社会学视角下的大数据方法论及其困境

文/鲍　雨

摘要：大数据不仅是一种庞大数据资料的称谓，也是一种获取和分析资料的方法。在社会学研究领域，大数据的引入被认为是定量研究的范式下所做出的收集和分析资料方法的创新。但是大数据在多大程度上反映人群的状况、能否解决被研究者的主观性问题、如何洞悉变量间关系的真伪、如何解决数据缺失问题等方面，还存在一定程度的方法论困境。这就要求研究者在使用大数据进行研究时，能够在研究对象与问题的选择、变量的选取、数据的清理等一系列工作中更加谨慎，切勿犯简化社会生活、盲目推广分析结果的错误。

关键词：大数据；方法论；定量研究；方法论；困境

中图分类号：C91-03　**文献标识码**：A　**文章编号**：1006-0138(2016)03-0048-05

一　引　言

近年来，“大数据”（big data）作为一个热门话题在社会各领域被广泛讨论。一般认为拥有“4V”的特征的数据集合是大数据：1）规模性（Volume），数据的体积巨大；2）高速性（Velocity），数据产生、处理和分析的速度快，数据具有实时性，且数据流量大。3）多样性（Variety），大数据的类型复杂，除传统的结构化数据之外，还包括大量非结构化数据。4）价值性（Value），数据的整体价值及商业利润高，潜力巨大。[1] 李天柱等在大数据的“4V”特征的基础上进一步指出：“假设存在规模庞大、类型多样、动态产生且价值巨大的‘特殊数据’集合，那么大数据可以定义为，在此‘特殊数据’集合的基础上，借助计算科学及其它新兴技术来实现特定功能的整体架构。”[2] 大数据的魅力除了“大”，更在于它将多学科、跨领域的数据结合在一起，开辟了传统方法所不能实现的、更为创新的研究路径。

需要注意的是，大数据不仅是用于研究的经验材料，也是一种获取材料的方式和运用材料的方法，具有独特的方法论逻辑。从已有的文献来看，大部分是对大数据的赞扬之声，一些学者甚至断言大数据带来了社会科学范式的革命，[3] 超越了定量与定性研究的对立，[4] 少数对大数据方法的质疑也仅仅停留在数据获取的伦理问题上。[5] 然而，大数据作为一种收集和分析资料的方法，能不能用于社会学研究之中，它在方法论上的基础是什么，其适用

作者简介：鲍雨，江苏省社会科学院社会学所助理研究员，社会学博士，南京市，210003。

性和边界又在哪里，这些问题依然没有得到充分探讨和回答。本文正是以此为起点，试图在梳理社会科学领域大数据研究的基础之上，阐述大数据在社会学研究中的方法论逻辑，并说明大数据在应用过程中的方法论困境。

二　大数据的方法论逻辑

大数据是将社会生活数字化、数据化、变量化，再通过测量这些变量，提取量化信息，得到关于这个变量的描述以及多变量关系的分析。大数据方法与定量研究范式有着共同的认识论基础，但又有其特有的方法论特征。

（一）样本趋近总体

社会学定量研究为调查某一社会现象在总体中的情况，在抽样框中随机抽取一定比例的样本，通过描述和统计等一系列方法，得到样本之中不同变量间的关系，并以此为依据推论总体之中这些变量之间的状况。而以大数据为经验材料的研究不涉及随机抽样的问题，而是将全部总体纳入到分析的框架内，得出的变量之间的关系也无需参数检验，而直接用于反映总体趋势。例如，2015 年 12 月淘宝联合第一财经商业数据中心推出的《淘宝大数据解读中国消费趋势》的系列报告，[6] 并不是将用户做随机抽样形成样本，然后调查他们的性别、地理位置、购物倾向等，而是直接将 3.86 亿淘宝用户这一总体作为分析的对象，描述总体的消费状况并预测未来消费趋势。

（二）利用非传统方式获取数据

一方面，大数据不同于传统问卷调查依靠被研究者的主诉来获得资料，而是直接利用技术手段对被研究者的行为进行检测。另一方面，大数据的获取不依赖传统的社会统计部门。例如，全国性人口普查是关于人口的最“大”的数据，理论上涵盖每一个个人的信息，体量巨大，却不能被称为“大数据”，因为人口普查数据通过接触被调查者直接获取，并依赖传统的社会统计部门逐级上报。大数据往往基于现代科技手段，采取实时监控、测量、存储的方式整合海量信息，例如交通流量监控、气象水文监测数据、环境监测数据、商业物流的数据记录，尤其是近年来基于互联网的浏览、搜索引擎、上传下载等行为的大数据，更是成为了大数据的主流——甚至几乎成为了“大数据”的同义词。

