编号：A0351

我国大数据产业的投入产出效率研究

目录

表格和插图清单

摘要

随着新经济发展战略模式的兴起，经济发展焕发了新活力。传统要素市场已不能全面且准确的分析我国经济的发展状况。近年来，关于经济新动能的研究已成热门话题。顺应时代的发展，5G、大数据、云计算等相关新兴产业迅速兴起，数字信息时代已经到来，人们生活的方方面面早已离不开数据。因此本文将对数据要素的统计测度展开研究。

首先，为了反映我国各地区生产要素对经济的推动作用，我们选择了数据源指数、数据开发成本指数、网络经济指数、大数据产业劳动要素投入指数、 资本要素投入指数作为投入指标；转型升级指数、软件业务收入指数作产出指标；经济活力指数作环境指标。为了考察指标体系中各指标的重要程度，采用层次分析法与熵值法结合对指标赋予综合权重，利用蒙特卡洛综合评价法对各地区大数据产业发展状况进行排序，结果表明东部及中部地区发展较好，西部地区大数据产业整体状况有待提高。

其次，为了分析数据要素对地区经济的影响，我们选取大数据产业发展排名前 15 的地区为决策单元，采用三阶段 DEA 法与 Malmquist 指数法分析含有数据要素与不含数据要素时各地区大数据产业投入产出的效率情况，结果表明发展大数据产业在一定程度上推动了经济的发展，加入数据要素后的模型效率更高，数据要素的加入使投入指标体系更加完善，即表明数据要素与传统生产要素融合，共同推动经济的发展。

关键字：三阶段 DEA 模型、Malmquist 指数模型、蒙特卡洛综合评价、层次分析、熵值法

一、 绪论

(一) 研究背景

党的十九大指出，我国经济已从高速发展阶段转向高质量发展阶段。随着大数据产业的飞速发展，传统的生产要素已不能满足经济发展的新要求，加快构建更加完善的要素市场已成为我国社会热点问题。在当今数字经济迅速发展的时代，互联网已成为人们生活的必需品，与此同时产生了大量数据，企业通过收集数据生产产品与服务，维持自身的发展，人们通过共享数据获得一些显性或隐性服务，数据成为企业竞争的重要战略资源以及社会关注的新焦点。数据作为一种新型生产要素，完善数据要素市场、发挥数据资源的价值、分析数据要素对经济发展的影响是推动经济新动能研究的关键，也是实现经济高质量发展的重要举措。

(二) 研究意义

与传统生产要素相比，探究加入数据要素是否可以推动大数据产业的发展及数据要素推动大数据产业发展的程度，对完善数据要素市场具有深远意义。分析我国部分省市大数据产业发展现状，并针对其存在的问题提出合理的解决方案，对深化社会主义市场化经济体制改革具有举足轻重的作用。

(三) 创新点

1. 构造了新的数据要素指标，包括数据源指数、大数据开发指数，它们与网络经济指数共同构成大数据产业数据生产与开发指标。本文选取计算机设备量、集成电路等指标构成数据源指数，它与数据的产生量正相关。因为数据的开发利用过程离不开专业人员与投入经费，所以选取 R&D 人员专利授权数、R&D 内部经费支出额等构成大数据开发指数。

2. 对加入数据要素后的指标体系赋权，再利用蒙特卡洛综合评价法就大数据产业发展状况对我国 31 个省市进行排序，克服了传统综合评价的绝对性，使评价结果可信度更高。

3. 从不同角度分析地区的大数据产业投入产出效率，利用三阶段 DEA 模型分析 2017-2019 年各年的地区投入产出效率状况，再用 Malmquist 指数法处理 2017-2019 年的面板数据，得到投入产出效率变化。

4. 分别讨论含有数据要素与不含数据要素情况下地区的投入产出效率，加入数据要素时大部分地区的投入产出效率提高，表明数据要素对地区经济发展具有明显的推动作用。

二、 文献综述

(一) 关于数据要素的相关概念

大数据是互联网与计算机产生的数据集合，具有大量、可再生、更新快、通用、即时等特点。耶鲁大学 Viktor Mayer-Schönberger[1]教授在《大数据时代:生活、工作与思维的大变革》中提出，在如今大数据迅速发展的时代，我们可以分析大量的数据，甚至可以处理特定产业的所有数据，不再局限于样本的分析。大数据时代提出者麦肯锡称，数据已经渗透到各行各业，成为极其重要的生产因素。关会娟等[2]学者提出信息技术、数字经济与国民经济不断融合，相互作用，更加高速的推动社会经济发展，数据已成为数字经济时代至关重要的战略性资源。纵观世界的发展，不可忽视的是在当今大数据时代，数据要素已成为与土地、劳动力、资本、技术并列的生产要素。

(二) 关于三阶段 DEA-Malmquist 指数模型

邵明振等[3]提出利用层次分析及熵值法对指标进行综合赋权。Jondrow ，Knox Lovell[4]提出了三阶段 DEA 模型的相关概念，权应杰[5]在其论文中详细的介绍了三阶段 DEA 模型及 Malmquist 指数法的理论方法，罗璐[6]基于 DEA 与 Malmquist 指数对江西省各地区的投入产出效率进行分析。同时叶世绮等[7]学者提出了确定三阶段 DEA 投入与产出指标的相关准则。张发明等[8]用蒙特卡罗洛模拟

方法分析被评价对象的优先排序概率，为决策者提供数据信息。易平涛等[9]分析了随机模拟型综合评价的具体方法步骤，对被评价对象进行相对客观的排序。

三、 相关模型及方法介绍

(一) 蒙特卡洛综合评价

1. 利用极值处理法（具体算法见附录 1）对投入指标数据进行标准化处理，得到标准化后的数据矩阵：

$$X=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}$$

其中 x_{ij} 表示地区 i 关于指标 x_j 经标准化处理后的值。

2. 自主优势量矩阵为：$C=\left[\lambda_{ij}\right]=\left[\mu\alpha_{ij}+\eta\beta_{ij}\right]$

λ_{ij}：被评价对象 d_i 在指标 x_j 上的优势程度；

α_{ij}：被评价对象 d_i 的第 j 项指标与其他 $n-1$ 个被评价对象之间的差异；

β_{ij}：被评价对象 d_i 的第 j 项指标与其他 $m-1$ 个被评价指标整体的优势差异；

μ：竞争性目标偏爱系数；η：发展性目标偏爱系数，μ、$\eta\in[0,1]$，$\mu+\eta=1$，通常取 0.5。

3. 将各地区设为被评价对象，记为 $D=\left(d_1,d_2,\cdots,d_n\right)$，任选 2 个 $\left(d_i,d_j\right)$ 进行分析，由于共有 n 个被评价对象，故需进行 n^2-n 次比较。

4. 设置仿真次数计数变量 count（初始化 count=0）。

5、运用随机发生器产生给定区间上的不确定比值 $r_k(k=2,3,\cdots m)$，r_k 服从均匀分布，可知指标 x_{k-1} 与 x_k 的重要程度之比 $r_k=\frac{w_{k-1}}{w_k}$。当 r_k=1 时，x_{k-1} 与 x_k 同等重要；当 r_k=1.2 时，x_{k-1} 比 x_k 稍微重要；当 r_k=1.4 时，x_{k-1} 比 x_k 明显重要；当 r_k=1.6 时，x_{k-1} 比 x_k 强烈重要；当 r_k=1.8 时，x_{k-1} 比 x_k 极端重要。

6、根据 $w_k=\left(1+\sum_{i=2}^{k}\sum_{j=i}^{k}r_j\right)^{-1}, w_{k-1}=r_k w_k, k=m,m-1,\cdots,2$

确定优势权向量 $w=(w_1,w_2,\cdots,w_m), i\in N$，按照自主优势向量对应排序 $\lambda_i=(\lambda_{i1},\lambda_{i2},\cdots,\lambda_{im})$，$i\in N$。

7、设置计数变量 r_s,r_e,r_f（初始化值均为 0），分别表示 $d_i^{'}>d_i^{*}$、$d_i^{'}=d_i^{*}$、$d_i^{'}<d_i^{*}$ 的次数。

若 $\sum_{j=1}^{m}x_{ij}^{'}w_j^{*}>\sum_{j=1}^{m}x_{ij}^{*}w_j^{*}$，则 $r_s=r_s+1$

若 $\sum_{j=1}^{m}x_{ij}^{'}w_j^{*}=\sum_{j=1}^{m}x_{ij}^{*}w_j^{*}$，则 $r_e=r_e+1$

若 $\sum_{j=1}^{m}x_{ij}^{'}w_j^{*}<\sum_{j=1}^{m}x_{ij}^{*}w_j^{*}$，则 $r_f=r_f+1$

8、count= count+1，若 count=sum（sum 随机仿真总次数，sum 初始值为 0，一般地，方案个数 n 越多，sum 的值应越大），转步骤 9，否则转步骤 2。

9、利用公式

$$s\left(d_i^{'}<d_i^{*}\right)=\frac{r_s+0.5r_e}{sum}$$

优先排序概率 $s\left(d_i^{'}<d_i^{*}\right)$ 的模拟值并保存，通过模拟仿真，得到 n 个被评价对象的优胜度矩阵 S：

$$S=[s_{ij}]=\begin{pmatrix}s_{11}&s_{12}&\cdots&s_{1n}\\s_{21}&s_{22}&\cdots&s_{2n}\\\vdots&\vdots&\ddots&\vdots\\s_{n1}&s_{n2}&\cdots&s_{nn}\end{pmatrix}$$

用 $\left|s_{ij}+s_{ji}-1\right|$ 表示误差值，规定临界值 ε（一般为 0.001），若误差值不高于 ε，说明模拟次数足够，模拟结果可信，否则，需要增加模拟次数。

综合评价计算：

各地区（即被评价对象）的相对综合评价值为 z_i，可通过 n 个地区之间的整体比较函数取最小求得：

$$\begin{cases} \min F(a)=\sum_{i=1}^{n}\sum_{j=1}^{n}\left[z_i-z_j-\left(s_{ij}-s_{ji}\right)\right]^2 \\ s.t.z_1+z_2+\cdots+z_n=c \end{cases} \quad (1)$$

构造 Lagrange 函数，并对其求导，可得：

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\left(s_{ij}-s_{ji}\right)\right]+\frac{1}{n}\sum_{j=1}^{n}\left(s_{ij}-s_{ji}\right)+\frac{c}{n} \quad (2)$$

因为 $\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\left(s_{ij}-s_{ji}\right)\right]=0$，最终 $z_i=\frac{1}{n}\sum_{j=1}^{n}\left(s_{ij}-s_{ji}\right)+\frac{c}{n}$

c：常数，且 c 的选取对 z 中大小顺序无影响，为使 $z_i\geq 0$，一般取 $c=n$

(二) 三阶段 DEA 模型

1. 关于效率的含义

本文研究的大数据产业投入产出效率是数据要素的管理效率，用 DEA 中的技术效率来衡量这一效率值，而技术效率等于纯技术效率乘以规模效率。本文的纯技术效率反映了每个地区利用数据要素的能力，规模效率反映了增加大数据产业规模对增加产出的影响能力。

