大数据视角下中国游客赴贝加尔湖地区时空分布与现状研究

Wendy Zhang

摘要：受中俄交流深化及免签政策推动，中国赴贝加尔湖旅游人数快速增长。然而，然而，受制于贝加尔湖所在的俄罗斯西伯利亚地区旅游开发程度有限、基础设施建设滞后及配套服务水平不足，该区域的旅游承载能力相对薄弱。本文基于大数据方法，爬取并分析主要旅游平台上的游客评论与游记，结合文本分析、旅游地理分析及 GIS 可视化技术，揭示中国游客赴该地区旅游的时空分布特征。研究结果表明，近年来中国游客数量显著增加，但当地旅游接待能力难以满足需求，存在恶性循环风险。本文建议改善交通、优化线路、发展伊尔库茨克旅游产业，并完善立法以兼顾发展与环境保护。

关键词：数据分析；贝加尔湖；旅游产业；发展对策

1. 研究背景

1.1 贝加尔湖概况

贝加尔湖是世界上最大，最深的淡水湖，被称为“西伯利亚之眼”。 该湖位于俄罗斯西伯利亚南部的伊尔库茨克州和布里亚特共和国，拥有美丽的风景，奇特的景观和丰富的物种。湖总容积 23.6 万亿立方米，最深处达 1637 米，是世界第一深湖、欧亚大陆最大的淡水湖。湖长 636 千米，平均宽 48 千米，面积为 3.15 万平方千米，由地层断裂陷落而成，湖面海拔 455 米左右。

随着贝加尔湖旅游名声鹊起,越来越多的中国游客前往贝加尔地区旅游,据统计在 2019 年一二月就有超过 36000 名游客探访贝加尔湖地区。游客传统上一般分南线与北线进行游览，但随着人数的增多，传统旅游路线已无法满足游客需求。随着互联网的不断发展，游客习惯于在相关的旅游网站与社交媒体上发表评论与游记。根据这些评论与游记，我们可以大体上评判中国游客对贝加尔湖地区旅游业发展变迁与现状的感知。总体而言，中国游客对贝加尔湖风景评价偏向于正面，但缺少对当地基础设施与旅游体验的评价，就我们进行的先期调研表示，贝加尔湖地区基础设施建设速度较慢，在交通状况、网络通畅度和居住条件方面等游客体验可能不尽如人意。本次研究主要聚焦于贝加尔湖落后的旅游接待能力是否遏制了中国游客的增长，并希望就研究结果能给贝加尔湖旅游产业健康发展提出一些建议。

1.2 调研方法

本次研究主要采用的是大数据分析，运用 python 爬取中国主要旅游网站中用户对贝加尔湖旅游的评论与游记，进行文本分析，结合旅游模型与 GIS 分析，得出贝加尔湖旅游时间与空间变化。

我们主要聚焦于三方面，分别为：（1）评估，即贝加尔湖旅游景点演变的客观综合评价（2）变迁，即通过爬取大数据，分析贝加尔湖地区中国游客的时空分布变化（3）建议，即提出关于如何在贝加尔湖地区发展旅游业的建议，以便能够满足不断增长的游客需求。

2. 研究理论与方法

2.1.python 爬取数据

根据现有条件与网络接口情况，考虑到现有的文本分析能力，为了平衡评论与博客数量与质量，在本次研究中，我们着重挑选了新浪博客，携程网，去哪儿网，艺龙与大众点评这五个平台作为数据源网站。本研究使用 Python 设计的分布式爬虫系统。我们按年份和月份搜索社交媒体上的简短评论，旅行笔记。经过爬取，我们在这些网站上共获得了 1710 个有效评论。其中，包括新浪网 756 篇博客，携程 400 条评论与 223 条旅游短评，去哪网 160 条旅游短评，艺龙 95 条旅游短评，大众点评 75 条评论。

对所爬得的数据进行三方面的分析，主要包括：

（1）对网络文本直接进行内容分析；

（2）使用自然语言处理工具对旅行网站中的文本数据进行爬行和分析，即文本分析。

（3）构建数学模型以分析旅行特征

2.2 旅游分析模型

（1）季节强度指数

季节性强度指数主要反映贝加尔湖地区旅游需求强度的变化，其中需求由数据收集的评论数量决定。

$$\mathbb{R} = \sqrt{\sum (x_{i} - \mathrm{x})^{2}/_{12}} \quad \text{.....(1)}$$

