·学科前沿·

《学术界》(月刊)
总第161期,2011.10
ACADEMICS
No. 10 Oct. 2011

精准抽样是量化分析推论的基础

○ 沈明明[1]，王蕴峤[2]
(1. 北京大学　中国国情研究中心，北京　100871;
2. 北京大学　政府管理学院，北京　100871)

〔摘　要〕随着近三十年来我国社会的快速发展与深刻变迁,大量农村、小城镇人口向大中城市集中,造成高比例的"人户分离"现象。而城市居民也因为城区改造和扩张,造成城市内部大量的"人户分离"。这使得传统的依赖户籍信息进行抽样的社会科学调查,在样本的抽取方面产生极大的困难与系统偏差。因此,如何以精准抽样为基础,从而避免造成导致"社会整体误解"的调查结果,已成为近年来中国社会科学量化研究中一个不容忽略的问题。本文旨在一方面对那些缺乏精准抽样的研究进行评论,另一方面也向读者介绍依托GIS/GPS最新技术发展而来的"GIS/GPS辅助的区域抽样方法",作为转型期中国社会大规模人口流动条件下精准抽样的解决方案,以供学界参考备用。

〔关键词〕代表性概率样本;覆盖偏差;空间抽样

在社会科学的量化研究当中,如何对目标总体进行科学、精确的抽样,是决定研究成果优劣的先决要件。只有经由"概率抽样"(probability sampling)才能获得对总体有"代表性"的样本(representative sample)。否则任何后续的分析都将不具有真正的科学归纳意义,更勿论科学推论意义。

在现实世界当中,对任何社会问题的研究,大多状况下,人们不太可能从事整体社会的全面普查。事实上,普查不仅不经济,也连带产生缺乏即时性的问题(Kish, 1995)[(1)]。由于我国人口基数庞大,几乎任何一个既有的或新生的社会问题,其涉及的公众往往都是海量数字。因而许多问题的解决方案,都必须建立

作者简介:沈明明,北京大学中国国情研究中心主任、北京大学政府管理学院教授、博导;王蕴峤,北京大学政府管理学院博士。

在透过严谨的科学抽样调查所获取的第一手的个人层面的数据的基础之上。而通过概率抽样获取的“有代表性的样本”是进而获得关于目标总体的可信的科学资料的前提保证。理想状态下,研究者希望获得“等概率”样本,也就是,目标总体中每一个体都具有相同的入选概率。然而,研究者在现实中往往囿于种种客观条件不足而无法达成等概率样本,因此,只要总体中个体的入选概率是已知的,不等概率抽样设计也是可以接受的。已知入选概率允许研究者通过事后分层和加权处理等工具来调整不等概率样本,从而获得可以推论总体的概率样本。所以,已知入选概率是概率抽样的底线。值得注意的是,抽样调查的后分层和加权处理非常重要,但当前的许多研究却忽略了上述工作。

抽样设计缺乏科学基础,样本的抽取不规范,那么调查结论就只能是对“样本”(被调查受访人)本身的某种态度或行为的归纳,而绝不可以就此推论“总体”(目标社会人群的整体)的态度或情况。换句话说,如此抽样方式得出的结论只能是样本分析而非科学研究所需要的总体推断。因之有时研究者会出现极端的判断错误。典型的例子有,2004 年,台湾地区的领导人选举,盖洛普台湾分公司举办了全面性的选举投票站的“出口民调”(exit poll),调查结果却与选举结果截然相反。此外,1992 年的英国大选中,两项“出口民调”结果都显示,没有任何党派能在下议院占绝对多数,但实际选举结果,保守党却在下议院取得了多数席位。之所以会出现这种不准确的预测现象,主要是由于没有考虑到或不知道如何解决抽样设计中存在的覆盖偏差(coverage bias),导致样本不能有效代表总体,从而使对总体的推论建立在不真实的基础上,得出错误结论的概率大大提高。

至于完全不考虑“代表性样本”的基本原则的“方便抽样”(如偶遇拦访、“随机”入户)、“志愿参与”(如平面和网络媒体常见的自填问卷调查)等形式的调查,当然也就不存在推论总体的条件。

