

Scatterplotting

Antony Unwin
University of Augsburg, Germany
antony.unwin@math.uni-augsburg.de

1 Introduction

In a recent column in this series Duncan Murdoch [2000] described the benefits of writing your own scatterplot code. He mentioned:

- the importance of specifying good scales;
- the flexibility of having your own software;
- the advantages of overlaying lines (or curves);
- the need for multiple plots;

and, in an aside,

- the possibilities of having interactive tools for scaling and reformatting.

Although that is a longish and a good list, it is in some ways more striking for what it leaves out than for what it includes and this column discusses what else every self-respecting scatterplot should have.

Predominantly interactive features are suggested, something that is sadly missing from most statistical software. Many of the additions described would be advantageous in static displays for presentation graphics too, but it is much more effective if they are implemented interactively. Writing fully interactive software is challenging, but rewarding. The interactive tools which software should offer in general are discussed in [Unwin, 1999]. Principles of GUI design relating to statistical software are outlined in [Unwin et al, 1999]. John Chambers has pointed out how awful GUIs can be [Chambers, 1999], but many bad examples should not discourage us from aspiring to do better. Although Duncan discussed scatterplots in isolation, this article assumes that the scatterplots can be part of a larger graphics system, and that all graphic displays are linked. This puts a greater burden on the programmer, but results in a much more powerful tool.

One topic not touched on in this article is how scatterplots look. It would be simple, but trite, to fill these pages with ugly scatterplots and make scathing remarks about them. Aesthetics is a contentious subject, where it is easier to criticise something for being bad than to agree on what makes something good. We can only encourage everyone to try to make their scatterplots look well — clear, informative and unencumbered with unnecessary chart junk.

2 Additional scatterplot features

2.1 Drawing overlapping points

Even in small data sets you can have to draw multiple points at the same screen location, either because the plot resolution is not high enough or because the points have exactly the same (X,Y) values. Suggestions like jittering, sunflowers or transformations are all impractical in general and it is

surprising that otherwise reputable books on graphics recommend them at all. That Lee Wilkinson [1999] can write in his excellent book on graphics “If points overlap we can use transparency to prevent occlusion” is more than surprising.

One solution is to use density estimation with a shading (or colouring) scheme to distinguish areas of different density. This has been suggested for very large data sets with hexagonal binning (Carr et al 1987) but would work for small data sets too, if necessary. Two issues should be clarified. Firstly, there is no need to estimate density where there are no points, if you are interested in data display as distinct from model estimation. (Or as Hyndman [1996] neatly puts it “Besides, with few observations and no prior knowledge of the underlying density, there seems little point in attempting to summarize the sample space.”) Secondly, interactive control of the kernel window-width and interactive zooming are valuable for checking local details. Heike Hofmann’s software MANET (Hofmann [2000], Unwin et al [1996]) uses a simple but fast approach that was first used in Graham Wills’ REGARD (Unwin [1994]). The brightness of each pixel depends linearly on the number of points overlapping the pixel up to a user-defined limit L . The points are circles whose size can be changed directly using the arrow keys. Figure 1 shows an example using the Old Faithful geyser data. More sophisticated two-dimensional density estimation should be better and would certainly be more elegant, but the graph shows the data concentrations clearly. Hyndman [1996] describes a bootstrapping approach for identifying HDRs (high density regions), where “high” has been predefined. His Figure 5 for Old Faithful also shows the two main modes, but is not interactive and takes much more calculation effort.

(Figure 1 about here)

2.2 Panning and Zooming

Zooming has already been mentioned above for inspecting details where overlapping arises, but obviously has more general application. It should be direct, though. Resetting limits in a command line or in a dialog box to redraw a display is a necessary option for precise predefined work, but is frustrating and discouraging for exploratory analysis. Some form of magnifying glass, as may be found in graphics programmes, is intuitive and swift. Figure 2 shows an example where it is needed. The plot on the left shows the apparent lack of association between profits and turnover in a data set from a bank. The plot on the right shows the same data set concentrating on an area 10^{-7} of the size of the one on the left. The maximum of the X-axis (turnover) has been reduced by a factor of 10^4 and the Y-axis maximum by 10^3 .

(Figure 2 about here)

Zooming without panning is like Tom without Jerry (or Kendall without Stuart?). You need to be able to move around to check what you have found and to make valid comparisons. And since we are following the tenor of Duncan’s original column and specifying what capabilities your own scatterplot should have, we need to include an orientation plot which tells you where your zoom is currently in the context of the whole display. Games

(e.g. Sim City) and geographic software (e.g. Graham Wills' REGARD) often have this.

