# 中央财经大学

Central University of Finance and Economics


课　　程　大数据探索性可视化

论文题目　地球自然环境可视化分析

学　　校　中央财经大学

姓　　名　甘鹏程 王思雨 冯洋洋

指导老师　吴翌琳

论文日期　2018/01/04

# 一、资源与地球环境可视化分析

## 1.1 资源数据变量说明

本报告部分使用了 World Bank Data 世界银行数据关于地球环境的部分的污染状况数据，本部分报告使用的数据全部来着该数据集的五个子部分，分别是：3.4_Deforestation_and_biodiversity 数据表、3.5_Freshwater 数据表、3.6_Energy_production_and_use 表、3.7_Electricity_production_sources_and_access 数据表。通过对数据表上的数据进行基本预处理和获取，对表中所记录的淡水、电力生产、生物多样性、能源等关于资源使用状况的可视化描述统计分析。在这些表中所涉及的变量有如下表格：

表 1-1 地球环境与污染主要变量说明表

| 变量类型 | 变量名 | 取值范围 | 详细说明 | 备注 |
|---|---|---|---|---|
| 可再生淡水资源 | **国内淡水流总量** Flows | [0, 42810] **单位：亿立方米** | 数值型变量 | 代表某年该国可再生的淡水资源总量 |
| | **国内人均持有淡水资源量** Per.capita | [0，519, 265] **单位：立方米** | 数值型变量 | 代表某年该国可再生的淡水资源人均量 |
| 森林退化和生物多样性 | **年平均森林退化率** deforest | [-6.68，614] **单位：%** | 数值型变量 | 衡量一段时间内某一地区森林面积变化的平均速率 |
| | **受威胁品种数** Threatened species | [0，2058] **单位：种** | 数值型变量 | 衡量某一年某一地区受威胁物种数 |
| 电力的生产与来源 | **电力生产总量** production | [0，5, 665.7] **单位：亿千瓦时** | 数值型变量 | 某一年该国家的地理生产总值 |
| | **电力生产来源** Sources | Coal<br>Natural gas<br>Oil<br>Hydropower<br>Renewable sources<br>Nuclear power | 分类变量/百分比 | 衡量各个电力生产来源的比例 |
| 能源生产与使用 | **能源消耗总量** Production | [0，8569.7] **单位：百万吨油当量** | 数值型变量 | 衡量某年某地区能源消耗总量当量总值 |
| | **能源消耗人均量** Per. capita | [0，1893] **单位：千克油当量** | 数值型变量 | 衡量某年某地区能源消耗总量当量人均值 |

| | | | | |
|---|---|---|---|---|
| 国家信息 | 国家代码 Country Code | 240 国家英文简称 | 文本信息 | 作为数据表的键值，用来连接数据表以及索引表 |
| | 收入分类 IncomeGroup | Upper middle income<br>High income<br>Lower middle income<br>Low income | 分类变量 | 不同收入水平的国家对能源的利用效率、资源消耗不太 |
| | 所属大洲 Region | East Asia & Pacific<br>Europe & Central Asia<br>Latin America & Caribbean<br>Middle East & North Africa<br>North America<br>South Asia<br>Sub-Saharan Africa | 分类变量 | 所处的大洲地理位置不同，由于地理环境的不同资源保有和使用也不同 |

## 1.2 地球资源描述性统计分析

### 1.2.1 2014 年全球多国家水资源排位图

2014 年全球各个国家，内部可再生淡水资源的分布，为了可视化各个国家淡水总量和人均量的特点，将总量和人均量进行排位处理；横轴表示淡水总量的排位，纵轴表示人均拥有淡水量的排位，从图 1-1 中看出。

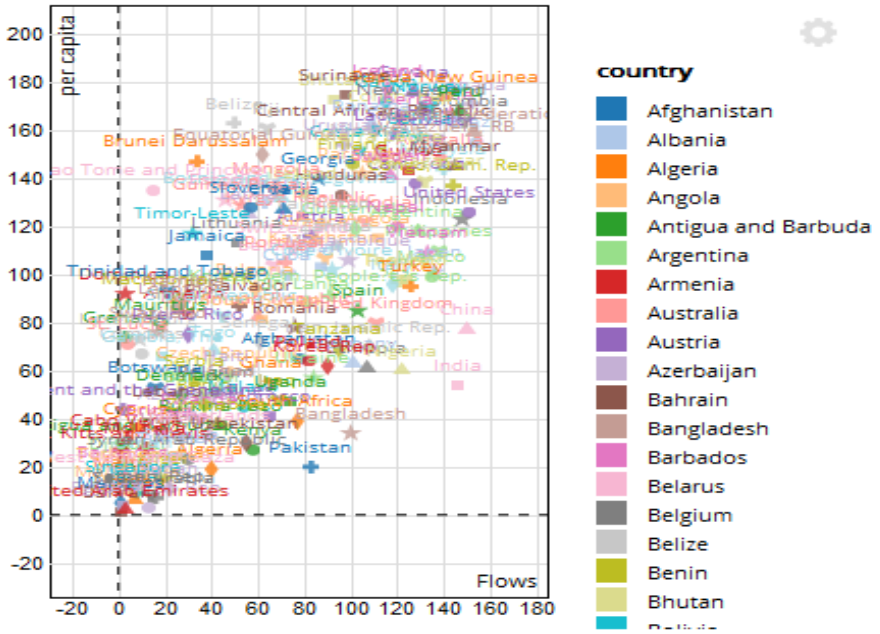


图 1-1 2014 年全球 200 多个国家可再生淡水资源的排位情况

可以看到像 Bahrain,United Arb Emirate 等国家可再生水资源和人均资源均处于最少的状况，而 Dominica,Grenada,Sao Tome and Principe 等国家可再生淡水的总量少但是人均在所有国家中位列前列。Chnia,India 等国家相反，有着丰富的淡水资源但是人均持有量缺排位十分靠后。两个指标都比较靠前的国家有 Canada,Peru,Colombia,Russia 等。整体的分布大概呈菱形，即总量和人均量的排名接近，但可以看出不少国家在左下角聚集，总量和人均都十分紧张。

### 1.2.2 2014 年世界不同大洲地区水资源保有和使用情况热力图

该图反应的 2014 世界不同大洲的水资源的保有和使用情况，该散点图包含了三个指标的信息，横着代表可再生淡水资源总量，纵轴代表人均量，图中各个类别散点的颜色代表每年使用的水量占当年可总量的比率：

