

财政大数据分析应用的方法论初探

●谢波峰 刘亚敏

【摘要】本文认为财政大数据分析应用需要方法论的指导,这是由财政大数据的特点、发挥财政大数据更大效益的要求所决定的,也是大数据分析应用递进式发展的必由之路。论文从循环式上升和统一框架的角度探讨方法论中的主要内容,认为财政大数据应用包括递进式三阶段、周期性四步骤的发展过程,统一框架需要提供静态和动态功能支持,论文最后提出在实践中加强方法论指导、借鉴成功经验推广和进行若干应用探索等建议和思考。

【关键词】财政大数据 大数据方法论 数据分析

2019年5月,财政部发布了《关于推进财政大数据应用的实施意见》,要求各部门推进财政大数据应用,以支撑建立现代化财政制度,这也是第一次从顶层设计层面对财政大数据的应用作出要求。财政大数据领域虽然也有不少研究和实践,不仅有提出大数据在财政方面应用的设想,包括运用大数据加强财政信息化管理(连家明,2017),通过大数据提升财政管理效率等(马洪范,2017;刘文卿,2017),而且也有研究进一步介绍了在重点领域的应用,如绩效方面的大数据应用(王敏等,2019),以及更基础的指标体系等(马蔡琛,2019),但与其他领域的大数据应用,尤其是税务领域应用相比(刘尚希等,2017;谢波峰,2017),需要更多的探索。

财政大数据的发展,不仅和其他领域存在着同样的问题,例如对于大数据的不同理解(朱扬勇,2015),还有些亟需解决的问题,方法论就是其一。虽然,大数据本身就是一种新的方法论(齐磊磊等,2020),但对于方法论的理解也各有不一,根据ISO国际标准化组织术语库、术语标准和术语条款中关于“方法论”的定义,并结合方法论相关研究发现,不同学科领域关于方法论的定义和描述都存在共通点,即方法论往往与研究范式和理论框架相关联(安小米等,2010),本文也同意这一观点,即在一般化的层面上,根据现有

的其他领域和目前的探索,归纳财政大数据应用的方法。本文拟针对以上意义的方法论进行探索,首先对研究财政大数据应用的方法论意义进行阐述,然后探讨方法论中的主要内容,最后提出相应的建议和思考。

一、研究财政大数据应用方法论的意义

(一)财政大数据的特点决定需要应用方法论

自从2008年《Nature》(自然)杂志专刊讨论大数据之后,大数据开始在包括经济学领域在内的各个领域流行(Varian,2014),与天文、生物等大数据概念起源的行业相比,财政大数据不仅有着各行业大数据共同的特点,也有着其自身领域的特点。

财政大数据与其他领域共同的特点就是都具有所谓的“4V”。第一,在规模容量(Volume)方面,虽然依照《自然》杂志所讨论的大数据至少是PB级的数据,但从目前来看尚未到达,然而随着财政大数据的发展,规模(Volume)将不是问题。第二,在数据积累速度(velocity)方面,在财政信息化程度较高的情况下,数据前所未有地不断增长。第三,可获得的数据种类和变量(Variety)越来越多。第四,对于财政大数据的价值(Value)基本上获得了共识。

然而,财政大数据同样也具有一些该领域的独有特点,或许可以称之为

“大而分散”。所谓“散”，主要指财政大数据来源众多而差异性较大，财政信息化包括的系统众多，以典型的省级平台为例，一般包括项目库、预算编制、预算执行、工资统发、政府采购、财政惠民一卡通、非税收入征缴、会计核算、财政资金动态监控等各个业务子系统，通过这些系统收集的财政大数据，无论按经济性质分类，还是按预算部门分类，或者按项目分类，各种不同的财政支出之间都既有一定可比性，又存在着不可比的因素，在这种大数据的基础之上，进行包括财政收支、宏观经济、政府债务、民生支出、三农支出等(李奇云,2018)多方分析应用数据，大数据应用的基础和前提就存在着一定的问题，而这种“大而分散”的现实格局，需要从方法论层面来统一。

