Springer

R语言应用系列

新视野·数据科学

统

ggplot2：数据分析与图形艺术

ggplot2. Elegant Graphics for Data Analysis

〔美〕哈德利·威克姆 著

统计之都 译

殷腾飞 审校

Hadley Wickham

西安交通大学出版社

XI'AN JIAOTONG UNIVERSITY PRESS

ℛ语言应用系列

ggplot2. Elegant Graphics for Data Analysis

ggplot2:数据分析与图形艺术

[美] 哈德利·威克姆 著
Hadley Wickham

统计之都 译
殷腾飞 审校

西安交通大学出版社
Xi'an Jiaotong University Press

陕西省版权局著作权合同登记号 图字 25-2012-216 号

图书在版编目(CIP)数据

ggplot2：数据分析与图形艺术/(美) 威克姆(Wickham，H.)著；统计之都译. —西安：西安交通大学出版社，2013.5
书名原文：Ggplot2. elegant graphics for data analysis
ISBN 978-7-5605-4969-9

Ⅰ.①g… Ⅱ.①威… ②统… Ⅲ.①统计分析-应用软件 Ⅳ.①C819

中国版本图书馆 CIP 数据核字(2013)第 006344 号

书　　名	ggplot2：数据分析与图形艺术
著　　者	(美) 哈德利·威克姆
译　　者	统计之都
审　　校	殷腾飞
出版发行	西安交通大学出版社 (西安市兴庆南路 10 号　邮政编码 710049)
网　　址	http://www.xjtupress.com
电　　话	(029)82668357　82667874(发行中心) (029)82668315　82669096(总编办)
传　　真	(029)82668280
印　　刷	陕西宝石兰印务有限责任公司
开　　本	720mm×1000mm　1/16　　**印张** 15.5
印　　数	0001～3000　　**字数** 253 千字
版次印次	2013 年 5 月第 1 版　2013 年 5 月第 1 次印刷
书　　号	ISBN 978-7-5605-4969-9/C·107
定　　价	46.00 元

读者购书、书店添货、如发现印装质量问题，请与本社发行中心联系、调换。
订购热线：(029)82665248　(029)82665249
投稿热线：(029)82665397
读者信箱：banquan1809@126.com

中译本序

每当我们看到一个新的软件，第一反应会是：为什么又要发明一个新软件？ggplot2 是 R 世界里相对还比较年轻的一个包，在它之前，官方 R 已经有自己的基础图形系统 (graphics 包) 和网格图形系统 (grid 包)，并且 Deepayan Sarkar 也开发了 lattice 包，看起来 R 的世界对图形的支持已经足够强大了。那么我们不禁要问，为什么还要发明一套新的系统？

设计理念

打个比方，想想我们小时候怎样学中文的。最开始的时候我们要识字，不认识字就没法阅读和写作，但我们并不是一直按照一个个汉字学习的，而是通过句子和具体的场景故事学习的。为什么不在小学时背六年字典呢？那样可能认识所有的汉字。原因很简单，光有单字，我们不会说话，也无法阅读和写作。我们缺的是什么？答案是对文字的组织能力，或者说语法。

R 的基础图形系统基本上是一个“纸笔模型”，即：一块画布摆在面前，你可以在这里画几个点，在那里画几条线，指哪儿画哪儿。后来 lattice 包的出现稍微改善了这种情况，你可以说，我要画散点图或直方图，并且按照某个分类变量给图中的元素上色，此时数据才在画图中扮演了一定的中心角色，我们不用去想具体这个点要用什么颜色 (颜色会根据变量自动生成)。然而，lattice 继承了 R 语言的一个糟糕特征，就是参数设置铺天盖地，足以让人窒息，光是一份 `xyplot()` 函数的帮助文档，恐怕就够我们消磨一天时间了，更重要的是，lattice 仍然面向特定的统计图形，像基础图形系统一样，有直方图、箱线图、条形图等等，它没有一套可以让数据分析者说话的语法。

那么数据分析者是怎样说话的呢？他们从来不会说这条线用 `#FE09BE` 颜色，

那个点用三角形状，他们只会说，把图中的线用数据中的职业类型变量上色，或图中点的形状对应性别变量。有时候他们画了一幅散点图，但马上他们发现这幅图太拥挤，最好是能具体看一下里面不同收入阶层的特征，所以他们会说，把这幅图拆成七幅小图，每幅图对应一个收入阶层。然后发现散点图的趋势不明显，最好加上回归直线，看看回归模型反映的趋势是什么，或者发现图中离群点太多，最好做一下对数变换，减少大数值对图形的主导性。

