大数据应用对中国企业市场价值的影响*

——来自中国上市公司年报文本分析的证据

张叶青　陆　瑶　李乐芸

内容提要: 推进大数据与实体经济的深度融合成为中国新一轮的经济增长点。本文通过对A股上市公司的年报进行文本分析,构建了衡量公司层面"大数据"应用程度的指标,探讨了企业大数据应用的发展状况及决定因素,检验了大数据应用对公司市场价值的影响。研究发现: 第一,规模较大、有形资产比例较低、盈利能力较强,以及所在地区市场化程度较高的公司更可能在生产经营过程中应用大数据; 第二,大数据的应用可以显著提高公司的市场价值; 第三,主要的影响机制在于大数据的应用显著提高了公司的生产效率和研发投入,而相关技术和人才供给的不足可能会阻碍大数据对市场价值的积极影响。本文结论对中国未来大数据相关的政策设计具有参考价值,为推动实体企业生产经营与大数据的高效融合提供了经验证据和指导建议。

关键词: 大数据　文本分析　市场价值　生产效率　研发投入

一、引　言

以大数据、云计算、人工智能等为代表的信息技术迅猛发展,数字化的知识和信息已成为实体经济中关键的生产要素。如何抓住新工业革命的历史性机遇,提高大数据与实体经济的融合发展水平,也随之成为我国经济发展中亟待解决的重要问题。从商业实践层面上看,自2013年起,微软、IBM、阿里巴巴、腾讯等全球知名企业开始迅速收购大数据上、下游厂商,布局大数据战略; 根据中国信息通信研究院发布的《中国数字经济发展白皮书(2021)》,我国2020年数字经济增加值占GDP比重已达38.6%,数字经济增速为GDP增速的3.2倍。①从政策层面,2014年以来,"大数据"相关内容已连续8年被写入国务院《政府工作报告》; ②党的十八届五中全会正式将大数据发展上升为国家战略; 党的十九大报告进一步明确将"推动互联网、大数据、人工智能和实体经济深度融合"列为深化供给侧结构性改革和建设现代化经济体系的重要举措。在技术发展和政策鼓励的双重推动下,大数据已成为我国经济转型的重要支点。2020年以来,新冠肺炎疫情在全球大流行,给人们的生产生活带来了巨大冲击,并将产生持久影响。疫情期间经济活动非聚集性的需要进一步激发了实体经济对数字化、智能化技术的应用需求。面对大数据蓬勃发展的趋势和疫情冲击导致

*　张叶青,中央财经大学财经研究院,邮政编码: 100081,电子信箱: yeqingzhang@cufe.edu.cn; 陆瑶(通讯作者),清华大学经济管理学院,邮政编码: 100083,电子信箱: luyao@sem.tsinghua.edu.cn; 李乐芸,哥伦比亚大学傅氏基金工程与应用工程学院,邮政编码: 100020,电子信箱: ll3357@columbia.edu。本研究得到国家自然科学基金优秀青年科学基金项目(71722001)、清华大学中国现代国企研究院项目(iSOEYB202006)、清华大学自主科学计划文科专项项目(2021THZWYY09)和国务院国资委委托智库项目(2021WTJF0456)的资助。作者感谢匿名评审专家的宝贵意见,文责自负。

①　详细内容参见 http://www.caict.ac.cn/kxyj/qwfb/bps/202104/t20210423_374626.htm。

②　此外,2014年以来,国务院、发改委和工信部相继出台了《促进大数据发展行动纲要》、《关于组织实施促进大数据发展重大工程的通知》和《大数据产业发展规划》等一系列纲领性文件,鼓励大数据产业的发展以及大数据与实体经济的融合。随后,各省市纷纷出台了地方性的大数据工作部署。

的数字化需求，我们亟需思考在数字化技术与传统产业深度融合过程中，大数据应用对微观企业的影响及其作用机理，从而为后疫情时期的高质量经济发展奠定基础。

然而，截至目前，探索大数据应用及其对公司实际影响的研究非常匮乏。大数据在中国上市公司中的应用普遍吗？哪些因素影响了公司应用大数据的决策？大数据应用如何影响公司的市场价值？其背后的影响机制是什么？上述影响在不同的公司、行业间是否存在着异质性？上述问题的回答，从短期来看有助于为发展数字经济、助推实体经济与传统产业的数字化转型提供启示；从长期来看，有助于为疫情常态化和构建新发展格局的大环境下，我国抓住新一轮科技革命、实现经济高质量发展提供实证支撑。

本文基于 2006—2017 年间 A 股上市公司的样本，从公司年报的文本信息中抓取“大数据”相关的关键词，构造大数据应用程度的衡量指标，并进一步探究大数据应用对公司市场价值的影响。本文的分析结果表明：第一，规模较大、有形资产比例较低、盈利能力较强，以及所在地区市场化程度较高的公司更可能在生产经营过程中应用大数据。第二，大数据应用能够显著提高公司的市场价值。本文采用了工具变量两阶段回归的方法来缓解内生性问题，利用 2009 年启动的“基础学科拔尖学生培养试验计划”构造工具变量。本文还执行了更换大数据应用衡量指标、更换研究样本、更换聚类方式等一系列测试，结果均稳健。第三，机制分析表明，大数据的应用显著提高了公司的生产效率和研发投入，但现实中的技术和人才供给不足则会限制大数据应用对企业的积极作用。第四，异质性分析表明，大数据应用对公司市场价值的促进作用在小规模公司、非国有公司和所在行业竞争较为激烈的公司中更为显著。

综上，本文的创新点主要体现在以下几个方面：第一，利用非结构化的文本数据构建了公司层面的大数据应用程度的衡量指标。采用文本分析的方法抓取了中国 A 股上市公司年报中与“大数据”相关的关键词，构造了大数据应用程度的变量，并检验了其有效性。与以往的研究相比，本文对大数据应用的度量更为直接、有效、准确，且全面覆盖了 A 股所有行业的公司样本，因此能够直观地刻画中国上市公司大数据发展水平的动态变化过程及其影响因素，为后续研究奠定了扎实的数据基础。第二，实证检验了大数据应用对公司市场价值的积极意义，并深入探索了其影响机制，为大数据应用对于公司竞争力提高、宏观经济增长的意义构建了理论基础，提供了实证依据。① 更为重要的是，本文的分析立足于中国企业的数据，充分结合中国的大数据政策优势、数字基础设施建设、人才供应条件等宏观要素，从而提供了根植于中国实际的政策建议，助力中国经济高质量发展。第三，通过探索大数据对公司市场价值影响的异质性，发现规模不同的公司、股权性质不同的公司、竞争激烈程度不同的行业采用大数据对公司市场价值的提高效果存在差异。上述发现既补充了相关文献，又为政府制定大数据相关政策提供了参考。第四，丰富了公司市场价值影响因素的相关研究。本文发现大数据的应用是影响公司股票市场价值的重要因素，这为数字经济趋势下公司市场价值影响因素的研究提供了新证据。基于此，政策设计也需要关注如何帮助企业更好地克服大数据应用过程中可能面临的问题，以推动大数据应用高效地转化为公司市场价值，助推经济的数字化转型。

二、大数据与实体企业融合的作用机理

“大数据”一词最早出现于 20 世纪 90 年代中期，直至 2011 年才开始受到社会各界的广泛关

① Wu et al.（2020）和 Tambe（2014）基于美国数据研究相关问题，但他们的研究聚焦于数据分析技能而非大数据要素或资产。此外，Srinivasan & Chen（2020）发现美国上市公司的大数据应用与其估值存在正相关关系，而本文利用外生政策冲击识别了因果关系。

注。在展开实证研究之前,本文首先需要清晰地定义“大数据”。大数据之“大”体现在哪些方面呢? 以往研究(Chen et al., 2012; McAfee & Brynjolfsson, 2012) 通常以三大特征来定义大数据: 大规模(volume) 、高速度(velocity) 和多样性(variety) ,也称“Three Vs”。① 例如,高德纳咨询公司(Gartner, Inc.) 将大数据定义为“海量、高速、多样化的信息资产”,它需要高效的、创新型的信息处理形式来提高洞察力和决策水平。国务院 2015 年印发的《促进大数据发展行动纲要》中指出“大数据是以容量大、类型多、存取速度快、应用价值高为主要特征的数据集合”。从大数据与实体企业融合的角度,Farboodi et al. (2019) 总结了大数据的四个重要的特征: (1) 大数据是经济活动的副产品; (2) 企业利用大数据来提高经营效率; (3) 大数据与技术的不同之处在于它本质上是一种信息; (4) 积累的大数据是企业的一种有价值的资产。借鉴上述大数据定义,本文将企业应用的“大数据”定义为企业收集、处理与利用的海量、高速、多样化的数据要素或资产。

在大数据与实体经济深度融合的进程中,大数据的应用影响企业生产经营和最终市场价值的作用机理是什么?

