参赛队号：

2021年（第七届）全国大学生统计建模大赛

参赛学校：中南财经政法大学

论文题目：中国数据中心产业布局的影响因素研究
——基于2010-2019年省级面板数据的实证分析

参赛队员：骆铜林 郑文森 谢京洋

指导老师：姜旭初

中国数据中心产业布局的影响因素研究

——基于2010-2019年省级面板数据的实证分析

目录

表格与插图清单

摘要

在数字经济蓬勃发展的今天，数字基础设施的建设作为“新基建”等概念的重要组成部分，得到了越来越多的关注。而数据中心产业（IDC 产业）就是数字基础设施的重要一类。为了更好地服务数据中心产业的快速发展，有必要对数据中心的产业布局影响因素这一问题进行较为深入的考察与研究。

考察现有文献发现，对数据中心的产业布局问题，大多数研究还停留在定性的经验分析阶段，有数据支撑的模型计算和实证研究还比较缺乏。

本文首先使用层次结构模型，对数据中心选址布局的影响因素进行了初步的探索和分析，并从中选取出六个代表性指标。紧接着，我们使用计量经济学的方法，建立 2010 年-2019 年全国 31 个省级行政区的面板数据模型，对数据中心选址布局的影响因素进行了实证分析。

我们通过多种面板数据模型的估计策略对有关数据进行了回归拟合，发现差分 GMM 法对数据的解释与拟合效果较好。通过对回归拟合结果的分析，我们发现，数据中心产业呈现出较强的集聚性和发展惯性，且数据中心产业较为依赖当地的经济文化产业基础。我们同时发现，西部地区因其得天独厚的自然环境和产业现状，有较好的发展数据中心产业的空间和条件。我们认为，这对我国在制定如“东数西算”等数据中心产业布局的整体规划等方面有着较为重要的借鉴意义。

关键词：数据中心；产业布局；影响因素；面板数据模型

一、 引言

（一） 研究背景

2017 年 12 月 8 日，习近平总书记在中共中央政治局第二次集体学习时指出：“要推动实施国家大数据战略，加快完善数字基础设施……”与此同时，“十四五”规划和 2035 年远景目标纲要也在强调，要做好准备，去“迎接数字时代”。根据有关研究[1]，2025 年，我国数据资源总量预计将增加到 25ZB，在全球达到领先地位。我们认为，数字经济建设的火热，是中国在新时代高速发展的一个切面的形式存在，其特殊处在于因为切合时代关键点，而具有广泛的应用基础和推动产业升级的能力。随着数字经济的迅猛发展，数据已经逐渐成为一种重要的新兴生产要素。而包含数据挖掘与分析、数据的存储与管理、数据的交换与传输等环节的数据产业供应链和价值链也正在快速形成。我们应该深刻地认识到，在数字经济腾飞的今天，作为数据产业链重要一环的数据中心产业（IDC 产业）有其独特的重要性，我们应该抓住机遇培育壮大数据中心产业。

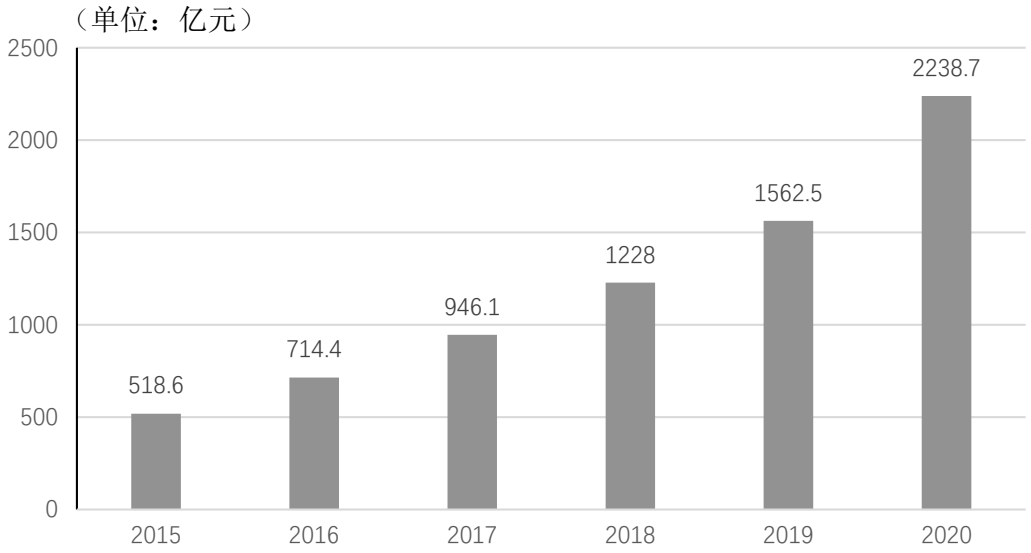


图1　2015-2020年中国数据中心产业规模

随着第四次工业革命掀起时代洪流，如果说科学技术作为第一生产力，那么数据就是第一生产力的时代主流。在先进的技术水平与科技方向指标下，信息在

互联网上的流通必然更加的流畅，信息在末端的收集与上传必然更加的繁荣，也必然伴随着大数据技术的兴起。当今时代，数字经济加速向各行各业融合渗透。数据中心产业作为数字时代数据存储与管理的中心，交换与传输的枢纽，在推动当地传统产业转型升级、提高当地产业的竞争力等方面起到越来越重要的作用。数据中心产业正如一颗冉冉升起的明星，显现出无穷的活力。

相对而言，目前研究方向的主流注重对于数据本身信息挖掘获取与处理的多样化，以图谋求更加有效的信息，更加准确的决策，而数据新动能其载体本身的价值发挥却成为火热研究方向里的角落，结构化的进步纵使再细微也应该被给予足够的重视，成为一个新的思考与研究方向。

（二） 研究意义

“数据中心建设需求与影响因素”就是众多可以深挖的方向中较为特别的一个。

数据中心的产业分布在时间与空间上历经变迁，乃至呈现出今天的分布状况。其总体呈现出“逐数据而居”的特性，大多集中在东部沿海发达地区。但王建冬、于施洋等人的研究指出，大型数字经济企业已经开始着眼于全国性的算力资源总调配，形成了“一路向西”的总趋势[2]，并提出“东数西算”的产业布局框架。2021 年 5 月 26 日，有关部委联合印发《全国一体化大数据中心协同创新体系算力枢纽实施方案》，正式推动“东数西算”战略开始落地实施。

作为数字经济的基石，数据中心代表的不仅仅是需要值得信任、安全的机房网络环境，大规模的场地——相对高昂的建造条件需求与建造成本。更多的是它构成了网络基础资源的一部分，让有所参与的网络用户能够享受更强的数据传输服务和高速接入服务，从基础上进行全局的提升。

我们认为，在面对各地区产业转型发展的需求、数字产业发展分布的极度不

平衡和自然环境资源等多方面矛盾的情况下，若是仍然期待达到长期创造合理价值的经济效益，满足投入产出合理的经济期待，就必须进行相关的建设需求影响因素分析，以图在建造之初就达到结构合理化，从而实现对数据资源长远地合理运用，并最大程度地服务于全国发展的整体布局。

