**山东科技大学**

**实验报告**

**课程名称： 大数据分析方法及应用实验**

**实验项目：基于DBSCAN与谱聚类的全球生态足迹模式识别研究**

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

# 目录

[引言​​ 3](#_Toc26469)

[文献综述 4](#_Toc24067)

[1. 算法介绍 5](#_Toc25510)

[1.1 DBSCAN算法 5](#_Toc4981)

[1.1.1基本概念： 5](#_Toc2641)

[1.1.2 DBSCAN算法实现 5](#_Toc31120)

[1.2谱聚类 6](#_Toc15438)

[1.2.1 基本概念： 6](#_Toc24279)

[1.2.2 谱聚类算法实现： 6](#_Toc8799)

[2.实例分析 7](#_Toc21516)

[2.1 数据来源 7](#_Toc4390)

[2.2数据描述 7](#_Toc20134)

[2.3数据预处理 7](#_Toc13872)

[2.4 数据分析 8](#_Toc25637)

[2.5聚类分析 9](#_Toc7803)

[2.5.1 DBSCAN聚类 9](#_Toc3339)

[2.5.2谱聚类 10](#_Toc27156)

[2.5.3聚类结果分析 11](#_Toc5240)

[3.结论 12](#_Toc24002)

[参 考 文 献 13](#_Toc16823)

[附录 14](#_Toc19152)

引言**​**​

随着全球资源消耗持续上升和生态环境问题加剧，各国政府和研究机构愈发关注生态可持续性的量化评估与比较分析。尤其在全球气候变化、土地退化和生态系统服务失衡等问题日益突出的当下，如何科学评估一个国家对自然资源的依赖程度，衡量其可持续发展的潜力，已成为环境科学与可持续发展研究的重要方向。

生态足迹（Ecological Footprint）作为衡量人类对自然资源需求的重要指标，自20世纪90年代提出以来，已成为全球通用的生态可持续性评估工具。该指标通过将人类对耕地、林地、牧场、水域、碳吸收土地等不同生态功能土地的占用需求统一换算为“全球公顷”（GHA）单位，进而与该地区的生物承载力（Biocapacity）进行比较，揭示资源使用是否超过生态系统的自然再生能力。

本研究基于 National Footprint Accounts 2018 数据集中关于全球多个国家的生态足迹与生物承载力的人均数据，结合社会经济变量如人均GDP，构建多维度的国家生态特征向量。相较于单一指标评估，本研究更注重整体生态结构的空间特征与差异模式。为此，本文引入 DBSCAN（基于密度的聚类算法）和谱聚类（Spectral Clustering） 两种无监督学习方法，对各国生态资源使用行为进行聚类分析。

其中，DBSCAN 算法适合发现不规则形状的簇，并能有效识别生态异常国家或孤立型国家（如生态占用极高或极低的小型国家）；而谱聚类则通过构建相似度矩阵和图论分割的方法，从整体结构上揭示国家间生态模式的潜在聚集性，适合处理高维、多变量和非线性关系的数据结构。通过两种方法的互补分析，可以更全面地识别出“生态承压型”、“资源高消耗型”、“资源盈余型”、“碳足迹集中型”等不同类别的国家群体，为生态战略划分和可持续发展政策制定提供有价值的参考依据。

本研究不仅从数据建模的角度探讨生态足迹的聚类特征，也尝试为全球环境治理提供数据驱动的支持路径。通过对全球国家在同一时间截面下的生态表现进行无监督分类，有助于揭示全球生态不平衡的空间结构，推动各国在生态保护与经济发展的平衡中实现更科学的资源配置和绿色转型路径。

文献综述

生态足迹模型由 Rees 和 Wackernagel（1996）提出，是衡量人类对自然资源需求与地球生态承载能力之间关系的经典工具。该方法目前已广泛用于国家比较、城市规划、资源管理等多个领域，并与多种机器学习方法结合，用于深入挖掘生态模式背后的数据结构。

在方法研究方面，传统的生态评估多基于单一指标或线性分析，难以揭示国家间在生态特征方面的复杂结构差异。随着大数据分析技术的发展，聚类分析逐渐被引入生态足迹研究。DBSCAN（Density-Based Spatial Clustering of Applications with Noise）是一种基于密度的聚类算法，能识别任意形状的簇，并自动剔除异常值，适用于生态足迹这类存在“极端国家”（如资源极丰富或极匮乏）的数据结构。例如，Liu 等（2017）使用 DBSCAN 对亚洲国家的碳足迹进行了分层识别，发现了典型的“高碳密集型经济体”聚集群。

