中国数字经济发展水平测度分析

闫鑫泉

北方民族大学数学与信息科学学院 宁夏银川

【摘要】目的 随着智能化时代的到来，中国经济发展水平受到数字经济的影响。在新发展格局下，数字经济占据着主导地位，它不仅象征着我国经济发展速度的快慢，而且是我国综合实力的具体体现。因此，研究数字经济发展水平对我国做出重要性的决策、及时调整总体布局以及合理评价各项相关性指标具有深远的意义。**方法** 首先，依据熵权 TOPSIS 法确定各项指标的相应权重并得到各省份（除西藏、港澳台）综合得分，基于上述所得结果利用 AGNES 聚类算法对我国 30 个省的数字经济发展水平进行梯队划分，然后，利用 XGBOOST 算法进行特征重要性分析，最后，运用 LSTM 模型对 2021 年各省地区生产总值进行预测。**结论** 在我国数字经济发展中，科技创新占据较大比重，它在经济发展进程中起着关键作用；对比全国、东部、中部和西部地区的核密度估计图，西部地区的数字经济发展水平在逐年提升；通过对地区生产总值的预测，可以看出我国的数字经济发展水平在稳步提高。

【关键词】熵权 TOPSIS；数字经济；AGNES 层次聚类；LSTM

【基金项目】北方民族大学研究生创新项目“大数据背景下新零售行业数据质量评价模型的建立及实证研究”（YCX22090）

【收稿日期】2023 年 10 月 31 日 **【出刊日期】**2023 年 12 月 8 日 **【DOI】**10.12208/j.ssr.20230017

Analysis of measurement of China’s digital economy development level

Xinquan Yan

School of Mathematics and Information Science, North Minzu University, Yinchuan, Ningxia

【Abstract】Objective With the advent of the intelligent era, the level of China's economic development is influenced by the digital economy. In the new development paradigm, the digital economy occupies a dominant position. It not only symbolizes the speed of our country's economic development but also represents the concrete embodiment of our comprehensive strength. Therefore, studying the level of digital economy development is of profound significance for making important decisions, adjusting overall layout in a timely manner, and reasonably evaluating various relevant indicators in China. **Method** Firstly, the weights of various indicators are determined based on the entropy weight TOPSIS method, and the comprehensive scores of each province (excluding Tibet, Hong Kong, Macau, and Taiwan) are obtained. Based on the results obtained above, the AGNES clustering algorithm is used to classify the digital economy development levels of the 30 provinces in China. Then, the XGBoost algorithm is employed to analyze the feature importance. Lastly, the LSTM model is utilized to predict the regional gross domestic product (GDP) of each province for the year 2021. **Conclusion** In the development of China's digital economy, technological innovation plays a significant role and occupies a large proportion. Comparing the kernel density estimation maps of the national, eastern, central, and western regions, the digital economy development level of the western region has been improving year by year. Through the prediction of regional gross domestic product (GDP), it can be seen that China's digital economy development level is steadily increasing.

作者简介：闫鑫泉（1997-）男，汉族，山西阳泉，北方民族大学数学与信息科学学院硕士研究生，研究方向：经济与社会统计。

【**Keywords**】Entropy TOPSIS; Digital Economy; AGNES; LSTM

1 引言

近年来数字经济蓬勃发展，数字经济持续高速增长，已经成为我国应对经济下行压力的关键抓手，数字经济是推动国民经济持续稳定增长的关键动力，我国的数字经济发展方面存在着巨大潜力，随着时代的进步，大数据、物联网、5G 等新型技术的迅飞速发展，中国数字经济便成了新发展格局的首要推动力[1]。数字经济是通过直接或间接的方法来使各类资源发挥相应作用，从而达到推动生产力发展的最终目的[2]。面对目前的大环境，在我国的经济发展中数字经济日益成为中坚力量。2021 年，数字经济在“十四五”规划中明确认定为未来推动中国经济发展的重要手段。在第六届数字中国建设峰会发布的《数字中国发展报告(2022 年)》的报告中指出，2022 年我国数字经济规模达 50.2 万亿元，总量稳居世界第二，占 GDP 比重提升至 41.5%。对于数字经济发展水平的测度分析在我国判断当前的经济形势以及规范其产业发展中发挥重要作用,可以提供促进我国数字经济发展的核心基础数据[3]。数字经济不仅给我国带来经济效益方面的增加，也使我国生产企业模式改革与创新的不竭动力。

数字经济主要包括数字产业化和产业数字化两大模块，数字产业化可以将消费、生产、服务中的重要性技术所产生的数据变为生产要素，从而用于新领域的发展；产业数字化借助科研创新将传统的产业向数字化转变，实现农业、制造业的数字化升级，以及互联网的普及，从而有效应用数字技术来改造三次产业。在新发展格局下，数字经济占据着主导地位，它不仅象征着我国经济发展速度的快慢，而且是我国综合实力的具体体现。因此，研究数字经济发展水平对我国做出重要性的决策、及时调整总体布局以及合理评价各项相关性指标具有深远的意义。

2 指标体系构建及数据来源

2.1 指标体系构建原则

为了全面科学合理的测度我国各省（西藏、港澳台除外）的数字经济水平发展状况，本文依据以下原则进行指标的选取：

• 科学性原则

在构建指标体系中所选取指标必须具有科学性，才能使得分析结果更具准确性。在充分考虑指标准确性的条件之下，科学的选取各项指标进行指标体系的构建。

• 客观性原则

构建的数字经济水平测度指标体系是客观反映我国数字经济水平发展的状况，因此选取的指标要从客观条件出发，提高整个分析结果的客观真实性。

• 整体性原则

为科学准确的进行数字经济水平测度研究，在构建指标体系时要有全局思维，应该从整体出发，使指标体系能够全面反映我国的数字经济发展水平。

• 可得性原则

为了准确反映我国数字经济发展水平，要求所选取的指标数据是易于获得且真实的数据，对于不能统计或不易收集的数据，本文不予研究。

2.2 指标设计

综合对比各种文献研究成果，本文选取数字基础设施、经济产业和数字融合等一级指标，在该指标下设立相应的 8 个二级指标，并且在二级指标下选取了 23 个三级指标来构建指标体系[4-6]。

（1）数字基础设施指标选取

在数字经济发展中基础设施发挥了一定程度的作用，综合对比各结论之后，本文在数字基础设施指标下设立终端基础设施和移动段基础设施等二级指标，在终端设施下设立 IPv4 地址数、互联网宽带接入端口和光缆线路长度这 3 个三级指标；IPv4 表示 IP 协议的第四个版本，在互联网上相当数量的通信流量都是以 IPv4 数据包的格式来进行封装的；光缆线路长度在经济发展中是重要基础。在移动端设施指标下设立移动电话用户和电话普及率这 2 个三级指标，这些指标可以很好反映出地区经济发展潜力。