R 指季节性强度指数，x 是平均需求量；xi 是该月贝加尔湖旅游需求的年度占比.R 值趋于零，月度分布越均匀；R 值越高，季节差异越大。

（2）旅游地生命周期理论

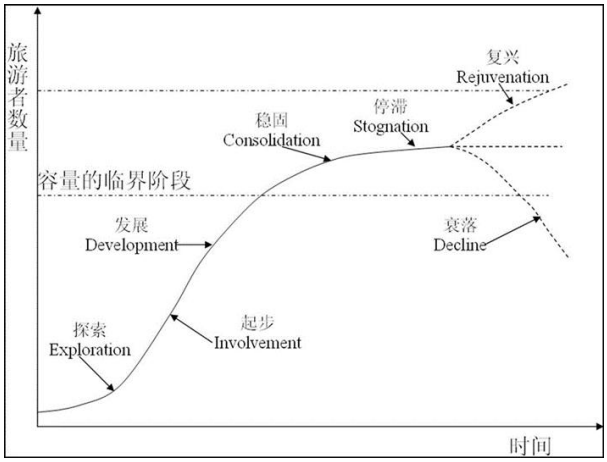


图 1

旅游目的地有其兴衰的模式，即旅游地的生命周期。在“参与”（involvement）阶段，由于充足的旅游设施供给以及随后的广告宣传，使旅游者数量不断增加。在“发展”（development）阶段，旅游者数量增加更快，而且对旅游经营实施控制的权力也大部分从当地人手中转到外来公司的手中。在“巩固”（consolidation）阶段，尽管旅游者总人数仍在增长，但增长的速度已经放慢。至于“停滞”（stagnation）阶段，旅游者人数已经达到高峰，旅游地本身也不再让旅游者感到是一个特别时髦的去处了。而到了“衰退”（decline）阶段，因旅游者被新的度假地所吸引，致使这一行将衰亡的旅游一日旅游者和周末旅游者的造访来维持其生计。

（3）旅游需求预测模型

该模型用于评估贝加尔湖地区的旅游承载力，并根据现有数据进行预测，为贝加尔湖地区的旅游规划和建设提供建议。

旅游承载力是由环境承载力发展而来，结合旅游地服务水平，环境质量等多方面因素，综合评价得来。由实地体验可以得知，贝加尔湖旅游业发展并未跟上其日益扩大的需求，因而可以粗略的估计，贝加尔湖地区旅游承载力较低，这为规划提出更高的要求。

