大数据时代与经济史计量研究*

陈争平

内容提要: 经济史计量分析大致有统计学、计量经济学、计量史学三大类方法。大数据时代来临,让我们更清楚地看到了样本无法揭示的细节信息,以统计为基础的经济史计量研究比小数据样本加数量模型更有效,统计学派更加贴近大数据时代的主要特点,更符合时代要求。我们现在进行数据库建设,既要注意数量,使得规模尽可能大; 又要抓好质量,要建设经得起检验的数据库。经济史计量研究也可以帮助创立经济学的新的论点。

关键词: 大数据时代　量化历史　计量史学　统计学　经济史

目前世界上正在开启的大数据时代,正在升温的量化历史研究都与近代中国经济数据库建设有着密切的关系,本文就此做些探讨。

一、为何要抛弃"计量史学"?

近年来国内关于量化历史(Quantitative History) 的研究正在升温,著名经济学家陈志武先生主办了四届"量化历史讲习班"以培养从事量化历史研究的青年学人,《中国经济史研究》则开辟有关专栏,努力推动这方面的研究。笔者在20 世纪80 年代初攻读硕士研究生时受吴承明先生启发,对计量史学(Cliometrics) 做过一些探讨,后来在清华大学任教时也开设了《历史统计与计量史学》的研究生课程。那么,20 世纪80 年代开始在国内流行的"计量史学"与近年来国内流行的"量化历史"有何区别? 笔者曾就这个问题请教李伯重教授。李伯重教授对国外学术动态有较多了解,他认为"计量史学"与"量化历史"就是一回事。笔者后来又进一步对 Cliometrics 与 Quantitative History 的定义及发展历程等进行检索,发现二者确实是一回事。

从名称上看,Cliometrics 比 Quantitative History 更简洁更有学术性。那么,近年来推动量化历史研究的一些学者为何要抛弃 Cliometrics 这一名称而宁愿采用 Quantitative History? 联系到吴承明先生指出计量史学"只曾盛行于美国。在欧洲虽有短暂反应,但不成气候",'在中国则无响应",而在美国"进入21 世纪,计量史学已消失生气"。①想来由于计量史学发展业绩不好,名声坏了,所以后来那些学者宁愿抛弃"计量史学"而改用"量化历史"(Quantitative History) 这一名称。

总结以往计量史学发展业绩差的原因,对于更好地开展经济史计量研究有重要意义。笔者在清华大学为《历史统计与计量史学》备课时,曾经对以往计量史学发展业绩差的原因做过一些分析,可概括为以下四个方面。

第一个方面是以往一些计量史学方法的倡导者过分夸大了历史数据的客观性及代表性。细考中国历史上一些数据来源,往往会发现它们来自于某个官员或士子的估算,后来又有一些研究者再根据这些估算作进一步推论,使得结论的主观性更强,客观性更低。一些中国近代农史研究者推崇

[作者简介]　陈争平,山东大学经济研究院教授,济南,250100; 清华大学人文学院教授,北京,100084。

*　本文为国家社科基金重大项目"中国近代经济统计研究"(批准号: 12&ZD149) 的阶段性成果。

①　吴承明《经济史: 历史观与方法论》,上海财经大学出版社2006 年版,第242 页。

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

民初卜凯的调查，实际上这一调查在地区的选择、指标的规定等方面都有较大主观性。① 卜凯所用调查人员多是年轻学生，他们一般出生于富足人家，所以才能上大学，回乡调查也多是问自家长辈和管家等，有关数据就会偏向富人，对于当时农村总体而言代表性较差。卜凯的著作中提到的贵州遵义平均单位面积产量，大大高于另一外国教授(M. N. Jen) 在实地调查中得到的数字。该教授认为，造成这种较大差异的原因在于，卜凯的著作仅以优质土地为样本，而实际上这种土地在遵义的耕地中只占非常小的比例。②

第二个方面是夸大计量方法的作用，甚至断言用计量方法就能把历史学变成真正的科学。这种夸大不但不能提高真正的业绩，还会引起其他史学家的反感，导致计量史学一再遭遇质疑和批评，一些计量史学倡导者热情冷却后又回归传统叙事方法。笔者认为，计量方法只是史学走向科学的必要条件，而不是充要条件。把必要条件当做充要条件，就会使人狭隘，所得出的研究成果也会有偏差。计量方法仅是众多研究方法中的一种。正如吴承明先生所言: “研究经济史应根据不同对象和史料条件，采取不同方法。”③史学研究还是要走定性分析与定量分析相结合之路。