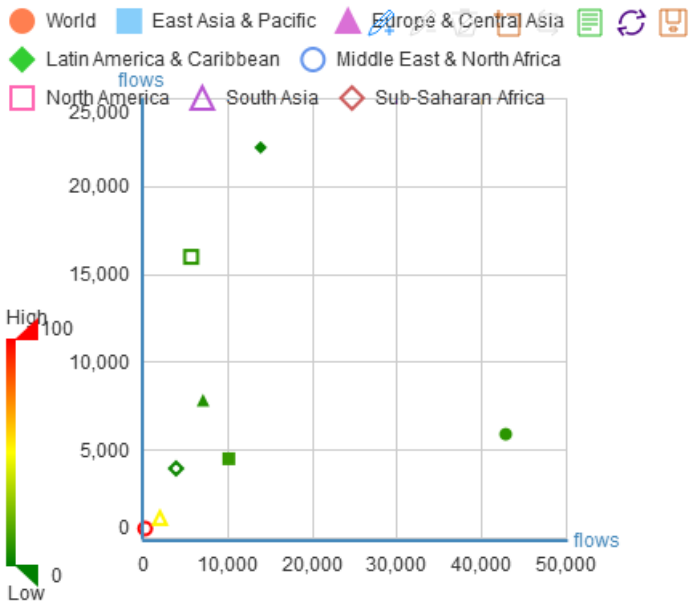


图 1-2 七个地区与全球平均的可再生淡水资源保有量和使用程度

从图 1-2 中可以首先可以看到各大洲水资源的保有情况，Middle East & North Africa 和 South Asia 的总量和人均量绝对值都比较低，特别是前者。世界平均水平在右下角，总量高而人均低的情况，Latin America & Caribbean 地区的总量和人均量均处于最高的位置，总量约占 1/3，但人均远高于世界平均水平和其他地区。Sub-Saharan Africa，North America，Europe & Central Asia，East Asia & Pacific 的总量依次递增，最低的月 4000 billion cu. m，最高约 10000 billion cu. m，但人均差距显著，North America 要远高于其他三个地区。散点的颜色热度表示使用情况，世界平均水平是 9.3%，只有三个地区超过了这个比率,East Asia&Pacific 略高于世界平均水平,但是 South Asia 超过 50%，Middle East & North Africa 更是达到了 138%远高于其当年可再生淡水总量。

### 1.2.3 2014 年不同收入地区水资源保有和使用情况

图 1-3 反应的是不同收入等级的地区，在淡水资源保有量和使用情，总共有四个类别，低收入地区、中的收入地区、中高收入地区、高收入地区。横轴代表 2014 可再生淡水总量，纵轴代表人均保有量，图型面积代表使用率。通过 R 软件绘制地图结果如下图 1-3 所示：

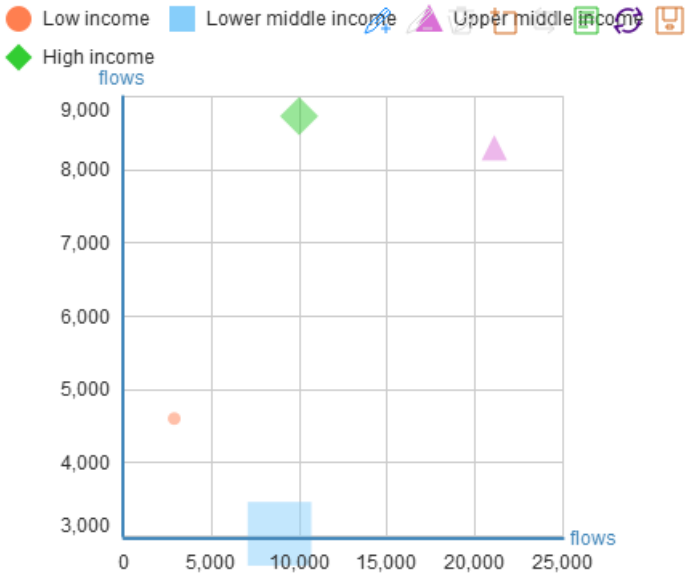


图 1-3 2014 年四个不同收入等级的可再生淡水资源的保有和消耗

从图 1-3 中可以看到，中高收入地区位于右上角，高总量和人均量，使用率也属于中度。中低收入地区和高收入地区有接近的总量，但是高收入地区的人均远高于前者，两者在人均保有量上处于两级分化状态。低收入地区在总量上和人均上均处于较低水平，但是使用度也是最低。中低收入地区的使用度是最高的达到了 18%，明显高于其他三类

### 1.2.4 1990 年与 2013 年全球各个国家的能源消耗人均量热力图

对比 1990 年与 2013 年内，全球各个国家的能源消耗人均量，美国、俄罗斯有所减少，中国则明显增加，大部分地区基本没有大变化，基本处于世界中等。如图 1-4 所示。

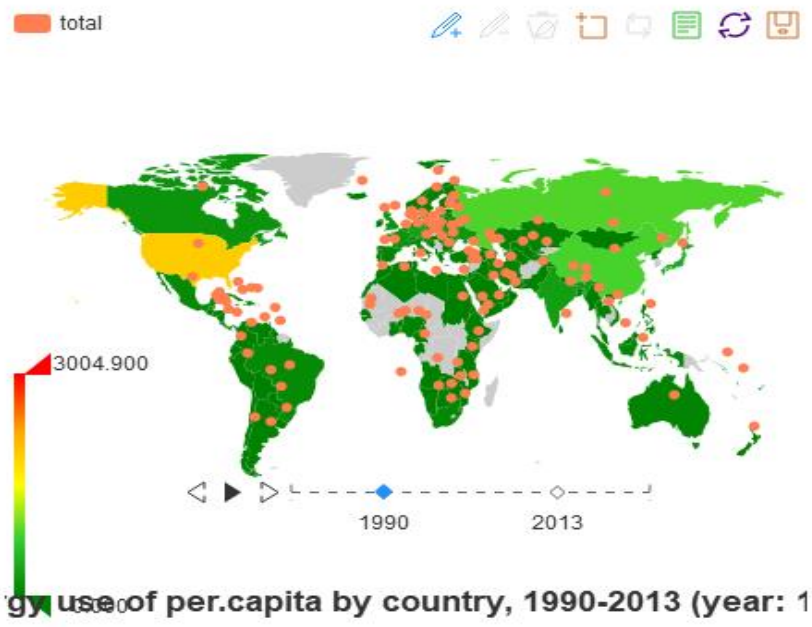


图 1-4 1990 年与 2013 年能源消耗人均量热力图

## 1.2.5 1990 年和 2013 年内全球各个国家的能源消耗总量热力图

从图 1-5 中可以看到，对比 1990 年和 2013 年内，全球各个国家的能源消耗总量，所以国家中，中国增幅明显：从绿色变成了红色。其次在 1990 年，美国的能源消耗总量远高于其他国家。

