《公共管理与政策研究方法论》
第9讲
2016年11月14日

二手数据分析

马亮博士 副教授
中国人民大学公共管理学院

#	课程模块	核心内容	教师	时间
1	方法论导论 科学哲学	介绍本课程的主要内容，并探讨研究方法论的关键概念。 对本体论和认识论等科学哲学概念进行解读。	马亮	9/12 9/14
2	方法论流派	介绍逻辑经验主义、行为主义、实证主义、后实证主义、后现代主义、建构主义等方法论流派。	马亮	9/19 9/21
3	研究问题、研究综述与研究评价	探讨如何提出一个有趣、重要和可行的研究问题。 探讨如何收集文献、述评文献，以及如何评价一项研究的质量和贡献。	马亮	9/26 9/28
4	变量的概念化	讨论概念的定义，以及怎样概念化。	钟玮	10/10 10/12
5	变量的测量	介绍变量的尺度、类型、测量方式，以及测量的信度与效度。	钟玮	10/17 10/19
6	理论与理论构建 因果关系论证	介绍什么是理论，以及如何建构理论。 比较相关关系和因果关系，探讨因果关系的论证方法。	马亮	10/24 10/26
7	实验设计	介绍实验室实验、准实验、自然实验、调查实验等实验设计。	马亮	10/31 11/2
8	调查设计	介绍问卷调查的设计策略和技巧。	马亮	11/7 11/9
9	二手数据分析	探讨二手数据的收集和分析策略。	马亮	11/14 11/16
10	质性研究设计	介绍民族志、扎根理论、叙事学、现象学、案例研究等质性研究方法。	马亮	11/21 11/23
11	混合研究设计	探讨如何将量化研究和质性研究结合起来使用。	马亮	11/28 11/30
12	研究计划书 论文写作与发表 研究伦理与学术规范	介绍研究计划书和开题报告的撰写技巧。 讨论如何写作、报告和发表论文。 介绍研究伦理注意事项，以及学术规范要求。	马亮	12/5 12/7
13	学生成果展示	请课程学生口头汇报研究设计或开题报告。	马亮 钟玮	12/12 12/14

多元的方法—管理学

Internal Validity: Time Frame of Studies[a]

Time Frame	1985–87	1995–97
Cross-sectional	77.40%	85.60%[+]
Longitudinal	22.60	14.40[−]

Research Strategy	1985–87	1995–97
Formal theory/literature review	22.90%	18.70%
Sample survey	6.90	3.60[b−]
Laboratory experiment	10.70	4.90[b−]
Experimental simulation	0.60	1.70
Field study		
Primary	38.00	40.90
Secondary	16.10	26.60[b+]
Field experiment	3.90	2.20
Judgment task	0.60	0.20
Computer simulation	0.30	1.20

资料来源：
Scandura & Williams (2000)

多元的方法——管理学

表 1　主要管理学期刊发表论文采用二手数据统计,2007—2011

年度	SMJ			AMJ			OS			JIBS		
	实证论文的总篇数	采用二手数据的论文篇数	比例	实证论文的总篇数	采用二手数据的论文篇数	比例	实证论文的总篇数	采用二手数据的论文篇数	比例	实证论文的总篇数	采用二手数据的论文篇数	比例
2008	70	53	0.757	54	35	0.648	41	27	0.659	68	43	0.633
2009	66	54	0.818	56	28	0.5	42	30	0.715	66	43	0.652
2010	63	56	0.889	62	37	0.597	56	36	0.643	65	47	0.724
2011	53	48	0.906	36	20	0.556	51	35	0.687	45*	31	0.688
加总	252	211	0.837	208	120	0.577	190	128	0.674	199	133	0.672

*2011 年第 5 期的 JIBS 乃案例研究专刊,故未纳入此统计之中。

对2008-2011年间涉及战略管理研究的四本顶尖学术期刊SMJ、AMJ、OS、JIBS进行分析，其所发表的实证论文中，二手数据的占比达到“半壁江山”。
资料来源：周长辉 (2012, p. 217)。

定量研究方法的分支及占比：
二手数据共计210篇，合计44. 8%。

Number (Percent) of Journal Articles by Type of Quantitative Research Method, 2001–2010 ($N = 469$)

Type of Research Method	*Number (Percent) of Journal Articles*
Survey	169 (36.0%)
Statistical analysis of primary data	54 (11.5%)
Statistical analysis of secondary data: survey	126 (26.9%)
Statistical analysis of secondary data: non-survey	84 (17.9%)
Both survey and statistical analysis of primary or secondary data	36 (7.7%)
Total	469 (100%)

Groeneveld, S., Tummers, L., Bronkhorst, B., Ashikali, T., & Van Thiel, S. (2015). Quantitative Methods in Public Administration: Their Use and Development Through Time. *International Public Management Journal*, 18 (1).

