基于局部 45° 原则的折线图坐标轴网格生成

包琛, 汪云海*

(山东大学计算机科学与技术学院 青岛 266237)
(wang.yh@sdu.edu.cn)

摘要: 坐标轴网格是辅助人们观察统计折线图的重要参考结构。文中研究了有利人们感知直线斜率的坐标轴标签。首先通过一系列用户实验探索折线图中网格局部长宽比对于估计线段斜率这一任务的影响, 通过用户估计的倾斜角度与直线真实的倾斜角度的误差大小反映不同实验条件的影响程度。实验得出的结论是适当的网格密度、接近正方形的网格格子局部长宽比有利于准确的斜率估计。基于该实验结论, 提出了新的为折线图数据自动生成坐标轴的算法, 考虑最大标签、最小标签和间隔步长 3 个变量, 参考前人关于标签简洁性、覆盖性、重叠程度的定义, 将正方形的局部长宽比这一因素加入水平和垂直 2 条坐标轴的标签联合优化之中, 该算法能够生成既美观又提高图形感知准确性的坐标轴网格。

关键词: 坐标轴生成; 背景网格; 图形感知; 斜率估计

中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2021.18660

Local Banking to 45°: Axis Grid Generation for Line Charts

Bao Chen and Wang Yunhai*

(School of Computer Science and Technology, Shandong University, Qingdao 266237)

Abstract: Axis grid is a basic reference structure that assists people in reading line charts. In this paper, we focus on axis labels that help people perceive slopes in line charts. A series of user studies are conducted to explore the influence of local aspect ratio on the task of slope estimation in line charts. We evaluate the influence of different experimental conditions by calculating the errors between the user estimated slope and the ground truth. The results show that appropriate grid density and squarified grid aspect ratio increase the estimation accuracy of line segment slopes. Accordingly, we propose a new algorithm for generating axis grid considering the maximal label, the minimal label and label intervals for line charts. Following the previous criteria of simplicity, coverage and overlap, our algorithm applies the squarified local aspect ratio to the joint optimization of the tick marks of both horizontal and vertical axis to generate aesthetic axis grid that helps increase graphical perception accuracy.

Key words: axis generation; background grid; graphical perception; slope estimation

统计图表, 如折线图、散点图和柱状图等, 在数据分析的专业人群中有着非常广泛的应用。研究者致力于产生有效的参考辅助结构, 以提高人们在人机交互实验中的感知能力。Bertin^[1], Cleveland 等^[2]和 Tufte 等^[3]的开创性工作中包括对统计图表

中代表数据的标记本身进行的研究, 如点和线等。

本文研究的是作为背景的坐标轴标签网格。网格对人们的图形感知可能存在主动或被动的支持帮助作用。到目前为止, 仅有极少数的研究者关注过这一方向。Bartram 等^[4-5]探究了不同透明度的背

收稿日期: 2020-09-13; 修回日期: 2020-11-16。基金项目: 国家自然科学基金(61772315, 61861136012)。包琛(1995—), 女, 硕士研究生, 主要研究方向为信息可视化; 汪云海(1984—), 男, 博士, 教授, 博士生导师, 论文通讯作者, 主要研究方向为数据可视化、人机交互。

景网格的影响。另外,还有一些研究者为如何在坐标轴上安放标签设计出了若干指导策略和自动生成的支持算法^[6-8]。这些研究内容与背景网格息息相关,但是对于网格本身在认知中扮演了什么角色尚未定义。Talbot等^[9]可视化科学家曾提出过这样一个假设:附加的背景线作为基准线可能对于从折线图中估计斜率有较大的影响。

本文研究了折线图中坐标轴网格的不同密度以及局部长宽比对斜率估计任务的影响,并做了一系列定量实验。首先排除了网格线与直线交叉位置的干扰,最终得出的结论是适宜密度、单个格子长宽比接近1:1的标签网格能够更好地支持斜率估计任务。本文将这一结论命名为“局部45°原则”(local banking to 45°),在这一原则的指导下,提出了新的折线图坐标轴标签自动生成算法,并展示了一个实例。

1 相关工作

图形感知是对蕴含在图中的信息进行视觉解码的过程^[10]。在图形感知过程中,可视化组件和文本组件共同协作,将数据的值映射成图形表示^[2],并呈现出定量信息^[11]。这些辅助结构的设计需要能够呈现想要表达的潜在信息,但又不能太突兀显眼,以致影响了人们对数据的常规感知^[5]。这种对图表参考辅助结构的精巧设计称为“组件级别的设计”^[11]。本文将关注点放在坐标轴标签和背景网格这2种参考组件上。

1.1 坐标轴标签和背景网格的生成

标签和网格是一幅图表中最常见的基本辅助结构。标签包括2个部分:(1)与坐标轴交叉以标记缩放后的值的细小刻度线;(2)与刻度线位置对齐的文字标注,通常为数字^[11]。网格线是由这些刻度线生成的并且是与相应坐标轴上的刻度线对齐的^[12]。标签和网格互为参考,互相补充,增强了人们对可视化数据的感知。

