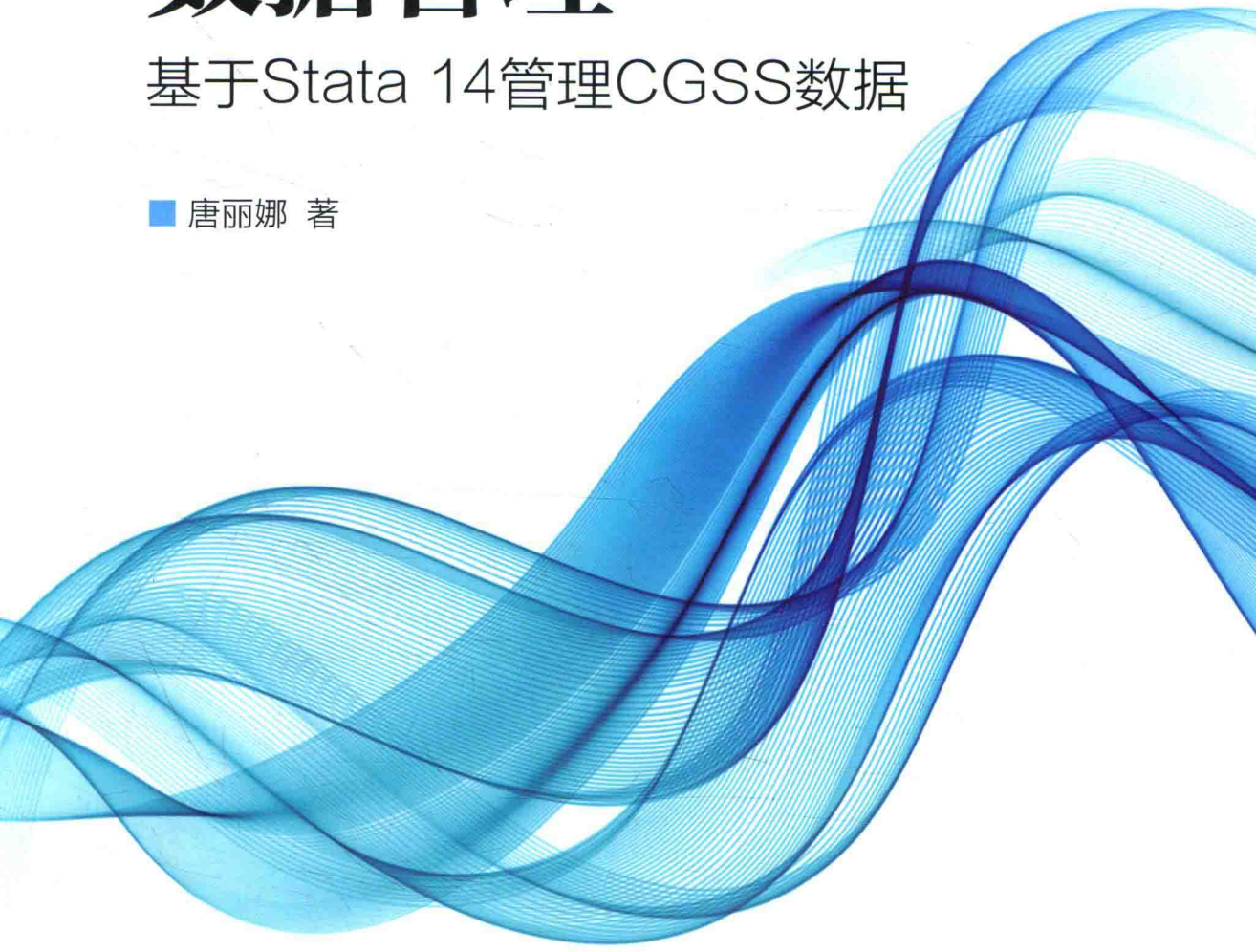
SOCIAL SURVEY DATA CURATION

精装版

社会调查数据管理

基于Stata 14管理CGSS数据

唐丽娜 著



中国工信出版集团

人民邮电出版社
POSTS & TELECOM PRESS

SOCIAL SURVEY
DATA CURATION

社会调查数据管理

基于Stata 14管理CGSS数据

唐丽娜 著

人民邮电出版社
北京

图书在版编目（CIP）数据

社会调查数据管理 : 基于Stata 14管理CGSS数据 / 唐丽娜著. -- 北京 : 人民邮电出版社, 2016.6
ISBN 978-7-115-42174-6

