Advances in Applied Mathematics 应用数学进展, 2023, 12(6), 2915-2923
Published Online June 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.126293

Hans汉斯

基于数据处理的江西省数字经济规模预测研究

刘星星[1*]，张延飞[1#]，颜七笙[2]，丁木华[1]，陈　萍[1]

[1]东华理工大学理学院，江西　南昌
[2]东华理工大学抚州师范学院，江西　抚州

收稿日期：2023年5月26日；录用日期：2023年6月21日；发布日期：2023年6月28日

摘　要

数字经济规模是衡量数字经济发展水平的量化指标。在参考和梳理数字经济规模影响因素基础上，收集整理江西省2011~2020年的指标数据，通过对样本缺失数据采用线性回归进行月份多重插补扩充，采用相关性分析及多重共线性分析实现冗余指标删除，通过主成分分析实现数据降维，同时构建基于数据处理的PCA-BPNN机器学习预测模型及对比模型，对江西省数字经济规模预测进行实证分析。实验结果表明：PCA-BPNN模型预测结果MAE、MAPE分别为240.0181、0.0233，预测精度相较于构建的对比模型均提高了30%以上，最高达73.8%，论证了基于数据处理的机器学习预测模型的有效性以及准确性。该方法为区域科学制定数字经济发展战略具有重要的理论与现实价值。

关键词

相关性分析，主成分分析，BP神经网络，数字经济规模预测

Research on Jiangxi Province's Digital Economy Scale Prediction Based on Data Processing

Xingxing Liu[1*], Yanfei Zhang[1#], Qisheng Yan[2], Muhua Ding[1], Ping Chen[1]

[1]School of Science, East China University of Technology, Nanchang Jiangxi
[2]Fuzhou Normal College, East China University of Technology, Fuzhou Jiangxi

Received: May 26th, 2023; accepted: Jun. 21st, 2023; published: Jun. 28th, 2023

Abstract

The scale of the digital economy is a quantitative indicator to measure the level of development of

[*]第一作者。
[#]通讯作者。

文章引用: 刘星星, 张延飞, 颜七笙, 丁木华, 陈萍. 基于数据处理的江西省数字经济规模预测研究[J]. 应用数学进展, 2023, 12(6): 2915-2923. DOI: 10.12677/aam.2023.126293

the digital economy. On the basis of referring to and sorting out the factors affecting the scale of the digital economy, collect and sort out the indicator data of Jiangxi Province from 2011 to 2020, use linear regression for monthly multiple imputation expansion of sample missing data, use correlation analysis and multi-collinearity analysis to delete redundant indicators, use principal component analysis to achieve data dimensionality reduction, and build a PCA-BPNN machine learning prediction model and comparison model based on data processing. Conduct empirical analysis on the prediction of the scale of digital economy in Jiangxi Province. The experimental results show that the MAE and MAPE predicted by the PCA-BPNN model are 240.0181 and 0.0233, respectively. Compared with the constructed comparative model, the prediction accuracy has been improved by more than 30%, with a maximum of 73.8%. This demonstrates the effectiveness and accuracy of the machine learning prediction model based on data processing. This method has important theoretical and practical value for regional science in formulating digital economy development strategies.

Keywords

Correlation Analysis, Principal Component Analysis, BP Neural Network, Prediction of Digital Economy Scale

1. 引言

数字经济[1]是人类通过大数据的处理，实现资源的快速优化配置与再生、实现经济高质量发展的经济形态。近年来，随着数字经济在全球经济发展中的地位不断提升，数字经济的发展水平和规模预测等问题受到国际组织、政府机构和国内外学术界的高度关注。国际上主要形成了美国 BEA 数字经济统计测算、经合组织(OECD)数字经济统计指标、欧盟数字经济与社会指数(DESI)等测量指标和方法。我国目前最主流的数字经济增加值测算方式为中国信通院数字经济核算方法。还有余丽等[2]给出了区域数字经济发展水平影响因素及定性评价模型，对构建数字经济规模预测模型影响指标选取具有参考价值；李栋等[3]综合运用相关性分析、主成分分析等多种统计方法对收集数据进行处理，并构建了中国数字经济规模预测模型；李英杰等[4]从数字基础设施等方面构建了数字经济发展水平量化指标体系，分别基于测度法和灰色预测模型对 2019~2028 数字经济发展走向进行预测；鲜祖德[5]根据《数据经济及其核心产业统计分类(2021)》构造了数字经济测算框架，测算并预测了我国数字经济核心产业规模。

