追踪调查中的追踪成功率研究*

——社会转型条件下的追踪损耗规律和建议

梁玉成

提要: 本研究将追踪调查的追踪结果成功与否作为研究对象,借助生态学中的标定再捕获理论,使用贝叶斯统计方法,将追踪成功概率分解为样本存活概率和存活样本的追踪成功概率。本文将该方法应用于 1989 – 2009 年 8 轮的追踪调查数据(CHNS) ,结果从家庭层面和个体层面,均证实了作者提出的中国追踪调查中追踪损耗的城乡差异假设、生命周期假设和社会转型假设。最后作者提出了减少追踪损耗的有关方法和建议。

关键词: 追踪调查　追踪损耗　标定再捕获　样本生存概率　存活样本追踪成功概率

一、前　言

追踪调查(panel survey) 作为数据收集的手段,其区别于横剖面调查(cross-sectional survey) 的优点在于研究者通过在不同时间点重复观测同一群体,可以连续收集到被观察者随时间发展的行为和心理的连续变化数据,从而可以分析被研究者的行为演变模式,以及各类现象之间的因果机制(风笑天,2006) 。

中国社会学界开展的调查,一直以横截面的时点数据为主,连续多年并广为学界使用的经典调查有中国人民大学社会学系与香港科技大学社会科学部合作的“中国综合社会调查(CGSS) ”、中国社会科学院

---

* 本文初稿曾在北京大学中国社会科学调查中心主办的“社会调查年会·2010”,以及香港中文大学社会学系举办的“2011 年珠三角社会研究研讨会”上进行报告。本文得到了中国卫生部疾控中心张兵教授的帮助,霍普金斯大学郝令昕教授对本文初稿给予了详细评论,与密西根大学谢宇教授的讨论使作者得到很大启发,匿名评审人的建议使本文的研究品质得到了提升,在此表示感谢。当然,本文的任何错误均由作者负责。本文受到中山大学“211”、“985”专项基金及 2009 年度高校基本科研业务费中山大学青年教师培育项目的资助。本研究是教育部 2010 年度人文社会科学一般项目(项目编号: 10YJA840021) 的阶段性成果。

社会学研究所的“中国社会状况综合调查(CSS)”等。近年来,随着研究的深入,中国学术界也开展了多项追踪调查,如中国疾病与预防控制中心与北卡罗来纳大学人口研究中心合作开展的“中国健康与营养调查(China Health and Nutrition Survey, CHNS)”,北京大学社会科学调查中心主持的“中国家庭追踪调查(CFPS)”,中山大学社会科学调查中心主持的“中国劳动力动态追踪调查(CLPS)”等。

在中国进行的追踪调查大致可以分为两种类型:一类是对家庭户/家庭成员进行跟踪,跟踪范围不限——无论家庭成员迁移到国内任何地点,均对其进行追踪调查,这类调查以北京大学中国社会调查中心的中国家庭动态调查为典型代表;另一类是对家庭户/家庭成员进行跟踪,但跟踪范围限定在被调查社区——只有被调查的家庭成员迁移没有离开本社区才对其进行追踪调查,这类调查以中国卫生部疾控中心和北卡罗来纳大学联合开展的中国营养健康调查为典型代表。

虽然多项追踪调查正在开展,但由于追踪调查在国内刚刚起步,因而关于追踪调查本身的理论和研究较少,目前这方面基础性的工作,可查询到的仅有风笑天教授(2006)关于追踪调查的概括性综述、郝令昕教授等(2010)关于发展中国家追踪调查的研究性综述,以及邱泽奇教授主持的“中国家庭动态追踪调查”项目组(2011)关于追踪调查中家庭拒访行为的研究。郝令昕等指出,由于中国社会正处于工业化、市场化和城市化的多重急剧变迁之中,在中国实施追踪调查所面临的困难和问题远较西方社会为多,而西方社会是在完成转型之后兴起的追踪调查,因此其经验未必适合处于转型条件下的发展中国家。追踪调查由于需要长期追踪,其成本远较横截面调查为高,一旦设计不当,偏误又会随时间的发展而不断放大。如果没有充足的在中国开展追踪调查的理论和实践研究为基础,只是照搬西方相关追踪研究经验,不但极易造成严重损失和浪费,还可能导致基于这类数据的研究结论产生严重偏误。因此,开展对追踪调查的研究,尤其是在中国积累追踪调查所应注意的本土社会学知识,是社会学研究方法领域中的当务之急。

二、追踪调查中的追踪损耗

在追踪调查中,最重要的调查质量问题是追踪损耗问题。一般而

言,追踪调查由首轮(first wave)调查和追踪调查(consecutive waves)(一次或多次)两部分构成。首轮调查样本的设计,要对研究的目标总体(target population)有很好的代表性。然而在后续调查中,样本会不可避免地出现追踪缺损(panel attrition),即追踪对象的流失。也就是当样本在第一次被征召参与调查之后的后续调查轮次中,部分样本会因各种原因而无法被追踪到。风笑天(2006)认为,追踪损耗或追踪对象流失,指的是在进行第二次(或第 N 次)调查时,第一次调查时的全部对象中有一部分因去世、搬迁、拒绝再次调查等原因而无法追踪到。风笑天报告的两个追踪调查一个在 3 年后的第二轮调查时的追踪损耗为 30% 左右,另一个在 2 年后的第二轮调查中的追踪损耗为 33.6%。

追踪调查中的追踪损耗往往不是随机发生的,而是一个包含由拒绝追踪、去世、搬迁、分家等因素构成的样本选择(sample selection)过程(Van den Berg & Lindeboom,1998; Peracchi & Welch,1995; Fitzgerald et al.,1998)。很明显,如果追踪损耗严重,也就是样本选择性严重(non-ignorable sample selection),追踪到的对象都是具有某种特征的群体,就会使得基于这种选择性的追踪样本的研究结果不能正确反映原始总体的变化情况,从而影响对现象之间因果关系的推断。

最近十年来,有学者提出(Hirano et al., 2001; Bhattacharya, 2008),后续轮次的追踪样本的代表性不应以其对首轮的目标总体的代表性为标准,而应以该轮次执行时的目标总体的横截面代表性(representative cross-section of the target population)为标准。在此基础上,金(Kim,2009)将缺损划分为两种类型:样本缺损(sample attrition)和总体缺损(population attrition)。样本缺损是指总体组成部分中的部分样本无法进入到收集的样本中;总体缺损是指第一轮次中总体组成部分中的部分样本不再存在于以后轮次的总体中。根据平野等(Hirano et al.,2001)和巴达恰亚(Bhattacharya,2008)的观点,总体缺损是可以接受的,并不会导致样本选择问题。例如,我们追踪一组 60 岁以上的老人,10 年后的一次追踪调查,如果某些老人还在世,但无法追踪到,就属于样本缺损;如果其中一部分老人已经不在世了而无法追踪到,就属于总体缺损。如果在此时点进行一个横截面调查,那些在世的老人和其他老人有同样的概率进入样本,而去世的老人进入样本的概率为零。因此第 2 轮追踪样本虽然也无法抽到去世的老人,但由于此时的横截面样本也无法抽取到这部分老人,所以并不损害其对横截面样本的代

