现代统计图形（删减章节）

赵鹏　　谢益辉　　黄湘云

2021-06-21

目录

表格

插图

附录 A　删减章节

A.1　中美对比

案情回放

2011 年 1 月 19 日，英国《卫报》发表了一篇关于中美国力对比的文章[1]。此时正是中国国家主席胡锦涛访美期间，《卫报》称此次访问是“硬实力与软实力的会晤”。文中分析了中美一些经济社会指标的对比，配图 A.1。

总体来说，这幅图形比较有冲击力，能突出两国各自的优势劣势，尤其值得称道的是作者将各个指标进行了排序，这样便于读者阅读。然而，如果我们再细看一下，就能发现一个问题：由于指标数值都是用扇环表示，因此容易造成一个误会，即数值大小究竟体现在扇环的半径上还是面积上？不同读者的视角自有不同，但这幅图如果按照面积大小去解读数据就错了。该文提供了部分原始数据[3]，见表 A.1，从中可以看到，中国人口是美国人口的 4.3 倍，但图中中国扇环（红色）的面积显然不止是美国的 4.3 倍，这样可能给读者造成一个印象，似乎中国人口是美国的七八倍。

那么，中美国力的实际对比到底是怎样的？我们应该如何用图形恰当地表达其差异？

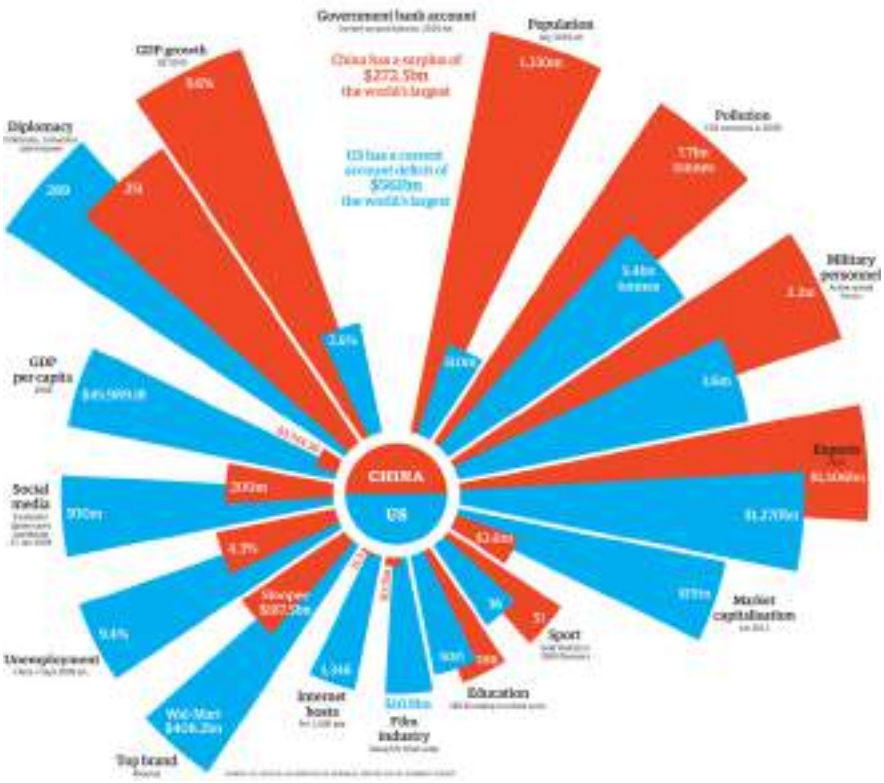
探案过程

我们利用这些数据重画了《卫报》的图形，如图 A.2 所示。

由于我们的主要目的是比较两国的指标，因此简单的条形图足以完成这个任务。我们大致上按中国指标与美国指标的比率大小排列这些条

[1] https://www.theguardian.com/news/datablog/2011/jan/19/china-social-media

[3] https://docs.google.com/spreadsheets/d/1Fs0B5uCUwt6p7CsjI5CwiZViHspS83pCJ6i-l3OiL7I



2

图 *A.1: 2010* 年中美国力对比：中国人口多、污染重、*GDP* 增长快；美国人均 *GDP* 高、市值高、失业率高

形图。虽然这幅图看上去比《卫报》的图形逊色，但不会造成任何误解，也能更客观反映两国差异。当然，更好的解决方案可能是保持原有的“放射性”设计，但把扇形替换为矩形条。

我们在检查《卫报》提供的数据表时，发现表中“市值”（Market Capitalisation）一栏中美数据颠倒了，对比原图和我们重画的图让我们很快注意到了这个差异。数据的录入错误无处不在，而“看数字不如看图形更直观”这一点也让我们相信图形在检查错误方面有其特殊价值。数据表中还有两处错误，你能通过对比两图找到它们吗?