(二)财政大数据发挥更大效益需要在方法论层面进行统一

一方面是财政大数据大而散的性质，另一方面是大数据分析利用需要集中进行的现实紧迫需求，方法论层面统一则是协调好这一矛盾的可行途径。

一方面是需要统一，而在具体的分析项目和系统上统一，例如所应用的模型要求一致，但是所研究分析的对象却存在较大差异，效果可想而知，而统一所谓的分析模块和流程，从目前信息化的实践来看，实质性价值大于实施的负担，现阶段不可行，或许将来也存在着困难。另一方面是财政体制所决定的，中央和地方不同的财政运行主体，各自需要进行财政大数据的应用分析，而相互之间既有相同之处，也有不同之处，在具体的分析应用项目和系统中进行统一，至少在目前阶段存在着一定的困难。

(三)财政大数据方法论指导是大数据分析应用递进式发展的必由之路

从目前来看，财政大数据的实践主要体现在两个层面，一是运行层面(Operational)，通过大数据的链接将财政运行扩展到财政以外的相关部门，实现数据自动交换，将财政运行相关的数据汇集到财政大数据平台，通过扩大数据域，

实现数据在网上多流动，办事人员在网下少走动；二是分析层面，在运行层面的基础上，将财政数据与跨部门数据联系起来进行分析，进一步挖掘财政大数据运行的规律。

这两个层面的实践也符合信息化建设规律，信息化建设一般都经过电子数据处理、管理信息系统、决策支持等三阶段，大数据领域的信息化应该也不例外，从实践来看，EDP阶段相对容易完成，而MIS和DSS阶段则更难，但正是后两个阶段对数据的分析利用，进一步体现了财政大数据的价值，将数据变成“宝藏”。当然，在实践中，由于不同部门的分析应用基础不同，可能也会出现某些局部应用比另一些应用超前的现象，使得大数据应用全局看起来有些“犬牙交互”。目前从我国财政大数据的实践来看，中央及不少省市已经着手将财政数据和一些领域的其他数据打通，如扶贫资金监管(朱良华,2019)、预算绩效监督(刘国永,2018)，并开始了一些初步分析，已经产生了一定的效用。

对不同阶段的应用进行整体性考虑，从而节约成本、发挥效益，离不开方法论的指导，从方法论层面高屋建瓴，不仅在实施阶段上有前瞻性，而且在考虑财政大数据应用全局中具有系统性，这样才能将分开的各阶段、各层面、各业务领域的应用分析一体化，不断推进和提升财政大数据分析的水平。

二、大数据应用方法论的主要内容

在认识到方法论之意义的基础上，大数据应用方法论要解决的问题，是要指导从具体应用到抽象框架的搭建，从单项应用到集成应用的汇总，从简单应用到成熟应用的发展，本文拟从统一框架和应用周期两方面进行论述展开。

(一)财政大数据应用的循环式上升

从某种意义来看，大数据应用是一个学习过程，不仅数据范围没有边界，随着大数据的积累，数据的范围和容量不断增长，所应用的模型方法也没有边界。虽然理论认为，数据是DGP过程^①，但财政大数据的DGP过程是复杂系统产生的，因此大数据则是通过不断找到新数据、新