表性，但样本缺损则会降低追踪样本对横截面样本的代表性。这样，当存在总体缺损的情况下，第 2 轮追踪调查的总体是第 1 轮追踪调查时总体的一个非随机子集(nonrandom subset)。因此第 2 轮的横截面数据也不可能对第 1 轮时点的总体有代表性。换句话说，随着时间变化，在以后轮次的追踪调查中，社会实体已经发生了变化，我们无法要求追踪样本仍然反映没有变化之前的社会实体。所以，追踪损耗中的主要问题是样本缺损，而非总体缺损。

根据以上追踪损耗理论，在中国进行追踪调查的追踪损耗情况如何？有多少是样本损耗，多少是总体损耗，它们的演变规律怎样？中国的多重社会转型如何影响追踪损耗？如何提高追踪调查质量，怎样设计好的抽样方案和样本替换方式来应对追踪损耗？这些都事关追踪调查的成败，是迫切需要回答的问题。

幸运的是，在中国有一个已经延续多年的追踪调查——中国健康与营养调查(CHNS)，该调查是由中国疾病与预防控制中心、北卡罗来纳大学人口研究中心和美国国家营养与食物安全研究所(The National Institute of Nutrition and Food Safety)合作开展的调查项目。该调查旨在检验健康、营养和计划生育政策的影响以及研究中国经济社会转型如何影响整个人口的健康和营养状况。到目前为止，该调查一共进行了 8 次，分别是 1989、1991、1993、1997、2000、2004、2006 和 2009 年。本研究试图通过分析该数据回答上文提出的研究任务和研究问题。

三、CHNS 的追踪损耗情况

我们首先分析 CHNS 的追踪情况。以下是 8 个追踪轮次、不同省份的样本分布(见表 1)。由于省份 21 在 1997 年中断之后，于 2000 年重新抽取样本，所以我们将其剔除在研究范围之外。省份 23 在 1997 年进入样本，不需要剔除。

我们在删除了 21 这个省份之后，仅仅以 1989 到 2000 年的 5 次追踪情况为例(之所以不一直分析到 2009 年，在于 2000 年的总样本量为 14891，而到 2004 年下一轮次时总样本量陡然下降到 11555)。我们分析了 4 种追踪情况：第一次观测、此后轮次的追踪调查中连续观测到、曾经观测不到后又重新被观测到，以及进入追踪样本后不能再次被观

测到(见表2)。

表1 CHNS各省样本数量一览表(个体)

年份 / 省份	1989	1991	1993	1997	2000	2004	2006	2009	合计
21	1514	1433	1347	0	1541	1232	1170	1120	9357
23	0	0	0	1473	1451	1228	1159	1104	6415
32	1543	1499	1494	1630	1637	1261	1200	1319	11583
37	1638	1581	1515	1564	1459	1150	1189	1195	11291
41	1924	1806	1720	1816	1717	1440	1285	1352	13060
42	1823	1801	1729	1796	1742	1227	1104	1118	12340
43	1795	1607	1649	1611	1580	1194	1279	1265	11980
45	2070	1952	1964	1963	1959	1473	1445	1574	14400
52	2041	1997	1904	1922	1805	1350	1330	1250	13599
合计	14348	13676	13322	13775	14891	11555	11161	11297	104025

表2 1989–2000年CHNS追踪类别分布表(个体)

	第一次观测	连续观测	中断后又被观测到	应观测而未被观测到(样本损耗)	当年次调查合计总人数
1989(开始年份)	12951				12951
1991	1072	11315		1636	12387
1993	888	10848	345	2830	12081
1997	4175	9268	445	5198	13888
2000	1511	11099	760	7227	13370

注:中断后又被观测到的损耗也称为非单调损耗(nonmonotone attrition)(Sergi & Perac-chi,2002),应观测到而未被观测到的损耗称为单调损耗(monotone attrition)。

从表2不难看出:(1)曾经观测不到后又重新被观测到的比例很低,大致有样本损耗数的10%;(2)样本损耗规模逐年增加,从1991年第2轮追踪开始到2000年,样本的总和追踪成功率为62.13%,每一轮次的几何平均成功追踪率为 $\sqrt[4]{1-\frac{7227}{11099+760+7227}}=0.888$,因

此，样本轮次几何平均损耗率为 11.2%。①

为进一步搞清追踪损耗的演变细节，我们根据个体样本的开始追踪年和结束追踪年，将各个年份进入和结束追踪的样本进行交互分类分析，得到其分布状况，见表 3。

表 3　　1989 – 2009 年 CHNS 追踪损耗一览表(个体)

		结束年								
		1989	1991	1993	1997	2000	2004	2006	2009	合计多年招募总数
开始年	1989	8.46%	8.27%	14.47%	8.25%	14.20%	4.89%	7.96%	33.50%	12819
	1991		13.17%	12.78%	4.45%	32.72%	8.52%	8.81%	19.55%	1033
	1993			17.32%	5.52%	26.08%	17.86%	12.66%	20.56%	924
	1997				19.42%	19.08%	9.31%	12.79%	39.40%	4145
	2000					25.81%	11.84%	18.92%	43.42%	1596
	2004						25.87%	23.45%	50.68%	1612
	2006							35.10%	64.90%	943
	2009								100.00%	1695
	合计当年结束总数	1085	1196	2147	1959	3602	1872	2769	10137	24767

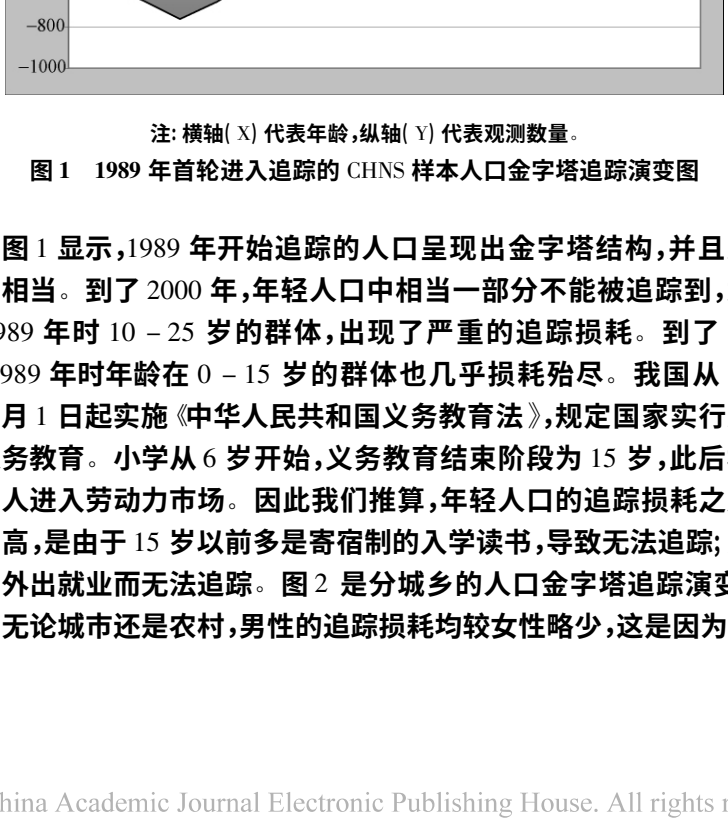
注：2009 年不是结束年，在该年结束的样本，在未来会继续被追踪。

对表 3 的仔细分析，可以发现追踪损耗的一些规律：(1) 随着时间的延续，越是后面进入的样本，其追踪缺损的比例越高(以当年次即追踪结束年份的比例计算)，这表明在中国追踪调查的难度越来越高；(2) 总体而言，进入追踪的样本的追踪损耗会逐年下降；(3) 一次不好的追踪调查安排，可能导致历次进入的样本均同时出现高的追踪缺损