（2）数字经济产业指标选取

本文在数字经济产业下设立产业规模、科技创新、经济增长和社会发展这 4 个二级指标，产业规模指标下包括：分地区技术市场成交额、软件产品收入和信息技术服务收入，本文主要选取信息技术方面的产业来反映数字经济发展的产业规模。

在科技创新指标下设立 R&D 人员全时当量、R&D 项目数、专利发明数和新产品开发项目数等三

级指标。R&D 人员全时当量表示的是每个地区从事科技创新研究行业的相应人员数目，科技创新是数字经济发展中的关键要素，科技创新发展能力越强表示数字经济发展的潜力更大。

经济增长指标下包括：网上零售额绝对值、电子商务销售额、第三产业增加值和工业资产总计。在数字经济发展过程中，第三产业和工业都是重要组成部分，与此同时数字经济的发展也带动了电子商务产业的相应发展。

在社会发展指标下设立外商投资总额和城镇登记失业率等三级指标，外商投资不仅包含国外资本、产品和服务的投入，还有国外先进的理念与技术，在各地区数字经济发展与建设中发挥着重要作用。城镇登记失业率也可以在一定程度上体现该地区的社会发展水平，通过失业率可以看出该地区的就业机会，提供就业机会的企业越多表示地区的经济越发达，也代表地区数字经济发展有更加夯实的基础。

（3）数字经济融合指标选取

本文在数字经济融合指标下设立个人应用和企业应用 2 个二级指标，个人应用指标下包括：移动互联网用户和互联网宽带接入用户，通过这 2 个指标来反映个人在数字经济融合方面的应用情况。企业应用指标下包括：有电子商务交易活动企业数等相应指标，数字化发展已经渗透到企业的各个方面，本文选取这 3 个指标来反映地区的数字经济融合情况。

表 1　数字经济发展水平测度指标体系

一级指标	二级指标	三级指标	单位
数字基础设施	终端基础设施	IPv4 地址数 x1	万个
		光缆线路长度 x2	公里
		互联网宽带接入端口 x3	万个
	移动端基础设施	移动电话用户 x4	万户
		电话普及率（包括移动电话）x5	部/百人
数字经济产业	产业规模	分地区技术市场成交额 x6	万元
		软件产品收入 x7	亿元
		信息技术服务收入 x8	亿元
	科技创新	R&D 人员全时当量 x9	人年
		R&D 项目数 x10	项
		专利发明数 x11	件
		新产品开发项目数 x12	项
	经济增长	网上零售额绝对值 x13	亿元
		电子商务销售额 x14	亿元
		第三产业增加值 x15	亿元
		工业资产总计 x16	亿元
	社会发展	外商投资总额 x17	亿美元
		城镇登记失业率 x18	%
数字经济融合	个人应用	移动互联网用户 x19	万户
		互联网宽带接入用户 x20	万户
	企业应用	有电子商务交易活动企业数 x21	个
		每百人使用计算机数 x22	台
		每百家企业拥有网站数 x23	个

2.3 *数据来源*

本文在国家统计局上截取 2016—2020 年中国 30 个省的面板数据作为研究对象，鉴于数据可得性，剔除西藏、港澳台的数据。本文所有数据均来自国家统计局网站、《中国统计年鉴》《中国工业统计年鉴》《中国信息产业年鉴》《中国科技统计年鉴》以

及各个省份的历年统计年鉴。

3 研究方法

3.1 熵权 TOPSIS 法

熵权法作为一种综合评价方法它可以适用于多个对象和指标，它是根据各指标分散程度来测定指标的相应权重，分散程度较大的变量，其本身蕴含的信息量越丰富，相应的指标权重就越大。与此同时确定各个指标权重时使用信息熵来进行，信息熵值越小，权重越大。此算法别名为优劣解距离法，它是通过对评价对象与最优解、最劣解的相对距离进行排序，从而得到对评价对象的相对优劣的评价。本文将熵权法与 TOPSIS 法相融合，用于新发展格局下的数字经济测度，使得测算结果更加客观可靠[7]。步骤如下：

（1）各指标数据的归一化处理

之前数据处理阶段已经对数据进行标准化处理

（2）计算信息熵

$$E_r = -k\sum_{i=1}^{n}\left(Z_{ir}/\sum_{i=1}^{n}Z_{ir}\right)\ln\left(Z_{ir}/\sum_{i=1}^{n}Z_{ir}\right)$$

其中，$k=\frac{1}{\ln n}$。

（3）计算各指标的熵权

$$\omega_r = (1-E_r)/\sum_{r=1}^{R}(1-E_r)$$

其中，R 表示指标的总数。

（4）计算各指标的加权指数

$$Q_{ir} = \omega_r Z_{ir}$$

（5）确定相应评价对象及最优解、最劣解的相对距离 D_i^g 和 D_i^b

$$D_i^g = \sqrt{\sum_{r=1}^{R}\left(Q_{ir}^g - Q_{ir}\right)^2} \quad D_i^b = \sqrt{\sum_{r=1}^{R}\left(Q_{ir} - Q_{ir}^b\right)^2}$$

其中，$Q_{ir}^g = \max\{Q_{1r},Q_{2r},...,Q_{nr}\}$，$Q_{ir}^b = \min\{Q_{1r},Q_{2r},...,Q_{nr}\}$

（6）判断各对象与理想值之间的相对接近程度 C_i

$$C_i = \frac{D_i^b}{D_i^g + D_i^b}$$

这里，C_i 越大，表明接近程度越高；C_i 越小，表明接近程度越低。

3.2 AGNES 聚类

AGNES 聚类算法是先确定初始聚类簇，然后计算每个簇之间的距离，通过距离计算结果进行相应的合并，在整个构建过程中不断重读此项操作，当达到设定的阈值时，模型立刻停止训练，因此该算法的核心关键在于计算距离公式的选取。在该算法中每个簇都可看作一个相应集合，从而关键问题转换为集合距离公式的选取，通常有以下公式：

最小距离：$d_{\min}\left(C_i,C_j\right) = \min_{x\in C_i, z\in C_j} dist(x,z)$

最大距离：$d_{\max}\left(C_i,C_j\right) = \max_{x\in C_i, z\in C_j} dist(x,z)$

平均距离：

$$d_{avg}\left(C_i,C_j\right) = \frac{1}{\left|C_i\right|\left|C_j\right|}\sum_{x\in C_i}\sum_{z\in C_j} dist(x,z)$$

AGNES 算法先对仅含一个样本的初始聚类簇和相应的距离矩阵进行初始化，然后不断合并距离最近的聚类簇，并对合并得到的聚类簇的距离矩阵进行更新，不断重复此过程直至达到预设的聚类簇数。

