Designing science graphs for data analysis and presentation

The bad, the good and the better

Dave Kelly, Jaap Jasperse and Ian Westbrooke

DEPARTMENT OF CONSERVATION TECHNICAL SERIES 32

Published by
Science & Technical Publishing
Department of Conservation
PO Box 10–420
Wellington, New Zealand

Department of Conservation Technical Series presents instructional guide books and data sets, aimed at the conservation officer in the field. Publications in this series are reviewed to ensure they represent standards of current best practice in the subject area.
Individual copies are printed, and are also available from the departmental website in pdf form. Titles are listed in our catalogue on the website, refer www.doc.govt.nz under Publications, then Science and research.

DOC Science & Technical Publishing gratefully acknowledge the permissions we have been granted to reproduce copyright material in this report. We would particularly like to thank the following publishers: Blackwell Publishers (Figs 15 & 22); British Ecological Society (Figs 21 & 29); New Zealand Ecological Society (Figs 9 & 13); Ornithological Society of New Zealand (Figs 12 & A1.3); and Royal Society of New Zealand (Fig. 14). The sources of all material used in this publication are cited in the reference list. We have endeavoured to contact all copyright holders. However, if we have inadvertently reproduced material without obtaining copyright clearance, please accept our apologies; if you contact us, we would be happy to make relevant changes in any subsequent editions of this report.

ISSN 1172-6873
ISBN 0-478-14042-8

This report was prepared for publication by Science & Technical Publishing; editing by Sue Hallas and layout by Amanda Todd. Publication was approved by the Chief Scientist (Research, Development & Improvement Division), Department of Conservation, Wellington, New Zealand.

In the interest of forest conservation, we support paperless electronic publishing. When printing, recycled paper is used wherever possible.

CONTENTS

Designing science graphs for data analysis and presentation

The bad, the good and the better

Dave Kelly[1], Jaap Jasperse[2] and Ian Westbrooke[3]

[1] School of Biological Sciences, University of Canterbury, Private Bag 4800, Christchurch

[2] Research, Development & Improvement Division, Department of Conservation, PO Box 10 420, Wellington (current address: Secretariat of the Pacific Regional Environment Programme, PO Box 240, Apia, Samoa)

[3] Research, Development & Improvement Division, Department of Conservation, PO Box 13049, Christchurch

A B S T R A C T

Graphs use spatial arrangement on the page or screen to convey numerical information; they are often easier to interpret than repetitive numbers or complex tables. The assumption seems to be made that creating good graphs is easy and natural; yet many bad or sub-optimal graphs encountered in the literature disprove this. We aim to help you to produce good graphs, and to avoid falling into traps that common computer software packages seem to encourage. A scientific culture is one where good graphs, and innovative and specialised approaches, are valued. Hence we explain some relevant psychology behind the interpretation of graphs. We then review a variety of graph formats, including some less common ones. Their appropriate uses are explained, and suggestions are given for improving the visual impact of the message behind the data while reducing the distraction of non-essential graphical elements. We argue against the use of pie charts and most three-dimensional graphs, prefer horizontal to most vertical bar charts, and recommend using box plots and multipanel graphs for illustrating the distribution of complex data. The focus of this publication is on preparing graphs for written communications, but most principles apply equally to graphs used in oral presentations. The appendices illustrate how to create better graphs by manipulating some of the awkward default settings of Microsoft Excel (2002 version) and illustrate the S-PLUS programming language (both programs are currently available on the computer network of the New Zealand Department of Conservation).

Keywords: Science graphs, graphical displays, graphic methods, Excel, S-PLUS

Kelly, D.; Jasperse, J.; Westbrooke, I. 2005: Designing science graphs for data analysis and presentation: the bad, the good and the better. *Department of Conservation Technical Series 32*. Department of Conservation, Wellington. 68 p.

1. Introduction

1.1 AIMS AND TARGET AUDIENCE

Graphs (or charts, another less common word for the same thing) are visual representations of numerical or spatial information: everybody knows that, and most people find them easier to read than repetitive numbers or complicated tables. There is an assumption that creating good graphs is easy and natural, not needing much study. For example, one of us (DK) taught in a 48-lecture third-year university course in biological statistics, which was designed to prepare the students for thesis research. The course content was entirely on methods of analysis. Observing that many thesis students were drafting poor graphs, DK decided that four lectures (8% of the course) on graphing theory and practice would be well worthwhile. The other course teachers were not enthusiastic, apparently believing that good graphs did not need to be taught. Undeterred, DK gave the lectures (which ultimately formed the structure of this publication), and the students said that they found them extremely helpful. This is because human visual cognition and perception, although very powerful, are complex processes. It takes suitable approaches to communicate the relationships inherent in data.

Modern computing allows the ready production of graphs—both good and, all too often, bad. This publication aims to help you to produce clear, informative graphs, and to avoid falling into traps that common computer software packages seem to encourage. Further, we aim to help create a culture where good graphs, and innovative and specialised approaches, are valued.

This guide is primarily targeted at staff of the New Zealand Department of Conservation (DOC), with examples taken from New Zealand conservation publications where possible. However, we believe that the application of the ideas herein goes well beyond that audience. We trust that the application of the proposed recommendations will help science communicators, students and established scientists alike, to improve the ways of conveying that message lying behind data.

The focus of this publication is on preparing graphs for written communications, but most principles apply equally to graphs used in oral presentations (see section 6).

1.2 WHY USE GRAPHS?

A graph uses a spatial arrangement on the page (or screen) to convey numerical information. This has several advantages:

- Graphs can have very high information density, sometimes with no loss of data. By contrast, stating only the mean and standard deviation provides a summary that loses information about, for example, the number and position of outliers.
- Graphs allow rapid assimilation of the overall result.

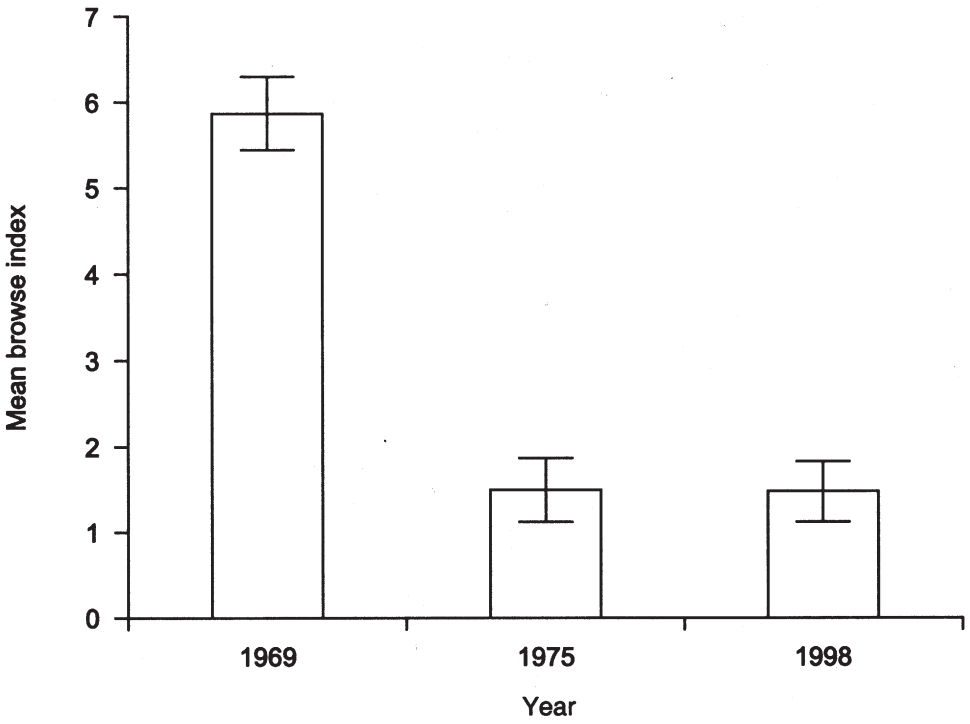
- The same graph can be viewed at multiple levels of detail (e.g. overall impression, close-up and exact location of several adjacent points).
- Graphs can clearly show complex relationships among multivariate data (in two, three, four, or even more dimensions).

For these reasons, good graphs are an important part of almost any experiment- or field-based thesis, research report, scientific paper or conference presentation.

However, graphs also have some disadvantages, especially if done badly:

- Graphs take up a lot of space if showing only a few data points. Hence they are best not used if there are only a few numbers to present.
- A graph may misrepresent data, for example by plotting regularly spaced bars for irregular data intervals (Fig. 1).
- A line may suggest interpolation between data points where none applies.
- It can be hard to read off exact numeric values, especially if badly chosen axis scales are used. If exact numeric values are required, a table is best.

Therefore, it is important to understand how to make the best of graphs. Note that it may not be necessary to display all available data in your graph. The key requirement is that the graph *honestly* and *accurately* represents the data you collected or want to discuss.



YEAR	INDEX (± SEM)
1969	6 ± 0.5
1975	1.5 ± 0.4
1998	1.5 ± 0.4

Figure 1. This simple bar graph misrepresents data by visually suggesting an equal interval between sampling dates: 6 and 23 years, respectively. The meaning of the error bars (standard error? 95% confidence interval?) was not explained in the accompanying caption, although it was in its source. The data are much more effectively and efficiently given in a tiny table (as shown to the left), or simply by the following sentence: 'The mean browse index (± SEM) was 6 (± 0.5) in 1969, 1.5 (± 0.4) in 1975, and 1.5 (± 0.4) in 1998'.

Original caption: Mean browse index on plots in the Murchison Mountains over the last 30 years ([from] Burrows et al. 1999).

1.3 TYPES OF DATA

Graphs are used to plot data, so it is useful to look at types of data first. There are two main types of variables: categorical / qualitative and numeric / quantitative.

Within these types are sub-categories that run along something of a continuum. Categories can be pure and unordered, e.g. species present at a sample site; complementary, e.g. male / female; or they can be in an ordered sequence, e.g. chick / juvenile / adult.

By contrast, numbers can have continuous values (e.g. for height or time) or discrete values (such as counts).

Quantitative variables can be grouped into categories, with some loss of information, but the reverse process is not generally possible. Table 1 illustrates this principle for the following dataset:

North Island: 284, 287, 296, 300, 302, 302, 304, 310, 310, 313, 315, 317, 319

South Island: 251, 264, 265, 265, 270, 271, 273, 273, 274, 275, 276, 276, 277, 277, 278, 279, 279, 280, 280, 280, 281, 282, 282, 283, 284, 284, 284, 284, 285, 285, 285, 285, 285, 287, 289, 289, 289, 291, 291, 292, 293, 301, 302, 304

TABLE 1. THE ABOVE DATASET GROUPED BY LENGTH CLASS.

REGION	LENGTH CLASS (mm)						
	251–260	261–270	271–280	281–290	291–300	301–310	311–320
North Island	0	0	0	2	2	5	4
South Island	1	4	15	17	4	3	0

1.4 GRAPHS VERSUS TABLES

The first decision to be made when presenting numeric data is when to use a graph and when to use a table. A table is an array of regularly spaced numerals or words. Again there is a continuum, which we can split into three types:

- Sentences listing a few numbers in the text (best for 1–5 numbers, where all are values of the same variable; e.g. as in the caption to Fig. 1).
- Text-tables, i.e. indented text lines (3–8 numbers, for one or two variables; often shown as a list of bulleted points such as this one).
- A full table that can cope with 5–100 numbers (see previous paragraph). Tabling over 100 items or so becomes unwieldy; if the items need to be included, it may be best to put them in an appendix.

Figure 1 illustrates the mistake of graphing simple data where presenting data in a table or as a sentence would have been much better. By contrast, a well-planned graph can, at a glance, give insight into many hundreds or thousands of bits of information (Fig. 2).

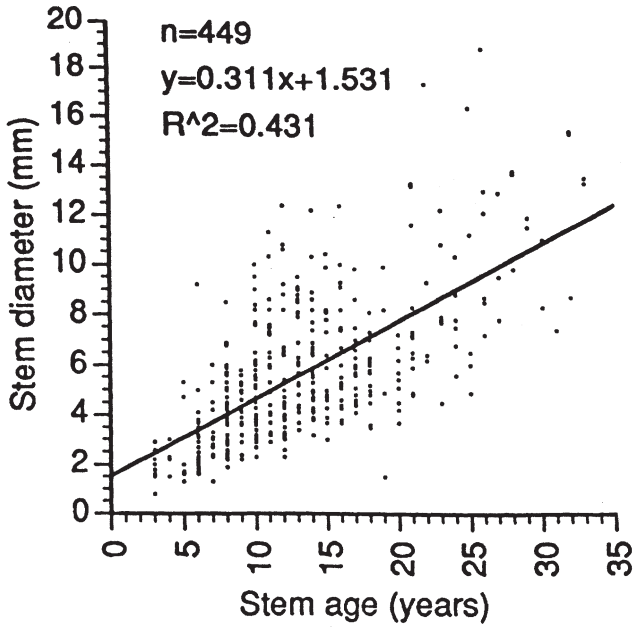


Figure 2. This graph summarises 449 data points. When drawing a regression line, it is best to show also the data points on which the regression is based (see section 4.7.2). Note that the graph could have been improved by having fewer ticks on the *x*-axis, the *x*-axis values reading horizontally, and it may have been appropriate to constrain the line to go through the origin.

Original caption: Simple regressions between stem age[, stem length,] and stem diameter of heather sampled in 64 plots on the north-western ringplain of Tongariro National Park.

Graphs and tables have different uses. In an oral presentation you would usually emphasise graphs, which get the main idea across more rapidly. In a thesis or research report, the detail, precision and archival value of tables may be more important. In a published paper, a mix of both will be used for different sets of data. It is considered bad practice to present the same data in two modes unless a very good reason warrants using the extra space. One instance where both may be justified is when highlighting a difference between the two modes of presentation (Fig. 1). In addition, occasionally it is advisable to have a graph in the main text showing the key points, and a full detailed table in an appendix giving the exact values for archival purposes.

1.5 HISTORY OF GRAPHS

Have graphs always existed in scientific works? The answer is no. Despite a wealth of classic literature describing the world around us, graphs of abstract empirical data were rarely published or non-existent before the 18th century (Tufte 1983). The diagrams that did exist represented physical space—maps, or maps of the heavens (Fig. 3).

By the late 1700s, with the rise of industry and trade, large quantities of economic and social data were accumulating that needed to be studied. Some of the earliest known data graphs were those of William Playfair (Fig. 4). This represented a huge intellectual leap—the representation of abstract numbers in physical dimensions, taking advantage of humans' highly developed visual processing abilities.

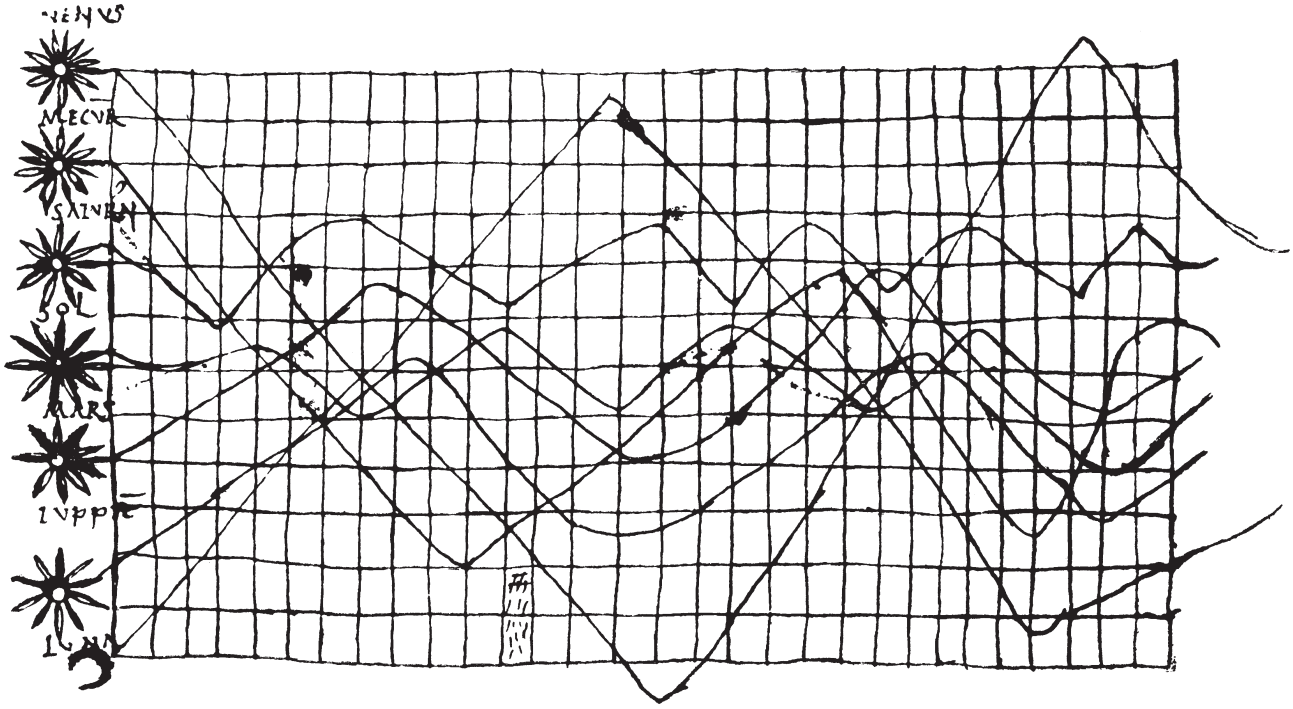


Figure 3. Illustration of planetary orbits, c. 950. This is the earliest known attempt to show changing values graphically.

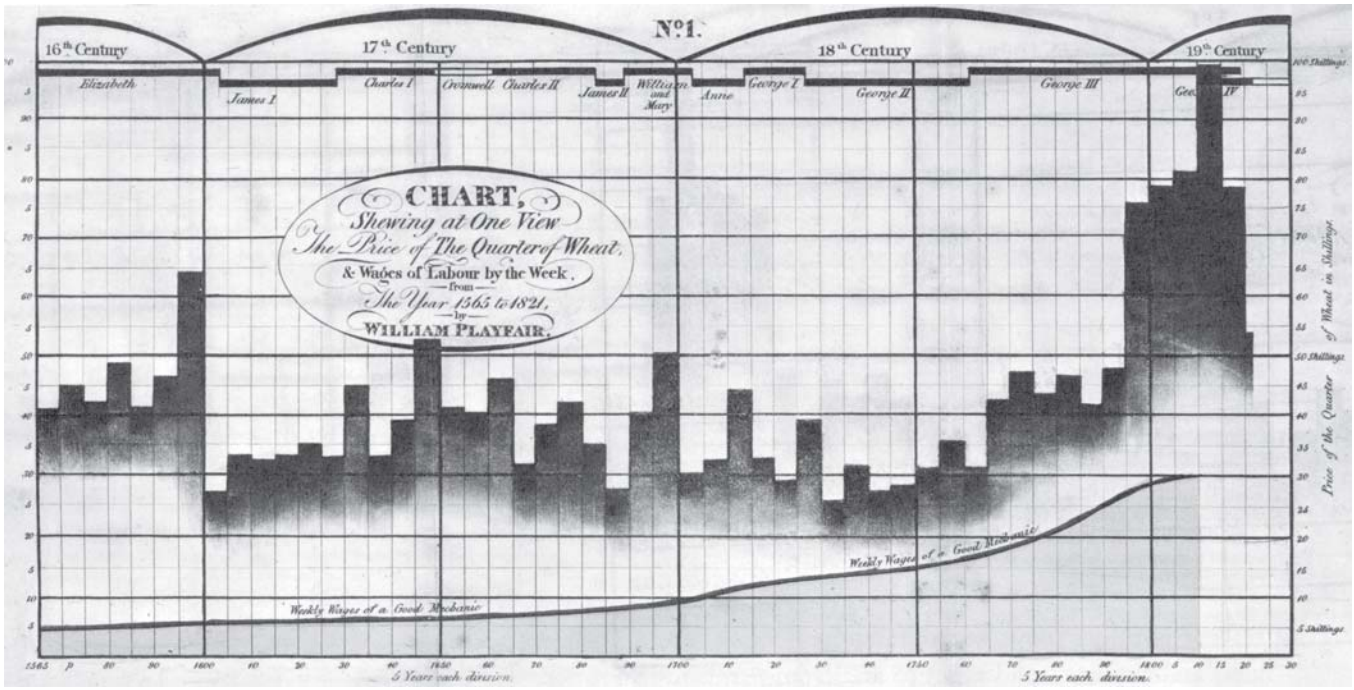


Figure 4. Playfair graph showing 200 years of wages and the price of wheat.

1.6 THE WAY FORWARD

The biggest development in the last 30–40 years has been the advent of computing, which allows both the storage of vast amounts of data, and the easy creation and modification of graphs—both good and, all too often, bad. The present publication aims to help you to produce good graphs, and to avoid falling into common traps. This guide outlines some key principles derived from two excellent books: William Cleveland's 'The Elements of Graphing Data' (1994, 2nd edition) and Edward Tufte's 'The Visual Display of Quantitative Information' (1983). Either makes a very good read. The latter is an entertaining work, beautifully laid out in colour, and in a coffee-table format. Another excellent brief guide is 'Editing Science Graphs' published by the Council of Biology Editors (Peterson 1999), and the website 'The Best and Worst of Statistical Graphics' (www.math.yorku.ca/SCS/Gallery/, viewed 23 March 2005) is also useful. A new book 'Creating More Effective Graphs' (Robbins 2005) provides a very readable guide to creating graphs based on the principles laid out by Cleveland (1994) and Tufte (1983).

The present text is based on a lecture handout from the University of Canterbury (DK), DOC Science Publishing editorial guidelines (JJ), and material prepared for a series of Graphs workshops in DOC (IW and JJ).

We include appendices that will help you to create publication-quality graphs by manipulating some of the awkward default settings of Microsoft Excel (2002 version) and by using the S-PLUS programming language. Both programs are available on the computer network of the New Zealand Department of Conservation.

2. Principles of graphing

Before even planning a graph of your data, you need to consider several general points. Some of these are derived directly from psychological principles, but most are plain common sense.

2.1 ASSUMPTIONS ABOUT YOUR TARGET AUDIENCE

Assume your audience is intelligent. Expect, in particular, that they will read and understand axis labels: hence there is no need to always have zero marked on each axis if the data all span a short range a long way from zero. However, don't overestimate your audience either: what may be a patently obvious relationship in your graph may need careful explanation in the caption (more about captions in section 4).

If the graph is to be published, assume that it will be reproduced at the smallest possible size to convey its information. This limits your choice of detail, font

size, line weights, etc. (See section 4 for more on such graphical elements.) For oral presentations, assume that your audience may have difficulty reading and understanding lots of small print or complex data (section 6).

2.2 WE RECOMMEND: CLEAR VISION, NO COLOUR AT FIRST

Our recommendations for creating clear graphs are:

- **Make the data stand out.** It is the most important part of the graph. Anything that distracts from data is undesirable.
- **Use clearly visible symbols**, which are more noticeable than any other text on the graph, such as axis labels.
- **Reduce clutter** on the graph. For example, use relatively few tick marks: 4–6 per axis is usually sufficient.
- **Labels on the graph should be clearly offset from the data** or even outside the axes, to ensure that they are not confused with the data; appropriate abbreviations can help to keep labels short.
- **Keep notes and explanations outside the data region** where possible.
- **Overlapping symbols or lines must be visually separable.**
- **Allow for reduction and reproduction**, since most printed graphs will be reduced and photocopied at some stage: sometimes through several generations! If you can reduce a graph to 0.71 twice (i.e. reduce by 50%) and it is still readable, it will suit most presentation purposes.
- **Try to design your graph without the use of colour.** If it reproduces well in black and white it will be able to be reproduced in any medium. For example, while a colour graph may look impressive on a web page, pdf files are likely to be printed to a monochrome printer, or photocopied, which may result in lost detail. In some situations, you may add colour to your graph later for emphasis (e.g. for an oral presentation).

2.3 PERCEPTION AND ACCURACY

There are several features of human perception that affect the relative accuracy with which different graph types can be read. Ignoring these principles may lead to incorrect perception and incorrect decoding of the data by the end user.

2.3.1 Weber's Law

According to Weber's Law, the probability that an observer can detect an increment of a certain size in a line depends on the *percentage increase of the increment, not its absolute size* (Cleveland 1993). Figure 5 illustrates the principle.

Therefore, based on Weber's law, you should arrange graphs to show data with the largest relative changes possible. That means you can leave zero off the axis scale unless the numbers are close to it: see Fig. 6.

Figure 5A. Can you tell the difference in length between: 1. The black parts of both bars? 2. The white parts? Weber's law predicts that it is much easier to tell the difference between the white areas because their percentage difference is bigger—even though both the black and the white pairs differ by the same absolute value.

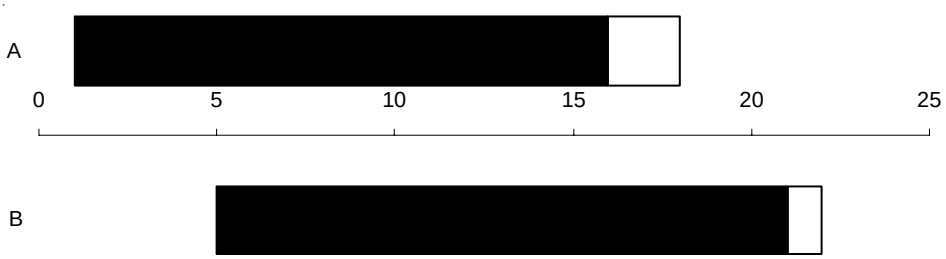


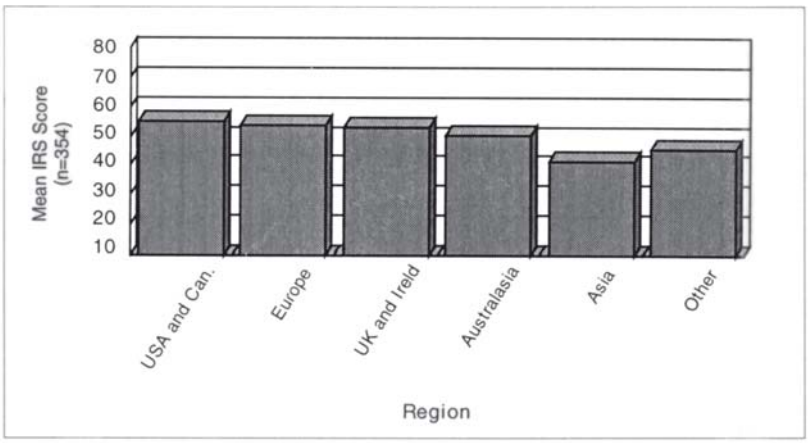
Figure 5B. Adding a reference axis to Fig. 5A allows the comparison of each bar with the tick marks. Instead of trying to compare the lengths of whole black bars directly, you can compare smaller segments of the bars against the tick-mark segments.

