Operations Research and Fuzziology 运筹与模糊学, 2024, 14(1), 1021-1032
Published Online February 2024 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2024.141095

Hans汉斯

聚类分析与判别分析在智慧旅游中的应用与探索

金婷婷，尚雨浩，陈涵怡

南京信息工程大学数学与统计学院，江苏 南京

收稿日期：2023年12月18日；录用日期：2024年1月8日；发布日期：2024年2月29日

摘　要

近年来，我国旅游业一直保持着高速稳定的发展趋势。在近两年疫情的影响下，旅游业发展受到影响，但我国旅游业目前仍处于持续发展增长的时期。并且，随着信息技术的持续发展，大数据时代已经悄然来临，社会各产业与大数据技术进行了深度的融合，智慧旅游应运而生。由于全国各地区之间差异等各种因素的影响，各个地区的旅游业发展水平呈现出不一致性。本文从Tableau可视化以及智慧旅游概念出发，利用Tableau作图工具对全国近十年的旅游发展趋势进行分析。由于影响智慧旅游发展水平的指标有很多，本文选取2019年全国各省份的具有代表性的若干指标进行研究，即选取旅游总收入、总人次、旅游类居民消费价格指数等9个指标。对于各省份智慧旅游的发展，利用聚类分析和判别分析模型对全国各省份的智慧旅游发展现状进行分析。最后，利用多元统计分析软件SPSS得到分析的结果，将各省份归为不同的类别，寻找原因并给出相应的对策。经研究可得，将全国各省份归为5类。结合各地区智慧旅游的发展现状，给出相应的建议和对策。对于北京、上海发展较成熟的地区，应更加注重高级智能化的旅游产品；对于贵州，由于地区、环境等的原因导致现阶段该区域的智慧旅游相对落后，政府的相关策略应向该地区偏斜。

关键词

智慧旅游，Tableau数据可视化，聚类分析，判别分析

The Application and Exploration of Cluster Analysis and Discriminant Analysis in the Smart Tourism

Tingting Jin, Yuhao Shang, Hanyi Chen

School of Mathematics and Statistics, Nanjing University of Information Science & Technology, Nanjing Jiangsu

Received: Dec. 18th, 2023; accepted: Jan. 8th, 2024; published: Feb. 29th, 2024

文章引用: 金婷婷, 尚雨浩, 陈涵怡. 聚类分析与判别分析在智慧旅游中的应用与探索[J]. 运筹与模糊学, 2024, 14(1): 1021-1032. DOI: 10.12677/orf.2024.141095

Abstract

In recent years, China's tourism industry has maintained a high-speed and stable development trend. Under the influence of the epidemic in the past two years, the development of tourism has been affected, but China's tourism industry is still in a period of development and growth. In addition, with the continuous development of information technology, the era of big data has quietly come, and various social industries and big data technology have been deeply integrated, then smart tourism has emerged at the historic moment. Due to the influence of various factors such as differences among different regions in the country, the level of tourism development in different regions is inconsistent. Starting with the Tableau visualization and the concept of smart tourism, this paper uses the Tableau mapping tool to analyze the development trend of national tourism in the past decade. As there are many indicators affecting the development level of smart tourism, this paper selects several representative indicators of various provinces in 2019, that is, nine indicators such as total tourism revenue, total person-time and tourism consumer price index, etc. For the development of smart tourism in each province, the cluster analysis and discriminant analysis model are used to analyze the development status of smart tourism in each province. Finally, the analyzed results are obtained by using the software SPSS to perform multivariate statistical analysis, classifying the provinces into different categories, finding the reasons and giving corresponding countermeasures. After the study, the provinces in the country were classified into five categories. Combined with the development status of smart tourism in each region, the corresponding suggestions and countermeasures are given. For Beijing and Shanghai, more attention should be paid to advanced and intelligent tourism products; For Guizhou, the smart tourism in this region is relatively backward at present due to regional and environmental reasons, and the government's relevant strategy should be skewed to the region.

Keywords

Smart Tourism, Tableau Data Visualization, Cluster Analysis, Discriminant Analysis

1. 引言

1.1. 研究背景与意义

1.1.1. 研究背景

在我国国民经济飞速发展的背景下，旅游业作为一个极具发展潜力的支柱性产业也得到了很大发展，同时，传统的旅游业体系已经不能满足旅游业信息化程度不断加深的整体需求。国务院在 2009 年出台《关于加快旅游发展的意见》，在此文件中提出，发展建立和健全的旅游信息服务平台，可以更好地促进旅游信息的资源共享，以及可以更好地加快发展旅游行业[1]。自此，我国传统旅游业开始进行产业升级改造，以满足游客个性化需求作为主要目的，将大数据分析技术与旅游业相结合，为打造高满意度的智慧化服务型旅游业提供了新的发展方向，智慧旅游时代正式开启。在当下信息化和大数据的时代背景下，如何利用数据分析手段监测行业发展新趋，并准确分析游客满意度并对景点客流量进行预测，针对游客需求提供个性化服务一直是智慧旅游发展的关键问题。

随着我国社会发展，旅游市场潜力将得到充分释放，信息技术的发展也会成为旅游业发展的潜在推

动力，因此将新一代信息技术与旅游业进行深度融合，使旅游业更加信息化、智能化，这不仅是对人们传统消费方式的改变，更是使景区经营方式、管理模式、资源整合能力得到快速升级。

