

人口流迁

两种调查视角下流动人口结构的对比分析*

周 皓

【内容摘要】文章以 2015 年小普查数据为基准,比较了 2015 年流入地和流出地两种调查地点下流动人口监测调查数据反映的流动人口的结构差异,并从调查方法论的角度,探讨了各种可能的调查误差对调查结果的影响。分析表明:流出地调查得到的年龄、婚姻和流动范围结构以及流入地调查得到的居住时间结构更接近小普查;流入地调查中的女性人口以及流出地调查中的短期流动人口和男性人口可能被过度抽样;两种调查都偏重于低受教育程度的流动人口,相对忽略了大专及以上学历的流动人口。从调查方法论角度看,不同的应答者是调查误差(特别是测量误差)的重要根源。除覆盖误差外,无应答误差也是影响调查结果的重要因素之一。文章建议重视样本的结构性偏差对分析结果的潜在影响,以正确地使用调查数据。

【关键词】流动人口;人口结构;流出地;流入地

【作者简介】周皓,北京大学社会学系教授、北京大学中国社会与发展研究中心研究员。北京:100871

A Comparative Study on the Population Structure of Migrants from Two Perspectives

Zhou Hao

Abstract: Based on data of China Migrants Dynamic Survey (CMDS) conducted in the origin and destination in 2015, this paper compares the structural differences of the floating population under the two survey perspectives, with the mini-census in 2015 as the benchmark. The results suggest that the survey conducted in origin is closer to the results of the mini-census in age structure, marriage, and distance of movement. The survey conducted in destination is closer to the mini-census in terms of residence time. Women in the destination areas and short-term (1 year or less) migrants in the origin areas are over-sampled. Both surveys focus on the population of low education level while neglecting the population with college degree or above. From the perspective of survey methodology, recruiting different respondents is one of the most important sources of survey errors, especially measurement errors. In addition to coverage error, non-response error is also one of the important factors affecting the survey results and sample structures.

Keywords: Floating Population, Population Structure, Origin, Destination

Author: Zhou Hao is Professor, Department of Sociology, and Senior Researcher, Center for Sociological Research and Development Studies of China, Peking University. Email: zhouh@pku.edu.cn

* 本研究受到国家社会科学基金项目“中国流动人口调查的比较研究”(19BRK039)的资助。

1 引言

但凡抽样调查,都必然存在其无法避免的各种问题。比如从装有黑白两色乒乓球的黑箱中抽取样本,只要样本规模足够大,抽取一次的结果对总体是有代表性的;抽取两次的话,尽管每次抽样结果可能会存在一定的差异,但却各自都有代表性,且内部的结构分布与总体也应该基本接近。这种实验性的差异可以归结为纯粹的抽样误差。但现实的社会调查或人口抽样调查则不仅包含抽样误差,而且从测量到代表性的各个方面都可能存在误差:测量方面的误差包括效度问题、测量误差和过程误差;代表性方面的误差则包括覆盖误差、抽样误差、无应答问题和调整误差(Groves et al., 2009)。因此,在现实的抽样调查过程中,抽样误差只是众多误差的来源之一。如何尽可能地缩小或降低包含抽样误差在内的所有误差是每个社会调查都需要考虑的基本问题。

相比于一般社会调查,流动人口的抽样调查更为复杂,它不仅涉及流出地与流入地两个方面,还涉及流量与存量的问题(即流动人口内部是不断循环与替代的)。从调查地点来看,不论是流出地调查,还是流入地调查,其对应的总体是一致的,即全国的流动人口,犹如黑箱中所有的球。但抽取的结果可能是不同的,当然即使不完全相同,样本仍然能够稳定地体现总体的各种特征。这种不同的结果,既有可能是因为抽样误差引起的,也有可能是因为上述其他各种调查误差引起的。以往研究多关注覆盖误差或抽样误差,而未从调查方法论的角度考虑其他各种调查误差对流动人口抽样调查结果的可能影响。

一般情况下,针对流动人口的抽样调查,或是在流入地进行,或是在流出地进行,几乎没有同时在流入地与流出地开展的调查。有幸的是,原国家卫计委在2015年组织的流动人口监测调查中,同时开展了流入地和流出地调查,得到了丰富的流动人口信息,这不仅为相关研究提供了很好的数据基础,也为我们同时从流出地和流入地两种调查视角考察与探讨流动人口抽样调查的方法论提供了宝贵的机会。

本文无意评价这两次调查的数据质量。有关流动人口监测调查数据的质量评估可参见王广州、王天忆(2015)以及郭志刚(2009)的相关研究。本文的所有讨论都是基于:假设这两个调查结果都为“真实的”随机抽样结果。本文的目标是:比较两种视角(调查地点)下的流动人口调查结果之间是否存在系统性的结构差异,进而从调查方法论的角度,讨论各种抽样误差和可能的社会选择性对这种结构差异的影响。

已有一些研究讨论了流动人口抽样调查中存在的问题,内容包括:抽样方案设计(齐嘉楠等, 2014);测量(吕利丹、段成荣, 2012);抽样过程,特别是抽样框设计过程(沈明明、李磊, 2007;陈传波等, 2012;庄亚儿、李伯华, 2014);亦有讨论新的抽样方法(如被访者驱动法等)对流动人口抽样调查方法的改进(赵延东、Jon Pedersen, 2007)。这些讨论都有助于正确认识流动人口抽样调查的难度及各种可能存在的方法论问题。梁玉成等人(2015)还讨论了流入地调查中存在的问题以及流出地调查的相对优势,指出流入地调查存在难以克服的抽样框问题、样本选择性问题、追踪调查难的问题和系统误差问题,并从理论和实践两个维度,说明流出地调查在减少调查系统误差、提高样本代表性、更具操作性以及可追踪性等方面具有巨大的优势。流出地调查诚然在某些方面存在一定的优势,但它同样可能会存在一些劣势。任何一个抽样调查都会存在这样或那样的问题,关键在于如何看待以及如何应用这些调查数据进行正确的分析。

2 调查数据的基本情况及其说明

本文所用数据为原国家卫计委2015年组织的流动人口监测调查(此为流入地调查)和流动人口流出地调查这两个数据。有关这两个调查的具体信息请参阅相关的技术文件。本研究在对比分析之前特别注意以下5点:

第一,在流入地与流出地调查中,调查对象“流动人口”的定义都是以“1个月及以上”为时间标准的。这与人口普查的定义有所区别。

第二,流出地调查是“在河北、辽宁、吉林、黑龙江、江苏、浙江、安徽、河南、广东、四川等十个省份抽取县/区级监测点,开展流出地监测调查,调查结果对该十个省份均具有代表性”。为了能够剔除由于调查省份的限定而导致的结果偏差,作者对比了流出地十省份调查数据和2015年1%人口抽样调查(后文简称“2015年小普查”)数据,发现流出地调查数据与2015年小普查数据中对应的十个省份流出人口及全国流动人口的结构基本相同。这从一个侧面说明流出地调查可以代表全国性的流出地调查结果。

第三,流动人口监测调查自2010年以来每年都进行。基于调查时点的考虑,本文选用了2015年的调查数据,它既可以与当年的流出地调查数据相对比,也可以与2015年小普查数据相对比。当然,流入地监测调查于2015年5~6月进行,流出地调查于2016年1月进行,这与小普查的标准时点(2015年11月1日)存在一定的差异。但这几个月的时间差异不足以改变整体结构。因此,本文并未对标准时点予以调整。