(一) 大数据应用与生产效率

首先,大数据应用可以通过信息渠道来影响企业的生产效率。大数据应用的核心是对海量数据的生产、采集、存储、加工、分析等,而数据本质上是一种信息(Farboodi et al., 2019) 。大规模数据的积累和分析可以为企业提供更加细致、信噪比更高、传播性更强的海量信息。根据信息理论,充足的信息可以提高企业的决策效率。具体而言,企业预测各种宏观经济变量、行业供求环境和微观生产条件的能力增强,而准确的预测结果能够帮助企业做出更优决策,提高生产效率(Brynjolfsson et al., 2011; Brynjolfsson & McElheran, 2016; Agrawal et al., 2019; Tanaka et al., 2020; Babina et al., 2021) 。比如,从“开源”的角度而言,公司通过分析销售大数据,更精准地预测消费者行为和偏好,从而调整自己的生产和销售计划,降低存储成本,提高利润率(Bajari et al., 2019) ; 从“节流”的角度而言,很多企业通过积累那些能耗监控相关的大数据,预测每天消耗的水电量,据此对高能耗设备重新排产,从而降低能耗成本,提高生产效率。

其次,大数据降低企业的劳动力成本,从而提高企业的生产效率。大数据应用实现了传统产业生产过程的自动化转型,可以替代一部分原本由劳动力承担的生产工作,降低生产经营过程中人工参与的程度,从而降低每单位产出的劳动力成本(Brynjolfsson & Mitchell, 2017; Agrawal et al., 2019; Babina et al., 2021) 。而工业化时代,对于采用高水平技术的企业而言,劳动力和机器设备的相对投入份额更低,生产效率更高(Midrigan & Xu, 2014) 。更为重要的是,随着劳动力的供给不断减少,劳动力雇佣成本不断提高,但大数据应用的成本却会随着技术的发展、制度的完善不断降低。因此,大数据应用通过推动劳动力的自动化转型而降低单位产出的生产成本,提高生产效率。Agrawal et al. (2019) 也指出,生产过程中的许多任务都是“预测性任务”,基于大数据的预测边际成本低于劳动力,且预测效率更高,因此大数据替代常规劳动力能有效提高生产效率。

再次,大数据可以从组织管理的角度降低企业生产运营的成本。大数据应用可以使企业内部各组织之间信息传递的成本更低,效率更高,从而缓解公司内部的组织管理问题,实现更加扁平、高效的管理模式,降低管理成本,提升生产效率(Brynjolfsson et al., 2011) 。Mikalef et al. (2018) 指出企业的大数据分析能力可以优化组织管理流程和决策结果。以物流管理系统为例,数据系统可以通过对产品库存、发货情况进行监控,以实现迅速调拨决策,有效平衡供应链的供需两端,降低物流费用,还可以为生产、业务部门及时捕捉市场需求,为下一期的生产销售提供决策依据。

① 其中,大规模指变量数与观测数多导致的海量的数据量; 高速度指数据的收集、更新、分析速度快; 多样性指结构化、半结构化和非结构化等多种大数据类型。

(二) 大数据应用与研发创新

大数据应用也会对广义的技术创新过程产生深远影响。创新是企业保持长期竞争力和获得市场认可的关键,因此大数据应用可能通过促进研发创新,提高公司的市场价值。大数据应用促进企业研发创新的影响渠道主要体现在如下三点。

首先,大数据应用有利于企业更精准地把握消费者的需求,明确当前生产过程的不足和技术创新的方向,进而提高研发投入的需求。与传统数据相比,大数据从数据类型、产生速度和数据规模上都体现出明显优势。受益于大数据的积累和分析,企业在研发前期对市场需求的分析更具前瞻性,使得研发方向更加针对于符合消费者需求结构的新产品或新服务以提高产品或服务的竞争力,从而大大降低研发成果商业化失败的风险(Cockburn et al., 2019)。例如,大数据不仅包括结构化的交易类数据,还包括规模更大的非结构化的消费者行为数据,因此可以更精准和迅速地把握消费者的需求和偏好,提高研发成果转化为公司业绩的效率,进而大大提高企业的研发动力。

其次,大数据应用降低了企业研发过程的不确定性和成本,增强了研发的动机。研发活动是长期的、不确定性较大的高风险活动,前期投入成本巨大,很多公司会为了规避风险而对研发创新的投入不足(Holmström, 1989)。而大规模、多样化、高速度数据的积累和分析能够大大提高预测能力(Brynjolfsson et al., 2011),从而降低了企业研发过程中的不确定性和研发成本。

此外,大数据应用为研发过程积累了丰富的资源和信息,增强了企业研发效率。一方面,大数据的海量规模和多样性扩大了企业信息搜索的空间,高速度提高了信息处理的效率,进而提高了企业从已有技术中提炼创新成果的能力。而过程创新往往建立在密集的信息处理和搜索的基础之上,对已有技术重新改进或组合进而获得技术的新应用或新产品的创造。另一方面,大数据应用还可能催生对创新方法和产业结构的根本性革新。Cockburn et al. (2019) 发现大数据应用为研发创新提供了新的方法,可以改善创新进程的本质,从而对创新和经济增长产生更为深远的影响。相比于发明创造出某一个产品,形成一种创造新产品的新方法往往可以更广泛而深远地影响各个生产领域(Griliches, 1957)。例如,基于非结构化大数据的模式识别(包括识别图片、影像等)催生了新一代自动驾驶技术的发明,可能会彻底改变汽车制造业未来的发展方向。

(三) 大数据应用过程中的摩擦

大数据应用于企业生产经营并不一定会展现立竿见影的积极影响,大数据应用过程可能存在摩擦,这使得大数据无法在短期内全面发挥为企业增值的效果。大数据的应用转化为公司价值过程中可能存在的限制和挑战主要体现在两方面。一方面,应用大数据相关技术的门槛较高,难度较大。以保险行业为例,据沃思信息技术服务公司的统计,①80% 的赔付数据都包括非结构化数据,例如手写的记录、视频和图片等,而非结构化数据的处理和分析具有较高的难度;此外,合并多个数据源系统所获得的数据也是极大的挑战。另一方面,大数据对企业的增值作用离不开一系列互补性资产或资源的投入。大数据应用需要与生产、运营、管理、销售等企业的基本活动紧密结合,因此除了大数据本身,企业还需要其他方面的投入加以配合,包括生产经营活动的重构、新的商业模式的建立、组织结构的调整、管理经验的升级、相关技术的员工培训、特定软件的定制开发等(Bloom et al., 2012; Dranove et al., 2014; Brynjolfsson et al., 2021)。在引进大数据相关设备、人才的过程中,公司往往面临着内部资源的重新分配、劳动力和组织结构的重组、新的生产线的调试等调整成本。因此,只有上述互补性投入与大数据应用齐头并进,大数据对企业的积极影响才能充分发挥,大数据与实体经济的有机融合才能顺利实现。

现实中大数据与实体经济的融合可能受到相关人才、技术等方面匮乏的限制,大数据的价值创

① 资料来源: https://www.wns.com/insights/blogs/blogdetail/407/the-big-leap-battling-challenges-in-adopting-big-data。

造难以充分发挥。从人才的角度,企业应用大数据所需的人才或来源于内部资源,或来源于外部市场。从企业内部而言,企业在职的熟悉大数据技术或研发的人员有助于大数据与原有生产活动的融合,反之员工技能与需求不匹配则会为企业的数据赋能带来障碍(Dranove et al., 2014) 。从外部而言,劳动力市场上的人才需求尚存在较大缺口,因为技术或研发相关的人力资本的积累需要高等教育的培养或多年工作经验的积累(Dranove et al., 2014; Babina et al., 2021) 。因此,与大数据应用相匹配的技术人员或高素质劳动力的供给不足是大数据应用的瓶颈之一。

从技术支持的角度,所在地区的数据类服务供给和数据中心等新型基础设施建设是企业利用大数据实现增值的重要保障。前已述及,大数据的应用存在较大的技术门槛,因此当地的技术环境和相关基础设施能够为企业跨越障碍提供助力。Dranove et al. (2014) 发现公司所在行业会影响其采用新技术的增值效率,如果公司所在行业的技术使用密度不高,那么采用新技术对绩效的改善效果则难以实现。随着大数据应用在行业和地区范围内的普及和程度加深,会逐渐形成规模效应,进而由大数据到企业价值的转化效率可能会有所提高。

