DOI:10.19715/j.tiyukexueyanjiu.2024.04.003

美国户外休闲产业发展特征及预测模型构建研究

——基于PLS和PLS-DA方法的分析

雷 雯[1],魏德样[2]
(1.闽江学院公共体育教学部,福建 福州 350108;2.福建师范大学体育科学学院,福建 福州 350117)

摘 要:户外休闲产业是美国经济的支柱性产业,分析其发展特征和构建预测模型对于我国健身休闲产业健康、可持续发展有重要借鉴价值。研究收集了美国50个州2021年户外休闲产业相关数据,选取27个影响指标,运用PLS和PLS-DA方法,对美国户外休闲产业发展特征进行分析并构建预测模型。研究表明:(1)美国州域户外休闲产业可分为4种发展模式,即总量大占比小型、总量小占比大型、总量较大占比较小型、总量较小占比较大型,4种发展模式在空间上呈现一定集聚分布特征。(2)构建的预测模型有效且精度较高,并筛选出12个VIP指标,经济、人口和社会因素似乎更多地影响美国户外休闲产业发展,而自然环境因素的影响相对较弱。

关键词:户外休闲产业;偏最小二乘法;偏最小二乘判别分析;美国
中图分类号:G 80-05 **文献标识码**:A **文章编号**:1007-7413(2024)04-0017-07

Research on the Development Characteristics and Prediction Models of American Outdoor Leisure Industry
——Analysis Based on PLS and PLS-DA Methods

LEI Wen[1], WEI Deyang[2]
(1. Department of Physical Education, Minjiang University, Fuzhou 350108, China;
2. School of Physical Education and Sport Science, Fujian Normal University, Fuzhou 350117, China)

Abstract: Outdoor recreation is a pillar industry of the American economy. Analyzing its development characteristics and constructing prediction models will have important reference value for the healthy and sustainable development of China's fitness and leisure industry. We collected data on the outdoor leisure industry in the 50 states of the United States in 2021, selected 27 impact indicators, and used the PLS and PLS-DA methods to analyze the development characteristics of the outdoor recreation in the United States and build predictive models. Studies show that: (1) The U.S. state outdoor recreation can be divided into 4 development modes, namely, high total amount, low proportion; low total amount, high proportion; total amount relatively high, proportion relatively low; total amount relatively low, proportion relatively high. The four development models present certain spatial distribution characteristics. (2) The prediction model is effective and highly accurate, and 12 VIP indicators are screened out; economic, demographic, and social factors seem to affect the development of outdoor recreation in the United States more, but the impact of natural environmental factors is relatively weak.

Key words: outdoor recreation; partial least squares method; partial least squares discriminant analysis; United States

2016年,《国务院办公厅关于加快发展健身休闲产业的指导意见》(国办发〔2016〕77号)中提出,到2025年,形成布局合理、功能完善、门类齐全的健身休闲产业发展格局,产业总规模达到3万亿元的发展目标[1]。为实现上述目标,文件中提出要大力普及日常健身(如徒步、骑行、钓鱼等),发展户外运动(如冰雪、山地、水上、汽车摩托车、航空等)[1]。上述项目与美国户外休闲项目高度重合。据美国

收稿日期:2023-06-10
基金项目:2023年度福建省社会科学基金项目"福建省公共体育设施与人口空间匹配研究"(FJ2023A004)
第一作者简介:雷雯(1977—),女,福建建阳人,副教授。研究方向:体育产业。

户外产业协会统计,美国开展的户外休闲项目主要有露营、垂钓、狩猎、摩托车运动、汽车越野、雪上运动、水上运动、越野跑、自行车及滑板运动、野生动物观赏等十大类[2]。由于美国属于联邦制国家,各州拥有较大的自主权,使得其在经济、社会发展方面存在不同方式。同时,美国幅员辽阔,州域间在自然环境等方面也存在较大差异,因此在评价美国户外休闲产业发展状况时,还要考虑经济、社会、自然环境等因素。

