DOI:10.13718/j.cnki.xsxb.2020.10.012

基于百度指数时空分布的旅游趋势预测研究
——以上海市为例①

康俊锋[1]，　郭星宇[1]，　方　雷[2]

1. 江西理工大学 建筑与测绘工程学院，江西 赣州 341000；
2. 复旦大学 环境科学与工程系，上海 200433

摘要： 科学、准确、便捷、低成本地预测旅游趋势对提高景区的科学管理能力及避免因旅游人数过多导致的公共安全问题具有重要意义. 研究选取 2011－2018 年中国各省级行政区(港澳台除外)与上海市旅游相关的百度指数数据和上海市国内游客数据构建旅游趋势预测模型. 通过 Granger 因果检验、ARIMA 模型挖掘公众网络搜索行为与现实旅游行为的映射关系；依据百度指数数据的时空分布规律，采用支持向量机方法对百度指数数据进行聚类，解决不同省份百度指数因变化趋势近似而造成的多重共线问题，优化后的预测模型平均预测精度提升 23.36%. 研究发现：①昨天的搜索者就是今天的旅游者；②基于地理位置的旅游空间距离与旅游出游率呈反比、百度指数的地理位置属性有助于提升预测精度.

关　键　词： 百度指数；旅游预测；时空分布；ARIMA 模型；支持向量聚类；地理信息系统

中图分类号： F59　　**文献标志码：** A　　**文章编号：** 1000－5471(2020)10－0072－10

随着经济水平的提高，我国旅游业蓬勃发展，旅游市场规模与日俱增. 旅游趋势预测不仅有助于景区管理者根据预测的游客数量动态调整景区接待能力，平衡景区服务质量与运营成本间的关系；也能帮助游客规避人流高峰，提高旅游品质[1-4]. 面对游客数量突然的爆发，不少景区都发生过接待能力超载的事件. 例如，2013 年的国庆长假第二天九寨沟游客人数就已超过 5 万，致使数以千计的游客不得不滞留在景区内直至后半夜[2]；2014 年“国庆”期间故宫游客量最高竟达 17 万人次，远超故宫规定的每日最大承载量(8 万人次)；2017 年武汉东湖生态旅游风景区，春节期间累计接待游客 30 万人次，单日客流量最高同比增长达到 650%. 过多游客涌入景区导致景区无法提供应有的游览体验，此类事件不仅伤害了游客，更使景区的声誉受损、接待能力被质疑. 对于城市而言，游客超过预期的汇聚易引发公共安全事件，上海市在 2014 年 12 月 31 日爆发了震惊世界的“外滩踩踏”事件，致 36 人死亡，严重损害了上海市的旅游声誉. 为避免游客数量超过景区接待能力及防范公共安全事件，建立准确的旅游趋势预测系统是每个景区及城市迫在眉睫的任务.

旅游趋势预测一直是旅游研究领域的重点研究方向，传统的旅游趋势预测依托于对景区历史游客数量进行建模预测，但传统的预测模型受限于历史数据量少、数据时效性差等因素，导致预测模型难以满足实际需要[1]. 互联网搜索数据源于网络用户搜索行为，在搜索时就能进行数据统计而非事后统计，较传统统计数据更具时效性；并且互联网搜索数据公开免费使用，获取便捷，相比传统抽样问卷调查和访谈，可以

① 收稿日期：2019－06－20
基金项目：国家重点研发计划项目(2016YFC0803105)；国家留学基金资助项目(201808360065)；江西省教育厅科学技术研究项目(GJJ150661)；国家自然科学基金青年基金资助项目(41701462).
作者简介：康俊锋(1978－)，男，副教授，博士，主要从事高性能 GIS 算法及应用研究.
通信作者：方　雷，博士.

节省大量经济成本和时间成本且数据更具代表性[1,5-6]. 利用互联网搜索数据构建预测模型，其预测精度及可用性已被很多研究证明[5,7-9]. 本研究通过挖掘互联网空间与现实世界的映射关系，分析百度指数数据的时空分布特征，结合 ARIMA 模型设计出基于百度指数的旅游趋势预测方法，可为景区及城市旅游管理部门提供旅游管理决策依据.

1　文献综述

利用互联网搜索数据开展预测已成为各行业的研究热点，利用互联网搜索数据可以进行气候预测[10-11]、疾病预测[12]、房价预测[13]、失业率预测[9,14]、经济预测[15]等研究. 近年来，国内外众多学者也开始研究互联网搜索数据在旅游中的应用. 黄先开等[16]以北京故宫为例，发现在 ARMA 模型加入百度指数能有效提升预测精度；李山等[1]通过对一批 5A 级景区的百度指数及旅游人数进行统计和分析研究百度指数的前兆效应；Yang Xin 等[5]发现百度指数和谷歌趋势上不同关键词对应不同的旅游时间滞后期；Oscar Claveria 等[17]对比多种预测模型得到旅游预测精度，发现自回归积分移动平均模型(Autoregressive integral moving average model，ARIMA)在整体上预测结果最优；马莉等[18]研究了旅游客流与网络关注度的时空特征；何小芊等[19]通过对旅游网络关注度分析发现国内温泉旅游是一种非天然温泉依赖旅游活动；Bing Pan[20]人发现不同级别的旅游网络关注度具有不同的幂律分布. 已有研究大多采用多个与预测对象相关的搜索关键词作为自变量，收集互联网搜索数据的历史数据形成时序数据，然后运用线性回归、灰色预测等方法构建预测模型[5,7-8,21-22]；在基于更大数据量或更高时间分辨率数据基础上，利用互联网搜索数据的时间特征来提高预测精度[16].

