TECHNIQUES TO AVOID: WHAT MOMMA SHOULD HAVE TOLD YOU ABOUT SAS/GRAPH™

Arthur L. Carpenter

ABSTRACT

Several available graphic options and PROCedures produce graphs and charts that are either misleading or do not convey the information intended by the creator of the graphic. This paper demonstrates a few of these commonly applied SAS/GRAPH graphics and options, and discusses their pitfalls and why they should be avoided.

The ability to distinguish values and to interpret graphics depends on our perceptions, and we do not always perceive what we think we do. Research has shown that some types of graphs and some combinations of graphical elements are more easily and accurately interpreted than others. Correct selection of the elements can lead to a strong presentation, while other combinations of these elements can be misleading or even deceptive.

INTRODUCTION

The following lists, in decreasing order of our ability to discriminate, various graphical elements (Cleveland, 1984):

1 Position along a common scale
2 Position along nonaligned scales
3 Length, direction, and angle
4 Area
5 Volume and curvature
6 Shading and color saturation

As the numbers indicate there is a gradation in how well we can interpret graphics which utilize the above elements. The values of two points positioned using a common scale might be easy to compare, while the same values would probably be more difficult to compare if color shading was used to distinguish them.

SAS/GRAPH can easily be used to produce graphics that ignore the failings of our minds. The following are examples taken from a section in my book *Visualization of Scientific Data: Using SAS/GRAPH Without Annotate*.

STACKED BAR CHARTS

Stacked bar charts are used to show the values of two or more variables (or categories) on each bar. Both PROCs CHART and GCHART allow the use of the SUBGROUP= option to easily create this kind of graphic. The problem with this type of chart is that the stacked values are perceived as being on a common scale, when actually the scales are nonaligned. Not only are they nonaligned, all but the first set of vertical values do not have a common base.

```
proc gchart data=vol1.lampara;
title j=l h=1 f=simplex 'Example 5.6.1';
title2 h=2 j=c f=duplex 'Stacked Histogram';
vbar length / subgroup=sex
        midpoints = 100 110 120 130 140 150
                    160 170 180 190 200;
pattern1 value=l1 c=black;
pattern2 value=x1 c=blue;
pattern3 value=r1 c=green;
run;
```

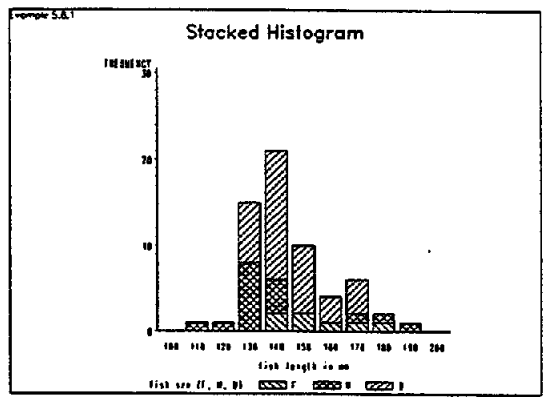


Figure 5.6.1

Because the stacked segments do not have a common base it is difficult to compare values such as the number of Males in the 130mm category and the number of Undeterminedss in the 150mm category

Plots in layers

The PROC GPLOT equivalent to the stacked values discussed in Example 5.6.1 uses the AREAS = option in the PLOT statement. Unlike the stacked histograms, the AREAS = option is designed to plot two or more variables that always take on monotonically greater than values i.e. for any value of the horizontal variable y1 < y2 < y3 etc. While this provides an aligned (vertically and horizontally) scale, our eye perceives a non-aligned scale. The plot in Figure 5.6.2 shows two curves with a constant vertical difference, however the use of the AREAS= option (aggravated by the choice of PATTERNs) makes it very difficult to see the constant difference. Our eye sees the horizontal spacing as well, thereby totally hiding the constant difference.

```
data inverse;
do x = .10 to 2 by .05;
  y1 = 1/x;
  y2 = 1/x + 1;
  output;
end;

proc gplot data=inverse;
title j=l h=1 f=simplex 'Example 5.6.2';
title2 h=2 j=c f=duplex 'Layers Plot of an Inverse Function';
title3 h=2 j=c f=simplex 'Vertical Width Of The '
```

```
'Upper Layer Is Constant';
plot y1*x y2*x / overlay haxis = 0 to 2 by .5
            areas=2;

symbol1 i=join v=none l=1;
symbol2 i=join v=none l=2;

pattern1 value=m3n90 color=black;
pattern2 value=m3n0 color=red;
run;
```

