文章编号：2095-2163(2021)02-0201-06 中图分类号：TP311.13 文献标志码：A

基于 PCA 降维的大数据可视化应用研究

马佳琪，滕国文
（吉林师范大学 计算机学院，吉林 四平 136000）

摘 要：由于近些年来火灾时有发生，被称为“地球之肺”的最大雨林区亚马逊也不断面临着威胁。因此，了解和分析火灾发生的时间和空间势在必行。基于此，在亚马逊火灾的分析评价中，试采用主成分分析法（PCA）建立数学模型，从时间、空间的不同维度对亚马逊火灾的发生情况进行了可视化分析。最终得出具体的时间和地点是火灾的高峰期。为预防更多火灾的发展，阻止全球气候变暖的发展提供参考方向。

关键词：数据可视化；PCA；亚马逊火灾

Research on big data visualization based on PCA dimensionality reduction

MA Jiaqi, TENG Guowen
(College of Computer, Jilin Normal University, Siping Jilin 136000, China)

【Abstract】 The Amazon, the rainiest forest in the world and known as the "lungs of the world", is under constant threat because of irregular fires in recent years. Therefore, it is imperative to understand and analyze the time and space of fire. Based on this, in the analysis and evaluation of Amazon fire, a mathematical model is established by principal component analysis (PCA), and a visual analysis is conducted on the occurrence of Amazon fire from different dimensions of time and space. So it is concluded that the specific time and place is the peak of the fire. To prevent the development of more fires and the development of global warming, the research in the paper could provide reference direction.

【Key words】 data visualization; PCA; Amazon fire

0 引 言

在人工智能发展的今天，可视化凭借计算机和数字图像处理方法，把批量高维数据转换为图表后进行展示和处理。当处理科研问题及其数据时，人们往往遇到甚至会达到数百万维度的真实数据[1]。尽管在其原来的高维结构中，数据能够得到最好的表达，但有时就可能需要给数据进行降维。降维的需求往往与可视化有关（减少两三个维度，方便人们绘图），但这只是原因之一。有时候，人们认为性能比精度更重要，那么就可以将 1 000 维的数据降至 10 维，从而让人们可以更快地对这些数据进行操作（比如计算距离）。综上可知，对降维的需求是存在的并且有很多应用。

1 数据可视化

可视化分析作为大数据分析的一个重要分支，已经广泛应用于科学计算研究和商业智能[2]。因此，数据可视化分析是大数据分析不可缺少的手段和工具[3]。可视化分析（Visual analytics）是科学可视化、信息可视化、人机交互、数据挖掘等研究领域交叉集成而产生的一种新的研究方向[2]，也是一种通过交互式可视化界面帮助用户分析和推理大规模复杂数据集的科学技术[4]。分析过程在数据和知识转化的过程中不断循环，可将大数据分析和挖掘方法与视觉信息处理过程相结合，将计算机的处理能力和人类的认知能力相结合，最终挖掘出大规模高维数据集所包含的价值[1]。

大部分存储的原始数据都是没有价值的，只有在提取信息后，才能发现价值。人类处理视觉信息的速度非常快，可以立即捕捉到隐藏在数字中的关键信息。因此，数据可视化已成为提取关键信息的最佳途径。

2 主成分分析法

主成分分析（Principal Component Analysis，PCA）[4]将包含冗余信息的高维数据转化为少量的低维数据，即主成分，每个主成分包含原始数据几乎所有的有效信息[5]。这将复杂的数据分析问题转化为只需要几个主成分的问题，不仅能够对问题进

基金项目：吉林师范大学研究生创新项目（201955）。

作者简介：马佳琪（1995-），女，硕士研究生，主要研究方向：数据可视化；滕国文（1963-），男，教授，硕士生导师，主要研究方向：人工智能。
通讯作者：滕国文 Email：1368478853@qq.com

收稿日期：2020-11-22

行更深入的分析,而且使分析过程更加容易[4]。基本思想是在最小均方误差的约束下,寻找一个最能代表原始数据主要特征的投影变换矩阵。在新的投影空间中,可以降低原始数据的维数,保留大部分信息[5]。整个转换过程遵循2 个原则。一个是近期重构,即:利用无量纲数据重构原始数据时误差之和最小。另一个是最大可分性,即:数据要在低维投影空间中尽可能分离[5]。其实可以证明,这两个原理是等价的[5]。

