

基于网络搜索指数和 EMD-ARIMA-BP 组合模型的游客量预测

——以张家界为例*

陆利军^{1,2,3}

(1. 中南林业科技大学 旅游学院, 湖南 长沙 410004; 2. 湖南工学院
经济与管理学院, 湖南 衡阳 421008; 3. 衡阳师范学院 湖南省
人居环境学研究基地, 湖南 衡阳 421008)



摘要:科学的客流量预测有利于完善旅游安全预警体系和优化旅游资源配置体系。为进一步提高游客量预测的准确度,提出一种基于网络搜索指数的 EMD-ARIMA-BP 组合模型,以探究互联网时代旅游消费者出行行为规律。该模型首先对网络搜索行为数据进行指数合成,其次利用 EMD 算法对游客量和网络搜索数据进行去噪处理,最后将 ARIMA 模型和 BP 神经网络进行组合,对游客量进行预测。实证分析以张家界为例。研究发现:(1)运用网络搜索数据预测旅游消费者出行行为切实可行,接近于实时的网络数据可以大幅提升预测的时效性;(2)经过 EMD 去噪算法对游客量与网络搜索行为数据进行去噪处理后,游客量的预测精度有较大程度提高;(3)基于网络搜索指数和 EMD-ARIMA-BP 组合模型的预测误差显著低于 ARIMA 模型和 BP 神经网络等基准模型。

关键词:网络搜索指数;ARIMA 模型;EMD 算法;BP 神经网络;游客量预测

中图分类号:F590.84

文章编号:1007-4074(2019)01-0138-13

基金项目:国家自然科学基金项目(61772192);湖南省人居环境学研究基地开放基金项目(RJ18K03);湖南省高等学校科学研究项目(14C0308)

作者简介:陆利军,男,中南林业科技大学旅游学院博士研究生,湖南工学院经济与管理学院讲师。

近年来,我国居民的旅游消费需求增长强劲。《2017 年中国居民消费发展报告》显示,2017 年,我国居民全年人均出游次数达到 3.7 次,旅游已经成为衡量现代生活水平的重要指标之一^[1]。然而,伴随着我国旅游需求的快速增长,加之因我国

休假制度设计尚未完善导致的节假日旅游需求集中释放,旅游目的地游客投诉、甚至旅游安全事故层出不穷,极大地影响了旅游者的出行体验和旅游产业的健康发展。通过对游客量进行及时、准确地预测,让旅游经营和管理者提前对游客量有

* 收稿日期:2018-10-19

一个清晰的了解,进而合理调度和配置有限的旅游资源,有助于旅游目的地服务质量的提高和旅游产业的可持续发展。

一、文献综述

(一)游客量预测模型选择的相关研究

在既有的游客量预测研究文献中,预测方法大致可以分为定性和定量两大类别^[2]。其中,“定量预测”的运用较“定性预测”而言更为普遍,主要包括时间序列模型、计量经济模型、人工智能方法和混合预测模型。在上述四种定量预测模型当中,时间序列模型因其对数据的低要求,有更广泛的运用。其中包括自回归移动平均模型(ARIMA)^[3]和广义自回归条件异方差模型(GARCH)^[4],以及由这两种模型衍生出来的季节性自回归移动平均模型(SARIMA)^[5]、多变量季节性自回归移动平均模型(MSARIMA)^[6]、X-12-ARIMA 模型^[7]和多变量广义自回归条件异方差模型(MGARCH)等^[8]。计量经济模型因其对旅游变化的高解释效度,也能够很好地为行政管理部门和旅游企业所用。在游客量预测研究中经常用的计量模型主要有:自回归分布滞后模型(ADLM)^[9]、误差修正模型(ECM)^[10]和向量自回归模型(VAR)^[11]。近年来,由于人工智能方法在解决非线性问题的优良表现,也被引入到游客量预测研究。用于预测游客量的人工智能方法主要有:人工神经网络(ANN)^[12]、遗传算法(GA)^[13]和 BP 神经网络算法^[14]。

以上三种预测方法既各有局限,又各具优势。“组合预测”的策略应运而生。Bates 和 Granger 便是这策略的提出者。1969 年,上述两位学者首次提出了“组合预测”思想。他们认为,针对一个研究问题采用两种或者两种以上的单项预测方法,并且给代表不同独立有效信息的预测结果分配以适当的权重,可以得到精度更高的预测^[15]。

由此,“组合预测”的研究方法也正式成为学者们致力探索的新方向。较有代表性的预测模型如下:将时变参数模型(TVP)和误差修正模型(ECM)结合,成为 TVP-ECM 模型^[16]、几乎理想需求系统模型(AIDS)^[17]、ARIMA-BP 组合模型^[18],以及 SVR-ARMA 组合模型^[19]。

综上,在游客量预测的各种方法中,时间序列模型胜在对数据的低要求,计量经济模型优在对自变量和因变量的因果分析,人工智能方法则强在对结果的精准预测。而组合预测的方法既规避了上述各种方法的不足,又吸纳了各种方法的精髓,相对更为科学合理,得到的预测结果也更为准确可靠。要获得精准且具解释力的预测,构建组合预测模型是大势所趋。

(二)网络搜索数据的旅游行为预测相关研究

对游客量预测研究而言,历史统计数据的获取非常关键。在现有的文献中,对游客量进行预测研究的数据类型主要有结构化数据(比如:国家统计局公布的官方数据)和非结构化数据(比如:网络日志、图片、音视频文件)。不容否认,近年来,旅游统计为我国旅游业的发展起到了良好的支撑与促进作用,然而,历史数据的空白与笼统、旅游统计实践中纷至沓来的问题都让形势变得更为严峻。当前,国内诸多旅游目的地仍被视为“灰色系统”^①。

近年来,大量文献证明,通过利用互联网搜索行为数据进行经济与社会行为预测是切实可行的,且基于网络搜索行为数据的经济与社会行为预测已成为一个学术热点,并在国内外取得了较为丰富的研究成果。搜索引擎成为了人们搜索网络信息的主要渠道,亦成为反映人们消费行为、消费偏好的镜像^[20]。国际上有关网络搜索行为数据的预测应用已然涉及到了经济、社会以及健康领域。在经济领域中,搜索数据可用于消费以及失业率的预测;在社会和健康领域中,搜索数据可