综上所述，国内外运用互联网搜索数据进行旅游趋势预测已经成为目前旅游研究的热点[6,13,19]，但研究大都停留在使用互联网搜索数据的大数据量和实时特性结合传统统计模型建立预测模型这一阶段. 已有研究对互联网搜索数据本身具备的其他特征研究较少，重视数据的数值却忽视了数据属性的价值，如较少对地理位置属性进行深入挖掘. 另外已有研究未深入讨论搜索关键字选择的原因及可能会带来的差异. 因此，本研究选取 2011 年至 2018 年来上海市旅游的中国旅游人数数据，在选择出最优的互联网搜索关键词后，依据省级行政区划分别收集同一个搜索关键词在不同地区的百度指数数据，采用 Granger 因果检验方法挖掘百度搜索数据和实际旅游人口的因果关系，分析互联网搜索数据的时空分布规律，基于百度指数的空间特征及时间特征进行旅游趋势预测.

2　研究区域及研究数据

上海市是中国也是全球最著名的旅游城市之一，2017 年数据显示上海拥有 99 个 A 级景区，34 个红色旅游基地，229 家星级宾馆和 1 578 家旅行社；上海市南濒杭州湾，北面、西面与江苏、浙江 2 省相接，地理位置优越，旅游市场广阔.

在确定构建预测模型所需的网络搜索关键词为“上海旅游”之后，通过编写爬虫程序收集从 2011 年1 月 1 日至 2018 年 12 月 31 日的每日百度指数数据. 由于百度搜索服务在香港、澳门和台湾地区市场占有率较低，因此将上述 3 个地区的数据排除，以减少不具代表性数据带来的干扰. 同时，通过上海市旅游局网站获取 2011 年至 2018 年上海市国内旅游人数月报数据. 通过降频处理将每日百度指数数据转变为月度百度指数数据，从而匹配上海市旅游人数月报的时间频率. 图 1 为中国不同省级行政区下“上海旅游”的百度指数图，对图 1 分析可知：旅游相关的网络搜索量呈伴随旅游距离增加而减少的特征，是旅游空间距离与旅游出游率呈反比的有力体现.

3　研究方法

3.1　格兰杰因果关系

格兰杰因果关系检验能够检验变量之间是否存在统计学上因果关系，其判断结果是建立旅游趋势预测模型的前提条件[5,16]. 格兰杰因果关系检验运用于时间序列数据时，2 个变量 X,Y 之间的格兰杰关系定义

为：若在包含了变量X,Y的过去信息的条件下，对变量Y的预测效果要优于只单独由Y的过去信息对Y进行的预测效果，即变量X有助于解释变量Y的将来变化，则认为变量X是引致变量Y的格兰杰原因[23].

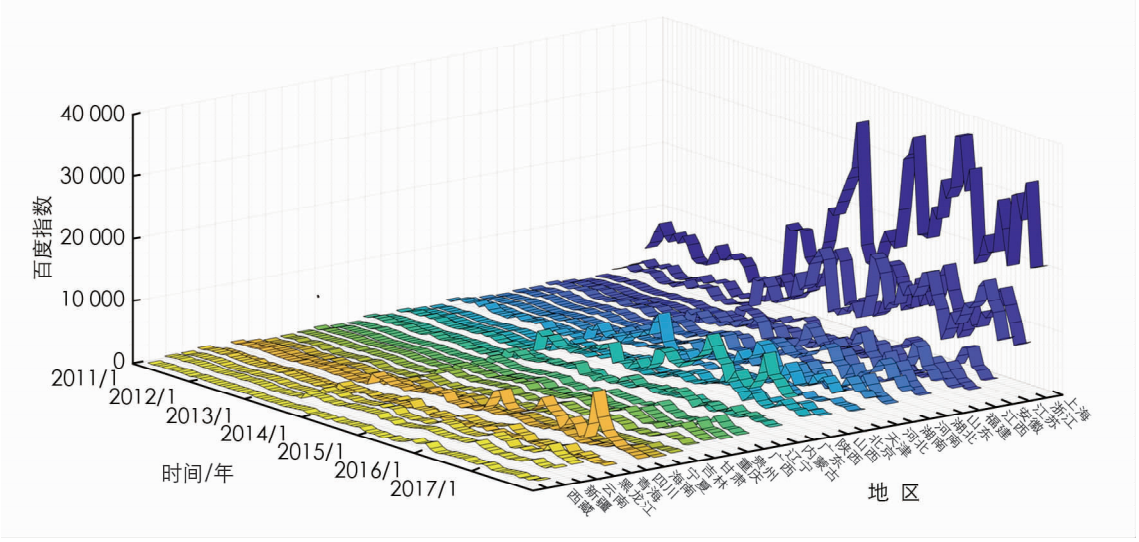

图1 中国不同省份关于“上海旅游”的月度百度数据(2011－2018年)

3.2 ARIMA预测模型

ARIMA模型能够更准确地预测季节性变化，而旅游活动是明显的季节性活动，因此本研究采用ARIMA模型构建旅游趋势预测模型，其公式如下：

$$(1-\sum_{i=1}^{P}\phi_i L^i)(1-L)^D x_t=(1+\sum_{i=1}^{Q}\theta_i L^i)\varepsilon_t \tag{1}$$

式中：P代表自回归项阶数；D代表序列差分阶数；Q代表移动平均项数.

3.3 支持向量聚类

支持向量聚类是一种使用支持向量机作为工具的无监督非参数型的聚类算法，其基本思想是将数据集中的数据样本通过非线性变换映射到高维特征空间中，在高维特征空间中一个超球面使其能包围全部样本点，超球面映射数据集时可以把数据集分割成任意几类[24]. 支持向量聚类的数学模型如下：

$$\begin{cases}\min R^2+C\sum_i\xi_i\\ s.t.\ \|\phi(v_i)-a\|^2\leqslant R^2+\xi_i\\ \forall i,\ \xi\geqslant 0,\ i=1,2,\cdots,N\end{cases} \tag{2}$$

式中：$\|-\|$是欧式范式；a是超球体的球心；ϕ是从原始空间到高维空间映射的非线性Y映射函数；ξ_i是松弛变量，允许一些样本点位于超球体的外部；R是超球体的半径，而$C\in[0,1]$是一个惩罚参数.