以上研究在围绕数字经济规模测算时主要集中于生产法、支出法、回归模型和增长核算框架四大类，目前国际上对数字经济的概念认知和测度宏观层面仍缺乏统一标准。在指标体系构建上，由于没有一致的规定，不利于测算数字经济规模和增速，多集中于定性研究，定量研究较少，难以准确量化评估数字经济发展水平；在测算方法上，各个国家和区域来源于对数字经济概念及核算范围理解的偏差，对数字经济总量的测算存在或多或少的高估或低估。根据数字经济发展水平和数字经济规模呈正相关[6]，本研究参考和梳理了数字经济规模的影响因素[7]，选取具有代表性的相关指标并完成数据收集，对样本缺失数据采用线性回归进行月份多重插补扩充，同时进行多重共线性诊断以降低数据扩充的影响，对筛选后的影响指标进行主成分分析以实现数据降维[8]，同时构建基于数据处理的机器学习预测模型及对比模型；利用 BP 神经网络对 2018~2022 年江西数字经济规模预测进行实证分析[9]，以期预判江西未来数字经济发展趋势，对客观且全面地认识江西数字经济发展的现状及发展前景具有重要的现实意义[10] [11] [12]。

2. 研究方法

2.1. 相关性分析

在数字经济规模预测前，需要先确定数字经济的影响因素，这种影响因素(自变量)与数字经济规模(因变量)的关联度判断可以用相关性分析方法。假设从众多影响因素中随机抽取一个为自变量 X 与因变量 Y 组成变量组，这组变量之间的相关性可用 Pearson 相关系数表示：

$$\rho_{X,Y}=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E\left[\left(X-E(X)\right)\left(Y-E(Y)\right)\right]}{\sigma_X\sigma_Y}$$

其中，$\rho_{X,Y}$ 是 Pearson 相关系数，cov 表示 X 与 Y 之间的协方差，δ_X、δ_Y 分别表示 X、Y 的标准差，E 表示数学期望。$\rho_{X,Y}$ 是反映两个变量 X 与 Y 的相关性强弱，介于−1 到 1 之间，$\left|\rho_{X,Y}\right|$ 越接近于 1 表明相关性就越强；反之，$\left|\rho_{X,Y}\right|$ 越接近于 0 则表明相关性就越弱。

2.2. 主成分分析

通过相关性分析对筛选后的影响因素有时需要数据降维处理。主成分分析(Principal Component Analysis, PCA)就是在“信息量”损失较少情况下，将高维数据转换为低维数据，降低预测模型的复杂度。主成分分析的原理是将原来变量通过线性组合重新组合成一组新的、相互无关的几个综合变量，这些新的综合变量就是成分，再根据成分的方差从大到小进行排序，按实际需要从中取出几个成分尽可能多地反映原来变量信息构成主成分，预测模型的构建可依据主成分作为模型的输入。主成分分析基本步骤如下：

① 原始数据的标准化。确定 P 个数字经济规模影响指标，构成 P 维随机向量 $x=\left(x_1,x_2,\cdots,x_p\right)^{\mathrm{T}}$，其中 n 个不同年份样本数据可表示为 $x_i=\left(x_{i1},x_{i2},\cdots,x_{ip}\right)^{\mathrm{T}}$，$i=1,2,\cdots,n$，$n>p$，构造样本矩阵 $X=\left(x_1^{\mathrm{T}},x_2^{\mathrm{T}},\cdots,x_n^{\mathrm{T}}\right)^{\mathrm{T}}$；对样本数据进行标准化变换得到标准化矩阵 Z，并计算得到相关系数矩阵 R。