① 此处计算的数字来源为表 2，其中 2000 年第一次观测的个案由于还未被追踪，因此不需要计算；因此，用 2000 年的应观测而未被观测到的个案数(7227) 代表损耗，其与 1989 到 2000 年期间连续观测的个案数(11099) 与中断后又被观测到的个案数(760) 一起构成了总体的个案数(19086)。当然，计算时应该将 1989 年观测到的 12951 人次到 2000 年的损耗情况加以计算才更为合理。本处的计算方法是一个简便方法，希望包含各轮次的情况，存在误差。提醒读者注意。

(2000 年后历次进入的样本均出现高追踪损耗);(4) 一旦出现(3) 的情况,会导致次年的追踪损耗也被提高。

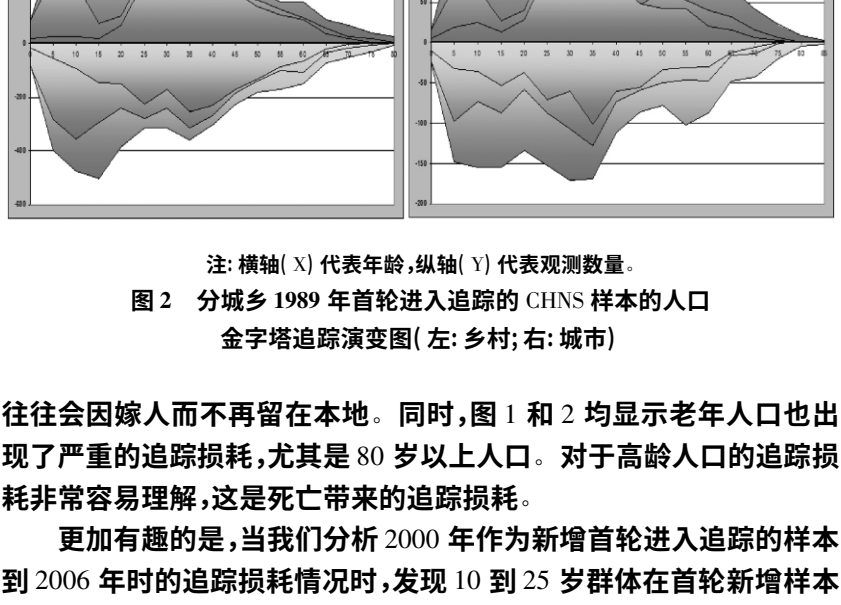
为进一步搞清到底是哪些年龄的人发生了追踪损耗,我们使用人口结构图来分析追踪损耗的对象及其特征。根据追踪损耗的定义,我们以 1989 年第 1 轮调查的样本为基础,分别计算了 2000 年、2006 年仍然被追踪到的第 1 轮进入追踪的样本的分布,绘制出人口年龄金字塔图(见图 1,历次样本的年龄均转换为 1989 年的年龄,以便对比)。



注:横轴(X) 代表年龄,纵轴(Y) 代表观测数量。

图 1 1989 年首轮进入追踪的 CHNS 样本人口金字塔追踪演变图

图 1 显示,1989 年开始追踪的人口呈现出金字塔结构,并且男女比例相当。到了 2000 年,年轻人口中相当一部分不能被追踪到,尤其是 1989 年时 10 – 25 岁的群体,出现了严重的追踪损耗。到了 2006 年,1989 年时年龄在 0 – 15 岁的群体也几乎损耗殆尽。我国从 1984 年 7 月 1 日起实施《中华人民共和国义务教育法》,规定国家实行九年制义务教育。小学从 6 岁开始,义务教育结束阶段为 15 岁,此后相当多的人进入劳动力市场。因此我们推算,年轻人口的追踪损耗之所以非常高,是由于 15 岁以前多是寄宿制的入学读书,导致无法追踪;后期由于外出就业而无法追踪。图 2 是分城乡的人口金字塔追踪演变图,显示无论城市还是农村,男性的追踪损耗均较女性略少,这是因为女性



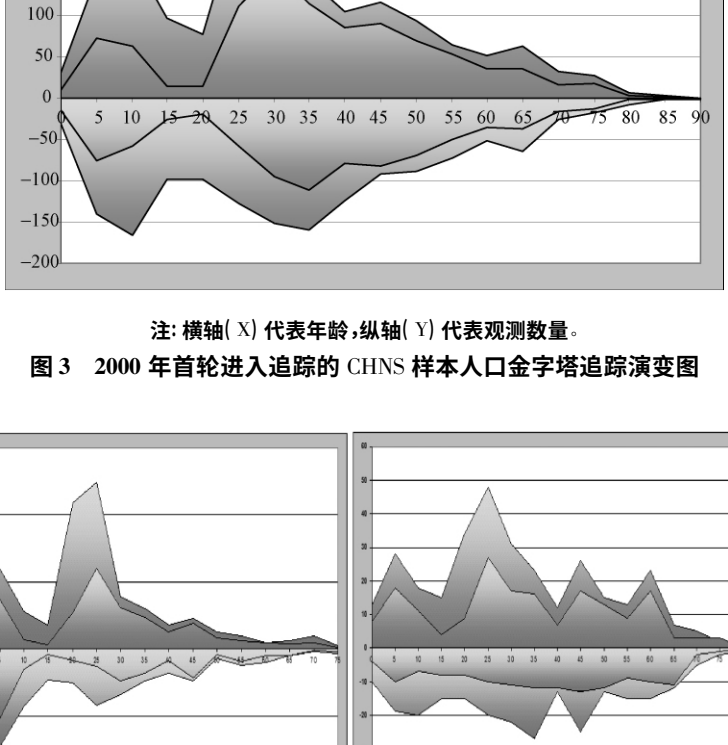
注: 横轴(X) 代表年龄,纵轴(Y) 代表观测数量。

图 2　分城乡 1989 年首轮进入追踪的 CHNS 样本的人口金字塔追踪演变图(左: 乡村; 右: 城市)

往往会因嫁人而不再留在本地。同时,图 1 和 2 均显示老年人口也出现了严重的追踪损耗,尤其是 80 岁以上人口。对于高龄人口的追踪损耗非常容易理解,这是死亡带来的追踪损耗。

更加有趣的是,当我们分析 2000 年作为新增首轮进入追踪的样本到 2006 年时的追踪损耗情况时,发现 10 到 25 岁群体在首轮新增样本的人口金字塔中显得很少,也就是说很难进入样本中(见图 3、4) 。考虑到 CHNS 的新增样本是以家庭为单位的,因此,这个年龄段的招募与其在人口结构中的不一致应该由招募机制来解释。我们推断是因为这个年龄段的人口更多地生活在宿舍里,而非家里。有关资料显示,随着中国农村留守儿童的增加以及学校减少导致学生上学路程变远,国家实施了"农村寄宿制学校建设工程"。① 此工程已覆盖了中西部地区的 23 个省市、自治区。西部农村初中寄宿生比例已达 53. 6%,在西藏、广西和云南 3 个省份甚至超过 70%。同时,2007 年西部农村小学寄宿生比例达到 11. 6%,西藏、内蒙古、云南、青海 4 个省份超过 20%(贾小娜,2007) 。这解释了我们看到的 10 到 16 岁青少年,年龄越大在家里追踪到的比例越低的状况。另外,农村比城市情况严重,也与国家寄宿制工程主要在农村推广有关。15 到 25 岁年龄群的征募比例不足,我

