Modeling and Simulation 建模与仿真, 2025, 14(6), 304-318
Published Online June 2025 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2025.146499

Hans汉斯

基于数字足迹与深度学习的城市旅游流量预测

潘嘉贝，朱　莹*，莫书迪

南京邮电大学理学院，江苏　南京

收稿日期：2025年5月25日；录用日期：2025年6月18日；发布日期：2025年6月25日

摘　要

随着智慧旅游的发展，如何实现对城市旅游流量的精准预测成为热点课题。本研究以杭州市为案例，结合2021~2024年百度指数关键词搜索数据与过夜旅客量数据，构建了一套基于数字足迹与深度学习的旅游预测体系。通过Spearman相关系数与随机森林算法，对关键词进行筛选与排序，选出8个与游客量高度相关的搜索词。在此基础上，构建BP神经网络与多特征LSTM神经网络模型，分别对游客量进行预测，并与传统ARIMA模型进行对比分析。研究结果显示，LSTM模型在拟合精度与波动趋势捕捉方面表现最优，MAPE最低为0.099，优于其他模型，验证了融合数字搜索数据与深度学习算法的有效性。研究为智慧旅游背景下的游客行为理解与资源调度优化提供了方法参考与实践支撑。

关键词

随机森林，旅游流量预测，多特征LSTM神经网络，BP神经网络，百度指数

Urban Tourism Flow Forecasting Based on Digital Footprints and Deep Learning

Jiabei Pan, Ying Zhu*, Shudi Mo

College of Science, Nanjing University of Posts and Telecommunications, Nanjing Jiangsu

Received: May 25th, 2025; accepted: Jun. 18th, 2025; published: Jun. 25th, 2025

Abstract

With the advancement of smart tourism, accurate forecasting of urban tourist flows has become a critical research topic. This study takes Hangzhou as a case city and constructs a tourism flow prediction framework based on digital footprints and deep learning by integrating Baidu Index keyword search data and overnight tourist volume data from 2021 to 2024. Using Spearman correlation analysis and random forest algorithms, eight keywords most closely related to tourist volume

*通讯作者。

文章引用: 潘嘉贝, 朱莹, 莫书迪. 基于数字足迹与深度学习的城市旅游流量预测[J]. 建模与仿真, 2025, 14(6): 304-318. DOI: 10.12677/mos.2025.146499

were selected. Based on these, a BP neural network and a multi-feature LSTM neural network model were developed to predict tourist flows, and their performance was compared with a traditional ARIMA model. Results show that the LSTM model achieved the best performance in terms of fitting accuracy and trend capturing, with the lowest MAPE of 0.099, outperforming the other models. This study demonstrates the effectiveness of integrating digital search data and deep learning algorithms, and provides methodological and practical support for understanding tourist behavior and optimizing resource allocation under the smart tourism paradigm.

Keywords

Random Forest, Tourism Flow Forecasting, Multi-Feature LSTM Neural Network, BP Neural Network, Baidu Index

1. 引言

近年来，我国旅游市场重新焕发生机，尤其在节假日高峰期，热门城市和景区出现“人满为患”现象，暴露出旅游接待能力与游客数量不匹配的问题，进一步凸显了对旅游流量精准预测的现实需求。与此同时，以人工智能为代表的数字技术不断赋能旅游产业，推动“智慧旅游”成为热点。网络搜索引擎作为游客决策的重要入口，沉淀了大量可挖掘的“数字足迹”，如关键词搜索量、需求图谱等，具备良好的前瞻性与代表性，为旅游需求预测提供了新的数据基础和研究视角。

传统的 ARIMA 等时间序列模型在旅游量预测中虽有一定优势，但难以捕捉非线性动态变化[1] [2]。近年来，BP 神经网络、LSTM 等深度学习方法凭借强大的拟合能力，逐步应用于旅游预测研究中[3]-[6]。与此同时，百度指数等网络搜索数据由于具备“前兆效应”[7]，可有效反映公众出行意向，被认为是增强模型预测精度的重要外部变量。已有研究表明，网络搜索数据与旅游流量存在显著相关性[8] [9]；但关键词选择标准不统一、特征建模不足、模型评估不系统等问题依然存在。尤其是在特大城市如杭州，其旅游需求受季节、节假日、网络热点等多因素交互影响，急需构建多源融合、结构清晰的智能预测体系。

基于此，本文以杭州市为例，提出一种融合网络搜索指数与人工智能模型的旅游流量预测框架。首先，采集 2021~2024 年杭州过夜旅客量及百度指数多维关键词搜索数据；其次，结合 Spearman 秩相关系数与随机森林算法，筛选出 8 个最具预测价值的关键词特征；在此基础上，分别构建基于 BP 神经网络、LSTM 神经网络的旅游预测模型，并与传统 ARIMA 模型进行误差对比，验证融合模型的优越性。本文的研究目标不仅在于提升旅游预测模型的精度与稳定性，更希望为智慧旅游管理者在高峰时段的接待规划、资源调度、游客分流等方面提供科学决策支持。

2. 网络搜索关键词的选择

百度指数搜索趋势研究是百度指数最基本的功能[10]。用户可以搜寻任意关键词在相关时间范围内的搜索趋势，对其搜索量的变化情况进行了解，帮助用户了解某个话题、产品或事件的热度变化。如图 1 为以“杭州旅游”为关键词进行百度指数趋势可视化，展示从 2021 年 1 月 1 日到 2024 年 5 月 5 日间杭州旅游的搜索热度。

为了确保研究结果的准确性，必须进行全面拓展和进一步优化核心关键词，从而涵盖可能影响因变

量变化的一切信息。信息覆盖不全面或核心关键词数量过少会影响研究结果。关键词选择的步骤依次为：选定初始关键词、拓展关键词、选定最终关键词以保证数据的精准性和完整性[8]。

Figure 1. Daily search trend of the keyword “Hangzhou travel” on Baidu Index
图 1. “杭州旅行”关键词百度指数日搜索量趋势

2.1. 基于 Spearman 系数的网络搜索关键词择优

在选取初始关键词时，将采用主观取词的方法从食住行游四个方面考虑，最终确定了以下八个初始关键词如表 1。

Table 1. Original keywords
表 1. 初始关键词

要素	初始关键词	初始关键词
食	杭州美食	杭州小吃
住	杭州住宿	杭州酒店
行	杭州地铁	杭州游玩路线
游	杭州旅游	杭州景点

接下来将基于八个初始关键词进行关键词拓展，由于百度需求图谱的相关时间跨度为一周，将取 2023 年 10 月~2024 年 5 月中“国庆假期”、“元旦”、“五一劳动假期”三个时间段中搜索热度前十的关键词作为拓展关键词并剔除重复关键词以及无关关键词，如表 2 所示。

Table 2. Expanded keywords
表 2. 拓展关键词

初始关键词	拓展关键词
杭州美食	西湖醋鱼、杭州特产、杭州小吃、杭州、杭州旅游攻略必去景点推荐、葱包烩、杭州旅游景点、定胜糕
杭州小吃	杭帮菜、杭州小吃街、青团、杭州小笼包、鸡蛋灌饼、片儿川
杭州住宿	杭州旅行社、杭州天气、武林广场、河坊街

