《学术界》(月刊)
总第240期,2018.5
ACADEMICS
No.5 May. 2018

中国实证社会科学的演进及使用大数据研究之现状与挑战〔*〕

○ 何晓斌,李　强
(清华大学　社会学系,北京　100084)

〔摘　要〕我国的实证社会科学是改革开放以来在开展全国性社会调查和学习借鉴国外实证社会科学研究方法及技术的基础上发展起来的。这些实证社会科学的数据来源包括:研究者自己组织实施的社会调查,中央和地方党政机关公布的统计数据,以及国内外科研机构公开的数据库,还有最近越来越受到关注的大数据。通过对中美两国三大社会科学代表期刊使用传统数据和大数据的实证研究论文的分析,发现中国社会科学采取实证方法研究的程度与美国相比还有一定差距,但是在使用大数据的实证社会科学研究方面,中美两国没有显著的差别。使用大数据的中国实证研究论文也大多结合了传统数据来源,采用了传统计量模型来研究和探讨因果机制。因此,使用大数据的实证社会科学的研究目前还只是基于传统数据的实证研究的一个有益补充,还没有产生完全突破传统实证研究方法的颠覆性研究方法和范式。导致这个结果的原因主要在于大数据的不够开放,简易经济的大数据获取和分析工具的缺乏,以及大数据本身的代表性缺陷。因此,要推动大数据实证社会科学的发展,需要加强政府和大型高科技公司与高校和科研机构的合作,积极开发大数据应用的合适工具,建立大数据实证社会科学的媒介和平台,不断改进使用大数据的实证研究的方法。

〔关键词〕实证社会科学;大数据;定量研究;研究范式
DOI:10.3969/j.issn.1002-1698.2018.05.002

实证社会科学研究是指基于实际调查或者访谈资料来验证理论假设或者构

作者简介:何晓斌,清华大学社会学系副教授,主要研究方向:经济社会学、组织社会学、计量社会学;李强,清华大学社会学系教授,主要研究方向:社会分层与流动、城市社会学、应用社会学。

〔*〕本文系国家社科基金重大项目"大数据时代计算社会科学的产生、现状与发展前景研究"(项目编号:16ZDA085)的成果。

建理论的研究范式。区别于纯理论思辨式的传统社会科学研究，实证社会科学研究的重要基础是获得有代表性的研究对象的详实数据。改革开放以来，我国的实证社会科学是在开展全国性社会调查和学习国外实证社会科学研究方法及技术的基础上发展起来的。

一、改革开放以来中国实证社会科学的演进及其数据来源

1978 年以后，随着“实践是检验真理的唯一标准”的确立，随着国家统计机构的建立，我国开始实行普查制度，先后开展了四次人口普查：1982 年的第三次全国人口普查，1990 年第四次全国人口普查，2000 年的第五次全国人口普查，以及 2010 年的第六次全国人口普查，获得了一些重要的基础数据。[1] 除了人口普查，在其他专题领域如农村、经济、企业、住房等也展开了各种各样的普查和社会调查，比如 1981 年对全国农业资源的调查；1982 年春对工人阶级状况的全国范围的大规模调查；1985 年和 1986 年两次生育力的抽样调查；1986 年和 1995 年第二次、第三次全国工业普查；1984 年开始历时两年完成的第一次城镇房屋普查；1987 年和 2006 年的第一次和第二次全国残疾人抽样调查；1997 年和 2007 年进行的两次全国农业普查；2004 年、2008 年和 2013 年分别进行了三次全国经济普查；1993 年和 2003 年的两次全国第三产业普查。[2] 这些普查和调查都为新时期党和国家战略、方针、政策的制定提供了重要依据。同时，在社会科学界也重新兴起了社会调查之风。以社会学界的调查为例，改革开放以来，在国家相关部门和机构的支持下，一大批社会学者针对中国社会的方方面面做了详细深入的研究，比如 1982 年费孝通先生倡导的对小城镇的实地调查研究；1992 年到 20 世纪末，中国人民大学社会学系组织的多次全国规模抽样问卷调查；1993 年，复旦大学社会学系和上海浦东新区社会发展局合作开展的社会变迁研究；2004 年，北京市社会科学院组织的“城区角落”的调查；1999 年，陆学艺教授主持的中国社会科学院社会学所对中国社会分层和流动问题的大规模专题调查，产生了一系列有影响的有关国家社会经济问题的重要报告、实证论文和专著。[3] 此外，由国家和知名高校科研机构主导的一些社会调查，特别是过去十几年来一些大型综合性全国社会调查的开展和数据免费对外开放，为中国实证社会科学研究提供了重要数据来源（参见下页表 1）。

