



Journal of Statistical Software

May 2021, Volume 98, Book Review 1.

doi: 10.18637/jss.v098.b01

Reviewer: Ulrike Grömping
Beuth University of Applied Sciences Berlin

Data Visualization: Charts, Maps, and Interactive Graphics

Robert Grant

Chapman & Hall/CRC, Boca Raton, 2019.

ISBN 978-1-138-70760-3. 243 pp. USD 47.96 (P), USD 140.00 (H), USD 53.96 (e).

<http://robertgrantstats.co.uk/dataviz-book.html>

Data visualization – or dataviz – goes back a long way, with many sources quoting historical visualizations like Charles Joseph Minard’s map of troop flows in Napoleon’s war against Russia, Florence Nightingale’s “Coxcomb diagram” on the causes of death in the Crimean war, or John Snow’s map of London street pumps that ties Cholera case locations to one contaminated pump. John Snow’s data have also made it into Robert Grant’s little book.

Robert Grant approaches dataviz in a fundamental and non-technical way. He promises **no algebra** (bold face in the original), and the book also refrains from covering any software or technical matters. The preface explains, what sets this book apart from many other books in the market: “provide a brief overview of techniques and tools, while all the time emphasizing statistical reasoning”. With this approach, Robert Grant targets a quite diverse audience: managers of data teams, statisticians without prior education on dataviz, designers and journalists who want to visualize in a statistically sound way, or consumers of visualizations who want to make the most of what they see and want to avoid being confused or misled. As a statistician who has already spent some energy and thought on professional visualizations, I am not part of the immediate target audience, but I nevertheless like the book.

The preface explains the three parts of the book: Part 1 (“Basics”) is for everyone, Part 2 (“Statistical building blocks”) can be skipped by statistically educated readers, and Part 3 (“Specific tasks”) contains “the real fun” with a series of short chapters that can be read during a commute – remember the time before the pandemic, when we used to commute? Part 1 starts with a chapter that makes the case for visualizations and emphasizes that a visualization should be planned with a target message and recipient in mind, ideally starting from the idea of what should be shown rather than from the possibilities of some software. Always using an example, the chapter explains even very basic aspects of a scatter plot, such as the markers, the axes, the legend, and how to read the information from a point in a scatter plot. The datasaurus by Alberto Cairo and the famous Anscombe quartet are used for driving home the message that visualization adds valuable information over and above descriptions in terms of numbers. Chapter 2 discusses visual encoding of quantitative information, again

heavily relying on a real life example: a time series of percentages of train cancellations and delays in consecutive four week intervals from 1997 to 2017 is visualized in several ways, demonstrating scatter plots, line charts, overlaid annual scatter plots and line charts with coloring for calendar year, a heat map like plot that highlights the worst three periods of each year etc. This introductory example already comes loaded with implicit advice, before Robert Grant demonstrates the visual effects of classical encodings such as position, length, area, volume, angle, color hue, color saturation, marker shape, marker features, line width by small more technical examples. The book also demonstrates more unconventional aspects of visualizing numbers, such as showing 200 people in cars, without the cars, on bikes, on buses or on a light rail train – clearly a visualization with a message. Chapters 1 and 2 hold some good and very general advice, like “Be prepared to sketch and discard”; Robert Grant means this literally, sketching on paper, but it is of course equally true for sketches created with software.

Part 2 (“Statistical building blocks”) consists of four chapters. Besides providing plot types for different types of data (continuous and discrete numbers, percentages and risks, raw data and statistics, differences or ratios, correlations), Part 2 emphasizes aspects like independent (unmatched) or matched groups of data, nesting of categories, robustness (quantiles, winsorizing, smoothing, medians with median absolute distances from medians), showing reference information, using adequate denominators (like person years at risk) or showing the information of interest normalized by a baseline value. Robert Grant also criticizes the use of odds and warns against cherry-picking or omitting important information.

After this statistical part, Part 3 provides ten brief chapters (7 to 14 pages). The initial of these, “Visual perception and the brain”, provides general insights that would – in my opinion – be better placed in Part 1. The other chapters in Part 3 do indeed cover specific aspects: “Showing uncertainty”, “Time trends”, “Statistical predictive models”, “Machine learning techniques”, “Many variables”, “Maps and networks”, “Interactivity”, “Big data”, and “Visualization as part of a bigger package”. The book closes with Part 4 (“Closing remarks”) that provides pointers to further reading for all chapters. References include influential statistically-oriented dataviz classics (like the Cleveland books from the eighties), as well as influential more current dataviz or machine learning sources like works by Gelman and Unwin, Tufte, Hans Rosling or Nate Silver, to name a few; these often come with `goo.gl` short links. There is no overall list of references; this is an omission I do not like. There is an index, however.