续表

杭州酒店	杭州两日游、美团、杭州东站、浙江音乐学院、咸亨酒店、西湖国宾馆、清河坊、杭州酒店价格
杭州地铁	杭州地铁线路图、杭州地图、杭州地铁运营时间、杭州公交车、杭州自驾游、杭州出租车
杭州旅游	杭州、杭州旅游攻略三日游、西溪国家湿地公园、浙江旅游、杭州旅游景点、杭州茶文化、杭州旅游网
杭州游玩攻略	杭州市旅游攻略、杭州宋城、雷峰塔、杭州西湖旅游、杭州景点最好玩的排名
杭州景点	杭州西湖、杭州千岛湖、杭州乐园、灵隐寺门票、杭州海底世界、杭州灵隐寺、杭州动物园门票、杭州黄龙洞、浙江景点

本文使用 Spearman 相关分析计算关键词与游客量相关性，Spearman 相关分析通过将观察值转换为秩次，该方法不受数据分布影响，更适用于不同类型的数据。最终的相关系数值将指示两组数据之间的相关性程度，有助于深入理解关键词与游客量之间的关系。下面是 Spearman 相关分析的具体计算步骤：

假设第 i 天内南京市的游客量以及某个关键词的百度搜索指数为 $(x_i, y_i), i = 1, \cdots, n$，将所有的 x_i 和 y_i 分别进行由小到大的排序后取秩次 W_i 和 Q_i。则 Spearman 相关系数的计算公式如下所示：

$$r_s = 1 - \frac{6}{n\left(n^2 - 1\right)} \sum_{i=1}^{n} \left(W_i - Q_i\right)^2 \tag{1}$$

其中，表示 r_s 秩相关程度，取值范围在 $[-1,1]$，$|r_s|$ 越大相关程度越大；当 $r_s > 0$ 时为正相关，当 $r_s < 0$ 时为负相关，当 $r_s = 0$ 时不相关。将通过计算得到的 Spearman 相关系数筛选其中大于 0.77 的对应数值，确定最终关键词，相关计算结果见图 2 所示。

Figure 2. Heatmap of Spearman correlation coefficients for selected keywords
图 2. 部分关键词的 Spearman 相关系数的热力图

根据相关系数的计算原理，对所有数据进行计算，仅在此展示部分选定关键词(27 个)的相关系数见表 3。

2.2. 基于随机森林的关键词重要性估计

随机森林的特征选择算法能够为每个特征提供一个重要性得分，依据得分情况从原始特征空间中选

择出一些最有效的特征以降低数据特征维度，提高学习算法性能。最终计算 27 个关键词的指标重要性，选取最优的八个关键词作为最终的网络搜索关键词，参与到杭州市游客量的预测当中。

如图 3，在一个原始数据集中，进行有放回的随机抽样 n 次，每次抽取形成一个子样本集，对应着一个决策树分类器。最终会得到 n 棵决策树。当有一个输入样本时，每棵树都会对其进行分类，因此会得到 n 个分类结果。接着，计算每个特征在所有 n 棵树中的得分，以确定每棵树中最具影响力的特征。然后，通过随机森林集成所有决策树的投票结果，找出投票次数最多的特征，将其输出为关键特征[11]。

Table 3. Spearman correlation coefficients of partially selected keywords
表 3. 部分选定关键词的 Spearman 相关系数

关键词	相关系数	关键词	相关系数	关键词	相关系数
西溪国家湿地公园	0.773	杭州地铁线路图	0.797	杭州地铁运营时间	0.789
咸亨酒店	0.777	杭州地图	0.770	杭帮菜	0.791
浙江景点	0.779	杭州东站	0.814	杭州	0.773
西湖国宾馆	0.772	杭州旅游	0.755	杭州茶文化	0.783
雷峰塔	0.776	杭州公交车	0.776	杭州特产	0.807
杭州宋城	0.785	杭州海底世界	0.795	杭州西湖旅游	0.793
河坊街	0.801	杭州乐园	0.795	杭州地图	0.770
杭州小笼包	0.794	杭州酒店价格	0.854	杭州小吃街	0.772
杭州千岛湖	0.786	杭州酒店	0.786	杭州小吃	0.789

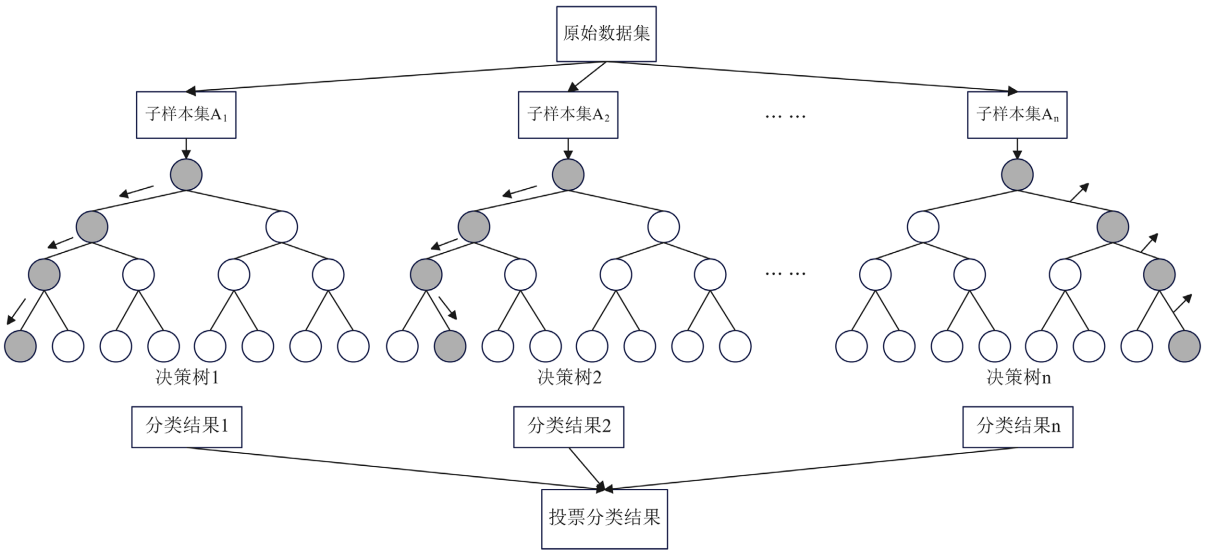

Figure 3. Schematic diagram of the random forest algorithm
图 3. 随机森林原理图

本研究中探究了杭州市过夜旅客量日数据与自变量之间可能存在的复杂关系，并且假设这种关系不仅仅是简单的线性关系。选择使用随机森林算法来进行特征选择，旨在从选择的关键词中挑选出最佳的特征变量，以构建游客量网络搜索指数。

具体而言，利用包括 27 个关键词的百度搜索指数以及杭州市的过夜旅客量日数据。通过随机森林

回归算法对这些数据进行分析，并根据结果选择了最为显著的特征变量，用于构建游客量网络搜索指数。最终得到以下的部分重要性得分表格数据(保留三位小数)并将所有特征变量的得分进行可视化展示如图 4。