三、数据来源及变量描述

(一) 样本选择与数据来源

本文以 2006—2017 年中国 A 股所有上市公司作为初始的研究样本,依次剔除如下样本: (1) ST、*ST 和 PT 公司; (2) IPO 当年的观测值和已退市的公司; (3) 净资产为负的观测值; (4) 主要变量缺失的公司; (5) 信息传输、软件和信息技术服务业的上市公司。大数据产业本身属于信息技术类行业,因此大数据直接相关的行业可能与其他行业受到的影响有所不同。根据本文的理论分析,大数据应用对公司市场价值的影响不应该只局限于与大数据直接相关的行业,而是在非大数据直接相关的行业中也存在显著影响,大数据与实体经济的融合以及非大数据直接相关的企业的数字化转型是本文关注的核心,因此本文将信息传输、软件和信息技术服务业剔除后进行后续分析。最终,获得 2501 家上市公司共 20623 个样本。公司大数据应用的相关变量来自于对公司年报的文本分析,其他市场交易和财务数据主要来自于国泰安(CSMAR) 数据库,CPI 数据来自于国家统计局。为了避免极端值的干扰,本文对连续变量进行上下 1% 的缩尾处理。

(二) 公司层面的大数据应用的测度

1. 大数据应用指标的构建

本文的一个核心变量是公司对大数据的应用程度。大数据应用程度是一个较为抽象的概念,现实中难以找到一个指标能够准确地反映出这一个概念。以往文献衡量企业对大数据的应用程度往往局限于某种大数据应用形式或特殊行业(Zhu, 2019) 。为了尽可能准确地刻画出公司对大数据的应用程度,且尽可能广泛地覆盖中国上市企业的整体水平,本文借鉴 Saunders & Tambe (2013) ,提出了一种新的衡量方式: 基于上市公司披露的年报的文本信息,通过 Python 程序批量抓取年报中与“大数据”应用相关的关键词,根据所有关键词在年报中出现的总次数来构造大数据相关变量。这种衡量方式的基本假设是: 上市公司披露的年报是基于公司实际运营情况客观的陈述,年报中大数据相关关键词的出现次数能够较好地反映公司的大数据应用程度。下文将采取多重验证方法,进一步检验该衡量方式的有效性。①

具体而言,本文利用关键词在公司年报中出现的次数来度量公司的大数据应用程度。关键词的选取借鉴了以往文献(Chen et al., 2012; McAfee & Brynjolfsson, 2012; Farboodi et al., 2019) 、政府文件以及业界报告等。我们一方面紧扣大数据的定义; 另一方面则尽可能地避免了选择的随机

① 本文对上市公司年报进行了抽样分析,发现大数据关键词的出现确实反映了与大数据相关的业务转型和战略规划。

性,按照普适性原则进行筛选。表 1 展示了本文构造变量所依据的“大数据”相关关键词,并详细阐释各个关键词的定义及其与大数据应用之间的紧密关联。①

本文将最核心的大数据应用的衡量指标(ln*Bigdata*) 具体定义为: 公司年报中提及表 1 中大数据相关关键词的次数加一后取对数。由于大数据应用情况随年份增长趋势明显,本文将 ln*Bigdata* 按照“公司—年份”的观测值确定每年缩尾(winsorize) 上下极值各 1% 。

表 1 “大数据”相关关键词及定义

关键词	定义
大数据	企业收集、处理与利用的海量、高速、多样化的数据要素或资产。
海量数据	根据高德纳公司对大数据的定义,海量规模是大数据的重要特征之一。
数据中心	安置计算机系统及相关部件的设施,用于在网络基础设施上传递、加速、展示、计算、存储数据信息。信息时代下,大数据需要安全可靠、高效率的数据中心进行存储、计算和交换。
信息资产	指由企业拥有或者控制的能够为企业带来未来经济利益的信息资源。根据高德纳公司的报告,大数据本质上是一种信息资产。
数据化	将均匀、连续的数字比特结构化和颗粒化,形成标准化的、开放的、非线性的、通用的数据对象,并基于不同形态与类别的数据对象,实现大数据的应用。
算力	也称哈希率,指比特币网络处理能力的度量单位,也是计算机计算哈希函数时输出的速度。

2. 大数据衡量指标的有效性验证

为了验证大数据衡量指标(ln*Bigdata*) 的有效性,我们考察该指标与真实的大数据相关投入之间的正向关联。大数据应用伴随着较大规模的机器设备等有形投入和人才、技术引进等无形投入,因此我们检验该指标与企业数字化投资的相关性。数字化投资明细项包含“软件”“网络”“客户端”“管理系统”“智能平台”等与数字化相关的固定资产购买支出的对数值和无形资产购买支出的对数值。我们首先用相关图(binned scattered plot) 直观展示了 ln*Bigdata* 与上述数字化投资的相关性,并发现了 ln*Bigdata* 与上述大数据相关的有形和无形投资之间均存在强相关性(p 值 < 0.01) 。② 进一步地,将上述数字化投入指标回归到 ln*Bigdata* 上,并全面控制一系列变量(依照模型 3) ,结果不变。因此 ln*Bigdata* 能较好地反映公司实际的大数据投入。

(三) 其他变量的定义

本文主要关注的被解释变量是公司的估值指标: 托宾 Q 值(*Tobin's Q*) 。托宾 Q 值定义为公司总市值与总负债之和除以公司总资产,是常见的衡量公司绩效和成长性的指标(Morck et al., 1988; Bharadwaj et al., 1999) 。与财务会计类的业绩指标相比,公司的市场价值与本文的研究主题更为契合。原因主要有两点: (1) 财务指标通常只反映过去的信息,不具有前瞻性。大数据应用更多地与企业的无形资产和长期价值相关联(Brynjolfsson et al., 2021) 。股票市场价值这一指标包含了公司已有资产的估值和对未来增长潜力的预期,因此与大数据应用的研究更为契合。(2) 财务类绩效指标仅反映账面数值,而公司在股票市场上的价值是基于股票市场的整体风险、通胀水平和公司的系统性风险暴露程度等风险因素调整之后的指标。大数据投入可能影响公司股价的信息含量和资本成本(Begenau et al., 2018; Zhu, 2019) ,因此股票市场价值指标的信息更为关键。

① 在稳健性检验中,本文采用了更广义的“大数据”词典,结果不变。

② 本文也分别验证了大数据应用指标与公司层面研发投入的对数值、研发技术类人员占比之间的正向关联。限于篇幅,大数据衡量指标的有效性验证的图和回归结果及分析留存备索。

本文选取的控制变量包括: 公司规模(ln*Assets*) 、公司杠杆率(*Lev*) 、固定资产比率(*PPE_TA*) 、公司年龄(ln*Age*) 、国有性质的虚拟变量(*SOE*) 、销售收入增长率(*SalesGrowth*) 和总资产收益率(*ROA*) 。其中公司规模和销售收入增长率是经过 CPI 调整之后的指标。

(四) 变量描述性统计

表 2 报告了主要变量的描述性统计结果。在 2006—2017 年间,共有 20623 个公司—年度观测值,覆盖了绝大部分 A 股上市公司。其中大数据应用程度指标(ln*Bigdata*) 的均值为 0. 180,中值为 0,标准差为 0. 508,说明样本中公司的大数据应用程度存在很大差异。托宾 Q 值(*Tobin's Q*) 的均值是 2. 485,与其他文献中上市公司的数据统计量相吻合(吴超鹏和唐菂,2016) 。

表 2　　主要变量描述性统计

变量	变量定义	均值	标准差	中位数	观测数
ln*Bigdata*	表 1 中大数据相关关键词在年报中出现的次数加一后取对数	0. 180	0. 508	0. 000	20623
Tobin's Q	公司总市值与总负债之和除以公司总资产	2. 485	1. 689	1. 946	20623
ln*Assets*	总资产取自然对数	21. 778	1. 241	21. 609	20623
Lev	总资产除以股东权益	0. 457	0. 204	0. 461	20623
PPE_TA	固定资产除以总资产	0. 246	0. 173	0. 213	20623
ln*Age*	当年年份减去上市年份加 1,再取自然对数	2. 183	0. 694	2. 303	20623
SOE	按公司实际控制人性质确定	0. 468	0. 499	0. 000	20623
SalesGrowth	当年营业收入除以上一年营业收入后减一	0. 194	0. 442	0. 119	20623
ROA	净利润除以总资产	0. 0371	0. 0519	0. 0336	20623

四、中国上市公司大数据应用的描述与决定因素

本文利用中国 A 股上市公司年报的文本信息测度了中国上市公司的大数据应用水平。那么中国实体经济整体的大数据应用程度究竟如何? 哪些公司更多地应用了大数据呢? 这一系列问题的解答有助于理解大数据与实体经济融合的进程,并为完善大数据产业发展政策提供依据。本部分描绘了中国上市公司大数据应用的动态变化过程,并探讨了大数据应用程度的决定因素。

