专题地图及其应用

草稿

黄湘云

2022-05-02

摘要

在美国和日本的国家统计局官网，地区分布图用于展示各类指标。衡量一个部门、一个业务、一个公司、一个行业，乃至一个国家都有一套金字塔式的指标体系，而国家每年发布的统计年鉴就包含一套衡量经济和社会发展情况的指标体系，涵盖人口、土地、生产、消费等等专题，省、市、区县以及自治区、州、县等各级地方统计局每年也会发布一份地方统计年鉴。

目录

插图目录

列表目录

> **重要**
>
> 本文引用的所有信息均为公开信息，仅代表作者本人观点，与就职单位无关。

1 本文概览

空间地理可视化的内容非常丰富，涉及空间坐标投影、操作空间数据、选择图形种类、选择绘图工具等。就图形种类而言，对标鼎鼎大名的收费 BI（Business Intelligence）工具Tableau，至少包含最常见的面量图、比例符号地图、点分布图、流线图、蜘蛛图（飞线图）、热图。其中的「面量图」通常又叫专题地图、地区分布图、统计地图，英文一般为Choropleth map，典型样例是基于统计年鉴的各类专题数据的地理可视化，国家地理信息公共服务平台提供了专题图层服务，可以快速地查看各个统计指标。在美国和日本的国家统计局官网，地区分布图用于展示各类指标 (Meyer, Broome, 和 Jr 1975)。衡量一个部门、一个业务、一个公司、一个行业，乃至一个国家都有一套金字塔式的指标体系，而国家每年发布的统计年鉴就包含一套衡量经济和社会发展情况的指标体系，涵盖人口、土地、生产、消费等等专题，省、市、区县以及自治区、州、县等各级地方统计局每年也会发布一份地方统计年鉴。

接下来，本文分四个部分展开介绍地区分布图，分别是单变量情形、多变量情形、本文小结和未来展望。

单变量情形中以 **latticeExtra** 包 (Sarkar 和 Andrews 2019) 内置的数据集 USCancerRates 为例，以地区分布图形式展示美国 1999-2003 年度各郡的年平均癌症死亡率，此处专题的含义是「人口死亡率」，显而易见，癌症死亡率只是一方面，还有流感死亡率等，癌症可以分类型，如乳腺癌、子宫癌等，人又可以分属性，如性别、年龄、种族等等。在数据操作、指标计算和分面绘图等方面从零开始介绍绘制地区分布图的过程，包括基础数据操作以及六个绘图工具 **maps** 包 (Becker 和 Wilks 1993)、**latticeExtra** 包、**ggplot2** 包 (Wickham 和 Girlich 2022)、**tmap** 包 (Tennekes 2018)、**sf** 包 (Pebesma 2018) 和 **mapsf** 包 (Giraud 2022)，阐述数据指标「年平均癌症死亡率」的实际含义、指标口径和计算过程，从易到难，层层深入，以期达到出版级的水准，探索出最佳实践。

多变量情形中以美国人口调查局发布的调查数据为基础，分析北卡罗来纳州各郡社区普查级的家庭年收入与白人占比的空间相关性。先以单变量的地区分布图描述各个普查区域里家庭年收入的空间分布，接着和二元变量的地区分布图形成对比，展示相关性的空间分布。

本文小结部分给出了 7 种不同绘图方案间的关系和一些绘图经验，希望帮助读者加深理解和学习。

未来展望部分从应用场景和绘图技术方面继续提供一些示例，供读者继续探索。

2 单变量情形

2.1 美国各郡的年平均癌症死亡率分布

下面以 **latticeExtra** 包(Sarkar 和 Andrews 2019) 内置的 USCancerRates 数据集为例介绍分面，同时展示多个观测指标的空间分布。USCancerRates 数据集来自美国国家癌症研究所（National Cancer Institute，简称 NCI)。根据 1999-2003 年的 5 年数据，分男女统计癌症年平均死亡率（单位十万分之一)，这其中的癌症数是所有癌症种类之和。癌症死亡率根据 2000 年美国标准人口年龄分组调整，分母人口数量由 NCI 根据普查的人口数调整，即将各年各个年龄段的普查人口数按照 2000 年的**美国标准人口年龄分组**换算。因 **latticeExtra** 包没有提供数据集的加工过程，笔者结合 NCI 网站信息，对此数据指标的调整过程略加说明，这里面其实隐含很多的道理。