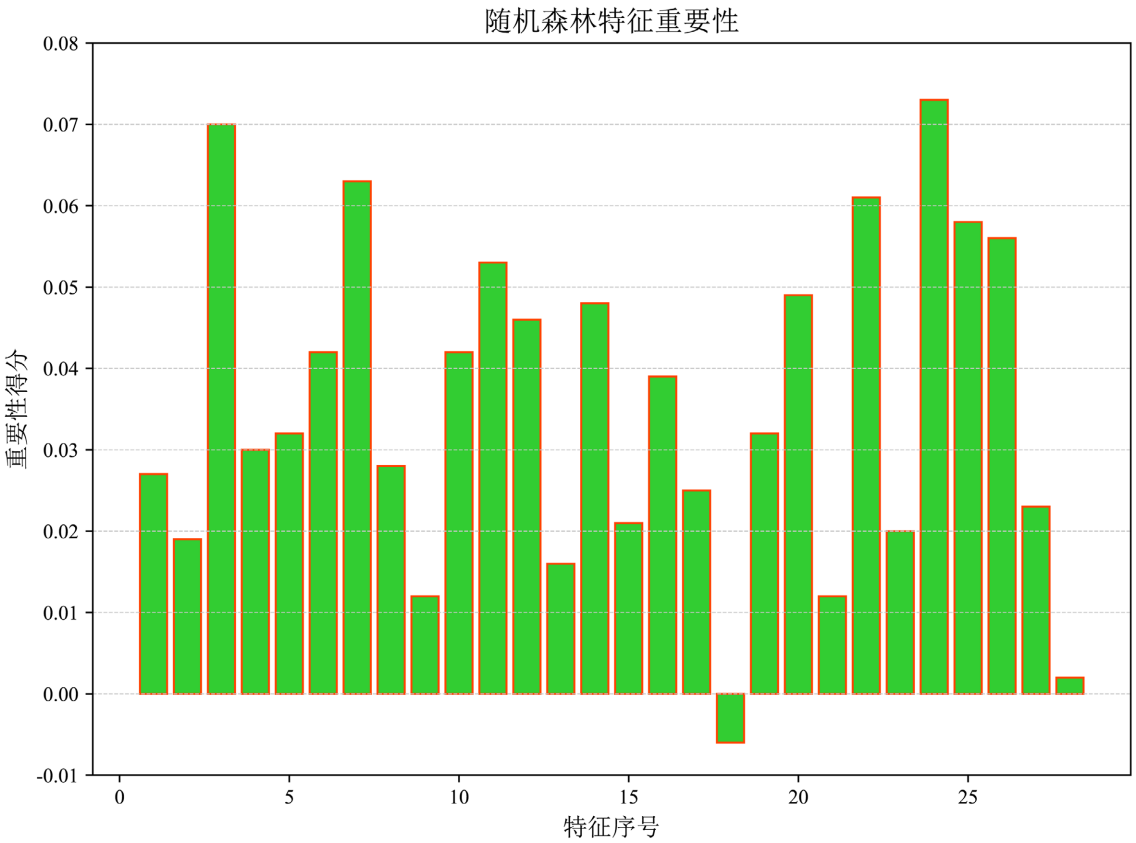

Figure 4. Feature importance scores of all keywords based on the random forest algorithm
图 4. 所有关键词的随机森林特征重要性得分

可以看出，正得分表示该特征对模型性能有正向影响，即增加该特征的重要性有助于提高模型的性能；而负得分表示该特征对模型性能有负向影响，即增加该特征的重要性会降低模型的性能。正负得分的存在反映了特征与目标变量之间的线性或非线性关系的方向。得分高的特征意味着它们对模型性能的贡献程度更大，而得分低的特征则对模型性能的影响较小。高得分特征通常具有更强的预测能力，可能与目标变量之间存在更为显著的关联关系，因此选取表格中的 8 个关键词为最终的网络搜索关键词参与到神经网络预测模型的建立当中(见表 4)。

Table 4. Importance scores of the final selected keywords
表 4. 最终选定关键词的重要性得分

特征(关键词)	Grade	特征(关键词)	Grade
西溪国家湿地公园	0.071	杭州宋城	0.060
杭州东站	0.0611	杭州小吃	0.039
杭州地铁线路图	0.074	杭州地铁运营时间	0.058
杭州特产	0.052	杭帮菜	0.057

3. 基于网络搜索数据的游客量预测模型

3.1. 游客量 BP 神经网络预测模型

3.1.1. BP 神经网络介绍

BP 神经网络(Back Propagation Neural Network)在传统的神经网络上加入了 BP 算法而形成的神经网络，其主要特点是输入的数据沿网络向前传播，而误差则反向传播。BP 神经网络可以用来解决分类、回归、模式识别、数据挖掘等多种问题，具有优良的非线性拟合能力以及强大的容错和逼近能力[5]。鉴于以上基本原理与特性，将首先选择使用 BP 神经网络构建基于网络搜索数据的游客量预测模型。

3.1.2. BP 神经网络的计算原理与步骤

假设输入向量 $X=(X_1,\cdots,X_n)$，节点个数为 n；隐藏层节点个数为 m；输出层向量为 $Y=(Y_1,\cdots,Y_m)$，节点个数为 m；期望输出向量 $Z=(Z_1,\cdots,Z_m)$。输入层与隐含层的连接权重是 W_{ij}，隐含层与输出层的连接权重是 W_{jk}，隐含层各神经元的阈值是 α_j，输出层各神经元的阈值 β_k。学习速率为 υ，激励函数为 $\varphi(x)$。根据原理图，其具体计算公式如下：

(1) 信号正向传播

隐含层第 j 个结点的输出：

$$y_j=\varphi(net_j)=\varphi\left(\sum_{i=1}^{n}W_{ij}x_i+\alpha_j\right) \tag{2}$$

输出层第 k 个结点的输出：

$$O_k=\psi(net_k)=\psi\left(\sum_{j=1}^{p}W_{jk}+\beta_k\right) \tag{3}$$

误差公式：

$$e=1/2*\sum_{k=1}^{m}(Z_k-Y_k)^2 \tag{4}$$

(2) 误差反向传播

BP 神经网络中将数据集分为训练样本与预测样本，训练样本的数量记为 p，p 个训练样本的总误差准则函数为：

$$e_p=1/2*\sum_{p=1}^{p}\sum_{k=1}^{m}(Z_k-Y_k)^2 \tag{5}$$

并根据以下的权重和阀值更新公式来更新参数值，经过反复地迭代最终损失函数收敛，此时的拟合效果是最佳的。

输出层权重的修正量：$\Delta W_{jk}=\upsilon\sum_{p=1}^{p}\sum_{k=1}^{m}\left(Z_k^p-Y_k^p\right)\psi'(net_k)y_j$ (6)

输出层阀值的修正量：$\Delta\beta_k=\upsilon\sum_{p=1}^{p}\sum_{k=1}^{m}\left(Z_k^p-Y_k^p\right)\psi'(net_k)$ (7)

隐含层权重的修正量：$\Delta\alpha_i=\upsilon\sum_{p=1}^{p}\sum_{k=1}^{m}\left(Z_k^p-Y_k^p\right)\psi'(net_k)W_{jk}\varphi'(net_j)$ (8)

隐含层阀值的修正量：$\Delta W_{ij}=\upsilon\sum_{p=1}^{p}\sum_{k=1}^{m}\left(Z_k^p-Y_k^p\right)\psi'(net_k)W_{jk}\psi'(net_j)x_i$ (9)

(3) 模型性能评价指标

为了更好的评价 BP 神经网络模型的效果，选取一些恰当的评价指标对模型进行效果对比。一般对绝对百分比误差(APE)取平均得到平均绝对误差百分比(MAPE)进行计算得分，该值可以更直接反映模型预测误差的大小，误差值越小表明该模型的预测精度越高[9] [10]。

$$\mathrm{MAPE}=\frac{1}{n}\sum_{t=1}^{n}\left|\frac{\hat{y}_t-y_t}{y_t}\right| \tag{10}$$