第四,所有结果都根据原始数据提供的权重进行加权。

第五,流出地调查的原始数据中提供了跨县(poprsdp)与跨乡(poprsdpt)的流动人口判断结果。为与人口普查数据相对应,本文均采用跨越乡镇的标准。

从抽样方案来看,2015年流入地调查“以31个省(区、市)和新疆生产建设兵团2014年全员流动人口年报数据为基本抽样框,采取分层、多阶段、与规模成比例的PPS方法进行抽样。”在抽样方案设计部分又强调“在流动人口全员信息系统2014年年报数据基础上建立的抽样框中选取”,“在保持对全国、各省有代表性的基础上,增强对主要城市、均等化重点联系城市的代表性。”“流动人口目标总体为全国在调查前一个月前来本地居住、非本区(县、市)户口且2015年5月年龄在15周岁及以上的流入人口。目标总体中……不包括在校学生与“半边户”^①流动人口。对符合抽样总体要求但在非正规场所(临时工地、废弃厂房、草棚、涵洞等非正式居所)居住的流入人口,编制次级抽样框时应避免遗漏。”^②

2015年流出地调查是“以全国第六次人口普查分县数据、2015年上半年分县全员流出人口数据为初级抽样框”分县(市、区)、村/居委会、调查小区(或称“网格”)和住户共四阶段,“采取分层、多阶段、与户籍人口规模成比例的PPS方法进行抽样”,“末端抽样采用地图抽样法。通过绘制抽中村/居房屋分布图,编制住户清单列表,在线抽取样本,并遵循同质化和随机化原则,由国家级专家组负责样本替换,有效避免样本有偏性问题。”调查对象为“抽中村/居的户籍人口”^③。

显然,这两个调查的抽样设计是完全不同的。流入地调查是以“流动人口全员信息系统年报数据”为抽样框,而流出地调查则是以行政区域加地图为抽样框。在具体分析中要注意不同的抽样框对抽样结果的重要影响。

从案例数来看,流入地调查数据共有499406个案例,均为流动人口。其中,64.17%为跨省流动人口,25.45%为省内跨市流动人口,市内跨县的比例为10.36%,另有约0.03%为跨境人口。

流出地调查数据共有149011个案例,其中未流动的常住人口和流出人口的比例分别为65.62%和28.11%,流入人口为4.24%,另有2.03%为不确定案例。流出与流入人口之总和(占比为32.35%,共

① “半边户”是指流动家庭中有1个及以上的成员户口为本地户口的家庭户。

② 引自:国家卫生与健康委员会.2015年流动人口卫生计生动态监测调查技术文件(内部文件)2015年5月。

③ 引自:国家卫生与健康委员会.2015年流动人口卫生计生服务流出地监测调查工作方案(内部文件)2015年12月。

48214^①人)即为下文分析与讨论的流出地调查中的流动人口。

本文的分析思路为:首先着重对比两个不同调查地点的流动人口年龄结构的差异;其次再比较诸如性别结构、流动时间、流动范围、受教育程度、婚姻结构和家庭户规模等其他特征上的差异;最后对结构差异的原因进行分析及探讨。

3 年龄结构的对比分析

年龄结构是人口学特征中最重要的维度之一。对全国总人口年龄结构的检验,既可以用前后几次普查的数据进行比对,也可以利用各种间接估计方法进行估计。但流动人口的年龄结构并没有一个“标准人口”可以作为参考,而且也不能按照 Rogers 和 Castro(1979)或杨云彦(1992)所给出的迁移人口年龄模式来检验调查数据,因为随着时代的变迁和国内外社会背景的不同,目前我国的流动人口状况与国外或国内以往的结果相比肯定会有所变化。因此,本文对年龄结构的分析主要是比较两个不同地点调查得到的结果,并将其与 2015 年小普查数据进行对比,以检验是否存在结构性差异。

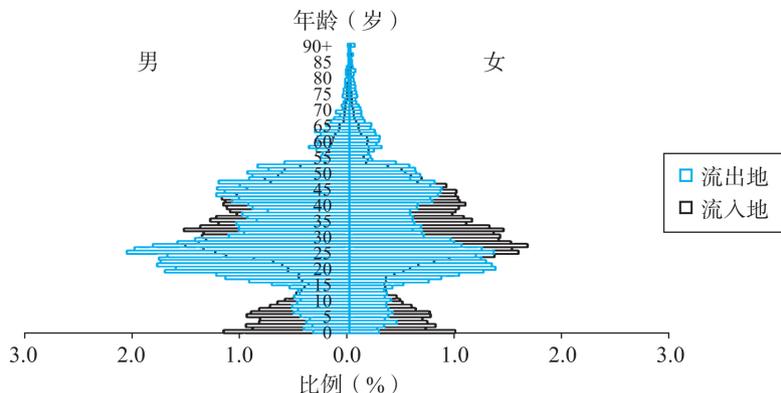
3.1 流出地调查与流入地调查的对比

图 1 将流出地与流入地的两个流动人口年龄金字塔叠加在一起,以便进行比较分析。从整体来看,不论是流出地调查,还是流入地调查,其年龄金字塔都反映了流动人口的年龄选择性特征,即分年龄人口比例在 20 多岁开始迅速上升至最高点,之后再逐步下降,而 50 岁左右还有个小的凸起,可能与我国总人口的年龄结构相关。

从流入地角度看,0~10 岁的低年龄段流动人口比例相对较高,特别是 0 岁组人口,相对高于 1 岁及以上的低年龄段流动人口。这可能是因为有较多的流动女性在流入地生育子女,但在其后则会将部分子女送回老家抚养,而流动母亲则继续在流入地工作。这种情况一直持续到 6 岁左右,然后其比例逐步下降,这与流动儿童就读小学有密切联系。在 10 岁之后,该比例再次出现递减趋势,特别是在 12 岁和 15 岁左右的下降,这与流动儿童本身的受教育过程相关联。流入地流动人口年龄结构的特征与家庭式迁移流动有着密切的关系。随着年龄的增加,流动人口分年龄人口比例先是逐步降低,到 20 岁左右时又迅速上升,并持续到 45 岁左右,之后再开始慢慢下降,且下降呈现出年龄的阶段性,即 40 多岁时下降一部分,50 多岁时又下降一部分,到 55 岁左右出现断崖式下降。这种年龄结构的变化过程与流动人口在劳动力市场上的就业形势相关联。

图 1 流动人口年龄金字塔:流入地和流出地比较

Figure 1 Age Pyramids of Migrants: A Structural Comparison of Origin and Destination



资料来源:图中数据为本文根据 2015 年流动人口监测调查和 2015 年流动人口流出地调查数据计算得到。

① 此为加权后的频数。

如果比较两个叠加的年龄金字塔,则可以看到两种视角下的调查结果存在许多不同之处:

首先,在 0~10 岁的低年龄段中,流入地的流动儿童比例远高于流出地调查的结果,特别是在 6 岁以前。这可能是因为更多的流动人口将婴幼儿流动人口带在身边抚养。从调查的角度来看,流入地调查得到更多的流动儿童可能与流入地调查的抽样过程(特别是抽样框)有关。那些在流入地成长的流动儿童,作为家庭式流动的结果,是流动人口中不可忽视的重要部分,因而在以家庭户为末端抽样单位时,必然会导致样本中包含更多的流动儿童。流出地调查则无法“捕捉”到那些在流入地成长的流动儿童,而且家庭式流动使整个家庭未能被纳入流出地的被调查对象之中。