1.1.2. 研究意义

随着大数据行业越来越被重视，而旅游行业拥有巨大的发展潜力，二者的融合发展将充分发挥大数据分析技术和人工智能的产业优势，实现传统旅游向智慧旅游的产业升级，为游客提供全面的信息化服务的同时，通过发掘近年来旅游大数据，以帮助旅游企业实现完成旅游产业整体智慧化转型的改革与创新。但在发展的过程中，智慧旅游还面临技术不完善、基础设施缺乏，人才缺失等问题。本文则以旅游大数据作为切入点，从数据分析与可视化以及智慧旅游概念入手，分析数据分析与 Tableau 可视化在智慧旅游中的应用。

1.2. 研究目的与方法

1.2.1. 研究目的

此旅游业提出了与信息技术进一步融合的发展战略。在大数据技术支持下的智慧旅游研究可以共享旅游信息资源、深入挖掘旅游数据资源，促进旅游行业固有模式的升级改造[2]。

由于各地区域的环境、气候等的差异，同时，也由于全国各个省份的地理位置、景区数量、景区等级等的不同，导致全国各省份的旅游发展、智慧旅游投入有些微不平衡。因此，本文主要研究全国各省份的智慧旅游发展的趋势及差异。可根据模型结果给出相对应的建议以及改进措施。

1.2.2. 研究方法

本文从基本概念入手，分析数据分析与 Tableau 可视化在智慧旅游中的应用，并利用数据可视化技术与聚类分析对智慧旅游发展现状进行分析，最后利用判别分析选择合适的判别规则，研究对聚类分析的结果进行分类判别。深入探究数据分析与可视化目前在智慧酒店的应用过程，并根据聚类分析和判别分析的研究结果对全国各省份的相关问题提出相应的对策，使各个省份更好的发展与推进智慧旅游，同时也推进数据分析与可视化在智慧旅游发展中的应用。

1.3. 国内外研究现状

“智慧旅游”的概念是在 IBM 首先“智慧世界”之后出现的，智慧旅游将是传统旅游业发展进步的必经之路。尽管研究人员已经定义了该术语，但仍未形成社会共识[2]。Yunpeng Li [3]等人通过比较了传统旅游信息服务与智慧旅游相关服务的特点，并基于这些假设提出了建议并讨论了未来的研究方向。在中国旅游市场，智慧旅游代表着一条新路径，对旅游企业和游客本身产生重大影响。作为一项社会科学，旅游研究已成为多学科研究的重点，并融入了许多不同的研究领域。Smart Tourism Research (STR)寻求更好的设计和流程、有效的方法和有效的资源管理，在综合旅游者、旅游业和酒店业的过程中，采取有效的资源管理方法。基于智能旅行体验，多样化技术(如应用技术、AI、数据分析、可视化)已成为现实，并在创造更高水平的社会价值和经济影响方面发挥着重要作用[4]。陈建民和徐苏莉[5]利用 Python 软件采集旅游者行为数据并构建数学模型，进行大规模智能数据分析数据建模。

2. 相关概念与基础理论

2.1. 智慧旅游

旅游业作为我国服务业的支柱之一，在我国经济发展中，旅游业作为我国服务业的支柱之一一直发挥着重要作用。在推动旅游业可持续发展的过程中，智慧旅游是旅游业发展的重要方向，大数据技术的合理应用对于促进有效的智慧旅游管理、提升旅游体验、提升个性化服务具有重要作用[6]。智慧旅游所

需要的挖掘的数据包括着视频、图片、评价等大量的非结构化数据，有效的筛选分析并形成科学的结果至关重要[7]。智慧旅游中，数据分析与可视化要求将旅游大数据形成分析语言以及对海量非结构数据特征识别模型的构建与分析[8]。

智慧旅游的研究最早是由 IBM 公司提出的“智慧地球”概念，而后拓展应用到了旅游行业，这就是智慧旅游的开始[9]。所谓智慧旅游，通常意义上是指在一定程度上利用云计算、结合大数据、结合物联网等目前新型先进的信息科学技术，而后通过互联网，接着借助便携智能高效的终端，然后自发主动的去感知一些旅游资源、经济、活动、以及旅游者等相关方面的相关信息。在这之后，通过及时发布，可以使人们及时的了解到这方面的相关信息，而后根据信息可以及时的安排、调整相关的工作以及对应的旅游计划，这也就达到了很好的利用智能化旅游的效果[10]。

社会经济发展的今天，人们的生活水平得到了很大的提升。而智慧旅游利用先进的信息技术，促进旅游景区的转型创新，并通过旅游大数据准确分析旅游者个性化需求，提供个性化服务。智慧旅游的不断深化将极大的解决传统旅游管理模式下人力成本投入大、管理效率低下、管理质量参差不齐等一系列问题，在旅游景区智能化设备的引入中将形成基于信息技术的新型管理模式，实时监控景区运营状况，全面提升旅游服务质量，及时发现并解决问题[11]。