Although the book itself omits technical detail, interested readers find data and code (often R, sometimes Stata, occasionally other tools) for various of the book’s visualizations on the book’s website. This enables readers to play with visualization choices, especially as examples are necessarily brief. The more exciting figures often do not come with code, so that it takes a lot of effort to reproduce them. One nevertheless gets inspired by reflecting on them.

With book reviews, I often pick a few aspects for having a detailed look. Here, I focus on the book’s coverage of time series visualization, which starts in Chapter 2 with line chart representations of the afore-mentioned train delays data (Figures 2.2 and 2.3). The book presents a plot without structuring elements (reproduced as plot (a) in Figure 1) and two versions with structuring elements that Robert Grant criticizes as being too overwhelming (reproduced as plots (d) and (e) in Figure 1). Figure 1 also shows three related attempts of mine (created with the support of R package `preplot`, which was inspired by [Rahlf 2017](#)). Plot (a) would rarely convince me, because time series with potential seasonality should always

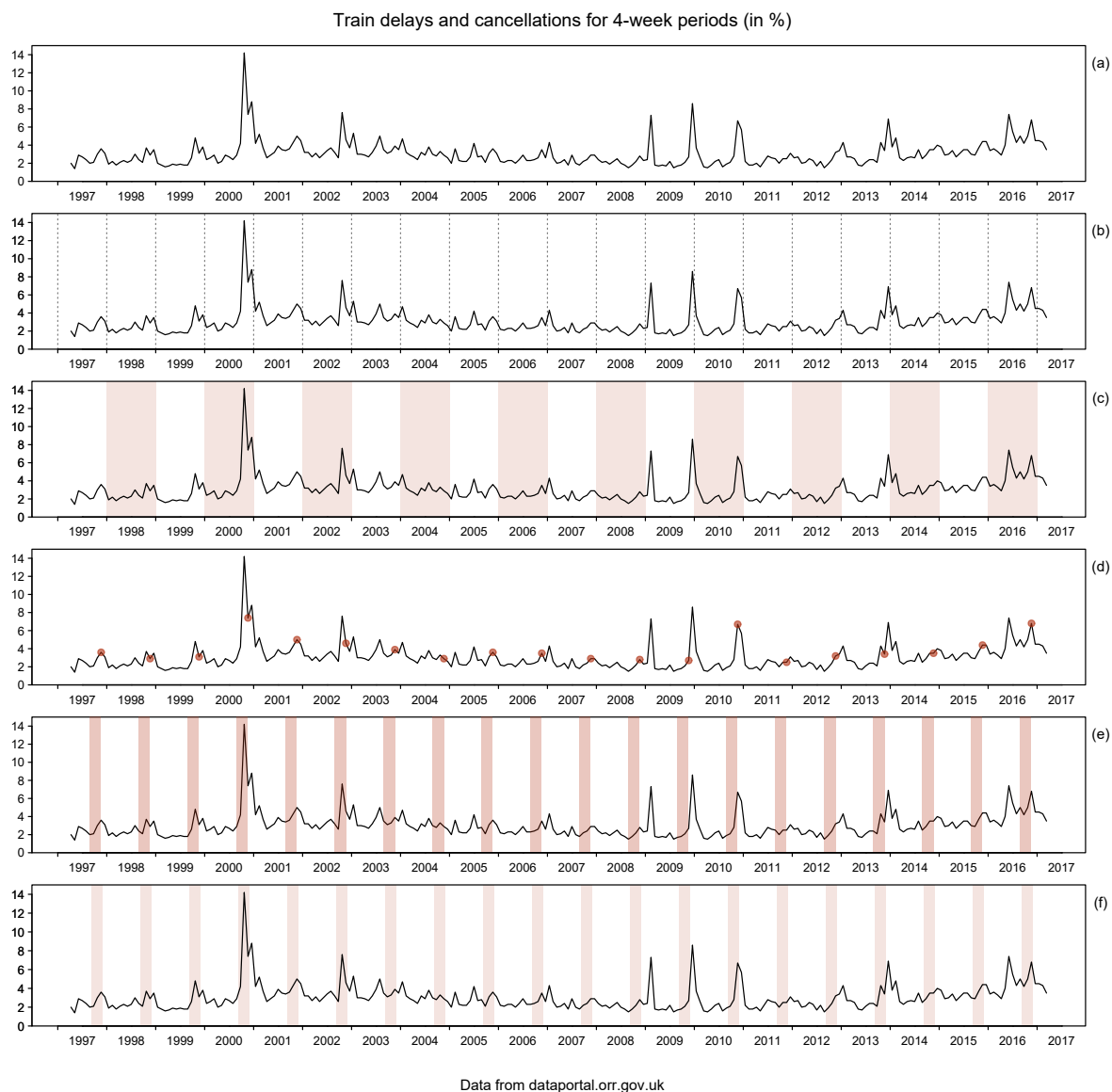


Figure 1: The question behind the train delays data: Are autumn leaves on rails an important contributor to train delays? The data are percentages of delayed or cancelled trains during 4-week periods from 1997 to 2017. Plot (a) has no structuring elements, plots (b) and (c) provide indications of year boundaries. Plots (d) and (e) mark specific 4-week periods related to falling autumn leaves (colored dot for mid November to mid December, highlighted period between mid September and mid December). Figure (f) provides more unobtrusive highlighting of the same period.