方法的学习过程。根据目前已有的实践,笔者将这一学习过程归纳为图 1 所示的三阶段四步骤(见图 1)。

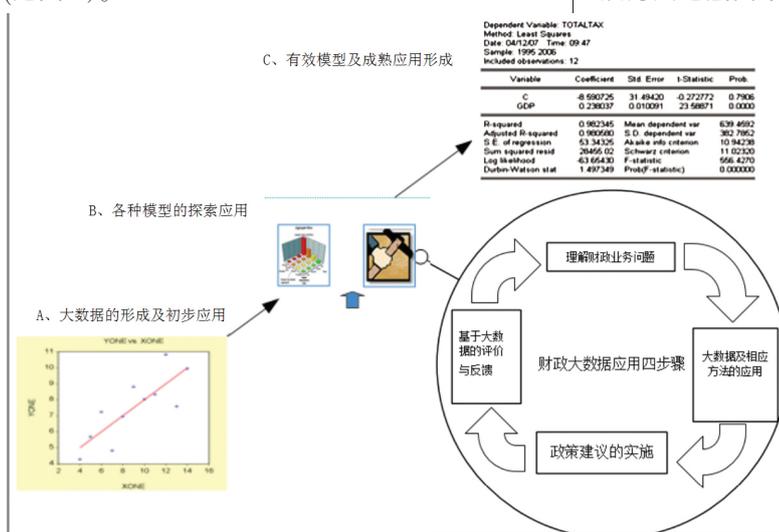


图 1 财政大数据分析应用的三阶段四步骤

1、大数据应用的递进式三阶段。大数据分析应用一般要经历三个阶段:大数据的形成及初步应用、各种模型的探索应用和有效模型确定及成熟应用形成。三个主要阶段的特征大致如下。

第一阶段,大数据的形成及初步应用。这一阶段主要通过收集形成财政大数据,并辅助以简单的方法对大数据进行描述和加工,以验证收集大数据的有效性,这一阶段中的分析方式简单,结果往往可以直观的图形方式表达出来。其中值得一提的是,在这一阶段,会有所谓探索性数据分析过程(Exploratory Data Analysis, EDA),对数据进行简单描述与加工,这往往是大数据项目成功重要的第一步,并且也会是最初大数据项目产生效应的起点。在这一探索过程中,加上业务常识,往往会起到较好的辅助决策作用,因此,形成有效的 EDA 平台往往也成为不少行业大数据项目第一期(第一阶段)的目标。

第二阶段,各种模型(方法)的探索应用。在大数据形成的基础之上,由于分析问题本身具有复杂性,以及试图进一步找到内部关联性的内在驱动,笔者会探索应用各种模型和方法。在不同的模型方法中,除了传统的模型和方法之外,随

着大数据的积累,所谓的机器学习方法层出不穷,是大数据应用第二阶段最令人期待的方向。从方法论层面来看,模型的形成机制和作用领域主要分为两大类:第一类是探测型模型(detection model),建模思想主要来源于财政管理经验,通过主动选择大数据中的特定变量,解决想了解的变量之间的关系(we know we don't know);第二类是发现型模型(discovery model),其主要用途是解决大数据中大家也不了解的规律(we don't know we don't know),模型的主要生成机制是数据挖掘或机器学习机制。

第三阶段,有效模型及成熟应用的形成。在第二阶段对不同方法的尝试中,总结出了不同模型的应用特点,并且得到了对于特殊对象及问题的有效模型。与其他模型相比较,所得到的模型可以刻画所分析领域的主要矛盾和重点,逐渐形成有效的分析应用,就进入了大数据应用的第三个阶段。在这一阶段将集成前两个阶段的经验,形成适用于不同层次的模型。值得一提的是,这一阶段又为下一个应用周期的开始奠定了基础。

2、大数据应用的周期性四步骤。每一阶段中,大致都有四个步骤,这四个步骤大致内容如下。

第一步,理解业务问题。理解业务问题,其实质是将财政业务问题建立模型,无论哪一个阶段的哪一类模型,核心无外乎:主要矛盾的取舍(变量)、主要方面的联系(谁是因变量,谁是自变量),而具体的模型形式和参数则是下一步骤的任务,业务问题模型化的目的都是反映数据(指标)之间的有机联系和主要规律。在理解业务问题进行模型化的过程中,一方面需要防止将模型片面化地理解成数学公式,另一方面要积极推动大数据下的“小应用”。