表 A.1: 《卫报》提供的 2010 年中美国力数据。未经订正

country	metric	value	units
China	Current account balance	272.5	$ billion
America	Current account balance	-561.0	$ billion
China	Population	1330.0	millions of people
America	Population	310.0	millions of people
China	Pollution	7.7	billion tonnes
America	Pollution	5.4	billion tonnes
China	Active armed forces	2.2	million military personnel
America	Active armed forces	1.6	million military personnel
China	Exports	1506.0	$ billion
America	Exports	1270.0	$ billion
China	Market capitalisation	15.0	$ trillion
America	Market capitalisation	3.5	$ trillion
China	2008 Olympics	51.0	Gold medals in 2008 Olympics
America	2008 Olympics	36.0	Gold medals in 2008 Olympics
China	Reading ability	556.0	OECD reading score in schools
America	Reading ability	500.0	OECD reading score in schools
China	Diplomacy	251.0	number of embasies, consulates and missions
America	Diplomacy	289.0	number of embasies, consulates and missions
China	Unemployment	4.3	% of working population
America	Unemployment	9.4	% of working population
China	GDP growth	9.6	GDP growth for % on previous year
America	GDP growth	2.6	GDP growth for % on previous year
China	$ GDP per capita	374436.0	$ GDP per capita
America	$ GDP per capita	459891.2	$ GDP per capita

探案工具

本案例使用了条形图（详见 4.1 节）和玫瑰图（详见 1.3 和 5.5 节），所使用的数据收集在 **MSG** 包中，前文所述的《卫报》数据表中的错误经过了订正。加载数据和绘制图形的代码如下：

```
# ggplot2 绘制 2010 年中美国力对比图
library(MSG)
library(ggplot2)
data(cn_vs_us, package = "MSG")
## 恢复使用卫报的错误数据
```

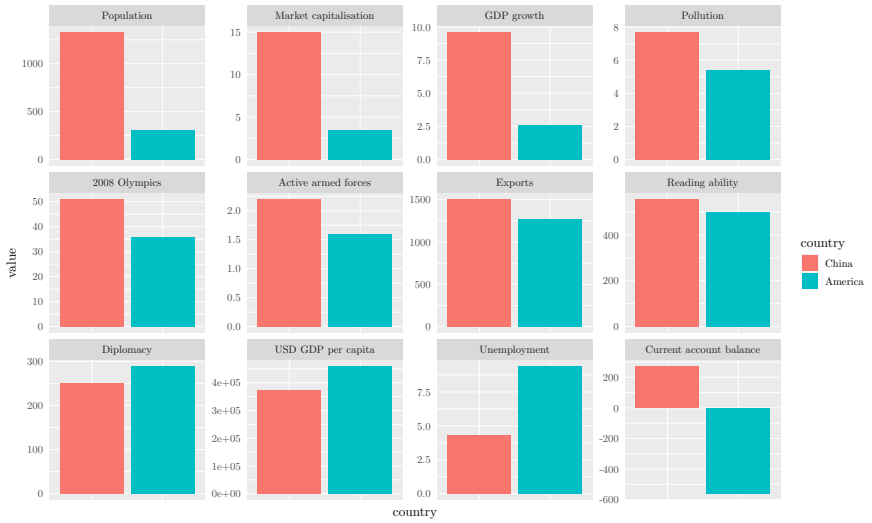


图 *A.2:* 重新调整后的 *2010* 年中美国力对比图：条形图的高度表达数值大小更清楚（图中红色代表中国）。各子图显示的指标详见表 *A.1*。注意：本图的“市值”（*Market Capitalisation*）数据与图 *A.1* 所示的数据中美两国颠倒了，这源于《卫报》给出的数据表中的错误。数据表中还有两处错误，你能找到吗？

```
cn_vs_us$value[11] = 15
cn_vs_us$value[12] = 3.5
cn_vs_us$value[23] = cn_vs_us$value[23] * 100
cn_vs_us$value[24] = cn_vs_us$value[24] * 10
## 按两国比例重新排列顺序
ratio = cn_vs_us$value[seq(1, 24, 2)] / cn_vs_us$value[seq(2, 24, 2)]
cn_vs_us$ratio = rep(ratio, each = 2)
cn_vs_us$metric = gsub("$", "USD", cn_vs_us$metric, fixed = TRUE)
cn_vs_us$metric =
  factor(cn_vs_us$metric,
         unique(cn_vs_us$metric)[order(unique(cn_vs_us$ratio),
                                       decreasing = TRUE)])

print(
  ggplot(cn_vs_us, aes(x = country, y = value, fill = country)) +
    geom_col() +
    theme(axis.text.x = element_blank(),
          axis.ticks = element_blank()) +
```

```
    facet_wrap(~metric, scales = "free_y", ncol = 4)
)
```

A.2 新冠疫情

A.2.1 案情回放

2020 年春节前后，新冠疫情爆发，亿万人的生活和工作都或多或少受到疫情的影响。笔者在关注疫情统计数据的过程中，遇到了一些有意思的问题。这里试举两例。

第一个例子是，笔者看到朋友分享的一幅疫情地图（如图 A.3），显示了 1 月 22 日全国各省确诊和疑似病例，只见沿海和中部地区一片火红（红色表示新冠肺炎确诊病例），唯有江苏省是白色。图上搭配文字："在吗，救我"，在地图上格外抢眼。笔者即将赴江苏就职，对此格外关注的同时，不免产生疑问：江苏和周边省份差别真的有这么大吗？为什么？