分析美国户外休闲产业发展特征及预测模型构建需选取多个指标形成指标体系,这必然带来指标间的多重共线性与多维等问题,偏最小二乘法(Partial Least-Squares,PLS)能够解决上述问题[3],其在化工、机械、生物、医学、社会学以及经济学等领域已得到较为广泛应用。PLS 方法与判别分析(Discriminant Analysis,DA)方法联合使用,产生了一种多变量统计分析方法,即偏最小二乘判别分析 (Partial Least-Squares Discriminant Analysis, PLS-DA),是模式识别的重要工具[4]。本文运用 PLS 和 PLS-DA 方法,对美国州域户外休闲产业发展模式进行识别,并分析其特征,构建出美国户外休闲产业发展的预测模型,旨在为我国健身休闲产业健康、可持续发展提供理论参考。

1 研究方法

1.1 PLS 的建模思路

设有 q 个因变量和 p 个自变量,n 个样本点。PLS 的建模思路是分别在 X 与 Y 中提取出成分 t_1 和 u_1,并且要求两者尽可能多地携带 X 与 Y 的变异信息和相关程度能够达到最大,这样既保证 t_1 和 u_1 可以很好地代表数据表 X 和 Y,也能使自变量的成分 t_1 对因变量的成分 u_1 有很强的解释能力[4-5]。

1.2 PLS 变量投影重要性指标

自变量 x_j 在解释因变量集合 Y 时,可用 VIP(Variable Importance in Projection,VIP)指标来反映其作用的重要性。当所有的自变量在解释 Y 时的作用相同时,则 VIP_j 都等于 1。否则,VIP_j 大于 1 的自变量对解释 Y 的重要性更大[3]。

1.3 PLS-DA 方法

PLS-DA 是一种集合了主成分分析、典型相关分析和多元线性回归特点的数据分析方法,是一种有监督的判别分析统计方法[4]。它运用 PLS-DA 建立变量与样本类别之间的关系模型,来实现对样本类别的预测。在 PLS-DA 模型中,以 $R2Y$ 和 $Q2$ 来分别表示累积的自变量(X)对 Y 的解释能力和模型的预测能力,$Q2 > 0.50$ 且 $0 < R2Y - Q2 < 0.2 \sim 0.3$,表明建立的 PLS-DA 模型稳定、有效[5]。$R2Y$ 和 $Q2$ 的值越接近于 1,表示自变量(X)对 Y 的解释能力和模型的预测能力越强。采用置换检验(Permutation)判断模型是否过度拟合。

2 指标选取及数据来源

2.1 指标选取

2.1.1 被解释变量

众所周知,评价某一产业的发展状况,涉及的因素较多,单一指标显然无法胜任。本文综合了张广海(2013)[6]、赵彦云(2006)[7]、胡日东(2011)[8]等学者的研究成果,选取行业收入(产值)、就业人数(工作岗位)、税收 3 个指标来评价户外休闲产业的发展水平。为考察不同区域户外休闲产业从业人员的收入水平,将行业的薪酬指标也纳入评价指标体系,同时为了增加可比性,再增加 2 个相对量指标,即工作岗位占比和产值占比,具体如表 1 所示。

2.1.2 解释变量(影响因素)

美国户外产业协会把户外休闲项目划分为十大类,分别是露营、钓鱼、狩猎、摩托车运动、机车越野、雪上运动、水上运动、越野跑、自行车及滑板运动、野生动物观赏[2]。这些项目的开展与所在区域的自然资源、经济社会发展水平、人口等密切相关。本文选取解释变量时主要考虑自然、经济、社会及人口 4 个方面因素,选取 27 个指标作为解释变量,具体如表 1 所示。

2.2 数据来源

美国各州户外休闲产业的基础数据来源于美国户外产业协会(Outdoor Industry Association)发布的户外休闲经济(The Outdoor Recreation Economy)报告[2],相对量指标(工作岗位占比、产值占比)通过计算而得。解释变量各指标基础数据通过 4 个渠道整理而得,分别为数字美国官网、美国商务部经济分析局官网、美国健康排行榜 2021 年度报告、维基百科,部分相对量指标通过基础数据计算而得。