2.2. 聚类分析

所谓聚类分析，是指将物理或抽象对象，将类似的对象分为一类，组成的多个类的集合分组的分析过程。通俗意义上，聚类分析可以用“物以类聚”这一简单的词来表达，是一种现代统计的分析方法，在众多领域中，都可以运用聚类分析作为分类研究的方法，例如对有相关数据的样品进行分类处理，以及变量进行分类处理。聚类分析可以分为系统聚类分析和 K-means 分析(K-均值聚类分析)。其中，本文选择系统聚类分析。系统聚类分析的基本思想是：先将样品或变量距离相近的聚成类，距离相差远的后聚成类，一直到所有的样品或变量聚成一类为止。假设共有 n 个样品，第一步先将每一个样品聚成一类，即每个样品自成一类，共有 n 类；第二步确定距离公式，一般用欧式距离，再根据距离公式将距离最近的两个样品聚成为一类，共聚成 $n-1$ 类；第三步是将距离最近的两个类聚成一类，一共聚成 $n-2$ 类，……，重复进行以上步骤，以所有的样品都聚成一类为止。

步骤一：聚类分析首先应计算距离矩阵 $\boldsymbol{D}$：

$$\boldsymbol{D}=\begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{pmatrix}=\begin{pmatrix} 0 & d_{12} & \cdots & d_{1n} \\ d_{21} & 0 & \cdots & d_{2n} \\ \vdots & \vdots & & \vdots \\ d_{n1} & d_{n2} & \cdots & 0 \end{pmatrix} \tag{1}$$

步骤二：离差平方和法使得两个大的类倾向于有较大的距离，因而不易合并；相反，两个小的类由于倾向于有较小的距离，从而易于合并。实际运用到聚类分析中可将 n 个样品分成 t 类 $P_1,P_2,\cdots,P_t$，用 X_{iK} 表示 G_K 中的第 i 个样品，n_K 表示中样品的个数，$\overline{X}_K$ 是 G_K 的重心，则 G_K 的样品离差平方和为：

$$S_t=\sum_{i=1}^{n_K}\left(X_{iK}-\overline{X}_K\right)'\left(X_{iK}-\overline{X}_K\right) \tag{2}$$

2.3. 判别分析

而所谓判别分析，是在分类确定的前提下，根据某一研究对象的各种特征值判别其类型归属问题的一种多变量统计分析方法。同时，判别分析也可称为“分辨法”。

采用 Fisher 判别分析 Fisher 判别法借用了一元方差分析的思想，通过将多维数据投影到某个方向上，

投影的原则是将总体与总体之间尽可能的分开，然后选择适当的判别原则，将新的样品进行分类判别[12]。设有 n 个总体 $P_1,P_2,\cdots,P_n$，其均值为 μ_i，协方差矩阵为 $\sum_i(>0)$，其中 $i=1,2,\cdots,n$，对于新样品 X，判断其来自哪个总体，并借助方差分析的思想，进一步构造一个线性判别函数：

$$Y(X)=u_1X_1+u_2X_2+\cdots+u_pX_p=u'X \tag{3}$$

其中，系数 $u=(u_1,u_2,\cdots,u_p)'$ 确定的原则目的是使得总体之间区别尽可能的大，同时使总体内部的离差尽可能的小。

步骤一：设从 n 个总体分别取得 n 组维观察值：

$$V_1:X_1^{(1)},\cdots,X_{n_1}^{(1)} \tag{4}$$

$$\vdots$$

$$V_n:X_1^{(n)},\cdots,X_{n_k}^{(n)} \tag{5}$$

其中

$$n=n_1+n_2+\cdots n_k \tag{6}$$

步骤二：设 u 为任一 p 维向量，$Y(X)=u'X$ 为 x 向 u 为法线方向的投影，由此，上述数据投影为：

$$V_1:u'X_1^{(1)},\cdots,u'X_{n_1}^{(1)} \tag{7}$$

$$\vdots$$

$$V_n:u'X_1^{(n)},\cdots,u'X_{n_k}^{(n)} \tag{8}$$

步骤三：组成一组方差分析数据，其组间平方和及其组内平方和分别为：

$$SSA=\sum_{i=1}^{k}n_i\left(u'\overline{X}^{(i)}-u'\overline{X}\right)^2=u'\left[\sum_{i=1}^{k}n_i\left(\overline{X}^{(i)}-\overline{X}\right)\left(\overline{X}^{(i)}-\overline{X}\right)'\right]u=u'Nu \tag{9}$$

$$SSR=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(u'X_j^{(i)}-u'\overline{X}^{(i)}\right)^2=u'\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(u'X_j^{(i)}-u'\overline{X}^{(i)}\right)\left(X_j^{(i)}-\overline{X}^{(i)}\right)'\right]u=u'Ru \tag{10}$$

其中 N 是组间平方和及交叉乘积和，R 为组内平方和及交叉乘积和。

步骤四：依据两个重心距离越大越好，两个组内的离差平方和越小越好的原则，找到使得 $\Phi(u)$ 值最大的 u 值，得到最佳的线性判别函数。若组均值有显著差异，则 F 应充分大，或者 $\Phi(u)$ 应充分大，其中公式为：

$$F=\frac{SSA/(k-1)}{SSR/(n-k)}=\frac{n-k}{k-1}\cdot\frac{u'Nu}{u'Ru}=\frac{n-k}{k-1}\Phi(u) \tag{11}$$