① 可参考教育部、发改委、财政部、国土资源部、建设部给国务院的《关于进一步做好农村寄宿制学校建设工程实施工作的若干意见》,该意见在 2005 年 6 月 27 日经由国务院办公厅转发各省、自治区、直辖市人民政府以及国务院各部委、各直属机构。



注: 横轴(X) 代表年龄,纵轴(Y) 代表观测数量。

图 3　2000 年首轮进入追踪的 CHNS 样本人口金字塔追踪演变图

注: 横轴(X) 代表年龄,纵轴(Y) 代表观测数量。

图 4　分城乡 2000 年首轮进入追踪的 CHNS 样本人口金字塔追踪演变图(左: 乡村; 右: 城市)

们认为可能是由于外出就业所致。

因此,2000 年新增首轮进入追踪的样本显示出与 1989 年很不一样的情况。我们对以家庭为追踪单位导致 10 – 25 岁年龄群显著低概率的捕获做了解释: 青少年群体从读书开始,其中一部分人因进入宿舍等集体住宿体制中,而无法在家庭中捕获; 尤其是国家实施 “农村寄宿

制学校建设工程”，使得农村比城市严重。① 这些不再属于追踪损耗问题，而属于追踪调查的抽样问题。

至此，我们总结一下上面的发现：(1) 追踪损耗比例很高，平均轮次在 10% 以上；(2) 随着中国社会的转型，追踪损耗不断上升；②(3) 随着中国社会的转型，以家庭为单位的抽样方案中，对 10 到 25 岁年龄群人口的抽样误差不断增大，其中农村的误差大于城市；我们推断，该群体的追踪误差也会呈现出城乡差异，但我们无法推断是否也是农村的误差大于城市；③(4) 年轻人口由于外出就业和分家，最终会不再被追踪到；(5) 老人由于死亡，最终也不再会被追踪到。

对于以上总结，我们将在后面研究中通过更加严谨的实证数据和模型进行证伪。我们将(2) 称为社会转型影响假设(H1)，(3) 称为城乡差异假设(H2)，(4) 称为生命周期假设(H3)。

四、追踪损耗的分解方法

我们在第二部分对 CHNS 追踪损耗的分析是根据风笑天(2006) 的定义，将“因去世、搬迁、拒绝再次调查等原因而无法追踪到的”均定义为追踪损耗。严格说来，该定义不严谨，其中包含了不同性质的两类

① 此处简单的分析可能隐藏了更为复杂的情况，例如农村人口进城打工可能带走孩子。这种情况我们认为属于群体损耗，而非抽样问题。

② 考多斯(Kordos，2005) 通过对 14 个转型国家 1991 – 2000 年的家庭抽样调查(household sample surveys) 的分析，发现随着这些国家的转型，不回答率(non-response rate) 在每个国家均逐年上升，他分析原因是由于家庭社会经济结构的变化。斯科特等(Scott et al.，2005) 对 9 个转型国家在 1997 – 2001 年的居住情况调查(living standards measurement study surveys) 的不回答率差异进行了分析，提出当国家内部发生过冲突，或社会信任变低的情况下，不回答率均会增加。考虑到我国在 CHNS 执行期内基本符合以上分析的情况，因此，本研究将追踪损耗的增加归结为社会转型的影响。

③ 推论一致的误差方向是冒险的。一般而言，调查中的无法观测误差(non-observation error) 有两个完全不同的误差来源：一个是无覆盖误差(non-coverage error)，指合格的被调查者没有机会被抽取到；另一个是无回答误差(non-response error)，指被抽取到的被调查者没有参与调查(Lepkowski，2005)。这是两个完全不同的概念；前者在本研究中指抽样误差，后者在本研究中指追踪损耗。在本研究中，以家庭为抽样框导致的 10 到 25 岁年龄群人口的抽样误差不断增大，其中农村的误差大于城市属于无覆盖误差。将无覆盖误差的城乡差异模式推论为追踪损耗导致的无回答误差也存在着同样的模式是很冒险的，需要更多的信息才可以做此推论。因此，此处提出的城乡差异假设仅指追踪误差也呈现出城乡差异，并不涉及方向。

损耗。那些因去世、搬迁而无法追踪的样本，属于无论如何努力都无法成功追踪的样本；那些拒绝的样本，则属于仍然在追踪范围、并可以接触的样本，也就是存在调查可能性的样本。按照金（Kim，2010a，2010b）关于追踪损耗的新理论，前者属于总体损耗，后者属于样本损耗。为了搞清追踪损耗的情况，需要更加深入地将追踪损耗分解为样本损耗和总体损耗。鉴于此，我们尝试提出以下几个概念：（1）追踪成功率，指曾经被纳入调查的样本，在日后的追踪成功率。（2）存活样本[①]追踪成功概率（recapture），指仍然在追踪调查的样本框范围内的存活样本成功被调查的概率。（3）样本存活概率（survival），指曾经被纳入调查的样本，存在于追踪调查的样本框范围内且存活的概率。

显然，“追踪成功率”等于“存活样本追踪成功概率”乘以“样本存活概率”。存活样本追踪成功概率的相反概念是样本追踪损耗率；样本存活概率的相反概念是样本死亡率，即样本死亡、搬迁等不再存在于被观察总体之中的情况，因此，样本死亡率（1－样本存活率）也即总体损耗率。显然，追踪成功率是样本损耗率和总体损耗率的函数。

为了将追踪成功率分解为样本损耗率和总体损耗率，本研究借鉴了生态学中的标定再捕获法（mark-recapture methods），该方法除了可以估算群体的规模，还可以估算群体的参数，如生存率（survival rate）、捕获率（recruitment rate）、群体生长率（population growth rate）等。封闭群体（closed population）的标定再捕获法对群体要求很严格，要求群内没有出生、死亡、迁入、迁出的个体，这往往与现实不符。我们常常遇见的是开放群体，其模型由乔利－塞贝尔（Jolly-Seber）提出，称为乔利－塞贝尔模型。对于开放群体的标定再捕获法分析，有以下几个前提要求：（1）调查期间，标记具有永久性，且再调查时可以正确记录在案；（2）加以调查标记处理，并不会影响再次接受调查的概率——每次被捕获的概率不变；（3）调查本身并不影响死亡率和迁移率；（4）调查后，该样本与其他样本或非样本的接触，并不影响其再次接受调查的概率；（5）每次捕获时，每个个体的被捕获概率相同；（6）两次捕获之间，群体中个体的生存概率相同；（7）标记不会丢失或者看错；（8）每次捕获和

① 存活样本所对应的概念是死亡样本。死亡在这里并非一定指样本真的死亡，而是泛指样本死亡、搬迁等不再存在于被观察的总体之中的情况。相应的，存活是指样本仍然存在于被观察总体之中的情况。

释放(调查时间)相对于调查期(整个追踪期)而言是短暂的;(9)群体中的移出率是固定的。

我们将追踪调查与开放群体的标定再捕获法对照,基本可以肯定,追踪调查就是开放群体标定再捕获法在人类社会中的一个应用。① 因此,可以使用标定再捕获法的分析方法来获得生存率、捕获率等参数。追踪成功是生存率和捕获率的联合分布。为了分解生存率和捕获率,我们使用贝叶斯(Bayesian)方法,其优点在于可以方便地将联合分布分解为各自独立的分布,并计算有关参数。