人口数每年都会变的，为使各年数据指标可比，人口划分就保持一致，表格 1 展示 1940-2000 年各个年龄段（共 19 个年龄组）的标准人口数，各个年龄段的普查人口数换算成年龄调整的标准人口数，换算公式为:

$$\text{某年龄段标准人口数} = \text{某年龄段普查人口数}/\text{总普查人口数} * 1000000.$$

以 2000 年的 10-14 岁年龄段标准人口数为例，即:

$$73032 = 20056779/274633642 * 1000000.$$

表格 1: 1940-2000 年美国标准人口分组

Age	2000	1990	1980	1970	1960	1950	1940
00	13,818	12,936	15,598	17,151	22,930	20,882	15,343
01-04	55,317	60,863	56,565	67,265	90,390	86,376	64,718
05-09	72,533	72,772	73,716	98,204	104,235	87,591	81,147
10-14	73,032	68,812	80,523	102,304	93,538	73,785	89,208
15-19	72,169	71,384	93,439	93,845	73,717	70,450	93,670
20-24	66,478	76,476	94,103	80,561	60,231	76,191	88,007
25-29	64,529	85,694	86,168	66,320	60,612	81,237	84,277
30-34	71,044	87,905	77,516	56,249	66,635	76,425	77,789
35-39	80,762	80,267	61,644	54,656	69,601	74,629	72,495
40-44	81,851	70,829	51,510	58,958	64,689	67,712	66,742
45-49	72,118	55,778	48,951	59,622	60,670	60,190	62,697
50-54	62,716	45,638	51,689	54,643	53,568	54,893	55,114
55-59	48,454	42,345	51,271	49,077	47,009	48,011	44,383
60-64	38,793	42,685	44,528	42,403	39,830	40,210	35,911

Age	2000	1990	1980	1970	1960	1950	1940
65-69	34,264	40,657	38,767	34,406	34,897	33,199	28,911
70-74	31,773	32,145	30,008	26,789	26,427	22,641	19,515
75-79	26,999	24,612	21,160	18,871	17,028	14,283	11,422
80-84	17,842	15,817	12,956	11,241	8,811	7,467	5,881
85+	15,508	12,385	9,888	7,435	5,182	3,828	2,770

年龄调整的比率（Age-adjusted Rates）的定义详见NCI 网站，它是一个根据年龄调整的加权平均数，权重根据年龄段人口在标准人口中的比例来定，一个包含年龄 x 到年龄 y 的分组，其年龄调整的比率计算公式如下：

$$aarate_{x-y} = \sum_{i=x}^{y} \left[(\frac{count_i}{pop_i}) \times (\frac{stdmil_i}{\sum_{j=x}^{y} stdmil_j}) \times 100000 \right]$$

一个具体的例子可见网站，篇幅所限，此处仅以 2000 年举例，一个年龄段 00 years 死亡人数 **29**（可看作婴儿死亡人数），总人数 **139879**，则年龄调整的死亡率：

$$aarate_{0-0} = \frac{29}{139879} * \frac{3794901}{274633642} * 100000 = 0.2864$$