① 通过国内外研究旅游预测常用方法模型对比发现,国内使用灰色系统理论(GST 或 GS)方法较多,但该方法在国外的类似研究中应用甚少。有学者将这一现象出现的原因归为:近年来,我国旅游产业发展迅猛,但与旅游产业相关的统计数据却仍有很多是空白或较为笼统。而灰色预测方法则更加擅长解决“样本量小、信息贫乏、确定性低”问题。

用于流感等疾病,以及公众注意力的预测^[21]。在国内,不少研究人员针对医学、经济学、传播学领域,利用网络搜索进行了相关研究,取得较多研究成果,主要在居民消费^[22]和金融市场^[23]方面有了一定程度的突破。

就旅游消费者出行行为预测研究而言,鉴于旅游者在做出旅游决策之前都会运用互联网工具大规模搜索旅游信息^[24],旅游者利用搜索引擎搜寻包括景点、交通和住宿等在内的旅游消费决策过程各阶段的相关信息,并最终进行旅游决策,完成旅游活动^[25],网络搜索行为数据为旅游决策者和学者们提供了一个分析旅游消费者出行行为的信息基础。学者们则开始关注如何运用网络搜索行为数据探寻旅游消费者出行行为^[26],网络搜索行为数据在旅游预测领域中的价值也逐渐凸显^[27]。然而,尽管现有研究成果从相当程度上验证了运用网络搜索行为数据预测旅游消费者出行行为的可行性;但是,鉴于网络搜索各关键词之间存在相关性,及其导致的各搜索词之间的共线性现象^[28],关键词的选取和选择方法成为构建有效预测消费者行为的关键环节。

在对网络搜索与目标变量间的关联机理进行细致探究的基础上,有学者提出,以“先选择单个关键词、再合成网络搜索指数”的局部视角从繁杂、庞大的网络搜索关键词中发掘大众“关注”的行业发展趋势不失为一条“两全其美”的有效路径^[29]。

(三)去除数据中的噪声因素的相关研究

非结构化信息的数据来源和表现形式多种多样。其主要优点在于信息更新快,数据实时可得;缺点则是信息噪声大、数据质量差^[30]。对于游客量预测研究来说,尤其是在以信息噪声大和数据质量差为基本特征的网络搜索行为数据被越来越广泛运用的背景下,噪声的干扰成为构建预测模型之前需要解决的难题。在运用这些数据进行游客量预测之前对数据进行预处理是有必要的。事实上,数据的预处理方法有很多种。其中,傅里叶变换和小波变换在信号处理领域中最理想,应用最广泛,效果最好^[31]。近年来,EMD 去噪方法也开始被大量地运用到数据处理和分析当中^[32]。

在游客量预测研究中,目前国内仅有少量文献对原数据去噪进行了应用论证。马丽君依据小波分析理论,以近 20 年来中国入境及国内旅游客流量成长过程为统计样本进行了滤波分析,提取了旅游成长过程中的趋势项和不同时间尺度的周期项^[33];陈玲玲引入 EMD 算法探讨入境旅游客源和外汇收入要素的变化特征、影响关系及其作用机制^[34];李晓炫提出利用 EMD 算法对数据进行噪声处理,利用去噪处理后的网络搜索行为数据对旅游客流量进行预测^[35]。上述尝试均取得了不错的效果。

可见,给游客量预测提供更高质量的数据内核,运用去噪算法对数据进行预处理很有必要。然而,旅游是一个由自然资源、社会环境、政治局势、宏观经济、消费心理等多种因素共同作用的复杂系统,单就游客量预测而言,偶然因素的发生也必不可免,因此,传统的诸如傅立叶分析和小波变换等擅长处理平稳过程的去噪算法并不适合游客量数据预处理。鉴于 EMD 去噪算法能够从数据本身出发,进而根据数据自身特性进行自适应分解,且并不需要先验条件,其在处理原始序列信号分解方面具有更为显著的优势。

二、构建组合模型

传统的游客量预测多依托于统计学模型,其数据收集需考虑统计分析目的,其过程包括设计调查方案、严格控制调查流程,故难以规避低效率、高成本的缺点。近年来,互联网广泛运用于各行各业,也为研究者提供了网络搜索行为数据、社交媒体数据以及卫星遥感数据等诸多搜寻信息渠道,数据来源的日益丰富也为游客量预测提供了新的分析范式^[36]。尤其是 Ginsberg 等学者证明了网络搜索行为数据在社会经济等领域的预测能力之后,获取大数据不再是重心,如何利用“旅游大数据”,获得更加贴近消费者行为的高质量数据才是当前提高景区游客量预测精准性的着眼点。鉴于旅游是一个由自然资源、社会环境、政治局势、宏观经济、消费心理等多种因素共同作用的复杂系统,大量的噪声干扰可能会影响游客量预测

效果,甚至造成严重偏差,且因互联网搜索行为数据存在信息噪声大、数据来源和形式多样化特征,无论是历史数据还是网络搜索指数往往是非线性和不稳定的,噪声干扰不可避免。没有了“降噪处理”这先行的一步,“预测精度”则变得无从谈起。

仅从游客量预测精度的角度看,在既有的游客量预测研究中,任何情况下都表现最优的模型并不存在^[37]。学者们往往会在充分考虑时间序列模型、计量经济模型和人工智能方法等各类预测模型和方法的基础上,为预测地区挑选切合其具体特征的预测模型。当然,现有研究表明,通过组合预测的方法,建立的模型是科学的、合理的,得到的预测结果是可靠的、准确的。基于不同预测手段和历史统计数据,进行组合预测,是有效改

善游客量预测模型精度的方式^[38]。综上,本文充分依托网络搜索行为数据的时效性,运用将网络搜索关键词进行指数合成的方法对与预测对象相关的网络搜索关键词进行有效的筛选和降维处理,并在预测游客量前运用 EMD 去噪方法对游客量和网络搜索行为历史数据进行去噪处理,再运用 ARIMA 模型和 BP 神经网络形成组合预测模型,探究互联网时代游客出行新特征,进而指导旅游经营和管理部门完善旅游安全预警体系和优化旅游资源配置体系,对促进旅游目的地服务质量的提高和旅游行业的可持续发展具有重要的理论探索与现实指导意义。

