相同出生队列小学及以上人口比例统计的一致性*

——基于亚太十三国人口普查数据的分析

Robert McCaa　Lara Cleveland　Patricia Kelly Hall　Steven Ruggles　Matthew Sobek

【摘　要】文章根据国际微观数据系列整合共享数据库中的 13 个亚太国家各自多次人口普查微观数据，检测了不同出生队列的小学毕业及以上受教育程度的人口比例在历次普查之间的一致性。结果发现，中国、越南、蒙古和印度尼西亚 4 个国家的一致性很高，平均差异不到 0.5 个百分点，回归系数为 0.93～1.07，R^2 高达 0.99。但是，另外一些国家的一致性较差，有的国家与平均值的绝对差异高达 16 个百分点。这 13 个亚太国家的回归系数的变化范围为 0.62～1.44，R^2 为 0.65～0.99。总体而言，这 13 个亚太国家小学及以上受教育程度的人口比例在各自历次普查中的统计一致性较高。最后文章就如何专业化地使用数据库中统一整合过的微观数据提出了一些建议。

【关键词】初等教育　统计一致性　人口普查样本　微观数据

【作　者】Robert McCaa　美国明尼苏达大学人口中心，教授；Lara Cleveland　美国明尼苏达大学人口中心项目主任、博士；Patricia Kelly Hall　美国明尼苏达大学人口中心，研究助理；Steven Ruggles　美国明尼苏达大学人口中心主任、教授；Matthew Sobek　美国明尼苏达大学人口中心，首席科学家。

一、引　言

鉴于人口普查数据在社会经济发展应用中的巨大效用，其质量自然至关重要。一个质量较高的数据应包含相关性、准确性、及时性、可获得性、可解释性和一致性 6 个方面（Baffour 等，2012），人口普查也不例外。所谓人口普查数据的统计一致性，通常是指同一普查中不同相关问项的数值之间符合一定的逻辑关系，即同一出生队列在某些固定特征（如性别、成年

* 本研究为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）（编号：HD047283）的阶段性成果。

人口在儿时的一些特征）上的登记数，或在相同出生队列中的比例在不同普查之间的吻合程度，或符合一定内在逻辑关系的程度，也就是在相同或相似问项上，普查数据与其他来源数据之间的吻合和接近的程度。经合组织（OECD）界定的普查的一致性，除同一套数据内部、不同数据之间、不同时间上的一致性外，还包括不同国家间的一致性。

在中国，已有很多文献对不同人口普查之间的一致性进行了较为透彻的分析。但绝大多数只是比较相同年份的出生队列在各次人口普查中登记数之间是否存在一致性，并以此来判别前后不同人口普查中可能存在的漏报或重报问题，特别是低龄儿童的漏报问题（崔红艳等，2013；胡耀岭、原新，2013；王广州，2003；于学军，2002；张为民、崔红艳，2002）。部分学者用外部数据（如户籍登记）或其他专题调查数据，对人口数量及结构、生育和死亡等指标进行过一致性分析（郭志刚，2004；黄荣清、曾宪新，2015）。所有这些研究无疑促进了我们对中国人口普查数据一致性及其质量的认识。

然而，目前除了一些学者（翟振武、陈卫，2007）使用中国教育部的在校学生数分析普查中的低龄人口漏报和生育率外，尚未有学者使用中国人口普查中的受教育程度数据对出生队列在各次人口普查间的一致性进行剖析。受教育程度是人口的一个基本特征，也是人口普查的一个主要问项。同一队列除非有大规模的全国性扫盲运动，否则接受初等教育的可能性很小。因此，我们可以分析成年人口中相同年份出生队列的人群在不同普查时点上申报的小学及以上（本文指小学毕业及以上人口）受教育程度人口比例之间的一致性。此外，现有文献中的另一不足是，几乎很少有研究使用多个国家的普查数据对普查数据质量进行系统的比较分析，尚未有研究基于多个国家各自历次普查数据对相同出生队列中接受过初等教育及以上的人口登记数或比例进行统计上的一致性分析。主要原因是数据的不可获得性和不同国家间受教育程度问项的不可比性。