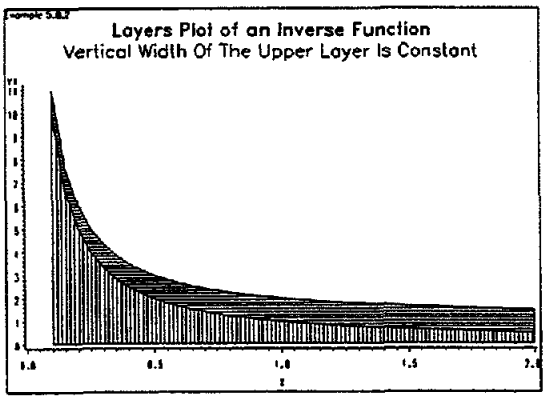


Figure 5.6.2

BLOCK CHARTS

Block charts are created by using the BLOCK statement in PROC PLOT. SAS™ will create a series of 'pseudo' three dimensional towers with the height of each representing the value of the selected variable. When two classification variables are selected, the towers are arranged like a small city of skyscrapers.

In addition to being a problem of nonaligned scales, our eye wants to see a three-dimensional picture, and will attempt to use perspective where none has been added. This can cause a tower in the back row to appear taller than one of equal height in the front row, because our eye wants to see the back one as being farther away. Notice that the parallel grid lines running from the front to the back seem to diverge in Figure 5.6.3, this is an optical illusion, and is an artifact of the eye trying to put the graph into perspective.

```
options ps=55 ls=140;

proc gchart data=vol1.lampara;
title1 h=2 j=c f=duplex 'Block Chart';
footnote j=l h=1 f=simplex 'Example 5.6.3';
block length / group=sex
        midpoints = 110 140 170 200;
run;
```

OPTICAL ILLUSION AND THE PATTERN STATEMENT

Some combinations of lines and especially patterns just do not go well together. As we design our graphic and choose our fill patterns, we must guard against occurrences of charts such as the one shown in Figure 5.6.4. Although the vertical bars are parallel and equally spaced, they appear to be leaning toward or away from their neighbor.

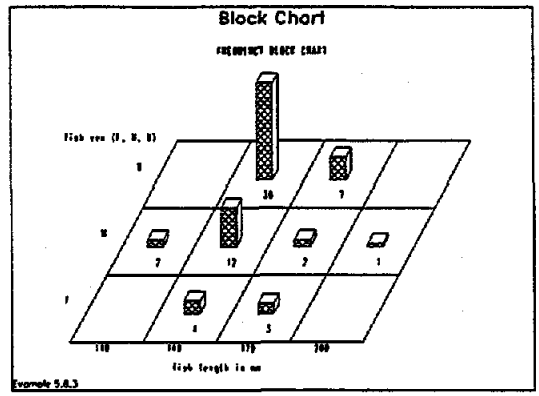


Figure 5.6.3

```
proc gchart data=vol1.lampara;
title j=l h=1 f=simplex 'Example 5.6.4';
title2 h=2 j=c f=duplex 'Patterns Seem to Create Leaning
Towers';
vbar length / patternid=midpoint
        midpoints = 120 140 160 180 ;
pattern1 value=l1 color=black;
pattern2 value=r1 color=black;
pattern3 value=l1 color=black;
pattern4 value=r1 color=black;
run;
```

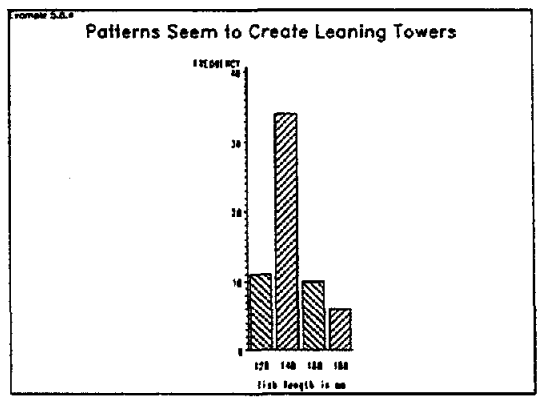
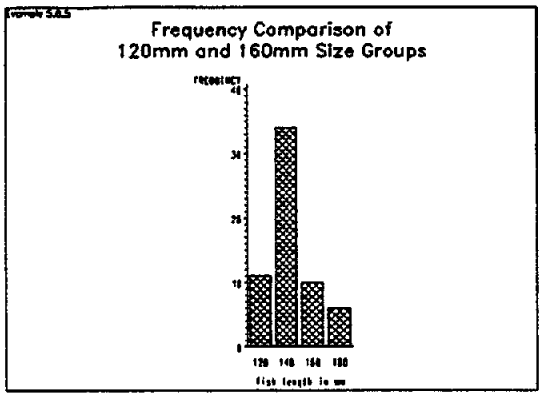