1.1.1 坐标轴标签

刻度标签能够支持基础的数据分析任务,如数值查找、范围估计和斜率估计^[7]。Heckbert^[6]依据他观察所得的一系列人们日常生活习惯用的数字以及数据在单一坐标轴的最大值和最小值,生成该坐标轴上合适的间隔步长,并返回给用户若干与用户输入的期望标签数相近的一组标签。R语言

库^[13]中的pretty的功能^①是Heckbert经典算法的一个变体。除了这些基于公式的方法,Wilkinson^[8]提出了一个从一组可能的标签生成方案中进行优化搜索的方法,定义了简洁性、覆盖性和间隔度3个变量属性,通过搜索找到一个优化方案使加和得分最高。

上述方法都是基于数据数值范围计算标签值的,都没有考虑图形组件和视觉空间的因素。因而有时候会产生从感知角度看并不理想的结果,如很多细长的背景网格聚在一起,干扰了人们对数据信息的获取。

Talbot等^[7]提出了一种拓展Wilkinson方案的算法,考虑了标签形状的3个相关因素:标签数字表示格式(如使用科学计数法表示以缩短标签,节省空间)、字体大小和标签方向;并且修改了Wilkinson的公式,增加了一些参数,产生了一种更稳健的搜索方法,能够得到具有更好的数据覆盖性和密度的标签组。然而,这种方法也避免不了可能会出现一些细长的网格的结果图。

另外,上述方法的设计者都缺乏对所提出公式及其影响因素的用户实验验证。目前还不清楚哪种坐标轴标签和网格有利于感知,以及先前设计者的设计方法在感知层的原理。除此之外,人们忽视了背景网格的作用。最后,这些算法在计算折线图坐标轴标签的时候都是在x轴和y轴上分别进行优化计算,这违背了折线数据本身的二维性质,也没有考虑需要综合2个方向的网格密度,以及网格局部长宽比的问题。

1.1.2 背景网格

网格线绘制是图表设计过程中的重要一步,其能够在查找数值、比较数值和比较局部数据模式等方面辅助图形感知^[11]。Simkin等^[14]提出参考辅助结构(如背景网格)能够直接提高人们对数据的定位和推断这种初级感知过程,Kong等^[15]据此提出了一个自动化系统,为现有的图表添加像网格线这样的辅助补充图示。

Bartram等^[5]首次研究了网格线的透明度问题。他们通过实验确定了一个人们可以接受的网格线透明度的大致范围,并且在不同条件下重复实验探索这个范围;之后,又将研究继续扩展到网格线颜色的角度^[4]。

Heer等^[16]通过众包网站进行实验,首先重复并印证了Bartram等实验的结论,随后做了一个关

① <https://svn.r-project.org/R/trunk/src/appl/pretty.c>

于网格线间隔的实验,结果证明给图表添加网格线能够提高估计的准确度.但是他们也同时提出了另一个猜测:太密的网格线会有损人们的感知能力,因为过密的网格不方便人们阅读标签.

1.2 长宽比的选取

感知过程中图表长宽比的选取一直是一个热点问题^[3],至今已经有很多这方面的研究.

1.2.1 折线图

Cleveland 等^[17]是最早开始研究折线图的整体长宽比的.他们提出了一个“向 45°倾斜”的理论,通过用户实验证明了用户判断斜率的误差在 2 条线段的中间角对齐在 45°时最小.

然而, Talbot 等^[9]通过一个经验实验重复检查了 Cleveland 等的实验设计,发现斜率估计的误差并不总是在 45°时最小.这个结论暗示了在那些自动选择长宽比的系统中可能相对扁平的长宽比是一个更好的选择.

尽管有 Talbot 等^[9]的实验结果,目前大多数的长宽比自动选择算法仍然是基于向 45°倾斜理论的. Cleveland^[18-20]提出中位数绝对斜率、绝对平均斜率、绝对平均方向和弧长加权平均绝对方向等方法,都试图将线段的中间角集中在 45°. Heer 等^[21]提出一个全局方向分辨率和局部方向分辨率的方法,能够直接计算出“方向分辨率”.

上述方法都是研究图表的整体长宽比,本文研究的是网格中由网格线交叉形成的单个格子的局部长宽比.

1.2.2 可视化中的正方形

正方形经常被视作一个“单位”或组成其他更复杂图形的“基石”^[22].而细长的长方形往往会聚集在一起,成为数据分析任务过程中的阻碍因素.在可视化的实践中,近似正方形的形状也经常发挥着重要作用. Bruls 等^[23]最早应用了近正方形的矩形创造正方化的树图,矩形的长宽比被设计成尽可能靠近 1.作者认为正方形不仅更容易被检测和识别到,而且能够帮助人们比较面积大小.但是, Kong 等^[24]做了一系列实验,证明了严格的绝对正方形在做面积比较时是没有优势的,长宽比为 2:3 或 3:2 的矩形才是最理想的.他们认为不仅是那些极端细长的长宽比不妥当,绝对的正方形也应该避免在设计中出现.