美国明尼苏达大学人口中心国际微观数据系列整合共享数据库团队经过 15 年的努力，在经各国国家统计局授权后，截至 2015 年年底，共收集了 82 个国家的 270 多套人口普查中的微观数据。这些微观数据（或称普查样本）都是从原始的所有普查个体数据中按一定比例（1%、5%或 10%）随机抽取出来的个体数据（Minnesota Population Center，2014），并根据被国际接受的标准概念、定义和分类，对 270 多套样本数据进行整合，标准化了这些数据中的所有变量。鉴于此，本文拟使用国际微观数据系列整合共享数据库（Integrated Public Use Microdata Series-International）中包括中国在内的 13 个具有较大样本规模的亚太国家的各自历次普查中相同出生队列的小学及以上人口的比例，考察其在各国历次普查之间的吻合程度，即统计一致性。我们对初等教育及以上的人群进行考察是由于普及初等教育是一个千年发展目标和持续发展目标，也是因为大多数亚太国家的普查都搜集这一数据，而且在国际微观数据系列整合共享数据库的样本中，该指标被广泛使用。具体来说，我们的研究问题是，同一个国家某一出生队列在最近一次的人口普查样本中的小学及以上人口的比例与该出生队列在前一次普查中该比例的吻合程度如何？虽然数据的准确性与一致性存在明显的相

关关系，但准确性并不是本文所关注的主要方面。本文所关心的问题不是人口普查数据是否准确，而是相同队列的小学及以上人口比例在两次普查中一致性的程度。通过比较各国各自历次普查样本数据中相同出生队列的小学及以上人口比例之间的一致性，可以帮助我们了解有关各国人口普查质量的状况，寻找导致一致性较低或较高的原因，以期各国相互借鉴，共同提高人口普查的质量。同时也为其他研究人员在使用普查数据的同类指标时提供警示和参考。

二、数据和方法

（一）数据来源

本研究中所用的各个国家的数据均来自国际微观数据系列整合共享数据库。该数据库中的微观数据是指样本数据，即从普查的全部个人数据中，一般以县级行政区划为单位，按一定比例（一般为1%、5%或10%等）随机抽取。为了使每一套微观样本数据能够正确地反映全部人口的数据结构，每一套样本数据通常会附带一个抽样权数变量。本研究中也使用权数变量，以使基于样本数据得到的比例能够反映总体。

该数据库的基本目标是将全球各国的历次人口普查数据进行整合。其理念是在确保各国各次普查中在不遗漏任何有用信息的情况下，征集各国历次人口普查中的微观数据样本，根据尽可能统一的界定、概念和口径，对经各国国家统计局授权的各套普查微观数据中的每一个问项进行统一规范化、匿名化、存档化等一系列管理和整合，以便于各国普查数据的直接比较。也就是说，任何一个单一概念，比如本文所研究的小学及以上受教育程度，在该数据库的所有各国的普查微观数据中的编码都是一致的。为了避免遗漏那些更详细样本中的重要信息，该数据库使用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，研究人员可以进行时空对比，并甄别其差异。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。微观数据整合后，每套样本中相同的概念（变量、类别）就有了相同的代码。

除了收集原始数据外，该数据库还收集了包括普查表、现场调查指南、编码本、技术手册和官方出版物等在内的所有原始文件，并将它们整理和文档化，并与相应的样本数据一起公布。该数据库项目使用所有这些文件，对每一套高精度的微观数据中的每一个变量和编码进行整合。原始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所包含的信息（Esteve 等，2003）。通过对综合性原始源文件的研究，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6类：编码、总体概述、可比性讨论、适用性问题、概念的可用性、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）和链接到用于创建各个整

合变量的“源变量”。另外，为了使研究者能最大限度地使用普查数据中所包含的信息，微观数据系列整合共享数据库除保留原有普查数据中的一切信息外，该数据库团队通过数十年使用微观数据的经验，对每一套普查数据还研发了 30 多个附加变量。这些附加变量可以归为 3 组：技术性（类别、国家、年份、微观样本整合共享数据库数据身份码、户码、所在省州码）、汇总性（户类别、每户的家庭数、每户的已婚夫妇对数、每户的父母亲数、每户中的亲生子女数等）和指示性（用来甄别同住的配偶、子女和父母）（Sobek 等，2009）。也就是说，该数据库都是个体数据，数据框架结构包括两大部分：各国历次普查或调查中原有的并经过标准化整合过的问项（变量），以及后来衍生而成的 30 多个变量。需要说明的是，该数据库公布的整合过的每一套数据经过深度测试和强化。耗费了数千小时来分析、讨论、辩论、测试和再测试，直至整合的微观数据被验证可以向研究人员公布为止（McCaa，2013）。研究者根据这些数据在经过一些简单的自我处理后可直接进行数据分析，而无需在分析前进行各种各样的编码。