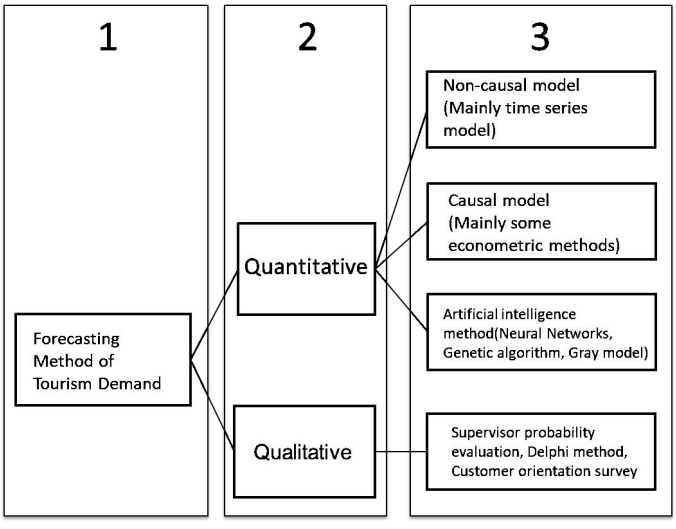


图 2

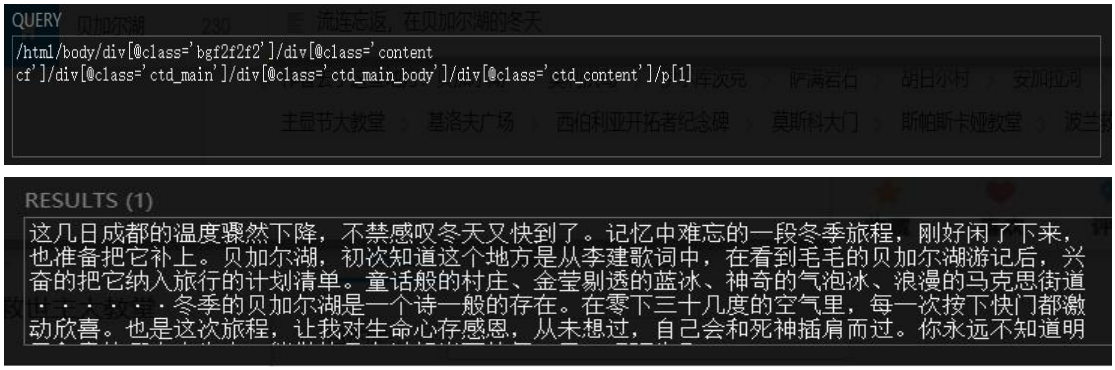
3. 研究过程与结果

3.1 文本爬取

我们挑选了新浪博客，携程网，去哪儿网，艺龙与大众点评这五个平台作为数据源网站。首先在被爬取网站搜索关键词“贝加尔湖”，得到相关词条的搜索结果。对搜索页面之间的翻页跳转关系进行观察，得出各个页面 url 之间的关系，完成爬虫代码中搜索结果翻页功能的基础实现。

查看搜索页面源代码，发现大部分网站的搜索结果对应文章 url 都能在源代码中找到，少数网站采用了动态网页，需要在 JS 动态请求接口后对请求结果中寻找。由于游记的 url 格式比较固定，如携程中关于贝加尔湖的游记 url 统一为“https://you.ctrip.com/travels/lakebaikal4726/*******.html”，“*”代指每篇游记不同部分。因此可以使用正则表达式，定向地选择出网页源代码中的下一级 url。这样，我们就获得的所有与贝加尔湖有关的游记的 url。

打开一篇游记，对网页内容进行审察。这里使用了浏览器插件 xpath helper 进行辅助，这一插件能定位网页中特定元素在这一网页的 html 文本中的 xpath，从而能在爬取过程中定向爬取正文内容。



在爬取过程中，我们发现爬取下来的正文中干扰项较多，主要有各种标点、数字以及乱码符号。为了使分词结果更加准确，我们在代码中加入了文本筛选功能，通过判断字符的 unicode 编码，定向地保留出正文中的中文字符部分。

最后，我们通过 python 中的 jieba 库对获得的原始文本文件进行分词，完成文本的爬取与预处理阶段。

3.2 词频分析

通过文本分析，对所爬取的大篇幅的游记进行词频统计，获得了以下十个高频词汇：

1.贝加尔湖 2.奥利洪岛 3.利斯特 4.木屋博物馆 5.喀山大教堂 6.马克思列宁大街 7.萨满石、萨满柱 8.叶尼塞 9.安加拉河 10.胡日尔镇 11.二战纪念广场

具体频数如下表所示：

贝加尔湖	1151
奥利洪岛	368
利斯特	249
木屋博物馆	143
喀山大教堂	132
马克思	105
列宁大街	39
萨满石、萨满柱	113
叶尼塞河	24
安加拉河	257
胡日尔镇	159
二战纪念广场	48

运用高德开放平台 Map Lab 对其进行可视化处理。由于词频提取得到的高频词汇中有部分地名涵盖区域重合，所以通过筛选得到了最后共七处地名，如下表。

所在州	地名	频数
伊尔库兹克州	贝加尔湖	1151
伊尔库兹克州	安加拉河	263
伊尔库兹克州	利斯特维扬小镇	257
伊尔库兹克州	木屋博物馆	147
伊尔库兹克州	喀山大教堂	139
伊尔库兹克州	列宁大街	41
伊尔库兹克州	胡日尔村	165
伊尔库兹克州	萨满石柱	120
伊尔库兹克州	二战纪念广场	48

之后通过 Google 地图找到每个地点的经纬度坐标，要注意的是 Google 地图上获得的都是以度分秒为单位的经纬度，需要将其转化为小数形式，如 52°31′要转成 52.5167。将地名、频数、经纬度制成图表导入到平台中，如下图所示。

