社会
2015·2
CJS
第35卷

必要与如何:基于历史资料的量化数据库构建与分析

以大学生学籍卡片资料为中心的讨论

梁　晨　董　浩

摘　要:随着"大数据"时代的到来,依靠大规模系统历史资料构建量化数据库并进行定量分析成为一种新的、行之有效的研究方法。如何将这类历史资料进行合理有效的编码和数据库化,并通过实证分析更好地帮助我们了解社会经济发展的历史经验和对当下的启示,成为学界需要加强探索和讨论的关键技术课题。本文试图借助笔者多年来整理、分析近现代中国高校大学生学籍卡资料的经验,说明这种新方法论视角用于定量分析历史资料的重要价值与必要性,以及可能存在的诸多挑战和可供参考的应对办法。希望藉此引起社会科学与人文学科学者对这种研究方法的关注、讨论、尝试与合作。

关键词:量化数据库　定量分析方法　历史资料　学籍卡
DOI:10.15992/j.cnki.31-1123/c.2015.02.005

Construction and Analysis of Big Historical Micro-Level Data: A Brief Discussion with Examples of Data Gathered from University Student Registration Cards

LIANG Chen　DONG Hao

Abstract: Along with boosting public and professional interests in "big data", construction and analysis of large-scale micro-level data from voluminous historical sources become available and promising. Big historical micro-level data facilitate inter-disciplinary and longitudinal social scientific research, of which implications are far beyond historical but related to a better

* **作者1**:梁　晨　南京大学中华民国史研究中心(Author 1: LIANG Chen, Center for the Republic of China History Research, Nanjing University) E-mail: liangchen@nju.edu.cn;
作者2:董　浩　香港科技大学社会科学部(Author 2: DONG Hao, Division of Social Science, The Hong Kong University of Science and Technology)

** **本文是国家社会科学基金青年项目(10CZS023)和香港优配研究金项目(16400714)的成果之一**。[The research was supported by the Youth Project of the National Social Science Fund (10CZS023) and Hong Kong Research Grants Council General Research Fund Project (16400714).]

understanding of change and continuity in human behavior and society. While China has one of the world's best and largest collections of historical documents surviving to date, practice in construction and analysis of historical micro-level data remain limited. We therefore share our experience from an ongoing research project that uses more than 150 000 individual student registration cards from two Chinese elite universities to study the evolution of social inequality in higher education between 1950 and 2000. We hope to stimulate broader academic interest, discussion, exploration, and collaboration in research using big historical micro-level data for the betterment of social sciences and humanities.

Keywords: micro-level data, quantitative methodology, historical, college student registration

20 世纪 90 年代以来,构建和分析大规模量化微观数据是全球学界蓬勃发展的一种社会科学研究方法的新趋势(Ruggles, 2014)。与二战后以计量史学(cliometrics)或新经济史(new economic history)为代表、侧重数理模型和高级统计分析方法的历史定量研究,或是许多以当代社会统计抽样调查数据为基础的微观定量研究相比,这个新的方法论视角更关注数据本身,尝试通过以人口总体(population)而非样本(sample)数据、以微观层面(micro-level)而非集合层面(aggregate-level)数据来提高社会科学定量分析在方法选择上的灵活性和在多层次社会信息利用上的有效性。[1] 除了较为经典的横截面(cross-sectional)人口普查微观数据库外,近年来一些基于长时段(longitudinal)行政记录的大规模量化微观数据库,如人口户籍数据、经济交易数据和医疗数据等,大大推动了人们对社会不平等、人口与健康、教育与经济发展等许多社会科学经典议题的新认识。[2]

1. 有关大规模量化微观数据对学术研究和社会科学方法论的促进作用,参见鲁格尔斯(Ruggles,2014)的研究。

2. 有关近年来国际上包括人口统计微观数据和人口户籍微观数据在内的横截面和长时段大规模量化微观数据库的发展及对人文、社会和自然科学的学术贡献,参见:Dong, Hao, Cameron Campbell, Satomi Kurosu, Wenshan Yang and James Z. Lee. "New Sources for Comparative Social Science: Historical Population Panel Data from East Asia." *Demography* (Forthcoming).