2 坐标轴标签网格的感知实验

在二维坐标轴的图表中,最常见的就是散点

图和折线图.为了能够进行更精确、细致的研究,本文只将折线图作为研究对象,而斜率估计是人们经常需要通过折线图完成的任务^[7].如前文所述,影响人们对线段斜率估计的因素有很多,如折线图整体的长宽比.本文主要研究背景网格线对折线图斜率估计的影响.

2.1 影响因素分析及假设

本文研究人员事先与一些从事设计相关职业的专业人员和经常绘制折线图的普通人进行了讨论交流,根据他们的经验推测,以下 2 个因素可能会影响斜率估计任务.

(1) 密度. 网格的密度在散点图数值读取时能够发挥很大的作用.通常,越密的网格越有利于准确读取点的数值.但是过密的网格也可能导致“背景抢镜”的效果,弱化了数据的展示,甚至妨碍人们读取数据.因此,推测不同的网格密度对折线图的斜率估计任务也会有影响.

(2) 局部长宽比. 局部长宽比指的是网格线交叉形成的单个格子的长宽比.目前所有的坐标轴自动生成算法生成的网格线都有可能出现极其狭长的情况.被采访者均表示狭长网格让他们感觉不适,甚至会影响数据的读取.本文重点探究局部长宽比改变导致的网格的不同形状是否会对折线图的斜率估计任务造成影响.

针对密度和局部长宽比这 2 个影响因素,本文提出 3 个假设.

(1) 网格线的密度和网格局部长宽比对斜率估计的准确度和效率有影响.

(2) 网格线密度过大或过小都会干扰斜率估计,但是适度增大的密度有利于提高斜率估计的准确度,节省时间.

(3) 网格线局部长宽比越接近 1,斜率估计准确度越高,所花时间越短.

2.2 实验设计

在一幅折线图当中,数据通常会出现非常复杂的趋势变化.如果研究在一幅完整的折线图中进行某个单独线段的斜率估计,很有可能会参考周围的数据线条或受到周围线条的干扰,将无法专注于网格线影响的研究.因此,为了简化问题,实验中仅在背景网格上展示一条单独的线段.

本节所有实验中,用户需要完成的单个任务都是相同的.用户首先看到的是有背景网格和一条倾斜直线的原图,如图 1a 所示.参与者将观察这幅图 3 s,随后此图消失,出现无网格的空白界面,以及一条相同长度的水平线段,用户需要凭借

图 1 用户实验的界面

自己脑海中的印象用鼠标点击这条线的右端将该直线旋转至原图中直线的倾斜角度, 如图 1b 所示。

本文将任务设计为直接旋转直线而非如 Talbot 等^[7]和 Cleveland 等^[17]研究中要求用户估计斜率比率的原因是, 研究核心是网格局部长宽比对固定倾斜角度的直线的感知影响, 而非如前人工作那样调整整体长宽比以寻求最佳直线倾斜角度。而且, 因为“斜率”代表的是直线的正切而非倾斜角, 斜率与倾斜角之间呈非线性关系。对用户来讲, 准确估计正切值是非常困难的一件事情。因此, 选取了更为直观的旋转直线的任务。

为了防止用户通过记忆倾斜直线在竖直方向上的投影顶点完成实验, 本文对于用户需要旋转的这条线段进行了沿 x 轴方向的平移, 使其离开原图中的初始位置, 如图 1b 中线段并非出现在图 1a 相同位置上。位移量是随机数产生的, 避免了用户在多次反复训练后记住位移大小。同时, 根据 Talbot 的实验结果^[22], 做任何 y 轴方向上的移动都将导致用户在旋转时所利用的基准线与原图中的基准线出现了偏差, 从而造成误差, 故不进行 y 轴方向上的平移。

参与者观察的原图大小为 600×600 像素, 图中直线长度为 200 像素, 倾斜角度为 $0^\circ \sim 90^\circ$ 。根据实验目的的不同, 背景网格设置为不同的长宽比和密度, 但都遵循 Bartram 等^[5]在其研究中提到的较理想的 0.2 透明度。

本节所有实验的参与人员均在相同的四核处理器台式机上完成实验, 使用鼠标和键盘作为输入。配有 58.6 cm 的宽屏液晶显示器作为输出显示设备, 分辨率为 1920 像素 $\times$ 1080 像素, 屏幕经过了校准。参与实验的人员坐在距离屏幕前 60 cm 处, 在一个安静且有足够照明的房间内完成实验。鉴于实验需要熟练使用计算机的输入设备进行操作, 从当地大学的计算机专业招募了志愿者进行实验, 年龄为 20~29 岁; 每组实验 20 人, 共 3 组志愿者, 各组人员无重复。所有志愿者确认视觉正常, 没有色盲色弱, 能够熟练使用鼠标、键盘等输入设备。