Figure 6. Mean scores for individual responsibility by region, from a survey regarding hazard warning signs of visitors to Franz Josef and Fox Glaciers: original (A) and regraphed (B). The differences in bar lengths in the original are difficult to distinguish, made all the more difficult by false 3-dimensional representation. Values in the original dataset were between 11 and 77, so the axes, and the length of the bars, are slightly misleading. The new version below highlights the relative values for the different groups and gives a much tidier appearance by using horizontal, not oblique type.

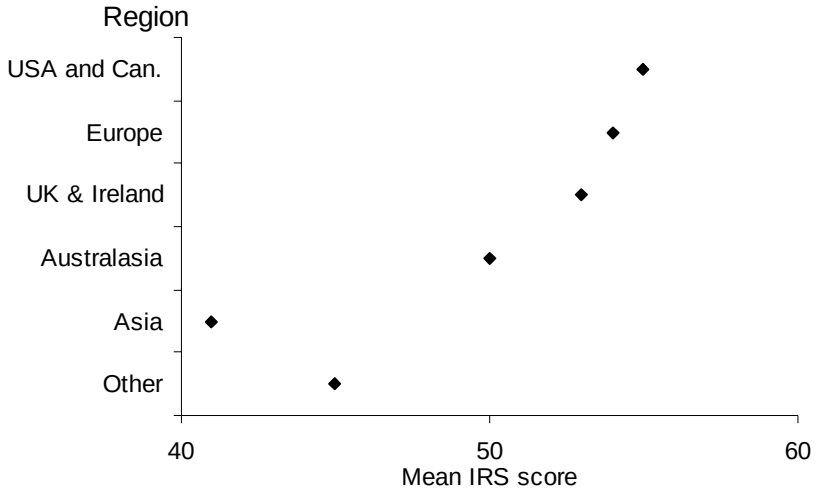
See Box 2, section 3.6.1 for guidelines on how to change the graph.

Original caption for A: Mean scores for individual responsibility by region.

A



B



2.3.2 Stevens' Law

Our perceptions of shapes and sizes are not always accurate, and our brains can be misled by certain features. Psychologists have found a general relationship between the perceived magnitude of a stimulus and how it relates to the actual magnitude (Stevens 1957).

Stevens' empirical law states that $P(x) = Cx^a$, where x is the actual magnitude of the stimulus, $P(x)$ is the perceived magnitude, and C is a constant of proportionality. Note the power relationship between magnitude and perceived magnitude, with the value of this power (a) varying with the task: for length, a is usually in the range 0.9–1.1; for area, a is usually 0.6–0.9; and for volume, a is usually 0.5–0.8. So, lengths are typically judged more accurately than either area or volume (the latter being judged least accurately).

Aspects of perception other than length, area and volume are also biased.

Angles: We tend to underestimate acute (sharp) angles and overestimate obtuse (wide) angles.

Slopes: Our eyes are affected by the *angle of the line to the horizontal* rather than its slope. (Slope or gradient is the vertical rise per unit of horizontal distance.) If asked to estimate relative slopes, we usually judge the ratio of the angles, meaning that slope is judged with considerable distortion. For example, a slope or gradient of 1 is equivalent to an angle of 45° ($y = 1x$) but a gradient of 4 has an angle of 76° ($y = 4x$). Hence, the slope increased fourfold, while the angle increased by only 69%.

2.3.3 Cleveland's accuracy of decoding

Graphs communicate quantities best if they use the methods of presentation that people perceive most accurately, and which allow the viewer to assess the relationships between the values represented without distortion.

Cleveland & McGill (1985) provide a hierarchy of features that promote accuracy of decoding:

1st (best)	Position on a common scale / axis
2nd	Position on identical non-aligned scales / axes
3rd	Length
4th equal	Angle
	Slope
6th	Area
7th equal	Volume
	Density
	Colour saturation
10th (worst)	Colour hue

For example, the same data may be much more readily interpreted as position on a single scale than as angles in a pie diagram. This is one reason why pie diagrams are best avoided (see section 3.1).

In honest illustrations, you also need to avoid distortion, i.e. when visual representation is not equal to the actual numeric representations. The most common graphical distortion is using area or volume to show change in length (Fig. 7).

This leads to another rule: do not use more dimensions in the graphical representation than are present in the data. If a series, for example level of funding, is one-dimensional (1D) then it should not be shown as area (two-dimensional, 2D) on a graph. By contrast, data on leaf areas (2D) should be represented in one dimension if total area is of key interest, or by showing the length and breadth on separate axes if they are of separate interest. Volume data (three-dimensional, 3D) could be shown as volume, area or length—in the last instance, possibly on a logarithmic scale (but the caveats above about biased estimation of areas and volumes compared to length apply here too).

The rules of visual perception apply primarily to representing *quantitative* data. These data should be represented by methods with high perceptual decoding accuracy. *Qualitative* measures, which do not require quantitative decoding, can be represented using methods that are lower in the accuracy hierarchy, like shading, or non-quantitative representations, such as symbol shape or line type.

Box 1 illustrates an exercise in the DOC Graphs workshop series (held in 2003) in which participants were asked to rank the best ways of representing data. Although some personal preferences showed up in the middle section, rankings of the extremes were quite clear-cut. Also illustrated here is how the exercise was analysed using box plots (discussed in section 3.4).

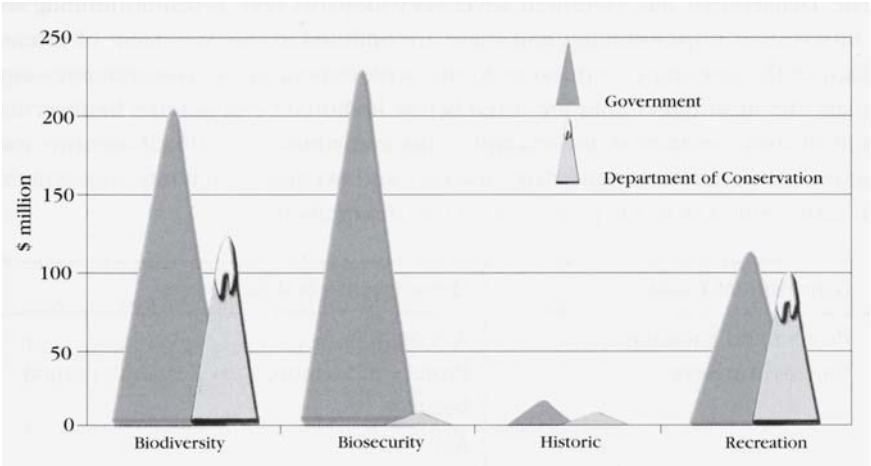
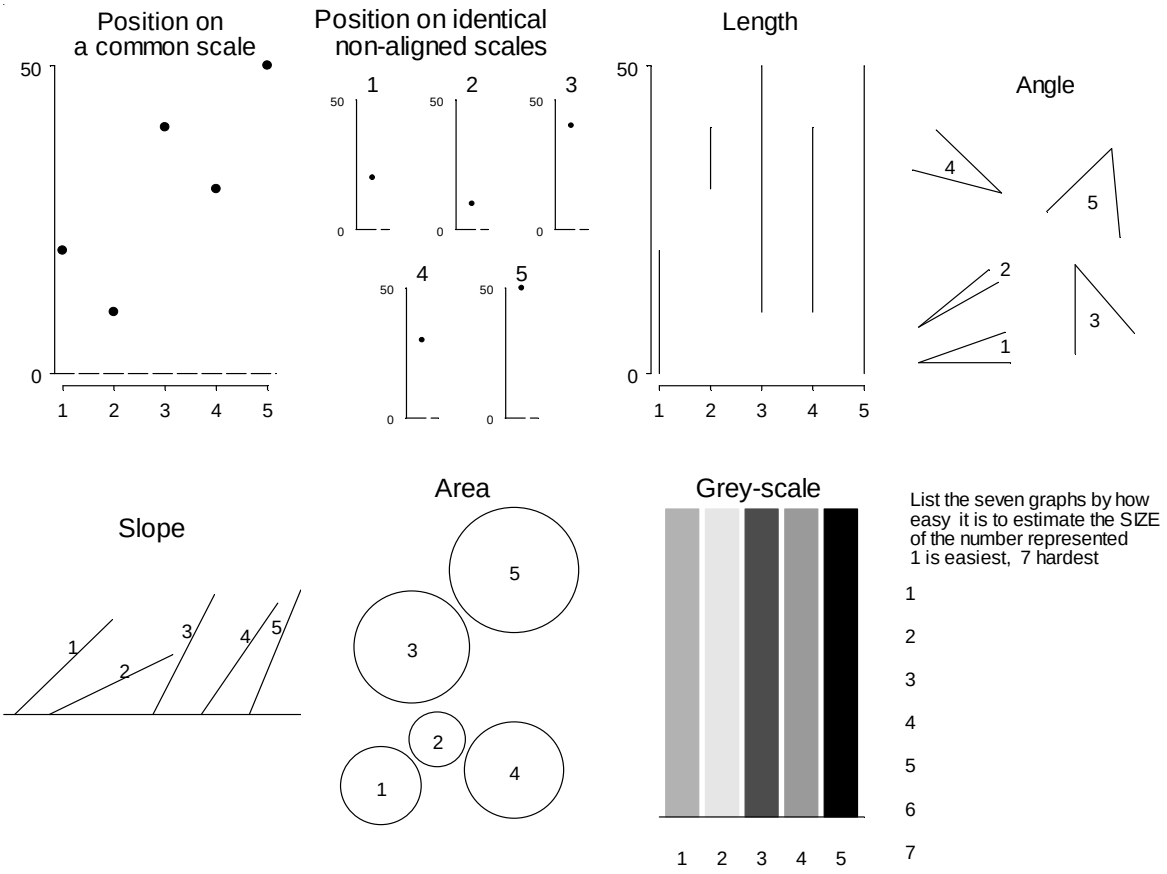


Figure 7. Although DOC receives more than half ($125 million) of the Biodiversity funding spent by Government (c. $200 million), it appears much less on the graph by representing the linear dollar variable as a triangle (decoded as area: 2-dimensional) or even as a mountain (decoded as volume of cone: 3-dimensional).

Original caption: Department funding as a proportion of Government spending.

Box 1: DOC Graphs Workshop exercise

As an exercise, participants carried out an informal assessment of Cleveland's recommended order of accuracy in graphical perception (Cleveland & McGill 1985), during a series of workshops in DOC in 2003. Colour and volume were excluded as too difficult to reproduce readily. An example of the exercise given out at the workshops is shown in the composite figure below, although the format varied between workshops. Participants, in groups of 2 to 4, were asked to order the seven graph types by how easy it was to estimate the size of the numbers represented. The results of the average ranking at each workshop are shown opposite.



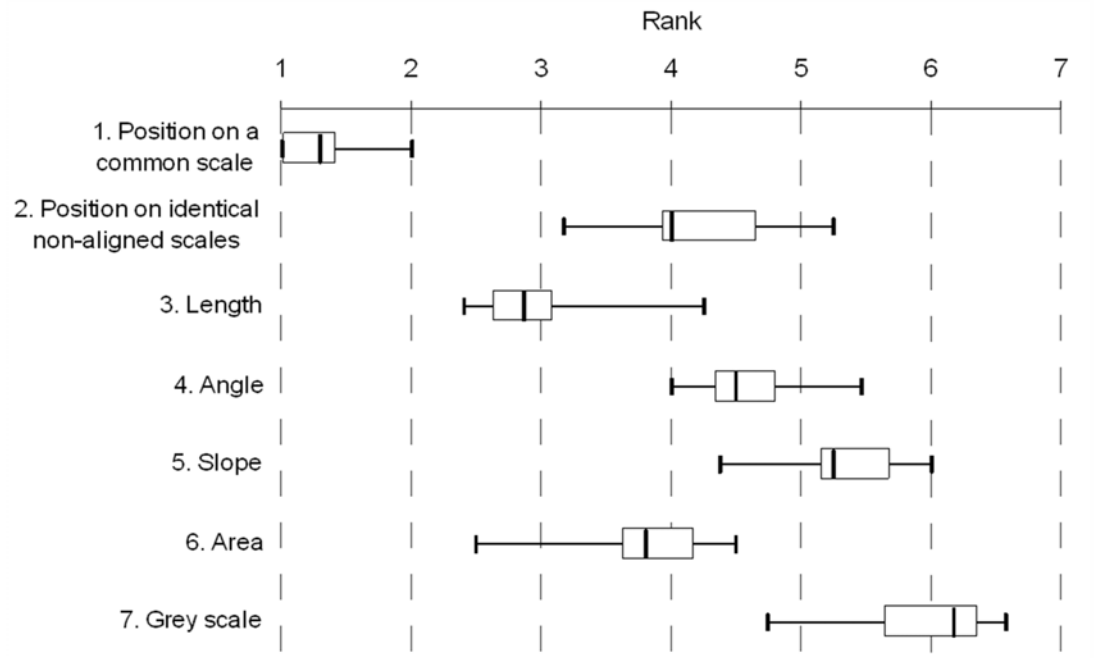
Example of the exercise given, with varying formats, at a series of graph workshops in DOC in 2003. The seven graph panels are ordered from top right across, and then down, in the order advocated by Cleveland (Cleveland & McGill 1985). The bottom right panel gives instructions on ranking to participants. All panels attempt to represent, in order, the values 20, 10, 40, 30, 50.

Continued on next page

Box 1—continued

This exercise worked well as a teaching tool about ways of representing quantitative data. However, the results are limited by the lack of scale for comparison except in the first three panels, although participants at most workshops were advised that each panel attempted to represent the same set of numbers. Further, it is important to note that the exercise assesses the respondents' perceptions of how easy it to estimate size of the number represented, rather than directly the accuracy of estimation.

The key outcome of the test was that position on a common scale was universally ranked the highest at each workshop. The consistency in this ranking was very clear—and underlines the importance of using position on a common scale as the preferred method for representing quantitative variables. The participants' rankings generally followed Cleveland's ranking, but with some discrepancies, which may be due to the limitations of the exercise.



Box-and-whisker plot of participants' preferences from the visual perception exercise in the figure opposite. The graph types are listed in order of decreasing accuracy of decoding according to Cleveland & McGill (1985). Each box shows the upper and lower quartile, the central bar represents the median (midpoint), while the whiskers show the minimum and maximum for each category. The ranks displayed are the average for that type of graph at each of 11 DOC workshops on 'Using Graphs to Analyse and Present Data', held in April to July 2003.

3. Types of graph

Having explained the principles underlying perception, we can apply these to the various types of graph.

3.1 PIE GRAPH (UNIVARIATE)

Pie graphs, pie charts or pie diagrams have no right to exist in science: the job they do can always be done much better in other ways. They are generally used for data with one numeric and one categorical variable, and display only a few data but take up a lot of space. Moreover, they represent the information as angles, which is low on the scale of decoding accuracy (section 2.3.3). Even worse are 'mock 3D' pies (Fig. 8A), which add insult (distortion) to injury (inaccuracy); they violate the stated rule that the number of data dimensions in a graph should not exceed the number of dimensions in the source data.

Generally speaking, pie-graph data are much better presented in a small table or as horizontal bar graphs (Fig. 8B). Note that many pie-graph designers admit the limitations of pies by adding numeric values and / or percentages to the individual pie segments, thus creating clutter. Pie graphs also often require a detailed key, which more often than not creates extra confusion: colours or shadings are often too similar to clearly identify the segment to which they belong. Generally, a key 'starts at 12 o'clock' and subsequent categories are then listed in clockwise order… but not many readers know that!

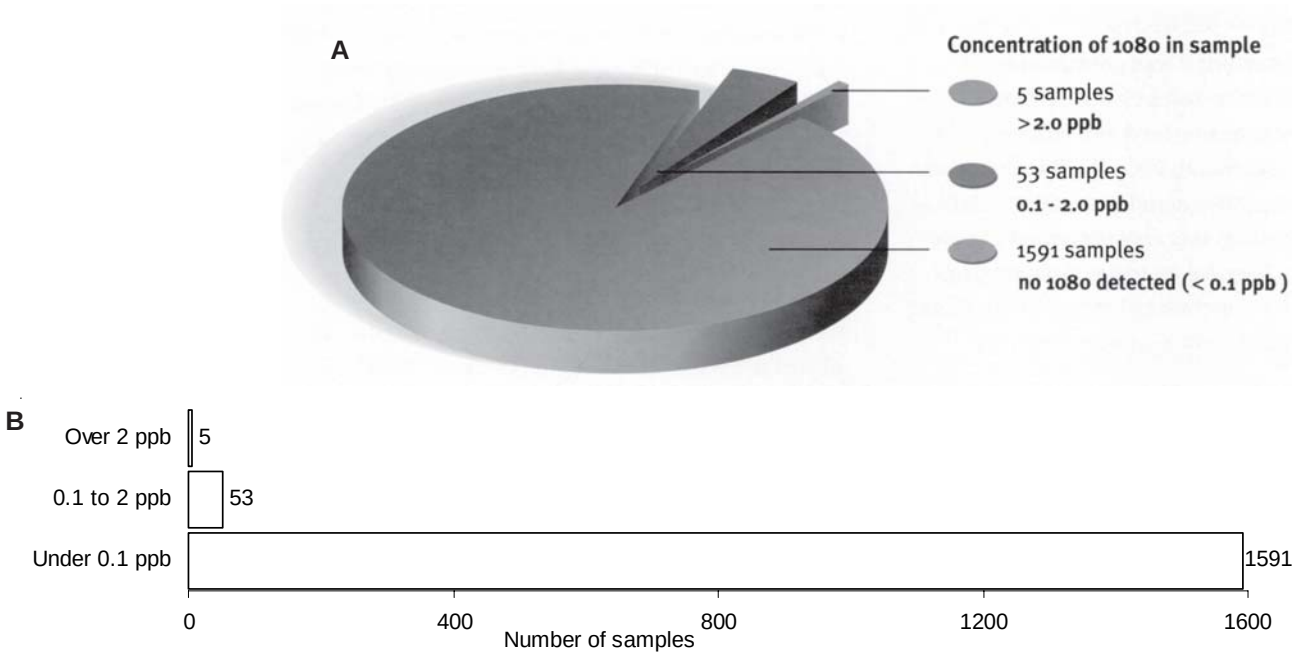


Figure 8. A classic example of a space-wasting pie graph (A), which still requires a table to explain its values. In B, the data from A, particularly the relative sizes of the samples, are much more accurately represented by horizontal bars.

Original caption to A: Results of water monitoring after aerial 1080 operations (1991–2003).

Bigwood & Spore (2003) agree that 'despite their mass popularity, pie charts do not communicate well' (but these authors 'offer some advice on designing and presenting them' in order to 'use them as effectively as possible').

You sometimes see linked pie graphs, where there are several in a row. Instead, if you have three categories that each add to 100%, scored at a number of different sites or samples, consider using a triangular diagram, sometimes called a 'ternary plot'. An example is given in Fig. 9.

In most instances it may be best to represent data from linked pies as a series of column graphs where each column adds up to 100% (Fig. 10). The columns represent the data as length, not angle, and you can run your eye across the values for each category more easily than if they are in pies. Column graphs (bar charts) are discussed in much more detail below.

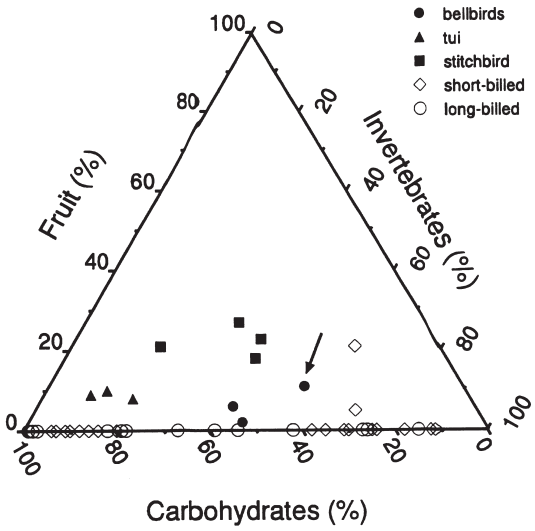
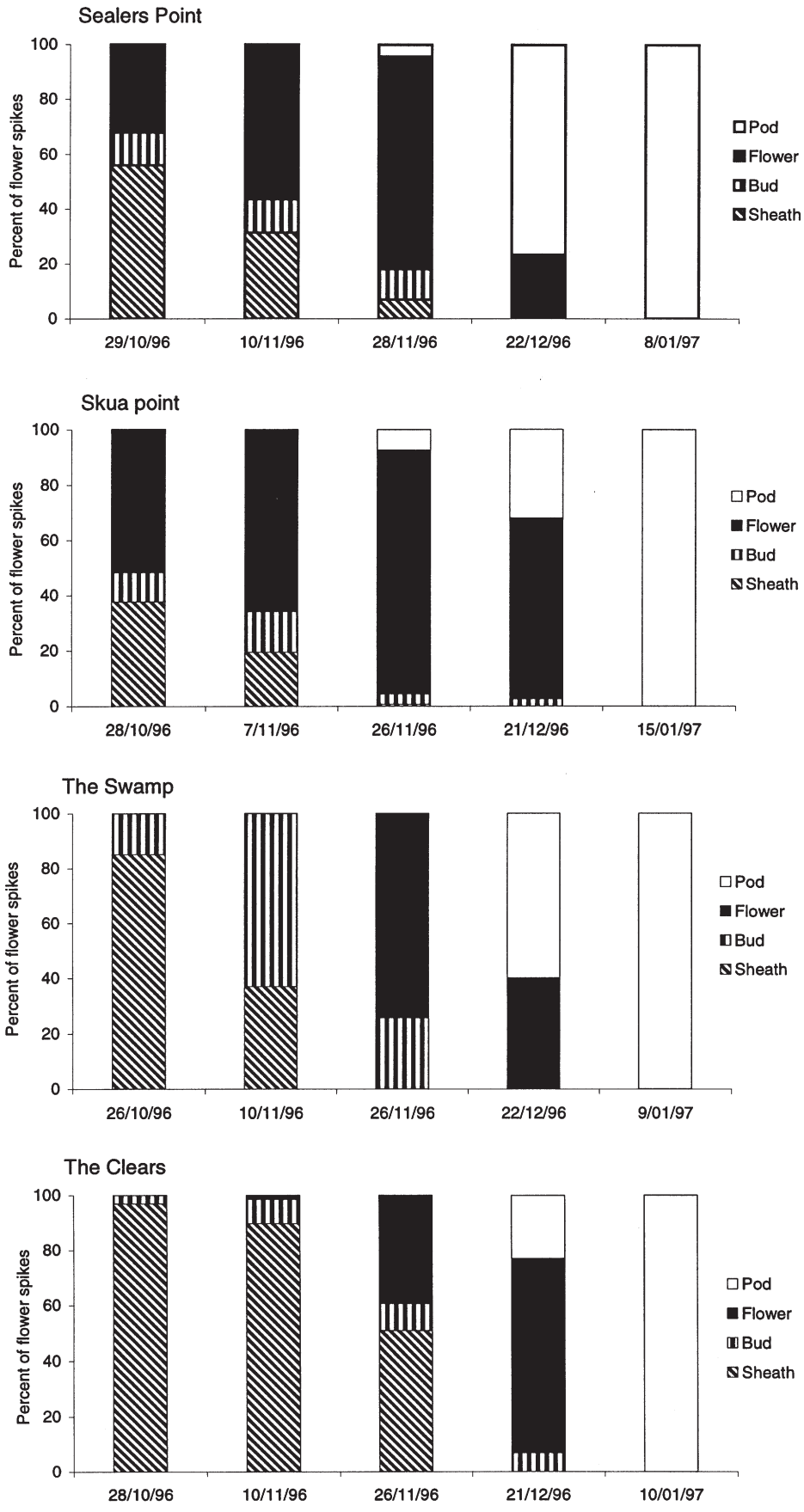


Figure 9. A ternary (triangular) graph, useful for three variables that sum to 100%. These graphs can be difficult to interpret on first encounter. It is easily grasped that the three corners represent 100% of one of the variables and 0% of the other two. In contrast, it is much less obvious that the point dead centre does not represent 50, 50, 50 for a sum of 150%. The reason it actually represents 33, 33, 33 to sum to 100% is that the gridlines run on different angles for the three axes. The left axis (in this case Fruit) gridlines run horizontally; the right axis (Invertebrates) gridlines slope downwards to the left, parallel to the Fruit axis line; and the lower axis (Carbohydrates) gridlines slope upwards to the left, parallel with the Invertebrates axis line. It helps to indicate this if (a) the axis tick mark labels are angled, as here; and (b) the graphs use long angled tick marks (in this case, they are angled, but perhaps too short).

Original caption: Annual mean diet composition of different New Zealand (solid symbols) and Australian (open symbols) Meliphagidae species. Each point on the graph represents the annual mean diet for a species from a single study or site, comprised of the annual mean percentages of the three major Meliphagidae food groups: invertebrates, fruit, and carbohydrates (nectar, honeydew, lerp and manna). Australian species are classified as long-billed or short-billed to distinguish between the two main feeding guilds in the Australian Meliphagidae. The Craigieburn bellbird data are marked with an arrow.

Figure 10. Example of a good 4-by-5 grid of split bars. However, the fills used in the bars run some risk of Moiré effects, see section 4.7.4. Also, the duplication of vertical labels and keys is unnecessary, and the *y*-axis label should read 'Percentage of flower spikes'.

Original caption: Flax flowering at selected plots on Rangatira Island.



3.2 VERTICAL AND HORIZONTAL BAR CHARTS / DOT GRAPHS

Bar graphs can be very clear, but they are overused and there are often better alternatives. Bar graphs tend to have a fairly low information density. They are easy to create using computer software packages such as Microsoft Excel; however, Excel tends to produce graphs that are not readily publishable to a high standard. Appendix 1 describes how such default graphs can be modified to meet science journal publication needs. More than 50% of graphs in DOC Science publications of 2002/03 were bar graphs, mostly produced in Excel—hence our concern with improving them (Appendix 1).

