Spatiotemporal Dynamics of Chinese Tourists to Lake Baikal Based on Big Data

Wendy Zhang

Abstract: Driven by deepening China-Russia exchanges and visa-free policies, the number of Chinese tourists visiting Lake Baikal has rapidly increased. However, the tourism carrying capacity of the Russian Siberian region where Lake Baikal is located remains relatively weak due to limited tourism development, lagging infrastructure construction, and insufficient supporting service levels. This paper employs big data methods to crawl and analyze tourist reviews and travel blogs from major tourism platforms, combining text analysis, tourism geographical analysis, and GIS visualization techniques to reveal the spatiotemporal distribution characteristics of Chinese tourists visiting the region. Research results indicate that Chinese tourist numbers have significantly increased in recent years, but local tourism reception capacity struggles to meet demand, creating risks of a vicious cycle. This paper suggests improving transportation, optimizing routes, developing Irkutsk's tourism industry, and improving legislation to balance development with environmental protection.

Keywords: data analysis; Lake Baikal; tourism industry; development strategies

1. Research Background

1.1 Overview of Lake Baikal

Lake Baikal is the world's largest and deepest freshwater lake, known as the "Eye of Siberia." Located in the Irkutsk Oblast and Republic of Buryatia in southern Siberia, Russia, the lake features beautiful scenery, unique landscapes, and rich biodiversity. With a total volume of 23.6 trillion cubic meters and a maximum depth of 1,637 meters, it is the world's deepest lake and the largest freshwater lake in Eurasia. The lake stretches 636 kilometers in length, with an average width of 48 kilometers and a surface area of 31,500 square kilometers. Formed by tectonic plate fractures and subsidence, the lake surface sits at approximately 455 meters above sea level.

As Lake Baikal's tourism reputation has grown, increasing numbers of Chinese tourists visit the region. Statistics show that over 36,000 tourists explored the Lake Baikal region in January and February 2019 alone. Tourists traditionally divided their visits between southern and northern routes, but with increasing numbers, traditional tourism routes can no longer satisfy visitor demand. With continuous internet development, tourists habitually post reviews and travel blogs on relevant tourism websites and social media. Based on these reviews and blogs, we can

generally assess Chinese tourists' perceptions of the tourism industry development changes and current status in the Lake Baikal region.

Overall, Chinese tourists evaluate Lake Baikal's scenery positively, but lack evaluation of local infrastructure and tourism experience. According to our preliminary research, infrastructure construction in the Lake Baikal region progresses slowly, with potentially unsatisfactory tourist experiences regarding transportation conditions, network connectivity, and accommodation facilities. This research primarily focuses on whether Lake Baikal's lagging tourism reception capacity restricts Chinese tourist growth and hopes to provide suggestions for the healthy development of Lake Baikal's tourism industry based on research results.

1.2 Research Methods

This research primarily employs big data analysis, using Python to crawl user reviews and travel blogs about Lake Baikal tourism from major Chinese tourism websites, conducting text analysis combined with tourism models and GIS analysis to determine Lake Baikal tourism temporal and spatial changes.

We focus on three main aspects:
(1) Assessment - objective comprehensive evaluation of Lake Baikal tourist attraction evolution;
(2) Changes - analysis of Chinese tourists' spatiotemporal distribution changes in the Lake Baikal region through big data crawling;
(3) Recommendations - suggestions on developing tourism in the Lake Baikal region to meet growing tourist demand.

2. Research Theory and Methods

2.1 Python Data Crawling

Based on existing conditions and network interface situations, considering current text analysis capabilities, and to balance review and blog quantity and quality, this research focuses on five platforms as data source websites: Sina Blog, Ctrip, Qunar, eLong, and Dianping. This research uses a distributed crawler system designed in Python. We searched for short reviews and travel notes on social media by year and month. After crawling, we obtained 1,710 effective reviews from these websites, including 756 Sina blogs, 400 Ctrip reviews and 223 travel short reviews, 160 Qunar travel short reviews, 95 eLong travel short reviews, and 75 Dianping reviews.

Three aspects of analysis were conducted on the crawled data:
(1) Direct content analysis of online texts;
(2) Text analysis using natural language processing tools to crawl and analyze text data from travel websites;
(3) Construction of mathematical models to analyze travel characteristics.

