

大数据

在政府统计中的
探索与应用

马建堂 主编

 中国统计出版社
China Statistics Press

014036502

C829.2
11

大数据

在政府统计中的 探索与应用

马建堂 主编

图书馆

 中国统计出版社
China Statistics Press



北航

C1723594

C829.2

11

图书在版编目(CIP)数据

大数据在政府统计中的探索与应用 / 马建堂主编.

—— 北京 : 中国统计出版社, 2013. 10

ISBN 978-7-5037-6997-9

I. ①大… II. ①马… III. ①数据处理—应用—国家
行政机关—统计—研究—中国 IV. ①C829.2

中国版本图书馆 CIP 数据核字(2013)第 238053 号

大数据在政府统计中的探索与应用

作 者/马建堂

责任编辑/张 赏

装帧设计/张 冰

出版发行/中国统计出版社

通信地址/北京市丰台区西三环南路甲 6 号 邮政编码/100073

电 话/邮购(010)63376909 书店(010)68783171

网 址/<http://csp.stats.gov.cn/>

印 刷/河北天普润印刷厂

经 销/新华书店

开 本/710×1000 mm 1/16

字 数/140 千字

印 张/13.5

版 别/2013 年 10 月第 1 版

版 次/2013 年 10 月第 1 次印刷

定 价/39.00 元

版权所有。未经许可,本书的任何部分不得以任何方式在世界任何地区
以任何文字翻印、拷贝、仿制或转载。

中国统计版图书,如有印装错误,本社发行部负责调换。

《大数据在政府统计中的探索与应用》

编委会

主 编 马建堂

副 主 编 鲜祖德

编写人员 潘 璠 刘富江 许剑毅

杨京英 许亦频 姜 澍 王 萍 任全忠

何 强 丛雅静 董 倩

前言：迎接大数据挑战

随着科学技术的进步和生产力水平的提升,在世界范围内,一个大规模生产、分享和应用数据的时代——大数据时代正在到来。积极主动制定大数据策略,是每个政府、企业及其他社会组织谋求长远发展的必然选择。一些以美国为首的发达国家,都已从政府层面将大数据发展提升到战略高度。而各个国家、地区的各级政府统计机构,作为统计数据生产、管理和发布的政府职能部门,面对大数据带来的影响和挑战,也应该认真研究,冷静分析,积极应对,制定出切实可行的在政府统计中应用大数据的策略和措施,推动统计生产力的发展和统计生产方式的变革。

从生产力发展与变革的角度考察,人类社会的发展阶段可以划分为农业社会、工业社会和信息社会。在农业社会和工业社会,物质和能源是主要资源,社会生产活动主要以大规模的物质产品生产为主。而在信息社会,信息成为重要的资源,以开发和利用信息资源为目的的信息经济活动迅速扩大,逐渐成为国民经济活动的重要内容。

1946年,由美国军方定制的世界第一台电子计算机——电子数字积分计算机在美国宾夕法尼亚大学问世,揭开了以电子信息业高速发展为标志的第三次科技革命的序幕。1975年,比尔·盖茨与保罗·艾伦创立了微软公司。1976年,史蒂夫·乔布斯、斯蒂夫·沃兹尼亚克

和罗·韦恩成立了以创新闻名的苹果公司。1995年,美国麻省理工学院教授及媒体实验室创办人尼葛洛庞帝所著的《数字化生存》一书问世,该书被认为是20世纪信息技术及理念发展的圣经。它通过使用全球大量详实的数据资料分析表明,数字化、网络化、信息化使人的生存方式发生了巨大的变化,并由此带来一种全新的生存方式。美国《连线》杂志的创始主编凯文·凯利在其1994年出版的《失控》一书中准确预言了“互联网未来5000天”的发展,预言云计算、物联网、网络社区将成为未来科技发展方向。