能够达成“有代表性的样本”的方法是“概率抽样”。概率抽样的基础是总体中每一个体都有同等的(至少是已知的)入选概率。这些都是在教科书中就讲明了的。那么,以“随意”抽样去“随意”推论总体的研究已没有必要在学术上予以评论。

所以,本文讨论的是,在抽样数据(户籍、人口普查数据)不足和更新滞后的约束条件下,如何设计和实施概率抽样,以获得对总体有代表性的样本。

一、传统抽样的困境

概率抽样的必要条件是需要一个反映总体组成的“名单”,进而依据名单以适当的方法抽取样本(list - assisted sampling)。传统上,这个名单主要是公开的、可以获得的人口普查数据或户籍资料。

具有悠久历史的我国户籍管理制度,曾经为抽样调查提供几近完美的人口分布信息——无论城乡,户籍按行政区划和层级管理,井井有条,而且相当稳定

(人口几乎不流动);户口簿上户主和家庭所有成员从姓名、性别、出生日期、受教育程度,到迁入、迁出的记录,一览无遗。抽样设计人员在这样优越的抽样数据资料的基础上,几乎可以根本不需考虑抽样覆盖偏差问题。然而,这样的"完美状态"已不复存在。

改革开放以来,特别是上世纪九十年代以来,我国进入社会结构的快速变迁时期,数量庞大且单向的向都市倾斜集中的人口流动,使得传统的户籍制度已经无法提供精确的人口资料作为"概率抽样"的抽样框。因之,如何针对中国的特殊国情,设计一种具有科学、可行、成本适宜的抽样方式,是中国社会科学研究在研究方法上应当给以高度重视的关键课题。

随着工业和服务业的快速发展,大量的农村人口涌向城市地区寻求工作机会,形成空前规模的人口流动,从而造成相当高比例的"人户分离"现象。依据2000年第五次全国人口普查,我国约有1.4亿流动人口,占人口总数的11.6%。在北京、上海及沿海地区,流动人口的规模已经超过了当地人口总数的20%[(2)]。2000年至今,我国城市化进程进一步加快,流动人口的规模更是大幅增长。按已公布的2010年第六次全国人口普查的数据,全国流动人口超过2.6亿,占总人口的19.5%,十年增长达81%。与十年前相比,人口从乡村向城市流动的规模和速度明显加大加快。如六普数据显示,2010年北京的流动人口比例已达36%,十年增长44.5%;而上海则突破了39%,十年增长高达159%,年平均增长率为9.9%![(3)]

不仅如此,由于老城改造和城市建成区的扩展,城市户籍人口在城市内迁徙的规模很大,户籍登记更新的滞后不可避免,从而也造成大量的"人户分离"。

大规模人口流动现象严重地挑战了传统的户籍管理制度,也对所有依据户籍人口资料从事的相关研究,在样本的抽取上带来了严重的困难,尤其是高比例的"人户分离"直接导致了"覆盖偏差"已大到不能容忍的地步。因此那些依赖户籍进行抽样调查的研究者在推论时面对一个严重的问题:大至四分之一实际意义上的城市居民是没有正式登记的,因此在户籍基础上的概率抽样只是对城市居民一个子集进行的概率抽样,这样的抽样已不可能涵盖抽样范围内的所有人口。短期流动人口在试图量化流动规模的研究中常常被系统性地排除在外。

二、"GIS/GPS辅助区域抽样方法"作为解决方案

面对抽样框不能真实反映人口分布状况的困境,抽样设计人员必须寻找新的工具,开发新的方法。令人鼓舞的是,GPS空间定位技术和GIS地理信息数据管理技术的飞速发展为我们提供了新的选择。针对我国的特殊国情,笔者供职的北京大学中国国情研究中心依据多年的调查经验和反复试验,发展了一套"GIS/GPS"辅助的"区域抽样方法"(以下简称"GIS/GPS抽样方法"),作为走出因社会快速变迁所带来的抽样困境的解决方案,从而为转型社会中的社会科学研究提供了一个科学、可行且有效的抽样工具。该方法的主要优点是能够系