图 *A.3: 新冠疫情下的江苏：2020 年 1 月 22 日，除了江苏省，东部和中部地区一片火红*

第二个例子是，有一段时间，一种酷似鹦鹉螺的玫瑰图非常流行，如图 A.4 所示，显示了 2020 年 3 月 27 日除我国以外的世界各国新冠疫情。图中用不同颜色、相同圆心角的扇形来代表确诊 707 例以上的不同国家，各国新冠肺炎的确诊和死亡人数以数字标注。显然，图形经过了精心设计，令人过目不忘。然而，仔细观察却不免心生困惑：与 A.1 小

节所举的例子类似，如果以半径代表数值，那么单从图形看，会给人一种印象，即美国确诊数是德国的 2 倍，而德国大约是英国的 2 倍；但一看数字，德国却大约是英国的 4 倍。那么，确诊数量究竟体现在扇形的哪个属性呢？从图上如何看出各国的实际差异呢？

探案过程

对于第一个关于疫情地图的问题，笔者获取了当天（1 月 22 日）以及前一天（1 月 21 日）的疫情数据。这是为了避免数据统计时间的不同而出现与图 A.4 在统计数字上的差异。在这些数据的基础上，笔者绘制了这两天的疫情地图，如图 A.5 所示。

在这两张地图上，我们同样是用红色来表示疫情的地理分布；不同的是，我们用红色的深浅来区分确诊病例的多少。同时，由于各省确诊病例数量极差较大，跨三个数量级，我们使用了对数坐标，以便显示低值之间的差异。

从图中可以看出，江苏跟周边省份差别并不是很大：1 月 21 日，除了浙江 5 例确诊（浅红色）外，江苏和安徽无确诊，山东只有 1 例，在地图上都是白色或接近白色；而 1 月 22 日，浙江共 10 例，而江苏、安徽、山东各增加 1 例，均是极浅的浅红色。

其实，我们可以看到，在图 A.3 中，很多省份其实只有 1 例确诊，却涂上了跟几百例确诊的省份同样的颜色，夸大了江苏和其它省份的差别。当然，笔者作图的角度是比较各省确诊病例的数量级；如果是为了区分“零确诊”与“有确诊”，那就另当别论了。在这个案例里，颜色的选择和图例的重要性可见一斑。

为了直观地看到疫情在地理上的变化，笔者将每天各省的确诊病例做成一张地图，然后连起来做成一幅动画，可以在线观看[4]。

对于第二个关于玫瑰图的问题，玫瑰图的本质是条形图，只不过是把条形图的笛卡尔坐标转换成了极坐标：用极坐标的角度代表笛卡尔坐标的 x，而半径代表 y。从图 A.4 来看，各国确诊病例对应的肯定不是扇形半径，因为论确诊数量的话，排在第一位的美国是排在末位的菲律宾的一百多倍，而对应的扇形半径凭肉眼判断，肯定相差不到 100 倍。有没有可能是圆心并不代表 0 病例，而是某个其它数值呢？也不是，因为如果那样的话，两者半径相差的倍数只会更多。

会不会是用扇形面积表示确诊数量呢？也不是，因为美国确诊数量

[4] https://ncov2020.org/animation/map_animation.gif

图 *A.4: 2020* 年 *3* 月 *27* 日世界各国新冠疫情的玫瑰图：图中用不同颜色的扇形来代表不同国家

表 A.2: 全国省级行政区新冠肺炎确诊人数（未发现确诊病例的地区并未列出）

名称	01-21	01-22
湖北	270	444
广东	17	26
北京	10	14
浙江	5	10
上海	6	9
重庆	5	6
河南	1	5
四川	1	5
天津	2	4
湖南	1	4
海南	0	4
辽宁	0	2
江西	0	2
山东	1	2
广西	0	2
河北	0	1
山西	0	1
江苏	0	1
安徽	0	1
福建	0	1
贵州	0	1
云南	1	1
宁夏	0	1
台湾	1	1
澳门	0	1

是德国的大约两倍，而其扇形面积显然超过这个比例。此外，选用扇形来对比两者大小的时候，要么是令圆心角相同而比较半径的不同（如 1.3 节的极坐标面积图和 5.5 节的风玫瑰图，此时面积与半径的平方成正比），要么是令半径相同而比较圆心角的不同（如 4.9 节的饼图，此时面积与圆心角成正比）。由于人眼对扇形的角度和半径比较敏感，对面积不敏感，一般不会在半径不同的情况下用扇形面积来区分大小。

那么，有没有可能是将确诊数量取了对数之后作图呢？让我们来看

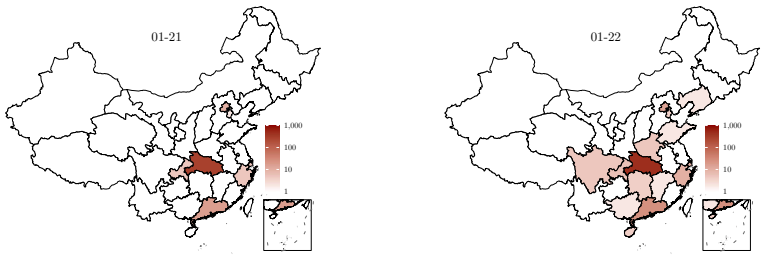


图 A.5: 新冠疫情中国地图。图中用红色的深浅来表示确诊病例的多少。由于各省确诊病例数量极差较大，数量跨三个数量级，使用对数坐标来显示低值之间的差异

一下：以美国、德国和英国为例，美国确诊病例是德国的大约 2 倍，而德国确诊病例是英国的大约 4 倍。假定是以扇形半径来表示确诊病例的对数，那么德英两国的扇形半径差应该是美德两国扇形半径差的 2 倍，但图 A.4 显然不是这样，甚至相反：德英两国的扇形半径差大约只有美德两国的一半。