我们对数据进行了相应处理,处理方法和依据如下:(1)每轮次的追踪样本如果观测得到则记录为1,否则为0。这样追踪样本的观测历史就是一系列的柏努利(Bernoulli)分布事件。当样本在第一个观测时点被观测到,且个体存活(指继续在原家庭中生活)时间长于下次追踪观测的时间,那么在下次追踪时点可被观测到的概率就是存活样本的追踪成功概率。(2)当然,在个体存活的情况下,才存在存活样本的再次追踪成功概率。如果个体死亡,则再次观测成功率为0。因此,还需模拟逐次观测时间点的死亡形式。(3)我们根据样本进入到追踪的轮次,将第一次被观测时的生存状态设为存活,并在第一次进入追踪样本之后才将其放入模型。(4)我们运用基于吉布斯(Gibbs)抽样的马尔可夫链蒙特卡洛(Markov Chain Monte Carlo,MCMC)方法动态模拟出参数后验分布的马尔可夫链,生成存活样本的再次追踪成功概率(recruitment)和样本生存概率(survival)的相关参数。有关贝叶斯模型的WinBUGS② 分析程序见下:

---

① 这里需要提醒读者注意,生态学中的标定再捕获方法,其分析的群体有两种类型:一种是开放群体的标定再捕获(open population mark – recapture method),另外一种是封闭群体标定再捕获(closed population mark-recapture method)。追踪调查是将上一轮进入追踪调查的对象看作是全体,在下一轮中,我们仅对这个全体进行再次捕获。因此,我们研究的是一个非开放群体的标定再捕获。

② WinBUGS 是 Bayesian Inference Using Gibbs Sampling 的缩写,它是英国剑桥生物统计学研究所(Biostatistics the Medical Research Council,Cambridge,United Kingdom)与圣玛丽(St. Mary's)皇家学院医学分院(the Imperial College School of Medicine)共同开发的,用马尔可夫链蒙特卡洛方法进行贝叶斯推断的专用软件包。它将所有未知参数都看作随机变量,然后对此种类型的概率模型进行求解。WinBUGS 所使用的编程语言非常容易理解,允许使用者直接对研究的概率模型作出说明。WinBUGS 可以免费下载使用,下载官方网址为:http://www.mrc – bsu.cam.ac.uk/bugs/welcome.shtml。

```
# WinBUGS 程序①
model {
   for ( i in 1: N) {    #对于每一个样本
     alive [i, First [i] ]  < - 1   #该样本第一次被观测时,状态标定为存活
     for ( j in First [i] +1: Years) { #从第二次观测开始,直到追踪结束
          palive [i,j]  < - survival *  alive [i, j - 1] #观测到的概率是样本存活概率和再次观测成功概率的联合分布
        alive [i,j]  ~ dbern( palive [i,j] )   #上次观测时的存活概率是二项式分布
        psight [i,j]  < - recapture *  alive [i, j]  #下一次观测时的存活概率是上年存活状态与存活概率的联合分布
        Y [i, j]  ~ dbern( psight [i,j] ) #( i 个体 j 次) 是否观测得到是 Bernoulli 分布事件
        }  } #结束每一个样本每一轮观测的循环
#对存活率和存活样本的再次观测成功概率的先验分布设为均匀分布
   survival  ~ dunif( 0 , 1)   #设定存活率是均匀分布
   recapture  ~ dunif( 0 ,1)   #设定存活样本的再次观测成功概率是均匀分布
} #结束模型定义
```

五、对 CHNS 追踪损耗的分解分析

(一) 追踪损耗的层次

追踪调查方案,往往被设计成多层次数据收集方案。一般而言,社区层次不会出现大规模追踪损耗。②因此,追踪损耗主要发生在家庭和个体这两个层次。

北京大学"中国家庭动态追踪调查"项目组曾在其 2008、2009 年进行的两轮试调查的基础上,报告并分析了拒访行为的发生规律(孙

① 本程序主要参考了麦卡锡 (McCarthy,2007) 的《Bayesian Methods for Ecology》一书的有关程序改写而成。"#"符号之后为注释,不影响程序执行。数据整理使用了 Stata 软件,并通过有关程序传递给 WinBUGS 程序,有关技术文献见 Thompson et al. ,2006。

② 作者曾以 2005 年中国综合社会调查全国协调人的身份负责该年的社区样本更换。2004 年国家为了进行社区建设,在全国范围内进行了村居的社区整合,对较小社区进行合并、对过大的社区拆分、散建楼住户就近纳入社区管理、破产企业的家委会改建居委会或并入就近社区、连片开发居民小区组建新居委会、棚户改造社区重新界定管辖范围,导致 2003 年抽取的部分社区消失或变更管辖范围。但即使这么大的调整,从全国范围而言,不可恢复的社区仅占 2.1%。因此,在常态下,社区损耗的比例应该是极低的。

妍等,2011)。虽然该研究旨在研究"追踪调查受访对象的丢失现象呈现出的特点和模式",但该研究仅仅分析了追踪调查中的家户层次的拒访行为,即家庭户整户追踪损耗的发生规律,而未注意到个体追踪损耗发生的规律。以该项目组2008和2009年在广东进行调查的情况为例:家庭层次上,家庭户整户追踪损耗占全部户数的17.48%,家庭户中部分成员的损耗占全部户数的24.59%;从个体层次来看,家庭户整户追踪损耗导致的个体追踪失败占全部个体追踪失败的53.04%,家庭户部分成员的损耗导致的个体追踪失败占全部个体追踪失败的46.96%。也就是说,以家庭层次追踪损耗来分析追踪损耗,只涉及个体损耗的一半略多(见表4)。不难看出,项目组由于仅将关注焦点放在家庭户的拒访上,因而其研究仅涉及140户332位个体的追踪损耗分析,而忽略了另外197户294位个体的追踪损耗。

表4　2008、2009年两轮中国家庭动态追踪调查广东子调查追踪损耗一览表

		户损耗情况			合计
		全户损耗	户部分成员成功追踪	全部成员成功追踪	
		140户(17.48%)	197户(24.59%)	464户(57.93%)	801
个体追踪情况	损耗	332(53.04%)	294(46.96%)	0	626
	追踪成功	0	393	1115	1508
	新增加	0	0	2	3
合计		332(15.54%)	687(32.15%)	1118(52.32%)	2137

显然,个体的追踪成功必然导致家庭的追踪成功,但家庭的追踪成功并不意味着对个体的追踪成功。同时,即使新的追踪调查轮次中没有新的家庭被招募进来,但却可能因为家庭成员结婚、生育、亲属迁入等原因导致新的家庭个体被招募进来。因此家庭和个体层次的追踪损耗会表现出不同的规律。下面我们将分别对个体和家庭层次的追踪损耗进行分解和分析,并对其进行对比分析。

(一)个体追踪损耗的分解分析

为了在个体层次验证追踪损耗的社会转型影响假设、城乡差异假设和生命周期假设,我们以年龄、追踪时期和城乡分别计算不同情况下个体存活样本的再次追踪成功概率和样本生存概率(见表5和表6)。

表 5 25 岁以上 CHNS 样本追踪成功概率和生存概率分布(个体) (单位:%)