统地覆盖到流动人口群体,解决了传统抽样中的"人户分离"问题以及行政区域抽样中的"边界不清"问题。

北京大学中国国情研究中心最早于 2001 年"北京地区年度调查(1995 至今)"中实验了"GIS/GPS 抽样方法"。实验发现,"GIS/GPS 抽样方法",相较于过去使用的基于行政区划的户籍资料的传统抽样方法,能够避免"覆盖偏差",更精准的抽取得到有效样本,以利于后继的研究推论。[4]

2001 年的北京实验表明,应用"GIS/GPS 抽样方法"能显著改善样本的代表性。相对依据户籍名单抽取的样本,"GIS/GPS 抽样方法"在主要的人口统计学变量上,均显著地改善了对真实人口分布的统计描述。举例来说:

首先,"GIS/GPS 抽样方法"的样本中,女性受访人比例为 52%,在 95% 的置信区间上,显示男女性别的分配是平衡的。反之,户籍抽样的样本则往往出现男性比例过高的现象。也就是说,纠正了样本中因覆盖偏差造成的男女比例失衡问题。

其次,户籍抽样的样本中,受访人平均年龄为 43.1 岁;而"GIS/GPS 抽样方法"取得的样本平均年龄为 38.4 岁。这显示"GIS/GPS 抽样方法"的样本,已然包括了更加年轻的流动人口(主要是 20 – 30 多岁的年轻人);以及虽然同样持有本市户口,但却在城市内迁移的人口。因此,该样本的平均年龄,比过去依据户籍名单抽取的样本,平均年轻了六岁,能够比较逼近社会人口分布的真实状况。

再者,对于"收入变量","GIS/GPS 抽样方法"的样本,平均收入的估计(每个月全家的收入/依靠这些收入生活的人数,以元为单位计),相对于依据户籍名单抽取的样本,在不同组间,显示有很大的差别。不仅如此,当设计效应(Deff)纳入考虑时,收入上的差距则相应地拉大。统计结果显示,处于流动中的家庭是最穷困的群体,每月人均为 912 元(当不用设计效应进行修正时为 965 元)。相反,对于那些市内人户分离者而言,该数字为 1208 元,表示该群体在经济上具有优势,有更多的居住选择及迁徙的机会及能力。而处于中间水平的是那些居住在户口所在地的居民,每月人均为 990 元。

"GIS/GPS 抽样方法"作为样本的覆盖偏差的解决方案,使样本更为接近社会人口分布的实际状况,是一个能够适应我国当前的社会变迁形态的科学、有效的抽样工具。在开发之初,"GIS/GPS 抽样方法",主要用于解决在中国城市中存在的流动人口问题,以适应当前我国的特殊国情。对不同国家或地区,我们认为,"GIS/GPS 抽样方法",对于解决抽样调查中所面临的系列挑战,同样也可以是一个有效的方法。在绝大多数发展中国家,政府往往都缺乏一整套完整而准确的人口统计数据。在诸如喀布尔、金边等没有人口普查数据的地方,除非研究者自己先在当地进行普查,否则没有获得一套"概率样本"的可能。任何既有的其他抽样技术显然都是不切合那里的社会真实状况(social reality),更遑论后继研究进行的科学推论。

本文强调的“GIS/GPS 抽样方法”，与传统的区域抽样方法有很大的不同。该方法不再假定社会总体的人口分布比例是可预知的。“GIS/GPS 抽样方法”严格遵循为数不多的几项实施原则，发展出来一套依据“空间抽样”的程序，可以在非预知总体中抽取“概率样本”。

当然，有关总体的人口统计学特征的信息仍然有用，但是已非必要。既有的人口资料固然不太精准，可是依然可以帮助研究决定项目实施的总量和成本，以及帮助预测样本的规模。如果实施得当，空间抽样法可以不必依赖普查数据的质量或者任何“官方”人口统计，从而成为针对任何类型的社会群体进行“概率抽样”的通用工具。