3.2.1 Notes on terminology

What most people call a bar chart has vertical bars, in distinct categories usually separated by white space. Microsoft products, however, call this type a 'column chart'. They are the most commonly seen graph type in all sorts of publications. The vertical arrangement often forces labels on the *x*-axis to be squashed or turned (up to 90°), which makes the axis hard to read, and looks ugly.

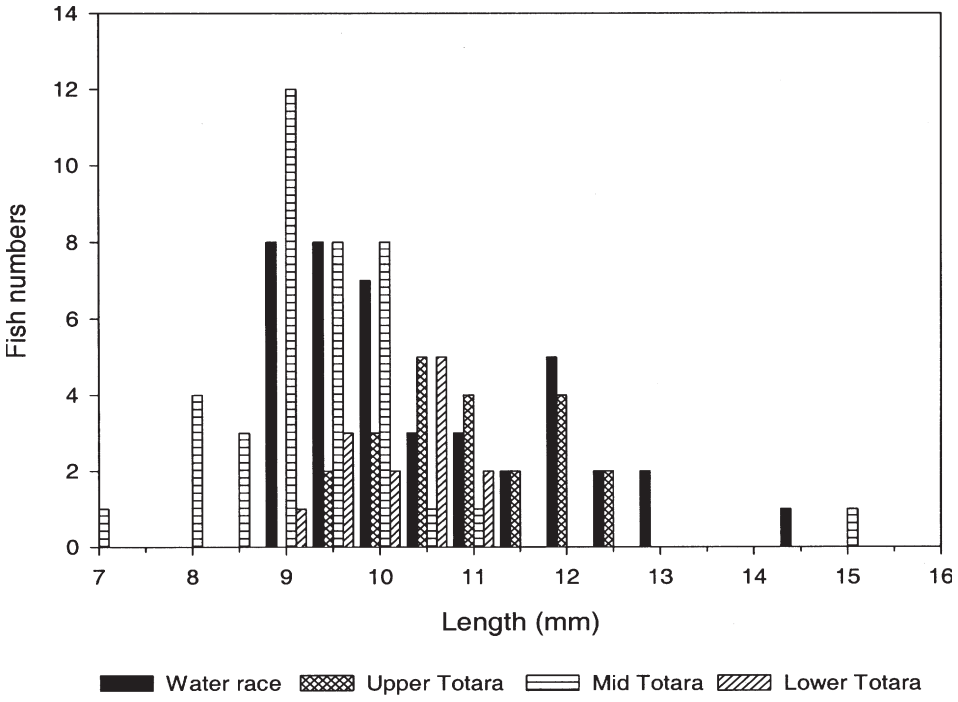
Horizontal bar graphs (bar charts in Microsoft lingo) are especially suitable for wordy categories, avoiding the need for vertical text labels or abbreviations. In this work, we will add the words 'horizontal' and 'vertical' to 'bar graph' where required to avoid confusion. The terms 'graph' and 'chart' appear to be used interchangeably.

Related to the vertical bar graphs are histograms, which display continuous variables with columns touching each other: more about these in section 3.3.

3.2.2 Vertical bar graph

A vertical bar graph displays one numeric variable, on the *y*-axis, against a categorical variable on the *x*-axis (site, species name, etc.). Such bars have a very low information density, and they implicitly present information as the length of the bar. This puts them low on the scale of decoding accuracy, and requires that you include zero on the *y*-axis. For bigger values, this can compromise resolution, and where the *y*-axis has a log scale, this is impossible—which poses a conundrum for good graph design. The information density is slightly higher if you add error bars (Fig. 1), use stacked bars (Fig. 10) or multiple bars (Fig. 11). When full dates do not fit on the *x*-axis, it may be best to abbreviate to the sequence of first letter of the months (i.e. JFMAMJJASOND) or just the day number, and show month and year in the caption.

Figure 11. Multiple vertical bars are not a very good way of presenting data accurately. It is difficult to gain a view of the distribution for each location because the bars are intermingled. Also, in this case, the *x*-axis should show the subdivisions of the length used for the counts. The presentation of an apparently continuous length variable creates distortion and does not clearly reveal that lengths were measured in intervals of 0.5 mm. Fig. 12A shows a more effective example (where the *x*-axis represents categories instead of a continuous variable), but even so better alternatives are available (Figs 12B & 13B).



Original caption: Length frequency distribution of larval galaxiids collected from four sites in Totara Creek on 5 December 1998.

3.2.3 Stacked and multiple bar graphs

Stacked bars (several values one above the other making a single column per category on the *x*-axis) are the general form of the column graph we recommend to replace linked pies (see Fig. 10). Multiple-bar graphs (several variables plotted as adjacent columns next to each category on the *x*-axis) can become hard to read (see Figs 11–13). The bars pile up together and discrimination is difficult, especially in black-and-white representation, where you must use stripes or stipples and a key to identify the various bars. Colour can make a multiple-bar graph easier to discriminate, but when the graph is photocopied in black and white it will be hard or impossible to interpret.

How to improve such graphs? If the *x*-axis is actually a continuous variable (e.g. length in mm or time in years) rather than a categorical one, then draw a standard *x*–*y* graph instead (see section 3.5). The use of different symbols and / or lines allows more than one series to be displayed readily. If you have a complex multiple-bar graph, data may be better represented as a table, where readers can run their eye down each column easily, or as multiple panels (multipanels) in the graph, often with identical axes, depending on the context (see section 3.7 and Fig. 12).

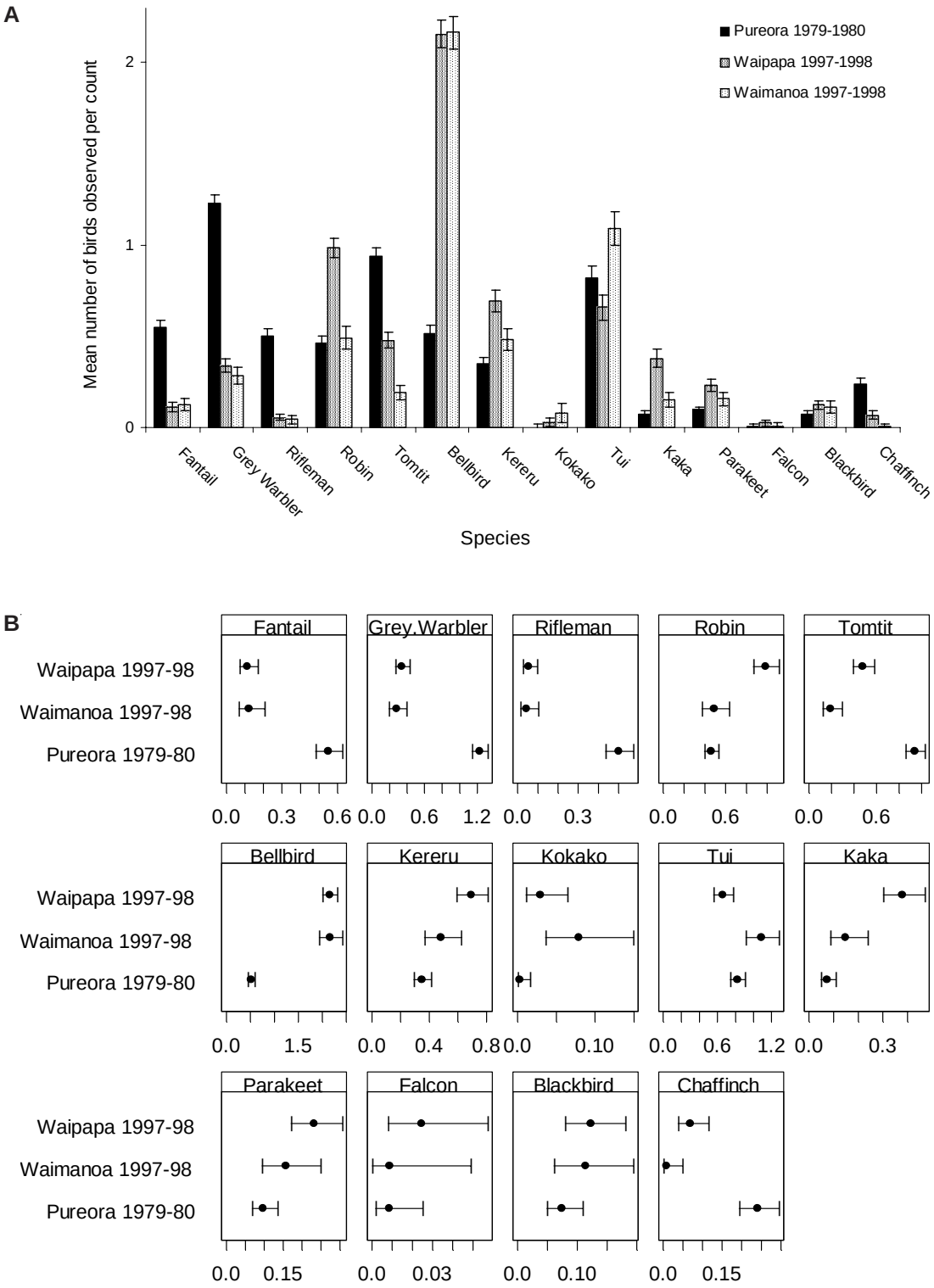


Figure 12. Comparison of a bar chart (A) with a dot chart finally designed for publication (B). The values represent average counts of birds in five-minute observation periods, with 95% confidence intervals. The use of varying scales for different panels is noted in the caption in the original.

Original caption to B: Winter (May and June) mean bird conspicuousness in two studies in Pureora Forest Park, with confidence intervals based on the assumption that the counts have a Poisson distribution. Note that the panels for different birds have varying scales.

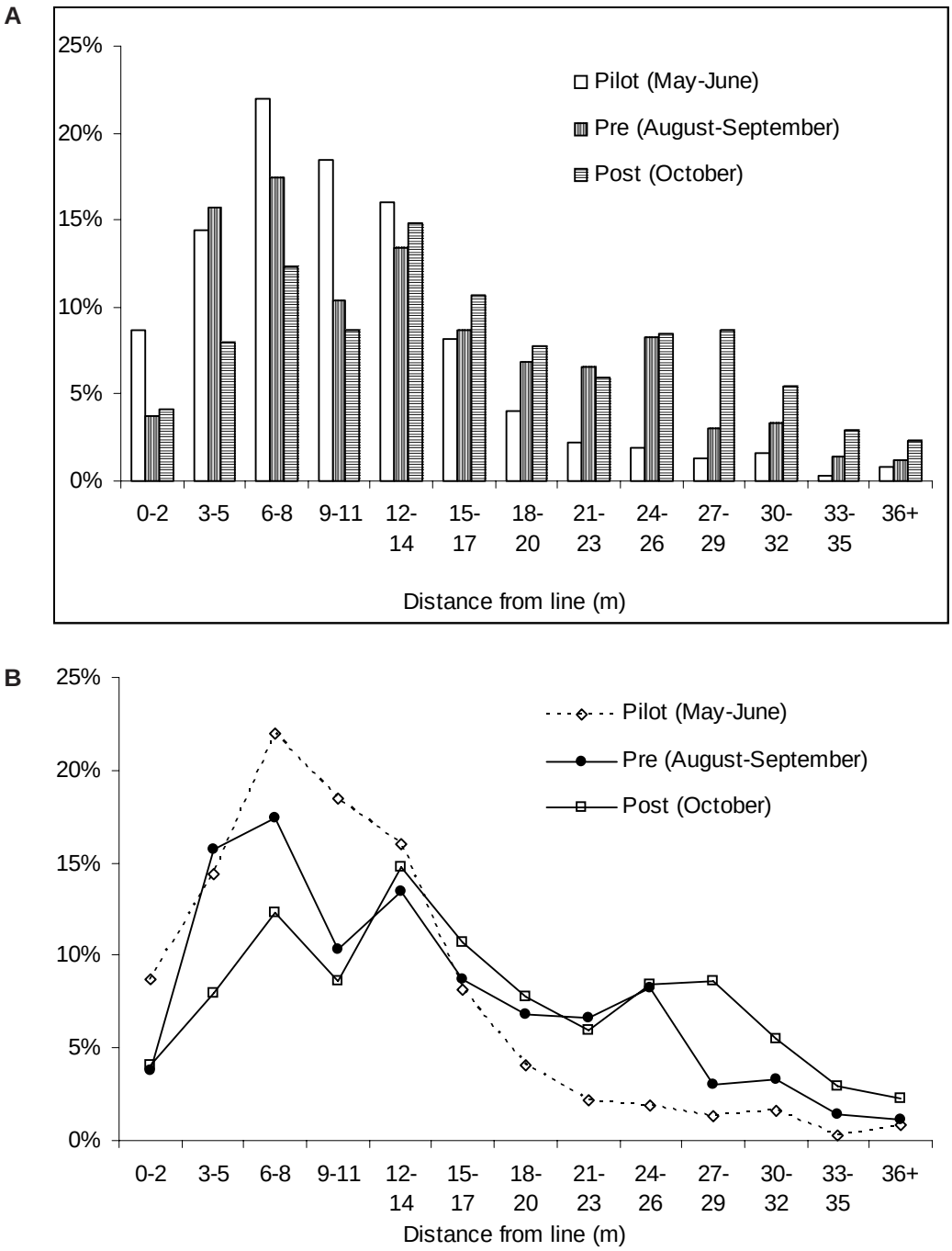


Figure 13. A bar chart (histogram) (A) and a relative frequency polygon (B—as published), based on the same data. Comparing several groups in one histogram destroys the continuity of the *x*-axis. The frequency polygon uses lines joining points to represent a distribution: it can show a modest number of related distributions clearly on one chart.

Original caption to B: Percentage of distance sampling observations in 3-metre distance classes, for three phases of the study: pilot (May-June 2001, *n* = 368), pre-treatment (August-September 2001, *n* = 439), and post-treatment (October 2001, *n* = 425).

3.2.4 Horizontal bar graph

When category labels are too long to reproduce in horizontal type on the *x*-axis of a vertical bar graph, it is better to use a horizontal chart rather than print oblique or vertical type. This graph shape is particularly well suited to categorical data with long names: results of questionnaires, etc. Figure 8B is an example.

3.2.5 Dot chart

A dot chart (dot plot) is a special type of horizontal bar chart, developed by Cleveland (Cleveland 1993). It uses a minimum of ink to optimum effect (which, according to Tufte (1983), indicates good design). The other strength of this design is that by using a dot it is clearly indicating the value by the position of the dot relative to the *y*-axis scale, rather than by the length of the bar, as in a normal bar chart. It may be better in technical works (Fig. 12B), although some authors and readers appear to have difficulty in letting go of the more familiar bars (Fig. 12A).

Dot charts feature:

- Horizontal arrangement (with plenty of room for long labels); usually categorical data.
- A dot marking the data point, not a bar.
- Optionally, light dots on left only (if zero baseline) or, more usually, all the way across to link the dot to its label.

3.3 HISTOGRAM AND FREQUENCY POLYGON

A histogram always has two numeric axes, but the *x*-axis is always a continuous variable, divided into an arbitrary number of categories—usually to show distributions. When drawn for a single variable, the bars of continuous variables by convention touch each other (see Appendix 1); bars for true categorical variables are better presented with spaces between them. Histograms have rather few, fairly specialised uses. They are fine for showing distributions within a large dataset. However, comparing several groups destroys the continuity of the *x*-axis (Fig. 13A), and there is some loss of information compared with showing the scatter, or a cumulative frequency curve, both of which can show the entire dataset.

A frequency polygon (Fig. 13B) is like a histogram, but uses lines joining points to represent a distribution, instead of bars. Its big advantage is that it can show a modest number of related distributions clearly on one chart, using different symbols and / or lines. It has also been shown to be technically superior (Scott 1992).

Histograms and frequency polygons can be based on numbers, or on relative frequencies (relative frequency = the frequency at each point in each category divided by the total for that category), depending on which is more useful.

According to Cleveland (1994), box-and-whisker plots and quantile plots are often better alternatives for assessing distributions. We discuss box-and-whisker plots in section 3.4, but readers are referred to Cleveland (1994: 136) for more on quantile plots.

3.4 BOX-AND-WHISKER PLOT (BOX PLOT)

The box-and-whisker plot (or box plot) is an excellent exploratory graph for summarising the distribution of one continuous variable, possibly broken up into several categories. It is very useful for picking up key aspects of the distribution of samples of modest to very large size.

The most common text-based summary of data involves either just the mean, or the mean and standard deviation, i.e. only a one- or two-number summary. While the mean and standard deviation are very good at summarising data with a normal distribution, most real datasets are not so well behaved.

By contrast, a simple box plot is based around a five-number summary of the data: these are derived by taking all the data and putting the values in order. The derived values are:

- The median (midpoint value in the data, i.e. 50th percentile)
- The upper and lower quartiles (the points midway between the median and the extreme values, i.e. 25th and 75th percentiles)
- The minimum and the maximum

Box plots may also add the positions of potential outliers.

The median and quartiles are used because they are robust: they will not be affected much, if at all, by some odd values in the data. In contrast, the mean, and especially the standard deviation, are very sensitive to the addition of a single extreme value to the data. A box plot example is shown in Fig. 14.

A box plot will show very clearly where the odd extreme values are, and also skewness—where values are systematically further from the middle in one direction than in the opposite direction. The box plot in Box 1, section 2.3.3, illustrates the decoding accuracy of various kinds of data presentation; it shows very clearly the winner: 'position on a common scale' was rated the best for decoding the value of numbers. Not only was the middle value highest, but it was also recorded as the best at every session, and the average ranking varied

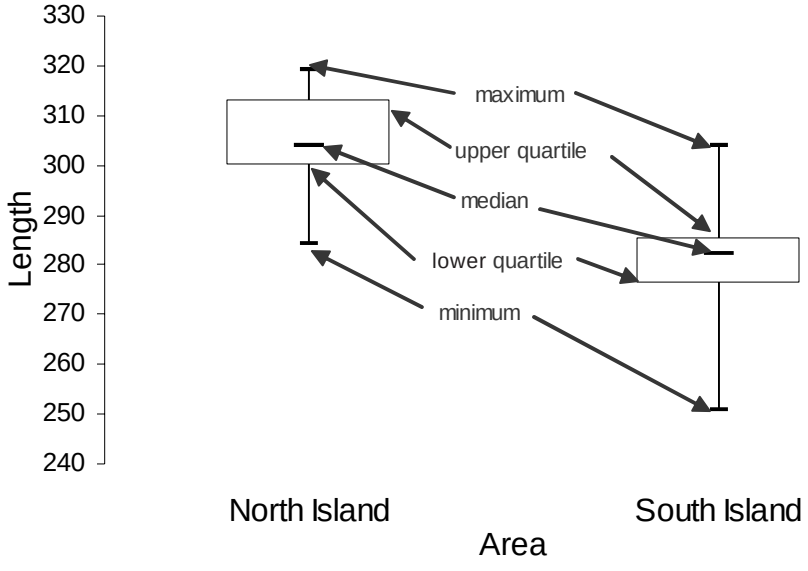


Figure 14. Example of a vertical box plot showing the distribution of Hector's dolphin data for North Island and South Island populations and the various box plot parts.

Original caption: Distributions of five measurements ... for the North and South Island populations, demonstrating the clear morphological separation between them...D, condylobasal length; ... Scale axes ... are in millimetres.

less than any of the alternatives. In contrast, some of the other methods were much more spread, and 'position on identical non-aligned scales' appeared to be skewed—with a median of 4, but many more values well above 4, and few much below.

Unfortunately, Excel does not provide facilities for creating a box plot as a standard type of graph, but there is a file developed at DOC that allows creation of simple box plots for up to 20 groups. The file can be requested from the third author (IW, DOC; email: iwestbrooke@doc.govt.nz).

Small datasets (say fewer than about 10 data points in each category), and some larger ones, may be better plotted as the individual values directly. An example is shown in Fig. 15.

Box plots do not work as well with integer data (e.g. counts) as they do with continuous variables (e.g. length); for integer data, for example, the 25th and 50th percentiles may both be on the same value, which messes up the box plot. This is illustrated in the evaluation data of the 2003 Graphs workshops, which applies the DOC spreadsheet for table format (Fig. 16A) and box plot (Fig. 16B). A simple, Excel-generated dot plot is provided for comparison (Fig. 16C).

More sophisticated box plots are available in statistical packages such as SPSS and S-PLUS. The key difference is that they go beyond the simple box plot by establishing 'fences' (usually 1.5 times the interquartile range—the range between the upper and lower quartiles) beyond the upper and lower quartiles. The whisker at each end stops at the extreme values of the data if within the fence, as in the simple box plot. However, if there are extreme values (possible outliers) outside these fences they are shown individually, with the whisker stopping at the closest data value within the fence. These more complex box plots are even more useful for exploratory data analysis. Because different implementations of box plots display different parts of the distribution with their lines and whiskers, it is always helpful to define these in the caption, e.g. 'The box plot indicates the median, interquartile range, maximum and minimum'.

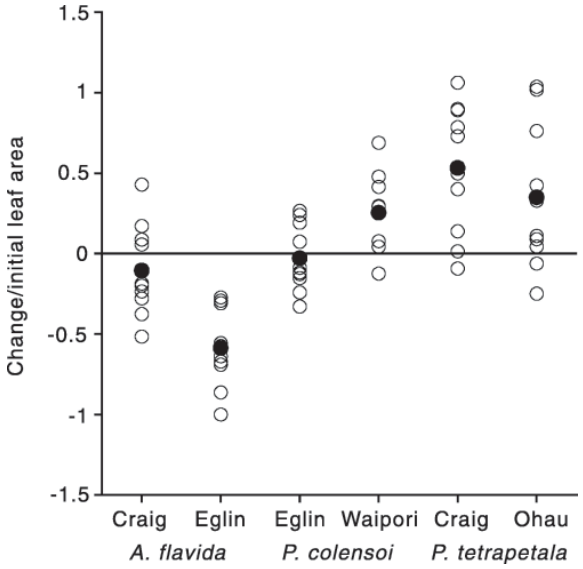


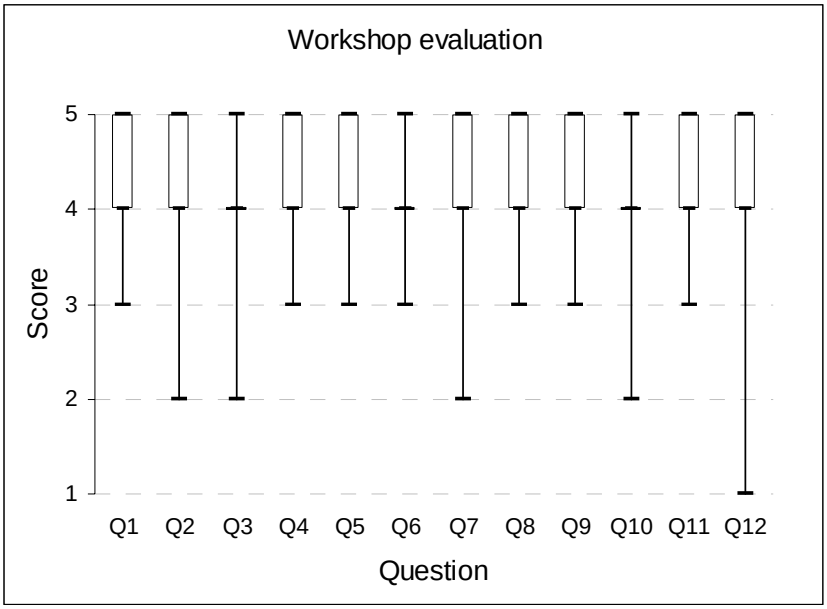
Figure 15. A dot plot showing the data for six categories that are tested statistically elsewhere with one-way ANOVA. Frequently this might be shown as a bar graph with six bars representing the means, and perhaps error bars. However, such bar graphs have a low information density, representing only 12 numbers (6 means and 6 SEMs / CIs). A somewhat more informative version uses boxplots (see Fig. 14). In the plot shown, the same space is used to display the number of data points and their full distribution, along with the means for each group.

Original caption: Overall annual leaf flux (net change in leaf area divided by the initial leaf area) between February 1997 and February 1998 on mapped branches in six populations of New Zealand mistletoes. (○), values for each plant; (●), population means. For full site names see Fig. 1.

A

Statistic	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12
maximum:	5	5	5	5	5	5	5	5	5	5	5	5
upper quartile:	5	5	4	5	5	4	5	5	5	4	5	5
median:	4	4	4	4	4	4	4	4	4	4	4	4
lower quartile:	4	4	4	4	4	4	4	4	4	4	4	4
minimum:	3	2	2	3	3	3	2	3	3	2	3	1
number of obs:	143	143	134	142	143	143	115	143	143	142	141	143
mean	4.3	4.3	4.0	4.2	4.4	4.1	4.2	4.4	4.4	4.1	4.3	4.2
standard deviation	0.6	0.6	0.6	0.5	0.5	0.6	0.6	0.5	0.5	0.6	0.5	0.7

B



C

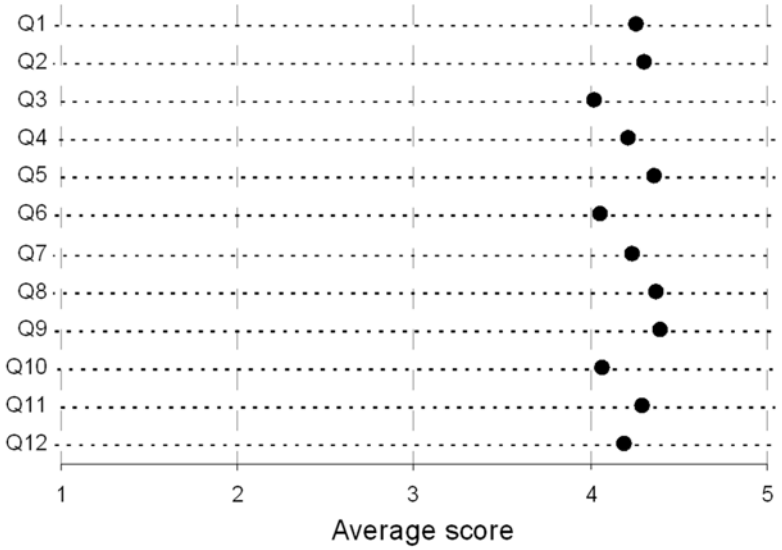


Figure 16. Participants' responses to graph workshop evaluation questionnaires. Scores: 1 Disagree strongly; 2 Disagree; 3 Neutral; 4 Agree; 5 Agree strongly. Figure 16A shows the results in table format, B shows a simple box plot, and C a dotplot of the average. The box plot does not work very well here with only a few response categories.

