

《为 21 世纪经济统计的大数据》述评

——“美国官方经济统计基础结构”的演化及启示

邱 东¹, 刘婷祎²

(1.江西财经大学 统计与数据科学学院,江西 南昌 330013;2.武汉大学 经济与管理学院,湖北 武汉 430072)

[摘要] 关于大数据背景下21世纪的官方经济统计如何改革,2019年美国收入与财富研究会(CRIW)召开了一次专题研讨会,会议论文集《为21世纪经济统计的大数据》收录了15篇论文,其序言是一篇综述性论文,主要概述相关研究背景和15篇论文的主要思想。结合美国官方经济统计基础结构的演化,对该论文集从十个方面展开述评,包括论文集概况、序言的主要内容、核心概念“基础结构”、“数据格局”的演变、大数据与经济统计结合的探索及启示、经济统计变革与时局的高度相关性、大数据与经济统计结合的方法论挑战与推进、推进经济统计中大数据应用的挑战和应对措施、方法论研究的持久性、经济统计与大数据交叉和结合研究的学科意义。

[关键词] 《为21世纪经济统计的大数据》;经济统计;大数据应用;数据基础结构;数据格局

[中图分类号] F222

[文献标识码] A

[文章编号] 2096-7411(2024)01-0001-11

Review on *Big Data and Economic Statistics in the 21st Century*

——Evolution and Enlightenment of “the Basic Structure of American Official Economic Statistics”

QIU Dong¹, LIU Ting-yi²

(1.School of Statistics and Data Science, Jiangxi University of Finance and Economics, Nanchang 330013; 2. Economics and Management School, Wuhan University, Wuhan 430072, China)

Abstract: In regard to how to reform official economic statistics in the 21st century under the background of big data, the International Association for Research in Income and Wealth (CRIW) held a symposium in 2019, and published a book of papers, named *Big Data and Economic Statistics in the 21st Century*. The book of papers contained 15 essays, and adopted a review paper as its preface. Based on the evolution of the basic structure of American official economic statistics, the paper conducted the review of this book from 10 aspects, including the overview of the book, the main content of the preface, the “basic structure” of core concepts, the evolution of “data pattern”, the exploration and enlightenment of the integration of big data and economic statistics, the high relevance of economic statistics reform and the times, the challenges and advancement of the methodology of the integration of big data

[基金项目] 国家社科基金“加快构建中国特色哲学社会科学学科体系、学术体系、话语体系”研究专项项目“新时代中国特色统计学基本问题研究”(19VXK08)

[作者简介] 邱 东(1957—),男,江苏张家港人,江西财经大学讲席教授,1993年获国务院学位办批博士生导师,1995年国务院特殊津贴获得者,2008年被聘为教育部长江学者特聘教授,研究成果集中于国民经济核算、经济测度、国际比较、可持续发展等领域,代表作为“当代经济统计学批判系列”,现任全国哲学社会科学规划项目学科评审统计学组召集人、全国统计教材编审委员会副主任等;刘婷祎(1992—),女,河南开封人,武汉大学经济与管理学院博士后,主要研究方向是经济测度、国民经济核算,本文通讯作者。

and economic statistics, the challenges and countermeasures in the process of applying big data in promoting economic statistics, the persistence of methodological research, as well as the discipline significance of the integration.

Key Words: *Big Data and Economic Statistics in the 21st Century*; economic statistics; information about big data application; basic structure; data pattern

一、《为21世纪经济统计的大数据》^①概况

在大数据背景下21世纪的官方经济统计应该如何改革?2019年3月在美国马里兰,斯隆基金会资助“美国收入与财富研究会(CRIW)”召开了一次专题研讨会,会议论文集《为21世纪经济统计的大

数据》由美国“国家经济研究局(NBER)”于2022年正式出版,该文集收录了15篇会议论文,其序言是一篇综述性论文,主要概述相关研究背景和15篇论文的主要思想。^②

表1 文集论文题目与作者

章节	论文题目	作者
序言	为21世纪经济统计的大数据:时不我待	Katharine G Abraham, Ron S Jarmin, Brian C Moyer, Matthew D Shapiro
第一部分 经济统计中推进 大数据的综合利用	关键国民经济指标的再构造(reengineering)	Gabriel Ehrlich, David Johnson, Matthew D Shapiro, John Haltiwanger, Ron S Jarmin
	美国消费者价格指数中的大数据:经验与计划	Crystal G Konny, Brendan K Williams, David M Friedman
	用“可替代数据源(Alternative Data Sources)”提升零售贸易数据产品	Rebecca J Hutchinson
	从交易数据到经济统计:为消费者开支构造实时、高频的地域测度	Rebecca J Hutchinson
第二部分 分类的大数据利用	将“自然发生文本数据”转换进入经济统计:“线上职位空缺公告(Online Job Vacancy Postings)”的案例	Arthur Turrell, Bradley Speigner, David Copple, Open Jobs Data at Nesta Jyldyz Djumalieva, James Thurgood
	2017年经济普查中特许经营问题的自动回应评估	Joseph Staudt, Shawn Klimek, Andrew Baer, Yifang Wei, Lisa Singh, Jensen J Bradford
	用公共数据生成产业分类代码	John Cuffe, Justin C Smith, Nevada Basdeo, Sudip Bhattacharjee, Ugochukwu Etudo, Wayfair Nathaniel, Shawn R Roberts
第三部分 部门测度的 大数据利用	地方经济的即时预报:使用Yelp数据测度经济活动	Edward L Glaeser, Michael Luca, Hyunjin Kim
	为进口和出口价格指数的单位价值:概念的证明	BLS Don A Fast, Susan E Fleck
	使用保险索赔和管理数据定量分析保健医疗计划中实施重要护理的生产率增长	John A Romley, Dana Goldman, Neeraj Sood, BEA Abe Dunn
	在大数据时代估价房屋服务:充分利用Zillow微观数据的使用者成本法	BEA Marina Gindelsky, Scott A Wentland, Jeremy G Moulton
第四部分 方法论挑战和进展	不可躲避的竞赛:机器学习与可替代数据预测经济指标的比较	Jeffrey C Chen, Abe Dunn, Kyle Hood, Alexander Driessen, Andrea Batch
	周扫描数据中的季节和周期销售额的机器学习分析	Rishab Guha, Serena Ng
	估计新产品收益	Erwin W Diewert, Robert C Feenstra