2.1 PCA 算法步骤

将“发生年份”、“发生月份”、“发生地点(州名)”、“平均纬度”、“平均经度”、“火灾发生次数”等原始数据 $X_0,X_1,X_2,X_3,X_4,X_5,X_6$,依次排列行向量,分别进行数据标准化,代入如下公式[6]:

$$Z_{ij}=\frac{x_{ij}-x_j}{s_j},\ i=1,2,3,\cdots\cdots,n;\ j=1,2,\cdots\cdots p, \tag{1}$$

$$\overline{X_j}=\frac{1}{n}\sum_{i=1}^{n}X_{ij};S_j^2=\sum_{i=1}^{n}X_{ij}(X_{ij}-\overline{X_j})^2, \tag{2}$$

接下来,将结果代入协方差矩阵,可得:

$$\boldsymbol{C}=(c_{ij})_{n\times m}=\begin{bmatrix}c_{11} & c_{12} & \cdots & c_{1n}\\ c_{21} & c_{22} & \cdots & c_{2n}\\ \cdots & \cdots & & \cdots\\ c_{n1} & c_{n2} & \cdots & c_{nn}\end{bmatrix}. \tag{3}$$

其中,$c_{ij}=Cov(X_i,X_j),i,j=1,2,3,\cdots,n$。

分别计算出 $X_0,X_1,X_2,X_3,X_4,X_5,X_6$ 的协方差矩阵。再将 X 的每一行(代表一个属性)置为0。在此基础上,得到协方差矩阵、协方差矩阵的特征值和相应的特征向量[6]。至此,将特征向量按照对应的特征值大小自上而下排列成矩阵,取前 K 行形成矩阵 $\boldsymbol{P}$,即降维为 K 维后的数据[4]。

2.2 PCA 主成分分析降维

亚马逊雨林区是世界最大的雨林区,可以消耗大量二氧化碳,阻止气候变暖;林区还藏有丰富的动植物资源,种类高达300 万种。但不容忽视的是,雨林生态系统却正不断面临着众多的威胁,越来越多的森林砍伐导致雨林面积逐年缩小。同时,全球变暖也增加了发生野火的可能性和频率。本文对1999 ~2019 年、总共20 年间的亚马逊雨林火灾数据进行探索分析与可视化。

本次研究将基于在 kaggle 下载的巴西国家太空研究所(INPE)公开的卫星图像检测数据,该数据中详尽记录了亚马逊地区火灾的情况。研究中,还将用到主成分分析,其目标是旨在找到数据中最重要的元素和结构,去除噪声和冗余,降低原始复杂数据的维数,揭示隐藏在复杂数据背后的简单结构[7]。混沌数据通常由3 部分组成:噪声、旋转和冗余[7]。区分噪声时,可以用信噪比或方差来衡量。方差是主要信号或主要成分。小的方差被认为是噪声或次要成分;对于旋转,旋转基向量,使得具有大信噪比或方差的基向量是主分量方向。在判断观测变量之间是否存在冗余时,可以用协方差矩阵来度量和判断[7]。

3 数据分析

将样本集 PCA 降维后进行数据分析。amazon_fires. csv 是按州、月份和年份统计在从 1999 ~2019 年巴西亚马逊地区发生的火灾次数文件。数据共计2 104 条,各数据字段含义见表1。

表1 数据集部分字段

Tab. 1 Data set partial fields

字段	含义
year	发生年份
month	发生月份
state	发生地点(州名)
latitude	平均纬度
longitude	平均经度
firespots	火灾发生次数

3.1 导入所需的库并读取数据

研究中可得统计量图表见表2。由表2 可以看到所有字段均为数字型,且不存在缺失值。对此,研究拟通过描述性统计函数 *describe*( )检查数据中有无明显异常值。年份、月份的最小最大值分别为(1999,2019),(1,12),且经纬度数据、火灾发生次数均不存在明显异常,说明降维后的数据较为“干净”。

