

参赛队号：（由大赛组委会办公室填写）

2021 年（第七届）全国大学生统计建模大赛

参赛学校： 浙江工商大学

数据新动能驱动中国经济增长的统计研究——
论文题目： 一基于“数字产业化”和“产业数字化”的
经济贡献测度

参赛队员： 刘嘉昕 孙艺恺 陶佳莉

指导老师： 李凯 朱贺

数据新动能驱动中国经济增长的统计研究

— 基于“数字产业化”和“产业数字化”的经济贡献测度

摘要：伴随着数据要素化进程的不断加深，对于数据如何作用于经济发展，数据与其他要素结合产生的动能应该如何测度的研究愈发重要。本文将数据新动能分解为“数字产业化”与“产业数字化”两个角度来对其进行统计测度，以更好地去理解数据是如何赋能与其他要素，助力我国数字经济高质量发展。

本文首先进行了文本分析，利用爬虫从知网、百度资讯上抓取相关文献，经过预处理、清洗、分词后，从词云图获取到文献聚焦的热点。在对分词后的建立“文档-词项”矩阵与 TI-IDF，并根据 TI-IDF 的结果建立 lda 主题模型，最终得到“数据要素”“技术创新”“宏观政策”“转型升级”“交易模型”“数据监管”六大主题，为后续问题的探讨明确了方向。

之后我们选择依据经济原理，通过综合评价来构建与“数字产业化”与“产业数字化”相对应的指标体系并使用柯布-道格拉斯生产函数，对“数字产业化”进行投入产出分析，通过 bootstrap 方法拟合方程计算得出各种生产要素在数据产业中的贡献率，并计算出相对应的“产业数字化”生产函数，通过数学变换分析得出了相应的数据要素对于劳动力，资本，科技的作用，来使得我们以更好地理解、分析“数字产业化”和“产业数字化”共同作用下的经济合力——数据新动能。

关键词：产业数字化 数字产业化 数据赋能 文本分析 生产函数

目录

一、引言.....	1
(一) 研究背景及意义.....	1
1、研究背景.....	1
2、研究意义.....	2
(二) 研究问题.....	2
(三) 研究内容.....	3
(四) 研究思路及方法.....	3
1、研究思路.....	3
2、方法介绍.....	4
二、数据新动能驱动经济增长的机理探索.....	7
(一) 基于文本挖掘方法的数据新动能解析.....	7
(二) 数据新动能测算及相关概念界定.....	9
(三) 数据新动能、“数字产业化”和“产业数字化”的作用分析.....	10
1、数据新动能作用机制分析.....	10
2、“数字产业化”作用机制分析.....	10
3、“产业数字化”作用机制分析.....	11
4、数据新动能与相关概念之间的作用机理.....	13
三、数据要素赋能中国经济增长的模型构建.....	13
(一) “数字产业化”指标体系构建.....	13
(二) “产业数字化”指标体系构建.....	14

(三) 基于“数字产业化”和“产业数字化”的生产模型构建.....	15
四、 数据要素赋能中国经济增长的测度分析.....	17
(一) 数据及变量说明.....	17
(二) 基于“数字产业化”新动能的经济贡献测度.....	18
(三) “数字产业化”新动能的灰色预测.....	20
(四) 基于“产业数字化”新动能的经济贡献测度.....	21
(五) “产业数字化”新动能的灰色预测.....	23
五、研究结论与建议.....	24
(一) 研究结论.....	24
1、“数字产业化”研究结论分析.....	24
2、“产业数字化”研究结论分析.....	25
(二) 数据驱动中国经济高质量发展的建议.....	26

表目录

表 1 LDA 主题模型结果.....	8
表 2 “数字产业化”指标设计表.....	14
表 3 “产业数字化”指标构建.....	15
表 4 变量说明.....	17
表 5 三级指标.....	18
表 6 “数字产业化”生产模型拟合结果.....	19
表 7 “数字产业化”生产模型统计检验量.....	19
表 8 “数字产业化”生产模型参数.....	19
表 9 2003 年-2019 年数据产品服务价值指标平均相对残差与平均级比偏差表.....	20
表 10 2020 年-2023 年数据产品服务价值预测值.....	20
表 11 “产业数字化”生产模型所用指标.....	21
表 12 “产业数字化”生产模型拟合结果.....	21
表 13 “产业数字化”生产模型统计检验量.....	22
表 14 “产业数字化”生产模型参数.....	22
表 15 2003 年-2019 年总产出指标平均相对残差与平均级比偏差表.....	23
表 16 2020 年-2023 年总产出预测值.....	23
表 17 各要素贡献率.....	24

图目录

图 1 技术路线图	4
图 2 Bootstrap 方法步骤图	6
图 3 词云图	7
图 4 数据新动能要素分解	10
图 5 数字产业化价值链图	11
图 6 产业数字化示意图	12
图 7 2003 年-2019 年数据产品服务价值指标相对残差与级比偏差图	20
图 8 2003 年-2019 年总产出指标相对残差与级比偏差图	23

一、引言

（一）研究背景及意义

1、研究背景

当下社会已进入数字经济时代，由于其提升全要素生产率的作用，数据已逐步成为较为关键的生产要素，新兴的数据处理与管理方式“共同推动数据要素向各行业全面渗透”。但是我国的数据要素化市场仍处于起步阶段，仍然具有很大的不完善性。

在这样的形势下，2020年4月，我国正式发布《构建更加完善的关于要素市场化配置》的文件，将数据作为一种新型生产要素，首次与传统要素并列。文件着重强调，“要推进政府数据开放共享，提升社会数据资源价值，统筹建立城市大数据平台”，充分发挥数据在政务管理方面的作用；“积极开展数据资产管理在各垂直行业中的应用研究”，扎实推进大数据技术与实体经济的深度融合；“要加快培育发展数据要素市场，创新交易机制”，不断优化营商环境，加强“数据资源整合和安全保护”，突破“数据自由流动障碍的瓶颈”。

在此基础上，文件继续指出应着力突破“数据确权、自由流动、隐私安全等方面的瓶颈制约”，完善相关配套措施，制定“数据确权定价和流动交易的标准、规范和共识”，推动“多方安全计算、可信执行环境等数据隐私保护新技术新标准”的应用，不断培育和发展数据要素市场，加快我国数据资产化的进程，积极构建和完善“数据治理监管体系”，使数据要素能够充分参与市场配置，推动我国经济高水平高质量发展。

数据要素市场化配置不仅在当下意义重大，在可预见的未来也会持续发挥重大作用。在此过程中需要着力完成有效配置，确保数据要素资源长期发挥价值，

拉动我国数字经济发展。在这样的背景下，对于数据这种新生产要素中所蕴含的新动能进行标准化，规范化，准确化，精确化的统计测度变得愈发重要。

2、研究意义

数据要素中所蕴含的信息，具有增加决策正确率，减少未知风险等一系列战略性作用。而数据的价值在需求方表现为数据要素中的这种信息熵对于使用者的价值；在供给方表现为挖掘，储存，交换这些数据所花费的成本。在经济生活中具体可以表现为数字经济中所提供数据分析，数据挖掘服务的价值。但当下我们对于数据要素的经济影响还缺乏完整的认知，经济学界还是统计学界至今仍未就如何测算数据要素的经济价值达成共识。而数据如何赋能于其他生产要素，促进经济发展经济进而产生的这样一种数据新动能更是研究中的重点。

