SESUG Paper 66-2021

Tips for Customizing Graphs Using Real Coronavirus Testing Data

Dennis J. Beal, Leidos, Oak Ridge, Tennessee

ABSTRACT

SAS® has many ways to generate high quality statistical graphics that include the older SAS/GRAPH® module to the latest SG plots using the Output Delivery System (ODS). Often times your clients may request very specific customizations to their plots that are not easily handled by simply changing an existing option within SAS. The annotation facility can be a powerful tool to customize your graphs. This paper shows examples of customized graphs using macros and the annotation facility on real publicly available coronavirus testing data. SAS code that generates the graphs is provided and discussed. This paper is for beginning or intermediate SAS users of Base SAS and SAS/GRAPH.

Key words: graphics, annotate, macros, SAS/GRAPH, trends

INTRODUCTION

Statistical graphics are one of the many ways SAS excels compared to its competitors. SAS has continued to improve its graphing capabilities from the older SAS/GRAPH module to the latest SG plots using ODS. Combining the macro language with the annotation facility are powerful tools to make specific customizations to plots. This paper shows how to use macros and annotation to customize bar charts and line graphs for trends in coronavirus testing data.

The daily data are county level data from the state of Tennessee beginning in March 2020. The data includes the cumulative number of positive and negative tests conducted by county within Tennessee and the cumulative number of new cases on a daily basis. The purpose of the charts is to show if new coronavirus cases are trending down over time so the next phase of resuming normal operations can be implemented for the U.S. Department of Energy installations in Oak Ridge, Tennessee. The data are publicly available for download online at *https://www.tn.gov/health/cedep/ncov.html*. The Excel file "TN_COVID19_CountyDaily.xlsx" shows the cumulative number of new cases, positive tests, negative tests and total coronavirus tests by day for each of the 95 counties in Tennessee. The graphs were produced using SAS/GRAPH in SAS 9.3 and the graphics editor on a Windows® 10 Enterprise platform.

CUMULATIVE PERCENT POSITIVE TREND TESTS

The client wanted to see both the cumulative number of coronavirus COVID-19 tests for each day in a bar chart and the cumulative percent of positive tests as a line chart on the same graph. The data for the 10 Tennessee counties Anderson, Blount, Campbell, Cumberland, Knox, Loudon, Morgan, Roane, Scott and Union were aggregated together since these counties surround the U.S. Department of Energy installations in Oak Ridge. In addition to the historical cumulative tests and percent of positive cases trends beginning March 31, the client wanted to see the previous 14 day trends on the same graph. This required having the number of tests shown in blue on the left vertical Y axis and the cumulative percent of positive tests shown in red on the right vertical Y axis. Linear regression trend lines for the historical cumulative percent of positive tests and the previous 14 days were also shown in red. The historical data and the previous 14 days data were to be separated by a thick vertical black line to distinguish them.

First a SAS macro is defined that counts the number of observations that are in a data set.

```
%macro obsnvars(ds);
 %global nobs;
 %let dsid = %sysfunc(open(&ds));
 %if &dsid %then %do;
   %let nobs =%sysfunc(attrn(&dsid,NOBS));
   %let rc = %sysfunc(close(&dsid));
```

```
     %end;
   %else
      %put Open for data set &ds failed - %sysfunc(sysmsg());
  %mend obsnvars;
```

The following SAS code extracts the data for the 10 counties and sums the number of positive, negative and total tests for each day.

```
 data co10;
  set cov.covid19_county_daily;
    where county in ('Knox' 'Anderson' 'Loudon' 'Roane' 'Morgan' 'Campbell'
'Scott' 'Union' 'Blount' 'Cumberland') and test_pos > 0 and d_collected <=
'02may2020'd;

      proc sort data=co10; by d_collected;
      proc summary data=co10;
        var test_pos test_neg test_tot;
        by d_collected;
        output out=co10sums sum=test_pos test_neg test_tot ;
      proc print data=co10sums; run;
```

This SAS code calculates the cumulative percent positive tests. The variable X is simply an integer for the number of days from March 31 through May 2 (1 through 33).

```
 data pos;
  set co10sums;
  pct = test_pos / test_tot * 100;
  X = _N_;
  keep d_collected pct test_pos test_neg test_tot x; run;
```