（三） 文献综述

对数据中心的选址布局问题，国内外目前有关的研究还比较少。总的来说，可以分为从地理信息科学角度出发和从经济学等社会科学角度出发两大类。

基于地理信息科学的研究，对数据中心选址问题的影响因素“是什么”比较重视。陈如波（2004）根据在深圳电信数字运营中心选址时的经验，提出应当考虑当地土地使用情况，周边环境，与供电电源距离远近等在当时看来较为新颖的因素。张子仪、蔡泽祥等（2020）通过能量流、业务流和信息流的关系，利用优化选址策略对电力物联网的分布式云数据中心的选址这一具体问题进行了讨论。赵锐（2005）在此基础上，使用空间聚类和计算几何的方法研究了有关的选址问题。王楚伊（2018）则系统地总结了近 30 年国内数据中心选址的研究。他的分析将既有的影响因素概括为自然环境因素、市场需求因素、市政条件因素、政策制度因素四个方面，对国内数据中心的选址问题有较强的指导作用和实际意义。

除了上文叙述的从信息科学角度进行的研究以外，许多学者在从社会科学角度入手的研究中也取得了一定的成果。钟肖英、谢如鹤（2021）采用贝叶斯模型比较法和极大似然估计法，通过空间权重矩阵和空间计量模型的比较，系统地研究了经济增长与数据产业发展的关系。张艺馨（2020）使用有关的省级面板数据，对信息基础设施的建设和经济增长的关系进行了分析。综合他们的研究结论，数据产业，信息基础设施和经济增长三者之间存在互为因果的关系。李英杰（2021）着重分析了地域因素对数字经济和产业结构的影响，发现不同地区的数字经济发

展存在较大差异。王倩、杜卓雅（2021）通过省级面板数据的实证分析，研究了中国区块链产业投资环境的省际差异。他们研究发现，中国的区块链产业存在极不平衡的“东优西劣”现象，且这一不平衡还在继续增强。这对本文的研究方法和理论有一定的指导意义。邓欢（2019）则着重考察在数字经济时代下，中国通信基础设施的产业布局。但邓欢的研究只考虑了政策层面，对数据基础设施选址的客观因素没有进行全面准确的分析。吴蝉丹（2015）从新经济地理学这一交叉学科的视角，初步定量地考察了中国互联网的市场潜能和产业布局。

在梳理现有的相关文献后，我们发现，有关文献大多发表在 2018 年及以后，对数据中心这一新兴产业的研究尚还处于起步阶段，缺乏对它的专一、深入的考察。已有的一些研究文献也缺乏定量的分析，结论大多建立在对经验的分析总结和定性分析的基础上。较少有学者从计量经济学的角度对数据中心产业的布局进行研究。

（四） 研究过程

1. 以 2010 年至 2019 年我国各省兴建的数据中心作为本阶段研究主体，对于可能引起数据中心落址位置与建造规模，特点构筑产生影响的建设需求属性分析后进行提取，查阅资料与探讨后进行相应阶段的属性总结与归纳，建立起层次结构模型，并在开展小组讨论通过后进入下一步的研究。

2. 以上一轮提炼出的建设需求属性作为本阶段研究主体，进行对应建设需求属性的数据查找，通过各种数据渠道收集从 2010 年至 2019 年对应的各项属性信息，并进行数据的初步探索，在开展小组讨论通过后转入下一部分的研究。

3. 以能够对数据中心建设影响因素进行定量分析为目标，选择合适的模型与计量方法，并以模型拟合结果为依据，给出相应的结论和建议。

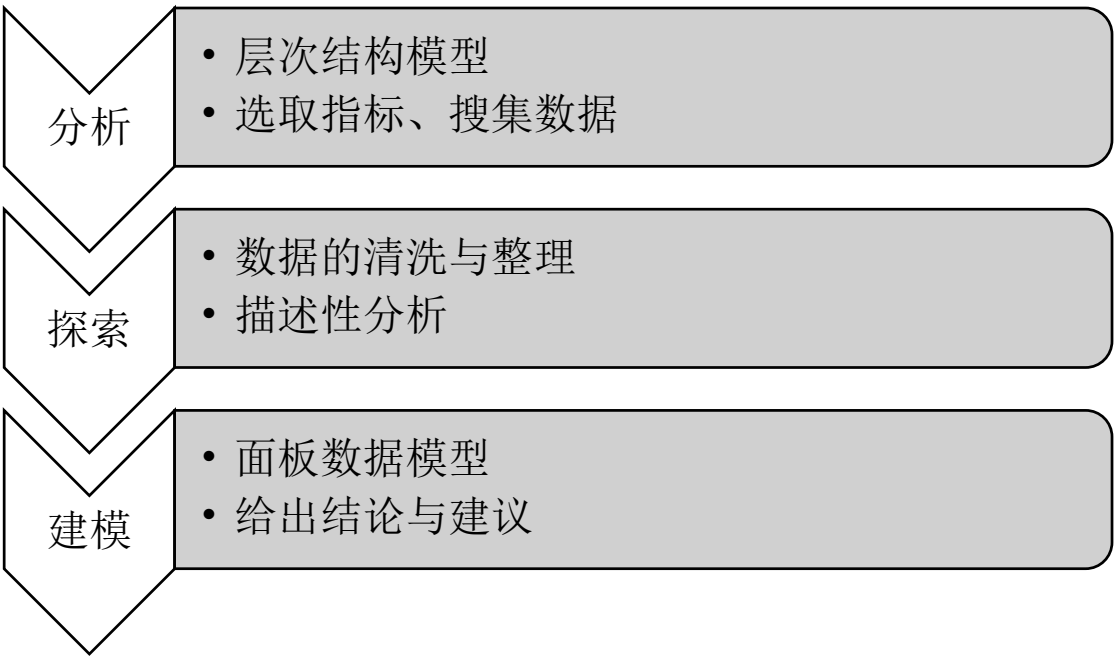


图 2　研究过程图解

（五） 研究问题与方法

我们决定采取先定性，再定量的研究方法。定性分析方面，我们首先采取在社会地理学和计算机科学研究中常见的层次结构模型法。层次结构模型是一种传统上被广泛应用的决策模型。其层层推进并发散的特性对变量的选择有较好的指导作用。具体到本文，我们从目标问题“数据中心选址的影响因素”出发，通过准则层、方法层等层级结构，逐级深入地进行了变量选取的分析。

在进行了变量选取后，接下来，我们在这一问题上创新地使用了计量经济学的办法，利用面板数据模型对数据中心产业布局的影响因素进行实证分析。我们通过各种渠道，搜集，整理数据，最终得到了涵盖中国 31 个省级行政区在 2010-2019 年 6 个有关指标的面板数据。面板数据是计量经济学中的一类重要研究对象。根据我们研究问题的特点，我们构建了动态面板数据模型。计量方法上，考虑到变量内生性的问题，我们采取差分 GMM 法进行参数估计。同时我们也计算了其余几个常见的静态面板数据模型，希望通过对多个模型回归拟合效果的对比，

能得到描述更为准确的计量模型。

本文共分为五个主要部分。

第一部分是引言部分，主要用于总的阐述本次研究的背景、问题、意义和方法。通过对既有文献研究的总结和归纳，提出本次研究的创新点和研究方法。

第二部分主要是对所得数据的初步分析。该部分第一节通过层次结构模型系统而条理清晰地阐述了本次研究变量选取的方法和理由。紧接着叙述本次研究数据的来源和整理的有关过程。第二部分还包含对数据的初步探索性分析。该部分旨在通过一些描述性统计量和可视化图表，为后文的模型建立打好基础。

第三部分叙述本次研究模型的建立过程。第三部分包含对面板数据的预检验，和以此为基础的面板数据分析模型的选择。

第四部分是对模型的求解及其结果的分析。该部分通过一些统计检验方法（如 Hausman 检验）等和回归拟合效果表征量（如 t 值、p 值，回归系数）对模型的拟合效果进行评估、比较、描述和解释。