一方面,这些长时段、大规模、多层次的新型微观数据库为复杂统计分析方法的发展提供了动力和试验田;另一方面,更为重要的是,其全面或系统覆盖被研究人口总体的特点降低了基础定量分析研究的门槛。比如,通过构建相应量化数据库,一些不是特别精通高级统计分析技术的学者也可能理解和进行基本描述性统计和相关分析,通过"更全、更细致"的数据而非"更精、更复杂"的方法得出一些基础但重要的新发现,这有利于把定量分析在研究广度与定性分析在视野深度上的优势结合起来,促进社会科学的整体发展。

中国史学家同样有重视考证方法的科学化和系统性扩展研究材料的学术传统。当代中国史学家发现,历史资料大扩展在推动历史研究的同时,也会成为历史研究的障碍和壁垒。历史学家如果不能用有效的研究方法进行武装,那么,在历史资料大扩展的同时,历史学研究可能会走入"愈发琐碎的考证的泥潭"。借助社会科学化的研究方法,例如建设大规模数据库,则可能是逾越"大历史观"、整体史研究与繁芜历史资料间鸿沟的有效办法(金观涛、刘青峰,2008),这为今后历史学研究的发展指明一个新的方向。

我们在为大规模微观数据库和定量分析方法给社会科学和史学研究打开一扇新大门而欢欣鼓舞的同时,更要清楚地意识到,没有一种新方法的运用和发展是简单、顺畅的。诚如许多同仁指出的那样,与其他任何社会科学定量和定性研究一样,我们在运用这种方法进行分析时,需要特别注意正确把握数据材料的代表性(或者说选择性)、编码的可靠性、分析方法的合理性,以及结论的特殊性与一般性等。分析历史数据的复杂性往往更甚于分析当代数据,所以,在这条通过运用"新"方法研究"老"数据以加深认识个人社会长期互动发展的学术道路上,还需要更多社会科学界和史学界的同仁贡献才智经验,跳出学科壁垒,共同努力探索。本文将以过去 10 多年从事 20 世纪中国大学生学籍卡资料收集与研究分析的经验,简要探讨对此问题的认识,并回答一些学界同仁对这种研究的质疑或困惑。

一、历史资料社会科学化研究的逻辑

历史资料能留存至现在是有很强的"随意"性的。许多有重大学术价值的历史资料和数据甚至是"碰运气"得来的。当下社会科学定量研

究和数据收集方法论体系基本上是20世纪中后期学科迅速发展的产物。绝大部分历史资料都早于这个阶段,而且绝大部分历史记录都并非为学术研究而存在,而往往是行政管理、商业交易、宗教风俗等各种社会经济制度的衍生品。这些现实共同决定了基于历史数据的学术研究几乎无法像许多当代定量研究那样事先制定详细的研究计划和步骤、设计具体变量和指标、确定特定研究对象和科学化收集相应信息。但是,无论是对历史数据,还是当代数据的研究,都是同等重要且互补的,其目的也更为一致,即如何通过有限的数据最大化地巩固已知或构建新知识。

显然,历史数据的局限比当代数据要大得多。一方面,由于不是专为学术研究设计和收集,历史数据所包含的信息往往未必尽善尽美,并且无法重新收集。另一方面,更为现实的是,时至今日,很多历史资料都因各种原因没有得到完整留存或公开,导致数据缺失。可供分析的有效数据往往只局限在一定时期、一定地区或其他类型的局部个案。

个案研究的逻辑往往成为应对历史数据所具有的各类现实问题的折中之策。首先需要说明的是,我们并不认同有学者强调的"标准的定量分析的基本前提是随机抽样"。随机抽样只是通过利用有限"个案"信息合理推论"总体"情况的一种收集信息和统计推论的方法,远非定量分析的全部或"基本前提"。各种基于"总体"数据直接进行分析的定量研究,如常见的基于人口普查微观数据的研究,就是反例。就受限于数据缺失的大规模量化微观数据"个案"研究而言,只要数据能全面系统覆盖这个"个案"(如某些地区、某些组织等)内的所有分析个体(如人、事件等),那么,这些数据记录的就是该"个案"自身所定义的"总体"。基于"个案"数据应用各种定量分析工具得出的结果,对于这个"个案"本身是可靠和有效的。如需进一步基于"个案"数据结果对由"个案"组成的更高层面的"总体"(如一个包含很多地区的国家或包含很多组织的行业)进行推论,则需要研究者继续附加一定的分析和讨论,不能直接将"个案"的经验等同于"整体"的经验。前文提到,留存下来或可供分析的历史数据往往比较"随意"且原因复杂,若将这些可供分析的历史数据作为"个案",那么这些"个案"往往不具有对"整体"的统计代表性,因为它们并非由严格意义上的随机抽样来确定,但这不等于不可以用个案分析的结论来推论总体状况。实际上,有学者已经指