再通过创建可视化项目得到效果图，即图 3。

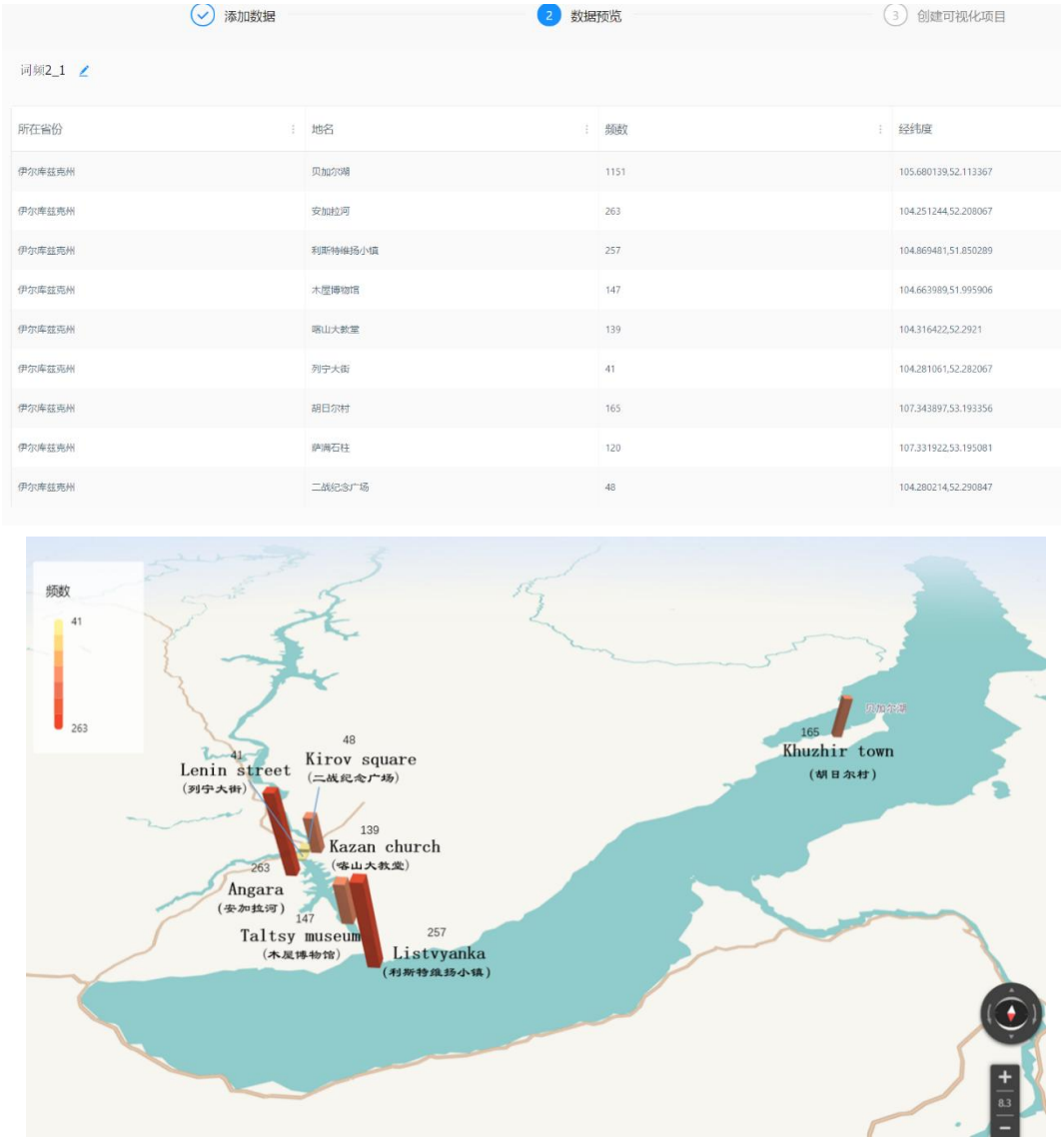


图 3

图中圆柱高度即为词频数，圆柱越高，即为此地名提及的次数越多，此地在中国游客的游览路线中越重要。

我们可以看出，游记中提到的贝加尔湖地区只是一个大的范围，并没有我们预想中根据游记可以推测出其旅游路线，同时也可以看出对于中国游客来说，贝加尔湖地区旅游具体的景点集中在几个岛上，且被提到的景点并不多，更多的可能是在湖畔游玩。根据我们的实地考察，我们发现在伊尔库兹克市区的景点我们遇到了很多中国人，但我们实地科考的路途中基本没有遇到中国人，我们去科考的很多地方景色都非常美，但游客很少，甚至本地游客都很少，根本看不到中国游客的身影。优越的自然风光适宜发展旅游业，但是相应的周边服务休闲场所基本没有，并且交通不便，未完工的泥土路面限制了行驶速度且非常颠簸，更大的限制因素是网络信号的不便，在很多地方根本没有信号，更不用说是用流畅的 4G 网络来导航了，这就限制了很多中国游客偏爱的自驾游的路线。