3.3 相关分析

相关分析即度量两个数值型变量之间的相关性，以及计算相关程度的大小。相关系数是专门用来衡量两个变量之间的相关程度的指标，常见的相关系数有以下几种：

Pearson 积差相关系数

测度两个变量 X 和 Y 之间的线性相关程度。它具有+1 和-1 之间的值，其中 1 是总正线性相关性，0 是非线性相关性，并且-1 是总负线性相关性。Pearson 相关系数的一个关键数学特性是它在两个变量的位置和尺度的单独变化下是不变的。其公式为：

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y}$$

Spearman 秩相关系数

该算法不要求原始数据的分布状况，是非参数统计中的一种重要方法，正因如此，该算法的应用范围更加广阔，它是通过计算数据之间的秩次大小来测度数据之间的线性相关程度，算法结果的取值范围为[-1,1],它的统计效能相对较低。

在样本容量为 n 的样本中，n 个原始数据被转

换成等级数据，相关系数 ρ 为：

$$\rho=\frac{\sum_i(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_i(x_i-\bar{x})^2\sum_i(y_i-\bar{y})^2}}$$

Kendall 秩相关系数

当数据为有序分类的情况下，可以考虑使用该算法，一般用 τ（tau）来表示其结果，结果取值范围为[-1,1]，取值为 1 表明数据的等级相关性高度契合，为-1 表示数据的等级相关性是截然不同的，取值为 0 表示数据之间没有任何线性相关关系，具体计算公式如下：

$$\tau=\frac{\left(number\ \ of\ \ concordant\ \ pairs\right)-\left(number\ \ of\ \ discordant\ \ pairs\right)}{\frac{1}{2}n(n-1)}$$

3.4 LSTM 算法

LSTM 算法又称长短时记忆，其设计的核心思想之一就是通过控制一种类似普通的递归神经网络（RNN）单元中增加或引入了门的概念来控制 RNN 单元。其中，RNN 单元一般视为用来存储短暂记忆的单元 h，LSTM 在 RNN 的基础上增加了存储过去信息的记忆单元 C。

LSTM 算法中涵盖不同类型的门：遗忘门、输入门和输出门。在遗忘门中，它决定了 C_{t-1} 是否要被遗忘；在输入门中，它决定是否把 $\tilde{C}_i$ 带入到记忆中来进行记忆的更新换代操作；在输出门中，它决定更新换代之后的记忆是否要被传输至下一个隐藏层。

遗忘门的工作机理是处理上一个状态中所涵盖的信息。其计算公式为：

$$f_t=\sigma\left(W_f\cdot\left[h_{t-1},x_t\right]+b_f\right)$$

在这进行传输的是上一个单元的状态 h_{t-1} 以及本次传输内容 x_t，σ 表示 sigmoid 函数。即将 t 时刻的输入 $X=[x_1,x_2,...,x_t]$ 与上一个隐藏层的数据 h_t 进行结合，然后用 W_f 矩阵将其调整为与 t 时刻隐藏层相同的维数，再加一个偏置 b_f，最后用上述函数进行 0~1 之间的分类。

输入门的工作机理是测定给记忆单元 C 增加新信息量的多少。新的信息由两部分计算得出。第一部分是输入 i_t：

$$i_t=\sigma\left(W_i\cdot\left[h_{t-1},x_t\right]+b_i\right)$$

第二部分是 C_t：

$$\tilde{C}_t=\tanh\left(W_C\cdot\left[h_{t-1},x_t\right]+b_C\right)$$

利用遗忘门以及输入门的工作机理，可以更新记忆单元 C 的状态：

$$C_t=f_t*C_{t-1}+i_t*\tilde{C}_t$$

迭代更新记忆单元状态之后，在输出门中利用其工作机理确定 LSTM 单元的输出。输出门的输出是：

$$o_t=\sigma\left(W_o\cdot\left[h_{t-1},x_t\right]+b_o\right)$$

使用 o_t 与 C_t 得到 h_t：

$$h_t=o_t*\tanh\left(C_t\right)$$

该模型流程图如下：

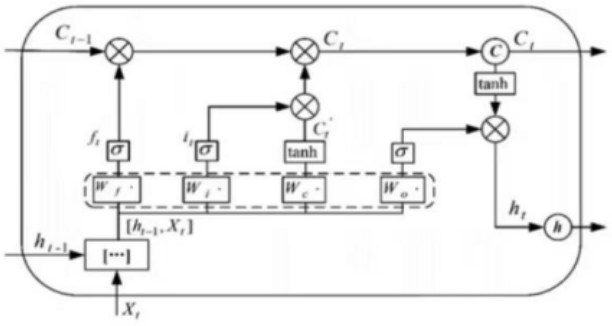

图 1 LSTM 模型流程图

4 计算指标权重及综合得分

4.1 数据处理

由于各数据之间的单位以及量纲的不同，各指标间数量级相差较大，无法直接使用原始指标数据进行测度研究，因此本文对指标数据进行标准化处理，本文选用极差化法对数据进行处理：

$$y_i=\frac{x_i-\min\limits_{1\le i\le n}\{x_j\}}{\max\limits_{1\le j\le n}\{x_j\}-\min\limits_{1\le j\le n}\{x_j\}}$$

4.2 指标权重及综合得分确定

由于本文选取 22 个基础指标进行数字经济水平测度研究，所选取的基础指标较多，而每一个基础指标的影响程度不尽相同，为了科学准确的进行分析研究，本文首先测定各指标的对应权重。本文选用熵权 TOPSIS 方法来测定指标权重并计算个指标的综合得分，利用 Python 软件计算 2016—2020 这 5 年间的各指标权重以及综合得分如下：

表 2 三级指标对应权重

三级指标权重	2016 年	2017 年	2018 年	2019 年	2020 年
IPv4 地址数 x1	0.0499	0.0604	0.0605	0.0448	0.2582
光缆线路长度 x2	0.0273	0.0258	0.0234	0.0164	0.0212
互联网宽带接入端口 x3	0.0233	0.0237	0.0245	0.0192	0.0219
移动电话用户 x4	0.0229	0.0218	0.0224	0.0164	0.0205
电话普及率（包括移动电话）x5	0.0259	0.0210	0.0254	0.1861	0.0273
分地区技术市场成交额 x6	0.0856	0.0783	0.0653	0.0482	0.0604
软件产品收入 x7	0.0704	0.0704	0.0393	0.0491	0.0725
信息技术服务收入 x8	0.0701	0.0695	0.0746	0.1486	0.0752
R&D 人员全时当量 x9	0.0543	0.0554	0.0621	0.0471	0.0604
R&D 项目数 x10	0.0598	0.0591	0.0610	0.0475	0.0608
专利发明数 x11	0.0691	0.0702	0.0728	0.0542	0.0677
新产品开发项目数 x12	0.0615	0.0660	0.0653	0.0485	0.0616
网上零售额绝对值 x13	0.0763	0.0719	0.0669	0.0492	0.0666
电子商务销售额 x14	0.0558	0.0561	0.0538	0.0407	0.0502
第三产业增加值 x15	0.0286	0.0283	0.0282	0.0213	0.0279
工业资产总计 x16	0.0262	0.0268	0.0281	0.0200	0.0615
外商投资总额 x17	0.0648	0.0697	0.0665	0.0441	0.0580
城镇登记失业率 x18	0.0069	0.0067	0.0062	0.0050	0.0107
移动互联网用户 x19	0.0229	0.0239	0.0230	0.0169	0.0215
互联网宽带接入用户 x20	0.0284	0.0268	0.0248	0.0175	0.0221
有电子商务交易活动企业数 x21	0.0398	0.0404	0.0429	0.0320	0.0409
每百人使用计算机数 x22	0.0252	0.0217	0.0259	0.0194	0.0244
每百家企业拥有网站数 x23	0.0051	0.0061	0.0068	0.0074	0.0084