4 实证研究

4.1 搜索关键词的确定

百度指数(http://index.baidu.com)是百度公司提供的关于网民网络搜索行为的数据分享平台. 百度搜索作为全球最大的中文搜索引擎，使得百度指数能够收集到海量反映网民行为的互联网搜索数据. 百度指数提供2011年至今的不同关键词的在线每日搜索量数据，本研究将上述互联网搜索数据称之为百度指数. 百度指数的趋势变化反映网络关注度的变化，通过分析百度指数能够发现网民的关注和需求变化.

以往搜索关键词的确定多采用主观取词法[16]、范围取词法[5]、规则取词法[6]来确定预测模型所用的网络搜索关键词，上述方法使用方便，但可能遗漏关键词. 本研究采用的是技术取词法，它是利用网络技术尽可能获取所有搜索关键词的每日数据，并计算词与预测目标之间的相关性，然后基于相关性分析结果确定模型关键字.

以上海市旅游趋势预测为例选取搜索关键词，本研究通过收集与上海市旅游相关的多个搜索关键词的百度指数数据，选择语义上与上海旅游趋势预测关联性较为紧密的“上海旅游”为基准挑选建模用数据，

从众多搜索关键词中挑选出在搜索量上与“上海旅游”近似或与上海旅游人数的相关性与“上海旅游”近似的网络搜索关键词作为建模用数据[5-6,16,25]. 图 2 为基于“上海旅游”的建模用搜索关键词筛选图，以“上海旅游”的百度指数数值为基准展示了不同关键词百度指数数值的多寡. 由图 2 可知与上海旅游预测关联度最大的词为“上海旅游”“上海旅游攻略”和“上海迪士尼”，其余关键词与上海旅游相关程度较小且搜索量少，在进行预测时可以忽略.

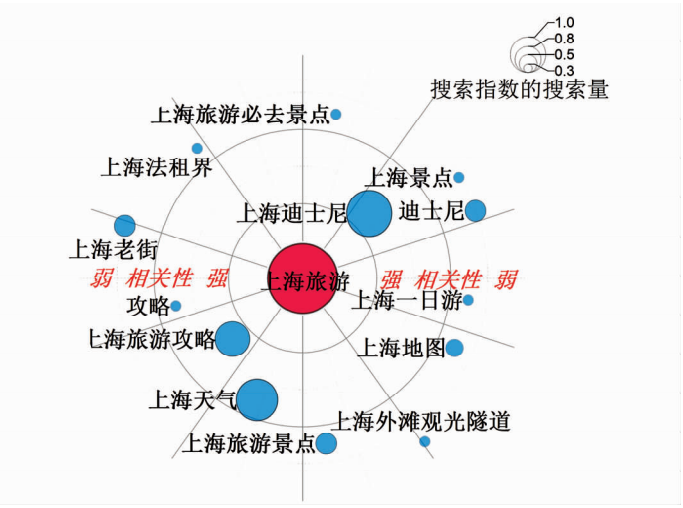

图 2　搜索关键词筛选

依据上述数据绘制图 3，用于分析上海市月度国内旅游人数与百度指数趋势变化关系. 由于“上海迪士尼”的变化趋势在 2016 年后多次出现爆发性增长从而产生了极大值，为了更好地对比不同数据的变化趋势，图 3 的 Y 轴采用了对数坐标轴. 由图 3 可知，上海市月度国内旅游人数变化呈平稳变化趋势，将旅游人数与其他关键词的趋势变化进行对比，发现“上海旅游”整体上能够提前一个月的时间反映上海市国内旅游人数变化趋势，且“上海旅游”在整体上比“上海迪士尼”能更稳定地反映国内旅游人数趋势变化；在搜索量方面“上海旅游”比“上海旅游攻略”大，显示“上海旅游”为更多人使用，更具代表性. 综合上述分析，本研究确定使用“上海旅游”作为上海市旅游趋势预测模型的搜索关键词，并使用“上海旅游”月度百度指数与上海实际月度国内旅游人数数据构建上海市国内游客趋势预测模型.

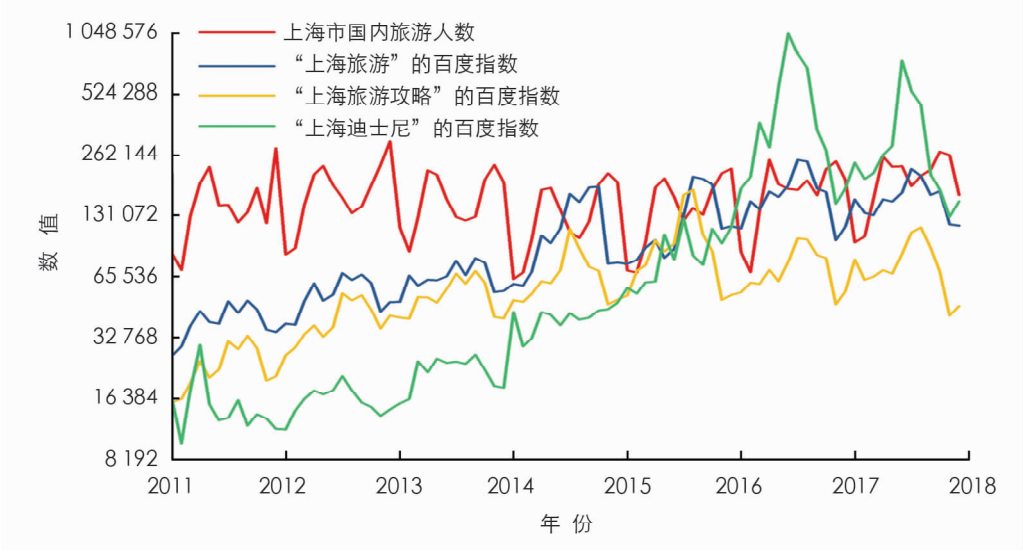

图 3　上海市月度国内旅游人数与百度指数趋势变化对比图(2011—2018 年)

