Statistics and Application 统计学与应用, 2025, 14(3), 160-172
Published Online March 2025 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2025.143068

Hans汉斯

基于ARIMAX-GA-XGBoost组合模型的景区客流量分析

——以象鼻山景区为例

李浩清[1]，涂江韬[2]，胡新豪[2]，廖　秀[3*]

[1]桂林信息科技学院商学院，广西　桂林
[2]桂林信息科技学院电子工程学院，广西　桂林
[3]桂林信息科技学院数学教研部，广西　桂林

收稿日期：2025年2月14日；录用日期：2025年3月7日；发布日期：2025年3月18日

摘　要

旅游作为绿色经济推动了地区经济社会的发展。本文以象鼻山景区为例，利用百度指数分析游客对该景点的网络关注度，并针对单一模型对景区日客流量预测精度不足的问题展开研究。提出将ARIMAX模型与GA-XGBoost模型采用残差法进行组合，将数理统计模型和机器学习模型组合，实现优势互补，提高预测精度；首先使用ARIMAX对数据进行预测分析，称预测结果为 y_1，再把ARIMAX模型的残差放入XBGoost模型进行学习，基于GA算法对XGBoost的超参数进行优化，解决了ARIMAX模型难以对非线性数据预测的问题，GA-XGBoost的预测结果为 y_2，组合模型的最终预测结果 $y = y_1 + y_2$。最后，根据预测误差评价指标对多个模型进行对比。研究结果表明，ARIMAX-GA-XGBoost组合模型预测精度更高，适应性及泛化能力更强，可为旅游相关管理部门的科学决策提供必要的参考，具有很高的经济效益与实际意义。

关键词

旅游客流量预测，百度指数，ARIMAX，GA-XGBoost，残差法

Analysis of Tourist Flow in Scenic Areas Based on the ARIMAX-GA-XGBoost Combined Model

—A Case Study of Elephant Trunk Hill Scenic Area

Haoqing Li[1], Jiangtao Tu[2], Xinhao Hu[2], Xiu Liao[3*]

[1]Business School, Guilin University of Information Technology, Guilin Guangxi

文章引用: 李浩清, 涂江韬, 胡新豪, 廖秀. 基于 ARIMAX-GA-XGBoost 组合模型的景区客流量分析[J]. 统计学与应用, 2025, 14(3): 160-172. DOI: 10.12677/sa.2025.143068

[2]School of Electronic Engineering, Guilin University of Information Technology, Guilin Guangxi
[3]Mathematics Teaching and Research Department of Guilin University of Information Technology, Guilin Guangxi

Received: Feb. 14th, 2025; accepted: Mar. 7th, 2025; published: Mar. 18th, 2025

Abstract

Tourism, as a green economy, drives regional socioeconomic development. Taking Xiangbi Mountain Scenic Area as a case study, this paper analyzes the network attention of tourists towards this attraction using Baidu Index. To address the issue of insufficient prediction accuracy of single models for daily tourist flow forecasting in scenic areas, a hybrid modeling approach is proposed. By integrating the ARIMAX model with the GA-XGBoost model through residual combination methodology, this study combines mathematical-statistical modeling with machine learning techniques to achieve complementary advantages and enhance prediction accuracy. Specifically, the ARIMAX model is initially employed for data prediction (denoted as y_1), followed by feeding its residuals into the XGBoost model for learning. The genetic algorithm (GA) optimizes XGBoost's hyperparameters, effectively resolving ARIMAX's limitations in handling nonlinear data prediction (GA-XGBoost prediction denoted as y_2). The final combined model output is formulated as $y = y_1 + y_2$. Comparative analysis of multiple models through prediction error evaluation metrics demonstrates that the ARIMAX-GA-XGBoost hybrid model exhibits superior prediction accuracy, enhanced adaptability, and stronger generalization capabilities. This research provides valuable decision-making references for tourism management authorities and holds significant economic benefits and practical implications.

Keywords

Tourism Flow Forecasting, Baidu Index, ARIMAX, GA-XGBoost, Residual Method

1. 引言

第五次全国经济普查数据显示，旅游及相关产业实现增加值 54,832 亿元，占 2023 年 GDP 比重为 4.24%，比上半年提高 0.57 个百分点。随着旅游业的快速发展，旅游人次的飞速增长，相比传统的月度、年度预测，日客流量的预测研究愈发显得重要，尤其对一些著名的旅游景点做出科学决策提供了可靠的依据，对景区智能优化管理、旅游资源合理开发与保护等都具有重要作用和意义[1]。

景区每日客流量数据是典型的时间序列数据，本文先基于传统数理统计的思想，提出了时序分析模型(ARMA, ARIMA)、灰色预测模型，这些传统统计模型的优势为精确捕捉到数据中的线性规律部分[2]。例如邓等[3]与冯等[4]使用 ARMA 与 ARIMA 对景区客流量进行预测，取得了不错的精准度，再有严[5]等和王[6]等使用灰色理论对客流量进行预测，也取得了较高的预测精度。

但不容忽视的是景区的日客流量数据通常易受到天气、政策、交通、节假日等多个方面的影响，在具有线性时序规律的同时也具有较强的波动性和复杂非线性，依靠单一的传统统计模型难以精确捕捉其内在变化，且其不具备很强的自学性和泛化能力，预测系统的鲁棒性也无法保证。在此情况下本文

引入了一些机器学习模型来与传统统计模型进行组合，同时各个模型之间进行组合排序，依据对各个预测模型评价指标的比较，选出针对景区日客流量数据预测能力较强的预测模型。例如邓[7]使用 ARIMA-ATT-LSTM 组合模型对景区客流量进行预测，相比对上文引用的单一预测模型的预测精度和泛化能力有了显著提升，孔祥源[8]等使用 ARIMA 与 SVR 进行组合模型的构建，也取得了出色的预测精度和泛化能力。在景区游客量预测领域，部分研究者采用 BP 神经网络进行建模分析。然而该算法存在固有缺陷：在模型训练阶段，其参数更新具有随机性特征，可能导致模型收敛于局部最优状态，并伴随欠拟合或过拟合现象的发生，最终致使预测结果产生波动并影响模型精度。