从始至终，数据分析者都在数据层面上思考问题，而不是拿着水彩笔和调色板在那里一笔一划作图，而计算机程序员则倾向于画点画线。Leland Wilkinson 的著作在理论上改善了这种状况，他提出了一套图形语法，让我们在考虑如何构建一幅图形的时候不再陷在具体的图形元素里面，而是把图形拆分为一些互相独立并且可以自由组合的成分。这套语法提出来之后他自己也做了一套软件，但显然这套软件没有被广泛采用；幸运的是，Hadley Wickham 在 R 语言中把这套想法巧妙地实现了。

为了说明这种语法的思想，我们考虑图形中的一个成分：坐标系。常见的坐标系有两种：笛卡尔坐标系和极坐标系。在语法中，它们属于一个成分，可自由拆卸替换。笛卡尔坐标系下的条形图实际上可以对应极坐标系下的饼图，因为条形图的高可以对应饼图的角度，本质上没什么区别。因此在 ggplot2 中，从一幅条形图过渡到饼图，只需要加极少量的代码，把坐标系换一下就可以了。如果我们用纸笔模型，则可以想象，这完全是不同的两幅图，一幅图里面要画的是矩形，另一幅图要画扇形。

更多的细节在本书中会介绍，这里我们只是简略说明用语法画图对用纸笔画图来说在思维上的优越性；前者是说话，后者是说字。

发展历程

ggplot2 是 Hadley 在爱荷华州立大学博士期间的作品，也是他博士论文的主题之一，实际上 ggplot2 还有个前身 ggplot，但后来废弃了，某种程度上这也是 Hadley 写软件的特征，熟悉他的人就知道这不是他第一个“2”版本的包了 (还有 reshape2)。带 2 的包和原来的包在语法上会有很大的改动，基本上不兼容。尽管如此，他的 R 代码风格在 R 社区可谓独树一帜，尤其是他的代码结构很好，可读性很高，ggplot2 是 R 代码抽象的一个杰作。读者若感兴趣，可以在 GitHub 网站上浏览他的包：`https://github.com/hadley`。在用法方面，

ggplot2 也开创了一种奇特而绝妙的语法，那就是加号：一幅图形从背后的设计来说，是若干图形语法的叠加，从外在的代码来看，也是若干 R 对象的相加。这一点精妙尽管只是 ggplot2 系统的很小一部分，但我个人认为没有任何程序语言可比拟，它对作为泛型函数的加号的扩展只能用两个字形容：绝了。

至 2013 年 2 月 26 日，ggplot2 的邮件列表 (http://groups.google.com/group/ggplot2) 订阅成员已达 3394 人，邮件总数为 15185 封，已经成为一个丰富、活跃的用户社区。未来 ggplot2 的发展也将越来越依赖于用户的贡献，这也是很多开源软件最终的走向。

关于版本更新

原书面世之时，ggplot2 的版本号是 0.8.3，译者开始翻译此书时是 0.9.0 版本；该版本较之 0.8.3，内部做了一些大改动。此后，ggplot2 频繁升级，目前版本号是 0.9.3，当然这也给本书的翻译过程带来了相当大的麻烦。因为译者不但要修正原书中大量过时的代码、重新画图，还要修正过时的理念，以及处理数次版本更新的影响。所幸，在翻译过程中，译者得到了本书审校殷腾飞博士、ggplot2 开发者 Hadley Wickham 和 Wistong Chang 的大力帮助。

如果你是老用户，那么可能需要阅读下面的小节。之后 ggplot2 有过多次更新，尤其是 0.9.0 之后，ggplot2 的绘图速度和帮助文档有了质的飞跃。关于 0.9.0 的更新，读者可以从 https://github.com/downloads/hadley/ggplot2/guide-col.pdf 下载一份详细的说明文档，但原文档比较长，而且有些内部更新问题我们也不一定需要了解，因此这里给一段概述。

- ggplot2 的帮助文档大大扩充了，过去头疼的问题之一就是一个函数里面不知道有哪些可能的参数，例如 theme() 函数，现在已经有了详细说明。
- 新增图例向导函数 guide_legend() 和 guide_colorbar()，前者可以用来指导图例的排版，例如可以安排图例中元素排为 n 行 m 列；后者增强了连续变量图例的展示，例如当我们把颜色映射到一个连续变量上时，过去生成的图例是离散的，现在可以用这个函数生成连续颜色的图例 (渐变色)。
- 新增几何对象函数 geom_map()(让地图语法变得更简单)，geom_raster