Next the linear regression trend line for the cumulative percent positive tests is calculated for the historical data March 31 through May 2. The number of days used in this trend line is stored in the macro variable N_DATES.

```
  proc reg data=pos;
    model pct = x;
      output out=regout p=predicted;
      run; quit;

  %obsnvars(pos); %let N_DATES = &nobs; run;
```

Now the linear regression trend line is calculated for the most recent 14 days. The variable X is updated to the next sequential 14 days.

```
 data pos14;
  set pos;
    where d_collected >= '19apr2020'd;
  x = _N_ + &N_DATES;
  rename pct=pct14;

  proc reg data=pos14;
    model pct14 = x;
      output out=regout14 p=predicted14;
      run; quit;
```

Since the X axis is crowded due to the large number of days shown on the plot, each day of the month does not need to be seen since the numbers would appear much too crowded. So a macro variable DAYTXT is created to store the day of the month that will be shown on the plot. Note that every other date will be shown to create some space between adjacent days.

```
data dates_txt;
 length DAYTXT $2;
 set pos pos14;
 day = day(d_collected);
 if x/2 ^= int(x/2) then daytxt = strip(put(day, 2.0));
   else daytxt = ' ';
 call symput('DAYTXT'||strip(_N_), strip(daytxt));
 run;
```

Since bars will be shown for the cumulative number of tests, both the minimum and maximum number of tests are needed per day in the data set to create the bars using the box plot interpolation option within PROC GPLOT. So the cumulative tests for each day are zeroed out and added back to the data set.

```
data neg;
 set pos pos14;
 TEST_TOT = 0;
 keep d_collected test_tot x; run;

data both;
 set regout regout14 neg;

 proc sort data=both; by x test_tot; run;
```

Next the FILENAME statement is used to name the plot. GOPTIONS is used to define the height, width and font for the plot.

```
filename plot 'Vbar chart ten counties percent positive tests all time.cgm';

goptions reset=all display gsfmode=replace noprompt gsfname=plot device=EMF
targetdevice=EMF rotate=portrait htext=0.95  cback=white ftext="Times New
Roman" horigin=0 in vorigin=0 in vsize=6.5 in hsize=6.9 in;
```

The thick vertical line separating the two areas of the graph is created using the annotation data set anno1. The two functions MOVE and DRAW are used to draw the line on the graph.

```
data anno1;
 length function color style $ 8 ;
 function='MOVE';
 size=2;
 l=1;
 xsys='2';
 ysys='2';
 hsys='4';
 x = &N_DATES + 0.5;
 y = 0;
 style=' ';
 color='black';
 when='a';
 position='6';
output;
 function='DRAW';
 size=2;
 l=1;
 xsys='2';
 ysys='2';
 hsys='4';
 x = &N_DATES + 0.5;
 y = 0;
```

```
 style=' ';
 color='black';
 when='a';
 position='6';
output;
 function='DRAW';
 size=2;
 l=1;
 xsys='2';
 ysys='2';
 hsys='4';
 x = &N_DATES + 0.5;
 y = 18000;
 style=' ';
 color='black';
 when='a';
 position='6';
output;
run;
```

The data set anno3 creates labels that are shown within the graph. The height, color, font and position are specified for each text label. The two annotation data sets are then combined together. Finally, the macro variable N_DATES_ALL is calculated as the total number of days shown on the X axis.

```
data anno3;
 length function color style $ 8 text $45;
 function='LABEL';
 size=1.5;
 xsys='1';
 ysys='1';
 hsys='4';
 x = 3;
 y = 96;
 style=' ';
 text="Cumulative Percent Positive COVID-19 Tests";
 color='black';
 when='a';
 position='6';
output;
 function='LABEL';
 size=1.5;
 xsys='1';
 ysys='1';
 hsys='4';
 x = 17;
 y = 90;
 style=' ';
 text="in 10 Tennessee Counties*";
 color='black';
 when='a';
 position='6';
output;
 function='LABEL';
 size=1.3;
 xsys='1';
 ysys='1';
 hsys='4';
 x = 25;
 y = 80;
 style=' ';
```

```
 text="March 31 - May 2";
 color='black';
 when='a';
 position='6';
output;
 function='LABEL';
 size=1.3;
 xsys='1';
 ysys='1';
 hsys='4';
 x = 77;
 y = 80;
 style=' ';
 text="14 Days";
 color='black';
 when='a';
 position='6';
output;
run;

data anno; set anno1 anno3; run;

%let N_DATES_ALL = %eval(&N_DATES + 14);
```