为了一探究竟，笔者获取了当天的全球疫情数据，自行绘制了一张类似的玫瑰图，如图 A.6 的上图所示，以扇形的半径来表示确诊病例的对数。图中的统计数字略有出入，可能是数据更新的当天时刻有不同，这不影响我们的对比。半透明的环状带标示了病例为 10、100、1000 和 10000 的位置。

从图中可以看出，玫瑰图用对数来展示确诊病例之后，相邻两个扇形的半径差别就没有那么大了。虽然视觉冲击力小了很多，然而我们可以清晰地从图中看出哪些国家的确诊病例是属于同一个数量级（即处于同一环状带），以及任意两国的病例数对比关系。例如韩国在最外围的半透明带边缘，而俄罗斯位于自外向内第二个半透明带的边缘，即使不看数字，也能知道两者是 10 倍的关系；类似的，美国跟韩国也是大约 10 倍的关系。而图 A.4 的扇形半径跟确诊病例数有何种换算关系，以及中心的白色半透明圆环表示的是什么意义，我们就不得而知了。

作为对比，笔者将同一组数据做了一张点图，见图 A.6 下图。与极坐标相比，在笛卡尔坐标系中，用不均匀的网格来表示对数坐标就会更容易理解。同时，读者可以很直观地从图中获取病例数的大概数值，以便做进一步的比较，这显然比玫瑰图更有意义。例如，前文所述，在对数坐标下，德英两国确诊病例之差应该是美德两国之差的 2 倍，在点状

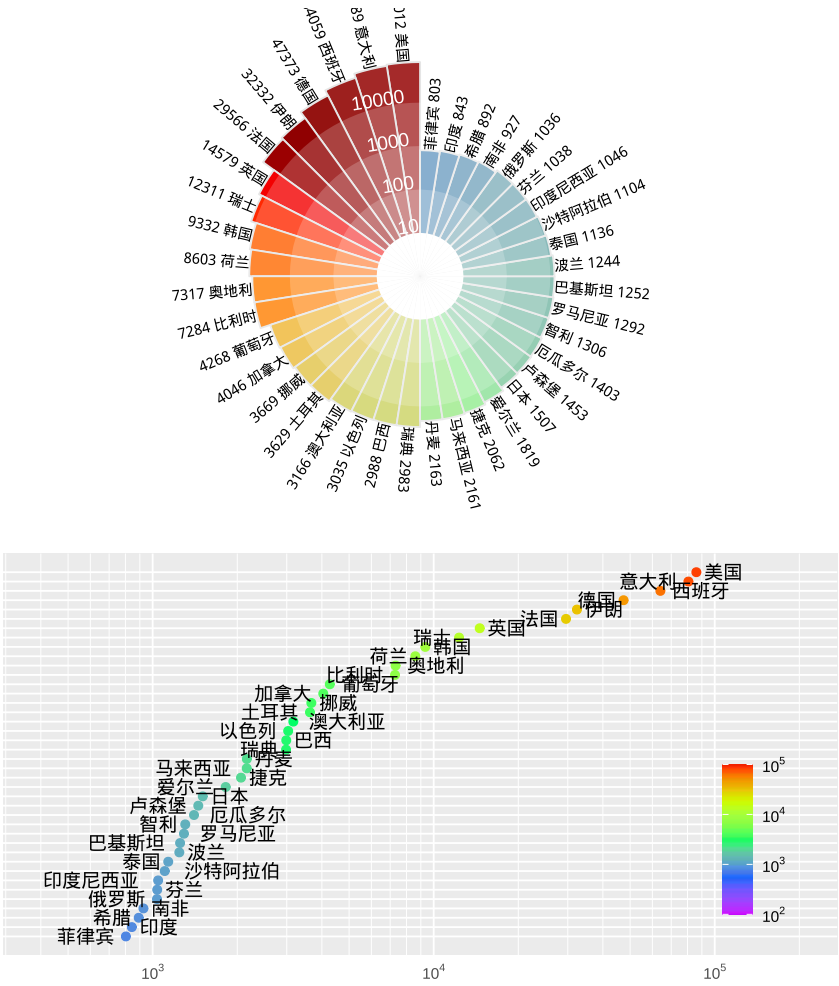


图 *A.6:* 对数坐标表示的世界各国新冠疫情的玫瑰图（上）和点图（下）：上图用不同颜色的扇形来代表不同国家

图上一目了然。注意：这张点图里，点的颜色其实是多余的元素；因为点的 x 坐标已经表达了其确诊病例数的数值。笔者在这里仍然标识了颜色，仅仅是为了视觉上的美观而已。

关于笔者参与的有关新冠疫情的进一步研究工作，参见 Zou et al. (2020)。

探案工具

本案例使用了玫瑰图（详见 1.3 和 5.5 节）、点图（详见 4.2 节）和地图（详见 A.3 小节）。用于绘制图 A.5 的 R 语言代码为：

```
# 绘制中国新冠疫情地图
library(ncovr)
library(dplyr)
library(scales)
covidchina = readRDS(
  system.file("extdata", "covidchina.rds", package = "MSG"))
for (choose_date in c("01-21", "01-22")){
  print(plot_ggmap(covidchina,
                   col_name = paste0("confirmed_", choose_date),
                   show_capitials = FALSE,
                   province_language = NA,
                   add_title = choose_date))
}
```