第二步,基于大数据及相应方法的应用。对业务问题进行大数据应用,强调的是要通过大数据理念和方法解决业务问题,要尽量跳出传统方法的思路,例如在税务领域的常见做法是新的变量、新的模型和新的理论(谢波峰,2017)。在大数据应用中,许多新的方法是属于研究变量之间相关关系的方法,而进一步探求因果关系的研究也在进行当中(susan,2017)。在相关模型和方法应用中,也会出现大数据环境下的相应问题,典型问题如,当解释变量数量增加时,不少方法会很快表现不佳,有些方法如随机森林、深度学习等方法则可以取得更好的表现,有时候甚至要考虑传统方法、机器学习等不同方法的融合(萧政等,2019)。

第三步,政策建议的实施。政策建议的实施就是将大数据分析结果用到实际工作中,如果不这样做,那么大数据分析就意义不大,大数据的价值也无法体现。针对财政政策效果评估、财政支出绩效评价等典型大数据分析应用,大数据分析结果的使用方式如下:

第一,分析结果的业务化。将大数据分析结果联系业务需求,加工成相应的业务应用,这种业务化通常以评分、排名等通俗易懂的形式,或者可视化的形式展现。

第二,分析预测与实践结果的对比。分析意义上的“预测”分成两种,一种是对已有客观结果的理论估计,另一种是对将来未知值的预计,分析结果在这两个层面上都可与实践进行对比。

第三,新的认知的验证。分析过程中发现的有关业务新知识,有助于理解财政运行的新规律,这些知识还需要在未来的活动中加以验证,例如,通过财政大数据对财政支出进行绩效评价,通过大数据找到更有效的数据指标、评估方式。

各种分析结果的应用都需要与实践互相验证,基于更大样本(甚至接近于总体样本)、更宽视野的大数据分析结论,与常见样本分析不同,更可能全方位地反映客观运行实践。大数据分析结论与实践验证的关系,不仅仅是实践(专家)对大数据分析结论的确认,还可以是对实践专家在

部分新领域、复杂情况下的纠偏,是“人类智能”与“人工智能”在大数据中的互动。

第四步,大数据应用的评价及反馈。大数据应用的评价和反馈与分析结果在财政业务实践中的验证是两个不同的概念,既有联系又有区别。应用的评价及反馈既要看大数据应用的效果,还要与常规方法比较,并且还要反思各个环节中的问题,辩证看待大数据分析结果与现有经验之间的关系,值得指出的是,要基于大数据进行应用的评价和反馈,要在大数据体系中找到合适的评价指标及体系,而不是局限于现有的评价指标。

另外,在评价实施环节,还可以对大数据应用中的数据问题进行数据修复。有时,通过评价及反馈可以发现现有的大数据中存在的问题,这些问题会大大影响财政大数据应用的科学性,通过修复数据,将大数据变得更加“干净”,从而进一步提升大数据分析利用的效果。

在财政大数据分析应用中,三阶段递进式发展,四步骤周期性互动,各阶段、各步骤缺一不可,每一步生成的结果不断向后传递,传递过程中不断产生信息与知识,信息与知识的增长又应用于业务过程,每一轮互动过程会形成一个周期,周期往复,不断上升到不同的大数据应用水平。

(二)基于应用路径和多维度功能要求的统一框架

大数据分析应用的统一框架不仅仅是大数据分析应用的IT基础要求的,就目前的大数据分析技术平台而言,从底层数据存储到分布式计算的硬件、软件结构都有相应的框架要求(程学旗等,2014),这也是实施大数据分析方法的逻辑要求,本文的讨论也偏重于这一方面。

1、建立统一框架是大数据分析应用路径的要求。建立统一框架的目标在于尽力改变财政大数据“大而分散”的局面,建立不同财政大数据应用的内在联系,这既是发挥大数据应用的长处,也是所谓大数据应用路径的内在要求。