（三）通过归纳建立模型

传统的定量研究先提出假设，并设计基于假设的待检验模型，进而通过分析数据证明或证伪假设及相关统计模型。而大数据方法是通过对海量的数据进行分析，寻找变量之间的关系，而后建立模型。正如张晓强等所说：“数据科学以海量的数据为研究对象，通过数据挖掘等手段来寻找海量数据中潜在的规律。它研究各个科学领域所遇到的具有共性的数据问题，通过对数据的规律的研究来实现对科学问题的解答。”[7] 这也就意味着通过大数据方法建立的模型并不反应必然规律，而是在非实验控制的条件下，一系列影响因素综合作用的结果。虽然这种模型具有模糊性与偶然性，却可以在一定程度上预测现象发展的趋势。

（四）呈现相关关系而非因果关系

舍恩伯格等认为，大数据“不是因果关系，而是相关关系”。[8] 定量研究目的是为变量的变异性提供因果解释，用其他变量解释所要研究的变量的变异性。而应用大数据，尤其在商业领域，其目的是销售商品，这种情况下商家只需要了解用户的行为与销售额的相关关系为何，并以此作为决策依据，向用户推荐可能需要的商品，指导商业活动，而不需要为这种相关性做出解释。例如，在美国沃尔玛在季节性风暴到来时，将蛋挞和飓风用品摆放在一起，以增加蛋挞的销售量。[9] 分析两个变量关系时，仅仅止步于相关关系，即使这个相关关系是虚假关系，或者通过其他变量相互关联。这并不是说大数据不能为因果解释提供数据基础，而是说因为在一些大数据广泛应用的领域，人们并不探究因果，而是利用大数据呈现现象，预测现象发展的趋势，为决策提供依据。

三　大数据的方法论困境

大数据的使用虽然也建立在统计与计算的基础之上，但是却有着不同于传统定量研究范式的方法论特征，这些特征使得研究者在运用大数据的过程中不可避免地遇到一些方法论困境，这就要求研究者在分析和结论推演时谨慎

行事。

（一）被研究者的主观性问题

大数据更多强调对个体活动的监测来获取数据，而非如传统的定量调查采取问卷的方式。其中隐含的预设是，个体的主诉是含混不清的，被研究者有意隐瞒或是无意遗忘都可能会影响到数据的信度与效度，而对个体活动监测得来的数据就可以避免被研究者主观意向的介入对数据质量的影响，例如我们在调查被研究者的收入情况时，如果采用问卷调查的方法，由被研究者自主填写，可能会出现由于记忆不准导致误报或者故意瞒报的情况，而如果有技术可以做到对被研究者银行账户收支的数据进行监测，那么后者将最为接近被研究者收入的“真实”情况。

但是在研究之中，排除被研究者的主观参与而对其行为的监测存在很多问题。首先，社会学的研究对象是社会现象，社会现象需要人的参与，但是个人的行为与社会现象是两个不同的概念。行为必须与外部世界中的他人发生联系并主观指向他人才能被称为社会现象，所以任何社会现象都有作为主体的人的主观参与，仅仅依靠观察个人的活动轨迹并不能说明发生了某种社会现象。例如，我们可以利用视频监控观察一定时间地点内的人群流动状况，但是我们无法区分人群是随机地在该地出现，还是发生了集会、游行等社会运动。所以用行动代替社会现象是一种概念的偷换。

其次，当我们利用行为监测来获得可供研究的经验数据时，数据的完整性依靠于我们监测的手段和方法的可及性与适用性。以调查收入为例，如果要完整地掌握一个人的收入状况，我们不仅需要将他名下的账户收支都调查清楚，还要考虑到他日常生活中的现金流动状况，即便我们可以使用银行的数据，但是还是难以监测他在日常生活中的收支状况，也就是说仅利用银行大数据也无法监测到此人的完整收入信息。所以在现有手段和技术的条件下，直接向被研究者询问的问卷法，依然是最有效的调查方法。

再次，虽然大数据的使用者声称用监测其行为的方式替代了被研究者的自我叙述，避免了其主观意志的干扰，但是在互联网的虚拟环境下，数据化的信息有很大一部分来自于被研究者的键入。例如社交网站上的性别、地点、爱好等信息，依然主要依赖被研究者的自我键入，他们有可能会胡乱填写一些错误信息，那么这种利用互联网大数据的调查方法比面对面的问卷调查更加难以保证材料的真实性。

（二）研究对象的局限性

大数据方法把总体作为分析的样本，直接分析总体的情况并建立模型，免去了随机抽样的过程。这种方法认为将总体作为样本避免了随机抽样过程中的抽样误差的产生，能够精确地反映总体的变化趋势。但是大数据方法能够获取的“总体”本身是存在偏差的。[10]