对于使用数据服务及数据产品的企业而言，如果缺乏对数据赋能产生的价值的认识与较为合理的测算，就可能会限制其数字化转型的进程，影响企业的相关要素投入与生产决策，进而影响数据资源配置的效率；对于提供数据服务与数字产品的企业而言，缺乏合理的数据动能测度有可能会造成错误的价值估计进而影响其服务定价，影响数据服务类企业稳定发展与行业市场的规范化；对于我国政府而言，缺乏合理的新动能测度，关于数据要素赋能的统计活动将无法准确反映我国实际经济运行情况，会影响到我国经济 and 数字竞争力的相关政策的制定，不利于充分发挥我国经济发展中数据新动能的潜力。

(二)研究问题

本次建模所要研究的主要问题，就是探索测度数据赋能经济而产生新动能的方式，希望能够寻找到部分具有较强代表性，较高准确性的测度指标，找到一种合适的方式，来对数据要素这一新兴生产要素在对经济赋能过程中所产生的价值进行测度，助力数字产品与服务市场规范化，激发数据产业新动能，更好的发挥

数据驱动新消费的作用，推动我国数字经济新发展。

(三)研究内容

本文接下来将进行数据新动能驱动经济增长的机理探索，基于文本挖掘的方法来对数据新动能进行解析，并对数据新动能及相关概念界定，分析数据新动能、“数字产业化”和“产业数字化”的关系与相互作用。之后我们将进行数据要素赋能中国经济增长的统计模型构建，并将其分为“数字产业化”评价指标体系和“产业数字化”评价指标体系进行设计。

基于这两个评价指标体系，我们将构建“数字产业化”和“产业数字化”的生产模型进行分析，并利用这一模型来进行数据要素赋能中国经济增长的测度分析，最后我们将根据我们研究的结果得出相关结论，并提出相关建议，助力数字化要素发展，让数据新动能使我国经济发展迸发出新的活力。

(四)研究思路及方法

1、研究思路

为了更好的了解和界定数据新动能，本文进行了文本分析，利用 Python 爬虫从知网、百度资讯上抓取相关文献，经过预处理后，共计获得文章 622 篇。在对文本进行清洗、分词后，按照词频制作词云图，从词云图获取到文献聚焦的热点。在对分词后的建立“文档-词项”矩阵与 TF-IDF，并根据 TF-IDF 的结果建立 lda 主题模型，以便从中获取主题，为后续问题的研究提供思路。

通过文本分析后我们发现数据新动能由于各方面存在的很大的宽泛性和不确定性，直接测度没有标准，测度难度较大，所以我们选择使用综合评价体系来进行指标设计，从侧面进行测度。通过数据新动能作用方式将其划分为“数字产业化”“产业数字化”两部分分别进行测度，并将其拟合成生产函数的形式，利

用数学变换来进行数据赋能的测度。

在拟合生产函数的过程中，由于数据量过少，部分指标统计近些年来才开始统计，我们使用了插值法来进行空值补全。为了解决模型拟合中由于数据量过少，大部分变量不显著的问题，我们通过 R 语言使用了对样本及其分布要求较低的 bootstrap 方法来近似扩大样本的过程，进行 1500 次重抽样来进行方程参数的估计。最终拟合出较为准确的回归方程，并进行下一步的灰色预测来对指标体系中的指标进行进一步预测。

最后，我们根据上述结果进行分析并提出了相关的建议，希望能助力于我国数字经济的发展，使我国经济迸发出更多的数据新动能。

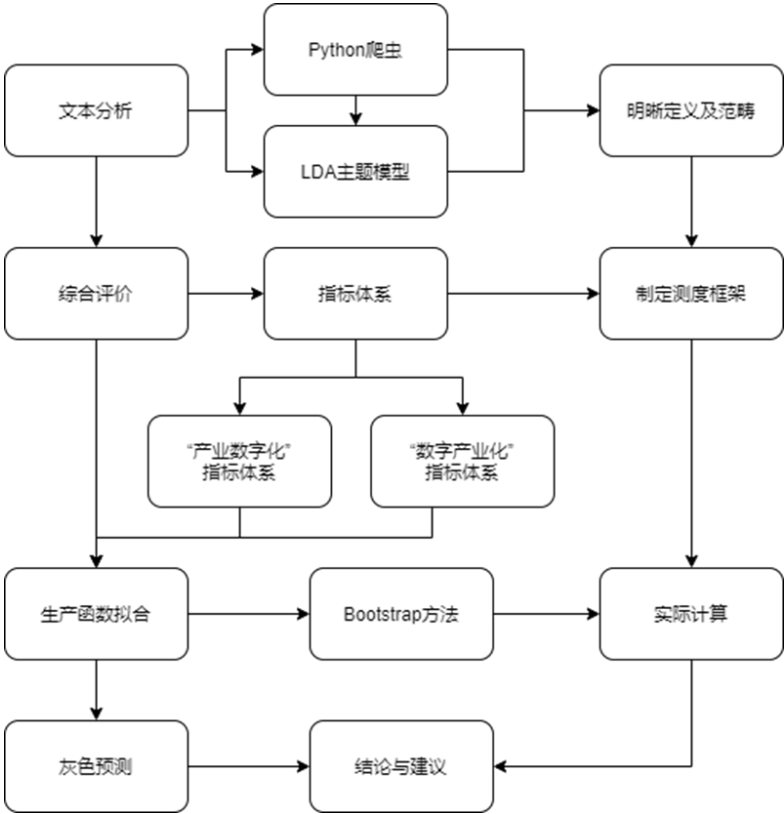


图 1 技术路线图

2、方法介绍

①LDA 主题模型

词袋模型

词袋模型是在 LDA 中采用的模型，在分析一篇文章时，我们不考虑词汇之间的顺序，而只是考虑该词汇是否在文章中出现。即，在词袋模型中“我喜欢你”和“你喜欢我”等价。

共轭先验分布

在贝叶斯概率理论中，假若先验概率 $P(\theta)$ 和后验概率 $P(\theta|x)$ 具有同样的分布律，那么，共轭分布的概念为先验分布和后验分布，并且，似然函数的共轭先验分布被称为先验分布。

$$P(\theta|x) = \frac{P(\theta, x)}{P(x)} \quad (1)$$

二项分布的共轭先验分布经过推导可知是 Beta 分布，同理，多项式分布的共轭分布经过推导可得知为狄利克雷 (Dirichlet) 分布。

Dirichlet 分布

Dirichlet 的概率密度函数为：

$$f(x_1, x_2, \dots, x_k; \alpha_1, \alpha_2, \dots, \alpha_k) = \frac{1}{B(\alpha)} \prod_{i=1}^k x_i^{\alpha_i-1} \quad (2)$$

其中，

$$B(\alpha) = \frac{\prod_{i=1}^k \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^k \alpha_i)}, \sum_{i=1}^k x_i = 1 \quad (3)$$