本文所包括的 13 个亚太国家按英文首字母排列分别为孟加拉国、柬埔寨、中国、斐济、印度、印度尼西亚、巴基斯坦、吉尔吉斯斯坦、马来西亚、蒙古、菲律宾、泰国和越南。各国普查的个体微观数据如表 1 所示。由于中国人口普查微观数据相当重要，本文在分析中使用了明尼苏达大学数据库中 1%的 1982～2000 年人口普查的微观数据，以及中国国家统计局网站上公布的 100%的汇总表格数据[①]。

（二）整合的受教育程度

对于一致性分析，本文关注数据库中使用最为广泛的变量，即受教育程度（变量名称为 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（联合国教科文组织统计研究所，2012）为依据，分为 4 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，国际微观数据系列整合共享数据库中的复合码包括 4 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，表示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）。

许多普查样本中含有如接受过初等、中等、甚至高等教育，但没有是否毕业等更多的信息。编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。国际层面上的数据整合必须记录这种区别，以使研究人员可以容易地了解这些细节（Ruggles，2006）。

表 2 列举了 13 个国家受教育程度变量的一般编码（由两位标准的国家或地区代码表示，ISO3166）。表 2 中的总码显示，所有样本都具有小学未毕业、小学毕业、中学毕业和大学毕业 4 个一般水平。表 2 中的细目码则表明，普查中搜集的受教育程度的信息在不同国家

① 中国国家统计局（2012）：中国 2010 年人口普查资料：表 4-1 全国分年龄、性别、受教育程度的 6 岁及以上人口（http://www.stats.gov.cn/tjsj/pcsj/rkpc/6rp/indexch.htm）。

表 1　亚太 13 国普查微观数据　%

国家	1980 年普查周期年份	样本比例	1990 年普查周期年份	样本比例	2000 年普查周期年份	样本比例	2010 年普查周期年份	样本比例
孟加拉国	–	–	1991	10	2001	10	2011	5
柬埔寨	–	–	–	–	1998	10	2008	10
中国	1982	1	1990	1	2000	1	2010[a]	–
斐济	1976	10	1986	10	1996	10	2007	10
印度	1983	0.091	1993	0.073	1999	0.065	2004	0.061
印度尼西亚	1980	5	1990	0.51	2000	10	2010	10
	1985	0.37	1995	0.43	2005	0.51		
吉尔吉斯斯坦	–	–	–	–	1999	10	2009	10
马来西亚	1980	2	1991	2	2000	2	–	–
蒙古	–	–	1989	10	2000	10	–	–
巴基斯坦	1981	10	–	–	1998	10	–	–
菲律宾	–	–	1990	10	2000	10	–	–
			1995	10				
泰国	1980	1	1990	1	2000	1	–	–
越南	–	–	1989	5	1999	3	2009	15

注：表中样本比例为数据库中个人数据占该次普查或调查中的全部个人数据的比例。a 为 100% 汇总列表数据；“–”表示没有数据。

和不同普查样本间千差万别。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，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。

除了中国 1982、2010 年普查和某些特殊说明外，本文中的小学及以上受教育程度均指小学毕业及以上。中国 1982 年人口普查的调查表中没有区分是否毕业，因此本文无法进行细分。中国 1990 年及以后的各次普查都对是否毕业进行了细分。但由于中国国家统计局尚未将 2010 年人口普查的个体数据授权于明尼苏达大学人口中心，本文无法生成小学毕业及以上人口的比例，只能借助中国国家统计局公布的 100%的汇总数据表格。该汇总表格中，受教育程度分为“未上过学”、“小学”、“初中”、“高中”、“大学专科”、“大学本科”和“研究生”，但并未区别各类教育是否毕业。也就是说，在使用中国 1982 和 2010 年人口普查数据时，本文用“小学及以上”（即 1982 年普查数据和 2010 年汇总表格中的初中及以上）来分析不同队列之间的一致性。

（三）分析方法

本文使用人口学中出生队列的概念，对某一套普查样本数据中 55 个出生队列（本文用 15～79 岁）各自的小学及以上人口比例与下次普查样本中相同队列（年龄为 25～89 岁）的同一个比例进行对比，分析其吻合的程度。若两次普查中 55 个队列各自小学及以上受教育