针对以上研究的不足和模型特性，本文提出了一种基于 ARIMAX-XGBoost 的组合模型对象鼻山景区日客流量数据进行预测，弥补了单一模型的不足，解决了 ARIMAX 模型难以对非线性数据预测的问题，提高了景区日客流量预测模型的泛化能力和预测的精度。并使用遗传算法对 XGBoost 模型进行参数寻优，进一步提高模型性能与泛化能力。

2. 研究方法

2.1. ARIMAX 模型

传统 ARIMA 模型在时间序列分析中存在一定局限性，因其仅能针对单变量数据进行建模和预测。但实际应用场景中，目标变量的变化往往受到多重外部因素的影响，这使得单纯依赖 ARIMA 模型难以获得理想的预测效果[1]。为克服这一缺陷，学界提出了多元时间序列分析方法，其中具有代表性的动态回归模型 ARIMAX 受到广泛关注。

Box 等学者在构建 ARIMAX 模型时强调，模型所涉及的因变量与各解释变量均需满足平稳性条件，否则可能引发统计推断中的伪回归现象。针对这一约束条件，Engle 等提出的协整理论提供了解决思路：当多个非平滑变量之间存在长期均衡关系时，即使各变量本身不平稳，只要其线性组合形成的残差序列保持平稳，即可有效避免伪回归问题[9]。这一理论突破不仅完善了计量经济学方法论体系，更显著提升了多元时间序列模型的预测效能，也推动该方法在其他领域的发展与应用。

设输入变量序列为$\{x_{1t}\},\{x_{2t}\},\cdots,\{x_{kt}\}$，输出变量序列为$\{y_t\}$，建立 ARIMAX 模型，结构为：

$$\begin{cases} y_t = \mu + \sum_{i=1}^{k} \dfrac{\Theta_{x_i}(B)}{\Phi_{x_i}(B)} B^{l_i} x_{it} + \varepsilon_t \\ \varepsilon_t = \dfrac{\Theta(B)}{\Phi(B)} a_t \end{cases}$$

式中：μ 是一个常数；$\Theta_{x_i}(B)=\theta_{i0}-\theta_{i1}B-\theta_{i2}B^2-\cdots-\theta_{iq_i}B^{q_i}$，$1\le i\le k$，为第 i 个输入变量的 q_i 阶移动平均系数多项式；$\Phi_{x_i}(B)=1-\varphi_{i1}B-\varphi_{i2}B^2-\cdots-\varphi_{ip_i}B^{p_i}$，$1\le i\le k$，是第 i 个输入变量的 p_i 阶自回归移动多项式；B^s 是回归残差序列$\{\varepsilon_t\}$的 p 阶自回归系数多项式；l 为 B 的指数：$\Theta(B)$是残差序列移动平均系数多项式；$\Phi(B)$为残差序列自回归系数多项式；$\{a_t\}$为零均值白噪声序列[2]。

ARIMAX 模型的建模过程为：

1) 对输出变量$\{y_t\}$和第 i 个输入变量序列$\{x_{it}\}$进行 ADF 单位根平稳性检验。若$\{x_{it}\}$非平稳，则对$\{x_{it}\}$建立 ARIMA 模型，产生白噪声序列$\{\varepsilon_{x_{it}}\}$，具体公式为[2]：

$$\varepsilon_{x_{it}} = \frac{\Theta_{x_i}(B)}{\Phi_{x_i}(B)} \nabla^{d_i} x_{it}, \quad i=1,2,\cdots,k$$

2) 对$\{y_t\}$进行同样的变化，得到白噪声序列$\{\varepsilon_{y_t}\}$，具体表达式如下：

$$\varepsilon_{y_t}=\frac{\Theta_y(B)}{\Phi_y(B)}\nabla^d y_t,\quad i=1,2,\cdots,k$$

3) 通过观测序列$\{\varepsilon_{x_{it}}\}$与$\{\varepsilon_{y_t}\}$的互相关系数，确定 ARIMAX 模型的结构，具体表达式如下：

$$y_t=\mu+\sum_{i=1}^{k}\frac{\Theta_{x_i}(B)}{\Phi_{x_i}(B)}B^{l_i}x_{it}+\varepsilon_t$$

2.2. XGBoost 模型

XGBoost 是一种基于梯度提升决策树的并行机器学习算法，特点在于对目标函数进行二阶泰勒展开并加入正则化约束。本研究将运用该算法构建象鼻山景区日游客量预测模型。以下是对 XGBoost 算法理论基础的阐述。

1) XGBoost 的目标函数包含两部分：训练损失函数和正则化项，表达式如下所示：

$$Obj=\sum_{i=1}^{n}\iota\left(y_i,\widehat{y_i^t}\right)+\sum_{i=1}^{t}\Omega(f_i)=\sum_{i=1}^{n}\iota\left(y_i,\widehat{y_i^{(t-1)}}+f_t(x_i)\right)+\Omega(f_t)+C$$

在上面的公式里，等号右侧的首项对应训练样本的损失函数，负责对应训练数据的拟合误差进行评估，次项则为正则化项，负责模型结构的复杂度控制。其中，预测值y_i表示模型对输入样本x_i的响应输出，ι是损失函数。以回归任务为例，最常采用的误差评估方法包括均方误差(MSE)和平均绝对误差(MAE)等[1]。f_t表示第t轮训练的树模型。$\Omega(f_i)$函数则量化了个基学习器的复杂度，正则化项即为此类复杂度指标的累加值。值得注意的是，在增量式训练过程中，因为前$t-1$棵树的结构已经完全确定，其对应的的正则化参数可视为参数C。因此第t棵树要学习的目标函数则为上式第二个等号后面所示的前两项组合[1]。

