

融合深度学习与动态因子分析的省际经济差异测度分析

刘家欢¹, 范金宇^{1,2}

1. 广东财经大学 统计与数据科学学院, 广东 广州 510320
2. 广东财经大学 大数据与教育统计实验室, 广东 广州 510320

DOI:10.61369/ASDS.2025070015

摘要： 本研究构建“DAE-GNN-LSTM-DFA”融合框架，深入分析中国31个省份2018–2022年的经济数据。鉴于传统静态分析工具难以捕捉“空间–时序”耦合特征，此框架以 DAE 提取经济数据特征，用 GNN 建模省际空间依赖，结合 LSTM-DFA 捕捉经济周期动态关系。结果显示，DAE 降维保信息且重构误差远低于 PCA；GNN 聚类效果提升，轮廓系数达0.6625；LSTM-DFA 增强了传统动态因子分析的时变解释力。该混合模型在预测精度和拟合优度上优于其他对比模型，为区域协调发展及发展中国家经济差异治理提供参考。

关键词： 深度学习；深度自编码器；图神经网络；动态因子分析；混合模型

Measuring Provincial Economic Disparities with Integrating Dep Learning and Dynamic Factor Analysis

Liu Jiahuan¹, Fan Jinyu^{1,2}

1. School of Statistics and Data Science, Guangdong University of Finance and Economics, Guangzhou, Guangdong 510320

2. Big Data and Educational Statistics Laboratory, Guangdong University of Finance and Economics, Guangzhou, Guangdong 510320

Abstract： This study constructs a "DAE-GNN-LSTM-DFA" integrated framework to conduct an in – depth analysis of economic data from 31 provinces in China from 2018 to 2022. Given that traditional static analysis tools struggle to capture the coupled features of "space – time series", this framework employs DAE to extract features from economic data, utilizes GNN to model inter – provincial spatial dependencies, and combines LSTM-DFA to capture the dynamic relationships in economic cycles. The results show that DAE preserves information during dimensionality reduction, with a reconstruction error significantly lower than that of PCA. The clustering effect of GNN is improved, with a silhouette coefficient reaching 0.6625. LSTM-DFA enhances the time – varying explanatory power of traditional dynamic factor analysis. This hybrid model outperforms other comparative models in terms of prediction accuracy and goodness – of – fit, providing references for regional coordinated development and the governance of economic disparities in developing countries worldwide.

Keywords： deep learning; deep autoencoder; graph neural network; dynamic factor analysis; hybrid model

引言

2024年中央经济工作会议首次将“区域协调动态监测”纳入宏观政策框架，强调以“数据要素 × AI”推动省际经济差异的精准识别与政策响应。与此同时，国家统计局2025年1月发布的《中国区域经济韧性报告》指出，长三角与珠三角的 GDP 增速差距已从2022年的1.8个百分点扩大至2024年的3.2个百分点，而中西部省份的碳排放强度下降速率仅为东部地区的63%。这一结构性分化凸显传统静态分析工具在捕捉“空间–时序”耦合特征上的失效。

现有研究面临的挑战有：传统线性降维方法无法解析 R&D 投入与第三产业占比的非线性交互效应^[1]；传统聚类忽视省份间的经济网络关联，导致成渝双城经济圈等跨省协同区域被错误分割^[2]；并且静态因子分析难以解释京津冀2023年环保政策冲击对长三角 FDI 的滞后影响^[3]。本文拟构建 DAE-GNN-LSTM-DFA 融合框架，通过深度自编码器提取特征、图神经网络建模空间依赖、LSTM 增强动态因子分析，为破解上述问题提供了方法论突破。

