Hans Journal of Data Mining 数据挖掘, 2022, 12(2), 133-151
Published Online April 2022 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2022.122015

Hans汉斯

基于网络搜索数据的北京市旅游需求预测

李静静[1]，李志新[1*]，陈继强[1]，李志国[2]

[1]河北工程大学数理科学与工程学院，河北 邯郸
[2]河北工业大学理学院，天津

收稿日期：2022年3月5日；录用日期：2022年4月5日；发布日期：2022年4月14日

摘 要

我国旅游业经过40年的高速度发展，现在进入了高质量发展新阶段。同时，随着疫情防控进入常态化和旅游市场逐步回暖，“互联网 + 旅游”新业态发展迅猛，海量网络搜索数据潜在反映着人们的旅游需求。因此，本文利用网络搜索数据(Internet search data, IS)用于北京市旅游需求预测。首先，利用Python爬取在线旅游网站的游记攻略，使用NLPIR分词系统提取高频词汇，并结合旅游六要素确定初始关键词词库。其次，采用需求图谱、百度指数相关词热度推荐、北京旅游网推荐等7种方法拓展关键词，经过Adaptive Lasso等方法筛选得到9个最佳预测变量，并引入季节性虚拟变量，然后结合网络搜索关键词和随机森林算法、极限梯度提升算法及支持向量回归算法对北京市旅游需求进行建模和训练。最后，借助多个预测性能指标，确定支持向量回归模型为最优模型。研究结果表明：网络搜索数据与旅游需求显著相关，具有很强的时效性，并且支持向量回归模型能够很好地解决突发事件和小样本问题，用于短期旅游需求预测是高效可行的。

关键词

互联网 + 旅游，网络搜索数据，Adaptive Lasso，支持向量回归，旅游需求预测

Tourism Demand Forecasting Based on Internet Search Data in Beijing

Jingjing Li[1], Zhixin Li[1*], Jiqiang Chen[1], Zhiguo Li[2]

[1]School of Mathematics and Physics Science and Engineering, Hebei University of Engineering, Handan Hebei
[2]School of Science, Hebei University of Technology, Tianjin

Received: Mar. 5th, 2022; accepted: Apr. 5th, 2022; published: Apr. 14th, 2022

[*]通讯作者。

文章引用: 李静静, 李志新, 陈继强, 李志国. 基于网络搜索数据的北京市旅游需求预测[J]. 数据挖掘, 2022, 12(2): 133-151. DOI: 10.12677/hjdm.2022.122015

Abstract

After 40 years of rapid development, China's tourism industry has entered a new stage of high-quality development. Meanwhile, with the gradual normalization of epidemic prevention and control and the gradual warming of the tourism market, the new format of "Internet + tourism" is developing rapidly, and massive Internet search data potentially reflects the tourism demand of people. Therefore, this paper attempts to apply Internet search data to the tourism demand forecast of Beijing. Firstly, Python is used to crawl the travel notes of online travel websites, NLPIR word segmentation system is used to extract high-frequency words, and six elements of tourism are combined to determine the initial keyword thesaurus. Secondly, seven methods, such as demand map, related word heat recommendation from Baidu index and recommendation from Beijing travel website, etc., are used to expand keywords. Nine predictive variables are selected by adaptive lasso and other methods, the seasonal dummy variables are introduced, then RF algorithm, XGBoost algorithm and SVR algorithm are combined to model and train the tourism demand of Beijing. Finally, the support vector regression model is determined as the optimal model with the help of multiple prediction performance indicators. The results show that there is a significant correlation between Internet search data and tourism demand, and Internet search data has strong timeliness. In addition, SVR model can well solve the emergency and small sample problems, and it is efficient and feasible to predict short-term tourism demand.

Keywords

Internet + Tourism, Internet Search Data, Adaptive Lasso, Support Vector Regression, Tourism Demand Forecasting

1. 引言

随着我国社会经济的快速发展和人们生活需求的日益丰富，强劲的旅游需求驱动旅游业的战略地位不断提升，旅游业发展势头迅猛。旅游在为人们提供幸福、美好生活方式的同时，也给各地带来了丰厚的收入，成为拉动经济发展的动力。同时，互联网和大数据持续快速发展，移动终端的广泛应用充实着人们的生活。中国互联网络信息中心(CNNIC)第 47 次中国互联网报告显示[1]，截止 2020 年 12 月，我国互联网的普及率达到 70.4%，网民规模接近 10 亿。使用互联网已经成为一种习惯，其中搜索引擎更是人们获取网络信息的重要工具和途径，深刻影响着人们生活的方方面面。“以搜索开始一天，以搜索结束一天”可以非常贴切地形容人们现在的网络生活，而搜索背后的数据则反映着人们潜在的旅游需求。2020 年伊始，新冠肺炎疫情爆发，全球旅游业处于停滞状态。在这样严峻复杂的环境下，我国成为全球唯一实现经济正增长的主要经济体[2]，其中互联网发挥了重要积极作用。借助大数据技术，旅游业充分展现了在国家战略中的责任担当，我国旅游业进入了智慧旅游和全域旅游时代。

北京市作为我们国家的首都，每年吸引大批国内外游客到京旅游观光，其旅游业发展与中国改革开放同步，始终处于首都核心功能的地位。近几年，北京市旅游人数和旅游收入稳步上升，2000 年北京市

国内旅游人数 1.02 亿人次，国内旅游收入 683 亿元，到 2019 年国内旅游人数和国内旅游收入分别达到了 3.18 亿人次和 5866 亿元，这些数据充分见证了北京市旅游业的高速增长，旅游已经成为新时期人们对美好生活和精神文化需求的重要组成部分。2020 年初，新冠肺炎疫情突袭，从 2020 年 1 月 23 日武汉封城到 7 月 20 日北京正式恢复跨省游，全年国内旅游人数和旅游收入直降为 1.84 亿人次和 2881 亿元，数据跌落回十年前水平，对当地旅游业造成很大的打击。2021 年开始北京接待旅游总人数 2.6 亿人次，实现旅游总收入 4166.2 亿元，旅游经济得到了复苏。

疫情防控常态化对于旅游业来说，是挑战，更是机遇。2020 年北京市人民政府下发《北京市推进全国文化中心建设中场及规划(2019 年~2035 年)》的通知[3]，对首都北京未来 15 年进行规划，目标将北京市建设成为彰显中华文化魅力的世界旅游名城，努力实现首都旅游业高质量发展。同时，“互联网 + 旅游”新业态发展迅猛，两者的融合发展已经成为不可逆转的趋势。基于此背景，本文以北京市旅游人数代表旅游需求，挖掘与北京市旅游相关的网络搜索数据，预测北京市旅游需求，为推动旅游业尽早恢复到疫前水平并保持高质量发展提供参考。

2. 基本概念及数据来源

2.1. 网络搜索数据的概念及数据来源

互联网用户利用搜索引擎对互联网上的信息进行搜索时，其中的网络搜索关键词和浏览痕迹都被搜索引擎记录下来，由此生成网络搜索数据。网络搜索的广泛使用以及它与人们日常生活的紧密交织，利用网络搜索数据研究某地的旅游需求成为现实。目前，在中国搜索引擎行业中，百度是国内使用率最高的搜索引擎，《2019 年中国网民搜索引擎使用情况研究报告》显示，百度搜索品牌的渗透率达到 90.9%，受到国内用户的青睐[4]。根据 Statcounter 网站流量数据统计，近几年百度搜索所占市场份额始终位列第一，占比 74.37%，用户规模最大，远远超过搜狗(13.7%)和谷歌(3.66%)。2006 年，百度推出了以百度海量网民行为数据为基础的数据分享平台——百度指数，用来收集用户通过百度检索留下的检索痕迹，包括“人群画像”、“需求图谱”和“趋势研究”三个模块。

本文以百度平台提供的百度指数为数据来源，分析网络搜索数据与旅游需求之间的相关关系。百度指数供用户查看的最早统计数据为 2006 年 6 月 1 日，其中 2006 年 6 月至 2010 年 12 月期间，百度指数仅提供 PC 端趋势数据；2011 年 1 月开始提供整体趋势数据，包括 PC 端数据和移动端数据。考虑到现在人们在移动设备上搜索信息更为频繁。同时，为了避免仅使用 PC 端数据造成的信息损失，选取 2011 年 1 月 1 日至 2020 年 12 月 31 日与北京市旅游人数相关的网络搜索整体趋势数据，以月为单位收集得到关键词的月平均搜索数据，每个网络搜索关键词的数据有 120 个。