二、文集序言《为21世纪经济统计的大数据:未来在即》的主要内容

文集的序言(Prefatory Note)是一篇综述性论文,共七个部分。

序言的开卷语阐述了三个问题:“美国官方经济统计基础结构(the infrastructure for official U.S. economic statistics)”的构建和演化过程、以“大数据

与经济统计结合”为主题的两次CRIW研讨会、大数据与经济统计结合的必要性和可行性。

序言第一小节阐述大数据对经济测度的潜在效益,在进一步阐述“数据基础结构”所面临的演变后,从时效性、数据修订和工作自动化三个方面做了概述。

序言第二小节探讨整个经济系统如何拓展大数

据在经济统计中的综合利用,主要述评论文集第一部分的五篇论文,分别探讨如何整合或细化零售价格和名义销售额的测度、开发CPI的可替代数据源、改进劳动力市场的实时测度等。

序言第三小节聚焦于统计分类细化中的大数据利用,主要述评论文集第二部分的三篇论文。统计指标的细致化处理,必然需要以正确分类为基础,而这又往往受制于经济社会现象中隐含的“模糊不确定性”。三篇论文分别探索如何利用大数据来估算“岗位空缺”、确定特许经营权的有无和产业分类系统的细化。

序言第四小节探讨部门测度的大数据利用,主要述评论文集第三部分的四篇论文。部门测度是国民核算的中观基础,论文集中改进部门测度的三个例子分别是外贸统计中进出口价格指数的编制、健康养护产出的测度和自有住房服务价值的测度。

序言第五小节阐述大数据与经济统计相结合的方法论挑战和进展,主要介绍论文集第四部分的三篇论文,探讨如何结合信用卡数据、谷歌流行数据和扫描数据等编制消费者支出和价格指数,以及如何应用机器学习方法对时间序列基础数据进行适当调整。

序言第六小节是对整个文集的总结,推进经济统计中利用大数据所面临的挑战和应对措施,主要包括以下七个方面:(1)不同类型数据间的匹配性;(2)大数据对统计机构数据供给的机制安置;(3)数据供给的连续性问题;(4)统计机构间合作及其与企业合作的整合;(5)统计数据细化与隐私保护;(6)经济统计的“计算基础结构”;(7)机构间合作进行统计方法论研究。

三、注重核心概念“基础结构(infrastructure)”

论文集综述论文中出现的第一个单词便是“基础结构(infrastructure)”,文中多次使用了这个核心概念,例如“美国官方经济统计基础结构(the infrastructure for official U.S. economic statistics)”“经济测度基础结构(economic measurement infrastructure)”“联邦数据基础结构(the federal data infrastructure)”“计算和分析基础结构(computational and analysis infrastructure)”“计算基础结构(computing infrastructure)”“现有基础结构(existing infrastructure)”“关键经济统计(数据)常规生产的基础结构(the infrastructure of the routine production of key official economic statistics)”等表述,以及“基本架构(the underlying architecture)”等相关表述。

应该注意到的是,此处“基础结构”这个概念的重心在“软实力(soft power)”,与我们通常所理解的“硬实力”方面的基础结构相比,并非处于同一个发展层级。国家所谓发达与欠发达,“基础结构”的软硬取向是一个重大的区别,基础结构建设的“脱实向虚”应该是社会发达的重要表现。

深入把握这个概念对新兴国家尤其重要,因为谋求发展时往往重视“实物基础结构(physical infrastructure)”,重视“硬实力”,重视实物消费品。例如最典型的“硬道理”就是“要致富先修路”,这是经济发展水平所决定的,发展上补课总是需要从硬件做起。由于客观物质条件的限制,“软件”建设一时还摆不上议事日程,很容易忽视“社会基础结构(social infrastructure)”,就“软实力”和公共品而言,实际上这是欠下了“工作债”。

人类社会发展的历史过程告诉我们,富强决不止于GDP总量,债务也未必都以资金的形式出现。一个国家的发展不能总是盯着以“实物基础结构”所表现出来的“硬实力”,高楼、高速、高铁固然重要,但“实力”的含义更广,即便“硬实力”的内涵也远不止这些。发展总要走高质量之路,“软实力”方面的“工作债”恐怕更为沉重,且不可长久拖欠。

政府统计数据的基础结构,正是“社会基础结构”的一个重要组成部分,是政府部门应该提供的公共品。让整个社会各方面都“知己知彼”,才可能在异常激烈的大国竞争中站稳脚跟,保持定力。对宏观数据的切实把握,正是国家提升“软实力”、提高发展质量不可或缺的基础。学习发达国家如何构建“经济统计基础结构”,应该成为当下国家统计系统的一项较为迫切的任务。

Abraham教授等(2022)^[4]开篇概要讲述了“美国联邦经济统计基础结构”的构建历史。尽管美国早在1810年就有了“制造活动普查(census of manufacturing activity)”,然而第一次“五年一度的综合经济普查(comprehensive quinquennial economic census)”开始于1954年,主要是为了应对大萧条和“二战”联邦政府强化宏观管理的需求。在20世纪40年代后期,美国整个经济仍然基于实物生产,1/3的劳工在制造部门就业。因此,美国国民核算对生产方面的指标一直比较重视。宏观经济统计的三个基本“测度领域(measurement domains)”就是生产、^③价格和就业。

从“数据基础结构”的角度看,美国经济普查提供了“国民收入和生产账户(NIPAs)”核算名义产出

的支柱,而“二战”之后发展的概率抽样,官方调查也提供了精确统计,相比“全面计数(complete enumerations)”而言成本更低,为高频经济测度做出了核心贡献。行政管理数据特别是税收记录中的收入数据,在构建NIPAs账户收入方以及在估算账户生产方的缺失数据上,发挥了重要作用。价格指数作为估算实际生产的缩减因子,独立于NIPAs发展。消费者价格指数(CPI)在1919年是以“生活成本指数(cost of living index)”的名义出现,用来缩减工资的通胀影响。普查、抽样调查和行政记录,以此三者为主体的“数据基础结构”主要构建于20世纪中叶,用户主要是联邦政府,还有其他决策者和研究人员。