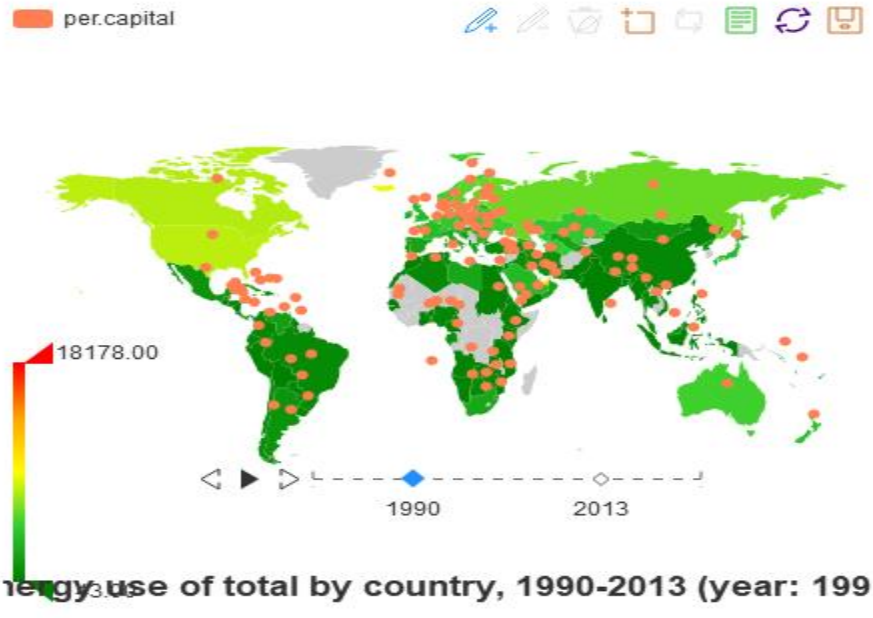


图 1-5 1990 年与 2013 年能源消耗总量热力图

## 1.2.6 1990 年与 2013 年世界各主要地区能源消耗总量和人均量的堆叠直方图

如图 1-6 所示，蓝色和红色代表 2013 年 1990 年的情况，左侧位能源消耗总量右侧为人均量。从图中观察，能源消耗总体量均增长明显，特别亚太地区。但是人均值差异明显，北美地区的人均增加值和保有值均是所以地区之最，远高于世界平均水准。

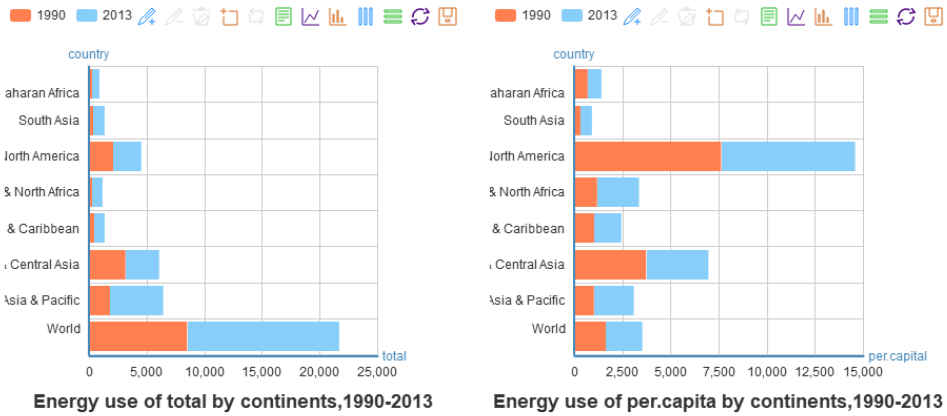


图 1-6 各大洲能源消耗总量和人均量

## 1.2.7 2014 年世界各国电力生产热力图

利用发电量是在一个站的所有交流发电机组的终端测量的生产量数据绘制，如下图 1-7 所示：

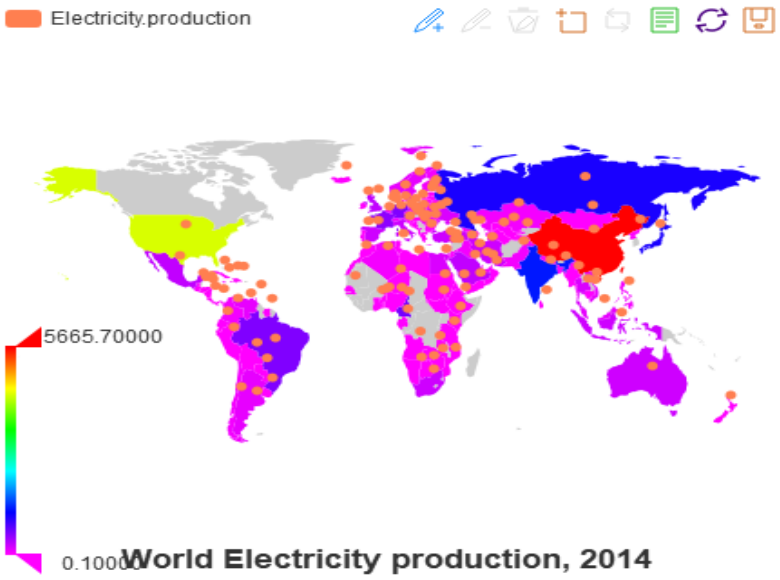


图 1-7 2014 年世界各国电力生产热力图

大部分国家低于平均水平，俄罗斯的生产量较低而美国与中国非常高。

## 1.2.8 2014 年不同收入地区的电力生产来源风玫瑰图

图 1-8 显示低收入地区数据缺失，包含中低收入、中高收入、高收入地区，来源包括：煤（火力）、水利发电、天然气、核电、石油、可再生能源发电。高收入地区煤电的比率要明显低于中高和中低收入地区，后两者的煤电比率均在50%左右。水利和天然气发电是中低和中高收入地区的次来源，天然气和核电是高收入地区的次来源：

