The Stata Journal (2010)
10, Number 4, pp. 670–681

Speaking Stata: Graphing subsets

Nicholas J. Cox
Department of Geography
Durham University
Durham, UK
n.j.cox@durham.ac.uk

Abstract. Graphical comparison of results for two or more groups or subsets can be accomplished by way of subdivision, superimposition, or juxtaposition. The choice between superimposition (several groups in one panel) and juxtaposition (several groups in several panels) can require fine discrimination: while juxtaposition increases clarity, it requires mental superimposition to be most effective. Discussion of this dilemma leads to exploration of a compromise design in which each subset is plotted in a separate panel, with the rest of the data as a backdrop. Univariate and bivariate examples are given, and associated Stata coding tips and tricks are commented on in detail.

Keywords: gr0046, graphics, subdivision, superimposition, juxtaposition, quantile plots, Gumbel distribution, scatterplots

1 Introduction

A common graphical problem—indeed for many researchers the key graphical problem—is to compare results for two or more groups or subsets of some larger group or set. Results might be measured responses, absolute or relative frequencies, summary statistics, parameter estimates, model figures of merit, or whatever else is worth plotting. We might seek comparisons according to treatment, disease, gender, ethnicity, industry, product, habitat, land use, area, time period, and so forth: you can multiply examples for yourself.

Various strategies, elementary but also fundamental, recur repeatedly in plotting results for different groups.

Subdivision of some whole is the principle behind pie charts, stacked bar charts, and layered area plots. Each group is represented by its own share, relative or absolute as the case may be.

Superimposition of differing points or lines is the principle behind scatterplots, line plots, and other plots in which different groups are denoted by (for example) distinct marker symbols, marker colors, line patterns, or line colors.

Juxtaposition of separate subpanels or panels within a display conveys group comparisons by what Tufte (2001) called small multiples: the same basic design is repeated for each subset, and possibly also for the total set. Other terms in use are trellis and lattice graphics, multipanel graph, and panel charts (Robbins 2010).

All these ideas are staples within statistical and scientific graphics, and their advantages and disadvantages have been much discussed in many texts and articles. The books of Tufte (1990, 1997, 2001, 2006) and Cleveland (1993, 1994) remain my own favorites as overviews of the field. Despite that large literature, with many preachers and many precepts, it often seems that only experimentation will show which idea is most effective for a particular dataset.

Subdivision is the strategy likely to be first encountered in graphical education through the pies and bars widely met in childhood. However, only some graphical problems reduce to comparing fractions of a whole.

Superimposition promises the advantage of a common scale that can be used for comparison, yet in practice superimposition can mean confusion. The mixture of different elements may appear mostly as a mess in which patterns are difficult to discern. Tangled line plots are often referred to as spaghetti plots, not always with affection or admiration. Some other term (muesli plots?) seems needed for classified scatterplots that convey much detail but in which systematic differences are hard to decipher.

Juxtaposition—in Stata terms often effected by a `by()` or `over()` option—provides separation, which clarifies what is being compared, sometimes at the expense of making that comparison harder work. To work well, juxtaposition still requires mental superimposition. Judging the fine structure of differences between adjacent panels can be difficult enough, and judging differences between panels at opposite ends of a display is evidently even more difficult.

In this column, we will look at a combination of superimposition and juxtaposition, in which subsets are shown separately, but in every case the set as a whole acts as a backdrop. An earlier Stata tip (Cox 2009) emphasized the notion that the whole of the data may serve as a graphical substrate for a particular subset. Repeating information might be criticized as redundant, but rather the idea is that repetition provides reinforcement. Consider a dictum of Tufte (1990, 37): "Simplicity of reading derives from the context of detailed and complex information, properly arranged. A most unconventional design strategy is revealed: to clarify, add detail."

In your own work, you are likely to be able to use color in your talks and possibly also within your reports or even published papers. However, many journals still prohibit or inhibit anything other than black and white and what shades of gray lie between. In this column, we follow such a restriction, using only contrasts discernible by varying gray scale.