进入21世纪以来,信息化对经济社会发展的影响愈加深远,“Big Data(大数据)”在行政管理、生产经营、商务活动等众多领域源源不断地产生、积累、变化和发展。从2009年开始,“大数据”逐渐成为互联网信息技术行业的流行词汇。大数据的战略意义不仅在于掌握庞大的数据信息,更重要的是体现在对海量的结构化和非结构化数据进行专业化处理,并快速获得有价值信息的能力方面。因此,IBM、EMC^①、Google、Microsoft等IT巨头都竞相投入到大数据的软硬件技术整合、大数据信息处理技术的研究开发之中,力求在新一轮的竞争中抢得战略先机与技术制高点。主要发达国家和国际组织也相继开展了大数据战略研究,以适应21世纪信息化时代发展带来的全新挑战。

近三年来,对大数据的认识在全球范围内得到前所未有的提升。2010年12月,美国总统科学技术顾问委员会在给奥巴马总统和国会的报告《规划数字化的未来》中,将大数据列为现有科技领域面临的挑战

^① EMC公司,1979年成立于美国马萨诸塞州霍普金市,1989年开始进入企业数据存储市场,提供信息存储及管理产品、服务和解决方案。截至2011年,EMC在中国的北京、上海、广州等地设立了16家分公司。

之首,并要求联邦政府的每一个机构和部门,都制定一个应对大数据的战略。2011年5月,世界著名的咨询机构麦肯锡公司发布的报告《大数据:创新、竞争和生产力的下一个前沿》指出,数据已经成为可以与物质资产和人力资本相提并论的重要生产要素,大数据可以发挥重要的经济作用,不但有利于私人商业活动,也有利于国民经济。2012年1月,世界经济论坛发布报告认为,大数据已经成为一种类似货币或黄金的资产,并将大数据处理技术誉为“改变世界的第一位科学技术”,必将推动世界经济的发展。2012年3月29日,美国奥巴马政府颁布了《大数据的研究和发展计划》。2012年5月,联合国“全球脉动计划”发布了题为《大数据促发展:挑战与机遇》的报告,指出大数据像纳米技术和量子计算一样带来了根本性的变革,将会塑造一个全新的二十一世纪。我国《“十二五”国家战略性新兴产业发展规划》明确提出,支持海量数据存储、处理技术的研发与产业化。

作为国家统计数据信息的官方生产和管理部门,统计部门不断加强统计信息化建设,对保证统计工作的顺利开展起到了至关重要的作用。国家统计局于2011年发布的《“十二五”时期统计发展和改革规划纲要》提出,建立现代统计体系,基本实现统计现代化是“十二五”时期统计发展改革的总体目标。近几年来,统计部门深入推进改革创新、规范统一、公开透明,积极应用现代信息技术。一是实施企业联网直报等统计四大工程。80万家“三上”企业、房地产开发经营企业和重点服务业企业,通过互联网直接向国家统计局报送数据。二是变革数据采集方式。消费价格调查由调查员手持电子终端现场采集数据并通过网络直报国家统计局,由70个大中城市房屋行政管理部门直接将房屋交易网签数据通过网络直接报国家统计局。三是全力推进统计信息化建设。建成以国家统计局为核心,连接各省级统计局和调查总队的国家

统计信息主干网络系统,已实现统计数据录入、汇总和存储的电子化、数据库化,基本实现统计数据上报的网络化。四是加大部门行政记录的应用。积极利用工商、民政、税务、质检、编制等部门的电子化单位登记资料,建立并更新基本单位名录库。五是加大空间信息技术应用力度。积极利用卫星遥感影像资料和相关地理数据,探索建设基本单位统计空间信息系统和人口地理信息系统,并利用多源遥感技术测量主要粮食作物产量,探索利用遥感和全球定位系统开展投资项目调查。政府统计通过应用这些信息技术努力变革统计工作模式,改进统计工作流程,提高统计能力、数据质量和政府统计公信力,既极大地推动了统计生产力的发展,也为进一步应用包括大数据在内的各种现代信息技术创造和奠定了更为适宜的工作环境和技术基础。

