企业应用

# 基于百度指数与机器学习的景区游客量预测研究

潘 辉

(徽商职业学院低空经济学院，合肥 230022)

**摘要**：基于百度指数，以黄山风景区为研究对象，选取 2014 年 1—3 月期间数据，通过六种机器学习方法构建了游客量预测模型。结果表明，百度指数与游客量关系密切，其可以作为游客量预测的依据；景区游客量不仅与其本身网络搜索量有关，还与周边景区搜索量存在距离衰减式的空间溢出效应；树模型能够较好地捕捉百度指数与游客量之间的非线性关系，其中，极限梯度提升模型在六种机器学习方法中表现突出。

**关键词**：百度指数；游客量预测；机器学习；空间溢出效应；黄山风景区

**中图分类号**：F592.6 **文献标志码**：A **文章编号**：1671－1807(2026)14－0207－09

游客量预测是旅游研究的重要内容，有助于目的地进行旅游资源规划与管理、旅游风险管理及政策制定与调整等。传统依靠政府统计数据进行游客量预测的方法存在一定的滞后性，游客是否前往目的地旅游受到多重因素的影响，而目的地的偶发性事件在网络发达的环境下更易被获悉，直接影响游客的出游决策。相较于传统统计数据，网络搜索数据实时可得，依靠网络搜索数据进行游客量预测，可以将偶发性事件对目的地游客量的影响考虑在内，更好地规避滞后性缺陷，便于目的地及时有效安排各项资源。网络搜索数据反映了游客对旅游目的地的关注度，对游客行为具有一定的引导力。旅游目的地的网络关注度是游客量的象征[1]，通过获取游客出行前的网络搜索数据，预测其之后的旅游活动成为互联网时代探究旅游出行特征的新方法[2]。

目前，关于游客量预测常用的方法有自回归移动平均模型[3]、季节性差分自回归滑动平均模型[4]等基于游客接待量历史数据进行游客量预测，以及在时间序列模型基础上引入解释变量构建灰色系统模型[5]、向量自回归模型[6]等。随着智能算法的普及，各类机器学习方法被广泛应用于游客量预测，如多层感知机[7]、支持向量回归(support vector regression，SVR)[8]、反向传播神经网络(back propagation neural network，BPNN)[9]等，用于捕捉游客量时序数据中的复杂特征。为进一步提升预测精度，部分学者尝试将卷积神经网络与长短期记忆网络相结合，或引入 Transformer 架构，以同时捕捉游客量序列中的短期波动与长期变化趋势[10]。数据获取渠道从传统的政府统计数据向网络搜索数据转变，百度指数作为网络搜索数据的主流平台，成为游客量预测的重要依据，目前已有学者基于百度指数对杭州[2]、水墨丹霞[11]、丽江[12]、天山[13]、九寨沟[14]等地进行了研究。

本文以黄山风景区为例，运用斯皮尔曼相关性分析方法，分析百度指数与景区每月接待游客量之间的关系。通过相关性分析选取了 11 个显著相关的关键词，接着采用遗传算法-随机森林-序列后向选择方法进行关键词降维处理，确定了 5 个关键词。在此基础上，运用 SVR、径向基神经网络(radial basis function neural network，RBFNN)、BPNN、决策树(decision tree，DT)、随机森林(random forest，RF)以及极限梯度提升(extreme gradient boosting，XGBoost)六种机器学习模型构建关键词指数与游客量之间的预测模型，以获取关键词与游客量之间的非线性关系，并基于多项评估指标确定最优预测模型。

## 1 实证研究

### 1.1 研究区概况

黄山风景区位于安徽省黄山市，自然资源丰

**收稿日期**：2026-03-21

**基金项目**：安徽省质量工程项目(2022tsxtz059)

**作者简介**：潘辉(1988—)，女，山东聊城人，硕士，讲师，研究方向为旅游经济。

富，文化底蕴深厚，拥有世界自然与文化双遗产、世界地质公园双重桂冠，是中国第一批国家AAAAA级旅游景区。随着大黄山世界级休闲度假康养旅游目的地建设行动方案的提出及营销推广，大黄山旅游目的地的影响力不断提升，而黄山景区作为大黄山旅游目的地的核心吸引物，其管理水平直接影响游客对大黄山旅游目的地的评价。因此，对黄山景区游客人数进行提前预测，有助于景区合理安排人力物力资源，为游客提供更好的旅游体验，更好地服务于“烟雨大黄山，心灵休憩地”的大黄山品牌建设。

## 1.2　数据来源

### 1.2.1　游客数量

为了更好地进行游客数量预测，选取2014年1月—2024年3月共123个月的黄山景区游客接待量数据进行分析。其中，黄山景区每月游客接待量数据来源于黄山市人民政府网站公布的政府数据（https://www.huangshan.gov.cn/zfsj/index.html）。除了2020—2022年期间新冠疫情对旅游造成的冲击，其他年份的黄山景区游客量呈现明显的季节波动性特征，约在每年的2月、4月、8月、10月出现四次波峰，如图1所示。2022年底新冠疫情管控全面放开后，2023年黄山景区游客量创历史新高，共接待游客457.5万人次，较之前的2019年增长了30.68%。为了后续机器学习模型的训练、预测及分析，将数据集划分为训练集和测试集，其中训练集用来构建模型，测试集用来检验模型，按照7∶3的比例划分训练集和测试集[15]。为避免2020—2022年新冠肺炎疫情冲击对模型参数估计的干扰，将研究期划分为常态与疫情两个阶段，其中2014年1月—2019年1月为常态阶段训练集，2020年1月—2022年1月为疫情阶段训练集。