同时通过这张图还可以进行旅游学分析，从图上可以清晰地看出，中国游客贝加尔湖的游览路线中都是以伊尔库兹克作为中转站，因而，伊尔库兹克的景点具有较高词频统计数。伊尔库兹克作为俄罗斯远东重要城市，也是重要的交通枢纽，中国去往贝加尔湖的游客一定会经过伊尔库兹克，所以在贝加尔湖的旅游发展中，伊尔库兹克是非常重要的一环，在贝加尔湖旅游规划过程中，应该注意伊尔库兹克的相关配套设施建设，并且可以抓住机会带动伊尔库兹克市的旅游业发展。

3.3 关联性分析

对提取出来的词进行关联性分析，如图 4；每个节点代表一个旅游景点，节点之间的连接代表不同旅游景点之间的相互关联程度，数字越大，所代表的两个节点一同出现在一篇游记的概率越大。

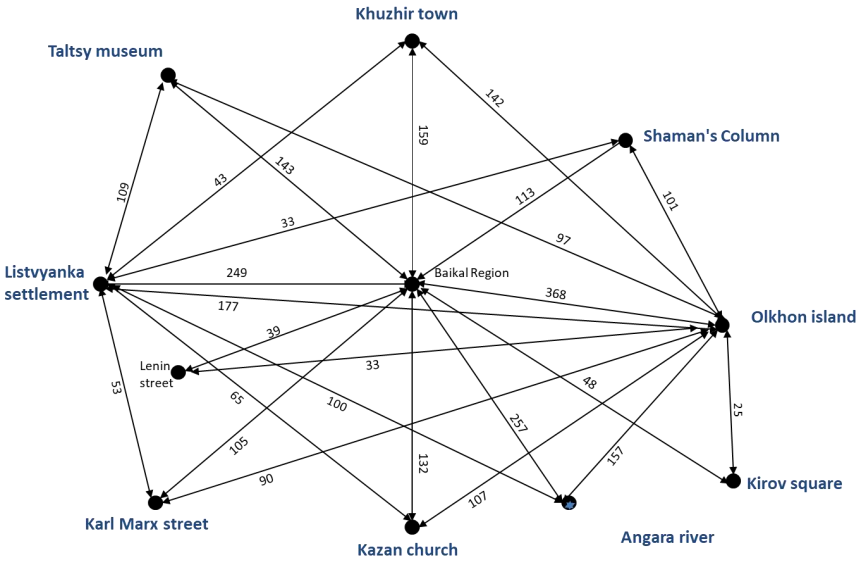


图 4

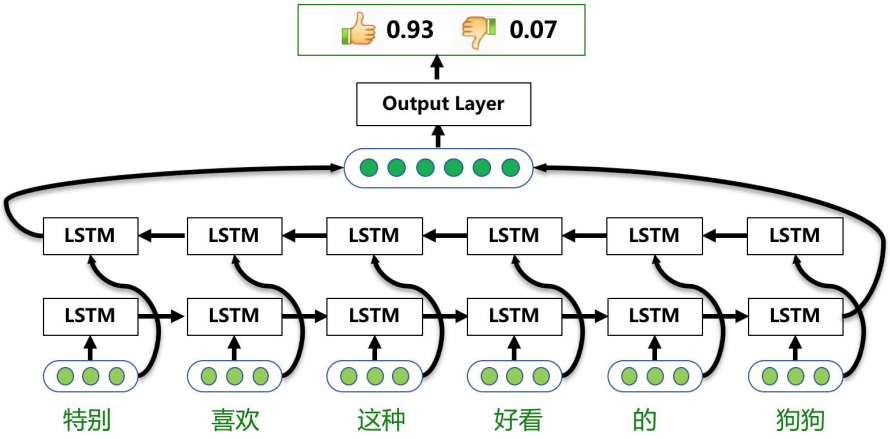
关联性分析可以有效分出两个旅游景点关联程度，即一同被游览的概率。旅游样本中旅游景点的频率反映了游客对景区的偏好，共现分布反映了游客对两个或两个以上景点的共同偏好。当游客首先期望的景点附近还有其他旅游景点时，根据时间成本考虑，游客将伴随着地理位置相对较近的旅游景点，以及远离彼此的景点。它具有显着的距离阻挡效果。而贝加尔湖作为旅游核心景点，对周围景点的发展具有极强的带动作用。以此作为依据，对团体旅游路线开发具有指导作用。同时，也能够对俄罗斯相关旅游产品的开发提供参考，比如两个高关联度的旅游点可以减少相似的纪念品售卖，从而最大程度程度上进行差异化服务。