2. 传统 DEA 模型——CCR 模型与 BBC 模型

相关概念：

决策单元：要评价分析的对象。

生产前沿面：在一定的生产要素和产出价格下，选择出的要素投入和产出的最优组合，是相对效率分析的基础。

$x_j=(x_{1j},x_{2j},...,x_{mj})$，$y_j=(y_{1j},y_{2j},...,y_{sj})$，

$v=(v_1,v_2,...,v_m)$，$u=(u_1,u_2,...,u_s)$

x_{ij}：第 j 个地区在第 i 种投入上的产出量

y_{rj}：第 j 个地区在第 r 种投入上的产出量

v_i ：第 i 种投入的权数

u_r：表示第 r 种产出的权数

其中，$i=1,2,\cdots,m, j=1,2,\cdots,n, r=1,2,\cdots,s$。第 j 个 DMU_j 的效率评价指数为：$h_j=\dfrac{\sum_{r=1}^{s} y_{rj}u_r}{\sum_{i=1}^{m} x_{ij}v_i}$。$h_j$ 越大，表示 DMU_j 的投入产出效率越高。

构造 CCR 模型：

$$\max h_j=\frac{\sum_{r=1}^{s} y_{rj}u_r}{\sum_{i=1}^{m} x_{ij}v_i}$$

$$s.t.\begin{cases}\dfrac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}}, j=1,2,...,n.\\ v_i,u_r\geq 0, r=1,2,...,s, i=1,2,...,m.\end{cases} \quad (3)$$

加入条件 $\sum_{i=1}^{m} x_{ij}v_i=1$，通过对偶理论将模型转化为对偶模型，最后加入非阿基米德无穷小量 ε。设投入松弛变量 s^-，产出松弛变量 s^+，构造基于投入导向的 CCR 模型：

$$min[\theta-\varepsilon(\hat{e}^T s^- + e^T s^+)]$$

$$s.t.\begin{cases}\sum_{j=1}^{n} x_{ij}\lambda_j + s^- = \theta x_{ij}\\ \sum_{j=1}^{n} y_{rj}\lambda_j - s^+ = y_{rj}\\ 其中\ \lambda_j\geq 0, s^-, s^+\geq 0, r=1,2,...s, i=1,2,...,m, j=1,2,...,n\end{cases} \quad (4)$$

θ 为综合技术效率（TE），λ_j 为各决策单元的权重乘数。

加入 $\sum_{j=1}^{n}\lambda_j=1$ 构造基于规模报酬变动的 BBC 模型：

$$min[\theta' - \varepsilon(\hat{e}^T s^- + e^T s^+)]$$

$$s.t.\begin{cases} \sum_{j=1}^{n} x_{ij}\lambda_j + s^- = \theta x_{ij} \\ \sum_{j=1}^{n} y_{rj}\lambda_j - s^+ = y_{rj} \\ \sum_{j=1}^{n} \lambda_j = 1 \\ 其中\ \lambda_j > 0, r = 1,2,...s, i = 1,2,...,m, j = 1,2,...,n \end{cases} \quad (5)$$

用排除了规模报酬变动影响的 BBC 模型可以求得最优的θ'。θ'为纯技术效率（PTE）。规模效率值为 SE=TE/PTE。当 SE=1，该地区的规模效率达到了最优。

3. 三阶段 DEA 模型方法

1. 第一阶段

（具体理论见附录 2）

此阶段要分析各个地区的技术效率值、纯技术效率值、规模效率值，可以由 BBC 模型分析得到。Fried 认为，决策单元的绩效受到管理无效率、环境因素和统计噪声的影响，因此有必要分离这三种影响，从而有了第二阶段的模型。

2. 第二阶段

用 SFA 回归剔除了环境因素和统计因素。

基于成本函数的 SFA 回归：

若现在有 K 个环境变量，我们以投入松弛S_{nk}作为因变量，环境变量Z_k作为自变量，得到 n 个回归方程：

$$S_{nk} = f(Z_k;\alpha_n) + \nu_{nk} + \mu_{nk}; i = 1,2,\cdots,K \quad (6)$$

S_{nk}：第k个决策单元第n项投入的松弛值

Z_k：环境变量；α_n：环境变量的系数

$Z_k\alpha_n$：环境变量对S_{nk}的影响；$\nu_{nk} + \mu_{nk}$：综合干扰项

ν_{nk}：随机干扰，$\nu_{nk} \sim N(0,\sigma_{\nu_n}^2)$

μ_{nk}：管理无效率，$\mu_{nk} \sim N^{+}(0,\sigma_{\mu_n}^{2})$

对S_{nk}进行极大似然估计，可以得到α_n，μ_{nk}，$\sigma_{v_n}^{2}$，$\sigma_{\mu_n}^{2}$。

设$X=(x_1,x_2,\ldots,x_K)$为$K\times N$的投入矩阵，S_{nk}为第 k 个*DMU*在第 n 个投入的松弛量，它的方程为：

$$S_{nk}=x_{nk}-X_n\lambda, n=1,2...,N, k=1,2,...,K \tag{7}$$

X_n：第 n 行向量

x_{nk}：第 k 个*DMU*在第 n 种投入的投入值

$X_n\lambda$为第 k 个*DMU*在第 n 种投入的效率前沿面的最优映射

根据 Jondrow 等论文的思路可以得到：

$$E\left(\mu_{nk}\left|\varepsilon_{nk}\right.\right)=\sigma^{*}\left[\frac{\phi(\lambda\frac{\varepsilon_{nk}}{\sigma})}{\Phi(\frac{\lambda\varepsilon_{nk}}{\sigma})}+\frac{\lambda\varepsilon_{nk}}{\sigma}\right] \tag{8}$$

其中，$\sigma^{*}=\frac{\sigma_\mu\sigma_v}{\sigma},\sigma=\sqrt{\sigma_\mu^{2}+\sigma_v^{2}}$，$\lambda=\sigma_\mu/\sigma_v$，$\varepsilon_{nk}=v_{nk}+\mu_{nk}$

然后计算随机误差项

$$E[v_{nk}\left|v_{nk}+\mu_{nk}\right.]=s_{nk}\text{-}f(z_k;\alpha_n)\text{-}E[u_{nk}\left|v_{nk}+\mu_{nk}\right.] \tag{9}$$

最后用公式

$$X_{nk}^{*}=X_{nk}+[\max(f(Z_k;\hat{\alpha}_n))-f(Z_k;\hat{\alpha}_n)]+[\max(v_{nk})-v_{nk}] \tag{10}$$

其中，$k=1,2,\cdots,K,n=1,2,\cdots,N$

剔除环境因素和随机因素。

其中，X_{nk}^{*}是调整后的投入；X_{nk}是调整前的投入；用$[\max(f(Z_k;\hat{\alpha}_n))-f(Z_k;\hat{\alpha}_n)]$剔除环境因素；$[\max(v_{nk})-v_{nk}]$用于剔除随机因素。

3. 第三阶段

用剔除了环境因素和随机干扰因素的投入量和产出量再次测算每个地区的效率，得到了更加真实的效率值。

(三) Malmquist 指数模型

（详细理论见附录 3）

构造全要素生产率变化指数模型：

$$TFPCH = M_{RD}(x^t, y^t, x^{t+1}, y^{t+1})$$

$$= \frac{D_V^{t+1}(x^{t+1}, y^{t+1})}{D_V^t(x^t, y^t)} \times \left[\frac{D_V^t(x^t, y^t)}{D_V^{t+1}(x^t, y^t)} \times \frac{D_V^t(x^{t+1}, y^{t+1})}{D_V^{t+1}(x^{t+1}, y^{t+1})} \right]^{\frac{1}{2}}$$

$$\times \left[\frac{D_C^t(x^{t+1}, y^{t+1}) / D_V^t(x^{t+1}, y^{t+1})}{D_C^t(x^t, y^t) / D_V^t(x^t, y^t)} \times \frac{D_C^{t+1}(x^{t+1}, y^{t+1}) / D_V^{t+1}(x^{t+1}, y^{t+1})}{D_C^{t+1}(x^t, y^t) / D_V^{t+1}(x^t, y^t)} \right]^{\frac{1}{2}}$$

$$= PC \times TC \times SC(\text{该TC指技术进步变化指数}) \qquad (11)$$

其中，技术效率变化指数$TC=PC\times SC$。

$x_j^t = (x_{1j}^t, x_{2j}^t, ..., x_{mj}^t)^T$ ：第 j 个地区在 t 时期的投入指标值

$y_j^t = (y_{1j}^t, y_{2j}^t, ..., y_{mj}^t)^T$：第 j 个地区在 t 时期的产出指标值，均为正数

$D_C^t(x^t, y^t)$：(x^t, y^t) 在 t 时期的距离函数

$D_C^{t+1}(x^t, y^t)$：在 $t+1$ 时期的距离函数

$D_C^t(x^{t+1}, y^{t+1})$：表示 (x^{t+1}, y^{t+1}) 在 t 时期的距离函数

$D_C^{t+1}(x^{t+1}, y^{t+1})$：表示在 $t+1$ 时期的距离函数

$D_V^t(x^t, y^t)$：表示 (x^t, y^t) 在 t 时期的距离函数

$D_V^t(x^t, y^t)$：表示 (x^t, y^t) 在 $t+1$ 时期的距离函数

$D_V^t(x^{t+1}, y^{t+1})$：表示在 (x^{t+1}, y^{t+1}) 在 t 时期的距离函数

$D_V^{t+1}(x^{t+1}, y^{t+1})$：表示在 $t+1$ 时期的距离函数。

将 Malmquist 指数运用结果得到的各指数记作 M，则 M 表示各时期到下一时期的效率动态变化。若 M>1，表示该时期到下一时期的效率提高；反之亦然；M=1，表示该时期到下一时期的效率不变。

纯技术效率变化（PC）指从 t 期到 $t+1$ 期技术的被利用程度；技术进步变化（TC）指从 t 期到 $t+1$ 期技术能力的变化状况；规模效率变化（SC）指从 t 期到 $t+1$ 期决策单元的投入产出规模与生产最大化时的比例关系。

四、 指标体系

（一）指标体系的构建

1. 指标选取

为了考察我国 31 个省市数据新动能的统计测度，参考国家统计局发布的测算经济发展新动能的统计指标体系以及数字经济相关产业指标体系，根据三阶段 DEA 模型的要求，本文将指标体系分为投入、产出与环境指标三个部分。大数据时代的到来推动了数据要素的发展，因此本文将指标的选取集中在大数据相关指标方向，同时由于数据要素与劳动力、资本等传统生产要素同等重要，本文将大数据产业数据生产与开发指标、大数据产业劳动力投入与资本投入指数作为一级指标。

互联网作为数据要素交易的最重要渠道，通过网络构建平台与用户的数据联系，产生海量数据；电子商务不受时间、地点、空间的限制，成为大多数人购物的第一选择；企业通过分析消费者偏好提供更有针对性的服务，消费者通过分享数据获得服务，不断推动经济的发展，故建立网络经济指数，与数据源指数、数据开发指数共同构成大数据产业数据生产与开发指标。

随着数字经济时代的到来，传统产业正逐渐向新兴产业转型，近年来高新技术产业迅速发展。国家高新区经济总量将占领中国工业增加值及出口创汇的 20%以上。因此高新产业产值、出口额及软件产业收入等数据在较大程度上反映了地区的数据产业发展状况，故设立转型升级指数作产出变量。