2 A univariate example

2.1 Annual maximum windspeeds

A first example concerns data on annual maximum windspeeds for various places in the southeastern United States. My source is Hosking and Wallis (1997, 31); their source is Simiu, Changery, and Filliben (1979). A recipe that has long been a standard in the

statistics of extremes is to focus on the maximums of a variable in each of several blocks of time. A year is a natural block for meteorology and climatology. The data are for varying numbers of years, ranging from 19 years (1958–1976) for Key West, Florida, to 35 years (1943–1977) for Brownsville, Texas, so that we should prefer a common basis for graphical comparison of these univariate samples. `windspeed.dta` is provided with the media for this issue of the *Stata Journal*. The dataset includes two variables, `windspeed` and `place`. The ordering of the places is by mean maximum windspeed.

Researchers accustomed to such data tend to reach first for quantile plots. The official command `quantile` is limited to one batch of data at a time. While more versatile user-written alternatives are available (Cox 2005a, 2010), the spirit of this particular column is that you can work out code for yourself using basic commands.

Given an ordered sample of size n for variable y, $y_{(1)} \leq y_{(2)} \ldots y_{(n-1)} \leq y_{(n)}$, the usual ordinate and abscissa for a quantile plot are $y_{(i)}$ and $(i-a)/(n-2a+1)$, respectively. The abscissa for some choice of a is, in effect, an empirical cumulative probability and is often called a *plotting position*. The naïve choice i/n for plotting position would imply probabilities $1/n$ and 1 at the ends of the data, while $(i-1)/n$ would just reverse the problem. Either choice would be awkward, implying that no value can be more extreme than those observed and because theoretical quantiles are often not defined for probabilities 0 or 1. We need, therefore, a slightly more complicated method. The choice of a is the subject of a small but contentious literature, to which Thas (2010) is one entry point. Given a choice, we can implement it for ourselves in Stata. Below we use $a = 0$—that is, $i/(n+1)$, as is common in statistics of extremes.

To calculate plotting positions, it is convenient, if not outstandingly efficient, to use `egen` functions. These functions take care of sorting issues, handling of any missing values, and separate calculations for separate groups:

```
. use windspeed
. egen rank = rank(windspeed), by(place) unique
. egen count = count(windspeed), by(place)
. generate pp = rank/(count + 1)
. label variable pp "fraction of data"
```

Figure 1 shows quantile plots, with a separate panel for each place.

```
. scatter windspeed pp, by(place) yla(, ang(h)) xla(0(.25)1)
```

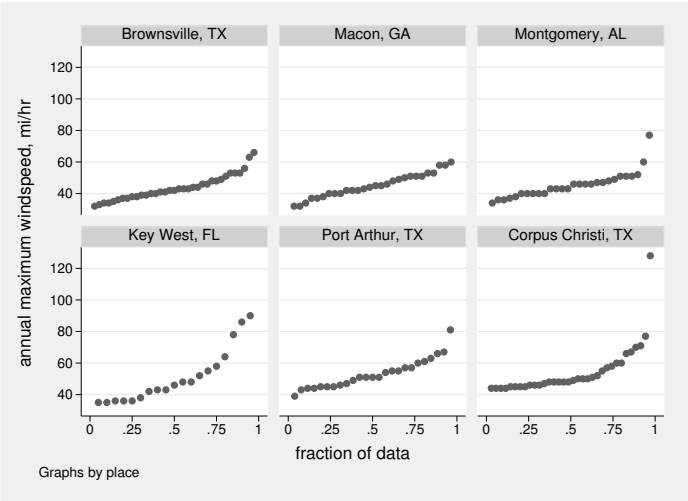

Figure 1. Quantile plots of annual maximum windspeed data for six places in the southeastern United States

The usual starting point with such data is the fitting of Gumbel distributions, with a distribution function for location parameter (mode) ξ and scale parameter α of

$$F(y) = \exp[-\exp\{-(y-\xi)/\alpha\}]$$

defined over the real line. Gumbel distributions are named for Emil Julius Gumbel (1891–1966), who did much to systematize knowledge of the statistics of extremes and wrote the first extended monograph on the subject (Gumbel 1958). For more detail on his scientific and political career, see Freudenthal (1967), Hertz (2001), and Brenner (2001).