其中 $\hat{y}_t$ 和 y_t 分别是实际值和预测值。

3.1.3. BP 神经网络预测模型的实现

本文选择使用 MATLAB 软件的神经网络工具箱建立游客量 BP 神经网络预测模型。

在 BP 神经网络的结构设计中，输入层节点数为 8 (对应 8 个关键词特征)，输出层节点为 1 (游客量)，隐藏层节点数参数取值通过经验法则与网格搜索确定为 0.772。学习率设置为 0.001，最大训练轮次为 1000 轮，目标误差设置为 0.001。

(1) 读取数据，划分训练集和测试集，对数据进行归一化：将所有数据按照时间顺序排列并存放在 data 矩阵中，按照 9:1 的比例将数据划分为训练集和测试集；最后根据归一化公式(使用 mapminmax()函数)对训练样本以及测试样本进行归一化。

(2) 构建神经网络并进行训练：将使用神经网络工具箱的 newff()函数来初始化神经网络，并设置相应参数以优化模型性能。在生成初始化游客量 BP 神经网络预测模型以后，使用 train()函数基于现有的样本数据进行网络的训练。

(3) BP 神经网络预测以及预测结果反归一化：使用 sim()函数可以进行神经网络的仿真，并使用反归一化函数对预测结果进行反归一化。

(4) 计算误差并绘制对比图：根据 MAPE 的计算公式计算平均绝对误差百分比作为模型优劣的评价指标，并使用神经网络工具箱绘制数据图以及结果图。

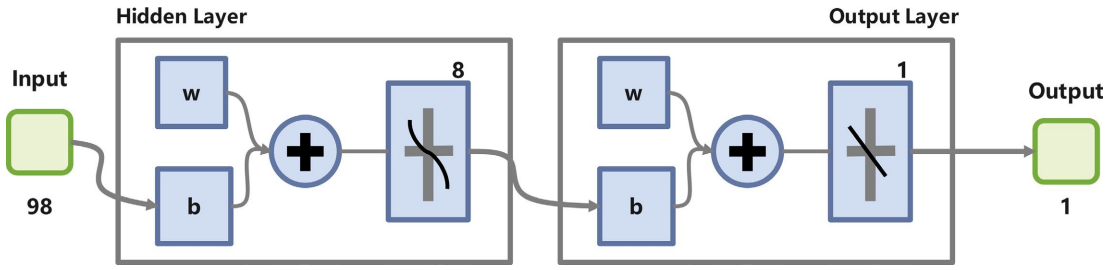

Figure 5. Training diagram of the BP neural network using Custom Neural Network
图 5. BP 神经网络训练 Custom Neural Network 图

如图 5 所示，采用 MATLAB 建模所得的神经网络结构图展示了该网络由具有 Sigmoid 激活函数的隐藏神经元和线性输出神经元构成的两层前馈结构。这种结构在隐藏层神经元数量充足的前提下，能够较好地拟合多维非线性映射问题。训练过程采用反向传播算法对网络权重和偏置进行迭代优化。

3.2. 游客量 LSTM 神经网络预测模型

3.2.1. LSTM 神经网络介绍

长短时记忆神经网络(LSTM)是一种改进的循环神经网络(RNN)，适用于处理序列数据。RNN 由于其时间序列数据处理的天然优势，在通过反向传播和梯度下降优化时能够校正错误。尽管如此，其在反向传播过程中面临梯度消失或梯度爆炸的挑战，此问题会随时间序列延长而加剧。LSTM 通过特别设计的结构，克服了这些困难，有效处理了长期依赖问题，从而改善了传统 RNN 的性能。LSTM 的反向传播算法特别适用于避免长期依赖问题，使其在许多应用场景中(如语音识别、图像描述、自然语言处理等)成为更受欢迎的 RNN 变体[2] [6]。RNN 神经网络和 LSTM 神经网络的结构示意图如图 6 所示。

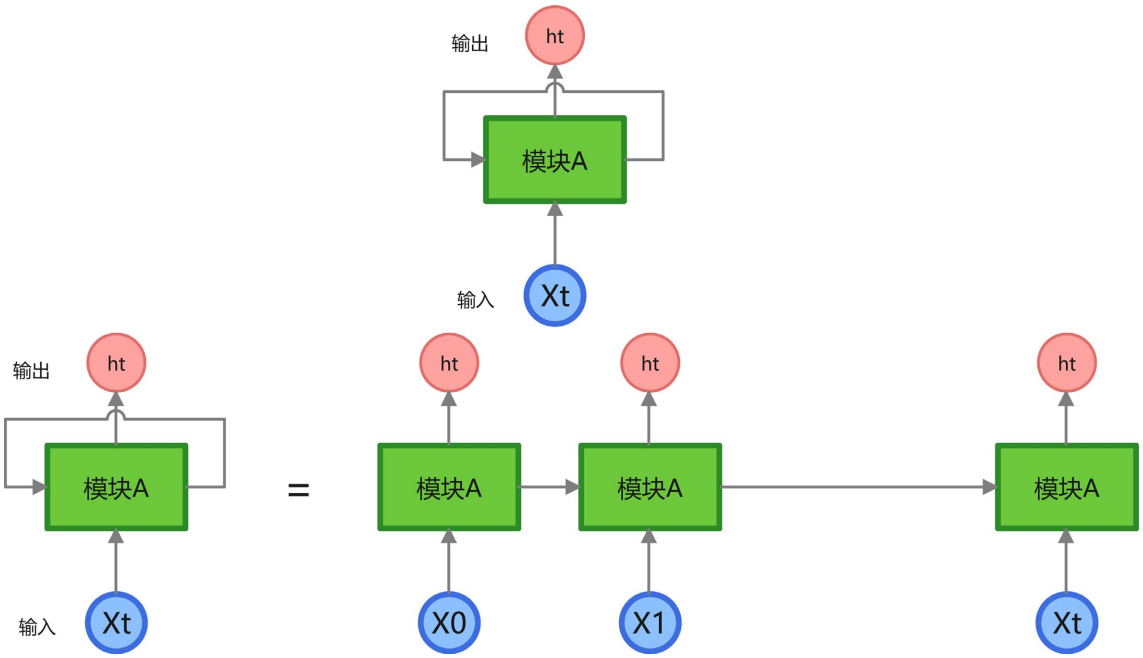

Figure 6. Structure diagram of the RNN and LSTM neural networks
图 6. RNN 神经网络和 LSTM 神经网络的结构示意图

3.2.2. 多特征 LSTM 神经网络预测模型的实现

本文选择使用 MATLAB 软件的 trainNetwork()函数建立游客量多特征 LSTM 神经网络预测模型。

LSTM 神经网络模型结构由以下部分构成：输入层 → LSTM 层(64 单元) → ReLU 激活层 → 全连接层 → 回归输出层。学习率设为 0.001，训练轮次为 1000 轮，优化器选择 Adam 算法。在构建网络时，使用了 MATLAB 中的 trainNetwork()函数，并启用 MiniBatchSize 为 32。所有输入变量归一化处理后进入网络训练。

(1) 读取数据，划分训练集和测试集，对数据进行预处理：将所有数据按照 9:1 的比例将数据划分为训练集和测试集，对数据进行归一化处理。

(2) 构建 LSTM 神经网络并进行训练：构建 LSTM 网络模型，包括输入层、LSTM 层、ReLU 激活层、连接层和输出层，制定网络训练的优化器和参数。

(3) 神经网络预测以及结果反归一化：使用训练好的模型对训练集和测试集进行预测，对预测结果进行反归一化。

(4) 计算误差并绘制对比图：根据 MAPE 的计算公式计算平均绝对误差百分比作为模型优劣的评价指标，并绘制数据图以及结果图。

4. 模型结果与分析

4.1. 数据来源

本文选取杭州市为主要研究对象，获取时间跨度为 2021 年 1 月 1 日至 2024 年 4 月 30 日的研究数据。为保证数据的真实性，研究基于杭州文化和旅游数据在线这一官方数据网站进行数据收集。