程度人口的比例均比较接近，两次普查之间的比例差异较小，说明两次普查数据中这些年龄上的人口数据很可能比较准确。本文除了计算 55 个队列的两次相邻普查之间小学及以上人口比例之间差异的均值和中位数外，还用最小二乘法对这 55 个队列在两次普查中的小学及以上人口比例拟合了线性回归方程。差异的均值是指某一普查 55 个队列中每一队列的小学及以上人口比例与相比较的另一次普查中各自相同队列的同一比例之间的差值

表 2　13 个国家微观数据中受教育程度未加权频数

代码	受教育程度	普查年份						
		2011	2008	1990	2007	2004–5	2010	2009
		孟加拉国	柬埔寨	中国	斐济	印度	印度尼西亚	吉尔吉斯共和国
总码								
0	NIU（不适用）	1117354	136274	1418185	–	–	2253453	72044
1	小学未毕业	3216705	766314	4383067	24403	316386	6117917	124184
2	小学毕业	2065976	376009	5069640	40755	172721	10135303	66697
3	中学毕业	639020	47837	915562	17684	85023	4394068	251330
4	大学毕业	166665	13010	49493	1468	28290	702308	48606
9	未知	–	677	–	13	413	–	2125
细目码								
0	NIU（不适用）	1117354	136274	1418185	–	–	2253453	72044
100	小学未毕业	–	–	–	–	–	–	–
110	未上过学	1961034	297550	2145035	10890	220227	1986754	–
120	上过一些学	1255671	468764	2238032	13513	96159	4131163	–
130	小学（4 年）	–	–	–	–	–	–	68016
	小学毕业但中学没有毕业							
	小学毕业							
211	小学（5 年）	1256266	–	–	–	88352	–	–
212	小学（6 年）	–	256570	2822479	24932	–	6539863	–
	中学未毕业							
221	正规或非特殊教育	809710	119439	2247161	15823	84369	3595440	46187
222	技校	–	–	–	–	–	–	20510
	中学毕业							
	正规或非特殊教育							
311	正规教育毕业	639020	41385	640916	10489	49669	3592138	208581
312	接受过一点大学教育	–	–	43450	380	29237	–	15106
320	技校教育	–	–	–	–	–	–	–
321	中专		2228	148554	–	–	400543	38242
322	大专		4224	82642	6815	6117	401387	–
400	大学毕业（本科毕业）	166665	13010	49493	1468	28290	702308	48606
999	未知或缺失	–	677	–	13	413	–	2125

表 2 13 个国家微观数据中受教育程度未加权频数 续表

代码	受教育程度	普查年份					
		2000 马来西亚	2000 蒙古	1998 巴基斯坦	2000 菲律宾	2000 泰国	2009 越南
总码							
0	NIU（不适用）	–	35396	1944463	935577	43640	1517591
1	小学未毕业	223334	84105	7900912	2132120	284685	4675806
2	小学毕业	166637	54743	2722128	1967457	179347	6140145
3	中学毕业	10486	55050	276019	1689518	69705	1316274
4	大学毕业	25456	14431	236278	305054	20933	527774
9	未知	9387	–	22224	388084	6209	–
细目码							
0	NIU（不适用）	–	35396	1944463	935577	43640	1517591
100	小学未毕业	56168	–	40263	·	–	–
110	未上过学	100909	–	6375658	466783	55479	892633
120	上过一些学	122425	–	1525254	166533	229206	3783173
130	小学（4 年）	–	43842	–	–	–	–
	小学毕业但中学没有毕业						
	小学毕业						
211	小学（5 年）	–	–	–	–	–	–
212	小学（6 年）	80005	–	2038363	1967457	116450	267120
	中学未毕业						
221	正规或非特殊教育	86632	47742	683765	–	62897	3468942
222	技校	–	7001	–	–	–	–
	中学毕业						
	正规或非特殊教育						
311	正规教育毕业	8878	40677	260127	814182	24371	1074774
312	接受过一点大学教育	–	–	–	715722	17237	151141
320	技校教育	–	14373	–	–	–	–
321	中专	27643	–	–	–	13106	90359
322	大专	–	1608	–	15892	14991	–
400	大学毕业（本科毕业）	25456	14431	236278	305054	20933	527774
999	未知或缺失	9387	–	22224	388084	6209	–

注：表中所列年份仅为示例，文中所用其他年份没有列出。印度 2004–5 是指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。“–”表示不适用。

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

的简单算术平均数。差异中位数是指这 55 个差值大小顺序中，处于正中间的那个差异值。若两次普查的一致性较高，则均值和中位数都应该趋近于 0。拟合的线性回归方程为$\hat{y}=a+bx$。回归系数（b）和解释系数可以用来测量样本数据间的一致性程度。若 R^2 为 1，则 y 完全可以