For context, note that the mean of a Gumbel distribution is $\xi + \alpha\gamma$ and the standard deviation is $\alpha\pi/\sqrt{6}$. Here $\gamma \approx 0.57721+$ is Euler's constant (in Stata, `-digamma(1)`) and $\pi \approx 3.14159$ (in Stata, `_pi` or `c(pi)`) is the even better known constant. Without venturing into numerical fits, the distribution function can easily be inverted, giving

$$y(F) = \xi - \alpha \ln(-\ln F)$$

so that a plot of y against $-\ln(-\ln F)$ should be approximately linear with intercept ξ and slope α if y is drawn from a Gumbel distribution. The quantity $-\ln(-\ln F)$ is thus

often called a Gumbel reduced variate, "reduced" implying unit-free and dimensionless. The resulting plot is a Gumbel plot. For another example and literature references, see Cox (2007a).

Because the plotting position variable `pp` has already been calculated separately for each place, we can apply functions as just stated algebraically:

```
. gen gumbel = -ln(-ln(pp))
. label var gumbel "Gumbel reduced variate"
```

There are two obvious versions of the corresponding graph. Figure 2 separates out different places into different panels:

```
. scatter windspeed gumbel, by(place) yla(, ang(h))
```

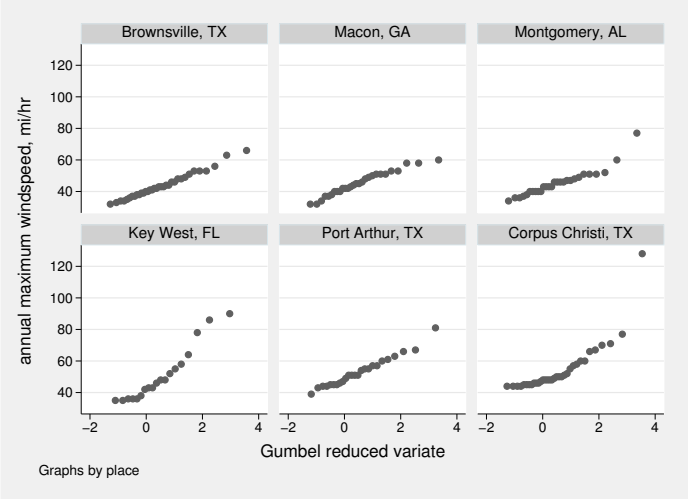

Figure 2. Gumbel plots of annual maximum windspeed data for six places in the southeastern United States, one panel for each place

Figure 2 is clearly ideal for considering individual places. How easy does it make comparison, however? Conversely, figure 3 separates places by using different point symbols (Cox 2005b). The `separate` command (see [D] **separate**) makes this step easier but is not essential because a series of `if` qualifications could produce the same result.

```
. separate windspeed, by(place) veryshortlabel
. scatter windspeed? gumbel, ytitle("`: var label windspeed´") yla(, ang(h))
> legend(pos(11) ring(0) order(6 5 4 3 2 1) col(1))
```

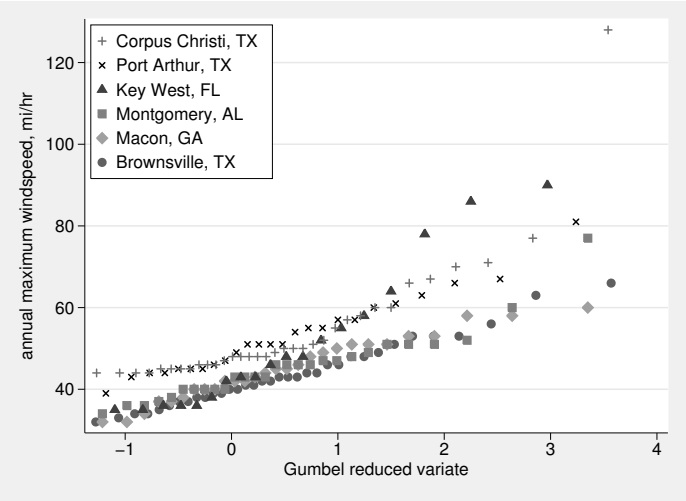

Figure 3. Gumbel plots of annual maximum windspeed data for six places in the southeastern United States, places being superimposed