本节所有实验的参与者都遵循相同的实验流程。

- (1) 阅读实验说明, 观看实验操作教学视频。
- (2) 首先完成若干训练任务, 熟悉操作。
- (3) 完成实验的所有任务, 最后提交。

(4) 实验人员对该参与者进行一个简短的采访, 主要是为了了解参与者采用的策略和实验过程中的感受。

实验结果测量的指标主要是准确率, 即旋转直线的角度与观察的直线倾斜角度的误差。一个实验完成的时间为 10~15 min。涉及底层认知的实验本身是非常易受干扰的, 而且误差差别微小而不易捕获。在对实验进行测试时发现, 当用户受到干扰或走神时, 他们往往会错过观察图片的 3 s, 进而在旋转时产生非常大的误差, 远大于其他正常观察并且操作的图片。为了排除这样的异常值, 本文选用了绘制盒须图(boxplot)时使用的计算异常点的方法: 假设当前误差值为 x , Q_3 为所有误差值的上四分位数, Q_1 为所有误差值的下四分位数, 如果 $x > Q_3 + 1.5 \times (Q_3 - Q_1)$, 则 x 被认为是异常值, 需要排除。对剩下的数据计算绝对误差的平均值并绘制盒须图和 95% 置信区间的统计图。

2.3 预备实验

本节预先排除了线段与网格线交叉这一因素的影响。本节所有实验的流程和前文提到的实验设计一致。

2.3.1 实验因素分析

在邀请用户进行实验之前, 实验人员为了调试实验程序以及生成实验数据, 在实验室内进行了小范围的实验测试, 发现直线和网格线的交叉位置不同似乎会导致完成任务的难度不同。

经观察, 这个猜测是合理的, 因为直线和网格线的交叉位置决定了网格线是否能够充当一个合格的“基准线”。例如, 当直线下方的端点恰好落在

一条水平网格线上时,这条网格线对于估计直线斜率是一条非常有利的基准线,提供了一个清晰的与水平方向的夹角.根据 Talbot 等^[9]的实验结果,这条基准线对于斜率估计的影响较大.如果直线下方的端点没有落到任何一条水平网格线上,而是落在了网格中间的空白处,人们估计斜率的难度看起来会上升.

在实验人员的调试和局部小范围实验室实验之后,在预备实验前做出如下假设:(1) 网格线与直线交叉的位置会影响斜率估计的效率;(2) 当直线的下方端点恰好落在一条水平网格线上时,最有利于提高斜率估计的准确度,节省时间.

对于网格线与直线交叉位置的量化描述,参考之前 Talbot 等^[9]的实验及结果,主要考虑直线与水平方向上的网格线的交叉位置,只考虑与直线交叉位置最低的一条网格线,本文认为这条线充当了基准线的角色.本文定义了一个变量—— x 方向上的投影比例,即直线在最低交叉网格线上的投影长度与该直线在 x 轴方向上的投影长度的比例,如图 2 所示.在生成的数据中均匀采样,控制 4 种不同的投影比例: 1.00, 0.75, 0.50, 0.25, 全部采用正方形的局部方格形状,正方形的边长度相同(即网格线密度相同),全为 70 像素.每种投影比例产生 6 个样本,这 6 个样本具有相同的一组倾斜角度.

图 2 4 种线段与网格线交叉位置的说明

在初步测试中发现,当倾斜角度非常大时,直线往往和网格线有多次交叉,因此也就拥有了多

根基准线,进行估计时的难度不大.主要阻碍人们完成估计任务的是那些倾斜角度较小、仅和一条网格线有交叉的直线.因此,生成数据图片时所有直线的倾斜角度均为 $0^\circ \sim 45^\circ$.

训练部分包含 4 张图片,每张图片各对应了一种投影比例数据,从 1.00 到 0.25 对应的倾斜角分别是 13° , 18° , 26° , 31° , 38° .数据图片随机出现,每个图片重复 3 次,因此在这个实验中用户需要完成 $5(\text{角度}) \times 4(x \text{ 方向上的投影比例}) \times 3(\text{重复}) = 60$ 次观察和旋转操作.

2.3.2 实验结果

关于线段与网格线交叉位置的实验结果如图 3 所示.图 3a 为盒须图,图 3b 为对应的 95% 置信区间估计.可知当 x 轴方向上的投影比例为 1.00 时平均误差是最小的,随着投影比例的减小,平均误差逐渐增大.因此,结论是网格线与直线的交叉位置的确会影响斜率估计的准确度.

图 3 关于线段与网格线交叉位置的实验的结果

为了研究密度和局部长宽比时控制住网格线与直线交叉位置造成的误差,在正式实验中采取了所有数据图片统一同一种交叉方式:直线的下端

点落在网格线上,即 x 轴方向上的投影比例为1.00.

2.4 实验1

本节设计并完成了关于网格线密度影响的实验.