()(更高效的geom_tile())，geom_dotplot()(一维点图，展示变量密度分布) 和 geom_violin()(小提琴，实为密度曲线)。

- 新增统计变换函数 stat_summary2d()(在二维网格上计算数据密度)，stat_summary_hex()(在六边形“蜂巢”上计算数据密度)，stat_bindot()(一维点图密度)，stat_ydensity()(密度曲线，用于小提琴图)。
- facet_grid() 支持 x 轴和 y 轴其中一者可以有自由的刻度 (根据数据范围而定)，以往要么所有切片使用同样的坐标轴刻度，要么所有都自由。
- geom_boxplot() 开始支持画箱线图的凹槽 (notch)，就像 R 基础图形系统中的 boxplot() 函数。
- 新增函数 ggmissing() 用来展示缺失值的分布，ggorder() 按照数据观察顺序先后画折线图，ggstructure() 展示数据热图。

另外此次更新涉及到一些函数参数名称的变化，如果旧代码在这个版本中报错说有未使用的参数，那么用户需要再次查看帮助文档，确保输入的参数在函数中存在。在所有这些表面的更新背后，实际上 ggplot2 很大程度上被重写了，例如开始使用 R 自带的 S3 泛型函数设计，以及将过去 ggplot2 的功能继续模块化为一些独立的包，一个典型的例子就是标度部分的功能被抽象到 scales 包中，从数据映射到颜色、大小等外观属性可以由这个包直接完成。这种分拆也使得其他开发者可使用过去 ggplot2 内部的一些功能函数。

0.9.1 版本主要解决了 0.9.0 版本中的一些漏洞。ggplot2 在 2012 年 9 月 4 日发布了新的版本 0.9.2，其中一些特性和更新有必要提及：

- 采用了全新的主题 (theme) 系统，opts() 函数已被标记为“不推荐使用” (deprecated)，将在未来版本中被取消，取而代之的是 theme() 函数，主题元素 (theme element) 由属性列表构成，支持继承，主题之间可以直接进行合并等操作。详情参见 wiki 页面: https://github.com/wch/ggplot2/wiki/New-theme-system。
- 依赖于新的 gtable 包。用来更方便地调整修改 ggplot2 图形中的图元，ggplotGrob() 会返回一个 gtable 类，这个对象可以利用 gtable 包中提供的函数和接口进行操作。
- 所有“模板”类型的图形函数，比如 plotmatrix(), ggorder() 等等，已被标记为“不推荐使用” (deprecated)，将在未来版本中取消。

在本书出版之际，ggplot2 更新到了版本 0.9.3，修复了 0.9.2 的一些漏洞，其主要更新包括

- 不再支持 `plotmatrix()` 函数。
- `geom_polygon()` 提速，比如世界地图的绘制快了 12 倍左右。
- 新增部分主题，比如 `theme_minimal()`, `theme_classic()`。

本书的所有代码和图片都是针对新版本 0.9.3 的，在内容方面也根据版本更新对原文做了适当的增删填补，以满足读者的需求。

本书把影响正文阅读的彩图集中放在附录后面，读者可以随时翻阅。

致谢

在听说我们翻译完这本书之后，本书原著 Hadley 很高兴，给我们发邮件说：

> I am excited and honoured to have my book translated to Chinese. ggplot2 has become far more popular than I ever imagined, and I'm excited that this translation will allow many more people to learn ggplot2. I'm very grateful that Yihui and his team of translators (Nan Xiao, Tao Gao, Yixuan Qiu, Weicheng Zhu, Taiyun Wei and Lanfeng Pan) made this possible.
>
> One of the biggest improvements to ggplot2 since the book was first written is the ggplot2 mailing list. This is a very friendly environment where you can get help with your visualisations, and improve your own knowledge of ggplot2 by helping others solve their problems. I'd strongly encourage you to join the mailing list, even if you think your English is not very good — we are very friendly people.

我们感谢这本书的译者，包括邱怡轩 (第 1~2 章)、主伟呈 (第 3~4 章)、肖楠 (第 5~6 章)、高涛 (第 7~8 章)、潘岚锋 (第 9 章)、魏太云 (第 10 章、附录以及翻译过程的协调安排和全书的 LaTeX 排版工作)。所有译者均来自于统计之都 (`http://cos.name`)。

爱荷华州立大学的殷腾飞博士、中国人民大学统计学院的孟生旺教授、浙江大学的张政同学通读了译稿，提出了很多有用的建议，殷腾飞博士还提供了大多数新版本中的解决方案，并担任本书的审校。肖凯老师和余光创博士分别对第 1~4 章、第 8~10 章以及附录提出了很多修改意见，此外，中国人民大学的陈妍、李晓矛、谢漫锜三位同学、中国再保险公司的李皞先生、百度公司的韩帅先生、eBay 公司的陈丽云女士、Mango Solutions 公司的李舰先生、京东商城的刘思喆先生、首钢总公司的邓一硕先生、新华社的陈堰平先生在此书的翻译过程中也曾提过不少宝贵的建议，在此一并表示感谢。