另一方面，谱聚类（Spectral Clustering）作为近年来广泛应用于高维非线性数据的图论方法，通过构造相似度矩阵并对图的拉普拉斯矩阵进行特征值分解，能挖掘出国家之间隐藏的生态相似结构。Zhou 等（2020）在“一带一路”沿线国家资源效率研究中引入谱聚类算法，成功划分出资源使用模式显著不同的国家群组，为区域可持续战略制定提供了依据。

综上，DBSCAN 和谱聚类作为两种互补的聚类方法，在生态数据建模中具备显著优势。前者擅长识别异常和自然边界，后者更擅长挖掘数据间深层结构关系。将这两种算法应用于国家生态足迹数据，有助于全面把握资源使用模式的空间分布和内在分层结构，为可持续发展提供有力的数据支持。

# 1. 算法介绍

1.1 DBSCAN算法

1.1.1基本概念：

（1）E邻域：给定对象半径为E内的区域称为该对象的E邻域。

（2）核对象：如果一个对象E邻域内的样本点数大于等于事先给定的最小样本点数MinPts，则称该对象为核对象。

（3）直接密度可达：给定一个对象集合D，核对象p的E邻域内的样本点q，称对象q从对象p出发是可直接密度可达。

（4）密度可达：对于样本集合D，给定一串样本点,,... ,, p=,q=，假如对象从直接密度可达，那么对象q从对象p密度可达。

（5）密度相连：对于样本集合D中的任意一点O，如果存在对象p到对象o密度可达，并且对象q到对象o密度可达，那么对象q到对象p密度相连。

1.1.2 DBSCAN算法实现

输入：输入：数据集D，参数MinPts, ε

输出：簇集合

(1) 首先将数据集D中的所有对象标记unvisited ；(2) do(3) 从D中随机选取一个unvisited对象p，并将p标记为visited ；(4) if p的ε邻域包含的对象数至少为MinPts个 (5) 创建新簇C ，并把p添加到c中； (6) 令N为p的 ε 邻域 中对象的集合；(7) for N 中每个点pi(8) if pi 是unvisited (9) 标记pi 为visited； (10) if pi 的ε邻域至少有MinPts个对象，把这些对象添加到N ；(11) if pi 还不是任何簇的对象。将 pi 添加到 簇C中 ；(12) end for(13) 输出C(14) Else 标记p 为噪声(15) Untill 没有标记为unvisited 的对象

1.2谱聚类

1.2.1 基本概念：

（1）图：由若干点及连接两点的线所构成的图形，通常用来描述某些事物之间的某种关系，用点代表事物，线表示对应两个事物间具有这种关系。

（2）损失函数：



（3）邻接矩阵:称权重矩阵，是由任意两点之间的权重值组成的矩阵。

（4）度矩阵：对于图中的任意一个点，它的度定义为和它相连的所有边的权重之和，利用每个点度的定义，可以得到一个nxn的度矩阵 。它是一个对角矩阵，只有主对角线有值，对应第i行第i个点的度数。

（5）Laplacian：L=D-W, L称为拉普拉斯矩阵，W为权重矩阵，D为度矩阵。

1.2.2 谱聚类算法实现：

Unnormalized Spectral Clustering

输入：样本及类别数K

1、根据样本建立权重矩阵W；

2、根据W，计算度矩阵D，进而计算拉普拉斯矩阵L；

3、计算L的特征值及特征向量；

4、取出前K小特征值对应的特征向量，并对矩阵的行向量进行聚类，得到K个Clusters

Normalized Spectral Clustering

输入：样本及类别数K

1、根据样本建立权重矩阵W；

2、计算拉普拉斯矩阵L及

3、计算的特征值及特征向量；

4、取出前K小特征值对应的特征向量，并对矩阵的行向量进行聚类，得到K个Clusters

# 2.实例分析

2.1 数据来源

本研究所使用的数据集来自于Kaggle平台上的 National Footprint Accounts 2018，由 Global Footprint Network 提供。该数据集涵盖了全球将近200个国家从1961年至2014年的生态足迹（Ecological Footprint）与生物承载力（Biocapacity）数据，广泛用于评估国家层面的生态可持续性。

