中国数据资本（V2）

2024.12

本数据为《Data as a Growth Factor: An Empirical Inquiry》（Taoxiong Liu, Ruofei Li, Working Paper）所用的主要数据，系在刘涛雄、戎珂、张亚迪（2023）的基础上，改进对中国各省份数据资本的估计而形成。本估算基于《数据资本估算及对中国经济增长的贡献——基于数据价值链的视角》中提出的成本法估算思路（刘涛雄等，2023），作出以下四点优化和改进。

第一，在文本资料来源上，本研究利用了 1999 年、2015 年、2022 年三版《职业分类大典》作为主要分析文本，全面地考虑了三版《职业分类大典》内容与结构的差异，能够更加充分地体现我国职业结构与工作内容的动态更迭。

第二，在估算各职业数据工作时长占比时，本研究借助 2022 版《职业分类大典》中提供的官方标注的数字职业集合，通过训练文本分类模型来识别数据相关的职业，让数据相关职业的界定标准更加客观。具体而言，本研究构建的分类任务训练集包括 2022 年《职业分类大典》中官方标注的 97 个数字化职业（标签为 1）以及本研究人工标注的 97 个传统非数字化职业（标签为 0）。本研究采用高斯朴素贝叶斯（GNB）、随机森林（RF）以及逻辑回归（LR）三类模型作为潜在的分类器，通过十折交叉验证法选取表现最佳的分类器模型。将分类器模型应用至所有的职业后，预测标签为 1 的职业被本研究识别为数据相关职业，剩余的职业则被认为其工作内容与数据价值形成无关。

第三，在挑选全时数据相关工作（即计算文本相似度的目标集合）时，本研究考虑到了现实中全时数据职业较少，而目标集合的选择又很大程度上会影响最终文本相似度计算的结果。为了克服全时数据职业挑选取不当可能带来的偏误，本研究先挑选了 20 个最接近全时数据工作的职业，聚焦于职业工作内容中的任务描述，将涉及数据工作的具体任务作为语义相似度计算的目标集合，尽量避免职业描述中行业等无关信息对结果的影响。表 1 展示了本研究最终采取的目标文本集合。