经济活力指数反映地区的经济发展状况，而地区本身的经济发展状况可能对分析数据投入推动经济的效率有影响，故设为环境变量。

2. 数据来源及预处理

本文选取了 2017-2019 年各地区关于大数据产业投入产出的指标数据，数据主要来源为《中国统计年鉴》和《中国科技统计年鉴》。其中从《中国统计年鉴》中获取网络经济指数相关数据，如电子商务平台销售额、移动互联网用户数等。从《中国科技统计年鉴》中获取科技前沿产业相关数据，如各地区高技术产业 R&D 经费内部支出、科技企业孵化器数量等数据。

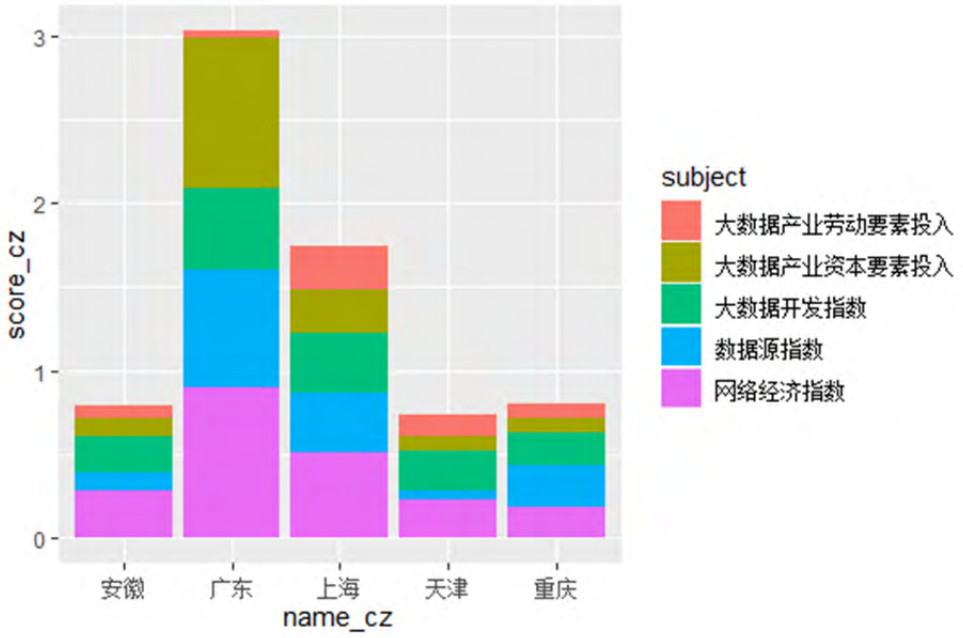
由于数据要素是一个全新的概念，许多数据没有全面的进行统计，导致部分指标缺失值较多，考虑到平均值填充法的局限，本文利用 SPSS 对数据集做多重插补，输出 5 个完整数据集，对其进行描述分析，从中选取标准误差最小的作为最终结果，进行建模分析。

3. 数据描述

本文选取 2019 年广东、上海、重庆、安徽、天津 5 个地区投入指标的数据进行描述性分析，他们分别处于大数据产业发展的不同阶段，广东省处于大数据产业发展的最前端；上海和天津发展位居前列；重庆大数据产业正在迅猛发展；安徽虽然起步晚，但是也奋起直追。

2019 年广东、上海、重庆、安徽、天津 5 个投入指标的堆积条形图如图 1 所示：（score_cz 为投入值）（算法见附录 4）

图 1　地区投入指标的堆积条形图



从图中我们可以看出大数据产业各投入指标的结构。其中，2019 年，广东省的大数据产业资本要素投入和网络经济指数非常高，两者占总投入一半以上，但是广东省在劳动要素投入方面极低，显示出了广东的大数据产业投入结构不平衡。相比之下，上海的劳动投入占总投入的比重相对较高。安徽、重庆、天津在投入规模上不及广东和上海，投入结构相比广东和上海较为平衡。可以发现，经济实力越强的省市，大数据产业规模更大，而投入产出结构重心发生了极度的偏移，生产要素市场化配置越加不合理，本文将详细阐述此类问题。

（二）指标赋权

层次分析法是主观赋权法，反映了决策者的主观判断；熵值法是客观赋权法，反映了数据信息量。本文综合两种方法对指标进行综合赋权，利用拉格朗日乘数法求解 AHP 赋权和熵值法赋权的最优组合权重，得到综合赋权权重。

（层次分析相关结果及算法见附录 5）

综合赋权：

设层次分析得到的权重为w_{1i} ，熵值法得到的权重为w_{2j}，最后得到综合权重

$$w_j = \frac{\left(w_{1j}w_{2j}\right)^{0.5}}{\sum_{j=1}^{m}\left(w_{1j}w_{2j}\right)^{0.5}} \qquad \mathrm{j} = 1,2,\ldots,\mathrm{m}\#(12)$$

最终得到指标体系及权重如表 1，投入指标中，数据源指数中集成电路权重为 0.5582，占比最高，表明集成电路是数据源指数中最重要的指标，在较大程度上反应数据量的多少。在大数据开发指数中，技术市场成交合同金额权重最高为 0.4446，表明技术市场成交合同金额更能反映数据开发程度。网络经济指数中权重为 0.3217 的电子商务平台销售额较大程度的反映了数据的开发与运用程度。大数据产业劳动力投入指数中经济活动人口中研究与开发机构 R&D 人员硕士及以上人数比例权重为 0.5375，体现出 R&D 人员硕士及以上学历人员是大数据产业劳动力的有力来源，有效的推动大数据产业的发展；各地区高技术产业 R&D 经费内部支出权重高达 0.7046，是大数据产业资本的重要来源。

产出指标中单位 GDP 能耗降低率是最能体现转型升级指数的因素。软件产品收入占软件业务收入的大部分，是较为重要的因素。

表 1　大数据产业指标体系及权重

	一级指标	二级指标	单位	AHP 权重	熵值法赋权	综合赋权
投入指标	数据源指数	集成电路	亿块	0.7049	0.3989	0.5582
		微型计算机设备	万台	0.0841	0.2157	0.1418
		移动电话产量	部	0.2109	0.3853	0.3000
	大数据开发指数	每万名 R&D 人员专利授权数	件	0.2062	0.2323	0.2379
		规上企业 R&D 经费支出与 GDP 之比	-	0.1111	0.2453	0.1794
		技术市场成交合同金额	万元	0.6186	0.2706	0.4446
		R&D 内部经费支出额	万元	0.0641	0.2518	0.1381
	网络经济指数	单位互联网宽带接入用户数	万户	0.0492	0.1537	0.0911
		移动互联网用户数	万户	0.2057	0.1067	0.1552
		移动互联网接入流量	万 G	0.2057	0.1077	0.1560
		电子商务平台销售额	亿元	0.3422	0.2756	0.3217

		有电子商务交易活动企业比重	%	0.0753	0.0988	0.0904
		网上零售总额占社会消费品零售总额的比重	-	0.1219	0.2574	0.1856
	大数据产业劳动要素投入	经济活动人口中研究与开发机构 R&D 人员硕士及以上人数比例	%	0.6491	0.4030	0.5375
		R&D 人员总数	人	0.0719	0.2766	0.1482
		每万名就业人员 R&D 人员折合全时当量	人年	0.2790	0.3204	0.3142
	大数据产业资本要素投入	各地区高等学校经费支出总额	万元	0.2500	0.3453	0.2954
		各地区高技术产业 R&D 经费内部支出	万元	0.7500	0.6547	0.7046
产出指标	转型升级指数	国家高新区单位企业总产值	万元	0.1047	0.2587	0.1750
		国家级高新区出口额占出口总额的比重	-	0.2583	0.4384	0.3578
		单位 GDP 能耗降低率	%	0.6370	0.3029	0.4671
	软件业务收入	软件产品收入	万元	0.7500	0.4935	0.6310
		软件技术服务收入	万元	0.2500	0.5065	0.3690
环境指标	经济活力指数	科技企业孵化器数量	个	-	-	-
		国家高新技术开发区企业单位数	个	-	-	-
		快递业务量	万件	-	-	-

五、 地区数据新动能统计测度的实证分析

(一) 蒙特卡洛综合评价实例分析

前文已经利用层次分析法及熵值法求得综合权重，在此基础上，利用蒙特卡洛综合评价法根据投入指标数据对我国各省市进行相对评价，以得到各地区的大数据产业相对发展水平。（具体算法及结果见附录 6）

结果表示，综合评价值最高的是广东，其次是北京；最低的是贵州。广东、北京等地大数据产业发展较好，贵州、宁夏等地资源匮乏、人才短缺，故大数据产业发展较为缓慢。

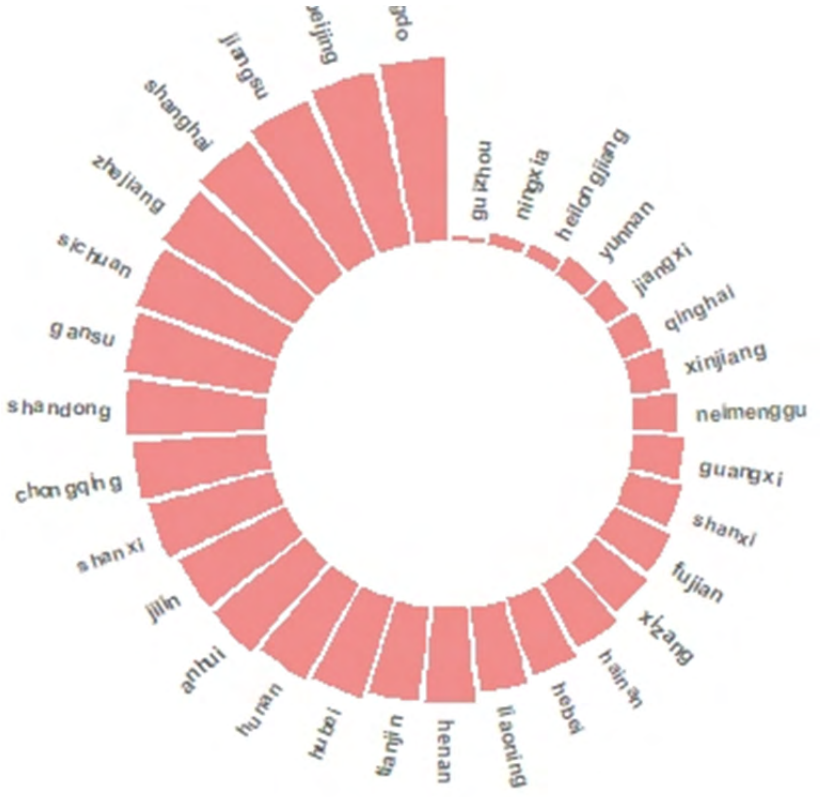


图 2 31 个省市蒙特卡洛相对综合评价值环形条形图

(二) 三阶段 DEA 结果分析

本文对大数据产业发展排名前 15 的地区进行三阶段 DEA 及 Malmquist 分析。

1. 第一阶段

DEA 第一阶段效率测度分析

①2017 年第一阶段测度结果如下表所示：

表 2　2017 年第一阶段测度结果

地区	技术效率	纯技术效率	规模效率	
广东	1	1	1	-
北京	1	1	1	-
江苏	1	1	1	-
上海	0.945	0.952	0.993	drs
浙江	1	1	1	-
四川	1	1	1	-
甘肃	1	1	1	-
山东	1	1	1	-
重庆	1	1	1	-
陕西	1	1	1	-
吉林	1	1	1	-
安徽	0.698	0.861	0.810	irs
湖南	1	1	1	-
湖北	1	1	1	-
天津	0.670	1	0.670	irs
均值	0.954	0.988	0.965	