第五部分是本文的结论和建议部分。这一部分旨在通过对前文的基本总结，试图在理论和实际相结合的基础上对本次研究的问题“中国数据中心产业布局的影响因素”作出初步回答。希望能对我国在数据中心建设的宏观布局上起到一定的作用。

二、 数据的初步分析

（一） 变量选择

1. 层次结构模型

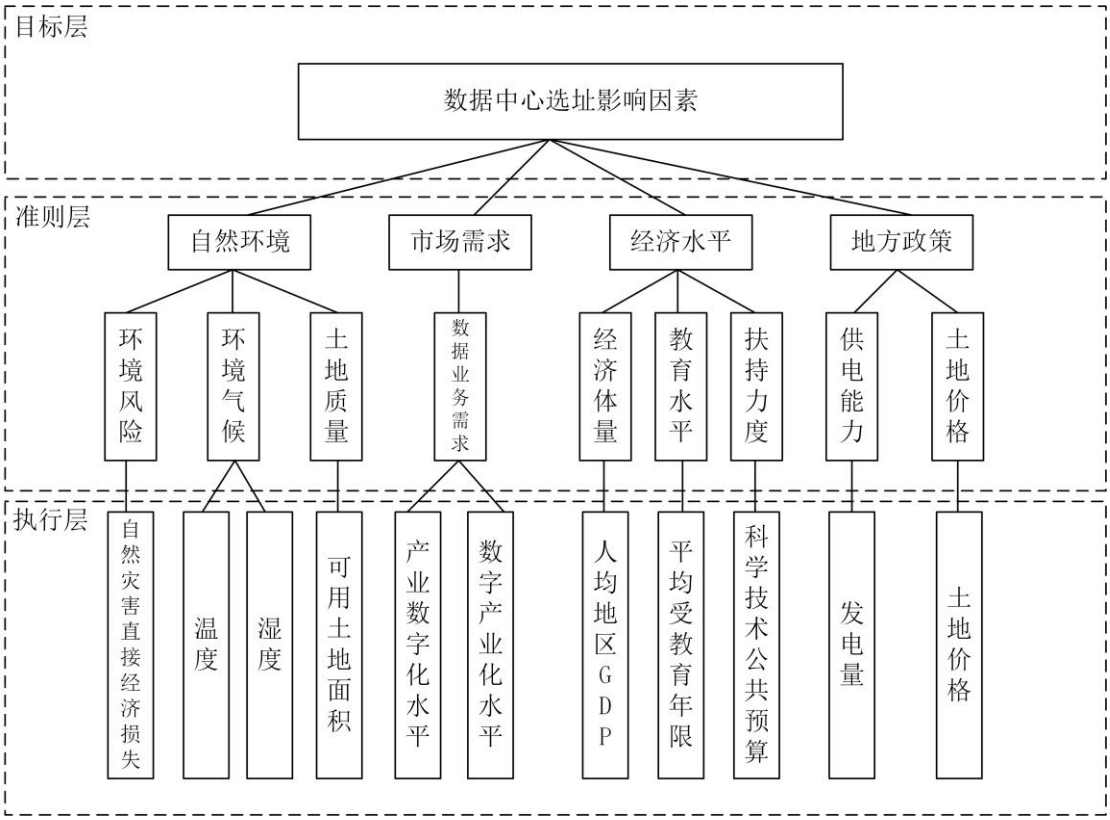


图 3　层次结构模型

自美国运筹学家 Saaty 教授提出层次结构模型以来，该模型已被广泛用于多属性决策领域。我们首先设定该层次结构模型的目标是“数据中心选址的影响因素”，在制定准则层时，为了保证模型发散的完整性和准确性，考虑到数据中心也属于一类公共基础设施，我们参考甘毅（2018）关于公共基础设施选址因素的研究，结合数据中心的特点，给出了如图 3 所示的四个一级因素和九个二级因素。紧接着，分别找到代表九个二级因素的具体指标，形成该层次结构模型的执行层。但是，如此形成的执行层指标体系过于复杂，不便于后续模型的分析。我们着重参考了行业典型项目“深圳电信运营中心”在选址时的考量[9]，在保证指

标代表性的基础上，最终选取了如表 1 所示的六项解释变量。经过专家评估和调整，我们认为这六项解释变量能够较为完整地代表准则层所体现的内涵。

2. 选取指标说明

被解释变量：数据公司保有量

我们借鉴王楚伊（2018）的研究，决定被解释变量是数据公司成立量。本文所称的数据公司，具体是指：行业分类为“信息传输、软件和信息技术服务业”，当前的登记状态为“在业/存续”，公司名中包含“数据”字样的公司。数据服务公司是数据中心产业的直接参与者，是数据中心需求最直观的上级链接点，我们认为，其年成立量能集中反映本省数据中心产业的活力。

市场需求：电信业务总量

互联网发展水平方面，考虑到数据中心的建立是为了给互联网的数据处理等提供必要与加强的数据服务，互联网经济发展水平直接关系到数据中心建立的目的性强弱。繁荣的互联网经济发展对于数据中心建立起到了强制性的导向与坚实的后盾，当互联网经济发展达到较高的水平时能产生对强数据处理能力的需求，以推动数据中心的建设要求，同时较强的经济实力也能够支撑数据中心的建设，从需求到供给形成推动闭环。

我们认为，互联网发展水平的直接体现就是数字产业化水平。而数字产业化的核心产业之一就是电信业，因此本次研究借鉴国内外常用的一种研究评价方式，选取 2010 至 2019 年各省的电信业务总量作为数字产业化水平的代表性指标[5]。

经济水平：人均地区生产总值

各省地域经济发展水平参差不齐，从省级的数据体量直接入手则更能体现经济发展在宏观视角对于数据中心的调控效果，仍符合我们最初对于数据中心的定义——经济发展中新动能之数据基石。同时，两者的强关联性又是否意味着经济

发展到达一定高度数据中心的兴起则成为必然？如此看来显得更加具有探索价值。因此本次研究选择各地区 2010-2019 人均地区生产总值作为变量之一。

地方政策：政府分地区科学技术公共预算

本文中地方政府预算支出额，具体指各省市财政一般公共预算支出中科学技术支出占比的部分，为了弥补单纯 GDP 精细度不足的缺陷，本次研究引入更加具有针对性的科学技术公共预算变量，尽可能减少地区发展结构差异带来的数据浮动，使得相关经济数据的作用范围更加明确，同时结合“关键”，“基础”，“必不可少”的选择方针进行思考，最后将政府分地区科学技术公共预算作为控制变量之一。

环境风险：自然灾害直接经济损失

数据中心产业具有投资大、使用年限长等特点，尤其是其存储与交换的数据，一旦出现丢失、损坏等问题，较造成较为重大的损失。较为频繁和严重的自然灾害将增加数据中心的运行成本，并严重干扰数据中心的正常运行。我们选取自然灾害直接经济损失这一指标以代表自然灾害的影响。

教育水平：平均受教育年限

教育是一切产业的发展基础。借鉴黄园（2014）和 Psachropoulos（1986）的方法，我们决定使用 6 岁以上人口平均受教育年限来作为当地的受教育程度的评估。计算公式为：

$$H = \frac{\sum_{i=1}^{n} r_i m_i}{R} \qquad \text{（式 1）}$$

式中，H 为所求的平均受教育年限，i表示《中国统计年鉴》中惯常划分的：未上学、小学、初中、高中、大专及以上 5 个阶段；r_i表示受相应阶段教育的人口数量；m_i表示不同教育阶段对应的受教育年限，具体地，未上学对应 0 年，小学对应 6 年，初中对应 9 年，高中对应 12 年，大专及以上对应 16 年。