多元的方法——中国公共管理学

TABLE 7 *Research methods used by academic rank of first authors and by journals*

		Empirical analysis	Historical research	Descriptive research	Logical analysis	Action research	Other method	No formal research method
By year of publication (%)								
Mainland China	1998–2001	1.2	1.6	1.0	0.0	0.0	2.4	93.9
	2002–2005	3.5	1.3	0.2	0.7	0.0	2.2	92.2
	2006–2008	7.3	1.4	0.7	0.3	0.2	3.5	86.7
	Total	4.8	1.4	0.6	0.4	0.1	2.8	90.0
Taiwan	1998–2001	13.5	35.6	19.2	2.9	2.9	10.6	15.4
	2002–2005	25.8	28.3	20.0	0.0	5.0	15.0	5.8
	2006–2008	40.2	19.5	19.5	2.3	2.3	6.9	9.2
	Total	25.7	28.3	19.6	1.6	3.5	11.3	10.0

TABLE 8 *Sources of data*

		Primary data	Secondary data	No data
By year of publication (%)				
Mainland China	1998–2001	1.2	17.1	81.8
	2002–2005	2.4	16.3	81.3
	2006–2008	5.6	20.0	74.4
	Total	3.6	18.1	78.2
Taiwan	1998–2001	13.5	26.0	60.6
	2002–2005	22.5	39.2	38.3
	2006–2008	37.9	37.9	24.1
	Total	23.8	34.4	41.8

资料来源：
Wu, He, Sun (2011)

不同学科的二手数据

- 经济学：主要是二手数据，特别是各类统计年鉴
- 金融学：证券市场股价、上市公司年报、各类二次数据库
- 社会学：主要是一手调查数据，包括问卷和访谈
- 心理学：主要是一手数据，通过调查或实验获取
- 政治学：各类政策、决策、报告、履历
- 管理学：
  - 宏观研究：主要是一手企业调查和二手案例分析
  - 微观研究：基本上都是一手员工调查
- 公共管理学、公共政策学、公共财政学
  - 约20-40%的数据来源是二手数据（调查或非调查）。

二手数据分析的意涵

- 什么是二手数据？倒几手才是二手？三手、四手？
- 定义：数据的目的与来源
  - 一手或原始数据（primary/first hand/original data）
  - 二手数据（secondary data）
- 孰优孰劣？孰主孰次？相互补充？“非你莫属”？
- 数据驱动还是理论驱动？数据在先还是理论在先？
  - Cook：the theory-data match
- 一手与二手数据的边界日趋模糊。
  - 商业调查中的“搭车调查”、与政府部门的合作实验
  - 网络搜索、API数据抓取、大数据

一手数据VS二手数据

维度	一手数据	二手数据
谁搜集的？	研究者本人或其委托的个人或机构	他人或机构搜集
什么目的？	直接用于研究者本人的研究	为了其他研究或其他目的，而不是专为本研究
是否接触研究对象？	通常直接介入和接触研究对象	通常不介入和接触研究对象
谁拥有？	一般为研究者本人所有	通常可以通过公开或公共渠道获取

资料来源：根据周长辉 (2008/2012)整理。

为什么要用二手数据？

- “在我看来，中国可以说遍地都是数据金矿。我这里说的数据金矿，就是指二手数据。”
- “但二手数据确如金矿，只不过丰富而珍贵的二手数据大多是以‘矿石’的形式存在着，它等待着有心人去探索、识别和开发。研究者要像淘金者一样去‘淘’。虽说‘淘’金的过程并不容易，但终归比问卷调研更能做到自主可控。”

周长辉. (2008/2012). 二手数据在组织管理学研究中的使用. 陈晓萍, 徐淑英, & 樊景立. 组织与管理研究的实证方法 (第九章). 北京:北京大学出版社.

二手数据的优势是什么？

- 总体（population）或大样本（large-N）
  - 通常来说，二手数据的样本量较大，且多数可以提供跨年、跨季的纵贯数据，进而有利于构造面板数据。
- 客观性、可复制和可重复性
  - 信度（validity）与科学本质
  - “他律性”与学术伦理，可以减少学术造假。
- 多源数据和三角测量（triangulation）
  - 避免共同方法偏误（CMB/CSB）
  - 多角度认识事物，从而更精确地观测并发现规律。
- 数据采集成本低廉
  - 许多情况下甚至是免费的
  - 成本是相对而言，因为数据清洗和处理成本未必低。
- 非侵入性或无干涉的研究

二手数据的局限与劣势

- 数据的可靠性（信度）可能欠缺
  - “官出数字，数字出官。”层层上报的遗漏、误差与蒙骗。
    - 中国官方GDP数据的水分、空气污染等环境数据造假。
  - 不同地区和国家的定义和测量方式不同。
- 数据不“解渴”，测量的效度不高
  - 没有问到最需要的问题，没有使用成熟量表提问，或者无法匹配到具体的地区（如县）、组织乃至个人。
  - 理论构念无法得到最佳的衡量，或者操作方式不同。
    - 例如，犯罪率的低估或低报、对交通事故的认定。
- 数据分析与处理都很“费劲”
  - 数据编码、清洗、匹配、合并、管理等需要大量工作。
- 数据的开发与再开发程度
  - 数据被“用烂了”，有人“捷足先登”，数据的再开发和再利用程度有限，需要“绞尽脑汁”和“独辟蹊径”。