首先，由于大数据抓取方式的特殊性，研究者的分析可能产生系统性偏误。也就是说，研究者仅能获得“能够被抓取”的信息，而大量不能被抓取的信息则被排除在了总体之外。到 2016 年初，中国有 6.88 亿网民，[11] 而中国总人口 13.68 亿，那么通过网络抓取个人信息的方式来获得的大数据依然无法涵盖不会在网上留下痕迹的 6.8 亿非网民。在实际的研究中，研究者往往仅依靠一个或几个网站的后台数据作为分析资料，那么这种大数据仅是使用该网站的用户的数据，不能将分析结果推论到其他网民，更不能推论到全体国民。因此，研究者必须注意，大数据中所谓的总体是有限的总体，大部分难以通过大数据方法抓取的个体并不被作为研究对象包含在总体中。

第二，即使在同一数据收集平台上，由于不同的个人活跃性不同，其信息被抓取的概率也不同，因此这些大数据的形成既不是抽样，也远非随机，而是具有极大的偶然性。英国广播公司在 2011 年通过互联网上自助填写问卷的形式进行了一项英国阶层调查，并以此为根据将英国社会分为 7 个阶层。[12] 该调查共有 161458 人参加，样本规模远超传统的问卷调查，有的学者将该调查所获得的数据界定为大数据，并认为该调查由被研究者填写，可以摆脱以往的大数据方法多是对行为进行观察而忽视个人主观性的方法论困境。[13] 然而网络调查的样本仅仅是在问卷投放网络的一段时间内发现并有

意愿填写问卷的网民，如果一个网民在这一时间并不活跃，那么他将不被纳入调查的样本之中，因此通过该网络调查得出的结果只是一种偶然关联，一种统计学上的相关，而不具有任何推论价值。

（三）变量关系的真伪问题

如前所述，一般情况下人们在使用大数据时仅仅关注两个事件的相关关系。而对社会学研究来说，研究者需要描述特定社会现象的变异，再通过了解该社会想象与其他社会现象（变量）的因果关系，为该现象的变异提供解释依据。所以当我们试图用大数据作为材料来进行解释社会现象时需要非常谨慎，可能两个具有高度统计相关的变量并不具有社会学意义上的因果关系。

第一，在一些领域应用的大数据变量间的关系为虚假关系。例如飓风用品的销售量和蛋挞的销售量呈现成比例相关，但是二者并不具有因果关系，而是共同的受到另一变量即飓风天气因素的影响，控制天气因素则二者相关关系自动解除。所以在使用大数据分析两个变量关系时往往需要控制其他变量，以达到辨别相关关系真伪，进而对社会现象进行因果解释的目的。

第二，由于大数据旨在一个较大的数量级上关注两个变量变异的总体趋势预测，所以往往忽视样本内部个体之间的差异，试图用一个变量完全解释另一个变量的变异，导致层次谬误。Artés对西班牙选举日的天气和投票大数据进行分析，发现如果天气状况不佳，则保守政党得票比例将增高。如果分析止步于此，就会得到结论：天气状况影响保守派的得票率。但是作者将投票者的社会经济地位作为变量带入分析之中，进而论述天气不佳会影响较低阶层的人们出行，而较低阶层的人群更多地将选票投给左翼政党，所以天气不佳，较低阶层的投票者数量减少，左翼政党得票比例下降，保守政党得票比例上升。[14] 控制社会经济地位这一变量，则可以看到天气仅对低阶层人群的投票率产生影响，所以仅关注数据的整体趋势则可能导致层次谬误的出现。

第三，利用大数据方法获取资料是否要进行显著性检验、应该如何进行相关检验，依然存疑。一方面，大数据方法声称其样本即总体，那么既然直接分析总体趋势的变化则可以免去显著性检验的过程。然而如前所述，大数据方法能够获取到的“总体”和作为研究对象的“总体”总是存在着错位，大数据“总体”（样本）的获取也并不是采用概率抽样，所以该不该进行显著性检验的问题就摆在了研究者面前。另一方面，即使暂且认为大数据需要进行显著性检验，通常显著性检验是针对正态分布的较小样本而进行的，样本规模会对显著性检验的结果产生影响，而大数据由于样本规模庞大，所以在分析的过程中很容易发现统计显著性，[15] 那么就有可能导致这样一种错误：被发现变量之间的关系是偶然出现的而并不具有任何规律性，却仅仅由于庞大的样本量而被证明具有显著性。

（四）数据缺失问题

不管是辨别相关关系的真伪，还是避免层次谬误，都离不开数据本身包含的变量的规模，只有在数据包括足够丰富的其他变量时，我们才能够引入或者控制这些变量，已达到解释因变量变异的目的。举例说明，当研究者对人群的收入进行研究时，不仅仅要看作为整体的收入变量如何分布，还需要收集人群的其他变量，例如性别、年龄、收入、职业等，分别分析收入在不同人群之中的变异情况，从而对收入不平等进行解释。也就是说，社会学研究需要的是多变量的数据矩阵，而不是样本庞大但变量单一的大数据。然而做到这一点并不容易。