2.2数据描述

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **变量名称** | **含义** | **类型** | **单位** | **示例值（Armenia, 1992）** |
| **country** | 国家名称 | 分类型 | - | Armenia |
| **ISO alpha-3 code** | 国家三位字母代码 | 分类型 | - | ARM |
| **UN\_region** | 联合国地区分类 | 分类型 | - | Asia |
| **UN\_subregion** | 联合国子区域分类 | 分类型 | - | Western Asia |
| **year** | 数据年份 | 数值型 | 年 | 1992 |
| **record** | 数据类别（指标类型） | 分类型 | - | EFConsPerCap |
| **crop\_land** | 耕地占用生态足迹 | 数值型 | global ha | 0.3909 |
| **grazing\_land** | 牧场占用生态足迹 | 数值型 | global ha | 0.1891 |
| **forest\_land** | 森林资源占用生态足迹 | 数值型 | global ha | 0.000001 |
| **fishing\_ground** | 渔业资源占用生态足迹 | 数值型 | global ha | 0.0041 |
| **built\_up\_land** | 建设用地生态足迹 | 数值型 | global ha | 0.0337 |
| **carbon** | 碳足迹（能源消耗类） | 数值型 | global ha | 1.1122 |
| **total** | 总生态足迹（各项之和） | 数值型 | global ha | 1.7301 |
| **Percapita GDP** | 人均国内生产总值（2010 美元） | 数值型 | USD | 949.033 |
| **population** | 人口总数 | 数值型 | 人 | 3,449,000 |

表1：数据变量描述

2.3数据预处理

在数据预处理阶段，依次筛选出五类核心指标（BiocapTotGHA：总生物承载、EFConsTotGHA：生态足迹总消费量、EFExportsTotGHA：生态足迹总出口量、EFProdTotGHA：生态足迹总生产量、EFImportsTotGHA：生态足迹总进口量）对应的全部年份数据。对于每类数据，首先剔除所有存在缺失值的记录，确保分析样本的完整性。随后，以国家为单位对各项生态足迹和经济特征进行均值处理，获得每个国家的标准化特征均值数据，为后续的聚类分析提供了统一、无缺失的数据基础。