其次,在 15~30 岁组中,流出地调查的比例远高于流入地调查。导致这一现象的原因可能比较复杂:一方面涉及比例的计算问题,即每个年龄段本身的规模在调查人口中所占的比例会随其他年龄段比例的变化而变化;另一方面可能也反映了流动人口的统计口径与调查方式的问题,如流入地调查由于抽样框问题而可能更关注已婚流动人口,而未婚人口的比例相对较低。分性别来看,这一年龄段里,流出地调查中男性相对较高的比例一直延续到 30 岁左右,而女性则仅到 26 岁左右。

再次,在 30~50 岁组中,流入地调查的比例显著高于流出地调查,且在男性和女性人口中都表现出同样的特征。这与 0~15 岁组流入地调查的比例较高是相关联的,既体现了家庭式迁移的影响,也体现了抽样框的特点。

最后,从 50 岁左右开始,流入地调查的结果出现断崖式下降,表明流动人口迅速退出流动大军。值得注意的是,这一年龄段中,流入地调查的比例远低于流出地调查。这说明,流入地调查可能对老年段的流动人口有所忽略。

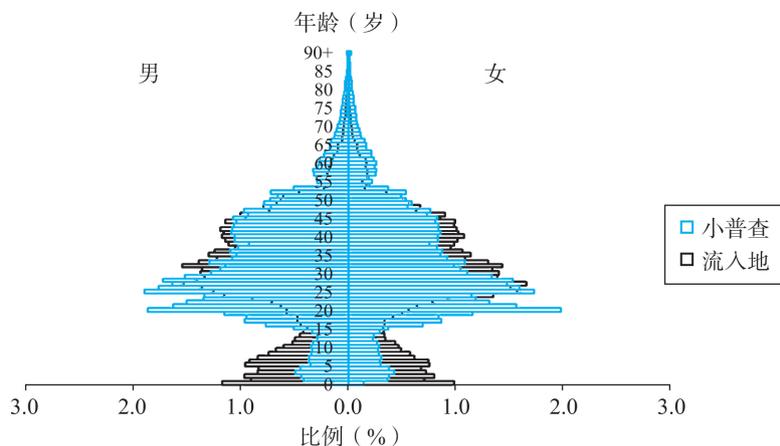
由以上比较可以看到,不同调查地得到的流动人口年龄结构存在显著的差异,且在不同年龄段表现出不同的特点。这可能既与年龄结构的计算方法有关,也与不同调查地点的抽样框及由此抽样得到的流动人口有关。

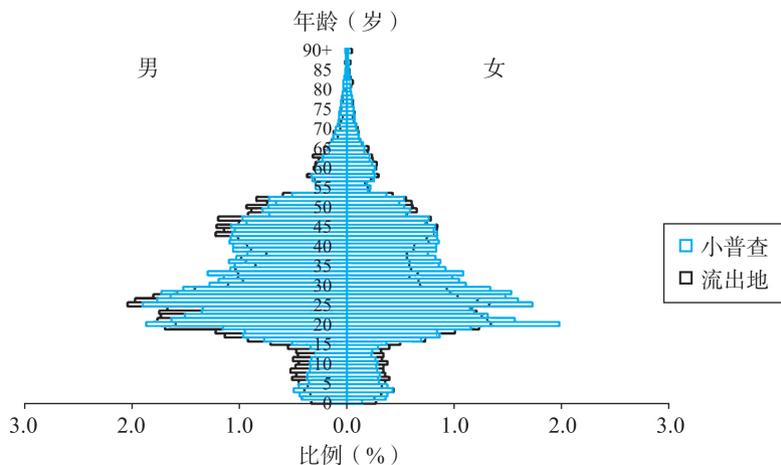
3.2 两个调查与 2015 年小普查的对比

2015 年小普查也是流入地调查,但其样本规模与抽样方法等完全不同于流动人口监测调查。尽管小普查也可能存在调查误差,但相对而言,其可靠性会更强一些。因此,本文将 2015 年小普查的调查结果作为标准,考察两种视角下流动人口监测调查得到的年龄结构。

图 2 两个调查与 2015 年小普查流动人口年龄金字塔的比较

Figure 2 Age Pyramid of Migrants: A Comparison to the 2015 Mini-Census





资料来源: 图中数据为本文根据 2015 年流动人口监测调查、2015 年流动人口流出地调查以及 2015 年小普查数据计算得到, 后文同。

注: 第一个图为 2015 年流动人口监测调查(流入地调查)与 2015 年小普查的比较结果; 第二个图为 2015 年流动人口流出地调查与 2015 年小普查的比较结果。

如图 2 所示, 与 2015 年小普查相比, 2015 年流入地监测调查中 0~15 岁流动人口的比例明显更高; 15~30 岁的比例明显更低; 30~50 岁的比例又明显更高; 而 50 岁及以上的比例则又明显更低。由此可见, 流入地调查的年龄结构与小普查数据截然不同。

相对而言, 流出地调查得到的人口年龄结构与 2015 年小普查结果更为接近。在低年龄段, 流出地调查结果与小普查基本一致, 只是 10 岁左右的几个年龄组略高于小普查。流出地调查得到的 20 和 25 岁左右的女性比例相对低于小普查数据, 而 45~55 岁的男性比例则相对高于小普查数据。在 55 岁以上, 流出地调查与小普查的结果基本相同。

上述比较分析表明, 流出地调查所得的流动人口年龄结构与小普查的结果更为相近, 而流入地调查与小普查之间的差异则更大。如果将小普查数据作为标准, 那么流出地调查得到的人口年龄结构可能相对更为可靠。

4 其他结构的对比分析

除年龄结构以外, 本文还分析了其他几个重要的人口学特征, 以考察各种人口结构在两种调查视角下的异同。

4.1 性别结构

一直以来, 流动人口都存在性别选择性。如 1990 年“四普”时, 全国 5 年间省际移民的平均性别比为 139.12, 高于省内移民(119.27), 更高于非移民(105.40) (张善余, 1992)。2000 年“五普”时, 迁移者的性别比为 111, 未迁移者的性别比只有 107 (唐家龙、马忠东, 2007)。2010 年“六普”和 2015 年小普查时的流动人口性别比分别为 113.46^① 和 111.68。历次调查结果都表明, 我国的流动人口中, 男性略多于女性。

从两种调查视角来看, 基于 2015 年流出地调查数据计算得到的流动人口性别比为 128.80 (男性占比 56.29%), 而基于 2015 年流入地监测调查数据计算得到的流动人口性别比为 109.19 (男性占比 52.50%)。流出地的调查结果远高于流入地。显然, 流入地调查更能捕捉到女性流动人口, 而流出地调查则覆盖更多的男性流动人口。