表 3 各省份指标综合得分

省份	2016	2017	2018	2019	2020
北京	0.5177	0.5282	0.5083	0.2733	0.5019
天津	0.1562	0.1294	0.1269	0.0867	0.1220
河北	0.1550	0.1527	0.1472	0.0869	0.1453
山西	0.0798	0.0783	0.0781	0.0951	0.2282
内蒙古	0.0698	0.0705	0.0684	0.0400	0.0717
辽宁	0.1825	0.1691	0.1501	0.0813	0.1389
吉林	0.0695	0.0693	0.0703	0.0345	0.0685
黑龙江	0.0743	0.0759	0.0724	0.0436	0.0798
上海	0.4199	0.3495	0.3459	0.1773	0.3283
江苏	0.6266	0.5636	0.5745	0.2877	0.5214
浙江	0.5290	0.4865	0.4984	0.2671	0.4743
安徽	0.1953	0.1821	0.1783	0.1050	0.1731
福建	0.2238	0.2068	0.2078	0.3981	0.1712
江西	0.0997	0.0988	0.1097	0.0679	0.1127
山东	0.4191	0.3956	0.3972	0.1915	0.3553

省份	2016	2017	2018	2019	2020
河南	0.1935	0.1866	0.1781	0.1038	0.1662
湖北	0.2036	0.1866	0.1848	0.1039	0.1697
湖南	0.1510	0.1448	0.1509	0.0885	0.1493
广东	0.7339	0.7303	0.7594	0.3778	0.7131
广西	0.0781	0.0742	0.0788	0.0498	0.0924
海南	0.0673	0.0613	0.0651	0.0357	0.2117
重庆	0.1280	0.1182	0.1210	0.0609	0.1191
四川	0.2458	0.2356	0.2367	0.1283	0.2243
贵州	0.0654	0.0632	0.0655	0.0383	0.0669
云南	0.0829	0.0772	0.0807	0.0528	0.0878
陕西	0.1492	0.1426	0.1436	0.0772	0.1404
甘肃	0.0458	0.0479	0.0477	0.0285	0.0500
青海	0.0421	0.0442	0.0555	0.4591	0.0539
宁夏	0.0505	0.0497	0.0547	0.0235	0.0506
新疆	0.0565	0.0517	0.0605	0.0292	0.0531

4 我国数字经济水平发展综合分析

4.1 数字经济水平发展总体情况分析

根据各指标综合得分结果，绘制每一年各省份的指标得分情况：

图 2　2016、2017、2018 年各省份指标综合得分

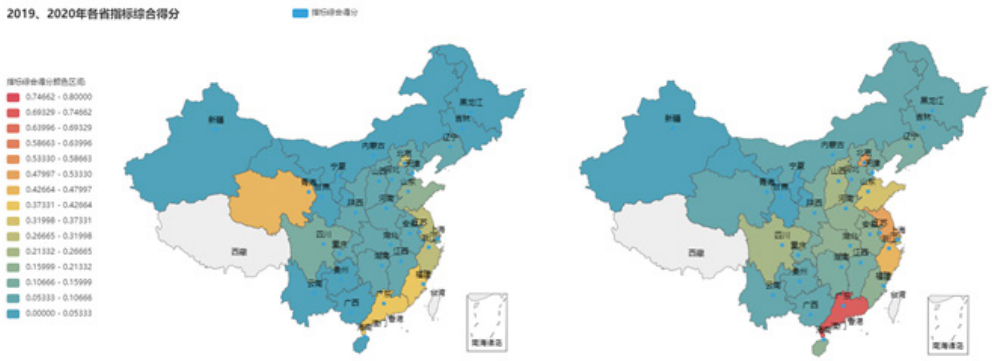

图 3　2019、2020 年各省份指标综合得分

在可视化结果中用颜色变化表示各省的指标综合得分，红色表示该省的指标综合得分较高，蓝色表示综合得分较低，没有取得数据的地区用白色来表示。从结果中可明显看出在 2016、2017 和 2018 年这三年间，江苏和广东的指标综合得分最高，其次是北京、山东、上海和浙江，其他城市的综合得分

较低。在 2019 年青海、广东和福建综合得分较高，其次是北京、江苏、上海和浙江。在 2020 年广东综合得分最高，之后是江苏和北京，其次是上海、山东和浙江。

4.2 数字基础设施情况分析

数字基础设施是数字经济发展的坚实基础，也是数字技术应用的必要条件，对于数字经济水平的发展具有一定的支撑作用。在计算各指标权重得分时，2016—2020 年这一时间段数字基础设施下各指标的权重之和分别为：0.1493、0.1527、0.1562、0.2830 和 0.1492，从权重结果可以看出数字基础设施在数字经济发展中不占据主导地位，不是促进我国数字经济发展的中坚力量。

在数字基础设施一级指标下，2016、2017、2018 和 2020 年中终端基础设施所占权重大于移动端基础设施权重，2019 年移动端基础设施所占权重大于终端设施所占权重。在终端基础设施指标下，2016—2020 这 5 年间 IPV4 地址数所占权重都大于其余两者的权重；在 2016 年和 2017 年光缆线路长度所占权重大于互联网宽带接入端口所占权重，而 2018、2019 和 2020 这三年中互联网宽带接入端口所占权重大于光缆线路长度所占权重。在移动端基础设施指标下，2017 年移动电话用户所占权重大于电话普及率所占权重，其余年份则相反。

4.3 数字经济情况分析

为了直观清晰地表示我国各个时期、各个地区的数字经济产业发展水平的情况（包括趋势、分布等），本文依据各省 2016 年至 2020 年的指标综合得分，利用核密度估计法进行可视化，绘制全国、东部地区、中部地区和西部地区数字经济产业发展水平的核密度估计图。