PROC GPLOT is used to produce the plot within the macro plotbar4 since the values of the dates are generated using a macro %do loop which is not allowed in open code. Using a macro allows this code to be used with very little modification as the number of days increases when the plot is updated with new data. Each bar is actually a box plot as specified in the i=box00tf in the symbol1 statement. The outline color co=blue is specified so the line for the median of each box is not shown since it is not necessary.

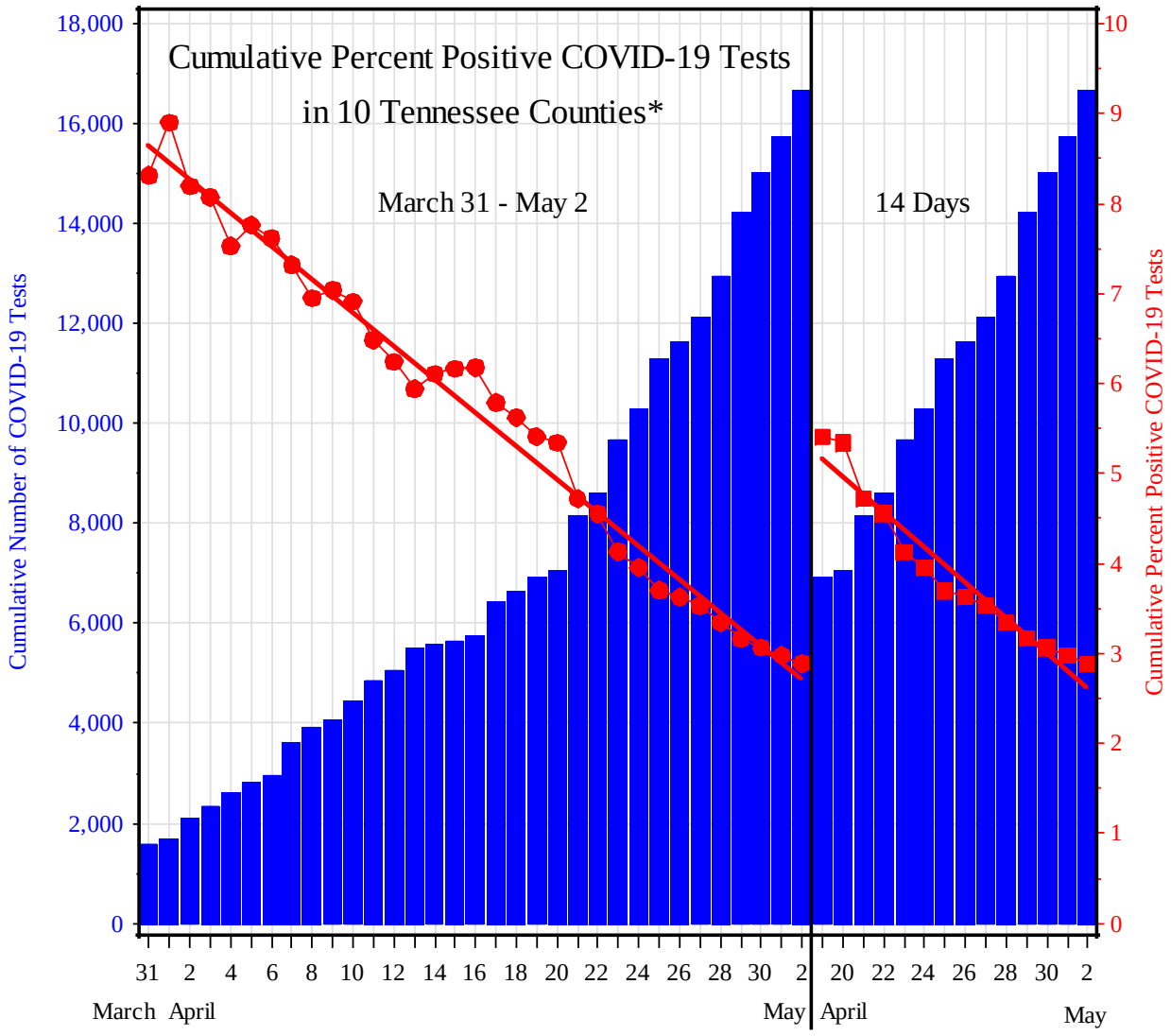
```
%macro plotbar4;

 proc gplot data=both annotate=anno;
   plot test_tot*x / vaxis=axis1 haxis=axis2 vm=1 hm=0 frame grid;
   plot2 pct*x predicted*x / overlay vaxis=axis3 vm=1;
   format test_tot comma7. pct pct14 5.0 x 2.0;
   label test_tot='00'x pct='00'x;
   axis1 width=2 order=(0 to 18000 by 2000) value=(c=blue h=1.05)
label=(a=90 r=0 h=1.05 j=c c=blue 'Cumulative Number of COVID-19 Tests')
c=black;
   axis2 w=2 length=5.8 in order=(1 to &N_DATES_ALL by 1) offset=(0.05 in)
label=none value=(  %do i = 1 %to &N_DATES_ALL;
                  t = &i. h = 1.05 j=c "&&DAYTXT&i." %end;   );
     axis3 w=2 order=(0 to 10 by 1) value=(c=red h=1.05) label=(a=90 r=0
h=1.05 j=c c=red 'Cumulative Percent Positive COVID-19 Tests') c=red ;
   footnote1 h=1.05 j=l 'March' j=c 'April' j=r 'May';
   footnote2 h=1.0 j=l '* The 10 counties include Anderson, Blount,
Campbell, Cumberland, Knox, Loudon, Morgan, Roane, Scott and Union.';
   symbol1 f=marker v=none bwidth=2 i=box00tf h=0.6 c=blue w=1 l=2 co=blue;
   symbol2 f=marker v='W' i=join h=0.8 c=red w=1 l=1;
   symbol3 f=marker v=none i=join h=0.8 c=red w=2 l=1;  * trend line;
   symbol4 f=marker v='U' i=join h=0.8 c=red w=1 l=1;
   symbol5 f=marker v=none i=join h=0.8 c=red w=2 l=1;  * 14 day trend line;
   run; quit;

%mend plotbar4;

%plotbar4
```