出，以少数研究对象来推理总体存在着两种不同逻辑或方法，一种可称为统计性扩大化推理，即从样本推论到总体的归纳推理形式，这也是统计调查的逻辑基础。另一种则可称为分析性推理，即直接从个案上升到一般结论的推理形式，这是个案研究的逻辑基础（王宁，2002）。所有个案都是自身个性和总体共性的统一体，如果“个案能较好的体现某种共性，那个案就具有了典型性”（王宁，2002），个案研究也由此能推论或揭示总体的某些普遍性。这两种逻辑非但没有高下之分，而且还可以互为补充。当然，不同于抽样调查数据相对直接客观的“总体”代表性，在基于个案研究逻辑的学术讨论中，到底研究者认为的“个案”代表性是合理还是臆测，其对“整体”的推论是恰到好处还是言过其实，则是一个“仁者见仁，智者见智”的具体实证问题，而非方法论问题。

“无声的革命”[3]便是这种基于大规模长时段微观数据库的“个案”研究。“无声的革命”研究以北京大学和苏州大学两个个案为基础，通过定量分析两校 1949 年后到 21 世纪初共 15 万多份的本科生学籍资料，探讨 50 年来两校学生社会来源的结构性变迁，并以此为基础探讨中国高等教育，尤其是精英高等教育的状况。对北京大学、苏州大学两校的选择，显然不是对全国精英大学随机抽样的结果，但我们认为，通过这两个高校的分析结果在一定意义上对了解中国高等教育整体情况和历史演变有很大帮助。

首先，自统招统考制度确立以后，全国高校的招生都是在此规则之下，较为被动地从教育主管部门划定的招生范围内按计划分批次招收符合规定的学生。统招统考的政策基本限定了学校招什么条件的学生，在哪里招学生以及招多少学生等等。同档次、类型相近的大学由于招生分数线或标准相似，新生来源具有相似性。因此，在这种制度下，北京大学和苏州大学个案具有一定的“普遍性”，依据它们的数据可以窥见全国或江苏同分数段或同等类型高校的情况。

其次，“无声的革命”研究使用的是多重个案，而不是单一个案，两个不同的个案可以帮助我们理解和确认结论的可靠性和普遍性。只有两所大学在数量上当然谈不上丰富，但北京大学和苏州大学几乎在各个方面都不尽相同，比如精英程度、隶属关系、历史沿革、影响范围与招

3. 这一研究包括一篇文章（梁晨等，2012）和一本著作（梁晨等，2013），以下简称“无声的革命”研究。

生区域等均存在差异,而且又与同层级大学有很多共性,据此推论不同层级精英大学的学生来源状况较为合适。[4]

最后,即使对于定量研究本身而言,学籍卡等个人历史信息资料也可以弥补相关随机抽样调查数据研究的不足。很多有关高等教育和生源社会背景的定量分析都依靠抽样调查数据。这类数据强调统计代表性,试图通过统计推论将样本所反映的信息推及到研究者关注的学生总体,但抛开回溯性社会调查所面临的如生存、迁移等选择性偏误和记忆偏误不谈,社会抽样调查资料除特殊设计的问题以外,往往无法提供更多被调查个体在当时所处的社会环境和同辈(peer)信息,如其他同校、同专业、同时期学生的家庭背景、性别构成等等。由此可见,基于具体学校的全部学籍卡建成的数据库的首要学术价值不在于统计推论全国整体,而在于它能够真实地反映不同时期人群(cohort)与其所处不同社会环境的互动。在这个意义上,结合调查数据的统计代表性与档案数据的深度与历时性,或许会成为未来一个很有价值的研究方向。

二、历史资料数据化处理的问题与应对

对各类历史资料所含信息进行系统、合理的分类与编码是开展数据库构建和进行最终定量分析的基础和前提,但历史资料并不是为既定的社会科学研究编码或分类体系而创建的,选用何种标准可以准确、合理地在定量分析中反映历史资料信息便成为难题,更何况大规模、长时期的历史资料还普遍存在体量庞大、填写混乱、内容缺失和不同年代同类信息含义有差异等诸多问题。采用灵活、有效的编码方法成为研究历史数据成败的关键。基于现当代大学学籍卡资料的研究,我们认为以下几条经验或原则可供参考:[5]

第一,通用、权威的分类或编码标准可作为历史数据编码的基础,但决不能为单一标准所囿,应充分考虑材料与研究对象的实际,尽量挖掘材料自身信息,采取多种方法,保证计算结果的准确性。比如,对于学生的专业院系,国家前后有不同的划分方法和标准,各学校又有一定的灵活性和差异性,因此,我们既要尊重各学校的实际,又要参照国家

