次普查中的各种数据。感谢联合国统计司出版的、被各国广泛采用的《人口与住房普查原则和建议》，使普查编码的统一化或多或少有了可能。

明尼苏达人口中心感谢亚太地区各国的国家统计局签署了国际微观样本整合共享数据库合作备忘录并将高准确的普查微观样本数据授权给中心予以散发。普查微观数据为各国事务繁重

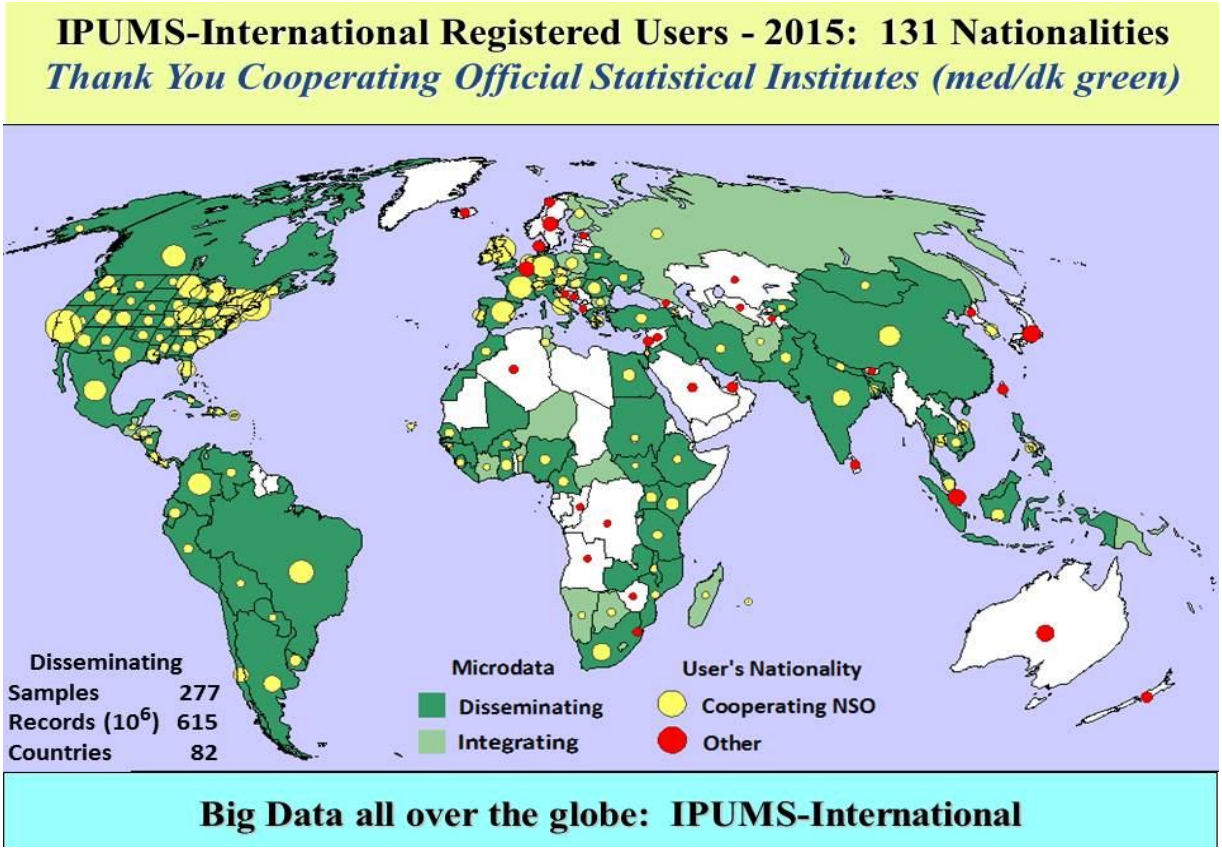
[6] https://ec.europa.eu/CensusHub2

的国家统计局和对这些专业化信息知之甚少的广大民众提出了挑战。与共享微观样本整合数据库合作从而推广整合过的国际普查微观数据具有成本低和风险小的较大优势。各国国家统计局也可从很多繁重的样本数据匿名化和文档化处理事务及责任中解脱出来。仅仅为了用户的较小回报，基于某种特别基础上发布普查微观数据，单个国家统计办公室会遇到较大的风险并付出巨大的人力资源成本。共享微观样本整合数据库项目在统一的处理方法、严格的安全保障以及较高数据质量和一致性下，为匿名化、整合和管理一系列普查微观数据提供了重要的规模经济效益。

研究人员对全球性整合微观数据的响应可用图5 表示。该图展示了各个国家注册的用户数量(在美国为各州的注册数）。目前国际微观样本整合共享数据库已有超过1万名注册用户，分别来自130个国家或地区的2000多个机构，包括各国国家统计局、各类高校、联合国各大机构和各研究中心 (联合国人口司、世界银行、世卫组织、经合组织，日本国家政策研究所等）。引用国际微观样本整合共享数据库的数据出版物文献已经超过1000个(https://bibliography.ipums.org)。在亚洲各国中，中国居首位，共引用64次。这是很突出的，因为数据库中最新整合过的中国微观数据是四分之一个世纪之前的历史数据。印度排在亚洲各国的第二位，共有58次。其次为越南49次和菲律宾36次。柬埔寨、印尼、马来西亚和泰国四国各为大约24次。

由于联合国统计司和各国国家统计局之间合作的不断加强，2020年周期的人口普查中的统计一致性可能会比2010年周期的人口普查中所反映的结果更好。

图 5 根据国籍和合作状态分的国际微观样本整合共享数据库注册的研究人员分布



参考文献

Baffou, B. and P. Valente. 2012. "An evaluation of census quality."*Statistical Journal of the IAOS* 28:121-135. DOI 10.3233/SJI-2012-0752.

Esteve, A. and M. Sobek.2003."Challenges and methods of international census harmonization."*Historical Methods* 36: 66-79.

Feeney, G. 2014. "Literacy and Gender: Development Success Stories." *Population and Development Review 40*:545–552. DOI 10.1111/j.1728-4457.2014.00697.

Lai, Qing. 2014. "Chinese Adulthood Higher Education: Life-Course Dynamics under State Socialism." *Chinese Sociological Review* 46:55-79.