2.2 Tourism Analysis Models

(1) Seasonal Intensity Index

The seasonal intensity index primarily reflects changes in tourism demand intensity in the Lake Baikal region, where demand is determined by the number of reviews collected in the data.

$$R = \sqrt{(\Sigma(xi - \bar{x})^2/12)} / \bar{x} \quad .....(1)$$

Where R represents the seasonal intensity index, $\bar{x}$ is average demand; xi is the annual proportion of Lake Baikal tourism demand for that month. The closer R value approaches zero, the more even the monthly distribution; the higher the R value, the greater the seasonal differences.

(2) Tourism Destination Life Cycle Theory

Tourism destinations have patterns of rise and decline, namely the tourism destination life cycle. In the "involvement" stage, tourist numbers continuously increase due to sufficient tourism facility supply and subsequent advertising promotion. In the "development" stage, tourist numbers increase more rapidly, and control over tourism operations largely transfers from locals to external companies. In the "consolidation" stage, although total tourist numbers continue growing, growth rates have slowed. In the "stagnation" stage, tourist numbers reach their peak, and the destination no longer feels particularly fashionable to tourists. In the "decline" stage, tourists are attracted to new vacation destinations, causing this declining tourism destination to maintain its livelihood through day-trip and weekend visitor visits.

(3) Tourism Demand Forecasting Model

This model evaluates tourism carrying capacity in the Lake Baikal region and makes predictions based on existing data, providing suggestions for tourism planning and construction in the Lake Baikal region.

Tourism carrying capacity developed from environmental carrying capacity, combining multiple factors such as destination service levels and environmental quality for comprehensive evaluation. From field experience, Lake Baikal tourism industry development has not kept pace with its expanding demand, thus we can roughly estimate that tourism carrying capacity in the Lake Baikal region is relatively low, placing higher demands on planning.

3. Research Process and Results

3.1 Text Crawling

We selected five platforms as data source websites: Sina Blog, Ctrip, Qunar, eLong, and Dianping. First, we searched for the keyword "Lake Baikal" on crawled websites to obtain relevant search results. We observed page-turning relationships between search pages to determine relationships

between page URLs, completing basic implementation of search result pagination functionality in crawler code.

Examining search page source code, we found that most websites' search result article URLs could be found in source code, while a few websites used dynamic web pages requiring searches in request results after dynamic JS requests. Since travel blog URL formats are relatively fixed (e.g., Lake Baikal travel blogs on Ctrip uniformly follow "https://you.ctrip.com/travels/lakebaikal4726/*******.html" where "*" represents different parts of each blog), we could use regular expressions to selectively extract next-level URLs from webpage source code, obtaining all travel blog URLs related to Lake Baikal.

Opening a travel blog, we examined webpage content using the browser plugin XPath Helper for assistance. This plugin locates specific elements' XPath positions in webpage HTML text, enabling targeted crawling of main content during the crawling process.

During crawling, we found numerous interference items in crawled main text, primarily various punctuation marks, numbers, and garbled symbols. To make word segmentation results more accurate, we added text filtering functionality to the code, selectively retaining Chinese character portions in main text by judging character Unicode encoding.

Finally, we used Python's jieba library to segment obtained raw text files, completing the text crawling and preprocessing stage.

3.2 Word Frequency Analysis

Through text analysis and word frequency statistics on crawled extensive travel blogs, we obtained the following ten high-frequency vocabulary words:

1. Lake Baikal 2. Olkhon Island 3. Listvyanka 4. Wooden House Museum 5. Kazan Cathedral 6. Karl Marx Street 7. Shaman Rock/Shaman Pillar 8. Yenisei 9. Angara River 10. Khuzhir Town 11. WWII Memorial Square

Specific frequencies are shown in the following table:

Location	Frequency
Lake Baikal	1151
Olkhon Island	368
Listvyanka	249
Wooden House Museum	143
Kazan Cathedral	132
Karl Marx Street	105
Lenin Street	39
Shaman Rock/Pillar	113

Yenisei River	24
Angara River	257
Khuzhir Town	159
WWII Memorial Square	48

Using Gaode Open Platform Map Lab for visualization processing, since some high-frequency vocabulary location names have overlapping coverage areas, we filtered to obtain seven final location names as shown in the following table:

Oblast	Location Name	Frequency
Irkutsk Oblast	Lake Baikal	1151
Irkutsk Oblast	Angara River	263
Irkutsk Oblast	Listvyanka Town	257
Irkutsk Oblast	Wooden House Museum	147
Irkutsk Oblast	Kazan Cathedral	139
Irkutsk Oblast	Lenin Street	41
Irkutsk Oblast	Khuzhir Village	165
Irkutsk Oblast	Shaman Rock Pillar	120
Irkutsk Oblast	WWII Memorial Square	48

We then found each location's latitude and longitude coordinates through Google Maps, noting that Google Maps provides coordinates in degrees, minutes, and seconds that need conversion to decimal form (e.g., 52°31′ converts to 52.5167). Creating charts with location names, frequencies, and coordinates for import into the platform resulted in the visualization shown in Figure 3.