2) 利用二阶泰勒展开式展开损失函数，如下式所示。

$$\left(y_i,y_i^{(t-1)}+f_t(x_i)\right)\approx\iota\left(y_i,y_i^{(t-1)}\right)+g_i f_t(x_i)+\frac{1}{2}h_i f_t^2(x_i)$$

公式中g_i和h_i分别是$\iota\left(y_i,y_i^{(t-1)}+f_t(x_i)\right)$的一阶导数和二阶导数。$\iota\left(y_i,y_i^{(t-1)}\right)$表示前$t-1$轮的训练损失，由于它对当前第$t$轮的训练损失没有影响，因此可以视为常量。由于常量不影响本轮训练结果，可以将其移除，最终目标函数值Obj为[1]：

$$Obj\approx\sum_{i=1}^{n}\left[g_i f_t(x_i)+\frac{1}{2}h_i f_t^2(x_i)\right]+\Omega(f_t)$$

3) 树的定义为叶子节点权重向量ω和叶子节点的映射关系q，q表达的是树的分支结构，如下式所示。

$$f_t(x)=\omega_{q(x)},\omega\in R^T,q:R^D\to\{1,2,\cdots,T\}$$

公式中ω度为T的一维向量，它的值是叶子节点的权重；q函数描述基学习器的拓扑结构，通过决策路径将输入样本划分至特定叶单元，这里假设这棵树有T个叶子节点[1]。

4) 树的复杂度，包括叶的节点的个数T和叶子节点权重向量ω_j的 L2 范数。

$$\Omega(f_t)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2$$

5) 由以上两个公式可以得到新的目标函数，如下式所示。

$$\begin{aligned} Obj &\approx \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \\ &= \sum_{i=1}^{n}\left[g_i \omega_{q(x_i)} + \frac{1}{2} h_i \omega_{q(x_i)}^2 \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \\ &= \sum_{j=1}^{T}\left[\left(\sum_{i:q(x_i)=j} g_i \right) \omega_{q(x_i)} + \frac{1}{2}\left(\sum_{i:q(x_i)=j} h_i + \lambda \right) \omega_j^2 \right] + \gamma T \end{aligned}$$

公式中 I_j 为每个叶子节点上的样本集合。

给定 C_i 和 H_j 的计算方法，如下式所示。

$$C_j = \sum_{i \in I_j} g_i, H_j = \sum_{i \in I_j} h_i$$

6) 将 C_i 和 H_j 的计算方法公式放入到新的目标函数中，再次得到新的目标函数，如下式所示。

$$Obj^{(t)} = \sum_{j=1}^{T}\left[C_j \omega_j + \frac{1}{2}\left(H_j + \lambda \right) \omega_j^2 \right] + \gamma T$$

对上述公式求导得最值点和最优目标函数，分别如下式所示。

$$\omega_j^* = -\frac{C_j}{H_j + \lambda}$$

$$Obj^* = -\frac{1}{2}\sum_{j=1}^{T}\frac{C_j^2}{H_j + \lambda} + \lambda T$$

7) 找出最佳分裂节点。

在 XGBoost 的弱学习器构建过程中，算法采用逐步最优的决策路径生成机制。具体而言，当执行特征空间划分时，每个候选分裂点都会触发目标函数的改进程度计算模块，通过严谨的数学评估量化当前划分对模型整体性能的优化贡献度。若改进值为正，则表明目标函数得到优化。用户可预先设定一个改进阈值，一旦达到该标准，训练过程将自动终止。目标函数的改进值计算公式如下所示：

$$\begin{aligned} Gain &= Obj_{L+R} - \left(Obj_L + Obj_R \right) \\ &= \left[-\frac{1}{2}\frac{G_{L+R}}{H_{L+R} + \lambda} + \lambda \right] - \left[-\frac{1}{2}\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} + 2\lambda \right] \\ &= \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{\left(G_L + G_R \right)^2}{H_{L+R} + \lambda} \right] - \lambda \end{aligned}$$

2.3. 基于遗传算法的 XGBoost 模型参数寻优

遗传算法是一种模拟生物遗传机制的智能优化算法，其核心思想源于自然界的选择与进化过程。该算法支持包括二进制在内的多种编码方案，其中 0~1 编码最为基础。在标准实现中，每个解被编码为二进制数字串，各位数值表征特定特征属性。算法通过适应度函数评估解的质量，依概率筛选优质个体。在重组阶段，两个父代个体的编码片段进行交换组合，生成新的子代个体。为避免陷入局部最优，算法会以预设概率对编码位进行随机翻转，维持种群多样性。

标准遗传算法的数学模型可以表示为：

$$SGA = f\left(C, E, P_0, N, \Phi, \Psi, \Gamma, T\right)$$

上式中，C 表示个体编码方法；E 表示个体适应度评价函数；P_0 为初始种群；N 为种群大小；Φ 为选择算子；Γ 为交叉算子；Ψ 为变异算子；T 为遗传算子终止条件[1]。遗传算法优化 XGBoost 模型参数的主要步骤如下：

Step 1：导入 ARIMAX 模型的残差数据集，按预设比例划分训练集与测试集；

Step 2：对特征变量实施标准化预处理；

Step 3：首先，初始化 XGBoost 回归模型，并配置基础超参数，包括学习率、树的最大深度、子样本比例等，以确保模型具备良好的起点和灵活性；

Step 4：定义遗传算法参数(编码方式、种群规模、交叉率及变异率)；

Step 5：解码染色体生成超参数组合，计算种群个体的适应度(模型预测性能)；

Step 6：若适应度满足终止条件(如误差阈值或迭代上限)，跳转至 Step9；否则执行 Step7；

Step 7：基于适应度筛选父代染色体，通过交叉重组与随机变异生成新种群；

Step 8：重复 Step 5~Step 7 直至满足终止条件；

Step 9：加载优化后的超参数训练 GA-XGBoost 模型，输出测试集预测结果。

2.4. 模型评价准则

预测模型的性能评估指标采用平均绝对值误差(MAE)和平均方根误差(RMSE)以及平均绝对值百分比误差(MAPE)，以评估模型的预测精确度。

预先假设预测值：$\hat{y} = \{\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n\}$ 真实值：$y = \{y_1, y_2, \cdots, y_n\}$，相关的计算原理为：

