Point of View

Basic principles of graphing data

Marcin Kozak

Warsaw University of Life Sciences – Dept. of Experimental Design and Bioinformatics – Nowoursynowska 159 – 02-776 – Warsaw, Poland. E-mail <nyggus@gmail.com>

ABSTRACT: Data visualization is a very important aspect of data analysis and of presentation. Focusing on the latter, this paper discusses various elements of constructing graphs for publications. Bad and good graphs are compared, and a checklist with graphical elements to be used while creating graphs is proposed.
Key words: visualization, graphs, statistics

Princípios básicos na confecção de gráficos

RESUMO: A visualização de dados é um aspecto importante da análise de dados e de sua apresentação. Nesse ponto de vista enfocam-se vários elementos da construção de gráficos para publicações. Gráficos ruins e bons são comparados, e uma lista de checagem de elementos de composição gráfica é proposta para ser utilizada durante a confecção de gráficos.
Palavras-chave: visualização, gráficos, estatística

Introduction

Agricultural scientists produce abundant results. Some of them are important, others may be less important, still others may be negligible. The way one presents the results, then, counts for several reasons. First, one needs to emphasize those important ones. Second, for most data, interpretation is easiest with graphs, and it is far easier based on a good than a bad graph. Just as statistical analysis can provide false conclusions (Huff, 1954; Kozak, 2009a), poor graphs can misinform, sometimes leading to serious misinterpretations.

Huff (1954) was probably the first to discuss that graphing statistical data can be a tool of misinformation. Later some of his notions found counter-arguments, yet the main ideas still hold: graphing must be correct and convey true information and interpretation. Scientific graphs are not to be "beautiful"; they are to be informative. Of course, ugly graphs will not convey any interesting message, so elegance should not be disregarded (Tufte 1991; 1997; 2001 and 2006).

Tufte (1991; 1997; 2001 and 2006), Cleveland (1993; 1994), Jacoby (1997; 1998) and Wilkinson (2005), among others, discuss graphing data in great detail. Harris (1999) is a useful reference for information graphics. Yet scientific literature in the 21st century is full of poor graphs, some of which are incomprehensible while others can even unintentionally falsify information. Data visualization is a developing research area, and what was considered a good graph 30 years ago does not have to be good these days. In this paper basic information is offered about graphing data with the help of which authors should be able to construct sufficiently good scientific graphs. Focus will mainly be placed on graphical rather than statistical aspects, so remember that when constructing a graph, you must take all care to ensure it delivers the message you intend it to, from both scientific and statistical points of view. Note also that visualization data with the purpose of data analysis and interpretation at the computer screen does not follow exactly the same rules as those described in this paper deals mainly with graphing for others, that is, for publications.

You will see good and bad graphs, what can go wrong when visualizing data, and why it is wrong. At the same time you will see how the problem can be overcome, and what to do to ensure the graph will be fine. In the next section, "General principles of graphing data", a general vision of graphing scientific data is presented. Next, in the section "Specific principles of graphing data" some detailed rules to be followed are presented to ensure if a graph is efficient. Do not forget, however, that these are really *basic* principles, as the title of the paper says; if one wants to learn more about this topic, one should refer to sources cited in this paper. The last section offers a short conclusion, while the Appendix lists data sets used in this paper and the information on the software employed to analyze and graph these data. Note that in figure captions more information than is normally needed is given, for example the type of graph. Each figure caption informs whether the figure (or its panels) represents a good or rather poor style.

General principles of graphing data

The general and most important principle of graphing data is to construct a graph so that it conveys the message in the most efficient way, and the message one wants it to convey. Thus, all elements of the graph should be helpful but not distracting, and important aspects should be emphasized but not hidden. This is the

most general rule, and all the following rules account for this general one. Remember that all details count: viewers of a graph will follow your ideas if you help them. So if you don't feel an expert in visualization, think twice - or rather thrice - before deciding not to follow any of the rules.

Before making a graph, think over what message it is to convey and whether it is needed whatsoever. In general, graphing three numbers is not advisable; no matter how amazing this can seem, one can find quite a few such graphs, especially piecharts and barplots. In fact, two numbers are sometimes graphed. Two numbers will be most efficiently presented within a sentence, while few numbers within a text-table (which is a simple table inserted within the text - Tufte, 2001; Kozak, 2009b) or a regular table. Tufte (2001) argues that even a dozen or so numbers are better represented by a table than a graph, but this is arguable - in general for 10 and even fewer numbers to be compared a graph can be preferred, although it depends on various things, including what the numbers represent.

Make graphs as simple and as complex as it is needed to deliver the message. But neither be afraid of a complex graph nor make it more complex than it should be. Remember that the human eye can work efficiently with very composite images, detecting large amount of information in small spaces (Tufte, 2001). However, always be careful to make the graph readable and understandable: avoid clutter (see, e.g., Reynolds et al., 2009; Silva, 2009c).

An efficient way to proceed while constructing a graph is to:

1) figure out the contexts and the message

2) figure out the way of presenting it, so the type, layout and style of graph to be used

3) construct the graph

4) check Tables 1 and 2 and revise the graph accordingly
5) check whether the revised version conveys the message you want it to convey

6) check Tables 1 and 2 and revise the graph accordingly

7) and so on...

Table 1, mentioned before, offers basic elements to check while constructing a graph - they will be discussed in the following section. They do not refer to every type of graph, but it will be best to check everything that is listed there, and decide what does and what does not apply to your graph. Table 2 cites Cleveland's (1994) very useful principles of graphing data.

Note how similar the process of graph construction is to the process of writing: figure out the message and style, construct, revise, revise, revise, revise... Revising cannot be ignored, and exactly like in writing, it can be a source of inspiration, leading to results and conclusions one could not even imagine before. The above "algorithm" of graph construction together with Tables 1 and 2 will be especially useful for non-experienced graph constructors, because those experienced ones follow it intuitively.

Specific principles of graphing data

Data points are very important elements of many graphs, especially scatterplot and line plot with lines superimposed on data points. Symbols representing them together with their size and color should be carefully chosen so that the points could be easily seen, and the patterns present in the data (or their lack) could be easily noticed. Figure 1, picturing logarithm of the volume of cherry trees against the logarithm of

Table 1 – Elements to check while constructing a graph. Refer to the text for details.

Data points	symbol, size, color, overlap
Lines	type, width, color, overlap
Color	check whether needed at all; check whether all elements are easily distinguished
Box and axes	type of box, aspect, minimum and maximum value, data rectangle and scale-line rectangle, tick marks (number of, location, direction, length and width), tick marks' labels (font type, length, rotation, numbering style, abbreviations of text labels)
Legend	check whether needed, position, box around, size, elements within (see above for "data points" and "lines", and "labels" within "box and axes")
Background	of the graph; of bars, cross-hatching, shading, color
Text inside the graph	check whether needed; font style, size and color
Captions	check whether informative, explaining everything that is needed to understand the graph and the message it delivers
Dimensions	check for over-dimensionality
Reference line	check whether needed, line type, width and color
Grid lines	check whether needed, line type, width and color
Error bars	check if the information is given on what they represent; line type, width and color, type of ending
Trellis display	all the above elements; layout (rows, columns, pages), choice of panel variables and conditioning variable(s), panel order

Table 2 – Clear vision principles given by Cleveland (1994).