I initially ordered the places lowest mean first. Now it becomes evident that the reverse order is needed here. More importantly, graphs like figure 3 appear frequently in books and journals. But how effective are they? Readers will find it easy to understand the principle: given detailed study of the legend, they could study the graph carefully to learn more about contrasts. But will they be encouraged to do that by the design? And how easy would that be? It is a characteristic of these data, like many others, that the groups overlap to some extent. With this design, however, that inevitably implies that some groups are partially obscured on the graph by others.

The resulting dilemma is clear. Figure 2 and figure 3 have corresponding advantages and limitations. Moreover, the example should strike many readers as modest if not minute in size, with just 6 groups and sample sizes between 19 and 35. The problem with more data can be much more serious.

A suggested compromise is this: Show each group separately, but with the rest of the data shown as a backdrop. Figure 4 is the result. Now, as is natural in many ways, the other data provide context for each subset.

```
. qui forval i = 1/6 {
>     scatter windspeed gumbel if place != `i´, ms(Oh) mcolor(gs12)
>     || scatter windspeed gumbel if place == `i´, ms(D) mcolor(gs1)
>     yla(, ang(h)) yti("") xti("") legend(off)
>     subtitle("`: label (place) `i´´", box fcolor(gs13) bexpand size(medium))
>     name(g`i´, replace)
> }
```

```
. graph combine g1 g2 g3 g4 g5 g6, imargin(small)
> l2ti("`: var label windspeed´") b2ti("`: var label gumbel´")
```

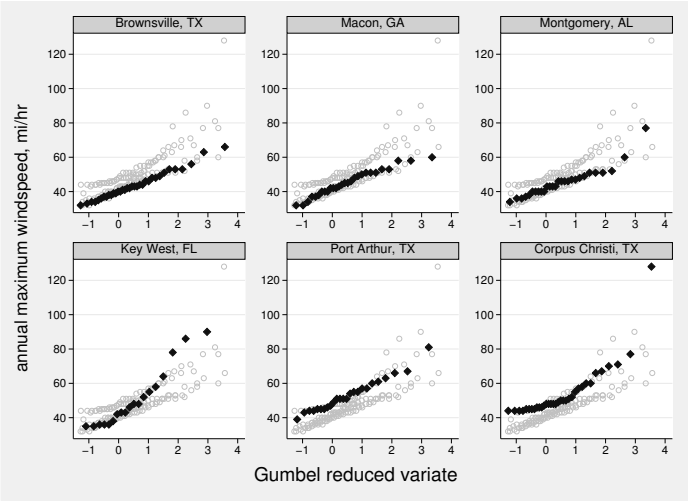

Figure 4. Gumbel plots of annual maximum windspeed data for six places in the southeastern United States. Data for the five other places are shown as a backdrop to the data for each place.

The code is more complicated than for previous graphs, but the logic is still straightforward. The next subsection gives a commentary for those who would like to see matters discussed in greater detail. The details of what you see printed depend on having previously set the *Stata Journal* graph scheme by typing

```
. set scheme sj
```

2.2 Comments on code

```
qui forval i = 1/6 {
```

1. We loop using `forvalues` over the distinct groups of a categorical variable (`place`). In this case, we know in advance that there are 6 groups, numbered 1–6. In more general situations, we might want to automate the looping. Various techniques exist for doing that. One method is to use `levelsof` (see [P] **levelsof**) to produce a list of the distinct groups, followed by a call to `foreach`. Another, perhaps easier, method is to use the `group()` function of `egen` (see [D] **egen**) to define a grouping variable with positive integer values (Cox 2007b).

```
scatter windspeed gumbel if place != `i´, ms(Oh) mcolor(gs12)
```

2. We first lay down the rest of the data as a backdrop. So, for example, the first time around the loop, when `'i'` evaluates to 1, we look for `place` not equal to 1. The data for that complementary subset are shown lightly. The suggestion here is that open circles `Oh` and a light color `gs12` are suitably muted.