We can see that Lake Baikal region mentioned in travel blogs represents only a broad range without our expected ability to deduce tourism routes from blogs. For Chinese tourists, specific Lake Baikal region attractions concentrate on several islands with relatively few mentioned attractions, suggesting more lakeside recreation activities. According to our field investigation, we encountered many Chinese people at Irkutsk city attractions, but encountered essentially no Chinese people during our research journey. Many locations we visited for research had beautiful scenery but few tourists, even few local tourists, with no Chinese tourists visible. Superior natural scenery suits tourism development, but corresponding peripheral service and leisure facilities are basically absent, with inconvenient transportation. Unfinished dirt roads limit driving speeds and cause significant bumping, with the greater limiting factor being inconvenient network signals—many areas have no signal, let alone smooth 4G networks for navigation, restricting many Chinese tourists' preferred self-driving routes.

This map also enables tourism analysis, clearly showing that Chinese tourists' Lake Baikal touring routes all use Irkutsk as a transit station, giving Irkutsk attractions high word frequency statistics. As an important Far Eastern Russian city and transportation hub, Irkutsk must be traversed by Chinese tourists going to Lake Baikal. Therefore, Irkutsk represents a very important link in Lake Baikal tourism development. Tourism planning processes should focus on Irkutsk's related

supporting facility construction while seizing opportunities to drive Irkutsk city tourism industry development.

3.3 Correlation Analysis

We conducted correlation analysis on extracted words as shown in Figure 4. Each node represents a tourist attraction, connections between nodes represent correlation degrees between different tourist attractions, with larger numbers indicating greater probability of two nodes appearing together in a travel blog.

Correlation analysis effectively distinguishes correlation degrees between two tourist attractions, namely the probability of being visited together. Tourist attraction frequencies in tourism samples reflect tourist preferences for scenic areas, while co-occurrence distributions reflect tourist common preferences for two or more attractions. When tourists' primary desired attractions have other tourist attractions nearby, considering time costs, tourists will visit geographically relatively close tourist attractions, with significant distance barrier effects for attractions far from each other. Lake Baikal, as the core tourist attraction, has extremely strong driving effects on surrounding attraction development. This provides guidance for group tourism route development while offering references for Russian related tourism product development—for example, two highly correlated tourist points can reduce similar souvenir sales to maximize differentiated services.

We found that various tourist attractions in the Lake Baikal region, while connected, are relatively dispersed with considerable distances between attractions, indicating that the Lake Baikal region remains a single-node tourism center without developing into a multi-node tourism center. However, based on our field investigation, we believe natural landscape conditions exist for transitioning from single-node to multi-node tourism centers. Strengthening services and infrastructure construction between different attractions has potential for developing into multi-node tourism centers, driving overall Lake Baikal region tourism industry development. Tourist attractions within Irkutsk city are relatively concentrated, requiring focus on standardized tourism area management.

3.4 Tourism Satisfaction Analysis

Based on text analysis of crawled reviews, we conducted tourist satisfaction analysis through sentiment analysis. This sentiment analysis used Baidu's senta library, with senta's open-source code defaulting to bi-LSTM models. Importing raw text, the senta library automatically executes word segmentation operations, then applies word segmentation results to bi-LSTM models for sentiment analysis. This model includes three layers: word semantic layer, sentence semantic layer, and output layer.

(1) Word semantic layer primarily converts each word in input text to continuous semantic vector representations (word embeddings).
(2) Sentence semantic layer transforms word semantic sequences into entire sentence semantic representations through bi-LSTM network structures.

(3) Output layer calculates sentiment tendency probabilities based on sentence semantics.

Combining tourist ratings, we obtained scores shown in Figure 5 (maximum score of 1):

According to the above sentiment analysis results, most tourists in the Lake Baikal region hold positive attitudes toward tourism experiences, effectively enhancing the Lake Baikal region's reputation. Tourist evaluations of Lake Baikal have remained basically unchanged over recent years without significant changes despite increasing numbers, indicating that Lake Baikal tourism products' core natural landscapes have not experienced significant quality decline due to reception number expansion, suggesting good environmental protection.