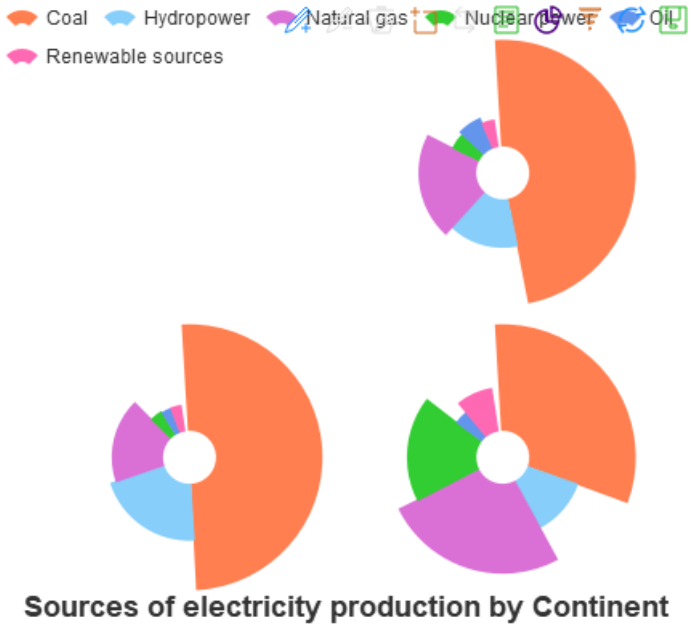


图 1-8 不同收入地区的电力生产来源分布情况

## 1.2.9 2014 年全球各大洲电力生产来源金字塔图

世界平均数据是第一个金字塔，形状偏紧瘦，煤电、天人气、水利、核电、可再生发电、石油依次递减，煤电依然是最主要的方式。East Asia & Pacific 和 South Asia、Sub-Saharan Africa 结构相近。煤电占比最重，其次是水利和天然气，其他来源不多，且后两者的石油发电更占比更多。Europe & Central Asia 是所有金字塔中最饱满的，各个来源占比差距更小。占比最高的是天然气发电，其次是煤电、核电、水利等 Latin America & Caribbean 是位数不多不以煤电为主的地区，其占比最重的来源是水利，其次是天人气。特别是石油发电排在第三位。Middle East & North Africa 地区电力来源最为简单，主要以不可再生能源类为主，占比最重的是天然气发电，其次是石油发电。North America 的比例最接近标准三角形，来源分布均衡。占比最重的是煤电，第二和第三是天然气和核电。

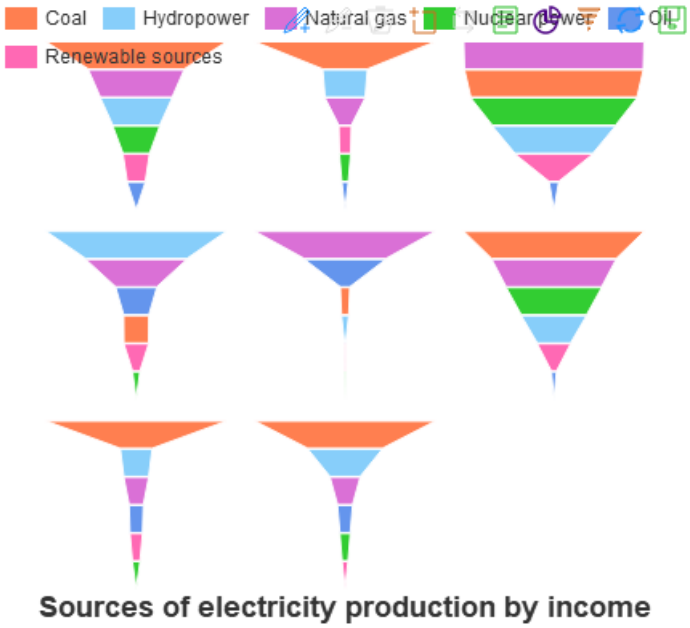


图 1-9 不同大洲地区的电力生产来源分布情况

## 1.2.10 1990-2015 年段年平均自然森林移动率 Top100 直方图

如图 1-10 所示，平均年自然森林增减移动率(变动绝对额最高的前 100 个国家)，左侧表示减少，右侧表示增加，两个系列分别是 1990-2000 和 2000-2015 两个时间段内的增减移动率。更多的国家都出现了更快的森林减少速率。在 1990-2000 和 2000-2015 期间，French Polynesia 森林年流失率均最高，2000-2015 期间 Togo 森林年增加率最高。

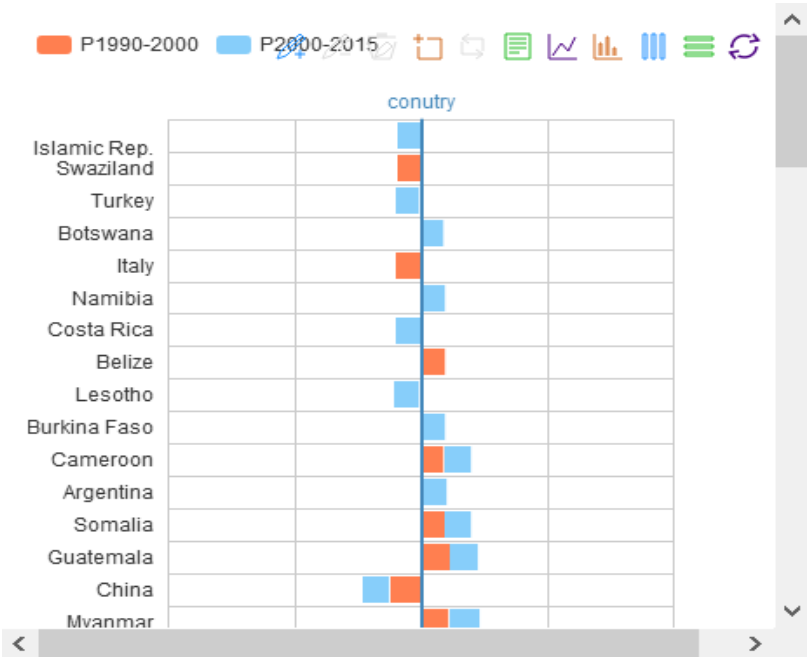


图 1-10 1990-2000 年与 2000-2015 年两个时段变化率最高的 100 各国家的森林变化情况

### 1.2.11 2016 受威胁物种数最高的 20 各国家对比直方图

图 1-11 中的统计数据包括 Mammal ，Bird ，Fish，Plant，species (higher) 四个类别的种类数。最多的国家是 Ecuador，数量是 2058 种，远高于第二名 Malaysia 的 927 种：

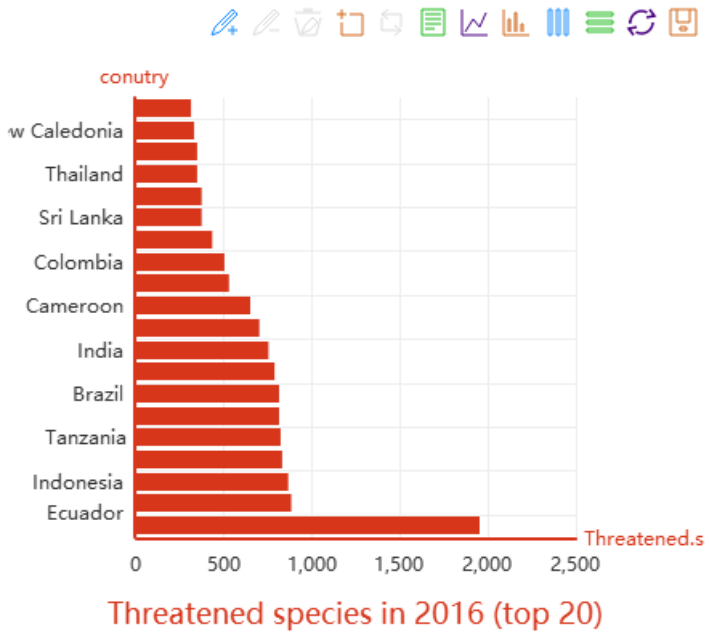


图 1-11 2016 年受威胁物种数最多的 20 各国家

## 二、污染与地球环境可视化分析

### 2.1 污染数据变量说明

本报告部分使用了 World Bank Data 世界银行数据关于地球环境的部分的污染状况数据，主要使用了 5 个 csv 表数据。分别是：CO2_emission.csv 数据表，CO2_emission_index.csv 索引表[1]，CO2_intensity.csv 数据表，CO2_intensity_index.csv 索引表，NO_emission.csv 数据表，NO_intensity_index.csv 索引表，Methane_emission.csv 数据表，Methane_index.csv 索引表，greenhouse gas emissions.csv 数据表，forestdata.csv 数据表以及 world 地理信息表[2]。利用可这些表格进行了地球大气污染和森林破坏等关于环境污染状况的可视化描述统计分析。在这些表中所涉及的变量有如下表格：

