Paper DG09-2011

SG Techniques: Telling the Story Even Better!

Chuck Kincaid, Experis, Portage, MI

ABSTRACT

The SAS™ Statistical Graphics (SG) procedures: SGPLOT, SGPANEL, SGSCATTER, and SGRENDER are exciting additions in 9.2 that give easy access to some of the power of the Graphics Template Language (GTL) in much the same way that macros can give a novice access to advanced Base SAS programming. As good as these procedures are in helping you tell your analytical story, and they are good, there are techniques that can add extra value to your graphs.

This paper will discuss three concepts – level ordering, banking, and slicing – introduced by William S. Cleveland in his book Visualizing Data. Ordering the levels and cells in a plot can add another level of information to the plots. Banking is a technique that enhances the reader's ability to visually judge lines in the plot. Finally, slicing is a technique that allows a quantitative variable to be used to define the cells of the panel. For these techniques, we will explain the concepts, provide examples and outline the basic algorithms in this paper. The audience for this presentation is the statistician or business analyst who wants to tell their story even better.

INTRODUCTION

By now, many have seen and, hopefully, used the SAS™ Statistical Graphics procedures. These procedures are easy to use for straightforward, professional quality graphics and, at the same time, can be used to build multi-dimensional graphics that convey complex information. This paper discusses three techniques that are not yet part of the SG capabilities. If you would like to see them incorporated, let the SAS Developers know.

Graphics are created to tell a story visually according to the analysis undertaken. These three techniques – slicing, ordering, and banking – help to tell that story better. Until they are incorporated into the SG procedures directly, each one requires preparation of the data creating new, likely temporary, datasets. These datasets are then used with the respective SG procedure. Each of these algorithms has been turned into a macro that is available upon request.

All three concepts can be found in Bill Cleveland's book, *Visualizing Data*. This seminal book was developed while he was at Bell Labs and published in 199X. He has many other exciting ideas in there and it is well worth a read.

ORDERING LEVELS AND CELLS

Cleveland widely introduced the coplot in his book, *Visualizing Data*. The examples first used were based on trivariate data, that is, measurements of three quantitative variables. These displays had a natural ordering on each axis and for the cells. The way the given levels were created used a technique called *slicing* that we'll discuss below. The move to multiway data, that is, one quantitative response variable and two categorical variables does not necessarily provide a natural ordering for the axes and for the cells. However, imposing an ordering based on a quantitative variable of interest can help in eliciting even more information from presentation. Let's look at the multiway example he begins with.