这里用到了笔者开发的 R 附加包 **ncovr** (Zhao and Cao, 2020) 中的 `plot_ggmap()` 函数。这个函数将中国地图的数据进行了封装，包含了九段线区域（即中国地图右下角的矩形区域）。**ncovr** 包发布在 GitHub 上，安装之前需要先安装 **remotes** 包，再使用其中的 `install_github()` 函数进行远程安装（或者使用 `MSG::msg("0")` 来一次性安装复现本书图形的所有附加包）：

```
install.packages("remotes")
remotes::install_github("pzhaonet/ncovr")
```

用于绘制图 A.6 的代码比较长，除了准备数据的代码外，大部分指令都是用作装饰和细微调整的。例如绘制其中的玫瑰图的核心代码其实只有以下四行：

```
# 数据映射：
ggplot(covid, aes(country, cum_confirm, fill=cum_confirm)) +
# 条形图：
  geom_col(width=1, color='grey90') +
# 对数坐标：
  scale_y_log10() +
# 极坐标：
  coord_polar()
```

其它代码用来添加参考线和文字、修改主题和配色方案、隐藏图例等。完整代码如下：

```
# 绘制世界各国疫情形势图
library(ggplot2)
library(patchwork)
library(scales)
covid = readRDS(
  system.file("extdata", "covidcountries.rds", package = "MSG"))
n_countries = nrow(covid)
covid = transform(
  covid, hjust = 1,label = paste(cum_confirm, country),
  angle = 1: n_countries * 360/n_countries - 90 - 180/n_countries)
second_half = (n_countries %/% 2):n_countries
covid$angle[second_half] = covid$angle[second_half] + 180
covid$hjust[second_half] = 0
covid$label[second_half] =
  paste(covid$country, covid$cum_confirm)[second_half]

p_polar =
  ggplot(covid, aes(country, cum_confirm, fill=cum_confirm)) +
  geom_col(width=1, color='grey90') +
  geom_col(aes(y=I(10000)), width=1, fill='white', alpha = .2) +
  geom_col(aes(y=I(1000)), width=1, fill='white', alpha = .2) +
  geom_col(aes(y=I(100)), width=1, fill='white', alpha = .2) +
  geom_col(aes(y=I(10)), width=1, fill = "white") +
  scale_y_log10() +
  scale_fill_gradientn(
    colors=c("steelblue", "lightgreen", "orange",
             "red","darkred", "brown"),
    trans="log") +
  geom_text(aes(label=label, y = cum_confirm,
                angle=angle, hjust = hjust), vjust= 0.5, size = 3)  +
  annotate(
    "text", x = 39, y = c(10, 100, 1000, 10000)*1.5, color = "white",
    label=c(10, 100, 1000, 10000), angle = 360 / 40) +
  theme_void() +
  theme(legend.position="none") +
  coord_polar()
```

```
p_point =
  ggplot(covid, aes(country, cum_confirm)) +
  geom_point(aes(color=cum_confirm), size = 2) +
  scale_color_gradientn(
    colours = rev(rainbow(5)),trans="log",
    limits = 10^c(2, 5),
    breaks = 10^(2:5),
    labels = trans_format("log10", math_format(10^.x)),
    minor_breaks = as.vector(sapply(2:10, function(x) x * 10^(2:6))))+
  geom_text(aes(label=country),hjust=rep(c(-0.2,1.2), 20), vjust=0.5)+
  scale_y_log10(
    breaks = 10^(2:5),
    limits = c(400, 200000),
    labels = trans_format("log10", math_format(10^.x)),
    minor_breaks = as.vector(sapply(2:10, function(x) x * 10^(2:6))))+
  scale_x_discrete(expand = c(0.05, 0.05)) +
  labs(x = NULL, y = NULL) +
  coord_flip() +
  theme(legend.title = element_blank(), legend.position = c(0.9,0.3),
        axis.ticks = element_blank(),axis.text.y = element_blank(),
        legend.background = element_blank())
print(p_polar / p_point + plot_layout(heights = c(4,3)))
```

A.3 地图

概述

毫无疑问，地图是展示地理信息数据时最直观的工具，尤其是当地图和统计量结合时，其功效则会进一步加强。在本书的第 1 章中曾经提到过 John Snow 的地图，注意图中不仅标示出了霍乱发生的地点，每个地点的死亡人数也用点的数目标示了出来。历史上还有不少类似的使用地图的例子，而在今天，地理信息系统（GIS）已经成为研究空间和地理数据的热门工具，地图的应用也是屡见不鲜。