(一) 中国上市公司的大数据应用整体情况与变化

图 1 反映了 2006—2017 年间公司层面的大数据应用程度的变化趋势。可以看出,在 2012 年之前公司年报中提到“大数据”这一关键词的平均次数较低,之后公司提及“大数据”的平均次数从 2012 年的 0. 07 次快速增长到 2017 年的 1. 81 次,涵盖的公司数量占比从不足 5% 上升到 38. 93%。“大数据应用”关键词指表 1 中列举的关键词集合,其出现次数也呈现出类似的趋势。上述趋势与中央、各省市出台大数据行动计划的时间较为吻合。① 本文将大数据指标与公司其他财务数据进一步匹配,最终保留 20623 条观测值用于后续分析。②

(二) 大数据应用程度的决定性因素分析

本文分别采用 probit(模型 1) 和 OLS(模型 2) 模型分析大数据应用程度的决定因素:

① 2012 年 7 月,国务院在《“十二五”国家战略性新兴产业发展规划》中首次明确提出对大数据的支持。从 2013 年开始,各省市陆续出台相关支持政策。

② 本文还发现了不同行业、不同所有制性质和不同地区的上市公司的大数据应用均呈上升趋势,但存在发展不平衡的情况,比如东部地区的总体应用水平最高。限于篇幅,此部分结果留存备索。

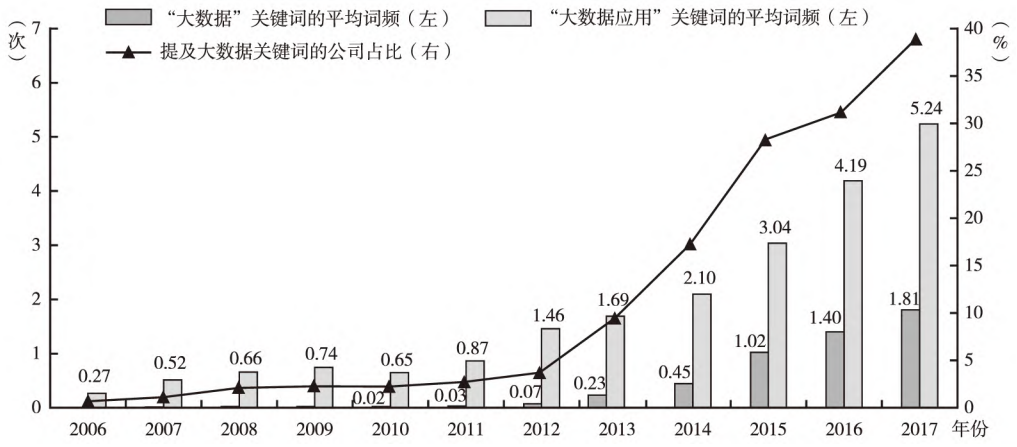

图 1　大数据应用情况的年度趋势

$$Prob(Bigdata\_dummy_{ijpt} = 1 \mid X_{ijp,t-1}, Ind_j, Year_t) = \Phi(\beta_0 + \beta_1 X_{ijp,t-1} + Ind_j + Year_t) \quad (1)$$

$$\ln Bigdata_{ijpt} = \beta_0 + \beta_1 X_{ijp,t-1} + \delta_{pt} + \gamma_{jt} + \mu_i + \eta_{ijpt} \quad (2)$$

其中，i 代表企业，t 代表年份，p 代表省份，j 代表行业；模型（1）的被解释变量为公司年报中是否提及大数据相关关键词的虚拟变量（*Bigdata_dummy*）；模型（2）的被解释变量为公司年报中披露大数据相关关键词的频率加一后取对数（ln*Bigdata*）。$X_{ijp,t-1}$ 代表一系列可能影响公司的大数据应用程度的滞后一期变量。其中公司规模（ln*Assets*）、年龄（ln*Age*）、杠杆率（*Lev*）、固定资产比率（*PPE_TA*）、销售收入增长率（*SalesGrowth*）、总资产收益率（*ROA*）、股权的国有性质（*SOE*）的定义与本文数据部分一致。公司治理变量包括第一大股东持股比例（*Sh*1），第一大股东持有股数占总股数的比重；董事会规模（*Board_Size*），董事会人数取对数。随时间变化的省级层面的因素包括：公司注册地所在省份的教育程度（*Education*），即大专以上学历人数占比；市场化程度（*Fangang_index*），即王小鲁等（2019）的地区市场化指数；GDP 年度增速（*GDP_ret*）。

模型（1）中的 Ind_j 代表行业固定效应，①$Year_t$ 代表年份固定效应；模型（2）还控制了省份—年的固定效应（δ_{pt}）和行业—年的固定效应（γ_{jt}）以控制潜在的公司所处地区经济环境和行业发展等因素的影响，控制了公司层面的固定效应（μ_i）以控制公司不随时间变化的特质性；η_{ijpt} 代表随机误差项，标准误在企业和年份层面上进行双向聚类。

表 3 汇报了大数据应用程度的决定性因素的分析结果。由于大数据投入在时间上可能存在序列相关性，第 1 列和第 2 列仅保留公司首次在年报中披露“大数据”相关关键词的年份及其之前年份的观测值，剔除了首次披露之后年份的数据，考察了公司开启大数据相关投资的决定性因素。被解释变量为大数据披露虚拟变量（*Bigdata_dummy*），采用 Probit 模型进行估计。第 1 列控制了行业层面和省份层面的固定效应，第 2 列仅控制了行业层面的固定效应，加入省份层面滞后一期的相关指标。第 3 列报告了基于大数据应用程度的连续变量（ln*Bigdata*）和全样本的分析。综合表 3 的结果发现：规模较大的公司、有形资产占比低、盈利能力强的公司更可能应用大数据；所在地区市场化程度越高，上市公司越可能应用大数据。

① 行业分类的依据是证监会行业划分标准：金融、公用事业、房地产、综合、工业和商业。表 3 的第（1）列没有加入省级层面的因素，因此还加入了省份层面的固定效应。

表 3　　大数据应用程度的决定性因素

	probit		OLS
变量	*Bigdata_dummy*		ln*Bigdata*
	(1)	(2)	(3)
ln*Assets*	0.0831***	0.0856***	0.0947***
	(0.0176)	(0.0175)	(0.0215)
ln*Age*	-0.0143***	-0.0142***	-0.0456
	(0.00312)	(0.00302)	(0.0716)
Lev	0.0559	0.0425	0.00855
	(0.110)	(0.108)	(0.0529)
PPE_TA	-0.656***	-0.675***	-0.109*
	(0.112)	(0.111)	(0.0576)
SalesGrowth	0.0422	0.0478	0.0141
	(0.0385)	(0.0382)	(0.0101)
ROA	0.921**	0.905**	0.239*
	(0.406)	(0.403)	(0.131)
SOE	-0.0994**	-0.0929**	0.00205
	(0.0418)	(0.0410)	(0.0282)
*Sh*1	0.0446	0.0555	-0.132
	(0.116)	(0.115)	(0.0759)
Board_Size	-0.0896	-0.0817	0.0133
	(0.0884)	(0.0878)	(0.0478)
Education		0.170	
		(0.221)	
Fangang_index		0.0285***	
		(0.0110)	
GDP_ret		-1.125	
		(1.275)	
固定效应	Year, Industry, Province	Industry, Year	Firm, Province-Year, Industry-Year
观测值	14962	14962	17567
pseudo/adjusted R^2	0.150	0.145	0.498

注: *、**、*** 分别表示在 10%、5%、1% 的水平上显著。以下同。第(1)列和第(2)列汇报的系数为边际影响,即解释变量每增加一单位对大数据应用概率的影响。

五、大数据应用对公司市场价值的影响

(一) 基本回归分析

本文采用以下基准模型检验大数据应用程度对公司在股票市场上价值的影响: ①

① 考虑到不同的大数据应用程度下,大数据对公司市场价值的边际影响可能存在差异。本文采用如下两种方法验证了基准模型设计的可靠性: (1) 选用面板门限模型(Hansen,2000) 判断大数据应用程度与公司市场价值之间不存在分段的线性关系; (2) 基于全样本将大数据应用指标的取值区间等分,验证不同区间内的影响是线性递增的。限于篇幅,该部分结果留存备索。

$$Y_{ijpt} = \gamma_0 + \gamma_1 BigData_{ijpt} + \gamma_2 Controls_{ijpt} + \delta_{pt} + \gamma_{jt} + \mu_i + \xi_{ijpt} \quad (3)$$