表 1　评价美国户外休闲产业发展状况及影响因素的指标体系及代码表

类 型	指　标
被解释变量	绝对量:工作岗位(*Y*1)、产值(*Y*2)、薪酬(*Y*3)、税收(*Y*4) 相对量:工作岗位占比(*Y*5) = 户外休闲产业提供的工作岗位÷所在区域总的工作岗位 × 100% 产值占比(*Y*6) = 户外休闲产业创造的产值÷所在区域的 GDP × 100%
解释变量	1、自然因素:国家公园数(*X*1)、国家森林公园数(*X*2)、州立公园数(*X*3)、游径数(*X*4)、区域面积(*X*5)、水域面积(*X*6)、水域面积占比(相对量)(*X*7)、森林面积(*X*8)、森林面积占比(相对量)(*X*9)、空气污染情况(*X*10) 2、经济因素:GDP(*X*11)、个体消费支出(*X*12)、个体收入(*X*13)、消费税率(*X*14)、家庭财产(*X*15)、房产拥有率(*X*16)、就业人数(*X*17) 3、社会因素:户外休闲参与人数(*X*18)、户外休闲参与人数占比(相对量)(*X*19)、贫困人口(*X*20)、贫困人口比率(相对量)(21)、基尼系数(*X*22)、肥胖人口(*X*23)、肥胖人口比率(相对量)(*X*24) 4、人口因素:总人口(*X*25)、美国公民人数(*X*26)、美国公民人数占比(相对量)(*X*27)

3　结果与分析

3.1　美国州域户外休闲产业发展模式识别及特征分析

3.1.1　美国州域户外休闲产业发展模式识别

在模式识别过程中,一般先进行无监督的 PCA(Principle Component Analysis,PCA),用前两个主成分(PC1,PC2)构建积分图(Score Plot),若样本之间有分类趋势,则采用 PLS-DA 进行有监督的模式识别。在 PCA 之前,先对原始数据进行预处理,采用中心化方式对数据进行标度化,再进行 PCA 运算。PCA(M1 模型)计算结果显示(表 2),*Q*2(cum)达到 0.911,说明构建的 PCA 模型预测能力很好。

表 2　美国户外休闲产业发展模式识别不同模型的参数值

模型代码	模型	样本量	*R*2*X*(cum)	*R*2*Y*(cum)	*Q*2(cum)
M1	PCA - Y	50	0.989	-	0.911
M2	PLS-DA	50	0.666	0.649	0.510

综合考虑 PCA 聚类分析中样品聚类的结果,本文将美国户外休闲产业发展模式分为四组,进行有监督的 PLS-DA 分析,以验证模式分类的可靠性。SIMCA 数据处理自动提取了 2 个主成分,其 *R*2*X*(cum)的值为 0.666、*R*2*Y*(cum)的值为 0.649、*Q*2(cum)值分 0.510(表 2)。PLS-DA 的得分图(Score Scatter)如图 1 所示。采用交叉验证法(Cross Validation,CV)判断模型是否过拟合。通过采用 200 次的响应排序检验(图 2),结果显示模型具有优秀的稳定性,即本文所提出的美国户外休闲产业发展模式的分类具有较高的可靠性。PLS-DA 得分图(图 1)显示,美国户外休闲产业可分为 4 种发展模式。

3.1.2　美国州域户外休闲产业发展模式的特征分析

PLS-DA 得分图(图 2)显示,美国州域户外休闲产业发展模式可分为 4 种,即(1)总量大占比小型;(2)总量小占比大型;(3)总量较大占比较小型;(4)总量较小占比较大型。

(1)总量大占比小型。该种模式所在地区的户外休闲产业产值和提供的工作岗位总量很大,但是,其产值占比和工作岗位占比却很小。从图 3 可知,共有 4 个州属于这种模式,分别为加利福尼亚、得克萨斯、佛罗里达和纽约。从空间分布看,该种模式呈零散分布,并未连成片,分别位于美国的东、西、南部。