示例

表 A.3 给出了 2005 年世界各国家地区的农业进出口竞争力指标数据 (Xie, 2007)。其中，我们将竞争力指标简单定义为（出口 − 进口）/

表 A.3: 2005 年世界各国农业进出口竞争力（部分。原数据为 97 行 2 列）

Index	Country
-0.7570701	Albania
-0.9667213	Algeria
0.8778870	Argentina
0.5264728	Australia
0.0062974	Austria

（出口 + 进口）。我们将这组数据在图 A.7 上标示出来。从图中可以看出，阿根廷、巴西等南美国家的农业进出口竞争力较强，而利比亚、阿尔及利亚等北非国家的竞争力较弱。

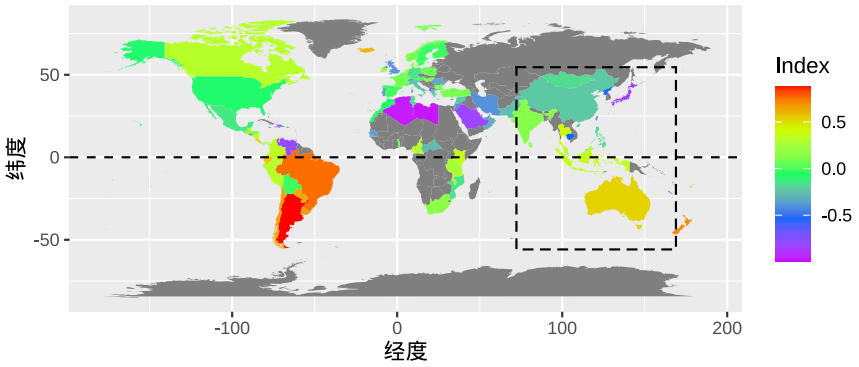


图 A.7: 2005 年各国农业进出口竞争力地图：农业出口强国在南美，弱国在北非

绘制方法

地图的本质是多边形（9.2 节），而多边形的边界则由地理经纬度数据确定。R 中绘制地图的传统附加包是 **maps** (Becker et al., 2018)，核

心的函数为 map()，它的用法如下：

```
map(database = "world", regions = ".", exact = FALSE, boundary = TRUE,
    interior = TRUE, projection = "", parameters = NULL,
    orientation = NULL, fill = FALSE, col = 1, plot = TRUE,
    add = FALSE, namesonly = FALSE, xlim = NULL, ylim = NULL,
    wrap = FALSE, resolution = if (plot) 1 else 0, type = "l",
    bg = par("bg"), mar = c(4.1, 4.1, par("mar")[3], 0.1),
    myborder = 0.01, namefield = "name", lforce = "n", ...)
```

该函数的两个主要参数为地图数据库 database 和地图区域 region。地图数据库中包含了所有区域的经纬度数据以及相应的区域名称。在指定一个数据库和一系列区域名称之后，这些区域的地图便可由 maps() 生成。其它参数诸如填充颜色、是否画边界、是否添加到现有图形上等这里就不再介绍，请读者参考帮助文件。

运行下面的代码可以将表 A.3 的数据绘制成地图：

```
# 世界各国农业进出口竞争力地图
source(system.file("extdata", "AgriComp.R", package = "MSG"))
demo("AgriComp", package = "MSG")
```

上述代码的大致制作过程为：首先我们用 world 数据库作出一幅空白的世界地图，地区边界用灰色线条表示，然后我们根据竞争力数据中的地区名称与地理数据库中地区名称的对应将数据以颜色的形式表示到世界地图中，最后我们在图中添加了赤道线以及东盟国家（ASEAN）的矩形区域，这是由于作为该图出处的会议论文 (Xie, 2007) 主题是中澳自由贸易区。

maps 包功能虽然比较完善，但绘制地图的过程仍然有些繁琐。近年来，**ggplot2** 发展迅猛，其地图绘制功能也越来越强大了。图 A.7 实际是用 **ggplot2** 包绘制多边形的方式作出来的，代码如下：

```
# ggplot2 绘制世界各国农业进出口竞争力地图
source(system.file("extdata", "AgriComp.R", package = "MSG"))
library(ggplot2)
worldmap = ggplot2::map_data("world")
worldmap2 = dplyr::left_join(worldmap, AgriComp,
                             by = c("region" = "Country"))
p = ggplot(worldmap2) +
  geom_polygon(aes(long,lat, group=group,  fill = Index)) +
```

```
  coord_quickmap() +
  scale_fill_gradientn(colours = rev(rainbow(5))) +
  labs(x = "经度", y = "纬度") +
  geom_rect(data = data.frame(xmin = 72.26818, ymin = -55.8565,
                              xmax = 168.93766, ymax =54.589),
            mapping = aes(xmin = xmin, ymin = ymin,
                          xmax = xmax, ymax =ymax),
            fill = NA, color = "black", linetype = 2) +
  geom_hline(yintercept = 0, linetype = 2)
print(p)
```