2.2. 旅游需求的概念及数据来源

需求是指人们有能力购买并且有意愿购买某种商品的欲望，当这种商品为旅游产品和服务时，就构成了旅游需求[5]，即在一定时间内，旅游者具有旅游的动机并且能够以一定货币支付能力和可支配时间购买旅游产品或者旅游服务的数量。反映旅游需求的指标有很多，以往研究中通常使用旅游人数、旅游收入、时空特征等方面来反映旅游需求指标，其中旅游人数是最常用的指标。

本文以北京市 2011 年 1 月至 2020 年 12 月的月度旅游人数作为另一种数据，具体数据来自北京市统计局官网，其中旅游区接待人数统计范围为 A 级及以上旅游区和其他主要旅游区，包括境外人数。百度指数基本反映的是国内搜索关键词的情况。因此，将旅游区总接待人数减去境外旅游人数后的境内旅游人数作为最终的北京市旅游人数，以此作为北京市旅游需求。

3. 网络搜索指标体系构建与筛选

3.1. 初选网络搜索关键词

目前关键词选取方法主要有三种：技术取词法、直接取词法、范围选词法。本文考虑到这三种方法在选取初始关键词上各有利弊，结合大数据环境下人们的网络搜索行为，决定从文本挖掘的角度入手，完成网络搜索关键词的初选。

在线旅游网站为旅游目的地拓宽了客源渠道，大量信息分享方面远远高于传统旅行社。它不仅可为游客提供产品预订一站式服务，还有提供用户点评、社区问答、旅游攻略等信息可以参考。然而互联网在线旅游网站以方便、快捷的绝对优势，受到越来越多人的青睐。人们出游前获取信息的主要渠道来自旅游网站的游记攻略栏目，并且在旅游前、中、后选择将自己的旅行经验和感受分享到旅游网站平台，实现了旅游者和旅游网站之间的信息交流和共享。根据 Alexa 网站统计，对各大旅游网站的注册时间、排名情况、网站搜索指数和与北京相关的游记数量进行深入研究，结果见表 1。

Table 1. Relevant information on the travel website
表 1. 旅游网站相关信息

网站名称	域名	注册时间	Alexa 排名	搜索指数	游记数量
携程旅行网	http://www.ctrip.com	2000/7/18	2755	77915	22287
穷游网	http://www.qyer.com	2007/11/6	3210	2558	1000
马蜂窝	http://www.mafengwo.cn	2007/11/29	4862	7408	3000
去哪儿网	http://www.qunar.com	2006/3/17	8141	59733	2000
驴妈妈网站	http://www.lvmama.com	2007/8/13	18662	3981	567
飞猪	http://www.fliggy.com	2004/5/23	37888	4235	-
艺龙旅行网	http://www.elong.com	2000/11/28	12464	1332	-
猫途鹰	http://www.tripadvisor.cn	2008/12/26	15524	767	-
途牛旅游网	http://www.tuniu.com	2006/12/18	6162	9171	155
同程旅游	http://www.ly.com	2004/3/10	234947	2773	15058

注：“-”表示该网站无相应游记栏目供人们搜索，无法获取游记数量。

表 1 中搜索指数是指网站名称在百度指数中的整体搜索量，指数越高说明该旅游网站的受关注程度越高。旅游网站排名靠前的有携程旅行网、穷游网、马蜂窝、去哪儿网；搜索指数靠前的有携程网、去哪儿网、途牛旅游网；北京游记数量较多的网站有携程旅行网、同程旅游、马蜂窝、去哪儿网，最终确定携程旅行网、马蜂窝、去哪儿网和同程旅游作为本文的目标旅游网站。

利用 Python 分别对携程旅行网、马蜂窝、去哪儿网和同程旅游网的游记攻略搜索“北京”关键词，希望通过文本大数据来了解用户对北京的旅游需求，按照网页推送对每个网站爬取 100 篇游记攻略，共获得 400 篇，总计 1541758 字。通过文本挖掘获得的 400 篇游记攻略是无结构的文本信息，需要将所有文字分割成有意义的词语。本文使用 NLPIR 汉语分词系统进行文本分词获取高频词汇，得到关键词列表及其对应的词性、权重和词频，如图 1 所示。可以发现，出现次数多的词语不一定是关键词，比如“可以”、“一个”、“开始”、“还有”等词语，虽然所占权重和词频比较靠前，但不具备完整意义，与本文研究对象“北京”没有绝对关联。因此，这些词语不被纳入初始关键词，筛选出高频词汇。

Word	Part-Of-Speech	Weight	Frequency
北京	ns	636.56	5908
可以	v	480.76	2948
一个	mq	355.62	2448
酒店	n	318.16	1835
时间	n	266.69	1816
故宫	ns	253.57	2058
建筑	n	247.93	1057
开始	v	228.81	621
景区	n	225.03	859
天安门广场	n_new	210.92	451
很多	m	210.85	1041
还有	v	202.22	870
恭王府	n_new	200.75	409
进入	v	197.91	589
看到	v	190.34	1024
胡同	n	179.8	950
选择	vn	174.82	648
烟袋斜街	n_new	174.76	175
司马台长城	n_new	170.85	183
需要	v	168.35	571
感觉	n	167.75	934
文化	n	165.57	640

Figure 1. Part of the high-frequency words extracted by the NLPIR system
图 1. NLPIR 系统提取的部分高频词汇

将高频词汇结合旅游六要素“吃、住、行、游、购、娱”，确定北京美食、北京特产、北京住宿、北京地图、北京旅游、北京购物、北京娱乐、北京这 8 个关键词，构建初始关键词词库，见表 2。将关键词在百度指数搜索栏中逐个进行搜索，发现每个关键词对应的百度指数都有记录，说明该初始关键词词库构建是合理的。

Table 2. The initial keyword thesaurus
表 2. 初始关键词词库

类别	关键词
吃	北京美食，北京特产
住	北京住宿
行	北京地图
游	北京旅游
购	北京购物
娱	北京娱乐
其他	北京

为了反映旅游人数变化，下面采用灰色关联分析方法检验 8 个关键词是否与北京市旅游人数相关。首先，对初始关键词词库中的关键词和北京市旅游人数进行无量纲处理，根据灰色关联系数计算公式：

$$\xi_i(t)=\frac{\min\limits_i\min\limits_t\left|y(t)-x_i(t)\right|+\rho\max\limits_i\max\limits_t\left|y(t)-x_i(t)\right|}{\left|y(t)-x_i(t)\right|+\rho\max\limits_i\max\limits_t\left|y(t)-x_i(t)\right|} \tag{1}$$

其中，$y(t)$ 表示因变量旅游人数，$x_i(t)$ 表示第 i 个自变量即关键词，ρ 为分辨系数，在(0, 1)内取值，不妨取 0.5。通过计算得出：

$$\min_i\min_t\left|y(t)-x_i(t)\right|=0.0007375 \tag{2}$$

$$\max_i \max_t \left| y(t) - x_i(t) \right| = 2.7417169 \tag{3}$$

得到各网络搜索关键词与旅游人数在不同时间的灰色关联系数。然后根据灰色关联度公式：

$$\gamma = \frac{1}{n}\sum_{t=1}^{n} \xi_i(t) \tag{4}$$

计算关联系数均值，其中，n 表示自变量时间序列个数，从而得到初始关键词与旅游人数之间的灰色关联度，见表 3。可以发现，北京美食、北京特产、北京住宿、北京地图、北京旅游、北京景点、北京购物、北京娱乐和北京的灰色关联度均大于 0.75，说明关键词与旅游人数之间具有较强的关联性，8 个初始关键词构成的关键词词库能够反映旅游人数的变化，可用其进行关键词拓展。