其中 irs 为规模报酬递增，drs 规模报酬递减，-为规模报酬不变。

由表 2 所知，2017 年所选取 15 个地区的数据要素的投入产出效率的平均值为 0.954， 说明了这 15 个地区整体的技术效率值较大；纯技术效率和规模效率的平均值分别为 0.988 和 0.965，说明这 15 个地区的大数据产业的投入产出结构和数据要素的市场化配置发挥出了比较好的经济效益。在样本选取的 15 个地区中，只有 3 个地区综合技术效率值小于 1，分别是上海、安徽、天津，同时这 3 个地区的规模效率也小于 1；在样本选取地区中，上海的规模报酬递减，安徽、天津的规模报酬递增，其余地区的规模报酬不变。

②2018 年第一阶段测度结果如下表所示：

表 3　2018 年第一阶段测度结果

地区	技术效率	纯技术效率	规模效率	
广东	1	1	1	-
北京	1	1	1	-
江苏	1	1	1	-
上海	0.908	0.934	0.973	irs

浙江	1	1	1	-
四川	1	1	1	-
甘肃	1	1	1	-
山东	1	1	1	-
重庆	0.894	1	0.894	irs
陕西	1	1	1	-
吉林	1	1	1	-
安徽	1	1	1	-
湖南	1	1	1	-
湖北	1	1	1	-
天津	0.742	1	0.742	irs
均值	0.970	0.996	0.974	

由表 3 所知 2018 年 15 个地区的数据要素投入产出效率的平均值为 0.970，说明了这 15 个地区整体的技术效率值较大；纯技术效率的均值为 0.996，说明这 15 个地区的大数据产业的投入产出结构具有协调性；15 个地区整体的规模效率均值等于 0.974，说明这 15 个地区的数据要素的市场化配置对大数据产业的产出有非常积极的影响。在这 15 个地区中，只有 3 个地区综合技术效率值小于 1，分别是上海，重庆，天津，这些地区处于非有效状态，其他地区都处于有效的生产前沿面，同时，这三个地区的规模报酬递增；在样本选取地区中，只有上海的纯技术效率小于 1。

③2019 年第一阶段测度结果如下表所示：

表 4　2019 年第一阶段测度结果

地区	技术效率	纯技术效率	规模效率	
广东	0.951	1	0.951	drs
北京	1	1	1	-
江苏	1	1	1	-
上海	0.859	0.867	0.991	irs
浙江	1	1	1	-
四川	1	1	1	-
甘肃	0.889	1	0.889	irs
山东	1	1	1	-
重庆	0.688	1	0.688	irs
陕西	1	1	1	-
吉林	1	1	1	-

安徽	0.789	1	0.789	irs
湖南	1	1	1	-
湖北	1	1	1	-
天津	0.703	1	0.703	irs
mean	0.925	0.991	0.934	

由表 4 所知 2019 年整体的综合技术效率、纯技术效率和规模效率都处于一个较高的水平，结合 2017 年和 2018 年的分析结果，可以看出这 15 个地区的大数据产业在发展，但各地区发展不均衡，使地区差异较显著。但是从地区来看，在 15 个地区中，2019 年有 6 个地区综合技术效率值小于 1，分别是广东，上海，甘肃，重庆，安徽，天津，这些地区都处于非有效状态。其中，在这 15 个地区中，只有上海的纯技术效率小于 1；广东、甘肃、重庆、安徽、天津的技术效率小于 1，其归结于规模效率小于 1，表明了在维持现有管理水平和技术水平下，这些地区应该合理调整大数据产业规模。另外，广东的规模报酬递减，上海、甘肃、重庆、安徽、天津的规模报酬递增，其余地区的规模报酬不变。

2. 第二阶段

由于高新区企业单位数与科技企业孵化器数量是直接影响地区经济发展的指标，而快递业务量是间接因素，所以本文主要分析前两者。

绘制 2017-2019 年高新区企业单位数与科技企业孵化器数量环形条形图。其中红色、蓝色、绿色分别表示 2017、2018、2019 年数据。 高新区企业单位数 2017 年与 2018 年最高的均为北京，最低为甘肃；2019 年最高为广东，最低为重庆。甘肃人才匮乏且重工业，故其经济发展较为缓慢。

科技企业孵化器数量三年来最高的均为广东，最低的是重庆。重庆大数据产业发展起步较晚，且为直辖市，产业规模无法与其他省匹敌。

第二阶段用SFA回归剔除环

图 3　2017—2019 年国家高新区企业单位数环形条形图

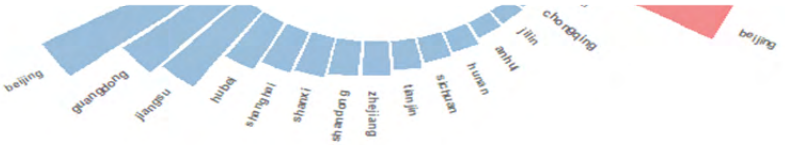
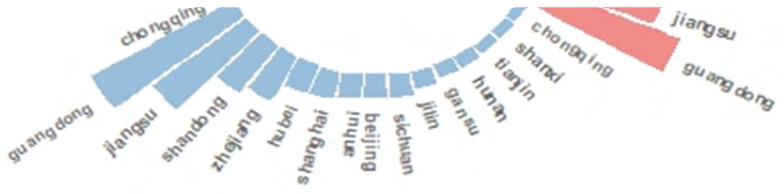


图 4　2017—2019 年科技企业孵化器数量环形条形图



境因素和随机因素。本文分别对每年每个投入松弛变量进行回归分析，得到了以下结果（具体回归过程及结果见附录 9）。

2017 年回归分析结果:

表 5　2017 年回归分析结果

参数	数据源指数松弛变量	大数据开发指数松弛变量	网络经济指数松弛变量	大数据产业劳动要素投入松弛变量	大数据产业资本要素投入松弛变量
α_0	-0.0287	0	0	-0.0663	0.0052
α_1	-0.0463	0.0280	-0.0230	0.0233	-0.0185
α_2	-0.0359	-0.3117	0.1007	0.0490	0.0170
α_3	0.0606	0.3202	0.0142	0.0034	-0.0037
$\sigma^2=\sigma_u^2+\sigma_v^2$	0.0065	0.0038	0.0003	0.0058	0.0007
γ	1.0000	0.0001	0.0000	1.0000	0.9500
LR 似然比	12.0033	——	——	10.1134	5.8385

其中，α_1：环境因素科技企业孵化器数量的系数

α_2：国家高新区数量的系数

α_3:快递业务量的系数

由于投入松弛 2 和 3 没有通过广义似然比检验，故本文采用多元线性回归的方法确定了参数值，上图的投入松弛变量 2 和 3 的参数值$\alpha_1,\alpha_2,\alpha_3$为 OLS 估计的结果。（具体算法及结果见附录 9）

当$\frac{\sigma_{\mu_n}^2}{\sigma_{v_n}^2+\sigma_{\mu_n}^2}$趋近于 1，说明管理因素是影响效率的主要因素，若趋近于 0，则认为随机因素是影响效率的主要因素。

2018 年回归分析结果：

表 6　2018 年回归分析结果

参数	数据源指数松弛变量	大数据开发指数松弛变量	网络经济指数松弛变量	大数据产业劳动要素投入松弛变量	大数据产业资本要素投入松弛变量

α_0	-0.0005	-0.0005	-0.0072	-0.0017	-0.0035
α_1	-0.0023	0.0017	-0.0110	-0.0012	-0.0055
α_2	-0.0017	-0.0006	0.0014	-0.0146	0.0007
α_3	-0.0055	-0.0017	0.0096	-0.0014	0.0049
$\sigma^2=\sigma_u^2+\sigma_v^2$	0.0018	0.0000	0.0002	0.0018	0.0000
γ	1.0000	1.0000	0.9300	1.0000	0.9300
LR 似然比	18.0485	17.3038	6.6096	16.5173	6.6268

2019 年回归分析结果：

表 7　2019 年回归分析结果

参数	数据源指数松弛变量	大数据开发指数松弛变量	网络经济指数松弛变量	大数据产业劳动要素投入松弛变量	大数据产业资本要素投入松弛变量
α_0	-0.0020	-0.0111	-0.0153	-0.0042	-0.0110
α_1	0.0151	-0.0120	-0.0162	-0.0012	-0.0124
α_2	-0.0151	0.0060	0.0068	-0.0138	0.0056
α_3	-0.0205	0.0058	0.0085	0.0093	0.0060
$\sigma^2=\sigma_u^2+\sigma_v^2$	0.0031	0.0003	0.0007	0.0023	0.0004
γ	1.0000	0.9300	0.9300	1.0000	0.9300
LR 似然比	17.2250	6.7492	6.7542	16.3689	6.7522

由表 6、表 7 的结果可知，2018-2019 年投入松弛变量均通过 LR 似然比检验。

3. 第三阶段

第三阶段我们用调整后了的投入产出（具体数据见附录 10）再次建立基于投入导向的 BBC 模型，并用软件 DEAP2.1 输出了 2017 到 2019 年的技术效率值、纯技术效率值和规模效率值。

表 8　2017 年第三阶段测度结果

地区	技术效率	纯技术效率	规模效率	e
广东	1	1	1	-
北京	1	1	1	-
江苏	1	1	1	-
上海	0.92	0.995	0.925	irs
浙江	1	1	1	-
四川	1	1	1	-
甘肃	1	1	1	-
山东	1	1	1	-
重庆	0.896	1	0.896	irs
陕西	1	1	1	-
吉林	1	1	1	-
安徽	0.775	0.997	0.778	irs
湖南	1	1	1	-
湖北	0.994	1	0.994	irs
天津	0.479	1	0.479	irs
均值	0.938	0.999	0.938	

从表 8 可以看出，2017 年，重庆在排除了环境因素和随机因素以后处于非有效状态，综合技术效率与之前相比较低，从侧面反映出国家级高新区对重庆的大数据产业有积极的影响。另外，从表可以看出重庆的综合技术效率值小于 1 主要是规模效率值小于 1 导致的，这说明重庆如果要促进大数据产业的发展，应该调整大数据产业规模。除此之外，相比没有排除环境因素和随机因素的情况，上海、湖北的综合技术效率小于 1，其主要是受规模效率的影响。天津的综合技术效率偏低，主要受规模效率偏低的影响。

（2018-2019 年结果见附录 11）由 2018 年的表可知，与 2017 年相比，上海的综合技术效率下降，仍然处于非有效状态，规模效率偏低仍然是导致其处于非有效状态的原因。相比加入了环境变量的分析结果，安徽的综合技术效率小于 1，处于非有效状态，其主要是受产业规模的影响。天津在去除环境变量以后，综合技术效率进一步下降。除此之外，上海、重庆、安徽、天津的规模报酬递增，其他地区的规模报酬不变。