   It is tempting just to plot all the data as substrate, on the grounds that we will just plot over each subset being emphasized. In principle, that is correct; in practice it is possible for small parts of the symbols in question to be visible even though they lie underneath the symbols to be plotted. So do use the `!=` constraint.

   ```
   || scatter windspeed gumbel if place == `i´, ms(D) mcolor(gs1)
   ```

3. The subset being emphasized is now plotted directly on top. More prominent symbols and colors are needed. (In other problems, line patterns, widths, and colors will be the elements to adjust.)

   ```
   yla(, ang(h))  yti("") xti("") legend(off)
   ```

4. About the options specified, note first that `yla(, ang(h))` is a personal choice, although the underlying logic that text is more readable horizontally is a point on which many will agree. More particular to the key themes here are suppression of `ytitle()`, `xtitle()`, and `legend()`. In this problem, the `ytitle()` and `xtitle()` would be the same on all six graphs, which is unnecessary. We will see in a moment how to put just one title on both the left and bottom sides of the overall graph. The appearance of a legend would be triggered by the double plotting of `windspeed`. A legend would not help, so we suppress it, too.

   ```
   subtitle("`: label (place) `i´´", box fcolor(gs13) bexpand size(medium))
   ```

5. Evidently, each graph needs some explanatory text. `place` is a numeric variable with value labels attached, so we use an extended macro function (see [P] **macro**) to look up the label concerned as we go around the loop. If no label were attached to the numeric value in question, the value itself would be shown instead. This approach does not extend to string variables, except that there is an easy work-around: just map the string variable to a numeric variable with value labels first, using `encode` (see [D] **encode**) or the `egen` function `group()` that was mentioned earlier.

   In the rendering of the text, the extra options `box`, `fcolor(gs13)`, `bexpand`, and `size(medium)` are doubly optional. In this case, they come from peeking at graphs produced with `by()` in the Graph Editor and so producing a similar overall style.

   ```
   name(g`i´, replace)
   ```

6. It is essential that we save each graph for later combining, which is the next step. There is a choice between using `name()` and using `saving()`. A side effect of using `name()` is that each resulting graph remains open in a separate Graph window, so that it can be checked. Although we have not previously used any of these graph

names, writing "`, replace`" will let you revise this code a little more easily—that is, assuming that you do not write perfect code the first time and every time.

```
graph combine g1 g2 g3 g4 g5 g6, imargin(small)
l2ti("`: var label windspeed´") b2ti("`: var label gumbel´")
```

7. `graph combine` is used to put the graphs together. In this case with six graphs, the default of two rows and three columns looks fine. We add titles to the combined graph using `l2title()` and `b2title()`. The option names denote titles on the `l`eft and `b`ottom of the graph (and may evoke nostalgia among longtime Stata users for the graphics syntax used before Stata 8). Notice further how we generalize a step beyond wiring in the particular variable labels: the extended macro option calls ensure that Stata looks up the current variable labels, so that the same code can be used even if we change the variable labels. (More general code yet would protect against the possibility that no variable labels have been assigned.)

3 Intermezzo: Advice from Edward Tufte

Make all visual distinctions as subtle as possible, but still clear and effective (Tufte 1997, 73).

Minimal contrasts of the secondary elements (figure) relative to the negative space (ground) will tend to produce a visual hierarchy with layers of inactive background, calm secondary structure, and notable content. Conversely, when *everything* is emphasized, *nothing* is emphasized; the design will often be noisy, cluttered, and informationally flat (Tufte 1997, 74).