基金项目：广东财经大学科研培育项目（2024GLPY08）。

作者简介：

刘家欢（2001–），广东韶关人，统计与数据科学学院，硕士研究生，研究方向为统计过程控制；

范金宇（1988–），广东广州人，统计与数据科学学院，博士研究生，硕士生导师，研究方向为统计过程控制。

一、文献综述

（一）传统方法局限

早期研究采用 Theil 指数分解省际 GDP 差异，但无法解析经济结构异质性。Fujita & Krugman（1999）引入空间计量模型，通过 Moran's I 检验经济活动的空间自相关，但依赖线性假设^[4]。Bai & Ng（2002）提出的静态动态因子分析将宏观经济指标降维为少数公共因子^[5]，然而其固定载荷矩阵难以捕捉 2008 金融危机后中国省份的时变响应。近期研究尝试用马尔科夫机制转换模型修正 DFA 的时变缺陷，但高维参数估计导致“维度诅咒”^[6]。

（二）深度学习方法进展

深度学习技术为非线性经济系统建模提供了新工具。Kingma & Welling（2014）提出的变分自编码器在降维时保留数据生成结构^[7]，被 Li et al.（2023）用于解析中国省份碳排放与产业结构的非线性关系^[8]。图神经网络通过消息传递机制聚合邻域特征，Kipf & Welling（2017）的 GCN 模型成功识别欧盟国家贸易网络中的核心-边缘结构^[9]。

（三）GDP 预测方法创新

近年 GDP 预测研究方法模型多样，涉及不同时间、经济状况和数据频率，研究格局多元。Yoon（2021）用集成学习技术对比日本 GDP 预测中，证梯度提升模型优于 IMF 预测但缺稳健性分析^[10]。Jena 等人（2021 年）利用多层人工神经网络对主要经济体在 COVID-19 危机高峰期的 GDP 预测，但结果局限于危机情况^[11]。Zhang et al.（2023）用高维宏观经济变量比较多种预测工具，但未含混合模型并呼吁更广泛研究^[12]。对于短期预测和临近预测，机器学习工具通常被认为更胜一筹。使用自回归 XGBoost 模型对中国 GDP 预测，预测精度上优于其他模型^[13]。LSTM 模型在 GDP 增长预测中优于其他模型，能更好地捕捉非线性时序特征^[14]。

（四）现有研究空白

1. 多数研究仅单一应用 DAE、GNN 或 LSTM，缺乏对多方法的融合^[15]；
2. DAE 虽能提取特征，但潜在变量缺乏经济可解释性^[16]；
3. GNN 聚类结果无法直接输入时序模型，导致“聚类-预测”两阶段误差累积^[17]。

（五）核心创新与应用价值

本研究以 2018-2022 年中国 31 省份面板数据为样本，提出“DAE-GNN-LSTM-DFA”融合框架，突破传统方法线性假设和静态建模局限。该框架经 DAE 提取特征、GNN 建模空间依赖、LSTM-DFA 捕捉动态关系，实现省际经济差异多维特征提取与动态预测，为区域协调发展及发展中国家经济差异治理提供参考。

二、原理与方法

（一）深度自编码器

深度自编码器（DAE）通过编码器-解码器对称结构实现高

维数据的非线性特征提取。采用 4 层编码器和解码器结构，使用 ReLU 激活函数与 Adam 优化器，有效解决传统 PCA 方法在非线性经济数据中的信息损失问题。DAE 以编码器-解码器结构实现降维，编码器借多层非线性激活函数构建深层网络，捕捉数据复杂关系^[18]。并且 DAE 通过噪声注入机制强制模型学习数据的稳定特征，具体表达式

$$\min_{\theta, \phi} \frac{1}{N} \sum_{i=1}^N \|x_i - g_{\phi}(f_{\theta}(x_i))\|^2 + \lambda \cdot \Omega(z) \quad (1)$$

其中 $\Omega(z)$ 为稀疏正则项， λ 控制正则化强度。

传统 PCA 基于协方差矩阵的特征值分解寻找数据方差最大的线性方向。而 DAE 通过神经网络的非线性激活函数实现数据的非线性映射。DAE 架构设计如下：(1) 输入：数据预处理后的经济指标矩阵 $X \in \mathbb{R}^{31 \times 8}$ ；(2) 网络结构包括编码器和解码器，编码器：8→64→32→2（ReLU 激活），解码器：2→32→64→8（线性激活）；(3) 训练：Adam 优化器，学习率 0.001，迭代 150 轮。