这里，coord_quickmap() 函数，专门用作地图坐标，确保地图上的经度和纬度之比符合常用的摩克托投影规则。此外，**ggplot2** 还提供了另外一个坐标转换函数 coord_map()，功能更复杂一些。如图 A.8 所示，上图是平面地图，下图是以北纬 20° 东经 90° 视角看到的球状地图。生成这两幅地图的代码如下：

```
# 世界地图的两个视角
library(ggplot2)
library(patchwork)
m0 = ggplot() +
  geom_polygon(data = worldmap,
               mapping = aes(long,lat, group=group,  fill = region)) +
  guides(fill = FALSE)
m1 = m0 + coord_quickmap()
m2 = m0 + coord_map("ortho", orientation = c(20,90,0))
print(m1 / m2)
```

在地理区域上标记大量的数值信息会遇到一个显而易见的困难，就是由于各个地理区域的面积不同而导致地图的解读失真或某些重要地理单元难以辨认。例如，我们在画中国省级地图时，北京和上海等直辖市相比其它省份显得面积太小，此时若用颜色来标记某个数值指标（如 GDP），就会使得各个直辖市的颜色几乎无法辨认。

一个更有趣的例子来自 2008 年美国总统大选，如图 A.9。若用红蓝两种颜色对各个州做标记，以表示该州支持麦凯恩或奥巴马，那么有些面积不大但是权重很大的州（如人口众多的加州）就会影响整幅美国地图。从原始地图上看，似乎麦凯恩会赢，因为他赢得了很多中部面积大的州（但人口稀少），整幅地图看起来以红色为主导；若我们保持州的相

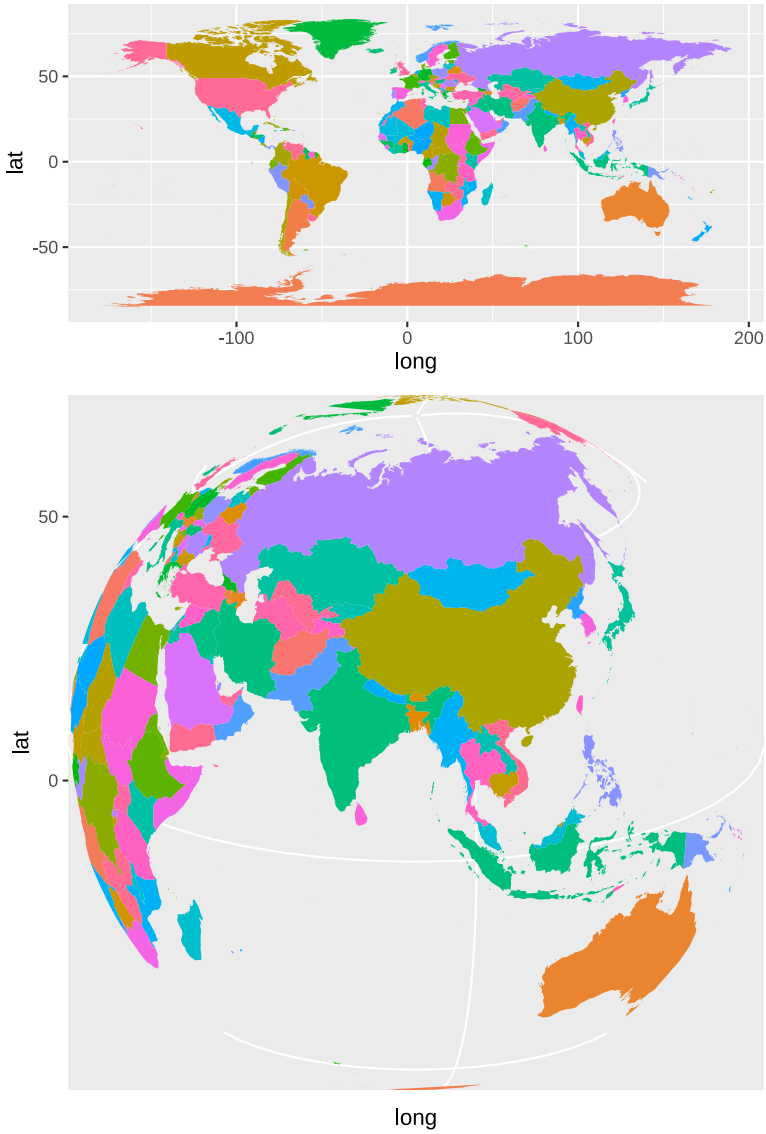


图 A.8: 不同视角和投影下的世界地图：上图是平面地图，下图是以北纬 $20°$ 东经 $90°$ 视角看到的球状地图

对地理位置不变，将各个州的形状进行大小的调整，使其面积与权重成正比，此时红蓝两色的局面就发生了逆转，地图以蓝色为主导色，地图传达信息的偏误才得到了纠正。我们把这种保持地理区域的相对位置不变、调整区域面积与某指标成比例的地图成为“变形地图”（Cartogram），详细内容可阅读笔者的博客[5]。