[1] 索引表是存储数据表中的国家信息的，用于表连接
[2] 所有表格都存在于压缩包中

表 2-1 地球环境与污染主要变量说明表

| 变量类型 | 变量名 | 取值范围 | 详细说明 | 备注 |
|---|---|---|---|---|
| 空气污染状况 | CO2 排放强度 CO2intensity | [2，3] 单位：kg | 数值型变量 利用每 kg 原油产生 CO2 | 由于各国的科学发展计数不同所以利用每 kg 的原油产生的 CO2 量不尽相同。 |
| | CO2 排放量 CO2emission | [1e2,4e6] 单位：kt | 数值型变量 每一年排放 CO2 总量 | 在 CO2_emission 表中统计了 1980-2014 年间每一年各国的 CO2 排量。 |
| | NO 排放量 NOemission | [1e2,3.5e5] 单位：kt | 数值型变量 每年排放 NO 总量 | 在 NO_emission 表中统计了 1980-2014 年间每一年各国的 NO 排量。 |
| | 甲烷排放量 Methemission | [5e1，8e5] 单位：kt | 数值型变量 各国每年排放甲烷总量 | 在 Meth_emission 表中统计了 1980-2014 年间每一年各国的甲烷排量。 |
| | 温室气体排放量 greenhouse | [2e2,12e6] 单位：kt | 数值型变量 每年排放温室气体总量 | gas_emission 表中统计了 1980-2014 年间各国的温室气体排放总量。 |
| 森林破坏 | 森林覆盖面积 Forest area | [0,8500] 单位：平方千米 | 数值型变量 | 可进一步转化为违约时长，用于判断违约程度 |
| | 濒危物种 Threatened species | [0，40] 单位：种 | 数值型变量 共有三个违约分类 | 代表着顾客的单词的违约程度 |
| 地理信息 | 纬度 Lat | N90—S90 | 度量变量 | 定位一个国家在地球位置信息 |
| | 经度 Long | W180—E180 | 度量变量 | 定位一个国家在地球位置信息 |
| 国家信息 | 国家代码 Country Code | 240 国家英文简称 | 文本信息 | 作为数据表的键值，用来连接数据表以及索引表 |
| | 收入分类 IncomeGroup | Upper middle income High income | 分类变量 | 不同收入水平的国家对能源的利用效率、 |

| | | | | |
|---|---|---|---|---|
| | | Lower middle income<br>Low income | | 资源量有所不同因此排放的污染气体不同 |
| | 所属大洲<br>Region | East Asia & Pacific<br>Europe & Central Asia<br>Latin America & Caribbean<br>Middle East & North Africa<br>North America<br>South Asia<br>Sub-Saharan Africa | 分类变量 | 所处的大洲地理位置不同，由于地理环境的不同排放的污染也就不同。 |

## 2.2 数据预处理

### 2.2.1 污染气体时间序列数据表

CO2_emission.csv 数据表，CO2_intensity.csv 数据表，NO_emission.csv 数据表，Methane_emission.csv 数据表皆为各国的排放的污染源的时间序列数据，在 1960-1969 年间缺失的数据量较大，又因为数据距离现在较为遥远，具有的代表性较差，所以决定删除 1960-1969 年间得所有数据。只研究 1970-2012 年的数据。对仍有缺失的数据进行记录的删除。

CO2_emission_index.csv 维度表，CO2_intensity_index.csv 维度表，NO_intensity_index.csv 维度表以及 Methane_index.csv 维度表中储存各国的信息但最后的两个描述列与本次的分析无关，故删除。

利用 country code 键值将数据表和对应的维度表相连接，得到的数据表并不是 R 软件能够处理的 tidy data 数据，所以利用 melt 函数将获得的数据框进行 tidy 数据处理，获得最终的数据分析数据框[3]，在绘制地图时此数据框还需要与 world 地理信息数据框进行关于 country 键的连接，并进行 country 不同键值的补全，最终得到相应的完整的地图处理数据框。

### 2.2.2 污染气体截面数据表

greenhouse gas emissions.csv 数据表和 forestdata.csv 数据表为面板数据，记录的是 2015 年各国的温室气体数据和森林覆盖状况。进行处理的方式较为简单，删除有缺失数据的记录以及无用的变量，最后和信息维度利用 country 键值进行相连，得到分析所用数据框。

[3] 具体处理过程见附录代码

## 2.3 地球污染描述性统计分析

### 2.3.1 CO2 排放强度时序折线图

CO2 排放强度值得是每获得 1kg 原油的能量所产生的 CO2 的千克数。四种不同收入类型国家，平均每消耗 1kg 的原油产生的 CO2 气体的年度变化状况可以从图 2-1 中看出。