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

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{y}_i\right)^2}$$

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right|$$

2.5. ARIMAX-GA-XGBoost 模型构建思路

本文提出的 ARIMAX-GA-XGBoost 混合预测模型旨在结合传统时间序列分析与现代机器学习方法的优势，以提升预测精度与鲁棒性。模型构建过程分为以下几个关键步骤：首先，利用 ARIMAX 模型捕捉时间序列的线性特征及外部变量的影响；其次，通过遗传算法(GA)优化机器学习模型参数，以提高预测性能；最后，引入 GA-XGBoost 进一步增强模型非线性特征的提取能力。具体步骤包括数据预处理、模型训练、参数优化及结果验证等环节，以确保模型的科学性与实用性。整个流程如下图 1 所示。

3. 实证分析

3.1. 数据来源

本研究以 2022 年 2 月 1 日至 2025 年 1 月 1 日期间“象鼻山”关键词的百度搜索指数为客流预测数据来源。鉴于该景区自 2022 年春节(1 月 31 日)后实行免票政策，为规避政策调整对数据完整性的干扰，特将样本起始点设定于免票实施次日。百度指数通过统计百度搜索引擎中关键词的搜索频次，作为衡量网络搜索热度的指标，综合加权计算得出，可以客观地反映出特定主题的大众关注度变化趋势。为全面

捕捉用户行为特征，本研究采用 PC 端与移动端搜索数据的综合指标，避免单一终端可能产生的统计偏差。

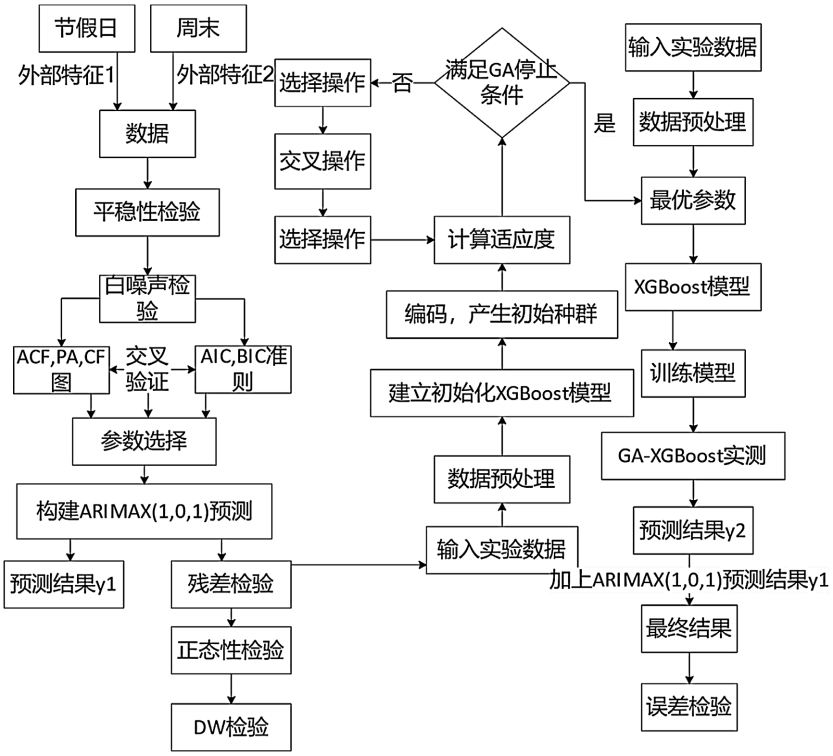

Figure 1. Flowchart of the ARIMAX-GA-XGBoost Hybrid Forecast Model
图 1. ARIMAX-GA-XGBoost 预测模型流程图

3.2. 数据预处理

3.2.1. 特征工程

景区的日客流量变化受到多种客观因素的影响，其中最为主要的还是周末和节假日这两大因素。为了检验这两个因素对景区日客流量的影响，首先查找出 2022 年至 2025 年的放节假日并标记在序列图中，如下图 2 所示。

其中标红的为 2022 年至 2025 年中国法定节假日，可以看出在标红处(节假日)数据明显高于其他非节假日数据，在五一和国庆等假期，游客数量显著增加，平均为 2070 人，标准差为 2805，显示出游客数量的极大波动，假期期间的游客数量可达几千人，远高于平日。说明节假日可以作为一个外部特征放入 ARIMAX 模型。对于周末数据，由于数据量过于庞大且波动较小，在趋势图里也不易直接发现特征，于是引入标准差进行特征检验，周末平均游客数量为 671 人，标准差为 203，说明周末的游客数量有一定波动，且一般高于平日。对于既是周末又是节假日的数据视为节假日数据，本文引入周末与节假日两个外部特征放入 ARIMAX 模型进行训练，提高预测精度。

3.2.2. ADF (Augmented Dickey-Fuller)平稳性检验

在时间序列分析的自回归建模框架下，当滞后项系数等于 1 时，系统呈现单位根特性。这种统计特性将导致残差序列具有非衰减记忆效应，其方差不会随观测数据量的增加而收敛，从而在变量间产生统计意义上的伪相关关系，这种现象在计量经济学中被称为虚假回归效应。若存在单位根，该过程即为随机游走[10]。

为识别此类非平稳特征，研究者常采用 ADF 检验方法，该检验通过构建统计量判断序列的平稳性：当拒绝原假设(存在单位根)时，可确认序列具有平稳性；若接受原假设，则表明序列存在单位根导致的趋势成分。基于 Python 的测试结果如表 1 所示：