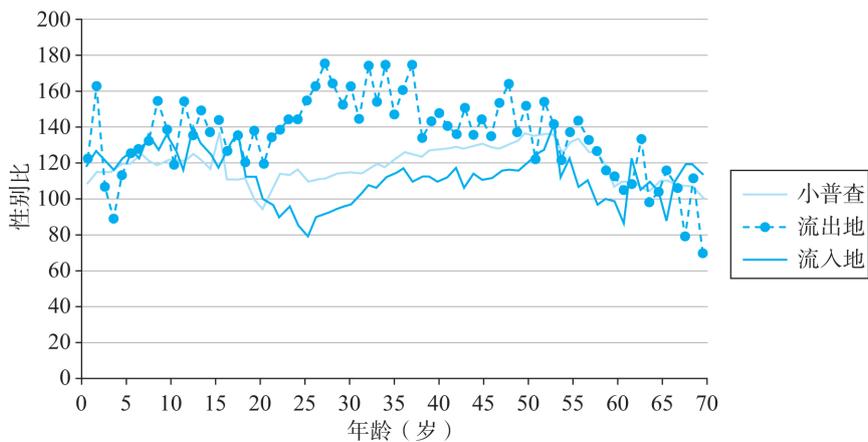
^① 根据第六次人口普查数据集: 男性流动人口为 1.18 亿人, 女性为 1.04 亿人, 分别占 53.15% 和 46.85%。

图 3 展现了两种调查视角下流动人口年龄别性别比的差异以及与小普查数据的比较结果,从中可以看出,尽管年龄别性别比曲线波折起伏,但总体来看,18 岁之前和 53 岁之后,流入地和流出地曲线的形状基本相近^①,二者的主要差别出现在 20~50 岁之间,可以看出,流出地调查的性别比远高于流入地调查。具体来说,在 20~35 岁,流入地调查的性别比相对较低,几乎都在 100 左右,甚至在 26 岁时,性别比低至 80;流出地调查的性别比则基本维持在 140~170 之间,有些年龄段甚至高达 175。这说明,在中间年龄段,流入地调查覆盖到更多的女性,而流出地调查覆盖到更多的男性。

如果与 2015 年小普查的结果相比较,在中间年龄段,小普查结果处于流出地和流入地调查结果之间,基本上稳定在 120~135 且随着年龄的增加缓慢上升。因此,如果以小普查数据为基准,则流入地调查中女性被过度抽样(Over-Sample)了,而流出地调查中则可能是男性被过度抽样了。这种结果仍然可能与流入地的抽样框有关。

图 3 流动人口年龄别性别比:两个调查与 2015 年小普查数据的比较

Figure 3 Age-Specific Sex Ratio of Migrants: A Comparison to the 2015 Mini-Census



4.2 在流入地居住时间

图 4 给出了按照不同流动时间长度划分的流动人口分布情况。两种调查视角下流动人口在流入地居住时间的分布呈现出不同的特点:其一,集中度不同。流出地调查所得结果主要集中在前 3 类,即半年到 1 年、1~2 年以及 3~4 年,3 类比例之和高达 70.89%;流入地调查的结果则主要集中于 1~2 年和 5~9 年。其二,分布趋势不同。流出地调查的结果呈现出明显的随时间延长而比例下降的趋势;流入地调查的结果则呈齿轮状,即低-高-低-高-低的走向,这可能与居住时间为 5~9 年的人群被过度抽样有关。其三,居住时间为半年到 1 年的流动人口在流入地调查中所占的比例远低于流出地调查;居住时间为 3~4 年的流动人口在流入地调查中所占的比例则略低于流出地调查。除这两组以外,其余各组在流入地调查中的比例一致性地高于流出地调查。如果按照这些分布计算平均居住时间,则流入地调查为 5.47 年,流出地调查为 4.56 年,两者几乎相差 1 年。

如果与 2015 年小普查数据^②进行对比可以发现:流入地调查的居住时间分布更接近于小普查,两

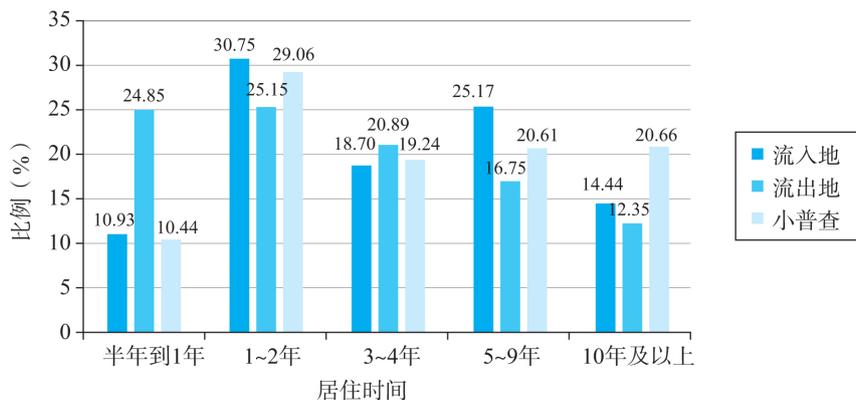
① 为避免高年龄段的年龄别性别比由于样本规模而出现大幅波动,本文将最高年龄设定为 70 岁,即 70 岁及以上的人口统一归入 70 岁组。

② 小普查数据中居住时间的分类标准与流动人口监测调查不同。本文将小普查数据中的“不满半年”和“半年至一年”这两类按照流动人口监测调查数据重新合并,然后将流动人口监测调查数据中 10 年及以上的 4 个类别(“10~14 年”“15~19 年”“20~29 年”和“30 年及以上”)合并成“10 年及以上”。同时,应该注意,流动人口监测调查中 1 年以下的流动人口是以“1 个月”为标准,而小普查则是以“半年及以上”为标准。

者的差异主要出现在 5~9 年和 10 年及以上,流入地调查中居住时间为 5~9 年的比例略高于小普查,而 10 年及以上的比例则低于小普查。流出地调查的结果则有所不同,居住时间为半年到 1 年的比例特别高,且除居住时间为 3~4 年以外,其他时间段的比例均显著低于流入地调查与小普查的结果。这种情况一方面说明,在居住时间的分布上,流入地调查与小普查数据较为接近,并未明显地表现出以往研究所认为的流入地调查的流动人口更稳定,具有更长的居住时间长度;另一方面则反映出,流出地调查更能捕捉到那些短期流动的人群,从而使整个结构发生变化。

图 4 流动人口在流入地居住时间:两个调查与 2015 年小普查数据的比较

Figure 4 The Distribution of Duration of Residence: A Comparison to the 2015 Mini-Census



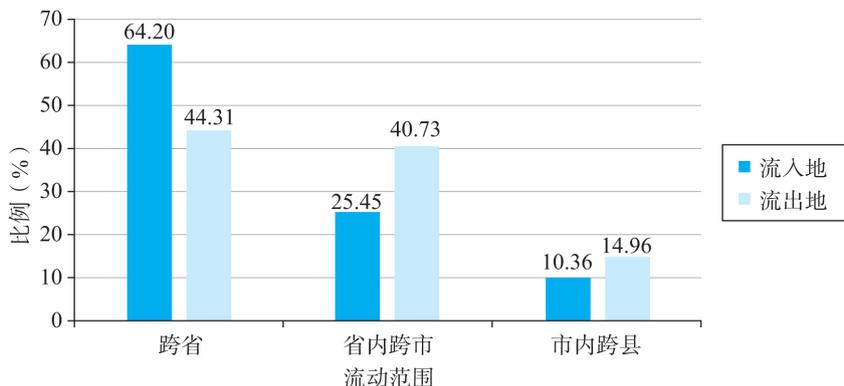
4.3 流动范围

两种调查视角下,流动人口的流动范围亦存在差异(见图 5)。从分布上来看,流入地调查中跨省、省内跨市和市内跨县的比例呈现出阶梯式下降,以跨省流动为主,比例高达 64.20%,而省内流动仅占 36%左右。流出地调查则以跨省与省内跨市两者并齐,如果以跨省和省内流动来区分,则跨省流动占比为 44%左右,而省内流动占比则接近 56%。由此可见,流入地调查更可能覆盖跨省流动的人口,而流出地调查更可能捕获近距离的流动人口。