分析图 3 还可知，百度指数的增幅大于实际旅游人数的增幅，而且伴随时间推移，两者差距越发加大，由此看出越来越多的游客在出发前使用互联网来搜索旅游目的地的相关信息，这是对民众互联网接入数量大幅增加和旅游市场受人口增长制约而小幅增长的客观反映；百度指数与游客数量呈现明显的月度变化趋势，波峰一直以来都在 9 月、10 月出现，波谷出现在 1 月，这体现了第三季度气候适宜旅游且“十一”国庆长假等假期是出游高峰.

4.2　旅游趋势预测研究

预测模型使用月度百度指数来预测上海市月度国内旅游人数，同时为了说明上海市旅游相关的百度指数与上海市的国内旅游人数存在相关性，本研究对二者进行了格兰杰因果关系检验. 完成上述步骤后再构建基于百度指数的 ARIMA 旅游趋势预测模型.

4.2.1 格兰杰因果关系检验结果

格兰杰因果关系检验结果能验证昨天的搜索者就是今天的旅游者这一理论假设是否合理[5,16]. 协整关系的存在是变量间能够进行格兰杰因果关系检验的先决条件，通过协整检验数据，确定“上海旅游”与上海市国内游客数量之间存在协整关系. 格兰杰因果关系检验结果如表1所示，其中SHLY代表“上海旅游”，SHYK代表上海市国内游客数量. 再依据赤池信息量准则(Akaike Information Criterion，AIC)和贝叶斯信息准则(Schwarz Criterion，SC)确定格兰杰因果关系检验的滞后期，AIC和SC结果显示格兰杰因果关系检验的滞后期为1时最优.

由表1可知，“上海旅游”的百度指数与上海市国内游客数量存在单方面的因果关系，结合图2，可以确定“上海旅游”的百度指数与上海市国内游客数量存在长期正相关性，即“上海旅游”的百度指数的增长可以预示上海市国内游客数量将会增长，而上海市国内游客数量不能用来预测“上海旅游”的变化. 格兰杰因果关系检验的结果是对“昨天的搜索者就是今天的旅游者”这一理论假设的有力证明.

表1 变量的格兰杰因果关系检验结果

滞后长度	格兰杰因果性	F值	F的p值	结论
1	SHLY不是SHYK的格兰杰原因	3.984 40	0.049 3	拒绝
	SHYK不是SHLY的格兰杰原因	1.887 11	0.174	不拒绝

4.2.2 ARIMA模型最优参数的抉择

参考前人研究[5-6,11,16,26-27]，经过序列平稳性检验，本研究确定变量数据无须进行差分处理，所以D取值为0. 而对于P和Q的取值，通过绘制上海市旅游数据的自相关函数(ACF)与偏自相关函数(PACF)图来确定旅游预测模型的最优参数. 分析图4发现，在滞后1阶后ACF图显示数据开始大部分落在边界值内，但在滞后3阶处存在越界现象；再观察PACF图，在滞后4阶后就少有值越界，因此确定P和Q的取值范围均在0～4之间. 为避免主观臆断的干扰，本研究选择贝叶斯信息准则BIC作为模型参数优劣的评判标准，并绘制出了图5. 图5显示P取0、Q取1时，BIC值最小、模型最优. 结合图4与图5，本研究最终确定预测模型为ARIMA(0，0，1)型.