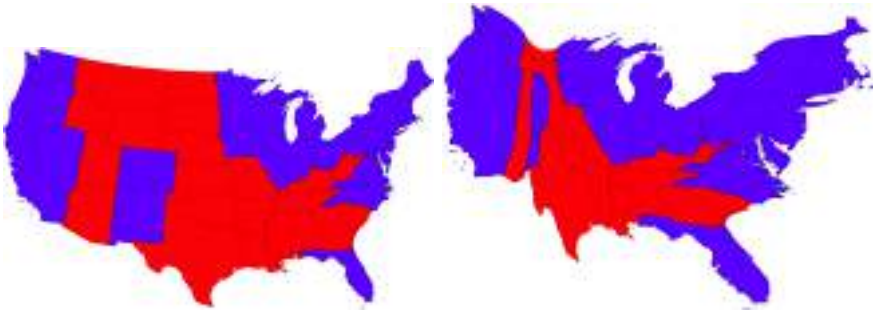


图 *A.9: 2008* 年美国总统大选各州投票情况：红蓝两种颜色分别表示该州支持麦凯恩或奥巴马。左图是正常地图，麦凯恩赢得了很多中部面积大的州（但人口稀少），整幅地图看起来以红色为主导；右图为变形地图，保持了地理区域的相对位置不变而调整区域面积与权重成比例，地图以蓝色为主导色

除了 **maps** 包和 **ggplot2** 包之外，R 语言还有更多的地图绘制系统，例如 **RgoogleMaps** 包 (Loecher and Ropkins, 2015)，将 Google Maps 提供的（卫星）地图数据引入 R 中，这里简单介绍一下。

首先，此包利用 Google Maps API，为 R 提供了一个十分便利的接口，以抓取 Google 服务器上的静态地图；其次，用户可使用获得的地图作为背景，在其上方自由叠加图形元素。对于一般的经纬度坐标数据，此包可计算包含这些数据点的矩形边界，以确定抓取地图的范围。其工作流程概括如下：

- 读取经纬度数据
- 通过计算确定获取图片所需参数
- 访问 Google Maps 服务器抓取图片
- 依据经纬度数据在图片上叠加图形元素

下面我们举一个利用 **RgoogleMaps** 包的例子。

[5] https://yihui.org/cn/2009/03/cartogram-as-special-maps/

表 A.4: 四川地区地震数据（部分。原数据为 354 行 3 列）

lat	long	ms
33.2	96.6	1.180
37.5	102.8	1.067
32.3	101.5	0.276
33.1	96.7	1.067
33.3	96.3	1.180

表 A.4 给出了来自中国国家地震科学数据共享中心的 354 条四川地区地震数据示例。3 个变量分别为震源的纬度、经度和震级大小（单位：面波震级 Ms），时间跨度为 2010 年 3 月 23 日到 2010 年 4 月 23 日。

图 A.10 显示了地震震源位置分布情况，背景采用了 Google Maps 提供的卫星地图数据。左图仅仅体现了震源位置的分布情况，不妨考虑将震级的大小映射为圆的半径大小。然而，图中存在着部分地震多发地带，如果使用圆来呈现震源的位置，这些区域的圆将出现严重的叠加现象，此处可以尝试使用 **??** 节中的透明度叠加来克服这类重叠问题，如右图所示。这里由于数据量不够大，这种透明度叠加的效果并不是非常明显。

本节具体的代码参见 `eqMaps` 演示：

```
# 在卫星地图上标记地震发生的地点和震级
demo("eqMaps", package = "MSG")
```

RgoogleMaps 包的潜力仍尚待挖掘。一方面，它可以展示空间分布信息，例如在 2010 年的 **ggplot2** 案例分析竞赛中，David Kahle 利用 **RgoogleMaps** 包和公开的犯罪信息数据，展示了休斯顿地区暴力犯罪的分布情况[6]；另一方面，如果数据包含时间属性，那么我们可以固定住抓取图片的边界，并保证叠加元素的坐标对应正确，便能制作出有用的动画。读者可以发挥想象力，拓展更多的应用情境。

本节只是介绍了一个非常简单的应用，但也引出了一个重要话题：统计图形如何与它要表达的问题的背景相融合？用通俗的话讲，就是要找“应景”的背景。在这方面，图 **??** 实际上做得很好，很有吸引眼球的效果，让人一看就明白要表达的主题。当然，背景元素也不能喧宾夺主，这一点在 8.2 小节中有详细论述。

[6] http://1t.click/Ksb

(a) (b)

图 *A.10:* 在卫星地图上标记地震发生的地点和震级：左图仅标记地点，右图用圆圈大小代表震级大小

参考文献

Becker, R. A., Wilks, A. R., Brownrigg, R., Minka, T. P., and Deckmyn, A. (2018). *maps: Draw Geographical Maps*. R package version 3.3.0.

Loecher, M. and Ropkins, K. (2015). RgoogleMaps and loa: Unleashing R graphics power on map tiles. *Journal of Statistical Software*, 63(4):1–18.

Xie, Y. (2007). *Visualization of Data and Statistical Models Using R*. Unpublished manuscript.

Zhao, P. and Cao, Y. (2020). *ncovr: Read and process nCoV data*. R package version 0.0.11.

Zou, Y., Pan, S., Zhao, P., Han, L., Wang, X., Hemerik, L., Knops, J., and van der Werf, W. (2020). Outbreak analysis with a logistic growth model shows covid-19 suppression dynamics in china. *PLOS ONE*, 15:1–10.