以上概要描述了美国政府统计“现有基础结构”,从中可以看出其三个重要特点。

第一,政府不同部门之间工作联动。典型地,税收记录对于NIPAs的编制发挥着不可或缺的作用。如果税务信息切实可靠,则这种联动将大大减轻统计部门收集基础数据的负担。

第二,联邦统计分散实施。例如,名义产出与其缩减因子(价格指数)的编制相互独立。这种独立性或许有利于避免数据操纵,同时也隐含了国民核算中所谓“核算”的基本蕴意。但是,跨越几个各负其责的政府部门,如何提供具备一致性的综合经济测度,这种架构可能也有其难以协调的弊端。可见,任何工作制度安排都是利弊相间,无法只取其一。在国际统计体制和方法的比较中,既要看到自己的优势和别人的劣势,也要看到别人的优势和自己的劣势,这才是知己知彼的要义所在。

第三,政府统计结果与宏观管理措施密切相关。例如,“生活成本指数”的高低直接关系到工资调整的幅度。所以,联邦政府统计的数据受到全社会的关注,常常引起社会争论和博弈。麦考夫教授2019年发表论文,专门阐述了这个历史过程(Rockoff, 2019; 邱东, 2021; He et al., 2019)^[2-4]。

应该注意到,美国的第一次经济普查比中国正好早了50年。尽管发展中国家有“后发优势”,有机会避免前人走过的弯路,但这种优势并不能自动实现,其前提在于对开拓者的经验和教训都有系统的认知,能够充分地准备少走弯路的系统条件,其中之一就是经济统计基础结构的构建。

四、关注“数据格局(the data landscape)”的演变

我们可以从必要性和可能性两个基本方面来梳理和剖析“数据格局(the data landscape)”的演变。

Abraham教授等指出,官方经济统计的“数据格

局”原本主要有经济普查、抽样调查和行政记录三大类数据方式或来源。在大数据和数字经济迅猛发展的全球化背景下,“数据格局”演变所带来的问题主要有以下三点:(1)统计部门得到“调查响应(survey responses)”越来越困难,调查的“响应率(response rates)”^④明显下降,人们对数据结果的质量越来越关注;(2)作为测度对象的经济变得越来越复杂;(3)用户希望得到更及时和“更细致(more granular)”的数据。

还要看到,原本政府统计机构在公共信息的生产过程中处于近乎垄断的地位,^⑤企业和居民户(household)履行社会责任,接受统计机构的“入户访问(household interview)”和“问卷调查(paper questionnaires)”,相关基础数据只是在进入国家统计机构的计算机系统之后才数字化。而今,其他经济主体如企业或非政府组织(NGO)等,也与时俱进地参与公共信息生产,给政府统计机构带来了竞争性压力。社会有信息需求,就可能形成相应的供给,私人部门甚至对公众发布了宏观经济统计信息。^⑥例如,美国自动数据处理公司(Automatic Data Processing, ADP)的月度就业报告、大企业研究会(the Conference Board,有的译为“咨商会”)的Help Wanted Online,以及the JPMorgan Chase Institute的生产统计信息。特别是,突发的社会危机也给这种供求匹配提供了契机。基于各种私人来源的数据,Opportunity Insight's Economic Tracker提供了就业、收入和消费支出的周指标(Chetty et al., 2020)^⑦。借公共服务而提升自身品牌价值,也是高端企业精明的垄断谋求(以趋势引领的名义)之道。

上述两个方面,充分地表现出政府“数据基础结构”谋求相应变化的必要性。那么,究竟如何应对这种“数据格局”演变带来的压力?怎么探索一种可行的路径?

Abraham教授等探索者注意到“可替代数据来源(alternative sources of data)”,从而可以让经济统计机构更好地满足用户需求。近年来,“原本数字化的数据(natively digital data)”具备了改进经济统计的潜力。它们主要包括三个主要类型:(1)来自零售扫描的详细交易数据(detailed transactional data from retail scanners)、公司内部(信息)系统(companies internal systems)、信用卡记录(credit card records)、银行账户记录(bank account records)、工资单记录 payroll records)、保险记录(insurance records),这些基础数据都基于私人商业目的而编辑;(2)由传感器

和移动装置自动记录的“痕迹数据”；(3)从网站和社交平台上获取的各种各样的数据。

统计方式变革的可行性还表现在,与20年前相比,使用大数据的现代数据科学方法取得了长足进步,使得更系统地将大数据纳入官方统计成为可能(徐宗本等,2022;朱建平等,2019)^[6,7]。

如果将这些“非设计大数据来源(non-designed Big Data sources)”结合到“经济测度基础结构”之中,有希望使得统计机构生产更精确、更及时和更多的“非总量统计(disaggregated statistics)”,同时降低“基础数据提供者(the data providers)”的负担,甚至降低统计机构自身的负担(罗良清,2019)^[8]。从基础建设的角度看,数据新来源的可用性为官方统计“基本架构(the underlying architecture)”的再设计提供了机会。

五、大数据与经济统计结合的探索与启示

联邦经济统计如何利用大数据,在美国是一个较早就进行探讨的课题。“收入与财富研究会(CRIW)”早在2000年就召开过主题为“扫描数据与价格指数(scanner data and price indexes)”的研讨会,专门研究利用零售交易数据进行价格测度。但由于种种原因,^②大数据与经济统计的结合研究进展比较缓慢。经历了20多年的努力,才召开第二次相同主题的研讨会,并发表了一系列工作论文(working paper)。

美国“收入与财富研究会(CRIW)”与“世界收入与财富研究联合会(IARIW)”不同,通常只是召开闭门会议,并不公开征集论文参会。这次特邀了荷兰统计局(Statistics Netherlands)局长(Director General) Tjark Tjin-A-Tsoi 博士做大会报告,报告了该局在统计项目实施中所取得的惊人进步,即用大数据来源补充和替代传统调查。

然而,大规模地充分实现大数据对经济测度的潜在助力,目前还只是一个愿景。^③尽管还有不少事项和挑战需要解决和应对,但与参会者对这种期许还是充满了热情。实验性研究的可行性并不等于常规统计中的可行性,所以,Abraham教授在文中专门提及“大规模地”实施,是比较清醒的态度。

2020年研讨会的论文给出一个信号,大数据与经济统计结合的时机已经来临。所谓时机成熟,恰是变革和交叉、结合的必要性和可能性兼备。实际上,考虑到现有测度模式因“调查响应率”下降而受到的威胁、业已增大的调查成本、追踪“速变经济”的困难日益加大,以及其他类型行为主体的新近数据生产