Gibbs Sampling 采样

将语料库定义为 \bar{z} ，语料库中的第 i 个词我们记为 z_i ， $i = (m, n)$ 代表一个二维下标， m 代表第 m 篇文档， n 代表该文档中的第 n 个词，用 $\neg i$ 表示去除下标为 i 的词。根据 Gibbs Sampling 算法，我们需要求任一个坐标轴 i 对应的条件分布 $p(z_i = k | \bar{z}_{\neg i}, \vec{\omega})$ 。假设已经观测到的词 $\omega_i = t$ ，则由贝叶斯法则，易得：

$$p(z_i = k | \bar{z}_{\neg i}, \vec{\omega}) \propto p(z_i = k, \omega_i = t | \bar{z}_{\neg i}, \vec{\omega}_{\neg i}) \quad (4)$$

借助于 Dirichlet 参数估计的公式最终得到的 LDA 模型的 Gibbs Sampling 公式为：

$$p(z_i = k | \vec{z}_{-i}, \vec{\omega}) \propto \frac{m_{m,-i}^{(k)} + \alpha_k}{\sum_{k=1}^K (n_{m,-i}^{(k)} + \alpha_k)} \cdot \frac{n_{k,-i}^{(t)} + \beta_t}{\sum_{t=1}^V (n_{k,-i}^{(t)} + \beta_t)} \quad (5)$$

LDA 模型生成

- 1、将语料库中的每个词 ω 随机的与某个主题 z 对应；
- 2、根据 Gibbs Sampling 公式对每个词 ω 采样，计算得到该词的主题，将计算得出的主题在语料库替换在语料库中的原主题；
- 3、重复复步骤 2，直到 Gibbs Sampling 公式收敛；
- 4、LDA 模型就是统计语料库后得到“主题-词”共现频率矩阵；

②Bootstrap 方法

Bootstrap 的基本思想为：因为观测样本包含了潜在样本的全部的信息，那么不妨把这个样本当做总体。那么相关的统计工作（估计或者检验）的统计量的分布可以从这部分样本中利用 Monte Carlo 模拟（蒙特卡罗模拟）得到。Bootstrap 方法可简单地概括为：样本从总体中抽取所得，从总体样本的中抽样获得的仍然是总体的样本，从样本中再抽样相当于间接的扩大样本。

其基本步骤为：

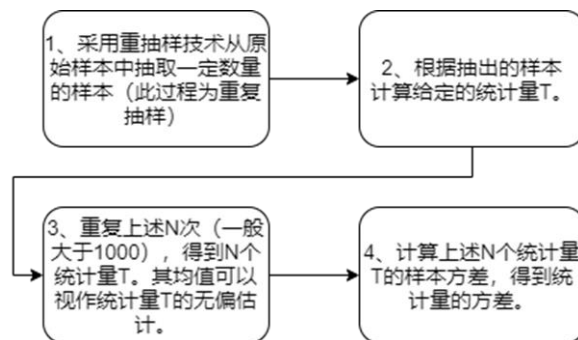


图 2 Bootstrap 方法步骤图

Bootstrap 对于分布理论要求极低，但同时也可以进行假设检验，获得置信区间。当数据分布未知，存在严重异常点，或者样本量过小的问题时 bootstrap

方法可有效解决上述问题。对于回归分析而言，bootstrap 对回归的正态性假设做了极大放松，便利了回归推断也更具说服力。

二、数据新动能驱动经济增长的机理探索

(一) 基于文本挖掘方法的数据新动能解析

首先为了进一步了解数据新动能，我们进行了文本分析。通过在知网，百度资讯，政府相关的规划报告等查找现有的以“数字经济”，“数据新动能”，“数据要素”，“数据资产”等关键词为主题的文献资料，我们使用 Python 获取这些文献资料中的高频词并制作出相对应的高频词云图。



图 3 词云图

以“数据新动能”“数字经济”“数据要素”等词语为关键词，在百度资讯、知网等平台搜索相关文章，并利用爬虫技术，爬取资讯、论文，在经过文本的整理与筛选后，共获得 622 篇文章。在对文本进行文本清洗后，利用 python 中的 jieba 对文本进行分词，并根据《四川大学机器智能实验室停用词库》《哈工大停用词表》等停用词表去除无用的词语，并根据词频绘制词云图。

在词云图中，词频通过词的大小来表示，词频越高，词越大。根据词云图可知，“增长”“要素”“发展”“权属”等为文本出现频率较高的词，这些词一定程度上反应了数据作为一种生产要素，在促进经济发展具有一定作用，同时当数据被认定为生产要素后，作为一种能够促进经济发展的重要动力，数据的权属问题也受到重视了。从上图中我们可以看出，“数据”、“经济”、“要素”、“增长”这四个关键词格外醒目。我们可以明显感受到数据要素化正在逐渐成为大趋势，数据已然成为创造经济增长的重要要素之一。

根据分词后的文本建立“文档-词项”矩阵，即 DTM 矩阵，DTM 矩阵记录着每个文档中每个词汇出现的次数。同时建立 TF-IDF，TF-IDF 用于评估一个词对于整个文档的重要性程度，是一种提取关键词的统计分析方法。该方法的原理为：假若一个词在文章中出现的次数越多，该词就越具有重要性，假若该词在文档中出现的次数越多，该词就越不具有重要性。该方法可有效避免常用词的影响，提高关键词与文章的相关性。

同时根据 TF-IDF 的结果进行 LDA 主题分析。LDA 主题模型可以给出文档中的文章所包含的主题。LDA 主题模型所得结果如下：

表 1 LDA 主题模型结果

主题	
数据要素	数字、资产、企业、发展、要素
技术创新	数字经济、数字化、创新、智能、科技、互联网、技术
宏观政策	中国、政府、国家、社会、构建、推动
转型升级	生产、产品、资源、转型、体系
交易模式	平台、业务、交易、存储
数据监管	数据管理、监管、隐私、保护、治理

根据 LDA 主题模型的结果，可将其分为六个板块：“数据要素”、“技术创新”、“宏观政策”、“转型升级”、“交易模型”、“数据监管”。“数据要素”板块反映了数据已经成为被广泛认可且逐渐受到重视的生产要素，数据在推动经济的过程中有着举足轻重的作用；“技术创新”说明了数据能够作为一种生产要素在经济活

动中发挥效用，是依托于技术的进步以及互联网、数字化的普及。数字化、互联网的普及产生了海量的数据，而技术的进步以及相关理论的出现才使得数据的收集、分析并创造价值成为一种现实；“宏观政策”板块说明了政府重视数据作为动能对经济的促进作用，并且在构建平台，整合资源，出台相关政策，期望能够更好的提升数据对于其他生产要素的放大作用，推动经济发展；“转型升级”说明了企业期望构建体系，让数据更好的为生产活动服务，助力企业转型升级；“交易模式”说明了当今市场出现了一些数据交易的平台，但数据交易的规范性，数据定价的合理性，仍是需要进行进一步讨论与规范；“数据监管”说明了数据在作为动能促进经济发展的过程中，存在着隐私泄露的风险，如何在最大限度运用数据的同时降低类似问题的发生也是亟待解决的问题，并且该问题的解决必然离不开政府的监管与治理。

（二）数据新动能测算及相关概念界定

数据新动能是指，在数字经济发展的过程中，数字产业化与产业数字化重塑生产力，所迸发出的带动产业转型、经济发展、社会进步的巨大动能。数字产业化代表了推进软件定义、数据驱动的新型数字产业体系加速形成的新一代信息技术的发展方向和最新成果。而产业数字化推动新一代信息技术与实体经济广泛深度结合，加速新型智能化生产方式的到来。数字经济在数字产业化与产业数字化的框架下，不断提高经济社会的数字化、网络化、智能化水平，加速重构经济发展与治理模式的新型经济形态。

