doi:10.3969/j.issn.1672-1292.2023.02.011

基于 EMD-VAR 模型的景区客流波动特征与预测研究——以南京夫子庙为例

丁　洁[1,2],丁春媚[3],张建新[3]

(1.南京旅游职业学院旅游管理学院,江苏 南京 211100)
(2.南京师范大学地理科学学院,江苏 南京 210023)
(3.南京大学地理与海洋科学学院,江苏 南京 210023)

[摘要]　网络搜索大数据为研究游客量预测提供了新的视角,而多数研究运用的传统计量经济模型难以处理网络搜索与客流时序中包含的大量非线性波动特征,导致预测精度不够理想. 引入经验模态分解方法(empirical mode decomposition,EMD)将向量自回归(vector autoregression,VAR)模型改进为 EMD-VAR 模型. EMD 方法分解夫子庙景区长三角日际网络搜索和游客量序列,得到不同频率尺度的分量,基于波动关联的视角将同一尺度的两类序列分量组合建立 EMD-VAR 模型进行预测. 结果表明:(1)网络搜索波动周期比游客量波动周期长. (2)网络搜索与游客量波动的关联紧密度在法定节假日时期最高. (3)EMD-VAR 模型比 ARMA 模型和 VAR 模型具有更高的预测精度.

[关键词]　百度指数,波动关联,即时预测,经验模态分解,EMD-VAR 模型
[中图分类号]F592.7　[文献标志码]A　[文章编号]1672-1292(2023)02-0077-10

Nowcasting Tourist Flow Volume of Tourist Attraction Based on the EMD-VAR Model:A Case Study of Nanjing Confucius Temple

Ding Jie[1,2],Ding Chunmei[3],Zhang Jianxin[3]

(1.School of Tourism,Nanjing Institute of Tourism and Hospitality,Nanjing 211100,China)
(2.School of Geography,Nanjing Normal University,Nanjing 210023,China)
(3.School of Geography and Ocean Science,Nanjing University,Nanjing 210023,China)

Abstract:Big data from network search provides a new perspective for the study of tourist flow volume prediction,but the traditional econometric models used in most studies are difficult to deal with the large number of nonlinear fluctuation characteristics in the timing series of network search and tourist flow,which leads to the unsatisfactory prediction accuracy. In this paper,empirical mode decomposition(EMD)is introduced to improve the vector autoregression(VAR)model to EMD-VAR model. EMD method is used to decompose the daily network search data and tourist flow volume of The Yangtze River Delta of Nanjing Confucius Temple Scenic Area,and a series of components with different frequency scales are obtained. Then,based on the perspective of fluctuation correlation,components of both network search data and tourist flow volume in the same scale are combined to establish a VAR model for prediction. The results show that:(1)The fluctuation cycle of network search is longer than that of tourist flow volume. (2)The compactness of correlation between network search and tourists flow volume is the greatest during the statutory holiday period. (3)The prediction accuracy of the EMD-VAR model is better than that of ARMA model and VAR model,respectively.

Key words:Baidu index,fluctuation correlation,nowcasting,empirical mode decomposition,EMD-VAR model

节假日时期旅游目的地客流量井喷与旅游危机和突发事件带来的冷清,常常对旅游目的地的生态和服务系统造成巨大的冲击,严重影响旅游业的可持续发展. 随着互联网的迅速发展,旅游者愈加依赖互联

收稿日期:2022-08-12.
基金项目:教育部人文社会科学研究一般项目(22YJA760106)、江苏省高职院校教师专业带头人高端研修项目(2022GRFX034)、江苏省高校哲学社会科学研究项目(2022SJYB0854).
通讯作者:张建新,硕士,副教授,研究方向:旅游地理和区域规划. E-mail:bokaimiwu@nju.edu.cn

网搜索引擎获取旅游信息以辅助旅游决策[1-2],这些搜索数据能够映射用户的关注意图和行为趋势[3],为游客量的预测提供了新的视角. 如何利用大数据精准预测客流量成为新常态下景区管理与发展的重要议题.

在线旅游信息流与旅游客流是旅游者的旅游行为分别在网络空间和地理空间中的投影,具有时序性和波动性等重要特征[4]. 这两种非线性时间序列内部蕴含的不同模式和关联因素增加了客流趋势预测的复杂性[5]. 当前,国内外基于网络搜索数据的游客量预测实证研究主要有两大视角:一是从数据端将网络搜索整体作为解释变量结合历史客流数据构建预测模型,主要涉及时间序列模型[6-14]、引力模型[15]、计量经济模型[16-20]等传统模型方法. 二是从方法端基于网络搜索数据探索预测模型的改进及组合,弥补不同预测模型的缺陷,其中以人工神经网络的组合模型居多[21-23].

目前,国内外学者在网络搜索大数据背景下对旅游需求预测做了大量研究,已取得一定成果,与此同时也存在一些不足. 在预测思路上,未能同时考虑网络搜索与客流数据内部的非线性趋势与波动关联性[13]. 传统计量经济模型如向量自回归(vector autoregression,VAR)模型能够很好地捕捉系统中网络搜索变量与游客量变量间的相互影响关系而被证实具有较好的预测能力[17],但忽略了时间序列的非线性趋势对预测效果的影响[24]. 经验模态分解(empirical mode decomposition,EMD)方法能够分层提取非线性数据中包含的重要关联特征,因此对于提升预测精度具有重要作用[25],然而现有研究仅将其应用在基于历史客流或网络搜索单维数据的游客量预测中[24,26]. 在研究尺度上,以月度[11,22-23]、全国[19-20]和入境[6,10,14]等大的时空尺度中游客量预测为主,基于日际数据的区域性旅游需求预测研究较为匮乏,更细时空尺度的预测模型可以提高客流量预测的峰值精度[26-29]. 基于此,本文以南京夫子庙景区为例,选取长三角地区客源为研究区域,采用颗粒度更精细的日际网络搜索和客流数据,基于波动关联的视角建立改进的 EMD-VAR 模型对区域游客量进行预测,并与传统 ARMA 模型和单一 VAR 模型对比预测效果,旨在为景区客流预测和管理提供新思路.

1 数据来源与研究方法

南京夫子庙景区地处长三角区域,是全国著名的历史文化街区. 据南京市文化和旅游局统计,夫子庙景区各法定节假日旅游接待人次在南京市总接待量中占比超过 60%. 兼具丰富多样的自然风光及文化旅游资源,受季节性因素的限制较小,具有稳定的客源及网络搜索. 游客对旅游目的地著名景区的网络搜索一定程度上能够反映游客拟到访景区的出游需求,具有较强的可预测性[17],故选择夫子庙景区作为研究案例具有代表性和典型性. 为了减小游客网络搜索行为的地域差异,划定当日可到达和游览夫子庙景区的长三角地区(包含江苏省、安徽省、浙江省和上海市)为研究范围.