由 x 确定。若各队列在两次普查中的小学及以上比例相同，则 $y=x$。因此，若两套数据之间的一致性很好，R^2 接近于 1；回归方程中的 a（截距）应趋于 0，回归系数（b）应接近于 1。

三、主要结果

由于篇幅限制，本文侧重展示中国、越南、印度 3 个国家的主要结果。其他所有的国家的结果只略微提及。

（一）中国的主要结果

尽管中国 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、1990 和 2000 年中国的 3 次人口普查的微观样本数据明显表现出近乎完美的一致性。由于 1982 年的人口普查没有区分毕业者与未毕业者，我们可通过根据 1982 年的界定对 1990 和 2000 年微观数据中受教育程度重新分组制表（见表 3）。研究结果证实了 3 套样本中的统计一致性很高。例如，1990 与 2000 年相比，相关系数和回归系数近乎完美，为 0.99，平均差异为 0.2 个百分点。中位数的差异为零。这是本文中涉及的所有国家数据中统计的一致性最好。1982 与 2000 年数据间的一致性虽不完美但一致性仍较高，回归系数为 0.96，平均差值比 1990 年与 2000年数据间的差异大近 30 倍。即使这样，1982 与 2000 年数据间的平均差异仅为 –3.5 个百分点。

中国 2000 和 2010 年的人口普查中，与教育有关的 3 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。

由于缺乏中国 1982 和 2010 年人口普查的微观数据中的详细的受教育程度分类指标，统计一致性可以通过采用一个略微不同的概念，即不用“小学毕业及以上”，而只用“小学以上”，再与 1990～2000 年各次微观数据进行比较。“小学以上”这个概念变得不单纯是指完成小学教育，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（如“小学以上”，即汇总表中的初中及以上），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，经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 1 所示，中国 1982～2010 年“小学以上”（即未包括那些只有小学毕业的人口）受教育程度的统计一致性很高。1982、1990 和2000 年 3 次普查与 2010 年普查两两间的回归系数均为 0.95，非常接近表 3 中的 0.99 和理想的1.0。相关系数 R^2 也几近完美，1982 与 2010 年两个数据样本之间的解释系数为 0.98，1990 与 2010 年、2000 与 2010 年之间的解释系数均为 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，本文考察了 1960 年出生队列。结果发现，这个队列在 1982～2010 年 4 次普查中的小学以上人口的比例分别为 64.8%、66.1%、68.6%和 69.8%。1925～1963 年出生的 39 个队列的总体平均比

表 3 13个亚太国家个体微观数据中初等教育的统计一致性

国家	年份	样本比例（%）	样本规模（×10⁶）	惠普尔指数及数值	SCHOOL 状态	YRSCHL 年数类别	EDATTAN 水平	人口比例** （%）	55个出生队列 百分比	均值	中位数	R^2	b
孟加拉国	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			0.99	0.96
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				

注：惠普尔指数：a表示数据申报质量很好（<105），b表示数据质量较好（105～109.9），c表示数据质量一般（110～124.9），d表示数据质量较差（125～174.9），e表示数据质量很差（≥175）。表中有关受教育程度的3个变量均由国际微观样本系列整合共享数据库整合后创建的。样本比例是公开发表的个体数据记录数占某次普查全部个体记录数的比例。“–”表示不适用。* 数据没有区分教育水平的开始或完成，** 表示1970年出生队列小学及以上人口比例。

资料来源：www.ipums.org/international。

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，也与在更大年龄阶段接受更多的

教育(如成人扫盲)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扫盲。1978～1998年,成人高等教育的招生数每年增长3.5%;1999～2010年,每年的增长率上升到10.4%(Lai,2014:62)。我们相信,2010年人口普查样本微观数据中的结果比仅基于中国国家统计局汇总表格数据中的比较更具一致性,因为在微观数据中,可以通过用有"仪式"象征意义的变量值,比如"小学毕业及以上",而非"小学以上"这一不直观的概念进行比较。