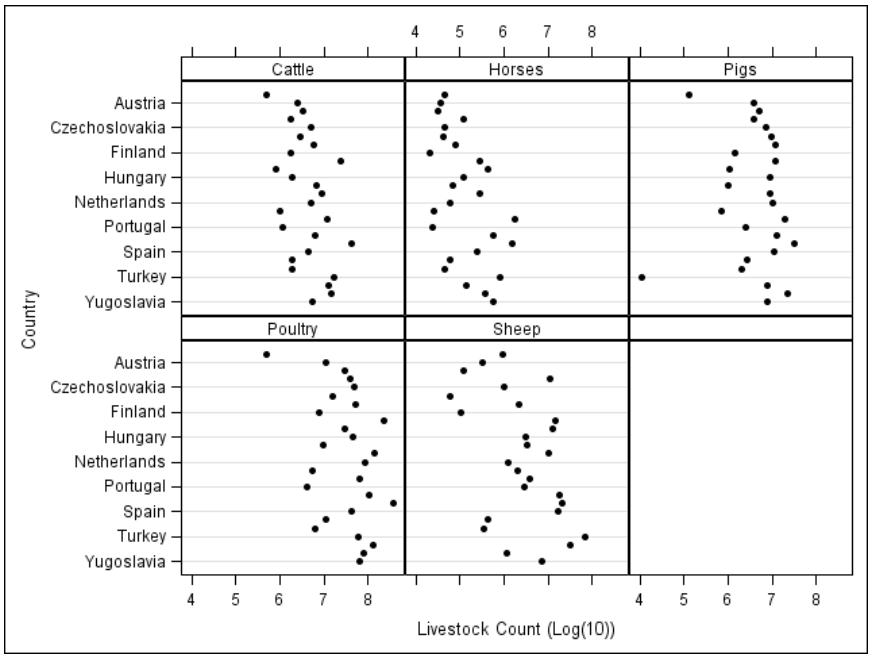

Figure 1

Figure 1 above "graphs the logarithms of livestock counts from a 1987 census of farm animals in 26 countries. … Russia *et al*. is the European part of Russia and the European countries that were formerly part of the Soviet Union." The data is graphed with a dot plot and the quantitative variable is displayed as log count because of the high variability in the counts. The default ordering of the countries on the vertical axis is alphabetical, and the order of the cells from upper left to lower right is alphabetical by Livestock Type, as well. Alphabetical ordering provides no information towards our understanding of the data. (Note that not all countries show on the axis because of the size of the graph.)

The data in this presentation appear very unassuming. The number of pigs in Turkey seems unusual, but otherwise there is not a lot that jumps out at us. However, if we order the levels of the class variable in the dot plots (country), then more information can be found.

The following table shows Country ordered by its median livestock count across all livestock type.

Country	Median	Country	Median
Albania	478,000	Czechoslovakia	5,131,000
Norway	971,000	Netherlands	5,241,000
Greece	1,107,000	East Germany	5,690,000
Finland	1,475,000	Yugoslavia	7,384,000
Sweden	1,902,000	Italy	8,928,000
Switzerland	1,954,000	Spain	11,263,000
Portugal	2,448,000	Poland	11,912,000
Austria	2,536,000	Romania	12,464,000
Denmark	2,873,000	United Kingdom	13,155,000
Hungary	3,183,000	France	14,346,000
Belgium	3,246,000	West Germany	15,098,000
Ireland	3,323,000	Turkey	16,983,000
Bulgaria	3,808,000	Russia et al.	33,100,000

This effectively converts country from a nominal variable to an ordinal variable and gives a rudimentary scatterplot.

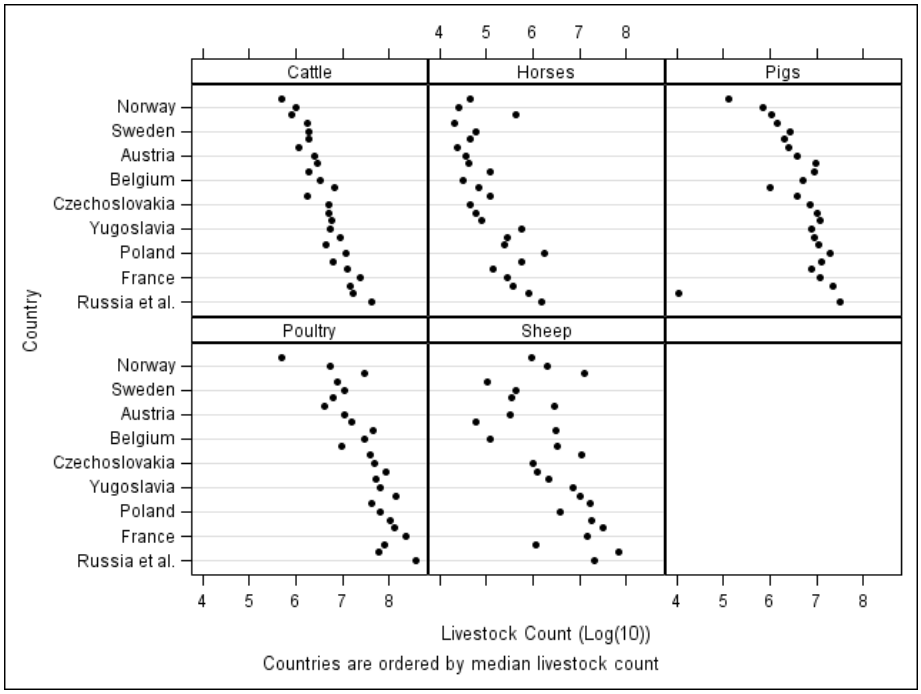

Figure 2

The new plot, Figure 2 above, tells the story better than the previous plot by giving us more information. We see that the number of Cattle and Pigs are strongly associated with the median number of livestock, and Sheep are weakly associated. The low value for Pigs in Turkey stands out even more than before, since it is now placed as the country with the second highest median number of livestock.