表2 统计量图表

Tab. 2 Statistics chart

	year	month	stats	latitude	longitude	firespots
0	1999	1	AMAZONAS	-2.371 113	-59.899 933	3
1	1999	1	MARANHAO	-2.257 395	-45.487 831	36
2	1999	1	MATO GROSSO	-12.660 633	-55.057 989	18
3	1999	1	PARA	-2.474 820	-48.546 967	87
4	1999	1	RONDONIA	-12.861 700	-60.513 100	1

3.2 火灾发生时间的可视化分析

研究中将按年份进行分组,计算 1999 ~2019 年间每一年的火灾发生总数,并通过折线图的方法进行可视化。仿真结果如图1 所示。

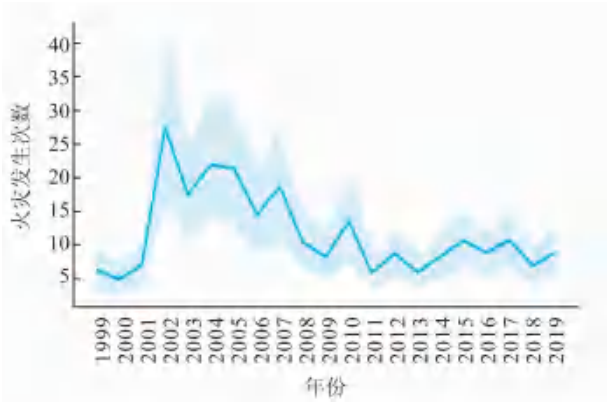

图 1　每年火灾发生次数

Fig. 1　Number of fires per year

由图 1 可以看到，亚马逊地区的火灾爆发在 2002 年达到了一个高峰，从 2002 年以来，火灾情况呈逐年减少态势。从 2010 ~ 2019 年，每一年的火灾爆发情况出现了小范围波动。在此基础上，本次研究又按月来统计了火灾爆发的情况，具体结果如图 2 所示。通过统计 12 月中每月的平均火灾数进行分析。

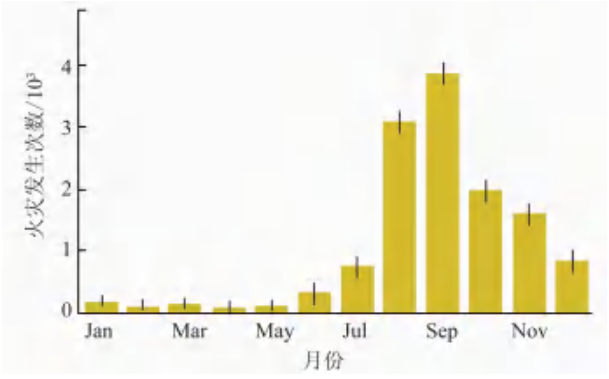

图 2　每月火灾发生次数

Fig. 2　Number of fires per month

由图 2 中可以明显看出，下半年平均受火灾的影响比上半年高很多，平均着火点数目位列前三的月份分别是 9 月、8 月和 10 月。

一般情况下，亚马逊的旱季从 7 月持续到 10 月，在 9 月底达到顶峰。在一年的其他时间里，潮湿的天气会将火灾的风险降到最低。但在旱季，降雨量的减少可能对火灾情况有较大影响。

3.3　火灾发生地点的可视化分析

巴西一级行政区划包括 26 个州和 1 个联邦区，亚马逊雨林分布在其中的 9 个州，这里拟通过计算每个州的火灾发生总数来分析哪个州受雨林火灾影响最大。研究后得到的仿真结果如图 3 所示。

由图 3 中可以看到，帕拉州（PARA）和马托格罗索州（MATO GROSSO）是受亚马逊河大火影响最大的巴西州，其火灾着火点总数是其他州加起来的至少两倍。后续可通过经纬度数据进行地理绘图，将火灾发生地点标记出来。