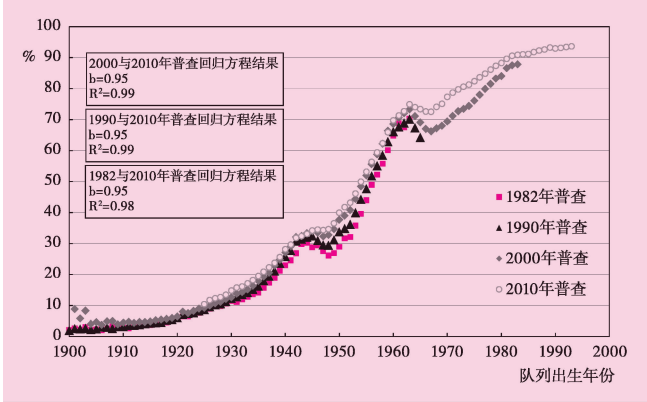

图1　1982、1990、2000和2010年中国4次普查样本中小学以上人口比例

注:2010年为100%汇总数据。其余年份数值由普查样本个体数据求得。1990年数据包含了80～89岁人口,2000年数据包含了90～99岁人口,因此与表3中的结果略有不同。

(二)越南的主要结果

图2描绘了越南1989、1999和2009年普查样本小学及以上受教育程度人口比例。这些比例曲线揭示了各次普查的一致性。比如,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1.3个百分点。1999与2009年、1989与2009年的全部队列的回归系数分别高达0.93和0.92,解释系数为0.99接近于等于1.0时完全吻合的状态。也许这些几乎完全一致的结果不值得奇怪,因为这些数据是来自同一个统计部门,而且3次普查数据的处理和编码均使用了不断完善的高技术,从而避免了很多可能产生的错误。

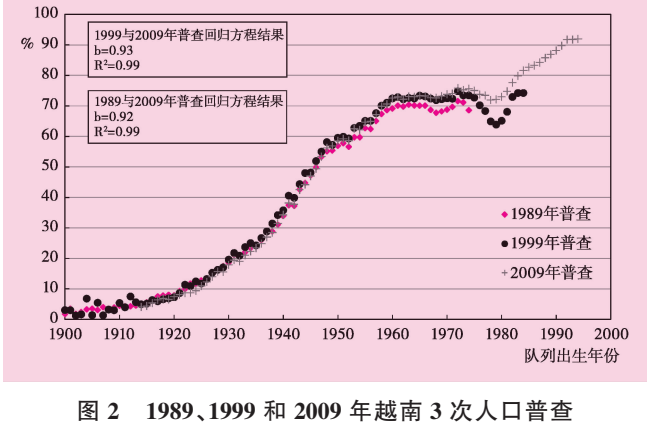

图2　1989、1999和2009年越南3次人口普查小学及以上受教育程度比例

注:1999年数据中包含了年龄90～99岁的人口,因此与表3中的结果略有不同。

尽管越南统计总办公室在3次普查中使用了统一的面对面实地访谈,但由于从3次普查各自的全部个人数据中抽选微观数据的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%①。

① https://international.ipums.org/international/sample_designs/sample_designs_vn.shtml。

（三）印度的主要结果

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，住户调查表 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数据系列整合共享数据库公布，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）3 套数据（见图 3）。印度全国抽样调查机构在其二十多年来的问卷中始终对小学毕业使用统一的定义，但调查问卷中没有受教育年限这一问项 。因此 EDATTAN 变量的编码与印度国家的惯例保持一致，即小学 5 年毕业，初中 8 年毕业，高中 10 年毕业。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，小学及以上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）。

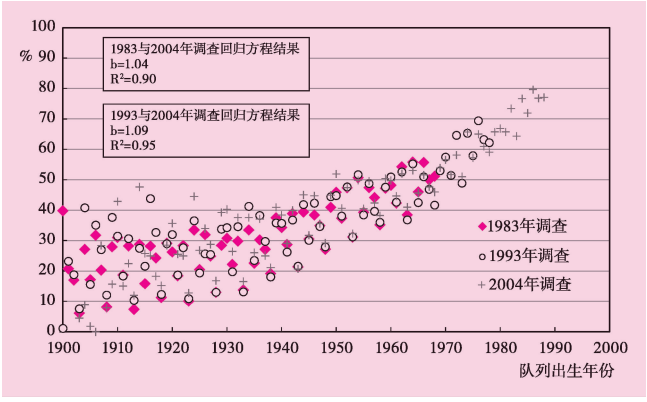

图 3　1983、1993 和 2004 年印度全国抽样调查机构组织的 3 个周期的调查

表 3 提供了这些调查及其他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，但只有中国的数据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。其数据规模虽然“微小”，但其统计的一致性却较高。如在印度全国抽样调查机构的调查样本、孟加拉国和巴基斯坦的人口普查样本中，未接受过教育的人群在年龄申报时存在的强烈数字偏好，歪曲了不同时间上的比较。表 3 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“非常低劣”。尽管如此，我们发现，1970 年的出生队列在 2004–5 年数据中申报小学及以上的比例为 53.9%，与 1993–4 年数据中申报的 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 3 总结了对目前由国际微观数据系列整合共享数据库公布的 13 个亚太国家的样本分析。中国、蒙古和泰国 3 个国家的一致性几近完美。这三国各自的平均差异小于 1 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，为 6～16 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解释系数都在 0.7 以下。