1.1 数据来源

本文的研究数据包括游客量数据和网络搜索数据. 其中游客量数据来源于南京智慧旅游大数据监测平台,该平台可以通过对移动通讯设备实时追踪游客活动的地理信息,因此提供的数据相较于历史统计数据具有规模大、针对性高、时效性强的优势[25,29]. 以景区内部基站地理围栏内停留信令间隔超过 30 min 的非本市用户作为游客识别的标准,本文利用 2017 年 1 月 1 日~2019 年 12 月 31 日(共计 1 095 d)的夫子庙景区日际游客量数据. 本文网络搜索数据运用百度指数来表征. 百度是全球最大的中文搜索引擎,百度指数通过加权计算用户在百度网页搜索的各个关键词频次,展示用户对关键词的关注程度及持续变化情况. 综合考虑旅游者收集旅游信息涉及的“吃、住、行、游、购、娱”六要素和相关文献关键词选取方法[9,18,26],选择“南京夫子庙”“夫子庙”“夫子庙游玩攻略”“夫子庙小吃”“夫子庙门票”“秦淮河”“乌衣巷”和“夫子庙灯会”作为基准关键词,找出它们的搜索量并遍历百度搜索引擎推荐的相关关键词,通过爱站网(www.aizhan.com)进一步验证以上 8 个关键词的搜索热度名列前茅,可以用于夫子庙的游客量预测研究. 采用关键词叠加算法计算得到每日剔除本市搜索后的长三角游客网络搜索数据.

1.2 研究方法

在建模之前,需要满足序列具有平稳性和因果关系的先验条件,因此本文先对网络搜索与游客量数据进行 ADF 单位根平稳性检验,运用格兰杰因果检验和滚动相关系数探索两类序列波动的关联性,在此基础上建立改进模型,具体通过 MATLAB 平台实现. 本文提出的改进模型采用“先分后合”的思路(如图 1 所

示):首先,引入 EMD 方法对网络搜索($\boldsymbol{B}$)与游客量($\boldsymbol{Y}$)序列分别进行波动分解,得到两组 IMF 分量及相应的残差 res. 然后,利用网络搜索分量($\boldsymbol{B}$_IMF、$\boldsymbol{B}$_res)作为解释变量,对应频率的游客量分量($\boldsymbol{Y}$_IMF、$\boldsymbol{Y}$_res)作为被解释变量,两两建立 VAR 模型进行预测. 最后,将每对分量预测结果加总作为 EMD-VAR 组合模型的预测值.

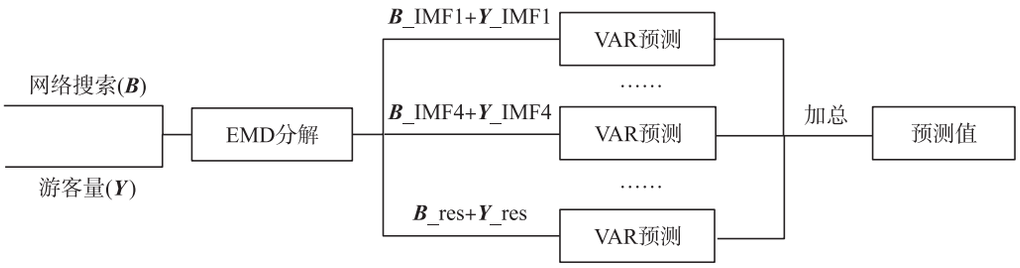


图 1　EMD-VAR 组合预测模型原理示意图

Fig. 1　Principle of EMD-VAR forecasting model

1.2.1　EMD 方法

EMD 方法的基本原理是根据局部特征自适应地对信号进行时频分解,得到一组本征模态函数(Intrinsic Mode Function,IMF)和残差 res. 由于 EMD 方法能够分解出非线性时间序列内部的波动特征和关联因素,故可用于处理夫子庙景区长三角网络搜索与游客量数据. 本文通过计算 IMF 的平均周期和方差贡献率两个指标来衡量两类序列的波动特征. EMD 分解过程的具体步骤[30]如下:

找出原始序列中 $x(t)$ 的全部极大值点与极小值点,利用三次样条函数分别拟合成该时间序列上包络线 $e_{\max}(t)$ 和下包络线 $e_{\min}(t)$. 计算任意时刻上、下包络线的均值 $m(t)$,并将原始序列 $x(t)$ 与均值 $m(t)$ 作差得 $d(t)$,

$$m(t)=\frac{e_{\max}(t)+e_{\min}(t)}{2}, \tag{1}$$

$$d(t)=x(t)-m(t). \tag{2}$$

检验 $d(t)$ 是否满足 IMF 的条件. 若满足则用残差 $r(t)=x(t)-d(t)$ 代替 $x(t)$,如果 $d(t)$ 不是 IMF,则用 $d(t)$ 代替 $x(t)$.

对 $x(t)$ 重复上述步骤,直到不能再分解出新的 IMF 为止,原始序列经过经验模态分解得到的最终结果可以表示为

$$x(t)=\sum_{i=1}^{N} d_i(t)+r(t). \tag{3}$$

式中,N 表示 IMF 的个数;$r(t)$ 为最后的残差项,反映 $x(t)$ 的变化趋势;$d_i(t)$ 为第 i 个 IMF 分量,IMF 的频率随着 i 的增大而逐渐减小.

1.2.2　VAR 模型

VAR 模型利用系统中每一个可以作为所有内生变量滞后值的外生变量函数来构造方程,被广泛应用于预测相互关联的时间序列变量系统. 本文建立的夫子庙区域游客量预测 VAR 模型形式为

$$\boldsymbol{R}_t=\boldsymbol{A}_1\boldsymbol{R}_{t-1}+\boldsymbol{A}_2\boldsymbol{R}_{t-2}+\cdots+\boldsymbol{A}_p\boldsymbol{R}_{t-p}+\boldsymbol{\varepsilon}_t. \tag{4}$$

式中,$\boldsymbol{R}_t$ 是由两个内生变量组成的向量,即 $\boldsymbol{R}_t=(\boldsymbol{Y}_t,\boldsymbol{B}_t)$;$\boldsymbol{Y}_t$、$\boldsymbol{B}_t$ 分别为夫子庙长三角游客量和网络搜索量;$\boldsymbol{\varepsilon}_t$ 是一个二维随机扰动向量;p 为滞后阶;$\boldsymbol{A}_1,\boldsymbol{A}_2,\cdots,\boldsymbol{A}_p$ 为待估计矩阵.

2　游前搜索与客流的波动分解与关联

2.1　时序波动分解

利用 EMD 方法分别对长三角客源市场的网络搜索与游客量时间序列进行分解,均得到 8 个 IMF 分量和 1 个残差分量(如图 2 所示). 波动频率由 IMF1~IMF8 逐渐减小,反映序列在不同时间尺度下的非线性变化特征.