作为,统计系统架构的基本改变将是必要的,否则就无法保持官方经济统计的质量和效用。

Abraham教授等明确指出,统计机构别无选择,只能从事这种艰难工作,并进行相当数量的必要投资,^④从而将本项研究所提出的“测度观(measurement approaches)”和不同“类型数据(types of data)”结合到官方经济统计的日常工作中。

我们应该充分注意到,这个信号对新兴国家具有格外的意义。这是一桩相当“耗时(time consuming)”的基础工作,对这种必须补课、不可躲避或放弃的基础性工作(慢活),还是需要及早开展,因为这种基础结构的构建不大可能靠短期突击来完成,也不大可能搞跨越式的发展。

反观经济统计现代化的历史过程,还有一个非常深刻的教训。对大数据技术的引进,不能浅尝辄止,不能仅仅停留在方法技术的掌握上,而应该真正与经济统计实务相结合,要在宏观经济分析中取得实效(肖红叶等,2022)^[9]。掌握了大数据技术,并不等于现实经济统计实务中就会用了,“套用”当然比较简单,有现成数据“套”进模型运算,就可以得出数据结构(张维群,2019)^[10],具备数学优势的中国学生对此不在话下。然而,实地操作需要的却是“应用”能力,达成这个“应”需要综合素质,需要脚踏实地的长期实践过程。如果以为“套用”便是“应用”,那么动起手来往往沦为“形式化动作”。

典型的例证便是投入产出表的编制和应用。自上个世纪80年代开始,我们学习、引进并试编了中国数据的投入产出表,现在本土也有学者在投入产出方法论研究上走在世界前沿。然而,中国编制投入产出表的分类详细程度却并不理想,40多年过去了,到目前为止分类程度仅达到150多个产品部门,而发达国家如美国和日本,投入产出表可以达到500多个部门,据说美国经济分析局的工作用表可以达到1000多个部门,可见,中国在此方法应用和核算实务上还存在相当的差距。

相比较而言,目前应用的投入产出技术并非特别复杂,而其实践价值恰恰在于分类程度详尽的实际数据,只有将基层的投入产出数据关系揭示出来,才能切实为宏观经济分析提供可靠的定量基础。投入产出核算是国民核算各个分支核算系统的基础,只有具有详尽的投入产出表数据,才能将生产过程中的中间消耗核算清楚,让GDP数据真正成为增加值的总和;才能编制出高质量的资金流向表,将实物与资金的部门间流向乃至存量关系核算得更为清

楚;才能采用“增加值法”计算不同国家间的进出口差额,进而明确国际收支关系,力求避免“全值法”计算带来的严重偏误。

2018年美国对中国发动贸易战,借口便是一个小小的部门经济统计指标——两国间巨大的贸易差额。然而,这是基于“全值法”口径数据做出的判断(邱东,2021)^[9]。中美两国间确实存在着贸易差额,但其规模究竟多大,还是一个需要深入探讨的经济统计疑题。如果我们平时就具备比较详尽的投入产出表实际数据,就能够对两国贸易差额的真实规模给出更具可靠性的数据,谈判过程中就可能取得更为主动的地位。

应该看到,投入产出表实际数据的分类并不是想做细就能做细的,这种细化需要一个历史过程,需要扎实的产业统计基础,需要实实在在的人力和资金投入。虽是“用兵一时”,但需要“养兵千日”,这种基础性工作很难取巧躲开,平时的积累功夫如何,关键时刻立刻显现,这便是我们呼吁“尽早补课”的根本原因所在。在将大数据与经济统计相结合的变革过程中,过去累积下来需要弥补的此类工作也应该一并加以考虑。

六、经济统计变革与时局的高度相关性

混合新数据来源以实现“联邦数据基础结构”的现代化,不仅可以从一般趋势来理解,而且突发的危机也会激发人们的认知,此次新冠疫情(the COVID-19 crisis)就使得这个议程变得极为重要。在危机应对的综合考虑中,及时和可靠的数据不可或缺。

私人研究者结合大数据生产的有关位置和活动类型的高频信息,引发了社会强烈的兴趣。例如,在采用集体行动(aggregate activity)应对公共卫生危机时,来自智能手机的“近实时位置数据(near-real-time location data)”提供了重要的参考与依据。不过,尽管开放新数据来源非常有价值,但这些“测度努力(measurement efforts)”彼此之间并没有合作,它们抓住了流行病经济影响的某个特定方面,但并没有提供一个综合画面。

统计机构也敏捷地对危机做出反应,例如引入两个新的“脉动调查(Pulse Surveys)”,提供居民户和小企业对危机做出反应的重要信息。^[9]除此之外,普查局(Census)基于行政数据发布了每周商业创新的新测度。劳工统计局(BLS)在现行就业和居民户调查中添加了问题,以了解商业运作如何变化以应对危机。

遗憾的是,对实时、细致的经济测度而言,统计

机构使用大数据只是一种“新生状态(a nascent state)”,基于稳健的、具代表性的大数据的“关键官方经济统计的常规生产的基础结构”尚待构建。Abraham教授等希望,当美国经济经历下一次危机时,统计机构能够应用“大数据的即时数据流(the ongoing flow of Big Data)”提供及时的、综合的信息,以帮助和指导政策制定者做出重要决策。^[10]

值得注意的是,在新冠疫情爆发前该项研究已经基本完成,所以各位专家所提交的15篇论文中并没有涉及此次流行病爆发的相关内容。但是在序言(综述性论文)中,Abraham教授等对此议题做了比较充分的补充。重点指出的是,在该文的31项参考文献中,与新冠疫情直接相关的就有8项,即2020年发表的专题论文。可见,美国学者和部分官员对保持经济统计的“相关性”高度重视,具备应有的职业责任感和时事敏感性。相比而言,中国统计学者也有不少对疫情发生状况的统计描述,但是在疫情对整个社会经济和福利影响的“即时描述”和“即时预报(nowcasting)”方面,还有待于深入拓展和提升。