表 1　作为目标文本的工作任务描述

工作类型	**工作任务描述**
第一阶段：数据采集	

信息采集	运用人造地球卫星及空间大地测量观测技术和方法,进行大地测量和数据处理;测量其他行星的形状及重力场,建立测量基准和控制网;进行大地测量成果检查与验收；监测、观测海洋水文、海洋气象、海洋大气、海洋水体、海洋底质、海洋生物体、海洋灾害及海上目标;进行区域地表水环境质量、城市饮用水源地水质、城市空气质量、酸沉降监测,农村环境和环境背景值监测,近岸海域、物理和生态环境监测;进行重点监控企业监督、监测和建设项目竣工环境保护验收、监测;使用验潮仪,进行潮位、潮时观测,填写潮汐观测记录簿;使用岸用光学测波仪,进行波形、周期、波高的观测和记录;使用表层水温表、盐度计等设备,测定表层海水温度、盐度并进行记录;观测和记录海冰的冰量、冰形、浮冰密度、漂流方向等特征;使用风速风向观测仪、气压计等仪器设备,观测和记录风向、风速、气压、温度、湿度和降水量等要素;观测和记录海面有效能见度和雾;使用卫星定位仪、水准仪、重力仪等仪器,进行天文、重力、三角、水准、精密导线测量的观测和记簿；进行全球定位系统(GNSS)接收机的观测、记录工作;进行外业观测成果资料整理、概算,提供测量数据;使用大中型飞行器观测平台,获取航空航天影像数据和遥感影像;布设野外控制点标志,进行野外控制点测量和地物、地貌等的调绘;区域网空中三角测量,加密供测图使用的控制点和数据;操作地面移动或固定的观测平台及遥感设备,获取目标的观测数据；收集地图制图的资料和数据;使用数字化仪、工程扫描仪等信息化设备,进行地图定向、地图数据采集、数据转换和比例尺变化等作业;使用工程测量仪器,进行控制测量、地形测量、规划测量、建筑工程测量、变形形变与精密测量、市政工程测量、水利工程测量、线路与桥隧测量、地下管线测量、矿山测量等专项测量;操作测量仪器,观测和记录海洋控制、水准、地形、水深、助航标志、障碍标志、障碍物、底质等；操作地面监控系统,操控无人飞行器或其他无人机设备,采集地表数据和航空影像数据;使用移动测量车、卫星定位仪、惯性导航系统等仪器和设备,行驶设计路线,采集地物的实景地理信息;使用激光扫描仪、立体测量摄影机等设备,获取地物的二维、三维及全景影像信息;使用卫星定位仪、数码相机和惯性导航系统,获取道路和导航兴趣点(POI)的位置信息和属性信息;采集、记录作业对象的地表自然要素、人文地理要素等属性信息;采集样品,处理、保存样品，分析样品,统计监测数据,编制实验报告;野外踏勘,观察、描述地质现象,进行地质剖面测量、样品采集、地质编录、地质填图;收集地质、物化探、遥感、测绘、测试等数据资料,进行资料综合解释;调查地下水资源情况,进行地下水动态监测;使用水准仪、全站仪、监测仪器仪表、自动化设备等监测仪器设备,观测水工建筑物变形、渗流、应力应变及环境量等；观测水位、降水量、蒸发量等项目,记录数据,计算和绘图;监测水质,现场采样处理,进行水质检测;采集、捕获数字文件及元数据；设计统计调查方案，组织实施统计调查;采集统计数据和资料;搜集和整理影像、资料和数据提取、固定电子数据;运用卫星、气象雷达、自动气象站等气象装备、设施获取大气以及陆地、海洋、空间等领域中与气象相关的物理过程、化学过程和生态过程信息;
采集设备维护	设计、研制、试验、计量、维修海洋调查、监测、观测、遥感遥测等海洋仪器设备；安装埋设变形测点装置、测压管、量水堰、渗压计、接(裂)缝和应力应变温度监测仪器等水工建筑物监测仪器及设施;保养、维护水工建筑物监测仪器仪表、自动化设备及设施;
采集方案设计	开发建设区域环境监测技术体系,提供区域环境监测网络和监测质量技术支持；根据作业要求,布设采集方案和线路
第二阶段：数据清洗与存储	
数据库建立与维护	研究、应用地理信息数据采集与集成的技术方法和工艺流程,指导作业人员对采集的数据进行标准化录入,建立地理信息数据库;维护、更新、管理地理信息系统(GIS)数据库;运行维护数据库系统;管理、维护并保障大数据系统稳定运行；安装配置数据库,并进行性能监控,故障诊断、排除等日常维护;提出并实施优化数据库性能及数据库集群方案；使用地理信息软件和工作平台,进行地理信息数据标准化录入,建立地理信息数据库,进行数据库逻辑检验和修改;进行数字文件的登记、分类、著录、编目、归档,转化为数字档案;标注和加工图片、文字、语音等业务的原始数

	据;
数据质量优化与数据资料整理	使用摄影测量工作站,进行影像数据的处理、几何纠正、影像判译、立体测图,绘制各种比例尺地形原图;使用遥感影像处理软件和图形工作站等,进行卫星遥感影像数据的纠正、配准、平差、融合、拼接和裁切等;进行外业观测成果资料的整理、概算,以及工程地形图数据的编辑处理等；进行航空遥感数据预处理或冲印处理;整理整编水工建筑物巡视检查及观测成果；进行降水量、蒸发量、水位、流量和含沙量等项目资料整编；检查获取影像、数据的数量和质量;检验测量成果资料,提供测量数据和测量图件；进行大地测量成果检查与验收；检验地理信息数据库准确性、精确性、完整性和逻辑性；检查数字档案的真实性、完整性、可用性和安全性汇总、整理统计数据；评估、分析统计数据和资料;统计、核算区域及企事业单位等碳排放数据
数据安全存储	研究信息系统加密与解密、认证、存取管理、机密信息管理、防火墙、安全协议、安全技术;分析信息系统安全性需求,制订信息系统安全规划;设计、开发、评估信息系统安全解决方案;指导或实施信息安全方案;制订信息安全政策、策略,实施等级保护、网络隔离、安全监控;制订安全危害预防策略,发现并解决信息系统中的泄密、病毒、攻击、信息篡改等安全问题;评估信息系统的安全性和安全等级；监控、管理和保障大数据安全；开发虚拟化、云平台、云资源管理和分发等云计算技术,以及大规模数据管理、分布式数据存储等相关技术;制订、实施与完善数据库的备份还原、复制、镜像等容灾方案;进行交换格式数据与所需的物理存储格式数据的转换；进行地理信息数据(库)的整理、存储、备份、维护管理和数据安全保密;进行数字档案的迁移、备份;
第三阶段：数据加工	
生产数据产品或提供数据服务	生产数字地面模型(DTM)、数字高程模型(DEM)、数字正射影像(DOM)等数字影像产品；应用地理信息系统(GIS)软件或工具,设计并组织实施地理信息数据库空间分析、数据建模；运用气象观测、预报预测信息和其他信息,研究、设计、加工制作气象服务产品；根据导航定位产品设计架构,集成、编绘、制作导航地理信息产品,提供位置监控、灾害预警、应急救援等导航与位置服务;定制开发地理信息系统(GIS),提供图表或数据服务；进行数据和信息处理,提供数据咨询服务；分析系统数据来源、数据应用需求;设计数据资源整合解决方案；提供大数据的技术咨询和技术服务；
具体业务中的数据分析应用	进行数据分析、数据挖掘、数据展现、决策支持；分析、运用财务、业务等信息,服务单位规划、决策、控制、评价等工作；分析统计信息,估计死亡率、意外事故发生率、疾病发生率、残疾发生率和退休率;监控、分析、管理人工智能产品应用数据；分析不同介质和智能终端的电子数据；运用网络信息技术和相关工具,对媒介和受众进行数据化分析,指导媒体运营和信息传播的匹配性与精准性；分析服务器、数据库及公有云的电子数据;分析物联网、工程控制系统的电子数据;采集相关数据,根据实时数据分析、监控情况,精准调整媒体分发的渠道、策略和动作；运用气象资料、技术和方法,进行天气分析,制作短期、中期和长期气象要素及灾害性天气预报；进行地形地貌、地表覆被、地理单元等地理国情要素动态、多时相的信息挖掘分析
基于数据的技术研发	研发、应用天气预报和数值预报等新技术、新方法；研究、开发、应用人工智能指令、算法及技术；研发、应用、优化语言识别、语义识别、图像识别、生物特征识别等人工智能技术；基于标注数据，分析提炼专业领域特征,训练和评测人工智能产品相关算法、功能和性能；研发气候监测预测、数值模拟与预估等新技术、新方法。