2019 年的输出结果表明上海、甘肃、重庆、安徽、天津为非有效单元，但纯技术效率都处于有效状态，并且规模报酬增加。其中，重庆的综合技术效率值最低，其次是天津。

(三) Malmquist 指数结果分析

未加入数据指标的 Malmquist 指数：

表 9 未加入数据指标的 Malmuist 指数

地区	技术效率变化	技术进步变化	纯技术效率变化	规模效率变化	全要素生产率变化
广东	1	0.851	1	1	0.851
北京	1	1.012	1	1	1.012
江苏	0.942	1.063	1	0.942	1.001
上海	0.938	1.103	0.937	1.002	1.035
浙江	1	0.953	1	1	0.953
四川	0.962	1.092	0.987	0.975	1.05
甘肃	0.811	1.314	1	0.811	1.065
山东	1	1.104	1	1	1.104
重庆	0.815	1.084	0.943	0.864	0.884
陕西	0.980	1.090	1	0.980	1.068
吉林	1	1.171	1	1	1.171
安徽	1.026	1.139	1.061	0.968	1.169
湖南	1	1.162	1	1	1.162
湖北	1.028	1.147	1.146	0.897	1.178
天津	1.193	1.103	1.134	1.051	1.315
均值	0.976	1.088	1.012	0.964	1.061

由表 9 得到如下信息。从技术效率变化来看，技术效率变化大于等于 1 的个数有 9 个地区，说明这些地区的技术效率有进步。其中，天津的技术效率增加得最快。技术效率变化小于 1 的个数有 6 个，分别是江苏、上海、四川、甘肃、重庆、陕西，这些地区的技术效率在 2017-2019 年期间有所退步。从技术进步变化来看，除了广东、浙江以外，其它地区技术进步变化都大于等于 1。由表可以看出，广东、北京、浙江、山东、吉林、湖南等地区的全要素生产率变化小于 1 主要是受技术退步的影响，因此，这些地区要关注技术进步的影响，而上海、四川、重庆、安徽、湖北、天津是受技术进步、纯技术效率变化、规模效率变化的三重影响。进一步可以看到，全要素生产率变化指数最大的是天津，最低的是广东。

加入了数据指标的 Malmquist 指数：

表 10　加入了数据指标的 Malmquist 指数

地区	技术效率变化	技术进步变化	纯技术效率变化	规模效率变化	全要素生产率变化
广东	1	0.848	1	1	0.848
北京	1	0.98	1	1	0.98
江苏	1	1.004	1	1	1.004
上海	1.004	1.015	1.003	1.001	1.019
浙江	1	0.926	1	1	0.926
四川	1	1.081	1	1	1.081
甘肃	0.811	1.315	1	0.811	1.067
山东	1	1.213	1	1	1.213
重庆	0.817	1.073	1	0.817	0.876
陕西	1	1.185	1	1	1.185
吉林	1	1.162	1	1	1.162
安徽	0.996	1.073	1.002	0.994	1.068
湖南	1	1.479	1	1	1.479
湖北	1.003	1.306	1	1.003	1.31
天津	1.155	1.059	1	1.155	1.223
均值	0.982	1.104	1	0.982	1.084

由表 10 可以看到：甘肃、重庆、安徽的技术效率变化<1，其中甘肃最低，说明甘肃的技术效率退步明显。从技术进步变化来看，广东、北京、浙江，技术发生退步，这直接导致了它们的全要素生产率变化<1。重庆的技术进步变化>1，但是全要素生产率变化<1，是因为受到了规模效率变化<1的主要影响。进一步可以看到，全要素生产率变化指数最大的是湖南，最低的是广东。

上图为不含数据指标和含数据指标的 Malmquist 全要素生产率变化指数图，

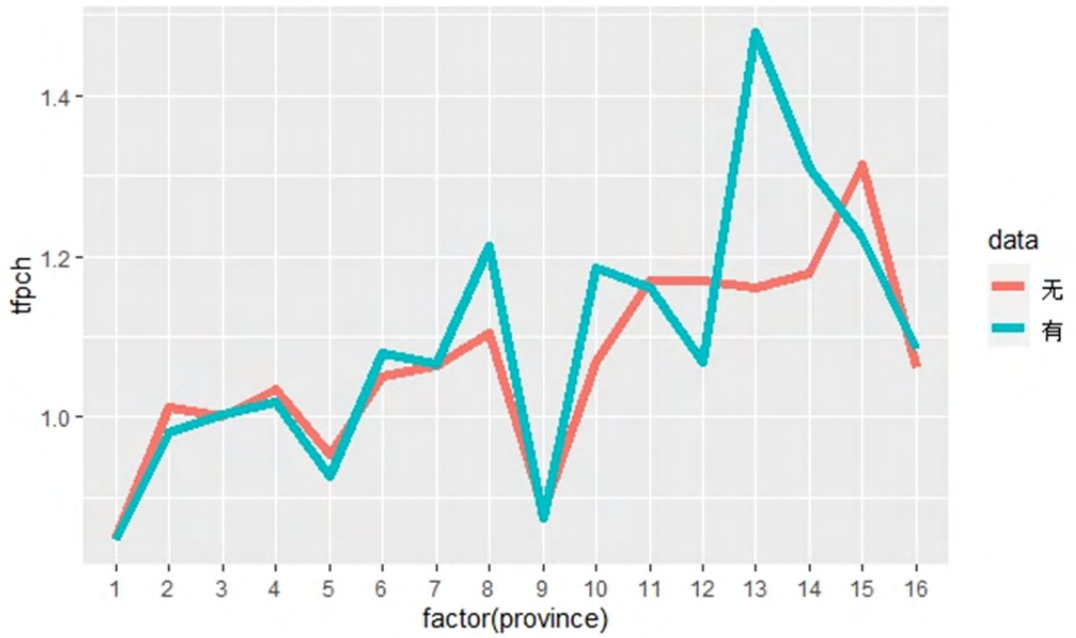


图 5　有无数据指标的全要素生产率变化指数折线图

由图 5　有无数据指标的全要素生产率变化指数折线图可知，除个别省市外，绝大多数省市加入了数据指标比没有加入数据指标的全要素生产率变化指数要高，说明了数据要素对效率的提高有较为明显的促进作用。

六、　结论与建议

(一)　结论

本文首先用蒙特卡洛综合评价法筛选出了大数据产业发展较好的 15 个地区，再用 DEA 第一阶段对 2017 年至 2019 年的 15 个地区的大数据产业的投入产出效率进行分析，得到了这 15 个地区的初步效率结果。我们分析出可能的环境因素

对效率的影响，得到了三个环境因素（科技孵化器数量、国家级高新区企业单位数、快递业务量），然后用 SFA 回归的理论对可能的环境因素进行分析，得出了各松弛变量对于各环境因素的回归方程，并得出了环境因素对投入松弛的影响情况。DEA 第二阶段排除了环境因素和随机因素的影响，然后再对调整后的数据进行 DEA 分析和全要素效率的分析。

1. 从蒙特卡洛相对综合评价值来看，我国各地区大数据产业发展状况存在较大差异，东部及中部地区发展现状较好，反观西部地区，西部地区深居内陆，较为偏僻的地理位置以及人才、资源的稀缺导致其发展速度不敌中部及东部地区。

2. 从 DEA 第一阶段的结果来看，2017 年至 2019 年所选的 15 个地区中，绝大多数地区大数据产业的投入产出效率处于最优化状态。但是天津、安徽等地的投入产出效率相对较低，其原因是规模效率较低，说明大量的对大数据产业的投入没有带来产出的有效增长。从 2017 年到 2019 年，重庆的投入产出效率逐年降低，降低的原因也是由于产业规模的增加没有带来有效的产出，新兴信息技术领域起步相对较晚，前端企业匮乏且产业生态尚未形成，可能对其大数据产业发展有所限制。

3. 从 DEA 第二阶段的结果来看，通过 SFA 回归剔除了环境因素和随机因素，排除了各地区的经济环境对大数据产业投入产出效率的影响，提高了效率的精确度。从 SFA 回归结果可以看出，科技企业孵化器数量、国家高新区的数量、快递业务量对五个投入松弛有显著的影响。从 2017 年的结果可以看到，国家高新区的数量对大数据产业劳动力、网络经济和大数据产业资本的投入松弛有正向作用，初步说明国家高新区的增加不是提高大数据产业投入产出效率的有效途径。从 2018 和 2019 年的结果来看，进一步说明增设国家高新区不是提高效率的最有效途径。同理，分析科技孵化器对各投入松弛变量的作用，我们可以看到，增加科技孵化器是提高效率的有效手段。

4. 从DEA第三阶段的结果来看，在2017到2019年期间，重庆、天津、安徽等地的投入产出效率较低，其很大程度是由规模效率较低导致的。其中，天津的效率最低。说明这些地区更应该重视大数据产业各投入产出的比例，否则，增加投入却没有得到有效的回报。

5. 从全要素生产率来看，未加入数据指标（网络经济指数、数据源指数、大数据开发指数）时，天津的全要素生产率变化最大，并且呈增长趋势，有比较大的潜力，而相比之下，广东的全要素生产率呈下降趋势；加入数据指标时，湖南的全要素生产率变化最大，有较好的发展前景，而广东最低。综合分析没有加入数据指标和加入了数据指标的全要素生产率变化，加入了数据指标的全要素生产率变化更大，说明数据要素相关变量对效率的提高有明显的促进作用，这正好体现数据要素对经济发展的积极影响。

(二) 建议

1. 缩减地区之间的发展不平衡。从区域层面，还是受经济发展的影响比较大，尽管有一些方面的突破，但是整体来说经济不发达的地区，它的大数据开发能力还是比较弱的，这主要集中在西部地区和少数几个东部地区，比如安徽、天津等地。

2. 加快数据资源产业化进程，调整数据要素与其他要素的市场配置比例。数据相对其他要素具有不可损耗性，怎么释放数据的巨大价值，是让数据要素成为经济新动能的关键。另外，要重视数据要素的投入产出结构及比例，让数据要素对其他要素发挥出乘数效应，是释放数据资源价值的关键。比如重庆、天津等地，大数据产业的投入产出的比例不相适应，导致增加数据要素的投入，没有得到有效的回报。

3. 加强科技创新，加快大数据技术的发展。科技创新既是国家战略要求，也是大数据产业适应时代发展的新要求。从数据的开发、存储和分析，再到数据

的交易和应用，整个过程都需要强大的科技实力作为支撑。因此，要重视科技企业孵化器、创新创业孵化园的运营，重视它们的数量和质量。比如在山东、江苏等地，科技孵化器数量远超全国平均水平，科技企业孵化器对它们的大数据产业的积极作用也比较明显。

4. 加快数据资源市场化，提升数据应用能力。数据要素要想以市场化方式进行合理配置，必须重视数据资源的市场化。另外，提升数据应用能力，是实现数据交易的关键。

参考文献

[1]Viktor Mayer-Schönberger.大数据时代:生活、工作与思维的大变革[M].周涛译版,浙江人民出版社,2013,1-261

[2]关会娟,许宪春,张美慧,等.中国数字经济产业统计分类问题研究,[J].统计研究,2020,37（12):4-16

[3]邵明振,马舒瑞,屈小芳,等.河南省经济新动能统计测度、经济效应及发展路径.[J].统计理论与实践 2021(03):9-14

[4]Jondrow , Knox Lovell , Materov , Schmidt. On the estimation of technical inefficiency in the stochastic frontier production function model[J].Economics, model[J]. Economics, 1982,19(2):233-2