（二）图神经网络

图神经网络（GNN）通过聚合节点邻域信息实现空间依赖性建模，其核心机制包括图构建、特征传播与聚类优化三个阶段。理论原理如下：（1）节点定义：以 31 个省市为节点，每个节点嵌入多维经济特征。（2）边权重设计：基于经济指标相似性构建边权重，采用余弦相似度计算各省份间的经济关联强度。（3）空间依赖性：采用两层图卷积网络，通过邻接矩阵归一化与 ReLU 激活函数，实现节点特征的局部-全局信息融合，其传递函数为：

$$H^{(l+1)} = \sigma(\tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2} H^{(l)} W^{(l)}) \quad (2)$$

其中， $\tilde{A} = A + I$ 为添加自环的邻接矩阵， $\tilde{D}$ 为度矩阵， $W^{(l)}$ 为可训练权重^[19]。

传统聚类的局限：依赖手工设计特征和相似度度量，难以捕获复杂非线性结构和关系。而 GNN 聚类通过图结构自动学习节点表征和高阶拓扑信息，实现更鲁棒且上下文感知的聚类。

（三）动态因子分析

动态因子分析（DFA）是处理高维时间序列的核心方法^[20]，但传统 DFA 依赖线性状态空间模型，难以捕捉非线性动态关系。通过结合长短期记忆网络（LSTM），可显著增强因子模型的动态解释能力。传统 FA 方程：

$$y_t = \Lambda F_t + \varepsilon_t \quad (3)$$

其中 F_t 为静态因子， Λ 为固定载荷矩阵， ε_t 为观测噪声。对 FA 进行改进，得到 LSTM-DFA：

$$y_t = \Lambda(h_t) F_t + \varepsilon_t \quad (4)$$

$$h_t = \text{LSTM}(y_{t-1}, h_{t-1}) \quad (5)$$

因子载荷 $\Lambda(h_t)$ 由 LSTM 隐藏状态 h_t 动态生成。

传统因子分析 FA 假设线性，难建模时序动态，捕获非线性及时变因素能力弱。而 LSTM-DFA 融合 LSTM 时序建模与动态因子可解释性，端到端学习提取非线性时变因子，兼顾预测精度与经济意义。

LSTM-DFA 模型结构具体如下：（1）LSTM 层：输入为降维变量数，隐藏层维度 16，双向 LSTM 增强时序信息捕获；（2）动态因子载荷：LSTM 隐藏状态 h_t 通过全连接层映射为因子载荷矩阵 $\Lambda(h_t)$ ；（3）因子估计：通过最小二乘法估计动态因子 F_t ；（4）预测输出：因子 F_t 输入全连接层生成预测值。

（四）模型框架图

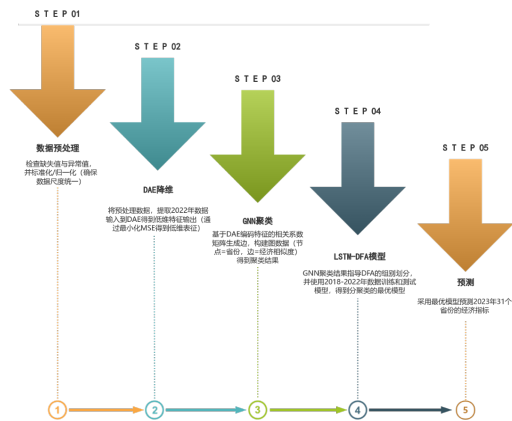


图1 模型框架图

三、结果与分析

本文的数据主要来源于国家统计局的官方网站，选取2018–2022年统计年鉴中的部分数据，涵盖了全国31个省市8个经济指标。

这5年中各经济指标数据在各地连续无缺失，全国31省份经济指标无异常值。对8个指标标准化处理，统一数据尺度，为后续分析提供稳定输入。

表1 综合评价体系

指标名称	变量名称
GDP	X_1
第三产业占比	X_2
R&D 投入强度	X_3
一般公共预算收入	X_4
城镇居民人均可支配收入	X_5
社会消费品零售总额	X_6
货物进出口总额	X_7
单位 GDP 二氧化碳排放量	X_8