② 确定主成分个数。求解样本相关系数矩阵 R 的特征方程得到 P 个特征值。按累计贡献率大于 85% 确定 m 值，即 $\sum_{j=1}^{m}\lambda_j\geq 0.85\sum_{j=1}^{p}\lambda_j$。其中 λ_i 为第 i 个主成分的方差贡献率；根据主成分总方差解释结果，选取涵盖影响指标绝大部分信息的主成分。

2.3. BP 神经网络

以降维后的数据作为机器学习预测模型的输入层，可构建 BP 神经网络(Back Propagation Neural Network, BPNN)预测模型来预测数字经济规模。BP 算法利用损失函数，每次向损失函数负梯度方向移动，直到损失函数取得最小值。反向传播算法是根据损失函数求出损失函数关于每一层的权值及偏置项的偏导数，用该值更新初始的权值和偏置项，一直更新到损失函数取得最小值或是设置的迭代次数完成为止。以此来计算神经网络中的最佳的参数。

以典型三层 BP 神经网络为例分为输入层、隐含层、输出层，具有如下结构：

如图 1 所示，$X=\left(x_1,x_2,\cdots,x_i,\cdots,x_n\right)^{\mathrm{T}}$ 是网络层的输入向量，隐含层第 K 个节点和输入层之间的权向量为 $V_k=\left(v_{k1},v_{k2},\cdots,v_{kn}\right)^{\mathrm{T}}$。隐含层的输入为 $T=\left(t_1,t_2,\cdots,t_q\right)^{\mathrm{T}}$，输出层的输入为 $Z=\left(z_1,z_2,\cdots,z_q\right)^{\mathrm{T}}$，$W_j=\left(w_{j1},w_{j2},\cdots,w_{jq}\right)^{\mathrm{T}}$ 为隐含层到第 j 个节点的权值向量，$S=\left(s_1,s_2,\cdots,s_m\right)^{\mathrm{T}}$ 为输出层的输入，网络的输出向量为 $\hat{Y}=\left(\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_j,\cdots,\hat{y}_m\right)^{\mathrm{T}}$。期望输出向量为 $Y=\left(y_1,y_2,\cdots,y_j,\cdots,y_m\right)^{\mathrm{T}}$。损失函数的期望值 $E=\frac{1}{2}\sum_{i=1}^{m}\left(y_i-\hat{y}_i\right)^2$。