1. Make the data stand out. Avoid superfluity.
2. Use visually prominent graphical elements to show the data.
3. Use a pair of scale lines for each variable. Make the data rectangle slightly smaller than the scale-line rectangle. Tick marks should point outward.
4. Do not clutter the interior of the scale-line rectangle.
5. Do not overdo the number of tick marks.
6. Use a reference line when there is an important value that must be seen across the entire graph, but do not let the line interfere with the data.
7. Do not allow data labels in the interior of the scale-line rectangle to interfere with the quantitative data or to clutter the graph.
8. Avoid putting notes and keys inside the scale-line rectangle. Put a key outside, and put notes in the caption or in the text.
9. Overlapping plotting symbols must be visually distinguishable.
10. Superposed data sets must be readily visually assembled.
11. Visual clarity must be preserved under reduction and reproduction.
12. Put major conclusions into graphical form. Make captions comprehensive and informative.

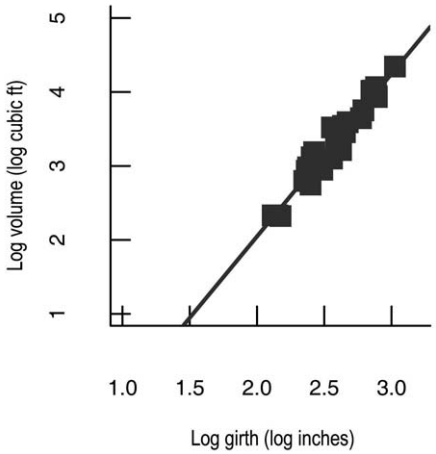

Figure 1 – A scatterplot with a regression line superimposed, representing trees' volume against girth, both after a logarithmic transformation (data set trees). The graph is **full of drawbacks**: filled squares as plotting symbols are difficult to distinguish; the box consists of two scale lines only; the data rectangle is too large so the points are clustered within a small area; the tick mark labels on the y-axis are vertical instead of horizontal even though they are just one-digit numbers; the tick marks point inward. Note that the regression line extrapolates the values outside the ranges observed in the experiment, which can be dangerous.

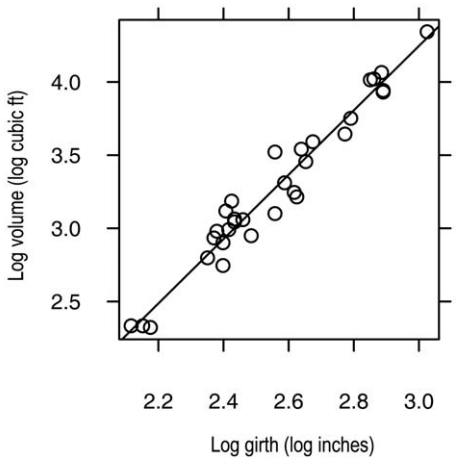

Figure 2 – A scatterplot with a regression line superimposed, based on the same data as in Figure 1, but all the drawbacks mentioned there are overcome. It is considered a GOOD GRAPH.

girth (data trees), is poor. The choice of filled squares as symbols is bad because it is difficult to distinguish the points. (By the way, a filled rhombus, which is the default symbol for a scatterplot in Microsoft Excel, is an equally poor choice.) Compare it with Figure 2, whose readability is greatly enhanced thanks to open circles used as symbols—in fact, open circles should be considered the best symbol in case of some overlap in scatterplots for non-grouped data. (Note that for dotplots, in which no overlap occurs, closed circles are usually used as plotting symbols—Figures 3 and 4). Sometimes, however, overlap can be a serious problem—see Figure 5.

When data are grouped, the choice of plotting symbols matters too: ○, ●, ⊙, ⦿ should work well for little overlap, and ○, +, <, s, w for serious overlap (Cleveland, 1994); but this will strongly depend on what is offered by software one uses. Varying color for groups can be even more efficient, not to mention combining plotting symbol and color. Sometimes decreasing the size of symbols will suffice, but a special technique to deal with overlapping points is jittering (Cleveland, 1994), which was applied in the right panel of Figure 5. This technique consists of adding a small amount of random noise to the data, thanks to which the overlapping points are usually easier to distinguish. There are also other techniques to deal with serious overlap when there are many points to draw, examples being sunflower plot (Cleveland and McGill, 1984), graphing binned data

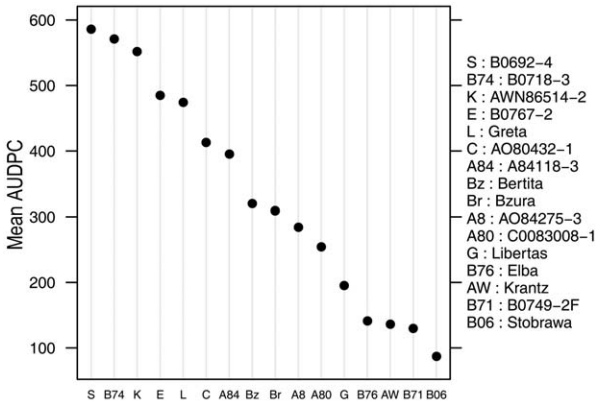

Figure 3 – A vertical dotplot representing mean area under the disease progress curve (AUDPC) of 16 clones in one environment (Hastings, FL; data set haynes). The clones are ordered by decreasing yield. A **drawback** of this graph is the legend: because the clones are presented in the x-axis, their names are too long to be presented without rotation or abbreviation. The abbreviated names are used here, which calls for the legend to explain the abbreviations.

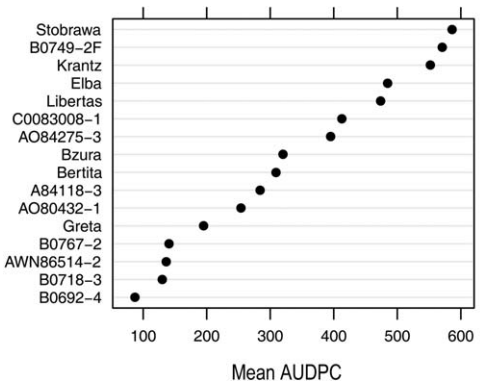

Figure 4 – A horizontal dotplot representing the same data as in Figure 3. The drawback of Figure 3 is overcome here: vertical dotplot enables one to use sufficiently long labels so that no legend is needed. Note that this Figure is smaller than the previous one, yet it is better readable. It is considered a GOOD GRAPH.

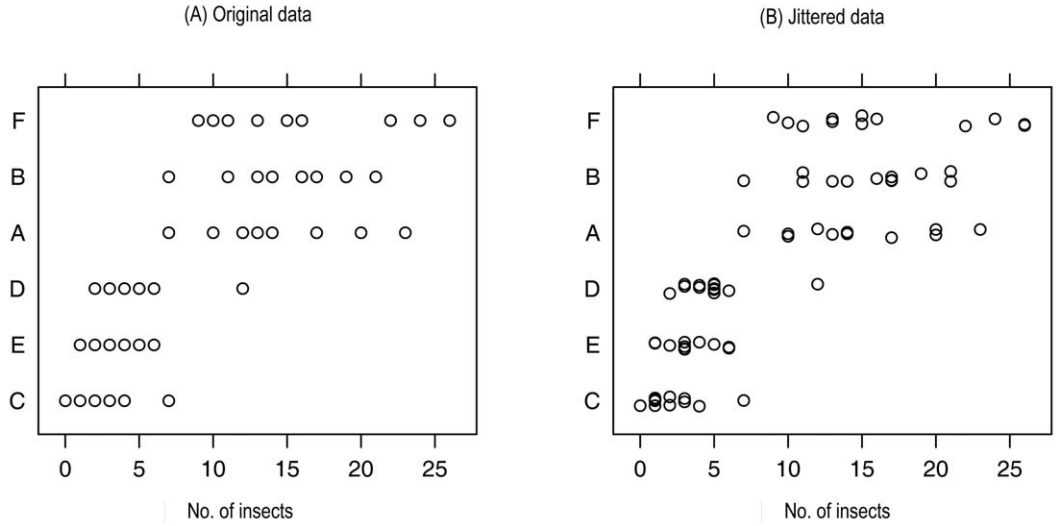

Figure 5 – Scatterplots representing replicated data of number of insects caught after applying various insecticides (data set InsectSprays). A **drawback of the left panel** is the overlap of points. This is overcome in the right panel, where jittering is employed (Observe the GOOD GRAPH IN RIGHT PANEL).