第四，针对行业中的职业构成，刘涛雄等（2023）采用了全国经济普查中行业中类就业人数统计资料，并将数据相关职业匹配至行业中类。相比起行业门类，部分行业中类与职业细类的相关性更强，例如“中药材的种植业”与“中药材种植员”，这种方法能够一定程度体现行业门类下不同职业的构成比例。然而，该方法的也存在着一些缺陷。一方面，行业是对企业等主体参与的经济活动进行分类，

而职业则是对劳动者从事的工作内容进行分类，两者从概念上而言并无绝对的对应关系。另一方面，该方法仅匹配了数据相关职业，当数据相关职业和数据无关职业在某些行业同时存在，这种处理方式就会导致数据资本形成过程中的劳动力成本被高估。为了改进这一问题，本研究参考了中国家庭追踪调查（Chinese Family Panel Studies, CFPS）、中国居民收入分配项目（Chinese Household Income Project Survey, CHIP）、全国人口普查等多来源的调查统计资料，对行业中的职业构成提供了较为准确的估算。

下面表 2 展示了 2003-2020 年基于上述方法得到的中国数据资本存量估算结果。表 3 展示了 2003-2020 年基于该方法计算的我国各省份数据资本存量。图 1 对比了刘涛雄等（2023）和本研究（即 V1 与 V2）中估算的中国数据资本存量及其增速。可以看到，两项研究中我国数据资本随时间积累和增长的趋势基本保持一致，但 V1 估算的数据资本规模要明显高于 V2，接近 V2 的三倍水平。这一方面是因为 V2 在估算各职业的数据相关工作时长占比时，采用了基于任务的目标职业挑选策略，能够一定程度上避免非数据相关因素的干扰导致各数据相关工作时长占比被高估[1]；另一方面，如前所述 V1 将职业层面的估算结果和行业劳动工资统计资料匹配时仅匹配了数据相关职业，从而使得数据资本形成过程中的劳动力成本被高估，而 V2 利用调查研究对行业中职业构成比例进行估算能够较好地解决行业与职业的匹配问题。因此，V2 的估算结果相对更为合理。

当然，在分析数据资本对经济增长的贡献时，关键是需要对数据资本存量的增长趋势进行准确衡量，数据资本存量绝对规模的大小并不会对结果造成太大的影响。具体而言，如果对数据资本存量规模高估的比例是相对固定的，那么在取对数值后，高估的部分将被回归的常数项所吸收，不会影响到数据资本产出弹性系数的估计。我们采用 V2 的估算结果分析数据资本对中国经济增长的贡献，也能够得到与《数据资本估算及对中国经济增长的贡献——基于数据价值链的视角》相似的结论。