The algorithm for the levelOrdering.sas macro is as follows

1) Compute the class ordering
    a. Use The MEANS Procedure to calculate the statistic for the levels of the class variable.
    b. Sort the PROC MEANS output according to the calculated statistic. The observation number will be the value of the new class variable.
2) Build the format for the new class variable.
    a. This format will associate the text values of the original class variable with the numeric values of the new class variable.
    b. Create a CTLIN data set to use with PROC FORMAT.
3) Create the final data set.
    a. The new class variable can be used to sort the records according to their value of the statistic. It will display according to the values of the original class variable via the format above.

The SAS code that generated Figure 2 is

```
%levelOrdering (
  dsIn        = livestock,
  dsOut       = livestock2,

  oldClassVar = country,
  newClassVar = countryNum,

  stat        = median,
  statVar     = count,
  fmtName    = cty_fmt
  );

ods graphics / imagename="multiway ordering";
```

```
footnote1 "Countries are ordered by median livestock count";
proc sgpanel data=livestock2 description="multiway ordering";
    panelby livestock_type / novarname ;
    dot countryNum / response=count;
    colaxis type=log logbase=10 logstyle=logexponent alternate label="Livestock
Count (Log(10))";
    rowaxis fitpolicy=thin label="Country";
run;
```

This concept can also be used for ordering the cells in the panel and other places a class variable is used.

SLICING

As mentioned above, Cleveland started his examples with trivariate data, i.e. three quantitative variables. In order to do so, he needed a way to slice one of the variable's values into segments. Each segment defined the data points of the other two variables that would be included in the plot. In his book, Cleveland defines the *equal-count algorithm*, which sliced the data according to the input criteria keeping two objectives in mind. First, to have, as much as possible, each of the intervals contain the same number of values. The second objective is to have "the fraction of values shared by each pair of successive intervals as nearly equal to the target fraction as possible."[1] The overlap of data from interval to interval is designed to smooth out the changing of the intervals. Without overlap the changes in the dependent panels from cell to cell would be choppier. We want to see how the graphic in the cells changes as we move along it in a smooth manner.

Consider watching the scenery while riding in a car. Suppose you were to look at the scenery on your stretch of the road, take a 15 minute nap, wake up and look at it again. If you're on a highway in a place like western Kansas or northern Michigan, then that may be enough, since the view does not change that much. However, as the scenery changes more rapidly or the route gets more complex, your naps will need to be shorter and shorter, until you have to pay attention all along the way. As you pay more attention, the overlap in what you see becomes greater and greater.

Cleveland's algorithm starts with choices by the user of the number of intervals in which to slice the data and the target fraction of data points to share by successive intervals. There is a tradeoff between the number of intervals, the number of points in each interval and the amount of overlap in the intervals. Our goal is to have enough data plotted in each cell to see the underlying relationship that we are exploring, but not so much so that we haven't gained anything by slicing. We want enough intervals so that the changes in the relationships being explored (histograms, loess plots, etc.) can be seen in each plot without being masked by data from other surrounding plots. To do this is a trial and error process, like much of exploratory data analysis, and is left to the reader.

Let's look at the algorithm now, again translated from Cleveland's book.

Consider N univariate data points that we are making into buckets, sorted in ascending order as $(x_1, x_2, x_3, ..., x_N)$.

Each of the intervals we determine will have endpoints as values of x_i. Let

1) $k =$ number of conditioning intervals
2) $f =$ target fraction of data points shared by successive intervals
3) $l_j =$ lower endpoint of the j^{th} interval; Note that $l_1 = x_1$.
4) $u_j =$ upper endpoint of the j^{th} interval; Note that $u_N = x_N$.

Then, the number of values in each interval can be determined as…

5) $$r = \frac{N}{k(1-f)+f}$$

Note that r does not have to be an integer, which will be accounted for later. Note also that as k, the number of intervals, or slices, increases, then r decreases. Also, as the target fraction, f, increases the number of values in each interval increases. Both of these relationships make intuitive sense. From this we calculate, that the index of the lower endpoint for the j^{th} interval is

6) $1+(j-1)(1-f)r$,

rounded to the nearest integer. Similarly the upper endpoint of the j^{th} interval is

7) $r+(j-1)(1-f)r$

rounded to the nearest integer.

Consider an example using Cleveland's rubber data. Following Cleveland, we have selected 6 intervals and a target fraction of 3/4.