3.5 *x*–*y* (BIVARIATE, LINE OR SCATTER) PLOT

Bivariate graphs are the bread and butter of scientific graphing. They make excellent illustrations, and you really cannot go wrong using more of them. *x*–*y* graphs display two numeric variables. We can recognise two slightly artificial subtypes: time series, where the *x*-axis is time (more than 75% of graphs in newspapers were like this in the late 1970s (Tufte 1983)), and relational, where neither axis is time (42% of graphs in the journal 'Science' 1978–1980 were of this form (Cleveland 1984)).

There are various types: line graphs (lines only), line-plus-symbol (Fig. 17), or scatterplots (symbols only, Fig. 18), which can apply different symbols for several different variables.

You can include error bars on points; this can be done one way (vertically, as shown in Fig. 17, or horizontally), or both ways (vertically and horizontally), as appropriate.

In a scatterplot, extra text labels to the data points may increase clutter and should generally be avoided. Sometimes you can use a text label as the data point (e.g. using capital letters A, B, C, etc. to mark locations and also identify sites—which gives labelling without increasing clutter: Fig. 18). Avoid letters overlapping.

You can plot a scatter with a fitted line, e.g. a regression line as in Fig. 2. Never show the regression only! It takes no extra space to put the data on and the scatter gives a lot of information about the data. Indeed, the data may well show that even though the r^2 value is close to 1, the interpretation may be suspect (Fig. 19).

A step function graph is a variant of the *x*–*y* graph, where the *y* value is constant over intervals then changes suddenly to a new value (e.g. the price of the daily newspaper over time), so the graph looks like series of irregular (square-edged)

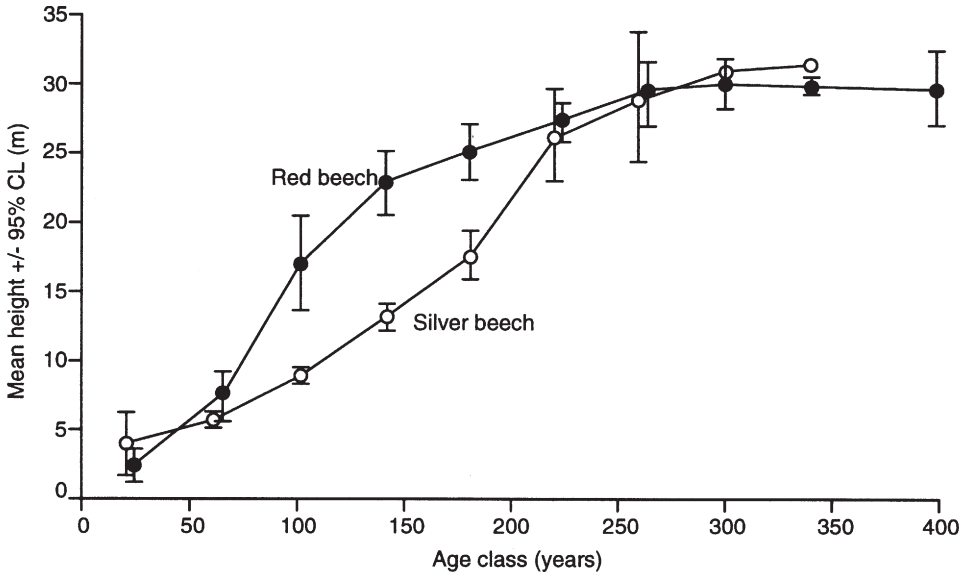
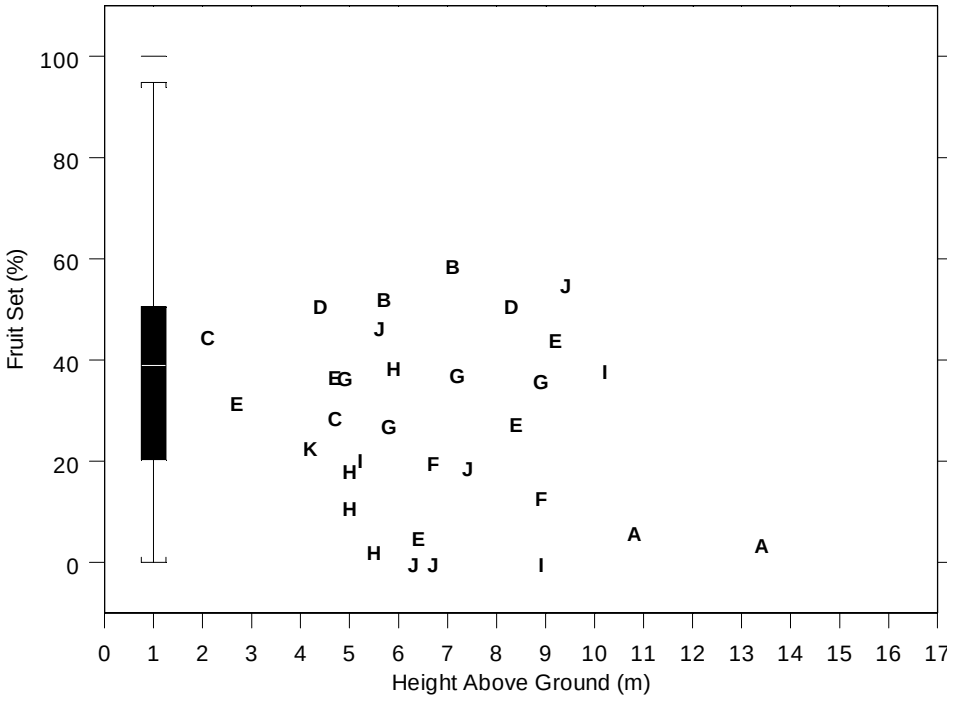


Figure 17. Good example of a clear *x*–*y* plot with suitable symbols, categories, and error bars with explanation (95% confidence interval).

Original caption: Average height growth of red and silver beech trees of different age classes in a stand in the Maruia Valley (After Stewart & Rose 1990).

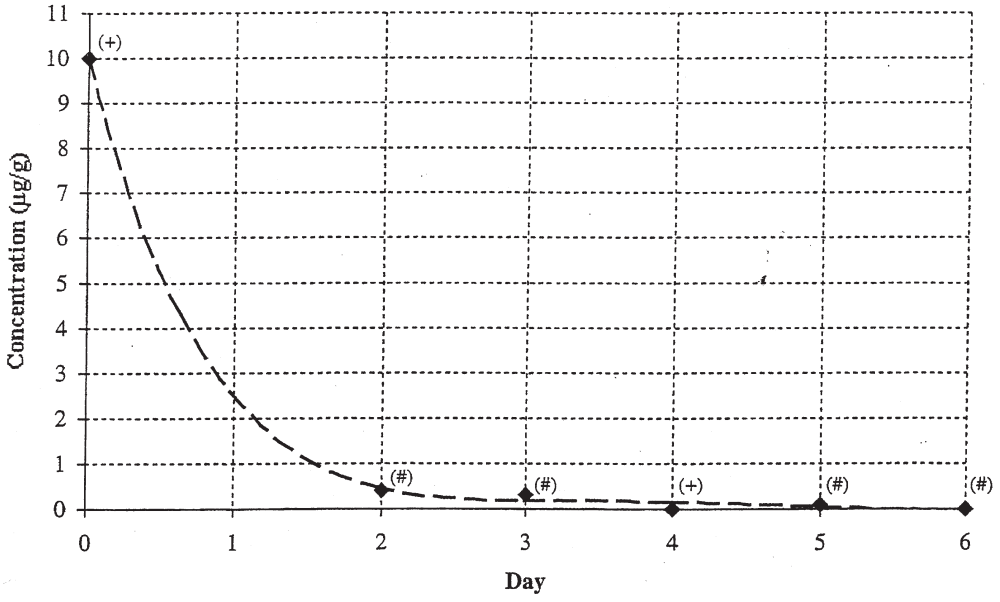
Figure 18. An example of text labels used as data points; the graph also uses a box plot on the left. The text labels serve to locate each point (mistletoe plant) both for height and fruit set rate; they also allow the reader to identify mistletoe plants that share a single host tree. It is not easy to decode the latter, but in this case the authors thought it not especially important to do so, as the overall message is that there is no effect either of height or of individual host tree. If it was important to easily link mistletoes on a single host, the points for mistletoes on the same tree could be joined by lines, but this would make it harder to see the overall picture (here of no relationship between height and fruit set).



Original caption: Fruit set in *P. tetrapetala* at Craigieburn Forest Park in the 1997/1998 flowering season. The box plot shows the range of fruit set values obtained from tagged plants used for our normal pollination treatments (all located within 4 m of the ground) while letters mark the 32 plants located up vertical transects accessed by climbing ropes. Shared letters indicate plants that are located on the same vertical transect.

Figure 19. Not only does the curve interpolate and extrapolate well beyond acceptable boundaries, e.g. curve between the first and second datapoints), it also incorrectly combines data from different sources according to the accompanying text. At best, the points could have been connected by two separate lines: one from (0,10) to (4,0) and another to connect the remainder, just above the *x*-axis.

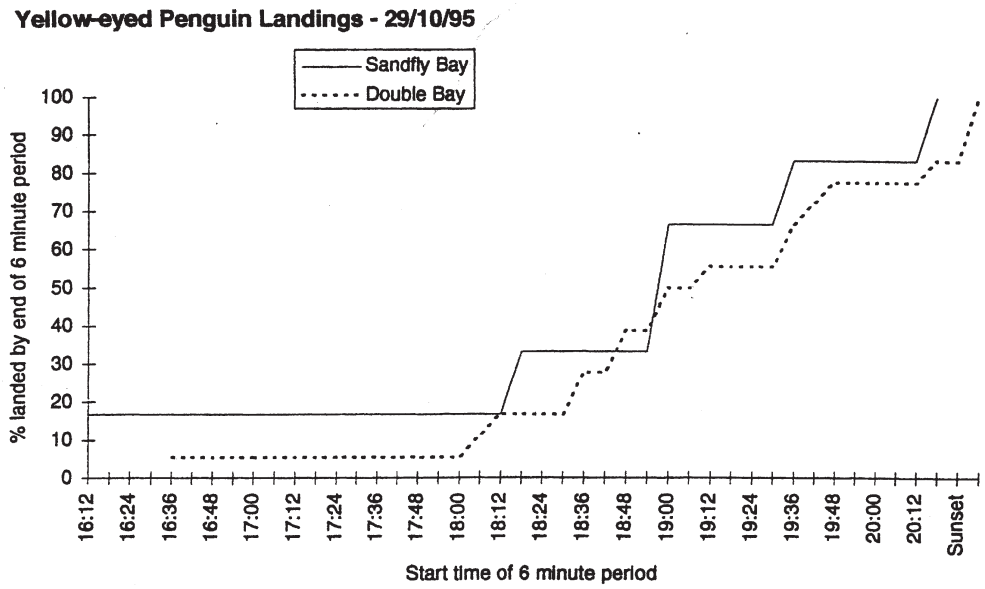
Original caption: Estimated excretion curve for brodifacoum in Orthopteran species. Based on data from this study (#) and Booth, Eason & Spurr 2001 (+).



steps. The step function graph is often used for cumulative proportion below a certain value in a sample, or for representing the estimated proportion surviving over time (Fig. 20).

You can plot several categories or classes on the same *x*–*y* graph, using symbols to separate them, as in Figs 17 and 18. The main concern is symbol / line separation; if this becomes a problem, you may have to present multiple small graphs rather than one large one (see section 3.7).

A



B

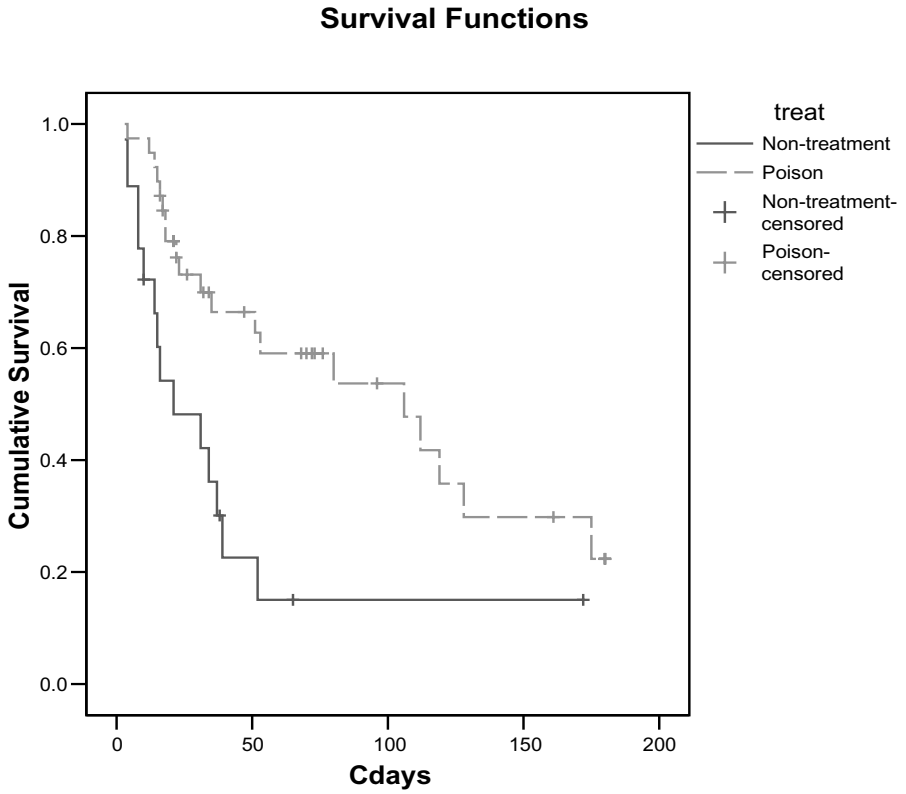


Figure 20. Step function graphs. A, with angled steps. Conventionally the steps drop vertically at each point to the level of the next point, as shown in B (survival rates estimated for two groups of kiwi chicks with and without pest treatment).

Original caption for A: Graph[s] showing comparison of daily yellow-eyed penguin landing times between Sandfly Bay and Double Bay.

3.6 THREE-DIMENSIONAL (MULTIVARIATE) AND TWO-DIMENSIONAL-PLUS GRAPHS

3.6.1 Three-dimensional (multivariate) graph

Multivariate graphs have high information content, which can be good, but is sometimes more than can be shown readily. The main problem with these graphs arises when they become too complex to be easily interpreted. Exercise caution. There are various ways to make such graphs more effective.

x–*y*–*z* graphs use a 'mock 3D' representation on the flat (2D) page. It is usually hard to represent three dimensions accurately on a flat sheet, and only some types of datasets will lend themselves to this treatment. The wire-frame style (i.e. a 3D line graph) is generally best, but its effectiveness depends on the exact shape of the data (Fig. 21); some data may be hard to see, with points hidden behind other points. Various approaches to graphing a 3D dataset are shown in Fig. 22. The points-on-a-stick graph (a 3D-scatter: Fig. 22D) can be difficult to make sense of. Never draw such points without the sticks as this makes it impossible to interpret them! For on-screen analysis or presentations, rotating graphs can be very useful and effective, but are unsuitable for publication.

3D histograms and 3D bars (Fig. 23A) are generally poor graphs: at best they are hard to interpret; at worst they create unnecessary distraction from the data and qualify as 'chartjunk' (see section 4.11). It is generally better to present these data as 2D multivariate bar graphs beside each other, or as *x*–*y* graphs (Fig. 23B) if appropriate (see Box 2 for an outline of methods to convert 3D bar charts to *x*–*y* line graphs using Excel).

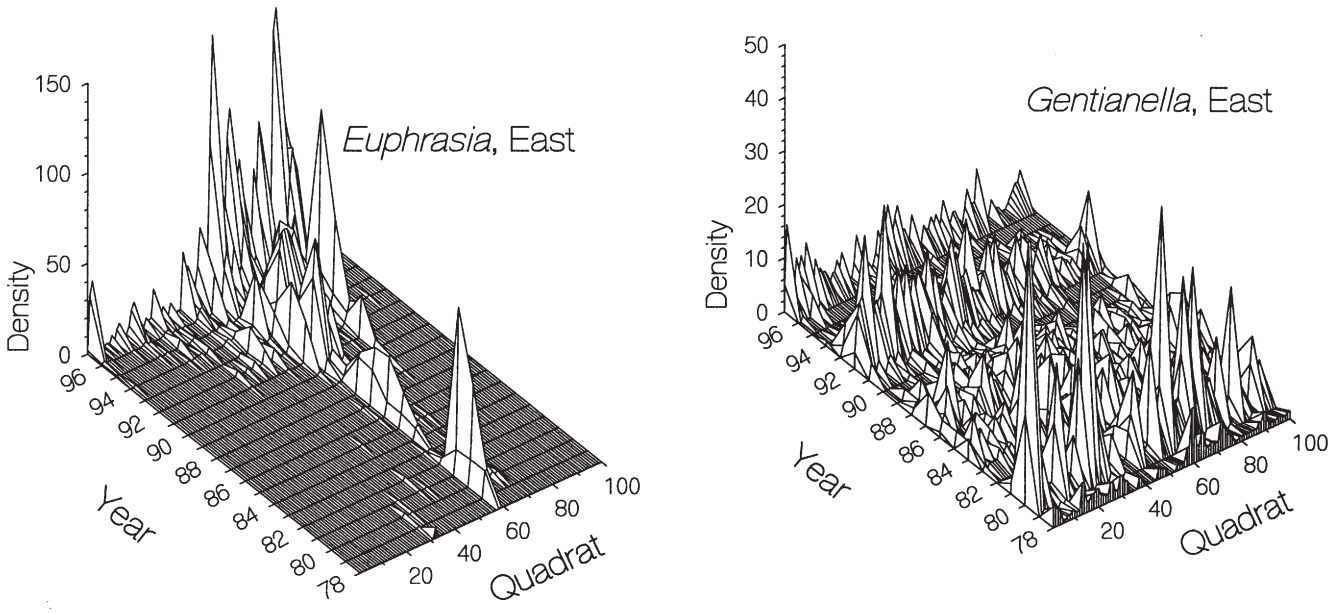


Figure 21. 3D graphs showing wire-frame representations, i.e. data are regularly spaced on a grid and each point is joined by lines. Note that this is much more successful where there is a clear, simple pattern (e.g. *Euphrasia*), preferably with a lot of short-scale autocorrelation (i.e. adjacent points tend to have similar values) than when there is a lot of scatter (e.g. *Gentianella*). This type of graph may be hard to draw if the original data are not collected on a regular *x*, *y* grid; interpolation of the values for the regular grid points may be necessary, which is undesirable and arguably misleading.

Original caption: Distribution of flowering individuals of four short-lived plants in the East transect, Castle Hill N.N.R., 1978–97. The total numbers of flowering plants seen in each 0.5 x 0.5 m quadrat in each year is shown. The species were *Gentianella amarella*, [*Rhinanthus minor*, *Medicago lupulina*] and *Euphrasia nemorosa*.

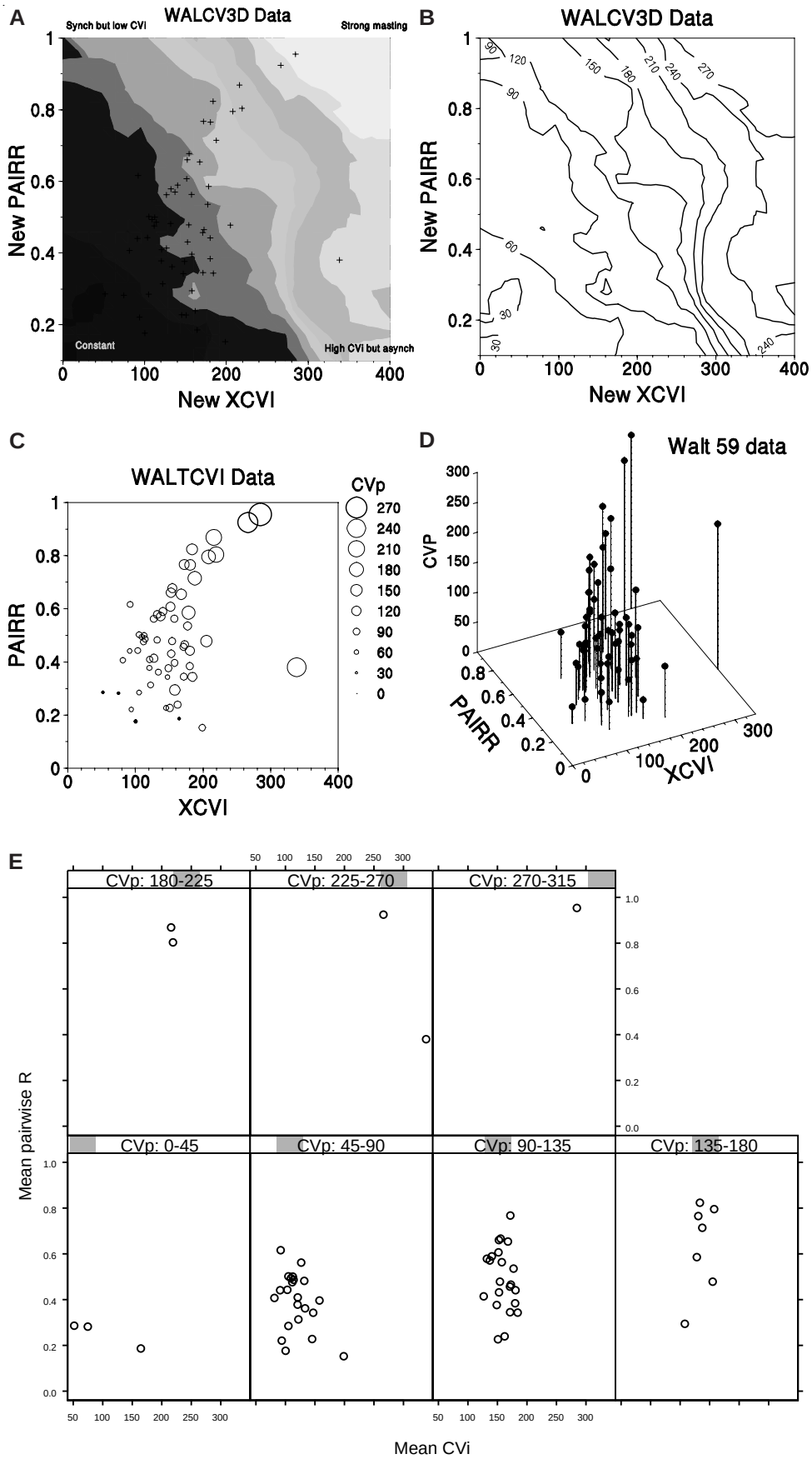
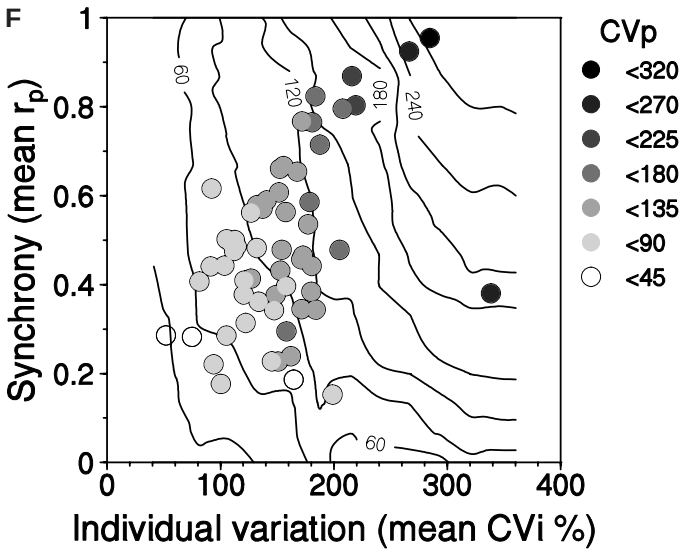


Figure 22. Six ways of graphing a 3D dataset to show the relationships among CVi (mean individual plant variation in seed crops), R (mean pairwise R—synchrony among plants), and CVp (population variation in seed crops). Four 3D representations of the same data (A–D) were tried before settling on the final published graph (F; overleaf); version E is an alternative approach, developed by IW. Version A shows the contours more clearly using colour and the location (but not the value) of the data points. However, as can be seen, this does not reproduce well in black and white. Version B shows only the contours, in monochrome. Version C shows the value of each data point (by the size of the circle) but not the contours. Version D shows the value of the 59 data points using a wire stick model, but not the contours. Version E shows the same 59 data points (but not the simulation contours), with nine panels for different levels of CVp, shown graphically in the strips at the top of each panel. The multipanel approach allows a clearer view of the relationship of CVi and R at the given levels of R, although it requires more space. Panel F (overleaf) was chosen because it shows both the actual values for each of the 59 real data points (by shading of the circles), and the contour lines for the simulations, all in monochrome as required by the journal. (The contour lines are more strongly smoothed in the final version (F) than in earlier versions (A, B), but should perhaps be smoothed even more to avoid giving an unwarranted impression of precision in the detailed contour patterns.)

Continued on next page

Figure 22—continued



Original caption: Interrelationships among mean rp, CV and mean CVi based on randomized permutations. For each of the 59 datasets used in the analyses, annual data for individual plants were reshuffled to vary synchrony from high to low while holding mean CVi constant, and calculating the resulting CVp. Plotted are CVp contours resulting from 270 different combinations of each dataset, giving a total of 16,000 reshuffled datasets. Circles are the actual CVp values for the 59 datasets.