3. 国内旅游发展现状

3.1. 数据来源

数据来源为：国内旅游相关数据来源于国家统计年鉴和中华人民共和国文化和旅游部中的统计信息(网址：http://zwgk.mct.gov.cn/zfxxgkml/447/465/index_3081.html)，未包含港澳台地区的旅游数据。智慧旅游属于旅游业的前进发展的时代产物，经济及高新技术的快速发展使旅游业向智慧旅游转化，分析智慧旅游的发展离不开分析旅游业的发展，近年来，旅游业已加入高新技术的使用，如智能感知信息、个性化游记、景区实时动态、智能语音解说等，从而旅游发展趋势反映出智慧旅游的发展。因此，找出并汇

总了近 10 年的旅游相关数据，即 2012~2021 年，选取其中可以直接分析旅游发展现状及趋势的相关指标的数据，其中，包括了国内旅游总人次(百万)、国内旅游收入(亿)、人均每次旅游消费(元)、城镇居民出游人次(百万)、农村居民出游人次(百万)、城镇居民旅游消费(亿)、农村居民旅游消费(亿)、城镇居民人均每次旅游消费(元)和农村居民人均旅游消费(元)，得到的具体相关数据如下表 1 所示。

Table 1. Domestic tourism data
表 1. 国内旅游数据

年份	总人次	总收入	人均每次消费	城居民出游数	农居民出游数	城居民消费	农居民消费	城人均次消费	农人均次消费
2012	2957	22706.2	767.9	1933	1024	17678	5028.2	914.5	491
2013	3262	26276.1	805.5	2186	1076	20692.6	5583.5	946.6	518.9
2014	3611	30311.9	839.7	2483	1128	24219.8	6092.1	975.4	540.2
2015	3990	34195.1	857	2802	1188	27610.9	6584.2	985.5	554.2
2016	4435	39389.8	888.2	3195	1240	32241.9	7147.9	1009.1	576.4
2017	5001	45660.8	913	3677	1324	37673	7987.7	1024.6	603.3
2018	5539	51278.3	925.8	4119	1420	42590	8688.3	1034	611.9
2019	6006	57250.9	953.3	4471	1535	47509	9741.9	1062.6	634.7
2020	2879	22286.3	774.14	2065	814	17966.5	4319.8	870.25	530.47
2021	3250	29191	899.28	2340	900	23644	5547	1009.57	61.56

3.2. 国内旅游现状可视化分析

根据 3.1 汇总的数据，用 Tableau 数据可视化工具分析近 10 年的国内旅游收入和国内旅游总人次等相关指标，得到填充气泡图和旅游趋势图，如图 1 填充气泡图和图 2 旅游趋势图所示。其中，填充气泡图是一种数据可视化工具，又称为填充型气泡散点图，结合了传统散点图和柱状图的特点，通过填充颜色来表示不同的数据类别或数据大小；趋势图用于展示数据随时间的变化趋势，在一张图上展示多个数据的趋势，可以直观地比较它们之间的差异和相似性，从而更好的把握旅游的发展趋势。