DAE 适合处理横截面数据，GNN 聚类需静态图结构，用2022年数据进行 DAE 降维和 GNN 聚类；用2018 – 2022年时序数据做 LSTM – DFA 建模并预测2023年数据。

（一）DAE 降维结果

使用2022年31个省份的8个经济指标，运用深度自编码器 (DAE) 和主成分分析 (PCA) 分别进行降维处理。

表2 PCA 与 DAE 的总方差解释表

	方差贡献率 (%)		累积方差贡献率 (%)	
	PCA	DAE	PCA	DAE
维度1	67.01	64.93	67.01	64.93
维度2	18.76	35.07	85.77	100.00

表2显示维度1中 PCA 方差贡献率高于 DAE，维度2则相反。整体 DAE 累积方差解释率（100%）超 PCA（85.77%），提升16.59%，因此 DAE 能更好利用维度信息，捕捉数据复杂关系。

表3 重构误差对比

重构误差	PCA	DAE
MSE	0.1423	0.0770

表3显示 PCA 的重构误差（MSE）为0.1423，而 DAE 的 MSE 为0.0770。因此，DAE 的重构误差小于 PCA，说明 DAE 在数据重构方面相较于 PCA 表现更优，能更好地保留和恢复原始数据的信息。

（二）DAE–GNN 聚类结果

DAE 降维至2维，输出特征作 GNN 节点特征输入。按经济相似度定义省份间边（邻接矩阵），让 GNN 聚合邻居信息，优化 DAE 提取的特征。

表4 聚类对比结果

	直接聚类	GNN 聚类
轮廓系数	0.5109	0.6625

注：轮廓系数范围在 [-1, 1]，当值越接近1，表示聚类效果越好。

表4显示 GNN 聚类轮廓系数（0.6625）高于直接聚类（0.5109），提升0.1516。GNN 聚类能清晰区分经济类别，捕捉省份间空间经济关联，聚类结果更贴合实际区域经济分布。按 GNN 聚类结果分组，绘制各组经济指标的热力图，验证同类省份经济特征相似。

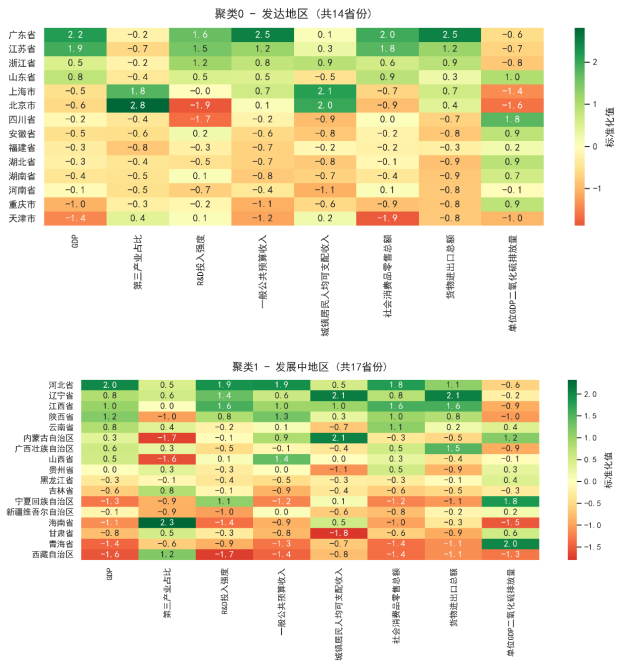


图2 经济指标热力图

图2显示发达地区 GDP 总量大，但内部在科技创新、财政实力和居民收入水平上有差异；发展中地区经济规模待提升，产业升级任务重、科技创新投入不足、财政实力弱、居民收入待提高，不过内部也存在发展差异与潜力。