### 1.2.2　百度指数

百度指数是基于百度海量网民行为数据的分析平台，其数据主要来源于网民在百度搜索引擎上所记录的主动搜索意愿和频次情况。百度指数可用于研究关键词搜索趋势，以了解网民需求变化，便于企业进行市场分析和决策。百度指数所提供的需求图谱功能，展示了关键词和相关词的关联程度及相关词自身的搜索需求，为关键词的进一步选取提供了参考依据。以百度指数搜索网站（https://index.baidu.com/v2/index.html#）作为数据来源，获取与黄山旅游相关关键词的百度指数，以此来探讨百度指数与黄山游客量之间的关系。

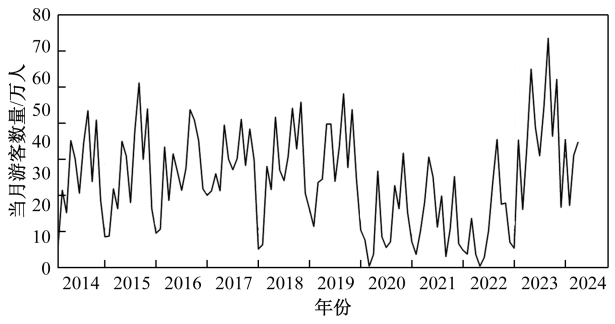



**图1　黄山风景区每月接待游客量变化**
**（2014年1月—2024年3月）**

## 1.3　数据处理

### 1.3.1　关键词选取及频次统计

（1）关键词选取。游客出行决策受到多种因素的影响，传统旅游六要素“食、住、行、游、购、娱”是吸引游客前往旅游目的地的重要因素。游客出行前，通过搜索引擎了解目的地的六要素情况。从六要素的角度出发，以“黄山美食”“黄山住宿”“黄山交通”“黄山旅游”“黄山特产”“黄山演出”作为初始的关键词进行百度指数搜索，选择需求图谱中“强”和“相关性”区域内的关键词句作为拓展关键词，再以拓展关键词进行二次搜索，去除百度指数未收录的关键词，再参考相关文献[12-14]所选取的关键词，如“天气”“地图”“旅游攻略”等，最终选取了75个关键词。

值得注意的是，在全域旅游背景下，游客往往将空间邻近的多个景区纳入同一行程，形成邻近景区联动游览的消费模式。因此，周边景区的网络搜索量可能通过空间溢出效应对黄山游客量产生影响。基于此，在关键词筛选中保留了西递、宏村、齐云山、牯牛降、瑶里古镇、庐山、千岛湖等周边景区关键词，以捕捉空间溢出效应。为验证这一效应的存在，在接下来的相关性分析中进行测算。

（2）词频统计。以2014年1月—2024年3月为统计时段，将75个关键词PC端和移动端的百度指数加总得到日百度指数，再汇总为月百度指数，以月份的时间粒度完成全部关键词的频次统计。图2为部分关键词词频统计。

### 1.3.2　相关性分析

（1）斯皮尔曼（Spearman）相关性分析法。为验证所选取的75个关键词的百度指数与游客量是否相关，须进行相关性分析。一般而言，可采用皮尔逊（Pearson）相关性分析法或斯皮尔曼相关性分析法，其中Pearson相关性分析主要针对服从正态分布的数据。采用柯尔莫戈洛夫-斯米诺夫（KS）检验

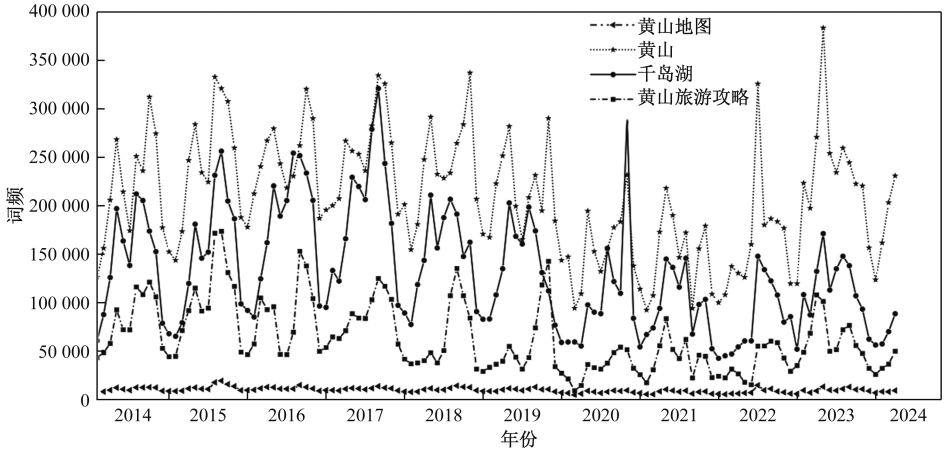



**图 2　黄山地图、黄山、千岛湖、黄山旅游攻略词频统计**

可以判别数据是否符合正态分布，若 $p>0.05$ 则说明符合正态分布，$p<0.05$ 则说明不符合正态分布。经检验，黄山风景区游客接待量的 $p$ 为 0.192>0.05，服从正态分布；75 组关键词的百度指数中 $p>0.05$ 的有 28 组数据，$p<0.05$ 的有 47 组数据，并不全部符合正态分布，因此采用 Spearman 相关性分析法来衡量游客量与百度指数之间是否存在相关性。