visualize periodicity in some way, e.g., like in plots (b) or (c) of Figure 1. Autumn stripes or autumn markers are helpful for an exploratory assessment whether the data support the idea that autumn leaves cause delays. The comparison between Grant's and my version of autumn stripes shows that the devil is in the detail: I agree with Robert Grant that the stripes in plot (e) of Figure 1 are a bit overwhelming; the stripes in plot (f) are more unobtrusive by

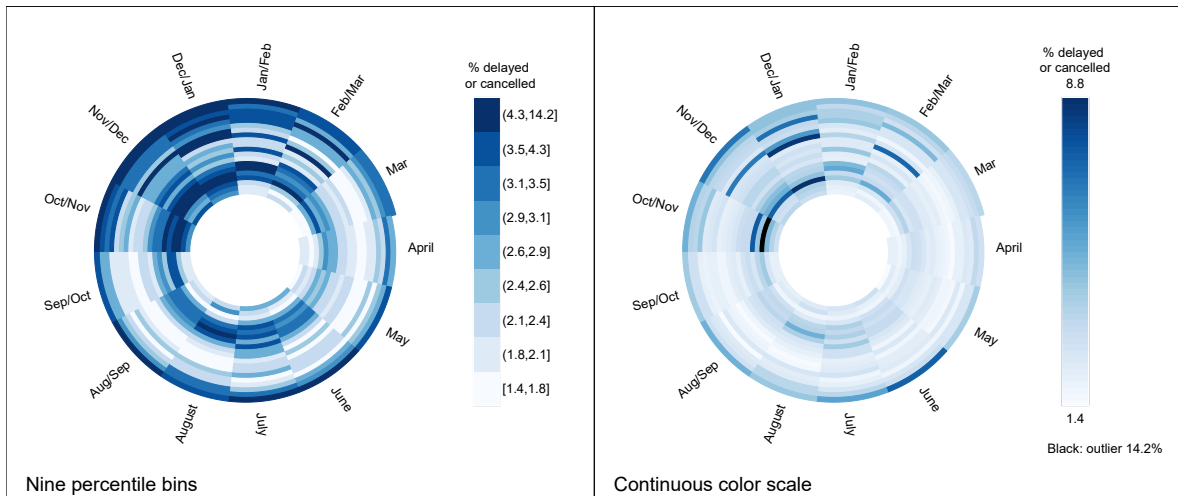
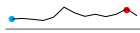



Figure 2: Two spiral heatmaps of the train delays data: time moves from the inside outwards in clockwise direction (1997 to 2017). Colors have been allocated by percentiles (left-hand side) or to 50 equally-spaced intervals (right-hand side; the severe outlier 14.2 was excluded from binning and is shown in black). The visual impression strongly depends on color choices, but most insights remain comparable.

design. In Figure 1, the stripes of plot (e) are not unacceptable, and on an economy mode printout they look quite OK; this may be due to the fact that Figure 1 shows the series in a more elongated format than the book does. Grant’s code for plots (a), (d) and (e) does not entirely work off the shelf, a few fixes here and there were necessary: most noteworthy, I had to omit `as.character` from a call to `as.factor`, because the ordering of 4-week periods was otherwise messed up on my machine. Presumably, another glitch in the code produced three autumn markers in Figure 2.3 of the book instead of the intended single marker as shown in plot (d) of Figure 1 (judging Robert Grant’s comments for the figure).