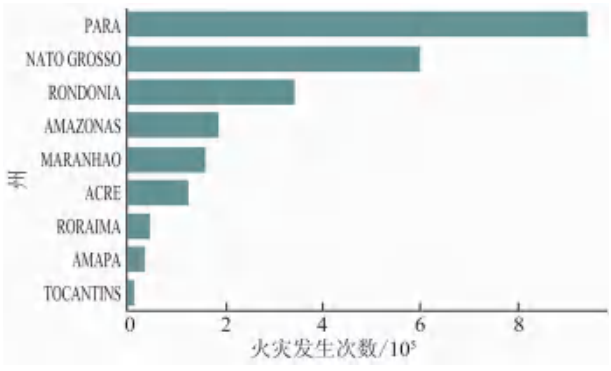

图 3　每个州火灾发生次数

Fig. 3　Number of fires per state

3.4　时间地点分析

为了更好地了解问题和当前状况，现将特征进行组合，更加深入地开展数据研究。在此，即根据州和年份进行组合，分析多年来每个州的火灾情况。由此得到的时间地点分析后的结果曲线如图 4 所示。对应地，也给出了该次研究编写的部分主要代码参见如下。

```
fig, ax = plt.subplots(3, 3, figsize = (14, 10), sharex = True)
sns.set_style("whitegrid")
ax = ax.flat
i = 0
for x in state_name:
    sns.lineplot(data = amazon_fires[amazon_fires['state'] == x], x = 'year', y = 'firespots', estimator = 'sum', ax = ax[i], color = 'teal', ci = None)
    ax[i].set_title(x, size = 'large')
    ax[i].set_xlabel("年份", size = 'large', fontproperties = font)
    ax[i].set_xticks([2000, 2005, 2010, 2015, 2020])
    ax[i].grid(False)
    ax[i].set_xticklabels([2000, 2005, 2010, 2015, 2020], fontsize = 'large')
    if i == 0 or i == 3 or i == 6:
        ax[i].set_ylabel("火灾爆发总次数", size = 'large', fontproperties = font)
    else:
        ax[i].set_ylabel("")
    i += 1
```

plt. subplots _ adjust (*wspace* = 0. 16, *hspace* = 0. 12)

plt. show( )

由图 4 可以看出,每个州在 2002 年左右都出现了火灾高峰,因此导致整体上 2002 年火灾数目非常高,2002 年后大部分州的火灾数目都逐渐减少。但是其他州也有例外,例如 AMAZONAS 州和 RORAIMA 州在 2002 年减少后又开始逐年增加,并且 RORAIMA 州在 2019 年达到了顶峰。

接下来再根据州和月份进行组合,分析不同月份下每个州的火灾情况,图 5 显示了每个州在每个月爆发火灾次数的平均值。

除罗赖马州(RORAIMA)之外,所有州的火灾都集中在下半年(7 ~ 10 月),即亚马逊雨林的旱季。综上研究后,则结合年份、月份和州三个属性进行可视化,分析火灾爆发的次数,研究得到的热力图如图 6 所示,该图显示了每年各州每月份的火灾爆发量,颜色越深代表火灾爆发次数越多。

由图 6 可以看出,几乎每个州在所有年份的火灾高峰期都在 7 ~ 10 月,这印证了之前的结论。并且在防范火灾方面,就需要在 1 ~ 4 月份格外注意 RORAIMA 州,因为只有该州的火灾高峰期不在 7 ~ 10 月。从 PARA、MATO GROSSO、RONDONIA、MARANHAO 和 TOCANTINS 五个州的数据来观察可知,随着年份的推移,火灾爆发的次数大大减少了。