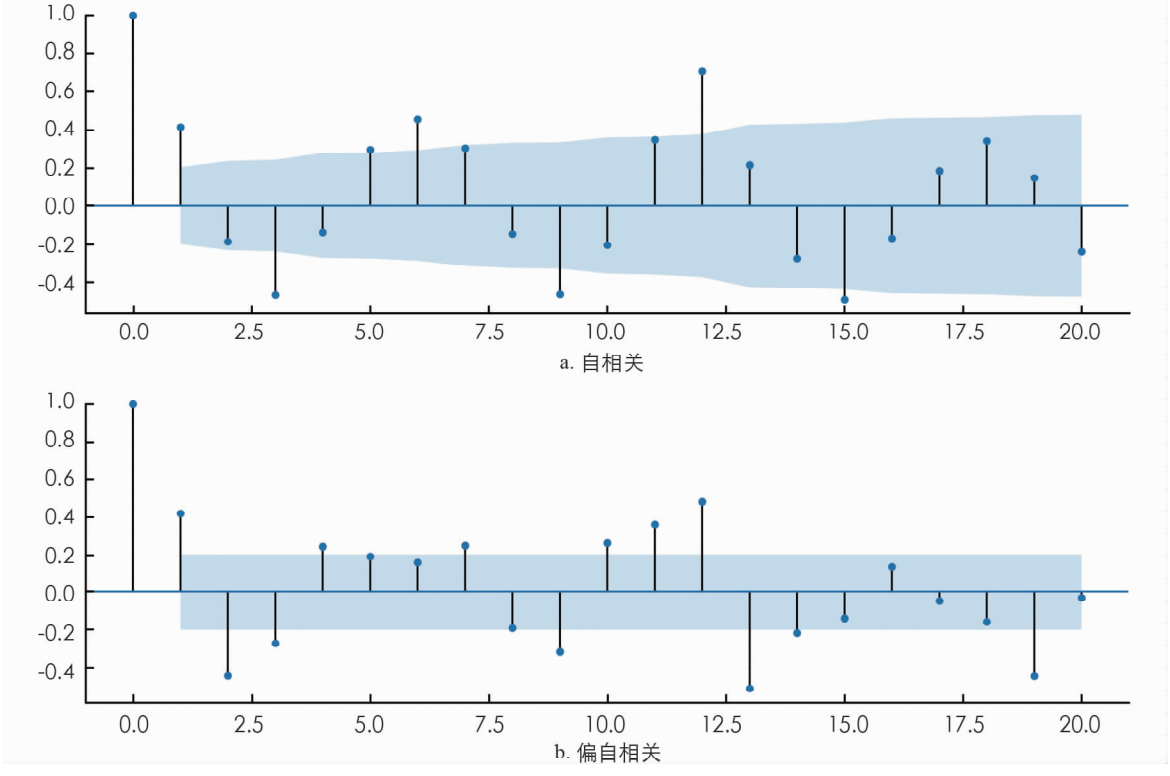

图4 上海市旅游数据的ACF与PACF

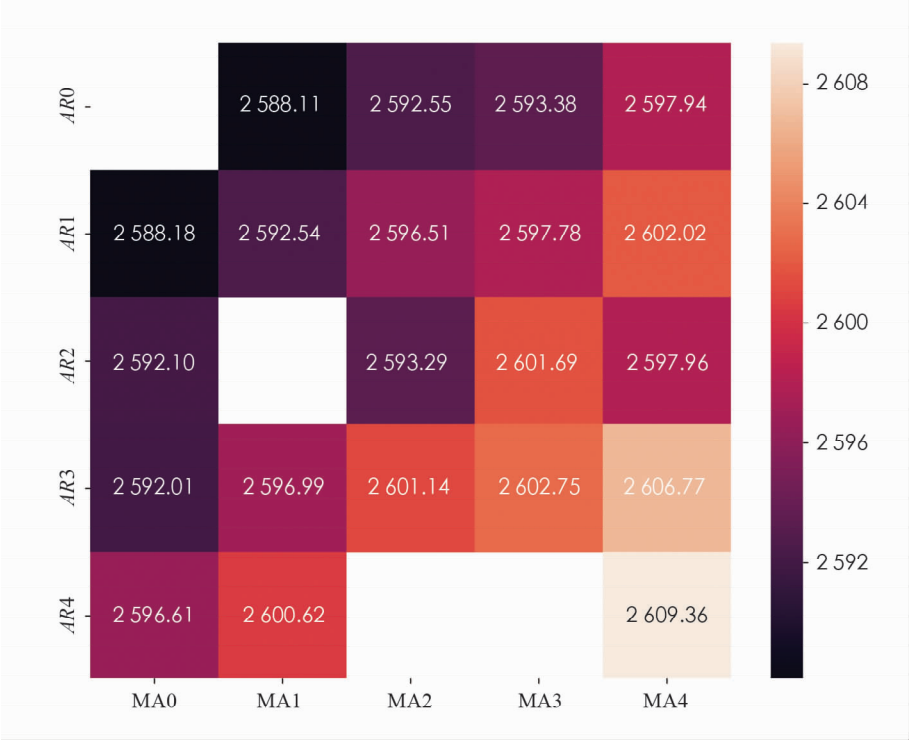

图 5　不同参数预测模型的 BIC 值热力图

4.2.3　基于支持向量机的多因素影响下的建模数据优化

具有相似时间趋势性的自变量共同存在于预测模型中会形成组合冗余，对预测模型造成干扰[5,16]. 由图 1 知，部分地区的百度指数(自变量)的时间趋势变化高度相似，如江苏与浙江、湖南和湖北、西藏与新疆，甚至可以根据这种相似性直观地将中国划分为几个具有同样趋势变化的区域集合. 而基于以上海市为旅游目的地的游客来自中国各地这一事实，原始预测模型将中国各省级行政区的百度指数都作为自变量进行全局预测，势必会使预测模型因为多重共线性变得不稳定和不可靠. 因此，基于数据的空间分布特征及旅游距离与旅游出游率呈反比，从空间结构出发，科学地将中国不同省份来源地的百度指数进行聚类和降维来缩减自变量数目是提高预测精度的有效手段[28].

不同地区关于上海市旅游的市场规模不仅由百度指数反映，影响旅游市场规模的因素还有地区人口规模及出行成本[4,29]. 出行成本由时间成本及经济成本构成，对中国游客而言上海市旅游为国内旅游，其经济花费在大众的承受范围之内，相反由于假期因素的限制，时间成本即旅游者前往旅游目的地的时间花费成为旅游者选择旅游目的地的关键因素[4]. 由于出行的交通工具不同及不同时间段路线拥堵程度不同，时间成本难以定量计算，因此以决定时间成本的游客所在地与上海市的直线距离作为时间成本的替代数据来决定聚类结果[30]. 本研究使用在地理信息领域广泛运用的支持向量机(Support Vector Machine，SVM)方法的扩展算法支持向量聚类(support vector clustering，SVC)对源自不同省份的百度指数进行初步的聚类并结合百度指数排名、人口规模排名(代表潜在旅游市场规模)及与上海市距离排名(代表时间成本)对聚类结果进行调整，经过聚类层数从 2～5 之间的实验，确定最优聚类结果如图 6 所示. 在图 6 中，不同球体间的连线代表在 SVC 算法下的其被聚为同类，为了避免图片过于混杂，只绘制了部分具有指示意义的连线，在连线构成的空间范围内的地区均从属于同类，并使用不同的颜色标示出最终的聚类结果.