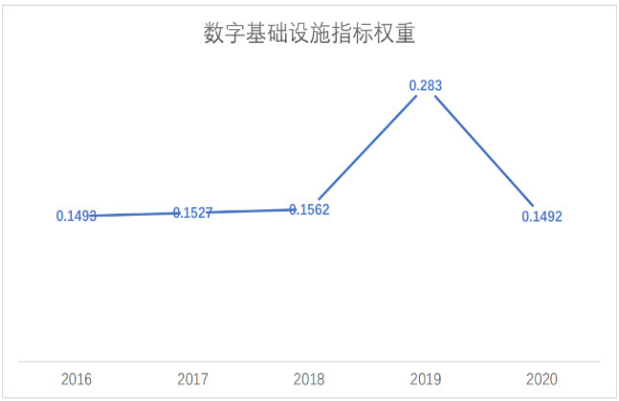

图 4 基础设施指标权重

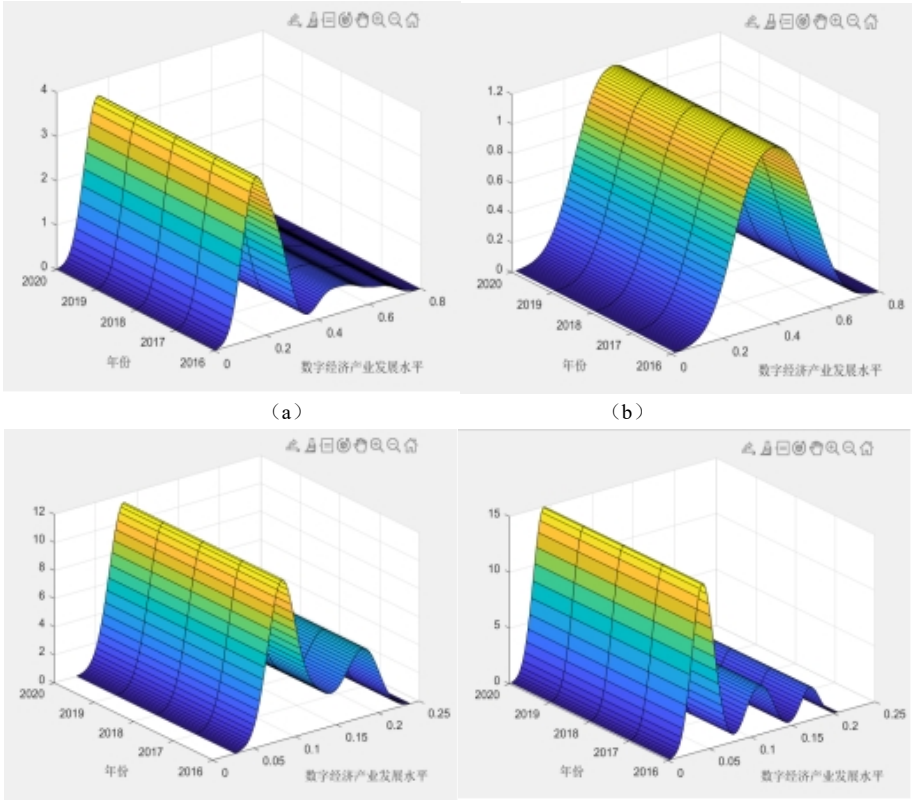

图 5 全国、东部、中部和西部指标得分核密度估计

图（a）表示全国核密度估计图，从中可以看出在样本观察期内核密度曲线保持着“单峰”的状态，峰宽的趋势为先缩小后增大的过程。但是总体峰宽有加大的趋势且分布右侧的拖尾显著，说明全国各个省份的数字经济产业发展水平的差距逐年被拉大。

图(b)表示我国东部地区核密度估计图，从中可以看出在样本观察期内核密度曲线“单峰”状态明显，峰宽始终保持着均匀分布。但是总体峰宽的分布左侧与右侧的拖尾程度接近相同，说明我国东部地区各省的数字经济产业发展水平的差距呈现平稳状态。

图（c）表示我国中部地区核密度估计图，从中可以看出在样本观察期内核密度曲线虽然“高峰”状态明显，但是同时伴随着一个“低峰”状态。峰宽的趋势为先增大后缩小的过程，右侧的“低峰”表示中部地区一些省份的数字经济产业发展水平正在提升，这就说明了中部地区某些省份的数字经济产业正在慢慢崛起。

图（d）表示我国西部地区核密度估计图，从中可以看出在样本观察期内核密度曲线虽然“高峰”状态明显，但是同时伴随着两个“低峰”状态。峰宽整体保持着稳定的分布，右侧的两个“低峰”表示西部地区一些省份拥有的内部高新产业正在慢慢增加，这就说明了西部地区的数字经济产业发展水平在逐年提升。

4.4 数字融合情况分析

在 2016—2020 这 5 年间数字经济融合总指标所占权重分别为：0.1214、0.1189、0.1234、0.0932 和 0.1173。在数字经济融合一级指标下，2016、2017、2018、2019 和 2020 年中企业应用所占权重大于个人应用权重，这表明企业应用在经济发展中起着关键的战略性支撑。在个人应用指标下，指标“互联网宽带接入用户”的权重大于指标“移动互联网用户”；在企业应用指标下，指标“有电子商务交易活动企业数”的权重大于“每百家企业拥有网站数”的权重，其中“有电子商务交易活动企业数”所占权重是三级指标中权重最高的。

在 2016 年全国范围内，有电子商务交易活动的总企业数为 102761 个，所占比重为 10.9%；2017 年全国企业数下降到 92122 个，所占比重为 9.5%，之后逐年递增至 2020 年 124552 个，所占比重为 11.1%，达到 5 年中最高比重。

4.5 数字经济发展水平梯队划分

通过上述分析，各个地区的指标综合得分是不尽相同的，在基础设施、数字经济产业和数字融合方面的发展程度也是不同的，因此本文进行 30 个省份的数字经济发展水平梯队划分，采用 AGNES 层次聚类算法进行相应的研究分析。利用 Python 软件对每一年的测度水平指标体系内所有指标采用 AGNES 层次聚类来进行梯队划分。