读者可能有疑惑，一系列复杂的调整是为什么？指标稳定性和可比性。稳定不是代表不变，稳定是不轻易受干扰。从各社区、各郡、各州乃至国家，从下往上聚合数据的时候，分年龄、种族、性别等下钻/上卷的时候，有的郡总人口可能相对很少，死亡人数也很少。可比性是指组与组间可比，且随时间变化依然可比，刻画因癌症死亡的相对风险。

```
# 加载死亡率数据
data(USCancerRates, package = "latticeExtra")
# 查看 Alabama 的 Pickens County 的数据
subset(x = USCancerRates, subset = state == "Alabama" & county == "Pickens County")
#                 rate.male LCL95.male UCL95.male rate.female LCL95.female
# alabama,pickens     363.7      311.1      423.2         151        123.6
#                 UCL95.female   state         county
# alabama,pickens        183.6 Alabama Pickens County
```

以 Alabama 的 Pickens County 为例，1999-2003 年平均年龄调整的男性癌症死亡率为 363.7（单位：十万分之一），在 95% 置信水平下，置信限为 [311.1, 423.2]。根据最新的五年数据显示 2014-2018 年男性癌症死亡率为 479.8，95% 置信水平下的置信区间为 [425.7, 539.3]。简单验证一下，就会发现有意思的现象，置信区间不是关于观测的癌症死亡率对称，且离置信区间中心尚有距离，$\frac{311.1+423.2}{2} = 367.1 \neq 363.7$。

一般来说，100000 人中有 363.7 人因癌症死亡，死亡人数较多（比如大于 100）的情况下，二项分布可用正态分布逼近，置信区间上下限应该分别为：

```
qnorm(p = 1 - 0.05 / 2)
# [1] 1.959964
# 置信下限
363.7 - 1.96 * sqrt(363.7 / 100000 * (1 - 363.7 / 100000) / 100000) * 100000
# [1] 326.389
# 置信上限
363.7 + 1.96 * sqrt(363.7 / 100000 * (1 - 363.7 / 100000) / 100000) * 100000
# [1] 401.011
```

而美国国家癌症研究所给的置信带更宽，更保守一些，显然这里面的算法没这么简单。以阿拉巴马州为例，将所有的郡死亡率及其置信区间绘制出来，如图 1 所示，整体来说，偏离置信区间中心都很小。

不难看出，女性癌症死亡率整体上低于男性，且各个地区的死亡率有明显差异。NCI 网站仅对置信区间的统计意义给予解释，这跟统计学课本上没有太多差别，没有提供具体的计算过程。可以推断的是必然使用了泊松、伽马一类的偏态分布来刻画死亡人数的分布，疑问尚未解开，欢迎大家讨论。

> ⚠ 警告
>
> 癌症死亡率相关数据仅可用于统计报告和分析，不可用于其他目的，请遵守相关法律规定。

参考文献

Becker, Richard A., 和 Allan R. Wilks. 1993. 《Maps in S》. 2. 卷 93. Statistics Research Report. AT&T Bell Laboratories. https://web.archive.org/web/20050825145143/http://www.research.att.com/areas/stat/doc/93.2.ps.

Giraud, Timothée. 2022. *mapsf: Thematic Cartography*. https://CRAN.R-project.org/package=mapsf.

Meyer, Morton A., Frederick R. Broome, 和 Richard H. Schweitzer Jr. 1975. 《Color Statistical Mapping by the U.S. Bureau of the Census》. *The American Cartographer* 2 (2): 101–17. https://doi.org/10.1559/152304075784313250.

Pebesma, Edzer J. 2018. 《Simple Features for R: Standardized Support for Spatial Vector Data》. *The R Journal* 10 (1): 439–46. https://doi.org/10.32614/RJ-2018-009.

Sarkar, Deepayan, 和 Felix Andrews. 2019. *latticeExtra: Extra Graphical Utilities Based on Lattice*. https://CRAN.R-project.org/package=latticeExtra.

Tennekes, Martijn. 2018. 《tmap: Thematic Maps in R》. *Journal of Statistical Software* 84 (6): 1–39. https://doi.org/10.18637/jss.v084.i06.

Wickham, Hadley, 和 Maximilian Girlich. 2022. *tidyr: Tidy Messy Data*. https://CRAN.R-project.