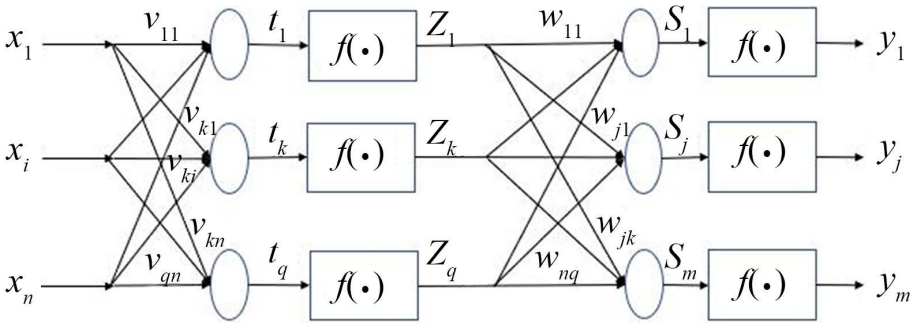

Figure 1. Schematic diagram of a three-layer neural network
图 1. 三层神经网络示意图

2.4. 基于数据处理的数字经济规模预测模型

基于上述研究方法构建了数字经济规模预测模型，并与其它预测模型作对比分析，进行模型评价。数据处理及模型构建流程如图 2 所示。

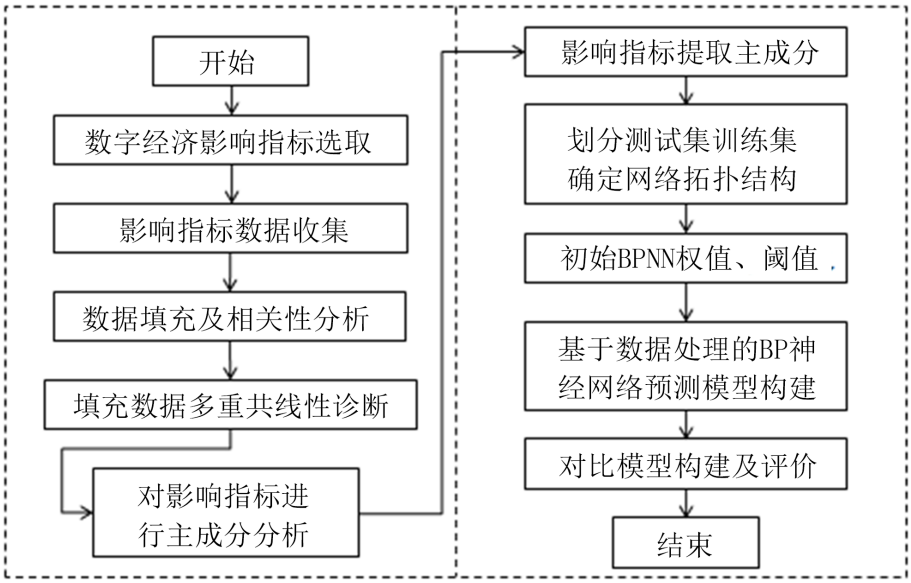

Figure 2. Data processing flowchart of the digital economy scale prediction model
图 2. 数字经济规模预测模型的数据处理流程图

3. 实证分析

3.1. 数据来源

数据收集

数字经济发展水平和数字经济规模呈正相关，通过查阅相关文献并进行分析总结数字经济发展水平的影响因素，论文从基础设施、数字金融产业、数字交易以及知识产权四个维度选取江西省数字经济规模的影响指标，见表 1。

本文收集了 2011~2020 年共计 10 组数据，其中数据主要来源于《中国统计年鉴》《中国第三产业统计年鉴》《中国信息产业年鉴》《江西省统计年鉴》、北京大学数学普惠金融指数、国家统计局等。

考虑到收集数据部分指标存在部分年份缺失的情况，对缺失数据采用线性回归方法进行月份多重插

补扩充，将影响指标的 2011~2020 年 10 组数据扩充到月份的 108 组数据。

Table 1. Impact Indicators of digital economy scale in Jiangxi Province
表 1. 江西省数字经济规模影响指标

一级指标	二级指标	单位	符号表示
基础设施	光缆长度	公里	x1
	移动电话基站数	万个	x2
	移动电话普及率	每百人部数	x3
	互联网宽带接入端口数	万个	x4
	互联网域名数	万个	x5
	互联网上网人数	万人	x6
数字金融产业	数字金融覆盖广度	-	x7
	数字金融使用深度	-	x8
	数字金融数字化程度	-	x9
	信息服务业从业人数	万人	x10
	规模以上工业企业 R&D 人员折合全时当量	人年	x11
	规模以上工业企业 R&D 经费支出	万元	x12
	规模以上工业企业 R&D 项目(课题)数	项	x13
数字交易	软件业收入	万元	x14
	网上移动支付水平	-	x15
	信息服务业产值	亿元	x16
	电信业务量	亿件	x17
知识产权	技术合同成交总额	万元	x18
	专利申请数	件	x19
	发明专利申请数	件	x20
	专利申请授权数	件	x21