其中，Y_{ijpt} 为第 t 年 p 省份 j 行业的 i 公司市场价值指标，即托宾 Q 值（*Tobin's Q*）；核心解释变量 $BigData_{ijpt}$ 表示公司层面的大数据应用程度（ln*Bigdata*），即公司年报中披露大数据相关关键词的频率加一后取对数；$Controls_{ijpt}$ 代表相关控制变量，包括企业的规模、杠杆率、固定资产比率、年龄、股权的国有性质、销售收入增长率和总资产收益率。δ_{pt} 表示“省份—年份固定效应”以控制不同地区随时间变化的特征，γ_{jt} 表示“行业—年份固定效应”以控制不同行业随时间变化的特征，μ_i 代表公司固定效应。ξ_{ijpt} 代表随机误差项，标准误在企业和年份层面上进行双向聚类。我们重点关注系数 γ_1 的符号及其显著性，其经济含义是大数据应用对公司市场价值的影响。

表 4　　大数据应用与公司市场价值的基本回归结果

变量	Tobin's Q			
	（1）	（2）	（3）	（4）
ln*Bigdata*	0.157***	0.154***	0.153***	0.151***
	（0.0375）	（0.0355）	（0.0354）	（0.0337）
控制变量	Yes	Yes	Yes	Yes
固定效应	Year，Firm	Firm，Industry–Year	Firm，Province–Year	Firm，Industry–Year，Province–Year
观测值	20623	20623	20623	20623
adjusted R^2	0.695	0.698	0.699	0.702

注：由于篇幅限制，本文省略控制变量系数的报告，有需要的读者可以联系作者索取，下表同。

表 4 报告了模型（3）估计结果。由于回归系数与标准误的估计结果可能较大程度上受到固定效应的影响，第 1 列至第 4 列考察不同固定效应的控制对研究结论的影响，以确保结果的稳健性。第 1 列至第 4 列中 ln*Bigdata* 的系数均在 1% 的水平上显著为正，即公司的大数据应用程度越大，其市场价值越高。可以看出，大数据应用显著提高公司的市场价值这一结论非常稳健。以第（4）列为例，披露了一个“大数据”相关关键词的公司的托宾 Q 值，比没有披露任何相关关键词的公司平均而言高出 4.21%（ln2 × 0.151/2.485）。这说明大数据应用对公司市场价值的影响在经济意义上也是显著的。总体而言，表 4 的研究结果表明公司应用大数据越多，其在股票市场上的价值越高。考虑到研究结果的可靠性，下文的回归模型都以控制最严格的固定效应为基准。

（二）内生性问题的处理

本文主要研究大数据应用对公司市场价值的影响，但利用上述回归模型来识别因果关系可能存在内生性问题。就本文研究问题而言，内生性主要有以下来源：第一，反向因果。绩效好、估值高的公司现金流充足，外部融资成本低，更有能力去付出较高成本来投入大数据技术和搭建数据平台等，进而导致回归估计系数被高估。第二，遗漏变量，可能存在难以观测的因素同时与公司的大数据应用和股票市场价值相关。例如，应用大数据的公司可能处在快速发展时期或者管理层有前瞻性进而带动公司的成长性，造成回归估计系数被高估；但公司如果面临较少的增长机会或者预期到业绩的负面冲击，也会因较低的机会成本而进行数字化转型，造成回归估计系数被低估。总体而言，OLS 回归系数估计偏误的方向从理论上并不明确。

本文构建工具变量来缓解可能存在的内生性问题。工具变量的设计基于 2009 年启动的“基础

学科拔尖学生培养试验计划”,又称“珠峰计划”,由教育部联合中组部、财政部发起,旨在培养拔尖创新人才,推动高等教育改革,推动创新型国家的建设。① 该计划第一批选择了 17 所高校②的数学、物理、化学、生物、计算机科学等理工科相关专业作为试点。宋弘和陆毅(2020)发现该计划提高了高校毕业生选择理工类职业的概率,有效增加了理工类人才供给。

本文主要关注的解释变量是大数据应用程度指标(ln*Bigdata*),人力资源成本是大数据决策的关键。理工科专业培养了与大数据应用息息相关的数理能力、编程能力和工程设计思维,因此理工科专业人才在大数据应用中发挥着至关重要的作用。以往的业界经验和学术研究都证实了理工类技术人才的匮乏是公司采用大数据或人工智能技术的瓶颈(Tambe, 2014; Babina et al., 2021)。“珠峰计划”第一批试点针对人群为理工科学生,它的实施增加了理工类人才供给,进而有效地缓解了试点大学附近的大数据人才供给不足的问题,降低了附近公司应用大数据的劳动力成本。

该计划的第一批试点于 2009 年正式启动,大部分高校于 2010 年正式开始实施,因此受该计划影响的第一批学生大部分是 2010 年入学的本科生,受影响学生毕业的年份为 2014 年及之后的年份。③ 2014 年大学生毕业的时点为年中,该年份难以完全被归为受影响之前或之后的时间段,因此本文在工具变量回归分析中均将 2014 年的观测值剔除。工具变量(IV_{it})构造如下:

$$IV_{it} = \frac{\ln\left(\sum_{k=1}^{n}\frac{1}{distance_{ik}}\right)}{N_c} \times Post_t$$

其中,i 代表公司,k 代表第一批试点高校,c 代表公司办公地点所在的城市;$distance_{ik}$表示通过上市公司 i 办公地点的经纬度与高校 k 主校区的经纬度计算的直线距离(单位为公里);n 为第一批试点高校的数目($n=17$);N_c 表示 2014 年公司 i 的办公地点所在城市 c 中的上市公司的总数量;$Post_t$ 为时间虚拟变量,2014 年之后设为 1,2014 年之前则设为 0。根据工具变量的设计,上市公司与试点大学之间的距离越近(即 $distance_{ik}$越小),该公司受到政策的辐射力度越大,越有可能提高大数据应用程度;而上市公司办公地点所在城市的上市公司数目越多(即 N_c 越大),则上述辐射力度越可能被削弱。因此,该工具变量从理论上满足工具变量的相关性要求。此外,该工具变量很难通过其他渠道对公司层面的市场价值产生影响,理论上满足排他性要求。

表 5 报告了工具变量回归的结果和相关检验。由于工具变量的变动很大程度上依赖于公司所在地,因此第 1—2 列没有控制“省份—年”固定效应;第 3—4 列控制了全面的固定效应。两者结果一致。以第 3—4 列为例,第 3 列报告了第一阶段的回归结果,即核心变量 ln*Bigdata* 回归到工具变量上,发现工具变量的系数显著为正。这表明在政策之后,距离试点高校越近的上市公司对大数据应用程度越高,与预期相符。④ 第一阶段的 Cragg-Donald F 统计量为 54.313,Kleibergen-Paap rk F 统计量为 26.778,远高于 Stock-Yogo 弱工具变量检验(零假设是弱工具变量)的 10% 临界值 16.38,说明工具变量满足相关性假设。此外,工具变量与模型(3)的残差项的相关性非常微弱(相

① 具体举措包括改善目标院校的师资配备,为学生提供一流的学习条件,打造创新型培养模式,开展国际合作等。请参考《基础学科拔尖学生培养试验计划实施办法》。

② 第一批的 17 所试点高校包括:北京大学、清华大学、北京师范大学、南开大学、吉林大学、复旦大学、上海交通大学、南京大学、中国科学技术大学、浙江大学、厦门大学、山东大学、武汉大学、中山大学、四川大学、西安交通大学、兰州大学。

③ 该计划的第二批试点开始于 2015 年,受影响的学生最早毕业于 2019 年,不在本文的样本范围内,因此本文只考虑第一批试点的影响。

④ 为了进一步检验工具变量的作用渠道,本文验证了“珠峰计划”对企业雇佣研发人员占比的促进作用,说明本文所采用的工具变量通过影响企业的劳动力来影响大数据应用。限于篇幅,没有汇报相关结果,留存备索。

关系数 0.018)。第二阶段回归中大数据应用指标的系数仍然在 5% 水平上显著为正,即处理了内生性问题后,大数据应用提高公司市场价值的基本结论不变。①

表 5　　工具变量回归结果

变量	第一阶段结果	第二阶段结果	第一阶段结果	第二阶段结果
	ln*Bigdata*	Tobin's Q	ln*Bigdata*	Tobin's Q
	(1)	(2)	(3)	(4)
IV	0.076***		0.060***	
	(0.013)		(0.012)	
ln*Bigdata*		1.333**		1.420**
		(0.593)		(0.630)
控制变量	Yes	Yes	Yes	Yes
固定效应	Firm, Industry-Year	Firm, Industry-Year	Firm, Industry-Year, Province-Year	Firm, Industry-Year, Province-Year
观测值	17001	17001	17001	17001
弱工具变量检验				
F 统计量 (Cragg-Donald)	106.734		54.313	
F 统计量 (Kleibergen-Paap rk)	32.584		26.778	