(Carr et al., 1987), a combination of these two (Dupont and Plummer, 2003), or graphing regions of bivariate kernel density (Hydman, 1996). Yet another idea is to use the trellis display (Cleveland, 1993) for grouped data, in which case the groups will be plotted in different panels, alleviating the overlap of points from different groups; this type of display will be discussed later.

Similar issues ought to be raised for **lines**: they should be easily distinguished, for which varying line type (e.g., solid, dashed, dotted), width and color can be used. Lines can be plotted together with data points, in which case their efficient combinations should be chosen. In Figure 6 three options are presented for this: in the top panel, each of the five lines representing the growth of a chicken is plotted with the same style of line and plotting symbol. Note that up to about 10 days the lines are difficult to be distinguished, and it is practically impossible to notice here what can be seen in the middle and bottom panels: that the lightest chicken at the end of the experiment was one of the heaviest during the first days. Color in the bottom panel seems to work better than varying line type and plotting symbol in the middle panel, but both succeed to convey the most important information. Note that not all the information can be extracted from these plots. In this

particular situation it does not seem a problem, but a simple solution might be to plot a logarithm of body weight on the y-axis - see Figure 7, where the lines are slightly easier to be distinguished for the first days of the experiment (although in many situations the gain in readability will be much larger).

Color can be powerful, but do *not* overuse it. Save color differentiation for situations when it helps, and don't use it when it adds nothing to the graph. For example, would it make sense to use different colors for the bars in Figure 8, picturing mean count of insects caught after applying six insecticides? Of course not—not helping to differentiate the insecticides, it would be redundant and distracting. But sometimes color can be very helpful, supporting a viewer by clear group differentiation (e.g., Figures 7, 10 and 15, and Lammel et al., 2007; Jaradat, 2007; 2009). Remember that color will help

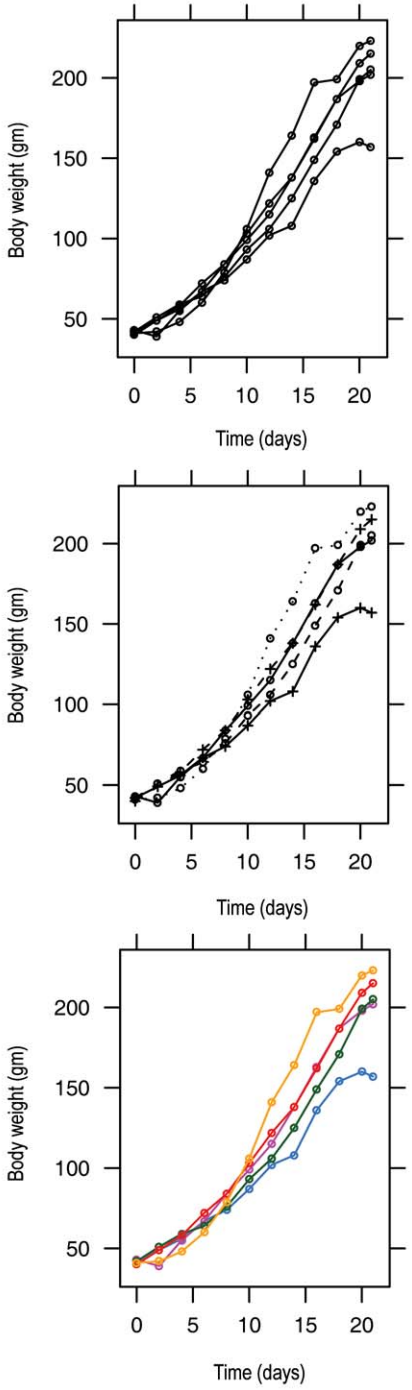

Figure 6 – Line plots representing changes in body weight of five chickens through time (data set ChickWeight). The **problem** in the top panel is that all five lines are plotted with the same line style and plotting symbol. This is overcome in THE MIDDLE AND BOTTOM PANELS (GOOD GRAPHS). In all graphs, the aspect ratio has been chosen according to the baking to the 45° rule.

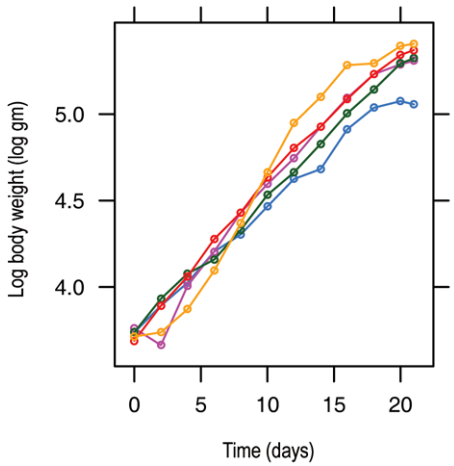

Figure 7 – A line plot representing the same data as in the bottom panel of Figure 6, but body weight is presented in natural logarithms. This helps one to compare the lines in the early days of growth. The aspect ratio has been chosen according to the baking to the 45° rule—note that the aspect ratio is different than in the plots in Figure 6, because of differentscales for variables in these two plots. It is a GOOD GRAPH.

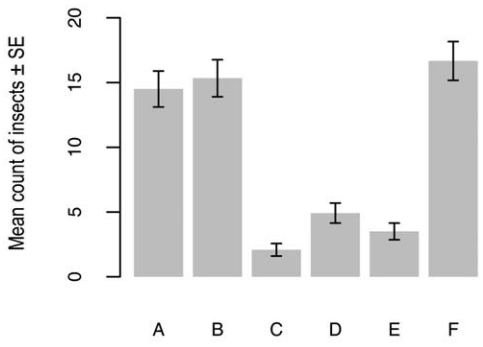

Figure 8 – A barplot representing mean count of insects caught after applications of six insecticides (data set InsectSprays). Error bars represent standard error (SE) of the mean estimated with generalized least squares (Pinheiro and Bates, 2000). **Drawbacks** The insecticides are ordered alphabetically even though their names have no meaning; tick mark labels at the vertical axis can be presented vertically.

only if other graphical elements are well arranged. For example, Vázquez et al. (2009) had eight groups to distinguish in a scatterplot, and they did that by varying plotting symbols and their color. Unfortunately, the large size of symbols made it practically impossible to compare the groups.

Box (frame) around a graph plays an important role: it consists of axes (scales) that work as rulers along which data are graphed (Cleveland, 1994). Box and axes help see the data and locate coordinates of data points. Whether the box should consist of two axes, which is a tradition, or four axes, which is more efficient (Cleveland, 1994), is a matter of discussion. A four-axis box is better, unless one is forced to do otherwise (for example, by a journal editor, which sometimes happens). Compare Figures 1 and 2 and note how well the full box in the latter defines the plotting region compared to the two-axis box in the former. A very important aspect of the box is that the data rectangle (so the region used for plotting) should be slightly smaller than the box itself (Cleveland, 1994) - otherwise the plotting symbols will interfere with the axes (see Figure 9 and Virdis et al., 2009), although Tufte's (1997, p. 45) graph is a very elegant with exception to this rule. Remember also that the data rectangle should be effectively used—for a variable ranging from 50 to 105 it would seldom be wise to start the axis at 10 and end at 170 (see also Figure 1). Of course, there are exceptions to this rule, when there are reasons to start and/or end the axes at particular values. In addition, the appearance of tick marks matters: most software has some algorithm of choosing their number and location, but remember not to use too many of them. They also should neither be too long nor too short, and they should point outwards. Refer to Table 1 for further details.