杭州文化和旅游数据在线：https://data.wgly.hangzhou.gov.cn/home/#/，是全国首个实时文旅数据在线查询、即时下载的数据开放平台。如图 7 是从网站中收集得到的 2021 年 1 月 1 日至 2024 年 5 月 5 日的杭州市过夜旅客量日流量数据可视化统计图。

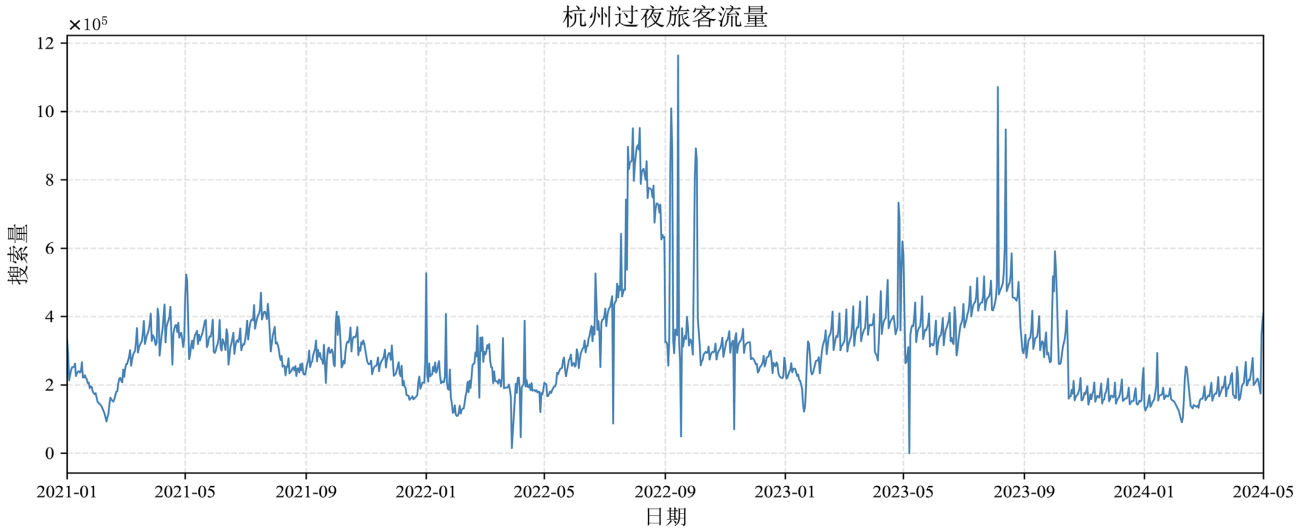

Figure 7. Tourist volume data from Hangzhou culture and tourism data online
图 7. 杭州文化和旅游数据在线关于游客量的数据

4.2. BP 神经网络预测结果

如图 8 展示了神经网络工具箱中 BP 算法的性能和相关性分析图像。性能图(左)展示了训练、测试和预测的均方误差(MSE)随着训练轮次的变化。在前四轮训练后，测试阶段的 MSE 达到最低点为 0.015035，表明模型在这一阶段的性能最佳。这种趋势可能表示模型在前几轮的训练中得到了有效的学习和泛化，但在后续轮次中可能出现了过拟合或其他问题，导致测试阶段的性能稍有波动。相关性分析图像(右)展示了回归任务中的模型性能和预测结果。通过观察这张图，可以评估模型的拟合程度和预测精度。可以看出大部分情况下模型与实际观测值紧密匹配且误差较小，则说明模型在回归任务中表现良好。

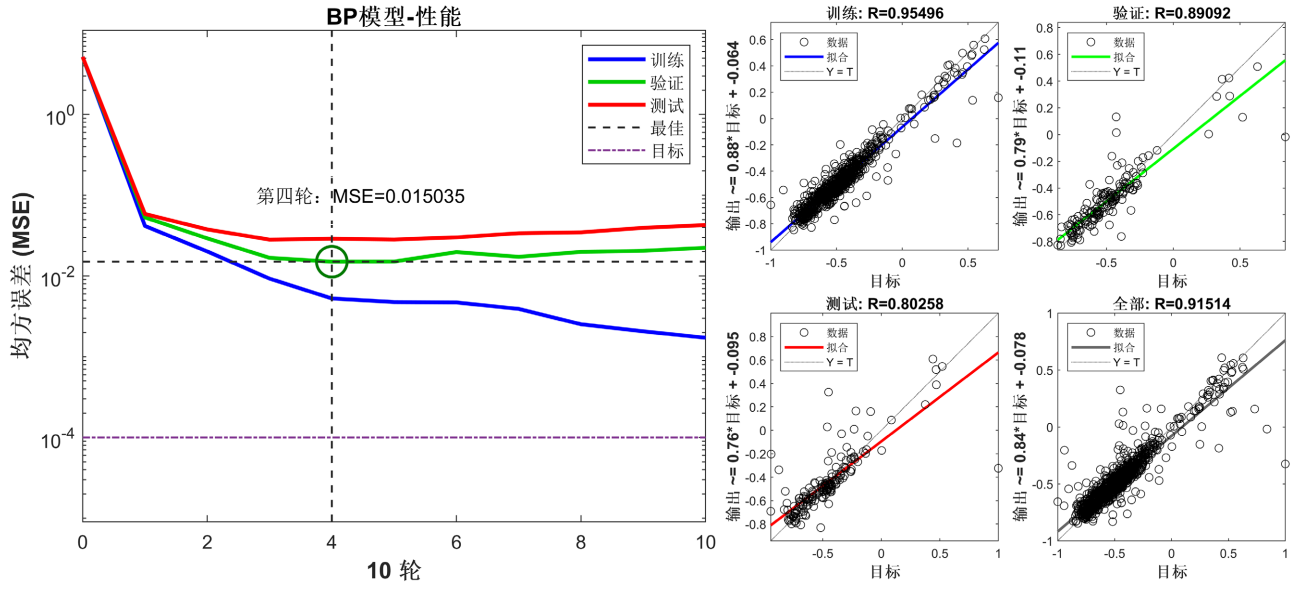

Figure 8. Performance and correlation analysis of BP neural network prediction model
图 8. BP 神经网络预测模型代码的性能和相关性分析图像

如图 9 是 BP 神经网络的训练状态图，展示了训练误差随着训练轮次的变化情况。这张图是监控神

经网络在训练过程中学习效果的关键工具，随着训练的进行，图中的曲线呈现下降趋势，表示网络在学习任务中逐渐减小误差。然而，曲线可能会在某些轮次出现波动，这可能是由于学习率的设置、训练数据的特点或网络结构等因素引起的。通过观察训练状态图，可以帮助调整网络参数和优化训练策略，以提高神经网络的性能和收敛速度。

利用 2021 年 1 月 1 日至 2024 年 5 月 5 日的杭州过夜旅客量数据作为本次的实验数据，并按照 9:1 的比例划分了训练集以及测试集，图 10 展示了模型在训练集和测试集上的预测结果与实际值的对比情况。