数字化的知识和信息的生产要素称为数据要素，数据要素贯穿于数字经济发展的全部流程，是数字经济全新的、关键的生产要素。数据要素并不直接等同于数据新动能，单一的数据要素无法发挥价值，但数据要素与其他生产要素不断组

合迭代，加速交叉融合，是数据新动能能够产生的必要条件。因而，数据要素是数据新动能关键组成部分，同时，数据新动能是对数据要素价值的进一步延展，使其能够作用于经济社会发展，重塑生产力结构。

（三）数据新动能、“数字产业化”和“产业数字化”的作用分析

1、数据新动能作用机制分析

数据新动能即数据要素激发经济增长新动能。数据新动能简单来说可以理解为数据作为一种新的要素通过各种方式所带来的经济高质量发展。数据新动能可以表现为数字经济的“四化”，分别为数字产业化，产业数字化，数字化治理和数据价值化。其中与经济生产息息相关的为数字产业化与产业数字化两方面。



图 4 数据新动能要素分解

2、“数字产业化”作用机制分析

数字产业化直接表现为信息通讯行业的发展。信息通讯行业是数字经济发展的先导产业，这一产业主要为经济发展提供服务，技术，电子信息产品等。数字产业化的主体具体包括电子信息制造业，电信业，软件和信息技术服务等。5G、集成电路、软件、人工智能、大数据、云计算、区块链等技术、产品及服务。

数字产业化作用经济发展的方式主要是通过信息通讯相关行业给其他行业提供数据产品和服务所带来的自身生产总值的增加对经济所产生的拉动。对于数据

对于其数据服务与产品生产者的经济赋能，我们选择依据最能体现数据赋能的数据生产价值链的框架来进行综合评价。联合国贸易和发展会议(UNCTAD)曾提出，数据价值链包括数据采集，数据整理，数据分析以及数据货币化四个阶段。通过进一步细化我们将其分为数据采集与存储，数据处理，数据分析，以及最后的数据使用与货币化四个部分。同时，数据服务生产活动的投入也对应着三个主要要素，分别是数据行业的从业者，数据储存以及相应的管理成本，以及这些互联网数据服务企业的平台建设维护的成本，产出方面即对应着数据产品及服务的价值量。

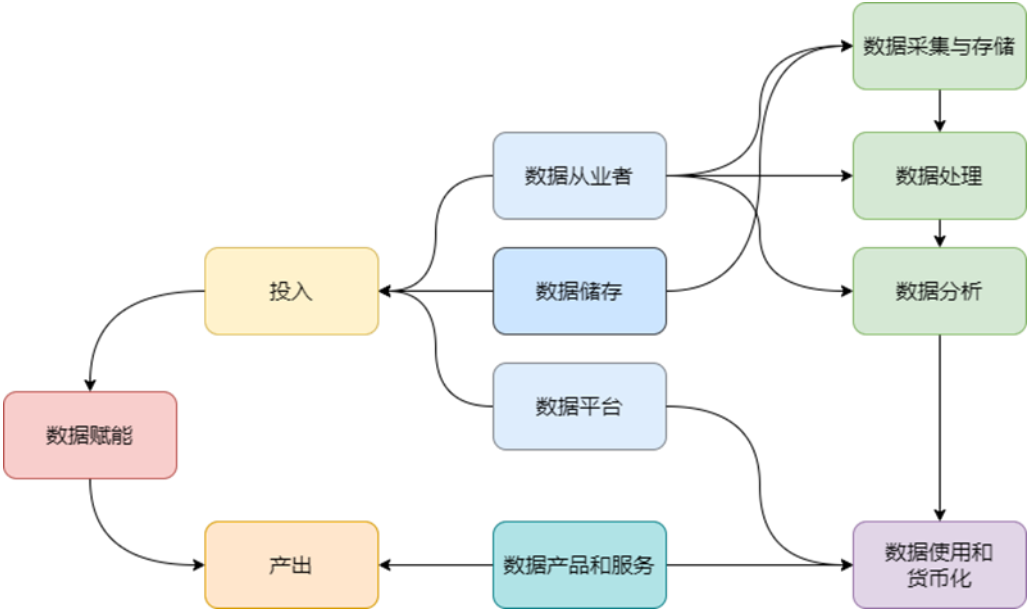


图 5 数字产业化价值链图

在这一价值链中，数据赋能的表现便是从投入到产出的全部增加值。因为这一整条价值链都是围绕着数据产生的，这一模式也被称为数据赋能型商业，数据是这一类商业模式运营的命脉，也是其核心创收活动的关键驱动因素。同时这也是数据新动能最显著的外在表现，即数字产业这一新兴产业对于经济增长的贡献。

3、“产业数字化”作用机制分析

除此之外，不止“数字产业化”应计入数据新动能的范畴，数据这种特殊的

生产要素与其他领域的生产活动结合后所释放出的巨大生产力，也就是“产业数字化”所产生的经济动能，同样也是数据新动能的重要组成部分。

产业数字化的含义是指数字技术与传统产业相结合所带来的生产力进步。其生产效率的提升是数据新动能的重要构成。实体经济是产业数字化的落脚点，数字经济是其发展的必要推动，高质量发展是产业数字化的必然要求。

其作用机制主要表现为：数据产业本身作为一种特殊的服务业，其生产出的数据服务作用于其他和数据结合的生产领域上，所产生的提升资源配置效率，提升决策准确度，预防未来可能出现的风险等等一系列具有广阔作用面的经济效益。这些效益的经济价值也应该计入产业数字化所创造的价值中。

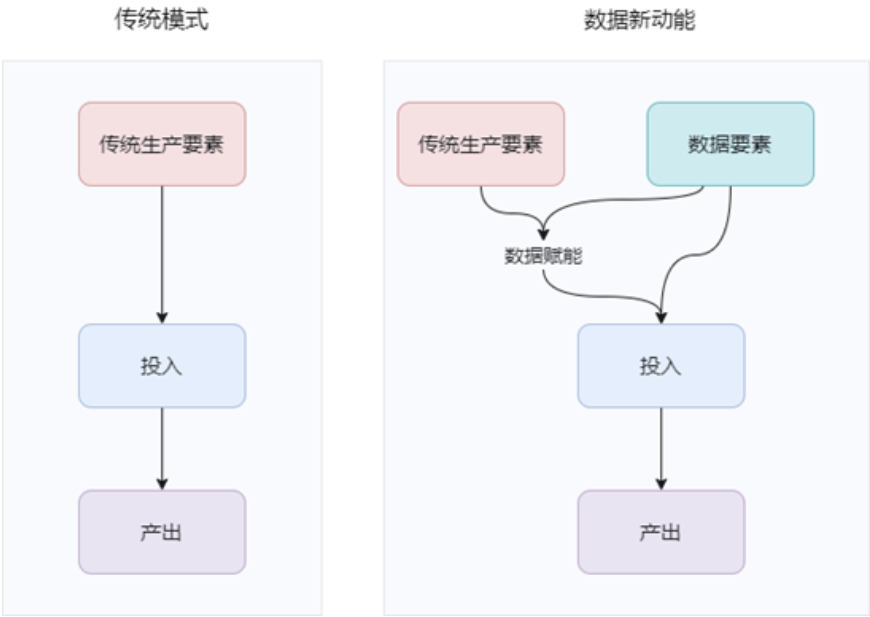


图 6 产业数字化示意图

但是产业数字化这部分的价值难以测算，我们很难去判断数据的要素作用到底对这些增加值起到了多大程度的影响。数据产生的新动能通常以与数据结合的对应产业的增加值的形式表现出来，不同的行业，不同的数据使用方式对于总增加值的赋能作用各不相同，数据赋能对于总体增加值的剥离系数具有很大的不确定性，如果采用同种剥离系数会产生较为严重的测量误差。数据要素可以作用的产业极多，范围广所以从宏观的角度去测定，可以更好的查看数据要素对于整个