在图 6 中，浙江、江苏较其他变量相比有明显的离群聚集现象，且各项排序均属前列，说明其对上海市旅游趋势影响巨大，称上述聚类为核心市场. 上海市除人口规模不足外其他均排名第一；虽然广东与上海市距离排名较后，但其他 2 项排名均在前五，说明广东较其他地区对上海旅游趋势有更多影响，因此将上海、广东归属为核心市场. 第二处明显聚集区域出现在右下角，由吉林、海南、宁夏、新疆、青海、西藏

构成，上述变量在所有的排序中均处于末尾，鉴于此将上述变量构成的区域称为边缘市场；其余离散地分布在图片中间区域的变量，由于其均处在不同排序的中间区域，将该聚类称为一般市场.

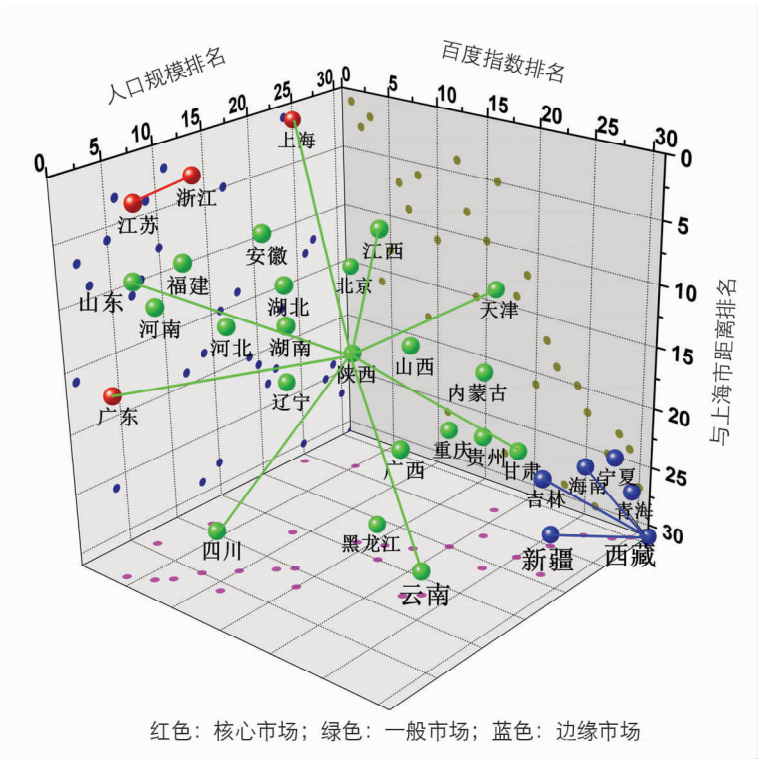

图6 最优聚类结果

对聚类后的3级市场进行主成分分析处理，实现减少预测模型中的自变量数目的，从而减少多重共线性带来的干扰. 计算显示，核心市场聚类的KMO(Kaiser-Meyer-Olkin)的值为0.841，一般市场的KMO值为0.949，边缘市场的KMO值为0.905. 上述KMO值显示聚类后的变量之间存在高度相关性和相似性，证明聚类结果优秀，分别对不同市场聚类结果进行降维处理从而减少共线性干扰，得到的不同市场的百度指数表达式如下：

$$x_{\text{核心市场}} = 0.251 \times x_{\text{广东}} + 0.254 \times x_{\text{上海}} + 0.256 \times x_{\text{浙江}} + 0.257 \times x_{\text{江苏}} \tag{3}$$

$$\begin{aligned} x_{\text{一般市场}} =& 0.044 \times x_{\text{四川}} + 0.045 \times (x_{\text{天津}} + x_{\text{北京}}) + 0.046 \times (x_{\text{安徽}} + x_{\text{湖北}} + x_{\text{辽宁}} + x_{\text{甘肃}}) + 0.047 \times \\ & (x_{\text{江西}} + x_{\text{福建}} + x_{\text{河南}} + x_{\text{湖南}} + x_{\text{河北}} + x_{\text{山西}} + x_{\text{陕西}} + x_{\text{内蒙古}} + x_{\text{广西}} + x_{\text{贵州}} + x_{\text{重庆}} + x_{\text{黑龙江}} + x_{\text{云南}}) \end{aligned} \tag{4}$$

$$x_{\text{边缘市场}} = 0.157 \times x_{\text{西藏}} + 0.17 \times x_{\text{宁夏}} + 0.176 \times x_{\text{吉林}} + 0.178 \times x_{\text{海南}} + 0.18 \times x_{\text{青海}} + 0.181 \times x_{\text{新疆}} \tag{5}$$

4.2.4 旅游趋势预测结果分析

将2011年1月1日至2017年12月31日期间的数据作为建模数据，并使用2018年的数据作为检验数据带入预测模型，使用原始的只区分地理位置属性的百度指数及经过SVC及降维处理优化的数据分别进行预测，得到的结果如图7所示，原始预测模型的平均预测精度为53.42%，而优化后的预测模型的平均预测精度为76.78%，相比提高了23.36%. 同时，可以发现原始预测模型的结果起伏大，波动变化剧烈，而优化后模型的结果与实际上海市国内游客数量变化拟合得更好. 上述证明本研究的模型优化思路确实有效，经过优化后的预测模型更具实用价值.

而对图7中优化后的预测结果在长期内都大于实际上海市国内游客数量，综合图2所绘制的数据趋势变化图，究其原因是：伴随着中国国民互联网接入率不断上升，越来越多的旅游者通过互联网提前搜索旅游目的地的相关信息，致使本研究涉及的百度指数的增长速度远大于上海市国内游客的增长速度，从而在后期出现预测结果偏大的情况.