(2)相关性分析结果。采用 SPSS 软件计算 2014 年 1 月—2024 年 3 月期间黄山景区接待游客量与 75 个关键词的月百度指数之间的相关系数，其中相关系数≥0.6 的关键词有 28 个，相关系数≥0.7 的关键词有 11 个，选取相关系数≥0.7 的关键词进行分析。相关性分析结果如表 1 所示。

(3)滞后性分析。由于百度指数为先行指标，对游客量的影响有一定的滞后性。将景区接待游客量与所选关键词的百度指数进行一阶滞后分析和二阶滞后分析，发现相关系数均小于 0.7，因此暂不考虑百度指数与游客量之间的滞后性。这可能与以月为时间跨度进行分析有关系，游客在一个月内完成从决策到出游的行为。有学者曾以杭州为研究对象，发现从百度搜索到实际旅游的滞后期为 1～23 d[2]，恰好验证了这一结论。

(4)空间溢出效应分析。在 11 个关键词中，黄山、黄山地图、黄山美食、安徽黄山 4 个关键词与黄山本身密切相关。西递、齐云山、牯牛降、宏村天气 4 个关键词与黄山周边景点相关，西递、宏村与黄山同为世界遗产，位于黄山市黟县；齐云山是中国道教四大名山之一，位于黄山市休宁县；牯牛降以“华东地区的最后一片原始森林”著称，位于池州市石台县，以上四个景区距离黄山景区 2～3 h 车程。瑶里古镇、庐山、千岛湖 3 个关键词为黄山周边跨省景点，瑶里古镇有“瓷之源，茶之乡，林之海”的美称，位于江西景德镇市；庐山为世界文化景观，位于江西九江市；千岛湖被誉为“天下第一秀水”，位于浙江淳安县，以上 3 个景点距离黄山景区约 4 h 车程。除庐山距离黄山景区较远外，宏村、西递、齐云山相距在 50 km 的范围圈内，牯牛降、瑶里古镇、千岛湖相距在 100 km 范围圈内，景区之间地理位置相近，在皖南、赣北、浙西三地之间形成了一个出行旅游圈。由此可知，黄山景区游客量除了受本身搜索量的影响外，还受周边景区网络搜索量的影响。在全域旅游背景下，实现省内市域之间及省域之间的协同联动显得尤为重要。

为验证周边景区搜索量的空间溢出效应，构建了引力模型。

$$I_j = \frac{B_j}{d_j} \tag{1}$$

式中：$I_j$ 为周边景区 $j$ 对黄山景区的空间影响力；$B_j$ 为周边景区 $j$ 的月均百度指数；$d_j$ 为景区 $j$ 与黄山景区的驾车距离，驾车距离数据取自百度地图。

经过逐月计算庐山、齐云山、千岛湖、西递、瑶里古镇、牯牛降等 6 个周边景区的综合影响力指数，将其与黄山月游客接待量进行相关性分析，结果如表 2 所示。

**表 1　Spearman 相关性分析结果**

| 变量 | 西递 | 黄山 | 黄山地图 | 瑶里古镇 | 庐山 | 黄山美食 | 齐云山 | 牯牛降 | 安徽黄山 | 宏村天气 | 千岛湖 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 相关系数 | 0.783** | 0.777** | 0.773** | 0.758** | 0.745** | 0.744** | 0.743** | 0.720** | 0.715** | 0.707** | 0.705** |

注：**表示在 0.01 级别(双尾)相关性显著。

**表 2　Spearman 相关性分析结果**

| 变量 | 西递 | 瑶里古镇 | 齐云山 | 牯牛降 | 千岛湖 | 庐山 |
|---|---|---|---|---|---|---|
| 相关系数 | 0.766** | 0.757** | 0.692** | 0.649** | 0.641** | 0.597** |

注：** 表示在 0.01 级别（双尾）相关性显著。

由表 2 可知，相关系数为 0.597～0.766，呈中度至强度正相关，且均在 0.01 级别（双尾）上显著。其中，西递的相关系数最高（0.766），其次是瑶里古镇（0.757），齐云山（0.692）位列第三，而庐山的相关系数最低（0.597）。排序结果表明，距离黄山较近的景区（如西递、瑶里古镇）其网络搜索热度与黄山游客量的相关性更强，而距离较远的景区（如庐山）相关性相对较弱，呈现一定的距离衰减特征，即距离越近的景区其百度指数与黄山游客量的相关性越强。这一结果验证了空间溢出效应的存在，说明在全域旅游背景下，邻近景区的网络关注度对本地游客量具有前兆预测价值，同时也为后续关键词降维保留了周边景区变量的必要性。

#### 1.3.3　关键词降维

以 11 个关键词作为解释变量进行建模，会增加模型的复杂度，不便于进行游客量预测，可进一步采用遗传算法-随机森林-序列后向选择（GA-RF-SBS）方法对关键词进行降维处理。

随机森林（random forest，RF）法是一种监督式学习算法，是由多个决策树组成的集成模型，训练集提取的随机性和随机选择的特征属性减少了决策树对训练数据的敏感性，从而提高了预测精度和泛化能力。其中决策树的数量和叶子数对模型精度有重要影响，本文采用遗传算法来优化决策树的数量和叶子大小。遗传算法（genetic algorithm，GA）是一种自适应的模拟生物遗传和进化过程的自适应全局优化算法。遗传算法会预先生成一组解，即要优化的超参数，随后经过遗传、交叉、变异操作，实现优胜劣汰，最后得到最优超参数。遗传算法的初始设置如下：种群数为 150，代数为 50，选择函数为“tournament”，锦标赛规模为 2，交叉函数为“scattered”，变异函数为“adapt feasible”，其他参数为默认参数[16]。序列后向选择（sequential backward selection，SBS）算法是一种基于包装器的特征选择方法，其基本思想是从一组候选特征开始，迭代地删除对模型适应度影响最小的特征，直到达到预定数量的特征。

