Methods in Statistical Graphics

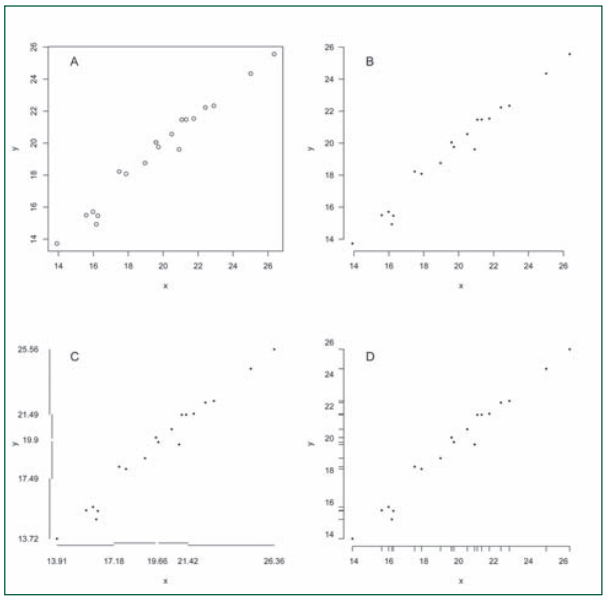
Tom Webb

To celebrate the launch of the new BES journal, *Methods in Ecology & Evolution* (see Freckleton *et al.*, this issue) we felt it was appropriate to instigate a regular 'methods' column in the ***Bulletin***. The aim of this series will be to discuss practical aspects of methods of particular relevance to BES members, with a particular focus on data analysis. However, in this first article, rather than focus on how to analyse data, we take a step back – or perhaps more accurately, a step forward in terms of the publication process – and think about how we might better *present* our data. This article draws heavily on the work of Edward R. Tufte (www.edwardtufte.com), who has spent his career exploring the most effective ways to present statistical graphics. Although one might not agree with everything he has to say – and he has some forthright views, for instance, "the only worse design than a pie chart is several of them…" – the exercise of thinking about our statistical graphics, of revising and editing them in the same way that we do text, is surely useful. The aim is that by combining quantitative expertise with graphical excellence, we might achieve "…a precision and grace in the presence of statistics."

To set this in context, in the 2001 edition of his seminal book *The Visual Display of Quantitative Information*, Tufte estimates that each year, somewhere between 9×10^{11} and 2×10^{12} images of statistical graphics are printed globally. The explosion of online content has surely increased this figure, but the point is that many of the graphics that we see in the technical literature are nowhere near as good as they could be. The news media is even worse ("…most news publications… operate at a pre-adult level of intelligence in graphical design"). Many computer packages excel (pun intended) in producing graphs full of what Tufte terms 'chartjunk' – unimportant clutter which detracts from the data. The central tenet of good design of statistical graphics might therefore be summed up as, 'above all, show the data'. Too often we don't trust our audience to find the data interesting, but in Tufte's words, "If the statistics are boring, then you've got the wrong numbers."

So, how might we go about improving the 'data density' of our figures? That is, maximising the proportion of ink (or pixels) used that actually communicates useful data. The first step – and this can be an intimidating one – is to step beyond the defaults of your statistical software. Even graphics produced directly from the best software are going to need some tweaking before they are ready for public consumption. One of the reasons that R is becoming so popular among ecologists (aside from its unparalleled statistical performance, and zero cost! See Petchey *et al*, p00, for more) is its graphical performance. In particular, even though its default settings are pretty sensible, it is possible to control programmatically every aspect of a plot, to produce precise, publication-quality graphics directly from your statistical software. Although R is capable of astonishingly sophisticated graphics, I consider here how Tufte suggests revising the humble (but immensely powerful) scatter plot. Again, this is not meant as a dogmatic argument in favour of any single plot – the outlet and intended audience of your work will influence your choice – but as a plea to consider the options, and to refine your graphics to maximise their impact. In the Open Source spirit of R, I will gladly provide the simple code used to produce these figures via email (t.j.webb@sheffield.ac.uk).

A default R scatter plot of some random data is shown in figure A. As you can see, the result is OK – everything looks reasonably clear, and there's little 'chartjunk' distracting from the strong positive association between the two variables. However, Tufte would not be happy – there is wasted ink in the box around the plot, the y-axis labels are not oriented for easy reading, and the open symbols are not very precise. All of that is easily rectified (figure B), but can we go further, and use some of the existing architecture of the plot to convey more data? Tufte suggests that the axes can communicate more than they are usually asked for. It is relatively easy to get them to extend to the data extremes, thus representing the range of each variable; in figure C, this concept is extended so that each axis is itself a quartile plot, indicating the data extremes, 25% and 75% quartiles, and median of each variable. Now we can examine the distribution of each variable in a reasonably efficient way at the same time as viewing all of the usual information portrayed in a scatter plot – here, for instance, the reasonably symmetrical distributions of both x and y are clearly apparent. Of course, the natural endpoint would be to represent the position of each data point on each axis, as in the dot-dash-plot (or rug plot) illustrated in figure D. For small datasets, it is even possible to replace the tick-marks on the axes with the actual values of each data point, squeezing pretty much every drop of data out of this simple plot.



Of course, any kind of plot can be subjected to the Tufte treatment. Figures E and F show, respectively, a default (well, almost – I tweaked it slightly) R histogram, and the 'Tufted' version, in which the x-axis labels are placed at the top of the relevant bar, which means we can really dispense with the bars – and the axis – altogether. Reading horizontally across from any value of x leads to a tick mark at the correct frequency on the y axis.

I suspect that the traditional scatter plot and histogram are safe for a while yet, and there will always be a compromise to be made between data density, visual clarity and aesthetic preference. In particular, I would argue that data density should be drastically reduced for any kind of oral presentation (incidentally, it may not be entirely surprising to learn that Tufte is rather scathing of a certain slide-making package, writing that "…the popular PowerPoint templates… usually weaken verbal and spatial reasoning, and almost always corrupt statistical analysis"). But now that science has entered, in the words of a Nature feature from September 2008, 'the petabyte era', effective graphical communication of huge datasets is becoming an imperative. Good graphics can make or break a paper, and even if you decide to disregard all of Tufte's principles of good design, the process of thinking about them must surely help in the process of communicating your science.

FURTHER READING

Tufte, E.R. (1990)	***Envisioning Information.* Graphics Press.**
Tufte, E.R. (2001)	***The Visual Display of Quantitative Information,* second edition. Graphics Press.**
Murrel, P. (2006)	***R Graphics.* Chapman & Hall / CRC**
Edward R. Tufte's website	**www.edwardtufte.com**
The R Graph Gallery	**http://addictedtor.free.fr/graphiques/allgraph.php**
Gallery of Data Visualization: The Best and Worst of Statistical Graphics	**http://www.math.yorku.ca/SCS/Gallery/**

Tom Webb is a Royal Society Research Fellow at the University of Sheffield and is Assistant Editor of the ***Bulletin***.