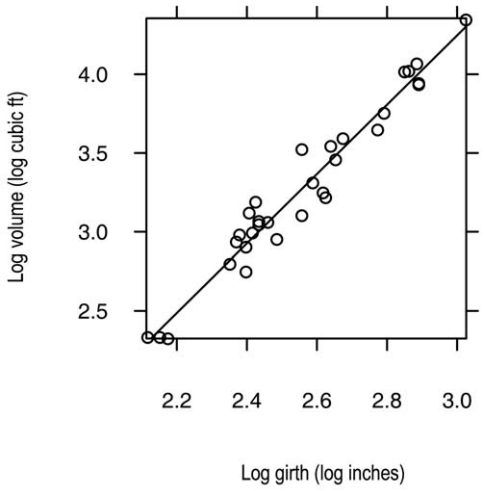

Figure 9 – A scatterplot with a regression line superimposed, based on the same data as in Figure 2, but the **drawback** of this graph is that the data rectangle is the same as the box, so four points are graphed at the axes.

Aspect ratio of a data rectangle is its height divided by width (Cleveland, 1994). It is the aspect ratio that decides about the shape of the graph, so it should be chosen with care. Try not to choose the aspect ratio so that the graph fills the space available in the publication, but choose it so that a graph conveys a message. Tufte (2001) suggested following the nature of data if there is any hint about the shape of the graph, and otherwise making the graph wider than tall about 50%. Cleveland (1994) proposed a rule of banking to 45°, which he showed to be very useful for line plots and scatterplots with a superimposed locally weighted regression line; this method optimizes judging the rate of change. Sometimes banking to 45° does not work because it makes the graph not readable, but the point is that the aspect ratio is important. In general, unless there are particular reasons for that, the box should be neither too narrow nor too wide compared to its height. It also strongly depends on the type of plot: for example, for scatterplots square boxes might be good - too many scatterplots in the scientific literature are far too wide. Sometimes a so-called isometric aspect ratio can be employed, in which the relation between physical distance on the device and distance in the data scale are forced to be the same for both axes (Sarkar, 2008) - see Figure 10 (which will be discussed in detail in the next section), in which petal length and width are in cm, so the isometric scale helps compare these two traits.

Order of elements matters. This is very important for tables (Ehrenberg, 1977), but equally important for groups compared by means of barplots, dotplots, boxplots and the like. It is generally acknowledged that

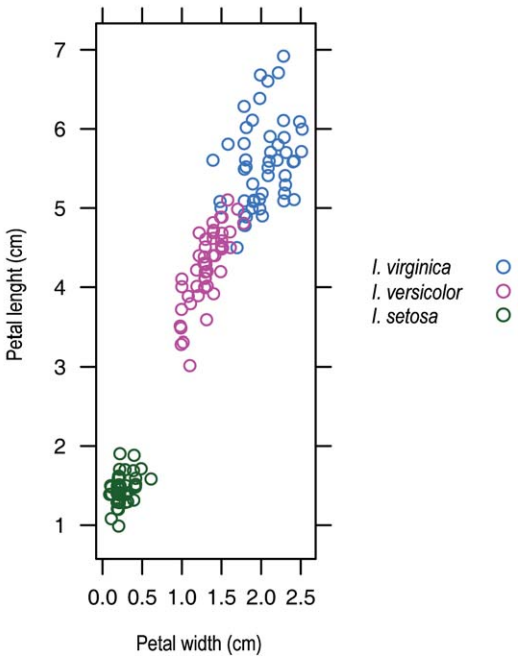

Figure 10 – A scatterplot representing petal length versus petal width of three iris species (data set iris). The shape of the graph results from the isometric scales. Look that it is a GOOD GRAPH.

comparing several numbers ordered alphabetically by labels, like here:

A 101.2
B 112.1
C 110.2
D 102.1

is more difficult than comparing these same numbers ordered by size, like here:

B 112.1
C 110.2
D 102.1
A 101.2

Compare the two boxplots presented in Figure 11. Clearly the ordering by median insect count makes the comparison of insecticides easier than the alphabetical ordering. Of course, you must decide if alphabetical ordering by labels does not matter indeed. Note also that for several graphs you may want to apply the same ordering to facilitate the general interpretation of the groups.

Text in graphs appears in various places: axis and tick mark labels, legend (sometimes), title (sometimes, usually for multi-panel graphs, like in Figure 11), and data point labels (see the comment below and Tables 1 and 2). It also appears in figure captions, but we will discuss them later. Use a font that will be readable for everyone; in most situations sans-serif fonts (e.g., Arial) work best for graphs. Horizontal labels are easier to read than vertical or rotated ones (cf. Figures 1 and 2)—so horizontal barplots and dotplots (cf. Figures 3 and 4) will usually be better than vertical ones. If numbers are too long to present them horizontally on the y-axis, consider dividing them by some reasonable factor. This does not mean that one should never present vertical tick mark labels at the y-axis, but that in most situations this will not be the best solution. Try to avoid putting text within the data rectangle when it can distract a reader's attention (e.g., Mohapatra and Kariali 2008); for example, regression equations are far too often put inside graphs (Fiorio and Dematte, 2009; Pahlavani et al., 2009; Simoes et al., 2009). However, Tufte's (2006, p. 120) graph is an excellent exception to this rule. Sometimes, labels are needed inside the data rectangle, examples being biplots (e.g., Lammel et al., 2007), but one must remember about Cleveland's (1994) rule of not allowing such data labels to interfere with the quantitative data or to clutter the graph (Table 2).

Do not overuse **legends**. If all the information necessary can be included in the graph itself, then there is no need of legend whatsoever. For example, Silva et al. (2009b) decided not to use a legend in a situation when most others would. Jonsson and Aoyama (2009) would have needed a legend had they used a horizontal version of the barplot instead of the vertical one, while Ambrosano et al. (2009) chose the vertical barplot and so had to use the legend. Compare Figures 3 and 4, representing the same data: mean area under the disease progress curve of 16 potato clones in one environment. In the former, the clones are presented at the horizontal axis (note that this is usually done for barplots). The clones' names are too long to be presented in full, so they are abbreviated and the legend is needed to explain the abbreviations; even despite that, the font size needed to be reduced. In Figure 4, a vertical version of the dotplot is drawn with full names, so no legend is needed. Of course, it does not mean that legends should never be used—sometimes they are indeed required, for example when the grouped data are plotted within one graph (e.g., Figure 10). Quite often the legend is included in a figure caption (which saves space), but placing it near the graph (like in Figure 10) facilitates reading. Nonetheless, remember Cleveland's advice of avoiding putting a legend within the box (Cleveland, 1994, Table 2; cf. Macedo et al., 2009). When the legend is presented, it should rather not be surrounded by a box, as it is sometimes done.