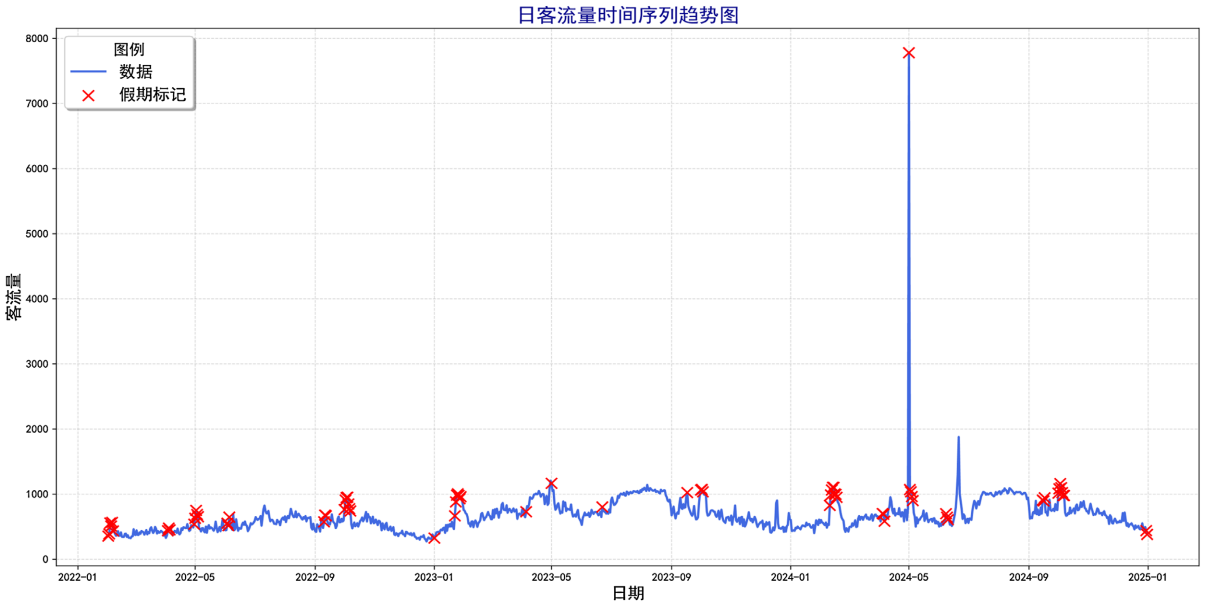

Figure 2. Time series graph of daily tourist passenger flow
图 2. 旅游日客流量时间序列图

Table 1. Augmented Dickey-Fuller (ADF) test table
表 1. ADF 检验表

置信度	临界值	ADF 检验统计量	p 值
10%	−2.56		
5%	−2.86	−3.92	0.001877
1%	−3.43		

ADF 检验统计量(−3.922233)小于 1%临界值(−3.436593)，而且 p 值(0.001877)小于常用的显著性水平(如 0.05)。由于检验统计量小于临界值，可以拒绝原有的假设(序列存在单位根，即非平滑序列)，因此该序列是平稳的。本文 ARIMAX 模型差分阶数为 0 即可。

3.2.3. 白噪声 Ljung-Box 检验

Ljung-Box 检验是用于检测时间序列中的自相关性。如果序列是白噪声，那么该序列中的各个滞后项应该不相关(即没有自相关)。具体检验结果如下表 2 所示：

Table 2. Ljung-Box test results table
表 2. Ljung-Box 检验结果表

滞后阶数	Ljung-Box 统计量	p 值
1	234.582123	5.97e−53
2	428.343086	9.69e−94
3	609.716592	7.89e−132

在 Ljung-Box 统计检验框架下，各滞后阶数的显著性分析显示，所有检验统计量对应的概率值均显著低于预设阈值(α = 0.05)。这一统计推断结果拒绝了序列服从白噪声过程的原假设，证实存在显著的时间序列自相关结构。检验结论表明该数据序列具有可辨识的动态依赖模式，其统计特性与随机无关联序列存在本质差异。

3.3. 建模分析

3.3.1. ARIMAX 模型参数选择

首先利用 ARIMAX 模型的自相关图(ACF)和偏自相关图(PACF)进行初步分析，基于 Python 实现如下图 3 所示：

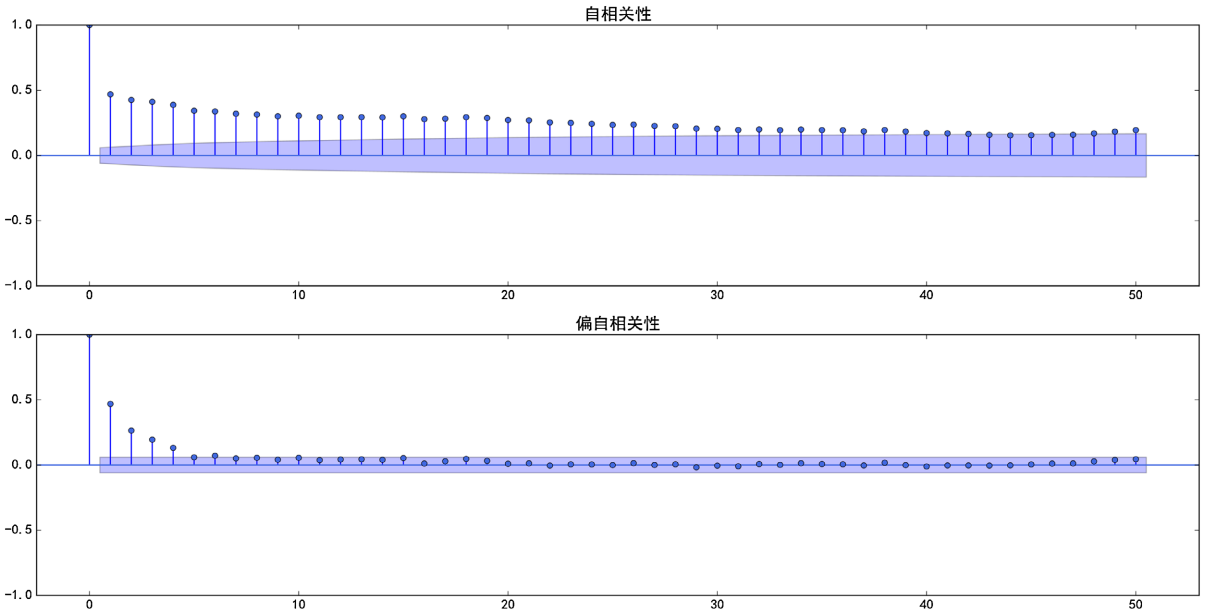

Figure 3. Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) graphs
图 3. 自相关图(ACF)和偏自相关图(PACF)