另外，考虑到数据中心对电力资源的高需求，我们选取该地区的年发电量作

为指标。

最后，我们将本节选取的变量及其符号总结见如下表 1：

表 1 选取的变量及其符号

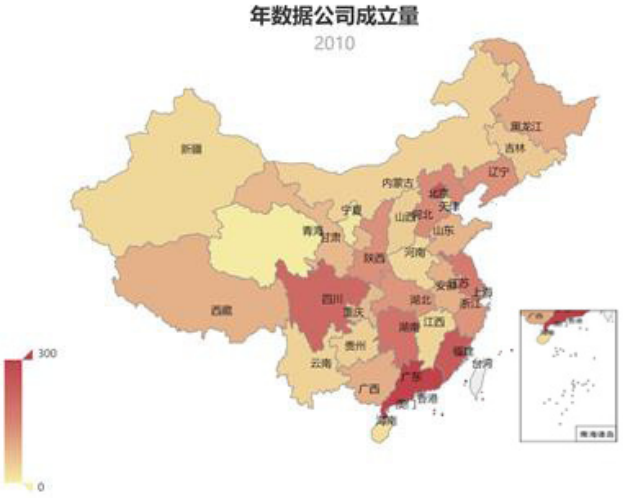
变量名称	变量符号
数据公司成立量	*company*
电信业务总量	*tts*
平均受教育年限	*edu*
发电量	*eep*
政府科技领域投入预算	*govern*
自然灾害	*disaster*
人均 GDP	*pgdp*

（二） 数据来源

限于香港、澳门、台湾地区的数据不全或难以获得，本文选取中国其余 31 个省级行政区的数据进行分析。年数据公司成立量来源于“企查查”网站，其余指标的数据均来源于 2010-2020 年度的《中国统计年鉴》，国家统计局和工业和信息化部官方网站。

（三） 描述性统计分析

对于中国数据中心分布的情况，我们可以暂时从人均 GDP，政府科技预算投入，平均受教育年限，年数据公司成立量等方面入手，其中不同地区的年数据公司成立量与中国数据中心产业布局的关系显然最密切。由此我们给出下文的初步分析。



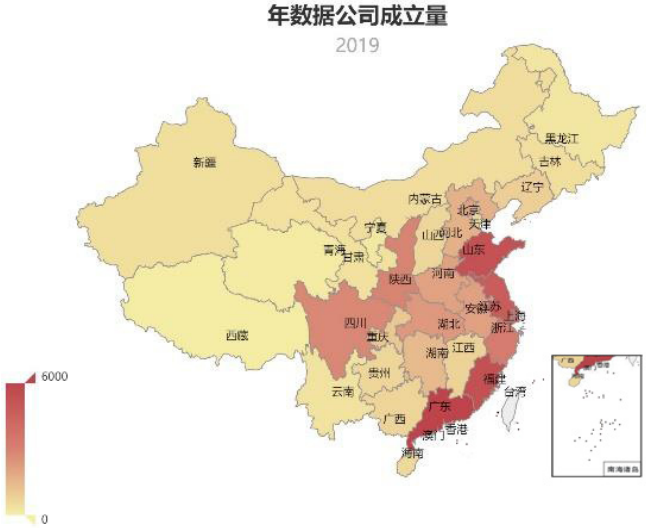


图 4 2010 年数据公司成立量　　**图 5 2019 年数据公司成立量**

图 4 给出的是 2010 年数据公司成立量，图 5 给出的是 2019 年数据公司成立量（由于台湾地区数据缺失，故以灰白色统计）。对比两图可以发现，2019 年数据公司的成立量较 2010 年数据公司成立量有大幅度上升。2010 年时数据公司成立量最高的广东省也仅仅刚超过 400 家，但到 2019 年大部分省级行政区已超过 500 家，广东省更是到达了 15000 家。由图 1 图 2 共同可得，年数据公司成立量高的省级行政区大多集中在东部沿海地区，且其分布具有地域性，所以我们之后将以东部，南部，西部，北部，中部五个地区各具代表性的省级行政区即江苏，广东，新疆，吉林，湖北加上处于西部地区但年数据公司成立量远高于其他西部省级行政区的四川作为典例省份作为观察样本（由于广东省部分数据远高于其他省份，故部分折线图不予显示）。

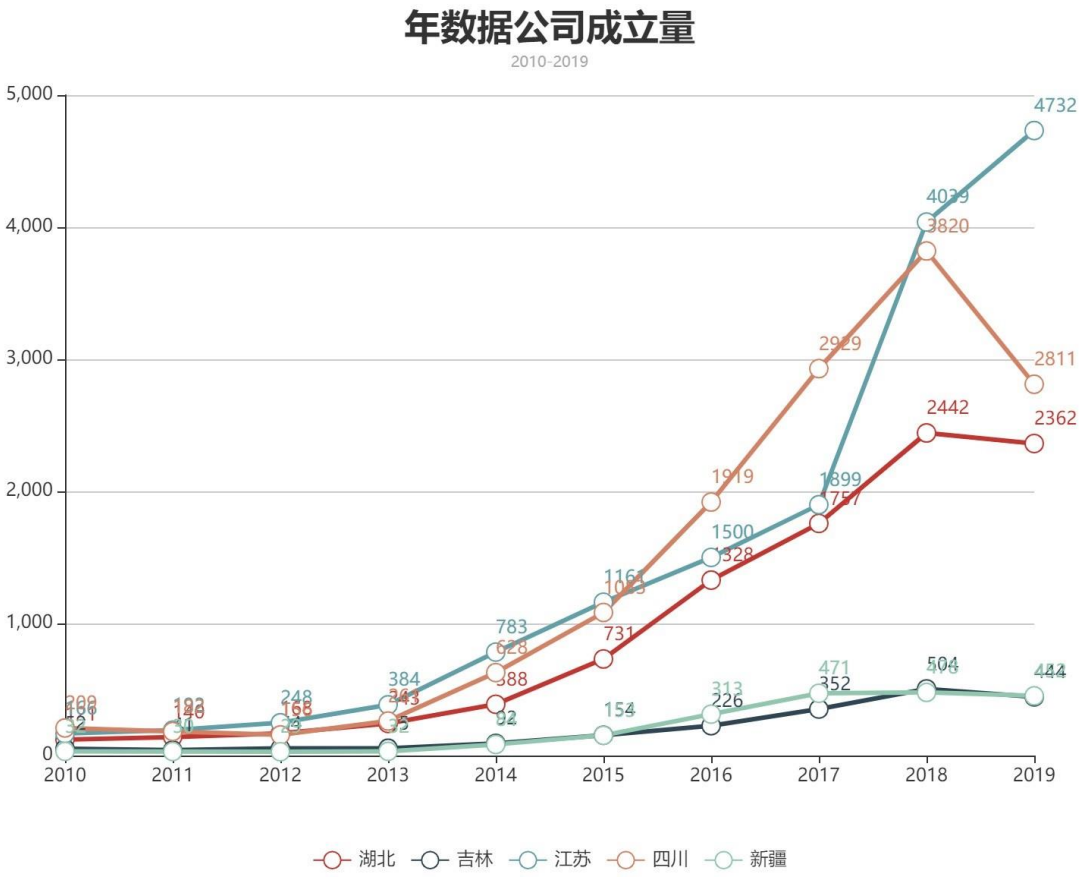


图 6 2010-2019 五省级行政区年数据公司成立量

由图 6 可以看出，总体而言各个省级行政区年数据公司成立量在十年中呈不断增长的趋势。但不同地区年数据公司成立量增长幅度有较大区别。在 2010 年各地相差最多不超过四倍，而在 2019 年已经相差十倍。由此可见年数据公司成立量基数高及增长快是一致的，且基本集中于中部，东部，南部，南部尤胜，这与我国经济发展地区分布有着较大关联。参考图 7 即可印证此观点。