2.4.1 实验因素分析

实验假设如下:(1) 网格线的密度过大或过小都会影响斜率估计的效率;(2) 适度增大的密度有利于提高斜率估计的准确度,节省时间.

为了与前人工作进行对比,使用通用的网格密度定义方式:将局部方格面积相同的网格线看做密度相同.

由于底层感知实验相对来说任务比较枯燥,对于志愿者,实验时间过长,他们的注意力会因疲劳而下降,影响实验结果.为了保证实验质量也控制实验复杂度,尽量减轻用户的实验负担,对于实验中的变量尽可能均匀地取一定数量的样本.生成的数据包括4种不同的密度(即控制4种不同的局部长方形面积):极密、中等、较疏和极疏,全部采用正方形的局部方格形状,正方形的边长分别为40像素、70像素、100像素和130像素.参考Talbot等^[9]的研究,为每种密度产生了6个样本,这6个样本具有相同的一组倾斜角度:17°, 25°, 38°, 40°, 62°, 70°.所有直线的较低端点均落在一条水平网格线上.用户训练部分使用了4个数据,每种密度各有一张图片.所有数据图片随机出现,每张图片重复3次,因此在本实验中用户需要完成6(角度) $\times$ 4(密度) $\times$ 3(重复)=72次观察和旋转操作.

2.4.2 实验结果

关于密度的实验结果如图4所示.当网格线密度极小的时候,误差最大,均值在4.1°左右;当密度为中等的时候误差最小,密度较大和中等时误差非常接近.总体来看,这4种情况的误差比较接近,但是在网格线极其稀疏的情况下误差会略微上升.这是合理的,因为稀疏的网格线提供的基准线更少.

2.5 实验2

本节研究的是网格的局部长宽比对于折线图的斜率估计这一任务的影响.

2.5.1 实验因素分析

根据前面实验的结果,生成数据时控制了直线交叉位置与网格线密度这2个变量.所有图片的网格线密度均控制为局部面积4900像素.所有的直线的较低端点均落在一条水平网格线上.

参考Heer等^[16]的研究,本文设计了5种不同的局部长宽比:5:2, 3:2, 1:1, 2:3和2:5.每种局部长宽比与实验1采取相同的一组倾斜角度:17°,

a. 盒须图

b. 95%置信区间估计

图4 关于密度的实验的结果

36°, 49°, 62°, 73°.所有数据图片随机出现,每张图片重复3次,因此在本实验中用户需要完成5(角度) $\times$ 5(局部长宽比) $\times$ 3(重复)=75次观察和旋转操作.

2.5.2 实验结果

由图5可知,当网格形状为正方形,即局部长宽比1:1时,斜率估计的误差最小.这与之前的假设相一致.

在实验1和实验2中,不同条件下用户花费的时间几乎是相同的.经观察,由于实验操作比较简单,衔接节奏较快,为了避免遗忘,用户往往操作迅速,每次旋转完成的时间基本都在1~2s.因此时间上几乎没有差别.

3 实验分析与讨论

根据以上实验结果,可以得到结论:(1) 网格线与直线交叉的位置影响较大,当直线的一个端点落在网格线上的时候最有利于减小斜率估计的误差;(2) 除了极其稀疏的情况下,不同密度对斜率估计的影响不大;(3) 当局部长宽比接近正方形的时候,斜率估计较为准确,将这一结论命名为

图 5 关于网格局部长宽比的实验结果

局部 45° 原则; (4) 网格线的形状对于斜率估计花费时间的影响不大. 接下来将分析误差产生的原因以及描述用户使用的策略.

3.1 误差产生的原因

研究者试图拆解实验过程中人们试图记忆以及重现直线的倾斜角度的过程, 推测出误差出现的原因. 人们观察屏幕上的直线, 在脑海中形成印象记忆, 然后根据印象记忆, 利用鼠标旋转重现脑海中倾斜的直线.

本文推测可能出现误差的地方如下.

(1) 观察并记忆的时候, 人们脑海中的印象与实际真实的被观察的物体存在偏差, 称为“感知误差”.

(2) 脑海中的印象随着时间的推移与初始状态产生偏差, 称为“记忆误差”.

(3) 人们利用鼠标旋转重现脑海中的倾斜直线时, 实际操作产生的结果和脑海中的印象不同, 称为“操作误差”.

本文希望能够检测到的误差是感知误差, 因此希望尽量减小记忆误差和操作误差. 减小记忆

误差的方式非常简单, 就是尽量缩短用户从观察到操作等待的时间. 在实验中, 用户观察 3s 之后立即进行旋转操作, 记忆误差已经被尽量减小. 但是操作误差就非常难以控制, 是受到操作者自身熟练程度的影响的.