根据 ACF 和 PACF 图的组合，可以考虑使用 ARIMAX (1, 0, 1)模型。ACF 显示该序列存在较强的滞后 1 的自相关，PACF 在滞后 1 上也显示较强的偏自相关，且滞后 2 之后衰减至零，暗示该模型可能是 ARIMAX 模型，阶数为 1。参数的正确选择是时间序列预测模型能够精确预测的重要前提，于是在自相关图和偏自相关图的阶数初步选择下，使用 AIC (赤池信息量准则)和 BIC (贝叶斯信息量准则)来进行参数寻优选择，并与上面的初步判断进行参数选择的交叉验证。最终得出最佳 AIC 为 14,743.31，最佳 BIC 为 14,763.20，最佳 ARIMAX 模型阶数为(1, 0, 1)，与上文的自相关图和偏自相关图的初步判断结果一致，交叉检验通过，基于 ARIMAX (1, 0, 1)开始预测建模与分析。

3.3.2. ARIMAX 模型预测结果

基于 ARIMAX (1, 0, 1)预测结果如下图 4 所示：

从图中可见，训练集的预测结果(红色曲线)与实际客流量(蓝色曲线)较为接近，表明模型在训练集上具有较好的拟合效果。此外，下图 5 的残差部分的正态性检验进一步从侧面展示了 ARIMAX 模型的预测性能。

可以看出 ARIMAX 模型的残差基本上符合正态分布的特征，残差接近正态分布意味着模型已很好地

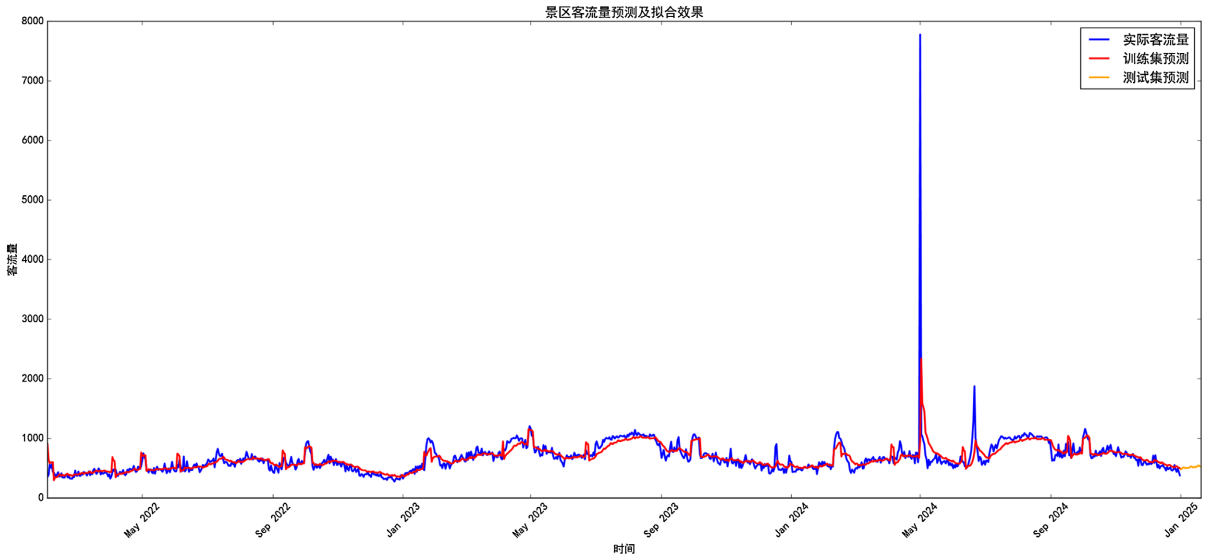

Figure 4. Tourist attraction passenger flow prediction and fitting performance graph
图 4. 景区客流量预测及拟合效果图

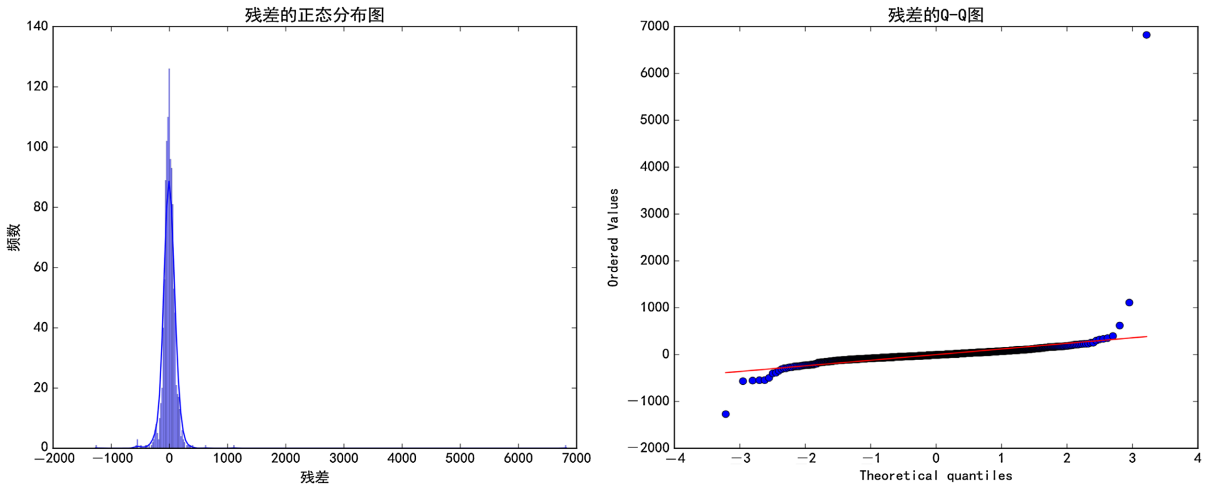

Figure 5. Normal distribution plot and Q-Q plot of residuals
图 5. 残差的正态分布图和 Q-Q 图