		1989 – 1993	1989 – 1997	1989 – 2000	1989 – 2004	1989 – 2006	1989 – 2009
全体	存活样本追踪成功概率	98.28	97.70	96.67	96.47	94.25	92.58
	样本生存概率	91.77	89.50	89.13	87.26	88.38	88.60
城市	存活样本追踪成功概率	97.59	97.28	95.62	96.00	93.92	92.23
	样本生存概率	88.19	84.79	84.67	82.86	84.22	84.57
农村	存活样本追踪成功概率	98.63	97.85	97.11	96.63	94.40	92.72
	样本生存概率	93.55	91.82	91.39	89.51	90.50	90.84

表 6 25 岁以下 CHNS 样本追踪成功概率和生存概率分布(个体) (单位:%)

		1989 – 1993	1989 – 1997	1989 – 2000	1989 – 2004	1989 – 2006	1989 – 2009
全体	存活样本追踪成功概率	94.57	94.41	94.24	92.46	89.30	84.86
	样本生存概率	89.09	86.05	85.58	79.87	79.99	81.12
城市	存活样本追踪成功概率	94.59	94.85	93.60	92.61	88.54	84.33
	样本生存概率	84.73	81.31	81.28	77.10	78.43	79.43
农村	存活样本追踪成功概率	94.61	94.24	94.47	92.28	89.55	85.12
	样本生存概率	90.60	87.74	87.11	80.92	80.56	81.70

根据表 5 和表 6 由追踪成功率分解出的样本存活率和存活样本的追踪成功率,我们有如下发现:

无论是样本的生存概率,还是生存样本的再次追踪成功概率,城市样本均低于农村样本。这说明相对于农村人口,城市人口样本的追踪更加困难,追踪缺失存在差异的城乡差异假设得到了与预期相反方向的验证:①具体而言,无论哪个年龄群,城市样本的样本损耗比农村样本都要大 1 – 2% 左右;而 25 岁以上群体的城市样本的总体损耗在不同追踪时期稳定地大于农村样本 6 – 7% 左右;25 岁以下群体城市样本的总体损耗在追踪时期,从高于农村样本 6% 逐渐降低到 2% 左右,表明追踪损耗的城乡差异在青少年群体中有趋同趋势。作者推论这与新生代人口劳动力市场的融合有关。

① 这里需要提醒读者注意,此处得到的追踪损耗的城乡差异,与通过对以家庭为抽样框导致的 10 – 25 岁年龄群人口的抽样误差不断增大,其中农村的误差大于城市误差的方向正好相反。这说明无覆盖误差与无回答误差存在着相反的城乡差异。

25 岁以上人口样本的生存概率，以及生存样本的再次追踪成功概率均稳定地高于 25 岁以下人口，这证实了生命周期假设，说明 25 岁以下人口的追踪更加困难。具体而言，25 岁以上人口的存活样本追踪成功概率比 25 岁以下人口在不同追踪时期均高 5%；随着追踪时期的延长，25 岁以上人口比 25 岁以下人口的样本生存概率从高 2% 逐渐增加到到 9%，说明 25 岁以下人口的总体损耗相对于 25 岁以上人口来说日益严重。

无论 25 岁以上还是以下群体，用 1989 到 2004 以后年份的数据拟合的参数普遍比用 1989 – 2000 年数据拟合的参数要低。因为本研究使用的模型建立在两个假设之上：(1) 假定每次捕获时，每个个体的被捕获概率相同；(2) 两次捕获之间，群体中个体的生存概率相同，因此是一个线性模型。由于用 1989 – 2000 年数据估算的 1989 – 2000 年的参数比用 1989 – 2004 数据估算的参数要高，所以我们应将这一下降趋势理解为一个非线性过程。也就是说，2004 年以后，真实的情况是下降的更多。这证实了社会转型影响假设，即随着中国社会的转型，追踪调查的困难越来越大。

CHNS 项目组在 1993 到 2006 年的调查中，对此前追踪不到或拒访的家户，并未尝试接触。2009 年，CHNS 试图重新联系这类家户，以减缓长期流失状况，维持样本的代表性。经过辛勤工作，在之前丢失的家户中尽可能找回了一些，因此 2009 年的样本生存概率有所上升，但存活样本的追踪存活概率并没有显著上升，依然维持了下降趋势(见图 5)。

(二) 家庭层次的追踪损耗分解分析

我们首先以家庭作为追踪层次，对不同年份招募和结束追踪的交互分布进行了分析(见表 7)。

表 7 所展示的家庭层次的招募与退出分布规律与个体层次类似：随着时间的推移，当年招募当年即损耗的比例逐年上升，表明在中国追踪调查的难度随着社会转型而越来越高；总体而言，进入追踪的家庭样本的追踪损耗会逐年下降；一次不好的追踪调查安排，可能导致历次进入的家庭样本均同时出现高的追踪缺损(2006 年历次进入的样本均出现高追踪损耗)。

同样，我们以家庭为追踪单位，将家庭追踪损耗分解为样本存活概率和存活样本的追踪成功概率(见表 8)。

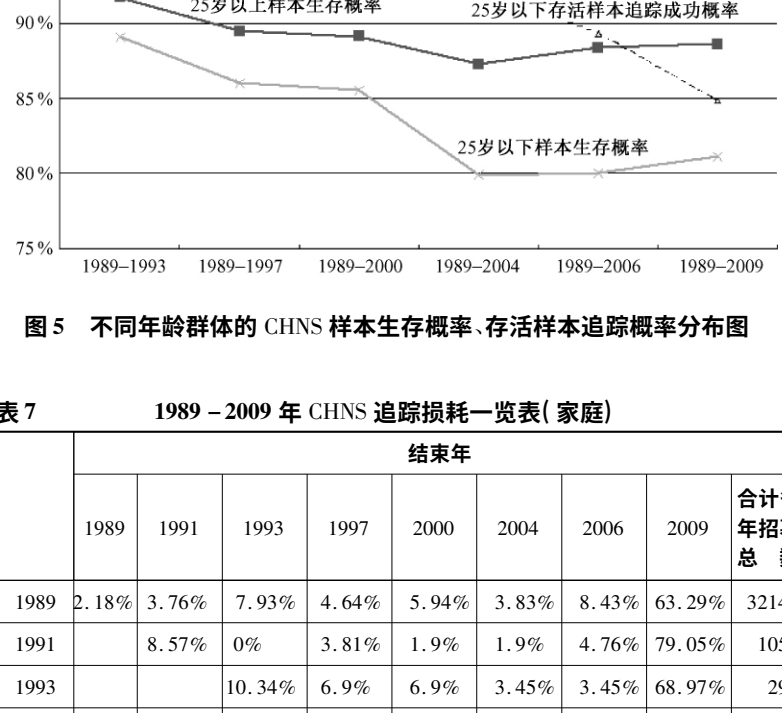


图 5　不同年龄群体的 CHNS 样本生存概率、存活样本追踪概率分布图

表 7　　1989 – 2009 年 CHNS 追踪损耗一览表(家庭)

		结束年								
		1989	1991	1993	1997	2000	2004	2006	2009	合计多年招募总数
开始年	1989	2.18%	3.76%	7.93%	4.64%	5.94%	3.83%	8.43%	63.29%	3214
	1991		8.57%	0%	3.81%	1.9%	1.9%	4.76%	79.05%	105
	1993			10.34%	6.9%	6.9%	3.45%	3.45%	68.97%	29
	1997				14.83%	9.19%	4.46%	12.1%	59.42%	1099
	2000					14.81%	8.05%	15.58%	61.56%	385
	2004						16.56%	18.2%	65.24%	489
	2006							25.32%	74.68%	233
	2009								100%	980
	合计当年结束总数	70	130	258	318	353	287	618	4500	6534