与云计算、物联网等新兴信息技术一样,大数据技术也是现代信息技术的重要组成部分。有研究指出,大数据是继互联网第二代(Web2.0)和云计算之后信息技术领域的又一次革命性浪潮。发达国家已经充分认识到大数据时代的发展趋势,并逐步在统计工作中应用大数据技术,而且取得了一定成效。但总体上看,各国政府统计应用大数据都还处于起步阶段。谁能抓住机遇,取得突破,就能够适应时代的发展,把政府统计工作提升到一个新的发展水平。所以,积极掌握大数据技术,提高其在政府统计工作中应用的深度和广度,既是统计部门深入贯彻落实科学发展观、建设服务型统计的迫切需要和必然选择,也是中国政府统计适应时代潮流、使统计工作走在世界前列的一个战略机遇和客观要求。

大数据就像春天的萌芽,即将破土而出;大数据就像东方的旭日,即将冉冉升起;大数据就像大海的浪潮,即将汹涌澎湃。深入研究大数据技术的特点以及在政府统计工作中的应用模式,是一项具有重要战

略意义的课题，既有重要的理论意义，也有迫切的现实意义。本研究通过对大数据理论研究和实践应用的观察与分析，探讨大数据在政府统计中可能应用的领域，明确大数据在政府统计中的应用方向，为统计改革和创新提供思路和参考。

马建堂

目 录

前言:迎接大数据挑战

第 1 章 大数据现象的观察与分析	1
第一节 大数据现象的源起与发展	1
一、20 世纪 80 年代至 90 年代中期:认知大数据的萌芽阶段	1
二、20 世纪 90 年代中期到 21 世纪前 10 年:大数据受到广泛 关注	2
三、2010 年至今:大数据战略应用被提上日程并迅速发展	3
第二节 对大数据现象的描述及观点	6
一、信息技术企业、咨询机构的观点	6
二、统计与数据专家、学者的观点	8
第三节 与大数据相关的几个问题	9
一、推动大数据发展的主体	9
二、大数据现象形成的客观条件	10
三、大数据数据量大的涵义	11
四、大数据与传统数据的区别	12
五、大数据的过滤与价值提取	14
第四节 大数据的特征和意义	15

一、大数据的特征	15
二、大数据的意义	17
第五节 大数据隐私与安全	20
一、大数据面临的几种风险	20
二、对大数据信息安全管理	23
 第2章 大数据在国内外统计中的应用实践	30
第一节 大数据在价格统计中的应用	30
一、美国的每日网上价格指数	31
二、淘宝网价格指数	31
三、基于 Twitter 数据对价格的调查	38
四、利用超市、商场和卖场的商品结算信息实现 CPI 采价	38
五、利用谷歌搜索数据计算 CPI	43
第二节 大数据在就业统计中的应用	45
一、联合国基于博客、论坛和新闻等媒体信息进行失业 调查	45
二、澳大利亚的在线职业空缺信息应用	46
三、谷歌利用搜索解析平台预测失业	46
第三节 大数据在经济活动分析和预测中的应用	48
一、基于百度搜索数据的消费者信心指数编制	48
二、利用谷歌搜索解析平台预测汽车销售市场走向	50
三、利用谷歌搜索解析平台预测 GDP	53
第四节 大数据在其他专业统计中的应用	57
一、大数据在人口统计中的应用	57
二、大数据在交通运输统计中的应用	58

三、大数据在公共卫生统计中的应用	60
四、大数据在商业零售服务统计中的应用	63
五、大数据在证券行业预测中的应用	65
六、大数据在网络社交统计中的应用	69
七、大数据在气象统计中的应用	69
第3章 大数据环境下政府统计的发展及新趋势	74
第一节 统计学和政府统计的发展历史概述	75
一、统计学的发展与演进	75
二、政府统计的建立与发展	77
三、政府统计的主要内容、数据来源和统计标准	79
四、政府统计的主要特征	80
第二节 中国政府统计的现状、发展与改革	80
一、中国政府统计的建立与发展	80
二、中国政府统计工作的现状	83
三、中国政府统计近期实施的重大改革	86
第三节 大数据背景下政府统计发展新趋势	91
一、大数据将促进统计科学的发展	91
二、大数据将促进统计法律的健全	92
三、大数据将促进统计标准的规范	93
四、大数据将促进数据来源的丰富	95
五、大数据将促进管理职能的转变	100
第4章 大数据背景下政府统计生产方式的变革	103
第一节 制度设计阶段	104