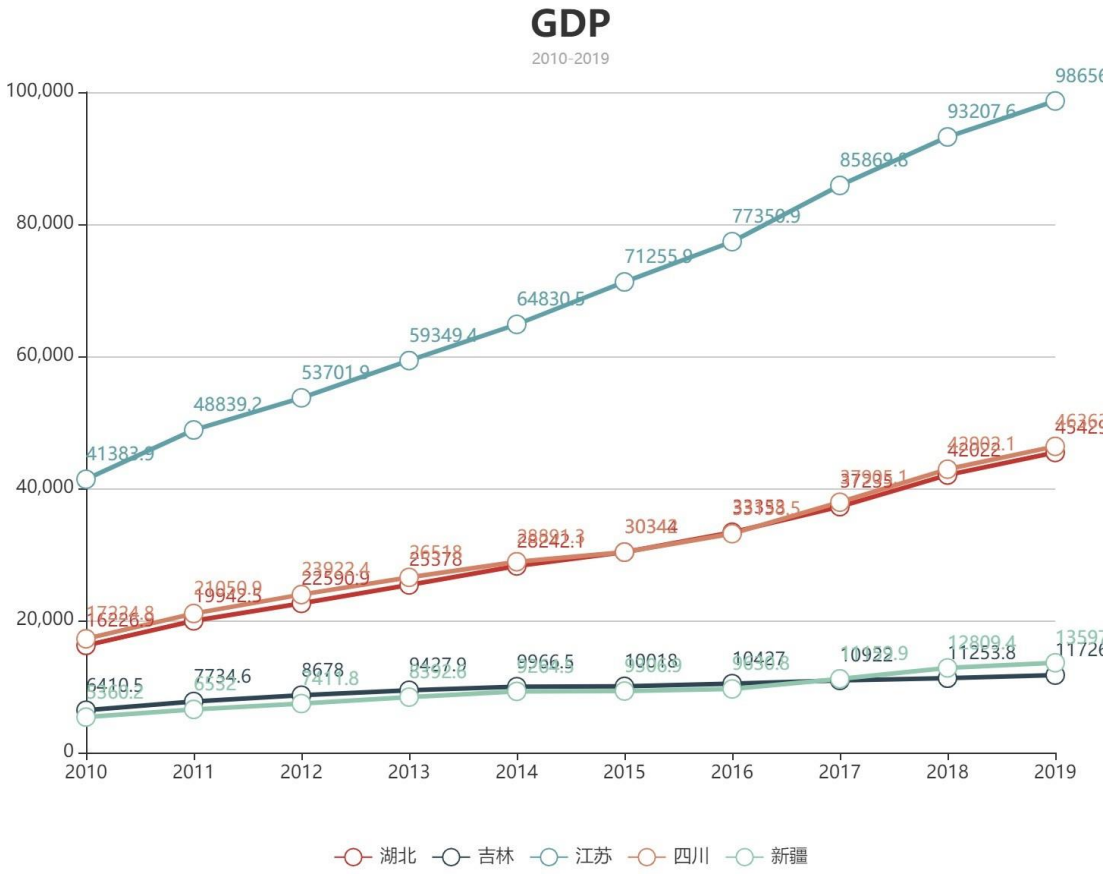


图 7 2010-2019 年五省区 GDP 增长曲线

结合图 7 与广东省 GDP 指标来看，年数据公司成立量增长曲线与各省份 GDP 增长曲线基本吻合，大致可分为四档，即西部和北部为一档，中部与四川为一档，东部沿海为一档，南部的广东省独为一档。将图 6 与图 7 对照也可得出，经济发展可以促进年数据公司成立量的增长。

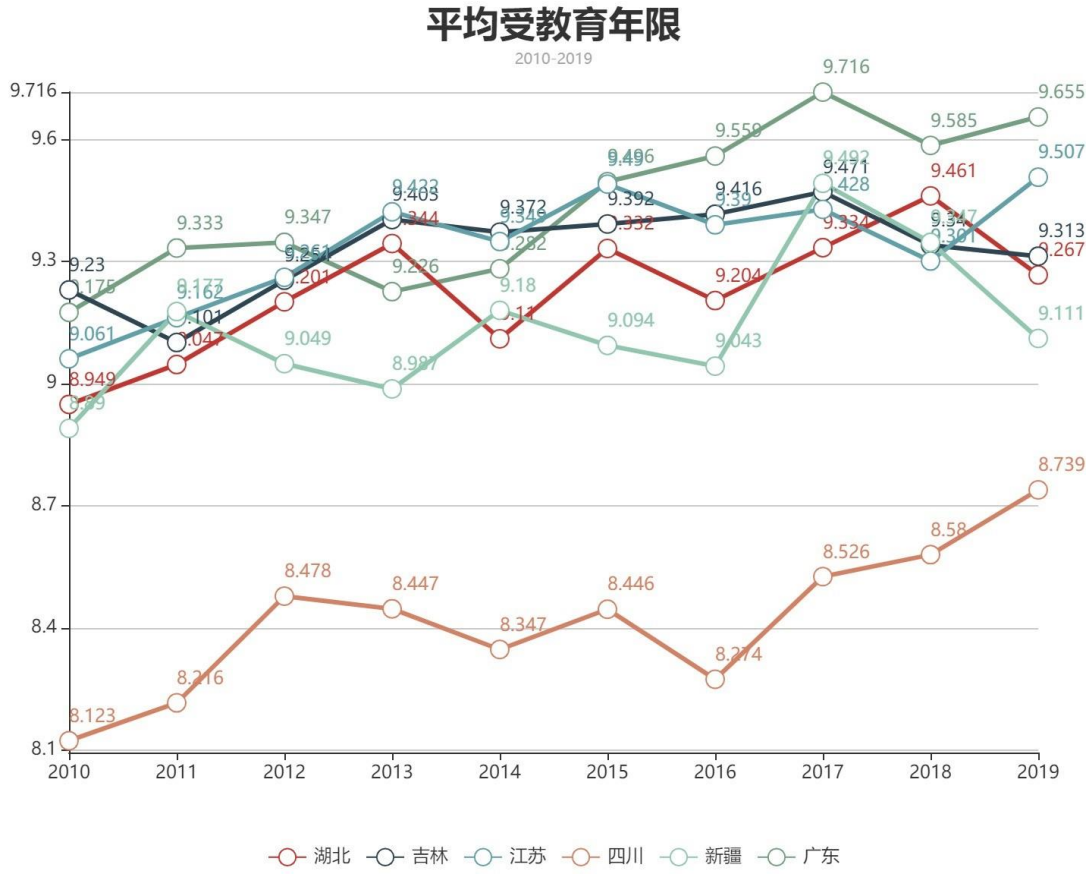


图 8　2010-2019 年五省区平均受教育年限

表 2 变量的统计性描述

变量名称	平均值	最大值	最小值	标准差
数据公司成立量	5.53	9.65	1.6	1.48
电信业务总量	10.55	120.46	0.23	13.69
平均受教育年限	8.98	12.08	4.22	1.05
发电量	18.67	58.97	0.19	12.82
政府科技领域投入预算	11.02	116.87	0.27	14.58
自然灾害	12.17	120.26	0	14.32
人均 GDP	4.95	16.45	1.28	2.63