GA-RF-SBS 方法的流程如图 3 所示。首先确定 1 个关键词进行 RF 建模，按照测试集均方根误差的大小逐步剔除另外 10 个关键词，随后依次增加关键词，最终获得最优关键词个数和最优关键词。

图 4 为引入关键词的个数对模型训练集和测试集的最小均方根误差（root mean square error，RMSE）的影响变化。模型测试集的 RMSE 并非随着关键词个数的增加而增加，而是呈现波动性变化。当关键词个数为 5 或 7 时，RMSE 数值较小，模型精度较高。为同时保证模型的简洁性，选择“西递、黄山地图、黄山美食、齐云山、千岛湖”5 个关键词的百度指数作为机器学习模型的输入变量。

## 2　游客量预测模型及评估

### 2.1　模型介绍

游客量受到多种因素的共同作用，百度指数和游客量之间的非线性关系还需进一步探索。为此，采用了六种非线性建模方法进行研究，RBFNN、BPNN、SVR、DT、RF 及 Extreme Gradient Boosting，XGBoost。

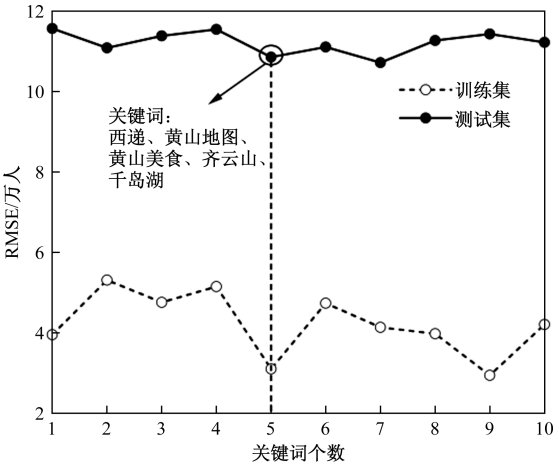



**图 4　RMSE 随关键词个数的变化趋势**

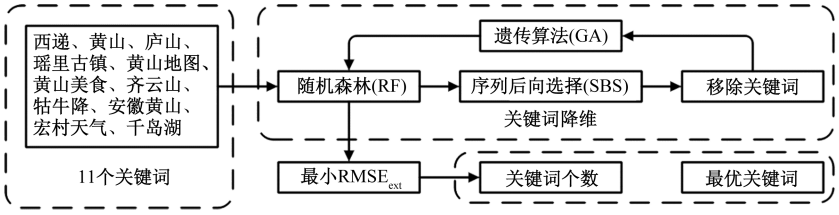



**图 3　GA-RF-SBS 关键词降维流程**

RBFNN是一种具有强大数据拟合和预测能力的人工神经网络[17],分为输入层、隐藏层和输出层,输入层与隐藏层之间使用径向基函数来确定连接权重。BPNN拥有优秀的非线性映射能力和良好的预测性能,一般来讲3层神经网络已有较强的预测效果,因此选择构建3层BPNN模型对游客量进行预测。SVR是常用的非线性回归算法[18],SVR的主要思路是在空间中构造一个超平面,使空间中的数据到该超平面的距离最短。在SVR中有linear、polynomial、RBF等核函数,由于游客量与百度指数之间存在明显的非线性关系,因此采用RBF核函数处理自变量与因变量之间的非线性关系[19]。DT是一种基于树结构的非参数监督学习方法,其基本思想是通过不断划分特征空间,使得每个叶节点内部样本尽可能相似,从而实现对目标变量的预测[20]。XGBoost是在梯度提升树基础上发展而来的高效集成学习算法,具有较强的非线性拟合能力和较优的计算效率。XGBoost通过对弱学习器进行逐步迭代提升,并在目标函数中加入正则化项,能够在提高预测精度的同时有效控制模型复杂度,减少过拟合风险[21]。本文所采用的建模方法涵盖树模型、神经网络、支持向量机等多种算法,各类方法在非线性建模中各有优势。通过对比这些模型在常态与新冠疫情阶段下的预测表现,可以更全面地评估不同算法在游客量预测中的适用性与稳健性。

在模型构建的过程中,需要不断调整超参数来提高模型精度。采用5折交叉验证算法对机器学习模型超参数进行优化,以提高模型的游客量预测精度。超参数范围如表3所示。

**表3　超参数和优化范围**

| 建模方法 | 超参数 | 范围 |
|---|---|---|
| RBFNN | gamma, n_centers | [0.01,2],[1,20] |
| BPNN | hidden_layer_sizes | [0.1,0.7],[1,60] |
| SVR | Cost, Gamma | $[2^{-7.5}, 2^{7.5}]$, $[2^{-7.5}, 2^{7.5}]$ |
| RF | n_estimators, max_depth, min_samples_split, min_samples_leaf | [100,300],[1,13],[2,6],[1,4] |
| DT | max_depth, min_samples_split, min_samples_leaf | [2,7],[2,4],[1,3] |
| XGBoost | max_depth, learning_rate, n_estimators | [3, 9], [0.01, 0.1], [50,200] |