三、“GIS/GPS 抽样方法”的实施过程

“GIS/GPS 抽样方法”的实现，主要必须归功于 GPS 全球定位系统的技术发展。因为 GPS 定位仪有能力以很高的精确度，可以基于经纬度准确标示地球上任何一个极其小的区域。比如我们可以随机抽选一个“一平方秒”(square second)的区域，并使用 GPS 定位仪指引，在几米精确度之内到达目的地。之后调查者可以登记在这一平方秒区域内的所有住户。一旦被列入，每一个住户(或者按照一定比例随机抽选的部分住户)都必须被访问到。不过，这里存在一个现实的缺陷：因为入选概率不依赖于住户的数量而是空间点(K/T)的数量，因此最后的样本规模无法预先确定。我们只能依靠对人口密度的“最佳估算”来决定各级样本单元的合理数量和规模。值得注意的是，这种抽样方法的一个难点就是排空“无人区域”，北京大学中国国情研究中心近年来正积极使用“Google Earth”提供的地图信息建立 GIS 数据库来配合实施现场抽样。

“GIS/GPS 抽样方法”。其具体抽样施行细节概述如下：

(一)建立初级抽样框架

“初级抽样框架”的建立必须依据既有的人口资料及区域地图。“GIS/GPS 抽样方法”是利用经纬度网格，比如“一分格”(a square minute)，选取“初级抽样单位”(PSUs：primary sampling units)。在被调查的区域内，绘制“一分格”网格，并对其分别编码。

(二)建立次级抽样框架

“次级抽样框架”的建立同样依据既有的人口资料及区域地图。“GIS/GPS 抽样方法”同样利用经纬度下级网格，比如“一秒格”(a square second)，作为“次级抽样单位”(SSUs：square seconds units)。在选定的 PSUs 区域内，绘制“一秒格”网格并编码。

(三)修正抽样样本偏差

由于既有的人口资料具有一定的偏差(主要是由于流动人口的统计偏差)，其次还有区域间人口密度的偏差，或是由于城乡人口密度的偏差，这些偏差均会影响样本对总体的统计描述，所以我们都必须一一通过不同的修正方式予以调

整,以期抽取纠正了覆盖偏差的“代表性样本”。同时,尽最大可能排除非居民区,以降低现场实施成本。

(四)确定最终抽样框架

经过上述的三个阶段的实施,“GIS/GPS 抽样方法”还要通过各种检定,决定最终的“抽样规模”。“抽样规模”基于成本考虑及效果考虑,必须具有“最小”及“最适”的两个要件。

四、“GIS/GPS 抽样方法”的研究成果

2001 年以来,“北京大学中国国情研究中心”在中心城市、区域和全国范围内多次应用“GIS/GPS 抽样方法”,这些调查研究的结果不仅发展、完善了“GIS/GPS 抽样方法”,同时也使我们对这种抽样方法的科学性和可行性更有信心。弥足珍贵的是,通过这些实际项目的检验,我们还在实际应用过程中发展并完善了一整套的执行程序和操作规范,确保了该抽样技术的规范性。

目前,应用“GIS/GPS 抽样方法”的调查也已经逐渐增多。如北京大学中国国情研究中心实施的北京地区调查(2001,2007,2009)、中国公民思想道德观念调查(2003)、中国公民价值观调查(2004,2009)、中国公民意识年度调查(2008,2009)、公民文化与和谐社会调查(2008)等。此外,若干国际上的大型调查项目,如“世界价值观调查”(World Value Survey),也开始考虑在一些抽样数据严重不足的发展中国家中试用“GIS/GPS 抽样方法”。

与此同时,中外学者还利用北京大学中国国情研究中心基于“GIS/GPS 抽样方法”获取的调查数据进行了大量社会科学实证研究,其中比较有代表性的研究成果主要包括沈明明和李磊对“GIS/GPS 抽样方法”的实验研究(Landry and Shen,2005;沈明明,李磊,2007)[5];沈明明、王裕华对中国农民经济纠纷解决偏好的研究(沈明明,王裕华,2009)[6];杨明和陈涓对中国人法律知识的分析(杨明,陈涓,2009)[7];严洁、李磊等对 GPS 抽样方法的新探索(严洁、李磊等,2009)[8]等。这都有助于读者对“GIS/GPS 抽样方法”的了解。