一、统计标准	105
二、统计内容	105
三、统计指标	106
四、统计范围	106
五、统计方法	107
第二节 数据采集阶段	108
一、搜索方式	109
二、购买方式	109
三、合作方式	110
第三节 数据处理阶段	111
一、直接组织数据处理	111
二、从简单汇总向数据挖掘方向转变	112
三、加工整理非结构化数据	112
第四节 数据存储阶段	113
一、建立国家统计云	113
二、虚拟化的云存储	114
三、数据备份与容灾系统	114
四、加工整理过程更加智能化	114
第五节 数据质量评估阶段	115
一、准确性	115
二、及时性	115
三、可比性	116
四、一致性	116
五、适用性	116
六、可获得性	117

七、经济性	117
第六节 数据发布与传播阶段	118
一、发布时间更加及时	118
二、发布媒介更加多样	118
三、发布内容更加丰富	119
四、发布形式更加灵活	119
第七节 数据分析阶段	120
一、实时分析	120
二、关联分析	121
三、可视化分析	123
第5章 大数据在部分专业统计中的应用探索	125
第一节 基本单位统计	127
一、数据来源和统计方法	127
二、可行性分析	127
第二节 价格统计	128
一、CPI 统计	128
二、PPI 统计	131
三、房价统计	134
第三节 工业统计	135
第四节 劳动力市场监测	138
第五节 人口普查	143
第六节 流动人口统计	145
第七节 农业统计	147
第八节 批发零售业统计	150

第九节 交通运输统计	151
第十节 收入统计	153
第十一节 工资统计	154
第十二节 主要产量指标预测	156
一、GDP 预测	156
二、煤炭产量预测	158
第 6 章 政府统计应用大数据的技术环境	161
第一节 大数据技术概述	162
第二节 大数据的采集技术	164
一、方便可靠的终端采集	164
二、智能分流的语音采集	164
三、穿透灵活的电子数据源采集	165
第三节 大数据的处理技术	165
一、技术路线分析	166
二、计算架构设计	172
第四节 大数据的存储管理	174
一、分布式云资源池	174
二、云资源池主机	174
三、云资源池存储	175
第五节 大数据的信息安全要求	176
一、终端的安全应用	178
二、配套的等级保护措施	179
第 7 章 应对大数据的几点对策	180

第一节	积极提高认识,树立大数据发展的新思维	180
第二节	完善法律法规,加强大数据应用法制保障	181
第三节	推动统计改革,重塑政府统计的生产流程	182
第四节	加强技术攻关,开发搜集工具和关联软件	184
第五节	强化培训交流,大力培养新型复合型人才	184
第六节	制定发展战略,推动大数据应用持续深入	185
参考文献		187
后 记		192

图 目

图 1:每日网上价格指数(2008 年 7 月为基期)与居民消费价格指数 对比	32
图 2:基于印度尼西亚 Twitter 用户交谈数据对价格的分析	38
图 3:上海钢联中国大宗商品价格指数与 PPI、CPI 走势对比	40
图 4:上海钢联中国大宗商品价格指数与工业增加值、预警指数 对比	41
图 5:泛亚稀有金属价格指数走势	42
图 6:基于谷歌搜索数据对 CPI 的预测	44
图 7:Google 搜索解析平台	47
图 8:2004 年 1 月—2009 年 5 月德国失业率预测	48
图 9:百度消费者信心指数与网络消费者信心指数对比	49
图 10:基于网络搜索数据构建的通货膨胀指数与 CPI 对比	50
图 11:大企业消费量景气指数与宏观经济先行景气指数对比	50
图 12:2006 年 1 月—2010 年 7 月智利月度汽车销售增长率移动 预测	51
图 13:2004 年第 1 季度—2011 年第 1 季度 OECD 的 GDP 季度增 长率	54
图 14:巴西登革热发病走势对比	61
图 15:Twitter 情绪指数与道琼斯工业指数对比	66
图 16:上证指数年收益率(实际值)与网络搜索指数(拟合值)的	