具体的游客量预测模型构建过程如下图 1 所示。

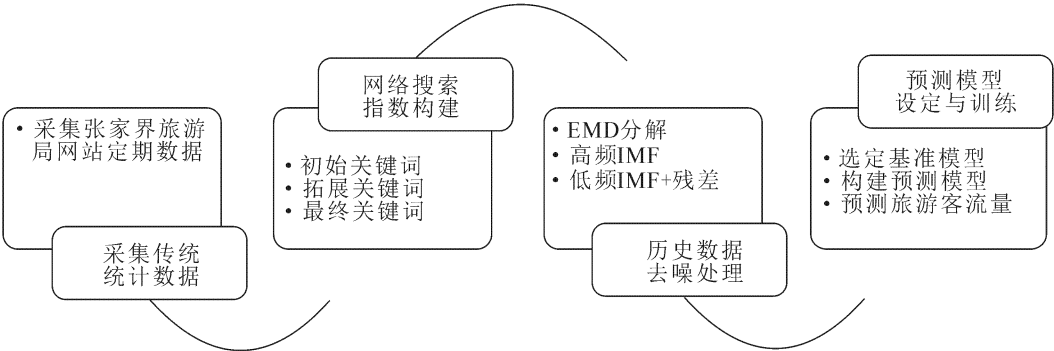


图 1 组合模型构建流程图

三、实证分析

(一)数据采集

张家界素有“扩大的盆景、缩小的仙山”美称。张家界以其得天独厚的旅游资源吸引了来自世界各地的大量游客前往旅游观光,这也由此引发了

游客安全和滞留的隐患^[39],根据媒体报道,张家界也曾多次发生因接待人数超载被投诉的事件。

因此,本文以张家界为例,基于百度指数^①和 EMD-ARIMA-BP 组合模型对张家界游客量进行预测。其中,张家界的游客量数据来自张家界统计信息网(<http://tjj.zjj.gov.cn>)发布的定期数据,本文以月为周期进行统计^②,截选 2011 年 1

① 因本文仅涉及国内游客量预测研究,根据 Statcounter 中的搜索引擎市场占有率排行榜(中国)历史统计数据,百度搜索在 2011.01—2018.03 的平均市场占有率达到了 70.05%。因此,考虑覆盖面和网民使用率,论文选择使用百度搜索引擎提供的关键词百度指数及搜索引擎使用率作为基础统计数据。此外,有学者运用 $I_{网络} = I_{百度} / R_{百度使用率}$ 公式对关键词的百度搜索指数进行修正以弥补关键词的百度搜索指数因百度搜索使用率的不断变化而无法全面反映关键词的网络搜索变化趋势的不足。本文经过测算认为,百度使用率的变化对关键词的网络搜索变化趋势的影响并不大,故未对关键词的百度搜索指数进行修正。

② 因张家界统计信息网发布的数据最小粒度为月,故本文以月为周期进行统计。

月至 2018 年 3 月共 87 个周期的游客量数据。网络搜索指数来自百度指数中的 PC 网络搜索指数,百度指数中的 PC 网络搜索指数来源于每天发布的前一天的 PC 网络搜索指数统计^①。考虑网络搜索指数的先行指标作用,搜索数据时间跨度包括 2010 年 8 月至 2018 年 3 月共 92 个周期的网络搜索指数^②。为保证模型的稳定性和可靠性,选择较大比重样本(75 个周期)作为训练集,剩余的作为预测集(12 个周期)。本文中综合考虑张家界旅游的淡旺季,选取 2011 年 1 月至 2017 年 3 月共 75 个周期为训练集,剩余的 2017 年 4 月—2018 年 3 月共 12 个周期为预测集。预测集包括一年内张家界旅游所有淡旺季。

(二)合成网络搜索指数^③

从百度给出的定义来看,用户关注度这一指标确实能够反映每类关键词对应的某类或某件事物的热门程度和被关注程度^[40]。本文网络搜索指数合成具体步骤如下:

第一步:关键词的选择直接影响后续的研究进展,甚至是研究的成败^[41]。根据黄先开等人的研究成果^[42],本文最终确定运用范围取词法选取关键词。根据马丽君等针对张家界游客信息需求相关研究^[43]和麻学锋等针对张家界旅游产品开发相关研究^[44],本文首先分别从餐饮、住宿、交通、游览、购物、娱乐以及其他 7 个类别出发筛选出若干个初始关键词,继而利用百度指数依赖于语义挖掘技术的需求图谱推荐功能,获得初始关键词的相关检索词作为拓展关键词,并设每一个相关关键词为 x_i 。

第二步:重复迭代上面的步骤,直到不再出现新的相关关键词推荐为止。将搜索量为零和未被收录到百度指数的关键词剔除,共剩余 27 个关键词(如表 1 所示);分别计算上述 27 个网络搜索关键词与游客量序列之间 0~5 期提前期的皮尔森相关系数。

表 1 网络搜索关键词

序号	关键词	序号	关键词	序号	关键词
1	张家界	10	张家界汽车站	19	张家界黄龙洞
2	张家界旅游	11	张家界门票	20	张家界金鞭溪
3	张家界地图	12	张家界旅游价格	21	张家界大峡谷
4	张家界天气	13	张家界旅游路线	22	天门山国家森林公园
5	张家界旅游地图	14	张家界旅游攻略	23	张家界特产
6	张家界小吃	15	张家界景点介绍	24	张家界自助游
7	张家界美食	16	张家界百龙天梯	25	张家界自由行
8	张家界住宿	17	张家界国家森林公园	26	张家界三日游
9	张家界酒店	18	张家界武陵源	27	张家界旅行社

① 百度搜索指数分为 PC 网络搜索指数、移动网络搜索指数和整体(PC 端+移动端)网络搜索指数。综合 PC 网络搜索指数的先行性优势和移动网络搜索指数发布时间(2011 年 1 月开始发布)与本文数据需求(自 2010 年 8 月开始搜集)考虑,本文的网络搜索指数特指 PC 网络搜索指数,不包括移动网络搜索指数。

② 谷歌搜索提供的趋势数据(Google Trends)针对特定关键词查询数量生成查询指数,将特定关键词网络搜索量与同一时段内最高网络搜索量比较,得出该词的相对搜索量,而百度指数反映的是通过百度进行搜索的关键词的绝对搜索量。且,谷歌趋势数据可以直接下载,百度指数只是在统计图中提供了动态数据,本文采用 webdriver 模拟用户登录,获取百度指数图片,再基于 tesseract 训练的指数识别程序得到具体百度指数数值。