四、讨论和结语

本文基于明尼苏达大学人口中心国际微观数据系列整合共享数据库中包括中国在内的13个具有较大样本规模的亚太国家的普查微观数据，考察了55个不同出生队列中的小学及以上受教育程度的人口比例在各国历次普查中的统计一致性问题。理论上，若各次普查的数据质量较高，相同出生队列所揭示的结果应该相似或一致。具体到本文，当普查界定、数据搜集和处理程序一致，且普查数据质量较高时，同一出生队列中的小学及以上受教有程度的人口比例在各次普查样本中应该相近。本研究发现，中国1990和2000年两次人口普查中，相同出生队列的小学及以上受教育程度人口比例的平均差异小于0.3个百分点，回归系数为0.96～0.99，其中位数的差值为0.0，R^2和b均为0.99。越南、蒙古和印度尼西亚3个国家的一致性也很高，平均差异不到0.5个百分点，回归系数在0.93～1.07，R^2高达0.99。但另外一些国家的一致性相对较差，有的国家与平均值的绝对差异可高达16个百分点。我们对非洲国家人口普查微观数据的研究显示，其分歧比亚太13个国家的差异还要大，R^2为0.38～0.99，回归系数为0.46～1.37（McCaa等，2015）。当然，欧洲国家一致性的差异要小得多，回归拟合效果也更好（McCaa等，2014）。中国、越南、蒙古和印度尼西亚4个国家的一致性比较高，可能主要归因于这些国家在各次普查之间确保了概念上的连贯性和受教育程度问项的一致性和严格沿用国际标准的定义的做法。当然，也可能与这些国家各自的历次普查的覆盖面及总体误差大致相同有关，还与他们较低的国际迁移率、人口内部不同群体之间稳定的死亡率和较小的差别死亡率有关。在那些一致性比较差的国家，可能与各次普查之间概念界定上的不连贯性和现场组织实施或数据处理过程不同，或者与各自普查之间成人扫盲项目的开展、各次普查间的覆盖面或误差等因素有关。产生这些不一致性的原因很多，需要对元数据、微观数据及人口变化进行深入分析才能给出答案。

在评估历次普查之间的一致性时，至少需要注意3个事项（Feeney，2014）：普查机构的组织实施方式，国际微观数据系列整合共享数据库的规范统一性和偏差。第一，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，鉴于本文分析所用的数据是经国际微观数据系列整合共享数据库整合过的，所以必须考虑数据库团队对微观数据进行的规范化的做法，以及对各次普查统一编码方案决定的正确与否。第三，评估时需要假定不同受教育程度人口的死亡率、迁移率及普查申报登记状况相同，且没有成人扫盲项目而使超过常规年龄后的小学及以上人口比例增加。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。

本研究结果所得出的中国、越南、蒙古和印度尼西亚4个国家较高的小学及以上受教育程度一致性，只能说明这55个出生队列（年龄为15～79岁）中的这一指标在各次普查之

间的一致性较高，并不表明在其他指标在各次普查之间的一致性，也并不表明其他年龄上的一致性。普查质量高，一致性一般就高，但一致性高并不一定说明普查质量高，因为普查质量还有其他方面的考量，比如准确性等。全面评价普查数据的质量不仅应该包含对多个指标一致性进行检验，还需要结合准确性等其他方面进行综合评价。

相同队列内部比较是测量不同时点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经过整合，否则这种方法因其应用上的困难而较少被使用。同时，为了实现统计上的一致性，定义、概念、框架和分类必须在国家甚至国际层面上做到明确和统一。理想的结果是，普查问项应保持长期不变，以便进行纵向比较（Baffour 等，2012）。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。然而，目前面临的挑战之一是，各国的国家统计局并不能提供历次普查中相同概念不同界定和编码统一标准化的问题。由于联合国统计司出版了《人口与住房普查原则和建议》，并被各国广泛采用使普查编码的统一化有了可能。但随之而来的问题是，因缺乏人力和资源，事务繁重的各国国家统计局很少对本国历次普查的概念和编码进行统一化和规范化处理，也很少对样本数据匿名化和文档化处理（联合国经社部统计司，2008）。在这种情况下，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）会遇到较大风险，并付出巨大的人力资源成本。