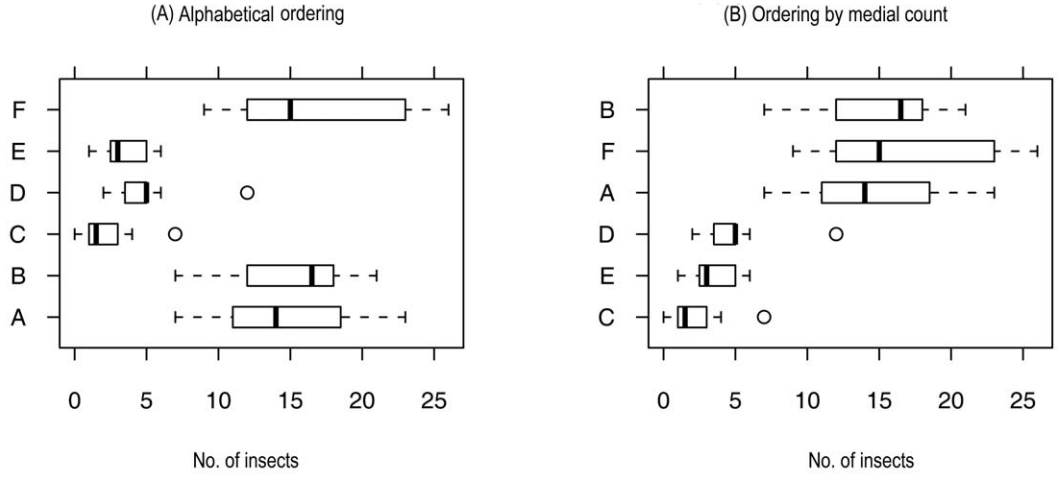

Figure 11 – Boxplots representing number of insects caught after applications of six insecticides (data set InsectSprays). In left panel, alphabetical ordering is employed, while in right ordering by count. Observe the GOOD GRAPH IN RIGHT PANEL.

Use **figure captions** not only to give the short title of a graph, but also to comment on the graph. It can be very efficient to draw viewer's attention to some interesting aspects of the graph, while the same information in the text has less power (the reader has to move between the text and figure). In fact, the more self-standing the caption, the better, but of course the balance needs to be struck, and repeating everything from the text does not have to be the best idea. (You also need to remember that not all journals will accept long captions.)

In general, use white **background** for graphs in publications: in most situations it will be the best color if the publication is printed on white paper. Thus for example grey, the default color of the background in MS Excel 2003 charts, is generally a poor choice.

With barplots one needs to be very careful. All-too-often do authors choose **cross-hatching** to help the viewer distinguish bars (e.g., Ambang et al., 2009; Czubaszek, 2009; El-Shafey et al., 2009; Ofosu-Anim and Leitch, 2009; Takahashi et al., 2009; Wiewióra, 2009), the effect of which is rather poor: instead of easy differentiation of groups the viewer is attacked with what Tufte (2001) calls "moiré" effects. Moiré effects are the optical noise that makes the bars difficult to compare, and the effects the bars represent difficult to grasp. Tufte (2001) calls such junk elements the chartjunk. Cross-hatching, then, should be replaced with shades of grey or color (Tufte, 2001); good examples are those by Karlidag et al. (2009) and Silva et al. (2009a).

Bars in barplots should normally start at zero. This is because in barplots we compare heights of bars, so they must be meaningful: unless they start at zero, they usually have no sensible meaning. Note that this does not apply for dotplots (Figure 4), in which the viewer compares the position of the points along the appropriate axis.

Error bars are very often used with statistical data. Some rules should be followed when presenting them. First of all, the information (usually provided in a figure caption) on what the error bars represent is critical—without it the reader will not be able to make any use of the bars. Is it a standard deviation, standard error, confidence interval, Tukey's HSD, or still something else? Figures 8 and 12 show standard errors (SEs) while Figure 13 confidence intervals (CIs)—note how lengths of SE and CI error bars differ. The choice is extremely important, but this topic is beyond the scope of this paper; refer to Cumming et al. (2004) and Cumming and Finch (2005) for a discussion on the use and interpretation of error bars. But one important aspect deserves to be mentioned: this is *not true* that two overlapping confidence intervals indicate a statistically significant difference at the corresponding significance level (see Cumming and Finch, 2005 for the details).

What also matters when including error bars is the way they are presented. Compare Figures 8 and 12 to see that error bars are easier to read in dotplots than in barplots. If there are too many error bars to show (e.g. Maciá-Vicente et al., 2009, Scarpari and Beauclair, 2009), a good idea might be to present them in grey (quite likely making the data points in black better visible), and in the case of overlap, to employ jittering. For symmetric error bars only a half of them can be presented, as Giacomini et al. (2009) did for standard errors of the mean in line plots. They had two lines that were superimposed on data points: for the upper line the error bars pointed up, while for the bottom line pointed down (in addition the lines and error bars for the two groups differed in color), which avoided overlap. However, if too many error bars make the graph be in such a clutter that nothing can be seen in it (including data points), the data should be presented in some other way. Note that error bars in Figures 12 and 13 have different endings. Both are good, although

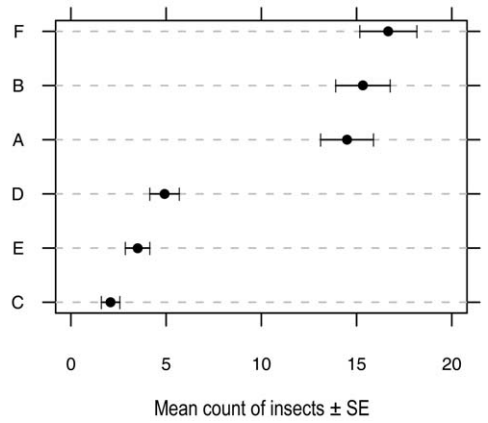

Figure 12 – A dotplot representing the same data as in Figure 8, but instead of the barplot, the dotplot is used (data set InsectSprays). The insecticides are ordered by decreasing mean count of insects. It is considered a GOOD GRAPH.

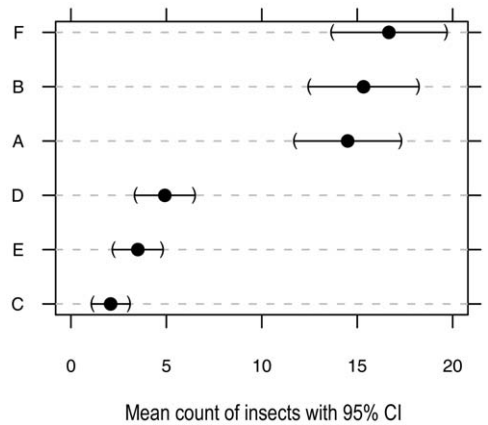

Figure 13 – A dotplot representing the same data as in Figure 12, but instead of standard errors of the means, 95% confidence intervals (CI) are presented. It is a GOOD GRAPH.

those in Figure 12 are preferred rather than those in Figure 13 because of easiness of comparison. The latter, however, have this advantage that they look like a confidence interval closed within parentheses, which is the way numeric confidence intervals are usually presented.

For multivariate data sets, a so-called **trellis display** (Cleveland, 1993, 1994) can be of help. It consists of panels arranged into rows, columns and pages, each panel containing a plot of the same type for subsets of the data set defined in a particular way. (Note that in publications trellis displays are seldom divided into pages.) This display can be very helpful in data analysis, interpretation and presentation. The trellis display works well with grouped data, in which case panels represent values of a categorical conditioning variable(s), but also with quantitative predictors, in which case a conditioning variable(s) is cut into intervals, and each panel represents those data points for which the value of the conditioning variable fits into the interval. Panels can represent various types of plots, including scatterplots, dot plots, boxplots, line plots and others. The trellis display can also be very useful in analyzing data from designed experiments (Cleveland and Fuentes, 1997). Most of the comments concerning graphing given before are equally important for this type of display, but there are many other rules that can or should be followed—the reader can refer for example to Cleveland (1993) and Cleveland and Fuentes (1997). Figure 14, which will be discussed later, is an example of the trellis display for a scatterplot of two quantitative variables and a categorical conditioning variable.