同时，在中国社会科学界的对外交流和合作研究中，特别是对国外社会科学研究方法的学习和推广，使得高级统计方法和工具在实证社会科学研究中得到大量应用，并形成了比较成熟的研究范式。[4] 这些实证社会科学的研究成果，基本上都是通过目前实证社会科学常用的数据收集手段如问卷调查法、访谈法、实验法和观察法等收集、清理之后，辅之以计算机相关统计软件来计算和建模完成的。这些实证社会科学研究论文使用的数据来源往往可以分为这么几类：一是研究者自己组织收集的大型社会调查数据（问卷、实验、量表等）。这类数据收集手段需要花费的经费和时间成本都很高，研究者常常只有得到国家基金和各

级政府部门的经费支持才能完成数据收集。二是中央、地方党和政府机构公开的数据，包括统计年鉴、年报、简报，会议记录等官方数据和资料来源。随着我国电子政务公开工作的推进，这类数据的获取来源也越来越多，成本变低。三是国内外学术科研机构公开的数据库，比如北京大学中国社会科学调查中心组织收集的中国家庭调查追踪数据，中国人民大学中国调查与数据中心组织收集的中国综合社会调查数据等。这种科研机构提供的数据质量高，而且是免费的，目前成为很多实证社会科学研究者的数据来源。四是市场上可以购买的数据库，比如国内外上市公司数据库，这些数据库成为经济管理类实证研究者的重要数据来源，但是要购买这些数据库的成本很高，往往在几十万甚至上百万元以上。

表 1　改革开放以来社会科学领域比较知名的中国综合性社会调查[5]

调查时间	调查频率	调查名称	调查地区	调查规模	调查组织者	调查内容
1983–1993	每年	天津千户居民调查	天津市 9 个区 36 个街道	1000 户	天津市社会科学院社会学所和市人民政府办公厅	家庭成员情况、婚姻、消费、家庭关系、职业、家务劳动及分工、闲暇时间利用、教育、居住环境和工作态度等
1988–1999	每年	百县市经济社会调查	全国 31 个省、市、自治区	119 个县	中国社会科学院社会学所	婚姻家庭、人口与生育、就业与流动、居民消费、社会关系、政治参与等
1997	当年	沿海发达地区社会变迁调查	沿海 14 个城市和 4 个特区	3013 个成年人	中国社会科学院社会学所	人口基本情况、工作和职业、婚姻家庭等
1988 年至今	每五年	中国家庭收入调查(CHIP)	全国 31 个省、市、自治区	19000-160000 户	北京师范大学中国收入分配研究院	中国城市、农村、流动人口的详细收入来源、资产来源数据
2003 年至今	每两年	中国综合社会调查(CGSS)	全国 31 个省、市、自治区、直辖市	10000 多户	中国人民大学中国调查与数据中心，香港科技大学	社会、社区、家庭、个人多方面的具体情况
2008 年至今	每两年	中国家庭追踪调查(CFPS)	全国 25 个省、市、自治区、直辖市	15000-16000 多户	北京大学中国社会科学调查中心	中国社会、经济、人口、教育、健康等的变迁
2011 年至今	每两年	中国健康与养老追踪调查(CHARLS)	全国 150 个县级单位，450 个村级单位	1.7 万人	北京大学中国社会科学调查中心	45 岁以上中老年家庭和个人的健康、社会关系等微观情况
2007 年至今	每两年	中国劳动力动态调查	29 个省、市、自治区	14000 多	中山大学社会科学调查中心	教育、工作、迁移、健康、社会参与、经济活动、基层组织等
2006 年至今	每两年	全国社会状况综合调查(CSS)	全国 31 个省、市、自治区的 151 个区市县	700-10000 多个	中国社会科学院社会学所	人口基础信息、劳动与就业、家庭结构、家庭经济状况、社会分层与流动、社会保障、价值观等
2010 年至今	每两年	中国家庭金融调查(CHFS)	全国除西藏、新疆、内蒙和港澳地区外的 2585 个市/县	8000-40000 多户	西南财经大学中国家庭金融调查与研究中心	家庭的资产与负债、收入与支出、保险与保障、人口与就业等