2.4 数据分析

表2：五类生态足迹/承载力指标特征均值汇总表

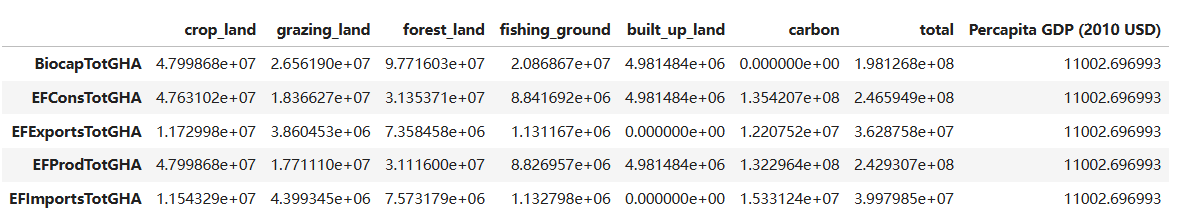


图1：各区域国家数量条形图

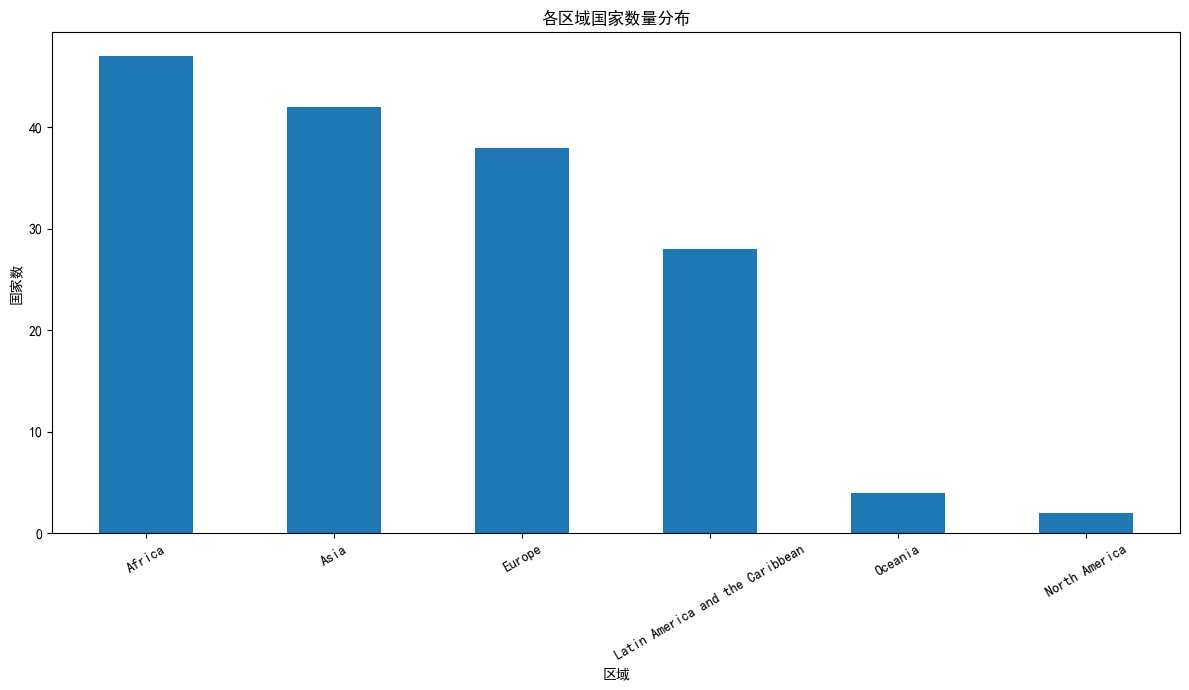
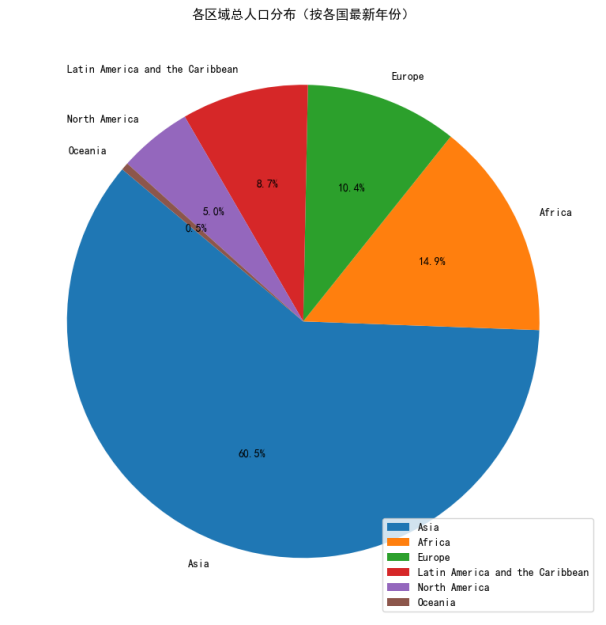


图2：各区域总人口分布扇形图



2.5聚类分析

2.5.1 DBSCAN聚类

表3：基于BiocapTotGHA指标的DBSCAN聚类国家分组

|  |  |  |
| --- | --- | --- |
| **​**​Cluster Label​**​** | **​**​Number of Countries​**​** | **​**​Country List​**​** |
| **​**​-1​**​ (Outliers)** | 8 | Australia, Brazil, Canada, China, India, Russian Federation, United States of America, World |
| **​**​0​**​ (Core Cluster)** | 154 | Afghanistan, Albania, Algeria,..., Zambia, Zimbabwe |

表4：基于EFConsTotGHA指标的DBSCAN聚类国家分组

|  |  |  |
| --- | --- | --- |
| **​**​Cluster Label​**​** | **​**​Number of Countries​**​** | **​**​Country List​**​** |
| **​**​-1​**​** | 6 | Brazil, China, India, Japan, ..., United States of America, World |
| **​**​0​**​** | 156 | Afghanistan, Albania, Algeria, Angola, ..., Zambia, Zimbabwe |

表5：基于EFExportsTotGHA指标的DBSCAN聚类国家分组

|  |  |  |
| --- | --- | --- |
| ****Cluster Label​**​** | **​**​Number of Countries​**​** | **​**​Country List​**​** |
| **​**​-1​**​** | 12 | Australia, Canada, China, Denmark, Ecuador, Germany, New Zealand, Norway, Russian Federation, Saudi Arabia, United States of America, World |
| **​**​0​**​** | 141 | Afghanistan, Albania, Algeria, Angola, Argentina, ..., Zambia, Zimbabwe |
| **​**​1​**​** | 5 | France, Japan, Netherlands, Spain, United Kingdom |
| **​**​2​**​** | 4 | Luxembourg, Qatar, Switzerland, United Arab Emirates |