在这种模式中,加利福尼亚最具代表性。例如,2021 年,加利福尼亚的户外休闲产值高达 920 亿美元,位居全美 50 个州首位(而位于末位的特拉华,其产值只有 31 亿美元,前面是后者的 29.68 倍),创造了 691 000 个工作岗位。但是,从相对量看,加利福尼亚户外休闲产业的产值只占该州 GDP 的 0.77%,所创造的工作岗位也只占全州工作岗位的 4.04%(纽约的相对量更低,分别只有 0.62% 和 3.48%)。可见,对于加利福尼亚而言,户外休闲产业总量很大,但在所在地区的经济发展中占比却很小,并不属于所在地区经济发展的支柱性产业。

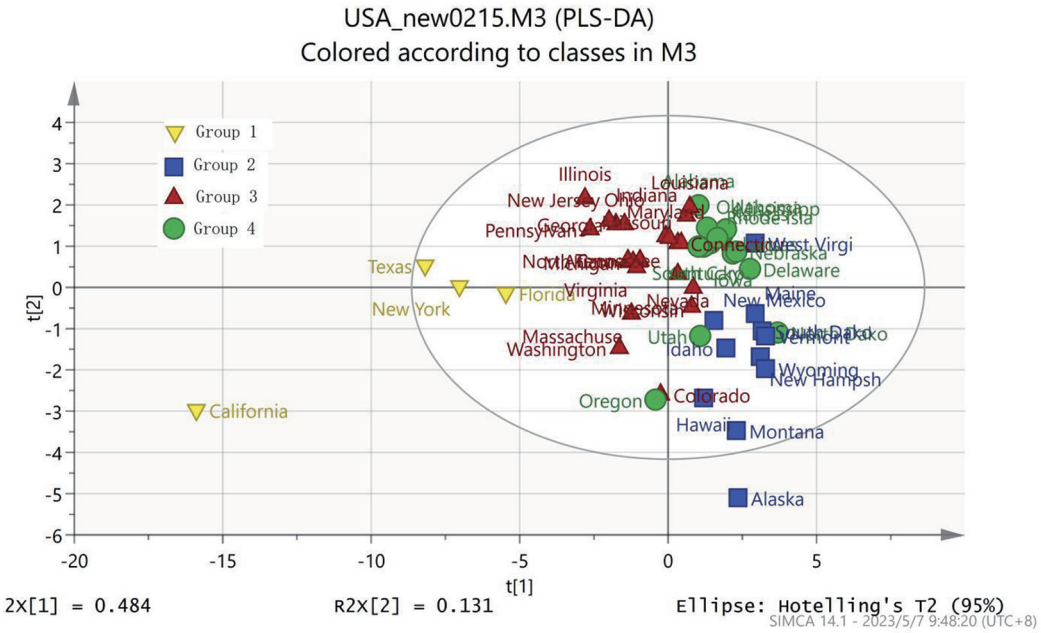

图 1　PLS-DA 的得分图

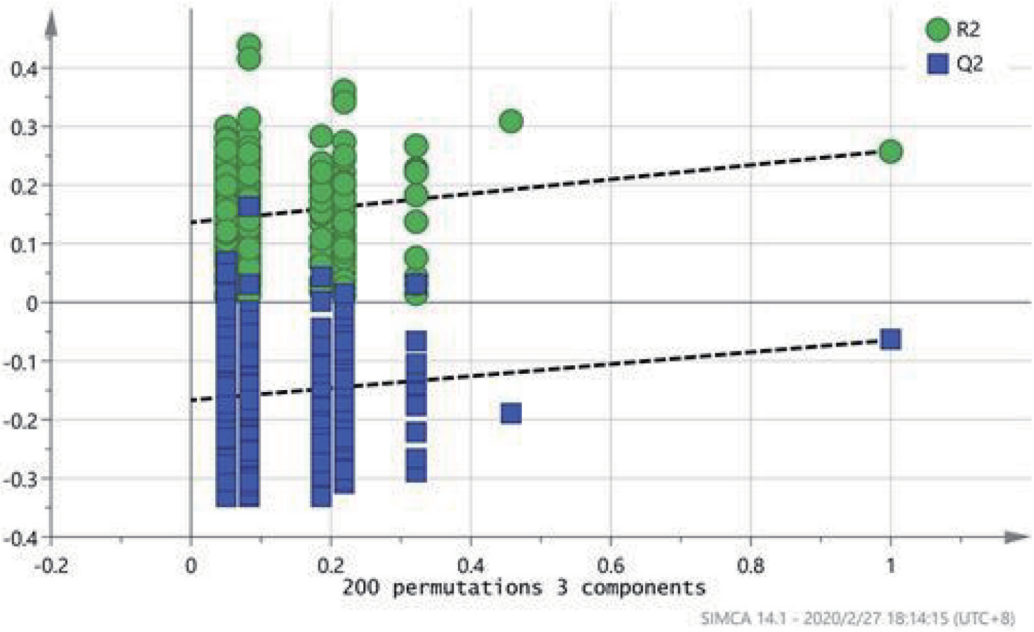

图 2　交叉验证法

从影响因素看，属于该种模式的州，都具有经济发达、人口众多、自然资源丰富等特点。例如，2021 年，加利福尼亚、得克萨斯、佛罗里达和纽约四个州的 GDP 总量、居民消费支出、居民收入、就业人数以及人口等 5 个指标的数值均位于全美 50 个州的前 4 名；州立公园数也位于前列，分别为 278 个（排名第一）、160 个（排名第四）、133 个（排名第七）、178 个（排名第三），这些因素都有力地促进所在地区的户外休闲产业发展。但是，各州也存在一些不利于所在地区户外休闲产业发展的因素。如居民户外休闲参与率低、美国公民比率低、房屋自有率低、肥胖人数多、空气质量差等。

从项目特征看，各州参与人数超过全美平均水平的项目分别是，加利福尼亚为徒步旅行和单板滑雪；佛罗里达为自行车和皮划艇；德克萨斯为越野跑和汽车越野；纽约为雪地摩托和高山滑雪。