Box 2: Conversion of 3D bar charts (Fig. 23A) to x–y line graphs (Fig. 23B) using Excel

*1. The key step is to **change the chart type**: select the Chart Area, click on 'Chart' on the menu bar, select 'Chart type', choose 'Line', and select the line with markers on it (usually the default).*

*2. **Maximise the size of the graph** within the overall chart window. Select the graph, by clicking in or near the graph itself until a dotted grey border appears. The graph can now be extended to take up as much as possible of the chart window by dragging on the black squares at the corners and in the middle of the sides. The default grey background is readily deleted by pressing 'Delete' on the keyboard while the graph area is selected.*

*3. **Labels for the axes** are essential, and are entered in the dialogue box from the 'Chart Options' option of the Chart menu, selecting the 'Titles' page. By default, the labels are in bold, so click on each label and change the bold format to regular. For publication, it is desirable to delete the overall title, except possibly for an on-screen presentation.*

*4. **Remove the horizontal gridlines** cluttering up the graph by clicking on one of them to select them, and again pressing 'Delete' on the keyboard.*

*5. **Remove the border on the chart window and the legend** by double-clicking on each in turn and, in the Patterns tab, setting 'Border' to none.*

*6. **Multiple lines need to be formatted** so they can be clearly identified, even when printed in black and white. For example, the pre-irruptions lines can be formatted as dotted lines with circles as markers, and the post-irruption can have solid lines with triangles. The core can be indicated with solid markers, while the buffer has hollow makers. To implement this, double-click on the line to get 'Format Data Series' dialogue box. Select 'Patterns', and under 'Lines' choose the line type and colour wanted. At this point, it is also possible to change the marker type, and foreground and background colours—to black, or no colour for example.*

*7. The **overall shape of the graph can be adjusted** if need be (provided it is embedded in a worksheet rather than in its own window) by selecting the whole graph window and changing it in the normal Windows manner.*

*8. **Final tidying up** may involve adjusting font size (in the tool bar), after selecting the whole graph or legend or an area of text. Then the graph area and legend may need to be resized, and the legend moved.*

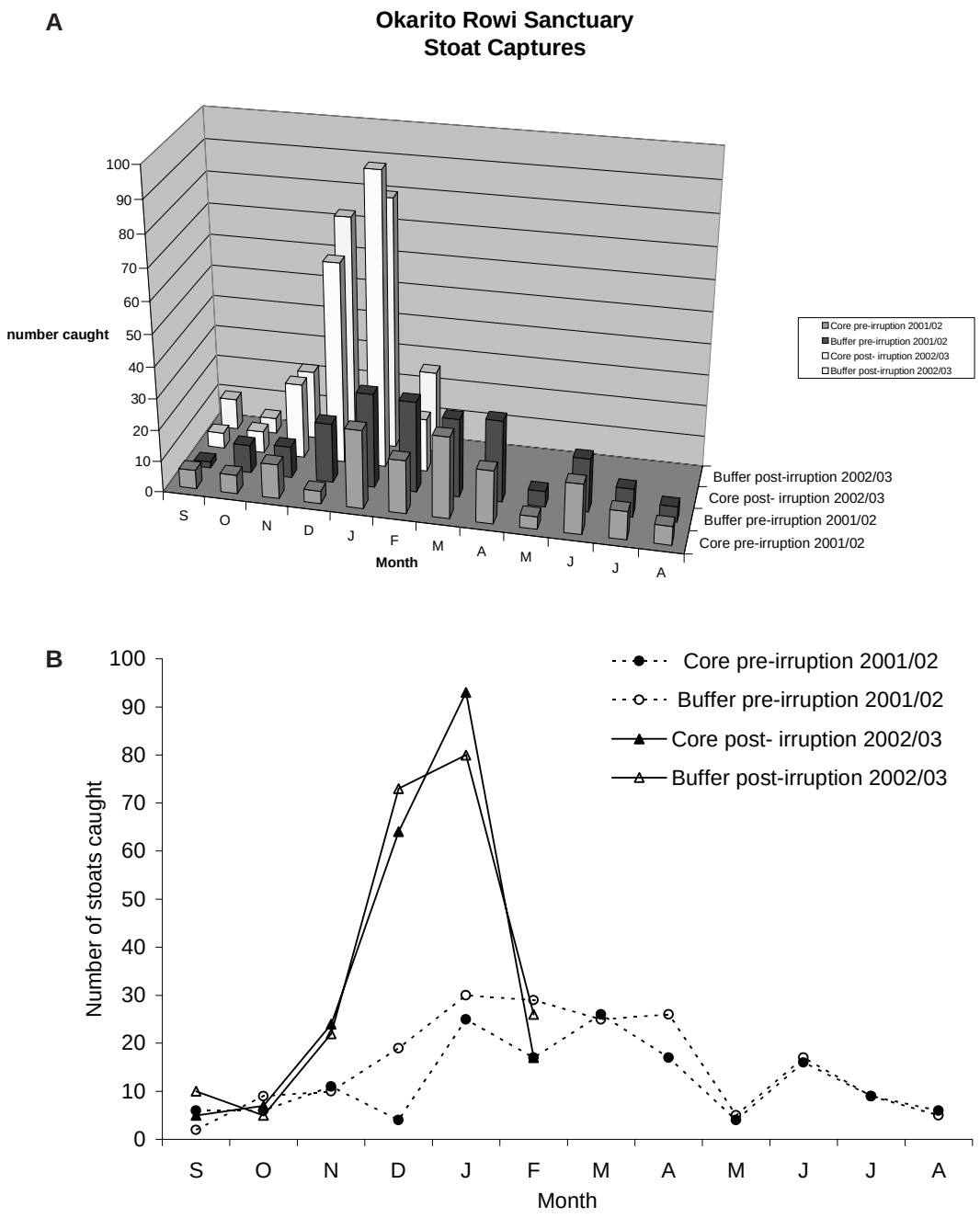


Figure 23. Stoat data from Okarito used in graph workshop. A. 3D bar chart intended to be published in colour. B. The same dataset presented as a monochrome *x*–*y* line graph (see Box 2 for methods on how to convert A to B).

3.6.2 Two-dimensional (2D)-plus graph

This is an *x*–*y* graph with various other ways of representing the value on the third axis. These usually work better than 'mock 3D' graphs. This type includes data maps, where the *x*- and *y*-axes show the physical location on the ground and the third dimension shows the value of some variable of interest at that spot. The third dimension can be shown using contours as in maps (this works only where the change in the third dimension is gradual); bubbles of varying sizes; shading; small pies or bars; or trajectories. An example using bars is shown in Fig. 24. Some of these methods work with particular data—for example, trajectories can be useful for showing a sequence of points through time; points are marked and successive points linked by lines, with labels or arrows to show the start and end.

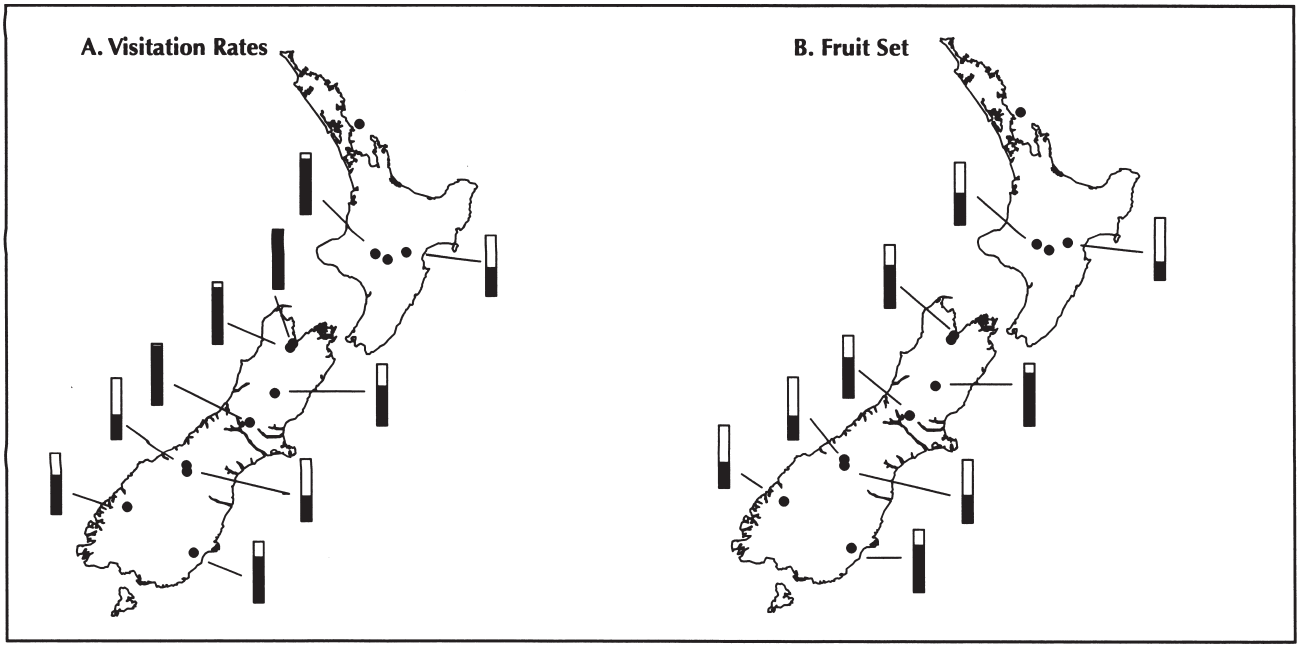


Figure 24. Two data maps, showing the value of a third variable (left, visitation rate by birds to mistletoe flowers; right, fruit set by mistletoe plants) at points located by geographic position (latitude, longitude). The third variable is shown by the fraction of a bar which is filled; this is better than simply having dark bars of variable length. Note that if the key message was the relationship between visitation rate and fruit set at each site, a plain *x*–*y* graph plotting the two directly against each other would be superior. The data could alternatively have been presented with the visitation rates (and possibly site names) on the map instead of the bars.

Original caption: Visitation rates for all mistletoe study sites during the 1997–98 season are shown in figure 9a. Visitation was quite low at Lake Ohau, the Temple and in the Kawekas. This signifies a lack of pollination in these areas. Fruit-set for all mistletoe study sites during the 1997–98 season are shown in figure 9b. Fruit-set generally correlated with visitation rates.

3.7 MULTIPANEL GRAPH

Multiple related graphs have become increasingly easy to prepare with the ready availability of high-powered computers. If you have more than two or three variables, this is really the only failsafe method to represent data: with care it can work for almost any dataset, as opposed to the tricks like those discussed above that work only for particular types of data. The use of multiple graphs (e.g. Fig. 10) has been systematised and extended by Cleveland (1993) to support graphs such as Figs 12B, 22E and 25. The simple but powerful underlying concept is to break the limitations of the two dimensions that are represented on two axes by using multiple panels of such graphs. Conventional graphs, typically displaying two to three variables, are systematically arranged in a series of panels that allow users to see a number of variables and their interactions simultaneously. With careful allocation of variables to axes, it becomes possible to view and even analyse complex interactions graphically, without needing to standardise, or to fit models that usually require strong assumptions to be made. Relationships within the data, obscured by the limitations of standard two- or three-variable graphs, are illuminated once other influential variables are controlled for within the multipanel graph structure. The power of this approach is shown in Fig. 25.

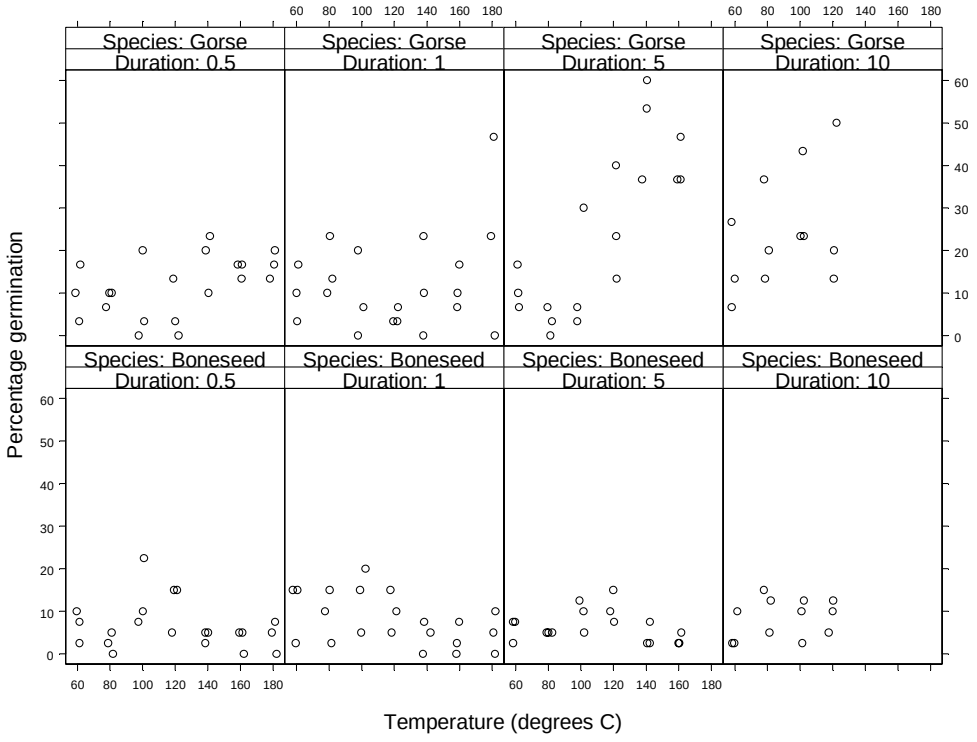


Figure 25. This set of graphs from S-PLUS shows, within each panel, germination rates (on the *y*-axis) for seeds exposed to different temperatures (on the *x*-axis). However, in addition to these two variables, an additional two variables are shown using panels. The upper row of panels is for gorse, and the lower for boneseed; the horizontal panels represent increasing exposure times to the raised temperatures. Thus four variables are presented. One is categorical—gorse / boneseed; one is continuous—percentage germination; and two represent discrete levels of potentially continuous variables—temperature (eight levels, but note that the temperature data has been jittered slightly to avoid points being hidden under each other) and duration of exposure (in minutes; four levels). This graph allows an almost immediate answer to a key research question for the experiment: 'Does gorse or boneseed germination respond to heat', and allows rapid exploration of further questions, like the effects of increasing temperature and duration of exposure to heat.

3.8 SCATTERPLOT MATRIX

This is a very good multipanel tool for exploring multivariate data, which plots all possible two-way scatterplots in a grid (Fig. 26). The reader can refer from one to another easily. Note that you plot each graph twice, once above and once (inverted) below the diagonal. This avoids having to mentally flip any graph and takes scarcely any more space on the page, using the otherwise vacant second half of the matrix on the other side of the diagonal. Scatterplot matrices are very good; they repay careful study.

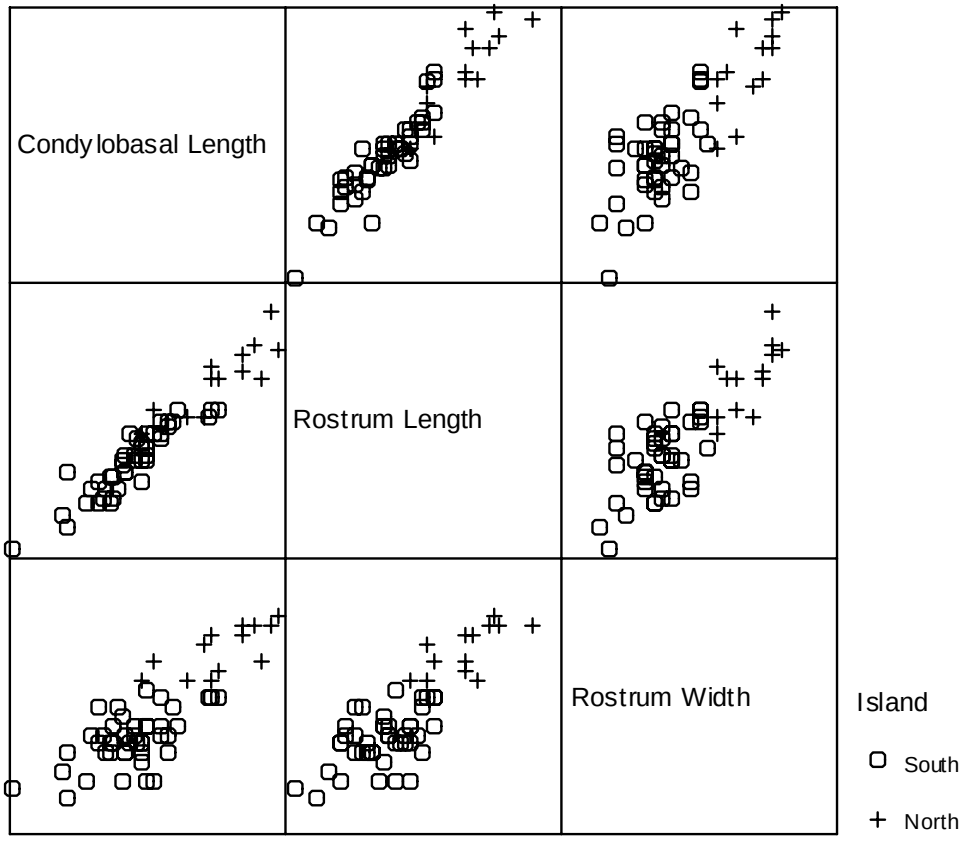


Figure 26. Scatterplot matrix of three measures of Hector's dolphin skulls, by island. Scatterplot matrices of many skull measurements were used during analysis of these data, and a plot of rostrum length and width was published in the paper, which established the North Island dolphins as a separate subspecies—Maui's dolphin. See also Fig. 14.

3.9 OTHER GRAPH TYPES

The Council of Biology Editors guide (Peterson 1999) lists many 'graph varieties', all of which have been covered above, except for area (band) and polar (circular) graphs.

Area (band) graphs are best avoided, except in specialist applications such as pollen diagrams (Fig. 27). The bands can suggest major variation in variables, often based on very few data points.

Polar (circular) graphs are again very specialised, used for example when plotting frequency of direction taken (Fig. 28).

Microsoft Excel offers some additional 'chart types'. Again, most of these have very limited, specialist applications. For example, the Stock chart is specifically for commerce (but we used it to mimic a box plot in Excel in section 3.4).

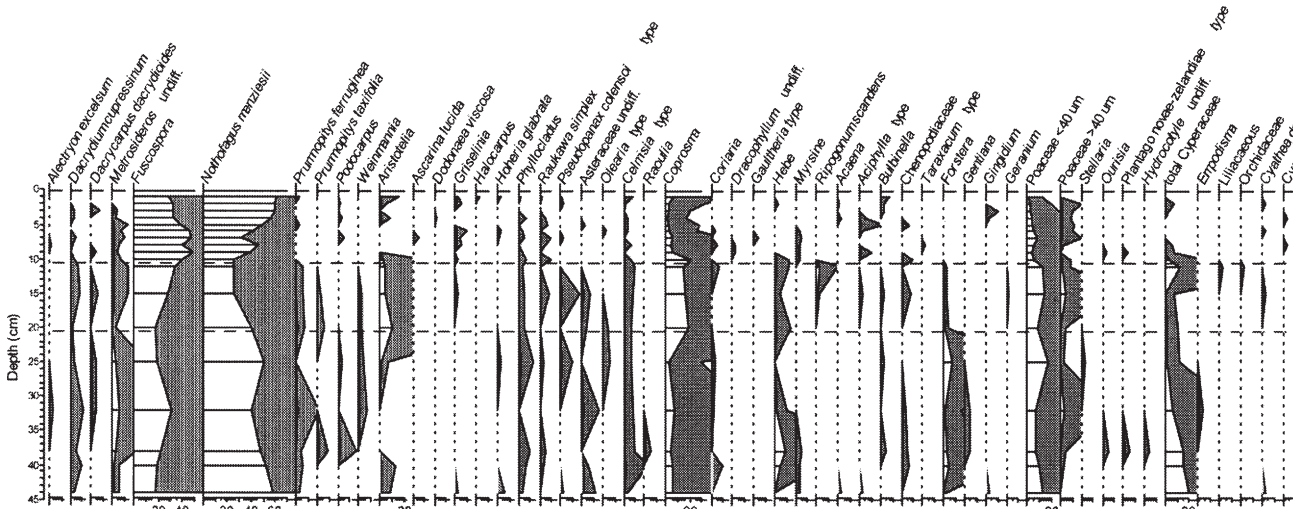
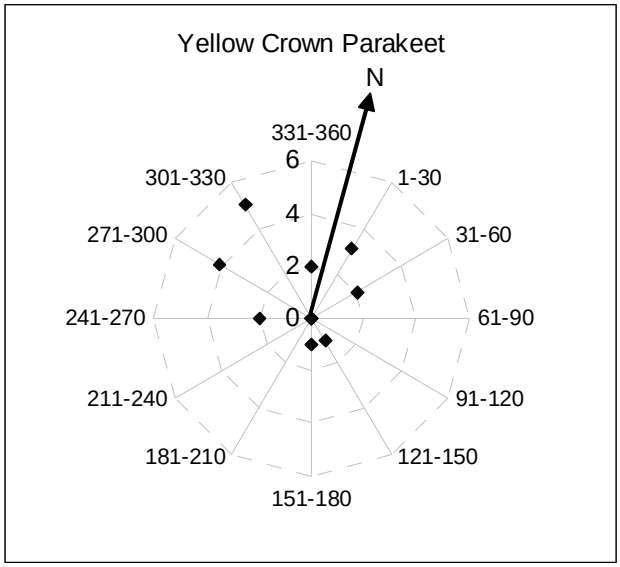


Figure 27. Example of an area (band) graph: percentage pollen diagram for soil cores [truncated at right].

Original caption: Percentage pollen diagram for core X00/2 from Eyles Upper Plateau Bog, forest margin. Shaded curves represent x10 exaggeration to highlight low values.

Figure 28. Polar or circular graph displaying the number of parakeet nests by bearing around the host tree, in 30-degree groups. North lies midway between the 331–360 and 1–30 groups, since the bearings are degrees from north, and 360 degrees is the same as 0 degrees. Note that far fewer nests face southward (91–270 degrees) than northward.



Some, such as the Donut, Cylinder, Cone and Pyramid 3D bar charts, are very prone to distortion (Tufte 1983); these are best avoided.

'Surface–3D surface' may be a suitable way to show a fitted surface in three dimensions, yet it suffers from an inability to show the points and so does not give a feel for the data spread. It is the 3D equivalent of plotting just a regression line without data, against which we argued in section 3.5. However, for complex data, with a number of variables being represented, it may become impossible to show the data directly.

4. Graphical elements

Now that we have defined the various graph types, let us look at the main parts of a graph and how to make sensible choices there. We list guidelines or rules for each element. The Council of Biology Editors (Peterson 1999) also provides an excellent detailed reference.

4.1 SHAPE AND SIZE

4.1.1 Shape

The standard shaped graph is slightly wider than tall (especially for a large dataset): tall formats may overemphasise or exaggerate change in the dependent variable (usually the *y*-axis). A horizontal orientation fits a computer or projection screen (e.g. for a computer-based presentation) better. In seminars, lectures, etc., the bottom of the screen may be cut off, so that you may lose the bottom of a tall graph.

One exception is the correlation graph: it is usually shown in a square since the two variables are treated equally and interchangeably. Similarly, a square may be best if you have the same units and ranges on both axes. It may be best to choose a shape that gives the same scale on both axes.

Cleveland (1994) recommends that the best shape is the one that provides a 45° angle for the key data or line, as this provides maximum visual resolution in both directions. However, there may be advantages in restructuring the data to have the key comparison or reference line being horizontal rather than at 45° (Fig. 31), since we are better at judging deviations from the horizontal.

4.1.2 Size

The graph must be large enough to be clear and allow for reduction. Bear in mind the graph's final destination. If it is intended for publication in a journal, check the page size minus the margins, and whether text is printed in one or two columns. Publishers will usually reduce graphs to the smallest possible size at which the data and labels are still (just) legible, and this is likely to be one page-width or one text-column width. When you design a figure panel for a full page, make sure that there is room for the caption to be inserted separately (see section 4.10).

4.2 AXES AND GRIDLINES

Data should more or less fill the data region, except for these constraints:

- Choose tick mark labels to include the entire range of data, using round numbers.
- Keep data clear of axis lines, especially the zero line (data get lost on the line); simply draw the *x*-axis a little lower than zero if necessary, or extend the plot area into negative *y*-values.
- Consider whether conventions or expectations may affect other choices, e.g. you may want to include a particular value such as zero or 1.
- In multipanel graphs, use the same axis range in all panels (or at least the same units per centimetre; Fig. 29) wherever appropriate to aid comparison.
- Avoid unnecessary repetition of labels (e.g. if all panels have the same *y*-axis, showing this only once will do); for an example see Fig. 29.

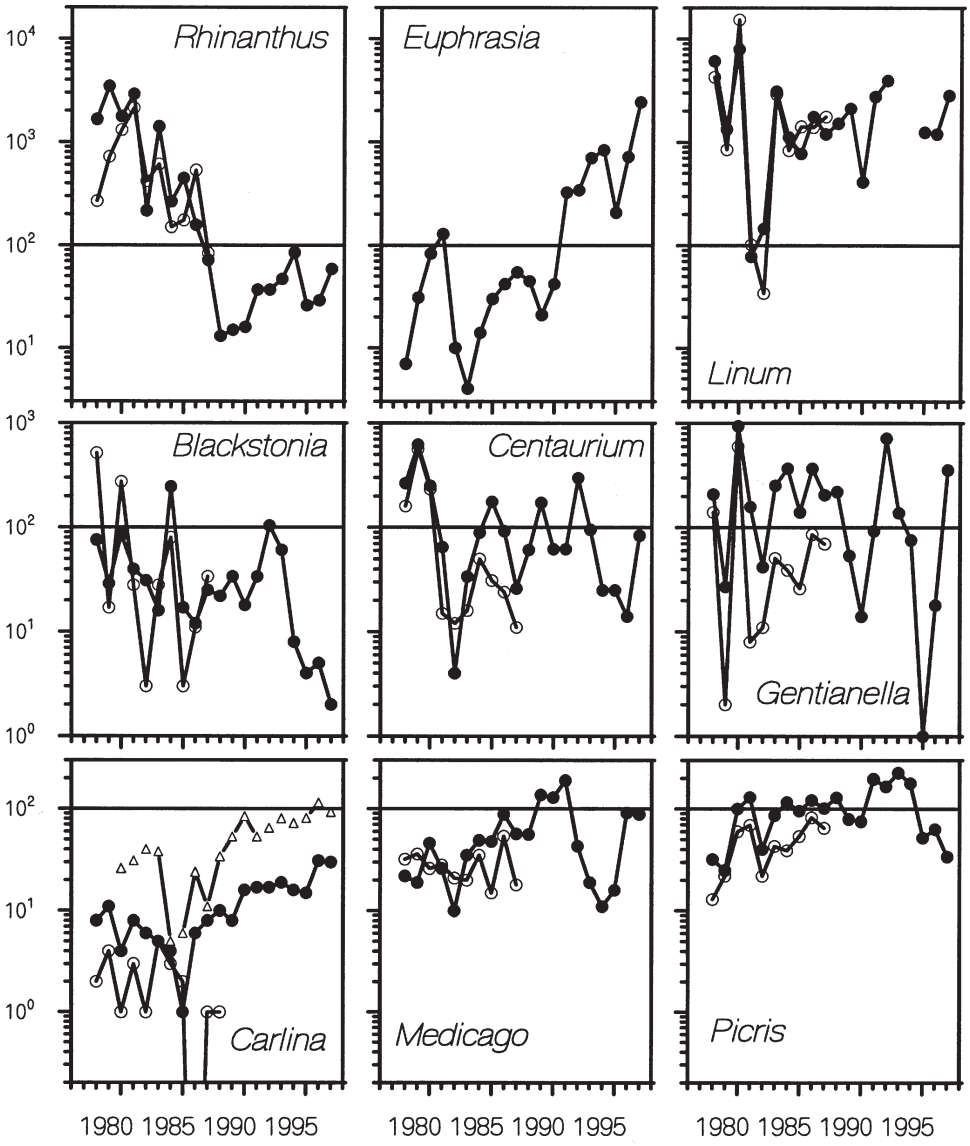


Figure 29. This multipanel graph uses the same *x*-scale on all panels, but to cope with widely varying plant densities, groups similarly common species in rows, with the same *y*-range in each row but different *y*-ranges between rows. Note, however, that firstly the *y*-scale (change per cm) is the same for all rows, and secondly the use of a reference line at $y = 100$ allows comparison among panels despite their different *y*-ranges.