## 2.2　模型预测与评估

考虑到2020—2022年受突发公共卫生事件影响,旅游出行行为、景区开放政策及游客搜索与出游转化关系均可能发生阶段性变化,若将该时期数据与常态时期样本直接混合建模,可能导致模型参数受到异常冲击干扰,从而削弱模型对常态规律的刻画能力。基于此,采用分段建模策略,将样本划分为常态阶段与疫情阶段分别建模,以检验不同阶段下模型拟合效果、参数特征及变量作用关系的稳定性。

以降维后所确定的5个关键词的百度指数作为输入变量,以游客量作为输出变量,分别采用RBFNN、BPNN、SVR、XGboost、RF、DT六种机器学习模型构建游客量预测模型。六种机器学习模型的最优超参数如表4所示,预测结果与真实值的对比情况如图5～图10所示。

为进一步分析6种模型在测试集和预测集上的效果偏差,采用平均绝对误差(mean absolute error, MAE)、均方误差(mean square error, MSE)及RMSE对机器学习预测精度进行评价,其表达式如式(2)～式(4)所示。MAE、MSE、RMSE的指标值越小,模型预测的误差越小,效果越好。六种机器学习模型在测试集上的统计学参数如表5所示。其

**表4　各模型最优超参数**

| 模型 | 阶段 | 最优超参数 |
|---|---|---|
| DT | 常态阶段 | max_depth=2, min_samples_split=4, min_samples_leaf=1 |
| DT | 疫情阶段 | max_depth=2, min_samples_split=3, min_samples_leaf=1 |
| MLP | 常态阶段 | activation='tanh', alpha=0.1, hidden_layer_sizes=(20,10), solver=adam |
| MLP | 疫情阶段 | activation='tanh', alpha=0.1, hidden_layer_sizes=(10, 10), solver='adam' |
| RBFNN | 常态阶段 | gamma=0.1, n_centers=15 |
| RBFNN | 疫情阶段 | gamma=0.05, n_centers=10 |
| RF | 常态阶段 | max_depth=7, min_samples_leaf=4, min_samples_split=6, n_estimators=300 |
| RF | 疫情阶段 | max_depth=5, min_samples_leaf=2, min_samples_split=6, n_estimators=100 |
| SVR | 常态阶段 | C=1 000, gamma=0.01 |
| SVR | 疫情阶段 | C=100, gamma=0.01 |
| XGBoost | 常态阶段 | learning_rate=0.01, max_depth=3, n_estimators=200 |
| XGBoost | 疫情阶段 | learning_rate=0.05, max_depth=3, n_estimators=100 |