③ 本文对运用范围选词法选定的 7 个关键词相互之间进行了相关性检验,发现,上述 7 个搜索关键词之间均存在强相关性(皮尔森相关系数大于 0.8),故采用合成网络搜索指数方法以消除数据间的共线性,提高模型的预测效果。

第三步:确定搜索词选择阈值^①。经过上一步的计算,每个搜索词均得到 6 个时滞相关系数,选取最大相关系数大于 0.8,并且该相关系数出

现在先行期大于 1 期的网络搜索关键词。根据两个筛选条件共选择 7 个网络搜索关键词(如下表 2 所示)。

表 2 最终关键词及其相关系数

关键词	相关系数	关键词	相关系数
张家界	0.866**	张家界自由行	0.815**
张家界美食	0.820**	张家界金鞭溪	0.806**
张家界景点介绍	0.807**	张家界黄龙洞	0.803**
张家界武陵源	0.816**		

注:**表示 5% 的差异显著性水平。

第四步:将上述 7 个网络搜索关键词加总,合成搜索指数 Index7,(如(1)式所示),

$$\text{Index7} = \sum_{i=1}^m x_i \quad (1)$$

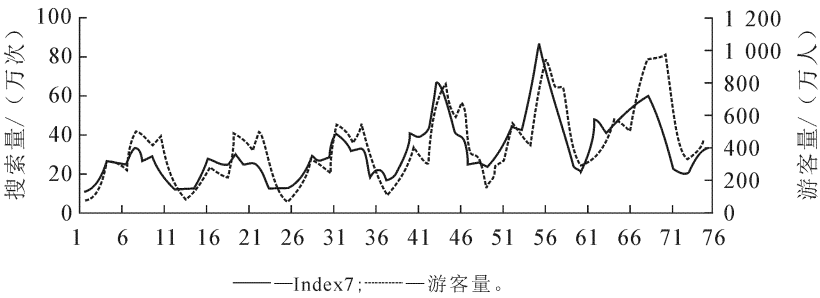


图 2 张家界客流量与搜索指数 Index7 相关图

图 2 中显示张家界游客量与合成搜索指数 Index7 的相关序列图,两个序列之间具有大致相同的变化趋势,合成搜索指数相对于张家界游客量有一期(一个月)的提前期。

(三)EMD 分解去噪

根据 EMD 去噪算法,对张家界客流量和网络搜索指数 Index7 分别进行 EMD 分解。本文利用 Python 语言实现张家界客流量和网络搜索指数 Index7 原始序列 EMD 分解。经过分解后

分别获得 4 条 IMF 函数序列(IMF-1-IMF-4)和一条残差,如图 3 和图 4 所示。其中 IMF-1 函数具有最高的频率,之后 IMF-2,IMF-3,IMF-4 等各条函数频率顺次降低,Residue 残差则接近单调函数。最高频率的 IMF-1 代表时间周期较短的随机波动,作为噪声处理。将剩余的各条 IMF 序列(IMF-2, IMF-3, IMF-4)与残差序列(Residue)做加总处理,分别作为低频客流量序列和低频搜索指数。

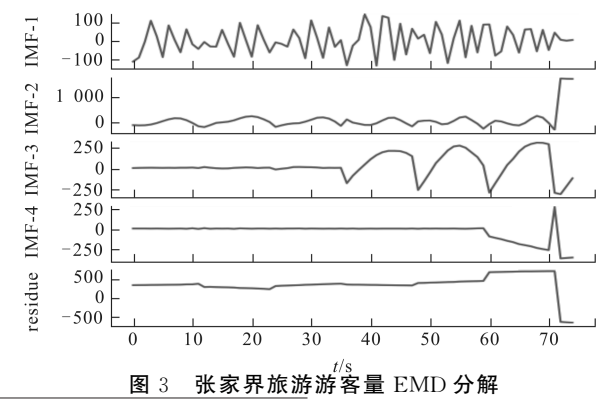


图 3 张家界旅游游客量 EMD 分解

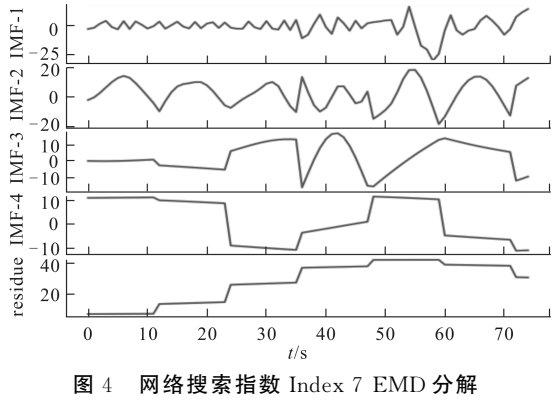


图 4 网络搜索指数 Index 7 EMD 分解

① 阈值的确定对于预测结果和噪声都有很大的影响,阈值过低会导致条件过于放松,从而降低搜索指数与游客量的相关性并包含过多的噪声干扰;而阈值过高则会导致关键词过于节俭,从而遗漏掉影响游客量的重要影响因素。此外,网络搜索关键词必须要有先行性,只有具有先行性的搜索词才具有预测能力。

(四)模型设定和训练

在游客量的预测中,大多学者选择使用时间序列模型和计量经济模型。本文则在考虑噪声干扰并对噪声进行分离处理的基础上,对张家界游客量进行预测。将客流量序列和网络搜索序列进行 EMD 分解,将 IMF-1 高频噪声从原始序列中分离出来,客流量低频部分由三条低频 IMF 序列和残差加总合成,并利用相应的低频网络搜索加总序列进行拟合预测。至于客流量高频部分,考虑到其非线性特点,利用 BP 神经网络进行预测。同时将张家界游客量时间序列 ARIMA 模型和未进行噪声处理的游客量与网络搜索模型,以及单纯利用 BP 神经网络预测模型作为三个基准模型:

$$\text{Travel}_t = c + \alpha_1 \text{Travel}_{t-1} + \mu_t, \tag{2}$$
$$\text{Travel}_t = c + \alpha_1 \text{Index7}_t + \alpha_2 \text{Travel}_{t-1} + \alpha_3 \text{Index7}_{t-1} + \mu_t, \tag{3}$$
$$Y_t = c + \alpha_1 X_t + \alpha_2 X_{t-1} +$$

$$\alpha_3 X_{t-2} + \mu_t. \tag{4}$$

其中,
 $Y = (\text{Travel} - L), X = (\text{Index7} - L)$ 。
 Travel_t 代表 t 时期张家界游客量序列,Index7 表示基于合成网络搜索指数的网络搜索序列, $(\text{Travel} - L)$ 和 $(\text{Index} - L)$ 则分别表示低频游客量序列和低频网络搜索序列。式(2)是张家界游客量的时间序列 ARIMA 模型,式(3)是网络搜索和客流量的计量回归模型,(2)和(3)均为基准模型,式(4)是低频网络搜索和低频客流量的计量回归模型。
通过对游客量序列进行单位根检验发现,该序列在 10% 显著性水平下通过单位根检验,即该序列是平稳序列。式(3)和式(4)的回归残差也均通过 1% 显著性水平的单位根检验,但是 DW 统计量显示残差存在序列相关,因此在回归基础上对残差进行序列相关的 ARIMA 调整,得到的计量回归结果如表 3 所示。