捕捉了数据的模式，未能捕捉到的数据(即残差)呈现出一种理想的分布。这是统计建模中一个常见的假设，特别是在许多假设检验和回归分析中，残差服从正态分布是理想的。并且 DW 检查值非常接近 2，说明残差自相关性不大，表明模型拟合很好，这个结果是理想的，意味着模型已充分捕捉了数据的时序性特征，没有遗漏自相关性的信息。

3.3.3. GA-XGBoost 建模

然而，尽管 ARIMAX 模型的残差没有自相关性，并且没有显著的时间依赖性，仍然可能存在一些复杂的非线性关系或影响因素，这些因素可能没有被 ARIMAX 模型捕捉到。此时，将这些残差作为输入特征放入 XGBoost 等机器学习模型进行预测，依然可能是有意义的。XGBoost 作为一种强大的树模型，可以捕捉到非线性关系，可能会进一步提取出残差中的某些潜在模式，从而提高预测性能。

本文基于 Python 的遗传算法工具箱 DEAP 对 XGBoost 模型的超参数进行优化择优。DEAP 箱易于操

作，适合进行遗传算法优化参数选择的问题。实验中 GA-XGBoost 模型参数所使用的 Python 库和函数详见表 3。

Table 3. Table of Python libraries, functions, and their implemented functions
表 3. 所使用的 Python 库和函数及实现功能表

Python 库和函数	实现功能
import,numpy,pandas	读取实验数据
fromsklearn.preprocessingimportMinMaxScaler,scale	数据归一化
fromxgboostimportXGBRegressor	XGBoost 模型
fromsklearn.metricsimportmean_absolute_error	MAE 函数
importdeap	遗传算法模块

首先对数据进行预处理和标准化，以消除量纲差异，并将 2022 年至 2025 年的前百分之 80%的数据作为训练集，后 20%作为测试机。采用遗传算法对 XGBoost 模型进行超参数选择优化，具体步骤如下：

Step 1：初始化 XGBoost 参数。

针对 XGBoost 集成学习模型，设定核心参数优化空间：树结构深度 max_depth ∈ [3, 10]，学习率 learning_rate ∈ [0.01, 0.2]，基学习器数量 n_estimators ∈ [50, 300]。为防止过拟合设置最大训练轮次阈值 maxTrappedCount = 10，初始化基准参数组合为(max_depth = 5, learning_rate = 0.1, n_estimators = 100)。

Step 2：配置遗传算法参数。

建立包含五维超参数的优化体系，个体基因编码涵盖 n_estimators、max_depth、learning_rate、样本抽样率 subsample 和特征采样率 colsample_bytree。通过遗传进化机制实现多维参数空间的协同优化，以寻找最优参数组合。

Step 3：采用 K 折交叉验证的 MAE 指标倒数作为适应度评价函数(Fitness = 1/MAE)，通过锦标赛选择、两点交叉和动态变异等进化操作迭代优化参数组合。

经多代进化后获得最优参数配置：n_estimator s= 30，max_depth = 5，learning_rate = 0.1941，subsample = 0.8，colsample_bytree = 0.9。将该参数集应用于时序残差预测模型，经反标准化处理后获得未来 20 步预测结果，GA 优化后的 XGBoost 模型具有良好的预测效果，具体 GA-XGBoost 对残差序列拟合结果图如下图 6 所示。

历史残差(蓝色)与预测残差(绿色)高度吻合，两条曲线在大部分时间步内几乎重叠，表明模型在训练集上能够有效捕捉数据规律，MAE 值优化显著。训练阶段残差值集中在[−500, 500]区间内，未出现极端偏差，说明模型稳定性较好。预测残差未出现明显偏离或发散，与历史残差波动范围基本一致，表明模型具备一定的泛化能力。

3.4. 模型结果对比

将本文构建好的模型往后预测 20 个时间步长，与真实值进行比较，基于 2.4 的模型评价准则进行误差评价。最后将本文模型误差结果与其他模型误差结果进行综合比较，得出的各模型平均误差值如下表 4 所示。

结合外部特征的 ARIMAX 模型预测精度普遍高于 ARIMA 模型。可见结合周末与节假日两个外部特征显著提高了模型的预测精度。引入 GA 优化后，基于机器学习的混合模型表现较好，特别是 BPNN 和 XGBoost。由此可见，合理选择超参数对机器学习的预测精确度是必不可少的。综合来看，ARIMAX-GA-BPNN 是表现最好的模型，具有最低的 MAPE、RMSE 和 MAE。说明它可以比较准确的拟合数据，误差