4. 关于以上两方面北京大学、苏州大学个案的可推理性更具体讨论,可参见:梁晨等(2013:29—36)。

5. 学籍卡研究时编码工作的具体考虑和操作可参阅梁晨等(2013:37—57)。

标准对院系专业进行两次编码和相应运算。对于职业编码，我们以《国家职业大典》为依据，[6]同时参考边燕杰等在“中国综合社会调查”中的办法（边燕杰、李路路、蔡禾，2006），对具体材料进行充分挖掘，尽量多和准确地对职业进行编码分类。我们不仅采用了能够找到的不同年代的国家或各省颁行的省重点中学名单，也根据实际情况，对部分市县重点，以及在江苏地区乡镇地方社会占据重要地位的“县中”进行了编码统计，使得研究不仅揭示了中国大学生的社会来源，也直接反映了中国各类中学的学生构成情况。

第二，长期、连续的大规模历史资料往往包含多种信息，在研究分析时需要仔细辨别，选择最贴切的信息构建研究变量，而不是选取便于归类、编码却可能失真的信息。例如，一般学籍卡资料都提供了学生户籍和家庭住址两项信息。以往的研究习惯注重分析由官方统一划定的户籍内容，忽视相对较难处理的具体家庭住址信息，但众所周知，改革开放以来，中国社会人户分离的现象日趋突出，特别是大量农业户籍的人口实际常年生活居住在城镇中，户籍对于居民城乡流动的限制越来越小，比如，湖南、四川、安徽等省的大量农民南下广州和深圳打工，江苏也有很多农民离开农村，进入附近的城市、集镇的乡镇企业工作等。户籍在反映一个人或家庭的实际生活与工作环境上并不如家庭居住地址直接。当我们的研究是关注学生真正的居住地的城乡分布时，以户籍来区分学生的城乡来源不如以家庭地址区分更加符合实际。[7]又比如，学籍卡中普遍存在的家长职业和家庭出身内容都可以在一定程度反映学生的家庭背景。家庭出身作为一项重要的官方身份区分标准，学生填写的标准程度很高。虽然家庭出身在1949年以后的30多年里确实起到一定的社会区分作用，也有一套完善和易于统计操作的分类标准，使得目前不少学术研究热衷于对其进行分析统计，但实际上，在20世纪80年代中期以后，家庭出身就基本失去了社会意义，根本无法

6.《中华人民共和国职业分类大典》（以下简称《职业分类大典》）于1999年5月颁布。作为中国“第一部对职业进行科学分类的权威性文献”（参见该书前言），《职业分类大典》将职业划分为8大类（1位码）、66中类（3位码）、413小类（5位码）和1838细类（7位码）。因应新职业的不断涌现，在保持总体框架不变的原则下，国家劳动和社会保障部先后组织专家对《职业分类大典》进行了增补修订。

7. 李强（2008）认为改革开放以后，中国社会有5个方面的重要转变导致原有社会身份体系的瓦解，其中第一个便是“农民开始突破户籍身份的限制”。

反映社会的变革与实际,所以我们选择分析家长职业而不是家庭出身。

因此,尽管户籍和家庭出身信息非常容易被分类,且被许多主流社会抽样调查问卷采用,我们也确实对其进行了定量分析,但最终并没有在主体研究中采用和汇报结果。尤其需要注意的是,由于社会抽样调查强调问卷对全国不同地区和不同年龄人口的普遍适用性,其变量设置不免会更加注意简明直接的分类。大部分可供分析的相关信息都是以类似的形式出现,大部分采用这些数据的定量研究文献就不可避免地采用了这些分类的方式,成为主流。但这并不代表在有其他形式的材料和更丰富的信息时,我们应该僵化地沿袭既有分类,放弃更加细致的信息所带来的学术价值。所以,我们认为研究者必须注重社会的实际变迁,选取历史资料中最能提供翔实信息的变量设计与编码方式,而不能局限于已有的标准或研究习惯。当然,在最大化利用丰富信息的同时,重视研究设计和结果与已有研究,尤其基于不同材料类型的研究的可比性,同样是非常重要的。

第三、对历史资料的灵活编码,还体现在根据具体研究问题的需要,对特定变量编码的宽严结合。学籍卡一类的材料,表格信息都由学生填写而来,大家对填写标准的认识程度参差不齐,内容自然五花八门,这给编码带来了很大难度。实际上,类似的情况非常普遍,自填型的社会问卷调查和其他信息收集方式同样不可避免,如果不谨慎处理,就可能会影响研究结果的准确性。就“无声的革命”研究而言,职业编码即是如此。学生在此处填写父母工作单位名称的有之,填写职务的有之,混合杂乱的亦有之。比如,对于职业编码分类中的第一大类“国家机关、党群组织、企业、事业单位负责人”或我们通俗称呼的“干部”,根据学生填写的信息种类,很多难以准确判断是负责人还是一般工作人员。很多学生填写了具体职务,如主任、队长、经理等,不结合更具体的信息根本无法弄清楚这些职务是否算得上负责人,因为太多不同等级和类型的单位可能存在这些头衔。如主任可能是中央机关的部门负责人,也可能是乡村的治保主任或妇女主任,性质完全不同。经理在改革开放后的经济活跃地区更是多如牛毛,任何一个单位的销售员、采购员都可以冠以此头衔,以便于他们的工作。