(2)总量小占比大型。该种模式所在地区的户外休闲产业产值和工作岗位总量虽比较小,但户外休闲产业的产值占比和工作岗位占比却很大。这表明,户外休闲产业对于这些地区而言,毫无疑问是其经济发展的支柱性产业。例如,2021 年,产值占比最高的 15 个州中,有 12 个州属于这种类型,其中佛蒙特的产值占比达到 4.08%。同样,工作岗位占比也存在类似情况,2021 年,工作岗位占比最高的 15 个州中,有 13 个州属于这种类型,其中工作岗位占比超过 15% 的有 4 个州,特别是阿拉斯加,其工作岗位占比甚至高达 20.78%。可见,户外休闲产业是解决这些地区就业问题的关键产业。从图 3 可知,共有 19 个州属于这种模式,主要分布于美国西部地区、中西部地区和新英格兰地区。从影响因素看,属于该种模式的州,大部分都具有经济总量小,人口少,个体消费支出和个体收入较低,但是,居民的户外休闲参与率却非常高等特点。例如,2021 年,全美户外休闲参与率排名前 15 名的州中,有 14 个州属于该种类型,有些参与率甚至高达 81%(阿拉斯加和蒙大拿)。

从项目特征看,徒步旅行是该种模式居民最喜爱的户外休闲项目,有 8 个州的居民参与人数超过全美平均水平。其次为露营项目,有 7 个州的居民参与人数超过全美平均水平。接下来的项目分别为,垂钓(5 个州)、皮划艇(3 个州)和野生动物观赏(3 个州)。

(3)总量较大占比较小型。该种模式所在地区的户外休闲产业的产值和提供的工作岗位总量较大,分别处于 189 亿 ~ 291 亿美元和 143 000 ~ 260 000 个之间,仅次于第 1 种模式,产业占比和工作岗位占比也比较小。从图 3 可知,共有 9 个州属于该种模式,分别为北卡罗来纳、宾夕法尼亚、佐治亚、密歇根、俄亥俄、伊利诺伊、弗吉尼亚、新泽西和亚利桑那,主要分布在美国五大湖周边和大西洋沿岸。从影响因素看(图 4),该种类型与第 1 种模式(总量大占比小型)相似,只是数值比后者小。例如,经济类指标(GDP、个体消费支出、个体收入、就业人数)、人口、户外休闲参与人数等指标的数值均较大(仅次于第 1 种模式),但户外休闲参与率也比较低,空气质量相对也比较差。