(三) 其他稳健性检验

为了进一步验证结果的可靠性,本文从如下多个角度进行了稳健性检验。②

更换大数据衡量指标。首先,词频数据可能存在一定的噪音。为此,本文将 ln*Bigdata* 更换为虚拟变量 *Bigdata_dummy*,代表公司在年报中是否披露了大数据相关的关键词。此外,我们构造分类变量,该变量取值为 0 表示公司年报中没有披露任何大数据相关关键词;如果公司年报中披露了相关关键词,则分年份按照词频由小到大排序并三等分,前 1/3 观测值的分类变量取值为 1,中间 1/3 取值为 2,后 1/3 取值为 3。其次,我们还关注了大数据词频密度,即"大数据相关关键词在公司年报中出现的总次数除以该年报的总词汇量(单位:百)",以反映大数据信息在年报中出现的密度。再次,还考量了更为广义的关键词集合的词频变量,③即除了大数据自身定义之外,还包括数据收集、处理和应用相关的技术、平台和资源(Saunders & Tambe, 2013; Srinivasan & Chen, 2020)。最后,为了避免所选关键词的行业特质性的干扰,进一步开展如下工作:一是在词典中剔除与行业相关的关键词,即公司所在行业出现的平均次数最高的一个关键词;二是剔除大数据应用指标中的行业趋势,将 ln*Bigdata* 减去所在行业当年的 ln*Bigdata* 均值(demean by industry-year),从而剔除"行业—年"层面的共同度量偏差。上述更换大数据衡量指标的结果均稳健。

更换研究样本。第一,为了验证信息技术类行业同样适用于本文的分析结论,我们将 A 股全部行业的上市公司作为研究样本。第二,由于金融业公司的资本结构和盈利模式与其他公司存在

① 本文对工具变量模型的系数估计值进行了更深入的解读,限于篇幅不再赘述,留存备索。

② 限于篇幅,稳健性检验的实证结果及其具体分析没有汇报,留存备索。

③ 限于篇幅,我们未报告广义的大数据关键词词典,留存备索。

较大差异,我们在基准样本的基础之上剔除金融业公司。最后,为了避免企业因为追逐热点而在年报中虚假披露大数据信息,我们仅使用大数据发展早期的样本(即 2014 年及之前的数据) ①进行分析。上述所有更换研究样本的方案取得的结果均与结论一致。

更换聚类方式。为了检验不同的聚类方式下回归结果的稳健性,本文进一步分别采用公司层面和"省份—行业"层面的聚类方式,结果均稳健。

六、进一步的讨论

(一) 大数据应用对公司市场价值影响机制的检验

1. 生产效率和研发投入

为了检验生产效率提高的渠道是否有效,借鉴以往文献(Liu & Lu, 2015; Liu & Qiu, 2016) ,本文采用全要素生产率(*TFP*) 来衡量公司的生产效率,考察大数据应用程度对公司生产效率的影响。本文采用当前主流的 LP 方法(Levinsohn & Petrin, 2003) 估算公司的全要素生产率。具体而言,按照证监会《上市公司行业分类指引(2012 年版) 》中的二级行业划分标准,本文利用销售收入、员工数目、固定资产和中间投入的信息,分行业来估计 *TFP*。将 *TFP* 作为被解释变量执行模型(3) 的回归,表 6 中第 1 列的结果表明: ln*Bigdata* 的估计系数显著为正。这表明大数据的应用帮助公司显著提高了自身的生产效率。

表 6　　大数据应用对公司市场价值的正向影响机制

被解释变量	*TFP*	*R&D_Exp/Assets*	*R&D workers*(%)
	(1)	(2)	(3)
ln*Bigdata*	0. 0245 **	1. 503 ***	1. 710 ***
	(0. 0108)	(0. 477)	(0. 298)
控制变量	Yes	Yes	Yes
固定效应	Yes	Yes	Yes
观测值	20542	20623	20621
adjusted R^2	0. 989	0. 699	0. 656

注: 所有回归均控制了公司固定效应、"省份—年份"固定效应和"行业—年份"固定效应。

表 6 中第(2) 列和第(3) 列汇报了大数据应用对公司研发强度的影响。本文从 R&D 的资金投入和人员投入两个角度来衡量公司的研发强度: 第(2) 列的被解释变量是 R&D 支出除以滞后一期的总资产(*R&D_Exp/Assets*) ,其中总资产以千计; 第(3) 列中的被解释变量是研发人员在所有员工中所占比例(*R&D workers*) 。同样采用模型(3) 进行回归,结果发现大数据应用程度的提高对公司的研发强度有显著的正向影响,从而印证了影响渠道: 大数据应用通过促进公司研发来提高其市场价值。

2. 大数据应用中的摩擦

根据本文对大数据与实体企业融合的作用机理的分析,大数据应用过程中的摩擦主要来源于两个方面: 一是从人才需求的角度,公司内部技术人员储备不足,外部劳动力市场的相关人才供给不足; 二是从技术支持的角度,公司所在地区的数据相关技术服务与基础设施建设不足。本部分分别对上述两类摩擦进行实证检验。

为验证人才需求方面的摩擦,我们分别从公司内部技术人员储备和公司所在地区的外部高学

① 参考何帆和刘红霞(2019) 的总结,大数据在中国的发展划分为萌芽期(2014 年之前) 、发展期(2015—2016 年) 和升华期(2017 年至今) 。因此,大数据在 2014 年之前尚未成为热点,这与本文图 1 的数据趋势相符。

历劳动力供给两个层面进行检验。首先,我们分析在不同的技术人员储备的子样本下,大数据应用对公司价值影响的差异,从而体现公司内部技术人员储备带来的摩擦。具体而言,基于公司在样本初期员工中技术或研发人员占比,行业内的中位数将公司样本划分为两组,并分别基于这两组子样本来估计模型(3)。结果在表7的第(1)—(2)列报告,第(1)列的 ln*Bigdata* 系数估计值为0.195且在5%置信水平下显著,而第(2)列的系数估计则不显著,两者之间的系数差异显著。因此,只有内部技术人员储备较为充足的公司能更好地将大数据转化为公司的市场估值。

表7　　大数据应用中的摩擦

被解释变量	Tobin's Q					
样本	内部的技术人员储备		外部的高学历人才供给		数据基础设施建设	
	>中值	<中值	>中值	<中值	国家大数据中心所在省份	其他省份
	(1)	(2)	(3)	(4)	(5)	(6)
ln*Bigdata*	0.195**	0.0911	0.185***	0.111**	0.204**	0.142***
	(0.078)	(0.057)	(0.050)	(0.040)	(0.0661)	(0.0344)
控制变量	Yes	Yes	Yes	Yes	Yes	Yes
固定效应	Yes	Yes	Yes	Yes	Yes	Yes
difference	0.104(p<0.01)		0.0741(p<0.01)		0.0618(p<0.01)	
观测值	6050	6776	9904	10512	1835	18778
adjusted R^2	0.667	0.709	0.699	0.701	0.737	0.700

注:所有回归均控制了公司固定效应、"省份—年份"固定效应和"行业—年份"固定效应。difference 为系数差异及对应的检验显著性结果,经验 p 值的计算基于 Cleary(1999)。

其次,我们利用不同地区之间的高学历劳动力供给情况的差异,分析基准效应的异质性。在大数据与实体经济的融合过程中,劳动力与组织管理的同步革新离不开当地劳动力市场中高素质人才的供给(Tambe, 2014; Babina et al., 2021)。因此,我们根据不同省份劳动力市场供给条件的差异,进行分样本检验。具体而言,我们计算了公司所在省份的大专以上学历的人口占比,按照每年的中位数将公司样本划分为两组,并分别基于这两组样本来估计模型(3),结果在表7的第(3)列与第(4)列中报告。结果表明,第(3)列的 ln*Bigdata* 系数估计值为0.185且在1%置信水平下显著;而高学历劳动力供给较少的地区第(4)列对应的系数仅为0.111。两者之间的系数差异显著(p值<0.01)。因此,高学历劳动力供给充足的地区的公司能够更好地将大数据转化为公司的市场价值。

为了验证技术支持方面的摩擦,我们以公司所在省份是否拥有"国家大数据中心(包括北京、贵州、乌兰察布)"作为所在环境的数据基础设施建设情况的度量,将所有样本划分为位于"国家大数据中心所在省份"和"其他省份"的两组子样本,并分别基于这两组样本来估计模型(3),结果在表7的第(5)列与第(6)列中报告。结果表明,位于数据基础设施建设好的地区的公司能够更好地实现大数据对公司的增值效果,两组子样本的系数估计值之间差异显著(p值<0.01)。

(二)异质性影响

不同类型公司的大数据储备有所不同,大数据应用的成本也存在差异,将大数据转化为企业价值的动力也不尽相同。本部分从公司规模、国有性质、行业竞争程度探讨大数据应用对公司市场价值影响的异质性。限于篇幅,异质性回归的结果留存备索。