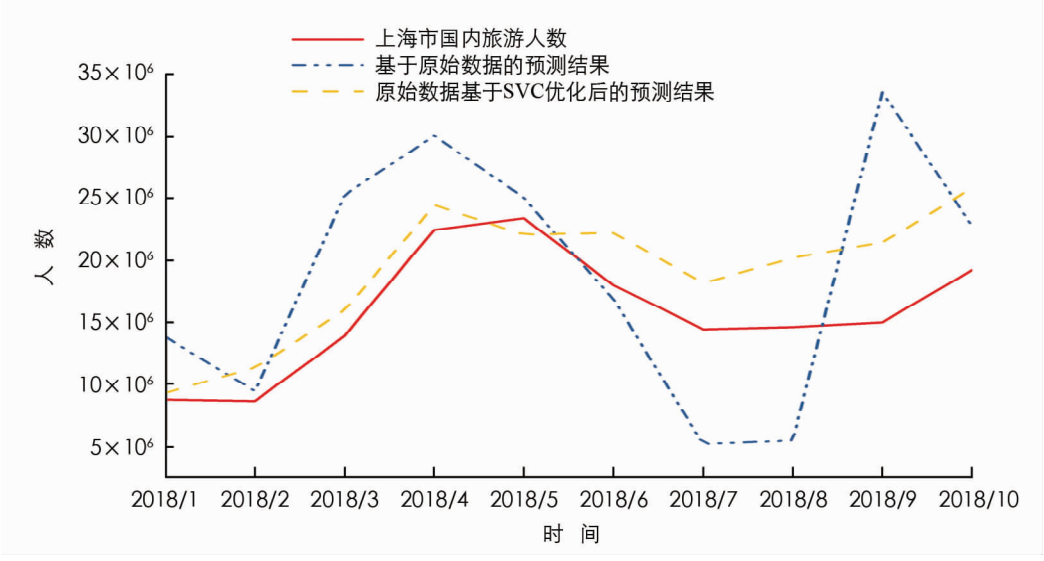

图 7　2018 年上海市实际旅游人数与预测人数对比图

5　结　　论

本研究选取基于 2011 年至 2018 年 10 月的上海市月度国内旅游人数及对应的百度指数数据，构建具有时空分布特征的百度指数与 ARIMA 模型结合的旅游预测模型，并在深入挖掘百度指数数据时空分布规律基础上，利用基于支持向量机的空间聚类方法解决了 ARIMA 预测模型的多重共线性问题.

1）昨天的搜索者就是今天的旅游者. 由格兰杰因果关系检验确定以上海市为旅游目的地的中国游客数量与上海市旅游相关百度指数存在长期正相关关系，随着上海市旅游相关百度指数数值的增大，上海市实际国内游客数量也会相应增加，确定旅游相关的互联网搜索行为与旅游行为存在密切相关性. 因此，旅游相关部门和产业可通过加大互联网广告的投入，努力将搜索人口转化为实际旅游人口.

2）旅游距离与旅游出行率呈反比. 由图 1 及图 6 知，总体而言互联网搜索关注度与旅游目的地的相关性会因空间距离增加而逐渐减小，但少数省份如广东的百度指数没有因为空间距离增加而减少. 上述特例产生的原因是该省份具有高居民收入和更多的人口，使得居民出游意愿强烈. 但从整体上分析百度指数的空间分布特征，仍是旅游距离与旅游出行率呈反比的强力佐证. 若上海市旅游市场增长放缓，旅游相关部门和企业则应该有针对性对不同旅游市场采取不同营销策略.

3）百度指数的空间属性有助于提升预测精度，依据百度指数的空间属性分析其数据分布规律，并利用空间分布特征进行空间聚类优化后的 ARIMA 模型平均预测精度提升了 23.36%. 百度指数易于获取且具有实时性，其蕴含的网络用户潜在消费欲望，使旅游趋势预测更具经济价值.

4）通过相关性分析科学地获得搜索关键词能够提高预测精度. 本研究发现前人的研究中使用随机的、过多的关键字看似全面，似乎可以获得旅游目的地相关搜索数据的全部，但数据之间的相互干扰会降低预测精度；依据搜索关键词与研究目的相关性分析结果，选择一个或者几个关键词更具可行性与科学性.

百度指数公开且易得，但实际旅游人数数据的获取制约着旅游研究的发展. 国内大部分景区都未对旅游人数数据进行公开，保守的数据策略亟待转变. 未来研究中，将注重每日的旅游人数数据获取，以更高的时间分辨率的数据开展搜索数据和实际旅游规模的研究，以此确定准确的搜索时间和实际旅游时间的滞后值，从而提高旅游预测模型的可用性. 在互联网中，不仅存在着以百度指数为代表的以表格形式存储的结构化数据，还存在着音频、视频、图像等非结构化数据，如微信的朋友圈、马蜂窝等旅游 APP 的旅游攻略等，对上述非结构化数据的研究将为旅游研究开辟新兴领域.

参考文献:

[1] 李 山,邱荣旭,陈 玲. 基于百度指数的旅游景区网络空间关注度:时间分布及其前兆效应[J]. 地理与地理信息科学,2008,24(6):108-113.

[2] 卢文刚. 景区容量超载背景下的旅游突发事件应急管理研究——以"10·2"九寨沟游客滞留事件为例[J]. 西南民族大学学报(人文社科版),2015,39(11):145-150.

[3] 韩 冰,路 紫,赵亚红,等. 旅游网站访问者行为的时间分布及导引分析[J]. 地理学报,2007,62(6):621-630.

[4] 林 青,晁 怡,杨 乃,等. 一种考虑时间成本的旅游线路推荐方法[J]. 地理与地理信息科学,2017,33(6):29-33,60.

[5] YANG X, PAN B, EVANS J A. Forecasting Chinese Tourist Volume with Search Engine Data[J]. Tourism Management, 2015, 46: 386-397.

[6] 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):158-166.