从项目特征看,越野跑是该种模式居民最喜爱的户外休闲项目,9 个州中有 4 个州的居民参与人数超过全美平均水平。其次是露营(2 个州)、垂钓(2 个州)和狩猎(2 个州)。

(4)总量较小占比较大型。这种模式所具有的特征与第 3 种模式(总量较大占比较小型)正好相反,即该种模式所在地区的户外休闲产业的产值和工作岗位总量比较小,但户外休闲产业的产值占比和工作岗位占比却比较大。从图 3 可知,共有 18 个州属于这种模式,主要集中在美国中部地区。从影响因素看,经济类指标、人口、户外休闲参与人数、空气质量等指标位于全美各州的中间水平,自然类指标整体偏低,但贫困率和肥胖率很高。例如,2021 年,贫困率最高的 10 个州,有 7 个州属于该种类型,其中密西西比的贫困率达到 20.80%(全美最高水平);肥胖率也有类似情况,2021 年肥胖率最高的 10 个州中,有 9 个州属于该种类型,其中西弗吉尼亚的肥胖率达到 38.10%(全美最高水平)。

从项目特征看,该种模式居民比较喜爱水上项目。例如,有 11 个州的居民参与垂钓人数超过全美平均水平;2 个州的居民参与皮划艇人数超过全美平均水平;帆船、站立式桨板、船后骑管 3 个水上项目也分别有 1 个州的居民参与人数超过全美平均水平,其次是露营项目,有 7 个州的居民参与人数超过全美平均水平。

3.2　基于 PLS 的美国户外休闲产业发展预测模型构建

3.2.1　美国户外休闲产业发展预测模型的交叉有效性与精度分析

根据 SIMCA - P + 14.1 软件自动拟合结果,共提取 3 个主成分 t_1、t_2 和 t_3,它们的 $R2X$、$R2X$(cum)、$R2Y$、$R2Y$(cum)、$Q2$、$Q2$(cum)等值如表 3 所示。

表 3　PLS 的交叉有效性检验与精度分析各值表

	$R2X$	$R2X$(cum)	$R2Y$	$R2Y$(cum)	$Q2$	$Q2$(cum)	临界值
第一主成分(t_1)	0.485	0.485	0.648	0.648	0.619	0.619	0.097 5
第二主成分(t_2)	0.122	0.607	0.135	0.783	0.293	0.731	0.097 5
第三主成分(t_3)	0.130	0.737	0.037	0.820	0.061	0.747	0.097 5

从表 3 可知，根据交叉有效性 $Q2$ 值，美国户外休闲产业发展 PLS 提取 3 个主成分是合适有效的，$R2X$(cum)、$R2Y$(cum)、$Q2$(cum)值分别为 0.737、0.820、0.747，说明构建的模型利用 X 解释变量组 73.7% 的信息能够解释 Y 集合 82.0% 的变异。同时，整个模型的预测能力达到 74.7%。这表明，所构建的美国户外休闲产业发展水平的预测模型有效，且精度较高。