表6：基于EFProdTotGHA指标的DBSCAN聚类国家分组

|  |  |  |
| --- | --- | --- |
| **​**​Cluster Label​**​** | **​**​Number of Countries​**​** | **​**​Country List​**​** |
| **​**​-1​**​** | 5 | Brazil, China, India, United States of America, World |
| **​**​0​**​** | 157 | Afghanistan, Albania, Algeria, Angola, ..., Zambia, Zimbabwe |

表7：基于EFImportsTotGHA指标的DBSCAN聚类国家分组

|  |  |  |
| --- | --- | --- |
| ****Cluster Label​**​** | **​**​Number of Countries​**​** | **​**​Country List​**​** |
| **​**​-1​**​** | 8 | China, Denmark, Germany, Japan, Russian Federation, Spain, United States of America, World |
| **​**​0​**​** | 150 | Afghanistan, Albania, Algeria, Angola, ..., Zambia, Zimbabwe |
| **​**​1​**​** | 4 | France, Italy, Netherlands, United Kingdom |

2.5.2谱聚类

表8：基于BiocapTotGHA指标的谱聚类国家分组

|  |  |  |
| --- | --- | --- |
| ****Cluster Label​**​** | **​**​Number of Countries​**​** | **​**​Country List​**​** |
| **​**​0​**​** | 58 | Albania, Algeria, Argentina, Armenia, Azerbaijan, ..., Ukraine, World |
| **​**​1​**​** | 60 | Afghanistan, Angola, Bangladesh, Benin, Bhutan, ..., Zimbabwe |
| **​**​2​**​** | 25 | Australia, Austria, Belgium, Brunei, Canada, ..., USA |
| **​**​3​**​** | 19 | Bahamas, Bahrain, Barbados, Croatia, Czech Republic, ..., Venezuela |

表9：基于EFConsTotGHA指标的谱聚类国家分组

|  |  |  |
| --- | --- | --- |
| **​**​Cluster Label​**​** | **​**​Number of Countries​**​** | **​**​Country List​**​** |
| **​**​0​**​** | 54 | Afghanistan, Bangladesh, Benin, Bhutan, ..., Zambia, Zimbabwe |
| **​**​1​**​** | 64 | Albania, Algeria, Angola, Argentina, ..., Ukraine, World |
| **​**​2​**​** | 19 | Bahamas, Bahrain, Barbados, Croatia, ..., Venezuela |
| **​**​3​**​** | 25 | Australia, Austria, Belgium, Brunei, ..., United States of America |

表10：基于EFExportsTotGHA指标的谱聚类国家分组

|  |  |  |
| --- | --- | --- |
| **​**​Cluster Label​**​** | **​**​Number of Countries​**​** | **​**​Country List​**​** |
| **​**​0​**​** | 42 | Argentina, Belgium, Brazil, Canada, China, ..., World, Yemen |
| **​**​1​**​** | 44 | Afghanistan, Bangladesh, Benin, Burkina Faso, ..., Zambia, Zimbabwe |
| **​**​2​**​** | 33 | Australia, Austria, Bahamas, Bahrain, ..., Uruguay, Venezuela |
| **​**​3​**​** | 43 | Albania, Algeria, Angola, Armenia, ..., Turkey, Turkmenistan |

表11：基于EFProdTotGHA指标的谱聚类国家分组

|  |  |  |
| --- | --- | --- |
| **​**​Cluster Label​**​** | **​**​Number of Countries​**​** | **​**​Country List​**​** |
| **​**​0​**​** | 65 | Albania, Algeria, Angola, Argentina, ..., Ukraine, World |
| **​**​1​**​** | 53 | Afghanistan, Bangladesh, Benin, Bolivia, ..., Zambia, Zimbabwe |
| **​**​2​**​** | 25 | Australia, Austria, Belgium, Brunei, ..., United States of America |
| **​**​3​**​** | 19 | Bahamas, Bahrain, Barbados, Croatia, ..., Venezuela |