Zooming neatly illustrates a general principle in designing statistical graphics: it is not the graphic elements that you want to magnify or shrink, but the statistical components. If 5 neighbouring, but different, points are represented at low resolution by a single screen point, then high resolution should show the 5 separate points and not just a bigger version of the original screen display. A useful term here is logical zoom, meaning that more details are shown at a higher level of resolution.

2.3 Large data sets

§2.1 and 2.2 have already described tools which would be useful for large data sets. It is a continual puzzle that more has not been done here. Surely all of us have to deal with large data sets often (and "large" for scatterplots starts at a few hundred, let alone the millions or more that can arise). There is a typical display in Derek Briggs' article on the effect of preparing for SAT in *Chance* recently (Briggs [2001]) involving around 3500 cases. This example draws attention to a challenging unsolved problem: even if we had a good density estimation display of the data, how should highlighted points be shown? Briggs' filled circles on unfilled circles are not very clear, but does anyone have anything better? Both MANET and Data Desk (Velleman [1997]) show the same highlighting at a point whether there is 1 case selected or 100 at that point. Furthermore, there is no information as to how many unselected cases are present. Data Desk offers "hot selection", a very powerful way of riffling through plots of selected subsets. This is effective, even with scatterplots of hundreds of thousands of points, because the redrawing is so fast. Wegman and Luo [1997] have implemented a hardware approach which they call saturated brushing in which colours of overplotted points are added. Something of this kind is clearly needed, but combined with density estimation and in an effective interactive form with direct controls over the density estimation parameters and the highlighting format.

2.4 Querying

Scatterplots may reveal many different features of the association between two variables. Querying points to find out which cases they represent, querying axes to find out where gaps or boundaries fall, querying clusters to find out how many cases are actually there are all useful facilities. Some software packages would claim they offer interactive scatterplots because you can type in a command which labels individual cases (e.g. S-Plus) or because coordinates of a point can be read off at some faraway corner of the screen when the cursor passes over the point. Querying should be immediate in place and time. If you query a point, then the information should appear where your focus of attention lies. You should get all the information you need, not just the coordinates, but the variables that the X and Y axes represent. Analysing a complex data set with many open windows, you need all the (well-designed) help you can get.

Querying works consistently when it is the same command in any window, but gives context-sensitive responses according to the window (be it a boxplot or a scatterplot or a mosaic plot or variable list or whatever) and according to where in the window has been queried. Responses to queries are generally

best left temporary. Options which permanently label points are fine for tiny data sets, but can only be used in limited special cases.

In MANET we have experimented with offering different levels of information. For instance, querying a point in a scatterplot gives the number of cases overlapping that pixel, the number currently highlighted, the names of the X and Y variables and the range of values for X and Y covered by the queried points, while deeper querying gives the names and values of any currently selected variables for the points queried.

No software that I know of permits querying of point clusters (unless Duncan has added this feature to his in the meantime), but it would be a useful capability. The same results can be achieved, albeit less directly and less interactively, using selection and linking.

2.5 Missing values

Ignoring missing values can be misleading. Axes can be drawn too short (cases missing on Y are not included in the determination of the limits of the X axis) and linking from another display to the scatterplot may suggest result patterns which would be discounted if all cases were included. MANET was first developed to see how interactive graphics tools could be extended to take account of missing values. Scatterplots are a particular problem, because values can be missing on X alone, on Y alone, or on both. Figure 3 shows an example from an exam marks data set that was published in the first article on MANET in 1995 [Unwin et al, 1995]. 15 of 110 students answered both questions, but only those are shown in the left hand plot. In the right hand plot the cases missing on one variable are plotted on the appropriate axis and are also included in a box representing all cases missing on that variable. There is a third box for the 35 cases missing on both variables. Note that the highest marks on each question were obtained by students who did not answer the other question.

(Figure 3 about here)

There are probably other solutions to dealing with missings, I look forward to seeing what people who write their own code suggest.

2.6 Weighted data

Variables like morbidity rates, financial averages or constituency voting percentages often have to be weighted to clarify their relative importance. A high rate of cancer incidence in a small population region should carry less weight than a high rate in a large population region. Bubble plots, where the size of the point represents a third dimension, are an obvious way to do this, but have rightly not found much favour in the statistical community, because there is usually too much overlapping for the resulting display to be easily interpretable. Coupled with interactive size control and density estimation, bubble plots could be much more informative. A version where a weighting variable can be dragged into the scatterplot has been implemented for testing purposes in MANET this year.