根据气泡图和趋势图的大小，可以直观地分析国内旅游收入和国内旅游总人次的趋势，从而可以得出国内旅游发展的现状，同时也为后文的数据选取提供依据。

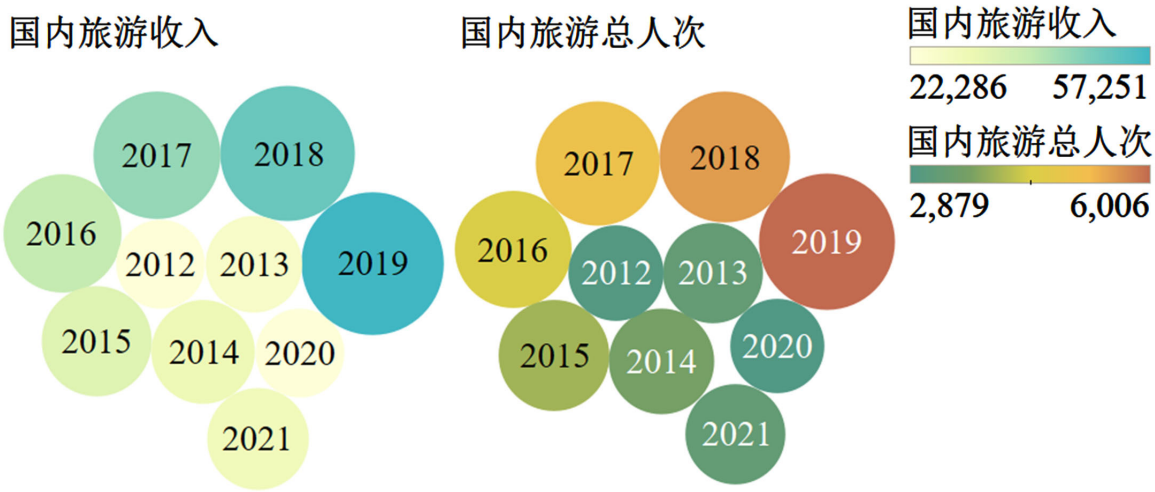

Figure 1. Filled bubble chart
图 1. 填充气泡图

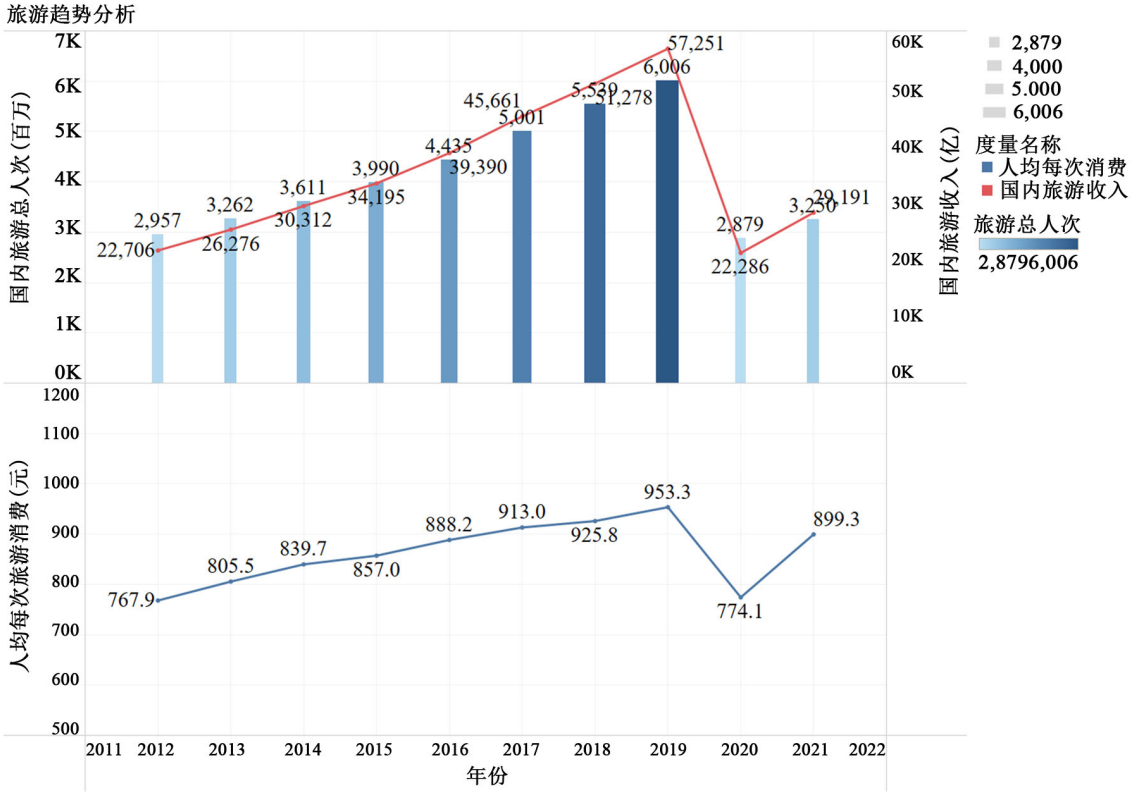

Figure 2. Tourism trend graph
图 2. 旅游趋势图

由填充气泡图和旅游趋势图可知，总体上，近十年来旅游呈现上升趋势。具体地，2012 年至 2019 年旅游呈现快速上升趋势，每年的旅游总人次和旅游收入显著增加；但 2020 年旅游趋势断崖式下降，2021 年较 2020 年有所上升，但仍与 2019 年及之前年份有很大差距，综合来看，短期内难以恢复至 2019 年旅游的发展状态。

结合现实情况分析，近年来国内经济快速发展，同时促使旅游业也欣欣向荣，但 2019 年年底疫情的突然爆发打破了旅游快速发展的趋势。由于疫情原因，自 2020 年以来，旅游数据急剧减少。剔除客观的疫情原因，结合相关的旅游发展研究的文献资料，旅游业发展会呈现上升趋势，用 2020 年至 2021 年的旅游数据研究旅游发展现状并不合适。因此，本文主要以 2019 年的全国各省份的旅游数据为支撑，研究全国各省份智慧旅游之间的差距对比。

4. 聚类分析和判别分析

4.1. 数据处理

本部分数据来源于中华人民共和国文化和旅游部的发展统计公报及国家统计局(文化和旅游发展统计公报的网址为：https://zwgk.mct.gov.cn/zfxxgkml/tjxx/202206/t20220629_934328.html；国家统计局网址为：https://www.stats.gov.cn/sj/ndsj/2021/indexch.htm)，未包含港澳台地区的旅游数据，因此，经过整理，得到中国(除港澳台地区)各省份在 2019 年的旅游数据，由于数据过多，此处不再列出。

对数据进行预处理，数据标准化处理过程：建立数据矩阵，设论域 $W=\{a_1,a_2,\cdots,a_n\}$ 为被分类对象，每个对象又由 m 个指标表示其形状：

$$x_i = \{a_{i1}, a_{i2}, \cdots, a_{im}\} (i = 1, 2, \cdots, n) \tag{12}$$

则得到原始数据矩阵为：

$$X = \left(a_{ij}\right)_{n \times m} \tag{13}$$

对于实际的问题，不相同的数据一般会有不相同的量纲，为了使有不同量纲的变量能进行比较，用标准差标准化将数据规格化[13]。本文中各省份的相关数据由于各指标的计量单位不同，故在聚类之前将原始数据进行标准差标准化处理。标准差标准化即是：对于第 i 个变量进行标准化，将 a_{ij} 换成 $a_{ij}^{'}$，即

$$a_{ij}^{'} = \frac{a_{ij} - \overline{a}_i}{S_i} (1 \le j \le m) \tag{14}$$

式中：

$$\overline{a}_i = \frac{1}{m}\sum_{j=1}^{m} a_{ij}, S_i = \sqrt{\frac{1}{m-1}\sum_{j=1}^{m}\left(a_{ij} - \overline{a}_i\right)^2} \tag{15}$$