五、结　论

过去,尽管理论上“空间抽样”可以做到“概率抽样”,但是社会科学家们始终没有充分利用 GPS 所带来的技术飞跃。直到最近,“空间抽样”的应用大多还仅限于自然科学。由于 GPS 技术的飞速发展,社会科学应用“空间抽样”调查成为了一个可能的事实,这促使我们开始尝试建立及应用“GIS/GPS 抽样方法”,试验的结果非常令人振奋鼓舞。我们可以得出以下阶段性的结论:

“GIS/GPS 抽样方法”可以在社会科学的量化研究中落实运用;

“GIS/GPS 抽样方法”抽取的样本可以覆盖其他抽样技术抽取的样本;

“GIS/GPS 抽样方法”可以极大地提高了流动人口的样本覆盖率;

单变量和多变量的分析均表明覆盖率的改善会产生与传统抽样调查不同的

结论。

然而空间抽样也不是万能的。在实施的过程也存在大量困难，尤其是如何在设计阶段正确地预测样本规模。从理论角度说，不准确的预测并不会使得这一方法失效，但它们会加大使用这一方法的成本。此外，无论是对样本规模的预测，还是基于以相对规模测度的 PPS 抽取 PSUs，“GIS/GPS 抽样方法”在相当程度上仍然要依靠官方数据。

但是，“GIS/GPS 抽样方法”能够在很大程度上纠正使用这些官方数据可能产生的问题，从而人们可以更好地掌握造成抽样偏差原因及可能后果。此外，该方法经过多年研发已经积累了大量经验和教训，大量缩减了实施成本，具备了经济适用性，具有进一步推广的潜力。近来，“北京大学中国国情研究中心”还将进行一系列利用 Google Earth 地图和卫星图片的明亮度等级来排空“无人居住区”和识别人口密度的实验，以进一步完善“GIS/GPS 抽样方法”。

在多数发展中国家和一些特殊的环境中，获得良好的人口统计数据绝对只是一种奢望。在“经济”及“有效”的双重考量下，“GIS/GPS 抽样方法”对社会科学研究而言不失为一个较好的解决方案，可以比较精准地抽取得到对目标总体“有效覆盖”的“代表性样本”。而精准的抽样一定是量化分析推论的基础！

注释：

(1) Leslie Kish, John Wiley & Sons, 1995, New York.

(2) 国家统计局《中国乡镇街道人口资料》，中国统计出版社，2002 年。

(3) 国家统计局《2010 年第六次全国人口普查主要数据公报》(第 1 号)。

(4) 沈明明、李晶《流动人口、覆盖偏差和 GPS 辅助的区域抽样方法》，《理论月刊》2007 年第 6 期。

(5) Pierre F. Landry, ShenMingming, “Reaching Migrants in Survey Research: The use of the Global Positioning System to reduce coverage bias in China”, Political Analysis, Vol. 13, No. 1, Winter 2005: pp. 1 – 22; 沈明明、李晶《流动人口、覆盖偏差和 GPS 辅助的区域抽样方法》，《理论月刊》2007 年第 6 期。

(6) Shen Mingming, Wang Yuhua, “Litigating Economic Dispute in Rural China”, The China Review, Vol. 9, No. 1, Spring 2009: pp. 97 – 121.

(7) Yang Ming and Chen Juan, “The Rule of Law in China: If It Has Been Built, Do People Know about It?”, The China Review, Vol. 9, No. 1, Spring 2009: pp. 123 – 145.

(8) Yan Jie, Pierre F. Landry, Ren Liying, “GPS in China Social Surveys: Lessons from the ILRC Survey”, The China Review, Vol. 9, No. 1, Spring 2009: pp. 147 – 163.

〔责任编辑：嘉　耀〕