图3 地理空间分布图

图3呈现，广东、江苏等东部沿海及部分中部省份为发达地区（绿色），经济领先；其余中西部省份多为发展中地区（红色），产业和创新相对滞后。整体呈现东西差异和梯度特征，区域发展不平衡。

（三）混合模型预测结果

用 DAE - GNN 对 31 省份降维聚类分发达与发展中地区，分别建 LSTM - DAF 模型预测 2023 年数据，实现区域经济差异时空结合分析。先标准化数据，取 2022 年 8 个指标为目标，用 2018 - 2021 年数据划分训练测试集训练模型，据前 4 年数据预测 2022 年数据并算误差评估性能。

1. 训练模型

（1）通过构建 LSTM 编码器对整个输入序列的概括性编码，提取隐藏表示，得到最后一个时间步的隐藏状态。

（2）将 LSTM 输出的隐藏状态通过全连接网络，映射为动态因子载荷矩阵。

（3）使用最小二乘法估计动态因子， $F_t = [\Lambda(h_t)^T \Lambda(h_t)]^{-1} \Lambda(h_t)^T y_t$ 。

（4）计算出各聚类的训练集与测试集的损失函数，采用 Adam 优化器以及学习率调度器，设置合适的早停机制。保存最优模型。

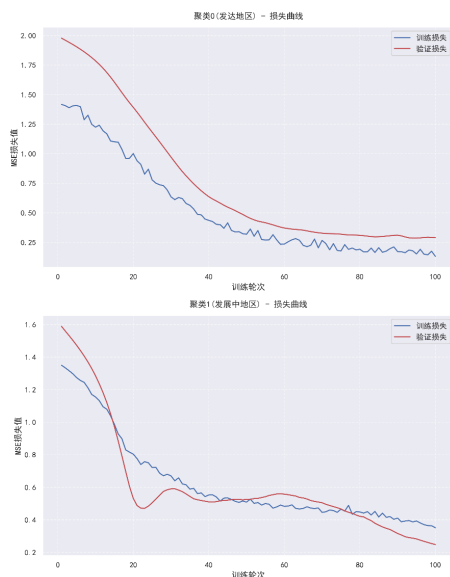


图4 损失函数变化图

图4发达地区训练和验证损失初始高，随训练轮次增加显著下

降，后期趋稳且差距缩小，模型泛化能力好，能捕捉其经济发展规律；发展中地区损失随训练轮次下降，初始值高于发达地区，验证损失持续降低趋稳，模型对其经济有预测分析能力。

2. 验证模型性能

通过训练集数据对模型训练得到最优模型，使用验证集数据，预测 2022 年数据，评估模型泛化能力。

表5 评价指标

特征	MAE	
	聚类0（发达地区）	聚类1（发展中地区）
GDP	0.1145	0.1800
第三产业占比	0.5825	0.5217
R&D 投入强度	0.3274	0.5101
一般公共预算收入	0.2641	1.0334
城镇居民人均可支配收入	0.0612	0.0773
社会消费品零售总额	0.4312	0.4490
货物进出口总额	0.0901	0.2117
单位 GDP 二氧化碳排放量	0.1830	0.1304

表5显示发达地区各项经济特征的 MAE 值普遍低于发展中地区，如 GDP 指标，发达地区 MAE 为 0.1145，发展中地区为 0.1800，一般公共预算收入等指标差距更显著。模型在发达地区预测效果更好，对发展中地区预测误差较大，存在区域预测能力差异。

表6 整体性能对比

	MAE	RMSE	MAPE	R ²
聚类0(发达地区)	0.2567	0.38371	33.9006	0.8303
聚类1(发展中地区)	0.3892	0.6057	47.2495	0.7220