考虑到大部分历史数据都无法如当代专门社会调查一样可以重新收集和确认信息,我们建议遵循的处理原则是在明确研究侧重点的基

础上,通过灵活编码分类实现对某些值得特别关注的类别的测量偏误(measurement error)方向进行大概的控制和估计。“无声的革命”研究关注的是精英大学的生源开放性,即关注较普通社会阶层如工农家庭等学生的比重,所以我们对“干部”采取宽松的编码标准,但对“工人”、“农民”则采取严格的标准。各种带“长”的职业、各类经理、意义含混的干部都被按负责人编码。这样做,会在分析时高估父母是干部的学生数量,使得干部子女比例比实际要大,同时低估父母是工农职业的学生数量,使工农子女比例可能比实际要小。按照这样的统计标准,在清楚测量偏误方向的情况下,我们仍发现,无论和中国过往的历史,还是和西方近 30 年的情况进行比较,高考制度下的两所精英大学仍有较高比例的工农学生。换句话说,在理想情况下,如果我们所担心的学生自填信息偏差得到纠正,真实干部子女比例可能更低,真实工农子女的比例可能更高,所以我们的研究设计和结果对于这类测量偏误是稳健的,只有这样我们对定量结果和由此产生的推论才比较有把握。

总的来说,由于时代背景错综复杂,历史资料建立和涵盖的时间长短不一,记录的内容可能不一致或不完整,且难以以今日的常识去直接理解等,使得历史资料的编码与量化分析存在一定困难和特殊性。将历史资料所记载的复杂信息灵活妥当地分类并设计变量编码方式,并非简单依靠电脑技术或其他模版即可实现。这种历史资料的复杂性一直是历史研究的难点,同时也是历史学学者学习、训练和研究的重点,因此,历史资料的编码与量化,需要社会科学化的分析技术与知识,但同样离不开历史学等人文学科的功底和经验。

三、历史数据分析结果的理解与推论

定量分析不仅能帮助发现很多未知现象,很多发现还与我们的预设、常识和已知有很大不同。依照我们的经验,不仅是学籍卡资料,包括很多人口记录、财产登记等长期、大规模历史资料在数据库化并进行定量分析后,非常容易有一些“不期而遇”的发现。“无声的革命”研究很多重要的新发现与我们最初的假设或前人研究的经验认识就不相符。对这些新发现的理解,仅靠数据本身或定量分析其实是不够的。历史资料的收集整理和量化数据库化是相当有难度的,而对历史数据定量分析结果的理解和诠释挑战更大,是当前和未来从事历史数据定

量分析的学者所必须面对的难题。因此,构建历史资料量化数据库和定量分析,要想真正有效地理解数据,对规律性现象形成深入认识,仅靠单纯的技术是无法完成的,必须将计算结果或新发现放回历史结构或背景中去,才能得出系统和合理的结论,从而不仅能通过数据挖掘新发现,更能解释和理解这些发现,进而改变我们固有的历史和社会科学理论与认知。正是因为社会历史环境与数据材料的双重复杂性,我们认为,类似基于长时段大规模历史数据的社会研究一定需要重视细化分析时空趋势,不可脱离环境,一概而论。

比如,我们曾普遍认为"文革"时期是中国农民教育状况最"革命"和"进步"的,但包含相应部分年份所有学生资料的学籍卡数据表明,"文革"时期两校学生中农民子弟的比例不仅低于 1965 年前,更是在 1949 年后首次出现下降。[8]这促使我们对"文革"宣称的将几千年科举制度和招生考试政策一下子废除,直接从工农兵中直接推荐"政治思想好"、"身体健康"和"有一定实践经验"的青年上大学的制度的真实效果产生怀疑(中央教育科学研究所,1992:433)。工农兵学员推荐制看起来是非常有利于工农子弟上大学的,在政策推行的 1970—1972 年期间似乎也确实达到了效果,但这两三年的情况并不能真实反映整个时期的现实。国内很多研究和回忆都指出,在制度推行了几年以后,"开后门"等腐败现象便越来越严重,各级干部子弟都可以方便地利用特权进入大学。有学者指出,"文革"中各种"走后门"入学取代了此前的"走后门"当兵,成为干部子女角逐竞奔的新热点(刘小萌,2009:217)。农村干部也纷纷利用自己掌握的推荐权力为亲友"开后门"。四川有知青回忆说:"(高校)招生几乎完全被公社及大队干部所垄断。有些地方将这些干部的娃娃依次排队,推荐的名单竟排到了 1987 年。"(刘小萌,2009:218)苏州大学学生的状况也对此有所体现。当时苏州大学学生中的工农比例虽然达到最高点,但如果分析学生家长的职业类别则会发现,干部子弟在这个时候出现强劲增长,工人子弟的比重还出现了下降。从家庭居住地来看,农村的学生出现大幅度下降,城市学生则上升。其他一些学者的研究也证明工农兵学员实际上被各级官员所掌