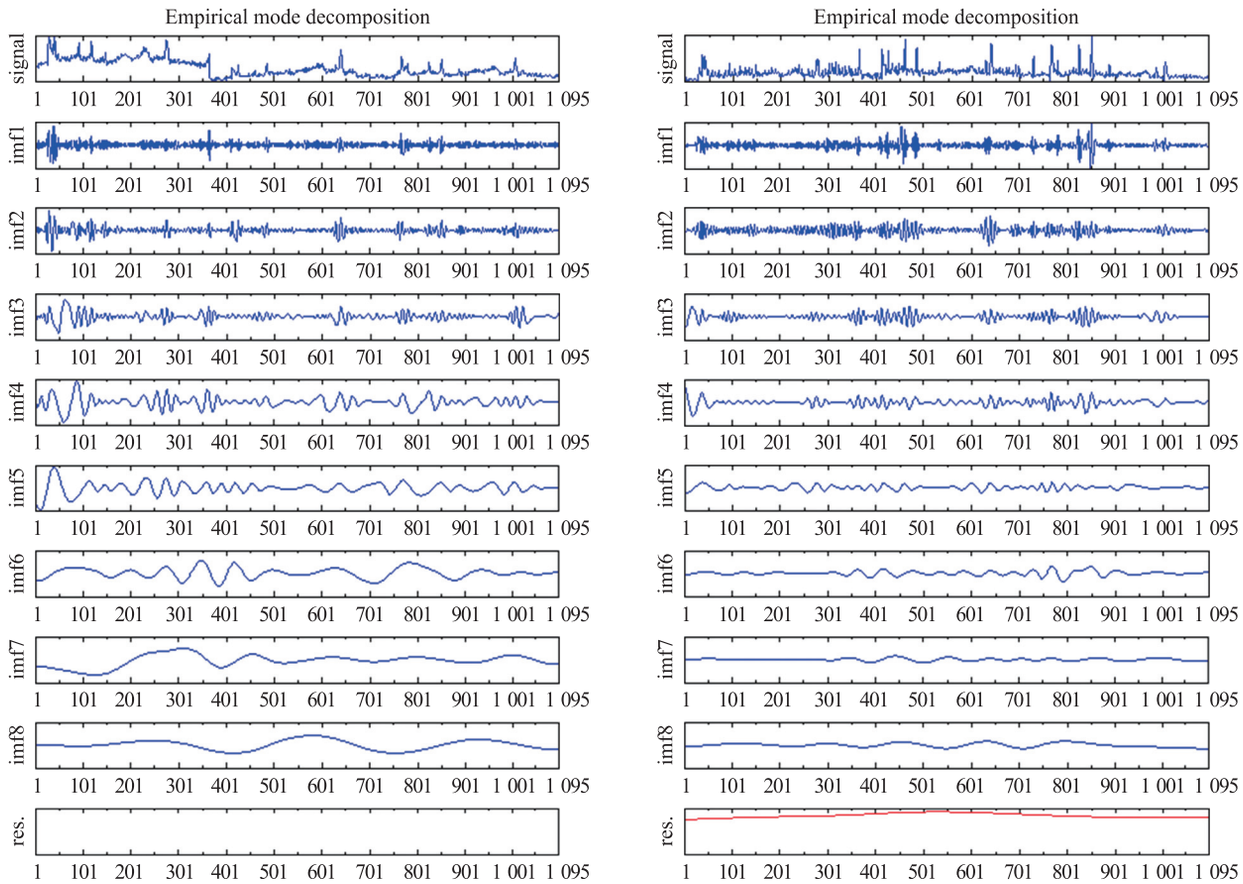


图 2　夫子庙景区长三角网络搜索与游客量 EMD 分解图

Fig. 2　The EMD decomposition of tourist volume and online search in the Yangtze River Delta Region for Confucius Temple scenic area

计算每个分量的平均周期、方差贡献率及其与原始序列的皮尔逊相关系数来透视波动特征,计算结果如表 1 所示. 长三角网络搜索在 IMF4～IMF8 阶段方差贡献率较高,以月度及以上周期波动为主,主要受到节假日和季节性因素影响. 长三角游客量则在 IMF1～IMF4 阶段的贡献率和相关系数更大,以周度波动为主,双休日的驱动作用明显. 长三角地区距离夫子庙景区近,游客出游能力强且重游率高,双休日亦是客流高峰,由于对景区比较熟悉,网络搜索一般在法定节假日等重大节事时期波动较大. 与游客量序列相比,网络搜索残差分量的方差贡献率及其与原始序列的相关系数均高于各 IMF 分量,说明其对网络搜索波动起着主导作用,具有较好的预测功能.

表 1　长三角游客量与网络搜索各分量统计分析表

Table 1　Statistics of all IMFs of tourist volume and online search in the Yangtze River Delta Region

	网络搜索(***B***)			游客量(***Y***)		
	平均周期/d	方差贡献率/%	皮尔逊相关系数	平均周期/d	方差贡献率/%	皮尔逊相关系数
IMF1	3.15	1.64	0.116*	3.62	20.41	0.363*
IMF2	6.27	1.91	0.203*	6.89	17.58	0.415*
IMF3	12.17	2.66	0.150*	11.00	14.15	0.318*
IMF4	21.91	5.15	0.214*	18.27	16.76	0.370*
IMF5	38.51	5.62	0.163*	28.51	6.86	0.324*
IMF6	100.70	4.53	0.234*	61.37	7.63	0.292*
IMF7	137.07	6.43	−0.162*	155.00	2.69	0.133*
IMF8	338.00	4.04	0.302*	381.00	6.39	0.276*
res		68.01	0.786*		7.52	0.297*

注:* 代表相关系数的显著性水平是 0.01.

2.2　波动关联性分析

2.2.1　滚动窗口动态关联

通过计算网络搜索与游客量序列的滚动相关系数探索两者的波动是否存在动态关联性. 鉴于上文 EMD 分解结果的周度波动平均周期主要为 10~12 d,且一般节假日的时间跨度为两周,因此 14 d 的滚动时窗长度进行计算,避免时窗太短难以充分表现序列的动态关联特征[31]. 为了减少偶然波动和便于展示,取滚动相关系数的周平均值绘图(如图 3 所示).

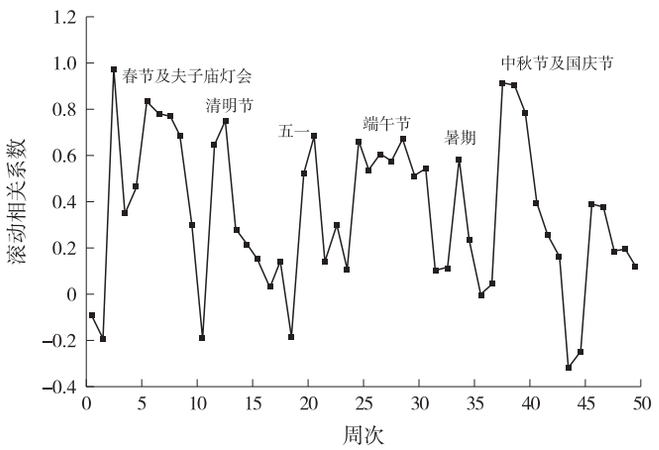


图 3　夫子庙景区长三角游客量与网络关注度的滚动相关系数

Fig. 3　Scroll correlation of tourist volume and online search in the Yangtze River Delta Region for Confucius Temple scenic area

除了个别时间段外,长三角网络搜索与游客量序列的波动均呈现正相关关系,随着季节性和节假日因素呈现多峰波动特征. 春节及夫子庙灯会、清明节、五一、端午节、8 月暑期以及中秋国庆期间网络搜索与游客量的滚动相关系数处于峰值,波动关联紧密度高,其余时间由于闲暇限制和气候的舒适抑制效应,潜在游客的网络搜索意愿及行为减少,网络搜索与游客量的关联性减弱.