七、大数据与经济统计如何结合——方法论挑战与推进

Abraham教授等强调,挖掘大数据提升经济测度潜能的一个主要障碍是,缺少完备的方法论,即如何将新型数据结合进入“测度基础结构”中。

在许多文本中,大数据应用采用“有监督的机器学习方法(supervised machine learning methods)”。在典型的应用中分析者处理两类观察数据,一类是符合“金标准测度(a gold-standard measure)”^[10]的相关数据,另一类是自信能够用于预估其他样本的大数据。通常的操作是将可用的数据分成三类,“训练数据集(a training data set)”用于估计大数据模型,“确认数据集(avalidation)”用于模型选择,“测试数据集(a test data set)”用于评估模型外样本的表现,后两类数据有助于应对“过拟合(overfitting)”问题。然而在具体应用中,还需要进一步研究“如何应用”或“如何结合”的方法论,以真正实现交叉和融合。所谓交叉和融合,一定是双向或多向的,不会是从工具到事例的单向动作,一定是一个过程,不可能一蹴而就。标准机器学习方法不能简单地引进和应用,Abraham教授等做出这种判断是基于一个事实,即适用于经济统计生产的大数据只是到了近期才变得可利用。其隐含的逻辑也很简单,毕竟大数据方法兴起已经多时,如果引进和应用非常简单,那么早就应该在经济统计生产过程中发挥效用了。

第一个方法论研究案例由经济分析局 Chen 等做出,在测度“个人消费支出(Personal Consumption Expenditure, PCE)”时,需要预测“季节服务调查(the Quarterly Services Survey, QSS)”系列数据。由于完整的系列数据需要等到该季后两个半月才能得到,而且会导致 PCE 数据的重大修订,因此 Chen 团队将信用卡和谷歌流行数据作为 QSS 系列数据的隐含预测因素,结合到 PCE 的早期估计中。在估算方法上他们采用“多元模型观(multiple Modeling approaches)”,不仅使用移动平均和回归模型,还使用不同的机器学习方法。然而,“有监督机器学习方法”隐含着对“样本长度”的要求,Chen 团队能拿到的 QSS 数据只有 31 个季度,再区分出训练数据集、确认数据集和测试数据集不大可行。根据数据可得性进行调整,Chen 团队采用 19 个季度的数据拟合多种模型,模型采用不同来源数据的组合、可变的选择规则和算法,然后比较余下 12 个季度 QSS 数据的预估与实际结果,籍此评价模型的表现。这种操作背后的直觉是:如果数据结果所表现的一致性较强,则遭受“过拟合”问题的可能性就较小。Chen 团队此项研究的结论是,与现行操作相比,随机森林等“集成方法(ensemble methods)”极有可能减少 PCE 数据修正的规模。

第二个方法论研究案例涉及消费者支出测度中对扫描数据的调整,挑战是如何从周数据中区分“季节和日历效应”的影响。与专项调查相比,扫描数据可以提供更为及时和丰富的信息,然而,这些数据并不能直接搬用,重大事件(如节假日等)在不同年份处于不同的周,从而会对消费者支出数据产生不同影响,而且这些影响在不同地区还可能不同。除非这种影响得到适当的调整,否则基于扫描数据的消费测度变化就难以适当解读。

本来经济时间序列中剔除“季节和日历效应”有很多方法,但这些方法的使用都是有条件的,需要大量的时间序列基础数据。并且,即便具备足够长度的数据,开发特定的调整模型也将是资源集约型的,当数据系列的数目非常多时不大可行。

Guha 和 Ng 两位教授分步骤、分类地剔除季节的影响,先是逐个剔除数据系列中“特定季节变动”的影响,然后剥离数据中的跨部门相关性因素,以剔除“共同性残余季节影响(common residual seasonal effects)”,之后使得“解释变量”(“年中日期”、“月中日期”、县域人口等)能够影响每个不同类别的支出。使用机器学习方法,Gu 和 Ng 两位教授从县域支出

系列的分类数据中剔除了两类季节变动影响,使得剩余的估计值能够用于追踪消费者支出的趋势和周期变动,而且是在“地域分量水平(a geographically disaggregated level)”上对“详细支出类别(detailed expenditure categories)”做出分解工作。

第三个方法论研究案例涉及扫描数据给经济测度带来的又一个问题——如何编制价格指数,以适当地测度商品出现和消失对消费者福利的影响。以橙汁数据为例,Diewert 和 Feenstra 两位教授比较了当下文献中针对此问题提出的几种实证方法。从理论背景看,使用“效用函数(utility function)”颇有吸引力,因为其与“费雪价格指数(Fisher price index)”相一致。然而从实践场景看,假定“固定替代弹性效用函数(a constant elasticity of substitution utility function)”更容易得出估计。试算结果表明,基于后者可能严重高估引入新产品的消费者福利获益。一个可能的结论是,假设“一个更为灵活的超越对数支出函数(a more flexible translog expenditure function)”将更好地在精确性和易处理性之间取得平衡。

应该注意到的是,上述三个案例仅仅是列示性的,大数据与经济统计如何结合的方法论研究应该全过程开展(耿直,2014)^[1]。

在论及部门测度时,Abraham 教授等指明了方法论研究的难点,例如使用“等值租金法(rental equivalence approach)”估算自有房屋服务价值时,尽管有了 Zillow 数据,但单位市场租金与自有住房服务价值切实可比仍是确保该方法可行的关键,需要对此机理做出深入阐释。“使用者成本法(a user cost approach)”在大数据条件下仍然举步维艰,需要估算的影响因素较多,特别是“住房服务实际预期资本获益(the real expected capital gain on housing services)”难以详细测度。在住房价格繁荣与萧条的周期中,所观察到的价格波动并不能反映住房服务实际价值的变化,Abraham 教授等指出,很难想象将基于交易数据的房价估算纳入“国民收入和生产账户(the National Income and Product Accounts, NIPAs)”。

八、研究项目总结:推进经济统计中大数据应用的挑战和应对措施

序言的这一部分不直接针对所收录的论文,而是对研究项目做出总结性阐述。推进大数据在经济统计中应用时,所面对的主要挑战有哪些?如何应对? Abraham 教授等主要列示了七个方面。

第一,不同类型数据间的匹配性。就联邦机构与私人数据提供者成功建立合伙关系所学到的经验来

看,其中之一就是应该接受大数据并非一种数据提供者以预先定义方式构建的存在。这意味着,这类数据原本并不打算用于统计分析,因而统计机构需要机敏待之。如本论文集所示,存在几种为商业和行政管理目的而生产信息的方式,它们并非直接与测度相匹配,就统计目的而言并非即刻有用。这种非匹配性主要表现在概念、分类、时间和纵向一致性上。