[1] V1 的估算过程选取了 20 个职业作为全时从事数据相关工作，将其工作内容作为目标文本来估算各职业的数据相关工作时长。但是在这些目标职业的工作内容描述中，可能包含了少部分与数据工作无关的工作任务。例如，地理信息采集员的工作内容包括了“维护保养仪器、设备、工具”，该项任务也出现在了很多其他职业的工作内容中。因此，在计算文本相似度时，各职业和目标职业的语义相似度包含了在这些数据无关任务上的相似，这使得对各职业数据相关工作时长的估计值偏高，最终也会导致数据资本存量估算结果存在高估。在 V2 中，我们采用基于任务的目标职业挑选策略，尽可能剔除了与数据工作无关的任务描述，避免此类任务对文本相似度结果构成干扰。

表 2　2003—2020 年中国数据资本存量估算结果（2003 年不变价）

年份	数据资本存量（亿元）
2003	4900.87
2004	5556.58
2005	6347.99
2006	7297.78
2007	8421.72
2008	9499.27
2009	10607.45
2010	11617.25
2011	13394.09
2012	16076.80
2013	19921.83
2014	23999.44
2015	28212.01
2016	32357.85
2017	36569.40
2018	40994.69
2019	46430.80
2020	52419.36

图 1　V1 与 V2 估算结果的对比

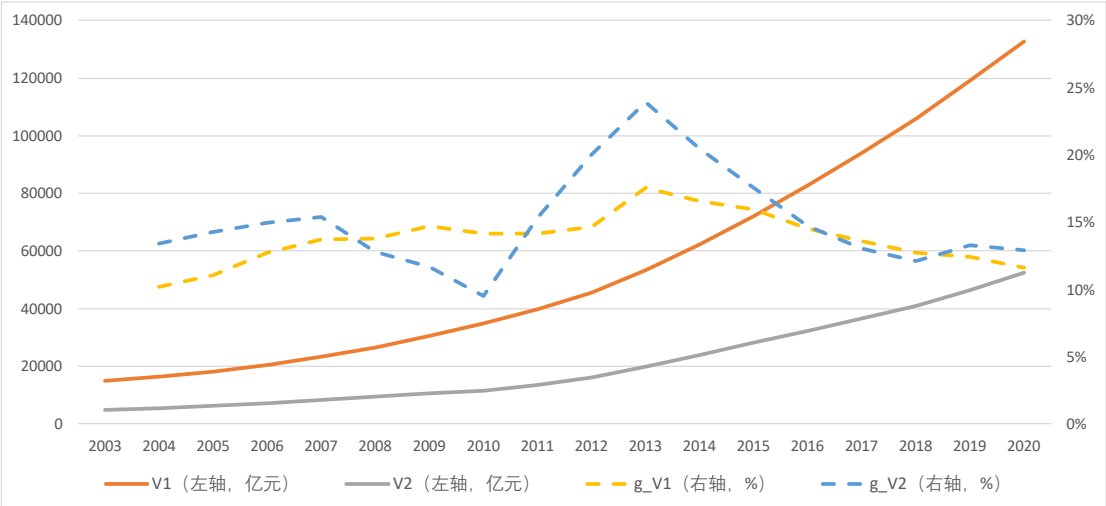

表 3　2003—2020 年中国各省份数据资本估算结果（2003 年不变价）

	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
上海	221.4	255.5	296.1	347.8	419.2	498.6	584.5	667.9	838.9	1098.5	1504.0	1952.6	2392.4	2803.9	3229.2	3695.0	4321.8	4959.8
云南	117.3	128.2	139.8	155.7	174.2	191.4	209.0	222.3	249.2	293.2	348.8	401.9	461.2	526.3	597.8	668.7	735.4	802.2
内蒙古	98.2	110.5	126.6	146.2	169.1	192.3	218.0	243.7	282.4	332.5	399.3	463.8	524.4	580.3	631.0	686.8	767.2	850.6
北京	351.5	413.1	485.3	571.7	668.1	816.8	1000.8	1218.2	1523.4	1937.4	2398.5	2915.1	3463.0	4014.5	4588.1	5202.8	5912.1	6589.6
吉林	116.4	127.1	138.3	151.5	166.7	179.5	191.5	201.0	217.6	242.5	281.5	323.1	370.0	418.0	466.9	517.0	574.5	623.2
四川	204.0	230.1	261.1	295.9	338.2	375.3	413.9	444.3	503.1	594.0	759.2	922.7	1098.6	1264.1	1424.7	1585.8	1763.4	2006.5
天津	98.5	114.1	130.4	151.2	178.2	203.7	229.1	253.2	299.9	374.5	463.0	555.1	665.9	779.6	882.7	969.9	1083.8	1199.6
宁夏	30.0	32.6	36.2	40.5	45.3	48.5	51.6	53.5	56.8	63.9	72.4	82.1	92.7	102.9	111.7	119.7	130.3	145.0
安徽	122.7	139.6	160.7	186.7	219.0	246.2	269.4	286.1	322.0	380.6	459.7	541.8	627.7	709.9	793.9	905.1	1027.6	1166.9
山东	312.0	350.3	405.0	465.4	536.8	589.1	635.2	669.7	749.7	877.8	1077.0	1282.3	1481.0	1682.6	1893.2	2097.2	2331.3	2633.2
山西	129.6	144.7	164.0	187.4	213.7	233.2	251.7	261.0	281.6	321.5	377.2	436.4	499.1	558.6	604.5	652.0	722.2	797.6
广东	449.6	526.1	615.9	718.7	835.1	942.4	1049.7	1154.0	1342.1	1625.0	2122.0	2632.8	3154.4	3691.5	4265.7	4977.3	5905.9	6976.5