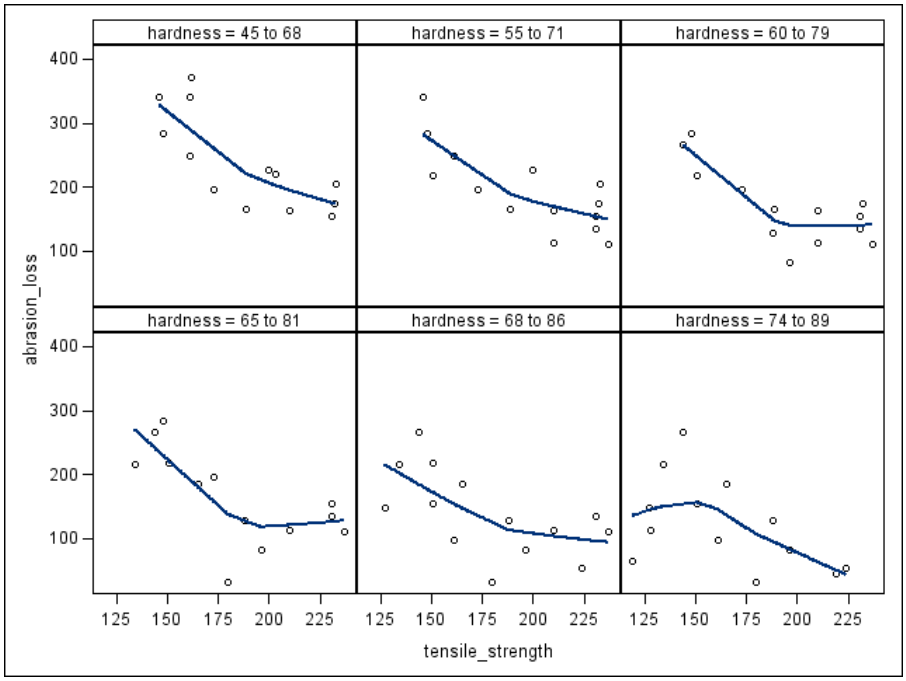

Figure 3

Figure 3 above was created with the following call to the macro and the SGPANEL procedure.

```
%slicing (
  dsIn          = rubber,
  dsOut         = rubber2,

  rankVar       = hardness,
  rankVar2      = tensile_strength,
  groupVar      = slice,

  noIntervals   = 6,
  overlapRatio  = .75
  );

proc sgpanel data=rubber2 noautolegend;
    panelby slice ;
    loess x=tensile_strength y=abrasion_loss;
run;
```

The macro, slicing.sas,

1) Calculates the indices according to the algorithm above
2) Creates an output data set with redundant records and a class variable, SLICE
3) Applies a format and a label to the class variable so that the output shows the values of the sliced variable, e.g. "hardness = 65 to 81"
4) The user can also provide a name and a library for the format.

Slicing is a great technique for creating panels according to the levels of a quantitative variable. The values for number of intervals and target fraction are often found by trial and error over a range for each.

BANKING

Banking is a technique “designed to maximize the discriminability of the orientations of the line segments in the chart.”[2] Edward Tufte also advocates banking in Beautiful Evidence with the following…

"… as William Cleveland discovered, for judging slopes and velocities up and down hills in time-series, best is an aspect ratio that yields hill-slopes averaging 45°, over all the cycles in the time-series. That is, variations in slopes are best detected when the slopes are around 45°, uphill or downhill. ... the aspect ratio should be such that the time-series graphics tend toward a lumpy profile rather than a spiky profile .. or a flat profile."[3]

The aspect ratio of a graph in Cleveland’s terminology is the ratio of the vertical scale (height) to the horizontal scale (width). For SAS procedures it is the opposite, that is, the ratio of the width to the height. However, in SAS we specify the height and width rather than directly specifying the aspect ratio. Therefore, we will use Cleveland’s terminology in our discussions.

To better understand the importance of the aspect ratio, consider the plot in Figure 4 based on one from Cleveland’s book.