Original caption: Total number of flowering plants recorded at Castle Hill, 1978-97. Each transect is 25 m^2. The East transect (filled symbols) was recorded until 1997, and the South transect (open symbols) until 1987. For *Carlina*, the totals from an adjacent area of 250 m^2 are also shown (small symbols).
(a) *Rhinanthus minor*
(b) *Euphrasia nemorosa*
(c) *Linum catharticum*
(d) *Blackstonia perfoliata*
(e) *Centaurium erythraea*
(f) *Gentianella amarella*
(g) *Carlina vulgaris*
(h) *Medicago lupulina*
(i) *Picris hieracioides*.

There are a few other niceties to observe. Put ticks on the outside of the *x*- and *y*-axis lines (so that the ticks will not overlap data). Labels on the *x*- and *y*-axis are usually horizontal at the bottom and vertical at the left, respectively, but units should always read horizontally. If necessary, both axis lines and axis labels may be repeated at the top and right to assist easier reading of values; this also sets the graph area aside from the text. However, some editors will not allow this for reasons of journal style, and the extra axes and labels may also create unnecessary clutter. Deciding whether or not to add the extras can also be a matter of personal preference. (When designing for journal publication, make sure to check the relevant journal's guide to authors and to check some back issues for the preferred style.)

You can double-label axes (e.g. year of birth with a set of labels on the left, and age in a particular year on the right). However, using two different scales on the same axes should be avoided: it can easily lead to misleading presentation (Fig. 30). Do *not* insist on zero being included if this ruins the resolution (remember: your audience will be intelligent enough to read the labels).

Scale breaks are an admission of failure: they violate the whole idea of graphs (position indicating the value of the variable), so avoid these whenever possible. A log scale (see section 4.3) may remove the need for a break by spanning a wider range of values. If you must use a scale break, make it a 'full axis break', not just a break in the data line. Such breaks must be obvious. Do not connect numbers across the axis break, i.e. make sure you 'interrupt' every line crossing the break.

Choose axes so that the reader is performing a comparison high on the order of decoding accuracy. Usually this will mean that the main point of comparison is with a straight or horizontal line. Figure 31 illustrates the principle further.

4.3 LOG SCALE

A log scale transforms an exponential function to a linear function. For example, a log scale with base 10 treats a 1 as zero, a 10 as 1, 100 (= 10^2) as 2, 1000 (= 10^3) as 3 and so on (as in Fig. 31). This allows a very wide range of values to be shown in one graph. Use a log scale when it is important to understand relative change across the whole range of data.

Log scales give lower accuracy of location for high values but much higher accuracy for low values. Log scales are useful for right-skewed data, i.e. data with many low values and a few very high ones (common in biology, e.g. plant weight, plant seed output, number of offspring per male bird). Plot the *full values* along the *y*-axis (with a few tick marks between orders of magnitude), i.e. list 1, 10, 100, etc., not 0, 1, 2, on a log 10-scale; or list 0, 1, 2, 4, 8, 16, etc., on a log 2-scale, which can be useful for a smaller range of numbers). Do not use bar graphs with vertical log scales, as bars need a zero value to start from, which on a log scale is not possible (the log of zero being negative infinity).

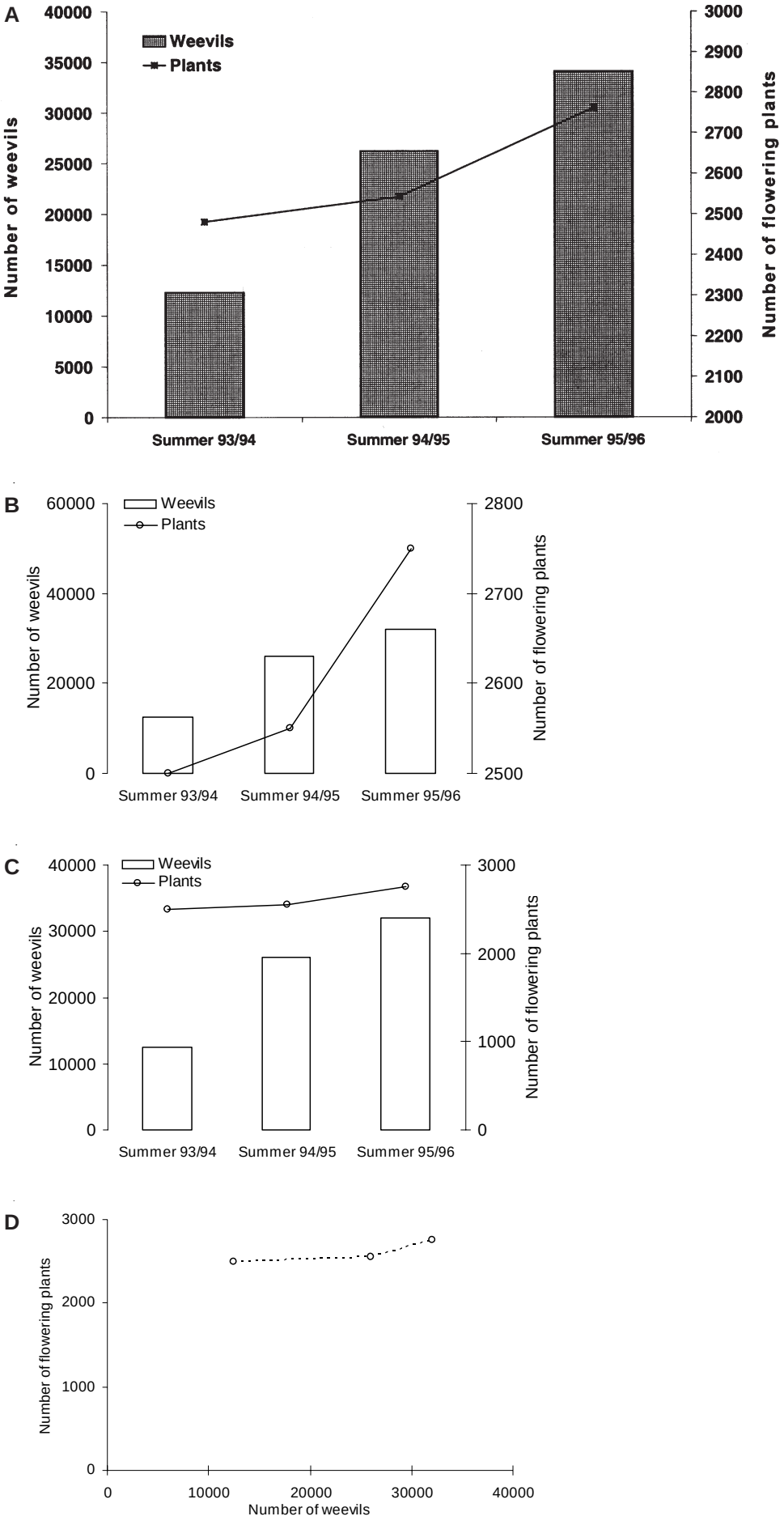


Figure 30. A–C illustrate the perils of using the same axes to present different variables. A, which follows closely the original published version, suggests visually that the two series are moving together but with less change in plants—but note that the weevil axis shows values from 0 to 40 000 while the plant axis shows 2000 to 3000. B makes plant numbers increase more than weevil numbers—by forcing the plant axis from 2500 to 2800. C shows both axes on a 0 base, avoiding distortion, and showing that plants change little proportionately, while weevils change a lot. D shows a direct comparison between the two series, which may be preferable.

Original caption: Totaled numbers of *H. spinipennis* and of flowering *A. dieffenbachii* plants for six selected patches on Mangere Island.

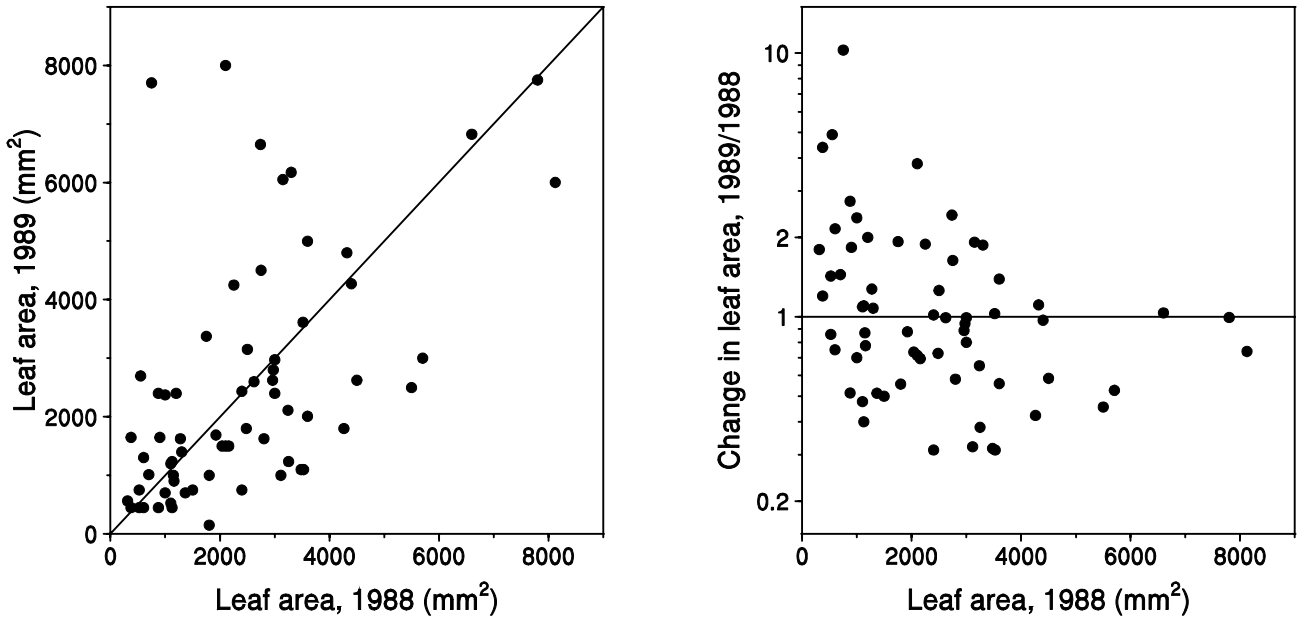


Figure 31. In these data, small plants tend to be larger the following year, and large plants tend to be smaller, but this is not very obvious in the raw plot on the left, which requires the reader to decode growth by seeing if the point is above the 45 degree diagonal line. On the right, the ratio of (size in year 2, i.e. 1989, divided by size in year 1, i.e. 1988) is plotted against size in year 1, so the reference line for no change in size becomes the $y = 1$ line. The ratio scale allows an estimate of position to answer the question, whereas the second case would require you to decode position up y compared to position along x. Note that the ratios are presented on a log-y scale, otherwise reductions in size ($y < 1$) are de-emphasised compared to equivalent increases in size ($y > 1$).

Original caption: Change in size (leaf area, in mm^2) of *Botrychium australe* plants at Cass between 1988 and 1989. Left panel, raw size in 1988 vs size in 1989, with the $x = y$ diagonal line (= no change in size) for reference. Right panel, size in 1988 versus change in size (1989/1988, on log scale) with the $y = 1$ line for reference (= no change in size).

4.4 TITLES AND LABELS

Show standard whole units on the axes. For example, if plotting data for successive 5-minute samples, do not label as sample 1, 2, 3... but as time (minutes 5, 10, 15...). Or, if you have density recorded as numbers per 0.125-m^2 quadrat, recalculate and label as per m^2.

Axis labels need at least three numbers per axis to tell linear from log scales, but probably no more than 5 or 6. Aim for even increments, such as 0, 20, 40, 60, which are much easier to divide mentally when interpolating than 0, 25, 50, 75.

Make the axis title clear about what is measured, and put the units in parentheses, e.g. 'Wasp arrivals (m^{-2} hr^{-1})', not 'Arrivals per plot per sample time' or 'Density' or 'Arrivals'.

Except for descriptive labels on the y-axis, vertical or oblique type is best avoided, particularly in graphs for oral presentations.

4.5 LEGENDS AND KEYS

The best graphs are self explanatory, with lines labelled on the graph rather than the reader having to refer to a key in the caption. However, there is some conflict here with the principle of reducing clutter (which obscures data); use your common sense. For choice of fonts, see below. Put the key within the graph area where possible as this optimises the use of space on the page or screen—see Appendix 1 for examples and how to avoid the problem of filling valuable data space with unnecessary clutter. There is no need for a box around the key unless there is potential confusion between real data and the symbols being explained.

4.6 TYPE FONTS

The best font for words and numbers anywhere in the graph (axis labels, titles and key) is a sans-serif font, such as Helvetica or Arial (Peterson 1999). Research on readability has shown that the serifs (the slight projections on strokes) on letters make letters more easily read if they are small or a long way away—which applies to solid pieces of text (such as in newspaper print). However, sans-serif type is easier to scale (i.e. it looks just as good when magnified or reduced, whereas the serifs can distort the type). Also, using sans-serif font distinguishes the labels clearly from surrounding explanatory text, which is usually in serif font.

Be aware that in some fonts, capital I, lower case L, and the number 1 are indistinguishable, as are the letter O (oh) and 0 (zero); so don't use these letters for graph symbols that have to be discriminated. Use normal upper and lower case rather than ALL CAPITALS or SMALL CAPITALS, again for maximum readability. Although many font choices are a matter of personal preference, usually reflecting familiarity with one over the other, make sure to always check the style requirements of the target journal when preparing for publication.

4.7 SYMBOLS, LINES AND FILLS

4.7.1 Symbols or lines

x–*y* graphs offer a choice between using symbols, or lines, or both.

Symbols-only graphs are best for showing an overall trend: they tend to downplay short fluctuations, as points are not linked to each other.

Lines-only graphs are best for showing a pattern of regular change when the exact time of sampling of individual values is not important; or when there are a large number of values, and symbols would obscure the exact line position (symbols are usually larger than the line thickness).

Plotting lines as well as symbols shows both the pattern from one data point to the next, and the values at particular points. We generally prefer this option (e.g. as shown in Fig. 17), although The Council of Biology Editors (Peterson 1999) suggests that frequency polygons should not have data points but only lines connecting the midpoints of histogram data.

4.7.2 Line types

There are few different line types available (solid, dashed, dotted, broken, etc). They tend to be easily confused, so it may be better to differentiate between data by using different symbols connected by superimposed solid lines: this is what the Council of Biology Editors (Peterson 1999) recommends. You could also use different line colours: this works very well for oral presentations but the difference may get lost when printing from a computer screen, for further replication of handouts, etc. (see section 2.2). Where data points are missing, either leave a gap in the solid line or join across the gap with a dashed line.

Fitting lines is usually inferior to joining the data points. If you have a regression line, or another complex equation, you can show the fitted line, especially if the equation has some biological meaning or interpretation. We recommend that the line's equation and statistics for its fit be provided in the caption (rather than on the graph) if possible—again, to reduce clutter, and to make sure it is properly typeset (e.g. superscripts can create a problem in graphics software).

Smoothed curves often seem little more than computer freehand doodling. They imply you have more data than you really have, and are best avoided. Figure 2 is a good example of a fitted line; Fig. 19 is a very bad one.

4.7.3 Symbol types

In your choice of symbols, emphasise visibility, discrimination and interpretation. Watch out for silly combinations: e.g. juxtaposed o and + can suggest church or female symbols which lead to unintentional distractions.

To optimise visibility, make the symbols a little larger than any text, and dark (bold) enough for good contrast.

Discrimination between multiple points where there may be overlap is best achieved by using hollow symbols: a circle is best; a triangle is next best; then a square. Only use filled symbols if you have few data points and little overlap; in this situation, filled symbols may be preferred as they stand out more clearly. Hollow symbols can still be identified when you have many overlapping points. Hollow symbols could also be used for pre-treatment or control, with the solid symbols giving more emphasis to treatment data.

Hollow symbols, especially circles, tend to be clearer than filled symbols when jittering is applied, i.e. when identical data points are slightly offset from each other; in technical terms, when you add a little 'noise' to the data, allowing overlying or overlapping points to be seen more clearly (Fig. 32).

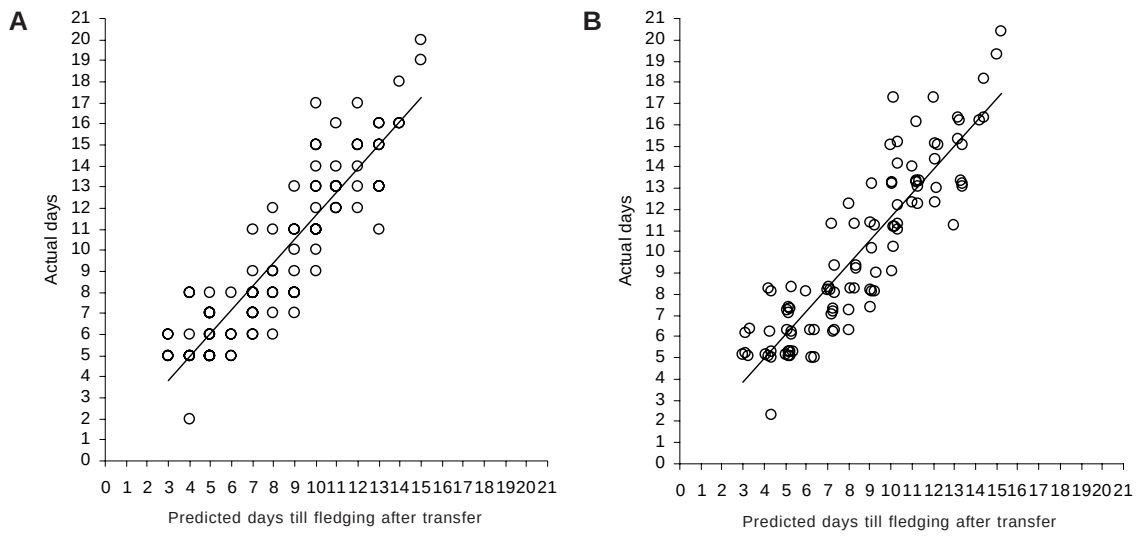


Figure 32. Jittering is often a good method for revealing hidden data points when graph points overlay each other. A has multiple points shown by just one symbol because they are recorded as whole numbers, and overlay each other in a number of cases. There are 100 values altogether, but there are 55 points on the panel, with up to 6 points on top of each other. B, which was published, represents the same data with a small amount of 'noise' added to each data cell, using the RAND( ) function in Excel. The line represents a simple linear trend through the original data. In Excel you need to use the RAND( ) function to create a new set of slightly more 'noisy' data to plot: take care that the background colour for the symbols is set to 'no color'. Note that the graph would have been better with many fewer tick marks, for example at 10-day intervals.

Original caption: Relationship between predicted and actual fledging dates of 100 fairy prions transferred to Mana I. in Jan 2004; $r^2 = 0.81$, $P < 0.001$. Multiple data points 'jittered' to reveal true sample size. All birds fledged within 7 days of their predicted fledging date.

4.7.4 Fill patterns and shading

Avoid Moiré effects, i.e. fine lines that seem to shimmer and thus hamper interpretation; they may also reproduce poorly in print. Be particularly careful with fill patterns in bar graphs: the bars may seem to bend.

If printing in black and white, it is much better to use cross-hatching or grey dots to indicate different categories. Note that grey shading does not survive computer printing or multiple generations of photocopying very well.

Again, the type of shading appears to have a strong personal preference component. Peterson (1999) recommends that clear bars should not be used as they suggest absence of data. However, DOC Science Publishing recommends for bar charts: open bars for the category that takes up most of the graph area; solid black for the category with the smallest area; and horizontal and vertical hatching for intermediate categories. If there are more categories still, then you probably need an extra graph.

Science journals published by the Royal Society of New Zealand allow shading provided that the categories are sufficiently differentiated, i.e. 20%, 40% and 60% tone (grey) can accompany white (clear) and solid black, but steps of 5% are too small.

Printing in colour can depict vivid differences, which is good, but there are problems (also see section 2.2). Colour printing is much more expensive, and a publisher may disallow it for journal articles (or make the author bear the extra charge). For categorical data, colours can be very good for indicating groups. However, numerical data have another problem: there is no agreed sequence of colours to represent low to high values. The least ambiguous numerical colour sequence is blue to red as a temperature analogy. Also, note that some readers will be colour blind (5%–10% of the population, particularly males, with red-green colour blindness being the most common type) so do not rely on red versus green for your key comparison. Finally, remember that a colour graph may be copied, or printed from a pdf file, into black and white at some stage, which may well render it completely unintelligible.

4.8 ERROR BARS

Error bars may be used for two main purposes:

1. They indicate the spread of data, such as by showing the standard deviation around the mean. Spread is generally much better illustrated by showing all the data points when the number is small, or with the box-and-whisker graph or box plot discussed in section 3.4. The standard deviation can provide a very good summary of spread when the data are normally distributed, but the usefulness of such an assumption should be checked. In any case, the standard deviation is probably best listed in the text rather than plotted.
2. Error bars indicate the reliability of the mean. For this, you can use a confidence interval (CI), or the standard error of the mean (SEM). The SEM is smaller, so it looks better, but it is rather less meaningful. SEMs are deceptively small when the sample size is very small because the multiplier of SEM to give CI half-width is much larger than 2 (the value achieved when sample size is 30 or more). We recommend that you stick to ± 95% CIs, but either way you *must state in the caption AND methods section which type of error* you are illustrating.

As error bars are usually symmetrical (unless means and errors have been back-transformed), you may wish to illustrate only one side, i.e. values + CI rather than ± CI. Make this clear in the caption.

The nature of comparisons of interest, together with the nature of the data, need to be considered when planning which error bars to present, especially if the data are correlated. For example, when measurements have been made twice on the same sample unit, CIs for the differences between measurements are likely to be of much greater interest than CIs for the mean of measurements for each unit.

4.9 SUPERIMPOSED VERSUS JUXTAPOSED

The limits on line and symbol discrimination mentioned above can mean that it is hard to superimpose many parts of a dataset on the same graph. In many instances, it may be better to show multiple panels, all to the same scale (Fig. 25).

Multiple panels work best when all use the same scale on the *x*- and *y*-axes; next best is equal units on axes (as in Fig. 29) where the *y*-axes are not identical, to save space and hence maximise the area occupied by data. Put the panels close together to maximise the data space in your graph area.

4.10 CAPTIONS AND HEADINGS

Make sure you describe every key element of the graph and give the main conclusions in the figure caption (or legend). Captions are much overlooked by authors, probably because they are usually written last, as an afterthought. But remember, when potential readers leaf through a paper, they usually look at the figures, and read their captions, first. Hence these should contain all essential background information (full species names, site, etc.), while still being brief.

If a complex graph needs careful interpretation you could 'talk the reader through it', so that the caption becomes a side story. Although this can sometimes work well (see Box 1, section 2.3.3), some journals will not like this and may put the graph with extended caption in an appendix instead.

In a thesis, report or science paper, captions are usually typeset separately, so do not design a caption as part of the graph. A graph heading takes up valuable data space, so avoid them in published work. Indeed, it may lead to inconsistencies of terms between caption and graphs.

Only when you prepare a graph for oral presentation is it helpful to have a short heading above the graph to help your audience; but this should be no more than a few words.

4.11 CHARTJUNK

'Chartjunk' is Tufte's (1983) term for any needless detail on a graph. The assumptions behind chartjunk are that graphs are boring and the audience is stupid, so we have to try to entertain and distract them. Chartjunk is often used to try to disguise low data density. You should expunge chartjunk at all costs.

Although we do not present any extreme examples of chartjunk, mild examples include the mountain decorations in Fig. 7, and the mock 3D pie representation in Fig. 8. As a general rule, avoid:

- **All false 3D representations** such as 'mock 3D' pie graphs (Fig. 8) and 'mock 3D' bar graphs (Fig. 23); these are particularly rife in data analysed with Excel.
- **Most fill patterns.**
- **All background fills.**

- **Most illustrations** on the graph (pictures of oil barrels, etc.). An exception to these rules is shown by Silvertown (1987), where each graph has a small line drawing of the organism to which the data relate; this may be helpful to the reader.
- **Corporate slogans and crests.** They add nothing to the understanding of the data. Although fine on an opening and closing slide of a presentation, they are usually totally redundant distractions.

5. Computer software for graphs

There are many options for producing graphs by computer. Graphing by hand may also be an option—but generally only for small datasets. The three main types of software available for graphing are:

- General spreadsheet packages, typically Microsoft Excel
- Statistical software packages, like SPSS or S-PLUS
- Specialist graphing packages, like SigmaPlot

At DOC, only the first two options are available with the packages mentioned, so we will concentrate on the strengths and weaknesses of these. But first let us look at storing the source data.

5.1 DATA SOURCING

Data are best accumulated in electronic form from very early on in a project. Unless data capture makes use of special software, Microsoft Excel is usually best for collation, checking and further manipulation, including exporting to graphs and other software packages.

To create the desired graph, you often need to reorganise the data in the spreadsheet—especially for graphs that Excel is not set up for, like multiple values of one continuous variable against a categorical variable. The ability to derive reorganised data is greatly enhanced if the raw data are in a standard format, where each line of data represents an observation, and the columns represent variables for each observation. The same or very similar identifying information may need to be repeated in blocks of observations. This means that new data are added at the bottom of the sheet, and datasets tend to be much longer than they are wide. This method of recording is in contrast to a common practise of adding new data in columns across the page. The key advantage is that Excel's excellent PivotTable facility can then be used to reorganise or summarise the data in just about any possible way. More information and guidelines on entering data in Excel can be found at a website created by the University of Reading[1] or, within DOC, in notes from a course 'Using Excel to Enter, Manage and Explore Data' (Cox 2001).