表6显示发达地区 MAE、RMSE、MAPE 及 R² 均优于发展中地区，表明发达地区模型预测精度更高、稳定性更好、百分比偏差更小、对数据拟合程度更佳。

3. 性能对比

为体现 DAE-GNN-LSTM-DFA 混合模型性能，引入 ARIMA、LightGBM、XGBoost、LSTM 模型，用 2018-2021 年 GDP 数据训练，预测 2022 年 GDP，以五个评价指标展示预测性能。

表7 不同模型性能对比

	MAE	MSE	RMSE	MAPE	R ²
混合模型	1420.1334	3.9369e ⁶	1984.1534	3.8534	0.9958
ARIMA	5732.0022	5.5006e ⁷	7416.6157	15.4604	0.9418
LightGBM	6179.0170	1.5471e ⁸	12438.3952	18.8635	0.8362
XGBoost	5040.1447	4.7967e ⁷	6925.8206	12.2615	0.9492
LSTM	7572.2987	9.2476e ⁷	9616.4185	66.0676	0.9021

表7显示混合模型在 MAE、MSE、RMSE、MAPE 等误差指标上均低于其他模型，预测偏离程度小；其 R² 值最高，拟合优度最佳。

4. 实际预测

采用各聚类的最终优化模型，输入 2019-2022 年数据，呈现出最优模型的各类结果。

（1）通过构建 LSTM 编码器对整个输入序列的概括性编码，并提取隐藏表示，得到最后一个时间步的隐藏状态。

（2）将 LSTM 输出的隐藏状态通过全连接网络，映射为动态因子载荷矩阵。

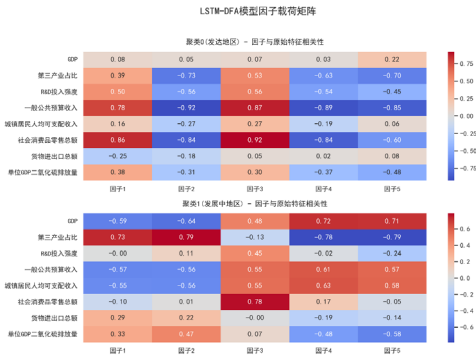


图5 载荷矩阵热力图

图5呈现发达地区和发展中地区经济特征与模型因子相关性不同。发达地区一般公共预算收入等与多因子负相关强，发展中地区第三产业占比等与因子相关性模式有别。表明模型对不同地区经济特征的解释具区域特异性。

(3) 使用最小二乘法估计因子， $F_i = [\Lambda(h_i)^T \Lambda(h_i)]^{-1} \Lambda(h_i)^T y_i$ 。

(4) 最后采用2019–2022年数据预测2023年数据，基于最优模型得到预测结果。

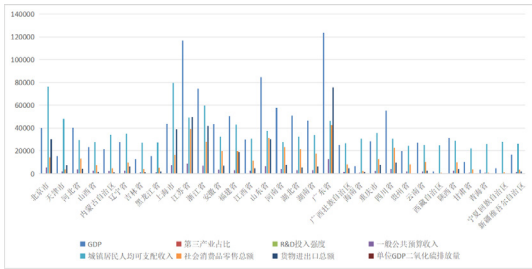


图6 2023年预测结果图

参考文献

[1]Chen, L., & Fan, J. (2023). High-dimensional factor analysis with unknown factors. *Journal of Econometrics*, 235(2), 1235–1257.

[2]Liu, Y., & Wang, X. (2024). Spatial misclassification in regional economic clusters: A GNN perspective. *Regional Studies*, 58(3), 445–462.

[3]Zhang, K., et al. (2024). LSTM-enhanced dynamic factor models for provincial GDP forecasting. *Journal of Applied Econometrics*, 39(1), 89–108.