我们可以发现贝加尔湖地区的各个旅游景点虽有联系但分散程度比较大，景点之间距离较远，说明贝加尔湖地区仍是单节点旅游中心，还并未发展成多节点旅游中心，但根据我们的实地考察我们认为从单节点向多节点旅游中心地转变的过程中已经有了自然景观的条件，还需加强不同景点之间服务和基础设施建设就有潜力向多节点旅游中心地发展，带动整个贝加尔湖迪区旅游业的发展。在伊尔库兹克市内的旅游景点相对集中，应该注重旅游区的规范

化管理。

3.4 旅游满意度分析

基于爬取评论的文本分析，进行游客满意度分析，即为情感分析。这里的情感分析采用了百度的 senta 库，senta 的开源代码默认使用了 bi-LSTM 模型。将原始文本导入，senta 库自动执行分词操作，再将分词结果套用 bi-LSTM 模型进行情感分析。这个模型包括三层：单词语义层，句子语义层，输出层。（1）单词语义层，主要是将输入文本中的每个单词转化为连续的语义向量表示，也就是单词的 embedding。（2）句子语义层，通过 bi-LSTM 网络结构，将单词语义的序列转化为整个句子的语义表示。（3）输出层，基于句子语义计算情感倾向的概率。



图源百度 AI

我们结合游客打分。得出如图 5 表的分数（满分为 1 分）

Year	Number of comments	Average score
before2015	44	0.9
2016	67	0.87
2017	79	0.92
2018	151	0.9
2019	142	0.92

根据以上情感分析结果，贝加尔湖地区的大多数游客都对旅游体验持正面态度，这有效的提升了贝加尔湖地区的声誉，并且这几年来游客对贝加尔湖评价基本保持不变，并未随着人数的增多有大幅变化，说明贝加尔湖旅游产品的核心即其自然景观并未随着接待人数的膨胀，而产生较大的质量滑坡，即其环境保护较好。

当然，以上的情感分析具有较大的局限性，如：抓取的评论数据较少、网上愿意评论的人并不能有效覆盖全体，满意度打分情况不能细化等因素都制约了我们的研究。考虑到游客评价体系不同，容易出现非常满意与非常失望这样对立的体验，这也是本次研究的缺憾所在，由于贝加尔湖行程问题，未能有效发放问卷，从而获得更加精准的数据，当由于分数较高，

不存在较大方差，因而具有一定的可信度与参考价值。

而差评主要集中于旅游过程中的各种消极感知：（1）坐车颠簸（2）景色一般（3）没有好酒店（4）极寒(5）交通不便。也为后续为贝加尔湖旅游产业提出建议提供了参考。

根据我们的实际考察，我们认为贝加尔湖地区最大的优势就是优美的自然风光与多样的自然景观，而缺点也很明显，即缺少有效和合理的规划开发，周边基础设施建设不够完善，休闲娱乐等服务设施也有待提高。

3.5 季节性强度指数变迁

我们根据评论与游记时间整理了不同时间段的数目（如图 6），虽然评论与游记存在滞后性，但长时间尺度下，滞后性可以忽略不记。从其数量可以从一定程度上反应赴贝加尔湖啊湖旅游人数变迁，并由此计算季节性强度指数。

图 6

根据季节性强度指数公式，分别计算出 2015-2018 年的季节性强度指数值（如图 7），R 值均较小，说明中国游客赴贝加尔湖旅游需求季节性变化较小，无明显淡旺季，不受中国假期影响。且整体呈下降趋势，说明各季节人数差异在进一步减小。