宏观的经济生产的作用。

4、数据新动能与相关概念之间的作用机理

数据要素是推动数据新动能作用的基础，脱离了数据，数据新动能就无从谈起；而产业数字化与数字产业化是数据新动能的作用方式，数据新动能主要以这两种方式为媒介作用于经济。其中产业数字化是其作用于实体经济的最主要途径，也是其在促进经济高质量发展上最为重要和深刻的内容。只有数据要素不断壮大，产业数字化和数字产业化共同发展，三方面共同作用，才能促进数据新动能全面协调的经济高质量发展。

三、数据要素赋能中国经济增长的模型构建

（一）“数字产业化”指标体系构建

为了更好对“数字产业化”进行测度，选择更好的指标来进行投入产出函数的构建，我们构建指标体系来对“数字产业化”对我国经济赋能的作用进行衡量。通过对数据新动能作用原理的分析以及查阅各类参考文献，我们构建出了数据赋能的指标体系。

我们选择将一级指标设置为投入与产出两项，其中投入的二级指标设置为数据行业从业者投入，数据储存及处理的硬件投入，数据平台投入这三个方面。指标体系如下图所示：

表 2 “数字产业化”指标设计表

一级指标	二级指标	三级指标	含义
投入	数据行业劳动者投入	数据行业从业者工资	当期数据从业者的工资性收入
	数据储存及处理的硬件投入	硬件设备总价值	当期数据储存设备如硬盘等硬件设施的销售量
	数据平台投入	平台营业额	当期正常营业的数据平台的营业额之和
产出	数据产品和服务产出	提供数据产品和服务的公司数量	当期提供数据产品和服务的法人单位数量
		数据产品和服务的价值总量	当期提供数据产品和服务的公司总营业额

投入层面，行业从业者的投入我们分为从业者数量和从业者工资两部分；数据储存的固定资本投入我们使用数据储存设备的销售量来进行描述；数据平台的投入我们分为数据平台数量，与数据平台的营业额，一并作为数字产业的中间投入进行计算。这样的指标设计更加贴合劳动力，资本的要素特点，可以更好的体现出互联网规模效应的作用。

而要素投入层面之所以没有涉及科技要素的指标设计，是考虑到数据要素和科技要素在数据产业这一特殊行业内部的融合率较高，大部分数据企业的科研投入领域可以算入数据科学范畴内，因此数据产业的科技投入可以说与数据要素投入绝大部分都是重合的，所以将科技的贡献率统一化为数据要素贡献率来进行计算。

产出层面我们采用了数据产品和服务产出来进行衡量，并将其分为提供数据产品和服务的公司数量与数据产品服务价值量两部分。

（二）“产业数字化”指标体系构建

其中产出我们选择使用中国 2003-2019 年的 GDP 来进行衡量，资本使用同一时间区间内固定资产投资额进行衡量，科技要素使用同一时期的科学技术研发投入

入总额进行计算；劳动力要素使用同时期劳动工资总额进行计算。

而对于数据要素，我们选择通过上文“数字产业化”中的计算结果来得出数字产业化贡献值来进行衡量。其经济意义在于体现由数据产业为主体依托所带来的可以与其他产业结合的经济动能。虽然随着数字要素化发展，除数字产业以外的其他企业也越来越重视数据在自身领域中的应用，有很大一部分企业自身的管理部门或者决策部门都已经开始大量利用数据要素来进行管理决策，也是数据新动能的重要体现。但是这方面的标准难以确定，不同的行业差别较大，而且数据的结合作用方式也使得其难以与其他要素的作用剥离。出于诸多考虑，我们使用“数字产业化”中的数据产业产值来进行衡量。指标构建如下表所示：

表 3 “产业数字化”指标构建

要素	指标
劳动力要素	劳动者报酬
科技要素	科技研发投入总额
资本要素	固定资产投资额
数据要素	数据产业产值

（三）基于“数字产业化”和“产业数字化”的生产模型构建

通过查阅现有文献及研究，我们选择采用柯布一道格拉斯生产函数来对数据产业的投入产出进行分析。我们假定数据行业的生产函数满足柯布一道格拉斯形式

$$y = F(x_1, x_2, x_3, \dots, x_n) = x_1^{z_1} \cdot x_2^{z_2} \cdot x_n^{z_n} \quad (6)$$

为了将模型转化为线性形式，对方程两边同时取自然对数可得：

$$\ln y = z_1 \ln x_1 + z_2 \ln x_2 \dots + z_n \ln x_n \quad (7)$$

为了计算要素贡献率，我们对两端对时间进行求导，变换可得：

$$1 = z_1 \frac{\frac{x_1'}{x_1}}{\frac{y'}{y}} + z_2 \frac{\frac{x_2'}{x_2}}{\frac{y'}{y}} \dots + z_n \frac{\frac{x_n'}{x_n}}{\frac{y'}{y}} \quad (8)$$

该方程的右侧各项即代表着要素 $x_1, x_2, x_3, \dots, x_n$ 的要素贡献率，但是由于要素 $x_1, x_2, x_3, \dots, x_n$ 中的数据要素 x_i 价值未知，我们无法直接测算，所以我们做以下变换：

$$z_i \frac{\frac{x'_i}{x_i}}{\frac{y}{y}} = 1 - z_1 \frac{\frac{x'_1}{x_1}}{\frac{y}{y}} - z_2 \frac{\frac{x'_2}{x_2}}{\frac{y}{y}} \dots - z_{i-1} \frac{\frac{x'_{i-1}}{x_{i-1}}}{\frac{y}{y}} - z_{i+1} \frac{\frac{x'_{i+1}}{x_{i+1}}}{\frac{y}{y}} \dots - z_n \frac{\frac{x'_n}{x_n}}{\frac{y}{y}} \quad (9)$$

我们可用上式推导出数据要素 x_i 的贡献率。依据同样的原理，我们同样可以构建“数字产业化”与“产业数字化”赋能下的生产函数模型。我们以资本，劳动力，科技，数据这四种要素建立方程：

$$y = K^\alpha L^\beta S^\gamma D^\mu \quad (10)$$

K, L, S, D 分别代表资本，劳动力，科技，数据这四种要素，取对数可得：

$$\ln y = \alpha \ln K + \beta \ln L + \gamma \ln S + \mu \ln D \quad (11)$$

作上述同种变换可得：

$$\mu \ln D = \ln y - \alpha \ln K - \beta \ln L - \gamma \ln S \quad (12)$$

上式即为我们本次建模计算数据要素贡献率的生产函数模型。其中 $\alpha, \beta, \gamma, \mu$ 分别为资本，劳动力，科技，数据的贡献率。通过该模型，我们将进行“数字产业化”和“产业数字化”数据赋能贡献率的计算。