具体到本文问题中，可知，全国各省份为被分类对象，即：$a_1, a_2, \cdots, a_n$ 分别指代北京、天津、河北、山西等 31 个省份。选取指标变量：旅游总收入(亿元)、旅游总人次(亿)、接待外国人游客(百万人次)、接待国际游客(百万人次)、居民人均可支配收入(元)、居民人均消费支出(元)、旅游类居民消费价格指数(上年 = 100)、互联网宽带接入用户(万户)、恩格尔系数(%)，即：$m = 9$。对原始数据进行标准差标准化处理之后，得到描述统计结果如下表 2 所示。

Table 2. Statistical table
表 2. 统计表

	N	最小值	最大值	均值	标准 偏差
旅游总收入	31	340.03	15200.00	7220.1387	4010.77162
接待外国游客	31	0.040	8.570	2.08790	1.999830
居民人均消费支出	31	7028.0	45605.0	21145.690	8067.1241
接待国际游客	31	0.07	37.31	3.9265	6.56559
旅游总人次	31	0.001257	75081.580000	2427.08101968	13484.10474941
居民人均可支配收入	31	19501.0	69442.0	30563.619	12380.7307
旅游类居民消费价格指数	31	95.2	106.5	101.494	2.7450
互联网宽带接入用户	31	174.54	4063.10	1597.3239	1069.96679
恩格尔系数	31	22.32	38.02	29.2468	3.50236
有效个案数(成列)	31				

4.2. 聚类分析

针对本文的具体问题研究，为了之后的判别分析，并为了判别分析与聚类分析的结果进行比较，本文选取 4 个省份(江苏、广东、四川、宁夏)为待判组，即被选取为待判组的江苏、广东、四川、宁夏 4 个省份不参与系统聚类。下面对剩下的 27 个省份进行系统聚类分析，利用统计分析软件 SPSS 读取已预处理的数据的原始数据矩阵，进行系统聚类分析，得到谱系图及进行分类，如下图 3 谱系图所示。

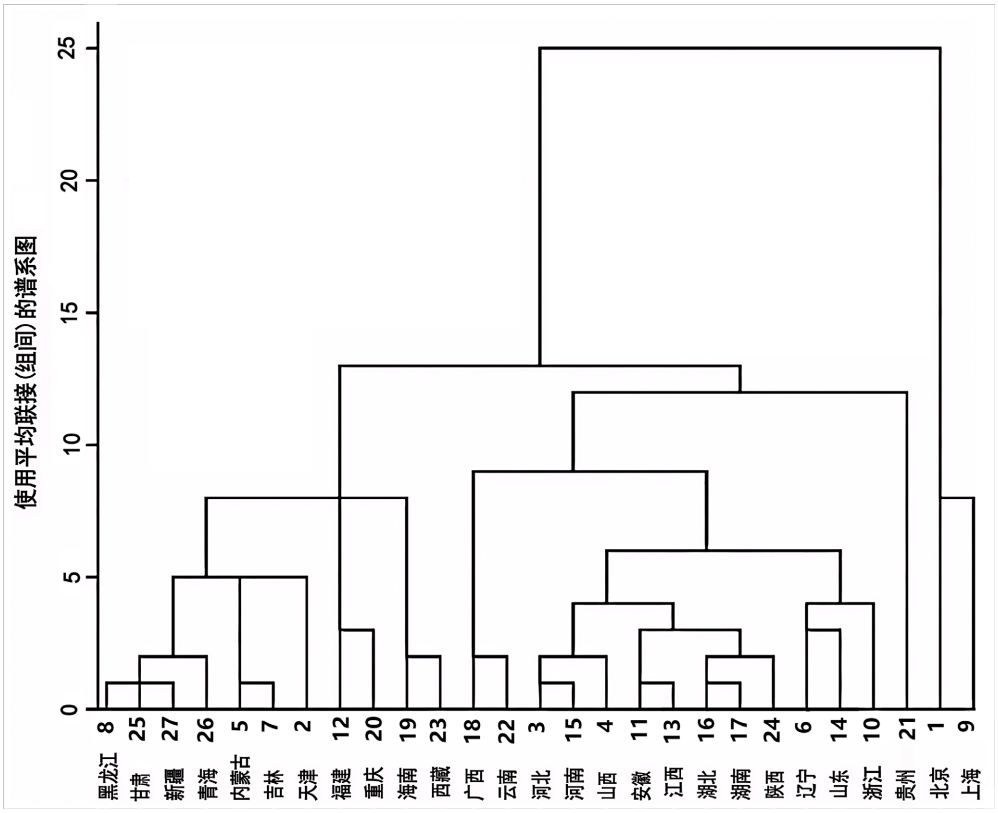

Figure 3. Pedigree chart
图 3. 谱系图

由谱系图得到分类结果，如下表 3 分类结果所示。

Table 3. Classification results
表 3. 分类结果

北京	天津	河北	山西	内蒙古	辽宁	吉林	黑龙江	上海
1	4	5	5	4	5	4	4	1
浙江	安徽	福建	江西	山东	河南	湖北	湖南	广西
5	5	4	5	5	5	5	5	3
海南	4 重庆	贵州	云南	西藏	陕西	甘肃	青海	新疆
4	4	2	3	4	5	4	4	4