概念(数学中的定义域)上,从企业和“居民户”数据中产生的变量通常与官方统计所内嵌的经济和统计概念不相匹配。不能认定统计调查的反馈都完整或正确,但与大数据结合时,统计机构需要找出路径将所引进的数据纳入所期望的“测度构建(measurement constructs)”中。论文集论文面临着这样的议题:将自然[®]发生的大数据演变为与“经济统计范式(the paradigm of economic statistics)”相匹配的变量。

分类上,商业目的编码数据可能未必与官方统计生产所要求的类别相对应。例如,扫描数据包括了产品水平的价格信息,但为了符合CPI项目的需要,其各个项目必须与CPI所公布的类别相对应。

时间匹配上,存在许多与观察值实际时间间隔相关的复杂难题。例如,周销售数据并非与月度和年度数据相匹配。再例如,工资单数据的支付期可能与期望的日历期相异,劳工统计局的“居民户”和基本单位调查要求回复者给出“参考期(a reference period)”,将实际状况与官方统计匹配的责任加之于调查回复者,但利用大数据时,这种责任就回到了统计机构身上。

纵向一致性上,作为内部过程结果的数据在不同时期未必一致,这意味着数据缺乏连续性,需要统计机构进一步加工。即便观察值在分类上前后一致,单位营业额或产品营业额也可能在利用大数据时面临重大挑战。测度名义销售额或消费总额在概念上比编制价格指数容易,后者需要将名义值转换为更具基本意义的实际数量。产品营业额在价格指数编制中引起了特别困难。当价格指数样本中选中的某些产品不再可用时,劳工统计局开发了一套“产品替换(product replacement)”的方法,然而当采用扫描数据编制价格指数时,这套办法就失效了,因为扫描数据包含了成千上万的独特项目。经济学家对此展开了深入探讨,处理持续的产品周转额需要新方法,即能够利用“支出模式(expenditure pattern)”去推断消费者在产品间替代的意愿。

第二,大数据对统计机构数据供给的机制安置。

到目前为止,利用大数据促进经济统计所做的研究还是开发性探索,或者评估其可行性,或者弥补数据缺口。在推广大数据的经济统计利用时,必然会涉及到数据供给机制的一系列安置问题。(1)如果基于条款有限的协议,系统外提供数据的稳定性和连续性就难以保障;(2)有的数据提供者更愿意提供一个更大的数据文件包,以减轻其数据整理的负担;(3)有的数据拥有者包括公共网站,明确排除了政府机构对其数据的使用权;(4)如果要扩展大数据的利用,统计部门还面临着预算约束。有的学者建议,现行法规需要做出改变,以鼓励私人部门的数据用于统计目的;还有的提出,由第三方数据提供者替其客户向联邦统计机构报告,采用市场化服务的路径。

第三,数据供给的连续性问题。宏观经济统计的一个本质特征是“常规性”,与典型调查或一次性研究不同,宏观经济统计内在地要求其数据结果的系统一致性,这决定了其实施过程的标准化处理,也就要求相应的基础数据在供给上保障连续性。

如何保持统计机构现行数据供给模式的可持续性,这是一个开放性问题。主要原因在于:(1)调查响应率的急剧下降,或“非强制性调查”难以得到回复;(2)大数据可能以前后非一致性的模式出现,甚至不再可用,破坏进行宏观统计估计所需要的连续性;(3)意外事件如COVID-19危机,将中断计划安排的调查数据搜集,或减弱关键行政管理数据的及时性;(4)统计机构与单一数据提供者的合同可能遭遇数据供给的“逼停问题(a hold-up problem)”,或者关键时刻的数据要价过高。

相比而言,大数据由于其商业机制支撑,破坏数据供给连续性造成的损失较小,而“滚动多年合同(rolling multiyear contracts)”有助于缓解数据断供。扩展可替代数据源并签订“滚动多年合同”,则有助于缓解“数据供给逼停问题”。

第四,统计机构间合作及其与企业合作的整合。挖掘大数据潜能的中心挑战在于,统计机构间的合作及其与企业合作的构建方式。在大数据时代,三大统计机构(普查局、劳工统计局和经济分析局)需要采取更为一体化的数据采集和加工过程,统计机构与企业合作也应该采用标准化的多目的数据使用协议,促进谈判,减少分歧,以促进大数据的利用。大数据的发展使得美国统计机构重组出现转机,而法律改变在没有重组的议程中需要优先考虑,以便更有效地用大数据改善经济统计。

第五,统计数据细化与隐私保护。大数据为经济

统计提供了生产“细致统计量(*granular statistics*)”的机会,然而详尽统计也导致私人和企业隐私保护上相应的困难,大数据给经济统计带来了最大的吸引力,也带来了最大的挑战,统计从私人数据集得到的结果越精确,个人隐私流失就越多(Dinur and Nissim,2003)^[12]。“形式化方法(*formal methods*)”如“差别隐私(*differential privacy*)”,允许数据发布者在隐私保护和数据效用之间做出精确的选择。然而在相互竞争的目标之间必须恪守平衡(Abowd and Schmutte,2019)^[13],普查局采取“差别隐私”作为隐私保护的方法论,^④从其2020年普查产品所产生的争议看,就“何者为适”达成一致是一个困难的过程。采用现代泄密防护技术,就可能生产出足够精确且能保护隐私的数据产品,可以期望在效用和隐私之间达成可接受的妥协,发布从详细交易、地理位置和其他敏感来源计算出来的新经济统计产品,更广泛地被数据用户接受。

第六,经济统计的“计算基础结构”。尽管此议题并非本论文集的主题,但“计算基础结构”对经济统计的改进也非常重要,这涉及到“安全的多方计算”“云上计算基础结构”等。在大数据时代,持续改进和投资于现代计算能力是成功的必要条件。此外,还有统计机构人员的业务能力培训,职员具有不同的学历背景,如统计学、经济学等,需要补充数据科学知识,否则大数据与经济统计的高质量结合就无法达成。

第七,机构间合作进行统计方法论研究。与学术界和其他研究机构的合作关系,对统计机构而言非常重要。“国家科学基金会普查研究网(*the NSF-Census Research Network,NCRN*)”将普查局与数所大学联系在一起,收入与财富研究会(*CRIW*)和国家经济研究局(*NBER*)长期以来是学术界与统计机构合作研究“测度问题(*measurement issues*)”的纽带。^⑤新近“全国企业经济学家联合会(*National Association of Business Economists*)”主持“技术经济学研讨会(*the Tech Economics Conferences*)”,将经济学家、数据科学家、公共部门和私人部门的相关人员联系在一起,合作研究数据与方法的应用。

