Methodologies to Create Multiple Symbols on a Single Line Plot
Sanjiv Ramalingam, Octagon Research Solutions, Inc., Wayne, PA
Lelia McConnell, SAS Institute, Cary, NC

ABSTRACT

The GPLOT procedure is widely used to create line plots. Situations may arise wherein multiple symbols are required within a single line plot to illustrate differences within the data. A typical example is to illustrate percent survival (or failure) as a function of time. Current options in SAS limit the use of symbols to one symbol per line plot. A possible solution is to create multiple lines per symbol and then overlay each plot. This solution is however data dependent and requires that there are no time points common to each of the individual plots. Two different approaches are presented. The first methodology uses the annotate feature and the other uses the GPLOT procedure.

METHODOLOGY

A graph of the percent survival (or failure) as a function of time is used to illustrate the effectiveness of the algorithm. The LIFETEST procedure is used to obtain the survival distribution function estimate. The dataset (srcdata) used to create the annotate dataset should be of the structure below.

<table>
<thead>
<tr>
<th>TIME</th>
<th>SURVIVAL</th>
<th>SYMBTYP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>100</td>
<td>2</td>
</tr>
<tr>
<td>0.28</td>
<td>99.53</td>
<td>2</td>
</tr>
<tr>
<td>0.57</td>
<td>98.59</td>
<td>1</td>
</tr>
<tr>
<td>0.61</td>
<td>96.26</td>
<td>1</td>
</tr>
<tr>
<td>2.71</td>
<td>95.45</td>
<td>2</td>
</tr>
<tr>
<td>4.68</td>
<td>86.34</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>52</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

TIME is representative of the X-axis.
SURVIVAL is representative of the Y-axis and reflects the survival function distribution estimate.
SYMBTYP is representative of the number of different symbols required. In the above example, the values 1 and 2 represent subjects at end of treatment and those who discontinued, respectively.

Case 1: Using Annotate

Only three different types of functions are used, namely the MOVE, DRAW and SYMBOL functions. The objective here is to automatically create the MOVE and DRAW statements for each of the symbols used. The algorithm to create the annotate dataset is discussed below.

Step 1
Each of the observations in the dataset used to create the annotate dataset are identified by an id variable. This is implemented by assigning a variable(I) to _N_.

<table>
<thead>
<tr>
<th>TIME</th>
<th>SURVIVAL</th>
<th>SYMBTYP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>100</td>
<td>2</td>
</tr>
<tr>
<td>0.28</td>
<td>99.53</td>
<td>2</td>
</tr>
<tr>
<td>0.57</td>
<td>98.59</td>
<td>1</td>
</tr>
<tr>
<td>0.61</td>
<td>96.26</td>
<td>1</td>
</tr>
<tr>
<td>2.71</td>
<td>95.45</td>
<td>2</td>
</tr>
<tr>
<td>4.68</td>
<td>86.34</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>52</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
**Step 2**

As a general rule, whenever drawing a line, the function MOVE is assigned first. In the algorithm the steps for the first record (assignment of the function MOVE) are different than those followed for the rest of the records. Hence the function assignment is done separately for this record. Among the different co-ordinate systems in annotate, only the absolute data area will be used hence XSYS and YSYS will be equal to 2. Two different symbols are used in this example. A circle is used to illustrate subjects who discontinued and a dot (filled circle) is used to illustrate end of treatment for those subjects.

```plaintext
data anno1;
  length function $8 text $1;
  retain line 1 xsys ysys '2' color 'black';
  set srcdata;
  if _n_ eq 1 then do;
    function="symbol";
    if symbtyp eq 1 then do;
      style="special"; text="J"; size=1.5;
    end;
    else do;
      style="cartog"; text="Q"; size=0.55;
    end;
  x=time; y=survival;
  output;
  function='move'; x=time; y=survival; case=1; output;
end;
if i ne 1 then output;
run;
```

The output dataset anno1 will differ from the source dataset (srcdata) by only one observation. The function assignment will only be made to the first two records of this dataset (srcdata).

**Step 3**

For records with the id variable (I) greater than one, only the DRAW statement needs to be assigned. The code is implemented below.

```plaintext
data anno2;
  set anno1;
  by i;
  xlag=lag(x);
  if i eq 1 then output;
  if i gt 1 then do;
    if first.i then do;
      function='draw'; x=xlag; y=survival; output;
    end;
    function="symbol";
    if symbtyp eq 1 then do;
      style="special"; text="J"; size=1.5;
    end;
    else do;
      style="cartog"; text="Q"; size=0.55;
    end;
  x=time; y=survival; output;
  function='draw'; x=time; y=survival; output;
end;
run;
```
The code in step 3 can be integrated into step 2 but has been explained as a separate step for more clarity. The conditional statement involving FIRST is primarily used to reproduce a step-wise pattern typical of PROC LIFETEST.

**Step 4**
The annotate dataset is now ready. The legends are also created using the annotate facility. The G PLOT procedure can now be used to create the required plot.

```sas
proc gplot;
   axis1 label=(height=1.2 "Weeks of Treatment")
      order=(0 to 60 by 10) minor=none;
   axis2 split="," label=(angle=90 height=1.2 'Percentage of Subjects on Treatment') order=(0 to 100 by 10)
      offset=(0,0) ;
   symbol1 i=none v=none;
   plot y*x/anno=anno2
      haxis=axis1
      vaxis=axis2
   ;
run;
quit;
```

The final figure will resemble the figure below.
Case 2: Using PROC GPLOT

An alternative approach is by using two PLOT statements. The first PLOT statement creates a scatter plot with a unique symbol for subjects that discontinued (SYMBOL1) and another symbol for subjects at end of treatment (SYMBOL2). Note that SYMBOL1 and SYMBOL2 have interpolation equal to none. The second plot statement is then used to connect all the points. The structure of the source dataset (srcdata) should be the of the same structure as illustrated before.

```
proc gplot data =srcdata;
    plot survival*time=symbtyp;
    plot2 survival*time;
    symbol1 i=none c=black v=dot;
    symbol2 i=none c=black v=circle;
    symbol3 i=join c=black v=none;
run;
quit;
```

The final figure will resemble the figure below.
CONCLUSION

Two methodologies to create a single line plot with multiple symbols have been shown. A noteworthy difference between the two approaches is that a step-wise graph typical of PROC LIFETEST can be reproduced using the annotate feature whereas the second approach is more straightforward to connect multiple symbols on a single line plot.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Please contact the authors at:

Sanjiv Ramalingam
Octagon Research Solutions, Inc.
585 E Swedesforde Road
Wayne, PA 19087
Email: sramalingam@octagonresearch.com

Lelia McConnell
SAS Institute
SAS Campus Drive
Cary, NC 27513
Phone: (919) 677-8008
Fax: (919) 531-9449
E-mail: support@sas.com
Web: support.sas.com/ts