四、 数据要素赋能中国经济增长的测度分析

(一) 数据及变量说明

本文使用的所有数据均来源于国家统计局，国家数据，我国统计年鉴。

表 4 变量说明

变量	变量说明
x_i	第 <i>i</i> 个要素
z_i	要素系数
y	因变量：产出
α	资本要素贡献率
β	劳动力要素贡献率
γ	科技要素贡献率
μ	数据要素贡献率
x_1	数据行业从业者工资
x_2	数据行业硬件设备总价值
x_3	数据平台数量
q	提供数据和产品服务的公司数量
q_1	数据平台数量
y_d	数据产品服务价值总量

变量	变量说明	代理变量
k	资本要素	固定资产投资额
l	劳动力要素	劳动者报酬
s	科技要素	科技研发投入额
d	数据要素	数据产业产值

(二) 基于“数字产业化”新动能的经济贡献测度

通过查询现有数据,我们将三级指标所对应的计算数值查找整理,我们选择信息传输、计算机服务和软件业城镇单位就业人员工资总额作为数据行业从业者工资 x_1 的计算数值。信息传输、计算机服务和软件业新增固定资产作为数据储存及处理的硬件投入 x_2 ,虽然这两个数值在实际的经济生产中可能存在较大的差异,例如信息传输、计算机服务和软件业新增固定资产可能包括办公场所等其他固定资产,但是考虑到现有数据中只有这部分最符合该行业数据储存及处理的硬件投入,且这部分新增固定资产内有大部分为电脑,服务器,硬盘等数据处理及储存投入,故我们依旧选择新增固定资产对该指标进行近似的代替。

电子商务销售额主要是由这些企业的网站平台创造的,故平台营业额 x_3 我们近似的使用信息传输、软件和信息信息技术服务业电子商务销售额进行代替。

表 5 三级指标

三级指标	变量名	对应统计指标
数据行业从业者工资	x_1	信息传输、计算机服务和软件业城镇单位就业人员工资总额(亿元)
硬件设备总价值	x_2	信息传输、计算机服务和软件业新增固定资产(亿元)
平台营业额	x_3	信息传输、软件和信息信息技术服务业电子商务销售额(亿元)
提供数据产品和服务的公司数量	q	信息传输、计算机服务和软件业法人单位数(个)
数据产品服务价值量	y	软件和信息服务业收入(亿元)

产出方面,提供数据产品和服务的公司数量 q 我们使用信息传输、计算机服务和软件业法人单位数进行计算。这一指标主要是为了计算换算各类平均指标,并不直接放入模型。数据产品服务价值量 y 我们使用软件和信息服务业收入进行计算。其中存在大量统计缺失值,故我们采用插值法来对缺失值进行填补。我们将数据带入生产函数模型进行函数拟合,结果如下:

表 6 “数字产业化”生产模型拟合结果

变量	系数	t 统计量	P 值
c	0.87227	0.666	0.518
$\ln x_1$	1.41337	7.523	0.0000
$\ln x_2$	-0.03194	-0.148	0.885
$\ln x_3$	-0.05276	-1.530	0.152

其中该生产函数的各种检验统计量如下图所示：

表 7 “数字产业化”生产模型统计检验量

统计量	数值
Adjusted R-squared	0.9894
F-statistic	373.6
Prob(F-statistic)	0.000022

从上述结果来看，除 x_2 在 95%的置信度下显著之外，其他三项指标均表现为不显著。其中调整后的 R 方为 0.9894，说明解释变量模型拟合度良好， F 统计量为 373.6，其 P 值为 0.000022。

考虑到有可能是样本数据过少导致的大量变量虽然具有显著的经济意义，但其数学意义在模型中并不显著，故我们选择使用 bootstrap 方式进行方程拟合，我们将重抽样次数设置为 1500 次，得出的具有 95%置信度的方程参数估计为：

表 8 “数字产业化”生产模型参数

变量	系数	标准化后系数
c	2.1681	0.6105248
$\ln x_1$	1.1663	7.523
$\ln x_2$	-0.2474	0.2073762
$\ln x_3$	3.8648	4.6669927

故最终得出方程：

$$\ln y = 0.61052 + 7.523 \ln x_1 + 0.20738 \ln x_2 + 4.667 \ln x_3 \quad (13)$$

（三）“数字产业化”新动能的灰色预测

通过建立 GM（1，1）预测模型，采用 2003 年-2019 年数据产品服务价值指标，预测未来四年的变化趋势，模型预测效果评估如下：

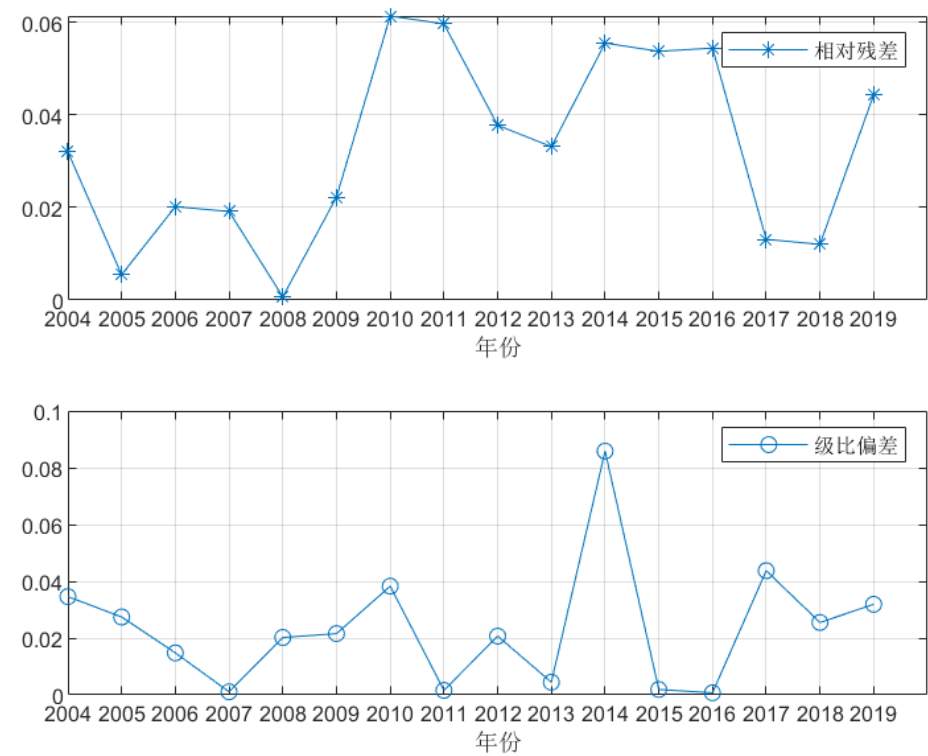


图 7 2003 年-2019 年数据产品服务价值指标相对残差与级比偏差图

表 9 2003 年-2019 年数据产品服务价值指标平均相对残差与平均级比偏差表

平均相对残差	平均级比偏差
0.032813	0.023393

该模型的平均相对残差小于 20%，平均级比偏差小于 0.1，说明 GM（1，1）对原数据的拟合效果较好，预测效果较为可信，最终所得预测结果见下表：

表 10 2020 年-2023 年数据产品服务价值预测值

年份	2020年	2021年	2022年	2023年
产品服务价值	19.9981	20.7105	21.4482	22.2121

(四) 基于“产业数字化”新动能的经济贡献测度

与数字产业化部分的模型建立类似我们首先找到部分计算数值来进行指标的计算，如下表所示：

表 11 “产业数字化”生产模型所用指标

要素	代理变量	对应统计指标
劳动力要素	劳动者报酬	城镇单位就业人员工资总额(亿元)
科技要素	科技研发投入总额	研究与试验发展经费支出(亿元)
资本要素	固定资产投资额	全社会固定资产投资(亿元)