[5]权杰庆.研发投入产出效率的国际比较研究——基于三阶段 DEA 模型分析,[D].山西财经大学, 2016

[6]罗璐.基于 DEA-Malmquist 指数的江西省城镇化效率研究. [D].江西财经大学,2015

[7]叶世绮,颜彩萍,莫剑芳.确定 DEA 指标体系的 B -D 方法. [J].暨南大学学报（自然科学版）2004,25(3):249-255

[8]张发明,郭亚军,易平涛. 一种基于蒙特卡罗模拟的群体协商评价方法及其应用.[J].运筹与管理,2010,19(2):63-67

[9]易平涛,李伟伟,郭亚军. 随机模拟型综合评价模式及其求解算法.[J].运筹与管理,2014,23(6):222-228

[10]罗登跃.三阶段 DEA 模型管理无效率估计注记[J].统计研究,2012(04),104-107

[11]葛文婷,戚戬,徐豪威.基于 DEA-Malmquist 模型的中部省域数字经济效率测算.[J].科技和产业,2020,20（9）:68-110

[12]陈诗一.刘文杰.为什么要素市场化配置对经济高质量发展如此重要？.财经问题研究. https://kns.cnki.net/kcms/detail/21.1096.f.20210414.1012.002.html

[13]何玉长，王伟.数据要素市场化的理论阐释.[J].当代经济研究,2021,(4):33-44

[14]任毅,丁黄艳. 我国不同所有制工业企业经济效率的比较研究———基于规模效率、管理水平和技术创新视角. [J].产业经济研究,2014,.68(1):103-110

[15]吴文江,单永华.对数据包络分析中被评价的决策单元的探讨.[J]武汉工业大学学报,1998.20(1)92-95

[16]程云洁,辛大国. 基于 DEA 模型的金砖国家生态效率研究. [J].湖南理工学院学报(自然科学版),2021,34(2):70-76

[17]易继承,张璐.基于三阶段 DEA 模型的创新型国家创新效率测度.[J].统计与决策,2021,(8):81-85

附录

附录 1：极值处理法

采用极值处理法对 2017-2019 年 15 个地区所有指标数据进行标准化处理，代码如下（Python）：

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import random
from sklearn import preprocessing
data = pd.read_csv("D:\\123.csv", sep=',',header=None)
print(data,'\n')
data_zs = pd.DataFrame(preprocessing.MinMaxScaler().fit_transform(data))
data_zs.to_csv("D:\\456.csv",sep=',',header=None)
print(data_zs)
```

附录 2：DEA 具体理论

$x_j=(x_{1j},x_{2j},...,x_{mj})$，$y_j=(y_{1j},y_{2j},...,y_{sj})$，

$v=(v_1,v_2,...,v_m)$，$u=(u_1,u_2,...,u_s)$

x_{ij}：第 j 个地区在第 i 种投入上的投入量

y_{rj}：第 j 个地区在第 r 种产出上的产出量

v_i：第 i 种投入的权数

u_r：表示第 r 种产出的权数

其中，$i=1,2,\cdots,m,\ j=1,2,\cdots,n,r=1,2,\cdots,s$。第 j 个 DMU_j 的效率评价指数为：$h_j=\dfrac{\sum_{r=1}^{s}y_{rj}u_r}{\sum_{i=1}^{m}x_{ij}v_i}$。$h_j$ 越大，表示 DMU_j 的投入产出效率越高。

接下来求第 j 个地区的效率评价指数最大值。构造 CCR 模型：

$$\max h_j = \frac{\sum_{r=1}^{s} y_{rj} u_r}{\sum_{i=1}^{m} x_{ij} v_i}$$

$$s.t. \begin{cases} \dfrac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}}, j = 1,2,...,n. \\ v_i, u_r \geq 0, r = 1,2,...,s, i = 1,2,...,m. \end{cases}$$

h_j=1 时，第 j 个地区处于有效状态，达到了生产前沿面；h_j<1 时，第 j 个地区处于非有效状态。根据 Charnes-Cooper 转换，把条件$\sum_{i=1}^{m} x_{ij} v_i = 1$加入到模型中，从而把问题转换成了线性规划问题，再通过对偶理论把模型转化为对偶模型。

最后向模型加入非阿基米德无穷小量 ε，设投入松弛变量s^-，产出松弛变量s^+，从而得到了基于投入导向的 CCR 模型：

$$min[\theta - \varepsilon(\hat{e}^T s^- + e^T s^+)]$$

$$s.t. \begin{cases} \sum_{j=1}^{n} x_{ij} \lambda_j + s^- = \theta x_{ij} \\ \sum_{j=1}^{n} y_{rj} \lambda_j - s^+ = y_{rj} \\ 其中 \lambda_j \geq 0, s^-, s^+ \geq 0, r = 1,2,...s, i = 1,2,...,m, j = 1,2,...,n \end{cases}$$

θ为综合技术效率（TE），λ_j为各决策单元的权重乘数。

接下来是进一步建立规模报酬变动下的 BBC 模型。在 CCR 中加入$\sum_{j=1}^{n} \lambda_j = 1$得到基于规模报酬变动的 BBC 模型：

$$min[\theta' - \varepsilon(\hat{e}^T s^- + e^T s^+)]$$

$$s.t.\begin{cases} \sum_{j=1}^{n} x_{ij}\lambda_j + s^- = \theta x_{ij} \\ \sum_{j=1}^{n} y_{rj}\lambda_j - s^+ = y_{rj} \\ \sum_{j=1}^{n} \lambda_j = 1 \\ 其中\ \lambda_j > 0, r = 1,2,...s, i = 1,2,...,m, j = 1,2,...,n \end{cases}$$

用排除了规模报酬变动影响的 BBC 模型可以求得最优的θ'。θ'为纯技术效率（PTE）。规模效率值为 SE=TE/PTE。当 SE=1，该地区的规模效率达到了最优。

用 SFA 回归剔除了环境的因素和统计因素。

①关于 SFA 回归：

定义生产函数$f(x)$为在给定x下的最大产出。定义产商 i 的产出为：$y_i = f(x_i,\alpha)\xi_i$，其中，α为待估参数；ξ_i为产商i的水平，满足$0 < \xi_i \le 1$。

但产出还会受到随机因素的直接冲击，故又有加入了随机因素e^{v_i}（>0）的生产函数：$y_i = f(x_i,\alpha)\xi_i e^{v_i}$.

设$f(x_i,\alpha) = e^{\alpha_o} x_{1i}^{\alpha_1} \cdots x_{ki}^{\alpha_k}$，则通过对$f(x_i,\alpha)$取对数得到生产函数的对数形式：$\ln y_i = \alpha_0 + \sum_{k=1}^{N} \alpha_k \ln x_{ki} + \ln \xi_i + v_i$。

由于$0 < \xi_i \le 1$，故$\ln \xi_i \le 0$。令$u_i = -\ln \xi_i \ge 0$，则：

$$\ln y_i = \alpha_0 + \sum_{k=1}^{K} \alpha_k \ln x_{ki} + v_i - u_i$$

$u_i \ge 0$：“无效率”项，说明生产商i与效率前沿面的距离。

$\varepsilon_i = v_i - \mu_i$：混合误差项，由于$\varepsilon_i$分布不对称，不能使用 OLS 估计无效率项$u_i$。若要估计无效率项$u_i$，须对$v_i, \mu_i$的概率分布作出假设，然后进行更有效率的最大似然估计。

本文无效率项的分布设为半正态分布。

随机前沿模型还可以估计成本，跟生产函数类似，可以得到以下加入了随机因素的 n 个成本函数：

$$\ln c_i=\alpha_0+\alpha_y \ln y_i+\sum_{k=1}^{N}\alpha_k \ln P_{ki}+\nu_i+u_i$$

c_i：产商 i 的成本；P_{ki}：要素 N 的价格；

y_i：产出；ν_i：成本函数的随机冲击；u_i：无效率项。

对于成本函数，$u_i=0$ 说明生产商实现了最低成本的效率前沿；反之，如果 $u_i>0$，则生产商需要支出更高的成本。

本文是基于成本函数做的 SFA 回归。

②基于成本函数的 SFA 回归

若现在有 K 个环境变量，我们以投入松弛 S_{nk} 作为因变量，环境变量 Z_k 作为自变量，得到 n 个回归方程：

$$S_{nk}=f(Z_k;\alpha_n)+\nu_{nk}+\mu_{nk};i=1,2,\cdots,K$$

S_{nk}：第 k 个地区第 n 项投入的松弛值；

Z_k：环境变量；α_n：环境变量的系数；

$Z_k\alpha_n$：环境变量对 S_{nk} 的影响；$\nu_{nk}+\mu_{nk}$：混合误差项；

$\nu_{nk}\sim N(0,\sigma_{\nu_n}^{\ 2})$：随机误差项，它表示随机干扰因素对投入松弛变量的影响；

μ_{nk}：管理无效率，它表示管理因素对投入松弛变量的影响，$\mu_{nk}\sim N^{+}(0,\sigma_{\mu_n}^{\ 2})$

对 S_{nk} 进行极大似然估计，可以得到 α_n，μ_{nk}，$\sigma_{\nu_n}^{\ 2}$，$\sigma_{\mu_n}^{\ 2}$。当 $\frac{\sigma_{\mu_n}^{\ 2}}{\sigma_{\nu_n}^{\ 2}+\sigma_{\mu_n}^{\ 2}}$ 趋近于 1，说明管理因素是影响效率的主要因素，若趋近于 0，则认为随机因素是影响效率的主要因素。

设 $X=(x_1,x_2,\ldots,x_K)$ 为 $K\times N$ 的投入矩阵，S_{nk} 为第 k 个 DMU 在第 n 个投入的松弛量，它的方程为：

$$S_{nk}=x_{nk}-X_n\lambda,n=1,2...,N,k=1,2,...,K$$

X_n：第 n 行

x_{nk}：第 k 个 DMU 在第 n 个投入的投入值

$X_n\lambda$：为第 k 个 DMU 在第 n 种投入的效率前沿面的最优映射。

根据 Jondrow 等论文的思路可以得到：

$$E\left(\mu_{nk}\left|\varepsilon_{nk}\right.\right)=\sigma^*\left[\frac{\phi(\lambda\frac{\varepsilon_{nk}}{\sigma})}{\Phi(\frac{\lambda\varepsilon_{nk}}{\sigma})}+\frac{\lambda\varepsilon_{nk}}{\sigma}\right]$$

其中，$\sigma^*=\frac{\sigma_\mu\sigma_v}{\sigma},\sigma=\sqrt{\sigma_\mu{}^2+\sigma_v{}^2}$，$\lambda=\sigma_\mu/\sigma_v$，$\varepsilon_{nk}=v_{nk}+\mu_{nk}$

然后计算随机误差项 $E[v_{nk}\left|v_{nk}+\mu_{nk}\right.]=s_{nk}\text{-}f(z_k;\alpha_n)\text{-}E[u_{nk}\left|v_{nk}+\mu_{nk}\right.]$

最后用公式

$X_{nk}^*=X_{nk}+[\max(f(Z_k;\hat{\alpha}_n))-f(Z_k;\hat{\alpha}_n)]+[\max(v_{nk})-v_{nk}]\quad k=1,2,\cdots,K;n=1,2,\cdots,N$ 剔除环境因素和随机因素，排除它们对效率测度的影响。

其中，X_{nk}^* 是调整后的投入；X_{nk} 是调整前的投入；用 $[\max(f(Z_k;\hat{\alpha}_n))-f(Z_k;\hat{\alpha}_n)]$ 剔除环境因素；$[\max(v_{nk})-v_{nk}]$ 用于剔除随机因素。

附录 3：Malmquist 指数具体理论

Malmquist 指数与 DEA 相结合，用于描述面板数据的信息。该模型通过当前时期到下一时期生产率的变化，测算 Malmquist 全要素生产率指数，是对各个地区不同时期数据的动态效率分析，包括综合技术效率变化以及技术进步指数。

Malmquist 指数原理

$x_j^t=(x_{1j}^t,x_{2j}^t,...,x_{mj}^t)^T$：第 j 个地区在 t 时期的投入指标值，$y_j^t=(y_{1j}^t,y_{2j}^t,...,y_{mj}^t)^T$：第 j 个地区在 t 时期的产出指标值，均为正数。

规模报酬不变：

$D_C^t(x^t,y^t)$ 表示 (x^t,y^t) 在 t 时期的距离函数，$D_C^{t+1}(x^t,y^t)$ 表示在 $t+1$ 时期的距离函数；$D_C^t(x^{t+1},y^{t+1})$ 表示 (x^{t+1},y^{t+1}) 在 t 时期的距离函数，$D_C^{t+1}(x^{t+1},y^{t+1})$ 表示在 $t+1$ 时期的距离函数。则有

$$M^{t}=\frac{D_C^{t}(x^{t+1},y^{t+1})}{D_C^{t}(x^{t},y^{t})}\quad M^{t+1}=\frac{D_C^{t+1}(x^{t+1},y^{t+1})}{D_C^{t+1}(x^{t},y^{t})},$$

分别表示t时期技术水平下，从t期到$t+1$期技术效率的变化值；$t+1$时期技术水平下，从t期到$t+1$期技术效率的变化值。

通过计算M^{t}与M^{t+1}的几何平均值，得到从t期到$t+1$期的综合技术效率变化值，公式如下：

$$M(x^{t},y^{t},x^{t+1},y^{t+1})=(M^{t}\times M^{t+1})^{\frac{1}{2}}=\left[\frac{D_C^{t}(x^{t+1},y^{t+1})}{D_C^{t}(x^{t},y^{t})}\times\frac{D_C^{t+1}(x^{t+1},y^{t+1})}{D_C^{t+1}(x^{t},y^{t})}\right]^{\frac{1}{2}}$$

规模报酬可变：

$D_V^{t}(x^{t},y^{t})$表示(x^{t},y^{t})在t时期的距离函数，$D_V^{t}(x^{t},y^{t})$表示(x^{t},y^{t})在$t+1$时期的距离函数；$D_V^{t}(x^{t+1},y^{t+1})$表示在(x^{t+1},y^{t+1})在t时期的距离函数，$D_V^{t+1}(x^{t+1},y^{t+1})$表示在$t+1$时期的距离函数。同理可得如下公式：

$$M^{t}=\frac{D_V^{t}(x^{t+1},y^{t+1})}{D_V^{t}(x^{t},y^{t})},M^{t+1}=\frac{D_V^{t+1}(x^{t+1},y^{t+1})}{D_V^{t+1}(x^{t},y^{t})}$$

$$M(x^{t},y^{t},x^{t+1},y^{t+1})=(M^{t}\times M^{t+1})^{\frac{1}{2}}=\left[\frac{D_V^{t}(x^{t+1},y^{t+1})}{D_V^{t}(x^{t},y^{t})}\times\frac{D_V^{t+1}(x^{t+1},y^{t+1})}{D_V^{t+1}(x^{t},y^{t})}\right]^{\frac{1}{2}}$$

本文根据 Ray 和 Desli 提出的 Malmquist 指数分解的 RD 模型分解形式，然后得到了全要素生产率变化指数公式：

$$\begin{aligned}&TFPCH=M_{RD}(x^{t},y^{t},x^{t+1},y^{t+1})\\&=\frac{D_V^{t+1}(x^{t+1},y^{t+1})}{D_V^{t}(x^{t},y^{t})}\times\left[\frac{D_V^{t}(x^{t},y^{t})}{D_V^{t+1}(x^{t},y^{t})}\times\frac{D_V^{t}(x^{t+1},y^{t+1})}{D_V^{t+1}(x^{t+1},y^{t+1})}\right]^{\frac{1}{2}}\\&\times\left[\frac{D_C^{t}(x^{t+1},y^{t+1})/D_V^{t}(x^{t+1},y^{t+1})}{D_C^{t}(x^{t},y^{t})/D_V^{t}(x^{t},y^{t})}\times\frac{D_C^{t+1}(x^{t+1},y^{t+1})/D_V^{t+1}(x^{t+1},y^{t+1})}{D_C^{t+1}(x^{t},y^{t})/D_V^{t+1}(x^{t},y^{t})}\right]^{\frac{1}{2}}\\&=PC\times TC\times SC(\text{该TC指技术进步变化指数})\end{aligned}$$

其中，技术效率变化指数$TC=PC\times SC$。

将 Malmquist 指数运用结果得到的各指数记作 M，则 M 表示各时期到下一时期的效率动态变化。若 M>1，表示该时期到下一时期的效率提高；反之亦然；M=1，表示该时期到下一时期的效率不变。

纯技术效率变化（PC）指从t期到$t+1$期技术的被利用程度；技术进步变化（TC）指从t期到$t+1$期技术能力的变化状况；规模效率变化（SC）指从t期到$t+1$期决策单元的投入产出规模与生产最大化时的比例关系。

附录 4：堆积条形图

广东、上海、重庆、安徽、天津 5 个投入指标的堆积条形图