大数据应用对不同规模公司的影响可能存在差异。Begenau et al.(2018)发现,大公司的数据更为丰富,因此基于大数据的预测更精确,从而降低生产的不确定性和融资成本。因此,大公司的大数据应用可能对市场价值的提升效果更好。但是,相对于大公司,小公司受到分析师的追踪分析

和投资者的关注不足,自身生产经营中和外部资本市场上的信息不对称程度更为严重(Botosan, 1997) 。因此,大数据的应用对于小公司获得更多信息来提高预测精度的意义更为重大。而 Farboodi et al. (2019) 则认为公司的初始规模并不是决定其是否成功的关键因素。为了厘清上述逻辑,我们按照总资产是否大于行业中位数构造虚拟变量,①并在模型(3) 中引入它与大数据应用指标的交乘项。结果发现,小公司更能从大数据应用中实现市场价值的提高。

大数据应用对国有公司和非国有公司的市场价值可能会产生不同的影响,原因在于这两类公司在经营目标、外部经营环境和公司内部治理等方面存在显著差异: (1) 在经营目标方面,非国有公司更加注重追求经济效益(姚洋和章奇,2001) ,而国有公司会更注重社会和政治目标(林毅夫等,2004) 。因此非国有公司会更有动力利用数据挖掘信息和提高生产效率,进而获得更高的市场估值; 而国有企业的大数据信息披露更可能是出于政策性目的,可能不会影响公司的市场价值。(2) 在外部经营环境方面,国有公司相比于非国有公司面临更优越的外部经营环境,包括更易获得的银行贷款和更多政策优惠(Allen et al., 2005) ,因此挖掘大数据背后价值的动力不足。(3) 从公司的内部治理角度,由于没有明确的所有者,国有企业容易被内部人控制,因此公司治理水平更弱(钱颖一,1999) ,进而导致国有企业的投资效率低,大数据无法发挥为公司增值的效果。我们在模型(3) 中加入大数据应用与国有性质虚拟变量的交乘项,结果发现大数据应用对非国有公司市场价值的提升作用较大且在统计意义上显著,但是对国有公司的影响很小且为负,与上述推论完全一致。因此,大数据应用可能会加大非国有公司和国有公司之间的效率和估值差异。

不同行业层面的竞争程度下,大数据应用对企业的影响也可能有所不同。一方面,根据 Coibion et al. (2018) 的研究,竞争更激烈的行业中的公司拥有更多市场环境的信息,也更愿意提高对信息的挖掘程度和利用效率,因此它们在应用大数据后能够更准确地预测各个经济变量。另一方面,Melville et al. (2007) 发现处于竞争压力下的企业会更多地利用信息技术类资产(例如大数据分析) 来开拓新的商业模式、决策过程等,从而实现创新能力和生产力的提高。因此,竞争激烈行业的公司可能更有动力利用大数据来提高生产效率,实现企业增值。我们以证监会分类的二级行业内所有样本行业主营业务利润率标准差的倒数来衡量该行业的竞争程度(Nickell,1996) ,按照中位数生成虚拟变量,与大数据应用指标交乘后加入模型(3) 中进行估计。结果与推论一致: 交乘项系数显著为正,即大数据应用对公司市场绩效的提升效果在竞争行业中更为显著。

七、结　论

本文利用中国 A 股上市公司披露的年报,抓取“大数据”相关的关键词,构造了公司层面的大数据应用程度的衡量指标,并基于此描述了大数据在中国上市公司中的应用情况,探究了公司大数据应用的影响因素。重要的是,本文研究了大数据应用对公司市场价值的影响及其作用机制和异质性,为大数据与实体企业融合的经济效应提供了实证依据。本文发现,不同公司在生产经营过程中应用大数据的概率有所不同。规模较大、有形资产比例较低、盈利能力较强,以及所在地区市场化程度较高的公司更可能应用大数据。而大数据应用能够显著提高公司的市场价值。本文通过利用 2009 年启动的“基础学科拔尖学生培养试验计划”来构造工具变量,缓解了内生性问题,得到了一致结论。机制分析表明,大数据的应用显著提高了公司的生产效率和研发投入,进而促进了公司市场价值的提升,而技术和人才供给的匮乏可能会阻碍大数据应用对企业的增值效果。大数据应

① 如果选择基期(2006 年) 作为标准会造成较大程度的样本缺失,而 2010 年及之前的年份大数据应用率极低,因此我们选取 2010 年作为刻画公司截面特征的时间点。我们也利用类似的方法检验了公司年龄维度的异质性,发现不同年龄公司之间的差异并不显著。限于篇幅,该结果不再赘述,留存备索。

用对不同公司市场价值的积极影响在小规模公司、非国有公司和竞争激烈的行业中尤为显著。本文的研究结论对于中国企业的大数据应用与政策设计有以下几点启示。

第一,大数据的应用切实提高了市场价值,推动企业数字化转型的意义深远。“十四五”时期的目标任务明确提出:“加快数字化发展,打造数字经济新优势,协同推进数字产业化和产业数字化转型,加快数字社会建设步伐,提高数字政府建设水平,营造良好数字生态,建设数字中国。”通过本文研究可知,大数据在实体企业生产经营中的应用程度逐渐提高,大数据应用切实提高了公司的生产效率和研发投入,其价值得到了投资者的认可,对公司长期竞争力的提高具有战略性意义。因此,以大数据为支点撬动生产方式和治理方式的变革势在必行,政策制定应当为企业提供充足的数字化转型动能。

第二,技术和人才的供给是企业高效利用大数据的关键所在。如何为企业提供充足的数字化转型动能,帮助企业克服利用大数据中存在的诸多摩擦,是政策改革需要关注的核心问题。一方面,技术水平的提高能够大大降低企业应用大数据的成本,例如云计算体系的推广降低了企业处理、分析和分享数据的成本。因此,加快新型基础设施的建设,降低应用大数据的技术成本,可以为企业的大数据转型之路保驾护航。另一方面,大数据相关人才的供给存在缺口。大数据应用融入生产环节需要在每个生产部门都配备数据建模人员、数据分析员等,但我国高等教育体系的数据相关人才培养却难以在短期内满足需求。因此,从政策层面要加快高等教育体系改革,调整人才培养结构,加大数据人才的培养力度,以满足日益增长的数字化人才的需求。

第三,不同公司与大数据的融合深度和价值转化效率存在差异。例如,竞争性不强的行业可能因为缺乏生产力提高的动力而难以发挥大数据的价值;国有企业因其优越的外部环境、经营目标的多样性和公司治理水平不足导致大数据的价值无法得到市场认可。有鉴于此,政府在进一步制定和实施大数据鼓励政策时,需要针对不同性质的公司给出更为细致的指导意见。

参考文献

何帆、刘红霞,2019:《数字经济视角下实体企业数字化变革的业绩提升效应评估》,《改革》第 4 期。

林毅夫、刘明兴、章奇,2004:《政策性负担与企业的预算软约束:来自中国的实证研究》,《管理世界》第 8 期。

钱颖一,1999:《激励与约束》,《经济社会体制比较》第 5 期。

宋弘、陆毅,2020:《如何有效增加理工科领域人才供给?——来自拔尖学生培养计划的实证研究》,《经济研究》第 2 期。

王小鲁、樊纲、胡李鹏,2019:《中国分省份市场化指数报告(2018)》,社会科学文献出版社。

吴超鹏、唐苟,2016:《知识产权保护执法力度、技术创新与企业绩效——来自中国上市公司的证据》,《经济研究》第 11 期。

姚洋、章奇,2001:《中国工业企业技术效率分析》,《经济研究》第 10 期。

Agrawal, A., J. S. Gans, and A. Goldfarb, 2019, "Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction", *Journal of Economic Perspectives*, 33 (2), 31—49.

Allen, F., J. Qian, and M. Qian, 2005, "Law, Finance, and Economic Growth in China", *Journal of Financial Economics*, 77 (1), 57—116.

Babina, T., A. Fedyk, A. X. He, and J. Hodson, 2021, "Artificial Intelligence, Firm Growth, and Industry Concentration", Working Paper.

Bajari, P., V. Chernozhukov, A. Hortaçsu, and J. Suzuki, 2019, "The Impact of Big Data on Firm Performance: An Empirical Investigation", *AEA Papers and Proceedings*, 33—37.

Begenau, J., M. Farboodi, and L. Veldkamp, 2018, "Big Data in Finance and the Growth of Large Firms", *Journal of Monetary Economics*, 97, 71—87.

Bharadwaj, A. S., S. G. Bharadwaj, and B. R. Konsynski, 1999, "Information Technology Effects on Firm Performance as Measured by Tobin's Q", *Management Science*, 45 (7), 1008—1024.