表12：基于EFImportsTotGHA指标的谱聚类国家分组

|  |  |  |
| --- | --- | --- |
| ****Cluster Label​**​** | **​**​Number of Countries​**​** | **​**​Country List​**​** |
| **​**​0​**​** | 51 | Albania, Angola, Argentina, Armenia, ..., Tunisia, Turkmenistan |
| **​**​1​**​** | 35 | Algeria, Belarus, Belgium, Brazil, ..., United States of America, World |
| **​**​2​**​** | 27 | Australia, Austria, Bahamas, Bahrain, ..., Uruguay, Venezuela |
| **​**​3​**​** | 49 | Afghanistan, Bangladesh, Benin, Bhutan, ..., Zambia, Zimbabwe |

2.5.3聚类结果分析

（1）总生物承载力（BiocapTotGHA）聚类分析  
DBSCAN方法将154个国家归为主类，8个国家（包括中国、美国、印度、俄罗斯、巴西、加拿大、澳大利亚）被识别为极端值，这些国家大多为国土面积大、资源丰富或者人口众多的全球性大国。  
谱聚类将国家分为4类，不同类别主要反映了资源禀赋与经济发展水平的分异。具体来看，一类主要聚集了西欧、北欧及北美等发达国家，展现出高人均生物承载力与高经济发展水平的特征；另一类以新兴经济体和部分资源型国家为主，表现出中等生物承载力与较高经济增长；还有一类主要涵盖中低收入发展中国家，资源禀赋有限，生物承载力和经济水平均处于较低水平；最后一类为小型或特殊经济结构国家，显示出独特的生态与经济特征。

（2）生态足迹消费总量（EFConsTotGHA）聚类分析  
在EFConsTotGHA指标下，DBSCAN同样将绝大多数国家（156个）聚为主类，仅有中国、美国、印度、日本、巴西等6国为极端点，说明大国在人类生态消费上具有显著的极端性。  
谱聚类下，国家划分为4类，展现出不同经济区位下的消费差异：一类为高消费型发达经济体，整体生态足迹消费较高，主要为欧美、部分亚洲发达国家；第二类为新兴工业化国家，经济增长快但消费总量与发达国家尚有差距；第三类为资源型经济体或出口导向国家，生态足迹消费受国际贸易影响显著；最后一类为普通发展中国家，生态足迹消费水平较低，主要受人口规模和经济基础限制。谱聚类结果揭示了消费模式和发展阶段的多样性。

（3）生态足迹出口总量（EFExportsTotGHA）聚类分析  
EFExportsTotGHA的DBSCAN聚类分布更为细化，主类包含141国，12个国家为极端异常点，且有多个小类。这些极端点主要为出口规模巨大或特殊经济结构国家（如中国、美国、俄罗斯、德国、沙特等）。  
谱聚类将各国分为4类：第一类为高度出口型经济体，通常拥有较强的制造业或能源出口能力（如中国、美国、德国等）；第二类为中等出口国家，经济和贸易较为均衡；第三类主要涵盖资源出口国和高收入小国，其生态足迹出口主要依赖单一或少数产业；第四类为经济体量有限、出口总量较小的国家。整体上，谱聚类较好反映了国际分工及出口导向特征，突出各国在全球生态贸易链中的地位。

（4）生态足迹生产总量（EFProdTotGHA）聚类分析  
EFProdTotGHA聚类结果显示，DBSCAN主类有157国，5国为极端异常点（包括中国、美国、印度、巴西），同样表现为全球生产端的极端国家。  
谱聚类将各国细分为四类：第一类为资源富集型发达国家，生产总量和经济发展均处于高位，如美国、加拿大、澳大利亚等；第二类为新兴工业化国家，生产总量逐年提升，主要包括亚洲和拉美经济体；第三类为普通发展中国家，生态生产总量处于中等水平；第四类涵盖特殊经济结构国家和经济体量较小的国家。各类国家在生产生态足迹上呈现显著分异，体现出全球产业格局的不均衡与多层次性。

（5）生态足迹进口总量（EFImportsTotGHA）聚类分析  
在EFImportsTotGHA指标下，DBSCAN识别出150个主类国家，8个极端点（如中国、美国、日本、德国、俄罗斯、西班牙等），还有小类聚集发达欧洲国家（如法国、意大利、荷兰、英国等）。  
谱聚类同样划分为四类：一类为高收入发达经济体，进口生态足迹总量大，对全球资源依赖性高；第二类为资源输出国及出口导向型新兴经济体，进口量相对有限；第三类为人口大国，进口规模大但人均进口水平有限；第四类为一般发展中国家，进口生态足迹总量处于低位。各类别反映了进口足迹对经济体量、产业结构和区域特征的敏感性，突显不同经济发展阶段国家间的差异。