```
name_cz <- c(rep(c("广东", "上海", "重庆",
                   "安徽", "天津"), each = 5))
subject <- c(rep(c("数据源指数", "大数据开发指数", "网络经济指数","大数据产业劳动要素投入","大数据产业资本要素投入"), 5))
score_cz <- c(0.6973 ,   0.4930 ,   0.9019 ,   0.0407,    0.9000,
              0.3537 ,   0.3603 ,   0.5087 ,   0.2586,    0.2584,
              0.2531 ,   0.1918 ,   0.1788 ,   0.0903,    0.0860,
              0.1057 ,   0.2208 ,   0.2826 ,   0.0729,    0.1082,
              0.0546 ,   0.2404 ,   0.2258 ,   0.1273,    0.0836
)

data_cz <- data.frame(name_cz, subject, score_cz)
print(data_cz)

ggplot(data_cz,aes(x = name_cz, y = score_cz, fill = subject))+
  geom_bar(position = "dodge", stat = "identity", colour = "black")+
  scale_fill_manual(values = c("#363433", "#aa6a4c",
"#b89485","#0000FF","#FFA500"))

ggplot(data_cz,aes(x = name_cz, y = score_cz, fill = subject))+
  geom_bar(position = "dodge", stat = "identity", width = 0.2 )+            #
width 更改条的大小
  scale_fill_manual(values = c("#363433", "#aa6a4c",
"#b89485","#0000FF","#FFA500"))
ggplot(data_cz,aes(x = name_cz, y = score_cz, fill = subject))+
  geom_bar(position = position_dodge(width = 1),
# colour
           stat = "identity" )+
  scale_fill_manual(values = c("#363433", "#aa6a4c",
"#b89485","#0000FF","#FFA500"))
```

```
ggplot(data_cz,aes(x = name_cz, y = score_cz, fill = subject))+
  geom_bar(stat = "identity" )
```

附录 5：层次分析

标度含义表

标度	含义
1	表示两个因素相比，具有同样的重要性
3	表示两个因素相比，一个因素比另一个因素稍微重要
5	表示两个因素相比，一个因素比另一个因素明显重要
7	表示两个因素相比，一个因素比另一个因素强烈重要
9	表示两个因素相比，一个因素比另一个因素极端重要
2,4,6,8	上述两相邻判断的中值
倒数	因素 i 与 j 的判断比较a_{ij},则，因素 j 与 i 比较的判断$a_{ij}=1/a_{ij}$

一致性指标 RI

n	1	2	3	4	5	6	7	8	9	10	11
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49	1.51

层次分析判断矩阵：

大数据开发指数的二级指标判断矩阵

	每万名 R&D 人员专利授权数	规上企业R&D经费支出与 GDP 之比	技术市场成交合同金额	R&D 内部经费支出额
每万名R&D人员专利授权数	1	2	1/3	3
规上企业R&D经费支出与 GDP 之比	1/2	1	1/6	2
技术市场成交合同金额	3	6	1	1/9
R&D 内部经费支出额	1/3	1/2	1/9	1

网络经济指数的二级指标判断矩阵

	单位互联网宽带接入用户数	移动互联网用户数	移动互联网接入流量	电子商务平台交易额	有电子商务交易活动企业比重	网上零售总额占社会消费品零售总额的比重
单位互联网宽带接入用户数	1	1/4	1/4	1/5	1/2	1/3
移动互联网用户	4	1	1	1/2	3	2

数						
移动互联网接入流量	4	1	1	1/2	3	2
电子商务平台交易额	5	2	2	1	4	3
有电子商务交易活动企业比重	2	1/3	1/3	1/4	1	1/2
网上零售总额占社会消费品零售总额的比重	3	1/2	1/2	1/3	2	1

大数据产业劳动要素投入的二级指标判断矩阵

	经济活动人口中研究与开发机构R&D人员硕士及以上人数比例	R&D人员总数	每万名就业人员R&D人员折合全时当量
经济活动人口中研究与开发机构R&D人员硕士及以上人数比例	1	7	3
R&D人员总数	1/7	1	1/5
每万名就业人员R&D人员折合全时当量	1/3	5	1

大数据产业资本要素投入的二级指标判断矩阵

	各地区高等学校经费支出总额	各地区高技术产业R&D经费内部支出
各地区高等学校经费支出总额	1	1/3
各地区高技术产业R&D经费内部支出	3	1

转型升级指数的二级指标判断矩阵

	国家级高新区总产值占工业增加值比重	国家级高新区出口额占出口总额的比重	单位GDP能耗降低率
国家级高新区总产值占工业增加值比重	1	1/3	1/5
国家级高新区出口额占出口总额的比重	3	1	1/3
单位GDP能耗降低率	5	3	1

软件业务收入的二级指标判断矩阵

	软件产品收入	软件技术服务收入
软件产品收入	1	3
软件技术服务收入	1/3	1

层次分析（MATLAB）：