Table 3. Grey related degree of the initial keywords
表 3. 初始关键词的灰色关联度

关键词	北京美食	北京特产	北京住宿	北京地图	北京旅游	北京景点	北京购物	北京娱乐	北京
灰色关联度	0.8842	0.7642	0.7763	0.7620	0.7603	0.7684	0.7643	0.7667	0.7650

3.2. 拓展网络搜索关键词

基于文本挖掘得到的初始关键词中所包含的信息是有限的，并不能全面反映旅游目的地“北京”的相关情况，还需要进一步对网络搜索关键词进行拓展。本文采用以下几种方法对初始关键词进行拓展。

1) 需求图谱。根据周边圆圈距离搜索关键词的远近和自身圆圈大小对初始关键词进行选择性拓展，用户需求越多，相关关键词的搜索指数越高，圆圈的直径就越大；圆圈距离搜索关键词所在圆圈的位置越近，相关程度越大，反之越小。图 2 为初始关键词“北京美食”的需求图谱，可以拓展出“北京美食街”、“北京烤鸭”、“北京小吃”、“北京特色美食”等关键词。

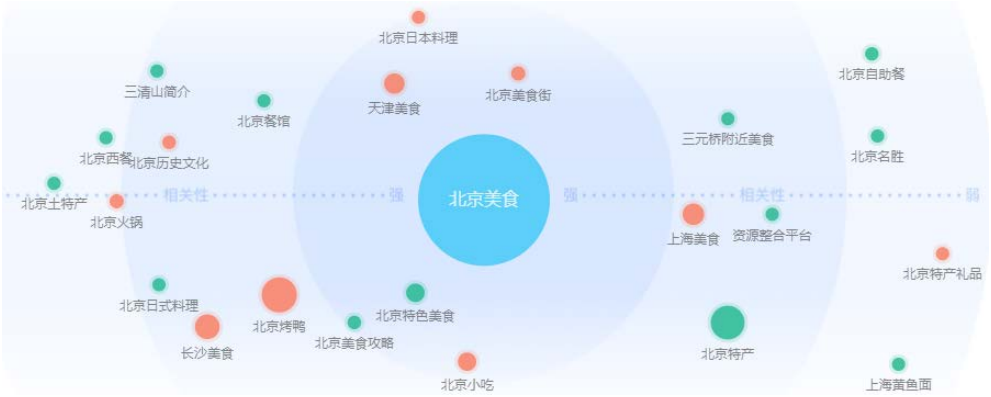


Figure 2. Interface of the demand map
图 2. 需求图谱界面

2) 百度指数相关词热度推荐。百度指数将所有与搜索关键词相关的需求进行排序区分显示，排名越靠前，表示相关关键词与搜索关键词的相关程度越大。比如由“北京美食”可拓展出“北京烤鸭”、“北京特产”、“北京小吃”等关键词。

3) 百度搜索引擎推荐。在百度搜索引擎输入关键词，搜索引擎会自动提示扩展关键词。比如对“北京旅游”进行搜索可拓展出“北京旅游排行榜前十”、“北京旅游攻略”、“北京旅游景点一览表”等

关键词。

4) 百度网页相关推荐。用户在百度搜索引擎输入关键词跳转到相关网页后，在搜索页面的底部会出现一栏“相关搜索”，可为用户提供一些相关关键词进行二次搜索。搜索引擎会记录下用户的搜索记录从而获得用户的搜索习惯，为后续搜索提供参考。比如“北京旅游”可拓展出“北京一日游”、“北京周边景点”等关键词。

5) 百度指数搜索推荐。在百度指数搜索栏输入关键词，会自动弹出相关关键词。以“北京旅游”为例，自动弹出收录在百度指数中的关键词“北京旅游景点”和“北京旅游包车”两个关键词。

6) 北京旅游网推荐。北京旅游网是北京市文化和旅游局监管的非营利性网站，会向用户提供国内最权威的北京旅游信息。根据搜索内容不同类别，可获得“天安门广场”、“故宫”、“八达岭长城”、“鸟巢”等关键词。

7) 网络游记推荐。根据去哪儿网、同程旅游、马蜂窝和携程旅行网收集到的游记攻略，挖掘有实际意义的词汇为扩展关键词，从中可以找到“烟袋斜街”、“司马台长城”、“北京朝阳公园”、“十三陵”等关键词。

经过以上 7 种方法，共收集到 258 个关键词。由于百度指数并不会记录所有词语的搜索量，因此需要将收集到的关键词依次输入百度指数检索栏进行搜索，选择剔除无记录搜索量的关键词，剩余 161 个关键词。使用 Python 程序爬取百度指数中 2011 年 1 月至 2020 年 12 月关键词数据，并使用月度汇总，发现有 11 个关键词在多个月份中搜索量为零，并且有 24 个关键词月搜索量整体明显偏低，所以将其剔除，最后剩下 126 个关键词构成拓展关键词词库，见表 4。

Table 4. The extended keyword thesaurus
表 4. 拓展关键词词库

类别	数量	网络搜索关键词
食	23	北京美食，北京特色美食，北京小吃，北京烤鸭，北京美食攻略，北京小吃街，茯苓饼，北京特产，北京稻香村，京八件，老北京小吃，六必居酱菜，驴打滚，稻香村，北京果脯，艾窝窝，北京饭店，全聚德烤鸭，炸酱面，卤煮火烧，二锅头，姚记炒肝，北京餐厅
住	10	北京住宿，北京青年旅社，北京酒店，北京希尔顿酒店，北京万豪酒店，北京王府半岛酒店，北京住宿攻略，北京四合院，北京宾馆，北京宾馆价格
行	22	北京地图，北京地铁线路图，北京地铁，北京交通，北京首都国际机场，北京地图地铁，北京公交，北京机场，北京汽车站，首都国际机场，南苑机场，北京机票，北京火车站，北京西站，北京南站，北京北站，北京地铁时刻表，北京旅游地图，北京景点地图，北京旅游，北京旅游攻略，北京自由行
游	44	北京旅游景点，北京旅游网，北京胡同，北京旅行，北京游玩，北京自驾游，烟袋斜街，司马台长城，八达岭长城，鸟巢，奥林匹克公园，毛主席纪念堂，北京一日游，北京三日游，北京夜景，北京动物园，北京自由行攻略，北京景点，北京北海公园，颐和园，恭王府，世界公园，十三陵，慕田峪长城，北京欢乐谷，北京南锣鼓巷，北京游乐园，后海，地坛公园，天坛公园，北京朝阳公园，景山公园，十渡，玉渊潭公园，石景山游乐园，北京景点大全，北京景点门票，天安门广场，故宫，北海公园，什刹海，前门大街，天坛，圆明园
购	11	北京购物，北京西单，北京购物中心，三里屯，王府井，国贸，西单大悦城，王府井百货，北京商圈，北京夜市，北京免税店
娱	8	北京娱乐，三里屯酒吧，北京酒吧，北京夜生活，798 艺术区，环球影城，北京展览，三里屯酒吧街
其它	8	北京，北京介绍，北京简介，北京历史，北京天气，北京温度，北京天气预报，北京文化

3.3. 筛选网络搜索关键词

为了更好地选择最优关键词，需要对这 126 个网络搜索关键词进行筛选。首先，计算斯皮尔曼等级相关系数，保留 45 个相关系数大于 0.6 的关键词。其次，从时间上来看，网络搜索关键词与旅游人数之间存在三种时滞关系：先行、同步、滞后，而绝大部分用户会在旅游当月或提前数月对旅游信息进行搜索。本文采用 K-L 信息量法确定网络搜索关键词变量与北京市旅游人数变量之间的时滞关系，计算北京市旅游人数与每个关键词之间的 0~6 期提前期和 0~6 期滞后期的相关系数，即对每 1 个关键词需要计算 13 次 K-L 信息量值，选择最小的 K-L 信息量值对应的时滞阶数作为最佳时滞阶数，剔除掉较旅游人数序列滞后变化的指标，保留先行和同步变化的指标，见表 5。

Table 5. Optimal delay order and K-L information value of the keywords
表 5. 关键词的最佳时滞阶数及相应的 K-L 信息量值