[7] HASSANI H, WEBSTER A, SILVA E S. Forecasting U. S. Tourist Arrivals Using Optimal Singular Spectrum Analysis[J]. Tourism Management, 2015, 46: 322-335.

[8] BANGWAYO-SKEETE P F, SKEETE R W. Can Google Data Improve the Forecasting Performance of Tourist Arrivals? Mixed-data Sampling Approach[J]. Tourism Management, 2015, 46: 454-464.

[9] FONDEUR Y, KARAMÉ F. Can Google Data Help Predict French Youth Unemployment?[J]. Economic Modelling, 2013, 30(1): 117-125.

[10] ARGIRIOU A A. Use of Neural Networks for Tropospheric Ozone Time Series Approximation and Forecasting &Ndash; a Review[J]. Atmospheric Chemistry and Physics Discussions, 2007, 7(2): 5739-5767.

[11] ZAFRA C, ÁNGEL Y, TORRES E. ARIMA Analysis of the Effect of Land Surface Coverage on PM10 Concentrations in a High-altitude Megacity[J]. Atmospheric Pollution Research, 2017, 8(4): 660-668.

[12] ALTHOUSE B M, NG Y Y, CUMMINGS D A T. Prediction of Dengue Incidence using Search Query Surveillance[J]. PLos Neglected Tropical Diseases, 2011, 5(8).

[13] LATINOPOULOS D. Using a Spatial Hedonic Analysis to Evaluate the Effect of Sea View on Hotel Prices[J]. Tourism Management, 2018, 65: 87-99.

[14] ASKITAS N, ZIMMERMANN K F. Google Econometrics and Unemployment Forecasting[J]. Applied Economics Quarterly, 2009, 55(2): 107-120.

[15] 张悟移,李 杰. 百度关注度指数与股票价格关系研究[J]. 西南师范大学学报(自然科学版),2019,44(02):75-83.

[16] HUANG K X, ZHANG L F, DING Y S. The Baidu Index: Uses in Predicting Tourism Flows-A Case Study of the Forbidden City[J]. Tourism Management, 2017, 58: 301-306.

[17] CLAVERIA O, TORRA S. Forecasting Tourism demand to Catalonia: Neural Networks vs. Time Series Models[J]. Economic Modelling, 2014, 36: 220-228.

[18] 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):93-99.

[19] 何小芊,刘 宇,吴发明. 基于百度指数的温泉旅游网络关注度时空特征研究[J]. 地域研究与开发,2017,36(1):105-110,126.

[20] PAN B. The Power of Search Engine Ranking for Tourist Destinations[J]. Tourism Management, 2015, 47: 79e87-87.

[21] CHAITIP P, CHAIBOONSRI C. International Tourists Arrival to Thailand: Forecasting by Non-linear Model[J]. Procedia Economics and Finance, 2014, 14: 100-109.

[22] CHU F L. Forecasting Tourism Demand with ARMA-based Methods[J]. Tourism Management, 2009, 30(5): 740-751.

[23] 张向宁,孙秋碧. 信息化与工业化融合有界性的实证研究——基于我国31省市面板数据[J]. 经济问题,2015(01):84-88.

[24] FINLEY T, JOACHIMS T. Supervised Clustering with Support Vector Machines [C]//ICML 2005-Proceedings of the 22nd International Conference on Machine Learning. 2005.

[25] WANG H, WANG W, MENG Y. Degree of User Attention to a Webpage Based on Baidu Index: An alternative to page view [J]. Journal of Experimental and Theoretical Artificial Intelligence, 2014, 26(2): 235-249.

[26] ETUK E H. A Seasonal Arima Model for Nigerian Gross Domestic Product [J]. 2012, 2(3): 46-53.

[27] BRIDA J G, GARRIDO N. Tourism Forecasting Using SARIMA Models in Chilean Regions [J]. International Journal of Leisure and Tourism Marketing, 2011, 2(2): 176.

[28] 孙晓蓓，杨晓霞，张枫怡. 基于百度指数的中国 A 级旅游洞穴景区网络关注度分布特征研究 [J]. 西南师范大学学报(自然科学版)，2018，43(4)：81-88.

[29] 殷　平. 旅游交通成本对旅游目的地空间竞争的影响研究 [J]. 地域研究与开发，2012，31(6)：97-101.

[30] 张　捷，李升峰，周寅康，等. 九寨沟风景区游客入游距离特征研究 [J]. 长江流域资源与环境，2002(1)：5-9.

Tourism Trend Prediction Based on Baidu Index Spatial and Temporal Distribution

KANG Jun-Feng[1], GUO Xing-Yu[1], FANG Lei[2]

1. *School of Architecture and Surveying Engineering, Jiangxi University of Science and Technology, Ganzhou Jiangxi 341000, China;*

2. *Department of Environmental Science and Engineering, Fudan University, Shanghai 200433, China*

Abstract: In this paper, the Baidu index and the number of Chinese domestic tourists (1.38 billion in total) of each consecutive monthly travel destination of Shanghai from 2011 to 2018 have been studied. Through the Granger causality test, ARIMA model, spatial clustering method and principal component analysis, the mapping relationship between Internet virtual space and the real world has been explored. With the help of Spatio-temporal distribution pattern analysis and seasonal trend analysis, the multicollinearity problem of a similar time trend of different sources has been solved, thus the average prediction accuracy of the optimized prediction model been increased by 23.35%. Moreover, it is concluded that "yesterday's searchers are today's tourists", "travel distance is inversely proportional to travel rate" and "the geographical location attribute of the search index is helpful to improve the prediction accuracy". Tourism forecast can provide scientific and accurate decision-making basis for the scenic spot management department to ensure the safety of the scenic spot and tourism experience.

Key words: Baidu index; tourism forecast; space-time distribution; ARIMA model; Support vector clustering; geographic information system

责任编辑　胡　杨