8. 相关统计结果请参见:梁晨等(2013:第三章)。

控,后期学生来源基本上是各类干部或干部子弟。一些国际同行的研究也印证了这一看法。周雪光等(Zhou and Hou,1999)对中国 20 个城市的调查发现,工农兵学员中干部子弟的比例异常得高是一个突出的现象。1972—1979 年间,父亲的职业对子女教育和职业获得的影响越来越大,干部子弟的优势也越发明显。考虑到这一现实和中国长期以来的社会文化,我们认为,这表明客观的考试让位于侧重各种表现的推荐是有极大风险的。

又比如,统招统考制度的确立在一定程度上使中学制度成为理解中国高等教育革命性转变的关键。两校的一个基本史实是,生源高度集中于少数重点中学,特别是省级重点中学。这促使我们以数据为基础,结合改革开放前后两校学生社会来源的总体状况、地区分布状况等反思如今被广为诟病的重点中学制度。一方面,我们当然认可重点中学制度存在诸多不合理的层面,比如重点中学主要分布在城市而农村很少设立,重点校与非重点校、城市与乡村的双重二元结构叠加成为精英大学中农村学生减少的重要原因。又比如,很多省重点的"县中"成为所谓"高考集中营",学生的学习压力巨大,素质教育无从落实等。另一方面,由于重点中学在招生时受"高考"导向影响,客观考试是主要方式,在一定程度上保证了出身照顾政策取消后,两校生源在相当长时期内还能保持多样性。重点中学重视分数和培养学生高考能力在这个角度看是有一定积极性的。我们发现,一个地区如果经济发展条件不是很好,可以通过集中加强教育投入,特别是对中学等基础教育予以重点投入,一样能取得超出经济水平的教育优势,这点对于经济水平中等的地区尤为适用。比如,江苏南通市在经济方面并不突出,但整个地区重视教育投入,不仅市区,甚至每个县(市)的中学都是重点,高考表现非常突出。而"县中"作为广大乡镇、农村子弟升入优秀大学的最关键渠道,如果某些省,如江苏、福建等,努力将全省范围内每个县至少一所中学打造成省级重点中学,其农村生进入北京大学的能力就会很突出。[9]

我们认为,应该重视重点中学向精英大学输送学生高比例的现实,更好地发挥其突出的输送能力,从而保证大学生源多样性的延续。简单废除重点中学的做法在现实中已经被证明不成功。近年来,中国社

9. 关于北京大学、苏州大学两校农村学生的地区分布可参见梁晨等(2013:85—88)。

会刮起了一股废除重点中学的风潮,重点中学作为教育不公平和社会不公平的重要标靶,成为舆论讨伐的对象,大多数省区教育部门也先后表态要取消重点中学。实际的情况是,大多数省区只是将重点中学的名号换成了“示范校”、“星级校”,真正的变化其实并不大。从北京大学、苏州大学的情况看,与其纠缠重点校的存废,不如更关注重点校招生的公平与多样化,逐步提高重点与非重点间的差距,让更多学校达到重点水准,才是在新世纪保持并拓展精英大学生源多样性的有效与现实的途径。

基于上述探讨可见,对历史资料进行研究和讨论,是需要定量研究与定性研究并重的。有学者认为定量与定性两种基本研究方法“逻辑各异”,在“无声的革命”研究中没有合理统一,我们认为这当然是值得继续探讨和提高的。针对我们前文所阐释的个案研究逻辑对整体认识的贡献,作为解决研究历史数据所面临的实际问题的折中之选,若有学者能提出合理统一的定性定量方法,解决这种“方法错位”的方式,我们非常愿意学习和改善。不过,我们认为,定量与定性研究虽然有诸多具体操作层面的不同,但在最基本的逻辑上是殊途同归的,本质都是基于问题,厘清事实,确定解释并排除替代解释(alternative explanation),最后形成或修正理论。因此,巴比(2012:24—26)认为,两者的主要区别仅在于资料的“数据化与非数据化”,“数量分析本身不是目的,只是认识的手段”(巴勒克拉夫,1987)。在学术研究中单独使用定性或定量方法都存在明显不足,作为两种最基本的研究方法,定量分析与定性分析应该互相融合而不是截然对立的(吴承明,1993,2006;Creswell,2003)。