我们使用城镇单位就业人员工资总额(亿元)来替代劳动者报酬，主要是因为劳动者报酬的统计间隔较大，每两年统计一次，而城镇单位就业人员工资总额(亿元)虽然和劳动者报酬的口径不一致，但是由于当下数据新动能在我国经济中的使用多聚集于城镇，数据对于农村的劳动者一是从业人数较少，二是作用效果不强，因而可以使用城镇单位就业人员工资总额(亿元)来替代劳动者报酬。科技要素的指标研发投入总额我们近似的使用研究与试验发展经费支出(亿元)来代替，所以使用这一数据来近似替代。资本要素的固定资产投资总额可以直接使用现有数据全社会固定资产投资。而这里的产出我们直接采用国民生产总值(GDP)来进行衡量。

我们将数据带入到生产函数中模型中拟合出方程结果如下表所示：

表 12 “产业数字化”生产模型拟合结果

变量	系数	t 统计量	P 值
$\ln k$	-0.04405	-0.639	0.5334
$\ln l$	-0.21495	-1.185	0.2566
$\ln s$	0.81524	8.011	0.0000
$\ln d$	0.12795	1.812	0.0916
c	-7.36948	12.843	0.0000

表 13 “产业数字化”生产模型统计检验量

统计量	数值
Adjusted R-squared	0.9987
F-statistic	3387
Prob(F-statistic)	0.0000

从以上两张表格中我们可以看出方程参数中仅有科技要素未通过显著检验，其余变量均通过显著性检验，从上述的结果中可以看出，调整过的 R 方为 0.9974，说明变量对方程拟合的程度良好，方程的 F 统计量为 582.74， P 值为 0.0017。故和之前的做法一样，我们采用 bootstrap 方法，对样本数据进行 1500 次重抽样并进行回归，拟合出的具有 95%置信度的方程系数结果如下图所示。

表 14 “产业数字化”生产模型参数

变量	系数	标准化后系数
c	1.7197	0.17672571
$\ln k$	-0.1129	0.10793980
$\ln l$	-0.1270	0.18963390
$\ln s$	0.5251	0.09375049
$\ln d$	0.1254	0.06629847

由上表可得，通过变换后我们可以得到测算数据要素价值的标准化方程：

$$0.662985 \ln d = \ln y - 0.1767257 - 0.10794 \ln k - 0.189634 \ln l - 0.0937505 \ln s \quad (14)$$

（五）“产业数字化”新动能的灰色预测

通过建立 GM (1, 1) 预测模型，采用 2003 年-2019 年总产出指标，预测未来四年的变化趋势，模型预测效果评估如下：

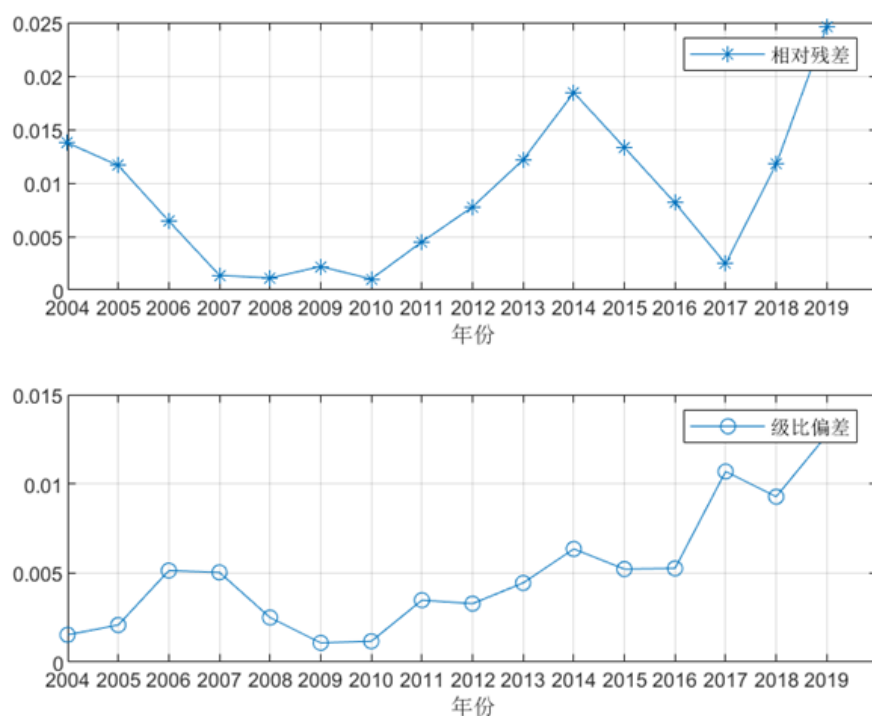


图 8 2003 年-2019 年总产出指标相对残差与级比偏差图

表 15 2003 年-2019 年总产出指标平均相对残差与平均级比偏差表

平均相对残差	平均级比偏差
0.008797	0.004951

该模型的平均相对残差小于 20%，平均级比偏差小于 0.1，说明 GM (1, 1) 对原数据的拟合效果较好，预测效果较为可信，最终所得预测结果见下表：

表 16 2020 年-2023 年总产出预测值

年份	2020年	2021年	2022年	2023年
总产出	2.5566	2.5916	2.6260	2.6607

五、研究结论与建议

（一）研究结论

1、“数字产业化”研究结论分析

通过模型构建中的公式(4)，(5)以时间作差分运算可以计算得出各要素2003至2019年的贡献率，其结果如下图所示：

表 17 各要素贡献率

要素	贡献率
数据行业劳动者	93.5%
硬件设备	2.7%
数据平台	3.6%

从上表的要素贡献率中，我们可以看出，数据行业劳动者的要素贡献率最高，为93.5%，这一点非常符合数据产业本身的服务业性质，服务业最重要的要素便是劳动者，劳动者创造了服务产品中的绝大部分价值，数据行业也不例外，从价值构成上来讲，现在大部分的数据挖掘与数据分析依旧是由数据行业从业者完成的，这一要素贡献率高正体现了这一点。

其次两部分为数据挖掘，处理与储存所需要的硬件投入和数据平台，贡献率分别为2.7%和3.6%。这说明这两部分投入仍有很大的改进空间，硬件设备提升对于数据挖掘的效率，重要数据储存的安全程度都十分重要，对于数据生产活动具有很强的保障作用，应得到重视。对于现有的数据行业的数据平台，从贡献率上来看仍需要优化，一个高质量的数据平台可以提升数据服务的水平，提升顾客的使用体验。在平台上获取数据的方便程度，平台建设的优劣等诸多问题都是影响数据服务业发展水平的重要原因，所以数据平台的建设依旧任重而道远，对于要素贡献率提升仍大有可为。