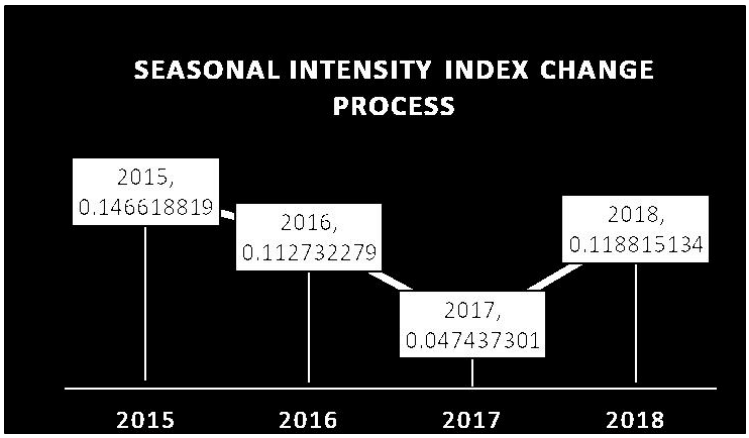


图 7

3.6 旅游生命周期评估

根据旅游生命周期理论，贝加尔湖旅游业的发展正处于发展阶段。 经过实地考察，我们发现发现目前有少数规模化的旅游服务设施，但也可以发现，旅游服务设施还很短缺，并且在快速建设过程中，还存在相当多问题，比如网络信号相当不稳定且速度较慢，正在施工的道路由于需要开山爆破，明显对当地环境产生一定影响。

旅游承载力是由环境承载力发展而来，结合旅游地服务水平，环境质量等多方面因素，综合评价得来。由实地体验可以得知，贝加尔湖旅游业发展并未跟上其日益扩大的需求，因而可以粗略的估计，贝加尔湖地区旅游承载力较低，这为规划提出更高的要求。同时旅游基础服务设施的开发可能会与环境保护相冲突，因而，在施工过程中应减少对环境影响，避免对生态环境造成无法修复的破坏，保护贝加尔湖独有的地理与地质现象与物种。

4．研究结论

本次研究主要聚焦于三方面，分别为：

(1) 评估，即贝加尔湖旅游景点演变的客观综合评价

(2) 变迁，即通过大数据爬取，分析贝加尔湖地区中国游客的时空分布变化

(3) 建议，即提出关于如何在贝加尔湖地区发展旅游业的建议，以便能够满足不断增长的游客需求

聚焦于以上三个研究目的，结合研究结果，我们可以得到结论：

近年来越来越多的中国游客到贝加尔湖旅游，但贝加尔湖地区的旅游能力无法满足他们的需求。贝加尔湖以其美丽的风景而闻名，这有助于它在中国赢得良好的声誉。 但是，应该迫切解决一些影响旅游体验的问题，否则，无法满足快速增加的游客需求，从而使贝加尔湖这个绝佳的旅游地进入恶性循环，最终陷入衰退，这不仅是游客的损失，也是俄罗斯政府的损失，这将使其失去大笔外汇收入，同时依附于旅游业的大量人员将承担失业与劳动报酬大幅下降的风险，对环贝加尔湖繁荣地区是致命打击。

针对我们发现的问题，提出以下建议：

(1) 改善道路状况，开辟新的交通方式。就实地考察而言，在伊尔库兹克东部与贝加尔湖之间由于地广人稀，现有公路能满足需求，但伊尔库兹克市内交通不够便利，对游客不够友好，公交系统新旧交织，有的公交没有英语或中文播报，增加了外国游客的出行难度。

(2) 根据景点相关性设置新的旅游线路和旅游区域，虽然贝加尔湖以自然风光著称，且偏向休闲疗养，但如果没有新的旅游线路投入，难以吸引一般游客二次游览与自发的宣传活动。

(3) 发展伊尔库茨克旅游产业，因为由我们数据分析发现伊尔库茨克是贝加尔湖旅游的主要的中转站，随着赴贝加尔湖游客的增多，对伊尔库兹克有明显的带动效应，如果能合理开发伊尔库兹克的旅游产业，可以将其打造另一个旅游节点，从而形成区域内的多节点旅

游模式，对区域性旅游有良性促进作用。

（4）立法促进当地旅游业发展，并且聚焦环境保护。在贝加尔湖某些地段，由于过度开发已造成特有地貌的破坏。过多的垃圾残留与环境破坏也遭到当地居民的反对，甚至有在贝加尔湖旁制造矿泉水的中国企业因民间压力而关闭。

参考文献：

[1] Buhalis D, Law R. Progress in Information Technology and Tourism Management: 20 years on and 10 years after the Internet-the state of etourism research[J]. Tourism Management, 2008,29(4):609-623.

[2] Sigala M, Christou E, Gretzel U. Social Media in Travel, Tourism and Hospitality: Theory, Practice and Cases[M]. Fanhamn: Ashgate Publishing, Ltd, 2012.