Figure 9.4 of the book, Andrew Elliott’s spiral chart of baby sleep times (Elliott 2016), inspired me to try out a spiral heatmap on the train delays data. The result (see Figure 2) does not look as impressive as Elliott’s, but gives an informative picture of the time pattern of delay and cancellation percentages, which is also in line with conclusions drawn in Chapter 16 of the book. I appreciate spiral heatmaps as an addition to my time series toolbox, because the time series is not interrupted at year boundaries, while one can nevertheless compare the seasonality behavior between years very well. (As neither the book nor Andrew Elliott provide code, I wrote the function `spiral.heatmap`, which is at the moment available in the single purpose non-CRAN R package `timespiral` from my website, but will be made available in Jim Lemon’s R package `plotrix`.)

Let us now have a look at selected chapters from Part 3 of the book. I particularly liked Chapter 9 (“Time trends”), which offers examples of various approaches for visualizing time related data, among them the afore-mentioned spiral heatmap: paths in a ternary plot show the development of compositions of three ingredients over time (Labour, Conservative and other in British elections), paths or average arrows of change direction in a Cartesian coordinate system show changes over time in a diagram that does not have a time axis, or the developments over time of several features is shown in a dashboard-like way without vertical

axes over a time scale that is annotated by a time line of events. The chapter also exemplifies Tufte’s concept of sparklines, i.e., little lines that show a development within a paragraph of text, e.g., the percentages of delayed or cancelled trains in the 13 four-week periods of 2016 (i.e., one year from Figure 1), shown as a sparkline line chart  and a corresponding sparkline bar chart ; here a cyan dot marks the first four-week period (coldest, from mid January) and a red dot marks the four week period from mid November to mid December (autumn). Contrary to the book’s demo of a sparkline, the zero level of the vertical scale is visualized by a bottom line, which I perceive as useful for this particular application. (The above two sparkline examples were created with L^AT_EX package **sparklines**.)

I was not so impressed by Chapter 11 (“Machine learning”), which is a 12-pager on a complex topic: the chapter touches upon ensemble methods, bagging and boosting, devoting three pages to classification and regression trees, and two pages or less to each of random forests, support vector machines, neural networks and deep learning, specifically convolutional neural networks for image data (supported by visuals from external sources). This chapter strikes me as rather superficial; the high-level statements are mostly accurate and reflect my own view on the relation between machine learning and statistics, and readers who were not aware of the different methods may gain a limited awareness of what the methods do; the explanation of support vector machines struck me as particularly tangible. Visualization content includes heatmaps of model predictions against certain features, several visualizations of a tree model for the iris data, and a scatterplot of response against a single feature, overlaid with scatter of random forest predictions against the feature. Partial dependence plots are mentioned but not explained; this was already the case in Chapter 10. Chapter 12 (“Many variables”) covers various principles such as small multiples, projection in general, principal components analysis, distance matrices, cluster analysis or t-SNE. The chapter has 14 pages, so that each topic can again only be superficially touched. The chapters in Part 3 of the book are of varying degree of detail, and readers will likely find valuable information in some and be disappointed by others. Which chapters are liked more or less is likely to depend on the reader’s background. Generally, Part 3 of the book is indeed most suitable for being read at leisure; deep-diving some of it can of course follow such a leisurely read.

In spite of some critical remarks, I do like this book. It is far from a manual or a reference book, and neither goes through principles in a systematic way. Instead, it covers many ideas by example. For instance, the color for visualizing the train delays data (also used in Figure 1) “was picked from a red maple leaf in real life using a smartphone app” (p. 25). The book has plenty of very good pieces of advice and does indeed make good on the promise of “all the time emphasizing statistical reasoning”. It also provides classical as well as current visualizations that are likely to inspire ideas. Robert Grant often gives advice that could be summed up as “Think!”, and thus criticizes dogmatic rules such as “All axes should start at zero”, of which he says that it misses the point (with which I agree). I like Part 2 of the book best, and was a bit disappointed with some chapters of Part 3, but nevertheless the book can be recommended as inspiring reading for data scientists, statisticians and other people who are interested in creating or consuming useful data visualizations.

References

Elliott A (2016). “My Daughters Sleeping Patterns for the First 4 Months of Her Life.”

Accessed 2021-04-11, URL https://www.reddit.com/r/dataisbeautiful/comments/5l39mu/my_daughters_sleeping_patterns_for_the_first_4/.

Rahlf T (2017). *Data Visualisation with R: 100 Examples*. Springer-Verlag, Switzerland.
doi:10.1007/978-3-319-49751-8.

Reviewer:

Ulrike Grömping
Beuth University of Applied Sciences Berlin
Department II
13353 Berlin, Germany
E-mail: groemping@bht-berlin.de
URL: <http://prof.beuth-hochschule.de/groemping/>