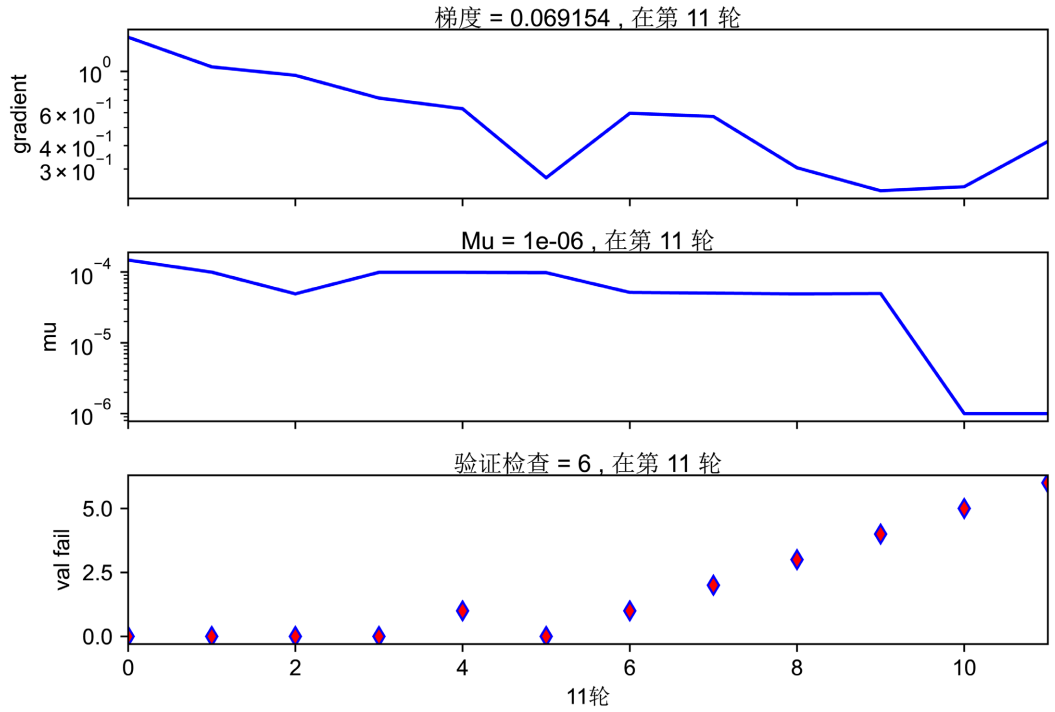

Figure 9. Training state of the BP neural network prediction model
图 9. BP 神经网络预测模型代码的训练状态

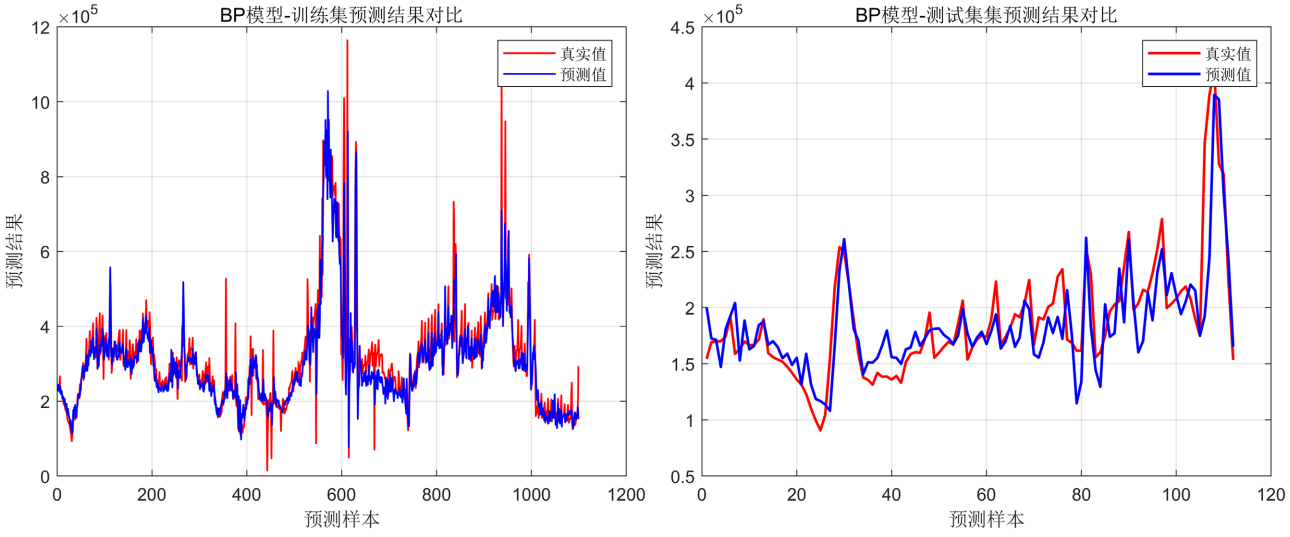

Figure 10. Comparison of predicted and actual values for BP neural network on training and testing sets
图 10. BP 神经网络训练集和测试集的预测结果对比图

4.3. LSTM 神经网络预测结果

在上文中，已经完成了 BP 神经网络预测模型，接下来介绍利用多特征 LSTM 神经网络的预测结果。如图 11 中展示了该场景下 LSTM 神经网络的 RMSE 和损失函数随部分迭代次数的变化。可以观察到模型的损失函数曲线非常陡峭并迅速下降，不到 100 次迭代，损失函数就快速降至 0.005 以下，并逐渐稳定。