第三个方面是各种数量模型的应用都有各自的前提条件，以往一些计量史学研究不论时空差异，盲目套用模型，以致扭曲历史真相，甚至会得出一些荒谬结论。当然，也并非任何模型都不能用，要视具体情况作具体分析。

第四个方面是历史数据缺失，使得计量分析面临极大的史料困难。吴承明先生认为，在计量方法中，必须有连续十年的系列数据才能建立一个模型。在中国，这种连续十年的系列历史数据严重缺失，以致在 20 世纪八九十年代国内那些计量史学的鼓吹者自己也始终停留在鼓吹阶段，没有做出什么业绩，身体力行的只有吴承明先生。

既然“量化历史”原本是改名换姓的“计量史学”，那么导致以往计量史学发展业绩差的四个方面的问题，值得现在从事量化历史研究的学者们警惕。

二、大数据时代来临

近年来国内关于量化历史研究正在升温时，恰遇一个新时代来临。英国学者维克托·迈尔－舍恩伯格、肯尼思·库克耶在《大数据时代: 生活、工作与思维的大变革》一书中宣告: 大数据时代来临。④

随着计算机技术全面融入社会生活，信息爆炸已经积累到了开始引发变革的程度。在云计算技术推动下，一个大规模生产、分享和应用数据的时代正在开启。大数据很可能成为发达国家在下一轮全球化竞争中的利器，而发展中国家依然处于被动依附的状态。大数据建设在加强国家治理能力、国际竞争力等方面将发挥日益重要的作用。现代历史上的历次技术革命，中国均是学习者。而在这次大数据新变革中，中国与世界的距离最小，在很多领域甚至还有着创新与领先的可能。只要我们以开放的心态、创新的勇气拥抱大数据时代，就一定会抓住历史赋予中国创新的机会。⑤

大数据时代的精髓与三个重大的思维转变有关，这三个转变是相互联系和相互作用的，这些转变将改变我们理解和组建社会的方法。

第一个转变就是，在大数据时代，我们可以分析更多的数据，有时候甚至可以处理和某个特别现

① 梁方仲《卜凯〈中国土地的利用〉评介》，《社会科学杂志》第 9 卷第 2 期(1947 年 12 月) 。

② 梁方仲《卜凯〈中国土地的利用〉评介》，《社会科学杂志》第 9 卷第 2 期(1947 年 12 月) 。

③ 吴承明《中国经济史研究方法杂谈》，《中国近代经济史资料》1987 年第 6 辑。

④ [英] 维克托·迈尔－舍恩伯格、肯尼思·库克耶著，周涛译《大数据时代: 生活、工作与思维的大变革》，杭州: 浙江人民出版社 2012 年版。

⑤ 田溯宁《推荐序一: 拥抱“大数据时代”》，维克托·迈尔－舍恩伯格、肯尼思·库克耶《大数据时代: 生活、工作与思维的大变革》(中译本) 。

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

象相关的所有数据，而不再依赖于随机采样。生活中真正有趣的事情经常藏匿在细节之中，而采样分析法却无法捕捉到这些细节。与局限在小数据范围相比，使用一切数据为我们带来了更高的精确性，也让我们看到了一些以前无法发现的细节——大数据让我们更清楚地看到了样本无法揭示的细节信息。①

第二个转变就是，研究数据如此之多，以至于我们不再热衷于追求精确度。当我们拥有海量即时数据时，绝对的精准不再是我们追求的主要目标。当然，我们也不是完全放弃了精确度，只是不再沉迷于此。适当忽略微观层面上的精确度会让我们在宏观层面拥有更好的洞察力。②

第三个转变因前两个转变而促成，即我们不再热衷于寻找因果关系。在大数据时代，应该寻找事物之间的相关关系，这会给我们提供非常新颖且有价值的观点。相关关系也许不能准确地告知我们某件事情为何会发生，但是它会提醒我们这件事情正在发生。因果关系只是一种特殊的相关关系。相关关系分析通常情况下能取代因果关系起作用，即使不可取代的情况下，它也能指导因果关系起作用。大数据的相关关系分析更准确、更快，而且不易受偏见影响。③

大数据绝不会叫嚣“理论已死”，但它毫无疑问会从根本上改变我们理解世界的方式。很多旧有的习惯将被颠覆，很多旧有的制度将面临挑战。④

三、经济史计量研究中的三大学派

吴承明先生希望在有关经济史的研究中“凡是能够计量的，尽可能作些定量分析”。⑤ 吴老指出，定量分析可以检验已有的定性分析，以尽量避免随意的定性判断，它还可以揭示多种变量相互之间的内在关系，揭示经济事物发展变化趋势，可以使人们对许多历史问题的认识不断深化。