统一框架所需要解决的问题是具体应用到

抽象框架的搭建,经过多个不同领域的大数据分析项目的实践分析(官思发等,2015),从应用路径范式来看,大数据分析无非是数据、指标、模型三者之间的转换和提升。

数据:所谓采集数据,结合目前的实践来看,主要指来自财政、税收、市场监管、国资管理、城乡建设、市场交易等各个部门、各个环节的数据项,从这个意义来看,因为并未局限于财政部门,取得了与财政运行和应用分析相关的数据,因此称之为财政大数据。

指标:基于(采集)的财政大数据,通过多种方法,笔者将数据转换成很多有利于财政业务管理的指标,因为这时有些形成的指标已经具有一定的分析功能,所以,又称之为所谓的(分析)指标,可以更进一步发挥数据的作用。

模型:最后,在指标之间找到关系,形成模型。实践应用分析中的模型并不完全等同于经济分析领域的模型,笔者认为,在大数据分析领域,模型是指根据管理的目的、能够反映财政经济的主要特征、抽象出主要矛盾的某种方法和形式,模型虽然具有不同形成机制,但是内在统一的目的都是反映数据(指标)之间的有机联系和主要规律。

任何分析应用都是从(采集)数据到(分析)指标,然后从指标联系形成模型,进行循环转换,因此在方法论指导之下,能够形成统一框架,不仅仅是IT技术框架的限制和要求,而统一框架包括的功能和内容就应该围绕这一路径展开。

2、统一框架对多维度功能要求的支撑。对统一框架的静态和动态功能支持,是方法论层面对统一框架设计的两个维度的要求,在静态的层次上,要求有数据库、指标库、模型库和应用库的支持,在动态转化上,要求有设计工具的功能支持。

对不同层级的静态库支持。虽然在技术实现上,低层都是IT意义上的数据库及其相应的管理系统,从方法论的要求上来看,统一框架需要有数据库、指标库和模型库、应用库等不同层级的库支持。数据库主要是针对采集数据的功能要求,要支持动态扩展,不仅是将固定结构的数据

采集到系统中,还要根据新数据的出现,不断将新的有价值的纳入大数据应用当中。指标库主要是针对指标的定义、生成、管理等功能要求,不仅可以由业务人员基于数据灵活定义相应的指标,而且业务含义一致的指标之间还可以互通互联。模型库和应用库共同组成面向具体业务的支撑能力,模型库重点在基本方法的支持,偏重通用性,而应用库重点在于专用的分析应用,偏重个性化。

对动态转化能力的设计工具支持。针对静态的不同层次的功能,以及对不同阶段、不同步骤的动态转化,需要有业务建模、交互优化、应用结果分发等设计工具提供能力支持。通过业务建模工具的支持,财政大数据应用可以从业务问题转化为大数据分析应用平台可运行的分析问题,而交互优化工具提供了不同方法和模型的选择比较,应用结果分发工具则根据常见的分析结果使用方式,对分析结果进行转化,并联系业务系统进行应用。

基于大数据分析的循环式上升周期,以及内在的应用路径转换规律,财政大数据分析应用统一框架才有可能存在,并且在方法论层面进行统一抽象,设计并提供相应的支撑功能。

三、推行方法论的若干实践建议

(一)在财政大数据分析应用实践中重视方法论的指导意义

推行方法论不仅仅是财政大数据分析发挥更大效益的需要,也是节约财政大数据建设资金的需要,更是积累大数据分析经验的要求。本文提出的方法论方面的观点虽然比较浅薄,但足以提示方法论的意义和重要性。当前,国家层面对大数据的战略方向已经确定,财政大数据应用推广的相关政策已经制定,有关部门应当适时重视方法论的价值,进一步完善其中的内容,要求进行相应的财政大数据分析项目的各个环节,对照方法论的要求进行检视,通过方法论的统一抽象作用,将中央和地方财政各层级、各领域的财政大数据项目形成合力,共同努力,发挥大数据分析的作用和效力。