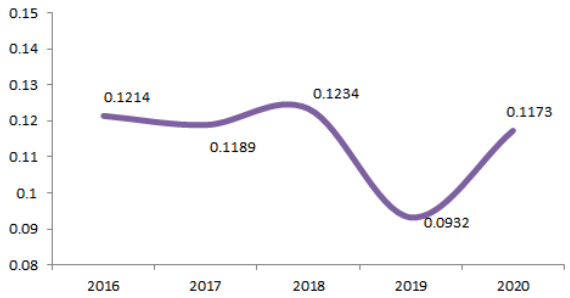

图 6 数字融合指标权重

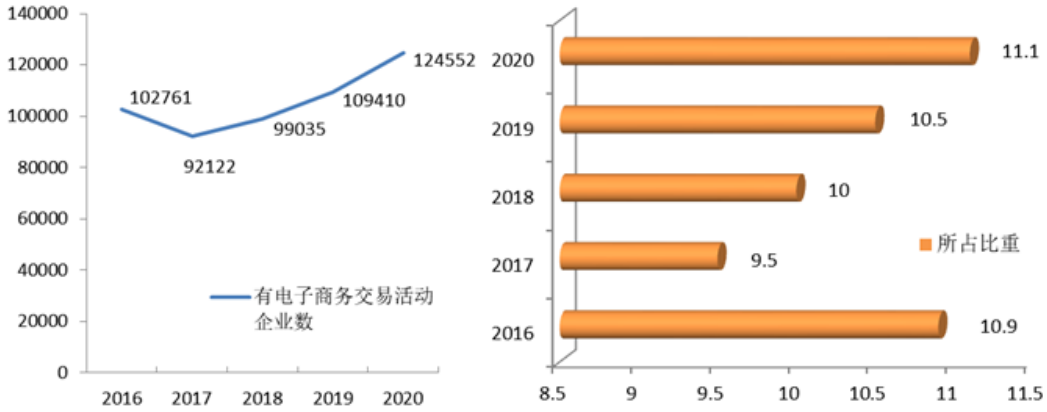

图 7 有电子商务交易活动企业数及所占比重

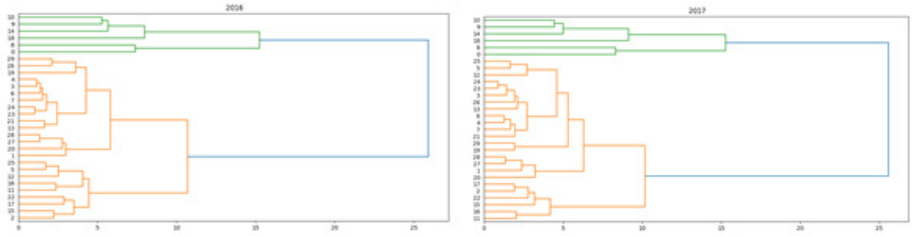

图 8　2016、2017 年层次聚类结果

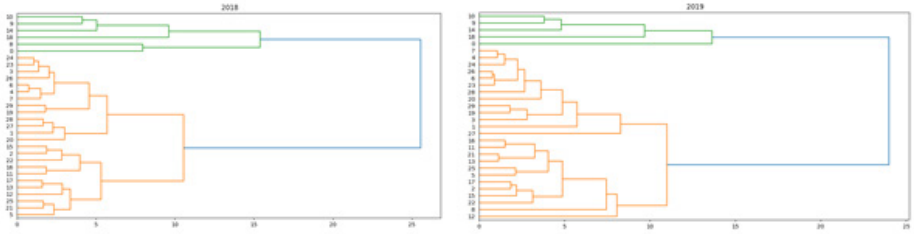

图 9　2018、2019 年层次聚类结果

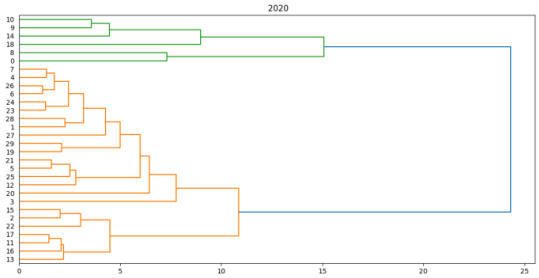

图 10　2020 年层次聚类结果

在 AGNES 层次聚类结果中，数字 0-29 分别表示 30 个省份，本文统一设定标准进行梯队划分，分析所得结果发现在 5 年结果中有大部分城市都被划定在同一梯队当中：

表 4　梯队划分

梯队一	江苏、浙江、山东、广东、北京
梯队二	吉林、黑龙江、山西、天津、广西、云南、海南、贵州、内蒙古、宁夏、青海、新疆、甘肃
梯队三	河北、安徽、湖南、河南、湖北、四川

本文根据聚类结果对 2016—2020 年 30 个省的数字经济发展水平类型进行了划分，将其分成了 3 个梯队。第一梯队里有浙江、江苏、山东、广东、上海五个城市，这 5 个城市的共同特征为都是沿海城市，城市中各个产业起步较早，资金链流通充足且数字经济随着社会的发展呈现上升趋势。

第二梯队的城市里面有黑龙江、吉林、山西、内蒙古、青海、天津、新疆、甘肃、广西、云南、海南、贵州和宁夏这 13 个城市，由于国家实施的西部战略计划，将各项方针政策、资源、人才引进和高新设备优先倾向于西部地区，从而使得西部地区的数字经济慢慢升高。

在第三梯队所属城市中，涵盖河北、安徽、湖南、河南、四川、湖北这 6 个城市，出现此情况的原因可能有：一部分人才的流失导致科研创新项目进度进行缓慢；部分政策落实得不是很到位。

但是，不乏会出现个别省份各年度划分情况变化的情形。重庆在 2016、2017 和 2020 年属于第二梯队，2018、2019 年属于第三梯队；陕西在 2016、2018、2019 和 2020 年被划分为第三梯队，2017 年被划分为了第二梯队；辽宁在 2016、2018 和 2019 年归属于第三梯队，2017、2020 年归属于第二梯队；

福建在 2016、2018 和 2019 年被归为了第三梯队，2017、2020 年被归为了第二梯队；江西在 2016、2017 年是第三梯队，2018、2019 和 2020 年是第二梯队；上海在 2016、2017、2018 和 2020 年为第一梯队，2019 年变为了第三梯队。

上述情况的发生可能是受到以下因素的影响：第一，各个省每年的数字经济衡量指标会做稍微的调整，从而导致梯队划分出现变化；第二，每个省的数字经济发展的侧重点不同，主要依据各省的地理位置及特色产业而决定；第三，部分城市会受到 2020 年疫情的冲击而导致数字经济发展水平的变化。

5 数字经济水平发展对 GDP 影响分析

5.1 数字经济水平与 GDP

近年来我国的数字经济产业不断发展，数字经济规模不断扩张，已经成为国民经济中核心内容之一。我国的数字经济规模由 2016 年的 22.4 万亿元扩张到 2020 年的 39.2 亿万元，同时数字经济占 GDP 的比重也在逐年提升，从 2016 年至 2020 年我国数字经济占 GDP 比重从 30.1%提升至 38.6%。

5.2 影响因素分析

5.2.1 相关性分析

通过上述分析可知数字经济规模在 GDP 中的比重日益增加，因此本文研究各地区的数字经济对该地区 GDP 的影响程度。利用上述构建的数字经济发展水平测度指标体系，采用灰色关联法分析每一年各个指标与地区 GDP 之间的关系。