表 3 模型拟合结果

数据类型	时间序列模型		Index7		EMD 分解	
模型	M(1)		M(2)		M(3)	
解释变量	Travel _{t-1}	0.752 6*** (10.03)	Index7	11.919 8*** (13.039 5)	(Index7 - L)	8.370 6*** (5.306 6)
	c	429.863 9*** (6.41)	AR(1)	0.658 9*** (7.200 3)	AR(1)	1.340 4*** (14.443 0)
					AR(2)	-0.653 9*** (-6.765 3)
					c	134.133 4*** (2.314 7)
R ²	0.583 0		0.751 9		0.906 0	
AIC 值	12.776 3		12.256 9		11.185 8	
SC 值	12.838 5		12.319 2		11.311 3	
DW 值	1.686 4		1.657 2		1.657 7	
残差平稳性 检验		-3.294 7*		-7.152 1***		-7.223 0***
	1%	-4.113 0		-3.522 9		-4.090 6
	5%	-3.483 9		-2.901 8		-3.473 4
	10%	-3.170 1		-2.588 3		-3.163 9

注:*,*** 分别表示在 10%和 1%显著性水平,()中为 t 值。
从表 3 中可以看出,加入网络搜索指数的拟合效果优于时间序列模型,而经过 EMD 分解去噪后的低频网络搜索指数的计量模型拟合效果则比未分解的模型效果有显著提高。

具体的模型拟合结果如式(5) — (7)所示:
 $\text{Travel}_t = 106.338 + 0.752\,6\text{Travel}_{t-1} + \mu_t,$
 $\text{Travel}_t = 11.919\,8\text{Index7}_t + 0.658\,9\text{Travel}_{t-1} -$

$$7.854\ 4\text{Index}7_{t-1} + \mu_t, \tag{6}$$
$$Y_t = 42.05 + 8.37X_t \pm 11.22X_{t-1} + 5.74X_{t-2} + 1.34Y_{t-1} - 0.65Y_{t-2} + \mu_t. \tag{7}$$

其中:

$Y = (\text{Travel} - L), X = (\text{Index}7 - L)$ 。

拟合结果显示,网络搜索对张家界游客量的影响是正向的,即网络搜索的增加将会引起未来时期内游客量的增加。式(6)中,网络搜索合成指数增加 1 单位将会引起未来客流量增加 11.919 8 个单位。式(7)是基于 EMD 分解基础上的张家界旅游低频率序列和搜索数据合成指数低频率序列的拟合。分离出噪声之后的低频合成指数序列对低频游客量序列的影响相比原序列稍有减弱,说明分离出的噪声中不仅存在负向干扰客流量的噪声,也存在正向的干扰噪声,并且这种干扰的总体影响是正向的。所有模型的残差均通过 1% 显著性水平下的平稳性检验,且 DW 值也都在 2 附近,说明经过残差的 ARIMA 调整已经消除了序列相关。

张家界客流量的高频部分序列则利用 BP 神经网络进行模型训练和预测。

步骤一:将游客量数据转为两列,第一列为 T 月的游客数、第二列为 $T + 1$ 列的游客数,以第一列的游客数作为一输出 X_i ,第二列 $T + 1$ 为预测值 Y ;

步骤二:以网络搜索指数 Index7 作为另外一个输入 X_2 ;

步骤三:数据进行归一化:公式 $Z = X_i - \min(X_i) / \max(X_i) - \min(X_i)$,其中, X_i 为第 i 个向量 $\min(X_i)$ 为最小值, $\max(X_i)$ 为最大值。输入层为 2 个神经元,隐藏层为 25 个,输出层为 1 个神经元。学习率为 0.001,迭代次数为 1000 次。

经过训练,BP 神经网络样本内(经过 EMD 去噪方法处理后分解出来的网络搜索行为数据 IMF-1 值与游客量历史数据 IMF-1 值)的预测值与原值(游客量 IMF-1 高频数据)的 MAPE(%)

数值为 175;RMSE 数值为 66.55。

此外,作为基准模型的 BP 神经网络,本文运用同上数据处理方法,采用三层结构的神经网络,输入层为 2 个神经元,隐藏层为 25 个,输出层为 1 个神经元,学习率为 0.001,迭代次数为 1000 次。经过训练,BP 神经网络样本内(未经 EMD 去噪方法处理的网络搜索行为与游客量历史数据)的 MAPE(%) 数值为 25; RMSE 数值为 103.71。

相比于基准模型(1)和(2),BP 神经网络在模型的拟合优度上更优,模型的训练效果优于传统的线性模型。训练集内的原值和拟合值拟合效果如图 6 所示。

(五)模型预测

通过上述模型拟合训练,发现加入网络搜索后的模型效果优于时间序列模型,而通过 EMD 去噪后的低频率序列模型效果又进一步优于未经过 EMD 去噪模型。

在训练模型的基础上对张家界预测期游客量进行预测。分别为基于模型 1 的时间序列预测、基于模型 2 的网络搜索预测、基于 BP 神经网络的人工智能预测和基于模型 3 和高频 BP 神经网络的 EMD-ARIMA-BP 预测。在预测研究中,预测效果评估是重要的一个部分,而预测效果评估指数有很多,常用的包括 MAPE 和 RMSE。本文选择这两个最常用的预测效果评估指数衡量 EMD-ARIMA-BP 算法和其他三个基准模型的预测效果。MAPE 和 RMSE 计算公式如下:

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - A_i|}{|A_i|} 100\%, \tag{7}$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - A_i)^2}. \tag{8}$$