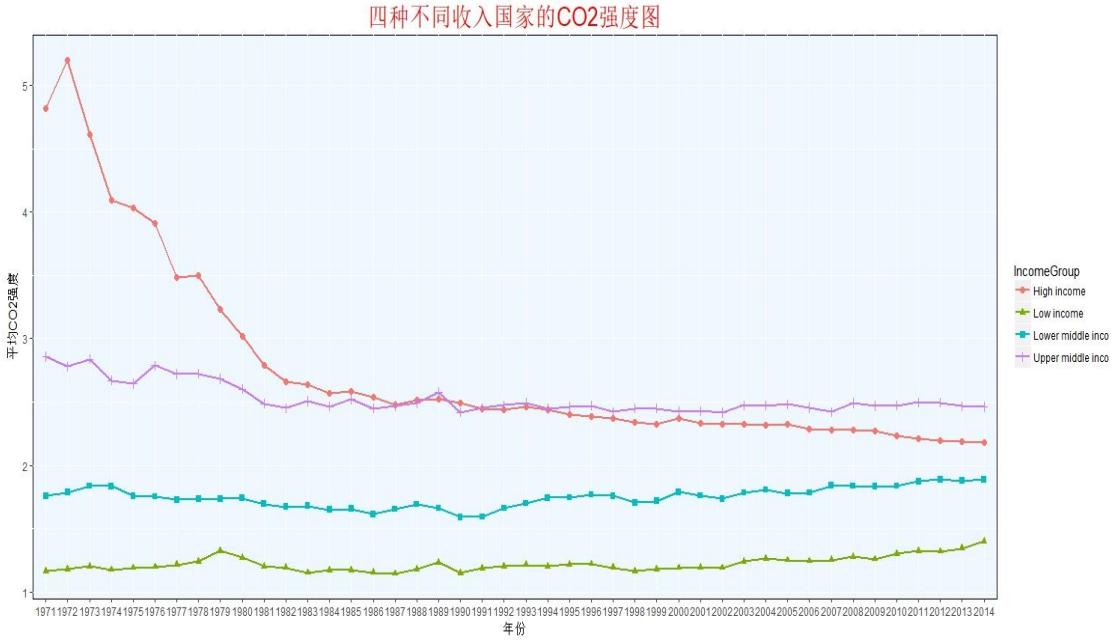


图 2-1：不同收入类型国家 CO2 排放强度变化

从 CO2 排放强度变化图中可以看出，从 1971 年到 2014 年间不同收入类型的国家中 CO2 排放强度基本稳定越高收入的国家 CO2 排放强度就越大。从变化趋势上来看低收入，中低收入和较高收入国家在这 40 年间得排放强度变化趋势不明显，而高收入国家的排放强度在 70 年代有了急速的上升并下降的过程，可能是由于在 70 年代中科学技术在发达国家中的急剧变化首先使得生产过程中的生产变大，但也有随之而来的污染后又经技术的革新使得污染气体的排放水平降低。进一步还可以看出，随着时间的不断推移，四种收入类型的国家的 CO2 排放强度在趋于相同水平，程度越来越接近。

### 2.3.2 CO2 排放量时序面积堆积图

CO2 排放量指的是不同类型收入国家的每一年 CO2 排放总量（单位：kt）。四种不同收入类型国家，1970 年—2014 年 CO2 排放总量状况可以有下图显示出：

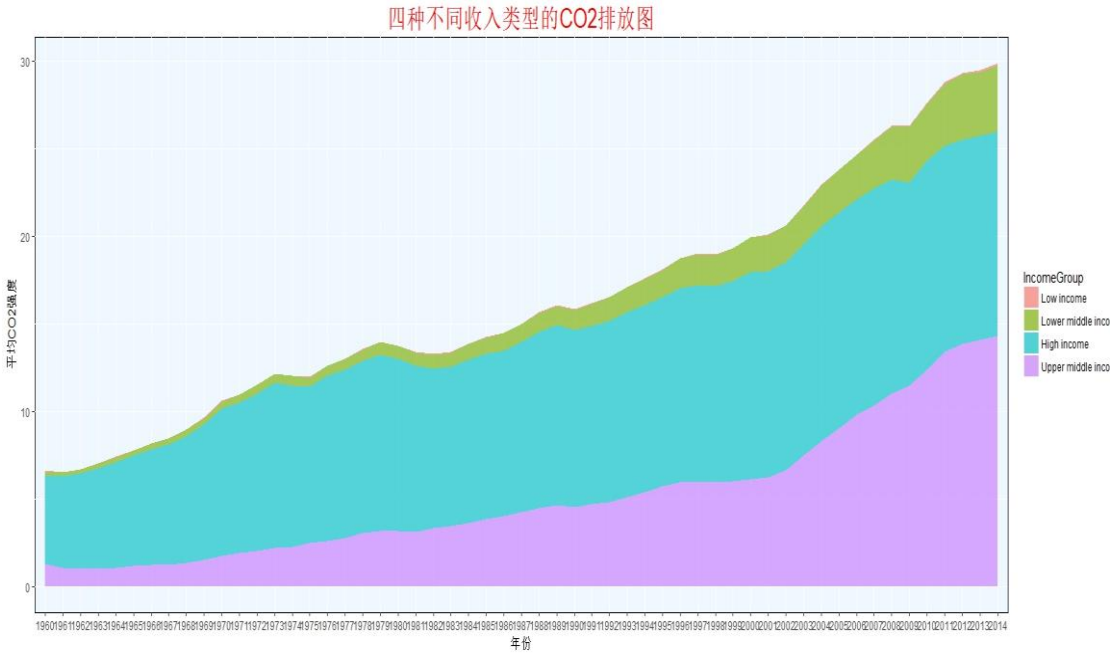


图 2-2：四种不同收入类型国家 CO2 排放总量

从图 2-2 中可以看出温室气体二氧化碳浓度达到近 50 年来的最高水平。而近三年来，全球 CO2 排放量已经趋于平稳，但近期又开始回升。在这 50 年来低收入国家和中低收入国家的碳排放量变化并不明显，在 CO2 全球排放总量上所占据的比例也并不大。相比来说，高收入国家的 CO2 排放总量增幅较大，而对于较高收入国家来说 CO2 增幅比例就非常大，增幅大于 1000%。并且高收入国家与中高收入国家在全球 CO2 排放总量中占比很大，对全球气候变暖负有最大责任。

### 2.3.3 2014 年各国 CO2 排放总量分布图

通过 R 的 ggplot 包以及 map 包，绘制关于 2014 年度世界各国 CO2 排放的总量图，直观的反映各国的 CO2 排放状况，通过 R 软件绘制地图结果如下图 2-3 所示：