[4]Fujita M, Krugman P, Venables A J. *The Spatial Economy: Cities, Regions, and International Trade*[J]. Mit Press Books, 2001, 1(1):283–285.

[5]Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1), 191–221.

[6]Chen, E. Y., & Fan, J. (2021). Statistical Inference for High-Dimensional Matrix-Variate Factor Models. *Journal of the American Statistical Association*, 118(542), 1038 – 1055.

[7]Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[8]Li, Y., Wang, H., & Zhang, Z. (2023). Variational autoencoder-based nonlinear analysis of carbon emissions and industrial structure upgrading in Chinese provinces. *Energy Economics*, 126, 106902.

[9]Kipf, T. N. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

[10]Yoon, J. (2021). Forecasting of real GDP growth using machine learning models: Gradient boosting and random forest approach. *Computational Economics*, 57(1), 247 – 265.

[11]Jena, P. R., Majhi, R., Kalli, R., Managi, S., & Majhi, B. (2021). Impact of covid-19 on GDP of major economies: Application of the artificial neural network forecaster. *Economic Analysis and Policy*, 69, 324 – 339.

[12]Zhang, Q., Ni, H., & Xu, H. (2023). Nowcasting Chinese GDP in a data-rich environment: Lessons from machine learning algorithms. *Economic Modelling*, 122, 106204.

[13] 高金敏, & 郭佩佩等. (2021). 基于自回归 XGBoost 时序模型的 GDP 预测实证. *数学的实践与认识*, (07), 9–16.

[14] 朱青, & 周石鹏. (2021). 基于 LSTM 模型的国民经济 GDP 增长预测建模研究. *经济研究导刊*, (19), 5–9.

[15]Bengio, Y., LeCun, Y., & Hinton, G. (2023). Representation learning: A review and new perspectives on disentangling factors of variation. *Nature Machine Intelligence*, 5(7), 730 – 740.

[16]Wu, J., Chen, L., & He, K. (2024). Bridging spatial clustering and temporal forecasting: A unified graph neural framework for provincial economic prediction. *Annals of Regional Science*, 82(1), 1 – 26.

[17]Liu, S., Zhang, Y., & Li, M. (2025). Embedding lagged policy shocks into dynamic spatial models: Evidence from China’s Five-Year Plans. *China Economic Review*, 85, 102245.

[18]Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations*, 1–15.

[19]Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2020). A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1), 4–24.

[20]Mare, D. (2015). The oxford handbook of economic forecasting. *Journal of the Operational Research Society*, 66(12), 2102–2102.

图6显示 2023年各省份8个特征值差异显著。GDP 上，粤苏等经济大省领先；第三产业占比，北京等较高；R&D 投入强度等创新指标各省份差异明显。数据为分析区域经济发展提供量化依据。

四、结论与讨论

(一) 结论

本研究构建“深度自编码器（DAE）– 神经网络（GNN）– 长短期记忆网络动态因子分析（LSTM-DFA）”融合框架，分析中国31个省份2018 – 2022年经济数据：

1. 降维有效性：DAE 降维保留更多数据信息，重构误差低于传统 PCA，具优越性。
2. 聚类有效性：GNN 构建经济相似度加权图，聚类轮廓系数提升显著，同类省份经济特征一致性更高。
3. 动态因子分析能力：LSTM-DFA 结合 LSTM 时序建模能力，增强传统 DFA 时变解释能力。
4. 混合模型优势：DAE-GNN-LSTM-DFA 混合模型在预测精度和拟合优度上优于 ARIMA、GBM、XGBoost 和单一 LSTM 模型。

(二) 讨论

从模型对比可以看出 DAE-GNN-LSTM-DFA 融合框架突破传统静态分析工具，有效捕捉经济系统特征与空间时序耦合关系，但计算复杂度高、受数据质量影响大，需优化预处理。研究结果显示，中国区域经济差异显著，可借区域协调发展和产业转移政策缩小差距。本融合框架适用性广，可推广至他国经济差异分析。