因为原始数据的各变量间数量级差别较大，例如：地区平均教育年限的取值

在 10 左右，而人均 GDP 的取值到达了十万级。变量间的数据级差别过于悬殊会导致计算机在计算模型时出现较大误差。为了解决这个问题，在后续的计算分析时我们对数据公司成立量作了取对数处理，并对所有变量使用归一法进行了标准化处理，以保证各指标之间的可比性。

从表 2 数据可以得出，取数据公司成立量最大值 9.65 与最小值 1.6 相差较大，由此可见数据公司成立量存在极大的地区发展不平衡性。再观察其标准差并结合图 4 与图 5 之间的变化，可以得出在 2010-2019 年期间数据公司成立量发展呈现井喷式增长，而结合其对数平均值可以发现，中位数远低于平均数，这表明最大值远高于其他地区，地区发展不平衡性在这里得到了充分的体现。浏览表格，电信业务总量与数据公司成立量比例基本相同，这也再次点明了数据公司成立量发展地区与地区之间存在巨大差距。由图 7 可以看出平均受教育年限相差不是很大，但明显和地区经济发展存在一定联系。而发电量与人均 GDP 可以结合起来，都是数据公司成立量发展的前提条件，自然灾害则是不可控变量，但大体上与地区仍存在一定联系（如位处地震带等）。最后由政府科技领域投入预算的数值可以看出，对不同地区的投入预算差别极大，这也就部分导致了数据公司成立量地区发展不平衡性。这种各变量在地区内部变化的一致性，为后文的研究奠定了基础。

三、 模型的构建

（一） 面板数据的构建

面板数据指的是这样一组数据，其包含了一段时间以内对一组个体的多个特征的跟踪。一般地，我们记其横截面维度（个体数目）为 n，时间维度（时间跨度）为 T[3]。

为了后续使用 Stata 等统计软件进行分析的便利性，本文将上述若干省际指

标在 2010-2019 年的取值排列生成面板数据。其结构如下表所示：

表 3　面板数据结构

	year	*company*	*Edu*	*GDP*	……	*disaster*
省 1	2010					
省 1	2011					
……						
省 1	2019					
省 2	2010					
省 2	2011					
……						

如果要建立对面板数据的分析模型，首先需要明确这个面板数据的分类。我们将本文源数据导入 Stata，执行有关命令，得到如下的结果：

```
. xtdes

    code:  1, 2, ..., 31                                    n =         31
    year:  2010, 2011, ..., 2019                            T =         10
           Delta(year) = 1 unit
           Span(year)  = 10 periods
           (code*year uniquely identifies each observation)

  panel variable:  code (strongly balanced)
```

图 9 面板数据结构信息

分析结果显示，本次研究预先所得到的面板数据其 *n* 值为 31，*T* 值为 10，符合 *n* > *T* 的条件，是一个“短面板”。“strongly balanced”的分析结果提示面板数据是一个“平衡面板”。

（二） 面板数据的估计策略

有关面板数据的分析模型，已有较多且比较成熟的研究。对于静态面板数据，有“个体效应模型”、“固定效应模型”（FE）、“随机效应模型”（RE）三种模型是

既比较常见，也比较适合本文的研究的。本节简要介绍它们的定义。

“个体效应模型”既考虑到了个体间的共性（相同斜率的回归方程），也考虑到了个体间的差异（不同截距的回归方程）。“个体效应模型”一般可以用式 2 来表示：

$$y_{it} = x_{it}\boldsymbol{\beta} + z_i\boldsymbol{\delta} + u_i + \varepsilon_{it} \quad (i = 1, \dots, n; t = 1, \dots, T) \qquad （式 2）$$

其中x_{it}代表了可以随着时间变化的个体特征,z_i代表了不会随着时间变化的个体特征。u_i代表了个体间的固定差异。ε_{it}代表了同时随个体与时间的扰动项，且该项与u_i不相关。

进一步地，如果u_i与某个解释变量相关，则为“固定效应模型”；如果u_i与所有解释变量都不相关，则称其为“随机效应模型”。

动态面板模型一般考虑在静态面板模型的基础上加入滞后的被解释变量，一般加入滞后一期，得到式 3：

$$y_{it} = \alpha + \rho y_{i,t-1} + \beta x_{it} + u_i + \varepsilon_{it} \quad (t = 2, \dots, T) \qquad （式 3）$$

对于动态面板数据的估计，因其组内估计量（FE）的不一致（Nickel,1981），需要另外的方法进行计算。现有研究一般采取以下三种方法：Arellano、Bond 在 1991 年提出的的“差分 GMM”法和他们在 1995 年提出的“水平 GMM”法，以及 1998 年 Blundell 和 Bond 的“系统 GMM”法。这几种方法在解决动态面板数据的变量内生性问题方面也有一定的作用。

（三） 数据的预检验

面板数据作为存在时间变量的数据类型，其平稳性是进行研究的基础。所以我们首先进行单位根检验。我们采取费雪式检验法，使用滞后两期的 ADF 回归检验 *lncompany*，同时认为该被解释变量存在漂移项。利用 Stata 内置的命令 xtunitroot fisher，我们得到如下结果：

```
. xtunitroot fisher lncompany, dfuller drift demean lags(0)

Fisher-type unit-root test for lncompany
Based on augmented Dickey-Fuller tests
------------------------------------------
Ho: All panels contain unit roots          Number of panels  =     31
Ha: At least one panel is stationary       Number of periods =     10

AR parameter: Panel-specific               Asymptotics: T -> Infinity
Panel means:  Included
Time trend:   Not included                 Cross-sectional means removed
Drift term:   Included                     ADF regressions: 0 lags
------------------------------------------------------------------------------
                              Statistic      p-value
------------------------------------------------------------------------------
 Inverse chi-squared(62)   P      189.9816       0.0000
 Inverse normal            Z       -8.6597       0.0000
 Inverse logit t(159)      L*      -9.0332       0.0000
 Modified inv. chi-squared Pm      11.4931       0.0000
------------------------------------------------------------------------------
 P statistic requires number of panels to be finite.
 Other statistics are suitable for finite or infinite number of panels.
------------------------------------------------------------------------------
```

图 10 单位根检验结果

检验结果显示，所有 4 个统计量均强烈拒绝面板数据存在单位根的假设（H_0），其相应的 p 值为 0.0000，故我们的被解释变量是平稳的，可以继续进行研究。

接着，我们对面板数据继续进行协整性检验。我们使用 Pedroni 检验法，对面板数据使用 Stata 内置命令 xtcointtest pedroni 后，得到结果如图 11 所示：

```
. xtcointtest pedroni lncompany edu pgdp tts eep govern disaster, trend lags(2) demean

Pedroni test for cointegration
------------------------------
Ho: No cointegration                       Number of panels       =     31
Ha: All panels are cointegrated            Number of periods      =      9

Cointegrating vector: Panel specific
Panel means:          Included             Kernel:           Bartlett
Time trend:           Included             Lags:             2.00 (Newey-West)
AR parameter:         Panel specific       Augmented lags:   2

Cross-sectional means removed
------------------------------------------------------------------------------
                                           Statistic         p-value
------------------------------------------------------------------------------
 Modified Phillips-Perron t                 10.0192          0.0000
 Phillips-Perron t                         -30.7329          0.0000
 Augmented Dickey-Fuller t                 -34.7155          0.0000
------------------------------------------------------------------------------
```