其中, Y_i 表示预测值, A_i 表示原值。

经过预测,未来 12 个周期各模型的原值、预测值、MAPE 和 RMSE 数值如表 5 所示,EMD-ARIMA-BP 算法和其他三个基准模型的 MAPE 值与 RMSE 值如表 6 所示。

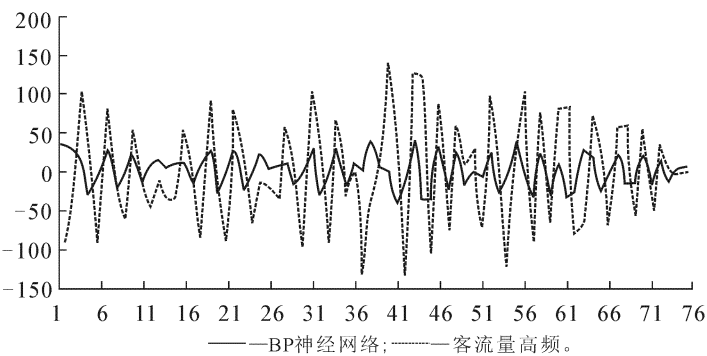


图 5 客流量高频部分 BP 神经网络训练拟合(单位:万人)

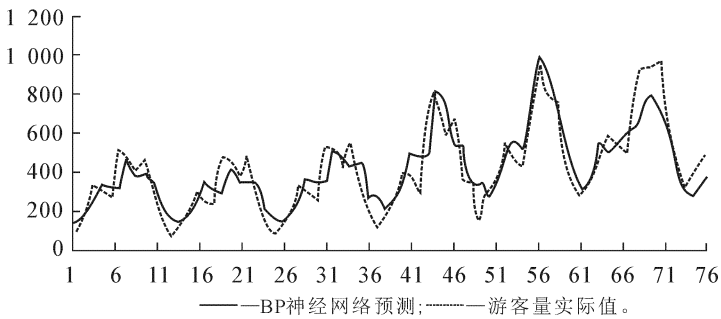


图 6 基准模型 BP 神经网络训练集拟合效果(单位:万人)

表 5 未来 12 期预测结果和预测误差

周期	2017.04—	1	2	3	4	5	6	7	8	9	10	11	12
原值	2018.03	588	547	500	764	938	935	970	630	401	321	393	483
预测值	时间序列	434	549	518	483	681	812	810	836	580	408	348	402
	Index7	471	642	593	618	752	744	672	617	472	350	388	455
	BP 神经网络	508	557	610	623	746	796	705	899	381	297	282	373
	组合模型	584	530	540	629	740	865	927	689	442	196	396	465
MAPE/%	时间序列	26	13	10	17	19	18	18	20	22	23	22	21
	Index7	20	19	19	19	19	19	21	19	18	17	16	15
	BP 神经网络	14	08	12	14	15	15	17	20	18	17	18	19
	组合模型	01	02	04	07	10	10	09	09	09	12	11	10
RMSE	时间序列	154	109	90	160	184	175	173	178	178	171	163	158
	Index7	117	107	102	115	132	144	174	163	156	148	141	135
	BP 神经网络	80	57	79	98	123	126	153	172	162	154	151	148
	组合模型	4	12	25	71	109	103	97	93	89	93	89	85

表 6 模型预测效果 MAPE、RMSE 评估结果

模型名称	MAPE/%	RMSE
时间序列	21.387 0	158.223 2
Index7	15.152 2	135.222 0
BP 神经网络	18.727 1	147.943 9
组合模型	10.458 9	85.246 2

表 6 中可以看出,在预测未来 12 个周期的张
家界游客量时,时间序列的 MAPE 和 RMSE 均
是最大的,说明时间序列模型的预测误差最大;其

次是 BP 神经网络和网络搜索指数。从 12 个周
期的预测结果发现(如表 5 所示),BP 神经网络在
短期预测和长期预测时表现较优,而在中短期预

测时,网络搜索指数的预测能力优于 BP 神经网络。

在四种预测模型中,基于网络搜索指数的 EMD-ARIMA-BP 组合模型表现显著优于其他基准模型,在 12 个周期内的预测误差均显著低于 ARIMA 模型和 BP 神经网络算法的预测误差。此外,相比于未进行 EMD 分解去噪的网络搜索预测模型,EMD-ARIMA-BP 组合模型的 MAPE

和 RMSE 均有显著下降,这也体现了 EMD 分解去噪对于提高游客量预测精度的重要作用。

图 7 表明了各预测模型的预测结果与原值的序列图。图中显示 EMD-ARIMA-BP 组合模型的预测值序列与原值最为接近,而相比之下,ARIMA 模型和 BP 神经网络基准模型的预测值与原值之间的预测误差均较大。

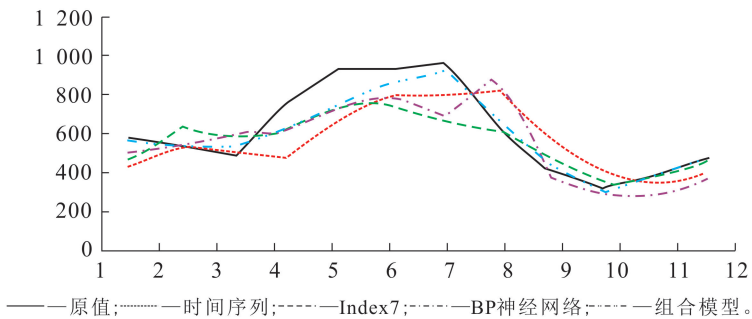


图 7 基准模型与组合模型预测值和原值序列图

注:在以上的分析中,确定网络搜索指数与预测目标张家界客流量之间的时滞性时,利用计算的 Pearson 相关系数中的最大相关系数确定网络搜索指数的先行性为提前 1 期。

五、结论与讨论

(一) 结论

1. 运用网络搜索行为数据预测旅游消费者出行行为切实可行。就游客量预测研究而言,历史数据的时效性极其重要。而与传统的游客量预测存在的几个月甚至更长滞后期相比,基于百度搜索的网络搜索指数仅有一周甚至更短的滞后期。本文研究发现,运用网络搜索行为数据预测旅游消费者出行行为切实可行,接近于实时的网络搜索行为数据可以做到提前一个月对游客量进行预测,其预测的时效性得到了大幅的提升。可以认为,网络搜索行为数据为旅游决策和学者们提供了分析旅游消费者出行行为的信息基础。