2.2.2　格兰杰因果检验

利用格兰杰因果检验进一步探索网络搜索与游客量原始序列及其不同频率尺度的分量间是否具有因果关系. 为确保序列的平稳性,首先进行 ADF 单位根检验(如表 2 所示),原始序列与所有分量均为零阶单整的平稳序列,可以进行格兰杰因果检验.

表 2　各分量的 ADF 单位根检验

Table 2　Unit root test of all IMFs

变量	ADF 值	临界值 Critical value			结论	变量	ADF 值	临界值 Critical value			结论
		1%	5%	10%				1%	5%	10%	
Y	−12.37	−3.436	−2.864	−2.568	平稳	***B***	−4.924	−3.966	−3.414	−3.129	平稳
Y_IMF1	−20.21	−2.567	−1.941	−1.616	平稳	***B***_IMF1	−23.27	−3.436	−2.864	−2.568	平稳
Y_IMF2	−18.11	−3.436	−2.864	−2.568	平稳	***B***_IMF2	−17.45	−2.567	−1.941	−1.616	平稳
Y_IMF3	−14.96	−2.567	−1.941	−1.616	平稳	***B***_IMF3	−9.887	−2.567	−1.941	−1.616	平稳
Y_IMF4	−12.46	−2.567	−1.941	−1.616	平稳	***B***_IMF4	−10.25	−2.567	−1.941	−1.616	平稳
Y_IMF5	−8.727	−2.567	−1.941	−1.616	平稳	***B***_IMF5	−9.564	−2.567	−1.941	−1.616	平稳
Y_IMF6	−6.723	−2.567	−1.941	−1.616	平稳	***B***_IMF6	−4.417	−2.567	−1.941	−1.616	平稳
Y_IMF7	−6.494	−2.567	−1.941	−1.616	平稳	***B***_IMF7	−2.932	−2.567	−1.941	−1.616	平稳
Y_IMF8	−5.812	−2.567	−1.941	−1.616	平稳	***B***_IMF8	−2.372	−2.567	−1.941	−1.616	平稳
Y_res	−7.535	−3.967	−3.414	−3.129	平稳	***B***_res	−5.365	−3.436	−2.864	−2.568	平稳

格兰杰因果检验对滞后期的选择较为敏感,一般依据赤池信息准则(Akaike information criterion,AIC)和施瓦兹信息准则(Schwarz criterion,SC)选择最优滞后阶数. 检验结果见表 3,除了分量 IMF8 外,其余网络搜索分量与相应频率的游客量分量均呈现双向格兰杰因果关系(p 值小于 0.05 表示存在格兰杰因果关系),而所有网络搜索序列均为相应游客量序列的格兰杰原因,即网络搜索原始序列及所有分量的变化均会分别格兰杰引起游客量原始序列及其对应频率分量的波动,证实网络搜索原始序列及其波动分解得到的分量序列均对游客量具有预测能力.

3　模型验证及预测结果比较

本文选择 2017 年 1 月 1 日~2019 年 11 月 30 日为样本期,共 1 064 组数据用于建模,对 2019 年 9 月 1 日~2019 年 11 月 30 日进行样本内静态预测. 为进一步验证模型预测能力,选择 2019 年 12 月 1 日~2019 年 12 月 31 日为测试期进行样本外动态预测. 同时将改进的 EMD-VAR 模型的预测效果分别与基于

历史游客量数据的 ARMA 模型和未对时间序列进行波动分解的 VAR 模型进行横纵向对比,采用误差均方根 V_{RMSE}、均方误差 V_{MSE}、平均绝对百分误差 V_{MAPE}和平均绝对离差 V_{MAD}评估预测性能.

表 3　各分量的格兰杰因果检验

Table 3　Granger causality test of all IMFs

项目	F 值	p 值	结论	项目	F 值	p 值	结论
Y 不是 ***B*** 格兰杰原因	37.360	0.000 0	拒绝	***Y***_IMF5 不是 ***B***_IMF5 格兰杰原因	2.318 4	0.041 6	拒绝
B 不是 ***Y*** 格兰杰原因	17.277	0.000 0	拒绝	***B***_IMF5 不是 ***Y***_IMF5 格兰杰原因	7.559 7	0.000 0	拒绝
Y_IMF1 不是 ***B***_IMF1 格兰杰原因	2.684 2	0.009 3	拒绝	***Y***_IMF6 不是 ***B***_IMF6 格兰杰原因	6.813 9	0.000 2	拒绝
B_IMF1 不是 ***Y***_IMF1 格兰杰原因	2.181 7	0.033 6	拒绝	***B***_IMF6 不是 ***Y***_IMF6 格兰杰原因	4.899 8	0.002 2	拒绝
Y_IMF2 不是 ***B***_IMF2 格兰杰原因	3.957 5	0.000 6	拒绝	***Y***_IMF7 不是 ***B***_IMF7 格兰杰原因	11.01 3	0.000 0	拒绝
B_IMF2 不是 ***Y***_IMF2 格兰杰原因	6.965 3	0.000 0	拒绝	***B***_IMF7 不是 ***Y***_IMF7 格兰杰原因	3.519 1	0.003 7	拒绝
Y_IMF3 不是 ***B***_IMF3 格兰杰原因	4.842 7	0.000 1	拒绝	***Y***_IMF8 不是 ***B***_IMF8 格兰杰原因	0.159 3	0.923 7	不拒绝
B_IMF3 不是 ***Y***_IMF3 格兰杰原因	2.172 0	0.043 4	拒绝	***B***_IMF8 不是 ***Y***_IMF8 格兰杰原因	3.323 7	0.019 2	拒绝
Y_IMF4 不是 ***B***_IMF4 格兰杰原因	2.710 9	0.019 2	拒绝	***Y***_res 不是 ***B***_res 格兰杰原因	17.536	0.000 0	拒绝
B_IMF4 不是 ***Y***_IMF4 格兰杰原因	5.612 2	0.000 0	拒绝	***B***_res 不是 ***Y***_res 格兰杰原因	5.826 1	0.000 0	拒绝