本文选取 Spearman 秩相关系数计算公式，用此公式测定指标之间的相关性,利用 Python 软件计算各指标间的相关程度，并绘制可视化热力图：

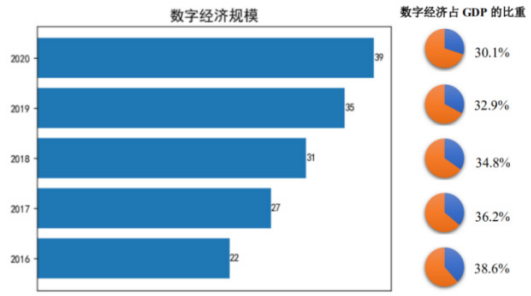

图 11 数字经济规模与 GDP

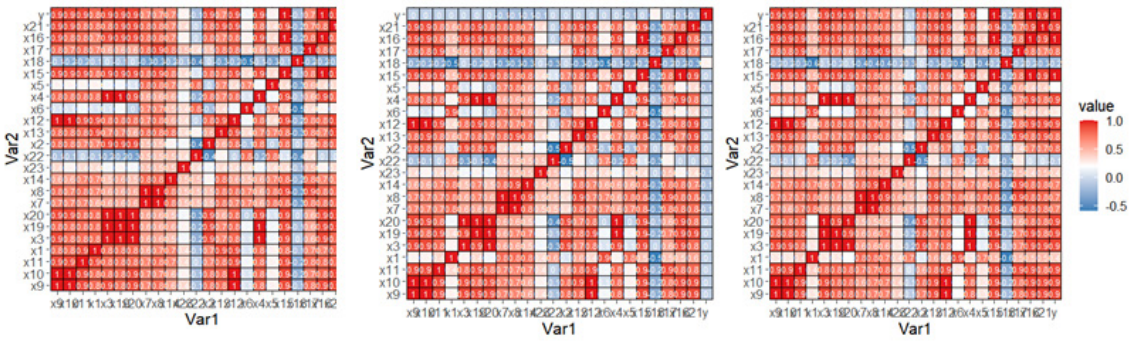

图 12 2016、2017、2018 年相关分析

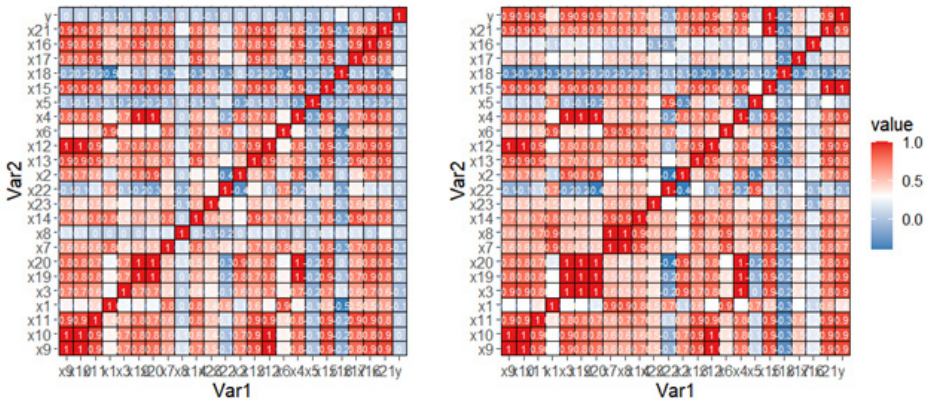

图 13 2019、2020 年相关分析

在可视化结果中用颜色变化表示变量间的相关程度，红色表示指标间相关性程度较高，浅色表示相关性程度较低，其中变量 x1-x23 表示数字经济水平测度指标体系内的各项指标，变量 y 表示地区生产总值，从相关性分析结果中可以看出部分指标与地区生产总值之间存在一定程度的相关性，因此研究数字经济发展水平对地区生产总值的影响是具有一定研究意义的。

5.2.2 XGBoost 模型求解

XGBoost 算法中可以得到每个指标的相应重要性得分。通常来说，一个指标在模型中建构决策树时应用频率越高，则可以认为它的重要性程度就相对较高。在单个决策树中，通过每个指标的分裂点来改进模型的性能度量以此来计算指标的重要性，节点用来确定加权和记录分裂次数。即一个指标对分裂点的改进性能越大，权重就越大。最后，将各个指标在全部提升树中所得的结果进行加权平均，从而得到重要性排序与得分。

本文选取 2016-2020 年这一时间段的各个指标以及地区生产总值的平均数作为样本集，利用 XGBoost 算法对数据集进行回归预测，并且得到各个指标的特征重要性：

表 5 特征重要性排名

排名	指标变量	特征重要性
1	信息技术服务收入	0.7277613
2	软件产品收入	0.24322948
3	R&D 人员全时当量	0.01571304
4	电子商务销售额	0.01329618

特征重要性结果表明在所有指标中，信息技术服务收入对地区生产总值具有显著性影响，而且影响程度较大；其次是软件产品收入这一指标，特征重要性分数为 0.24322948；最后是 R&D 人员全时当量和电子商务销售额这两个指标，对于地区生产总值也有一定程度的影响。

在筛选的具有显著影响的指标基础上，本文在选用 2016-2019 年的地区样本数据中进行 XGBoost 模型的训练，在训练好的模型上对 2020 年地区生产总值进行预测，将预测结果与真实结果进行对比：

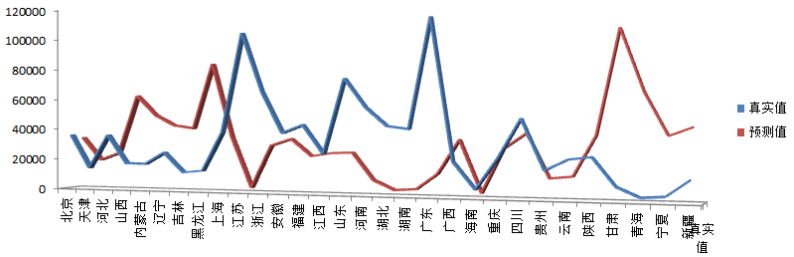

图 14 2020 年地区生产总值真实值、预测值对比

上图中线条颜色为蓝色的表示真实值，红色表示预测值，通过对比结果发现预测值与真实值之间存在较大差异，模型拟合效果欠佳。本文猜测造成此结果的原因可能是该数据涉及到时间这一维度，从而使得该模型的误差较大，因此本文采用 LSTM 模型对 2020 年地区生产总值进行预测。

5.3 LSTM 预测

本文首先对 2016—2019 年各省的数据进行划分，按照一定比例划分为训练集与测试集，在训练集上进行模型的训练与构建，然后用测试集来验证该模型，用 RMSE 来衡量模型误差，通过不断调整相应参数进行迭代，最终选取精度最高的相应参数来进行模型的最终训练：