[1] www.ssc.rdg.ac.uk/publications/guides/topsde.html, viewed 21 July 2005.

5.2 MICROSOFT EXCEL GRAPHS

Excel has a great strength—general accessibility. Most people have access to and familiarity with Excel and it is often where the data are. As demonstrated in Appendix 1, with a little work it is possible to produce publication-standard graphs of various types, particularly bar / column graphs and x–y scatterplots.

Excel has major disadvantages as well. The default settings and graphs have undesirable features, such as an apparent preference for bar graphs and pie charts, and a tendency to use too many cluttering lines and unnecessary 3D effects. Some types of plots are difficult or near-impossible to produce in Excel; for example, there is no box-and-whisker plot option provided, although a workaround to produce one has been developed by DOC (see section 3.4). It takes considerable work and knowledge of Excel to create plots of multiple values of one continuous variable against a categorical variable, or to mimic the dot plots of Cleveland (1994).

In Excel, each graph needs to be handcrafted, which can become a major issue when creating multipanel displays, where you often need identically formatted graphs with different data. The best—perhaps the only—way to get these is to create one graph that meets all your specifications exactly. This graph can then be copied the number of times needed, and then the data reference for each graph can be changed. If some overall aspect needs changing, typically all graphs need re-creating. By holding the Alt key down while moving or changing the size of a graph, the graph only moves or changes sizes in steps, to fit with the Excel gridlines.

In versions from about 2000 onwards, Excel includes PivotCharts, which allows you to make a graph from a pivot table. Excel selects the graph it thinks is best, with very limited flexibility. Copying values-only from the pivot table (using Edit > Paste special, and selecting either Values or Paste Link) to a conventional table allows the usual Excel operations.

With care, Excel can create excellent graphs—most of the graphs in this publication were produced in Excel—but its limits are easily reached. If you already have the need to use a statistical package, then SPSS and S-PLUS become attractive alternatives. Their use requires learning new tools and, typically, a different approach: less point-and-click, more menus, and sometimes programming code must be written.

5.3 STATISTICAL PACKAGES: SPSS AND S-PLUS

Almost all modern statistical packages include graph-producing facilities, and SPSS and S-PLUS are no exception: they have extensive graphing facilities. Statistical packages generally require data in the rows-as-observations and columns-as-variables format mentioned above. (This is another reason for using this approach in your spreadsheets; you can just copy and paste from the spreadsheet to the statistical package, or use the data import facilities provided by the package.)

SPSS has a full Microsoft Windows interface to its graphing: use the Graphs menu on the menu bar. Data are readily imported from Excel. Most of the graphs discussed here are available both directly on the Graphs menu, including full box plots and scatterplot matrices. Graphs > Interactive leads to another set of graphs, which are more flexible and allow for panel variables (allowing the creation of multipanel graphs).

S-PLUS also allows the creation of graphs from Windows using point-and-click. Data are readily imported from Excel or almost any other format. S-PLUS has the advantage of allowing linkage to the original Excel data to create graphs (or do other things) without creating a permanent copy of the data. The data continue to be accessed from their Excel source, and graphs reflect any changes that have been made there. Although the S-PLUS–Windows interface is not always easy to use effectively, S-PLUS comes fully into its own for graphs once you are familiar with the very powerful computing language S that underlies the package (see Appendix 2). It takes a considerable time investment to learn S, which may be recouped only with regular use. S-PLUS has full and sophisticated implementation of multipanel graphs—originally developed by Cleveland himself (Cleveland 1993). These are available using the menus, or by programming. The computer package R (cran.r-project.org) is an S-PLUS look-alike, without point-and-click graphs or an Excel linkage wizard, but with the major advantage of being free. It has excellent programmable graphing, including multipanel options.

5.4 SPECIALTY GRAPH PACKAGES

There are many dedicated graphing packages available, such as SigmaPlot. However, they are not available on the DOC computer network, and they carry the additional overheads of users needing to learn new tools and having to move data around. The output may also not be compatible with publishing packages such as Adobe PageMaker, meaning that graphs may need to be scanned before publication. The majority of DOC's graphing needs can be met using Excel or a statistical package.

6. Presentation medium and production strategy

6.1 MEDIUM

Once you have decided which type of graph best illustrates the point you are making from the data, spend some extra time thinking about the medium in which the graph is to be presented. Many people use the same graph over and over again, irrespective of whether it is presented in a paper, in an oral presentation or on a web page. In the latter two media, any vertical type should be avoided, as the graph cannot easily be turned around to assist reading; also, these two media do not allow detailed study, so the typefaces and data symbols should be easy to read, quick to interpret and not contain any unessential detail. Gridlines, formatted to be in the background, may help interpretation here, whereas in graphs for the print medium these tend to create undesirable clutter. When using log scales with few orders of magnitude, gridlines may be useful to show 'logginess' through the changing spacing between tick marks. Some other differences are shown in Table 2.

TABLE 2. CHARACTERISTICS OF GOOD GRAPH DESIGN FOR ORAL PRESENTATION (e.g. IN POWERPOINT) AND FOR PUBLICATION IN PRINT MEDIA.

The aim is for the audience to find it easy to see the key message, not for the author to impress the reader with unnecessary embellishments.

CHARACTERISTIC	ORAL PRESENTATION	PUBLICATION
Optimum use of space	Design for maximum emphasis on data, no clutter	
Type font sans-serif	Yes (make larger rather than bold for emphasis); e.g. Arial	
Type size	Max. 30 characters / line, 15 lines / slide; minimum 24 point type (e.g. see Wainer 1977)	Depends on final printed size; no smaller than 6 points at the final published size
Vertical / oblique type	No	OK for *y*-axis label, but not for data labels
Title caption	Yes (short)	No
Background colour	OK	No
Colour bars / lines / data points	Maybe	Not unless need / cost can be justified
Box around graph	Maybe	No, unless it is journal's style
Gridlines	OK	Generally not
Key	Maybe	Only where essential
Location of key	Where suitable	Within graph area if there is room
Box around key	Not unless there is confusion with actual data points	

6.2 ITERATION AND IMPROVEMENT

Iteration is trying out various ways of presenting the data (e.g. the different panels of Fig. 22). This should be done for two reasons: to search for patterns in the data, and to find the best way to present the data (e.g. does the graph work on the screen, or on the page?).

The aim of iteration is to produce a good graph, i.e. one that has:

- High information content
- Data that stand out clearly
- User-friendly labelling (minimum translation and reference to the key)
- Important points presented high on the scale of decoding accuracy
- No redundant or misleading elements

As with written and oral presentation, test the draft on your colleagues and a sample audience and see if its key points have come across well. Use the photocopier to reduce your graph to both half the size and twice the size and see if it is still legible and pleasing to the eye.

All of this will improve the impact of graphs in your thesis, paper or conference talk and will result in a much more receptive audience of people who are interested in your work.

7. Acknowledgements

Our guidelines are inspired by the work of William Cleveland and Edward Tufte, and by many errors and deficiencies noted in the scientific and popular literature (with an emphasis on New Zealand material). The text is based on a lecture handout from the University of Canterbury (DK), DOC Science Publishing editorial guidelines (JJ), and material prepared for a series of Graphs workshops in DOC (IW and JJ).

We thank the generations of published authors and students that wittingly or unwittingly have contributed to improvement of the lecture notes and this guide; the participants in the DOC Graphs workshops in 2003, and the editorial members of the DOC Science Publishing team (Lynette Clelland, Ian Mackenzie, Helen O'Leary, Geoff Gregory, Amanda Todd), all of whom have contributed ideas, examples and suggestions for improvement.

Special thanks to Lindsay Rollo, Helen O'Leary and Sue Hallas who ran their critical eyes over the submitted draft; as did two anonymous referees.

8. Sources

8.1 REFERENCES

Bigwood, S.; Spore, M. 2003: Presenting numbers, tables and charts. Oxford University Press, New York. 144 p.

Cleveland, W.S. 1984: Graphs in scientific publications. *The American Statistician 38*: 261–269.

Cleveland, W.S. 1993: Visualising data. Hobart Press, Summit, New Jersey. 360 p.

Cleveland, W.S. 1994: The elements of graphing data. 2nd edition. Hobart Press, Summit, New Jersey. 297 p.

Cleveland, W.S.; McGill, R. 1985: Graphical Perception and Graphical Methods for Analyzing Scientific Data. *Science 229*: 828–833.

Cox, N. 2001: Using Excel to enter, manage and explore data. Course prepared for Department of Conservation. Unpublished AgResearch report. 70 p.

Hosking, G.P. 2003: Rata litterfall and canopy condition, Whirinaki Forest Park, New Zealand. *DOC Science Internal Series 103*. Department of Conservation, Wellington, New Zealand. 12 p.

Imber, M.J.; West, J.A.; Cooper, W.J. 2003: Cook's petrel (*Pterodroma cookii*): historic distribution, breeding biology and effects of predators. *Notornis 50*: 221–230.

Peterson, S.M. 1999: Council of Biology Editors Guidelines: Editing science graphs. Council of Biology Editors, Reston, VA (USA). 34 p.

Robbins, N.B. 2005: Creating more effective graphs. Wiley, Hoboken, New Jersey. 402 p.

Scott, D.W. 1992: Multivariate density estimation: theory, practice, and visualization. Wiley, New York. 376 p.

Silvertown, J. 1987: Introduction to plant population ecology. Longman. 320 p.

Stevens, S.S. 1957: On the psychophysical law. *Psychological Review 64*: 153–181.

Tufte, E. 1983: The visual display of quantitative information. Graphics Press, Cheshire, Connecticut. 197 p.

Wainer, H. 1977: Visual revelations. Graphic tales of fate and deception from Napoleon Bonaparte to Ross Perot. Copernicus, Springer Verlag. 180 p.

8.2 ADDITIONAL RESOURCES

Anonymous 1997: Write, Edit, Print. Style Manual for Aotearoa New Zealand. Commonwealth of Australia, and Lincoln University Press. (Pp. 297–303.)

Booth, W.C.; Colomb, G.G.; Williams, J.M. 1995: Communicating evidence visually. Pp. 175–200 (Chapter 12) in: The craft of research. University of Chicago Press, Chicago.

Briscoe, M.H. 1996: The journal figure. Pp. 103–116 (Chapter 7) in: Preparing scientific illustrations. A guide to better posters, presentations and publications. 2nd edition. Springer-Verlag, New York.

Davis, M. 1997: Presenting data. Pp. 105–111 (Chapter 11) in: Scientific papers and presentations. Academic Press, London.

DOC (Department of Conservation). Guide for authors of DOC science publications. Department of Conservation website > Publications > Science and Research > Guide for Authors www.doc.govt.nz/Publications/004~Science-and-Research/Guides-for-authors/index.asp (viewed 15 July 2005).

Maindonald, J.H. 1992: Statistical design, analysis, and presentation issues. *New Zealand Journal of Agricultural Research 35*: 121–141.

Royal Society of New Zealand journals: www.rsnz.org/publish/instruct_auth.php (viewed 15 July 2005).

Statistics New Zealand: www.stats.govt.nz/about-us/policies-and-guidelines/data-use/graphics-guidelines.htm (viewed 21 July 2005).

Tufte, E. 1990: Envisioning information. Graphics Press, Cheshire, Connecticut. 126 p.

Tufte, E. 1997: Visual Explanations: Images and Quantities, Evidence and Narrative. Graphics Press, Cheshire, Connecticut. 156 p.

Wallgren, A.; Wallgren, B.; Persson, R.; Jorner, U.; Haaland, J-A. 1996: Graphing statistics and data. Creating better charts. SAGE Publications, Newbury Park, California. 94 p.

8.3 SOURCES OF FIGURES

Figure 1—Allan, R.B.; Bellingham P.J.; Wiser, S.K. 2003: Forest biodiversity assessment for reporting conservation performance. *Science for Conservation 216*. Department of Conservation, Wellington, New Zealand. 49 p. (Fig. 8.)

Figure 2—Rogers, D.M. 1996: Control, demography, and post-control response of heather in the central North Island: Part 2. *Science for Conservation 29*. Department of Conservation, Wellington, New Zealand. 35 p. (Fig. 2.)

Figure 3—Tufte 1983, p. 28. Also at: www.fi.uu.nl/wiskrant/artikelen/hist_grafieken/begin/images/planeten.gif.

Figure 4—Tufte 1983, p. 34. Also at:www.math.yorku.ca/SCS/Gallery/images/playfair-wheat1.gif.

Figure 5—IW originals.

Figure 6—A: Espiner, S.R. 1999: The use and effect of hazard warning signs. Managing visitor safety at Franz Josef and Fox Glaciers. *Science for Conservation 108*. Department of Conservation, Wellington, New Zealand. 40 p. (Fig. 19.)
B: IW original.

Figure 7—DOC (Department of Conservation) 2003: Department of Conservation Statement of Intent 2003–2006. 88 p. (P. 12.)

Figure 8—A: Green, W. 2004: The use of 1080 for pest control. A discussion document. Animal Health Board and the Department of Conservation. 60 p. (Fig. 7.)
B: IW original.

Figure 9—Murphy, D.J.; Kelly, D. 2001: Scarce or distracted? Bellbird (*Anthornis melanura*) foraging and diet in an area of inadequate mistletoe pollination. *New Zealand Journal of Ecology 25*: 69–81. (Fig. 6.)

Figure 10—Dilks, P.; Onley, D.; Kemp, J. 1998: Ecology and breeding of Chatham Island tui. Progress report October 1996–June 1997. *Science for Conservation 88*. Department of Conservation, Wellington, New Zealand. 39 p. (Fig. 3.)

Figure 11—Allibone, R. 2000: Water abstraction impacts on non-migratory galaxiids of Otago streams. *Science for Conservation 147*. Department of Conservation, Wellington, New Zealand. 43 p. (Fig. 4.)

Figure 12—A: ANH Smith and IW original.
B: Smith, A.N.H; Westbrooke, I.W. 2004: Changes in bird conspicuousness at Pureora Forest. *Notornis 51*: 21–25. (Fig. 2; corrected for misalignment of lines in original publication.)

Figure 13—A: IW original.
B: Westbrooke, I.M.; Etheridge, N.D.; Powlesland, R.G. 2003: Comparing methods for assessing mortality impacts of an aerial 1080 pest control operation on tomtits (*Petroica macrocephala toitoi*) in Tongariro Forest. *New Zealand Journal of Ecology 27(2)*: 115–123. (Fig. 2.)

Figure 14—Baker, A.N.; Smith, A.N.H.; Pichler, F.B. 2002: Geographical variation in Hector's dolphin: recognition of new subspecies of *Cephalorhynchus hectori*. *Journal of the Royal Society of New Zealand 32*: 713–727.www.rsnz.org/publish/jrsnz/2002/036.pdf (Fig. 3; adapted.)

Figure 15—Sessions, L.A.; Kelly, D. 2001: Heterogeneity in vertebrate and invertebrate herbivory and its consequences for New Zealand mistletoes. *Austral Ecology 26*: 571–581. (Fig. 3.)

Figure 16—Jasperse, J.; Westbrooke, I. 2003: Report on national series of workshops by Ian Westbrooke and Jaap Jasperse: Graphs for data exploration, presentation and publication (April–July 2003). Unpublished report to DOC Management. 6 p. (P. 3.)

Figure 17—Stewart, G.H. 1995: Stand development in the red / silver beech and mixed beech forests of north Westland. *Science for Conservation 8*. Department of Conservation, Wellington, New Zealand. 14 p. (Fig. 2.)

Figure 18—A.W. Robertson, J.J. Ladley and D. Kelly: unpublished graph.

Figure 19—Craddock, P.; Clout, M. 2001: Environmental risks of using brodifacoum at managed sites: toxicity and patterns of bait consumption by invertebrates. Unpublished report to Department of Conservation. Auckland Uniservices Ltd. (Fig. 1.)

Figure 20—A: Wright, M. 1998: Ecotourism on Otago Peninsula. Preliminary studies of yellow-eyed penguin (*Megadyptes antipodes*) and Hooker's sea lion (*Phocarctos hookeri*). *Science for Conservation 68*. Department of Conservation, Wellington, New Zealand. 39 p. (P. 33.)
B: Robertson, H.A.; Westbrooke, I.M. 2005: A practical guide to the management and analysis of survivorship data from radio-tracking studies. *DOC Technical Series 31*. 47 p. (P. 20.)

Figure 21—D. Kelly and P.J. Grubb: unpublished graph. [Kelly, D.; Grubb, P.J. submitted: Population fluctuation in short-lived chalk grassland plants. II. Spatial and temporal stability over 14 years. *Journal of Ecology*. (Fig. 1.)]

Figure 22—Koenig, W.D.; Kelly, D.; Sork, V.L.; Duncan, R.P.; Elkinton, J.S.; Peltonen, M.S.; Westfall, R.D. 2003: Dissecting components of population-level variation in seed production, and the evolution of masting. *Oikos 102*: 581–591. (Fig. 4.) (Panel F only; others are unpublished variations by the authors.)

Figure 23—Unpublished graph from report to Stoat Research Technical Advisory Group.

Figure 24—Unpublished graph drawn by A.W. Robertson, from data of D. Kelly, A.W. Robertson and J.J. Ladley.

Figure 25—Graphs unpublished, but for more information on the data see: McAlpine, K.; Timmins, S.M. 2002: Poster: The Effect of fire on bone-seed and gorse germination. DOC Science poster no. 56.

Figure 26—Unpublished graph, but see: Baker, A.N.; Smith, A.N.H.; Pichler, F.B. 2002: Geographical variation in Hector's dolphin: recognition of new subspecies of *Cephalorhynchus hectori*. *Journal of the Royal Society of New Zealand 32*: 713–727. (Fig. 2.)

Figure 27—Wilmshurst, J.M. 2003: Establishing long-term changes in takahe winter feeding grounds in Fiordland using pollen analysis. *Science for Conservation 228*. Department of Conservation, Wellington, New Zealand. 25 p. (Fig. 7.)

Figure 28—Unpublished parakeet data from May 2003 Christchurch graphs workshop.

Figure 29—P.J. Grubb and D. Kelly: unpublished graph. [Grubb, P.J.; Kelly, D. submitted: Population fluctuation in short-lived chalk grassland plants. I. The regeneration niche and relative abundance. *Journal of Ecology*. (Fig. 2.)]

Figure 30—Schoeps, K. 2000: Metapopulation dynamics of the coxella weevil (*Hadramphus spinipennis*) on the Chatham Islands. *Science for Conservation 134*. Department of Conservation, Wellington, New Zealand. 37 p. (Fig. 10.)

Figure 31—Graphs unpublished, but for more information on the data see: Kelly, D. 1994: Demography and conservation of *Botrychium australe*, a peculiar sparse mycorrhizal fern. *New Zealand Journal of Botany 32*: 393–400.

Figure 32—Miskelly, C.; Gummer, H. 2004: Third and final transfer of fairy prion (titiwainui) chicks from Takapourewa to Mana Island, January 2004. Department of Conservation, Wellington, New Zealand. 40 p. (Fig. 7.)

Figures in Appendix 1:

Figure A1.1—JJ original.

Figure A1.2—Hosking, G. 2003: Rata litterfall and canopy condition, Whirinaki Forest Park, New Zealand. *DOC Science Internal Series 103*. Department of Conservation, Wellington, New Zealand. 12 p. (Fig. 1; variations are JJ originals.)

Figure A1.3—Imber, M.J.; West, J.A.; Cooper, W.J. 2003: Cook's petrel (*Pterodroma cookii*): historic distribution, breeding biology and effects of predators. *Notornis 50*: 221–230. (Fig. 2.)

Figure A1.4—JJ original.

Figure A1.5—JJ original.

Figures in Appendix 2—Amanda Todd and IW originals.

Figures in Box 1—IW originals.

Appendix 1

MAKING MICROSOFT EXCEL DEFAULT GRAPHS SUITABLE FOR SCIENTIFIC PUBLICATION

By Jaap Jasperse

Microsoft Excel is an excellent package to create a variety of different graphs from data tables. Unfortunately, the package clearly betrays its origins in accounting. It tends to produce graphs that have all the hallmarks of bad presentation in science: lots of clutter (boxes, gridlines, shading, unnecessary and large, poorly placed legend, and inappropriate type), and a tendency to emphasise bar and pie graphs over *x*–*y* graphs; to propose 'line' graphs to have equal spacing on the *x*-axis rather than reflecting the numerical spacing of the values; and to encourage lurid colours and 'mock 3D' option, etc.

In this appendix, we create some graphs using the default settings of Chart Wizard, and then work through a series of changes that will make the graphs suitable for publication in a science journal. It reflects the sentiment expressed by Bingwood & Spore (2003): 'If you are working with a popular spreadsheet, such as Excel, you can produce good tables and graphs only with rigorous editing and deletion; you must ruthlessly remove features that the program provides'. Microsoft Excel graph terminology is shown in Fig. A1.1, and the following conventions are applied:

- In Excel it is usually best to change one of the characteristics of a graph element by bringing the cursor over it, then right-mouse clicking on the element (or left-click if you have the buttons swapped). This will bring up a named menu with various options.
- The descriptions below assume that you know where to click (experiment!), then indicate by the > symbol which option to select.
- Many of the functions described using menu items can also be done by just selecting the item and then deleting it, dragging it or double-clicking on it. To cancel a selection, press ESC. Whichever method you prefer, the result should be a much better graph!

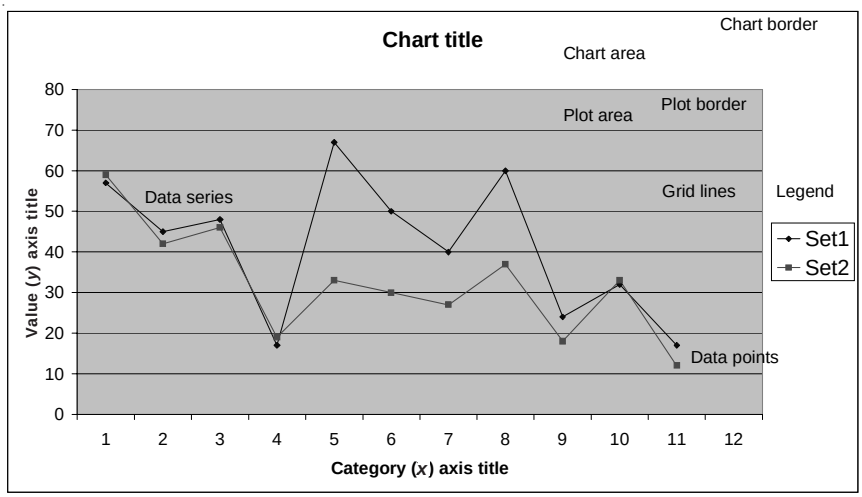


Figure A1.1. A generic graph showing Microsoft Excel terminology for parts of graphs (charts).

Example A1.1 Vertical bar chart of measurements taken for independent categories

TABLE A1.1. DATA FOR FIGS A1.2A–C (OPPOSITE).

Total leaf litter (g) from individual trees sampled three times 2-monthly (3 × 2-monthly) and twice 3-monthly (2 × 3-monthly). Crown Categories (CC) for trees are:
L, Long; S, short;
H, Healthy;
I, intermediate;
U, unhealthy.
(After Hosking 2003.)

CC	3 × 2-monthly	2 × 3-monthly
L / H	57	59
L / H	45	42
S / H	48	46
S / H	17	19
L / I	67	33
L / I	50	30
S / I	40	27
S / I	60	37
L / U	24	18
S / U	32	33
S / U	17	12

Consider the data in Table A1.1, of total litterfall for a six-monthly sampling period, by individual tree and sample interval (Hosking 2003).

The Chart Wizard in Excel allows us to create a graph, complete with title and key, very quickly from this—which is a major advance on manual creation. It is merely a matter of selecting the table in the spreadsheet, pressing the button for the ChartWizard, and answering a few questions. It takes no more than a minute (Fig. A1.2A).

However, for publication purposes, the resulting default graph shows serious deficiencies: unnecessary clutter, tiny type, and reduction of effective space for the data it purports to illustrate.

The graph can be made suitable for publication as follows:

1. First, **remove the title** (under Chart Options > Titles or select the title and delete it).

   Figures are usually published with a separately typeset explanatory caption; the graph should therefore concentrate on presenting data only. Indeed, more often than not titles end up being in a contrasting style to, or contradicting, the caption. Only when the figure is used for, say, a PowerPoint presentation or an overhead transparency can a title have any use—but it is better added later than at the start.

2. Next, **remove all boxes** from around the:

   • Chart area: Format Chart Area > Patterns > Border > None

   • Plot area: Format Plot Area > Patterns > Area > None

   • Key: Format Legend > Patterns > Border > None

   These extra lines only produce clutter that has nothing to do with the data.

   As noted before, many of the menu choices described can be made more easily by just selecting the item and then deleting it, dragging it or double-clicking on it.

3. Now **improve the key** (legend).

   First, move it to above the graph from the side, so that much more width is available for plotting the data horizontally: Format Legend > Placement > Top, or select and drag it.

   Next extend the graph upwards by clicking on the graph area once, and then dragging the upper limit up to where it overlaps the key.

   Then move the key to a corner area where it does not overlap any data.

4. Next, **remove unnecessary graph items** as follows:

   • Gridlines: Chart Options > Gridlines > (untick all), or select and delete

   • Background shading: Format Plot Area > Patterns > Area > None, or click on the background and press 'Delete'

   This will result in Fig. A1.2B, which makes much more effective use of space and has fewer distracting elements.

Figure A1.2. A graphical illustration of the data in Table A1.1.

A, Default Excel settings with title, key and colour.

B, Improving effective graph space by removing the title (a separate caption is better) and moving the key; the gridlines create unnecessary clutter.