3.2. 数据处理

3.2.1. 相关性分析

采用 SPSS(26.0)对上述指标扩充后的 108 组数据进行相关性分析，各指标与数字经济规模间的相关性热图见图 3，其中 x22 表示数字经济规模。

由图 3 第一行数据可知，各指标与数字经济规模的相关系数最小值为 0.847，最大值达 0.991，说明各指标均与数字经济规模的相关性极强。从相关性热图颜色辨识，各影响指标间存在多重共线性关系。用容差和方差膨胀因子进行多重共线性诊断分析，选取所有指标中的一个指标 x_i 为因变量，其他指标为自变量得到的线性回归模型的决定系数 Ri，计算容差 $TOL=1-Ri^2$，容差的倒数为指标的方差膨胀因子 VIF，对于容差小于 0.1 或者 VIF 大于 0，则表明有共线性存在。用 SPSS(26.0)得到排除影响指标的诊断结果，见表 2。

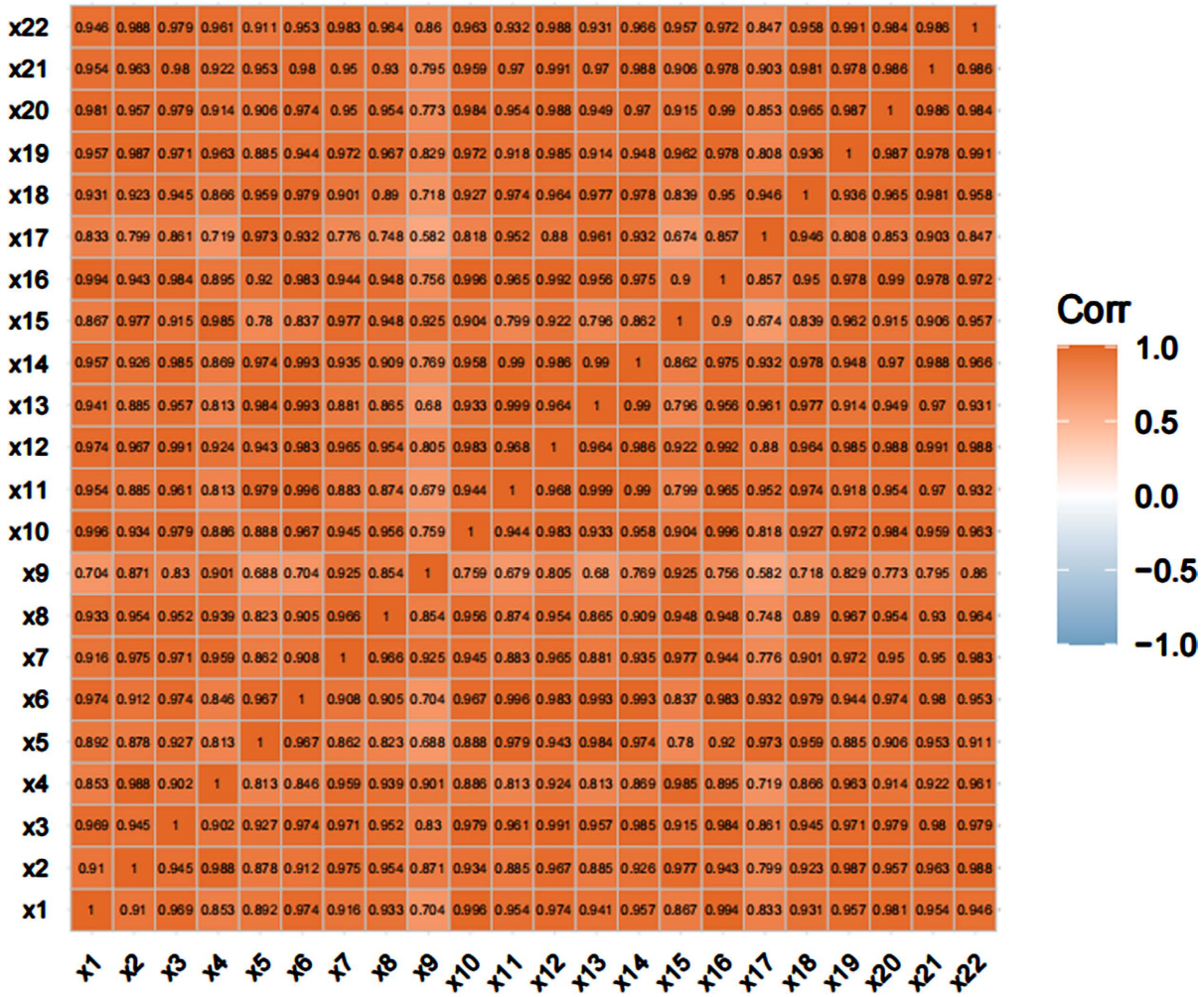

Figure 3. Correlation analysis heat map
图 3. 相关性分析热图

Table 2. Collinearity diagnosis results of excluded variables
表 2. 排除变量共线性诊断结果

排除变量	容差	VIF
移动电话基站数	3.521E−6	284010.22
移动电话普及率	3.726E−7	2683843.26
互联网上网人数	2.013E−8	49677098.96
软件业收入	4.578E−8	21843599.83
规模以上工业企业 R&D 经费支出	3.931E−10	2555583951
专利申请数	1.132E−7	8833922.26
发明专利申请数	2.019E−7	4952947.003
信息服务业从业人数	3.087E−6	323939.10
信息服务业产值	6.375E−8	14814814.81

通过多重共线性分析排除以上 9 个指标，选取光缆长度、互联网宽带接入端口数、互联网域名数、

数字金融覆盖广度、数字金融使用深度、数字金融数字化程度、规模以上工业企业 R&D 人员折合全时当量、规模以上工业企业 R&D 项目(课题)数、网上移动支付水平、电信业务量、技术合同成交总额、专利申请授权数等 12 个指标表征江西省数字经济规模。

3.2.2. 主成分分析

对排除后的 12 个指标进行主成分分析，KMO 统计量值为 0.745，大于 0.7，可看出变量间的相关程度无太大差异；巴特利特球形检验的结果小于 0.05，球形假设被拒绝，原始变量之间存在相关性，结果均表明上述数据适合进行主成分分析，得到样本数据主成分总方差解释，见表 3。