The final graph is shown in Figure 1 which shows the cumulative percent positive tests in the 10 counties decreasing over time from March 31 through May 2 for both the historical trend and the latest 14 days. The cumulative number of tests continues to increase. As testing increases the percent positive continues to decrease. Note the names of the months are shown only for the beginning of each month. The SAS graphics editor is used to extend the black vertical line between the historical trend and the latest 14 day trend and align the names of the months.



* The 10 counties include Anderson, Blount, Campbell, Cumberland, Knox, Loudon, Morgan, Roane, Scott and Union.

Figure 1. Time-trend bar chart of the cumulative number of COVID-19 tests and percent positive results in 10 Tennessee counties March 31 – May 2, 2020

CONCLUSION

The annotation facility within SAS/GRAPH is a powerful tool for customizing graphs. Real publicly available coronavirus testing data from 10 Tennessee counties from 2020 was used to produce a graph with bars for cumulative number of tests with separate regression lines showing trends for both long-term and the last 14 days trends of percent positive tests. The SAS code used to produce the graph is also shown.

CONTACT INFORMATION

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

Dennis J. Beal, Ph.D.
Senior Statistician/Risk Scientist
Leidos
301 Laboratory Road
P.O. Box 2502
Oak Ridge, Tennessee 37831
e-mail: beald@leidos.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.
Other brand and product names are registered trademarks or trademarks of their respective companies.