2. EMD 去噪方法对游客量与网络搜索行为数据进行去噪处理后,游客量的预测精度有较大提高。运用 EMD 去噪方法对历史数据做去噪处理在国外金融领域运用较为广泛^{[31]52-68},就旅游预测而言,在国内仅有少数几篇文章对运用 EMD 去噪方法对原始数据去噪进行过论证^{[34]106-118}。

如前文所述,旅游是一个由多种因素共同作用的复杂系统,偶然因素的发生不可避免。本文在时间序列等基准预测模型构建的基础上充分考虑噪声对网络搜索行为的干扰,并利用 EMD 分解将高频噪声从原序列中分离出来,进而利用经过噪声处理后的游客量以及网络搜索行为数据对张家界游客量进行预测,预测结果表明,经过噪声处理后的张家界游客量的预测精度有显著提高,这也充分证明了在预测中对噪声进行处理的重要性和 EMD 去噪方法在旅游大数据时代处理信息噪声大、质量不稳定、来源和形式多元的非结构化信息的有效性。

3. 本文构建的 EMD-ARIMA-BP 组合模型的预测误差显著低于 ARIMA 模型和 BP 神经网络等基准模型。通过组合预测的方法充分吸收各类单一评价模型的优点,同时又克服了单一评价模型自身的缺陷,进而建立更为精准的预测模型是一种趋势。本文将合成网络搜索指数方法与 EMD 去噪方法相结合,并利用 BP 神经网络对高频噪声进行预测,利用 ARIMA 对低频噪声进行预测;进而提出了基于网络搜索指数的 EMD-ARIMA-BP 组合模型,并将 ARIMA 模型和 BP

神经网络作为基准模型。通过计算各模型的各个阶段预测误差率,以及预测模型的 MAPE 和 RMSE 发现,基于网络搜索指数的 EMD-ARIMA-BP 组合模型的预测误差显著低于基准模型。该预测模型可以在未来为更精确地预测游客量和其他预测提供思路和参考,从而提高旅游目的地和相关支持产业的资源配置效率,并进一步保证旅游和其他行业的平稳运行。

(二)讨论

1. 本文基于黄先开等人的研究成果,直接采用范围取词方法选取关键词;并在参考马丽君等人针对张家界游客信息需求相关研究的基础上,最终合成了 Index7。从本文的预测结果看,用范围取词方法选用关键词构建的搜索指数对游客量预测具有较好的解释度。然而,在基于网络搜索的信息分析与预测研究成为学界广泛关注的研究内容时,对关键词选择的基本过程进行进一步优化对大数据时代的旅游预测显得极为必要。在下一步的研究中,应从关键词的“集中度”和关键词的“相关度”等角度出发,对面向信息分析与预测的网络搜索关键词选择的基本过程进行系统优化。

2. 本文在运用 EMD 去噪方法时,根据数据频率变化基本规律,直接默认 IMF-1 为噪声序列、之后的 IMF-2、IMF-3、IMF-4 和残差为有效数据序列,而未探索上述不同分量与原始时间序列的关联程度。考虑到相关度越高的 IMF 分量可能对原始序列的影响越大,且每个 IMF 分量所包含的原始序列的信息不同。在下一步的研究中,应该首先对不同 IMF 分量与原序列的关联程度进行分析探索,进而根据不同分量与原始时间序列的关联程度对 IMF 分量进行再组合,以达到既减少分量的个数,也保留原序列中重要信息的研究目的。

3. 本文在确认网络搜索行为数据与预测研究对象张家界客流量之间的时滞性时,只是利用网络搜索指数 Index7 与游客量之间的皮尔森相关系数选定网络搜索行为数据的先行性为提前 1 期。鉴于在预测中,预测因子的先行性与预测精度之间的取舍关系(即,先行性越低则预测精度越

高,相反先行性越高则预测精度越低),在接下来的研究中应该以原始序列周期时长为依据,对相应的利用网络搜索行为数据合成的搜索指数进行多期预测,以检验不同滞后期对预测误差的实际影响,进而根据实际预测误差情况选择最佳先行期数。

致谢:本文在写作过程中得到中南林业科技大学廖小平教授的悉心指导,在此谨致谢忱!

参考文献:

- [1] 国家发展和改革委员会. 2017 年中国居民消费发展报告[R]. 北京:人民出版社,2018:3-5.
- [2] HAIYAN SONG, GANG LI. Tourism demand modeling and forecasting—A review of recent research[J]. Tourism Management, 2008(2):203-220.
- [3] 李丹,赵媛. 对中国国际入境旅游市场的滚动样本预测——基于季节效应 ARIMA 模型[J]. 经济问题, 2008(6):124-126.
- [4] 刘霁雯,梁峰,向华. 基于 GARCH 族模型的中国旅游酒店板块指数收益率波动分析[J]. 统计与信息论坛, 2010(4):63-68.
- [5] 吴良平,张健,王汝辉. 基于 SARIMA 模型的汶川大地震对四川省旅游业影响研究[J]. 旅游论坛, 2012(3):56-60;86.
- [6] 杜勇宏,王汝芳. 季节时间序列理论发展综述[J]. 技术经济与管理研究, 2009(6):3-9.
- [7] 周成,冯学钢. 上海入境旅游市场的季节特征、波动周期与发展趋势——基于 X-12-ARIMA 和 HP 滤波法[J]. 旅游科学, 2016(6):39-53.
- [8] 王领,苟晓敏. 中国游客赴日旅游消费的影响因素研究——基于 X12-ARIMA 模型的分析[J]. 山东财经大学学报, 2018(2):86-95;104.
- [9] 吴良平,张健. 中国入境客流区域集散的结构差异与转移规律研究[J]. 旅游科学, 2017(6):14-29.
- [10] 魏瑾瑞,崔浩萌. 基于网络搜索行为数据的区域旅游指数及其微观动态:以西安为例[J]. 系统科学与数学, 2018(2):177-194.
- [11] 李维维,马晓龙. 国内旅游经济周期与宏观经济周期同步吗?——基于 MS-VAR 模型时变特征的验证[J]. 旅游学刊, 2017(11):49-59.
- [12] 古冰. 基于投入产出法及 ANN 模型的文化产业和旅游产业融合分析[J]. 商业经济研究, 2017(18):170-173.