3.2.2 美国户外休闲产业发展的影响因素重要性分析

通过对上述数据的求解，获得 VIP 得分图(图 3)。

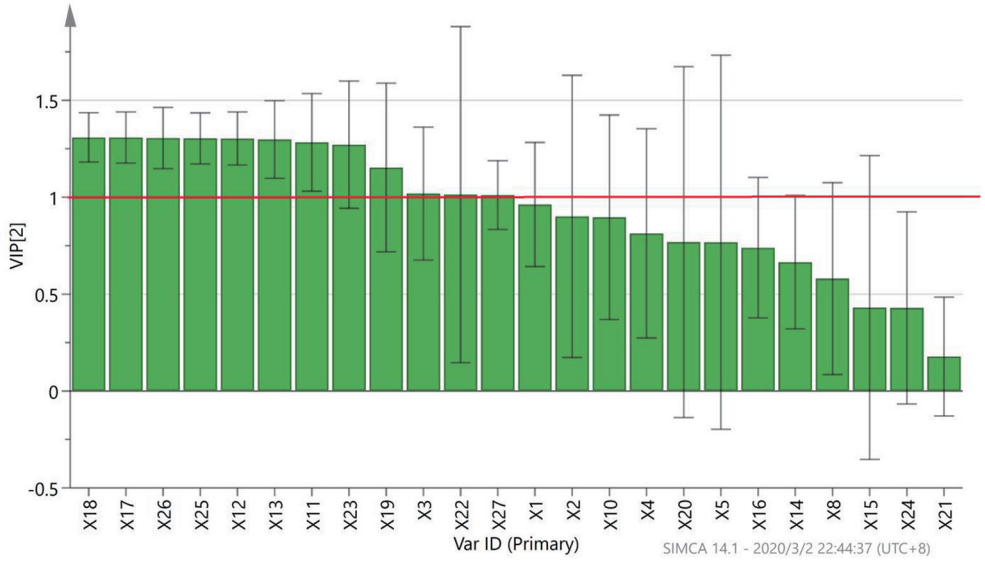

图 3 美国户外休闲产业发展的影响因素 VIP 得分图

从图 3 可知，在解释变量 X 集合中，共有 12 个变量的 VIP 得分大于 1，它们依次为户外休闲参与人数、就业人数、美国公民人数、人口、个体消费支出、个体收入、GDP、肥胖人口、户外休闲参与人数占比、州立公园数、基尼系数、美国公民人数占比。从 VIP 结果看，经济、人口和社会因素似乎更多地影响美国户外休闲产业发展，而自然环境因素的影响相对较弱，只有州立公园数这唯一指标进入 VIP。