九、大数据与经济统计“如何结合”的方法论研究是持久性议程

大数据相关企业及其业务的兴起,是经济统计开辟新数据源的现实基础。除了前文列举的数字化数据资源外,序言中还列示了十多项,如Nielson and the NPD Group的Retail transaction data、JD Power

的New Vehicle sales、Gas Buddy的a crowdsourced website、First Data的Credit card transactions data、The Kilts Center的Retail scanner data,还有The Yelp API、The Zillow data、Quick Books、Retailer等。只有新数据源发展到一定程度,大数据与经济统计的结合才能真正步入实践过程。

在大数据与经济统计的结合中,应该对可替代数据源的属性做出解释和说明,Aladangady等采用“过滤程序(*filters*)”控制数据库中特定项目的纳入和排除,这是“非设计数据(*non-designed data*)”可用于宏观分析的前提之一,^⑥这种数据“预处理”势必将估算过程复杂化。还需深入思考的问题是,过滤程序的使用能否加以一般化推广?

此外,私人信息产品纳入公共信息产品,不仅在宏观意义上可能“缺乏代表性(*lack representativeness*)”,例如外贸与内贸所能借鉴的外部数据资源就大有差异,而且“非设计数据”的质量也需要慎重考察。如果微观数据被用来做宏观测度和管理,就不能排除人造数据嵌入网络的可能。所谓网络痕迹未必自然发生,好多往往只是采用了一种客观或中性的包装。在信息生产中,企业和个人有意和无意的偏误都将产生数据的道德风险。进而需要思考的问题是:企业公布公共信息是否需要市场进入的准许?是否需要公共信息产品的“特许经营权”?是否需要政府机构对其数据生产过程进行监督?是否需要对其质量进行检验?是否需要制定公共信息产品标准?

从社会实践的基本事实和基层逻辑出发,进一步思考测度对象的定义域,这便是经济统计过程对经济科学基本概念的理论反馈作用(李金昌,2014)^[14]。例如,论文集提供了一个非常重要的信息,即美国也存在中央和地方不同层次的经济统计信息需求。原来我们存在误解,美国政府不干预经济,尤其是地方政府,并不需要统计信息。论文集第5页专门提到州政府与联邦政府在信息需求上的差异,如何解决这个矛盾,兼顾不同政府层级的信息需求,也是美国经济统计中一个需要深入探讨的问题。其实地方政府主要负责教育和卫生事务,这是广义经济范围中的重要内容,特别是SNA拓展了不同领域的卫星账户后,地方对信息的需求更为突出,例如Glaeser等论文中提到的官方“县域商业模式统计(*County Business Patterns statistics*)”。^⑦

无论在哪个现实领域、纳入何种“非设计数据”、采用何种数理模型,关键在于大数据与经济统计结合工作的机理何在(邱东,2014)^[15]?即系统地回

答:为什么可以这样做?如果只是“聊胜于无”的改进,那么就需要交代:究竟改进了多少?所谓改进是否可能带来新的缺陷?“此时此地”的改进是否可以用于“彼时彼地”?为什么此项改进在其他现实场景也具备推广价值?这些应该是方法论研究的主要内容,如果对这些没有系统的回答,大数据与经济统计结合的时机究竟是否成熟,可行性是否切实、全方位地具备,就值得进一步考虑。而且,再做深入的方法论研究会涉及到科学哲学层次的思考(邱东,2023)^[6],可见,大数据与经济统计“如何结合”的方法论研究是持久性议程。

十、经济统计与大数据交叉、结合研究的学科意义

从表1可以看出,关于大数据与经济统计相结合的方法论研究项目,其人力资源分布为:美国大学参与这个项目研究的有22位,占55位全部参与人员的40%;三大政府统计机构中普查局9人、劳工统计局5人、经济分析局8人,是此项研究的主体,加上美联储的研究人员9人,一共31人,占了56%以上;企业研究人员中,有3位参与了这项研究。

值得指出的是,大学参与人员中,有10位还特别标注为“国家经济研究局NBER研究员”,在欧美经济学界,这是一个非常令人尊敬的学术头衔。可见,大数据如何与经济统计相结合,是经济学领域知名的大学教授和政府部门中的学者型官员共同努力(合作并争辩)的结果,这是美国经济统计发展一个悠久的传统。在这些研究者中不乏全球知名的学者,经济统计研究虽然颇为繁琐,但并非等而下之。

“官方经济统计(official economic statistics)”这个概念屡屡被提及,是论文集序言的主题词。正如国际统计学学会佐尔坦·肯尼西先生在1994年所言,“官方统计同样是一个激动人心的学术领域”(邱东,2013)^[7]。这部论文集告诉我们,在探讨大数据与经

济统计结合的过程中,美国学者都做了些什么?因为他们是先行者,当然也指出了领域发展的趋势:这种结合究竟应该做些什么?当下能够做些什么?有哪些经验值得借鉴?有哪些教训值得吸取?还有哪些空白需要弥补?

论述21世纪经济统计的发展,当然必须以大数据、人工智能和数字经济为大背景(洪永森、汪寿阳,2021)^[8],必须进行交叉研究,必须将现代技术与经济统计相结合。但同时需要明确的是,在社会经济事务的定量研究中,学科间交叉和结合并不仅仅是且主要不是计算机科学,并不仅仅是且主要不是数理统计,而是以经济统计为主,来探索三者之间如何交叉、如何结合。学科间交叉和结合决不是取代经济统计(邱东,2021)^[9]。

请那些只讲数理统计的“国际接轨者”看看,国外统计存在的这些学术成就是否也应该引起足够的重视,否则将来如何培养能够与国际统计机构讨论方法与制度的专业人才。经济统计学当然需要与时俱进,但并不是搞历史虚无主义,并不是另起炉灶。如果确实需要“新经济统计”的话,它应该是什么样子,美国研究者对该专业“基础结构”的探讨不可忽视,毕竟人家已经探索了20多年,无论是经验还是教训,先行者都为我们指点迷津。