依据 2015 年小普查汇总资料中的数据计算出的跨省与省内流动比例分别为 39.42%和 60.58%^①。可见,在流动范围方面,流出地调查与小普查的结果较为接近,而流入地调查的结果则正好与小普查相反。

图 5 两种调查视角下流动人口的流动范围

Figure 5 Regional Scope of Migrants from Two Perspectives



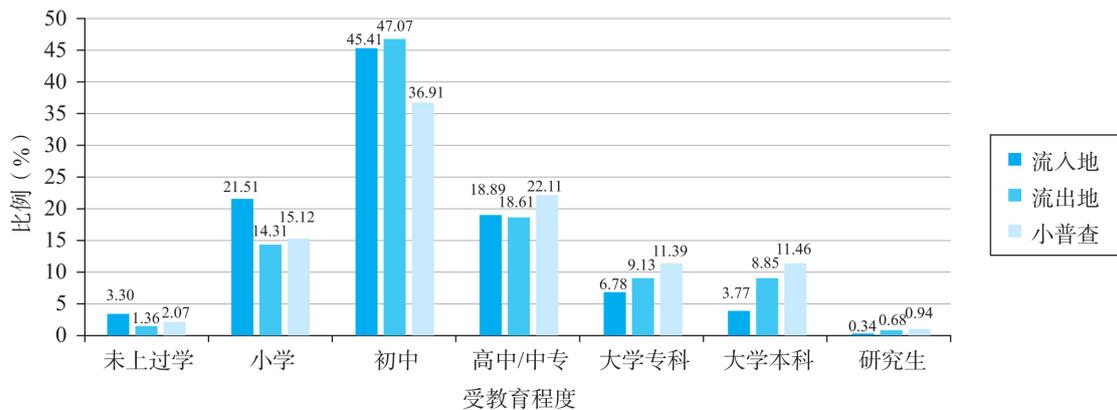
① 根据《2015 年 1%人口抽样调查数据集》表 12-1 中的第一行数据计算得到。

4.4 受教育程度

受教育程度是人口社会经济特征中重要的基础维度之一。尽管两种调查视角下流动人口的受教育程度都以小学和初中为主,但具体分布不尽相同,而且与 2015 年小普查的结果存在较大差异(见图 6)。流入地调查中,受教育程度为小学的流动人口比例远高于流出地调查与小普查数据;初中比例略低于流出地调查,但高于小普查结果。流出地调查中,受教育程度为小学的流动人口比例与小普查接近;初中比例最高,为 47.07%,远高于小普查数据(36.91%)。高中部分,两个流动人口调查的结果相近,但均低于小普查的结果。大专及以上学历的部分,则一致性地呈现为:流入地调查相对低于流出地调查,而两者又都低于小普查的结果,且都呈现出台阶式的分布。因此,从受教育程度来看,不论是流出地还是流入地调查,其结果都更偏向于低受教育程度的流动人口,而相对忽略了高受教育程度的流动人口,特别是大专及以上学历的流动人口。

图 6 流动人口受教育程度:两个调查与 2015 年小普查数据的比较

Figure 6 The Distribution of Education Level of Migrants: A Comparison to the 2015 Mini-Census

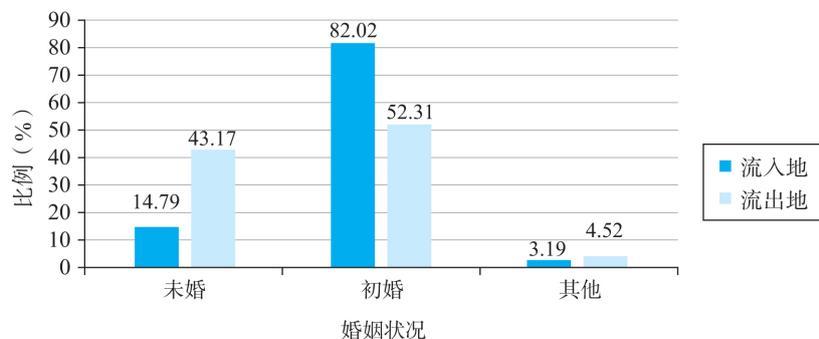


4.5 婚姻结构

两种视角下的调查结果在婚姻结构上都主要集中于未婚与初婚两个类别,其他(包括再婚、离婚和丧偶 3 类)群体比重相对较低。但在未婚与初婚这两类的分布上,两个调查结果正好相反。流入地调查中,初婚人口比例高达 82.02%,未婚比例仅为 14.79%;流出地调查中,未婚比例为 43.17%,初婚比例为 52.31%(见图 7)。这说明,流入地调查偏重于已婚人口,在某种程度上忽略或低估了未婚人口。这种婚姻结构与上述年龄结构存在关联,由于流入地调查以初婚为主,因而年龄结构普遍上移,且儿童部分也有相应的扩大,但未婚的年轻劳动力部分则有被低估的可能,其根本原因可能还是抽样框的问题。

图 7 两种调查视角下流动人口的婚姻结构

Figure 7 The Marriage Structure of Migrants from Two Perspectives



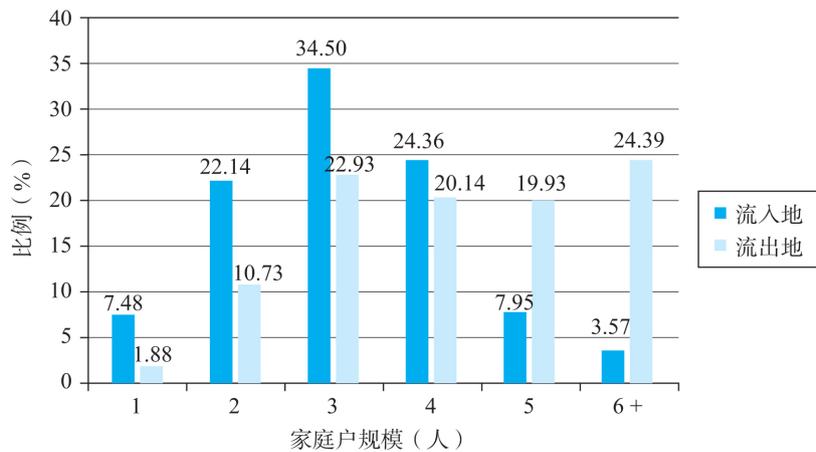
4.6 家庭户规模

以往研究在利用流入地数据讨论家庭因素对人口流动原因的影响时,总是认为在流入地调查得到的家庭结构并不是流动人口做出流动决策时所对应的真正结构,必须用流出地调查得到的家庭结构才有可能作为影响流动决策的真实因素。两种视角下的调查结果(见图8)也恰好说明了这一点。

流入地调查中,超过六成的家庭总人口在3人及以下,近九成的家庭总人口在4人及以下,81%的流动人口家庭户规模在2~4人之间,只有近12%的家庭总人口在5人及以上。流出地调查中,超过六成的家庭总人口在4人及以上。总体而言,5人及以上的家庭户比例,流出地调查远高于流入地调查,但4人及以下的家庭户比例则正好相反,流入地调查远高于流出地调查。可见,流入地调查以核心家庭为主(2~4人),流出地调查以扩展家庭为主。流出地与流入地调查得到的家庭结构存在明显的差异,这一点在以后的分析过程中需特别注意。