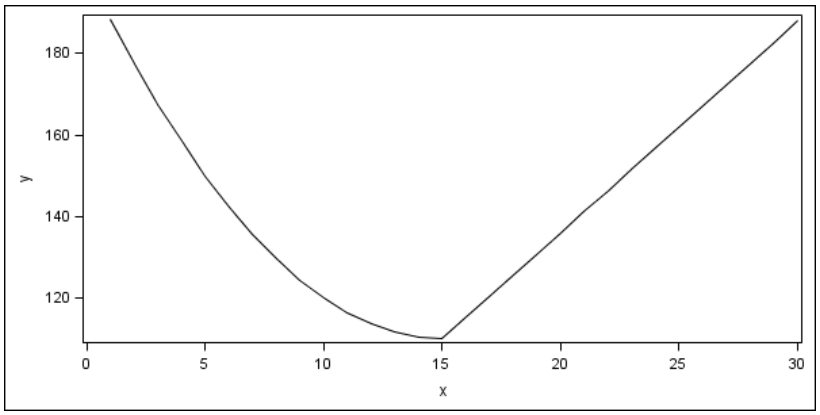

Figure 4

This aspect ratio clearly shows the curvature in the first half of the line. If we use the aspect ratios of 5 and 0.05, though, we get the following two plots, Figure 5 and Figure 6.

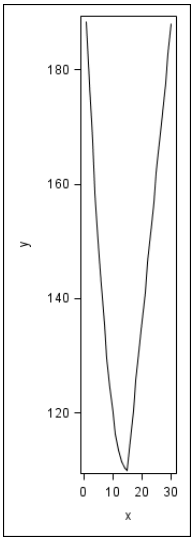

Figure 5

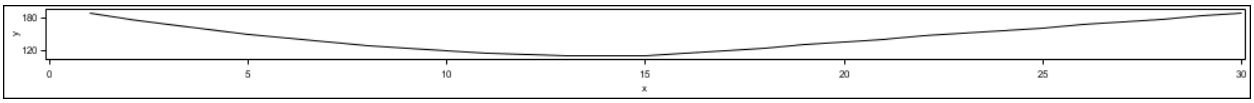

Figure 6

That curvature is not as readily available in either of these. To be fair the example is somewhat contrived. If we consider the default settings for SGPLOT, we get the following figure, Figure 7, in which the curvature is also readily apparent.

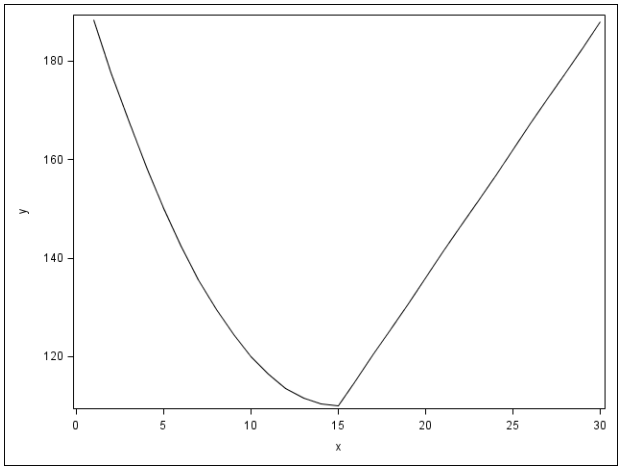

Figure 7

The aspect ratio can be controlled for Statistical Graphics by the ODS GRAPHICS statement and the WIDTH= and HEIGHT= options. The SAS default value for the aspect ratio is 0.75 (4/3, in SAS terminology) meaning that the height (480 pixels by default) is 75% of the width (640 pixels by default). If you specify only one of these two, then the other is automatically adjusted so that the default aspect ratio will be maintained. Most of the time this is the appropriate course of action, since, in many cases, the graph might have a specific layout. "For example, a plot that has multiple columns, or that has a statistics table on the side needs a wide aspect ratio. Changing the aspect ratio for this plot by specifying both width and height might produce unpredictable results."[4] However, for our purposes, this is exactly what we want to do.

Note that at this time, we will only consider changing the aspect ratio for the SGPLOT. The banking technique is very appropriate for multi-cell panels, where what we will see below for one cell is extended for all cells as if there were only one. All cells would then be set to the same aspect ratio. Because many factors make setting this aspect for multi-cell panels difficult, in this paper we will not consider the situation.

Another point to realize is that these options set the height and width of the entire plot and not just the data area. Whether the data area has the same aspect ratio depends on factors such as the size of the tick values, and whether and where a legend appears. Unfortunately, at this time, the SG procedures, nor GTL, provide more control. If this is of interest, let SAS know.