在这种背景下，国际微观数据系列整合共享数据库具有一定的规模经济效益，具有成本低和风险小的优势。这种整合是全球化趋势发展的要求，有利于各国之间人口普成数据之间的直接比较，促进了普查数据共享机制的良性发展。

目前该数据库共收集了世界上82个国家的270多套普查微观样本数据。为获取数据，研究人员需要在数据库网站[①]注册。注册成功后，登录自己的账户，在线递交在研究中所需要的数据申请，在数据库评审委员会审阅通过后，研究人员会被授权，然而可以免费下载所申请的微观数据（McCaa 等，2006）。截至2015年年中，该数据库注册用户已超过1万名，分别来自130多个国家或地区的2 000多个机构。引用国际微观数据系列整合共享数据库的数据出版物文献已超过1 000个。在亚洲各国中，中国居首位，共引用64次[②]。印度居第二位，共引用58次。其次为越南（49次）和菲律宾（36次）。微观数据系列整合共享数据库已经成为各国国家统计局发布普查数据的一种必要补充。

参考文献：

1. 崔红艳等（2013）：《对2010年人口普查数据准确性的估计》，《人口研究》，第1期。

① https://international.ipums.org/international。

② 随着2010年周期普查数据及原有积压但已授权明尼苏达大学人口中心的其他国家的微观数据样本的整合工作的完毕，到2020年，数据库的容量可能会翻倍。中国经过整合的2000年人口普查的微观数据样本计划于2016年公布。我们希望在最近1～2年内，中国2010年人口普查的样本数据也能被数据库收入；希望更多的中国学者使用或引用国际微观数据系列整合共享数据库，从而让人口普查最大潜能地服务于公众。

2. 郭志刚(2004):《对中国 1990 年代生育水平的研究与讨论》,《人口研究》,第 2 期。
3. 胡耀岭、原新(2013):《1982～2010 年期间四次全国人口普查数据一致性研究——基于出生人口队列的分析》,《人口研究》,第 1 期。
4. 黄荣清、曾宪新(2013):《"六普"报告的婴儿死亡率误差和实际水平的估计》,《人口研究》,第 2 期。
5. 联合国教科文组织统计研究所(2012):《国际教育标准分类法 ISCED 2011》,加拿大蒙特利尔。
6. 联合国经社部统计司(2008):《人口和住房 普查的 原则和建议》,第二修订版,纽约。
7. 王广州(2003):《对第五次人口普查数据重报问题的分析》,《中国人口科学》,第 1 期。
8. 于学军(2002):《对第五次全国人口普查数据中总量和结构的估计》,《人口研究》,第 3 期。
9. 张为民、崔红艳(2002):《对中国 2000 年人口普查准确性的估计》,《人口研究》,第 4 期。
10. 翟振武、陈卫(2007):《1990 年代中国生育水平研究》,《人口研究》,第 1 期。
11. Baffou, B. and P. Valente (2012), An Evaluation of Census Quality. *Statistical Journal of the IAOS*. 28:121–135.
12. Esteve, A. and M. Sobek(2003), Challenges and Methods of International Census Harmonization. *Historical Methods*. 36:66–79.
13. Feeney, G.(2014), Literacy and Gender: Development Success Stories. Population and Development Review. 40:545–552.
14. Lai, Qing(2014), Chinese Adulthood Higher Education: Life–Course Dynamics under State Socialism. *Chinese Sociological Review*. 46:55–79.
15. 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 Histórica*. 30:69–87.
16. McCaa, R., M. Sobek, L. Cleveland, A. Esteve, and A. Lopez(2014), Statistical Coherence in Secondary Education Completed: Census Hub Hyper Cubes and IPUMS/IECM integrated Census Sample Results Compared. Conference of European Statisticians, Group of Experts on Population and Housing Censuses Sixteenth Meeting, Geneva, 23–26 September.
17. 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.
18. 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.
19. Minnesota Population Center(2014), Integrated Public Use Microdata Series, International: Version 6.3 [Machine–readable database]. Minneapolis: University of Minnesota.
20. 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.
21. Sobek, M and S. Kennedy(2009), The Development of Family Interrelationship Variables for International Census Data. Minnesota Population Center.

(责任编辑:朱　犁)