图8 两种调查视角下流动人口的家庭户规模

Figure 8 The Distribution of Family Size from Two Perspectives



5 总结与讨论

5.1 分析结果总结

从流出地与流入地两种视角下调查得到的流动人口,在各种结构上存在着明显的差异:第一,在年龄结构方面,流入地调查捕获到更多的流动儿童,但在15~30岁组,流出地调查的比例远高于流入地调查结果,且整体来看流出地调查的结构与小普查结构基本相似;第二,在性别结构方面,以小普查数据为基准,则流入地调查中女性被过度抽样,而流出地调查中,20~50岁的男性比例相对较高;第三,从在流入地的居住时间来看,流出地调查更能捕捉到短期流动人口,而流入地调查更能捕捉到流入时间较长(5~9年)的流动人口,相对而言,流入地调查所得的居住时间分布与小普查更为接近;第四,在流动范围方面,流入地调查中省际流动的比例相对较高,流出地调查结果与小普查更接近;第五,在受教育程度方面,流入地和流出地调查都更偏向于低受教育程度的流动人口,如果以小普查为标准,二者都相对忽略了大专及以上学历的流动人口;第六,在婚姻结构方面,流入地调查中初婚部分显然被过度抽样了(高达82.02%),而未婚明显偏少;第七,在家庭户规模方面,流出地调查的家庭户规模明显大于流入地调查,二者的家庭结构存在显著差异。

比较结果表明,尽管在某些特征(如年龄结构、流动范围等)上,流出地调查的结果与小普查更为接近,但在另一些特征(如居住时间)上,却是流入地调查结果与小普查更为接近。在受教育程度上,以小普查为标准,两种调查结果都存在着较大的偏差。因此,我们无法得到有关两种视角下调查结果孰高孰低的一致性结论,只能说各有各的优势与劣势。

5.2 结构性差异的理论探讨

理论上讲,如果排除各种调查误差,那么针对同一总体(如流动人口)的抽样调查结果在某种程度上应该具有一致性,至少偏差不会太大。但本研究的比较结果表明,两种调查视角下流动人口的结构存在着显著差异。这说明两个调查中至少有一个存在偏差,甚至于两个同时存在偏差。我们既应该注意这种结构性偏差的可能来源,也应该注意结构性偏差对分析结果的可能影响,以及其对基于分析结果所得的相关结论与政策建议的可能影响。

以往许多研究已经注意到流入地调查中存在抽样框问题(齐嘉楠等,2014;庄亚儿、李伯华,2014)、样本选择性问题、追踪调查难问题和系统误差问题(梁玉成等,2015)。亦有研究注意到了流出地调查的相对优势:如流出地调查的“总量优势”(吕利丹、段成荣,2012);流出地调查更具可操作性、更能保证概率抽样原则、更适用于对农民工整体结构的研究(朱磊,2014);有充分的农村流动人口代表性、能有效解决样本选择性问题、能对农村流动人口开展追踪调查、有更高的访问应答率(梁玉成等,2015)。当然这些研究也都指出,两种视角下的流动人口调查都存在着各自的问题(朱磊,2014;梁玉成等,2015)。只是这些研究未能较为系统地讨论调查过程中的各种误差来源对调查结果的可能影响。

事实上,抽样调查的误差不仅仅来自覆盖误差,特别是研究总体与目标总体、抽样框总体等之间的差异,同时也来自测量(包括效度、测量误差和过程误差)与代表性(除覆盖误差以外的抽样误差、无应答和调整误差)等抽样调查过程(Groves et al., 2009)。因此,下文将主要从应答者、测量、代表性及其他这4个维度,讨论两种视角下的流动人口调查结果的差异来源,这实际上也反映了两种调查的优势与劣势(见表1)。

表1 流入地与流出地调查的优势与劣势

Table 1 The Advantages and Disadvantages of Survey on Migrants in Origin and Destination

维度	误差类型	流入地		流出地	
		优势	劣势	优势	劣势
应答者	—	本人应答	—	—	家人代答
	效度	相对较高	—	—	相对较低
测量	测量误差	利于了解现状类问题	不利于了解流出原因类问题	利于了解流出原因类问题	不利于了解现状类问题
	过程误差	两者相同			
	覆盖误差	—	地址或地图抽样困难,抽样框名单信息不全	可选用地址或地图等抽样方式	—
代表性	抽样误差	两者相同,仅受样本规模的影响			
	无应答	相对较低	—	—	相对较高
	调整误差	—	需注意权重设计	—	需注意无应答权重的调整和计算
其他	研究视角	以城市为中心	—	以农民工为中心	—
	研究问题	变量优势	无法了解总体规模	总体优势	无法深入了解后果
	问卷内容	流出后的状况及后果	无法了解原因	流出原因	无法了解当前状况
	样本选择性	—	需注意抽样框带来的样本选择性问题	相对较低	需注意无应答对样本代表性的损害
	追踪的难易	—	再流动后无法追踪	跑得了和尚跑不了庙	无法追踪举家流动

首先是应答者方面。两种视角下的调查对象虽然都是流动人口,但被访问的应答者(Respondent)是完全不同的。流入地调查的应答者确实可以是流动人口本人(至少可以是当前和流动人口本人共同居住、生活的家庭成员),但流出地调查的应答者则只能是流动人口本人在流出前所属家庭中的其他成员。这种应答者的不同对调查结果的影响将体现在调查过程的各步骤中,特别是在效度、测量误差、无应答等方面。事实上,应答者的不同是导致调查误差的重要根源。

其次是测量方面。测量包括了效度、测量误差和过程误差。过程误差是指数据录入等过程中产生的误差,在各种调查中都可能存在,在此不再讨论。效度是指测量工具能够测出所需测量的概念的准确程度。影响效度的主要是测量工具,即问卷设计(如问题的内容、提问方式、选项设置等),它与测量误差紧密关联。测量误差则是测量结果与真实值之间的差异,包括偏差(Bias)和方差两部分,它既与测量工具相关(如问题的提问方式、前后顺序等),也与调查过程相关(如访员的特征与行为、被访者的理解等)。在样本规模一定的情况下,测量误差更多是由偏差导致(Groves et al. 2009)。

效度方面,对于流动人口调查,由于应答者不同,问卷设计过程中,每道问题的提问方式及选项为适应被访者而需要做出相应的调整,进而可能影响到该问题的效度。最直接的就是流入地调查中的“您…”必须改成流出地调查中的“他…”。此外,留在流出地的家庭成员的受教育水平可能更低,问题的提问方式需要改用更通俗易懂的话语。这些改变都可能会影响到效度。

同时,应答者的不同可能会带来更大的偏差,进而扩大测量误差。在流出地调查中,一来应答者(流出地的原家庭成员)可能无法真正了解流动人口在流入地的真实状况;二来由于留在流出地的家庭成员的受教育水平可能更低,在“理解-回忆-调整-应答”这种应答的逻辑过程中,流出地的被访者可能无法准确理解问题的真实意义,且回忆过程也可能会有偏差,进而产生应答偏差及各种测量误差。比如流动时间,两个调查都规定了离开户籍地1个月以上为流动人口,并同时要求填写具体的流动时间。前者可能是在“1个月”的时间界定上存在理解偏差;后者则可能会由于回忆而导致偏差。因此,测量误差与应答者是完全相关联的。当然,这种测量误差同样也会出现在流入地调查中,但相对而言,流出地调查可能更为严重。