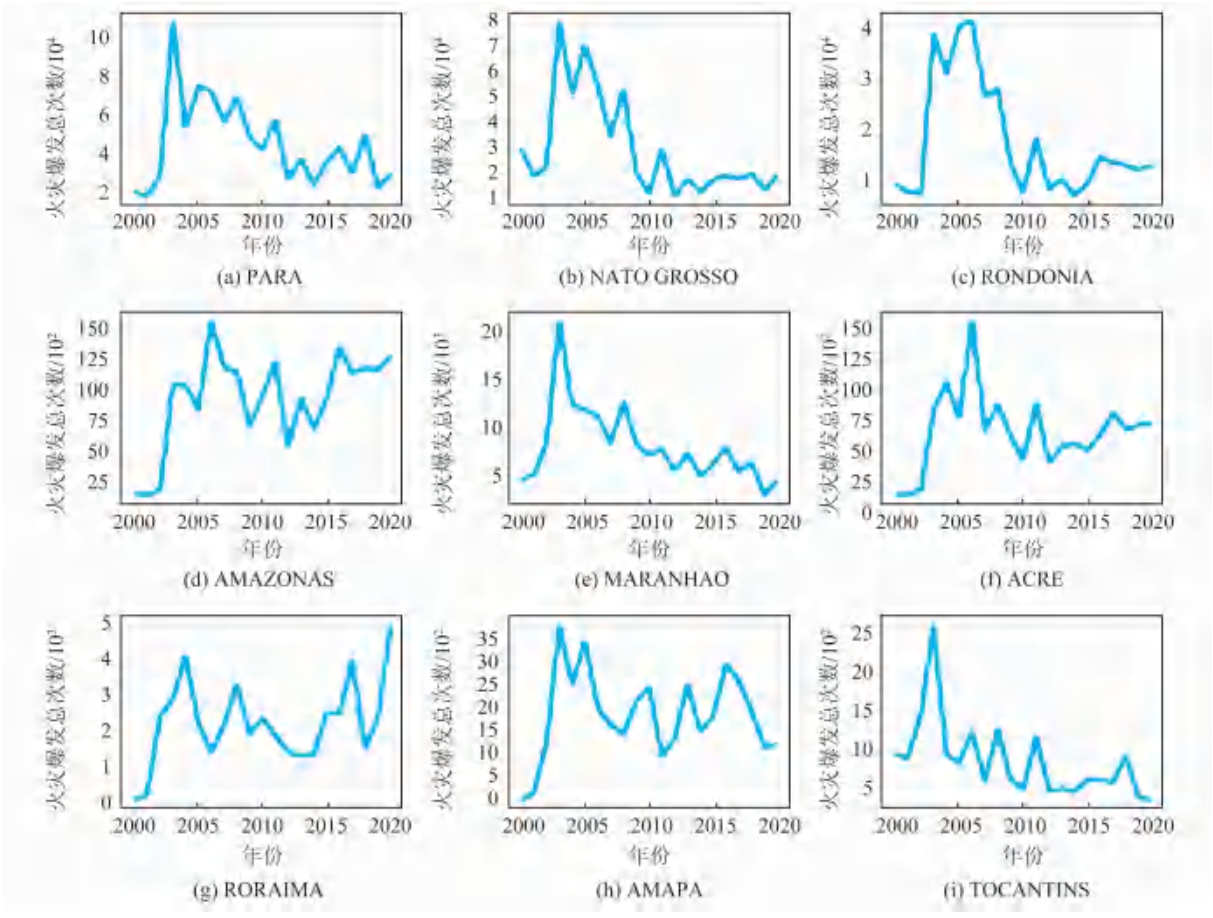

图 4 时间地点分析

Fig. 4 Time and place analysis

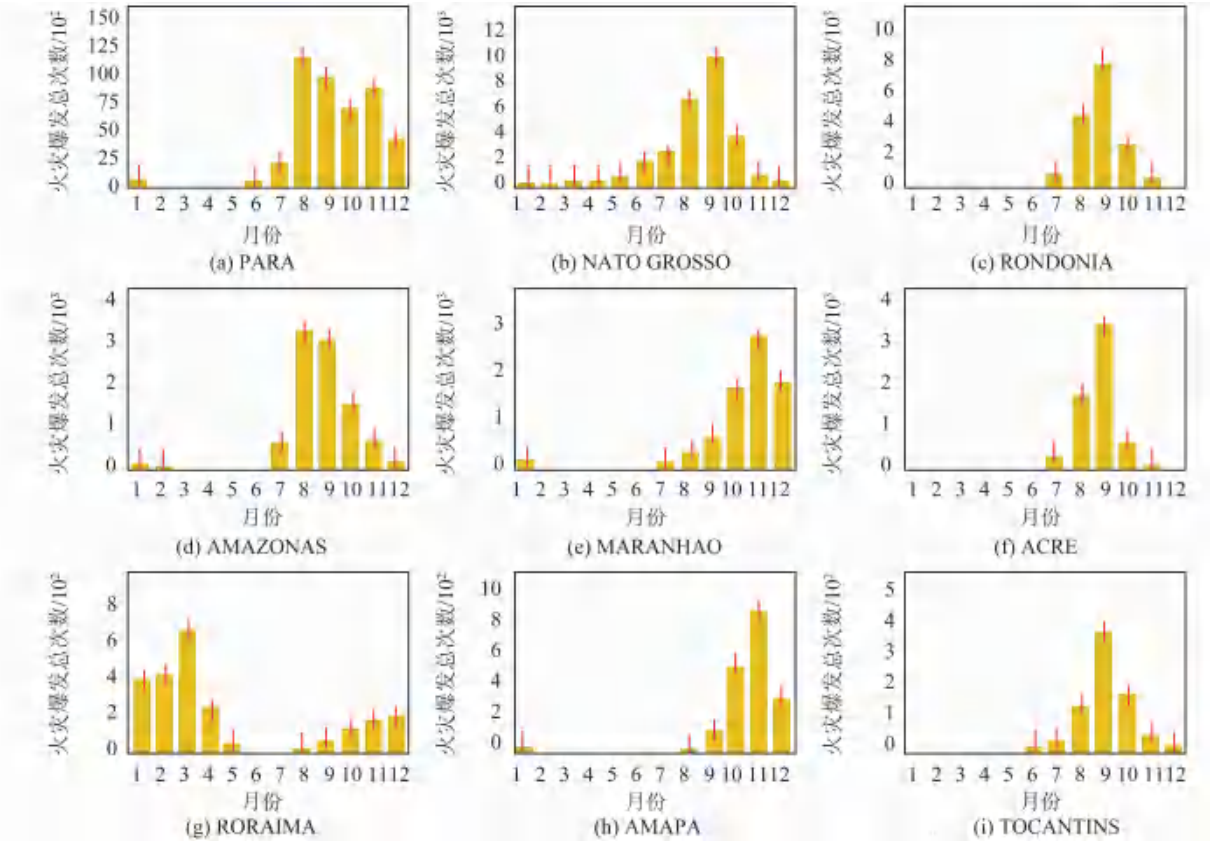

图 5　火灾次数平均值

Fig. 5　Mean fire frequency

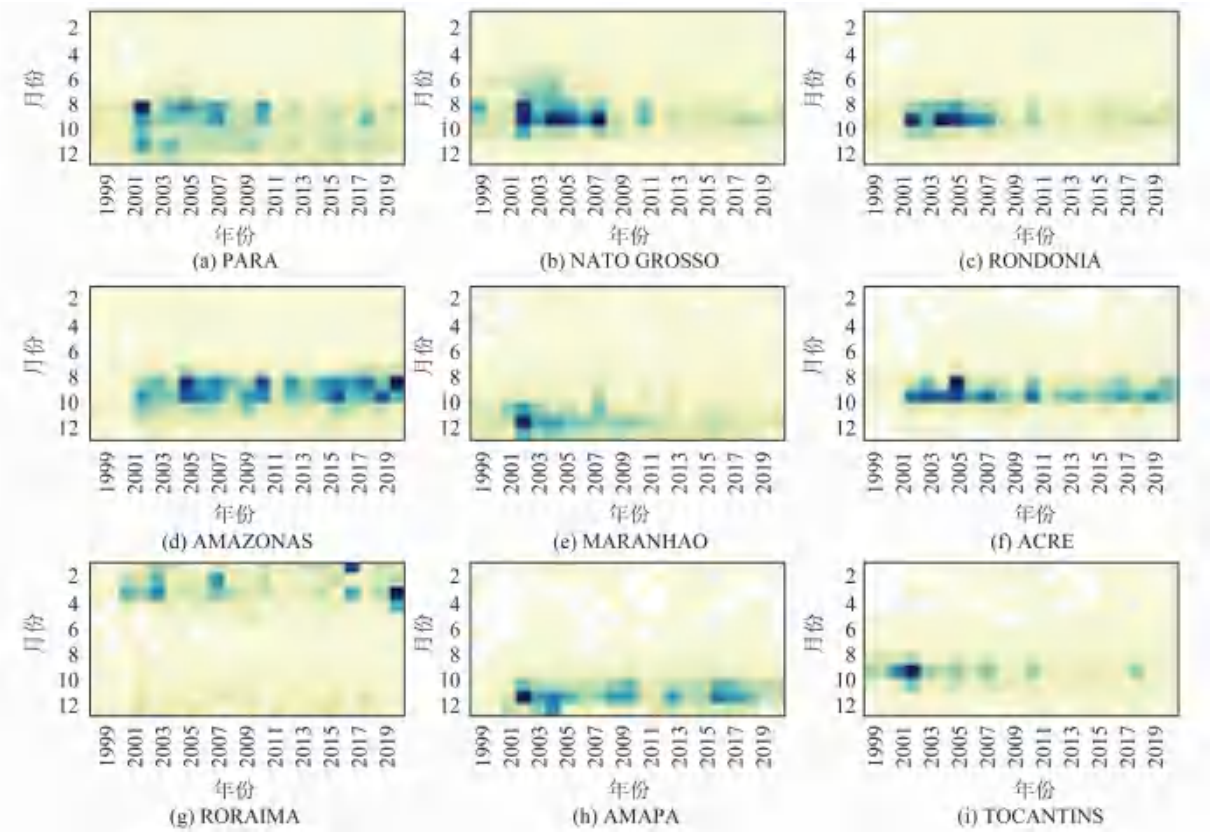

图 6　热力图

Fig. 6　Thermal diagram

4 结束语

近年来,数据可视化技术的发展日趋成熟,从结果图中研究者们能够直接找出自己所需要的信息。亚马逊雨林的面积约是印度的两倍,在调节全球气候和提供诸如水净化和二氧化碳吸收等其他服务方面发挥着至关重要的作用。在本文中,分别从时间、空间的不同维度对亚马逊火灾的发生情况进行了可视化分析,研究发现 7 ~ 10 月是火灾的高峰期。同时,本文绘制了丰富的可视化图形,对于数据的探索性分析可以提供有益参考。

参考文献

[1] 马佳琪,滕国文. 基于 Matplotlib 的大数据可视化应用研究[J]. 电脑知识与技术,2019,15(17):18-19.