关键词	时滞阶数	K-L 信息量	关键词	时滞阶数	K-L 信息量
北京美食	0	0.0313	北京北海公园	0	0.0300
北京特色美食	0	0.0575	颐和园	0	0.0557
北京小吃街	0	0.0257	恭王府	0	0.0440
北京住宿攻略	−1	0.0565	世界公园	0	0.0399
北京旅游地图	0	0.0641	十三陵	0	0.0358
北京景点地图	0	0.0270	慕田峪长城	0	0.0430
北京旅游攻略	0	0.0698	北京欢乐谷	0	0.0545
北京自由行	−1	0.0334	北京南锣鼓巷	0	0.0468
北京旅游景点	0	0.0487	北京游乐园	0	0.0497
北京胡同	0	0.0352	后海	0	0.0518
北京游玩	0	0.0281	天坛公园	0	0.0399
北京自驾游	−1	0.0391	北京朝阳公园	0	0.0364
烟袋斜街	0	0.0457	景山公园	0	0.0420
司马台长城	0	0.0266	十渡	0	0.1272
八达岭长城	0	0.0668	石景山游乐园	0	0.0481
鸟巢	0	0.0306	北京景点大全	0	0.1014
奥林匹克公园	0	0.0177	北京景点门票	−1	0.0371
毛主席纪念堂	0	0.0437	北海公园	0	0.0222
北京一日游	0	0.0660	前门大街	0	0.0348
北京三日游	0	0.1071	北京夜市	0	0.0239
北京夜景	0	0.0336	798 艺术区	0	0.0581
北京动物园	0	0.1071	北京展览	0	0.0340
北京景点	0	0.0205			

由表 5 可知，大部分关键词与北京市旅游人数是同步变化关系，少数关键词先行于北京市旅游人数变化，北京自驾游、北京景点门票、北京住宿攻略、北京自由行这 4 个关键词比旅游人数提前 1 期。各个关键词反映了旅游需求的动态变化，关键词变量的不同时滞阶数反映了用户在不同阶段的潜在旅游需求。由于旅游需求数据发布往往具有滞后性，而网络搜索数据具有时效性特征，所以计算得到的同步指标不予剔除，仍可作为后续关键词筛选。同时，经过 K-L 信息量均为先行或同步指标，表明计算斯皮尔曼相关系数进行筛选已经剔除掉了相对滞后的关键词。

网络搜索数据也经常存在异常点和振幅差异，对应的网络搜索关键词和旅游人数序列就会出现整体形状相似但在时间轴上发生错位和偏移现象，而 DTW 动态时间弯曲可以很好地避免这一点，筛选的关键词具有更好的预测效果[6]。因此，本文通过 DTW 进一步计算每个网络搜索关键词序列与旅游人数序列之间的最短弯曲距离。为防止关键词漏选，保留与旅游人数最为相似的前 20 个关键词，见表 6。

Table 6. Internet search keyword screened by three methods
表 6. 三种方法筛选后的网络搜索关键词

关键词	斯皮尔曼相关系数	时滞阶数	K-L 信息量	DTW 动态时间弯曲
北京游玩	0.6909	0	0.0281	9.8994
北京景点	0.8198	0	0.0205	10.0793
北京夜市	0.7482	0	0.0239	10.1692
北海公园	0.8313	0	0.0222	10.9350
北京小吃街	0.7656	0	0.0257	11.1606
奥林匹克公园	0.8634	0	0.0177	11.4660
北京美食	0.7347	0	0.0313	11.5092
天坛公园	0.7160	0	0.0399	11.6671
慕田峪长城	0.7605	0	0.0430	11.7371
北京景点地图	0.7877	0	0.0270	11.9060
北京景点门票	0.6352	−1	0.0371	12.0026
北京夜景	0.6869	0	0.0336	12.1014
烟袋斜街	0.7033	0	0.0457	12.3385
北京旅游景点	0.6722	0	0.0487	12.4432
北京自由行	0.6799	−1	0.0334	12.6311
景山公园	0.7101	0	0.0420	12.6688
石景山游乐园	0.7133	0	0.0481	13.0080
恭王府	0.7164	0	0.0440	13.2487
北京住宿攻略	0.6988	−1	0.0565	13.5083
北京一日游	0.6535	0	0.0660	13.9550

从表 6 可以看出，用户在对旅游六要素中的“食”和“游”方面的关注度明显高于其它方面，其中“奥林匹克公园”、“北海公园”、“天坛公园”、“慕田峪长城”、“烟袋斜街”、“恭王府”等是北京市著名旅游景点，基本上都会被记录在游记攻略中，而北京美食小吃众多，在关键词筛选过程中如

“驴打滚”、“北京烤鸭”、“炸酱面”等具体美食名称被过滤掉，保留了“北京小吃街”和“北京美食”等关键词。

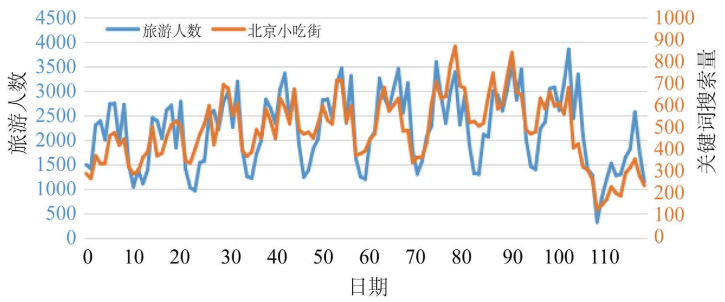
然而，经过斯皮尔曼相关系数、K-L 信息量、DTW 方法筛选保留的 20 个网络搜索关键词并非都能作为最终预测变量，这些关键词之间可能存在多重共线性，造成预测结果过拟合现象。因此，需要进一步对数据降维处理，本文采用 Adaptive Lasso 进行变量选择，结果见图 3。在 R 语言中直接调用 lasso.adapt.bic2 函数，调用格式为：lasso.adapt.bic2 (x=Data[,1:20], y=Data$旅游人数)，其中 x 为需要进行变量选择的 20 个变量集合，y 为因变量北京市旅游人数。同时，根据选取的关键词对应的最佳时滞阶数对数据进行月度的平移调整即错位对齐，其中关键词“北京景点门票”、“北京自由行”、“北京住宿攻略”比北京市旅游人数序列提前 1 期，将其 $t-1$ 期数据与北京市旅游人数的 t 期数据对齐，其它关键词与旅游人数为同步关系，数据保持不变。进行错位对齐后，所有关键词和北京市旅游人数的数据期间均为 2011 年 2 月至 2020 年 12 月，数据量为 119 个。

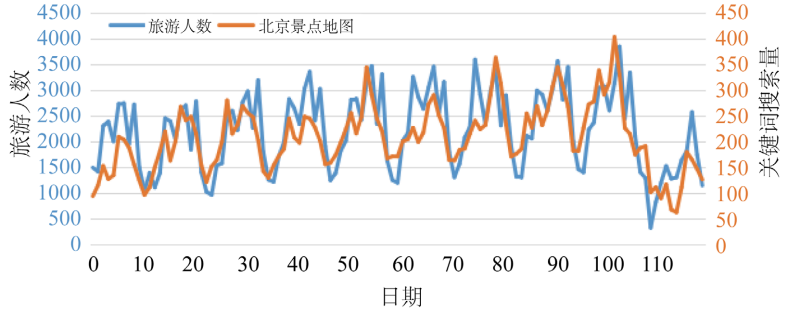
```
> out1$x.ind
[1]  2  4  9 10 11 12 17 18 20
> #保留五位小数
> round(out1$coeff,5)
     北京美食     北京小吃街  北京住宿攻略  北京景点地图       北京游玩
      0.00000      0.52365      0.00000     -1.22331      0.00000
     北京景点       北海公园       天坛公园  奥林匹克公园     慕田峪长城
      0.00000      0.00000      0.00000      0.75210      0.94070
     北京夜景  北京景点门票       烟袋斜街       景山公园  北京旅游景点
     -2.05442      1.66954      0.00000      0.00000      0.00000
   北京自由行       恭王府  石景山游乐园     北京一日游       北京夜市
      0.00000      0.20696      0.42577      0.00000      2.48378
```