四、历史资料数据化与定量分析的必要性

人文与社会科学研究往往会预先设定研究问题或理论模型和假设,然后去寻找相关材料,但部分研究因为材料收集有较强的主观性和选择性,往往倾向重复确认“已知”,忽略发现“未知”,很难促进对社会事物整体规律统一且有效的认知。在历史学领域,李伯重(2002)认为存在“选精”与“集粹”两种惯用却弊端很大的研究方法。这两种方法在本质上并无较大差异,其特点都是通过从历史资料中选取具有代表性的例证来推导出结论。但代表性本身并没有统一标准,研究者常将“某一或某些例证所反映的现象普遍化”,从而丧失真实性,研究结果自然

不可靠。[10]因此，众多基础的“史实”在当下中国可能都需要学者们努力“重建”一番（茅海建，2005：自序）。

构建量化历史数据库和定量分析的研究模式是在明确研究方向和研究问题的基础上，以“大数据”为驱动，发现社会实际与证实/修正社会理论并重。定量分析不应只是对数据单纯地简单描述统计和相关分析，而应以理论为导向从大规模数据中发现“已知”和“未知”规律并回应、补充或升华已有理论的有效工具（Boonstra，Breure and Doorn，2006）。西方主流学界已经意识到大数据分析是非常重要的学术工具（Big Data is a Big Deal），能提升众多学科的研究模式与深度（Shaw，2014）。量化数据库构建和定量分析使得部分历史研究从“解读分析特定历史文献和档案的历史研究经典模式向收集、整理和统计分析系统成规模的历史记录数据的社会科学研究模式转变”（Anderson，2007），通过全面的大数据分析寻找出相关现象或规律。同时，更多的大规模历史数据，以及更多历史学家在定量方面学术意识的提高，对社会科学更深入地研究各种社会、经济、人口方面的长期演变大有裨益。大规模微观数据库一般包含了某一范围内大部分或所有分析个体的状况，而且不同于抽样数据，可以允许通过个体层面数据对当时的宏观层面社会环境信息进行一定程度的复原。现有统计分析方法可以兼顾个体分布的影响与权重，还可以估计个体、组织和整体等不同社会层面特点的影响和互动，从而在帮助历史学者避免选择材料时的疏漏与偏废的同时，还可以帮助社会科学学者拓宽研究领域和视角，更加全面地理解许多经典议题和当代问题。

当然，必须强调的是，正如前文所论述的那样，构建历史数据库并开展量化研究绝非易事。一方面，构建量化数据库通常包括数据采集、数据分类、数据编码、数据存储、数据信息挖掘和定量分析等多个环节，数据库建成后还可能需要数据管理和维护等多种工作。相对于这种以数据为中心的“科学化”、“电子化”的研究方式，“传统史学研究多少显得有些手工艺式的陈旧”（Best，1991），这使得当下很多史学研究者很难掌握这一研究方法。因此，尽管很多历史学学者也承认量化分析可

10. 李伯重（2002：110－121）指出前者是指“从有关材料中选取一两种据信是最重要或最有‘代表性’的，以此为据来概括全面”；后者是指在研究“一个较长时期或一个较大地区中的重大历史现象时，将与此有关的各种历史资料尽量搜寻出来加以取舍，从中挑选出若干最重要或最有代表性的，集中到一起，合成一个全面性证据，以求勾画历史现象的全貌”。

以为“描写大人口群的历史提供了巨大机会”,但他们对待量化研究的态度却非常消极(Hudson,2000)。另一方面,作为一种研究方法,历史数据库构建和定量分析显然有其适用范围。李伯重(2013)认为,在历史研究中需要量化的只是其中一部分,不可夸大量化史学的适用范围。但他也指出,确实有一些历史研究问题可以,并且需要(甚至必须)量化研究。因此,离开了具体的研究对象去抽象地谈论量化方法是否适用于史学研究的问题是没有意义的。

我们想进一步指出,除了研究问题的需要外,历史资料本身的条件对应否以及如何数据库化和量化更为重要。只有根据历史资料的特点,依照“史无定法”原则“量体裁衣”,才能完成历史资料的量化分析。而对社会科学学者来说,尽管部分已经较好地掌握了数据构建、管理和分析的技术,并对各种社会现象和问题有良好的研究直觉,但对历史资料最根本的认识(以决定如何构建变量编码、理解选择及测量偏误和设计分析架构),以及对相应历史社会环境的整体把握,往往是令其望而却步的主要困难和挑战。鉴于我们研究组的自身经验,组成由不同学科背景的研究成员密切互动的专项研究团队,实现跨学科合作,是一条切实可行和较有成效的道路。