Table 3. Interpretation of principal component total variance
表 3. 主成分总方差解释

	初始特征值			提取载荷平方和		
成分	总计	方差百分比	累计%	总计	方差百分比	累计%
1	10.671	88.922	88.922	10.671	88.922	88.922
2	0.950	7.916	96.838			
3	0.207	1.721	98.559			
4	0.084	0.701	99.261			
5	0.043	0.352	99.619			
6	0.025	0.192	99.825			
7	0.013	0.109	99.930			
8	0.004	0.036	99.965			
9	0.004	0.032	99.995			
10	0	0.004	99.998			
11	0	0.001	99.999			
12	0	0.001	100			

根据主成分总方差解释，前四个主成分的累计方差贡献率超过 99%，这样提取前四个主成分(分别记作 F1、F2、F3、F4)可以涵盖选取 12 个指标的绝大部分信息，构建预测模型对数字经济规模进行预测。

3.2.3. 预测结果及分析

基于主成分分析提取的四个主成分 F1、F2、F3、F4 分别构建 BP 神经网络预测模型以及对比模型进行分析。设定滞后期数为 12，即用前 12 个月份的值对当前月份的数字经济规模进行预测，同时划分前 7 年的第 1~84 组数据为训练集训练神经网络，第 8 年的第 85~96 组数据为测试集评估神经网络训练效果，取主成分最后 12 组数据作为 BP 神经网络的输入得到输出即为预测结果。其中 BP 神经网络按照“4-9-1”的网络拓扑结构完成训练及预测，其中 4 为输入层神经元个数，9 为隐藏层神经元个数，1 为输出层神经元个数，网络结构如图 4 所示。

根据建立的基于数据处理的 BP 神经网络预测模型，对江西省 2011 年 12 月至 2020 年 12 月数字经济规模进行实证分析，为了进一步检验模型预测效果以及验证主成分分析实现数据降维的有效性，本文将基于主成分分析构建 PCA-BP 神经网络模型与 PCA-ARIMA 模型、PCA-MLP 模型、ARIMA 模型、MLP 模型、BP 神经网络模型做比较，其中 MLP 模型按照“4-9-1”的网络拓扑结构完成训练及预测，ARIMA 模型、MLP 模型、BP 神经网络模型使用的是经过数据填充以及多重共线性诊断筛选，但未经主成分分析进行数据降维的影响指标数据，对 2021 年度江西省数字经济规模进行预测，得到各模型预测结果(见图 5)。