2.6 Rescaling and reformatting

Duncan mentioned the importance of good default scaling, but no scaling algorithm can cope with every data set, so that interactive rescaling tools would be helpful. It would also make sense to link scales across graphics, though in two different ways. If the scale for a variable is changed in one display, we probably want to have that scale in all displays of that variable. Often it is useful to give a number of different graphics a common scale, which is calculated based on the ranges of all of them (which MANET allows). Data Desk offers another alternative, forcing one display to adopt the scaling of another.

Assigning colours or shapes to selected points or to points according to values on another variable can be extremely helpful, when groups are separated in a scatterplot. Switching interactively between different options encourages exploratory analysis.

Switching between many options or toggling between two are capabilities that should in general be included in interactive software. Switching the X and Y axes in a scatterplot can give new insights, but it has to be an easy (and easily reversed) operation.

3 More complex ideas

There are several other ideas that could be mentioned as well; subgroup plots or the more sophisticated conditioning of trellis graphics, overlaying and layering of plots, representing points by glyphs to incorporate additional data dimensions, and doubtless many more. These would involve substantial development effort, but would considerably enlarge the scatterplot's potential. One feature that can safely be ignored is false 3-d plots, i.e. adding an unnecessary third dimension for supposedly decorative reasons.

4 Conclusion

As readers will have gathered from this, I found Duncan's article very thought provoking. You do learn a lot from writing software yourself, but a second aim should be to do something innovative, to try out new ideas. I hope these comments will have convinced anyone thinking of writing their own scatterplots that there are many interesting open issues and encouraged them to think far beyond just making slight improvements on what is generally currently available.

References

Briggs, D. C. (2001). The Effect of Admissions Test Preparation. Chance, 14(1), 10-18.

Carr, D. B., Littlefield, R.J., Nicholson, W.L., Littlefield, J.S. (1987). Scatterplot Matrix Techniques for large N. JASA, 82(398), 424-436

Chambers, J. (1999). Computing with Data: Concepts and Challenges. American Statistician, 53(1), 73-84.

Hofmann, H. (2000). MANET www1.math.uni-augsburg.de/Manet/ Augsburg: Rosuda.

Hyndman, R. J. (1996). Computing and Graphing Highest Density Regions. American Statistician, 50, 120-126

Murdoch, D. J. (2000). Drawing a scatterplot. Chance, 13(3), 53-55.

Unwin, A. R. (1994). REGARDing Geographic Data. In P. Dirschedl and Ostermann, R. (Eds.), Computational Statistics (pp. 315-326). Heidelberg: Physica.

Unwin, A.R. (1999) Requirements for Interactive Graphics Software for Exploratory Data Analysis Computational Statistics, 14, 7-22

Unwin, A. R., Hofmann, H. (1999). GUI and Command-line — Conflict or Synergy? In K. Berk Pourahmadi, M. (Ed.), Computing Science and Statistics, Proceedings of the 31st Symposium on the Interface, 31 (pp. 246-253). Chicago: Interface Foundation.

Unwin, A. R., Hawkins, G., Hofmann, H., and Siegl, B. (1996). Interactive Graphics for Data Sets with Missing Values - MANET. Journal of Computational and Graphical Statistics, 5(2), 113-122

Velleman, P. (1997). Data Desk www.datadesk.com. Ithaca: Data Description.

Wegman, E., Luo, Q. (1997). High dimensional clustering using parallel coordinates and the grand tour. Computing Science and Statistics, 28, 352-360.

Wilkinson, L. (1999). The Grammar of Graphics. New York: Springer.