- [13] 陈荣,梁昌勇,谢福伟,等.基于自适应 GA-SVR 的旅游景区日客流量预测[J].中国管理科学,2012(S1):61-66.
- [14] 邓祖涛,陆玉麒.BP神经网络在我国入境旅游人数预测中的应用[J].旅游科学,2006(4):49-53.
- [15] BATES J M,GRANGER C W J. The combination of forecasts[J]. Journal of the Operational Research Society,1969,20(4):451-468.
- [16] 谈正达,唐琳,胡海鸥.我国国际资本流动与货币冲销的有效性研究:1999—2010[J].国际金融研究,2011(12):29-37.
- [17] 王国新,杨晓娜.基于 AIDS 模型的入境旅游者消费结构和消费选择分析——以杭州市为例[J].旅游论坛,2014(6):23-31.
- [18] 雷可为,陈瑛.基于 BP 神经网络和 ARIMA 组合模型的中国入境游客量预测[J].旅游学刊,2007(4):20-25.
- [19] 梁昌勇,马银超,陈荣,等.基于 SVR-ARMA 组合模型的日旅游需求预测[J].管理工程学报,2015(1):122-127.
- [20] PAN B,MACLAURIN T,CROTTS J C. Travel blogs and the implications for destination marketing[J]. Journal of Travel Research,2007,46(1):35-45.
- [21] GINSBERG J,MOHEBBI M H,PATEL R S,et al. Detecting influenza epidemics using search engine query data[J]. Nature,2009,457(7232):1012.
- [22] 张崇,吕本富,彭廉,等.网络搜索行为数据与 CPI 的相关性研究[J].管理科学学报,2012(7):50-59,70.
- [23] 刘颖,吕本富,彭廉.网络搜索对股票市场的预测能力:理论分析与实证检验[J].经济管理,2011(1):172-180.
- [24] 张捷,刘泽华,解杼,等.中文旅游网站的空间类型及发展战略研究[J].地理科学,2004(4):493-499.
- [25] 孙焯,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(3):152-160.
- [26] SMERAL E. Tourism Forecasting Performance Considering the Instability of Demand Elasticities[J]. Journal of Travel Research,2017,56(7):913-926.
- [27] LI X,PAN B,LAW R,et al. Forecasting tourism demand with composite search index[J]. Tourism management,2017,59:57-66.
- [28] 董现垒,Bollen Johan,胡蓓蓓.基于网络搜索行为数据的中国消费者信心指数的测算[J].统计与决策,2016(5):9-13.
- [29] 孙毅,吕本富,陈航,等.基于网络搜索行为的消费者信心指数构建及应用研究[J].管理评论,2014(10):117-125.
- [30] 刘涛雄,徐晓飞.互联网搜索行为能帮助我们预测宏观经济吗?[J].经济研究,2015(12):68-83.
- [31] SONG H,WITT S F,JENSEN T C. Tourism Forecasting: Accuracy of Alternative Econometric Models[J]. International Journal of Forecasting,2003,19(1):123-141.
- [32] ATHANASOPOULOS G,SONG H,SUN J A. Bagging in Tourism Demand Modeling and Forecasting[J]. Journal of Travel Research,2018,57(1):52-68.
- [33] 马丽君,孙根年,王洁洁.基于本底趋势线与小波函数的中国旅游成长及多周期分析[J].旅游科学,2009(6):21-27.
- [34] 陈玲玲,杨慧.我国入境旅游客源与旅游外汇收入的多尺度对比研究——EMD 方法的新应用[J].地域研究与开发,2011(6):94-97.
- [35] 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(1):106-118.
- [36] 蔡跃洲,张钧南.信息通信技术对中国经济增长的替代效应与渗透效应[J].经济研究,2015(12):100-114.
- [37] WITT S F,MOUTINHO L,HUANG K H,et al. Tourism demand modelling and forecasting[M]//CABI Publishing,Strategic management in tourism,2018:45-72.
- [38] YE Q,SONG H,LI T,et al. Tourism Management Perspectives[J]. Tourism Management,2018,26:89-96.
- [39] 黄炜,孟霏,徐月明.游客环境态度对其环境行为影响的实证研究——以世界自然遗产地张家界武陵源风景区为例[J].吉首大学学报(社会科学版),2016(5):101-108.
- [40] 任乐,崔东佳.基于网络搜索行为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.
- [41] 卢洪涛.面向信息分析与预测的网络搜索关键词集中度和相关度研究[D].武汉:武汉大学,2013.

[42] 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(11):93-100.

[43] 马丽君,郭留留. 基于网络关注度的北京市居民对5A级景区旅游需求时空特征分析[J]. 干旱区资源

与环境,2017(10):203-208.

[44] 麻学锋,杨雪. 张家界旅游产品开发的空間演化与乡村振兴分析[J]. 吉首大学学报(社会科学版),2018(3):48-58.

(责任编辑:栗世来)

On the Prediction of Tourist Volume Based on Network Search
Index and EMD-ARIMA-BP Combination Model:
A Case Study of Zhangjiajie

LU Lijun^{1,2,3}

(1. College of Tourism, Central South University of Forestry and Technology, Changsha 410004, China ;2. College of Economics and Management, Hunan Institute of Technology, Hengyang 421008, Hunan China ;3. Hunan Human Settlement Environment Research Base, Hengyang Normal University, Hengyang 421008, Hunan China)

Abstract: Scientific prediction of tourist volume is helpful to perfect the early warning system of tourism security and optimize the allocation system of tourism resources. In order to further improve the accuracy of tourist volume prediction, a combination model of EMD-ARIMA-BP neural network based on web search index is proposed to explore the new rules of travel behavior of tourism consumers in the Internet age. The model firstly synthesizes the web search behavior data exponentially, using the EMD algorithm to deal with the noise of the visitor volume and the web search behavior data, combining the econometric prediction model and the BP neural network model to predict tourist volume. The empirical analysis takes the prediction of tourist volume in Zhangjiajie as an example. The results are as follows: (1) it is feasible to predict the travel behavior of tourism consumers by using web search behavior data, and real-time network data can greatly improve the timeliness of prediction; (2) after de-noising the data of tourist volume and web search behavior with EMD de-noising method, the prediction accuracy of tourist volume is improved to a great extent; (3) the prediction error based on the combination of network search index and EMD-ARIMA-BP neural network model is significantly lower than the three benchmark models of ARIMA time series, econometric prediction model and BP neural network.

Key words: network search index; ARIMA model; EMD algorithm; BP neural network; prediction of tourist volume