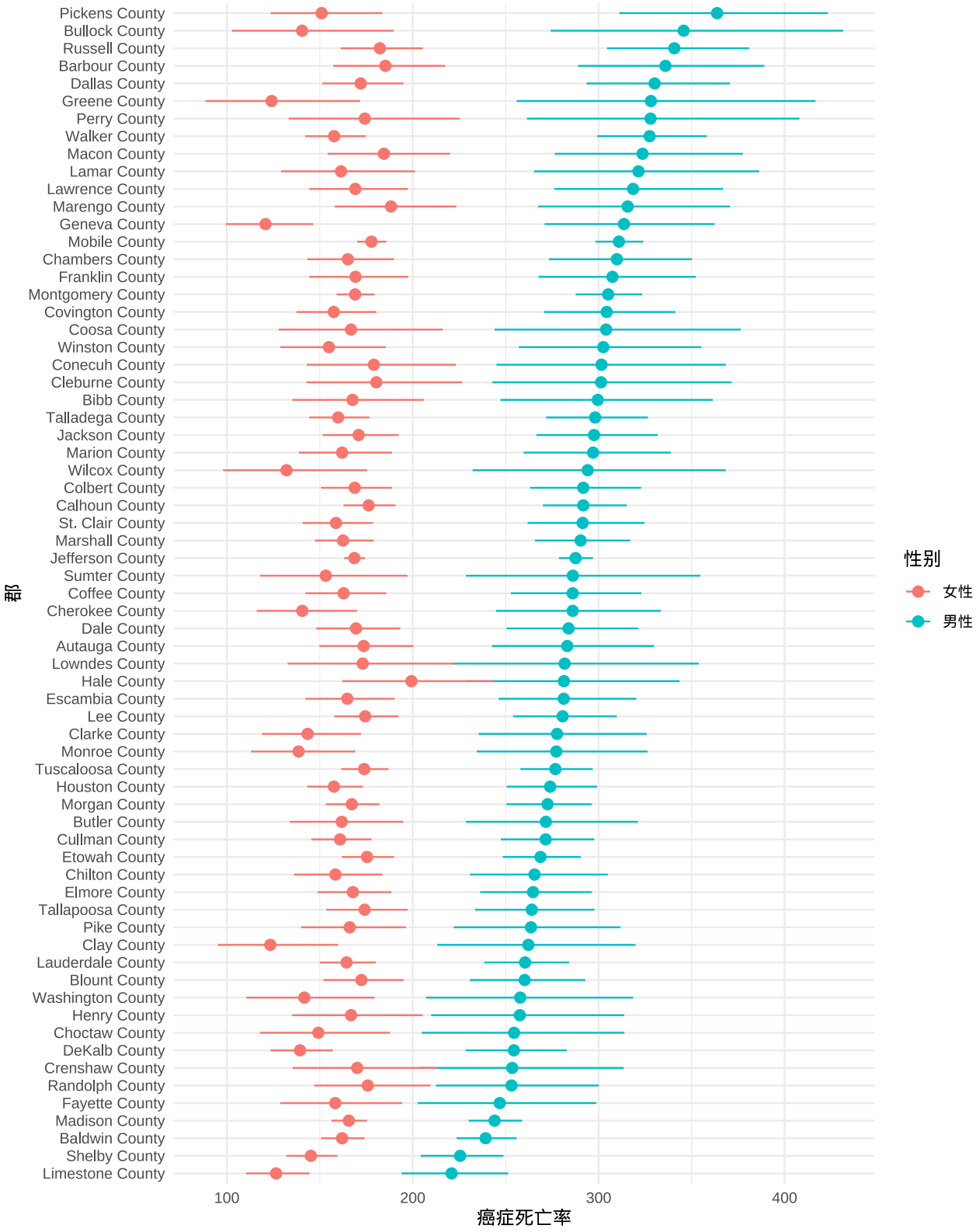

图 1: 1999-2003 年美国阿拉巴马州各个郡的年平均癌症死亡率

org/package=tidyr.