吴老指出，经济史计量分析大致有统计学、计量经济学、计量史学三大类方法。⑥ 他告诫我们，计量研究是一项要小心谨慎，要下苦功的工作，统计是经济史计量研究的基础。

对于计量经济学方法，吴老认为它可以用于“检验已有的定性分析，而不宜用它创立新的论点”。⑦ 计量经济学方法依赖于特定的经济学理论，而吴老认为至今仍“没有一个古今中外都通用的经济学”，“计量经济学方法用于经济史研究有很大局限性”。⑧ 他不主张用小数据样本加数量模型来研究经济史，还有一主要原因是数量模型里无“人”，看不见“人”的主观能动性。他曾批评说 “从司马迁起，写人物就是中国史学的优良传统。但近代史学，尤其是经济史，似乎丢掉了这个优良传统”。⑨

至于计量史学，吴老认为它“已消失生气”。所以吴老指出，经济史计量研究仍然“主要是统计学方法”。⑩ 实际上，从模型派、量化历史派已有成果看，他们所用的仍然是频率分析、回归分析等基本统计方法，主成分分析、判别分析与聚类分析等高级统计方法在史学界还很少有人用，更遑论灰色系统理论及 GM 模型的运用了。高级统计方法在中国史学研究中的运用，还有待年轻学者去努力实践。

---

① 维克托·迈尔－舍恩伯格、肯尼思·库克耶《大数据时代：生活、工作与思维的大变革》（中译本），第 1 章。
② 维克托·迈尔－舍恩伯格、肯尼思·库克耶《大数据时代：生活、工作与思维的大变革》（中译本），第 2 章。
③ 维克托·迈尔－舍恩伯格、肯尼思·库克耶《大数据时代：生活、工作与思维的大变革》（中译本），第 3 章。
④ 维克托·迈尔－舍恩伯格、肯尼思·库克耶《大数据时代：生活、工作与思维的大变革》（中译本），第 94 页。
⑤ 吴承明《市场·近代化·经济史论》，昆明：云南大学出版社 1996 年版，有关章节。
⑥ 吴承明《经济史：历史观与方法论》，第 242 页。
⑦ 吴承明《经济史：历史观与方法论》，第 248 页。
⑧ 吴承明《经济史：历史观与方法论》，第 214、215、219、221—224、282 页。
⑨ 许涤新、吴承明主编《中国资本主义发展史》第 1 卷，北京，人民出版社 1985 年版，第 12 页。
⑩ 吴承明《经济史：历史观与方法论》，第 242、250 页。

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

受吴老启发，对于经济史计量研究中三大学派之争，我们认为，统计学派更加贴近大数据时代主要特点，更符合时代要求，因为：1、大数据的“大”是相对而言，意思就是要分析与某事物相关的所有数据，而不是少量的数据样本。我们的国家社科基金重大项目《近代中国经济统计研究》工作就是要尽最大可能收集整理与近代中国经济相关的所有数据，在此基础上开展计量分析；2、《大数据时代：生活、工作与思维的大变革》一书中有这样的论断：“大数据的简单算法比小数据的复杂算法更有效”。据此可以推论以统计为基础的经济史计量研究比小数据样本加数量模型更有效；3、需要强调的是：大数据建设对于加强国际竞争力有重要意义，而本项目研究是中国大数据建设的一部分。笔者认为，在一定场合，方法仍然有优劣之分：“孤证”优于“无证”（细考以往已发表的计量史学成果，有不少数据来源“无证”，纯属研究者臆断），“罗列”优于“孤证”，“统计”优于“罗列”。所以笔者赞同吴老说的经济史计量研究仍然“主要是统计学方法”观点。

近几年经济史计量研究三大学派在我国发展形势有喜人变化：三大学派都有中青年学者参与。吴老注重统计的思想需要有人践行，我们的《近代中国经济统计研究》项目团队有数十位中青年学者正在披荆斩棘，努力做好这方面工作，我们这一拨算是统计学派；陈志武先生主办的四届“量化历史讲习班”吸引了一批又一批青年学人，他们以后在方法论上究竟会有什么走向还不好说，我们暂时按照讲习班的名称把讲习班师生这一拨归为年轻的计量史学派；广东外语外贸大学刘巍教授组建了中国计量经济史研究中心，编印了《中国计量经济史研究动态》学术通讯，发表了一系列重要成果。从他们所用方法来看，应属于模型派代表。新时期三大学派各自努力，互相激励，都在推动我国经济史计量研究。三大学派可以说现在都在打基础，尤其是我们统计学派打基础需要花费更多精力。笔者相信，三大学派各自会做出何种业绩，预计十年后可以初见分晓。