McCaa, R. 2013. "The Big Data Revolution: IPUMS-International. Trans-Border Access to Decades of Census Microdata Samples for Three-fourths of the World and more."*Revista de Demografía Histórica30*: 69-87.

McCaa, R., L. Cleveland, P. Kelly-Hall, S. Ruggles and M. Sobek. 2015. "Statistical coherence of primary schooling in population census microdata: IPUMS-International integrated samples compared for fifteen African countries." *African Population Studies* 29(1):157-180.

McCaa, R. and A. Esteve. 2006. "IPUMS-Europe: Confidentiality measures for licensing and disseminating restricted access census microdata extracts to academic users."*Monographs of official statistics: Work session on statistical data confidentiality*. Luxembourg: Office for Official Publications of the European Communities, 37-46.

Minnesota Population Center. 2014. *Integrated Public Use Microdata Series, International: Version 6.3* [Machine-readable database]. Minneapolis: University of Minnesota.

National Bureau of Statistics. 2012. *2010 Population Census of China: Tables*, educational level by age and sex. http://www.stats.gov.cn/tjsj/pcsj/rkpc/6rp/indexch.htm (accessed March 26, 2015).

Organization of Economic Cooperation and Development (OECD). 2011. *Quality Framework and Guidelines for OECD Statistical Activities*. Version 2011/1

Ruggles, S. 2006. "The Minnesota Population Center data integration projects: Challenges of harmonizing census microdata across time and place."*Proceedings of the American Statistical Association, Government Statistics Section*. Alexandria, VA: American Statistical Association, 1405-1415.