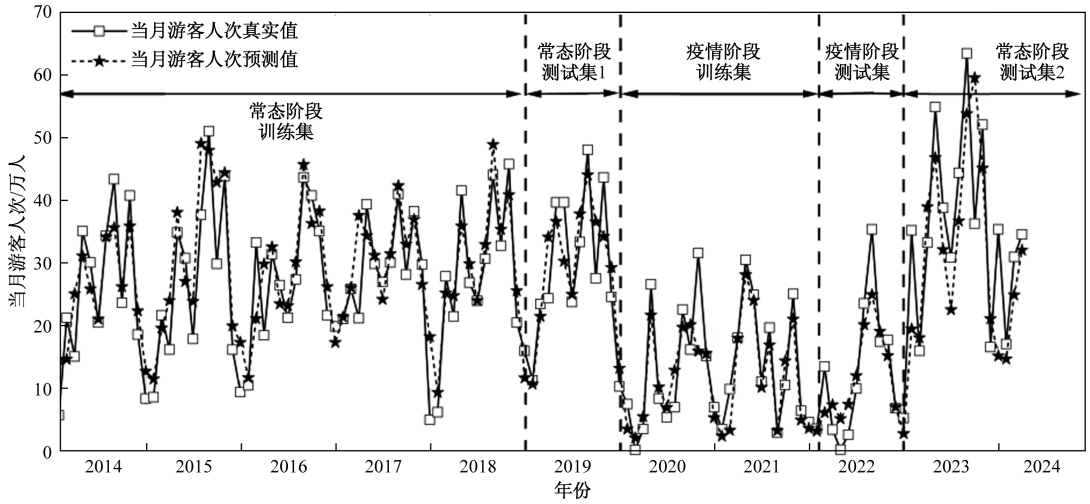



**图 5　RBFNN 模型游客真实值与预测值对比**

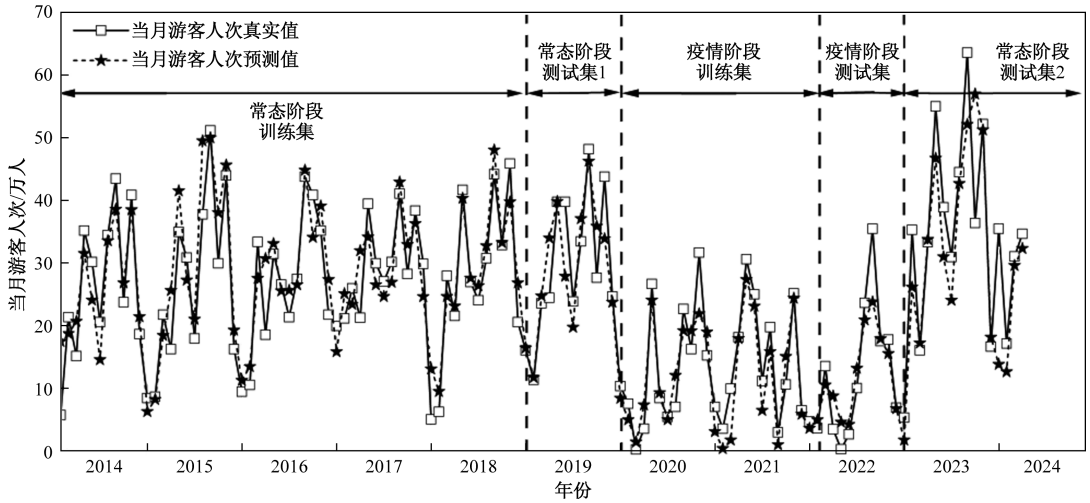



**图 6　BPNN 模型游客真实值与预测值对比**

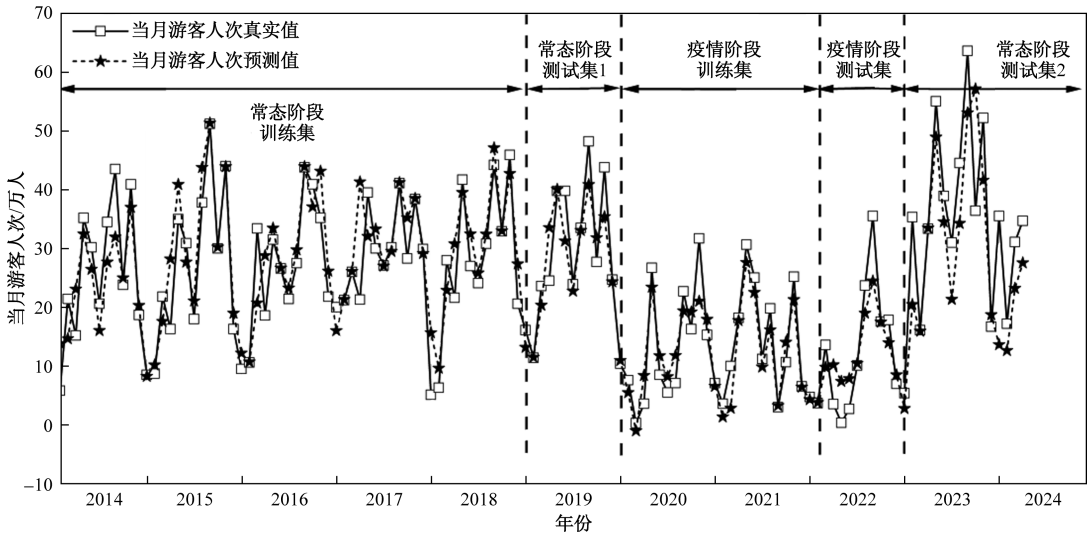



**图 7　SVR 模型游客真实值与预测值对比**

中在常态阶段，SVR 模型的预测误差最大，XGBoost模型的误差最小，其 MSE、MAE、RMSE 分别为 46.95 万、5.59 万、1.32 万人；在疫情阶段，XGBoost 模型仍然表现最优，其 MSE、MAE、