Ⅰ. ①社… Ⅱ. ①唐… Ⅲ. ①社会调查－数据管理－应用软件 Ⅳ. ①C915-39

中国版本图书馆CIP数据核字(2016)第089942号

内 容 提 要

这是一本关于社会调查数据管理的实务操作手册，以国内第一个、综合性、长期性的调查数据——中国综合社会调查（CGSS）数据的管理为例，基于最新版的 Stata 14 软件，全面讲解了一个社会调查数据管理的完整周期，重点演示了社会调查数据管理工作中的重点和难点。

本书适合社会调查者、在校大学生、学者、研究者及其他和数据管理相关的从业者阅读参考。为方便读者学习，书中所有示例数据及命令都可以从人民邮电出版社异步社区网站下载。

◆ 著　　唐丽娜
责任编辑　王峰松
责任印制　焦志炜

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京圣夫亚美印刷有限公司印刷

◆ 开本：800×1000　1/16
印张：21.25
字数：580 千字　　2016 年 6 月第 1 版
印数：1－2 000 册　　2016 年 6 月北京第 1 次印刷

定价：79.00 元

读者服务热线：(010)81055410　印装质量热线：(010)81055316
反盗版热线：(010)81055315

序 一

数据是这个时代的脉搏，和石油、矿藏等一样也是一个国家重要的资源。在学术领域，数据密集型驱动的学术研究范式日益盛行，数据已经发展成为一种重要的研究基础。在社会科学研究中，实证研究的前提是质量可靠的数据。在整个数据生命周期中，会有不同主体的参与，比如，研究者通常既是数据的生产者也是使用者，政府、学校和科研机构是产生数据的经费支持来源，也是推动数据开放和数据共享的重要推手。不同主体在数据生命周期中扮演着不同的角色，但是，有一点是所有参与主体必须关注的焦点——数据管理，这也是本书作者的研究重点。

在社会科学领域，我国的实证研究起步较晚，“数据素养”有待提高，对数据的关注点几乎都聚焦在数据分析上。然而，数据采集和数据分析之间还有一座重要的桥梁，那就是数据管理。有过数据使用经历的人都知道，研究者或机构采集到的原始数据很难直接用于数据分析，需要对其进行一定的数据清理，如进行选取样本、选择变量、重新编码、插补缺失值、逻辑检验、基于原始变量创建研究所需的新变量、数据格式转换、数据形状转置等大量繁琐、复杂、耗时的工作，之后才能开始数据分析。对研究人员而言，这些和数据管理相关的工作往往会占据一项学术研究的一半甚至更多的时间。考察目前国内社会科学界几大旗舰项目发布的数据不难发现，我国的数据管理工作亟待改进和提高，几乎每个旗舰项目发布的数据在基本要素上都各有“特色”，即使在同一个项目数据的内部对同一个要素的操作都不统一、不规范，如数据中的变量名，有的用题号做变量名，有的用该变量对应的题目的英文缩写做变量名，还有的把题号和英文缩写合在一起做变量名，确立合适的社会科学数据管理标准迫在眉睫。

本书正是应这样的数据需求而生，作者唐丽娜博士长期以来一直从事社会科学数据的采集、清理、管理、分析和挖掘工作，积累了丰厚的经验，而且全程参与中国国家调查数据库的建设和维护，对数据生命周期的各个环节都有着自己独到的见解。这本书是她对自己多年和数据打交道的一个阶段性总结，也是她对国内社会调查数据管理的大胆探索。