本次的用户实验中, 对于操作误差, 通过限制实验参与人员的专业进行控制. 但是这并不能完全或基本消除操作误差的影响. 如果用户仅仅快速旋转一次, 依然很容易造成操作误差. 通过多次反复调整可以减小操作误差, 但是同时随着时间的流逝, 记忆误差也会增大, 二者存在矛盾. 除非用户用来观察的图片不会消失, 一直可以供用户查看, 边调整边观察可能是记忆误差和操作误差相对较小时. 但是这种情况下该如何展示是此项工作尚未做好的. 用户在同时展示观察界面和操作界面的情况下更容易作弊, 而且这种情况下误差可能会更小, 是否会小到难以检测或难以说明问题尚且存疑.

本文采用实验室实验而非使用众包平台进行大规模用户实验的原因是这种涉及底层感知的实验对于实验环境依赖程度较高, 用户必须精力非常集中, 极容易走神或受到干扰这种问题在众包平台实验中尤为严重. 未来如果有研究者能有更好的实验方案解决这一问题, 那么在众包平台上重复这个实验获得更多的数据是一个值得完成的任务.

3.2 用户策略

每个用户完成实验后, 实验人员都会对他进行简短的采访, 询问的主要问题是“请问你在完成任务的过程中使用了怎样的策略来记忆、重现直线的倾斜程度?”“你认为背景网格对你的估计是否有影响?”

研究者预先对用户可能使用的技巧有一些推测, 但是有趣的是, 实际上用户仍然使用了一些未预见到的策略.

(1) 记忆直线的走势, 试图旋转出所观察直线的平行线.

(2) 估计并记忆 x 轴方向投影和 y 轴方向投影的长度比例(和直线本身构成了一个大三角形), 正方形网格时方便数格子计算长度.

(3) 对于角度比较大的角试图记忆它的余角.

(4) 观察直线和网格线交叉形成的最小的三角形, 试图记忆 2 条直角边的比例.

(5) 只观察 x 轴方向投影或者只观察 y 轴方向投影.

(6) 观察时与 30°, 60°, 45°等常用角度比较, 旋转的时候在常用角度的基础上下浮动调整。

(7) 利用正方形网格自身隐含的 45°进行校准。

值得注意的是, 有的用户提到他们不只使用了一个策略, 而是综合上述多个策略进行估计。

绝大多数用户都表示网格的存在对他们进行斜率估计有帮助, 有个别用户甚至明确提到感觉狭长的网格会干扰他们对斜率的估计。用户所采用的策略进一步印证了局部 45°原则。总结一下这一原则起作用的原因: 首先, 正方形包含一条 45°的隐含的角度基准线, 人们可以轻易地判断观察的倾斜角度与 45°的差距; 其次, 正方形网格在度量 x 轴方向和 y 轴方向上的长度时充当了一个单位长度的作用。

本文猜测, 狭长网格可能造成一个心理学上的错觉——“蓬佐错觉”(Ponzo illusion)^[25], 也叫作铁道错觉, 核心思想是人类的大脑根据物体所处的环境判断它的大小。狭长的网格线可能会使人们对直线斜率的估计产生扭曲。本文并未对此猜想进行验证, 有待后续研究。

4 自动生成标签算法

本节将描述利用前面实验得到的结论设计的一个自动生成标签算法。算法的核心思想是根据局部 45°原则, 控制网格的局部长宽比尽量接近正方形进行 x 轴和 y 轴的坐标分配的。后文公式中提到的所有变量及其含义如表 1 所示。

表 1 变量名及其描述

变量名	描述
Q	按照用户喜好排序的“好的标签间隔”集合
q	Q 中的元素
j	跳步大小
i	Q 中 q 的序号
$l_{\min}, l_{\max}$	最小标签和最大标签
$d_{\min}, d_{\max}$	数据最小值和数据最大值
v	如果生成的标签序列包含 0 标签, 则 $v=1$; 否则 $v=0$
k	一个标签方案中生成的标签个数

标签的计算关键在于把握 3 个量: 最大标签、最小标签以及标签间隔步长。这里参考 Talbot 等^[7]在 Wilkinson 方案的扩展算法中提出的“灵活标签”, 允许最大和最小标签在数据范围以内或之外; 但是后续采用了完全不同的、更为简洁的搜索方法。

经过初步分析可知, 最大标签和最小标签分别有 2 种可能的情况。最大标签可能小于数据最大值, 或大于或等于数据的最大值。最小标签情况类似。如图 6 所示, 在 q 控制确定的情况下, 最小和最大标签组合 $(l_{\min}, l_{\max})$ 有 4 种情况: (t_1, t_3) , (t_1, t_4) , (t_2, t_3) , (t_2, t_4) 。其中,

$$\begin{aligned} t_1 &= \left\lfloor \frac{d_{\min}}{q} \right\rfloor \times q, \\ t_2 &= \left(\left\lfloor \frac{d_{\min}}{q} \right\rfloor + 1 \right) \times q, \\ t_3 &= \left\lfloor \frac{d_{\max}}{q} \right\rfloor \times q, \\ t_4 &= \left(\left\lfloor \frac{d_{\max}}{q} \right\rfloor + 1 \right) \times q. \end{aligned}$$