4 A bivariate example: Cirque lengths, widths, and grades

For a bivariate example, we examine some data similar to a dataset used in Cox (2005b). The data (Evans and Cox 1995) refer to the lengths and widths of cirques in the Lake District, England. `cumbrian_cirques.dta` is provided with the media for this issue of the *Stata Journal*. Cirques are armchair-shaped hollows formerly occupied by glaciers. Length and width are basic quantitative measures of their size. Logarithmic scales are standard for such data. Here we also bring in grade, a judgment-based variable of how well developed each feature is on a five-point ordered scale from classic to poor.

Figure 5 is a standard scatterplot. Because we will have five scatterplots for each grade, we might as well use the `total` suboption to add a panel for all the data combined. An ad hoc refinement is to insist on an extra space in the axis label at 2000 meters to prevent the last digit from being elided in the combined display.

```
. use cumbrian_cirques, clear
. scatter width length, by(grade, total) xsc(log) ysc(log) ms(Oh)
> xla(200 500 1000 2000 "2000 ") yla(200 500 1000 2000)
```

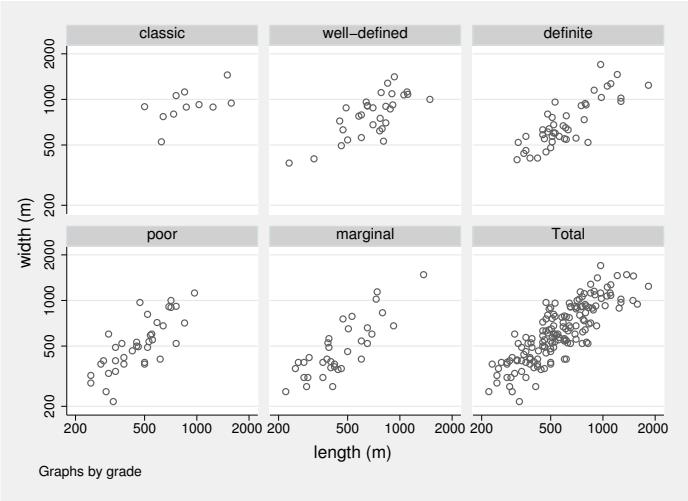

Figure 5. Scatterplots of cirque width and length by grade for the English Lake District

In the compromise design, most of the small code tricks are the same, but as with figure 5 we add a further display of all the data. So that you have a comparison of style with earlier graphs, we will leave the subtitle area unboxed. Figure 6 is the result.

```
. forval i = 1/5 {
>     scatter width length if grade != `i´, xsc(log) ysc(log) ms(Oh) mcolor(gs12)
>     xla(200 500 1000 2000 "2000 ") yla(200 500 1000 2000)
>     || scatter width length if grade == `i´, xsc(log) ysc(log) ms(D) mcolor(gs1)
>     yla(, ang(h)) yti("") xti("") legend(off)
>     subtitle("`: label (grade) `i´´", size(medium)) name(g`i´, replace)
> }
. scatter width length, xsc(log) ysc(log) ms(D) mcolor(gs1)
> xla(200 500 1000 2000 "2000 ") yla(200 500 1000 2000)
> yla(, ang(h)) yti("") xti("") legend(off)
> subtitle("all cirques", size(medium)) name(g6, replace)
. graph combine g1 g2 g3 g4 g5 g6, imargin(small)
> l2ti("`: var label width´") b2ti("`: var label length´")
```

(*Continued on next page*)