Some additional elements can be used to enhance a graph. These could be for example a **reference line** (see Table 2 for Cleveland's [1994] comment) or, for multipanel graphs (especially scatterplots), **visual grid**. Remember that a visual grid seldom has to do with table look-up of values (this could be the case for example for barplots); it usually aims to enhance between-panel visual comparisons (Cleveland, 1994) - Figure 14.

There is no space for **over-dimensionality**. If you are constructing a barplot, forget about 3D-barplots (e.g., Moussa and Abdel-Aziz, 2008; Abou Khalifa, 2009; Brunnings et al., 2009; Marcinkowska et al., 2009). Do note that regular barplots also add an unimportant dimension: the bars' width. Hence Cleveland's dotplots (Cleveland, 1994) will usually do better (cf. Figures 8, 12 and 13).

Last but not least, make it a rule: there is no space for **piecharts** in agricultural sciences. The piechart has been criticized by many (Cleveland, 1994; Tufte, 2001; Sarkar, 2008) - which does not mean it has not been defended (e.g. Spence and Lewandowski, 1990; Friendly, 1994). Nevertheless, for regular agricultural data piecharts should not be used, and this should be considered a rule with no exception. In every single case there will be a chart that works at least equally well as the piechart, and in the abundance of situations it will work much better, let it be a barplot, a dotplot and the like.

In summary, pay attention to every single element of the graph, checking its size, color, position and cooperation with other elements. After one has prepared quite a number of graphs, this will come unconsciously. But before this time comes, one might wish to follow the rules given in this paper.

Additional examples

Several examples of good graphs have been provided (Figures 2, 4, 7, 10, 12, 13, 14, 15; right panels of Figures

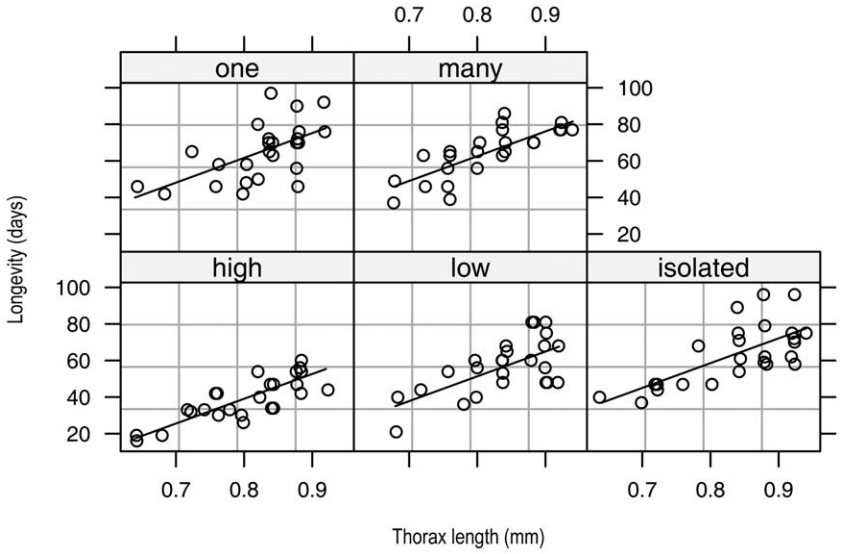

Figure 14 – A trellis display portraying relationships of fruitfly longevity and thorax length in different experimental groups (representing various sexual activities), with a regression lines superimposed (data set fruitfly). The regression lines were obtained through generalized least squares with a power variance function to deal with heterogeneity of residuals, and the final model included a common slope for thorax length intercept and varying intercept across activities. This is a GOOD GRAPH.

5 and 11; and middle and bottom panels of Figure 6). Here, three figures are discussed (two of which have already been referred to) in detail.

Example 1

Figure 10 shows an association between length and width of petals of three Iris species. The three species are plotted in the same panel, so three colors are used to distinguish them; a legend is needed, then. It was placed next to the graph, not in the figure caption: this facilitates quick recognition which species is represented by which color. Note that the species in the legend are presented in the same order as in the graph: *I. virginica* is in the top position, while *I. setosa* in the bottom one. This very subtle thing facilitates reading the graph. Jitter helps notice overlapping points. The box consists of four scale lines with tick marks pointing outwards, the top and right axes having tick marks without labels. The untypical shape of the box (narrow and high) is due to the isometric scales: petal length and width are presented in the same and comparable units (cm), and one cm has the same length at both scales. It works well by facilitating comparison of the two traits: we can immediately see that petal length is much more variable than petal width, information that would not be so obvious without the isometric scales. So do not be afraid of untypical aspect ratios of graphs if they are helpful indeed.

Example 2

Figure 15 presents a comparison of within-species distributions of petal length of the three Iris species. It has all properties of a good graph we have discussed above: color is used to differentiate the groups (varying line and point types might work too); the full box is used, with tick marks on each scale and tick mark labels on left and bottom ones; legend is added to introduce the colors corresponding to the species. But note also two additional interesting elements of this graph. This time the legend is placed above the graph, and the order of the species there is the same as in the graph itself: *I. setosa*, with the shortest petals, is to the left, while *I. virginica*, with the longest petals, is to the right of the legend; this small aspect helps read the graph. Second, the so-called rug is added (Chambers and Hastie, 1992), which shows individual points - this is useful for example for outlier detection. The rug is jittered.

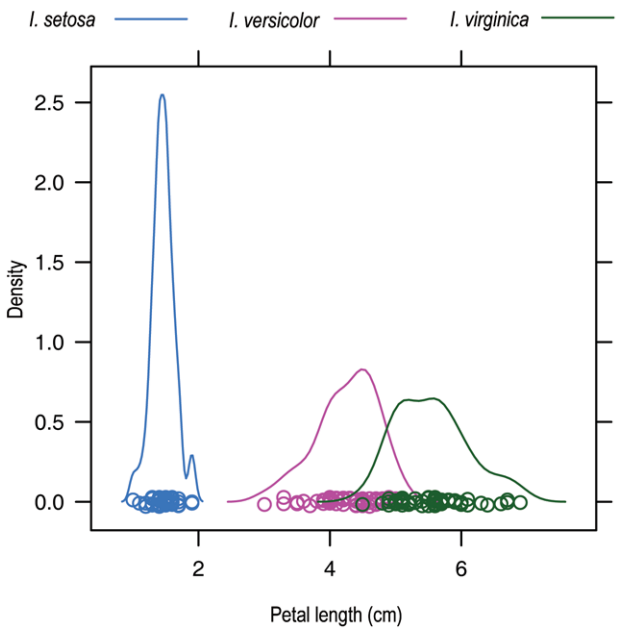

Figure 15 – Kernel density of petal length of three Iris species (data set iris). It is considered a GOOD GRAPH.

Example 3

The last example will be more biological than agricultural, describing association between fruitfly male longevity and thorax length, depending upon sexual activity. First we will model the relationship, and then we will visualize the results.

A linear model includes longevity as the dependent variable and thorax length and sexual activity as predictor variables, the former being continuous and the latter categorical. So, it is an analysis of covariance model.

Because thorax length ranged between 0.64 and 0.94, in the modeling 0.80 was subtracted from the variable to allow sensible inference for intercepts for activities. After fitting a model with varying intercepts and slopes it appeared that the residuals were heterogeneous, so this needed to be taken into account. The plot of standardized residuals versus fitted values suggested a power variance function structure, which indeed worked well after estimation with generalized least squares (Pinheiro and Bates, 2000). However, the slopes seemed quite similar, and the Akaike Information Criterion showed that the model with the common slope and varying intercepts showed the best fit; the power variance function still worked best to deal with heterogeneity of residuals. In this model both sexual activity and thorax length affected longevity of the fruitfly male.