国家是数据最大的生产者和使用者，数据管理更是一项国家战略，在数据开放和数据共享的大趋势下，建立规范、科学的数据管理变得愈发重要，而共享的前提是持续的、规范化的数据管理，否则结果只能是大量数据的无序集合而已。数据驱动型研究和数据导向型经济推动着社会各界对数据管理专业技术和专业人才的需要，我国在这方面仍处于起步、探索阶段，对涉及其中的主体和主体职责有待进一步的明确，社会调查数据的微观分享需要国家在宏观层面的政策支持和法律保护。

目前，国际上对此已经做出了很多有益探索并提供了大量可供借鉴的案例和经验，如英国早在 2000 年就通过了信息自由法，而且在 2004 年成立了全球第一个专门从事数据管理研究和探索的机构——数据管理中心（Data Curation Center），为英国的数据管理提供了很多成功的案例、实用的管理工具及必要的技术培训。美国的 NSF、NIH 不仅强制要求接受资助的研究人员提交项目数据，而且提供专项基金用于研究数据管理。我国目前的数据封闭独享意识仍然存在，数据交换和共享尚未得到广泛认可，这极大地抑制了数据的学术效用和社会效益。希望这本书的出版，能够为国内社会科学领域中的数据管理、数据开放和共享提供想象的空间和讨论的基础。

袁卫

中国人民大学统计系教授，中国国家调查数据库项目负责人

序 二

1972 年美国芝加哥大学的国家民意调查中心（National Opinion Research Center，NORC）启动了综合社会调查项目（General Social Survey，GSS），旨在收集能够反映美国社会变迁及社会态度的数据，为政策制定者和学者提供一套清晰无偏的数据。迄今为止，该项目仍然是美国国家自然科学基金支持过的最大的社会调查项目，在美国对 GSS 数据的使用率仅低于美国人口普查数据，GSS 数据的学术效应和社会效益已经形成气候，值得我们学习。

有感于中国改革开放以来，学术界想用数据来研究中国历史上的这一重大变迁，苦于没有数据可用，中国人民大学社会学系和中国香港科技大学多位志同道合的社会学家联合发起做中国自己的综合社会调查，定名为中国综合社会调查（Chinese General Social Survey，CGSS）。现在，CGSS 是国内社会科学领域持续时间最长的一项社会调查，到目前为止，成功访问过来自中国大陆的 102730 名居民，收集到的有效居民数据是 102730 条，社区数据为 2031 条，积累了 10 年的数据集，为国内外的学者提供了宝贵的、无以复制的研究中国社会变迁和居民社会态度、社会行为的数据。

社会调查是一项基础研究，不同于应用研究，如果做不到专业严谨的研究和应用，调查的意义和效用则大大降低。社会调查数据的广泛应用离不开数据开放和数据共享的意识氛围，也需要全面系统的数据管理规范和标准。在我国学术界，社会调查特别是大型、随机抽样调查的起步相对较晚，受实证主义研究范式的影响，大部分学者和研究人员在做社会调查时，更关注用数据做实证分析，对数据本身的管理不够重视。纵观最近几年国内知名度较高的几个社会调查品牌项目会发现，每个调查项目发布的数据都或多或少存在这样、那样的问题，这些问题都起因于对数据的管理不当或不够，导致用户无法使用某些数据，或者能使用但用起来很费劲。在社会调查领域，长期以来数据的采集、管理、分析和挖掘都由同一机构或研究团队完成，现在这些工作也面临着专业化、精细化的发展，数据的采集、管理和分析可以分别由三支专业团队执行，方能确保数据的质量和长期效益。国内急需建立起一套对社会调查数据进行管理的标准、规范、流程，为数据的开放和共享提供数据标准和技术支持。

随着 CGSS 调查的推进和数据的积累，CGSS 数据也面临着数据管理的困境和数据服务的难题，特别是当数据的使用范围和目标对象从当初的有限范围扩展到全社会范围内的数据使用者时，数据服务方面的各种问题也日益凸显。自 2015 年开始，CGSS 项目组专门组织研究人员探索 CGSS 数据的管理规范，致力于提高 CGSS 数据服务。这本书只是 CGSS 数据管理起步阶段的一个小总结，希望国内会有更多、更深入的和社会数