第一，如果一个大数据样本量巨大，但只包括单一的变量，那么只能对这个大数据中的这个变量进行描述统计，则这个大数据不能作为社会学分析的材料，例如单一的地区人均收入数据、流行病死亡率数据等。如果研究者想要使用这些数据，必须将时间、地点等内容作为中间变量，将该大数据与其他数据库进行对接，以此获得较多的分析所需的变量。

第二，一些网络大数据由于其获取形式的非常规问题，导致数据本身的模糊性和混杂性，致使不可避免地出现数据缺失的情况。假设要对一个匿名的不需要身份验证的社交网站上的

内容做分析，我们难以保证所有用户都键入了研究所要分析的社会特征信息，例如性别、毕业学校、薪资等内容，这就必然导致大量的数据缺失，面对海量的缺失数据，任何删除和填补的补救措施都会对分析的结果产生影响，因此关键变量的大量数据缺失使得任何统计分析的结果都存在偏误。

四 结 语

大数据近年来的广泛应用，大大拓展了社会学的理论视野。然而，任何单一范式中的理论与方法都有其适用性和解释边界。大数据将社会生活数字化、数据化、定量化，认为社会生活的本质是由信息构成的，[16]主张测量一切，实质上是用一种科学主义简化论的世界观来看待社会生活。社会学是一门多研究范式的学科，以统计分析为特征的大数据方法，充其量只能作为社会学诸多研究范式中的一种，并不会带来社会学的范式革命。

数据，是世界通过我们的感觉和工具呈现给我们的东西，而知识，是我们对数据的理解与诠释。大数据不是“告诉”了我们世界如何运作，而仅仅是呈现给我们需要解读的材料，如何对数据进行理解与诠释，还需要具有一定知识结构和理论背景的研究者发挥社会学的想象力。大数据是我们认识世界的工具，并不能代替研究者的理性思考，也不会带来社会学研究范式的根本转变，因此神化大数据的诸多论断，在本文看来都是不可取的。

注释：

[1]赵勇、林辉、沈寓实：《大数据革命——理论、模式与技术创新》，北京：电子工业出版社，2014年，第3页。

[2]李天柱、王圣慧、马佳：《基于概念置换的大数据定义研究》，《科技管理研究》2015年第12期。

[3]罗玮、罗教讲：《新计算社会学：大数据时代的社会学研究》，《社会学研究》2015年第3期。

[4]夏国美：《大数据时代的社会学审视》，《新疆师范大学学报》（哲学社会科学版）2016年第1期。

[5]静思英：《大数据及其反思》，《湛江师范学院学报》2014年第4期。

[6]《淘宝最全数据解读消费趋势》，2015年12月8日，http://roll.sohu.com/20151208/n430227084.shtml，2016年1月4日。

[7]张晓强、杨君游、曾国屏：《大数据方法：科学方法的变革和哲学思考》，《哲学动态》2014年第8期。

[8]维克托·迈尔－舍恩伯格、肯尼思·库克耶：《大数据时代：生活、工作与思维的大变革》，盛杨燕、周涛译，杭州：浙江人民出版社，2015年，第67页。

[9]维克托·迈尔－舍恩伯格、肯尼思·库克耶：《大数据时代：生活、工作与思维的大变革》，第73页。

[10]D.A.McFarland, H.R.McFarland, “Big Data and the Danger of Being Precisely Inaccurate”, *Big Data & Society*, vol.2, no.2 (2015), pp.1–4.

[11]中国互联网络信息中心：《中国互联网络发展状况统计报告》，2016年1月22日，http://cnnic.cn/gywm/xwzx/rdxw/2015/201601/t20160122_53283.htm，2016年2月4日。

[12]BBC Science: The Great British Class Survey Results, 2013年4月3日，http://www.bbc.co.uk/science/0/21970879，2016年1月22日。

[13]R.Burrows, M.Savage, “After the Crisis? Big Data and the Methodological Challenges of Empirical Sociology”,*Big Data & Society*, vol.1, no.1(2014), pp.1–6.

[14]J.Artes, “The Rain in Spain: Turnout and Partisan Voting in Spanish Elections”, *European Journal of Political Economy*, vol.34 (2014), pp.126–141.

[15]R.Tinati, S.Halford, L.Carr, C.Pope, “Big Data: Methodological Challenges and Approaches for Sociological Analysis”,*Sociology*, vol.48, no.4(2014), pp.663–681.

[16]维克托·迈尔－舍恩伯格、肯尼思·库克耶：《大数据时代：生活、工作与思维的大变革》，第125页。

责任编辑 刘秀秀