[2] 马佳琪,滕国文. 基于大数据的幸福感可视化技术研究[J]. 电脑知识与技术,2020,16(7):263-264.
[3] 王振宇,高东健. 智慧城市大数据平台[J]. 中国新通信,2018,20(19):30.
[4] little_angle. 主元分析 PCA 原理以及应用[EB/OL]. [2012-05-29]. https://blog. csdn. net/j123kaishichufa/article/details/7614234.
[5] 曲学超. 基于高分辨距离像的雷达目标识别算法研究[D]. 成都:电子科技大学,2018.
[6] 刘浩昌,林汇峯,张英,等. 基于 PCA 法的汽车产业竞争力的综合评价[J]. 科技经济导刊,2020,28(31):224-225.
[7] 黄潇. 基于聚类分析的专家分类方法研究[D]. 南京:东南大学,2017.

(上接第 200 页)

音把智能家居生态和本地生活服务连了起来,形成了闭环。相信随着人工智能技术的不断发展,人工智能与物联网相结合,可以创造出更多的“智能设备”。

参考文献

[1] 韩丽丽,潘炜,刘丰威. 基于人工智能语音识别客服稽查应用前景[J]. 电子测试,2020(15):118-119,95.
[2] 徐来,朱海昆. 浅谈人工智能家居在室内设计中的应用[J]. 戏剧之家,2020(29):174-175.
[3] 张夏明,张艳. 人工智能应用中数据隐私保护策略研究[J]. 人工智能,2020(4):76-84.
[4] 周坤,李小松. 人工智能与计算智能在物联网方面的应用探索[J]. 计算机产品与流通,2020(11):152.
[5] 刘娟宏,胡彧,黄鹤宇. 端到端的深度卷积神经网络语音识别[J]. 计算机应用与软件,2020,37(4):192-196.
[6] 张文霞,闫顺斌. 智联网家居控制系统设计[J]. 无锡商业职业技术学院学报,2019,19(6):97-102.
[7] 吕值敏,苏皓,曹志文. 面向物联网应用的人工智能技术[J]. 造纸装备及材料,2020,49(1):95.
[8] 李荪,范志琰. AI+趋势下智能语音产业多模态发展趋势研究[J]. 电信网技术,2019(6):17-21.