图 6 确定最大和最小标签的示意图

进而可以计算出跳步大小

$$j = \frac{l_{\max} - l_{\min}}{(k-1) \times q}.$$

由此可以计算出标签间隔的值 c 为 q 和 j 的乘积。注意, 为了将数据空间的距离映射到坐标轴上, 需要分别计算 x 轴和 y 轴上标签之间在图标上的实际距离, 对 c 进行缩放即可。对于特别大或特别小的步长, 如 $[0, 1]$ 或 $[100, +\infty)$ 的数据, 需要扩展 Q 数列到 Q' , 即

$$Q'_i = Q_i \times 10^{\lambda}, \quad \lambda = \lfloor \lg(d_{\max} - d_{\min}) \rfloor - 1.$$

根据之前的实验结果, 在优化坐标的过程中添加了一个表示长宽比的变量

$$R = - \left| \frac{c_x}{c_y} - 1 \right| \quad (1)$$

用来使生成的网格尽可能接近正方形。

这是一个同时针对 x 轴和 y 轴计算的联合优化变量, 如图 7 所示。

本文保留了 Talbot 等^[9]方法中的简洁性、覆盖性和可读性中的重叠程度。所有的公式都要分别对 x 轴和 y 轴进行计算。由此结合式(1)可以计算总分

$$s = S_x + S_y + C_x + C_y + O_x + O_y + \alpha \cdot R \quad (2)$$

其中, S_x 和 S_y 分别代表 x 轴和 y 轴上的简洁性项; C_x 和 C_y 同理代表覆盖性项; O_x 和 O_y 代表可读性中的重叠程度. 可以根据正方形化程度的需求修改 R 对应的参数 α 的大小. 计算公式见文献[7].

图 7 长宽比 R 的计算方法示意图

Talbot 等^[9]的计算方法包含 4 重循环, 其中和数据相关的循环共 2 层, 因此复杂度为 $O(n^2)$. 本文设计了更为简洁有效的求解方式. 已知输入是折线数据和图表绘制区域的长与宽. 对于 x 轴的数据最大值和最小值, 在已知 q 和 k 的情况下可计算 $(t_1, t_3), (t_1, t_4), (t_2, t_3), (t_2, t_4)$ 对应的 4 种简洁性项和覆盖性项; 同理计算 y 轴上的简洁性项和覆盖性项. x 轴和 y 轴共有 16 种标签的组合方式, 对于这 16 种方式分别计算对应的 R 值和可读性中的重叠程度, 并根据式(2)求最终的 s 得分. 遍历 Q 中元素, 对于其中每个 q , k 尝试取 $[2, +\infty)$ 的每个整数值, 计算上述 s , 取 s 最大时作为最终输出的标签序列. 注意 k 实际上并不会无限增大. 当 k 增大到 $(d_{\max} - d_{\min}) / (k - 1)$ 求得的步长小于 q , k 达到了当前数值范围和 q 条件下的最大值, 即当前 k 的搜索循环结束. 本文方法提前考虑标签的起始位置, 分 4 种情况讨论. 第 1 层循环中的跳步大小 j 可以通过起始位置以及 q 和 k 求得, 不必循环遍历, 将复杂度降至 $O(n)$. 同时, 除了对得分较低的情况的剪枝, 通过提前终止第 2 层 k 的循环进一步提高了计算的效率.

图 8 展示了一个使用本文自动生成算法生成的正方化坐标轴及网格的结果例子. 数据为 1980—2015 年每月大气中二氧化碳的浓度^①, 数据范围为 337~402 ppm. 与 Talbot 等^[9]的方法生成结果做了对比, 本文算法生成的横坐标间隔为 4, 最小标签为 1980, 最大标签为 2016; 纵坐标间隔为

10, 最小标签为 340, 最大标签为 400. y 轴的标签表现了灵活标签原则, 即数据范围超过实际标签范围. Talbot 等^[9]方法的横坐标间隔为 2, 纵坐标间隔为 10, 最小标签为 330, 最大标签为 410. 可以看到网格线相对狭长而密集, 形成了一定的干扰.

图 8 生成效果展示: 1980—2015 年
每月大气 CO_2 浓度

5 结 语

本文在前人对于网格线透明度的研究基础之上, 研究了折线图中网格线密度和局部长宽比对人们感知直线斜率的影响, 通过一系列实验室实验首先排除了网格线与直线交叉位置的影响, 进而证明了适当密度且局部长宽比接近正方形的网格有利于减少斜率估计的误差. 通过对前人发明的坐标轴标签生成方法的分析和探讨, 结合前面实验的结论, 开发了新的坐标轴标签的生成算法, 将局部长宽比这个前人没有考虑过的因素加入 2