Of course, the above sentiment analysis has significant limitations: insufficient crawled review data, people willing to comment online cannot effectively cover the entire population, satisfaction scoring situations cannot be detailed, and other factors all constrain our research. Considering different tourist evaluation systems easily produce contrasting experiences of extreme satisfaction versus extreme disappointment, this represents a shortcoming of this research. Due to Lake Baikal itinerary issues, we could not effectively distribute questionnaires to obtain more precise data. However, due to high scores without significant variance, results have certain credibility and reference value.

Negative reviews mainly concentrated on various negative perceptions during tourism: (1) bumpy car rides (2) ordinary scenery (3) no good hotels (4) extreme cold (5) inconvenient transportation. This provides references for subsequent Lake Baikal tourism industry suggestions.

Based on our actual investigation, we believe the Lake Baikal region's greatest advantage is beautiful natural scenery and diverse natural landscapes, while disadvantages are also obvious: lack of effective and reasonable planning development, insufficient peripheral infrastructure construction, and leisure entertainment service facilities needing improvement.

3.5 Seasonal Intensity Index Changes

We organized numbers for different time periods based on review and travel blog times (Figure 6). Although reviews and blogs have lag effects, lag can be ignored over long time scales. Numbers can reflect changes in tourist numbers visiting Lake Baikal to some degree, enabling seasonal intensity index calculations.

According to the seasonal intensity index formula, we calculated seasonal intensity index values for 2015-2018 (Figure 7). All R values are relatively small, indicating small seasonal changes in Chinese tourist demand for Lake Baikal tourism without obvious peak and off-peak seasons, unaffected by Chinese holidays. The overall declining trend indicates further reduction in seasonal number differences.

3.6 Tourism Life Cycle Assessment

According to tourism life cycle theory, Lake Baikal tourism industry development is currently in the development stage. Through field investigation, we found few large-scale tourism service facilities currently exist, but tourism service facilities remain severely insufficient. During rapid construction processes, considerable problems exist, such as quite unstable and slow network signals, and ongoing road construction requiring mountain blasting obviously impacts local environments.

Tourism carrying capacity developed from environmental carrying capacity, combining multiple factors such as destination service levels and environmental quality for comprehensive evaluation. Field experience indicates Lake Baikal tourism industry development has not kept pace with expanding demand, roughly estimating relatively low tourism carrying capacity in the Lake Baikal region, placing higher planning demands. Simultaneously, tourism basic service facility development may conflict with environmental protection; therefore, construction processes should minimize environmental impacts, avoid irreparable ecological damage, and protect Lake Baikal's unique geographical and geological phenomena and species.

4. Research Conclusions

This research primarily focuses on three aspects:
(1) Assessment - objective comprehensive evaluation of Lake Baikal tourist attraction evolution
(2) Changes - analysis of Chinese tourists' spatiotemporal distribution changes in the Lake Baikal region through big data crawling
(3) Recommendations - suggestions on developing tourism in the Lake Baikal region to meet growing tourist demand

Focusing on the above three research objectives and combining research results, we can conclude:

Increasing numbers of Chinese tourists visit Lake Baikal in recent years, but tourism capacity in the Lake Baikal region cannot satisfy their demands. Lake Baikal is famous for its beautiful scenery, helping it win good reputation in China. However, some problems affecting tourism experiences should be urgently addressed; otherwise, inability to satisfy rapidly increasing tourist demand will cause Lake Baikal, this excellent tourism destination, to enter a vicious cycle and ultimately decline. This represents not only tourist losses but also Russian government losses, causing loss of substantial foreign exchange income while large numbers of people dependent on tourism will face unemployment and significantly reduced labor compensation risks—a fatal blow to prosperous regions around Lake Baikal.

Based on discovered problems, we propose the following suggestions:

(1) Improve road conditions and develop new transportation methods. From field investigation perspectives, existing roads between eastern Irkutsk and Lake Baikal can meet demands due to sparse population and vast land, but Irkutsk city transportation is insufficiently convenient and unfriendly to tourists. Bus systems mix old and new, with some buses lacking English or Chinese announcements, increasing foreign tourist travel difficulties.

(2) Establish new tourism routes and tourism areas based on attraction correlations. Although Lake Baikal is famous for natural scenery and tends toward leisure recuperation, without new tourism route investments, attracting general tourists for repeat visits and spontaneous promotional activities becomes difficult.