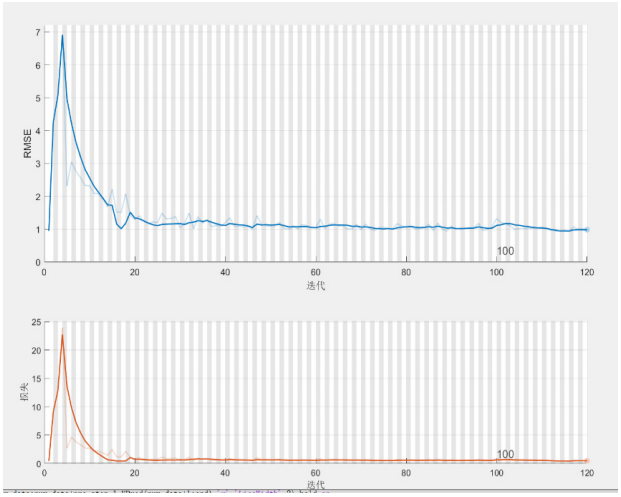

图 15 LSTM 模型训练

选用最佳参数对 2021 年地区生产总值进行预测，对比结果如下：

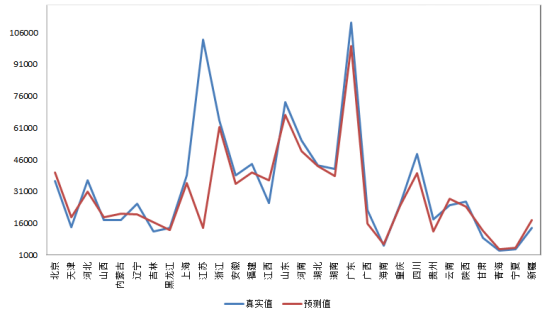

图 16 LSTM 模型预测值与真实值对比

预测结果表明大部分省市的预测值与真实值之间差异较小，有个别城市的预测值与真实值之间差异较大，总体增长趋势基本保持一致，表明该模型的拟合效果良好，可以用该模型对 2021 年各省的地区生产总值进行预测，并与真实值进行对比（数据真实值结果来源于智研咨询网整理所得）：

表 6 2021 年各省地区生产总值预测结果对比

省份	真实值	预测值
北京	40269.60	40245.70
天津	15695.05	15680.82
河北	40391.30	40379.90
山西	22590.16	22568.36
内蒙古	20514.20	20498.56
辽宁	27584.10	27570.35
吉林	—	12378.52
黑龙江	14879.20	14885.67
上海	43214.85	43210.20
江苏	116364.20	116385.62
浙江	73516.00	73522.39
安徽	42959.20	429611.23
福建	48810.36	48821.31
江西	29619.70	29615.80
山东	83095.90	83090.54
河南	58887.41	58892.56
湖北	50012.94	50012.22
湖南	46063.09	46065.26
广东	124369.67	124370.58
广西	24740.86	24738.92
海南	6475.20	6480.25
重庆	27894.02	27898.04
四川	53850.79	53851.88
贵州	19586.42	19583.78
云南	27146.76	27145.88
陕西	29800.98	29811.02
甘肃	10243.30	10243.96
青海	3346.63	3345.67
宁夏	4522.31	4523.43
新疆	16000.00	16005.61

6 结论与建议

6.1 结论

本文基于统计年鉴 2016 至 2020 年的数据，并且查阅相关文献构建数字经济发展水平测度指标体系，依据熵权 TOPSIS 法确定各项指标的相应权重并得到各省份（除西藏、港澳台）综合得分，利用 XGBOOST 算法进行特征重要性分析，并且采用 LSTM 模型对 2021 年各省地区生产总值进行预测。综上分析得到如下结论：第一，在数字经济发展中，科技创新占据较大比重，它在经济发展进程中起着关键作用；第二，通过对比全国、东部、中部和西部地区的核密度估计图，发现西部地区的数字经济发展水平在逐年提升；第三，通过对地区生产总值的预测，发现我国的数字经济发展水平在稳步提高。

6.2 建议

虽然我国经济在 2020 年受新冠疫情影响出现了波动和下滑，但在政府的积极应对和各方共同努力下，经济已逐渐恢复，为提高我国数字经济发展水平，基于此提出以下建议：

建立完整的数字经济机制体系，对数字经济发展的地区分布、重点项目尽量地落实到位，避免出现举措同质化的现象，要实现全方位、多样化的发展；

虽然我国的数字经济比大多数国家具有领先优势，但是也要持续改善数字经济的运营环境，加快构建有助于数字经济发展的高效机制；

要不断地推进数字经济发展的创新趋势，与时俱进地发挥数字重要性技术在数字经济中的创新作用。

参考文献

[1] 李震.数字经济赋能新发展格局：理论基础、挑战和应对[J]. 社会科学, 2022(03):43-53.

[2] Zhou Xiaoyong, Zhou Dequn, Zhao Zengyao, Wang Qunwei. A framework to analyze carbon impacts of digital economy: The case of China[J]. Sustainable Production and Consumption, 2022 (prepublish).

[3] 王军,朱杰,罗茜.中国数字经济发展水平及演变测度[J].数量经济技术经济研究,2021,38(07):26-42.

[4] 邝劲松,石校菲,杨祎,黄灿灿.中国省域数字经济发展水

平测度与空间演变格局研究[J].商学研究, 2022, 29(01): 94-102.

[5] 焦帅涛,孙秋碧.我国数字经济发展测度及其影响因素研究[J].调研世界,2021(07):13-23.

[6] 吴云. 数字经济可持续发展水平的测度研究[D].重庆邮电大学,2021.

[7] 陶长琪,徐茉.经济高质量发展视阈下中国创新要素配置水平的测度[J].数量经济技术经济研究,2021,38(03):3-22.

[8] Li Xiangyin,Liang Xueping,Yu Ting,Ruan Sijia,Fan Rui. Research on the Integration of Cultural Tourism Industry Driven by Digital Economy in the Context of COVID-19—Based on the Data of 31 Chinese Provinces[J]. Frontiers in Public Health,2022,10.

[9] 朱明爽. 中国数字经济规模统计测度：理论与方法[D].山东财经大学,2021.

[10] Wang Xueyang, Sun Xiumei, Zhang Haotian, Xue Chaokai. Digital Economy Development and Urban Green Innovation CA-Pability: Based on Panel Data of 274 Prefecture-Level Cities in China[J]. Sustainability, 2022, 14(5).

[11] Haita WANG, Xuhua HU, Najabat ALI. Spatial Characteristics and Driving Factors Toward the Digital Economy: Evidence from Prefecture-Level Cities in China[J]. The Journal of Asian Finance, Economics and Business (JAFEB),2022,9(2).

版权声明：©2023 作者与开放获取期刊研究中心（OAJRC）所有。本文章按照知识共享署名许可条款发表。
http://creativecommons.org/licenses/by/4.0/

OPEN ACCESS