表 8 显示,随着时间的推移,无论是样本的生存概率,还是存活样本的追踪成功概率均不断下降,这意味着转型假设不止在个体层次上得到了验证,在家庭层次也得到了验证。中国家庭的追踪难度在于,家庭迁移比率不断上升,家庭拒绝访问的概率也不断加大。

表 8　　CHNS 样本追踪成功概率和生存概率分布(家庭)　　(单位:%)

		1989 – 1993	1989 – 1997	1989 – 2000	1989 – 2004	1989 – 2006	1989 – 2009
全体	存活样本追踪成功概率	98.79	98.19	96.1	96.39	95.24	94.17
	样本生存概率	95.96	93.88	94.3	92.85	93.45	93.31
城市	存活样本追踪成功概率	97.82	98.01	95.31	96.18	94.70	
	样本生存概率	93.28	89.54	90.01	88.32	89.36	
农村	存活样本追踪成功概率	99.11	98.17	96.40	96.39		
	样本生存概率	97.24	95.92	96.35	95.06	95.43	

注:2009 年 CHNS 家庭层次的数据中不再有家庭城乡属性的变量。

同样,城乡差异假设也通过家庭层次追踪损耗的分解得到了验证,城市家庭的样本生存概率显著低于农村样本的生存概率(1989 年到 2006 年相差 6 个百分点),但存活样本的追踪成功概率城乡差异却并不明显。

(三)家庭层次和个体层次追踪损耗的比较

根据表 5、6 和 8 得到的家庭样本生存概率分布与 25 岁以上及以下个体样本存活概率分布显示,家庭样本的存活概率大大高于个体。个体中,25 岁以上个体样本的存活概率高于 25 岁以下个体样本。因此,以社区作为追踪单位而言,家庭层次的追踪效果好于个体层次;成年人的追踪效果好于未成年人和年轻人(见图 6)。

图 7 是根据表 5、6 和 8 绘制的家庭存活样本追踪成功率与 25 岁以上及以下个体存活样本追踪成功率的对比图。这里显示了与样本生存概率不同的图景。存活的家庭样本的成功追踪概率和 25 岁以上成年人的成功追踪概率差异并不是非常大,具体而言,在 1989 – 2004 年间非常接近,[①]在 2004 – 2009 年逐渐增大。25 岁以下存活个体样本则显著低于家庭存活样本,并在 2000 年以后呈现加速下降的趋势。

① 虽然分析结果显示,2004 年存活家庭的追踪成功率(96.10%)略低于 25 岁以上个体(96.67%),但由于家庭是由个体成员构成,只要家庭中有任意一名成员被追踪成功,则家庭就被追踪成功,因此一般而言,存活家庭的追踪成功率应该大于或等于个体成员的追踪成功概率。此处有所违背,我们认为是由于分解所使用的 MCMC 算法模拟中初始值、迭代收敛等造成的误差。我们建议应系统地综合判读使用 MCMC 算法分解的结果,而不宜对某一单独结果过度强调。

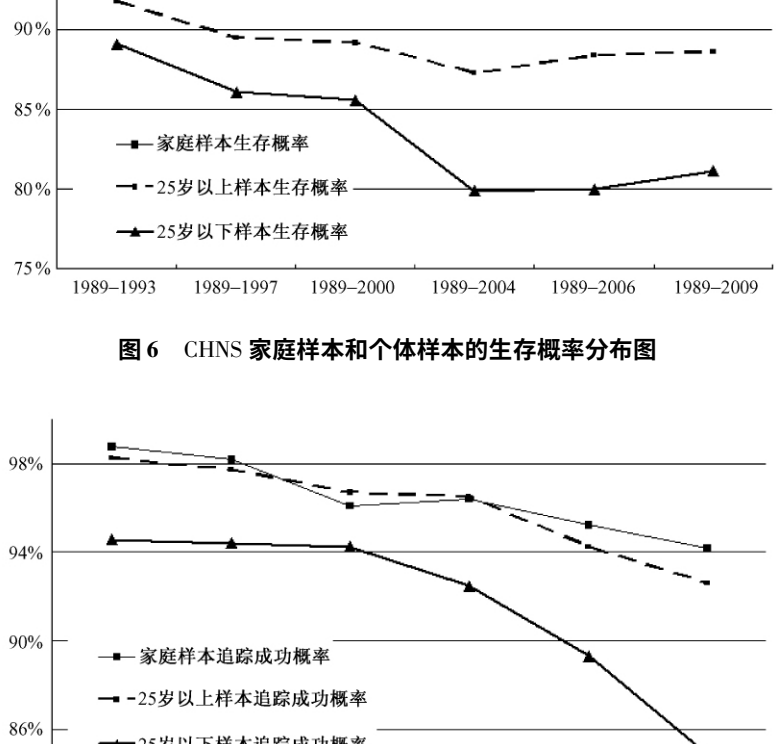


图6 CHNS家庭样本和个体样本的生存概率分布图

图7 CHNS家庭和个体存活样本的追踪成功概率分布图

六、追踪调查中的实践建议

所有社会科学抽样调查均要求样本具有随机性，而有多轮调查的追踪调查的样本随机性的保证，则不仅要求首轮样本具有随机性，更需要此后多轮的追踪样本具有比较低的样本损耗，并且在每一追踪轮次

中流失的家庭户/个人也是随机的。然而本研究显示,在以社区为追踪区域的追踪调查 CHNS 中,随着追踪轮次的增加,样本损耗呈现为较大规模的、非随机的和不断恶化的。其规模巨大表现在每轮损耗在 10% 到 20% 之间;其非随机性表现在城市的追踪损耗概率大于农村,25 岁以下青少年的追踪损耗概率大于 25 岁以上成人,个体的追踪损耗概率大于家庭;而且,随着中国社会的转型追踪损耗逐年增加。这意味着在追踪调查中,即便第一轮的样本是随机性样本,较高的追踪损耗也使得跟踪样本逐渐不再具有随机性。因此,追踪损耗是一个不能忽视的问题,对该问题的解决是在中国转型情况下开展追踪研究的一个非常关键和重要的问题。

尽力提高样本追踪成功率,是追踪调查方案设计中的一项重要任务。基于本文研究发现,我们试图提出以下建议,以期为在中国社会转型条件下开展追踪调查提供一些借鉴。

研究结果表明家庭追踪的损耗大大低于个体损耗,因此,如果我们采用基于家庭的追踪方案来收集个体的资料,会大大提高个体样本的追踪成功率。简单而言,我们应该允许家庭成员代为回答其他无法追踪的家庭成员的信息。此外,由于 25 岁以上成年人的追踪损耗大大低于 25 岁以下的成年人和年轻人,同时,一个家庭中年长者比年轻者更可能留在家中,而且往往是年长者更了解年轻者的情况,因此,由家中年长者代答年轻者的信息,也较为可靠。采用该方法,既能提高个体层次样本的存活概率,也能提高个体层次样本的追踪成功概率。根据上文数据进行测算,会使 25 岁以上个体样本在 2009 年的成功追踪概率提高到 1997 年的水平(2009 年家庭样本的总和追踪成功率为 94.17% ×93.31 = 88%;25 岁以上个体样本在 1997 年的总和追踪成功率为 97.7% ×89.5 = 87%);使 25 岁以下个体样本在 2009 年的成功追踪概率提高到比第一轮追踪的 1997 年还高的水平(94.57% × 89.09% = 84%)。