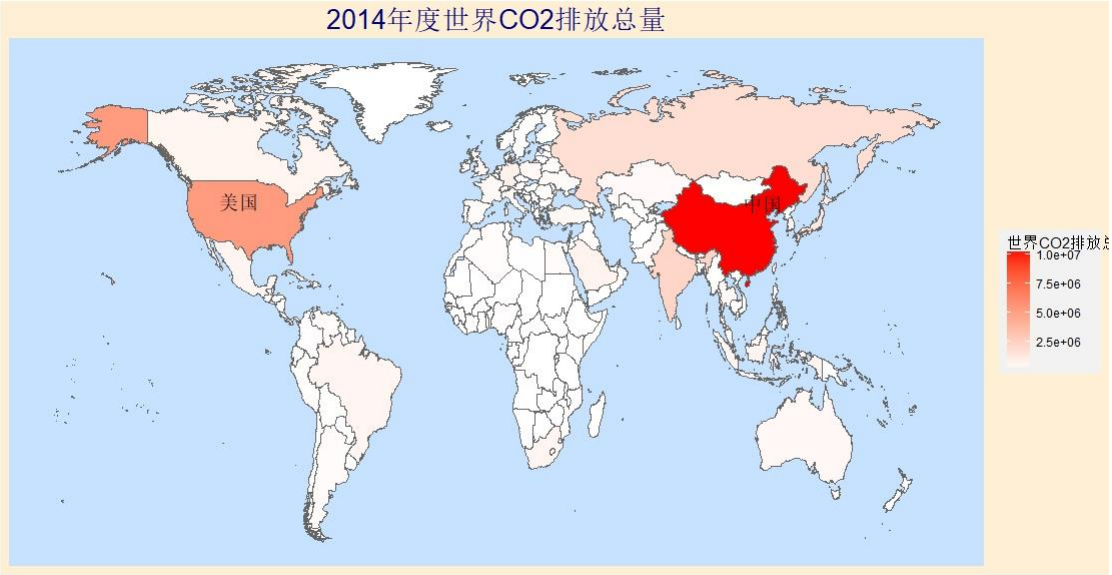


图 2-3 2014 年各国 CO2 排放总量

从地图中可以直观看出温室气体 CO2 的排放状况，在众多国家中，由于中国正处于经济发展的重要阶段，生产发展速度近年来达到了前所未有的程度，而由此带来的 CO2 气体排放状况也最为严重，环境状况不容乐观。和中国有着相类似的情况的也有金砖国家俄罗斯，巴西和印度。这些国家都处于能源消耗性经济发展时期，所以经济发展的同时带来的环境破坏也是比较严重的。相比之下还有美国，英国，德国，日本等传统的经济强国的 CO2 排放量也在世界排放总量了上占有了很大比例。

### 2.3.4 各大洲甲烷排放总量 gif 图[4]

甲烷也是造成全球气候变暖的重要温室气体之一，通过搜集 world bank data 的甲烷气体 1970 年—2012 年的变化情况来具体反映各大洲的甲烷排量变化。为了突出变化情况，利用 R 软件动画生成包 animation 并结合绘图软件 ImageMagick 来绘制各大洲甲烷排量年份 gif 动图[5]。

[4] 由于 Rmarkdown 无法进行 gif 动图展示，所以该本分代码被展示在附录中。
[5] 在此只展示截图，动图请见附件。

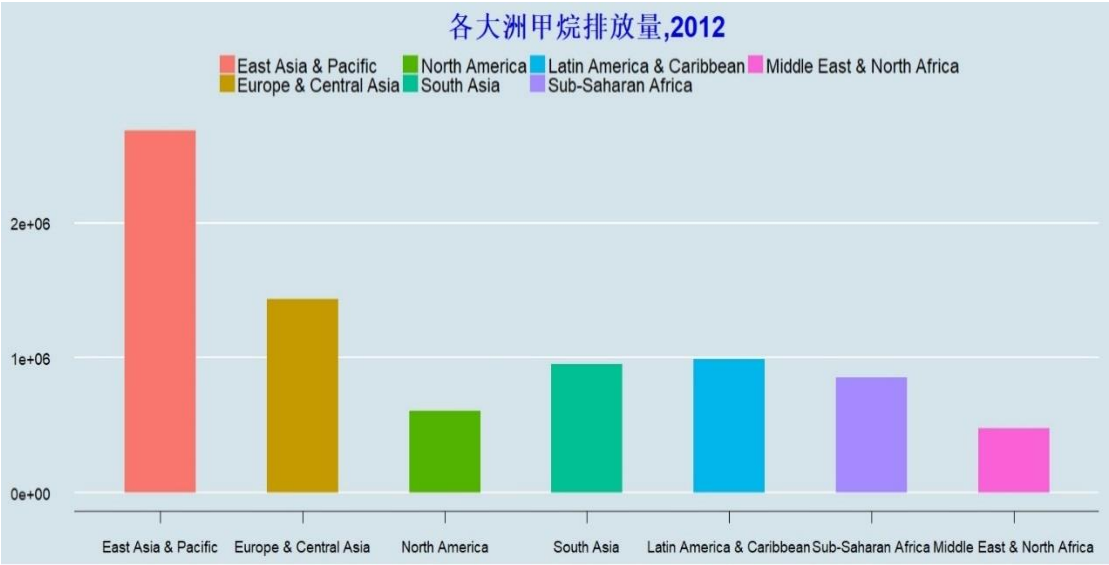


图 2-4 各大洲甲烷排放量 gif 动图

从 gif 动图中可以清楚的看到各大洲的温室气体排放量在不断的上升，其中东亚即太平洋地区的变化幅度最大，且一直是甲烷排放量最多的地区，而北美洲的变化较小，甚至在年份的变化过程中从 1970 年的甲烷排放量位于第三到 2012 年的排放量降到了仅次于东北非的低甲烷排放大洲。

## 2.3.5 各大洲 2012 年甲烷排放小提琴图

从甲烷排放时间序列动图中我们可已看出各大洲的温室气体排放的时间变化状况，但我们同样可以通过截面数据分析距现在更近的 2012 年的各大洲甲烷排放状况，如下图各大洲甲烷排放量小提琴图：

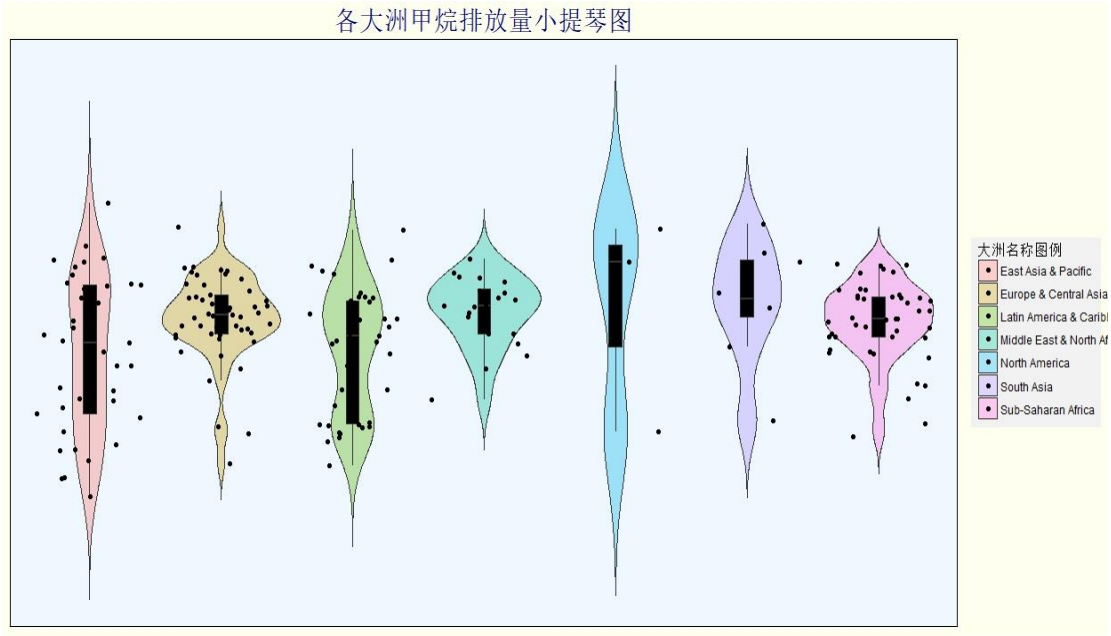


图 2-5 2012 年甲烷排放小提琴图

在小提琴图中我们可以清楚地观察到每个大洲国家甲烷排放状况的分布情况，在每个小提起图中还绘制了箱线图和散点分布图，更好的比较分布状况。其中北美洲和东亚的小提琴图和箱线图较为细长，说明这两个大洲的国家甲烷排放状况较为分散，而欧洲、北非和拉丁美洲分布较为集中，在这三个大洲的甲烷排放状况差距不大。