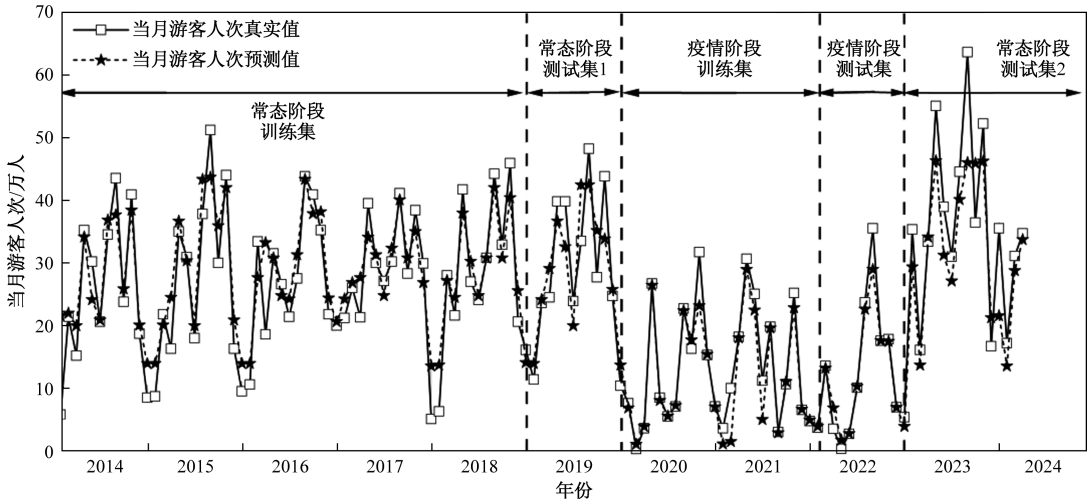



**图 8　XGBoost 模型游客真实值与预测值对比**

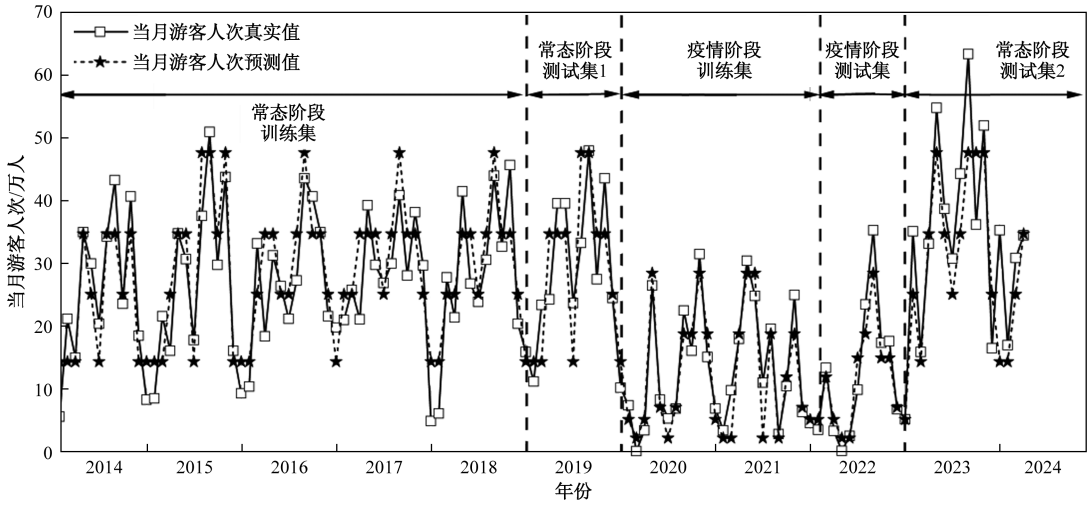



**图 9　DT 模型游客真实值与预测值对比**

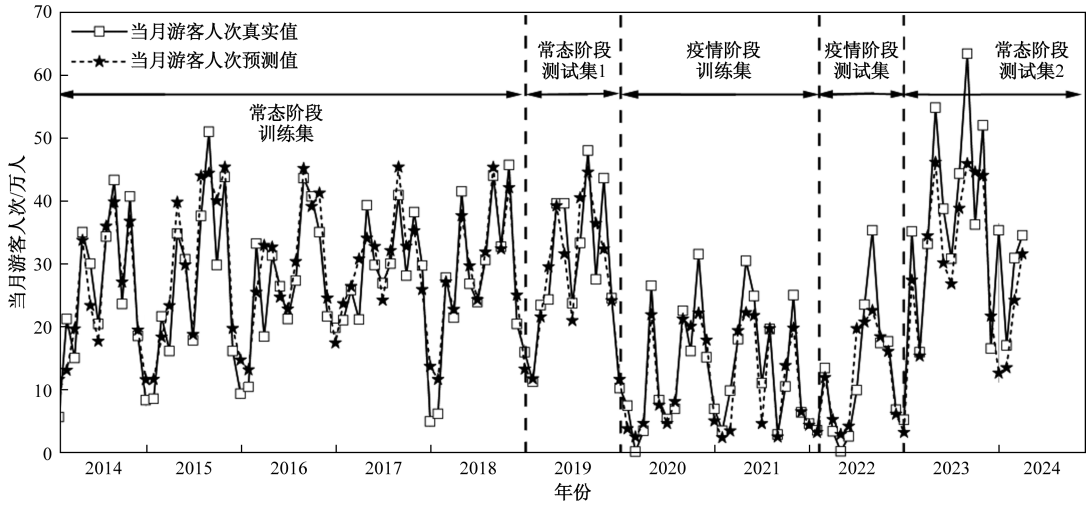



**图 10　RF 模型游客真实值与预测值对比**

RMSE 分别为 5.42 万、1.35 万、0.7 万人;SVR 模型的误差则相对最大。进一步从模型类别来看,树模型对百度指数与游客量之间的非线性关系有较强的捕捉能力。

表 5 六种模型测试集预测效果对比

| 模型 | 常态阶段 MSE/万人 | 常态阶段 MAE/万人 | 常态阶段 RMSE/万人 | 疫情阶段 MSE/万人 | 疫情阶段 MAE/万人 | 疫情阶段 RMSE/万人 |
|---|---|---|---|---|---|---|
| DT | 71.77 | 6.9 | 1.63 | 10.65 | 2.51 | 0.98 |
| MLP | 65.06 | 5.69 | 1.55 | 20.89 | 3.46 | 1.38 |
| RBFNN | 78.51 | 7.05 | 1.71 | 23.39 | 3.98 | 1.46 |
| RF | 62.23 | 6.02 | 1.52 | 26.35 | 3.48 | 1.55 |
| SVR | 76.07 | 6.46 | 1.68 | 27.55 | 4.26 | 1.58 |
| XGBoost | 46.95 | 5.59 | 1.32 | 5.42 | 1.35 | 0.7 |

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|(y_i-\hat{y}_i)\right| \tag{2}$$

$$\mathrm{RMSE}=\frac{1}{n}\sqrt{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{3}$$

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{4}$$

## 3 结论

本文基于 2014 年 1 月—2024 年 3 月的黄山风景区游客接待量以及百度指数数据，利用斯皮尔曼相关性分析方法，分析了游客接待量与网络搜索关键词百度指数之间的关系，并通过遗传算法-随机森林-序列后向选择的方法对关键词进行了降维处理，确定了与黄山风景区关系密切的 5 个关键词，将其百度指数作为输入变量，通过 RBFNN、BPNN、SVR、XGboost、RF、DT 六种机器学习方法构建了黄山景区游客量预测模型。通过六种模型预测效果的比较分析，得到以下主要结论。

(1)通过斯皮尔曼相关性分析发现，在 75 个关键词中有 11 个关键词的百度指数与黄山景区游客量的相关系数在 0.7 以上，分别是“西递、黄山、黄山地图、瑶里古镇、庐山、黄山美食、齐云山、牯牛降、安徽黄山、宏村天气、千岛湖”，呈现显著相关性，因此，通过关键词的百度指数进行游客量预测是可行的。