The idea behind banking is to consider the line segments defined by each pair of data points, their orientation and their length. First, we define some terms. Suppose we have N data points, (X_1,Y_1) … (X_N,Y_N). Then

8) h Horizontal length of the data in physical units (e.g. inches, cm)

9) v Vertical length of the data in physical units (e.g. inches, cm)

10) $h_i(h)$ Change in physical units when the horizontal length of the data rectangle is *h*

11) $v_i(v)$ Change in physical units when the horizontal length of the data rectangle is *v*

12) $a(h,v) = v/h$ Aspect ratio of the graph

13) $h_{span} = X_N - X_1$ Horizontal span of the data in data units (e.g. $, yrs, ft)

14) $v_{span} = Y_N - Y_1$ Vertical span of the data in data units (e.g. $, yrs, ft)

15) $\ddot{h}_i = |X_{i+1} - X_i|$ Absolute horizontal distance of the ith line segment in data units

16) $\ddot{v}_i = |Y_{i+1} - Y_i|$ Absolute vertical distance of the ith line segment in data units

17) $\bar{h}_i = \ddot{h}_i / h_{span}$ Relative width of the ith line segment

18) $\bar{v}_i = \ddot{v}_i / v_{span}$ Relative height of the ith line segment

19) $\theta_i(h,v) = \arctan(\frac{v_i(v)}{h_i(h)}) = \arctan(a(h,v)\bar{v}_i / \bar{h}_i)$ Orientation of the ith line segment

20) $l_i(h,v) = \sqrt{h_i^2(h) + v_i^2(v)} = h\sqrt{\bar{h}_i^{-2} + a^2(h,v)\bar{v}_i^{-2}}$ Physical length of the ith segment

To illustrate these concepts, consider the 5 points below. The horizontal data span is h_{span} = 8 (= 10 – 8) and the vertical span is v_{span} = 10 (= 18 – 8).

Data Point (X_i,Y_i)	Absolute Horizontal Data Distance $(\ddot{h}_i)$	Absolute Vertical Data Distance $(\ddot{v}_i)$	Relative Width $(\bar{h}_i)$	Relative Height $(\bar{v}_i)$
(2,8)				
(6,12)	|6 – 2| = 4	|12 – 8| = 4	4/8	4/10
(8,9)	|8 – 6| = 2	|9 – 12| = 3	2/8	3/10
(9, 16)	|9 – 8| = 1	|16 – 9| = 7	1/8	7/10
(10, 18)	|10 – 9| = 1	|18 – 16| = 2	1/8	2/10

Our focus will be how the aspect ratio affects the mean of the absolute orientations weighted by the line segment lengths. The optimal aspect ratio is one for which that mean is 45°. There are other criteria for finding the optimal aspect ratio. We'll only look at this one. The mean is

$$\frac{\sum_{i=1}^{N}\theta_i(h,v)l_i(h,v)}{\sum_{i=1}^{N}l_i(h,v)} = \frac{\sum_{i=1}^{N}\arctan(a(h,v)\bar{v}_i/\bar{h}_i)\sqrt{\bar{h}_i^{-2} + a^2(h,v)\bar{v}_i^{-2}}}{\sum_{i=1}^{N}\sqrt{\bar{h}_i^{-2} + a^2(h,v)\bar{v}_i^{-2}}}$$

This mean depends on *v* and *h* only through $a(h,v)$, that is, the aspect ratio. As stated, then, the optimal aspect ratio is one that makes this mean equal to 45°. Fortunately, the weighted mean is a monotone function of $a(h,v)$, since there is no closed-form solution. Typically, few iterations are needed to find the optimal aspect ratio. In SAS this ratio can be found using the OPTMODEL procedure with an objective function of

```
min bank_eq =
  (
    (
      sum{k in indx}
           atan(ar*vibar[k] / hibar[k]) * sqrt(hibar[k]**2 + ar**2 * vibar[k]**2)
    ) / (
           sum{k in indx} sqrt(hibar[k]**2 + ar**2 * vibar[k]**2)
```

```
        ) - &orient
)**2
   ;
```

Minimizing the square of the equation rather than the absolute value avoids certain issues in solving the equation.