四、建设经得起检验的数据库

现在中国经济史计量研究状况有两大问题，一是历史数据资料仍然很缺乏；二是已有的数据资料集存在较多问题，需要进行检验，不能拿来就用。笔者发现，在已有的中国经济史数据资料集中，严中平主编的《中国近代经济史统计资料选辑》学术价值较高但也存在较多问题。汪敬虞先生曾参加编写《中国近代经济史统计资料选辑》工作，他后来又曾委托笔者对《中国近代经济史统计资料选辑》中的错误部分进行核校修订。但是该项核校修订刚刚进行一个多月，汪老又突然下令停止。这方面的工作成果只能在我们《近代中国经济统计研究》项目成果中体现了。我们现在进行数据库建设，既要注意数量，使得规模尽可能大；又要抓好质量，要建设经得起检验的数据库。

我们的《近代中国经济统计研究》项目预期成果包括建成两大套数据库：一整套近代中国经济统计原始数据库（Primary Data），可供大家查询和检验，有较高资料文献价值，以及一整套经过我们努力考证、核校、插值形成的近代中国经济统计改进数据库（Improve Data），可供大家查询。

我们先要下大功夫做好以往研究综述及主要用史学方法广泛收集的相关统计资料，在此基础上合成一整套近代中国经济统计原始数据库，再运用相关史学、经济学、统计学方法，进行认真细致地考证，去伪存真；同时要整理关于中国各地近代计量单位的资料，切实解决各地各时期计量单位换算等问题；并结合其他资料，用科学插值法进行补充和修正，做出系列统计表，合成一整套近代中国经济统计改进数据库。这两大系列数据库是中国大数据建设的一部分，对于经济学、统计学、历史学学科建设都有着重要意义，也是我们进一步展开分析的基础。我们的数据库将按基金管理有关规定提供给社会各界使用。

据笔者初步了解，陈志武、李中清等学者也在从事有关中国历史数据库建设工作。这些数据库都是中国大数据建设的一部分，这些工作将有助于大大减少中国经济史计量研究面临的历史数据缺失困难问题，进一步推动经济史计量研究的开展。

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

五、经济史计量研究与经济学理论发展

吴老在给研究生讲课时曾经指出，定量分析可以检验已有的定性分析，以尽量避免随意性定性判断，它还可以揭示多种变量相互之间的内在关系，揭示经济事物发展变化趋势，可以使人们对许多历史问题的认识不断深化。他曾以清代江西景德镇制瓷业研究为例，告诉我们：从当时史料数量看，景德镇官窑留下的史料多，民窑的很少，不做计量研究则会给人以清代景德镇制瓷业是以官窑为主的印象。做了计量研究，才发现当时官窑的产量和占用的技术力量都不到民窑的1%。吴老还列举了其他一些案例，使我们对经济史研究中计量方法的重要性有了较深的印象。

吴老也告诫我们，定量分析要与定性分析相结合，“已有的定性分析常有不确切、不肯定或以偏概全的毛病，用计量学方法加以检验，可给予肯定、修正或否定”；而计量经济学方法可以用于“检验已有的定性分析，而不宜用它创立新的论点”。①

吴老肯定了经济史计量研究对检验已有的定性分析的作用。至于吴老的后一句，笔者要表示一点不同意见。笔者认为，经济史计量研究也可以帮助创立新的论点。诺贝尔经济学奖获得主M. 弗里德曼等人通过对1867—1960年美国货币史的统计研究，推导出了著名的货币层次理论及货币供应决定模型，就是这方面的一个典型案例。

19世纪中叶德国统计学家恩格尔的工作也是这方面的一个典型案例。

表1　　恩格尔对比利时三个阶层消费结构的统计

	食粮费	衣着费	住宅费	燃料费	文教卫生娱乐费
一般劳动者家庭	62%	16%	12%	5%	5%
中等阶层家庭	55%	18%	12%	5%	10%
高等阶层家庭	50%	18%	12%	5%	15%

他从这一统计表，推出了经济学上著名的恩格尔定律。表1显示，贫困家庭食粮费支出的比率反而高。随着家庭收入的增加，食粮费支出比率渐次减少，衣着费的支出比率先上升后持平，住宅费、燃料费的支出比率保持不变，文教卫生娱乐等杂项费用支出比率随家庭收入增加而明显增长。1868年，德国统计学家修瓦彭研究了柏林市民的所得额与住房支出的关系，推翻了恩格尔的关于住房支出比例相对不变的结论。但是恩格尔关于收入水平变化与食物支出比率变化的函数关系的推定，得到了广泛的认同。人们据此得出一个消费结构变化规律：一个家庭收入越少，家庭总支出中用来购买食物的支出所占的比例就越大，随着家庭收入的增加，家庭总支出中用来购买食物支出所占比例则会下降。推而广之，一个国家越穷，每个国民的平均收入中（或平均支出中）用于购买食物的支出所占比例就越大，随着国家的富裕，这个比例呈下降趋势。这一定律被称为恩格尔定律，反映这一定律的系数被称为恩格尔系数。其公式为：