(2)以 11 个关键词的百度指数作为输入变量构建预测模型较为复杂，通过关键词降维处理发现，使用“西递、黄山地图、黄山美食、齐云山、千岛湖”这 5 个关键词的百度指数来构建黄山景区游客量预测模型，既可以保证模型精度，又可以简化模型，方便后期使用。

(3)验证了周边景区搜索量对黄山游客量的空间溢出效应，且该效应随距离增大而衰减。这表明景区游客量不仅与自身搜索量有关，还受周边景区搜索量的显著影响。因此，全域旅游背景下应推动区域协同联动，尤其重视邻近景区的网络营销与信息共享。

(4)采用六种机器学习模型在游客量预测效果上呈现一定差异，比较发现，采用极限梯度提升对黄山景区游客量进行预测有优势，其测试集的均方根误差最小，预测效果最好。

## 参考文献


[1] 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.

[2] 任欢，刘婷，康俊锋，等. 一种基于百度指数的城市日游客规模预测方法[J]. 浙江大学学报(理学版)，2020，47(6)：753-761.

[3] 杨东方，高振会，于子江，等. ARIMA 模型在国内旅游市场预测中的应用[J]. 科学决策，2008(11)：151-152.

[4] 张娜，佟连军. 基于 SARIMA 模型的黑龙江省冰雪旅游国际需求预测[J]. 资源开发与市场，2012，28(7)：660-663.

[5] 欧蕾，张琼. 国内旅游客源市场的灰色预测模型——以河北省为例[J]. 燕山大学学报(哲学社会科学版)，2005(S1)：98-100.

[6] 杨蕊菡，温麒，农丽佳，等. 基于谷歌趋势的桂林入境游客量预测[J]. 北部湾大学学报，2024，39(2)：74-79.

[7] XU K，ZHANG J，HUANG J，et al. Forecasting visitor arrivals attourist attractions using the N-BEATS time series framework for sustainable tourism[J]. Sustainability，2024，16(18)：20711050.

[8] 陈荣. 基于支持向量回归的旅游短期客流量预测模型研究[D]. 合肥：合肥工业大学，2014.

[9] 陆利军，廖小平. 基于 EMD-BP 神经网络的游客量预测研究[J]. 统计与决策，2019，35(4)：85-89.

[10] ZHANG Y，TAN W H，ZENG Z. Tourism demand forecasting based on a hybrid temporal neural network model for sustainable tourism[J]. Sustainability，2025，17(5)：2210.

[11] 王雅，任静梅，王妍文，等. 基于统计与机器学习融合模型的水墨丹霞客流量预测优化研究[J]. 兰州交通大学学报，2025，44(6)：63-72.

[12] 柯蓉，王海源，余凯. 机器学习在旅游客流量预测中的应用研究——以丽江市为例[J]. 数学的实践与认识，2023，53(10)：71-80.

[13] 康梓蝶，李啸虎. 基于百度指数的天山天池景区客流量预测研究[J]. 职业技术，2024，23(2)：99-108.

[14] 伍银艳. 基于网络搜索数据的游客量预测研究[D]. 济南：山东工商职业学院，2023.

[15] 赵秋皓，金平斌，王冰冰，等. 基于多源数据与机器学习的乡村旅游竞争力评价研究——以杭州市临安区为例[J]. 地理科学进展，2023，42(8)：1541-1555.

[16] 杨富斌. 基于热经济性分析及人工神经网络建模的车用有机朗肯循环性能优化[D]. 北京：北京工业大学，2018.

[17] HUANG J W, XIAO Q T, LIU J J, et al. Modeling heat transfer properties in an ORC direct contact evaporator using RBF neural network combined with EMD [J]. Energy, 2019, 73: 306-316.

[18] KAZAKOV A, MUZNY C D, DIKY V, et al. Predictive correlations based on large experimental datasets: critical constants for pure compounds[J]. Fluid Phase Equilibria, 2010, 298(1): 131-142.

[19] ZHU T Y, TAO C C. Prediction models with multiple machine learning algorithms for POPs: the calculation of PDMS-air partition coefficient from molecular descriptor[J]. Journal of Hazardous Materials, 2022, 423: 127037.

[20] TAY Z Y. Predicting hydroelastic response of very large floating structures using decision tree methods: a comparative study of single regression tree and ensemble models (Bagging and LSBoost)[J]. Ocean Engineering, 2025, 341: 122636.

[21] ZHANG L F, LU G, YAN X Q, et al. A differential evolution optimized hybrid XGBoost for accurate carbon emission prediction[J]. Environmental Modelling & Software, 2025, 193: 106627.

# Research on Tourist Volume Prediction of Scenic Spot Based on Baidu Index and Machine Learning

PAN Hui

(School of Low-Altitude Economy, Huishang Vocational College, Hefei 230022, China)

**Abstract**: Based on Baidu Index and using Huangshan Scenic Area as the research subject, this study selects data from January 2014 to March 2024 to build visitor volume prediction models using six machine learning methods. The results indicate a strong correlation between Baidu Index and visitor volume, showing that the former can be a reliable predictor of the latter. Moreover, visitor numbers at a scenic spot are not only influenced by its own search volume but also by the search volume of nearby attractions, exhibiting a distance-decay spatial spillover effect. Tree-based models are effective in capturing the nonlinear relationship between Baidu Index and visitor volume, with the Extreme Gradient Boosting model standing out as the most performing method among the six machine learning methods.

**Keywords**: Baidu index; tourist volume prediction; machine learning; spatial spillover effect; Huangshan Scenic Area