对聚类结果进行分析，由分类结果表可知，可以将上述 27 个省份分为 5 类：

第 1 类：北京、上海；

第 2 类：贵州；

第 3 类：广西、云南；

第 4 类：天津、内蒙古、黑龙江、吉林、海南、重庆、西藏、福建、甘肃、青海、新疆；

第 5 类：河北、山西、辽宁、浙江、安徽、江西、山东、河南、湖北、湖南、陕西。

4.3. 判别分析

经过聚类分析之后，27 个省份被划分为 5 类。对于 9 个指标变量分别设为 X_1、X_2、X_3、X_4、X_5、

X_6、X_7、X_8、X_9，即：旅游总收入(亿元)为X_1、旅游总人次(亿)为X_2、接待外国人游客(百万人次)为X_3、接待国际游客(百万人次)为X_4、居民人均可支配收入(元)为X_5、居民人均消费支出(元)为X_6、旅游类居民消费价格指数(上年 = 100)为X_7、互联网宽带接入用户(万户)为X_8、恩格尔系数(%)为X_9。现将剩下没有参与聚类分析的 4 个省份进行判别分析，所以需建立判别函数来进行类别间的判定，判别函数建立如下：

$$Y = r_0 + r_1X_1 + r_2X_2 + r_3X_3 + r_4X_4 + r_5X_5 + r_6X_6 + r_7X_7 \tag{16}$$

观测样本的数据是构建判别函数的关键所在，由此来确定判别函数的系数r_j，运用 SPSS 进行运算分析[14]。检验各组协方差阵是否相等，输出结果如下表 4 检验结果所示。

Table 4. Test results
表 4. 检验结果

博克斯 M		229.286
F	近似	2.367
	自由度 1	45
	自由度 2	1159.699
	显著性	0.000

根据输出结果可知，检验协方差阵相等的 Boxs’ M 值为 229.286。所以，在 0.05 显著性水平下，认为各组协方差阵没有显著性差异。并且，根据 F 检验的显著性概率，说明判别是显著的，也就是判别出错的可能性很小。

利用 SPSS 依次来求解判别函数的系数，得到典型判别式函数系数，如表 5 标准化的典型判别式函数系数所示。

Table 5. Coefficients of the discriminant function
表 5. 判别函数系数

分类(组别)	1	2	3	4	5
旅游总收入	−0.005	0.002	0.001	0.001	$9.226*10^{-5}$
旅游总人次	33.377	32.205	29.967	28.737	32.421
接待外国游客	53.067	54.692	53.359	49.668	54.515
接待国际游客	−43.423	−49.806	−45.565	−43.966	−47.465
居民人均可支配收入	−0.012	−0.017	−0.017	−0.018	−0.017
居民人均消费支出	0.016	0.020	0.020	0.021	0.020
旅游类居民消费价格指数	50.389	53.090	54.085	53.675	55.253
互联网宽带接入用户	0.017	0.027	0.029	0.028	0.031
恩格尔系数	26.931	30.561	30.525	29.755	30.725
(常量)	−2865.239	−3175.468	−3245.654	−3183.358	−3383.590

输出结果表 5 是每组的分类函数的系数，也就是费希尔判别函数的系数。据此可以写出费希尔判别函数：

第 1 类：

$$\begin{aligned} Y_1 = &-2865.239 - 0.005X_1 + 33.377X_2 + 53.067X_3 - 43.423X_4 - 0.012X_5 \\ &+0.016X_6 + 50.389X_7 + 0.017X_8 + 26.931X_9 \end{aligned} \tag{17}$$

第 2 类：

$$Y_2 = -3175.468 + 0.002X_1 + 32.205X_2 + 54.692X_3 - 49.806X_4 - 0.017X_5 + 0.020X_6 + 53.090X_7 + 0.027X_8 + 30.561X_9 \tag{18}$$

第 3 类：

$$Y_3 = -3245.654 + 0.001X_1 + 29.967X_2 + 53.359X_3 - 45.565X_4 - 0.017X_5 + 0.020X_6 + 54.085X_7 + 0.029X_8 + 30.525X_9 \tag{19}$$

第 4 类：

$$Y_4 = -3183.358 + 0.001X_1 + 28.737X_2 + 49.668X_3 - 43.966X_4 - 0.018X_5 + 0.021X_6 + 53.675X_7 + 0.028X_8 + 29.755X_9 \tag{20}$$

第 5 类：

$$Y_5 = -3383.590 + 9.226*10^{-5}X_1 + 32.421X_2 + 54.515X_3 - 47.465X_4 - 0.017X_5 + 0.020X_6 + 55.253X_7 + 0.031X_8 + 30.725X_9 \tag{21}$$