3.2.3 美国户外休闲产业发展的预测模型分析

通过 SIMCA-P + 14.1 软件自动导出($\hat{Y}$, Y)散点图(图 4)。

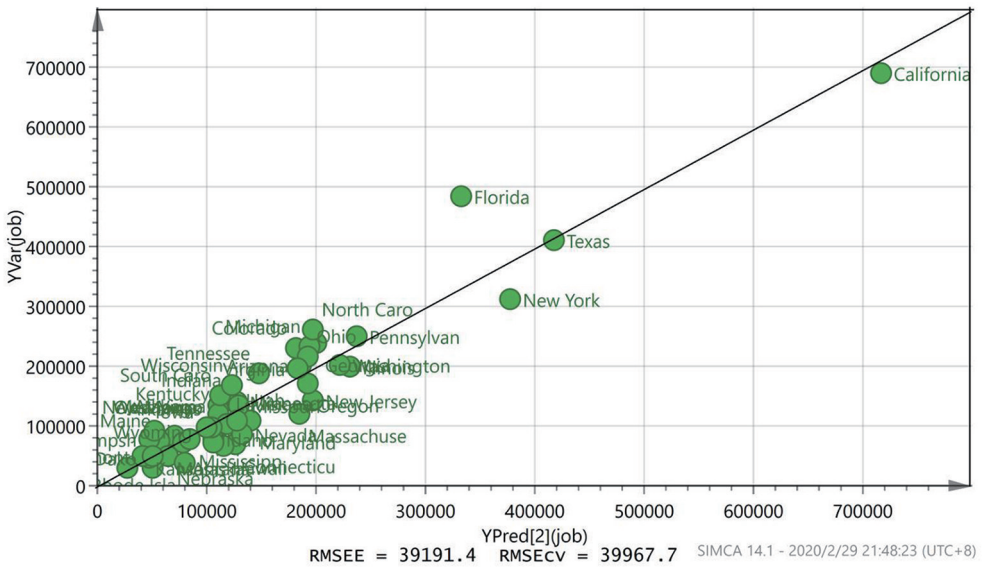

图 4 美国户外休闲产业发展预测值与原始值比较图

从图 4 可以看出,美国户外休闲产业发展预测值与原始值的样本点基本上分布在对角线上,说明构建的美国户外休闲产业发展的预测方程拟合效果比较理想。

4　结论

一、美国州域户外休闲产业可分为 4 种发展模式:(1)总量大占比小型,分别为加利福尼亚、得克萨斯、佛罗里达和纽约四个州,位于美国的东、西、南部,呈零散分布,并未连成片。四个州具有经济发达,人口众多,自然资源丰富,居民户外休闲参与率低,美国公民比率低,房屋自有率低,肥胖人数多,空气质量差等特点。参与人数超过全美平均水平的项目有徒步旅行、单板滑雪、自行车、皮划艇、越野跑、汽车越野、雪地摩托、高山滑雪。(2)总量小占比大型。共有 19 个州属于这种模式,主要分布在美国西部、中西部和新英格兰地区。这些州具有经济总量小,人口少,个体消费支出和个体收入低,但居民的户外休闲参与率非常高等特点。户外休闲产业是上述地区经济发展的支柱性产业,也是解决美国西部地区就业问题的关键产业。徒步旅行是该种模式居民最喜爱的户外休闲项目,其次是露营,再次分别为垂钓、皮划艇和野生动物观赏。(3)总量较大占比较小型。共有 9 个州属于该种模式,主要分布在美国五大湖周边和大西洋沿岸。这些州具有经济类指标(GDP、个体消费支出、个体收入、就业人数)、人口、户外休闲参与人数等指标的数值均较大(仅次于第 1 种模式),但户外休闲参与率也比较低,空气质量相对也比较差等特点。越野跑是该种模式居民最喜爱的户外休闲项目,其次为露营、垂钓和狩猎。(4)总量较小占比较大型。共有 18 个州属于这种模式,主要集中在美国中部地区。这些州具有经济类指标、人口、户外休闲参与人数、空气质量等指标位于全美各州的中间水平,自然类指标整体偏低,但贫困率和肥胖率很高等特点。该种模式居民比较喜爱水上项目(垂钓、皮划艇、帆船、站立式桨板、船后骑管)和露营项目。

二、构建的预测模型有效且精度较高,并筛选出 12 个 VIP 指标,即户外休闲参与人数、就业人数、美国公民人数、人口、个体消费支出、个体收入、GDP、肥胖人口、户外休闲参与人数占比、州立公园数、基尼系数、美国公民人数占比。可见,经济、人口和社会因素似乎更多地影响美国户外休闲产业发展,而自然环境因素的影响相对较弱。

参考文献

[1]国务院办公厅. 国务院办公厅关于加快发展健身休闲产业的指导意见[EB/OL]. (2016-10-28)[2023-06-28]. http://www. gov. cn/zhengce/content/2016-10/28/content _ 5125475. htm.

[2]美国户外休闲产业协会. 户外休闲产业[EB/OL]. (2021-11-28)[2023-08-18]. https://outdoorindustry. org/advocacy/.

[3]王惠,吴载斌,孟洁. 偏最小二乘回归的线性与非线性方法[M]. 北京:国防工业出版社,2006:1-7,102-104,117-120,140-141.

[4]柯朝甫,武晓岩,李康. PLS-DA 模型四种诊断统计量在代谢组学应用中的比较[J]. 中国卫生统计,2014,31(03):403-406.

[5]魏德样,黄彩华,雷雯. 中、外体育用品上市公司盈利能力的特征分析与预测模型构建——基于 PLS 回归和 OPLS-DA 方法的分析[J]. 体育科学,2012(10):12-19.

[6]张广海,刘真真,李盈昌. 中国沿海省份旅游产业发展水平综合评价及时空格局演变[J]. 地域研究与开发,2013,32(04):22-27.

[7]赵彦云,余毅,马文涛. 中国文化产业竞争力评价和分析[J]. 中国人民大学学报,2006(04):72-82.

[8]胡日东,李颖. 我国房地产业发展的综合评价——基于动态因子分析法[J]. 经济地理,2011,31(11):1862-1866.

[责任编辑　江国平]