Sobek, M and S. Kennedy.2009 "The development of family interrelationship variables for international census data."Minnesota Population Center.https://international.ipums.org/international/resources/misc_docs/pointer_working_paper_2009.pdf .

United Nations European Economic Commission (UNECE). 2014. Group of Experts on Population and Housing Censuses. "Quality Management" – Draft Text for the *Conference of European Statisticians Recommendations for the 2020 Census Round*. Geneva, Sep. 23-26.

United Nations Department of Economic and Social Affairs, Statistics Division (UNSD). 2008. *Principles and Recommendations for Population and Housing Censuses, Revision 2*. Statistical papers Series M. No. 67/Revision 2, New York.

UNESCO Institute for Statistics. 2012. *International Standard Classification for Education ISCED 2011*. Montreal.

附录 1：国际微观样本整合共享数据库的追加价值 www.ipums.org/international

微观样本整合共享数据库为散发普查微观数据提供了一种手段，作为各国国家统计局发布普查数据的补充。各国国家统计局向广大公众(居民、官方机构、媒体、分析人员等）发布官方统计数据和官方统计数据产品。而国际微观样本整合共享数据库则向一小部分但重要的群体，诸如本刊的读者等那些需要详细的个体和户的数据来度量和分析复杂关系且经常对各国和不同时间进行比较的研究人员在限制性访问基础上散发普查微观数据。

微观样本整合共享数据库从不散发原始和未经加工过的文档。相反，散发的那些微观数据都是经过处理的、统一化的和整合过的。这样，任何一个单一概念，比如小学毕业数，在本数据库的所有各国的普查微观样本数据中的编码都是一致的(参见上文的"受教育程度的整合"一节)。微观样本整合共享数据库也不是散发数据中的所有信息。相反，每一个研究人员根据自己的研究需要通过网上的菜单提取国家、普查年份、子人口和变量。每一次提取都形成一个独立的汇集数据。这个汇集数据将被登记以便于他人重复和防止欺骗。这种方法为用户确保微观数据的安全和遵守用户使用条例提供了很强的利益导向。由于完整的数据并不通过DVD或其他媒介散发，极大地降低了与未授权个人共享微观数据的可能。

微观样本整合共享数据库团队具有数十年使用微观数据的经验，研发了30多个追加变量，从而增扩了每一个样本数据。这些增加的变量可以归为三组：技术性、汇总性和指示性。

- **技术性变量**：包括类别、国家、年份、微观样本整合共享数据库数据身份码、户码、每户人数、户的权数、子样本数、集体户状态、各大洲洲码、国家所在地区、省份(州)和扩展因子（即样本数据中的个体和户的权数)。

- **户和家庭的汇总变量**：户类别、每户的家庭数、每户的已婚夫妇对数、每户的母亲数、每户的父亲数、户主在每户中的数据位置、每户的非家庭成员数、家庭单元隶属关系、每户中的自家人口数、每户中的亲生子女数、每户中不到 5 岁的亲生子女数、每户中最年长亲生子女的年龄和每户中最年幼亲生子女的年龄。

- **指示性变量**用来甄别同住的配偶、子女和父母。每户中母亲、父亲和配偶的数据位置、连接父母和配偶的规则、可能的继母、可能的继父、拥有两个或以上妻子的男子以及他们的这些妻子，等等。这类变量是整合样本数据中最具价值的追加内容之一， 因为它有利于根据母亲或父亲的特征来对儿童进行分析，或根据儿童特征对父母, 以及根据配偶其中一方的特征来对另一方进行分析 (Sobek and Kennedy 2009)。由于微观样本整合共享数据库数据中每户数据已经将母亲和她们的同住子女相联接，并且微观样本整合共享数据库数据的提取系统中的“追加特征”功能能将母亲的特征添加到每一个子女的记录中，因此用亲生子女法计算生育率变得容易了。