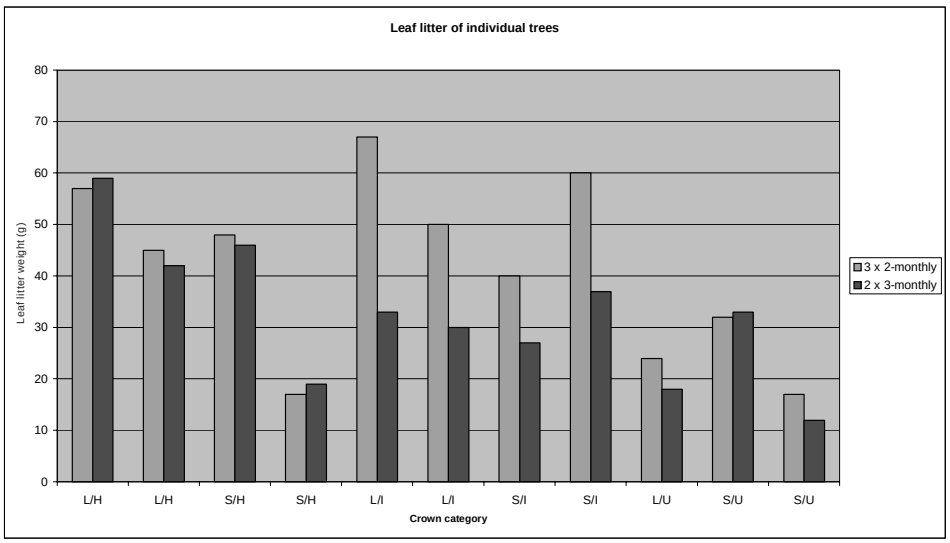
C, Tidy, clear and clutter-free graph.

Note that A–C are reproduced here deliberately at the same scale as produced in Excel (for best visual effect, *y*-axes are usually best vertically aligned on a page).

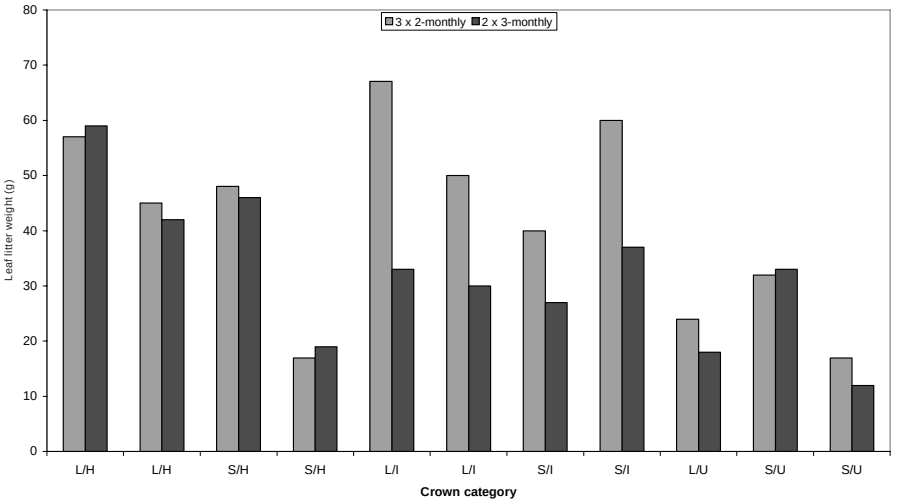
Fig. A1.2C allows reduction to at least 50% without loss of clarity (Fig. A1.2D), but the font size may need further adjustment.

Original caption: Total rata leaf litter in two-monthly (grey) and three-monthly (black) collections. Crown categories: L large, S small, H healthy, I intermediate, U unhealthy.

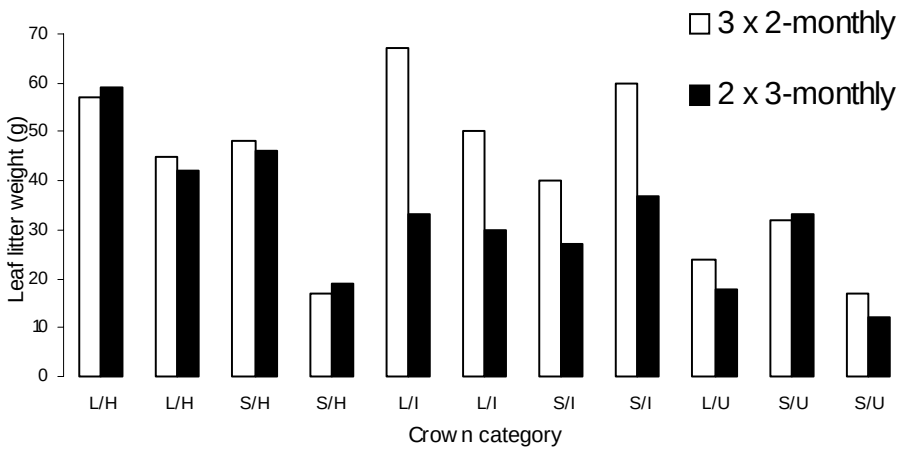
A



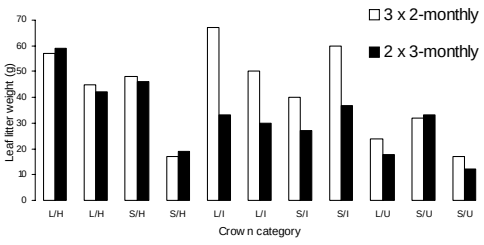
B



C



D



5. **Remove colour** from bars:
   - Use white (blank) for one set: Format Data Series > Area > None
   - Use solid black for the other: Format Data Series > Area > (black square, palette top left)
6. Finally, **improve the labels**.

   Remove bold type; instead make the type larger to allow greater reduction without loss of legibility: Format Axis > Title > Font > Regular (and increase point size somewhat).

   Sometimes it is better to have fewer labels on the axes, but in this instance all horizontal labels are needed. The *x*-axis may be removed as there is no relationship between the pairs of bars: Format Axis > Axis > None.
7. We can make the vertical axis more efficient by going to the true (rounded) maximum value (Excel usually goes higher than that, in order to draw the (unnecessary) boxes around the data). Format Axis > Scale > Maximum (here 70), Minor unit (here 10), Minimum 0.

   The key can now be further tidied up to show items one above another; place it so that it 'squares off' the figure at the top right-hand side.

The result: Figure A1.2C is a clear and clean graph with emphasis on the data, minimum clutter, and which reproduces well in black and white. With its separate caption, it has lost none of the original information shown in the ugly and inefficient graph produced by the Excel Chart Wizard default settings.

This graph can now be reduced to about 50% size (Fig. A1.2D) without losing clarity and legibility, provided the font for labels is big enough. (The minimum font size suitable for printing is 6 points. Note that Excel may produce label errors by truncating type.) Science journals generally reproduce at the smallest size that still transfers essential information—which is often smaller than authors expect! Avoid disappointment about lost detail by keeping the smallest final published size firmly in mind (Figs A1.3–A1.5).

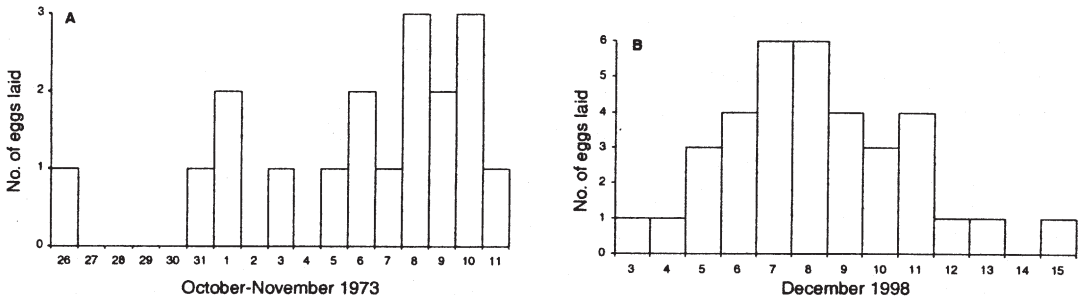


Figure A1.3. These histograms are reproduced here at the exact size published in the journal *Notornis*. How these graphs were derived from default Excel tables is described on the following pages and illustrated in Figs A1.4 and A1.5.

Original caption: Laying dates of Cook's petrels (*Pterodroma cookii*): A. Hauturu (Little Barrier I.) in 1973 until 11 November when 50% of eggs had been laid; B. Whenua Hau (Codfish I.) in 1998.

Example A1.2 Vertical bar chart of continuous variable (time) with measurements in discrete values (counts)

TABLE A1.2. COUNTING PETREL EGGS (FROM IMBER ET AL. 2003).

DATE (DEC 1973)	EGGS LAID
3	1
4	1
5	3
6	4
7	6
8	6
9	4
10	3
11	4
12	1
13	1
14	0
15	1

Histograms are an appropriate way to visualise the distribution of the data as shown in Table A1.2. By using the Excel Chart Wizard, the vertical bar chart shown in Fig. A1.4A was obtained.

The default graph suffers 'the usual' Excel deficiencies that can be tidied up as already described in Example A1.1: unnecessary boxes, shading and bold-type labels. Furthermore, the key (legend) is entirely redundant here as there is only a single data category. To remove it: Chart Options > Legend > (untick box Show legend), or select it and delete.

The main problem of the graph as a histogram is that the continuous series (time) is shown in discrete steps. Correct this by clicking on a bar, then: Format Data Series > Options > Gap width = 0.

Also adjust the *y*-axis label to have a maximum value of 6, and remove clutter by setting steps to 2: Format Axis > Scale > Maximum = 6, Minor unit = 1, Major unit = 2.

The result: Figure A1.4B shows again a clear and clean graph with emphasis on the data, minimum clutter, and which reproduces well in black and white and at small sizes.

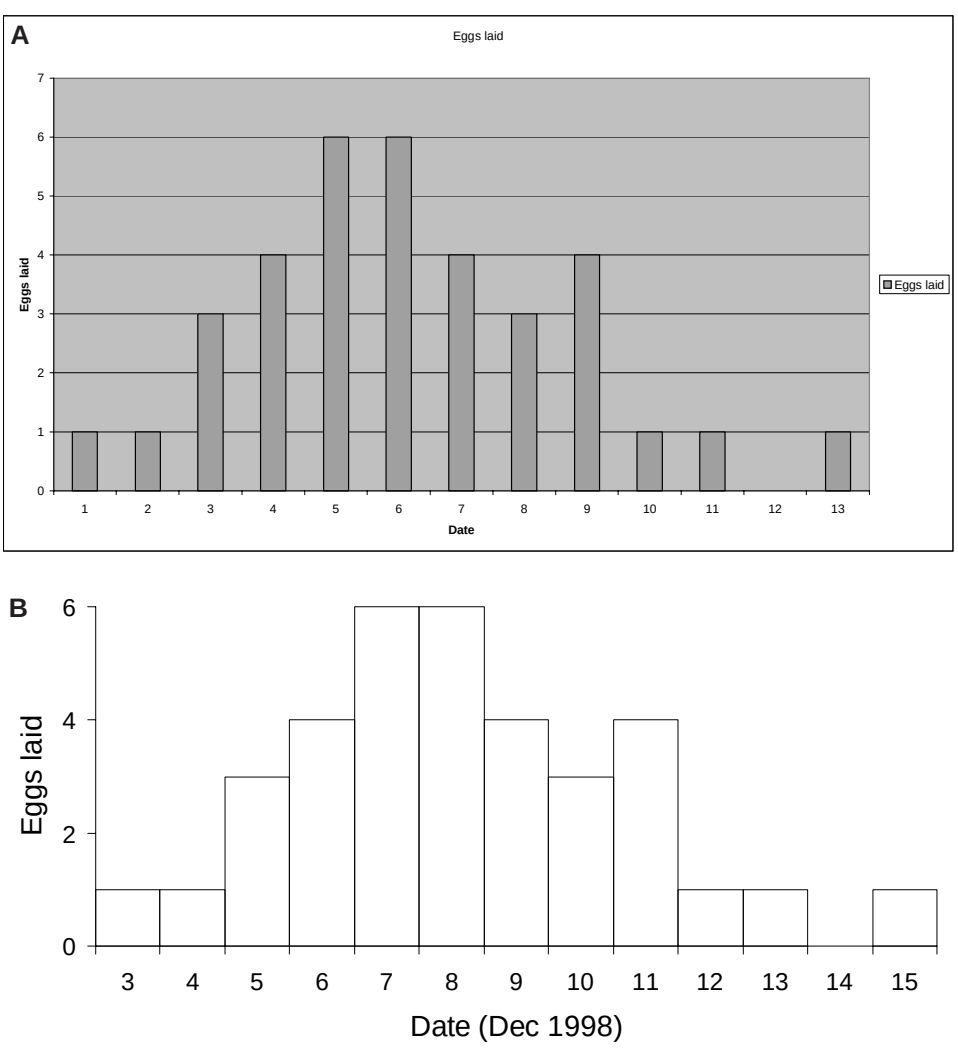


Figure A1.4. Egg-laying data (from Imber et al. 2003). A shows the default settings in Excel, including redundant key for one category, unnecessary clutter and colour. B is the graph improved as described. Its categories are continuous and it uses space much more effectively. It could be reduced to 50% or smaller without loss of clarity, provided the font size is increased accordingly.

Example A1.3 Histogram showing Excel errors in date and count representation

TABLE A1.3. MORE EGG-LAYING DATA FROM IMBER ET AL. (2003).

DATE (1973)	EGGS LAID	FIX DAY
26 Oct	1	26
27 Oct	–	27
28 Oct	–	28
29 Oct	–	29
30 Oct	–	30
31 Oct	1	31
1 Nov	2	1
2 Nov	–	2
3 Nov	1	3
4 Nov	–	4
5 Nov	1	5
6 Nov	2	6
7 Nov	1	7
8 Nov	3	8
9 Nov	2	9
10 Nov	3	10
11 Nov	1	11

The next dataset and Chart Wizard default graph in Fig. A1.5A illustrate the danger of incomplete date recording in a spreadsheet.

Although the dates are recorded for 1973, they are graphed as for the current year (= default setting). The problem was solved by creating an extra column, 'Fix day', for the days of the respective months and plotting these instead (the *x*-axis label indicating the true values). The 'corrected' dates must be selected under: Source Data > Series > Category (*x*) axis labels. If necessary, to avoid oblique or vertical type on *x*-axis, Format Axis > Alignment > Orientation (horizontal = 0 degrees).

The default graph was then tidied up as before. In addition, the fractions in the vertical axis were removed as there were no half eggs! Format Axis > Scale > Maximum = 3, Minor unit = 1.

The result: Figure A1.5B shows once more a clear and clean graph with correct dates, and which reproduces simply in black and white, even at very small sizes (see Fig. A1.3).

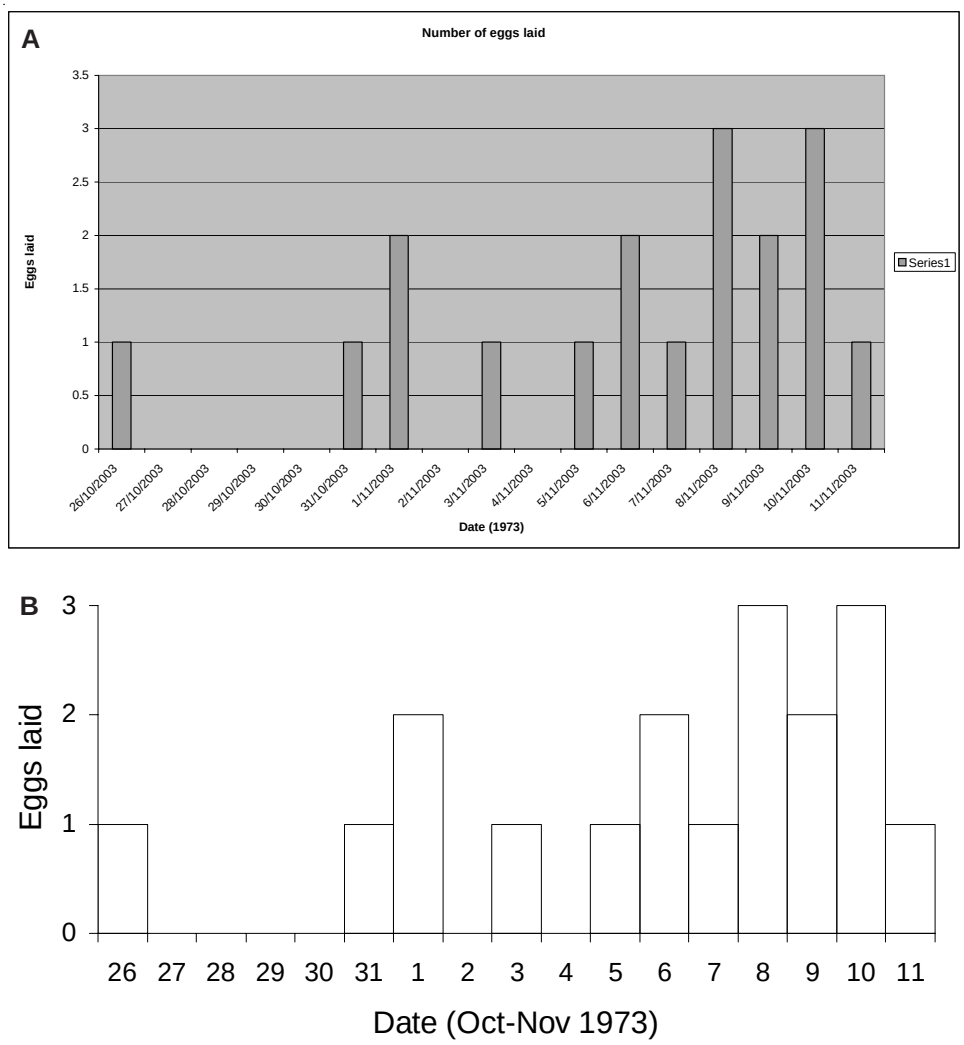


Figure A1.5. Graphs of the data in Table A1.3. A, The software defaults to the current year if incomplete dates are entered. B, the tidied-up graph, which displays 'whole eggs' only.

Appendix 2

CREATING BAR CHARTS USING S-PLUS

By Amanda Todd and Ian Westbrooke

S-PLUS can also be used to create bar charts. Although this package is not as user-friendly as Excel, it has the advantage that graphs can be more easily tailored to specific requirements, and are readily repeated in identical form for modified or new data; this is especially useful for creating groups of identical graphs for differing datasets in panels.

In this appendix, we demonstrate how the same graphs as presented in Appendix 1 can be created using S-PLUS. Although this can be done using drop-down menus (similar to the Excel chart wizard), these have their own set of default settings that need to be overridden to produce your final chart. Therefore, here we present the command language used to generate clean, easy-to-read graphs in one step.

The majority of datasets will probably be contained in Excel worksheets. Such data can easily be brought into S-PLUS using the Excel to S-PLUS Link Wizard.

To open the Excel worksheet from within S-PLUS:

```
File > Open > Excel worksheet (xls)
```

Highlight the entire data area and click on the Link Wizard button (top left of toolbar). If your first row contains column names, ensure you check this box. Re-name the file and click 'Finish'.

Open a script file in S-PLUS (this enables you to input the command language):

```
File > New > Script file
```

You are now ready to analyse your data and create graphs.

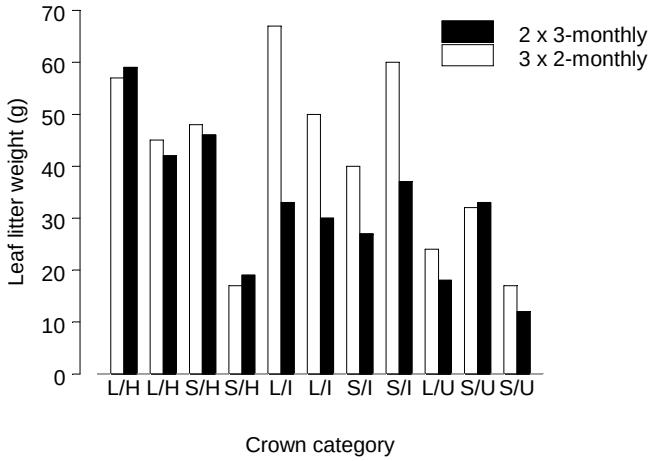
The commands below can be copied, pasted and edited from the electronic version of this manuscript rather than being entered from scratch if you prefer.

Example A2.1 Vertical bar chart of measurements taken for independent categories

The data presented in Table A1.1, named 'hosking' and with revised column names (‘cc’, ’two.monthly’ and ‘three.monthly’) for S-PLUS, can be graphed (Fig. A2.1) as follows:

```
attach(hosking)
    # enables S-PLUS to recognise column names in the dataset
    # ‘hosking’

graphsheet(color.scheme=‘trellis black on white’)
    # creates a black-and-white graphsheet

par(mar=c(7,7,4,7)+.1)
    # increases margins on the page to make sure all labels
    # are shown

barplot(
    # the basic command for bar chart

    t(hosking[,2:3]),
    # specifies the two data columns to be plotted
    # using t() to transpose them to be treated as rows

    beside=T,
    # places the two categories beside each other as opposed
    # to stacked

    col=c(0,1),
    # specifies the colours of each category

    space=c(0,1),
    # sets the space between the bars

    yaxs=‘s’,
    # specifies the extent of the y-axis

    las=1,
    # all labels written horizontally

    mgp=c(4,1,0),
    # the margin line for the axis title, labels and line,
    # to put the x-axis label away from the bar names

    cex=1.5,
    # increases the font size of axis labels

    xlab=‘Crown category’,ylab=‘Leaf litter weight (g)’,
    # x- and y-axis labels

    bty=‘n’,
    # removes box around legend

    legend=c(‘3 x 2-monthly’,‘2 x 3-monthly’
    # legend captions

    text(x=3*1:length(cc)-1,y=-2,labels=cc,cex=1.5)
    # adds label to bars

    )

n.gps1=nrow(hosking)+1

segments(x1=3*(1:n.gps1)-3,y1=rep(0,n.gps1),x2=3*(1:n.gps1)-
    2,y2=rep(0,n.gps1),col=0)
    # these last two commands remove the bottom lines
    # between columns
```

Figure A2.1. S-PLUS version of Fig. A1.2C.

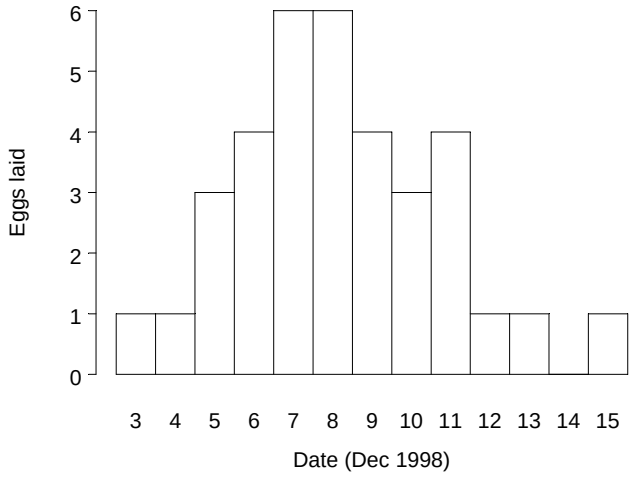


Example A2.2 Vertical bar chart of continuous variable (time) with measurements in discrete values (counts)

The data presented in Table A1.2, called 'imber' and with revised column names ('date.1973' and 'eggs.laid') for S-PLUS, can be graphed (Fig. A2.2) as follows:

```
attach(imber)

graphsheet(color.scheme='trellis black on white')

par(mar=c(7,7,4,2)+.1)

barplot(eggs.laid,names=as.character(date.1973),space=0,inside=T,col=0,yaxs='s',
    las=1,cex=1.5,mgp=c(5,1,0),xlab='Date (Dec 1998)',ylab='Eggs
    laid')
```

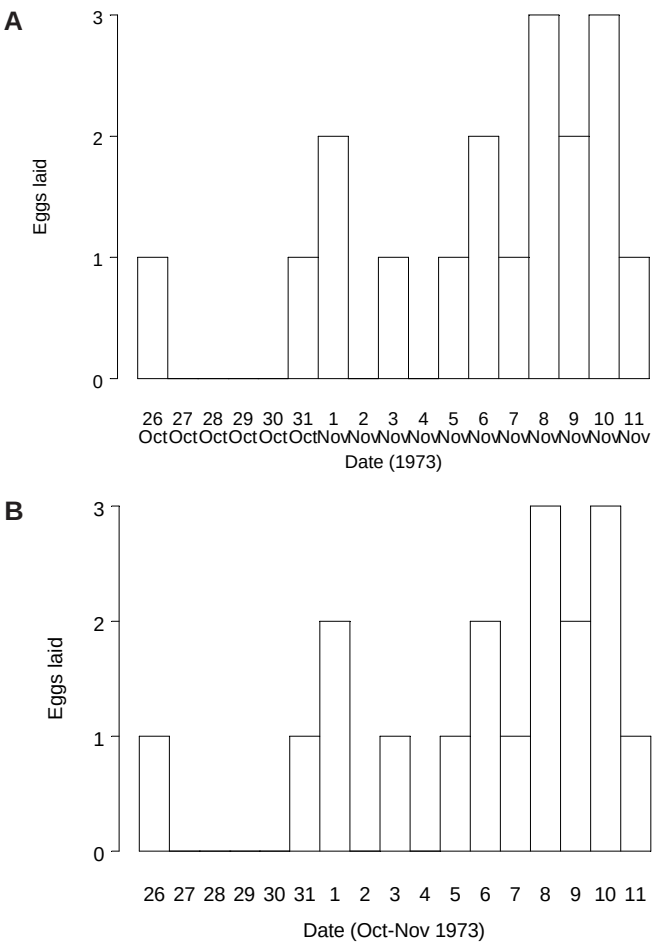
Figure A2.2. S-PLUS version of Fig. A1.4B



Example A2.3 Histogram with date information

The data presented in Table A1.3, called 'imber2' and with revised column names ('date.1973', 'eggs.laid' and 'fix.day') for S-PLUS, can be graphed in one of two ways (Figs A2.3A and B, equivalent to Figs A1.5A and B respectively). The second is preferable, as the *x*-axis is less cluttered. You will need to make sure that the values represented by '–' are correctly entered as 0 for S-PLUS.

```
attach(imber2)

graphsheet(color.scheme='trellis black on white')

date.formatted=timeDate(julian=date,format='%d\n%b')
    # creates a new date variable (datemod),
    # which outputs as day and month (month on a new line)

par(mar=c(7,7,4,2)+.1)

A:

barplot(eggs.laid,space=0,col=0,yaxs='s',las=1,lab=c(3,3,3),
    # the second value controls the number of tick marks on
    # the y-axis, allowing us to present only whole numbers

    mgp=c(5,1,0),cex=1.3,xlab='Date (1973)',ylab='Eggs laid',
    names=as.character(date.formatted))

B:

barplot(eggs.laid,names=as.character(fix.day),space=0,col=0,yaxs='e',
    las=1,lab=c(3,3,3),cex=1.3,xlab='Date (Oct-Nov
    1973)',ylab='Eggs laid')
```



Figures A2.3A and B. Two alternative S-PLUS versions of Fig. A1.5A and B. Note that Fig. A2.3A is not of publication quality, as the *x*-axis labels are too closely spaced.