资料来源：笔者根据水延凯主编的《中国社会调查简史》(中国人民大学出版社，2017 年)第 361 – 363 页及其他公开资料整理。

近年来，随着大数据概念的出现，[6] 大数据的重要性和应用前景随着各行各业的广泛讨论已经得到了商业、政府部门和科研机构的高度关注。[7] 大数据受到关注是过去二十多年来以互联网为基础的信息科技高速发展和广泛应用的结果，特别是移动互联网的发展和移动设备的普及使得人类每时每刻都在生产和储存数量惊人的数据。截至 2020 年，全世界每人每天平均将产生 1.5GB 的数据；每台无人驾驶车每天将产生 4TB 的数据；一家小型工厂平均每天能产生高达 1PB 的数据。[8]《大数据时代》的作者维克托认为大数据是一种可以绕过随机采样而处理分析全部数据获得认知的一种新的方法和思维模式，大数据并不是绝对意义上的数量“大”。[9] 本文所讨论的大数据，是指主要通过互联网渠道

自动收集的，包含全体研究对象的大量数据的集合。[10] 比如，所有手机用户某一时期的使用行为数据，春节期间所有中国人的出境旅游的基本数据，政府官方网站上的所有留言数据等。这些新的数据来源的出现，以及海量的图书、报纸、期刊、照片、绘本、乐曲、视频等人文资料被数据化，并在互联网上提供给研究者存取和利用，使得原来很难或者无法量化的社会科学问题的研究成为可能。

就像 20 世纪 60 年代计算机技术和统计分析工具的出现促进了社会科学的量化和实证研究一样，[11] 大数据时代的来临和新处理工具的逐步出现也可能会对目前社会科学的研究范式和方法带来新的冲击。虽然大数据在商业领域的研究和应用已经非常活跃，[12] 但是大数据在中国社会科学研究中的应用现状到底如何，碰到了哪些挑战，有何对策，这些问题却很少有人做深入具体的分析。

二、中国实证社会科学使用大数据的研究成果现状

为了全面深入把握使用大数据的实证社会科学研究在中国的发展情况，同时兼与美国实证社会科学研究作比较，笔者专门浏览了 2006—2017 年发表在国内三大著名社会科学期刊《经济研究》《社会学研究》《政治学研究》，以及美国三大著名社会科学期刊 *American Economic Review*（AER），*American Sociological Review*（ASR），*American Political Science Review*（APSR）上的所有研究论文，对这些研究论文的数量，是否采用传统数据开展实证社会科学的研究，是否以大数据作为实证研究的数据来源等情况作了认真统计。这里的传统数据是指使用社会调查、访谈、实验、量表等形式获得的数据，而这里的大数据指的是从互联网网站、银行交易系统、卫星传感器等渠道获得的以研究对象全部数据作为实证研究论文全部或者部分论证来源的数据类型。[13] 统计结果如表 2。

表 2　中国三大社会科学期刊实证研究论文统计[14]

	《经济研究》			《社会学研究》			《政治学研究》		
年份	文章总数	传统数据	大数据	文章总数	传统数据	大数据	文章总数	传统数据	大数据
2017	196	134	1	63	17	1	68	13	0
2016	180	119	2	62	19	0	65	5	0
2015	185	110	2	62	20	0	71	9	0
2014	202	127	1	65	21	0	76	4	0
2013	169	101	1	64	26	1	85	5	0
2012	179	143	0	68	17	0	82	7	0
2011	186	119	1	64	20	0	86	7	0
2010	170	114	0	64	18	0	88	8	0
2009	162	106	0	66	18	0	91	2	0
2008	166	100	0	70	14	0	110	4	0
2007	163	110	0	81	13	0	66	1	0
2006	142	105	1	72	19	0	66	0	0
总体	2100	1388	9	801	222	2	954	65	0