3.1　ARMA 模型与一般 VAR 模型拟合

ARIMA(p,d,q)中,AR 是“自回归”,p 为自回归项数,MA 为“滑动平均”,q 为滑动平均项数,d 为使之成为平稳序列所做的差分次数. 由于游客量原始序列(Y)为零阶单整的平稳序列($d=0$),因此选择 ARMA(p,q)模型. 根据自相关 ACF 图和偏自相关 PACF 图的拖尾现象,初步设定 p 与 q 最大值均为 3. 遍历 p 和 q 排列组合的 9 个 ARMA 模型,并根据上述信息准则筛选出最优估计模型 ARMA(2,0,2). 为进一步检验模型的稳定性,对 ARMA(2,0,2)模型进行残差序列相关检验,得到 p 值为 0.816,说明残差序列不相关,构建的模型比较优良. 最优估计的 ARMA(2,0,2)模型的回归结果可表示为

$$\boldsymbol{Y}=\sum_{i=1}^{2}AR(i)Y(-i)+\sum_{i=1}^{2}MA(i)\varepsilon(-i)+C. \tag{8}$$

式中,具体的参数估计结果如表 4 所示. 模型的拟合优度为 0.520, AIC 值为 21.716, SC 值为 21.739,样本内和样本外预测的误差均方根分别为 13 061.03 和 14 008.24.

表 4　游客量序列 ARMA 模型估计结果

Table 4　Regression results of ARMA model of tourist time series

解释变量	参数估计值	T 统计值
$AR(1)$	1.423 327	23.221 630
$AR(2)$	−0.444 389	−8.418 170
$MA(1)$	−0.596 934	−9.665 035
$MA(2)$	−0.308 688	−7.727 363
C	36 492.33	20.778 320

表 2 与表 3 检验结果显示,长三角网络搜索与游客量均为零阶单整的平稳序列,且互为格兰杰因果,根据上述信息准则确定最优滞后阶数为 3,因此建立 VAR(3)模型. 模型所有特征根(0.980,0.531,−0.501−0.483i,−0.051+0.483i,0.448,−0.139)均落在单位圆内,且通过变量内生性检验,说明建立的 VAR(3)模型稳定且可靠,得到游客量估计结果为:

$$\boldsymbol{Y}_t=\begin{bmatrix}0.735&-0.005\\5.525&0.984\end{bmatrix}\boldsymbol{Y}_{t-1}+\begin{bmatrix}-0.213&-0.003\\-1.401&-0.013\end{bmatrix}\boldsymbol{Y}_{t-2}+\begin{bmatrix}0.184&-0.002\\-3.970&0.008\end{bmatrix}\boldsymbol{Y}_{t-3}+\boldsymbol{\varepsilon}_t. \tag{9}$$

该模型的拟合优度值为 0.540,AIC 的值为 21.675,SC 的值为 21.708,样本内和样本外预测的误差均方根分别为 12 158.70 和 17 023.93.

3.2　EMD-VAR 模型拟合与比较

网络搜索与游客量原始序列经过 EMD 分解得到的 IMF 分量及残差分量均通过了平稳性检验、特征根检验和格兰杰因果检验等一系列检验,满足建立 EMD-VAR 模型的前提条件,因此可以将同一频率尺度的网络搜索分量和游客量分量,如 ***B***_IMF1 和 ***Y***_IMF1 作为自变量和因变量带入式(4),建立 9 个 EMD-VAR 模型进行估计,估计结果见表 5.

网络搜索 IMF1 分量与原始序列的皮尔逊相关系数较低,因此模型的拟合优度亦较低,而其余分量模型的拟合优度均在 0.9 以上,说明回归方程高度拟合. 在所有模型中,IMF1 的 AIC 值和 SC 值最大,分别为 20.824 和 20.894,整体上模型稳定性较好. 运用估计好的各分量模型独立进行样本内静态预测,最后将各

分量预测值加总得到 EMD-VAR 模型的最终游客量预测值,通过式(5)计算得样本内预测的误差均方根为 6 831.44. 再利用同样的方法对夫子庙 2019 年 12 月份的长三角游客量进行样本外动态预测的误差均方根为 3 962.61. 同 ARMA 模型和 VAR 模型相比,EMD-VAR 模型的拟合优度提高,AIC 值与 SC 值减小,从侧面说明 EMD 方法对数据的优化处理提高了分量 VAR 模型的稳定性.