具体到“无声的革命”研究来说,探索和尝试仍然还在进行中。更深入地探讨中国精英教育获得者的社会来源与结构转变这样的重大问题,两个个案和 50 年的期限还是显得单薄了点。由于每所大学都有自己的特性,个案少使得对一些具体统计现象的认识产生困难,即使可以反映一些普遍性的规律,但必然会影响对细节的理解认识。民国时期大学的生源状况和最近十几年中国大学生的社会来源的不同对于理解生源转变以及制度设计意义重大。因此,目前我们一方面在努力寻求和其他高校的合作,特别注重选取多样化的类型和不同分布地区的高校,建设更多的学籍卡数据库。另一方面,我们正试图利用各地官方档案馆系统所保存的基本都开放的民国学籍材料,努力构建全国性的民国时期大学生学籍卡数据库,从而从资料的广度上推动研究发展。

在新事物面前,人们可能会兴奋并产生憧憬,也可能会惶恐并形成排斥,但对于学术研究而言,每一种新方法的应用,又都需要一定时间扎实的工作与反复的论证。没有一种方法可以包打天下,但每一次方法的革新都可能推动我们的研究和认识。对于历史资料的数据库化和

定量分析，我们应该持开放心态，积极探索，推动史学研究和社会科学研究的融合和共同发展。

参考文献(References)

巴比，艾尔. 2012. 社会研究方法[M]. 邱泽奇，译. 北京：华夏出版社.

巴勒克拉夫，杰弗里. 1987. 当代史学新趋势[M]. 杨豫，译. 上海译文出版社.

边燕杰、李路路、蔡禾. 2006. 社会调查方法与技术：中国实践[M]. 北京：社会科学文献出版社.

金观涛、刘青峰. 2008. 历史的真实性：试论数据库新方法在历史研究的应用[J]. 清史研究(1)：90－108.

李伯重. 2002. 理论、方法、发展趋势——中国经济史研究新探[M]. 北京：清华大学出版社.

李伯重. 2013. 史料与量化[D]. "第二届清华大学量化历史国际学术年会"论文.

李强. 2008. 改革开放30年来中国社会分层结构的变迁[J]. 北京社会科学(5)：47－60.

梁晨、李中清、张浩、李兰、阮丹青、康文林、杨善华. 2012. 无声的革命：北京大学与苏州大学学生社会来源研究(1952－2002)[J]. 中国社会科学(1)：98－118.

梁晨、张浩、李中清，等. 2013. 无声的革命：北京大学、苏州大学学生社会来源研究(1949－2002)[M]. 北京：生活·读书·新知三联书店.

刘小萌. 2009. 中国知青史：大潮(1966－1980年)[M]. 北京：当代中国出版社.

茅海建. 2005. 戊戌变法史事考[M]. 北京：生活·读书·新知三联书店.

王宁. 2002. 代表性还是典型性？个案的属性与个案研究方法的逻辑基础[J]. 社会学研究(5)：123－125.

吴承明. 1993. 论历史主义[J]. 中国经济史研究(2)：1－9.

吴承明. 2006. 经济史：历史观与方法论[M]. 上海财经大学出版社.

中央教育科学研究所. 1992. 中华人民共和国教育大事记(1949－1982)[M]. 北京：教育科学出版社.

Anderson, Margo. 2007. "Quantitative History." In *The Sage Handbook of Social Science Methodology*, edited by William Outhwaite and Stephen Turner. London: Sage Publications: 246－263.

Best, Heinrich. 1991. "Technology or Methodology? Quantitative Historical Research in Germany." *Computer and Humanities* 25(2/3): 163－171.

Boonstra, Onno, Leen Breure and Peter Doorn. 2006. *Past, Present and Future of Historical Information Science*. Helsinki: Edita.

Creswell, John W. 2003. *Research Design: Qualitative, Quantitative, and mixed methods approaches* (Second Edition). Sage Publications Inc.

Hudson, 2000. *History by Numbers: An Introduction to Quantitative Approaches*. London: Arnold.

Morris, R. J. 1990. "History & Computing, A New Magazine." *Historical Social Research* 15(1): 118－120.

Ruggles, Steven. 2014. "Big Microdata for Population Research." *Demography* 51(1): 287－297.

Shaw, Jonathan. 2014. "Why 'Big Data' is a Big Deal: Information Sciences Promises to Change the World." *Harvard Magazine* 3(March-April): 30－35, 74－75.

Zhou, Xueguang and Liren Hou. 1999. "Children of the Culture Revolution: The State and the Life Course in the People's Republic of China." *American Sociological Review* 64(1): 12－36.

实习编辑：岳　芸
责任编辑：张　军