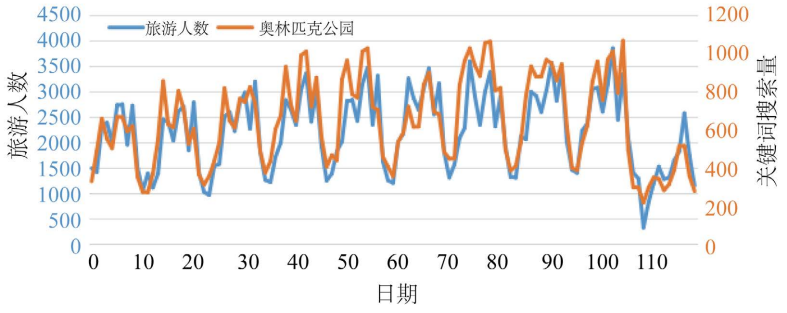
Figure 3. Run results of Adaptive Lasso in R
图 3. Adaptive Lasso 在 R 中运行结果

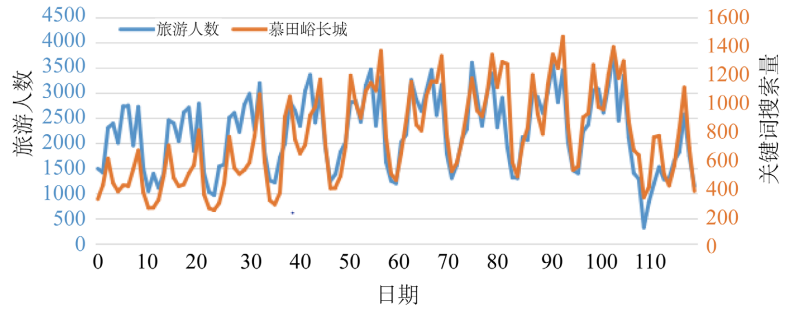
由 R 程序运行结果可知，最终只有 9 个关键词变量的回归系数非零，分别为北京小吃街、北京景点地图、奥林匹克公园、慕田峪长城、北京夜景、北京景点门票、恭王府、石景山游乐园、北京夜市。经过 Adaptive Lasso 变量选择，20 个变量减少为 9 个变量，减少了维度，从而可以降低计算难度，提高计算效率。

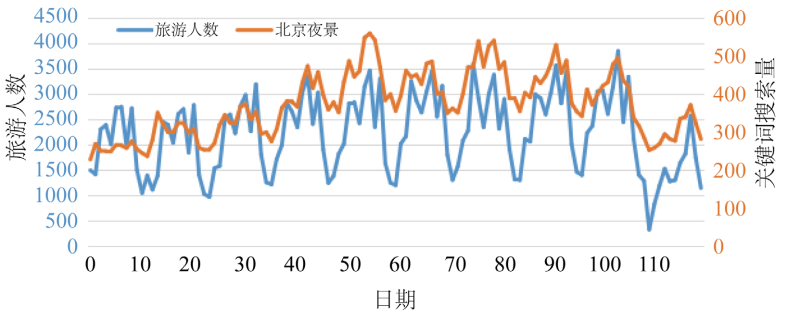
综上所述，经过多个步骤筛选得出的 9 个关键词变量中只有关键词“北京景点门票”比北京市旅游人数序列提前 1 期，绘制得到关键词变量与旅游人数的时间趋势走向，见图 4。可以看出，每个关键词对应的搜索量与旅游人数时间序列均呈现出一致的时间趋势和波动特征，并且波峰和波谷的特征大致相同，充分描述了北京市旅游的发展趋势和季节性。在每年的五一、十一节假日前后以及暑假期间，北京市旅游人数都会达到峰值，主要是在节假日期间吸引大批游客来京旅游，还有在个别年份的二月会出现较小的波峰，主要考虑到春节过后天气回暖，人们旅游需求旺盛。基于以上分析，最终筛选得到的 9 个关键词能够客观反映北京市旅游需求的动态特征，充分体现人们的潜在旅游需求。

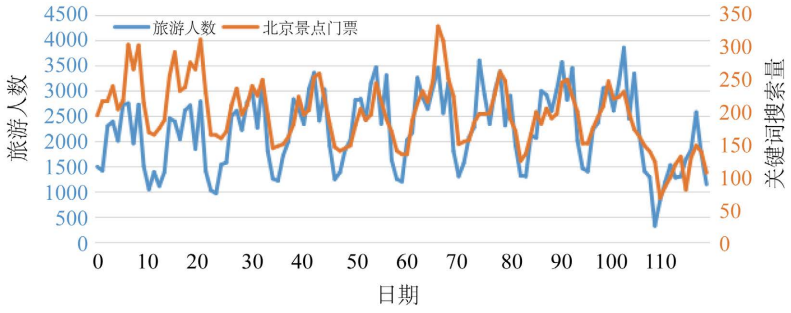



旅游人数
北京景点地图
旅游人数
关键词搜索量
日期


旅游人数
奥林匹克公园
旅游人数
关键词搜索量
日期


旅游人数
慕田峪长城
旅游人数
关键词搜索量
日期


旅游人数
北京夜景
旅游人数
关键词搜索量
日期


旅游人数
北京景点门票
旅游人数
关键词搜索量
日期

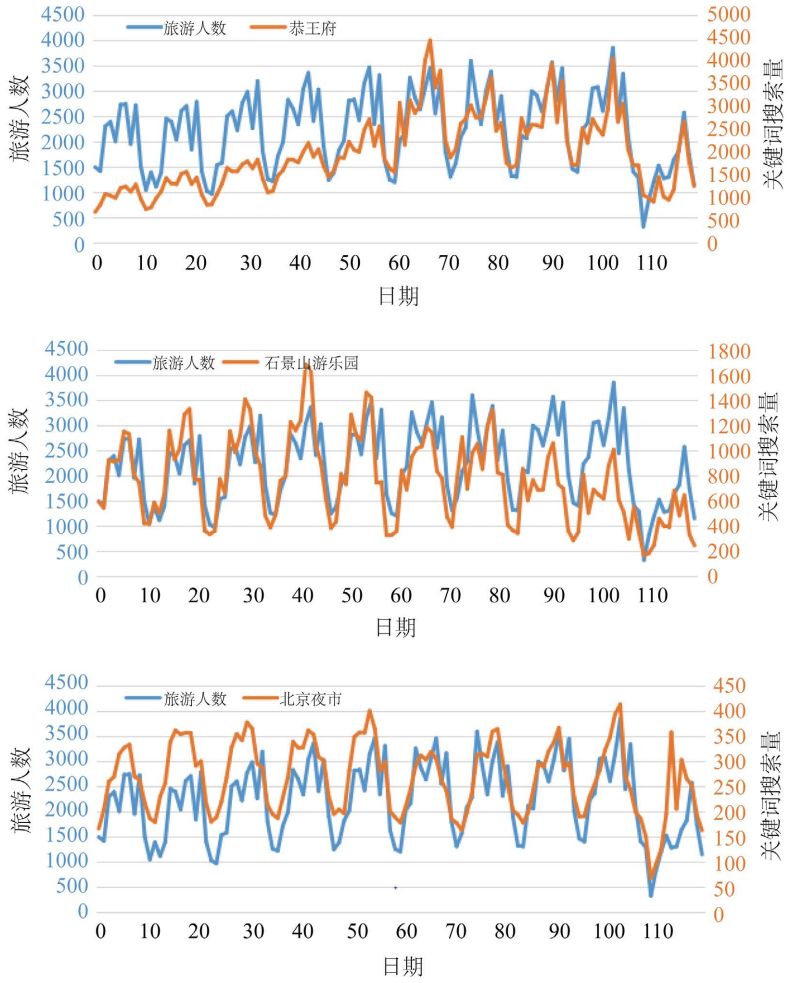


Figure 4. Trends of keyword variables and tourist arrivals
图 4. 关键词变量与旅游人数的趋势图

4. 旅游需求预测模型的建立

4.1. 变量集构造

考虑将 Adaptive Lasso 构造出的 9 个变量，即北京小吃街、北京景点地图、奥林匹克公园、慕田峪长城、北京夜景、北京景点门票、恭王府、石景山游乐园、北京夜市，依次命名为 $x_1, x_2, \cdots, x_9$。错位对齐处理后，所有关键词和北京市旅游人数的数据期间均为 2011 年 2 月至 2020 年 12 月，数据量为 119 个。在构造模型的变量集时，还要考虑旅游人数本身的季节性变化，针对北京市旅游人数序列每年中主要存在的三个客流高峰，本文引入虚拟变量 d_t 进行季节性调整，当 t 为每年的 4 月、8 月、10 月时取值为 1，其余月份 t 为 0。