图 11 协整性检验结果

其有关的三个 p 统计量值均为 0.0000，显示强烈拒绝原假设 H_0 “面板数据不协整”，也即面板数据通过协整性检验，可以继续进行分析。

（四） 数据中心选址影响因素计量模型的构建

参考侯继森、李英杰的研究，考虑到数据中心的选址是一个影响因素多，变

量关系复杂的问题，为了更加准确的研究影响数据中心选址的因素，我们将滞后一期的年数据公司成立量加入模型中.得到如下的动态面板模型：

$$lncompany_{it} = \alpha + \rho lncompany_{i,t-1} + \beta x_{it} + u_i + \varepsilon_{it} \quad （式 4）$$

式中，被解释变量为 *lncompany*，代表了省份数据中心的数量，x_{it}是解释变量，包含了上文 2.1 节所提到的电信业务总量（*tts*）、年发电量（*eep*）、政府科技投入预算（*Govern*）、自然灾害直接经济损失（*disaster*）、平均教育年限（*edu*）。*i* 代表本文研究的中国 31 个省级行政区，取值为 *i* = 1,2,⋯31，*t* 代表年份,取值为 *t* = 2010,2011,⋯,2019，u_i表示个体间的差异，ε_{it}代表同时随时间和个体改变的扰动项。

因为动态面板模型不可避免会涉及到内生性问题，我们采取 Arellano 和 Bond 在 1991 年提出的广义矩估计（差分 GMM）来对模型进行估计[8]。

在使用差分 GMM 法进行计算时，我们指定电信业务总量（*tts*）为内生解释变量，并最多使用二阶滞后值作为工具变量。我们设定使用允许存在异方差的稳健标准误，并针对两步 GMM 的估计进行了调整。

另外，为了选择出最优模型并评估各模型的拟合效果，本文也同时使用静态面板数据的多种计量方法，以作为参照系。静态面板数据模型的基础表达式为：

$$lncompany_{it} = \beta x_{it} + u_i + \varepsilon_{it} \quad (式 5)$$

四、 模型的求解与分析

（一） 模型的求解与检验

利用 Stata 的内置有关命令进行模型求解，得到的结果如表 4 所示：

表 4 模型计算结果

变量名称	静态面板			动态面板
	OLS	FE	RE	差分 GMM
lncompany$_{-1}$	-	-	-	0.717***
				(26.64)
edu	0.071	0.752***	0.092	0.236***
	(0.75)	(4.97)	(0.92)	(2.97)
pgdp	0.243***	0.464***	0.411***	0.220***
	(4.12)	(12.07)	(6.08)	(7.23)
tts	0.039***	0.005	0.016**	0.006***
	(4.24)	(1.21)	(2.51)	(5.17)
eep	0.031**	0.069***	0.060	0.040***
	(3.17)	(7.10)	(4.66)	(4.76)
govern	0.006	0.001	-0.007*	0.003*
	(-0.51)	(0.25)	(-0.79)	(1.42)
disaster	0.010*	-0.003*	-0.000*	-0.021*
	(1.80)	(-0.98)	(-0.22)	(-2.06)

表中括号内的数值为相应项系数估计的 t 统计量值，“***”、“**”、“*”分别表示在 0.01、0.05、0.1 的显著性水平上显著。

首先是对混合回归和固定效应模型计算结果的比较。在进行以普通标准误为基础的计算时，Stata 会自动输出一个 F 检验。F 检验其原假设 H_0 为“all $u_i = 0$”，即混合回归的假设前提成立。对于本次实验，Stata 输出的 F 检验结果为：

```
F test that all u_i=0: F(30, 273) = 19.73                     Prob > F = 0.0000
```

其 p 值为 0.0000，表示强烈拒绝原假设，我们可以据此认为 FE 的效果明显优于混合回归，每个个体应该有其自己的截距项。

其次是对固定效应模型和随机效应模型计算结果的比较。豪斯曼检验是对比这两个模型的经典方法。使用 Stata 内置的命令 hausman FE RE，我们得到如图 12 所示的结果：

```
. hausman FE2 RE2,constant sigmamore
```

	Coefficients (b) FE2	(B) RE2	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
edu	.7517136	.092291	.6594226	.1520127
pgdp2	.4643184	.4111015	.0532169	.02703
tts2	.0047404	.0163377	-.0115973	.0014193
eep2	.0692311	.0600296	.0092015	.0092206
govern2	-.0014447	-.0070655	.0056209	.0029513
disaster2	-.0025891	-.0006735	-.0019156	.0006936
_cons	-4.812883	1.461175	-6.274058	1.291252

```
                b = consistent under Ho and Ha; obtained from xtreg
 B = inconsistent under Ha, efficient under Ho; obtained from xtreg

Test:  Ho:  difference in coefficients not systematic

              chi2(7) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                      =       94.12
            Prob>chi2 =      0.0000
            (V_b-V_B is not positive definite)
```

图 12 豪斯曼检验结果

豪斯曼检验的 p 值为 0.0000，显示强烈拒绝原假设 H_0：“u_i与x_{it}, z_i不相关”，故我们认为固定效应模型对本问题的拟合效果显著优于随机效应模型。

最后是对差分 GMM 拟合效果的评估。差分 GMM 成立的前提是扰动项ε_{it}不存在二阶或更高阶的自相关。根据以上计算出来的结果，考察扰动项的差分是否存在一阶与二阶自相关，使用 Stata 内置命令 estat abond 进行计算，得到结果为：

Order	z	Prob > z
1	-3.3474	0.0008
2	.18841	0.8506

结果显示，AR(1)的 P 值为 0.0008，AR(2)的 P 值为 0.8506，也即扰动项的差分存在一阶自相关，且不存在二阶自相关，故接受原假设 H_0：“扰动项无自相

关”。

为了进一步地确认差分 GMM 法使用的前提条件，我们继续进行 Sargan 检验，以保证工具变量的外生性。使用 Stata 得到的计算结果如下：

```
. estat sargan
Sargan test of overidentifying restrictions
        H0: overidentifying restrictions are valid

        chi2(46)     =  28.72465
        Prob > chi2  =    0.9784
```

显示 Sargan 检验的 p 值为 0.9784，无法拒绝“所有工具变量均有效”的原假设，我们据此认为所有的工具变量是有效且合理的。

综合以上的分析讨论，我们认为对本研究使用差分 GMM 法是能够成立且有实际意义的。同时，考虑到差分 GMM 法的各项统计学指标均优于静态面板，各变量的回归系数都能较好地通过显著性检验，我们认为动态面板模型的建立是成功而有效的。

（二） 模型结果的分析

本节的讨论基于表 4 中差分 GMM 法的分析结果。

年数据公司成立量的一阶滞后项系数达到 0.717 且为正数，表明数据公司的成立量较强地受上一期成立量的影响。数据公司的选址在省级层面上表现出较强的惯性和集聚效应。数据公司的成立省份选择表现出特定的偏好。

平均受教育年限和人均 GDP 两项在回归方程中的系数也比较大，且显著为正。教育和人均 GDP 属于体现省级经济文化发展程度的两个指标。回归结果表明，经济文化发展程度较高的省份在数据中心产业上有更强的吸引力。这也与数据公司属于高新产业，他们的发展需要有较强的经济文化发展程度作为支撑。

年发电量、年自然灾害直接经济损失是数据公司所需客观环境的代表指标。在回归结果中，其系数分别为 0.040 和-0.021，且分别在 0.01，0.1 的显著性水平上显著。年发电量和自然灾害造成的损失对数据公司的选址分别起到正向和负向

作用，这一结果也较为符合常识。回归结果表明，发电量、自然灾害在数据公司的选址策略上有一定的影响，但不是很大。我们认为这与全国层面的电力资源统一调配和防灾减灾工作的持续发展有一定关系。这两项的研究同时表明，客观环境在数据公司的选址考量上 并不处于靠前的位置。