Bloom, N., R. Sadun, and J. Van Reenen, 2012, "Americans Do IT Better: US Multinationals and the Productivity Miracle", *American Economic Review*, 102 (1), 167—201.

Botosan, C. A., 1997, "Disclosure Level and the Cost of Equity Capital", *Accounting Review*, 323—349.

Brynjolfsson, E., and T. Mitchell, 2017, "What Can Machine Learning Do? Workforce Implications", *Science*, 358 (6370),

1530—1534.

Brynjolfsson, E., and K. McElheran, 2016, "The Rapid Adoption of Data-driven Decision-making", *American Economic Review*, 106 (5), 133—139.

Brynjolfsson, E., D. Rock, and C. Syverson, 2021, "The Productivity J-Curve: How Intangibles Complement General Purpose Technologies", *American Economic Journal: Macroeconomics*, 13 (1), 333—372.

Brynjolfsson, E., L. Hitt, and H. Kim, 2011, "How Does Data-driven Decision-making Affect Firm Performance", Workshop for Information Systems and Economics at the 9th Annual Industrial Organization Conference, April. 8—10.

Chen, H. C., R. H. L. Chiang, and V. C. Storey, 2012, "Business Intelligence and Analytics: From Big Data to Big Impact", *MIS Quarterly*, 36(4), 1165—1188.

Cleary, S., 1999, "The Relationship between Firm Investment and Financial Status", *Journal of Finance*, 54 (2), 673—692.

Cockburn, I. M., R. Henderson, and S. Stern, 2019, "The Impact of Artificial Intelligence on Innovation", The Economics of Artificial Intelligence: An Agenda, 115.

Coibion, O., Y. Gorodnichenko, and S. Kumar, 2018, "How Do Firms Form Their Expectations? New Survey Evidence", *American Economic Review*, 108 (9), 2671—2713.

Dranove, D., C. Forman, A. Goldfarb, and S. Greenstein, 2014, "The Trillion Dollar Conundrum: Complementarities and Health Information Technology", *American Economic Journal: Economic Policy*, 6 (4), 239—270.

Farboodi, M., R. Mihet, T. Philippon, and L. Veldkamp, 2019, "Big Data and Firm Dynamics", *AEA Papers and Proceedings*. 38—42.

Griliches, Z., 1957, "Hybrid Corn: An Exploration in the Economics of Technological Change", *Econometrica*, 25 (4), 501—522.

Hansen, B. E., 2000, "Sample Splitting and Threshold Estimation", *Econometrica*, 68 (3), 575—603.

Holmström, B. R., 1989, "Agency Costs and Innovation", *Journal of Economic Behavior & Organization*, 12 (3), 305—327.

Levinsohn, J., and A. Petrin, 2003, "Estimating Production Functions Using Inputs to Control for Unobservables", *Review of Economic Studies*, 70(2), 317—341.

Liu, Q., and L. D. Qiu, 2016, "Intermediate Input Imports and Innovations: Evidence from Chinese Firms´ Patent Filings", *Journal of International Economics*, 103, 166—183.

Liu, Q., and Y. Lu, 2015, "Firm Investment and Exporting: Evidence from China's Value-added Tax Reform", *Journal of International Economics*, 97 (2), 392—403.

McAfee, A., and E. Brynjolfsson, 2012, "Strategy & Competition Big Data: The Management Revolution", *Harvard Business Review*, 90 (10), 60—68.

Melville, N., V. Gurbaxani, and K. Kraemer, 2007, "The Productivity Impact of Information Technology across Competitive Regimes: The Role of Industry Concentration and Dynamism", *Decision Support Systems*, 43 (1), 229—242.

Midrigan, V., and D. Y. Xu, 2014, "Finance and Misallocation: Evidence from Plant-level Data", *American Economic Review*, 104 (2), 422—458.

Mikalef, P., I. O. Pappas, J. Krogstie, and M. Giannakos, 2018, "Big Data Analytics Capabilities: A Systematic Literature Review and Research Agenda", *Information Systems and e-Business Management*, 16 (3), 547—578.

Morck, R., A. Shleifer, and R. W. Vishny, 1988, "Management Ownership and Market Valuation: An Empirical Analysis", *Journal of Financial Economics*, 20, 293—315.

Nickell, S. J., 1996, "Competition and Corporate Performance", *Journal of Political Economy*, 104 (4), 724—746.

Saunders, A., and P. Tambe, 2013, "A Measure of Firms´ Information Practices Based on Textual Analysis of 10 - K Filings", Working Paper.

Srinivasan, S., and W. Chen, 2020, "Going Digital: Implications for Firm Value and Performance", Working Paper.

Tambe, P., 2014, "Big Data Investment, Skills, and Firm Value", *Management Science*, 60 (6), 1452—1469.

Tanaka, M., N. Bloom, J. M. David, and M. Koga, 2020, "Firm Performance and Macro Forecast Accuracy", *Journal of Monetary Economics*, (114), 26—41.

Wu, L., L. Hitt, and B. Lou, 2020, "Data Analytics, Innovation, and Firm Productivity", *Management Science*, 66 (5), 2017—2039.

Zhu, C., 2019, "Big Data as a Governance Mechanism", *Review of Financial Studies*, 32 (5), 2021—2061.

Effects of Big Data on Firm Value in China: Evidence from Textual Analysis of Chinese Listed Firms' Annual Reports

ZHANG Yeqing[a], LU Yao[b] and LI Leyun[c]

(a: Institute of Finance and Economics, Central University of Finance and Economics;
b: School of Economics and Management, Tsinghua University; c: Fu Foundation School of Engineering and Applied Science, Columbia University in the City of New York)

Summary: As the global data volume explodes and the data processing technologies advance, it becomes the new growth opportunity for the Chinese economy to use big data in firm operations. The growth scale of China's digital economy accounted for 38.6% of the GDP in 2020, which can be primarily attributed to many forward-looking policies for big data development by the government. However, the basic situations, determinants, and real effects of using big data in firms are still vague. Therefore, this paper attempts to gain insights into these important issues by providing empirical evidence based on Chinese firms.

This paper first analyzes the annual reports of Chinese listed firms from 2006 to 2017 and constructs a variable measuring firm-level use of big data. Specifically, we use the Python program to crawl keywords related to big data in annual reports and count the frequencies of these words. Based on the variable, we depict the dynamic changes of big data adoption in Chinese listed firms and examine its determinants by using both Probit and OLS regressions. We find that firms of larger sizes, with lower proportions of tangible assets and higher profitability, and firms located in provinces with higher marketization levels are more likely to apply big data in the production and operation.

Then, we show that the use of big data significantly improves firm value, measured as Tobin's Q. To release the potential endogenous issue, we use the experimental program for training top-notch students in basic subjects in 2009 as an exogenous shock and construct an instrumental variable. The baseline results are robust when we use alternative measures of big data, alternative samples, and alternative clustering options. We further explore the main mechanisms and show that the use of big data significantly increases firms' productivity and R&D investment. Meanwhile, the supply of related technologies and talents is the key to magnify the value-added effect of big data. Furthermore, heterogeneous analyses show that the positive effect of the use of big data on firm value is more significant for small firms, non-SOE firms and firms in highly competitive industries.

Our study makes four major contributions. (1) We use the unstructured text data to construct a measure of firms' use of big data. Compared with previous studies, our measure is more direct, more effective, and more accurate, and can be applied to firms in all industries. It depicts the dynamic trend and determinants of the big data development of Chinese listed firms, laying a solid data foundation for subsequent research. (2) We empirically test the positive effect of the use of big data on firm value and explore its mechanisms. Our study provides evidence for big data's contribution to increasing firm competitiveness and boosting economic growth. (3) We examine the heterogeneous effects of big data adoption on firms' market value. These findings complement the relevant literature and provide evidence for the design of big data-related policies. (4) This paper enriches the literature on the determinants of firms' market value. That is, in the digital economy revolution, the use of big data is one of the important determinants of firms' valuation on the stock market.

This paper has the following policy implications for China's application of big data. (1) The use of big data indeed raises firm value and is meaningful for firms to stay competitive. Therefore, policy design should provide firms with sufficient momentum for digitalization. (2) The supply of technologies and talents is the key to the efficient use of big data for firms. Thus, we should speed up the construction of digital infrastructure and reduce the cost of adopting big data. Moreover, we should accelerate the reform of the higher education system, adjust the talent training structure, and strengthen the training of data-related talents to meet the growing demand for data-related employees. (3) Firms with different characteristics differ in the effect of big data on firm value. Therefore, when designing and implementing policies to encourage the use of big data, the government should provide detailed suggestions tailored to firm characteristics.

Keywords: Big Data; Textual Analysis; Firm Value; Productivity; R&D Input

JEL Classification: C80, D21, L25

(责任编辑:陈小亮)(校对:王红梅)