① <https://cran.r-project.org/web/packages/locfit/locfit.pdf>

条坐标轴的标签联合优化之中, 该算法能够生成既美观又提高人们认知效率的带有网格背景辅助的折线图. 本文从图表辅助结构的角度研究如何提高折线图认知效率, 对算法的创新和优化都是经过用户实验验证的. 本文集中研究网格自身特性对斜率感知的影响, 出于实验复杂度的考虑, 将直线与网格线交叉这一因素简化为预备实验得出的一个控制条件, 针对各种交叉情况的具体影响的实验有待未来研究完成. 此外, 散点图是否受网格线的影响也是值得研究的问题. 特别是关于如何提高散点图的数值读取这一常见任务的效率, 非常值得探讨. 另外, 本文将折线图上的线段抽象出来进行了研究, 而实际上折线图本身的数据可能有独特的模式特征, 图表的展示应该考虑如何更好地表现数据特征. 这一点在散点图中更为明显, 如散点图点的聚类或趋势特征. 目前尚无研究将数据特征考虑到标签生成中.

参考文献(References):

- [1] Bertin J. Semiology of graphics; diagrams networks maps[R]. Madison: Wis University of Wisconsin, 1983
- [2] Cleveland W S, McGill R. Graphical perception and graphical methods for analyzing scientific data[J]. Science, 1985, 229(4716): 828-833
- [3] Tufte E R, Graves-Morris P R. The visual display of quantitative information[M]. Cheshire: Graphics Press, 1983
- [4] Bartram L, Cheung B, Stone M. The effect of colour and transparency on the perception of overlaid grids[J]. IEEE Transactions on Visualization and Computer Graphics, 2011, 17(12): 1942-1948
- [5] Bartram L, Stone M C. Whisper, don't scream: grids and transparency[J]. IEEE Transactions on Visualization and Computer Graphics, 2011, 17(10): 1444-1458
- [6] Heckbert P S. Nice numbers for graph labels[M] //Glassner A S. Graphics Gems. Cambridge: Academic Press Inc., 1990: 61-63
- [7] Talbot J, Lin S, Hanrahan P. An extension of Wilkinson's algorithm for positioning tick labels on axes[J]. IEEE Transactions on visualization and computer graphics, 2010, 16(6): 1036-1043
- [8] Wilkinson L. The grammar of graphics[M] //Handbook of Computational Statistics. Heidelberg: Springer, 2012: 375-414
- [9] Talbot J, Gerth J, Hanrahan P. An empirical model of slope ratio comparisons[J]. IEEE Transactions on Visualization and Computer Graphics, 2012, 18(12): 2613-2620
- [10] Cleveland W S, McGill R. Graphical perception: theory, experimentation, and application to the development of graphical methods[J]. Journal of the American Statistical Association, 1984, 79(387): 531-554
- [11] Few S. Show me the numbers: designing tables and graphs to enlighten[M]. El Dorado Hills: Analytics Press, 2004
- [12] Harris R L. Information graphics: a comprehensive illustrated reference[M]. England: Oxford University Press, 2000
- [13] R Core Team. R: a language and environment for statistical computing[M]. Vienna: R Foundation for Statistical Computing, 2018
- [14] Simkin D, Hastie R. An information-processing analysis of graph perception[J]. Journal of the American Statistical Association, 1987, 82(398): 454-465
- [15] Kong N, Agrawala M. Graphical overlays: using layered elements to aid chart reading[J]. IEEE Transactions on Visualization and Computer Graphics, 2012, 18(12): 2631-2638
- [16] Heer J, Bostock M. Crowdsourcing graphical perception: using mechanical turk to assess visualization design[C] //Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2010: 203-212
- [17] Cleveland W S, McGill R. Graphical perception: the visual decoding of quantitative information on graphical displays of data[J]. Journal of the Royal Statistical Society: Series A (General), 1987, 150(3): 192-210
- [18] Cleveland W S. A model for studying display methods of statistical graphics[J]. Journal of Computational and Graphical Statistics, 1993, 2(4): 323-343
- [19] Cleveland W S. Visualizing data[M]. Thousand Oaks: Hobart Press, 1993
- [20] Cleveland W S. The elements of graphing data[M]. Massachusetts: Wadsworth Publishing Co., 1985
- [21] Heer J, Agrawala M. Multi-scale banking to 45 degrees[J]. IEEE Transactions on Visualization and Computer Graphics, 2006, 12(5): 701-708
- [22] Pinna B. What is the meaning of shape?[J]. Germany: Gestalt Theory, 2011, 33(3/4): 383-422
- [23] Bruls M, Huizing K, van Wijk J J. Squarified treemaps[M] //Data Visualization 2000. Heidelberg: Springer, 2000: 33-42
- [24] Kong N, Heer J, Agrawala M. Perceptual guidelines for creating rectangular treemaps[J]. IEEE Transactions on Visualization and Computer Graphics, 2010, 16(6): 990-998
- [25] Brislin R W. The Ponzo illusion: additional cues, age, orientation, and culture[J]. Journal of Cross-Cultural Psychology, 1974, 5(2): 139-161