不管数理工具如何发达,都不能丢了经济统计主业。在新世纪的经济统计专业教育中,只讲数理统计课程,再补充一些计算机课程,是一种课程体系的极偏态分布,严重背离了经济统计的专业本意,无法培养能够切实建设“官方经济统计基础结构”的专业人才,无法提升和培育全社会的经济统计意识,对中国官方经济统计的质量提升危害极大,从而也会影响到国家高质量发展的进程。

注释:

① 由北京师范大学王亚菲教授在2021年推荐给笔者。

② 本文对这15篇论文的述评基于该文集的序言,由于拿到纸质版文集的时间不长,又希望快些推介这部非常重要的著作,笔者尚未来得及仔细阅读每一篇论文。因此,涉及到各篇论文的述评实际上假定了序言对各篇论文的述评正确,且笔者对序言述评的理解正确,至于述评质量究竟如何,还需要进一步的研究工作。

③ 现代生产测度的一个重要方法是生产率测度,Abraham教授等在序言中将其单独列示,与三大关键经济统计(key statistics)并列。

④ 安格斯·迪顿教授曾质疑过问卷响应的质量,应该也是已有“数据格局”所隐含的问题。

⑤ 基础数据的采集、加工和监管。

⑥ 其影响早就超越了国界,而是全球化的。

⑦ 原文没有列示其详,是否也有难言之隐?

⑧ 应该特别注意荷兰与美国在经济体量上的差异,荷兰的经验在美国未必适用。

⑨ 西方政府表面上没有发展计划,但却有很多项目预算,这实际上是具有资金保障的真正计划,所谓自由市场经济的标签,值得深入辨析。

⑩ 江西财经大学李晶博士对这两篇文献做了述评工作,可供追踪参考。

⑪ 注意,这里确实是在讲,统计机构用信息指导政策制定者做决策,专业指导的地位不容小觑。

⑫ 即基于调查和普查数据的估计。

⑬ 原文如此论述,但应该看到,大数据中的数据未必都是自然发生的,对其所谓的客观性议题还需要深入探讨。

⑭ 原文就是这个提法,就连“隐私保护”都存在方法论问题,可见笔者强调方法论研究是必要的。方法是具有生命力的,自身有成长过程,方法总是需要思考与应用场景如何匹配,因而需要持续地去研究,即需要方法论。

⑮ 本论文集(第79卷)就是一个典型证明,NBER的收入与财富研究系列(Study in Income and Wealth)尤其值得关注。

⑯ 参见论文集中 Aladangady 等撰写的第4篇论文。

⑰ 其实,联邦预算中的经济项目落地在哪个州,也是国会议员最为关心的问题。美国搞三权分立,对其政府概念要做广义理解,也就是说,应该切实领会 SNA 中“广义政府(general government)”的概念。国会是政府的重要组成部分,法律干预也就是政府干预,不能将政府与行政分支混为一谈,也不能以为只有行政干预才是政府干预。当然,美国白宫对经济事务也是直接插手的,最为典型的例证就是他们死命打压中国的华为公司,根本不顾及自由市场经济的信条。

⑱ 深究起来,美联储的地位非常特殊,虽然自称非政府机构,却实质性地起到了美国中央银行的作用,等于是美国政府把公权力让渡给了金融资本。更有甚者,由于美国在全世界已然具有超级帝国的实力地位,这又等于美国政府替全世界做主,把国际经济关系的决定权让渡给了美国金融资本。由此,经济统计不仅仅是美国联邦政府进行国内宏观管理的得力工具,也成了美国实施全球独裁“治理”的得力工具。

⑲ 广义的学者概念,包括政府部门和企业中的工作人员。一个人是否为学者,应该以其行为为标准,而不是以其头衔和工作单位为标准。

[参考文献]

[1] Abraham K G, Jarmin R S, Moyer B, et al. Big Data for Twenty—First—Century Economic Statistics[M]. Chicago: University of Chicago Press, 2022.

[2] Rockoff H. On the Controversies Behind the Origins of the Federal Economic Statistics [J]. Journal of Economic Perspectives, 2019, 33(01): 147-164.

[3] 邱东. 数据科学在社会经济领域应用的重心——兼评《十字路口的统计学, 谁来应对挑战》[J]. 计量经济学报, 2021, 1(02): 250-265.

[4] He X M, Madigan D, Yun B, et al. Statistics at a Crossroads: Who is for the Challenge [R]. Arlington: National Science Foundation, 2019.

[5] Chetty R, Friedman J, Saez E, et al. The Determinants of Income Segregation and Intergenerational Mobility: Using Test Scores to Measure Undermatching[R]. NBER Working Paper(w26748), 2020.

[6] 徐宗本, 唐年胜, 程学旗, 等. 数据科学——它的内涵、方法、意义与发展[M]. 北京: 科学出版社, 2022.

[7] 朱建平, 谢邦昌, 马双鸽, 等. 大数据: 统计理论、方法与应用[M]. 北京: 北京大学出版社, 2019.

[8] 罗良清. “大数据时代”的数据产品生产[C]. 对外经贸大学“统计发展与创新”研讨会, 2019.

[9] 肖红叶, 杨贵军, 尚翔. 数据技术应用概论[M]. 北京: 科学出版社, 2022.

[10] 张维群. 大数据对统计科学的影响及其应用实例[C]. 对外经贸大学“统计发展与创新”研讨会, 2019.

[11] 耿直. 大数据时代统计学面临的机遇与挑战[J]. 统计研究, 2014, 31(01): 5-9.

[12] Dinur I, Nissim K. Revealing Information While Preserving Privacy [C]. Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, 2003: 202-210.

[13] Abowd J M, Schmutte I M. An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices [J]. American Economic Review, 2019, 109(01): 171-202.

[14] 李金昌. 大数据与统计新思维[J]. 统计研究, 2014, 31(01): 10-17.

[15] 邱东. 大数据时代对统计学的挑战[J]. 统计研究, 2014, 31(01): 16-22.

[16] 邱东. 格局·相关性·方法论——领域应用中统计研究的聚焦[J]. 计量经济学报, 2023(01): 30-50.

[17] 邱东. 经济统计学学科论[M]. 北京: 中国财政经济出版社, 2013.

[18] 洪永淼, 汪寿阳. 大数据、机器学习与统计学: 挑战与机遇[J]. 计量经济学报, 2021, 1(01): 17-35.

[19] 邱东. 基石还是累卵——经济统计学之于实证研究[M]. 北京: 科学出版社, 2021.

[责任编辑: 李莉]