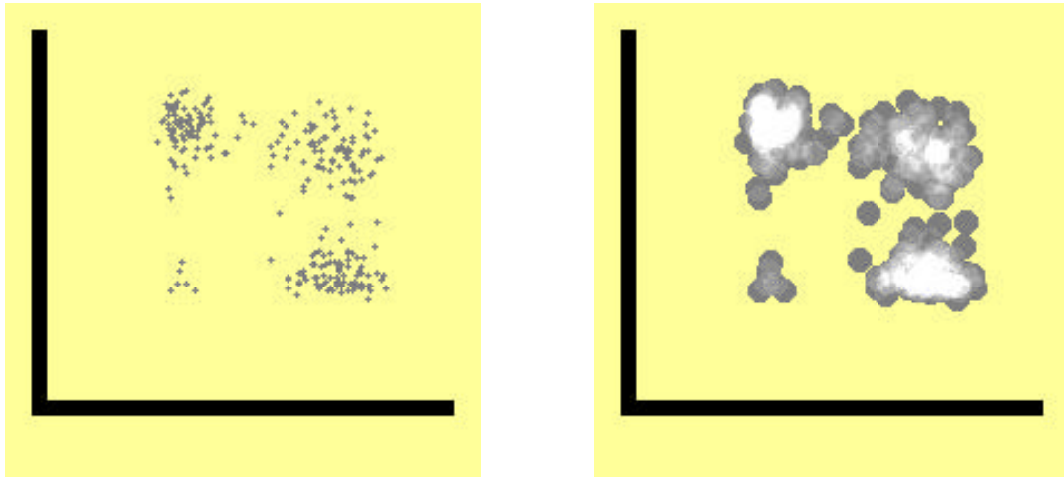


Figure 1 Scatterplots of eruption duration against previous duration from the Old Faithful geyser data. There is no adjustment for overplotting on the left. On the right, each point is drawn slightly bigger and pixels with 8 or more overlapping points are bright white.

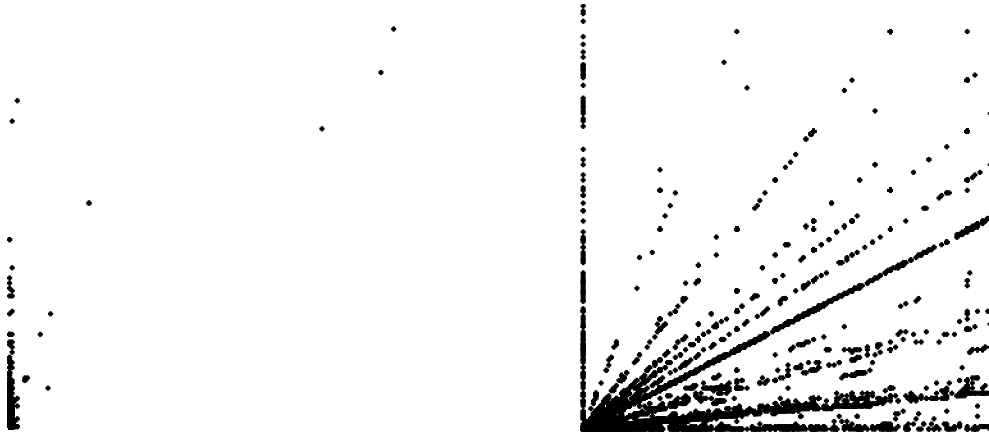


Figure 2 Profits plotted against turnover for bank transactions. The left hand plot is based on a subset of 17243 data points, the right hand one on 9095 points from that subset close to (0,0). The right hand plot was drawn by reducing the Y axis scale by a factor of 10^3 and the X axis by a factor of 10^4 .

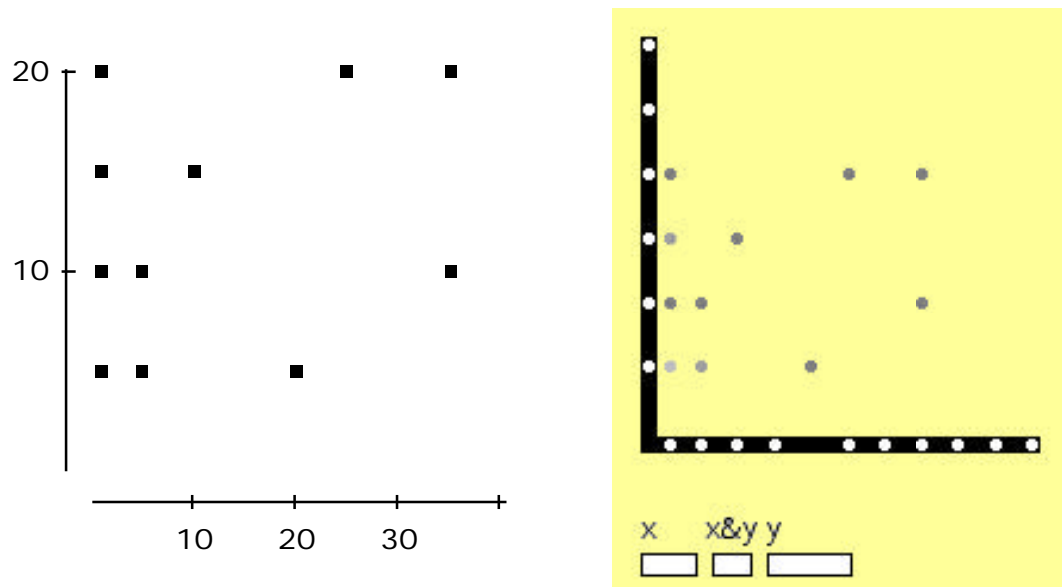


Figure 3 Scatterplots of marks of students on questions 5 and 6. Plot from Data Desk on the left, from MANET on the right.