再次是代表性方面。数据的代表性主要包括覆盖误差、抽样误差、无应答和调整误差4个方面。其中,覆盖误差是抽样框总体与研究目标总体之间的差异;抽样误差取决于具体抽样方法与样本规模;无应答包括无联系(No Contact)、单位无应答(Unit Nonresponse)和选项无应答(Item Nonresponse)等;调整误差则主要源自对权重的计算与调整。

针对流动人口调查中的覆盖误差已经有了许多讨论,涉及从抽样框的建立到抽样过程等各方面。事实上,研究者从抽样框的建立到抽样方法(如从随机抽样到结合GIS抽样,再到非随机的滚雪球抽样等)上都做出了不懈的努力,以求降低覆盖误差并使调查结果更为精确。但本文仍要提请注意,由于流动人口在流入地的分布是不均匀的(周婕等,2015),所以,在流入地调查中试图以地址或地图的方式建立抽样框并进行随机抽样的方式可能都是徒劳的。以现有的各部门流动人口登记信息作为抽样框则必然存在大量的未登记流动人口所带来的未覆盖现象(即未包含合格单元),以及由于信息系统更新滞后而导致的包含不合格单元的问题,进而使被调查的群体具有某种明显的样本选择性,如上文比较结果显示的流入地已婚女性被过度抽样以及流入地低龄儿童比例相对较高的情况。因此,从覆盖误差的角度来看,流入地调查相对会更为困难。流出地调查尽管也可能由于迁移流动的选择性(如流动人口家庭或个体特征的选择性)使结果存在偏差(如上文比较结果所显示的流出地男性劳动年龄人口相对较多的情况),但相对于流入地而言,至少能够较好地了解流动人口在总人口中的比例、规模等,这是“总量优势”(吕利丹、段成荣,2012)的体现。当然,同样不能忽视流出地调查所得的某些结构可能仍然由于覆盖误差而导致有偏的情况。

无应答是影响代表性的另一个重要因素,单位无应答与选项无应答是“无应答”的两种重要形式。目前各类社会调查的无应答率都相对较高,且随时间推移而提高(Meyer et al., 2015)。流动人口调查也存在同样的问题,但流出地调查的无应答率可能会相对更高(其根本原因还是在于应答者的不同)。从单位无应答来看,由于流动人口举家流动的可能性在逐步增大,因而流出地调查抽到举家流出的可能性也在增加,从而使流出地调查所得的流动人口比例可能会被低估,进而形成样本选择性并导致调查结果存在偏差。再者,由于应答者的不同,导致被访者可能会拒绝流动人口调查的访问,从而造成另一种形式的单位无应答。这部分误差在最后的結果中事实上会被样本替换所掩盖,因而相对比较难以测量。

除单位无应答以外,选项无应答同样是比较严重的问题。特别是在流出地调查中,应答者并不了解被访者的真实状况,无法真实地回答调查问题,从而使调查数据存在一定规模的缺失值。这是另一种隐性的无应答。在数据分析过程中,通常最简便的处理缺失值的方法就是将所需变量中有缺失值的案例全部排除在分析数据之外,从而会使分析样本(及其对应的总体)的各种结构与完整调查样本(及其对应的总体)之间存在一定差异。毕竟这种选项无应答并非完全随机的缺失,因此,从研究总体到抽样框总体,到调查样本总体,再到剔除无应答后的分析样本总体,由于误差的层层叠加,导致分析样本总体与研究总体在结构上可能存在极大的差异。这种结构性差异对分析结果可能产生重要的影响(Berk, 1983)。已有研究大都有意无意地忽略了由此产生的偏差及由此带来的对分析结果的影响,而未予以足够的重视。本研究认为,对缺失案例(变量)的处理应该予以高度重视。

调整误差的来源主要在于权重。既然流入地调查的抽样框和概率抽样过程存在质疑,那么其权重同样可能存在问题。流出地调查则应更注意无应答权重的调整,同时也应该特别注意无应答权重的计算,因为无应答误差可能存在选择性问题。

最后,本文试图在上述讨论的基础上,归纳总结并重新审视两种视角下的流动人口调查在其他方面的优劣势。

从研究视角看,朱磊(2014)将流入地与流出地调查区分为以城市为中心和以农民工为中心两种立场,体现了研究者的价值取向与研究立场。但从流动人口(或被访者)自身的角度看,研究视角需要与研究问题结合:流入地调查更侧重于流动的后果;流出地调查更侧重于流动的原因。这体现了“变量优势”与“总体优势”的区别。例如,上文分析显示,流出地调查所得的流动人口年龄结构、流动范围等与小普查的结果更为吻合,而流入地调查中的居住时间分布则与小普查的结果更为吻合。再如,以往的许多研究几乎都是利用流入地调查数据,以流入地当时的各种状况作为解释变量,讨论人口流动的原因,如家庭结构对流动倾向的影响,而流入地当时的各种状况(如当时的家庭结构)从根本上来讲并不是促使人口流出的真正原因,其真正的原因应该是流出地当时的家庭结构。从这个角度看,流出地调查有其独特的优势。因此,尽管研究视角与研究问题决定了问卷内容,但同样需要考虑相应的调查地点。

对样本选择性的讨论,以往文献更关注于流入地的抽样问题(如抽样框、抽样方法、抽样过程等),但事实上,样本选择性也与无应答有关。特别是在流出地调查中,由于访问通常是代答形式,从而使无应答,特别是选项无应答成为关键因素。同时,作为无应答的另一种形式,无法联系也将会由于“举家流出”原居住地而造成无应答。因此,尽管流出地调查的样本代表性可能会更好,但由于无应答的存在不可避免地会影响到样本选择性。归纳起来,流入地调查需要注意抽样(如抽样框)所可能带来的样本选择性问题,而流出地调查则更需要注意无应答对样本代表性的损害。从本质上来讲,样本代表性的损害(或不足)集中体现了抽样与调查实施过程中的所有误差。

当前越来越多的调查重视追踪。追踪调查作为纵贯数据有其独特的优势(任强、谢宇, 2011)。流

动人口的研究同样需要基础性的追踪调查数据。但在流入地进行追踪调查会由于流动人口的“流动性”而导致流失(Attrition)样本的比例相对较高。更关键的是,流失样本并不是随机的,而是带有显著特征的,这种非随机的流失显然会造成追踪样本代表性的损害。尽管流出地看似有利于追踪调查的开展,所谓“跑得了和尚跑不了庙”,但正如上文所讨论的,流出地的追踪调查同样会面临着由于“举家流出”而导致的样本流失,且这种流失同样也是有选择的而非随机的。因此,在针对追踪的案例进行分析时,需要特别注意这种非随机的样本流失及其对分析结果的可能影响。

5.3 建议

流动人口的调查本来就是极其困难的,能够开展流动人口调查实属不易。作为研究人员,在利用调查数据之前,应该首先了解数据的来源、调查过程等,并关注数据的代表性及其对分析结果的可能影响。从调查方法论的角度看,流出地与流入地的调查结果之间的差异是各类调查误差共同作用的结果。我们应该注意到,不同的调查误差在不同调查地点的实施过程中可能产生的作用与结果是不同的,相应的,在不同调查地点的实施过程中应该分别予以重视与规避。