据管理有关的文章或书籍问世，为国内社会调查数据的推广和走向国际化提供更有价值的建议。

李路路
中国人民大学社会学系教授，长江学者，CGSS 项目负责人

序　三

数据管理在数据的生命周期中是一个至关重要的方面，但是，在我国的社会科学研究中一直被忽视。长期以来，学界把实证量化研究方法的重点放在各种统计方法及模型上，出版了大量的书籍，也举办了大批的培训，但接下来遇到的问题是，当研究者们掌握了各种统计方法和技术后，发现自己学会的是屠龙术，但是无龙可屠，空有统计方法和技术，但缺乏可用来分析的数据。所以，近几年来，学术性社会调查兴起，国内各高校启动了多项覆盖经济与社会各个领域的、具有全国代表性的截面或追踪调查项目，产出了一批高质量的社会调查微观数据。但是，当研究者们终于获得了宝贵的数据时，却往往发现无从下手。这就如同菜谱上讲的是如何用洗净、切好、进行过预处理、符合要求的食材来做菜，但拿到手却是近似最初状态的蔬果肉蛋，大多数人就有点无所适从了。实际上，从实地阶段的数据采集到数据的分析与开发之间还有一个重要的中间环节，这就是数据管理。

数据管理的工作贯穿数据生命周期的全过程。进行实地数据采集前，就需要制定详尽的数据管理方案，调查问卷的设计与调查的执行需要与这个方案配合；完成实地调查后需要进行数据的录入（电子化）、编码、清洗、插补、转换、派生、建档等；当数据集及相关文档准备就绪后，还需要对其进行存储、发布、共享，并提供用户支持服务；而数据集本身也可做一个单元与其他的各种层次和类型的数据进行匹配、整合，如可以把 CGSS2010 的数据和 2010 年美国的 GSS 数据合并在一起，做国际比较分析，成为综合性研究资源的一个部分持续起作用；这些都是数据管理的内容。好的数据管理对于提高数据的投入产出率，延长数据的生命周期具有至关重要的作用。但是长期以来，数据管理一直是我国社会科学实证量化研究中的薄弱环节。正是由于对于调查数据管理的相对薄弱，才使得项目组内部对数据的分析与开发受到影响；而当数据开放之后，外部用户对于数据的使用则更是有诸多障碍。正是由于我国的社会调查数据在用户友好上做得非常欠缺，加上数据的开放共享不足，才造成数据的利用率较低，生命周期短暂。

数据管理的重要性一直被国际科学界所强调，美国国立卫生研究院（NIH）和美国国家科学基金（NSF）这两家美国最大的科学研究基金都把项目申请中的数据管理方案作为对项目申请书进行评价的重要方面。随着我国社会科学领域由数据驱动的实证量化研究的发展，对于微观调查数据的管理的重要性也逐渐被认识到，本书正是这一发展趋势的反映和重要标志。本书的作者唐丽娜博士长期以来作为主要成员参与了我国历时最长的全国性连续学术社会调查项目——“中国综合社会调查（CGSS）”的工作，同时也在我国第一个社会调查数据资源库——“中国国家调查数据库（CNSDA）”的建设和运行中发挥了重要作用。她对社会

调查数据的产出到利用的全过程有着深刻的了解与把握，对于社会调查数据管理工作的各个方面有着丰富的经验。她的这部著作重在实用性与工具性，对于社会科学领域从事与数据相关研究工作的各类人员有着切实的帮助。这本书是作者长期以来在社会科学基础数据工作领域里无私奉献的一个阶段性总结，也是对我国社会调查研究这些年来所取得进展的一个阶段性汇报。

王卫东
中国人民大学中国调查与数据中心
2016 年 2 月

作者序

机缘巧合，自2005年起就开始和调查及数据卯上了。坦白说，在这些方面，我从来没有接受过所谓的“科班”教育。但是，我一直在学习这方面的知识，也一直在实践这些知识。从一开始的“叶公好龙”，到后来的“爱不释手”，过程中充满了欢乐、恼恨、时不时的放弃、反复的质疑、失落、失望……基础由此得到了夯实，兴趣因此变得更加浓厚。