设模型的自变量集为：

$$X = \left(x_{1,t}, x_{2,t}, x_{3,t}, \cdots, x_{9,t}, d_t\right) \tag{5}$$

模型的因变量集为：

$$Y = \left\{y_t\right\} \tag{6}$$

同时，为避免异常点及不同变量间的数量级差异对预测结果造成影响，需要对变量集进行标准归一化处理，并为更好地对预测模型进行评估，将全部数据分为训练集和测试集两部分，通常按照 3/2~4/1 比例进行数据集划分，考虑到样本量只有 119 个，数据规模较小，测试集占比太大可能会影响训练效果。因此，将 2011 年 2 月至 2019 年 12 月期间的数据作为训练集，2020 年 1 月至 2020 年 12 月期间的数据作为测试集，所用数据见表 7。

Table 7. Internet search keywords and travel demand data
表 7. 网络搜索关键词与旅游需求数据

日期	北京小吃街	北京景点地图	奥林匹克公园	慕田峪长城	北京夜景	北京景点门票
2011.02	293.32	94.64	330.32	317	228.32	195.65
2011.03	271.26	115.55	488.58	415.94	270	217.43
2011.04	373.93	153.2	659.53	603.87	251.3	218.81
2011.05	339.32	127.61	554.61	431.45	249.52	240.7
…	…	…	…	…	…	…
2020.08	295.42	112.13	389.35	552.77	336.45	81.23
2020.09	320.83	180.4	516.4	805.97	340.77	130.48
2020.10	358.23	165.61	517.03	1106.84	372.1	149
2020.11	282.23	146.9	358.3	678.07	324	139.32
2020.12	239.03	126.55	279.03	373.94	282.29	107.53

日期	恭王府	石景山游乐园	北京夜市	季节性变量	境内人数
2011.02	681.29	594.79	166.68	0	1499
2011.03	839.45	545.81	204.58	0	1421
2011.04	1082.03	925.8	260.43	1	2308.1
2011.05	1043.23	913.35	270	0	2394
…	…	…	…	…	…
2020.08	1179.81	683.39	303.52	1	1649.3
2020.09	2091.23	485.03	266.1	0	1827.5
2020.10	2660.39	646.45	252.45	1	2579.3
2020.11	1789.13	332.8	194.07	0	1746.4
2020.12	1252.9	247.94	163.23	0	1151.4

4.2. 基于随机森林算法的旅游需求预测

随机森林(RF)是基于 bagging 框架的决策树模型，构造随机森林有三个步骤：确定用于构造的树的数量；对数据进行自助采样；基于新数据集构造决策树。本文调用 Python 中的 sklearn.ensemble 库，使用 RandomForestRegressor 函数构建随机森林模型。随机森林[7]的主要参数有五个：决策树的数量 n_estimators、树的最大深度 max_depth、最大选择特征数 max_features、最小叶子节点样本数 min_samples_leaf、最小分离样本数 min_samples_split。

训练随机森林时，从训练集中采取 Boostraping 方法随机有放回地抽取样本，借助 gridsearchcv 即网格搜索 + 交叉验证的方法进行参数寻优。由于本文样本量不大，将 n_estimators 的取值范围设置为[0, 200]，步长为 10，max_depth 的取值范围设置为[1, 11]，步长为 1，max_features 为最佳节点分割时要考虑的特征变量数量，不超过所有特征数量，而本文自变量个数为 10，取值设置范围为[1, 10]，步长为 1，min_samples_leaf 采用默认值 1，min_samples_split 采用默认值 2，每种参数组合方式在训练集上训练 9 次，一共训练 1800 × 9 = 16200 次。当模型训练结束后，通过 best_params_确定最优参数组合为 n_estimators = 110，max_depth = 8，max_features = 4，min_samples_leaf = 1，min_samples_split = 2，bootstrap = TRUE。

当获得随机森林预测模型的最优参数组合之后，利用训练的模型在测试集上进行预测检验，得到所建立的预测模型在训练集和测试集上的拟合值曲线，见图 5。

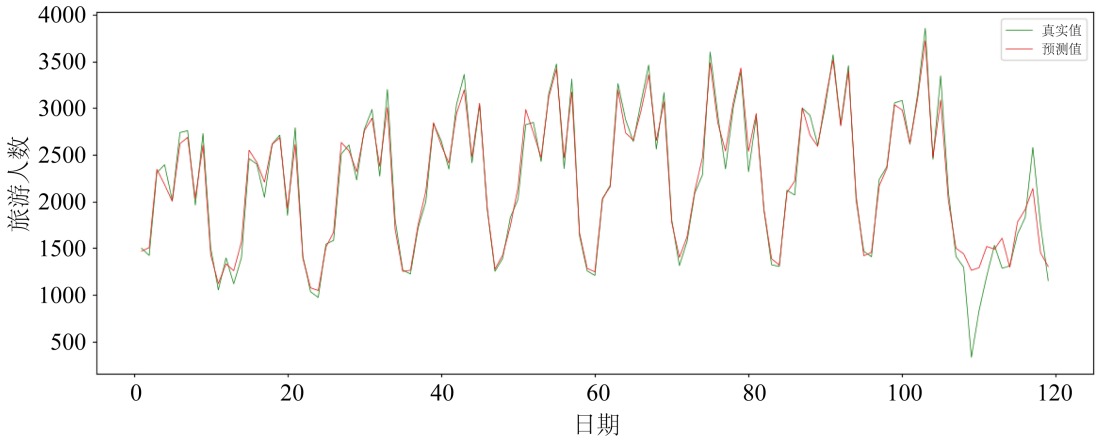


Figure 5. The true and predicted values of the RF model
图 5. 随机森林模型下的真实值与预测值

由图 5 可以看出，随机森林预测模型在训练集上的拟合效果较好，在测试集上拟合效果不太理想，模型泛化能力较差。预测北京市 2020 年 2 月旅游人数为 1262.4 万，而实际旅游人数为 330.6 万，偏差达到 931.8 万，在 3 月、4 月、6 月、10 月及 11 月均有较大偏差，造成训练集和测试集差异的原因可能是由于噪声过大，建模时出现了过拟合现象，预测时对峰值敏感度不高。

4.3. 基于极限梯度提升算法模型的旅游需求预测

极限梯度提升(XGBoost)算法包括通用参数、提升参数和学习参数三种类型的参数，由于 gbtree 的性能比 gblinear 的性能好，并且本文用于回归预测，因此通用参数选用 gbtree，学习参数选用 reg:linear，后续主要对提升参数调优，不同参数对结果的影响不同，按照影响从大到小的顺序调参[8]，运用网格搜索算法和 9 折交叉验证来确定最优参数组合，具体步骤如下：

Step1：n_estimator：42

n_estimator 表示迭代次数，本文设置范围[0, 1000, 10]，得到决策树数量为 40，然后缩小范围为[0, 300]，步长为 1，确定最优决策树数量为 42。

Step2：eta(learning_rate)：0.13

eta 的取值范围为[0, 1]，该参数越小，计算速度越慢，参数越大，有可能无法收敛到真正的最佳，典型取值范围为 0.01~0.2。本文设置范围[0, 0.2]，步长为 0.01，得到最优学习率为 0.13。

Step3：max_depth：5　min_child_weight：7

max_depth 的取值范围为[0, +∞)，该参数越大，模型的学习越具体，越容易过拟合；min_child_weight 的取值范围为[0, +∞)，该参数越大，算法越稳健，越不容易过拟合，但取值过高，会导致欠拟合。本文设置范围均为[0, 200]，步长为 10，得到最大深度为 10，最小权重和为 10，然后设置范围为[0, 20]，步长为 1，确定最大深度与最小权重和的组合点在(5, 7)时，测试集上的得分最高，确定最优最大深度为 5，最优最小权重和为 7。