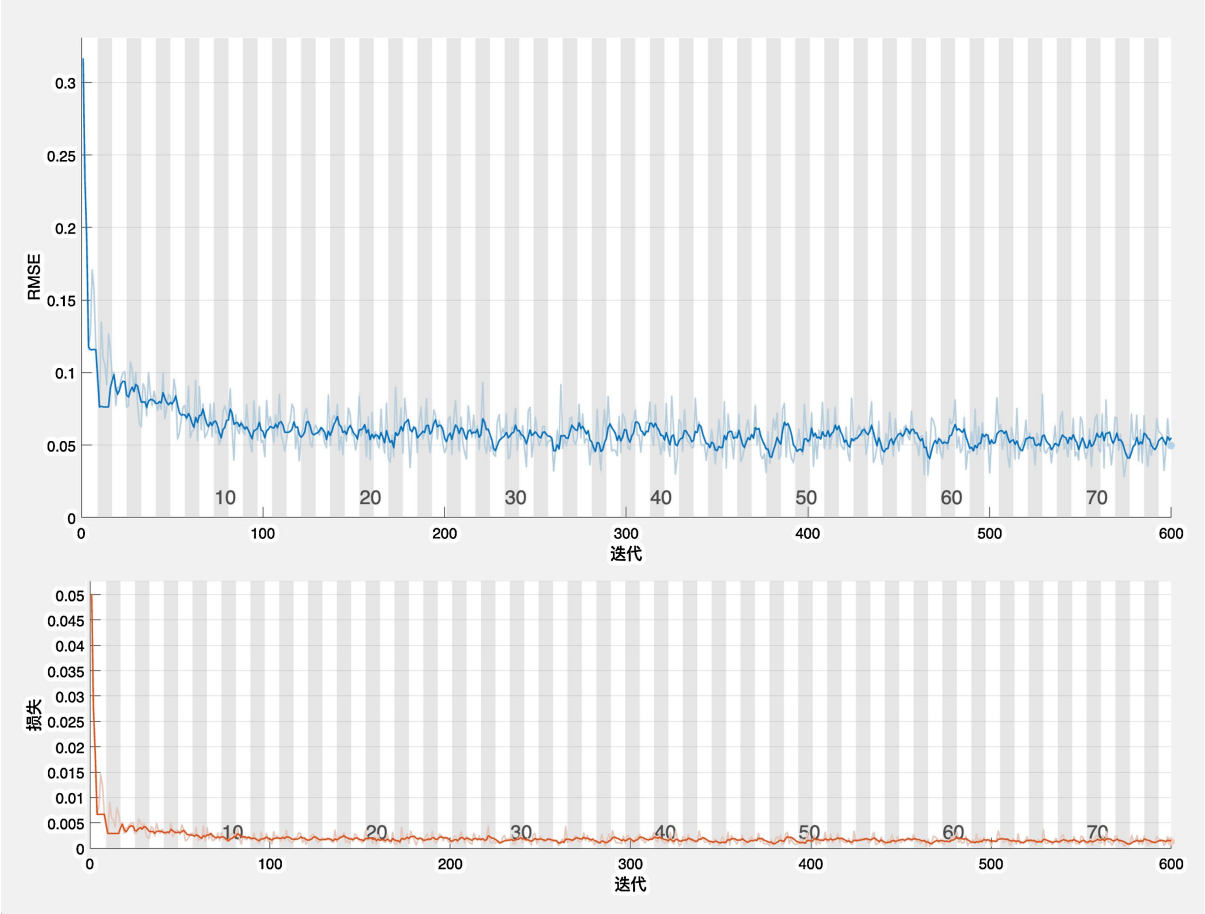

Figure 11. RMSE and loss variation of the multi-feature LSTM prediction model
图 11. 多特征 LSTM 神经网络预测模型 RMSE 和损失函数的变化图

将多特征 LSTM 模型训练好之后，将利用训练集和测试集分别对比实际值进行拟合预测，最后将测试集数值反归之后，输出结果如图 12 所示，更直观地观察模型的预测效果。

4.4. 预测结果对比

在将上述两个神经网络模型进行预测结果对比之前，对 2021 年 1 月 1 日至 2024 年 5 月 5 日的杭州市过夜旅客量数据建立传统的 ARIMA 预测模型，并初步给出 ARIMA 模型的预测结果对比见图 13。

将三个模型(ARIMA 预测模型、BP 神经网络预测模型、多特征 LSTM 神经网络预测模型)的预测结果同时与过夜旅客量实际值进行可视化对比分析，得到对比如图 14 所示。

可以看出，单一模型中，ARIMA 模型的预测效果相对良好，实际值上下波动较为接近，但其曲线仅表现出趋势性，缺乏波动性。这也进一步证实了 ARIMA 模型在线性关系分析方面的优势，但在非线性

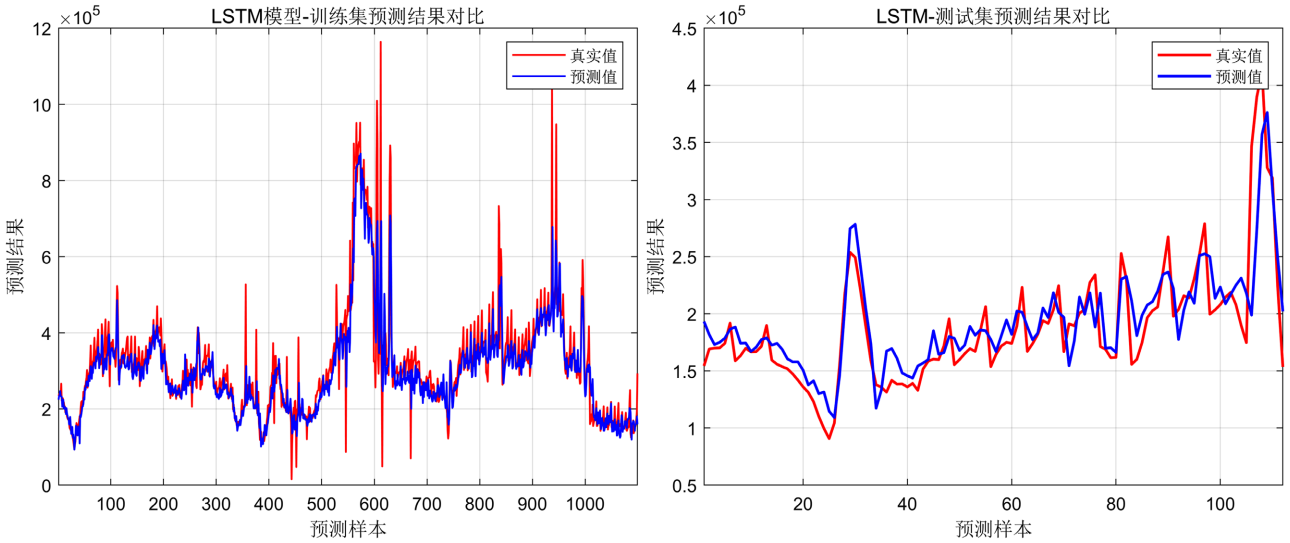

Figure 12. Comparison of prediction results on training and test sets using multi-feature LSTM neural network
图 12. 多特征 LSTM 神经网络训练集和测试集的预测结果对比图

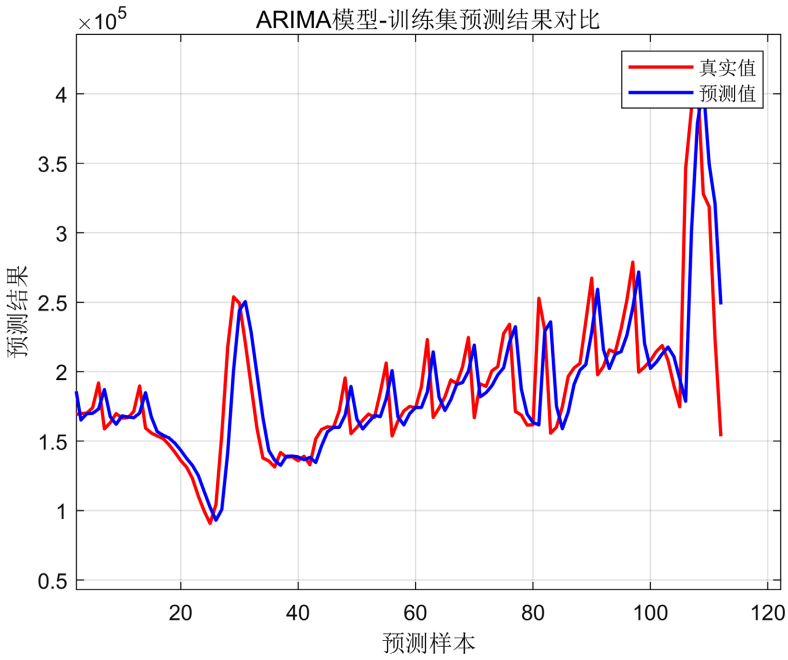

Figure 13. Comparison of prediction results on the training set using the ARIMA model
图 13. ARIMA 模型训练集的预测结果对比图

关系上的表现不足。相比之下，BP 神经网络和 LSTM 神经网络的预测结果具有较好的波动性，并且在实际值方面更为贴近，特别是 LSTM 神经网络。

4.5. 预测误差对比

通过上述图形对比，可以明显观察到基于百度指数的多特征 LSTM 神经网络在预测趋势与实际值曲线之间的贴合效果最佳，更为准确地反映了样本数据的真实情况，因此在预测模型中具有显著的优势。为了更全面地展现各模型的优劣，将通过比较它们的预测误差，并以数字形式呈现不同模型的预测准确