我不是“有意”要写这本书，于我而言，写一本关于调查和数据的书的念头由来已久，硬要说出个一二三来，那我认为：其一，数据作为一种生产力，值得现代社会的每个人都来了解和学习；其二，大众对调查和数据的认识存在种种误区，且这些误区既不利于大众，也不利于社会，更不利于国家；其三，目前国内相关书籍的编写方式过于“学术化”，或者说都更像是教科书，而不是科普书，可读性不足，不容易让人产生兴趣，还可能会导致感兴趣的人望而失趣。

那么，我写的这本书就能避免这些吗？我不敢说一定都能，但我想试试。我敢试试还是源于我的工作经验。理论上，我只有一年半的工作经验（截至写作之前，我在中国人民大学中国调查与数据中心以博士后的身份工作了一年半），但实际上这个中心自成立之日起，我就一直“浸染”其中。任何一个单位或公司在成立之初都会面临很多问题，中心也不例外。在五花八门的困难中，最大也最头疼的是人，特别是会管理数据的人。

鉴于种种考虑，最终决定招聘的标准是：品德过硬、对数据管理有兴趣、踏实肯学，对专业没有做任何限制。中心现在的员工在专业分布上也是“醉”了，如生物、文学、化学、法律、兽医、英语、国际关系、马克思主义、金融、国际贸易等。这样的专业分布好处是：所有员工进入中心时，起点几乎都是一样的，只要工作制度相对合理，那么就基本能够保证每个人在中心的成长和发展机会相对公平。这些专业没有一个和数据管理有关，要上岗，自然要对人家进行相关的培训，并经历一定的锻炼。正是在这些培训中，萌生了写一本关于数据管理的书的念头。我发现，数据管理说难也不难：文学等文科方面的员工都能学会基本的数据管理技能，还有一些优势，比如，在这一领域他们是白纸一张，恰恰更容易绘画，和那些已经被“乱画”过的相比，会学到更精准和更扎实的技能。

作为一个“非科班”出身的人，更能了解一个门外汉在学习时可能会遇到的问题，更能体会哪些知识和技能需要深入的理解和长久的锤炼。

作为一本关于社会调查数据管理的入门书，为使全书连贯可读，在写作的过程中，必须把通常比较复杂的问题进行某种程度的简单化处理，因此不可避免地要忽略某些问题和观

点，我对本书中所有可能的错误负责。本书的成书时间仓促，如果将来有机会，我会在别的书中对数据管理和数据分析做更多的讨论，望读者阅读拙作能如我初心。

唐丽娜

2015 年 11 月 2 日于中国人民大学明德国际楼

致　谢

本书源于 CGSS 项目，首先要感谢 CGSS 项目组的辛苦工作和无私支持！特别要感谢我的 3 位导师——袁卫教授、李路路教授（CGSS 项目 PI）和王卫东教授对我在学习上的指导和工作上的理解及帮助。他们不仅是我学习上的导师，更是我人生中的导师，遇到他们，是我的幸运，没有他们，就不会有本书的出版。

还有很多人从不同角度和层面为本书付出了辛勤的劳动，在此我对他们都表示诚挚的谢意：

感谢我的同事盖琴宝和刘斌帮助我绘制了数据合并的图示并编辑文字。

感谢我的同事葛欢、韩佳妤和孙立鑫对本书的认真编辑，他们纠正了书中的一些错误，并对本书的资料组织给出了中肯、宝贵的建议。

感谢忘年之交王天星老师在整个出版过程中的不吝赐教和鼎力相助。

他们的严谨工作和无私奉献为本书的质量保驾护航。

感谢我的闺蜜王昕，是她一直鼓励我、支持我、肯定我在数据管理和数据分析方面的认识，让我热情高涨、轻松愉快地写书。

我还要感谢我的先生，他是我的灵魂伴侣，没有他对我的理解、支持、包容和关爱，我无法在这么短的时间内成书。

最后，我要感谢我的父母和弟弟，我的一切都是他们赐予的，这本书是我给他们的献礼。我对生活的热爱、工作的执着，都源于我的家人。

目　录

第一部分　社会调查者的数据管理

第二部分　数据使用者的数据管理