数据来源：笔者根据三大期刊发表的论文人工统计。

从上述统计结果来看，以传统数据为基础进行的实证研究比例最高的是经

济学，2006—2017 年采用传统数据为基础发表的实证论文占所有发表论文总量的比例平均为 66%（最低年份的比例为 59%，最高年份的比例为 80%），也就是说，目前中国大部分经济学的研究都采用计量和统计模型为立论基础的实证主义研究范式。比例次高的是社会学，12 年中以传统数据为基础发表的社会学研究论文平均占到 28%（最低年份的比例为 16%，最高年份的比例为 41%）。比例最低的是政治学，平均只有 7%（最低年份的比例为 0%，最高年份的比例为 19%）。类似地，以大数据为基础发表在三大期刊上的实证研究论文可以说是屈指可数，12 年间在《经济研究》上共有 9 篇，《社会学研究》上共 2 篇，而《政治学研究》上 1 篇都没有，三大期刊在过去 12 年使用大数据的实证研究论文占所有发表论文的比例平均不到 1%。

再看看发表在美国三大著名社会科学期刊上的实证研究论文的情况（参见表 3），我们可以看到：

表 3　美国三大社会科学期刊实证研究论文统计

	《美国经济学评论》			《美国社会学评论》			《美国政治学评论》		
年份	文章总数	传统数据	大数据	文章总数	传统数据	大数据	文章总数	传统数据	大数据
2017	122	88	0	41	29	3	50	30	0
2016	130	76	0	47	43	1	55	31	0
2015	117	74	1	46	37	2	47	26	3
2014	150	87	1	49	40	3	50	34	0
2013	108	63	0	43	39	0	50	36	1
2012	123	71	0	43	32	0	44	32	0
2011	122	90	0	40	33	0	42	22	0
2010	106	69	0	40	33	2	43	28	0
2009	101	59	1	46	34	0	37	25	0
2008	100	47	0	45	29	0	35	20	0
2007	103	49	0	44	34	0	53	24	0
2006	101	47	0	47	31	0	59	22	0
总体	1383	820	3	531	407	11	565	330	4

数据来源：笔者根据三大期刊发表的论文人工统计。

《美国经济学评论》上以传统数据为基础的实证研究论文 12 年平均占比为 59%（最低为 47%，最高为 74%），比中国《经济研究》的相应比例还稍微低一些，但总体差别不大。但是在社会学和政治学领域，美国实证研究论文的比例要显著高于中国相对应的期刊。《美国社会学评论》和《美国政治学评论》上实证研究论文占全部发表文章数量的比例分别为 77%、58%，而中国对应期刊的所占比例分别为 28%、7%。从使用大数据的实证社会科学研究来看，美国的数量稍微多些，但是差别不大。美国三大期刊发表的大数据实证研究论文的总数为 18 篇，而中国三大期刊的总数为 11 篇。中美三大社会科学期刊上使用大数据的实证研究论文占所有论文的比重过去 12 年平均都不到 1%。因此，目前整个美国社会科学界和中国社会科学界如果单从大数据实证研究论文的数量上来看，使用大数据进行实证研究都处于早期发展阶段。

如果我们把《经济研究》和《社会学研究》上使用大数据发表的实证社会科学的论文再做仔细分析的话(参见表4),可以发现:中国经济学的研究继承了一贯的注重量化实证研究的传统,在使用大数据的实证研究创新方面也引领了整个中国社会科学界。

表4　中国社会科学期刊使用大数据的实证研究论文数据类型、计量模型和研究类别

期 刊	刊 号	题 目	数据类型	计量模型	研究类别
《社会学研究》	2017年第1期	代内“文化反授”:概念、理论和大数据实证	新浪微博和百度搜索2013-2015年网络热词的每日词频指标	时间序列和面板数据模型	因果机制
	2012年第4期	韩国人推特网络的结构和动态	2010年8月1日至9月30日期间的1133365个韩国人推特帐号数据	社会网络概念和模型	描述性和探索性
《经济研究》	2017年第2期	钱随官走:地方官员与地区间的资金流动	2006—2010年共中国人民银行大额支付数据+官员公开任职经历数据	OLS线性回归	因果机制
	2016年第2期	政治关联与经济增长——基于卫星灯光数据的研究	美国国家海洋和大气管理局(NOAA)的全球灯光数据+部长出生地、籍贯、工作地等信息	面板数据回归模型	因果机制
	2016年第8期	开发时滞、市场不确定性与城市蔓延	DSMP/OLS夜间灯光数据和Landscan全球人口动态分布数据	OLS回归和面板数据模型	因果机制
	2015年第12期	互联网搜索行为能帮助我们预测宏观经济吗?	互联网搜素大数据+官方统计数据	时滞回归模型	探索性研究
	2015年第7期	网络安全风险感知与互联网金融的资产定价	百度大数据、余额宝7日年化收益率数据+官方利率数据	OLS回归模型	因果机制
	2014年第7期	聪明的投资者:非完全市场化利率与风险识别——来自P2P网络借贷的证据	人人贷网络借贷平台的数据	Probit和OLS回归	因果机制
	2013年第1期	中国式拆迁、投资者保护诉求与应计盈余质量——基于制度经济学与Internet治理的证据	“中国式拆迁”在搜索引擎上的关注度数据+上市公司数据	OLS回归模型	因果机制
	2011年第1期	投资者关注与IPO异象——来自网络搜索量的经验证据	谷歌趋势提供的搜索量数据+上市公司数据	OLS线性回归	因果机制
	2006年第12期	信誉的价值:以网上拍卖交易为例	eBay电子商务网站网上拍卖交易数据	Probit回归模型	因果机制