Figure 14 portrays the data and the model fitted, with the help of the trellis display framework. We can see how powerful such a display can be when there are too many groups to be presented within one graph. (Here, 124 points and 5 lines within one panel would cause clutter, while the trellis display effectively compares the sexual activity groups.) All elements in the plot follow the rules given above and in Tables 1 and 2. Notice some additional aspects: the panels are ordered by the increasing intercept: the bottom left represents the "high" group (a fly kept with eight virgin fruitflies), which has the lowest intercept, and the upper middle panel represents the "many" group (a fly kept with eight pregnant fruitflies), which has the highest intercept. Mild jittering was employed for thorax length to deal with overlap of points, and grid lines facilitate panel-to-panel comparisons. Note that although the model was fitted for the thorax length subtracted by 0.80, this is not seen in the graph - the tick mark labels were simply changed from (–0.1, 0.0, 0.1) to (0.7, 0.8, 0.9).

Conclusion

Graphing data is not a simple matter. In fact, this "point of view" is just a small portion of the knowledge of visualizing scientific data, but it is believed to be sufficient for authors constructing simple graphs for their publications. Whole books have been devoted to the topic, best-known ones being the classics by Tufte (1991; 1997; 2001 and 2006), Cleveland (1993; 1994) and Wilkinson (2005). Jacoby (1997; 1998) also presents some interesting ideas, while Harris (1999) is a useful reference of information graphics.

Most principles presented in this paper are so basic that they apply to most graph types. However, be aware that there are other types of graphs that offer various possibilities of visualizing data. One should always choose that graph which is best for each data: best in terms of data presentation and interpretation, but also easiness of reading. When a type of graph is chosen, all its details should be carefully selected so that the graph conveys the message its author planned it to.

Appendix

Data sets used in the examples. All these data sets come from R (R Development Core Team, 2009), from the data sets of the same names.

ChickWeight
The data from a repeated-measure experiment on the effect of diet on early growth of chickens (Crowder and Hand, 1990).

fruitfly
The data on longevity and sexual activity of male fruitflies, originating from Partridge and Farquhar (1981), downloaded from the faraway package (Faraway, 2009) of R. See also the paper by Hanley and Shapiro (1994), which discusses interesting aspects of this data set and its usefulness in teaching statistics.

haynes
The data on phenotypic stability of resistance to late blight of potato clones; the variable considered is area under the disease progress curve (AUDPC) (Haynes et al., 1998). Data were taken from the agricolae package (Mendiburu, 2010) of R.

InsectsSprays
The data originating from Beall (1942). They include a number of insects counted after application of six insecticides, in 12 replications.

iris
A famous iris data set (Anderson 1935; Fisher, 1936), providing the measurements (in cm) of sepal length and width and petal length and width for 50 flowers from each of three species of iris: *Iris setosa*, *I. versicolor* and *I. virginica*.

trees
The data on measurements of the girth, height and volume of timber in 31 felled black cherry trees, provided by Ryan et al. (1976).

Software used for graphing and data analysis. The statistical analysis for Example 3 was performed with linear mixed-effects modeling framework offered by the nlme package (Pinheiro and Bates, 2000) of R (R Development Core Team, 2009). The graphs were constructed in the graphics (Murrell, 2005) and lattice packages of R (Sarkar, 2008).

References

Abou Khalifa, A.A.B. 2009. Physiological evaluation of some hybrid rice varieties under different sowing dates. Australian Journal of Crop Science 3: 178-183.

Ambang, Z.; Ndongo, B.; Amayana, D.; Djilé, B.; Ngoh, J.B.; Chewachong, G.M. 2009. Combined effect of host plant resistance and insecticide application on the development of cowpea viral diseases. Australian Journal of Crop Science 3: 167-172.

Ambrosano, E.J.; Trivelin, P.C.O.; Cantarella, H.; Ambrosano, G.M.B.; Schammass, E.A.; Muraoka, T.; Guirado, N.; rossi, F. 2009. Nitrogen supply to corn from sunn hemp and velvet bean green manures. Scientia Agricola 66: 386-394.

Anderson, E. 1935. The irises of the Gaspe Peninsula. Bulletin of the American Iris Society 59: 2-5.

Beall, G. 1942. The transformation of data from entomological field experiments. Biometrika 29: 243–262.

Brunnings, A.M.; Datnoff, L.E.; Ma, J.F.; Mitani, N.; Nagamura, Y.; Rathinasabapathi, B.; Kirst, M. 2009. Differential gene expression of rice in response to silicon and rice blast fungus *Magnaporthe oryzae*. Annals of Applied Biology 155: 161-170.

Carr, D.; Littlefield, R.J.; Nicholson, W.L.; J.S., Littlefield. 1987. Scatterplot matrix techniques for large *N*. Journal of the American Statistical Association 82: 424-436.

Cleveland, W.S. 1993. Visualizing Data. Hobart Press, Summit, NJ, USA. 360 p.

Cleveland, W.S. 1994. The Elements of Graphing Data. 2ed. Hobart Press, Summit, NJ, USA. 323 p.

Cleveland, W.S.; Fuentes, M. 1997. Trellis Display: Modeling Data from Designed Experiments; Technical Report. BellLabs, Paris, France.

Cleveland, W.S.; McGill, R. 1884. The Many Faces of a Scatterplot. Journal of the American Statistical Association. 79: 807-822.

Chambers, J.M.; Hastie, T.J. 1992. Statistical Models in S. Brooks/ Cole Wadsworth, Florence, KY. 608 p.

Crowder, M.; Hand, D. 1990. *Analysis of Repeated Measures*. Chapman and Hall, NY, USA. 257 p.

Cumming, G.; Finch, S. 2005. Inference by eye: confidence intervals and how to read pictures of data. American Psychologist 60: 170-180.

Cumming, G.; Williams, J.; Fidler, F. 2004. Replication, and researchers' understanding of conûdence intervals and standard error bars. Understanding Statistics 3: 299-311.

Czubaszek, A. 2009. The effects of genotype and environment on selected traits of oat grain and flour. Plant Breeding and Seed Science 60: 45-60.

Dupont, W.D.; Plummer, W.D.J. 2003. Density distribution sunûower plots. Journal of Statistical Software 8: 1-5.

Ehrenberg, A.S.C. 1977. Rudiments of numeracy. Journal of the Royal Statistical Society. Serie. A. 140: 277-297.

EL-Shafey, N.M.; Hassaneen, R.A.; Gabr, M.M.A.; EL-Sheihy, O. 2009. Pre-exposure to gamma rays alleviates the harmful effect of drought on the embryo-derived rice calli. Australian Journal of Crop Science 3: 268-277.

Faraway, J. 2009. Faraway: functions and datasets for books;. R package version 1.0.4. Available at: http://www.maths.bath.ac.uk/~jjf23/ [Accessed Dec. 18, 2009]

Fiorio, P.R.; Dematte, J.A.M. 2009. Orbital and laboratory spectral data to optimize soil analysis. Scientia Agricola 66: 250-257.

Fisher, R.A. 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics 7: 179-188.

Friendly, M. 1994. A Fourfold Display for 2 by 2 by K Tables York University, Psychology Department, Toronto, Canada. (Technical Repport 217).

Giacomini, A.A.; Da Silva, S.C.; Sarmento, D.O.L; Zeferino, C.V.; Da Trindade, J.K.; Souza Júnior, S.J.; Guarda, V.A.; Sbrissia, A.F.; Nascimento Júnior, D. 2009. Components of the leaf area index of marandu palisadegrass swards subjected to strategies of intermittent stocking. Scientia Agricola 66: 721-732.

Hanley, J.A.; Shapiro, S.H. 1994. Sexual activity and life span of male fruit flies: a data set that gets attention. Journal of Statistical Education 2.