数据公开与复制研究

（Replication）(King, 1995)

Data Access and Research Transparency (DA-RT): A Joint Statement by Political Science Journal Editors

Transparency requires making visible both the empirical foundation and the logic of inquiry of research. We agree that by January 15, 2016 we will:

- Require authors to ensure that cited data are available at the time of publication through a trusted digital repository. Journals may specify which trusted digital repository shall be used (for example if they have their own dataverse). If cited data are restricted (e.g., classified, require confidentiality protections, were obtained under a non-disclosure agreement, or have inherent logistical constraints), authors must notify the editor at the time of submission. The editor shall have full discretion to follow their journal's policy on restricted data, including declining to review the manuscript or granting an exemption with or without conditions. The editor shall inform the author of that decision prior to review.
- Require authors to delineate clearly the analytic procedures upon which their published claims rely, and where possible to provide access to all relevant analytic materials. If such materials are not published with the article, they must be shared to the greatest extent possible through institutions with demonstrated capacity to provide long-term access.
- Maintain a consistent data citation policy to increase the credit that data creators and suppliers receive for their work. These policies include using data citation practices that identify a dataset's author(s), title, date, version, and a persistent identifier. In sum, we will require authors who base their claims on data created by others to reference and cite those data as an intellectual product of value.
- Ensure that journal style guides, codes of ethics, publication manuals, and other forms of guidance are updated and expanded to include improved data access and research transparency requirements.

二手量化数据的处理

- 数据来源
  - 中国各级政府的统计年鉴、年报、季报、月报
  - 国际组织等的跨国数据、区域数据
  - 其他国家的统计年鉴和调查数据
  - 商业咨询公司数据（Compustat、国泰安、零点）
  - 其他研究者分享的数据
- 量化数据的清洗与合并
  - 一个数据库的内部清洗与整理
    - “情人眼里出西施”
  - 多个数据库的合并管理
    - “变废为宝”、“化腐朽为神奇”

质性数据的二次分析
（secondary analysis of qualitative data）

- 数据来源
  - 文本：访谈记录、问卷开放题、实地手记、日记、年谱、档案等。
  - 影像：录音、图片/照片、录像/视频。
  - 共享：英国质性数据中心（Qualidata，1994）
- 质性数据的二次分析（质性➡质性）
  - 区别于文档分析：是否曾被其他研究者使用过？
  - 分析类型：原始数据的收集者是否参与？
  - 区别于质性数据的量化分析：元分析/系统综述？
- 质性数据的转换（质性➡量化）
  - 从文本、图片、视频等质性数据转化到量化数据
  - 编码（coding）至关重要
    - 内容分析（content analysis）或扎根理论（grounded theory）

几个实例

- 治理研究的挑战：多层模型(Heinrich & Lynn, 2001)。
- 公务员调查数据的二次利用(Fernandez, et al., 2015)。
- 公共政策/计划/项目绩效的数据(Moynihan, 2013)。

大数据技术

- 数据开放、数据共享与云计算
- 科学2.0、研究2.0与政府2.0
- 原始获取的大数据
  - 电子病历、网上投诉、政府采购、审判文书等。
- 二次开发的大数据
  - 谷歌流感、百度迁徙、淘宝消费、微博热度等。
- 实例与趋势
  - 经济学的应用(Einav & Levin, 2014)

值得讨论的问题

- 在条件允许的情况下，尽可能开展一手数据收集，特别是通过调查和观察，抢救式记录中国公共管理的历史进程。
  - 中国公务员价值观、态度、动机与行为调查
  - 中国政府部门决策、行为、绩效调查
- 如果有可用的数据，为什么不用？培养数据敏锐性和嗅觉，让数据找你，而不是你找数据！
  - “好记性不如烂笔头”，随时随地记录和整理数据。
- “万事万物是普遍联系的。”
  - 研习如何嫁接和联系多个数据库，如跨层分析或分层线性模型的使用。
- 实证公共管理研究：理论与数据的水乳交融

下一次课程研讨的论文清单

1. Brower, R. S., Abolafia, M. Y., & Carr, J. B. (2000). On improving qualitative methods in public administration research. *Administration & Society, 32*(4), 363-397.
2. Gerring, J. (2004). What is a case study and what is it good for? *American Political Science Review, 98*(02), 341-354.
3. Ospina, S. M., & Dodge, J. (2005). It's about time: Catching method up to meaning - the usefulness of narrative inquiry in public administration research. *Public Administration Review, 65*(2), 143-157.
4. Cappellaro, G. (2016). Ethnography in public management research: A systematic review and future directions. *International Public Management Journal*, 1-35.
5. Suddaby, R. (2006). From the editors: What grounded theory is not. *Academy of Management Journal, 49*(4), 633-642.

谢谢！

http://liangma.weebly.com
Email: liangma@ruc.edu.cn