同时,我们更应注意到近年来流动人口内部结构的变化,特别是流出地的城乡比例的改变(朱宇等 2016)。如果以农村为纯粹的流出地来进行调查,则结果可能会存在一定的偏差,特别是随着城-城流动人口和家庭式流动人口比例的增加。因此,未来的流出地调查应该更加强调“城乡统筹”,即同时将城市与农村视为流出地。这从某种意义上也就意味着不再区分流出地与流入地,而是合并流入地与流出地,将全国视为一个总体,调查过程则应同时登记流入人口与流出人口。

从整体来看,流入地调查由于抽样框等原因而使流动人口结构可能存在偏差,因而在利用流入地调查数据时应通过调整其结构来尽量降低代表性问题;流出地调查尽管在样本代表性上可能相对较好(某些结构同样也会存在偏差),但由于其根本的“家人代答”的形式可能会带来更大的测量误差和无应答问题,因而在利用流出地调查数据时应充分考虑测量误差和无应答等问题,并运用相应的统计方法予以一定的调整。总之,在使用与分析包括流动人口在内的各种调查数据时,应该充分重视调查过程中的各种误差,特别关注样本与总体间的结构性偏差及其对分析结果的可能影响。

参考文献/References:

- 1 Berk, Richard A. 1983. An Introduction to Sample Selection Bias in Sociological Data. *American Sociology Review* 3: 386-398.
- 2 Groves, Robert M., Fowler, Floyd J., Couper, Mick, Lepkowski, James M., Singer, E. and Tourangeau, R. 2009. *Survey Methodology* (Second Edition). John Wiley & Sons 3: 10-11.
- 3 Meyer, Bruce D., Mok, Wallace K. C. and Sullivan, James X. 2015. Household Surveys in Crisis. *The Journal of Economic Perspectives* 4: 199-226.
- 4 Rogers Andrei and Castro Luis J. 1979. Migration Age Patterns: II. Cause-Specific Profiles. Working Paper of the International Institute for Applied Systems Analysis.
- 5 陈传波,白南生,赵延东. 适应性区群抽样: 研究流动农民工的方法与实践. *统计研究*, 2012; 5: 100-104
Chen Chuanbo, Bai Nansheng and Zhao Yandong. 2012. Adaptive Cluster Sampling: Research of Migrant Population. *Statistical Research* 5: 100-104.
- 6 郭志刚. 近年生育率显著“回升”的由来——对 2006 年人口和计划生育调查的评价研究. *中国人口科学*, 2009; 2: 2-15
Guo Zhigang. 2009. On the Notable “Bounce” of Fertility Rates in Recent Years: Evaluation on the 2006 National Population and Family Planning Survey. *Chinese Journal of Population Science* 2: 2-15.
- 7 梁玉成,周文,郝令昕,刘河庆. 流出地调查法: 农村流动人口调查的理论与实践. *华中科技大学学报(社会科学)*

- 版), 2015; 4: 113-123
Liang Yucheng, Zhou Wen, Hao Lingxin and Liu Heqing. 2015. The Origin Approach for Rural Migration Survey: Theory Versus Practice. Journal of Huazhong University of Science and Technology 4: 113-123.
- 8 吕利丹, 段成荣. 对我国流动人口统计调查的总结与思考. 南方人口, 2012; 3: 73-80
Lv Lidan and Duan Chengrong. 2012. An Evaluation of China's Floating Population Survey and Policy Suggestions. South China Population 3: 73-80.
- 9 齐嘉楠, 覃民, 李伯华. 流动人口监测调查抽样设计的思考. 统计与决策, 2014; 3: 14-16
Qi Jianan, Qin Min and Li Bohua. 2014. Considerations on Sampling Design of China's Migrant Monitoring Survey. Statistics and Decision 3: 14-16.
- 10 任强, 谢宇. 对纵贯数据统计分析的认识. 人口研究, 2011; 6: 3-12
Ren Qiang and Xie Yu. 2011. Statistical Analysis of Longitudinal Data. Population Research 6: 3-12.
- 11 沈明明, 李磊. 流动人口、覆盖偏差和 GPS 辅助的区域抽样方法. 理论月刊, 2007; 6: 5-15
Shen Mingming and Li Lei. 2007. Reaching Migrants in Survey Research: The Use of the Global Positioning System to Reduce Coverage Bias in China. Theory Monthly 6: 5-15.
- 12 唐家龙, 马忠东. 中国人口迁移的选择性: 基于五普数据的分析. 人口研究, 2007; 5: 42-51
Tang Jialong and John Ma. 2007. Migration Selectivity in China: An Analysis Based on the 2000 Census Data. Population Research 5: 42-51.
- 13 王广州, 王天忆. 流动人口监测抽样调查数据质量评估. 载于《对话: 人口流动迁移与数据开发共享——第一届新型城镇化与流动人口社会融合论坛演讲论文选编》. 中国人口出版社, 2015; 7: 72-82
Wang Guangzhou and Wang Tianyi. 2015. Quality Assessment of Sample Survey Data of Floating Population Monitoring. In Department of Floating Population. Dialogue: Migration and Collaborative Data Mining. China Population Publishing House 7: 72-82.
- 14 杨云彦. 人口迁移年龄模式的分解与合成. 人口研究, 1992; 4: 15-22
Yang Yunyan. 1992. Decomposition and Composition of Age Specific Pattern of Migration. Population Research 4: 15-22.
- 15 张善余. 第四次人口普查省际迁移数据分析. 人口与经济, 1992; 3: 13-19
Zhang Shanyu. 1992. Inter-provincial Migration in China: An Analysis Based on the 1990 Census Data. Population and Economics 3: 13-19.
- 16 赵延东, Jon Pedersen. 受访者推动抽样: 研究隐藏人口的方法与实践. 社会, 2007; 2: 192-205, 208
Zhao Yandong and Jon Pedersen. 2007. Respondent-Driven Sampling: Method and Practice of Studying Hidden Populations. Society 2: 192-205, 208.
- 17 周婕, 罗逍, 谢波. 2000~2010 年特大城市流动人口空间分布及演变特征——以北京、上海、广州、武汉等市为例. 城市规划学刊, 2015; 6: 56-62
Zhou Jie, Luo Xiao and Xie Bo. 2015. Spatial Distribution and Variation of Floating Population in Megacities: Case Study of Beijing, Shanghai, Guangzhou and Wuhan. Urban Planning Forum 6: 56-62.
- 18 朱磊. 流入地抽样抑或流出地抽样? ——对当前农民工研究中抽样方法的评析. 青年研究, 2014; 1: 87-93, 96
Zhu Lei. 2014. Inflow Sampling or Outflow Sampling? Comments on Sampling Methods in the Study of Rural Migrant Workers. Youth Studies 1: 87-93, 96.
- 19 朱宇, 林李月, 柯文前. 国内人口迁移流动的演变趋势: 国际经验及其对中国的启示. 人口研究, 2016; 5: 50-60
Zhu Yu, Lin Liyue and Ke Wenqian. 2016. Trends in Internal Migration and Mobility: International Experiences and Their Implications for China. Population Research 5: 50-60.
- 20 庄亚儿, 李伯华. 流动人口调查抽样的实践与思考. 人口研究, 2014; 1: 30-36
Zhuang Yaer and Li Bohua. 2014. Reflections on Sampling of Migrants Surveys. Population Research 1: 30-36.