Harris, R.L. 1999. Information Graphics: **A** Comprehensive Illustrated Reference. Oxford University Press, Oxford, UK. , 448 p.

Haynes, K.G.; Lambert, D.H.; Christ, B.J.; Weingartner, D.P.; Douches, D.S.; Backlund, J.E.; Fry, W.; Stevenson, W. 1998. Phenotypic stability of resistance to late blight in potato clones evaluated at eight sites in the United States. American Journal of Potato Research 75: 211-217.

Huff, D. 1954. How to lie with statistics. WW Norton, New York, MY, USA. 142p.

Hydman, R.J. 1996. Computing and graphing highest density regions. The American Statistician 50: 120-126.

Jacoby, W.G. 1997. Statistical Graphics for Univariate and Bivariate Data: Statistical Graphics. Sage, Thousand Oaks, CA, USA. 97 p.

Jacoby, W.G. 1998. Statistical Graphics for Visualizing Multivariate Data. Sage, Thousand Oaks, CA, USA. 103 p.

Jaradat, A.A. 2007. Predictive grain yield models based on canopy structure and structural plasticity. Communications in Biometry and Crop Science 2: 74–89.

Jaradat, A.A. 2009. Modeling biomass allocation and grain yield in bread and durum wheat under abiotic stress. Australian Journal of Crop Science 3: 237-248.

Jonsson, C.N.; Aoyama, H. 2009. Extraction, partial characterization and Susceptibility to Hg^{2+} of acid phosphatase from the microalgae *Pseudokirchneriella subcapitata*. Scientia Agricola 66: 634-642.

Karlidag, H.; Yildirim, E.; Turan, M. 2009. Salicylic acid ameliorates the adverse effect of salt stress on strawberry. Scientia Agricola 66: 180-187.

Kozak, M. 2009a. Analyzing one-way experiments: a piece of cake or a pain in the neck? Scientia Agricola 66: 556-562.

Kozak, M. 2009b. Text-table: an undervalued and underused tool for communicating information. *European Science Editing* 35: 103-105b.

Lammel, D.R.; Brancalion, P.H.S.; Dias, C.T.S.; Cardoso, E.J.B.N. 2007. Rhizobia and other legume nodule bacteria richness in brazilian *Araucaria angustifolia* forest. Scientia Agricola 64: 400-408.

Macedo, O.J.; Barbin, D.; Mourao, G.B. 2009. Genetic parameters for post weaning growth of Nellore cattle using polinomyals and trigonometric functions in random regression models. Scientia Agricola 66: 522-528.

Maciá-Vicente, J.G.; Rosso, L.C.; Ciancio, A.; Jansson, H.B; Lopez-Lorca, L.V. 2009. Colonisation of barley roots by endophytic *Fusarium equiseti* and *Pochonia chlamydosporia*: Effects on plant growth and disease. Annals of Applied Biology 155: 391-401.

Marcinkowska, J.; Boros, L.; Wawel, A. 2009. Response of pea (*Pisum sativum* L.) cultivars and lines to seed infection by Ascochyta blight fungi. Plant Breeding and Seed Science 59: 75-86.

Mendiburu, F. Agricolae: Statistical Procedures for Agricultural Research. R: 2010. Package Version 1.0-9. Available at: http://CRAN.R-project.org/package=agricolae [Accessed Dec. 18, 2009]

Moussa, H.R.; Abdel-Aziz, S.M. 2008. Comparative response of drought tolerant and drought sensitive maize genotypes to water stress. Australian Journal of Crop Science 1: 31-36. Murrell, P. 2005. R Graphics. Chapman Hall, New York, NY, USA. 328 p.

Ofosu-Anim, J.; Leitch, M. 2009. Relative efficacy of organic manures in spring barley (*Hordeum vulgare* L.) production. Australian Journal of Crop Science 3: 13-19.

Pahlavani, M.H.; Miri, A.A.; Kaziem, G. 2009. Response of oil and protein content to seed size in cotton(*Gossypium hirsutum* L., cv. Sahel). Plant Breeding and Seed Science 59: 53-64.

Pinheiro, J.P.; Bates, D.M. 2000. Mixed-effects models in S and S-Plus. Springer, New York, NY,USA. 528 p.

Partridge, L.; Farquhar, M. 1981. Sexual activity and the lifespan of male fruitflies. Nature 294: 580-581.

R Development Core Team. 2009. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Reynolds, M.; Manes, Y.; Izanloo, A.; Langridge, P. 2009. Phenotyping approaches for physiological breeding and gene discovery in wheat. Annals of Applied Biology 155: 309-320.

Ryan, T.A.; Joiner, B.L.; Ryan, B.F. 1976. *The Minitab Student Handbook*. Duxbury Press, Pacific Grove, CA, USA.

Sarkar, D. Lattice Multivariate Data Visualization with R. 2008. Springer, New York, NY,USA. 265 p.

Scarpari, M.S.; Beauclair, E.G.F. 2009. Physiological model to estimate the maturity of sugarcane. Scientia Agricola 66: 622-628.

Silva, S.C.; Oliveira Bueno, A.A.; Carnevalli, R.A.; Uebele, M.C.; Bueno, F.O.; Hodgson, J.; MAtthew, C.; Arnold, G.C.; M, J.P.G. 2009a. Sward structural characteristics and herbage accumulation of *Panicum maximum* cv. Mombaça subjected to rotational stocking managements. Scientia Agricola 66: 8-19

Silva, M.M.; Libardi, P.L.; Fernandes, F.C.S.. 2009c. Nitrogen doses and water balance components at phenological stages of corn. Scientia Agricola 66: 512-521.

Silva, R.B.T.R.; Nääs, I.A.; Moura, D.J. 2009b. Broiler and swine production: animal welfare legislation scenario. Scientia Agricola 66: 713-720

Simoes, M.S.; Rocha, J.V.; Lamparell, R.A.C. 2009. Orbital spectral variables, growth analysis and sugarcane yield. Scientia Agricola 66: 451-461.

Spence, I.; Lewandowsky, S. 1990. Graphical perception. p. 13–57, In: Fox ,J.; Long, J.S., eds. Modern methods of data analysis, Sage Thousand Oaks, CA, USA.

Takahashi, D.; ditt, R.F.; Lambais, M.R. 2009. Cloning of putative *ure*G genes from *Glomus intraradices* and urease activities in tobacco arbuscular mycorrhizal roots. Scientia Agricola 66: 258-266.

Tufte, E.R. 2001. The Visual Display of Quantitative Information. 2ed. Graphics Press, Cheshire, CT, USA. 199 p.

Tufte, E.R. 1991. Envisioning Information. Graphics Press, Cheshire, CT, USA. 126 p.

Tufte, E.R. 1997. Visual Explanations. Graphics Press, Cheshire, CT, USA. 157 p.

Tufte, E.R. 2006. Beautiful Evidence. Graphics Press, Cheshire, CT, USA., 213 p.

Vázquez, E.V.; Vieira, S.R.; De Maria, I.C.; González, A.P. 2009. Geostatistical analysis of microrelief of an Oxisol as a function of tillage and cumulative rainfall. Scientia Agricola. 66: 225-232.

Virdis, A.; Motzo, R.; Giunta, F. 2009. Key phonological events in globe artichoke (*Cynara cardunculus* var. *scolymus*) development. Annals of Applied Biology 155: 419-429.

Wiewióra, B. 2009. Long-time storage effect on the seed health of spring barley grains. Plant Breeding and Seed Science 59: 3-12.

Wilkinson, L. 2005. The Grammar of Graphics. 2ed. Springer-Verlag, New York, NY, USA. 690 p.

Received December 19, 2009
Accepted January 05, 2010