(3) Develop Irkutsk tourism industry. Our data analysis reveals Irkutsk as the main transit station for Lake Baikal tourism. With increasing Lake Baikal tourists, clear driving effects on Irkutsk emerge. Reasonable development of Irkutsk tourism industry can create another tourism node, forming regional multi-node tourism patterns with positive promotional effects on regional tourism.

(4) Legislate to promote local tourism development while focusing on environmental protection.** In certain Lake Baikal sections, excessive development has caused unique landform destruction. Excessive garbage residue and environmental damage have also faced local resident opposition, with Chinese enterprises manufacturing mineral water beside Lake Baikal even closing due to civilian pressure.

References

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

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

[3] Liang, Z., & Bao, J. (2012). Research on seasonal characteristics of theme park golden week tourist flow—Taking Shenzhen OCT theme parks as an example. Tourism Tribune, 27(1), 45-52.

[4] He, Y., & Ma, X. (2014). Analysis of causes and mechanisms of inverted "U" structure of tourist flow in Wulingyuan and Huanglong Cave scenic areas. Economic Geography, 34(5), 178-184.

[5] Zhang, T., & Sun, G. (2014). Analysis of tourist flow peak-forest structure and causes in tourism destinations—Comparison of inbound and domestic tourism in Fenghuang, Hunan. Tourism Science, 1, 64-73.

[6] Huang, X., & Ma, X. (2011). Research on tourist activity rhythm based on GPS data. Tourism Tribune, 26(12), 32-37.

[7] Huang, X. (2009). Research on tourists' spatiotemporal behavior patterns in scenic areas based on time geography—Taking Beijing Summer Palace as an example. Tourism Tribune, 24(6), 32-39.

[8] Zhang, Z., Huang, Z., Jin, C., et al. (2015). Research on spatiotemporal behavior characteristics of scenic area tourism activities based on Weibo check-in data—Taking Nanjing Zhongshan Scenic Area as an example. Geography and Geographic Information Science, 31(4), 59-65.

[9] Anna. (2016). Research on protective development of Lake Baikal tourism resources. Shenyang Normal University.

[10] Kirillov, S., & Sedova, N. (2014). Problems and prospects for tourism development in the Baikal region. Ecology and Environmental Protection, 14, 531-538.

[11] Kaplina, D. V. (2017). Tourism potential of the southern coast of Baikal. Irkutsk Agrarian University, 13-18.

[12] Qian, W., & Tang, K. (1994). Research on irreplaceability of tourism products and countermeasures. Journal of Beijing International Studies University, 6, 67-72.

[13] Zhu, K. (1998). Tourism products and their marketing problems. Regional Research and Development, 6, 80-85.

[14] Sun, Y. (2006). On seasonal characteristics of tourism markets and coping strategies—Taking Gansu Province as an example. Special Zone Economy, 3, 156-158.

[15] Tang, X. (2001). Customer Satisfaction Measurement. Shanghai Science and Technology Press.

[16] National Quality Supervision Bureau Quality Management Department & Tsinghua University China Enterprise Research Center. (2003). China Customer Satisfaction Index Guide. China Standards Press.

[17] Liu, X., Liu, Y., & Yang, Z. (2003). Constructing new customer satisfaction index models. Nankai Management Review, 5(6), 52-56.

[18] Feng, W. (2005). Analysis of current situation and future development of outbound tourism in China. Economic Geography, 2, 244-246.

[19] Song, H., Lv, X., & Jiang, Y. (2016). Effects of tourist characteristics on Chinese outbound tourism destination choice behavior: An empirical study based on TPB model. Tourism Tribune, 31(2), 33-43.

[20] Dai, L. (2011). Analysis of crisis event impacts in outbound tourism and coping strategies. Tourism Tribune, 26(9), 8-9.

[21] Xie, T. (2011). Relevant measures for outbound tourism safety and security. Tourism Tribune, 26(7), 7-8.

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

[23] Moutinho, L., Huarng, K. H., Yu, H. K., et al. (2008). Modeling and forecasting tourism demand: The case of flows from Mainland China to Taiwan. Service Business, 2(3), 219-235.

[24] Cortés-Jiménez, I., Durbarry, R., Pulina, M., et al. (2009). Estimation of outbound Italian tourism demand: A monthly dynamic EC-LAIDS model. Tourism Economics, 15(3), 547-565.

[25] Seetaram, N. (2011). Estimating demand elasticities for Australia's international tourism. Tourism Economics, 18(5), 999-1017.

[26] Chan, F., Lim, C., & McAleer, M. (2005). Modelling multivariate international tourism demand and volatility. Tourism Management, 26(3), 459-471.