《经济研究》上发表的以大数据为基础的计量经济学研究的数据类型包括:美国国家海洋和大气管理局(NOAA)公布的全球灯光数据;DSMP/OLS夜间灯光数据和Landscan全球人口动态分布数据;百度搜索词指数;余额宝七日年化收益率数据;人人贷网络借贷平台的数据;拍卖网站eBay公司的拍卖数据。而仔细分析这些论文可以发现,这些使用大数据的实证研究论文基本上都只是把大数据作为整篇论文实证论证的一部分,或者把大数据作为更好测量论文构念的一个来源,比如用灯光数据来测量经济总量,同时跟官方的一些统计数据相结合来验证理论模型。而只有少数论文的数据全部来源于大数据,比如人人贷的网站数据,ebay公司的拍卖数据。

《社会学研究》上的这两篇使用大数据的实证研究的论文,论证基础全都是大数据,一是百度搜索热词,二是社交网络数据。第一篇有关代内文化反授的文章以“网络热词”的传播为例,利用提取自新浪微博和百度搜索2013—2015年的网络热词的每日词频指标进行了流行文化传播规律的探索,利用时间序列的宏观分析和面板数据的微观分析证实了“文化反授”模式的存在。第二篇研究

者搜集了从 2010 年 8 月 1 日 0 时起到 2010 年 9 月 30 日 24 时止两个月内 1133365 个韩国人账户创建的 77452090 个推特(Tweet),对韩国人推特的内容进行了描述,对于内容传播的规律和特征进行了探索性的分析。

在计量模型的运用上,这些使用大数据的实证研究论文所使用的计量模型也都是为学术界所接受和熟悉的成熟的社会科学常用的统计模型,如线性和非线性回归、时间序列和面板数据分析等。[15] 在研究类型上,这些使用大数据的实证研究论文跟使用传统数据的论文一样,主要注重于社会科学领域的因果机制。

综上而言,尽管一些使用大数据的实证研究拓展和加深了我们对社会经济运行和人类行为规律的认识,但截至目前还没有产生对传统实证研究范式有重大突破的成果。目前使用大数据研究的实证研究论文大部分只是把大数据作为对传统数据来源的一个有益补充。按照目前的发展现状来看,这些使用大数据的实证社会科学研究短期内不可能取代传统的研究手段。这说明,大数据量化实证研究虽然在很多研究者看来有非常好的前景,但是目前还远远没有成为探索研究社会科学问题的主流研究手段和方法。[16]

三、使用大数据的中国实证社会科学研究发展的挑战及对策

总体而言,当前大数据作为一种新的数据来源,还只是以传统数据为基础的实证社会科学研究的一种补充。完全应用大数据做出原创性实证社会科学研究的还极少。实证社会科学研究的基础是高质量的数据,目前的中国社会科学,除了经济学,社会学和政治学在使用传统数据基础上的实证研究程度还远远低于美国的社会学和政治学学科。在使用大数据的实证社会科学发展程度上,我国目前跟美国没有显著差别。[17]

目前使用大数据的实证社会科学的发展还处于初步阶段,主要受制于以下几方面的原因:

一是在大数据的获得上还有很大的制度障碍。目前大数据的两个主要来源是政府和大型互联网高科技公司。而我国政府部门的大数据的整合和开放的程度较低,政府各个部门或出于各自的部门利益,或出于安全考虑,或由于开发成本问题,很多的大数据都没有公开,“信息孤岛”问题普遍存在。而大型互联网公司对于大数据的开放和利用的主要动力在于商业动机和短期利益,与学术研究工作者的关注点不一样。正如维克托在其《大数据时代》书里所说的,大型科技互联网公司的主要关注点在于大数据所反映出来的客户行为的相关关系,[18] 而实证社会科学希望通过研究互联网和物联网轨迹背后的人类行为能够构建行为变量之间,或者环境变量和行为变量之间的因果机制。当然这个制度障碍的背后还有我国相关信息大数据立法的滞后。对于政府部门的大数据而言,如何在保护个人隐私的基础上合理开放政府部门的数据,如何确立大数据使用的知识产权,这些问题目前都还处于探索阶段。

二是获取成本和技能障碍。上述的制度障碍其实也可以看成获取成本的一

部分。如果数据不开放，那么通过市场上科技公司去抓取，往往也要支付相当高的成本。对于大数据的获取、使用和分析目前还缺乏相应的技能普及。一些大数据分析工具，比如文本抓取和分析工具 Python、R 等软件学习成本较高，从而给大数据的分析和使用带来不小的障碍。正如 Gary King 已经意识到的那样，[19] 大数据必须依赖合适的分析工具才能发挥其重要价值。目前在商业领域虽然出现比较流行并可能成为大数据分析标准的软件系统 Hadoop，还有各种各样的大数据分析工具和软件包，[20] 但这些工具在商业领域的应用还处于早期阶段，使用起来非常复杂，大部分社会科学研究者都还不清楚这些工具。

三是大数据本身的代表性问题。大数据的获取来源是其平台或者设备的载体，但是没有一个平台或者载体能够记录和存取所有研究对象的所有活动。从某种程度上说，大数据只是全体研究样本的一个方便样本，不是一个随机抽样样本。比如，如果研究对象是全体中国城市居民，那么互联网用户只是中国城市居民的一部分，因为没有一个平台能够记录所有中国城市居民的行为。因此，以大数据为基础的实证研究论文在结论一般化方面会受到很大限制。

那么，如何推动大数据在实证社会科学领域的应用呢？最重要的还是要推动数据的公开和分享。首先，应逐步推动不涉及国家安全的大数据在脱敏后开放给社会公众使用。政府部门可以通过与高校和科研机构的合作，来更好地规划、处理和开发大数据的应用，无论是学术层面还是公共服务层面，让政府大数据真正为社会服务。同时推动互联网公司与高校和科研机构在建立相互信任的基础上开展深度合作，探索一种有效的互联网公司与科研工作者的合作模式。[21] 其次，应积极建立社会科学大数据应用和交流的平台，尽管目前不少高校已经建立了大数据研究院，但是这些研究院刚开始往往与企业合作较多，而很少有专门针对社会科学的媒介和平台。三是需要全社会加快对于大数据相关分析工具的开发和普及，推动大数据分析技能在社会科学领域的推广和应用，不断改进使用大数据的实证研究的方法。但是，要实现上述领域的进步，需要政府、企业界和学术界共同努力和长期协作，并不是一朝一夕能够实现的。

清华大学社会学系的吕浩、张新望、余涵为本文做了一些资料整理工作，在此谨致谢意。

注释：

〔1〕我国的第一次人口普查始于 1953 年，第二次在 1964 年，后来因为文化大革命中断。除了人口普查，国家统计局还分别于 1987 年、1995 年、2005 年、2015 年进行了全国 1% 抽样调查。

〔2〕水延凯主编《中国社会调查简史》，北京：中国人民大学出版社，2017 年，第 350 – 355 页；刘云：《我国社会调查研究历史的回顾》，《新疆大学学报（哲学社会科学版）》1994 年第 4 期。

〔3〕水延凯主编《中国社会调查简史》，北京：中国人民大学出版社，2017 年，第 356 – 361 页。

〔4〕有统计表明，1992 年以后，随着调查技术、分析手段的进步，以及社会研究方法的成熟，越来越多的社会学者用高级统计分析方法来进行社会科学问题的研究，而 1992 年之前则基本上是以描述分析的简单量化研究为主，参见水延凯主编《中国社会调查简史》，北京：中国人民大学出版社，2017 年，第 364 页。

〔5〕该表格只列举了根据公开参考资料和笔者多年实证社会科学研究所接触和熟悉的一些数据来源。囿于笔者的知识和接触面所限,该表并不能包括改革开放以来所有中国综合性社会调查的数据。