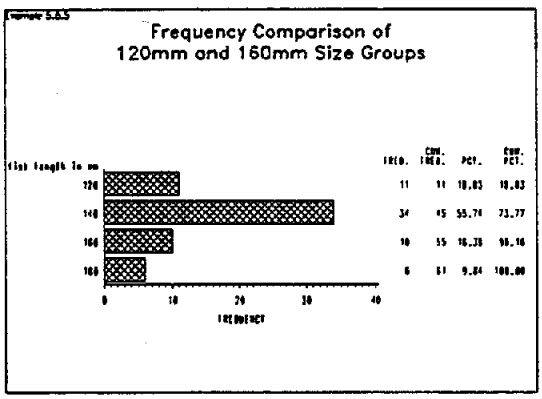
Figure 5.6.4

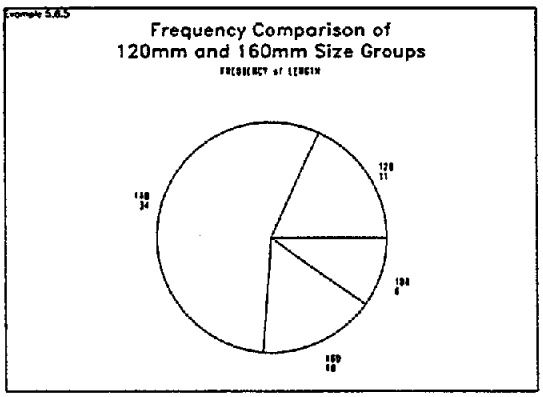
Pie chart

The pie chart is one of the classical graphical alternatives to the histogram (bar chart). Certainly within SAS, bar charts and pie charts are created by the same PROCedures. Pie charts are not bad, but the programmer (and wouldn't it be nice, the user of the graphic) needs to be aware that we can more accurately distinguish the length of bars with a common scale than we can the area of a circular sector. Exploded pie charts (EXPLODE= option in the PIE statement) are even worse because they trick the eye into seeing too much area.

```
proc gchart data=vol1.lampara;
title j=l h=1 f=simplex 'Example 5.6.5';
title2 h=2 j=c f=duplex 'Frequency Comparison of';
title3 h=2 j=c f=duplex '120mm and 160mm Size Groups';
vbar length / midpoints = 120 140 160 180 ;
hbar length / midpoints = 120 140 160 180 ;
pie  length / midpoints = 120 140 160 180 ;
pie  length / midpoints = 120 140 160 180 explode=160;
```







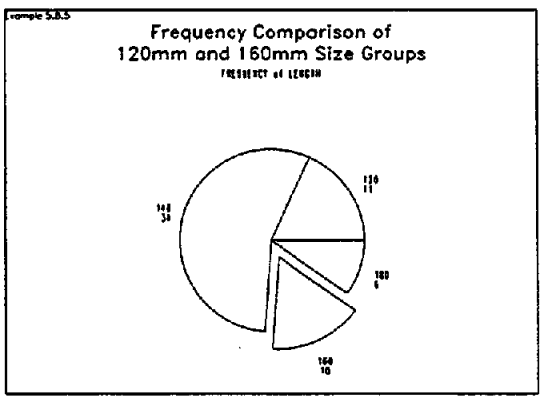


Figure 5.6.5

Each of the four graphs in Figure 5.6.5 depict the same data values. Although the size classes of 120mm and 160mm are close in size, it is fairly easy to distinguish which is larger in the two histograms (they have an aligned scale). It is much more difficult to determine the larger group by inspection of the two pie diagrams (comparison of area).

It is for this reason that some readers may perceive the exploded section (160mm) in the lower right corner of Figure 5.6.5 to be larger than the 120mm section when in fact it is not.

NONALIGNED AXES

The PLOT2 and BUBBLE2 statements in PROC GPLOT allow the user to specify a second vertical axis when overlaying two variables with a common horizontal axis. These two axes may or may not have values in common, however, our eye will tend to want to compare the magnitude of two values by looking at their height above the horizontal axis rather than by seeking out their true values.

```
proc gplot data=vol1.ca88air;
plot o3 * month = 'O';
plot2 co * month = 'C';
where station='SFO';
title1 j=l h=1 'Example 5.6.6';
title2 h=2 f=duplex '1988 Air Quality Data - SFO';
title3 h=2 f=simplex 'Two Non-aligned Axes';
run;
```

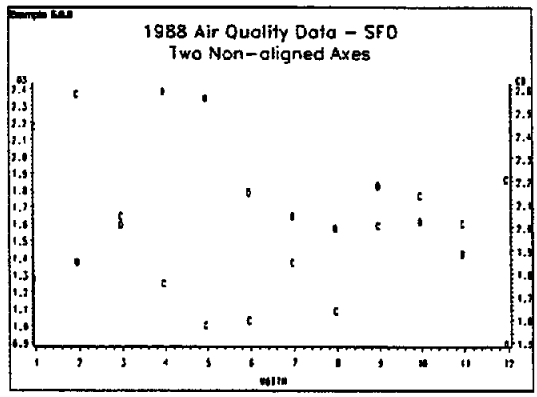


Figure 5.6.6

The non-aligned axes in Figure 5.6.6 make it difficult to compare concentrations of ozone and carbon monoxide. In September (month=9) it first appears that the O3 concentration is greater since it is higher on the plot. Closer inspection, however shows that O3 was equal to about 1.85 while CO was over 2.0. Our eye has been fooled by the non-aligned axes.