Step4：gamma(min_split_loss)：0

gamma 表示节点分裂时最小损失函数的下降值，取值范围为[0, +∞]，该参数越大，算法越稳健，能够避免过拟合现象。本文设置范围均为[0, 100]，步长为 10，得到损失阈值为 0，然后设置范围为[0, 1]，步长为 0.01，确定最优损失阈值为 0。

Step5：subsample：0.6　colsample_bytree：0.3

subsample 用于控制每棵树的随机采样比例，即训练样本数量占整体样本数量的比例，取值范围为(0, 1]，该参数越小，算法越稳健，越不容易过拟合，但取值过小，会导致欠拟合；colsample_bytree 用于控制树每一级的每一次分裂过程中对列数的采样比例，取值范围为[0, 1]。本文设置范围均为[0, 1]，步长为 0.1，确定最优样本采样率为 0.6，最优列采样率为 0.3。

Step6：lambda(reg_lambda)：0.01　alpha(reg_alpha)：0.87

lambda 为 L2 正则化参数，取值范围为[0, +∞)，该参数越大，算法越稳健，越不容易过拟合；alpha 为 L1 正则化参数，取值范围为[0, +∞)，该参数越大，越不容易过拟合。本文设置范围均为[0, 100]，步长为 10，得到 L2 正则化参数为 0，L1 正则化参数为 10，然后设置范围为 lambda 范围为[0, 1]，步长为 0.01，alpha 范围为[0, 20]，步长为 0.01，确定最优 L2 正则化参数为 0.01，最优 L1 正则化参数为 0.87。

依照上面的调参顺序，得到最优参数组合见表 8。

Table 8. The tuning parameters
表 8. 调优后的参数

参数	值	参数	值	参数	值
booster	gbtree	max_depth	5	lambda	0.01
n_estimator	42	min_child_weight	7	alpha	0.87
eta	0.13	subsample	0.6	objective	reg:linear
gamma	0	colsample_bytree	0.3		

当获得极限梯度提升算法模型的最优参数组合之后，利用训练模型在测试集上进行预测检验，得到预测模型在训练集和测试集上的拟合值曲线，见图 6。

4.4. 基于支持向量回归算法模型的旅游需求预测

支持向量机(SVR)的参数有三个：核函数 $K(x,x')$、惩罚因子 C、核参数 gramma，其中核函数主要包括线性核函数 Liner、多项式核函数 Polynomial、高斯核函数 Radial、神经网络核函数 Sigmoid。为了将低纬度非线性预测问题转化为高纬度线性可预测问题，减少预测过程中的内积运算量，本文选择引入高斯核函数，相比其它核函数，高斯核函数的灵活性比较高，应用更为广泛[9]，具体表达式为：

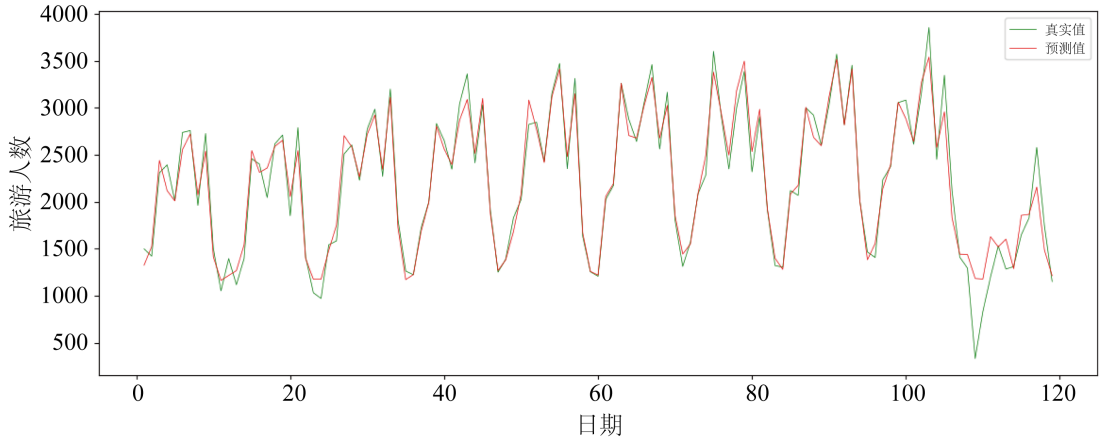


Figure 6. The true and predicted values of the XGBoost model
图 6. 极限梯度提升模型下的真实值与预测值

$$K\left(x_i, x_j\right) = \exp\left(-\frac{\left|x_i - x_j\right|}{2\sigma^2}\right)^2 \tag{7}$$

在确定核函数之后，需要寻找最优惩罚因子和核参数，考虑到本文参数较少，选择网格搜索的方法对参数寻优。网格搜索是一种穷尽式的搜索方法，通过遍历每一个点来得到最优解，实际上就是将需要寻找的参数在一定空间范围内按照拟定的坐标系划分成长短相同的网格，其中的每一个网格点代表一种参数组合，通过将每一种参数组合带入到支持向量回归中进行验证，使得结果最优的网格点即为最优参数组合。本文将惩罚因子 C，核参数 gamma 和距离误差 epsilon 的取值范围都设置为$[2^{-10}, 2^{10}]$，在 epsilon-SVR 模型中指定了模型的类别为 epsilon-tube。同时，根据本文所用数据的周期数，对每一种参数组合结合 9 折交叉验证方法对参数进行检验。在 Python 中调用 SVR 包，对归一化处理后的数据采用网格搜索和交叉验证进行循环调试，确定最优参数组合为 $C = 771.67$，gamma = 0.0386，epsilon = 0.0017。预测模型在训练集和测试集上的拟合值曲线，见图 7。

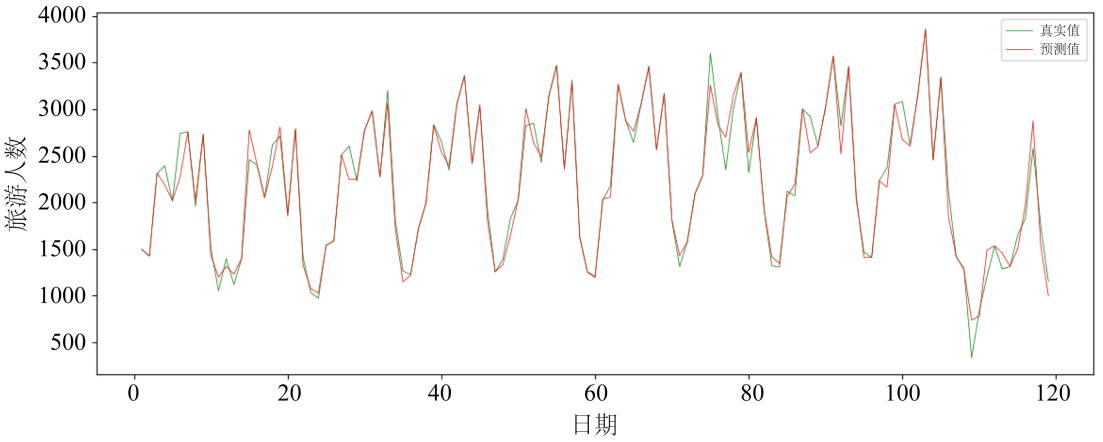


Figure 7. The true and predicted values of the SVR model
图 7. 支持向量回归模型下的真实值与预测值

由图 7 可知，支持向量回归模型在整个样本区间上的拟合效果较好，泛化能力较强。利用该模型预

测北京市 2020 年 1 月至 12 月的旅游人数时，测试集前期的预测值和真实值之间有所偏差，2020 年 2 月最为明显，其次是 10 月和 11 月，其余大部分月份预测值拟合效果较好，接近旅游人数真实值。造成测试集前期偏差较大的原因主要有两个：其一，本身支持向量回归模型前期对数据的灵敏度不高，且数据样本量较大时训练比较困难，随着训练时间的增加，模型的容错程度所有提升；其二，2019 年 12 月底全国爆发新冠肺炎疫情，在 2020 年 2 月 12 日每日新增确诊人数达到顶点，截止 16 日全国确诊病例超过 7 千人，疫情的爆发对全国旅游业造成了深重的影响，各地旅游均处于停滞状态，北京市前九年旅游人数走向基本一致，呈稳定上升态势，而 2020 年一整年的旅游人数都处于低迷状态，这在很大程度上影响了 2020 年的预测结果。