除此以外，政府科技预算的投入和省年电信业务总量在回归系数上仅分别为0.003 和 0.006，虽然在统计学上也有较好的显著性，但因数值过小，在经济学上的意义不是很大。有鉴于此，我们认为这只能表明它们能在对数据公司的吸引上起到一定的正向作用。

五、 结论与建议

（一） 结论

本文以我国 31 个省（直辖市、自治区）在 2010 到 2019 这十年间的有关指标出发，对数据中心选址布局的有关因素进行了简要分析。我们在经过初步的描述性统计分析和进一步的构建面板数据进行回归拟合过后，根据实证分析的结果，结合各省较为突出的特点，得出以下几点结论：

1. 数据中心产业的发展显示出集聚性和惯性

正如上文所言，滞后一期数据公司成立量的系数较大且显著为正。这表明数据中心产业的发展显示出较大的集聚性和惯性。事实上，在最近几年数据中心产业的发展实践中，产生了如贵州贵安新区、仙桃国际大数据谷等规模较大，影响力显著的数据中心产业集聚区[4]。我们认为，数据中心产业通过区域集聚，能降低附属设施建设成本，达到更好地吸引人才，形成产业协同发展的作用[6]，也即“空间优势能转化为算力优势”。

2. 经济文化的发展是数据中心产业的动力源

数据中心产业作为进入互联网时代以来的新兴产业，带有强烈的科技属性。而经济与文化的发展往往存在相互关联的关系。在我们的回归分析中，经济文化的发展程度与数据中心产业的发展有较高的关联性。我们认为，经济条件决定了数据中心产业的需求。可以预见的是，以数据中心为代表的数字经济在不远的未来将快速融入社会生产的各行各业。经济活力强的地区，将会产生旺盛的数据中心产业需求，能有力地促进数据中心产业的发展。另一方面，教育基础决定了数据中心产业的供给。数据中心产业对以信息科学为代表的一系列前沿热点领域有较强烈的依赖。其发展目前受有关专业的大专及以上人才数量的限制还较为显著。能不能吸引或留住高等教育人才，尤其是信息科学有关的高等教育人才，将成为未来各省份竞争发展数据中心产业的一大看点。

3. 西部地区有进一步发展数据中心产业的条件

长期以来，西部地区省份存在数字产业化水平不高，自然灾害造成经济损失较高等劣势，但也具有巨大的能源，尤其是可再生能源的储备的特点。通过我们的实证分析，我们发现，数据中心产业的分布与当地自然灾害和电信业务总量并无显著关系。我们认为，这是因为近年来，随着现代地理科学的发展和防灾减灾工作的持续推进，能直接造成数据中心损失的自然灾害已经很难发生。事实上，我们注意到，北京、上海、深圳等地均在近年发布文件，在环保节能，建设规模，建设区域等多方面提高数据中心产业的进入门槛。西部地区的特点事实上形成了承接东部地区转移的数据中心产业的基础，具备进一步发展数据中心产业的条件。

（二） 建议

根据以上结论，我们对数据中心产业的布局和发展给出以下建议：

1. 持续推进实体经济文化领域的建设，加速数字经济向全产业的融合发展

经济文化事业是一个省份综合实力的直接体现，需要从各个方面进行总体的提升。数据中心产业作为数字经济时代的“新基建”，其发展也高度依赖于实体经济的发展。在产业结构升级的大背景下，数据中心作为数据交换的中心和枢纽，其发展与上下游产业链的完善越来越呈现出相辅相成的关系。经济文化的发展能显著抬升数据中心产业的需求，对数据中心产业的发展能起到较强的促进作用。只有坚定不移地走“科教兴国”的路线，才能牢牢抓住数据中心产业发展的战略机遇。

2. 进一步推进“东数西算”战略建设，完善信息基础设施建设

正如前文所言，我国数据中心产业在不同地域的发展极不平衡。数据中心产业的选址策略和数字经济的发展趋势逐渐形成冲突，尽快推进“东数西算”的落地实施，成为化解矛盾的有利策略。一般来说，东部地区数据“算力”需求较大，但是能源成本较高，西部地区能源成本较低，但数据业务量较小。考虑到在回归分析中电信业务总量和自然灾害直接损失两项的系数不显著，我们认为在西部地区推动数据中心产业的发展有较大的空间和潜力。

同时，为了降低数据传输的成本和时延，也需要重点推进跨地区的信息基础设施建设。通过打通数据中心产业上下游，推动西部地区能更好地承接东部地区转移的数据中心产业的需求，实现数字经济的区域协调发展。

（三） 研究局限与展望

在本次研究接近尾声之际，我们研究小组也对整个研究过程进行了一些反思和总结：

第一是我们指标选取的方式上，我们着重采取了层次结构模型进行分析。层

次结构模型是一种定性的分析方式，存在较大的主观性，容易造成选取变量的遗漏。我们在进行研究时，缺少与有关方面的专家、从业人士的请教和交流，这也容易造成我们的研究结论的失真。

第二是在建立面板模型进行回归拟合分析时，我们只进行了全国层面的研究，没有进行分地域、分地区的讨论。考虑到中国巨大的地域差异性，我们研究得出的结论对特定地区是否存在适用性尚还需要进一步的讨论。

除此以外，限于时间和能力有限，本次研究还存在诸多遗憾与不足之处，恳请老师批评指正。

参考文献：

[1] 国家信息中心.培育壮大数据中心产业势在必行[EB/OL].http://www.sic.gov.cn/News/611/10926.htm, 2021-05-17

[2] 王建冬,于施洋,窦悦.东数西算：我国数据跨域流通的总体框架和实施路径研究[J].电子政务,2020(03):13-21.

[3] 陈强.高级计量经济学及 Stata 应用[M].北京:高等教育出版社,2014.

[4] 陈健,陈志.从规模增长走向价值增长——新基建背景下大数据中心产业发展的问题与思考[J].科技中国,2021(04):60-63.

[5] 李英杰,韩平.数字经济发展对我国产业结构优化升级的影响——基于省级面板数据的实证分析[J].商业经济研究,2021(06):183-188.

[6] 张艺馨.信息基础设施建设对我国经济增长影响研究[D].吉林大学,2020.

[7] 王楚伊.中国数据中心选址发展研究——以近 30 年国内数据中心选址布局为例[J].建材与装饰,2018(52):75-76

[8] 胡毅.GMM 估计矩条件的选取方法及其应用[M].北京:经济科学出版社,2016

[9] 陈如波.大型数据中心选址应注意的几个新问题[J].信息通信,2015(09):294-295.

致谢

终于到了敲下“致谢”二字之时，回望这将近两个月的论文撰写时光，不禁感慨万千。

作为缺乏比赛经验的我们，从比赛报名阶段开始就不断遇到麻烦。敲定选题，确定研究思路，下载整理数据……每一个阶段都经历了艰难的探索，现实中遇到的种种困难多得超乎我们的想象。但是，“头发少了”，数据集终于找到了；眼睛痛了，输出的结果终于达到预期了；手腕酸了，论文终于成型了……此中酸辛，唯有亲历者可知。

在此，我们首先想表达对我们的指导老师衷心的感谢，您在我们困顿之时给我们的帮助，我们将永远铭记。其次，我们感谢的是我们的好友汉唐，你慷慨的支持与工作在我们论文的完成上起到了重要的作用。还有，正是有我们各位参赛队员间相互的支持与理解，不断学习，我们的研究才得以顺利推进并按时结题。在这段宝贵的时光里，我们收获颇丰。除此之外，还有众多的老师和同学在我们的研究过程中给予了无私的帮助，在此一并向你们表示感谢！