广西	112.5	125.2	140.8	158.4	179.1	197.7	215.3	228.9	250.4	285.3	335.5	386.8	445.8	500.6	556.2	612.6	682.7	766.6
新疆	130.8	142.8	154.3	167.2	184.0	200.0	218.5	232.7	254.3	286.2	327.4	372.9	430.7	489.0	541.9	586.7	645.3	721.8
江苏	286.7	329.9	379.4	444.0	517.2	582.0	640.9	689.7	782.9	922.9	1291.6	1667.2	2024.9	2354.2	2688.4	3043.4	3439.6	3870.8
江西	95.3	106.6	120.5	135.7	153.0	167.1	182.4	194.5	216.5	256.7	318.0	382.0	452.3	521.6	594.0	662.7	753.4	848.9
河北	191.7	212.5	236.4	263.5	294.3	323.1	349.6	367.9	404.4	470.8	558.3	656.7	763.4	867.9	942.7	1034.9	1152.8	1287.9
河南	246.4	272.6	304.3	348.1	403.1	448.9	489.8	516.2	574.1	665.1	788.3	924.0	1066.1	1224.5	1391.1	1521.7	1692.0	1868.1
浙江	255.7	303.1	369.0	448.5	540.3	632.6	725.7	809.1	956.5	1188.1	1433.5	1694.2	1946.6	2205.7	2497.5	2780.0	3137.8	3585.1
海南	27.0	31.2	36.0	40.7	46.7	52.1	58.1	65.0	73.2	83.2	98.2	114.1	132.6	150.2	170.5	192.6	215.7	248.2
湖北	173.0	193.9	219.9	250.9	280.7	304.4	330.9	355.0	400.7	465.5	557.1	660.8	772.5	887.5	999.4	1106.9	1240.9	1380.5
湖南	173.1	191.1	212.7	238.2	271.6	301.3	329.7	352.4	394.6	462.3	541.7	625.4	706.6	790.0	876.4	969.3	1095.7	1231.6
甘肃	80.2	88.2	98.1	109.6	123.0	135.1	146.9	154.1	164.1	186.1	222.2	262.1	310.3	359.1	409.1	454.3	508.9	583.1
福建	147.3	169.3	196.7	232.7	273.9	312.8	345.3	373.7	434.4	531.8	640.6	752.5	873.7	991.8	1109.3	1243.6	1381.1	1523.3
西藏	16.3	18.3	20.2	22.2	26.3	29.3	31.8	34.1	37.5	43.5	52.1	59.9	74.9	86.1	96.9	107.5	126.7	141.1
贵州	71.8	80.8	92.1	106.8	124.7	138.2	150.2	159.0	179.0	211.5	252.7	297.3	340.6	384.7	429.4	471.2	525.6	588.8
辽宁	219.1	247.3	281.4	319.4	361.9	404.5	448.8	495.6	569.5	668.8	809.5	948.5	1079.0	1178.5	1267.0	1356.3	1478.4	1612.8
重庆	78.3	90.5	106.0	125.1	149.2	171.2	193.2	214.1	257.8	324.6	408.7	501.2	595.2	684.4	770.6	861.2	962.5	1085.9
陕西	124.4	139.1	154.4	170.9	194.2	215.3	238.9	259.2	292.1	339.7	411.5	494.5	584.7	678.6	769.7	852.3	952.1	1061.5
青海	23.7	26.3	29.3	33.0	37.1	40.9	45.7	49.6	57.0	67.5	77.6	88.8	99.7	110.1	121.8	133.5	150.1	169.6
黑龙江	196.6	215.9	237.2	264.2	297.7	325.8	361.3	391.7	428.1	475.9	534.7	600.7	681.9	761.1	848.2	926.8	1014.0	1093.2