4.5. 模型评价

模型性能可从模型的拟合程度和泛化能力两方面进行评价，具体来说就是利用训练集误差和测试集误差两个指标进行度量，经过模型拟合训练，得到三个模型在测试集上的预测值，见表 9。

Table 9. Comparison of the forecast results of tourist arrivals to Beijing of different models
表 9. 各模型对北京市旅游人数预测结果对比

日期	真实值	随机森林模型	极限梯度提升模型	支持向量回归模型
2020.01	1293.8	1438.4	1439.0	**1279.2**
2020.02	330.6	1262.4	1183.7	**735.3**
2020.03	830.3	1290.1	1176.9	**779.9**
2020.04	1201.4	1517.8	1628.2	**1484.8**
2020.05	1529.9	1488.5	1520.2	**1533.4**
2020.06	1284.1	1607.0	1603.1	**1454.7**
2020.07	1309.6	1294.4	1287.2	**1312.3**
2020.08	1649.3	**1779.2**	1855.8	1500.0
2020.09	1827.5	1913.6	**1866.1**	1970.5
2020.10	2579.3	2138.8	2157.0	**2875.4**
2020.11	1746.4	1450.1	1487.6	**1537.8**
2020.12	1151.4	1303.0	**1213.1**	995.5

由表 9 可知，在 2020 年 12 个月的预测中，支持向量回归模型的最优预测值最多，有 9 个月，而随机森林和极限梯度提升模型的最佳预测值较少，分别为 1 个月和 2 个月。根据每个模型在训练集和测试集上的拟合曲线，可以看到三个模型均能较好地拟合旅游人数训练集部分的时间序列，随机森林模型最为接近，预测值与真实值基本能够重叠，但只有支持向量回归模型能够较好地拟合测试集部分的时间序列，在整个时间序列区间上预测曲线与真实曲线最为接近，更好的拟合旅游人数曲线的动态特征。

为进一步验证结论，选取平均绝对误差(MAE)、均方误差(MSE)和拟合优度(R^2)三个评价指标模型拟合预测性能，具体评价指标值见表 10。

由表 10 可知，随机森林模型在训练集上的拟合效果最好，其中 R^2 高达 0.9826，但在测试集上的各种误差值均较大，预测精度最差，R^2 仅为 0.5152。极限梯度提升模型在训练集上各评价指标值与随机森林模型相差不大，在测试集上稍有进步，而支持向量回归模型在训练集评价指标值一般，还能保证在测

试集上的拟合优度 R^2 为 0.8595，拟合效果较好。总结来看，三种模型的拟合效果都较好，支持向量回归模型相比前两个模型具有较好的泛化能力。即使 2020 年旅游经济受到疫情影响，旅游人数数据变动趋势较往年变化太大，在这种影响下，支持向量回归模型仍然能够取得较好的预测效果。

Table 10. Comparison of the performance evaluation indexes of different models
表 10. 各模型的性能评价指标对比

模型	训练集			测试集		
	MAE	MSE	R^2	MAE	MSE	R^2
随机森林	0.0213	**0.0007**	**0.9826**	0.0789	0.0110	0.5152
极限梯度提升	0.0284	0.0014	0.9663	0.0735	0.0097	0.5712
支持向量回归	**0.0201**	0.0013	0.9683	**0.0445**	**0.0032**	**0.8595**

5. 结论

本文结合常态化疫情防控措施从旅游需求的实用价值出发，利用百度指数搜索量，结合考虑“吃、住、行、游、购、娱”旅游六要素，并通过使用文本挖掘方法，得到与旅游需求相关的网络搜索数据，构建初始关键词词库，通过灰色关联度方法得出网络搜索数据与旅游需求之间相关性较强。加之旅游时间序列自身的季节性特征，最后使用三种机器学习算法对北京市月度旅游人数进行预测。结论如下：

1) 网络搜索数据与旅游需求之间相关性显著。本文通过网络旅游攻略和旅游六要素的文本挖掘，构建初始关键词词库，并通过灰色关联度方法得出网络搜索数据与旅游需求之间的相关性很强。同时，通过旅游需求图谱、百度指数等推荐，构建拓展关键词词库，并通过计算斯皮尔曼相关系数、K-L 信息量和 DTW 动态弯曲距离进行筛选，最终保留北京小吃街、北京景点地图、奥林匹克公园、慕田峪长城、北京夜景、北京景点门票、恭王府、石景山游乐园、北京夜市这 9 个关键词。可以看出，人们更加关注美食、景点和娱乐方面。这些关键词对北京市旅游人数序列影响最大，与旅游需求具有同步波动特征，当地政府和旅游主管部门可以就此进行多渠道宣传来吸引游客，创造更多的旅游收入，从而推动旅游业恢复疫前水平并保持高质量发展。

2) 网络搜索数据时效性很强。通过计算斯皮尔曼相关系数和 K-L 信息量，可以发现相关性较高的关键词对应的时滞阶数均为 0 或−1，即大部分游客会在出游前一个月甚至当月进行网络信息搜索。对比官方公布的旅游人数数据，通过百度指数取得关键词搜索量较为便捷，并且更具时效性，只需要在每月月底获取关键词的百度指数搜索量，数据处理后就可对旅游需求进行预测。

3) 支持向量回归模型能够很好地处理突发事件和小样本问题，用于短期旅游需求预测。在本文中，实际用于模型拟合的样本量只有 119 个，是小样本问题，容易产生过拟合现象。本文使用随机森林、极限梯度提升和支持向量回归三种算法建立模型，结果发现，随机森林算法在拟合效果最优但泛化效果最差，出现过拟合问题。支持向量回归模型的拟合效果和泛化能力都很好，显著优于随机森林和极限梯度提升模型。面对突发的新冠肺炎疫情影响下，支持向量回归模型仍可得到较好的预测效果，说明该方法具有很强的可靠性和适用性。

基金项目

邯郸市科学技术研究与发展计划项目(No.21422304303)；河北省高等学校科学技术研究资助项目(Nos.QN2019064，ZD2020185，ZD2020130)；河北省自然科学基金资助项目(Nos.A2019402043，A2021402008)。

参考文献

[1] 中国互联网络信息中心. 第 47 次中国互联网发展状况统计报告[R/OL]. http://cnnic.cn/hlwfzyj/hlwxzbg/hlwtjbg/202102/P020210203334633480104.pdf, 2021-02-11.

[2] 清华大学中国经济思想与实践研究院(ACCEPT)宏观预测课题组, 李稻葵. 中国宏观经济形势分析与未来取向[J]. 改革, 2021(1): 1-17.

[3] 北京市人民政府. 北京市推进全国文化中心建设中长期规划(2019 年-2035 年)[EB/OL]. http://www.beijing.gov.cn/zhengce/zhengcefagui/202004/t20200409_1798426.html, 2020-04-12.

[4] 中国互联网信息中心. 2019 年中国网民搜索行为调查报告[R/OL]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/ssbg/201910/P020191025506904765613.pdf, 2021-02-11.

[5] 黄蓉. 中国城镇居民的国内旅游需求研究[D]: [博士学位论文]. 武汉: 华中科技大学, 2015.

[6] 彭赓, 刘金烜, 曾鹏志, 李晓炫. 时间序列相似性与基于搜索数据的预测研究——以九寨沟客流量预测为例[J]. 管理现代化, 2016, 36(2): 107-110.

[7] 张倩. 基于随机森林回归模型的住房租金预测模型的研究[D]: [硕士学位论文]. 长春: 东北师范大学, 2019.

[8] 龚洪亮. 基于 XGBoost 算法的武汉市二手房价格预测模型的实证研究[D]: [硕士学位论文]. 武汉: 华中师范大学, 2018.

[9] 王芳. 基于支持向量机的沪深 300 指数回归预测[D]: [硕士学位论文]. 济南: 山东大学, 2015.