〔6〕IBM 公司概括了大数据的 5V 特征,即数量(Volume)大、类型(Variety)多、速度(Velocity)快、准确性(Veracity)强、价值(Value)大。

〔7〕2009 年 Lazer 等人在《科学》杂志上发表的《计算社会科学》,标志着计算社会科学的诞生。Lazer, D, Pentland, A., Adamic L. A., et al., "Computational Social Science", *Science*, 2009, 323(5915), pp. 721 – 723;刘涛雄、尹德才:《大数据时代与社会科学研究范式变革》,《理论探索》2017 年第 6 期。

〔8〕数据来源:第 1 财经,http://www.yicai.com/news/5390789.html?xueqiu_status_id=99157680,2018 年 3 月 16 日登录。

〔9〕〔12〕〔18〕〔英〕维克托·迈尔 – 舍恩伯格〔英〕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,2013 年。

〔10〕这里的全体研究对象也是相对的。因为在实际的数据储存或提取过程中,受制于成本或者技术限制,获得全体研究对象的信息是非常困难的。比如研究婚恋行为的社会科学研究者,即使获得了一个大型婚恋网站的所有注册用户的网上活动资料,也很难获取一个大范围地域内所有经历过婚恋行为的人的行为数据,因为这些注册用户只是被研究总体对象的一部分。

〔11〕1960 年代末,美国斯坦福大学的一个政治学博士生 Norman Nie 和两个计算机系的博士生 Dale Bent 和 Tex´Hull 合作开发了一个专为社会科学统计分析使用的计算机软件 SPSS(Statistical Package for the Social Sciences),该软件界面友好,操作简单,为社会调查之后的数据清理和统计分析提供了方便,很大程度上推动了社会科学实证研究的发展。Norman Nie 因为对政治科学量化研究的贡献和对该软件的开发推广而获得了美国民意研究学会颁发的终生成就奖,并当选为美国艺术和科学院院士。

〔13〕我们对传统数据和大数据的划分也不是绝对的,我们这里的大数据是指随着互联网技术的发展出现的相对创新的数据收集手段。比如经济学者很早就开始利用上市公司全部股票交易数据来进行研究了,还有一些政治学者使用了瑞典所有政府登记的选民的数据,这些数据的获得也相对容易,因此在本研究统计过程中把这些类型的数据也归为传统数据。

〔14〕这里的实证研究论文是指使用大规模数据样本(含大数据)为理论基础的论文。经济学的一些论文只有基于理论模型和数学模型的推理,但没有用数据来验证或者计算这些模型的结果,这些没有算在这里的实证研究论文里面。在统计文章总数时可能包括了一些学术会议的综述,但是这部分文章在总体文章数量中占比很少,因此对我们计算实证研究论文比例不会产生太大影响。

〔15〕绝大部分论文对于数据处理和分析的计算机统计软件没有给出说明,因此笔者无法知悉和统计这些实证研究论文所使用的分析工具。但是根据笔者的经验判断,大部分这些论文所使用的大数据文件的大小都还在现有成熟计算和统计软件如 R、SPSS、Stata、SAS 能够处理的计算能力范围之内。

〔16〕由于篇幅所限,我们没有在表 4 中列出对美国三大社会科学期刊 18 篇使用大数据的实证研究论文的分析。但是对于美国三大期刊上使用大数据的实证研究论文的分析并没有使我们改变这个结论。

〔17〕不过,发表在国内这些期刊上的一些使用大数据的实证研究论文明确表明是受到了美国相关研究论文的启示,比如表 4 中发表在 2006 年《经济研究》上的论文就受到美国一篇 2000 年就发表的使用电子商务交易网站数据的启发。

〔19〕King, Gary, "Preface: Big Data is Not About the Data!", in *Computational Social Science: Discovery and Prediction*, edited by R. Michael Alvarez, Cambridge: Cambridge University Press, 2016.

〔20〕曾忠禄:《大数据分析:方向、方法与工具》,《情报理论与实践》2017 年第 1 期。

〔21〕笔者曾经参加过阿里巴巴研究院与研究者商谈合作的会议,但是向这些大企业获取数据的程序非常繁琐,这些大公司也对研究者非常谨慎。

〔责任编辑:刘姝媛〕