表 5　各分量 EMD-VAR 模型估计结果

Table 5　Regression results of EMD-VAR model of all IMFs

模型	估计方程	R^2
IMF1	$\boldsymbol{Y}_t=\begin{bmatrix}0.113 & -0.003\\ 0.431 & -0.034\end{bmatrix}\boldsymbol{Y}_{t-1}+\begin{bmatrix}-0.339 & -0.002\\ 0.446 & -0.211\end{bmatrix}\boldsymbol{Y}_{t-2}+\begin{bmatrix}-1.120 & -0.002\\ 0.602 & -0.128\end{bmatrix}\boldsymbol{Y}_{t-3}+\begin{bmatrix}-0.088 & -0.000\\ 1.848 & -0.125\end{bmatrix}\boldsymbol{Y}_{t-4}+\begin{bmatrix}-0.154 & -0.002\\ 2.637 & 0.051\end{bmatrix}\boldsymbol{Y}_{t-5}+\begin{bmatrix}-0.083 & -0.001\\ 0.957 & 0.081\end{bmatrix}\boldsymbol{Y}_{t-6}+\begin{bmatrix}0.023 & -0.002\\ 2.644 & 0.091\end{bmatrix}\boldsymbol{Y}_{t-7}+\boldsymbol{\varepsilon}_t$	0.162
IMF2	$\boldsymbol{Y}_t=\begin{bmatrix}1.773 & 0.001\\ 4.689 & 2.063\end{bmatrix}\boldsymbol{Y}_{t-1}+\begin{bmatrix}-2.115 & -0.005\\ -7.752 & -2.511\end{bmatrix}\boldsymbol{Y}_{t-2}+\begin{bmatrix}1.418 & 0.006\\ 7.201 & 1.920\end{bmatrix}\boldsymbol{Y}_{t-3}+\begin{bmatrix}-0.909 & -0.006\\ -4.188 & -1.181\end{bmatrix}\boldsymbol{Y}_{t-4}+\begin{bmatrix}0.355 & 0.004\\ 2.467 & 0.485\end{bmatrix}\boldsymbol{Y}_{t-5}+\begin{bmatrix}-0.166 & -0.002\\ 0.635 & -0.180\end{bmatrix}\boldsymbol{Y}_{t-6}+\boldsymbol{\varepsilon}_t$	0.927
IMF3	$\boldsymbol{Y}_t=\begin{bmatrix}3.089 & -0.003\\ 1.740 & 3.413\end{bmatrix}\boldsymbol{Y}_{t-1}+\begin{bmatrix}-2.115 & -0.005\\ -4.152 & -5.144\end{bmatrix}\boldsymbol{Y}_{t-2}+\begin{bmatrix}3.649 & -0.014\\ 4.722 & 4.447\end{bmatrix}\boldsymbol{Y}_{t-3}+\begin{bmatrix}-1.989 & 0.016\\ -3.904 & -2.442\end{bmatrix}\boldsymbol{Y}_{t-4}+\begin{bmatrix}0.727 & -0.010\\ 2.485 & 0.866\end{bmatrix}\boldsymbol{Y}_{t-5}+\begin{bmatrix}-0.158 & 0.003\\ -0.799 & -0.162\end{bmatrix}\boldsymbol{Y}_{t-6}+\boldsymbol{\varepsilon}_t$	0.995
IMF4	$\boldsymbol{Y}_t=\begin{bmatrix}3.678 & -0.001\\ 3.153 & 3.560\end{bmatrix}\boldsymbol{Y}_{t-1}+\begin{bmatrix}-5.609 & 0.002\\ -10.145 & -4.839\end{bmatrix}\boldsymbol{Y}_{t-2}+\begin{bmatrix}4.404 & -0.002\\ 12.966 & 2.956\end{bmatrix}\boldsymbol{Y}_{t-3}+\begin{bmatrix}-1.770 & 0.001\\ -7.791 & -0.662\end{bmatrix}\boldsymbol{Y}_{t-4}+\begin{bmatrix}0.286 & -0.000\\ 1.854 & -0.018\end{bmatrix}\boldsymbol{Y}_{t-5}+\boldsymbol{\varepsilon}_t$	0.999
IMF5	$\boldsymbol{Y}_t=\begin{bmatrix}3.925 & -0.001\\ 0.416 & 2.874\end{bmatrix}\boldsymbol{Y}_{t-1}+\begin{bmatrix}-6.211 & 0.003\\ -0.737 & -2.314\end{bmatrix}\boldsymbol{Y}_{t-2}+\begin{bmatrix}4.953 & -0.005\\ -0.035 & -0.431\end{bmatrix}\boldsymbol{Y}_{t-3}+\begin{bmatrix}-1.995 & 0.004\\ 0.706 & 1.293\end{bmatrix}\boldsymbol{Y}_{t-4}+\begin{bmatrix}0.326 & -0.001\\ -0.346 & -0.423\end{bmatrix}\boldsymbol{Y}_{t-5}+\boldsymbol{\varepsilon}_t$	0.999
IMF6	$\boldsymbol{Y}_t=\begin{bmatrix}2.966 & -0.001\\ -0.096 & 2.680\end{bmatrix}\boldsymbol{Y}_{t-1}+\begin{bmatrix}-2.951 & 0.003\\ 0.219 & -2.369\end{bmatrix}\boldsymbol{Y}_{t-2}+\begin{bmatrix}0.985 & -0.002\\ -0.124 & 0.687\end{bmatrix}\boldsymbol{Y}_{t-3}+\boldsymbol{\varepsilon}_t$	0.999
IMF7	$\boldsymbol{Y}_t=\begin{bmatrix}2.707 & -0.005\\ 0.047 & 1.620\end{bmatrix}\boldsymbol{Y}_{t-1}+\begin{bmatrix}-1.861 & 0.001\\ -0.049 & 0.046\end{bmatrix}\boldsymbol{Y}_{t-2}+\begin{bmatrix}0.696 & 0.017\\ -0.040 & -0.555\end{bmatrix}\boldsymbol{Y}_{t-3}+\begin{bmatrix}1.144 & -0.016\\ 0.053 & -0.510\end{bmatrix}\boldsymbol{Y}_{t-4}+\begin{bmatrix}-0.293 & 0.003\\ -0.012 & 0.399\end{bmatrix}\boldsymbol{Y}_{t-5}+\boldsymbol{\varepsilon}_t$	1.000
IMF8	$\boldsymbol{Y}_t=\begin{bmatrix}2.000 & -0.000\\ 0.120 & 1.997\end{bmatrix}\boldsymbol{Y}_{t-1}+\begin{bmatrix}-1.001 & -0.000\\ -0.122 & -0.997\end{bmatrix}\boldsymbol{Y}_{t-2}+\boldsymbol{\varepsilon}_t$	1.000
res	$\boldsymbol{Y}_t=\begin{bmatrix}1.242 & 0.036\\ -0.033 & 0.782\end{bmatrix}\boldsymbol{Y}_{t-1}+\begin{bmatrix}0.317 & -0.030\\ 0.055 & 0.513\end{bmatrix}\boldsymbol{Y}_{t-2}+\begin{bmatrix}-0.307 & -0.004\\ -0.015 & 0.199\end{bmatrix}\boldsymbol{Y}_{t-3}+\begin{bmatrix}-0.311 & -0.042\\ -0.044 & -0.063\end{bmatrix}\boldsymbol{Y}_{t-4}+\begin{bmatrix}0.059 & 0.040\\ 0.037 & -0.432\end{bmatrix}\boldsymbol{Y}_{t-5}+\boldsymbol{\varepsilon}_t$	1.000

通过上述预测过程,得到 3 种模型对夫子庙长三角客源市场日际游客量的样本内和样本外预测结果. 为便于观察,图 4 绘制出了 2019 年 9 月~11 月的游客量样本内预测结果,预测模型的样本内和样本外误差评价如表 6 所示.

表 6　预测模型误差评价指标

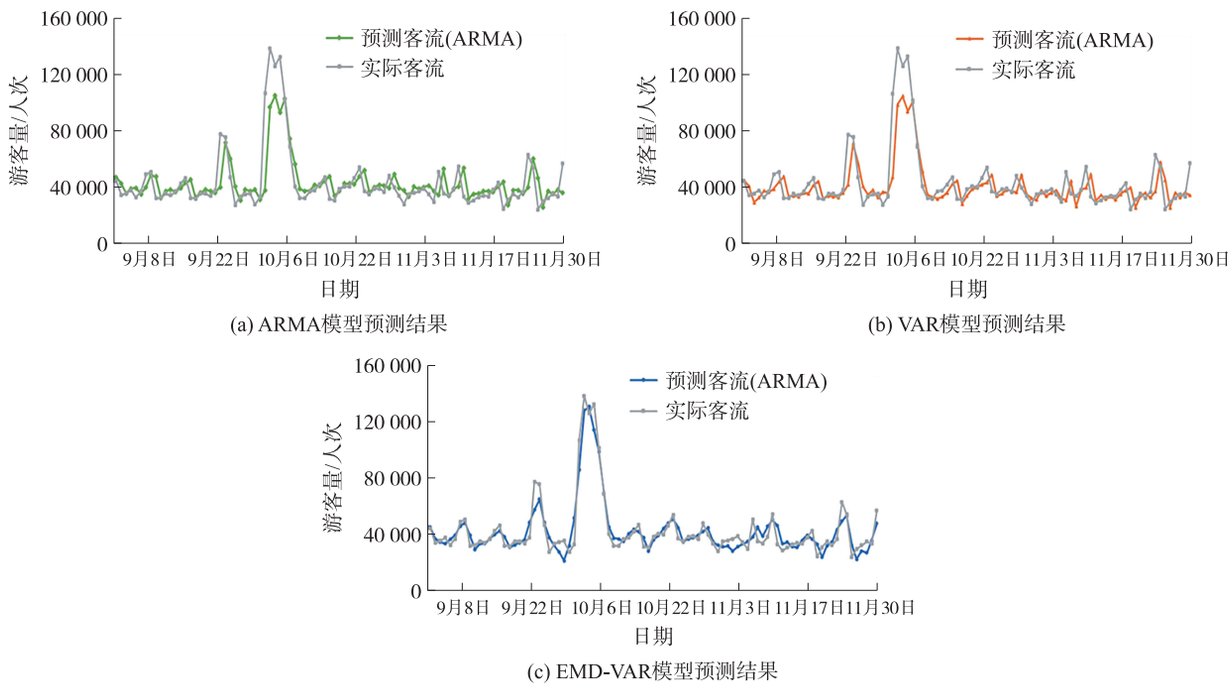
Table 6　Error evaluation indexes of prediction model