```
A=[    ];
W = prod(A, 2)
n = size(A, 1);
W = nthroot(W, n)
W = W / sum(W)
Lmax = mean((A * W) ./ W)
[V,D] = eig(A)
[Lmax,ind] = max(diag(D));
Lmax
W = V(:,ind) / sum(V(:,ind))
n = size(A, 1);
CI = (Lmax - n) / (n - 1)
RI = [0 0 0.58 0.90 1.12 1.24 1.32 1.41 1.45 1.49 1.51];
CR = CI / RI(n)
```

附录 6：蒙特卡洛综合评价算法及代码（Python）

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import random
from sklearn import preprocessing
#  导入数据
#  存在序关系的
#  原始评价数据矩阵（方案矩阵）
data = pd.read_csv("D:\\2019 年投入指标数据.csv", sep=',',header=None)
print(data,'\n')

# n 个评价对象
n = len(data)
print('评价对象个数:',n)
# m 个评价指标
m = data.shape[1]
print('评价指标个数:',m)
#  两两比较
list = []
for i in range(n):
    list.append([i,i])
    for j in range(i+1,n):
```

```
            list.append([i,j])
            list.append([j,i])
l = len(list)
print(list)
print(l)
#  初始化优胜度矩阵
S = pd.DataFrame(np.zeros((n,n)))
print(S)
for t in range(l):
    #  仿真次数
    count_ = 0
    #  仿真总次数
    sum = 1000

    r_s, r_e, r_f = 0, 0, 0

    while (sum != count_):
        #  不确定比值
        #  两两之间的重要性比值判断
        q =
[(0.968,1.146),(0.923,1.153),(0.932,1.125),(0.367,0.534),(0.823,0.904)]
        r_k = []
        for i in q:
            r_k.append(np.random.uniform(i[0], i[1]))
        #  优势权向量
        w = [0]
        b=0
        for n in range(2, m + 1):
            r = 0
            for i in range(2, n + 1):
                for j in range(i, n + 1):
                    r = r + r_k[j - 2]
            w.append(1 / (1 + r))
            b=b+1
        w[0] = r_k[0] * w[1]

        # ww = [i/pd.Series(w).sum() for i in w]

        #  得到优胜权向量
        # ww
        #  计数变量

        a = list[t][0]
        b = list[t][1]
```

```
            o_a = 0
            o_b = 0

            for j in range(m):
                o_a += data.iloc[a, j] * w[j]
                o_b += data.iloc[b, j] * w[j]

            if o_a > o_b:
                r_s += 1
            elif o_a == o_b:
                r_e += 1
            else:
                r_f += 1

            count_ += 1

            # 优胜度矩阵
            S[a][b] = (r_s+0.5*r_e) / sum

        t += 1
print(S)
# 相对综合评价值
z = []
for i in range(len(S)):
    c = 0
    for j in range(len(S)):
        c += S[i][j]-S[j][i]
    z.append(1+c/len(S))
print(z)
```

下表为输出的相对综合评价值表：

地区	相对综合评价值					
广东	1.967742					
北京	1.903226					
江苏	1.83871					
上海	1.774194					
浙江	1.709677					
四川	1.645161					
甘肃	1.580645					
山东	1.516129					
重庆	1.451613					

31 个省市蒙特卡洛相对综合评价值：

```
library(ggplot2)
library(tidyverse)
```

```
# install.packages('ragg')
library(ragg)
data = read.table("D:\\31 个省市蒙特卡洛相对综合评价
值.csv",sep=",",header=TRUE)
print(data)
data <-arrange(data,year,MonteCarlo)
print(data)
data$id <- seq(1, nrow(data))
head(data)


label_data <- data
number_of_bar <- nrow(label_data)
angle <- 90 - 360 * (label_data$id-0.5) /number_of_bar
label_data$hjust <- ifelse( angle < -90, 1, 0)
label_data$angle <- ifelse(angle < -90, angle+180, angle)
head(label_data)

#  绘制排序分组环状条形图
p <- ggplot(data, aes(x=as.factor(id), y=MonteCarlo, fill=year)) +
   geom_bar(stat="identity", alpha=0.5) +
   ylim(-100,120) +
   theme_minimal() +
   scale_fill_brewer(palette = "Set1") +
   theme(
      legend.position = "none",
      axis.text = element_blank(),
      axis.title = element_blank(),
      panel.grid = element_blank(),
      plot.margin = unit(rep(-1,4), "cm")
   ) +
   coord_polar() +
   geom_text(data=label_data, aes(x=id, y=MonteCarlo+10, label=province,
hjust=hjust), color="black", fontface="bold",alpha=0.6, size=2.5, angle=
label_data$angle, inherit.aes = FALSE )
p
```

附录 7：2017——2019 年国家高新区数量环形条形图

```
library(ggplot2)
library(tidyverse)
# install.packages('ragg')
library(ragg)
data = read.table("D:\\2017——2019 年国家高新区数量环形条形
图.csv",sep=",",header=TRUE)
print(data)
data <-arrange(data,year,Incubator)
print(data)
data$id <- seq(1, nrow(data))
head(data)


label_data <- data
number_of_bar <- nrow(label_data)
angle <- 90 - 360 * (label_data$id-0.5) /number_of_bar
label_data$hjust <- ifelse( angle < -90, 1, 0)
label_data$angle <- ifelse(angle < -90, angle+180, angle)
head(label_data)

#  绘制环状条形图
p <- ggplot(data, aes(x=as.factor(id), y=Incubator, fill=year)) +
   geom_bar(stat="identity", alpha=0.5) +
   ylim(-100,120) +
   theme_minimal() +
   scale_fill_brewer(palette = "Set1") +
   theme(
      legend.position = "none",
      axis.text = element_blank(),
      axis.title = element_blank(),
      panel.grid = element_blank(),
      plot.margin = unit(rep(-1,4), "cm")
   ) +
   coord_polar() +
   geom_text(data=label_data, aes(x=id, y=Incubator+10, label=province,
hjust=hjust), color="black", fontface="bold",alpha=0.6, size=2.5, angle=
label_data$angle, inherit.aes = FALSE )
p
```

附录 8：2017——2019 年科技企业孵化器数量环形条形图

```
library(ggplot2)
library(tidyverse)
# install.packages('ragg')
```

```
library(ragg)
data = read.table("D:\\2017——2019 年科技企业孵化器数量环形条形
图.csv",sep=",",header=TRUE)
print(data)
data <-arrange(data,year,Incubator)
print(data)
data$id <- seq(1, nrow(data))
head(data)

label_data <- data
number_of_bar <- nrow(label_data)
angle <- 90 - 360 * (label_data$id-0.5) /number_of_bar
label_data$hjust <- ifelse( angle < -90, 1, 0)
label_data$angle <- ifelse(angle < -90, angle+180, angle)
head(label_data)

# 绘制环状条形图
p <- ggplot(data, aes(x=as.factor(id), y=Incubator, fill=year)) +
  geom_bar(stat="identity", alpha=0.5) +
  ylim(-100,120) +
  theme_minimal() +
  scale_fill_brewer(palette = "Set1") +
  theme(
    legend.position = "none",
    axis.text = element_blank(),
    axis.title = element_blank(),
    panel.grid = element_blank(),
    plot.margin = unit(rep(-1,4), "cm")
  ) +
  coord_polar() +
  geom_text(data=label_data, aes(x=id, y=Incubator+10, label=province,
hjust=hjust), color="black", fontface="bold",alpha=0.6, size=2.5, angle=
label_data$angle, inherit.aes = FALSE )
p
```

附录 9：第二阶段 SFA 回归结果：

		投入松弛变量					环境变量		
DUM		1	2	3	4	5	1	2	3
1	1	-0.03496	0.143697	0.005946	6.74E-05	0.010519	1.0000	0.5897	1.0000
2	1	-0.03731	-0.2323	0.074536	-0.0001	0.031146	0.1381	1.0000	0.2244
3	1	-0.00435	0.040961	0.004486	-0.00032	0.000431	0.8088	0.2718	0.3548
4	1	0.223831	-0.06546	0.035529	0.201446	0.064927	0.2324	0.3804	0.3074
5	1	0.020862	0.202785	0.003557	-0.00442	0.017989	0.3108	0.2897	0.7827
6	1	-0.00549	-0.14666	-0.03938	-0.00215	0.014943	0.1886	0.2940	0.1093
7	1	-0.01487	-0.1174	-0.0073	-0.00866	0.019165	0.1102	0.2036	0.0071
8	1	0.007167	0.056945	-0.02125	-0.00467	0.066221	0.4011	0.3525	0.1495

附录 10：:2017 年的投入松弛 2 和 3 的 OLS 估计 R 代码：

```
a=read.table("D:\\123.csv",",",header=T)
print(a)
b=lm(y ~ x1 + x2 + x3, data = a)
A = b$coefficients["x1"]
B = b$coefficients["x2"]
C = b$coefficients["x3"]
print(A)
print(B)
print(C)
summary(b)
```

第二阶段调整后的数据：

	产出指标		投入指标			
地区	转型升级	软件业务	数据源指	大数据开	网络经济	大数据产
广东	0.1777	0.8414	0.700556	0.716769	0.997943	0.027067
北京	0.1718	1.0000	0.196036	1.044769	0.629943	1.002522
江苏	0.1139	0.7811	0.744979	0.681769	0.549943	0.125924
上海	0.1320	0.5116	0.347259	0.606768	0.556943	0.311956
浙江	0.1622	0.4804	0.109862	0.56177	0.647943	0.050086
四川	0.2098	0.3547	0.249805	0.552767	0.426942	0.290289
甘肃	0.1662	0.0051	0.406745	0.464768	0.189943	0.164392
山东	0.1067	0.4914	0.062966	0.632772	0.563943	0.070816

附录 11：2018-2019 年 DEA 第三阶段结果

2018 年第三阶段测度结果

地区	技术效率	纯技术效率	规模效率	
广东	1	1	1	-

北京	1	1	1	-
江苏	1	1	1	-
上海	0.886	0.993	0.892	irs
浙江	1	1	1	-
四川	1	1	1	-
甘肃	1	1	1	-
山东	1	1	1	-
重庆	0.77	1	0.77	irs
陕西	1	1	1	-
吉林	1	1	1	-
安徽	0.964	1	0.964	irs
湖南	1	1	1	-
湖北	1	1	1	-
天津	0.67	1	0.67	irs
均值	0.953	1	0.953	

2019 年第三阶段测度结果

地区	技术效率	纯技术效率	规模效率	
广东	1	1	1	-
北京	1	1	1	-
江苏	1	1	1	-
上海	0.927	1	0.927	irs
浙江	1	1	1	-
四川	1	1	1	-
甘肃	0.658	1	0.658	irs
山东	1	1	1	-
重庆	0.598	1	0.598	irs
陕西	1	1	1	-
吉林	1	1	1	-
安徽	0.768	1	0.768	irs
湖南	1	1	1	-
湖北	1	1	1	-
天津	0.639	1	0.639	irs
均值	0.906	1	0.906	

致谢

首先我们要尤其感谢指导老师对我们的悉心指导，从开题报告到实际操作，他们提出了许多建议。每当我们有疑问时，他们耐心解答我们的问题，为我们提出可供选择的解决方案。在我们有困惑时，他们细心分析，提出自己的见解；在我们怀疑自己时给予我们肯定，遇到困难时为我们出谋划策。在老师们的指导下，我们顺利的完成了此次论文。

其次要感谢我们自己，在前期确定选题时的迷茫状态下我们每天阅读大量论文，参考大量书籍，直到基本确定建模方向。后期实际操作时多次遇到困难，我们不曾放弃，逐个解决问题。我们在彼此的配合与帮助下完成了这篇论文。

最后要感谢参与此次论文评阅的所有老师，感谢您的阅读与建议，我们将以更严格的要求规范自己。