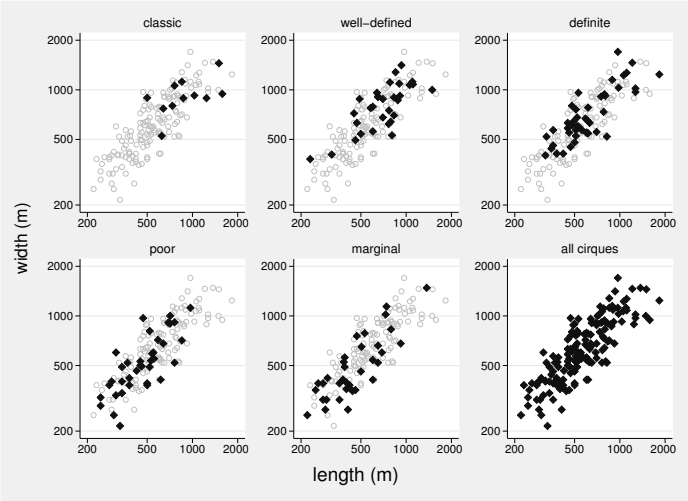

Figure 6. Scatterplots of cirque width and length by grade for the English Lake District. Data for the four other grades are shown as a backdrop to the data for each grade.

5 Conclusions

The conclusions lie with you, the reader. This column has a flavor of experiment. Do you think that the compromise design—which might be called a subset and substrate design—has any advantages over alternatives for the examples here? Do you think that you can use the ideas here to improve your own comparative displays? Many graph types lie open for exploration, including especially attempts to more easily and more effectively see fine structure in spaghetti plots.

6 References

Brenner, A. D. 2001. *Emil J. Gumbel: Weimar German Pacifist and Professor*. Boston: Brill.

Cleveland, W. S. 1993. *Visualizing Data*. Summit, NJ: Hobart.

———. 1994. *The Elements of Graphing Data*. Rev. ed. Summit, NJ: Hobart.

Cox, N. J. 2005a. Speaking Stata: The protean quantile plot. *Stata Journal* 5: 442–460.

———. 2005b. Stata tip 27: Classifying data points on scatter plots. *Stata Journal* 5: 604–606.

———. 2007a. Stata tip 47: Quantile–quantile plots without programming. *Stata Journal* 7: 275–279.

———. 2007b. Stata tip 52: Generating composite categorical variables. *Stata Journal* 7: 582–583.

———. 2009. Stata tip 78: Going gray gracefully: Highlighting subsets and downplaying substrates. *Stata Journal* 9: 499–503.

———. 2010. Software Updates: gr42_5: Quantile plots, generalized. *Stata Journal* 10: 691–692.

Evans, I. S., and N. J. Cox. 1995. The form of glacial cirques in the English Lake District, Cumbria. *Zeitschrift für Geomorphologie* 39: 175–202.

Freudenthal, A. M. 1967. Emil J. Gumbel. *American Statistician* 21(1): 41.

Gumbel, E. J. 1958. *Statistics of Extremes*. New York: Columbia University Press.

Hertz, S. 2001. Emil Julius Gumbel. In *Statisticians of the Centuries*, ed. C. C. Heyde and E. Seneta, 406–410. New York: Springer.

Hosking, J. R. M., and J. R. Wallis. 1997. *Regional Frequency Analysis: An Approach Based on L-Moments*. Cambridge: Cambridge University Press.

Robbins, N. B. 2010. Trellis display. *Wiley Interdisciplinary Reviews: Computational Statistics* 2: 600–605.

Simiu, E., M. J. Changery, and J. J. Filliben. 1979. Extreme wind speeds at 129 stations in the contiguous United States. Building Science Series 118, National Bureau of Standards.

Thas, O. 2010. *Comparing Distributions*. New York: Springer.

Tufte, E. R. 1990. *Envisioning Information*. Cheshire, CT: Graphics Press.

———. 1997. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, CT: Graphics Press.

———. 2001. *The Visual Display of Quantitative Information*. 2nd ed. Cheshire, CT: Graphics Press.

———. 2006. *Beautiful Evidence*. Cheshire, CT: Graphics Press.

About the author

Nicholas Cox is a statistically minded geographer at Durham University. He contributes talks, postings, FAQs, and programs to the Stata user community. He has also coauthored 15 commands in official Stata. He wrote several inserts in the *Stata Technical Bulletin* and is an editor of the *Stata Journal*.