此外,本研究还发现,由于就学和外出工作,使得 10 – 25 岁年龄段的个体招募情况明显不符合人口年龄结构。该年龄段的人口往往是在宿舍体制中度过而非在家庭中。而单独设计一个抽样方案来弥补该群体的抽样不足,会大大增加抽样的复杂性;追踪调查的多层次要求,又使得对该群体的单独抽样面临着家庭层次变量的数据收集困难。因

此，最好的解决方案还是请家中的年长者代为回答。①

CHNS 这类以社区为追踪单位的追踪调查，其最大问题是对年轻世代的追踪损耗过大，使得所收集数据无法用于分析代际之间的流动、教育获得和职业获得。如果采用上述改进方法，则可以在很大程度上弥补此类追踪数据的缺陷，使其与另一类以家庭为追踪单位的追踪调查一样，成为观察偏差较小、可以用于代际比较的数据。当然，我们无法推知家中 25 岁以上成年人所代答的未成年人和年轻人的信息，哪些是准确的、哪些是不那么准确的。这还有待于进一步的专门研究。

以社区为追踪单位的追踪调查如采用上述方法进行追踪，则家庭成员，尤其是家庭中年轻成员的迁移将不再成为一个造成严重观察偏误的问题。迁移离开社区的家庭，可以由新随机抽取的家庭取代。②因此，由流出年轻人的家人代答的社区范围的家庭户/个体追踪方案所获得的样本将是一个克服了样本损耗的生命周期差异的低偏误的追踪样本。这种样本与以全国范围家庭户为单位的追踪方案对青少年无偏的追踪样本相比，效果类似但追踪成本又远较后者为低，不失为一种在经费条件有限情况下进行追踪调查的好方法。同时，该方法还可以克服以全国范围家庭户为单位的追踪方案会逐渐丧失对社区的代表性这一不足，而满足跨“社区—家户—个体”的多层次数据收集和分析的要求。

最后，值得强调的是，发展中国家的社会研究本身就有很多独特议题，同时在发展、转型条件下的社会追踪调查又面临着比发达国家更为复杂的情况，所以不能盲目照搬发达国家追踪调查的经验。在中国积累追踪调查的本土社会学知识，也许正是中国社会学界的挑战和机遇之所在。

参考文献：

风笑天，2006，《追踪研究：方法论意义及其实施》，《华中师范大学学报（人文社会科学版）》第 45 期。

郝令昕、王卫东、谢桂华，2010，《发展中国家农村跟踪调查：回顾和前瞻》，工作论文。

① 在具体调查的实际操作上可以有多种选择：只由家人代答、由家人现场代访、家人事后代访、家人提供联络方式由访问员现场访问等，并且可以单独或混合使用以上方法。

② 只有仍然在社区的家庭户样本，却由于某种内在一致性的原因拒绝访问才有可能造成系统性的偏误。

孙妍、邹艳辉、丁华、严洁、顾佳峰、邱泽奇，2011，《跟踪调查中的拒访行为分析——以中国家庭动态跟踪调查为例》，《社会学研究》第2期。

贾小娜，2007，《农村寄宿制学校建设工程纪实》，《教育》第13期。

Bhattacharya，Debopam 2008，"Inference in Panel Data Models under Attrition Caused by Unobservables."*Journal of Econometrics* 144(2).

Fitzgerald，J.，P. Gottschalk & R. Moffitt 1998，"An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics."*Journal of Human Resources* 43.

Hirano，K.，G. W. Imbens，G. Ridder & D. B. Rubin 2001，"Combining Panel Data Sets with Attrition and Refreshment Samples."*Econometrica* 69(6).

Kim，Seik 2009，"Sample Attrition in the Presence of Population Attrition."University of Washington Working Paper.

—— 2010a，"Economic Assimilation of Foreign-Born Workers in the United States: An Overlapping Rotating Panel Analysis."University of Washington Working Paper.

—— 2010b，"Wage Mobility of Foreign-Born Workers in the United States."University of Washington Working Paper.

Kordos，Jan 2005，"Household Surveys in Transition Countries."In United Nations *Household Sample Surveys in Developing and Transition Countries*. New York，NY: United Nations Publications.

Lepkowski，James 2005，"Non-observation Error in Household Surveys in Developing Countries."In United Nations *Household Sample Surveys in Developing and Transition Countries*. New York，NY: United Nations Publications.

McCarthy，M. A. 2007，*Bayesian Methods for Ecology*. New York: Cambridge University Press.

Peracchi，F. & F. Welch 1995，"How Representative Are Matched Cross－Sections? Evidence from the Current Population Survey."*Journal of Econometrics* 68.

Scott，Kinnon，Diane Steele & Tilahun Temesgen 2005，"Living Standards Measurement Study Surveys."In United Nations *Household Sample Surveys in Developing and Transition Countries*. New York，NY: United Nations Publications.

Sergi，Jiménez－Martín & Franco Peracchi 2002，"Sample Attrition and Labor Force Dynamics: Evidence from the Spanish Labor Force Survey."*Spanish Economic Review* 4(2).

Thompson，J. T.，T. Palmer & S. Moreno 2006，"Performing Bayesian Analysis in Stata Using WinBUGS."*The Stata Journal* (4).

Van den Berg，G. J. & M. Lindeboom 1998，"Attrition in Panel Survey Data and the Estimation of Multi－State Labor Market Models."*Journal of Human Resources* 43.

作者单位：中山大学社会学与社会工作系、社会科学调查中心

责任编辑：杨　典

Abstract: Based on the data of 676 Chinese publicly-traded companies (2000-2007), this paper examines the causes and consequences of corporate diversification. Different from conventional efficiency-based studies on corporate strategy, this research examines diversification from an institutional perspective and investigates how institutional environment shapes corporate strategy. Specifically, the author focuses on the role of state and financial market in the construction and diffusion of diversification strategy in China. The study finds that both the high level of diversification and later de-diversification of Chinese firms are largely driven by state policies. The result also shows that Chinese institutional investors have substantial influence on firm strategy. This paper finds significantly negative effects of diversification on stock return and firm growth. Despite the negative relationship between diversification and performance, Chinese firms have diversified to the highest level among the major economies in the world, implying that diversification was spread more through institutional than economic processes in China.

Abstract: This paper focuses on the success tracking rates in panel survey in China. Based on the ecological mark-recapture theory, this research develops a method that decomposes non-ignorable sample attrition rates into survival rate and recruitment rate with Bayesian statistics. The proposed method is used to obtain the survival rates and recruitment rates of the eight-wave panel survey—China Health and Nutrition Panel Survey. The sample attrition urban-rural differences assumption, social transformation assumption and life cycle assumption are confirmed on both household level and individual level. At the end, this paper proposes suggestions on how to reduce sample attrition rates.

Abstract: In the past several decades, many researchers focused on individual level determinants of environmental concern while neglected the effects of regional level factors on environmental concern. In this article, the authors try to apply a two-level-hierarchical linear model to sort out the significant municipal level effects from the effects of individual characteristics. As expected, 5% variance can be explained by municipal level factors. The findings indicate that income, age, education and gender are significantly associated with environmental concern on the individual level; at the municipal level, post-materialist value, city type and GDP per capita