	样本内			样本外		
	EMD-VAR 模型	VAR 模型	ARMA 模型	EMD-VAR 模型	VAR 模型	ARMA 模型
V_{MAD}	5 192.09	7 320.42	7 748.00	3 125.78	16 522.75	13 454.33
V_{MSE}	46 668 552	147 834 028	170 590 462	15 702 260	289 814 350	196 230 769
V_{RMSE}	6 831.44	12 158.70	13 061.03	3 962.61	17 023.94	14 008.24
V_{MAPE}	12.67%	15.58%	16.95%	16.94%	93.71%	76.86%

由图 4 可见,与传统 ARMA 模型和单一 VAR 模型相比,EMD-VAR 预测模型的样本内预测值与实际客流的吻合度更好,一致性更高,尤其是“十一”黄金周节假日时期游客量的预测效果明显更优. 从数据来看,样本内预测的各项误差指标的大小排序为:ARMA 模型、VAR 模型、EMD-VAR 模型,样本外预测中 EMD-VAR 模

型的各项误差亦显著小于 VAR 模型和 ARMA 模型. 就误差均方根而言,EMD-VAR 样本内和样本外的预测精度较 VAR 模型分别提高了 43.81%和 76.72%,较 ARMA 模型相应提高了 67.86%和 71.71%.

由此可知,EMD-VAR 模型的预测效果显著优于单一 VAR 模型和传统 ARMA 模型,反映了经过 EMD 分解优化后的网络搜索序列对游客量序列具有更好的解释能力,能够有效提高游客量的预测精度.



(a) ARMA模型预测结果　　(b) VAR模型预测结果

(c) EMD-VAR模型预测结果

图 4　3 种模型样本内游客量预测结果

Fig. 4　Tourism prediction results in sample of three models

4　结论与展望

4.1　结论

本文以南京夫子庙景区为例,基于长三角地区 2017~2019 年每日百度指数和游客手机信令数据,探索性地建立 EMD-VAR 组合模型进行游客量预测,并与单一 VAR 模型和传统 ARMA 模型对比预测效果,实证考察了网络搜索与游客量时间序列的非线性特征和波动关联性对预测效果的影响. 研究表明,网络搜索的波动周期比游客量的波动周期更长. 网络搜索以月度及以上的中低频波动为主,主要受到节假日和季节性因素影响,一般在法定节假日等时期波动较大. 游客量以周度的高频波动为主,双休日驱动作用明显,周末亦是客流高峰. 网络搜索与游客量的波动具有格兰杰因果关系,两者波动关联性的季节性特征显著. 网络搜索原始序列及其所有分量的变化均会分别格兰杰引起游客量原始序列及其相应频率分量的波动,且两类序列波动的关联紧密度峰值均出现于法定节假日时期,明显高于非节假日. 基于网络搜索和游客量波动分解的 EMD-VAR 模型样本内预测精度较单一 VAR 模型和传统 ARMA 模型分别提高了 43.81%和 67.86%,尤其体现在关联紧密度大的“十一”黄金周时期,充分反映了经 EMD 方法优化处理的网络搜索序列对游客量具有更好的预测能力.

夫子庙这类都市型目的地以周内波动为主是广泛存在的特征,而网络搜索中低频波动与游客量高频波动之间的差异,这恰恰反映出了游客出游意愿和搜索的诸多特征,特别是夫子庙作为著名人文型目的地游客受网络关注程度是相对稳定的,同时搜索行为是在出游前持续一周或更长时间准备[5],实际到达游览则集中在周末. 这种出游意愿和行为的错配最终呈现为了两种波动的诸多关联特征.

4.2　展望

随着时代的发展,旅游信息交流方式从传统线下渠道转变为依托互联网人工智能技术的智慧旅游导览,为景区游客量预测提供了更多的数据源,这些旅游大数据之间的关联性与前兆效应对游客量预测精度的提升是传统数据无法比拟的. 本文基于网络搜索数据与游客量的日际波动关联性,使用 EMD 方法在非

线性信号分解中的时频分辨优势可以弥补传统线性 VAR 预测模型在非线性时间序列不同时间尺度波动关联性捕捉上的不足,从而实现更精准的即时预测,这对旅游景区管理部门而言具有重要的实践指导意义.

结合文章实证分析结果,景区管理部门一方面应积极响应旅游目的地智慧旅游大数据平台的建设,管理、维护并利用好景区的客流实时监测数据库. 另一方面,借助本文预测思路提前做好旺季客流和疫后解封等措施的即时预测、峰值管理和安全预警工作,据此完善基础服务供给和资源调度,预防拥挤踩踏等旅游安全事故的发生和景区超载对生态环境系统造成的波动冲击,促进景区的可持续发展.

同时,从预测结果可以看出,各种模型对于景区某些时刻的客流峰值预测存在较大误差,研究案例地属于发达地区历史文化街区类型景区,复游率与客源稳定性高,预测方法在应用到其他类型景区时需要进行相应调整优化. 尤其在经历新冠疫情 3 年管控后旅游业重归全面复苏态势,对于游客搜索和出游行为的短期和长期变化还需要进行预测方法的针对性研究.

[参考文献](References)

[1] 程绍文,张捷,梁玥琳,等. 我国旅游网站空间分布及动力机制研究[J]. 旅游学刊,2009,24(2):75-80.

[2] LIU P X,ZHANG H L,ZHANG J,et al. Spatial-temporal response patterns of tourist flow under impulse pre-trip information search:from online to arrival[J]. Tourism Management,2019,73:105-114.

[3] 李方一,肖夕林,刘思佳. 基于网络搜索数据的区域经济预警研究[J]. 华东经济管理,2016,30(8):60-66.

[4] 闫闪闪,梁留科,索志辉,等. 基于大数据的洛阳市旅游流时空分布特征[J]. 经济地理,2017,37(8):216-224.

[5] 刘培学,朱知沛,张捷,等. 旅游在线搜索与客流波动的动态关联研究——以南京钟山风景名胜区为例[J]. 旅游学刊,2021,36(11):95-105.