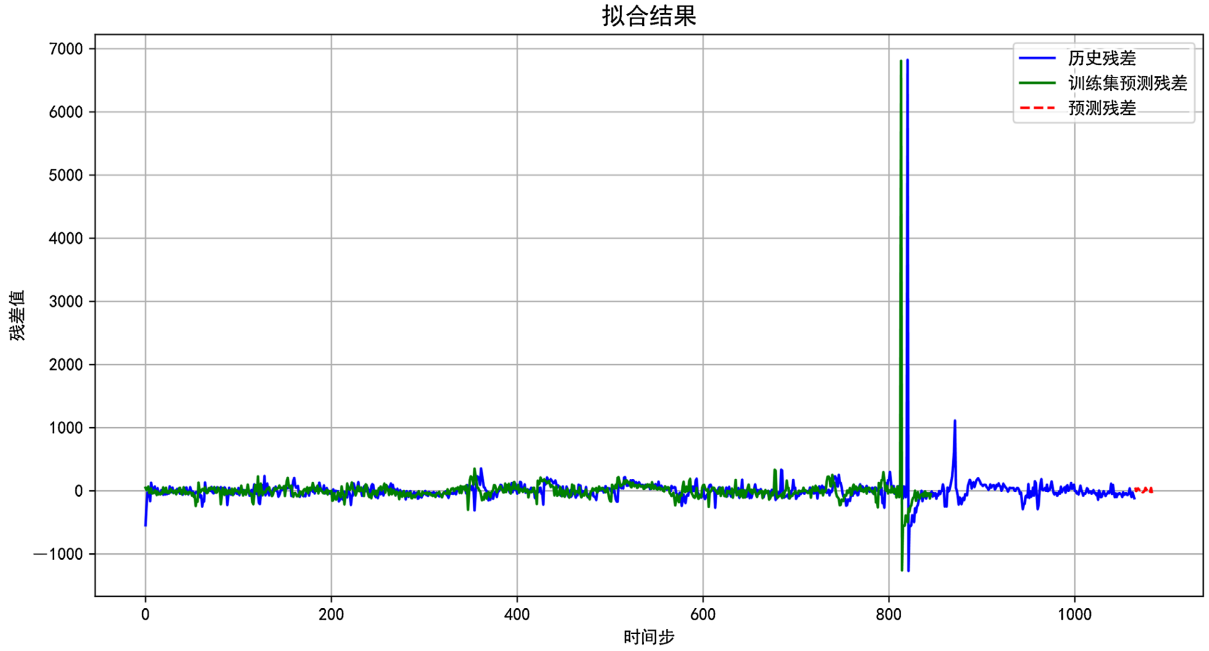

Figure 6. Fitting results graph of residual series by GA-XGBoost
图 6. GA-XGBoost 对残差序列拟合结果图

Table 4. Comparison table of mean error values across different models
表 4. 各模型平均误差值比较表

模型	MAPE	RMSE	MAE
ARIMA	12.85	66.49	60.18
ARIMAX	8.73	47.28	40.7
ARIMAX-BPNN	7.03	45.66	35.71
ARIMAX-GA-BPNN	6.42	34.88	30.77
ARIMAX-XGBoost	6.85	41.52	33.73
ARIMAX-GA-XGBoost	6.13	34.1	29.52
ARIMAX-LSTM	10.42	55.06	48.94
ARIMAX-CNN-LSTM	10.02	55.59	47.07

也比较小。ARIMAX-GA-BPNN 模型在所有误差指标中最为优秀，推荐作为景区日客流量预测的最优选择。

4. 结论

在本研究中，结合数理统计模型与机器学习方法，提出了一种基于 ARIMAX-GA-XGBoost 组合模型的景区日客流量预测方法，并应用于象鼻山景区。实验结果表明，单一的 ARIMA 模型无法充分捕捉景区日客流量数据中的复杂非线性特征，而通过引入具有外部特征变量的 ARIMAX 模型与集成机器学习 XGBoost 模型，将两者基于残差法进行组合，有效的解决了这一问题，且增强了模型的泛化能力。同时，采用遗传算法(GA)优化 XGBoost 模型的超参数，进一步增强了模型的预测性能。综合来看，ARIMAX-GA-XGBoost 组合模型在 MAPE、RMSE、MAE 等误差指标上均优于其他模型，特别是在处理复杂数据和非线性关系方面具有显著优势。因此，本研究提出的组合模型不仅为景区管理者提供了高效的客流量预

测工具，也为旅游行业的智能化管理提供了现实可行的解决方案，具有较高的经济和社会价值。

基金项目

2024 年大学生创新创业项目：基于传统统计模型与机器学习模型的组合模型的景区客流量分析——以象鼻山景区为例(编号：S202413644020)。

参考文献

[1] 李顺, 李君, 吴鑫, 梅碧舟. 基于 GA-XGBoost 的宁波港物流需求预测[J]. 浙江万里学院学报, 2021, 34(2): 71-77.

[2] 赵盼盼, 张迷, 焦力宾. 基于 ARIMAX 模型的中国税收收入预测与分析[J]. 高师理科学刊, 2024, 44(5): 22-27.

[3] 邓慧琼, 陈怀娜, 曾毓芬, 连宗胜, 周燕. 基于小波分析和 ARIMA 模型的假期客流量预测分析[J]. 中国新通信, 2019, 21(21): 60-61.

[4] 冯玉香. ARMA 模型在景区客流量预测中的应用[J]. 浙江统计, 2008(10): 8-10.

[5] 严安, 杨雨琪, 蒋鑫阳, 莫寒翔, 段博雅. 基于灰色预测和 BP 神经网络的高校食堂客流量预测与调度——以华北电力大学为例[J]. 自动化应用, 2022(3): 56-59.

[6] 王转敏. 基于灰色理论的城市轨道交通客流量预测研究——以兰州市轨道交通 1 号线为例[D]: [硕士学位论文]. 兰州: 兰州交通大学, 2022.

[7] 邓雨菲. ARIMA-ATT-LSTM 在旅游客流量预测中的应用研究[D]: [硕士学位论文]. 大连: 大连理工大学, 2022.

[8] 孔祥源. 基于 ARIMA 和 SVR 模型的景区客流量分析——以四姑娘山景区为例[D]: [硕士学位论文]. 桂林: 广西师范大学, 2021.

[9] Engel, R.F. and Granger, C.W.J. (1987) Cointegration and Error Correction: Representation, Estimation and Testing. *Econometrics*, **55**, 251-276.

[10] 卢佳. 中长期市场下梯级水电运行风险分析方法研究[D]: [博士学位论文]. 大连: 大连理工大学, 2021