为了更好地服务社区，我们还建立了翻译主页：`https://github.com/cosname/ggplot2-translation`，读者可以在这里得到最新的勘误和书中的代码，也可以随时提出任何问题。

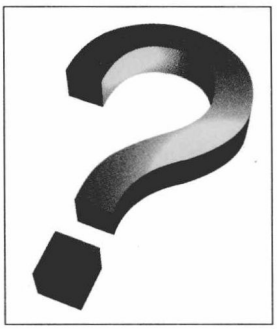
谢益辉

2013 年 2 月 26 日

即将出版

数据科学中的R语言——基础框架和行业应用

李　舰　肖　凯 著



本书是一本R语言实战类书籍，目标群体为缺乏编程或者统计基础，但希望能从零开始深入地理解并应用R语言的读者。全书分为编程篇、模型篇和应用篇，从R的语言特性和分析方法讲起，帮助读者快速入门，然后循序渐进地使读者跟随书中的例子进入到进阶训练，最后应用到实际的案例中。

本书的特点在于行业应用的真实案例。包含了大量从传统的统计分析领域如新药研发、金融分析到当前最热门的大数据、社交网络等应用的例子。作者把从业以来积累的R语言在各行业中的应用案例第一次公开出版奉献给读者。书中所有的案例和代码都会做成R包发布在CRAN上，供读者进行学习和重用。

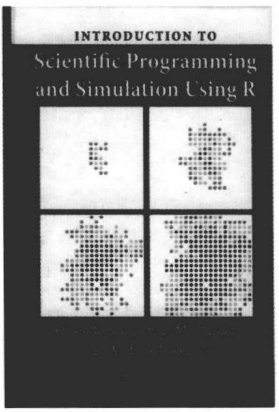
2013年12月出版　380页　R语言应用系列

R语言的科学编程与仿真

Owen Jones　Robert Maillardet　Andrew Robinson 著

王　亮　周丙常(西北工业大学)

王　亮(西安电子科技大学)　译



这本书主要介绍了科学编程与随机建模的一些技巧。旨在整合科学编程和概率论，尤其是通过数值模拟实现现代概率统计理论的应用。随书附带的spuRs包含了大部分提到的程序，可供读者调试使用。

不同于大部分R的指导书，本书不仅介绍了统计方法的应用，还介绍了如何将算法转化为代码。这一点对于那些不仅希望使用代码，而且想要设计代码的人是很有用的。

2013年11月出版　400页　R语言应用系列

目录

[①] 本节为适应 0.9.0 及之后版本而重写

第 1 章　简介

1.1　欢迎来到 ggplot2 的世界

ggplot2 是一个用来绘制统计图形 (或称为数据图形) 的 R 软件包。与其他大多数的图形软件包不同，ggplot2 是由其背后的一套图形语法所支持的。这一语法基于《Grammar of Graphics》(Wilkinson, 2005) 一书，它由一系列独立的图形部件组成，并能以许多种不同的方式组合起来。这一点使得 ggplot2 的功能非常强大，因为它不会局限于一些已经定义好的统计图形，而是可以根据你的需要量身定做。这听起来似乎有些困难，但实际上只需要掌握一些核心准则以及少许特例，ggplot2 还是很容易学习的 (尽管你可能需要花费一些时间来忘掉其他图形工具中一些固有的概念)。

ggplot2 可以绘制出很多美观的图形，同时能避免诸多繁琐的细节，例如添加图例等。用 ggplot2 绘图时，图形的每个部分可以依次进行构建，之后还可以进行编辑。ggplot2 精心挑选了一系列预设图形，因此在大部分的情形下你可以快速地绘制出许多高质量的图形。如果在格式上还有额外的需求，也可以利用 ggplot2 中的主题系统来进行定制。这样，你将无需花费太多时间来调整图形的外观，而可以更加专注地用图形来展示你的数据。

ggplot2 采用了图层的设计方式，你可以从原始的图层开始，首先绘制原始数据，然后不断地添加图形注释和统计汇总结果。这种绘图方式与分析问题中的结构化思维是一致的，它能缩短你“所思”与“所见”的距离。特别地，ggplot2 可以帮助学生锻炼结构化的分析思维，进而达到专业的水准。

学习图形语法不仅可以帮助你绘制出你已经了解的图形，甚至还可以启发你创作出更佳的方案。如果没有这一套语法体系，图形的绘制便失去了理论支持，这也就是为什么现有的很多图形软件包只是一系列特例的堆积。例如，在