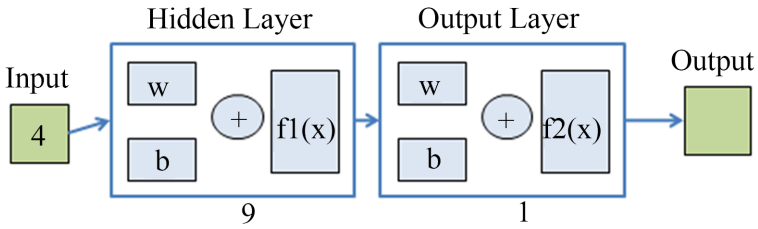

Figure 4. Topological structure of BP neural network
图 4. BP 神经网络拓扑结构

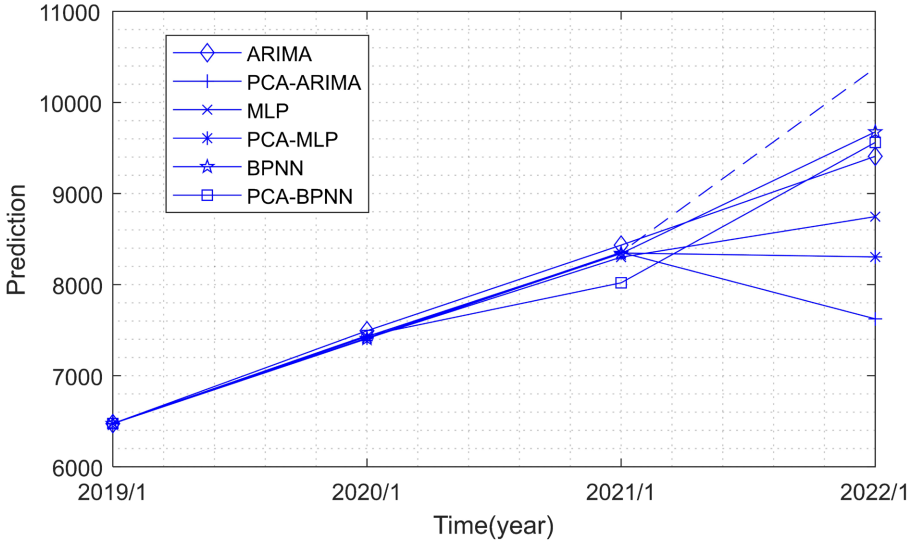

Figure 5. Comparison of prediction results of various models
图 5. 各模型预测结果对比

从图 5 中可以看出，2018~2020 年间除 BPNN 模型外其余模型预测结果与对应年份实际变化趋势基本吻合，2020~2021 年间不同模型预测结果与实际值变化趋势出现较大偏差，综合比较来说，PCA-BPNN 模型与实际值变化趋势基本吻合。同时通过计算得到各模型预测结果误差(见表 4)。

Table 4. Prediction results error of each model
表 4. 各模型预测结果误差

模型	ARIMA	PCA-ARIMA	MLP	PCA-MLP	BPNN	PCA-BPNN
MAE	918.4600	376.4163	699.3910	562.4627	392.8576	240.0181
MAPE	0.0885	0.0379	0.0677	0.0546	0.0408	0.0233

结合表 4 实验结果，分别比较 ARIMA 模型和 PCA-ARIMA 模型、MLP 模型和 PCA-MLP 模型、BPNN 模型和 PCA-BPNN 模型可知，经过主成分分析提取主成分构建的预测模型预测效果远优于未经数据降维构建的预测模型的预测效果，相较于 ARIMA、MLP、PCA-MLP 模型 MAE、MAPE 均提高了 50%以上，表明了主成分分析实现数据降维的有效性。综合比较来说，PCA-BPNN 预测效果最佳，进一步论证了 PCA-BPNN 预测模型的准确性及可行性。