[6] ARTOLA C, PINTO F, GARCIA P D P. Can internet searches forecast tourism inflows? [J]. International Journal of Manpower,2015,36(1):103-116.

[7] PAN B,WU D C,SONG H. Forecasting hotel room demand using search engine data[J]. Journal of Hospitality and Tourism Technology,2012,3(3):196-210.

[8] YANG X,PAN B,EVANS J A,et al. Forecasting Chinese tourist volume with search engine data[J]. Tourism Management,2015,46:386-397.

[9] 黄先开,张丽峰,丁于思. 百度指数与旅游景区游客量的关系及预测研究——以北京故宫为例[J]. 旅游学刊,2013,28(11):93-100.

[10] 沈苏彦,赵锦,徐坚. 基于"谷歌趋势"数据的入境外国游客量预测[J]. 资源科学,2015,37(11):2111-2119.

[11] 张斌儒,黄先开,刘树林. 基于网络搜索数据的旅游收入预测——以海南省为例[J]. 经济问题探索,2015(8):154-160.

[12] 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.

[13] HUANG B,HAO H. A novel two-step procedure for tourism demand forecasting[J]. Current Issues in Tourism and Practice,2021,24(9):1199-1210.

[14] PARK S,LEE J,SONG W. Short-term forecasting of Japanese tourist inflow to South Korea using Google trends data[J]. Journal of Travel & Tourism Marketing,2017,34(3):357-368.

[15] 刘少湃,田纪鹏,陆林. 上海迪士尼在建景区客源市场空间结构预测——旅游引力模型的修正及应用[J]. 地理学报,2016,71(2):304-321.

[16] GUNTER U,ÖNDER I. Forecasting city arrivals with Google Analytics[J]. Annals of Tourism Research,2016,61:199-212.

[17] LIU Y Y,TSENG F M,TSENG Y H. Big Data analytics for forecasting tourism destination arrivals with the applied Vector Autoregression model[J]. Technological Forecasting and Social Change,2018,130:123-134.

[18] 孙烨,张宏磊,刘培学,等. 基于旅游者网络关注度的旅游景区日游客量预测研究——以不同客户端百度指数为例[J]. 人文地理,2017,32(3):152-160.

[19] 殷杰,郑向敏,董斌彬. 基于 VECM 模型的景区网络关注度与旅游人数的关系研究——以鼓浪屿为例[J]. 福建农林大学学报(哲学社会科学版),2015,18(5):68-75.

[20] 陈涛,刘庆龙. 智慧旅游背景下的大数据应用研究:以旅游需求预测为例[J]. 电子政务,2015(9):6-13.

[21] 李晓炫,吕本富,曾鹏志,等. 基于网络搜索和 CLSI-EMD-BP 的旅游客流量预测研究[J]. 系统工程理论与实践,2017,

37(1):106-118.
[22] 谢天保,赵萌. 基于网络搜索数据的游客量预测模型研究[J]. 计算机系统应用,2018,27(7):199-204.
[23] 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.
[24] 余向洋,沙润,朱兴,等. 基于 EMD 的景区客流波动特征及其组合预测——以黄山风景区为例[J]. 地理科学进展,2012,31(10):1353-1359.
[25] 赵雪花,陈旭. 经验模态分解与均生函数——最优子集耦合模型在年径流预测中的应用[J]. 资源科学,2015,37(6):1173-1180.
[26] 高亚男,周彬,虞虎,等. 长江经济带入境旅游时空分异及趋势预测[J]. 资源开发与市场,2020,36(10):1153-1158.
[27] 姜鉴铎,张建新,吴国平,等. 基于加权方法的旅游流网络结构特征分析——以南京市为例[J]. 资源开发与市场,2019,35(5):706-711.
[28] 马莉,刘培学,张建新,等. 景区旅游流与网络关注度的区域时空分异研究[J]. 地理与地理信息科学,2018,34(2):87-93.
[29] 戴文,丁蕾,刘培学,等. 城市旅游流客源地分布及预测研究——以南京市为例[J]. 资源开发与市场,2018,34(5):676-681.
[30] 张衍广,原艳梅. 基于经验模态分解的中国生态足迹与生态承载力动力学预测[J]. 资源科学,2008(8):1212-1217.
[31] CANTIS D S,FERRANTE M,VACCINA F. Seasonal pattern and amplitude—a logical framework to analyze seasonality in tourism:an application to bed occupancy in Sicilian hotels[J]. Tourism Economics,2011,17(3):655-675.

[责任编辑:陈　庆]

(上接第 39 页)

[7] 王双. 时空叙事可视化理论与方法研究[D]. 郑州:解放军信息工程大学,2017.
[8] LU M,ARIKAWA M. Map-based storytelling tool for real-world walking tour[M]. Berlin,Heidelberg:Springer,2013.
[9] SLAVIN N. Map-based storytelling in spatial augmented reality:Projection of interactive layers[D]. Munich,Germany:Technische Universität München,2020.
[10] 俞哲旻,彭兰. 用街景地图讲故事:对弗格森案的沉浸式叙事报道[J]. 新闻界,2015(1):68-73.
[11] 彭韵筑,张一凡,叶瑞恒,等. “慰安妇”公共创伤记忆的数字化构建——以“南京地区侵华日军慰安所的 AR 故事地图”为例[J]. 图书馆论坛,2020,40(11):68-79.
[12] 胡昊宇,胡迪,程星华,等. 面向公众的三国历史 WebGIS 设计与实现[J]. 南京师范大学学报(工程技术版),2018,18(1):71-78.
[13] 陆佳莺,孙梓洋,汪亦铠,等. 基于地图故事的“图说随园”系统设计与实现[J]. 南京师范大学学报(工程技术版),2019,19(1):86-92.
[14] ILIES G,ILIES M. A storytelling map of the upper Mara Valley[J]. Cartography & Geoinformation,2018,17(30):16-27.
[15] 张保立. 民国游记中的重庆地理意象研究[D]. 重庆:西南大学,2018.
[16] 黄强. 简论徐霞客庐山地理考察的成就[J]. 江西师范大学学报(自然科学版),1998,22(1):77-82.
[17] 王立群. 中国古代山水游记研究[M]. 郑州:河南大学出版社,1996.
[18] 闾国年,俞肇元,袁林旺,等. 地图学的未来是场景学吗?[J]. 地球信息科学学报,2018,20(1):1-6.
[19] 朱梦泽,赵海英. 叙事式可视化综述[J]. 计算机辅助设计与图形学学报,2019,31(10):1719-1727.
[20] 苏世亮,张江玥,杜清运,等. 历史文化风貌区叙事地图设计——可读性框架与表达策略[J]. 测绘科学.2021(10):194-201.
[21] 苏世亮,王令琦,杜清运,等. 校园文化地图集设计——以《漫步珞珈地图集》为例[J]. 测绘科学,2020,45(12):153-160.

[责任编辑:陈　庆]