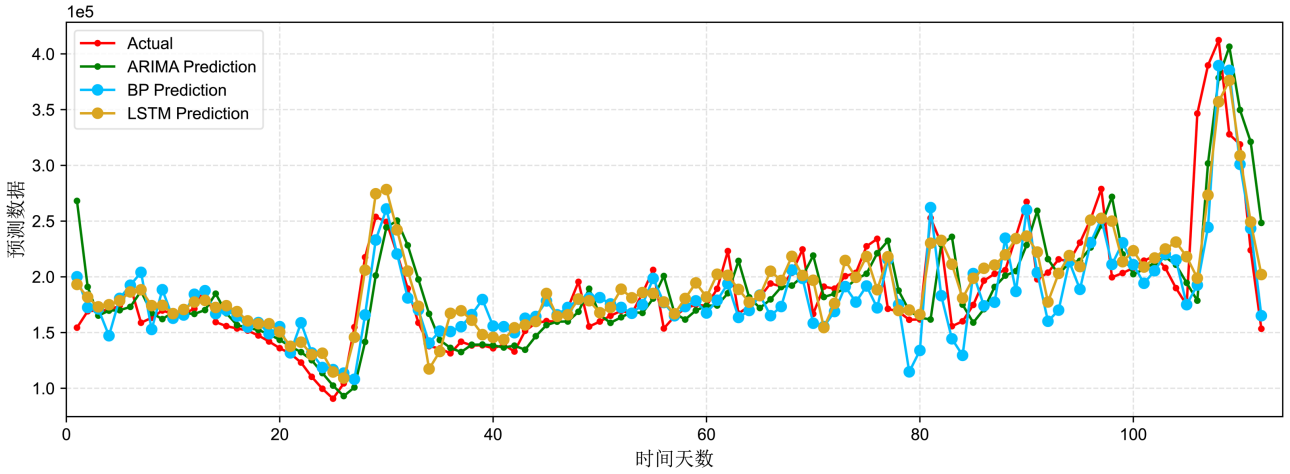

Figure 14. Comparison of prediction results from three models with actual overnight tourist volume
图 14. 三种模型预测结果与过夜旅客量实际值对比图

性。考虑到本研究中构建的五个模型所使用的样本量存在差异，决定不采用常用的 RMSE 和 MAE 指标，而是专注于比较各模型的 MAPE 值。在前文中，每个模型的 MAPE 值已经计算得出，现在将它们汇总如表 5 所示：

Table 5. Summary table of MAPE values for different prediction models
表 5. 不同预测模型的 MAPE 汇总表

模型	MAPE
ARIMA 预测模型	0.116955624
基于百度指数的 BP 神经网络游客量预测模型	0.106814971
基于百度指数的 LSTM 神经网络游客量预测模型	0.099464894

本文所构建的基于杭州历史旅游数据的 ARIMA 模型是传统时间序列预测中最为经典的模型之一。它能够很好地拟合历史数据并进行预测，特别是在线性关系的分析方面表现突出。而基于百度指数关键词的多特征神经网络，则将经过多次筛选的关键词数据和历史销量数据同时作为输入，分别输入到 BP 神经网络和 LSTM 神经网络中进行训练。

5. 结论

本文以杭州市为研究对象，基于 2021~2024 年游客过夜量和百度指数关键词数据，构建了一套集关键词优化筛选、特征重要性排序和深度神经网络预测为一体的城市旅游流量预测模型。在关键词遴选方面，结合 Spearman 相关系数与随机森林重要性评估方法，有效提取出与游客行为高度相关的搜索特征变量，在提升预测信号质量的同时，也增强了模型的可解释性。研究发现，“西溪国家湿地公园”等景点词与“杭州东站”等交通出行词具有较强的游客引导作用，表明网络搜索行为与实际旅游需求之间存在显著的因果关联。

在模型构建方面，本文分别建立了基于百度指数的 BP 神经网络与多特征 LSTM 神经网络预测模型，并与传统 ARIMA 模型进行了对比分析。从模型性能指标来看，LSTM 神经网络的预测误差(MAPE)最低，仅为 0.099，显著优于 BP 神经网络与 ARIMA 模型，具有更强的时间记忆能力和非线性拟合能力。在预测精度、拟合曲线平滑度与波动捕捉等方面，LSTM 模型表现更为优越，特别适用于旅游需求受多因素

交互影响的复杂预测场景。研究结果验证了将数字足迹数据与深度学习方法结合的可行性与有效性，为城市旅游流量管理与智慧旅游决策提供了数据支持和技术方案。

综上所述，基于数字足迹与深度学习的城市旅游流量预测方法，兼具前瞻性、实用性与可拓展性。通过深度挖掘网络搜索行为中蕴含的潜在出行意图，结合非线性神经网络的强大建模能力，本文实现了对旅游流量动态演化的精准刻画，有效提升了预测模型的时效性与适应性。该研究不仅为城市旅游管理者在高峰期接待能力评估、交通疏导与资源配置等方面提供了决策参考，也为推动智慧旅游的数据融合、模型优化与服务升级提供了方法借鉴，具有广泛的现实价值与应用前景。

基金项目

江苏省研究生科研与实践创新计划项目(SJCX230250)。

参考文献

[1] 刘洋. 旅游流时空分布特征及调控研究——以桂林为例[D]: [硕士学位论文]. 南宁: 广西大学, 2016.

[2] 李云飞. 基于 ARIMA 和 LSTM 神经网络对中国入境游客规模预测的比较研究[J]. 社会科学前沿, 2019, 8(7): 1291-1298.

[3] He, K., Ji, L., Wu, C.W.D. and Tso, K.F.G. (2021) Using SARIMA-CNN-LSTM Approach to Forecast Daily Tourism Demand. *Journal of Hospitality and Tourism Management*, **49**, 25-33. https://doi.org/10.1016/j.jhtm.2021.08.022

[4] 王雷雪. 基于机器学习的旅游需求量预测研究[D]: [硕士学位论文]. 西安: 西安石油大学, 2021.

[5] 邹红梅, 朱成涛. 基于 LSTM 和 BP 神经网络的水库入库径流中长期预测比较研究[J/OL]. 水文: 1-8. https://doi.org/10.19797/j.cnki.1000-0852. 20230413, 2024-05-14.

[6] 陈尚林. 基于 LSTM 神经网络和百度指数的新能源汽车销量预测[D]: [硕士学位论文]. 武汉: 湖北大学, 2024.

[7] 李山, 邱荣旭, 陈玲. 基于百度指数的旅游景区络空间关注度: 时间分布及其前兆效应[J]. 地理与地理信息科学, 2008(6): 102-107.

[8] 任乐, 崔东佳. 基于网络搜索数据的国内旅游客流量预测研究——以北京市国内旅游客流量为例[J]. 经济问题探索, 2014(4): 67-73.

[9] 谢天保, 赵萌. 基于网络搜索数据的游客量组合预测模型[J]. 计算机系统应用, 2018, 27(7): 199-204.

[10] 孟思聪, 马晓冬. 基于百度指数的连云港旅游网络关注度研究[J]. 旅游论坛, 2017, 10(5): 102-115.

[11] 戴贵洋, 綦秀利, 余晓晗. 融合人类知识的随机森林特征选择方法研究[J]. 计算机技术与发展, 2022, 32(7): 155-160.