## 2.3.6 NO 排放交互式密度分布图

探究拥有数据据今最近的 2014 年 NO 排放量交互式分布图，由于无法将交互式作图放到 word 中，所以只显示截图，交互式 html 在附件中显示。

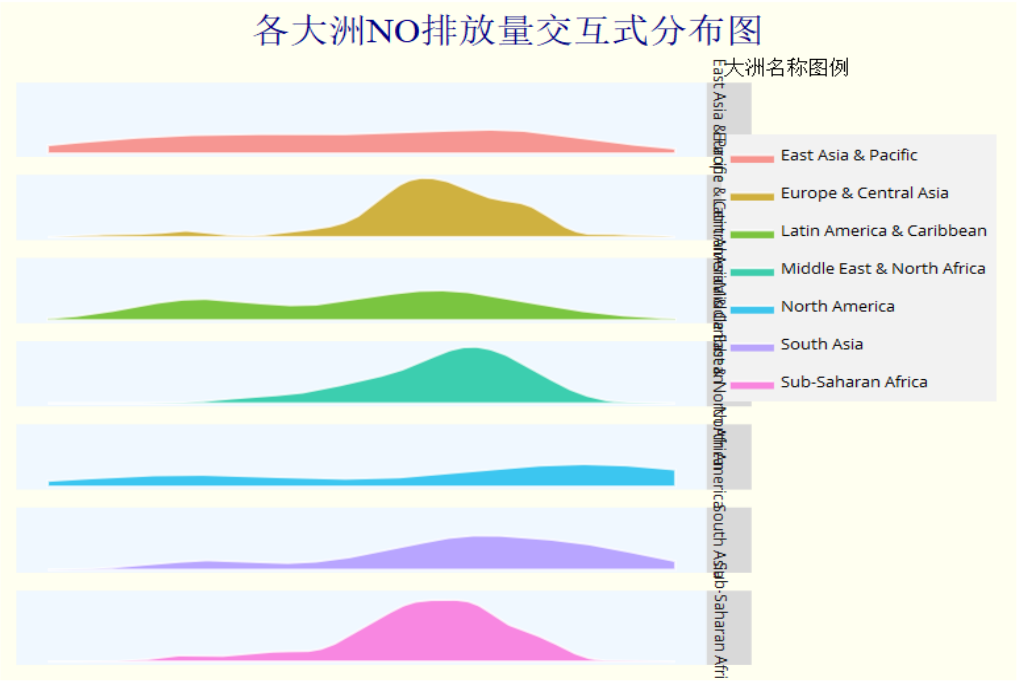


图 2-6 各大洲 NO 排放量交互式密度分布图

由于 NO 的排放量在各大洲国家内的差距比较大，所以利用 log 取对数使作图更加清晰，另外利用交互式作图也可以帮助阅读者更好的理解图中内容，了解图中信息。在图中可以清楚地看到东亚，南亚和拉丁美洲的 NO 气体的分布较为分散，国家之间的差距较大，而非洲和欧洲国家在 NO 排放上的差距就不怎么明显，交互图请见附件。

## 2.3.7 国家间污染状况聚类图

利用 2014 年的温室气体排放总表（greenhouse gas emissions.csv 数据表）中 CO2，NO，甲烷以及其他温室气体的排放量几个指标来进行污染气体国家间的聚类，如下图 2-7 所示：

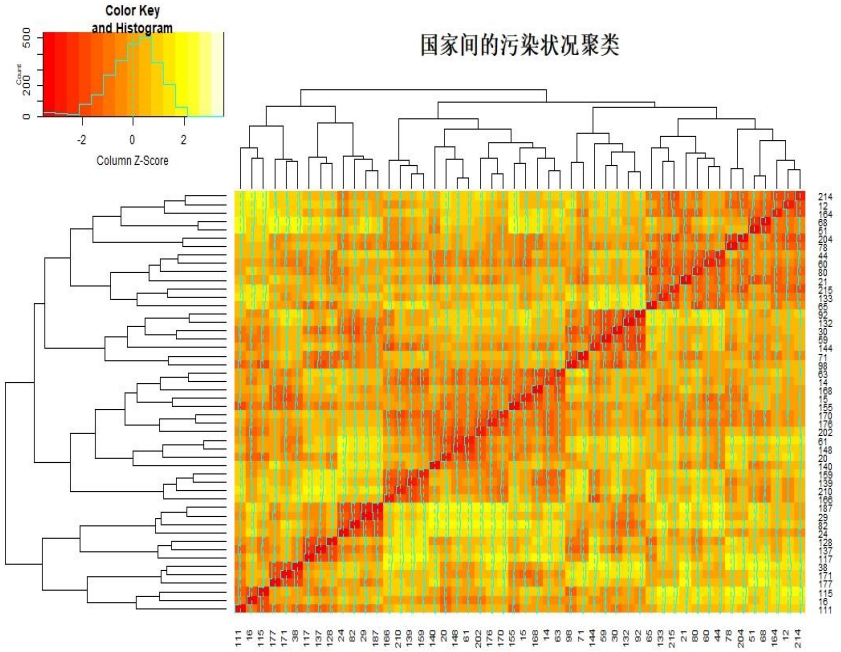


图 2-7 国家间污染状况聚类图

如图 2-7 所示根据不同种类的温室气体的排放状况先进行标准化再将世界各个国家进行聚类，得到的聚类结果可以在热图的边缘线上显示，在左上角相关关系分布直方图中还可以观察到国家间的距离分布直方图，可以看到各个国家间的距离分布状况是服从正态分布的，因此聚类结果比较可信。

## 2.3.8 森林覆盖面积交互式地图

一个国家的环境也表现在森林覆盖面积上，森林的覆盖面积越大对空气污染亦或是水体污染都有净化功能，在图 2-8 中显示了 2015 年的各个国家间得森林覆盖状况，以此来反映各个国家对环境污染的抵抗能力。且具有交互功能，在 word 中只放置了截图，交互图请见附件：


图 2-8 国家间森林覆盖状况

如图 2-8 所示，可以清楚地发现在国土面积（单位：平方千米）较大的拥有的更多的森林覆盖面积，像美国，中国，俄罗斯等但森林覆盖率并不是很高，相比来说，巴西和印度以及墨西哥的森林覆盖程度更加大。尤其是位于赤道附近的国家的森林覆盖面积大且这些国家的污染气体排放少，对环境的修复能力较强，环境状况比较乐观。高收入国家由于发展中产生更多的温室气体所以应该更加注意森林的保护和种植。