恩格尔系数（%）＝食品支出总额/家庭或个人消费支出总额×100%

在经济分析中常用恩格尔系数来衡量一个国家和地区人民生活水平的状况。根据联合国粮农组织提出的标准，恩格尔系数在59%以上为贫困，50%—59%为温饱，40%—50%为小康，30%—40%为富裕，低于30%为最富裕。

笔者认为，恩格尔定律仍有较大的拓展空间。可以推论：随着收入的增加，消费结构中食物支出比例（恩格尔系数）下降时，其他方面的支出所占总支出比例会相应上升。我们进一步要问的是：消费结构其他方面的变化又有什么规律？根据明太祖九世孙、“东方百科全书式的人物”朱载堉创作的散曲《山坡羊·十不足》，结合上述恩格尔关于比利时不同家庭消费结构的统计表第3列数据，可以

① 吴承明《经济史：历史观与方法论》，第248页。

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

提出以下假设。

《十不足》讲“逐日奔忙只为饥,才得有食又思衣”。贫民一旦填饱肚子,就要考虑穿衣问题。结合上述恩格尔关于比利时不同家庭消费结构的统计表第 3 列数据,可以假设:当恩格尔系数由 59% 移向 50% 时,人们由“糊口”走向“温饱”时,消费重心开始向“穿”的方向移动,衣着所占总支出比例会有较大幅度上升。《十不足》接着讲“置下绫罗身上穿,抬头又嫌房屋低”。据此可以假设:当恩格尔系数由 50% 移向 40%,人们由“温饱”走向“小康”时,消费重心开始向“住”和“用”的方向移动,住房及日用必需品等支出所占总支出比例会有较大幅度上升。我们还可以继续推论:当恩格尔系数由 40% 移向 30%,人们由“小康”奔向“富裕”时,消费重心开始向“文体娱乐”方向移动,文教卫生娱乐费(包括旅游交通费及雇佣仆人费用)等支出所占总支出比例会有较大幅度上升。当恩格尔系数由 30% 下移,人们由“富裕”迈向“最富裕”时,消费重心开始向“社会公益事业”方向移动,慈善活动费及社会公益费用等支出所占总支出比例会有较大幅度上升。

当然,上述关于恩格尔定律拓展的思考,只是受《十不足》前两句的启迪而做出的理论猜想,还有待经济史统计资料的证明。如果能得到证明,可以将其命名为“扩展型恩格尔定律”(恩格尔定律 + 消费结构其他方面变化规律),以向著名统计学家恩格尔致敬。

不过,正如恩格尔定律反映的是一种在熨平短期波动中求得的长期趋势那样,我们关于扩展型恩格尔定律的猜想也是在研究一种需要熨平短期波动的长期趋势,并且进行比较时还要考虑到剔除那些不可比因素,如消费品价格比价不同、居民生活习惯的差异以及由社会经济制度不同所产生的特殊因素等。

The Era of Big Data and the Quantitative Research on the Economic History

Chen Zhengping

Abstract: Cliometrics Analysis often uses three categories of methods, including: statistics, econometrics and quantitative method of history. The warming “quantitative history” and the “quantitative method of history” is one thing. Therefore, the four aspects, which have been leading to the poor development of the quantitative method of history, are worth to be vigilant in the current research on the quantitative history. Nowadays the era of big data is coming. Big data allows us to discover the details more clearly than individual samples provide. “Simple algorithm using big data is more effective than complex algorithm using small amounts of data”, thus we could infer that the statistics based quantitative research on the economic history is more effective than the quantitative model with small amount of data. The statistics school is closer to the main features of the big data era, thus it better fitted with the trend of the times. By building a database at present, it is necessary to keep the quantity of data as large as possible, and the quality and the robustness should be ensured as well. Cliometrics Analysis can also help creating new arguments. Nobel Prize Award Winner M. Friedman et al. is a classic case. The statistical research on the monetary history of the United States through 1867—1960 derived the famous Friedman rule and the money supply. The research work of germen statistician Engel in the middle19th is another classic case.

Key Words: Era of Big Data; Quantitative History; Cliometrics; Statistics; Economic History

(责任编辑:高超群)

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net