BUBBLE PLOTS ETC.

The BUBBLE and BUBBLE2 statements in PROC GPLOT, and the SCATTER statement with the SIZE= option in PROC G3D produce symbols whose size represents the magnitude of another variable. Unfortunately, the ability to compare areas is not very high on the list of accurate graphical tasks consequently size comparison is difficult.

At least with the BUBBLE and BUBBLE2 statements, comparisons are made among symbols that have the same shape. In a PROC G3D SCATTER plot, different shapes can be used to add another dimension to the plot, but our eye does not easily compare the volumes of a heart, spade, and star. The problem is compounded by the 'pseudo' three-dimensional effect

discussed above.

```
data air; set vol1.ca88air;
if station='AZU' then do;
   shapevar='club';
   colorvar='black';
end;
else if station='LIV' then do;
   shapevar='spade';
   colorvar='blue';
end;
else if station='SFO' then do;
   shapevar='heart';
   colorvar='green';
end;
* temperatures range from 40 to 80;
* create a scaled variable (0 to 3) based on temperature;
min=40; max=80;
sizevar = (tem-min)/(max-min)*3;
run;

proc g3d data=air;
scatter o3 * co = no3 / size=sizevar
                  shape=shapevar
                  color=colorvar
                  grid;
title1 j=l h=1 'Example 5.6.7';
title2 h=2 f=duplex '1988 Air Quality Data - SFO';
title3 h=2 f=simplex 'Temperature Relationship to CO, O'
           h=1 move=(+0,-.5) '3'
           h=2 move=(+0,+.5) ', and NO'
           h=1 move=(+0,-.5) '3';
title4 h=2 f=simplex 'Symbol size relates to Temperature';
run;
```

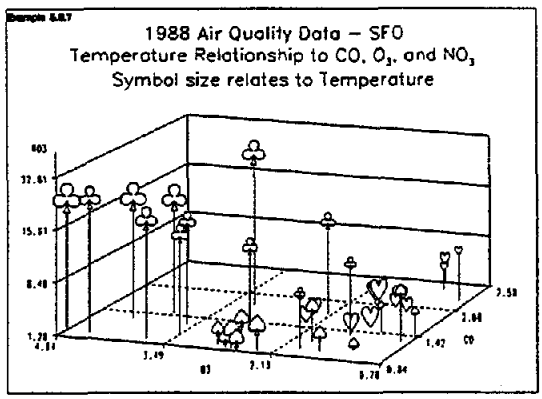


Figure 5.6.7

SUMMARY

It is not enough to be able to transfer what we visualize that a graphic should look like into SAS code. We must always be aware that how we construct the graph, with our choice of graphical elements, will determine how those who look at our graph will interpret what they see.

TRADEMARK INFORMATION

SAS and SAS/GRAPH are registered trademarks of SAS Institute, Inc. in the USA and other countries. ™ indicates USA registration.

REFERENCES

Carpenter, A. and C. Shipp. in press. *Visualization of Scientific Data: Using SAS/GRAPH Without Annotate*. SAS Institute, Inc. Cary, NC.

Cleveland, W. and R. McGill. 1984. "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods". *Journal of the American Statistical Association*. Volume 79. Number 387. pp. 531-554.

ABOUT THE AUTHOR

Arthur L. Carpenter has over seventeen years of experience as a statistician and data analyst and has served as a senior consultant with California Occidental Consultants, CALOXY, since 1983. His publications list includes a number of papers and posters presented at SUGI and has served as a steering committee member and president of the Southern California SAS User's Group, a Section Chair at the Western Users of SAS Software regional conference, WUSS, and in various positions at SUGI. He has developed and presented several courses and seminars on statistics and SAS programming and has taught for Colorado School of Mines, University of Redlands, and University of California at San Diego.

CALOXY offers SAS contract programming and in-house SAS training nationwide, including a three day course on SAS/GRAPH. This presentation is based on a section in his latest book, *Visualization of Scientific Data: Using SAS/GRAPH Without Annotate*, which was co-authored with Charlie Shipp.

AUTHOR

Arthur L. Carpenter
California Occidental Consultants
4239 Serena Avenue
Oceanside, CA 92056-5018

(619) 724-8579