The macro banking.sas works as follows

1) Expects user input on (among other things),
    a. either the desired height or the desired width, but not both
    b. the names of the macro variables in which the results will be returned
2) Calculates the terms indicated in the table above
3) Runs PROC OPTMODEL with the objective function above to determine the optimal aspect ratio
4) Calculates the new width or height, depending on which was entered, based on the optimal aspect ratio
5) Populates the specified macro variables with the width and height

The programmer can then use these macro variables in a statement like

```
ODS GRAPHICS / HEIGHT=&myVarHeight. WIDTH=&myVarWidth.;
```

This will change the aspect ratio of the (entire) graph area for all plots going forward. To reset the height and width to their defaults, the programmer would execute.

```
ODS GRAPHICS / RESET=all;
```

Because we don't have specific enough control, the height and width values may need to be modified to get the desired aspect ratio in the data area. In particular, one might need to account for the axes and legends.

The traditional example for banking, which works well because of the large difference between the optimal aspect ratio and the default aspect ratio, is the sunspot data. The default plot, Figure 8, provided by this code

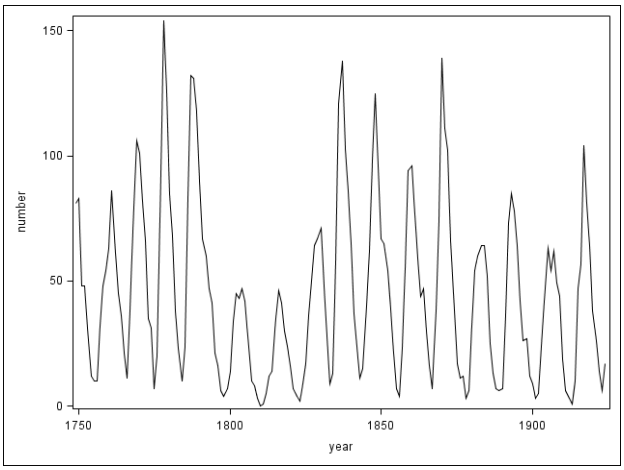

Figure 8

While this view is okay, let's look at it with the line segments banked to 45° as in Figure 9.

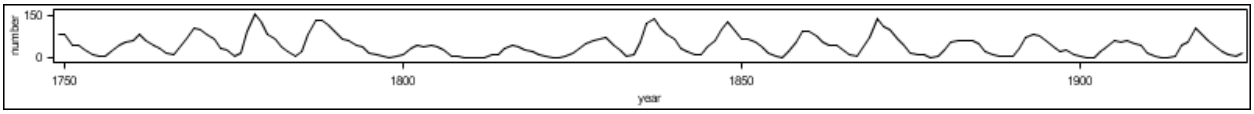

Figure 9

It is easier to compare the line segments to one another in this plot. We can now see that sunspot activity typically increases at a faster rate than it decreases.

Banking, as it is, is not as valuable as it could be. If there is enough interest and SAS adds it as an option to SGPLOT and SGPANEL, then it will be even more so. Note also, that there can be good reasons why the "optimal" aspect ratio isn't the desired one. The programmer should use their best judgment.

CONCLUSION

The Statistical Graphics procedures are a wonderful addition to the programmer's toolkit. They expand the programmer's ability to understand the stories that are in the data and then create persuasive graphs to share their work with others. The techniques described here are helpful additions to that toolkit.

The macros described in this paper are available by contacting the author. They are provided as is and we ask that you retain the header documentation.

REFERENCES

[1] Cleveland, William S., Visualizing Data, Hobart Press, 1993

[2] Heer, Jeffrey and Maneesh Agrawala, Multi-Scale Banking to 45°, IEEE Transactions on Visualization and Computer Graphics, Vol 12, No. 5, September/October 2006

[3] Tufte, Edward, Beautiful Evidence

[4] SAS Help Documentation: Managing Your Graphics With ODS: Using the ODS GRAPHICS Statement

ACKNOWLEDGMENTS

The first author would like to thank the second author, Jack Fuller, for his programming expertise in creating the macros for publication. Also, we acknowledge the R&D staff at SAS who developed these procedures and are very willing to talk about them at any time.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Chuck Kincaid
Experis
5220 Lovers Lane
Portage, MI 49002
(269) 553-5140
chuck.kincaid@experis.com
www.experis.us/analytics