综上所述，数据新动能在“数字产业化”这一方向上主要的作用主要体现在数据中的信息赋予了数据挖掘分析等一系列数据服务活动经济价值，数据要素自身的重要性使得这一产业不断蓬勃发展，使之成为当下新兴产业的重要组成部分，数字要素市场化进程的不断推进必然会推进数字产业化进程的发展，而数字产业化所提供的数据服务又必然推进产业数字化发展。继续推进我国数字产业化发展，与产业数字化形成合力，对于推动我国数字经济渗透深化，数据新动能进一步迸发具有深刻意义。

2、“产业数字化”研究结论分析

从变换过的推算数据要素的方程中我们可以看出资本要素和劳动力要素的系数分别为-0.21759 和-3.34625，均为负数，与数据要素呈反方向变化，这体现了部分数据要素的特征。我们可以得知数据要素化的发展一般会促进技术密集型经济发展，所以“产业数字化”一定程度上会导致劳动密集型经济下降（退化），也就进而导致方程中用于计算的劳动者报酬这一指标的水平下降。产业数字化的这一特点我们应该去辩证的看待，一方面产业数字化虽然会导致劳动者工资下降，但是同时会促进搞技术水平的劳动者发展，有利于提升我国劳动力的素质水平。但是我们也要注意其大量运用带来的另一方面影响，做好劳动者的教育与生活保障，实现产业数字化的扬长避短，实现劳动者发展保障和产业进步的双赢结果。

同时，资本要素在方程中的系数为-0.21759，也与数据要素符号相反，与劳动力要素符号相反的原因类似。数据要素对于促进生产要素的合理分配以及对于要素利用率具有显著的提升作用，在某些粗放型行业中，仍然是以“高资本投入，低水平的产出”的生产方式为主，而数据要素促进生产要素合理分配，提升要素使用效率的作用，会使得原本资本的大额投入更加趋于合理，减少不必要的资本投入，做到资本的高利用率，提升我国经济体中投资对于经济的作用效果，从而

达到资源的在我国市场的流动与合理配置。

与以上两个要素相反的是科技要素，从方程中我们可以看到科技要素的系数 0.22075 与数据要素均为正数，两变量呈同方向变动。这表明科技要素的投入与发展可以促进数据要素的进步发展，也可以解释为科学技术的进步会增加对于数据的需求。在信息科学与信息经济不断发展的今天，数据作为科学研究与经济发展的重要要素，数据的需求量与日俱增。当下的前沿科技，物联网，人工智能，等等一系列新信息科技的基础均为数据科学。信息技术的核心就在于数据，几乎每一项信息技术，一方面依赖数据，另一方面也产生数据，数据要素在这一过程中均能直接或间接地服务于这些科学技术。尤其是人工智能需要建立在处理大数据的能力基础上，对于人工智能水平的提升更是需要依赖大量的样本数据。从这些角度来看，数据对于当下的新信息科技时代更显重要。

（二）数据驱动中国经济高质量发展的建议

我们将数据要素驱动数字经济发展的作用机制分为两种类型：一是对数据资源分析运用，形成新数字产业和业态，促进产业升级；二是通过要素之间的协同联动，发挥数据要素的乘数效应从而促进数字产业化和产业数字化。

综上所述，对于“数字产业化”，我们需要加强平台建设，增强数据平台提供服务的能力，创造更好的用户体验，提升服务效率，以增加数据平台在“产业数据化”上的贡献。数字经济需要政府治理创新，需要数据要素的市场化配置，政府应将数据作为治理的核心对象，构建数据要素的供求机制、竞争机制和价格机制，针对数据要素市场的“失灵”现象进行合理规制，保障高质量发展。同时我们也要提升数据产业的固定资产的生产水平，提升其生产能力，助力数字服务的生产，实现数字经济的高质量发展。

而数据新动能在“产业数字化”层面的体现更加深刻，主要体现在与其他要

素的结合所产生的新动能上。数据要素化的发展对于促进技术密集型经济发展具有重要作用,“产业数字化”一定程度上会导致劳动密集型产业退化,促进我国经济转型升级实现高质量发展。对于资本要素,数据要素的使用可以促进资产要素的优化配置,提升其使用效率,减少不必要的多余投入,实现投资准确化,精确化,使得资本要素在市场中的流动趋于合理。当下科学技术的进步不断增加着对于数据的需求,而科技要素的投入又反作用于促进数据要素的进步发展。在新信息科学与信息经济不断发展的今天,对于数据要素的开采与应用也已成为推动信息科技发展的最重要因素。为了让各主体拥有充分的数据资源,政府应出台相应制度并加强落实,保障产业发展和政府治理的数据需求,促进数据在政府内部、企业之间、政企之间的融通,以多元共治为路径,平衡公共利益和商业利益之间的冲突,守住安全底线,并以政府统筹为引领,绘制数字经济高质量发展的蓝图。

数据赋能全要素生产力已成为新型经济增长的重要方式,在这样的背景下,我们更需要重视数据要素的作用,完善相对应的体制机制,促进数据要素化有序健康发展,让一切促进经济发展的要素充分涌流。

参考文献

- [1]任晓波. 如何认识和测算数据的经济价值[J]. 竞争情报, 2021, 17(02):2-10.
- [2]李晓华, 王怡帆. 数据价值链与价值创造机制研究[J]. 经济纵横, 2020(11):54-62-2.
- [3]崔俊富, 陈金伟. 数据生产要素对中国经济增长的贡献研究[J]. 管理现代化, 2021, 41(02):32-34
- [4]刘永阔. 基于 CAPM 模型的资产定价问题求解以及模型改进[J]. 商讯, 2021(07):174-175.
- [5]王佳楠, 程港桠, 校园齐, 李嘉伟, 程启先. 基于灰色模糊物元预测的新乡市大气环境质量评价[J]. 湖南理工学院学报(自然科学版), 2021, 34(02):57-62.
- [6]臧国全, 李哲. 数字保存的经济价值——以 American Memory 为例[J]. 图书馆学研究, 2014(11):22-29.
- [7]徐恒, 张梦璐, 孙德厂. 基于 LDA 模型的国内评论挖掘与情感分析领域主题分析与演化趋势[J]. 河南工业大学学报(社会科学版), 2021, 37(02):38-46.

致谢

统计建模暂时告一段落，感谢李凯老师和朱贺老师对我们的帮助和指导。

在确定主题和方向阶段，我们小组讨论了几次依然决定不下，也有些无从下手，最终在李凯老师和朱贺老师的指点下，得以确定方向并且顺利走出了第一步，成功搭建新动能的体系框架。后面整个完成过程也并非一帆风顺，从文本撰写到数据挖掘，再到后期的修改完善，李凯老师俱是尽力的给予帮助，并提供意见。李凯老师细致地和成员讲解关于文本分析和文本挖掘的相关方面，提供各种资料参考学习，力求让大家在比赛的同时，能够学到更多有意义的知识。老师的亲切与耐心，是我们得以成功完成整个成果的最大助力。特向两位老师致以最真挚的感谢。