(二)借鉴已有成功实践推行方法论工具

抽象的方法化要通过工具进行固化,可以借鉴互联网企业中的数据中台概念和实践(谭虎,2019),落实方法论的统一框架的要求,推行财政大数据中的方法论。数据中台实质也反映了伴随着大数据的积累,互联网企业应对数据应用需求的统一化应用的方法论工具。探索在财政大数据分析中建立数据中台与业务中台两个维度的平台,通过数据中台将采集数据统一标准和口径,加工成可用财政业务数据群,形成财政大数据资产层,为内部和外部提供高效服务;通过业务中台则是将财政分析业务与分析逻辑进行隔离,制定通用化的各类分析服务和功能,向具体的分析业务提供快速、低成本、可复制的分析功能。

(三)进行若干重点的分析应用弹性开发探索

通过若干重点应用,验证和完善方法论思想的可操作性,以及相应工具的实用性,在具体推进中可以先“小”后“大”,通过小范围的应用,建立大范围的统一框架,在统一的规划下,在敏捷开发等先进IT理念的指导下,先行建设具体应用中亟需的小工具、小平台,对业务重点的大数据分析进行主题域建设,以支撑实际业务应用的探索,并适用建设管理的制度资金需求,通过开放式接口,逐一整合纳入到统一的框架平台和工具包中,形成越来越完善的财政大数据分析应用平台。■

(基金项目:国家社会科学基金重大项目“新时代中国预算绩效管理改革研究”(19ZDA071))

(作者单位:中国人民大学金融与财税电子化研究所、中国人民大学财政金融学院)

注释:

①关于数据产生过程(DGP),可以进一步参阅王少平.宏观计量的若干前沿理论与应用[M].天津:南开大学出版社,2003:133-134.

参考文献

[1]安小米,王丽丽.大数据治理体系构建方法论

框架研究[J].图书情报工作,2019,(24).

[2]程学旗,靳小龙,王元卓,郭嘉丰,张铁赢,李国杰.大数据系统和分析技术综述[J].软件学报,2014,(09).

[3]官思发,孟玺,李宗洁,刘扬.大数据分析研究现状、问题与对策[J].情报杂志,2015,(05).

[4]刘国永.构建政府决策大数据 全面实施预算绩效管理[J].中国财政,2018,(01).

[5]连家明.大数据时代财政信息化建设回顾与前瞻[J].地方财政研究,2017,(12).

[6]李奇云.进一步推进财政信息化工作的基本思路[J].中国财政,2018,(14).

[7]刘尚希,孙静.税收治理:基于大数据的理论思考[J].经济研究参考,2016,(09).

[8]刘文卿.新时代对财政大数据应用的思考[J].地方财政研究,2017,(12).

[9]马蔡琛,赵笛.大数据时代的预算绩效指标框架建设[J].中央财经大学学报,2019,(12).

[10]马洪范.大数据时代的财政治理[J].地方财政研究,2017,(12).

[11]齐磊磊,贾玮晗.从复杂性科学哲学看大数据的方法论价值[J].系统科学学报,2020,(01).

[12]谭虎.详解阿里云数据中台[N].中国信息化周报,2019-10-28(014).

[13]王敏,彭敏娇.大数据时代全面预算绩效管理面临的机遇和挑战分析[J].经济纵横,2019,(05).

[14]谢波峰.面向大数据的税务管理应用模式及政策建议[J].国际税收,2017,(04).

[15]萧政,周波.一名计量经济学家对大数据的展望[J].财经智库,2019,(01).

[16]朱良华.大数据背景下提高财政扶贫资金绩效的策略研究——基于广西的考察分析[J].财政监督,2019,(11).

[17]朱扬勇,熊赟.大数据是数据、技术,还是应用[J].大数据,2015,(01).

[18]Susan, Athey. Beyond Prediction: Using Big Data for Policy Problems[J]. Science, 2017.

[19]Varian H R. Big Data: New Tricks for Econometrics[J]. Journal of Economic Perspectives, 2014,(02).