# 3.结论

本文基于全球生态足迹数据库，选取总生物承载力、生态足迹消费总量、生态足迹出口总量、生态足迹生产总量和生态足迹进口总量等五项生态与经济核心指标，综合应用DBSCAN与谱聚类算法，对196个国家的生态经济特征进行了系统的无监督聚类分析。

通过DBSCAN聚类结果可以发现，无论是生态承载力还是生态足迹相关指标，各项指标下均存在极少数“极端国家”，这些国家往往因国土面积大、人口众多或资源禀赋突出，形成了明显的异常点分布。大部分国家则表现为“中位数”特征，处于主类之中。DBSCAN算法在自动识别异常值和异质性群体方面表现突出，尤其适用于全球尺度下具有高度结构不均衡的数据集。

谱聚类算法则进一步揭示了各国之间生态经济特征的内部结构与分层。不同指标下，谱聚类普遍将国家分为四个类别，类别间反映了经济发展水平、资源禀赋、人口规模和产业结构等多重差异。高收入发达国家、新兴工业化国家、资源型或出口导向国家以及一般发展中国家在聚类结构中表现出各自鲜明的生态经济特征。谱聚类的结果揭示了全球生态经济格局的复杂性和多样性，为进一步的区域合作、分层治理和可持续发展战略制定提供了理论参考。

总体而言，DBSCAN与谱聚类两种方法在全球生态足迹数据分析中优势互补：DBSCAN善于识别边界与异常，谱聚类擅长刻画复杂群体结构。本文的聚类分析不仅揭示了全球范围内生态经济特征的异质性和空间分布格局，也为不同发展阶段国家的生态管理与政策制定提供了数据支持和分层依据。未来研究可结合更多社会、环境和政策变量，进一步丰富全球生态足迹的分层诊断与动态演化研究。

具体来说，高收入发达国家应重点关注生态足迹的减量管理和绿色转型，推动资源循环利用、清洁能源替代及技术创新，带动全球可持续发展进程。新兴工业化国家则需要兼顾经济增长与生态保护，在扩大工业产能的同时强化环境治理与资源高效利用，防止走上高污染高消耗的发展路径。资源型或出口导向国家建议强化对出口产品生态足迹的管理，提高产品附加值，减少对初级资源的过度依赖，同时引入绿色认证等机制参与全球生态治理。中低收入发展中国家则应在基础生态保障、人口压力缓解和生态修复方面加强国际合作，争取技术转让和资金支持，提升自身可持续发展能力。

未来，全球不同类型国家在生态政策制定上需要根据自身聚类特征，制定差异化、分层次的管理策略，实现全球范围内的生态环境协同保护和绿色发展。

**参 考 文 献**

【1】 Rees, W., & Wackernagel, M. (1996). Our Ecological Footprint: Reducing Human Impact on the Earth. New Society Publishers.

【2】 Liu, S., Xu, X., & Ma, J. (2017). Carbon footprint clustering and driving factor analysis of Asian countries based on DBSCAN. *Ecological Indicators*, 76, 154-163.

【3】 Zhou, Y., Ma, L., & Li, X. (2020). Resource efficiency and group identification in Belt and Road countries based on spectral clustering. *Journal of Cleaner Production*, 258, 120624.

**附录**

import pandas as pd

from sklearn.preprocessing import StandardScaler

from sklearn.cluster import DBSCAN, SpectralClustering

from IPython.display import display

def cluster\_by\_record(df, features, record\_type, eps=1.2, min\_samples=3, n\_clusters=4):

print(f"\n{'='\*30}\n分析类型：{record\_type}\n{'='\*30}")

# 1. 只保留指定 record 的所有年份数据

df\_sub = df[df['record'] == record\_type].copy()

# 2. 删除有空值的行

df\_sub\_filtered = df\_sub.dropna(subset=features)

# 3. 按国家对特征取均值

country\_mean = df\_sub\_filtered.groupby('country')[features].mean()

# 4. 标准化

scaler = StandardScaler()

X\_scaled = scaler.fit\_transform(country\_mean[features].values)