[3] 梁增贤,保继刚. 主题公园黄金周游客流季节性研究——以深圳华侨城主题公园为例[J]. 旅游学 刊,2012,27(1):

[4] 何颖怡,麻学锋. 武陵源与黄龙洞景区客流量倒“U”结构成因及机制分析[J]. 经济地理,2014,34(5):

[5] 张铁生,孙根年. 旅游地客流量峰林结构及成因探析——湖南凤凰入境旅游与国内旅游的比较[J]. 旅游 科学,2014,(1):

[6] 黄潇婷,马修军. 基于 GPS 数据的旅游者活动节奏研究[J]. 旅游学刊,2011,26(12):

[7] 黄潇婷. 基于时间地理学的景区旅游者时空行为模式研究——以北京颐和园为例[J]. 旅游学 刊,2009,24(6):.

[8] 张子昂,黄震方,靳诚,等. 基于微博签到数据的景区旅游活动时空行为特征研究——以南京钟山风景名 胜区为例[J]. 地理与地理信息科学,2015,31(4):

[9] 安娜. 贝加尔湖旅游资源的保护性开发研究. 沈阳师范大学(C) , 2016

[10] Sergey Kirillov， Natalia Sedova. Problems and prospects for tourism development in the Baikal region. Ecology and Environmental Protection（J）,2014(14):531-538

[11] Каплина Д.В. Туристический потенциал южного побережья Байкалаю.Иркутский аграрный университет,2017:13-18

[12] 钱炜,唐开康旅游产品的不可替代性及其对策研究北京第二外国语学院学报,1994(6)

[13] 朱孔山.旅游产品及其市场营销问题,地域研究与开发,1998(6):80-85

[14] 孙永龙.论旅游市场的季节性特征及应对策略一以甘肃省为例特区经济,2006

[14] Tang Xiaofen. Customer Satisfaction Measurement. Shanghai: Shanghai Science and Technology Press, 2001. [唐 晓芬. 顾客满意度测评. 上海: 上海科学技术出版社, 2001. 1-

[15] 国家质检总局质量管理司，清华大学中国企业研究 中心.中国顾客满意度指数指南. 北

京: 中国标准出版社, 2003. 2-21.

[16]. 刘新燕, 刘雁妮, 杨智 等. 构建新型顾客满意度指数模型. 南开管 理评论, 2003, 5(6): 52-56.

[17] FENG Wei. The analysis of current situation and future development of outbound tourism in China[J]. Economic Geography, 2005, (2): 244-246. [冯玮. 中国出境旅游现状及其未来发展思考[J]. 经济地理, 2005, (2): 244-246.]

[18] SONG Huilin, LYU Xingyang, JIANG Yiyi. The effects of characteristics of tourists on Chinese outbound tourism destination choice behavior: An empirical study based on TPB model [J]. Tourism Tribune, 2016, 31(2): 33- 43. [宋慧林, 吕兴洋, 蒋 依依. 人口特征对居民出境旅游目的地选择的影响——一个 基于 TPB 模型的实证分析[J]. 旅游学刊, 2016, 31(2): 33-43.]

[19] DAI Linlin. Analysis of the impact of crisis events in outbound tourism and its coping strategies[J]. Tourism Tribune, 2011, 26 (9): 8-9. [戴林琳. 出境旅游中危机事件的影响分析及其应对 策略[J]. 旅游学刊, 2011, 26 (9): 8-9.]

[20] XIE Ting. Relevant measures for the safety and security of outbound tourism[J]. Tourism Tribune, 2011, 26(7): 7-8. [谢婷. 出境旅游安全保障的相关措施[J]. 旅游学刊, 2011, 26(7): 7-8.]

[21] YANG Y, WU X. Chinese residents'demand for outbound travel: Evidence from the Chinese family panel studies[J]. Asia Pacific Journal of Tourism Research, 2014, 19(10): 1111-1126.

[22] MOUTINHO L, HUARNG K H, YU H K, et al. Modeling and forecasting tourism demand: The case of flows from Mainland China to Taiwan[J]. Service Business, 2008, 2(3): 219.

[23] CORTÉSJIMÉNEZ I, DURBARRY R, PULINA M, et al. Estimation of outbound Italian tourism demand: A monthly dynamic EC-LAIDS model[J]. Tourism Economics, 2009, 15(3): 547-565. [24] SEETARAM N. Estimating demand elasticities for Australia' s international tourism[J]. Tourism Economics, 2011, 18(5): 9991017.

[25] CHAN F, LIM C, MCALLER M. Modelling multivariate international tourism demand and volatility[J]. Tourism Management, 2005, 26(3): 459-471.

[26] SCHUBERT S F, BRIDA J G, RISSO W A. The impacts of international tourism demand on economic growth of small