模型结果分析：在观察值分组的时候，将每一个观测值代入 5 个组的分类函数，以函数值的大小来做比较，哪一组的分类函数值大，就将该观测值判入该组。经计算可得：可以将江苏归为第 5 类，将广东归为第 4 类，将四川归为第 1 类，将宁夏归为第 3 类。

5. 结论

通过 Tableau 可视化分析可知，排除 2020 年以来疫情的影响，近年来全国旅游趋势稳定且大幅度上升。在国家统计局、国家统计年鉴以及各省份国民经济和社会发展统计公报查找相关的智慧旅游数据，运用聚类分析和判别分析将全国各省份归为 5 类。结合各地区智慧旅游的发展现状，给出相应的建议和对策。对于北京、上海发展较成熟的地区，应更加注重高级智能化的旅游产品；对于贵州，由于地区、环境等的原因导现阶段该区域的智慧旅游相对落后，政府的相关策略应向该地区偏斜。

运用聚类分析和判别分析将全国各省份归为 5 类。根据各地区的实际情况，以及旅游发展的现状可知，聚类分析的结果有很大的可行性。其中，北京和上海为一类，处于领头的位置。北京是现代化国际城市，作为中国经济中心第三产业发展规模也居于全国领先地位。北京的景区数量、景区环境及旅游智能化程度都处于先列，是最重要的旅游城市。上海作为全国经济发达的城市，地理位置很优越，交通便利，是经济最发达的城市之一，如此各种因素都带动上海的旅游基础设施建设和旅游的智能化发展。广西、云南和宁夏归为一类，该地区的旅游发展水平良好，还有很大的提升改进空间。天津、吉林、广东等比较发达的城市为一类。江苏、浙江、湖北等发展趋势迅猛的城市为一类。这两类地区的旅游发展形式较好，处于稳速上升的阶段，旅游到智慧旅游的转型也很前进。下面重点分析贵州这一类别。

通过数据对比可知，贵州的互联网宽带接入量较其他省份差距很大，说明该地区的智能化发展较为落后，在网络、智能化管理等方面仍需要很大的改进和发展。另外，就贵州的历史和地理位置而言，其地理位置偏僻，且相较于其他城市，该地区的景区、博物馆等可观光游览的旅游区较少。也因其地理位置等的原因，该地区的综合实力及经济发展水平仍处于比较落后的水平，该地的智慧旅游较为欠缺。因此，政府应重视该地区的发展状况，给予一定的政策支持。

总的来说，聚类分析和判别分析的结果符合各地区的实际情况。所以，从一定程度上来说，通过聚类分析和判别分析对全国各省份的智慧旅游发展情况进行归类的结果是有一定的参考意义的。因此，可以根据本文的分析结果给出相应的政策调整。

参考文献

[1] 卢慧敏. 数据分析在智慧旅游中的应用研究[J]. 劳动保障世界, 2017(29): 63-69.

[2] 王强进. 基于大数据分析的智慧旅游研究[D]: [硕士学位论文]. 长春: 长春工业大学, 2021.

[3] Li, Y.P., Hu, C., Huang, C. and Duan, L.Q. (2016) The Concept of Smart Tourism in the Context of Tourism Information Services. *Tourism Management*, **58**, 293-300. https://doi.org/10.1016/j.tourman.2016.03.014

[4] Koo, C., Park, J. and Lee, J.-N. (2017) Smart Tourism: Traveler, Business, and Organizational Perspectives. *Information & Management*, **54**, 683-686. https://doi.org/10.1016/j.im.2017.04.005

[5] 陈建敏, 徐苏丽. 基于人工智能的智慧旅游大数据分析模型的构建[J]. 电脑知识与技术, 2019, 15(11): 189-190.

[6] 陈胜花. 基于大数据时代的智慧旅游开发策略探究[J]. 旅游纵览(下半月), 2020(2): 18-19.

[7] 郭珂. 智慧旅游中大数据的应用研究[J]. 旅游纵览(下半月), 2018(12): 15-16.

[8] 吴星星. 我国智慧旅游研究热点可视化分析——基于 CNKI 核心期刊载文数据[J]. 湖北文理学院学报, 2021, 42(2): 35-40.

[9] 张赞. 基于大数据的智慧旅游系统设计与实现[D]: [硕士学位论文]. 沈阳: 东北大学, 2016.

[10] 刘全才, 牛牧原, 刘秋雨, 梁瀚余, 张思纪. 数据分析与可视化在智慧旅游中的探索与应用[J]. 产业科技创新, 2020, 2(17): 44-45.

[11] 杨静. 关于大数据在智慧旅游管理中的应用[J]. 旅游纵览(下半月), 2019(22): 52-53.

[12] 陈龙. 基于判别分析的家庭外出旅游动机影响因素探究[J]. 攀枝花学院学报, 2014, 31(3): 105-107.

[13] 祝新亚, 李许坚. 基于聚类分析和判别分析的我国主要省市综合实力状况评价[J]. 北方经济, 2011(8): 16-18.

[14] 丁柳, 刘艳华. 我国各地区经济发展水平的实证分析——基于聚类分析及判别分析法的应用[J]. 赤峰学院学报(自然科学版), 2017, 33(15): 143-145.