# 5. DBSCAN聚类

dbscan = DBSCAN(eps=eps, min\_samples=min\_samples)

db\_labels = dbscan.fit\_predict(X\_scaled)

country\_mean['DBSCAN\_Label'] = db\_labels

# 6. 谱聚类

sc = SpectralClustering(n\_clusters=n\_clusters, affinity='nearest\_neighbors', random\_state=42)

spectral\_labels = sc.fit\_predict(X\_scaled)

country\_mean['Spectral\_Label'] = spectral\_labels

# 7. 输出聚类分布

print(f"{record\_type} DBSCAN 聚类分布：")

print(country\_mean['DBSCAN\_Label'].value\_counts())

print(f"{record\_type} 谱聚类分布：")

print(country\_mean['Spectral\_Label'].value\_counts())

# 8. 输出DBSCAN每类国家名单

for label in sorted(country\_mean['DBSCAN\_Label'].unique()):

print(f"\n{record\_type}-DBSCAN标签 {label}（国家数：{(country\_mean['DBSCAN\_Label']==label).sum()}）:")

print(country\_mean[country\_mean['DBSCAN\_Label']==label].index.tolist())

# 9. 输出谱聚类每类国家名单

for label in sorted(country\_mean['Spectral\_Label'].unique()):

print(f"\n{record\_type}-谱聚类标签 {label}（国家数：{(country\_mean['Spectral\_Label']==label).sum()}）:")

print(country\_mean[country\_mean['Spectral\_Label']==label].index.tolist())

# 10. 显示表头

display(country\_mean.head())

print("\n" + "="\*40 + "\n")

return country\_mean

# ----------- 主程序部分 ----------------

# 1. 读取数据和特征列

df = pd.read\_csv('NFA 2018.csv')

features = ['crop\_land', 'grazing\_land', 'forest\_land', 'fishing\_ground',

'built\_up\_land', 'carbon', 'total', 'Percapita GDP (2010 USD)']

print(f"原始国家数量：{df['country'].nunique()}")

# 2. 循环分析不同record类型

for record\_type in ['BiocapTotGHA', 'EFConsTotGHA', 'EFExportsTotGHA', 'EFProdTotGHA', 'EFImportsTotGHA']:

cluster\_by\_record(df, features, record\_type, eps=1.2, min\_samples=3, n\_clusters=4)

mean\_table = []

record\_types = ['BiocapTotGHA', 'EFConsTotGHA', 'EFExportsTotGHA', 'EFProdTotGHA', 'EFImportsTotGHA']

for record\_type in record\_types:

df\_sub = df[df['record'] == record\_type].dropna(subset=features)

row = df\_sub[features].mean()

row.name = record\_type

mean\_table.append(row)

# 合并成一个表

mean\_df = pd.DataFrame(mean\_table)

print("表1 五类生态足迹/承载力指标特征均值汇总表\n")

display(mean\_df)

import matplotlib.pyplot as plt

# 筛选各国，只保留每个国家的唯一值

unique\_countries = df\_cleaned[['country', 'UN\_region']].drop\_duplicates()

# 去掉 country 为 'World' 的记录

unique\_countries\_no\_world = unique\_countries[unique\_countries['country'] != 'World']

# 统计每个区域的国家数量

region\_counts = unique\_countries\_no\_world['UN\_region'].value\_counts().sort\_values(ascending=False)

# 画条形图

plt.figure(figsize=(12, 7))

region\_counts.plot(kind='bar')

plt.xlabel('区域')

plt.ylabel('国家数')

plt.title('各区域国家数量分布')

plt.xticks(rotation=30)

plt.tight\_layout()

plt.show()

# 对每个国家找到最新年份的数据，再取出人口与区域

idx = df\_filtered.groupby('country')['year'].idxmax()

latest\_pop\_info = df\_filtered.loc[idx, ['country', 'UN\_region', 'population']]

# 按区域汇总人口总量

region\_pop\_sum = latest\_pop\_info.groupby('UN\_region')['population'].sum().sort\_values(ascending=False)

import matplotlib.pyplot as plt

import matplotlib

matplotlib.rcParams['font.sans-serif'] = ['SimHei']

matplotlib.rcParams['axes.unicode\_minus'] = False

# 1. 去除 'World' 区域

region\_pop\_sum = region\_pop\_sum.drop(labels=['World'], errors='ignore')

# 2. 画扇形图

plt.figure(figsize=(8, 8))

region\_pop\_sum.plot(

kind='pie',

autopct='%1.1f%%',

startangle=140,

ylabel='',

legend=True

)

plt.title('各区域总人口分布（按各国最新年份）')

plt.tight\_layout()

plt.show()