4. 结论

① 相关性分析通过对主观选取因子的定量化分析，确定研究对象的影响指标。主成分分析是设法将

多影响因素构成的原始变量重新组合成一组新的、相互无关的几个综合变量，同时根据实际需要从中可以取出几个较少的总和变量尽可能多地反映原始变量的信息。本文初始选取 21 个数字经济规模的影响指标，通过筛选变为 12 个影响指标，再通过主成分分析提取出 4 个主成分，再分别构建基于数据处理的机器学习预测模型以及对比模型进行实证分析，验证了模型预测的准确性。

② 实证分析表明，当构建机器学习预测模型而样本数据过少时，可对相邻年份进行数据填充以实现数据扩充，变为大样本数据进行分析。同时对扩充数据进行多重共线性诊断以降低线性填充的影响，筛选得到 12 个影响指标并进行主成分分析以实现数据降维。在上述数据处理的基础上，本文分别构建了典型机器学习模型 MLP 模型、BPNN 模型以及传统预测模型 ARIMA 模型，实验结果表明，当样本数据足够多时，机器学习模型预测效果远优于传统预测模型，且文中给定机器学习模型中，基于数据处理的 PCA-BPNN 神经网络预测效果更佳。

③ 本研究在参考和梳理数字经济规模的影响因素，选取具有代表性的相关指标，对样本缺失数据进行月份多重插补扩充，同时进行多重共线性诊断以降低数据扩充的影响，对筛选后的影响指标进行数据降维，同时构建基于数据处理的 BPNN 预测模型，对定量测算数字经济规模和增速具有参考价值。

基金项目

国家自然科学基金项目(71961001)。

参考文献

[1] 李芃达. 数字经济发展动能强劲[J]. 中国外资, 2022(21): 64-65.

[2] 余丽, 冯瑶. 中国数字经济发展区域差异及影响因素分析[J]. 市场周刊, 2021, 34(3): 72-75.

[3] 李栋, 王珊, 任晓菲. 中国数字经济规模预测模型构建[J]. 统计与决策, 2022, 38(10): 5-9.

[4] 李英杰, 韩平. 中国数字经济发展综合评价与预测[J]. 统计与决策, 2022, 38(2): 90-94.

[5] 鲜祖德, 王天琪. 中国数字经济核心产业规模测算与预测[J]. 统计研究, 2022, 39(1): 4-14.

[6] 严武, 万良伟. 数字经济对实体企业金融化的影响机制研究[J]. 江西社会科学, 2022, 42(10): 44-53+206.

[7] 董心知. 中国数字经济发展水平综合评价及其对经济增长的影响研究[D]: [硕士学位论文]. 杭州: 浙江工商大学, 2022.

[8] 房汉国. 基于主成分分析法的宏观经济景气指数研究[J]. 当代经济, 2022, 39(1): 26-31.

[9] 赵鹏, 董倩. 基于改进的 PSO-BP 组合模型在经济预测中的应用——以河北省 GDP 为例[J]. 商展经济, 2022(1): 27-29.

[10] Tian, J.F. and Liu, Y.R. (2021) Research on Total Factor Productivity Measurement and Influencing Factors of Digital Economy Enterprises. *Procedia Computer Science*, **187**, 390-395. https://doi.org/10.1016/j.procs.2021.04.077

[11] Chohan, U.W. (2020) Some Precepts of the Digital Economy. *Critical Blockchain Research Initiative* (*CBRI*) *Working Papers*, 2020. https://doi.org/10.2139/ssrn.3512353

[12] Chinoracky, R. and Corejova, T. (2021) How to Evaluate the Digital Economy Scale and Potential? *Entrepreneurship and Sustainability Issues*, **8**, 536-552. https://doi.org/10.9770/jesi.2021.8.4(32)