Most Effective Use of PROC GMAP: Solutions for Professional-grade Statistical Mapping

LeRoy Bessler, Miller Brewing Company
Steven J. Subichin, Miller Brewing Company

Abstract and Introduction

Most SAS software sites have large amounts of data which include geographic unit area designators (in the USA, state code is probably the commonest). Though this data can be reported in various tabular formats, a geographic effect (such as a proximity effect) will not easily be revealed without a statistical map. Also, for presentation, a visual image is more interesting than a listing. But the map can be supplemented with detail look-up data. The detail can be a complete list in key sequence, a ranking report based on response level, or just a Top NN List of the NN most significant (i.e., highest response) geographic areas. A Top 10 table is concise enough to use as a slide. Better yet is to infill all the detail on the respective geographic areas of the map.

This paper provides complete program code to build well-designed SASGRAPH statistical maps. It presents two innovations:
(a) Automatic Rationale-based Response Range Assignment (ARbRRA)--the solution to what we call The Four Color Map Problem; and (b) "blinking" for map annotation--a feature long needed in SASGRAPH. ARbRRA eliminates the need to accept default assignment of response ranges or to inspect a sorted list of all responses to "hand pick" ranges, which may be arbitrary anyhow. Blanking assures that annotation is always readable, even when area fills are grey shades or dark colors. This paper also includes valuable adjustments to the vendor-provided USA state-center coordinates (the SASGRAPH data set MAPS.USCENTER). These improved centers enable one to infill more annotation without bumping into or running over the geographic area boundary.

The sample program is intended to be reusable and ready adapted--just point it at your data, customize a few details, and you're done. Though the example uses the USA map, the principles apply to county or international maps as well. Program details are commented in the code, not explained below.

Device Driver Details

The GOPTIONS statement (not shown here) specifies the device driver, device-related parameters, physical size of the graph display area, etc. An AFP-compatible, black-and-white laser printer, capable of printing grey shades, was used. In the programs, you will see familiar color names. Software translates them into grey shades. Upon request, we can furnish details of GOPTIONS, GDDM nickname, MVS JCL (Job Control Language), etc., and comments as to what works well, and what doesn't.

Exhibit 2: SAS/GRAFPH Defaults Unacceptable

Exhibit 2 is an unacceptable map of a real data set, using PATTERN statements, a COUTLINE, and otherwise a default invocation of PROC GMAP.

The adverse effect of the default ranges chosen (their midpoints are shown in the map legend) by the PROC GMAP default algorithm is due to two values, viz., 736 and 447; all other values are below 179.

Four Color Map Problem Solved

Even without an extreme result as Exhibit 2, it is better to make a deliberate choice of ranges, based on a rationale. In principle, that requires you to have knowledge of the data distribution. But first doing, e.g., a PROC SORT and PROC PRINT so data can be inspected is inconvenient, and can result in an arbitrary decision anyhow.

In a prior paper (Bessler, L.R., "Effective and Efficient Information Delivery for Executive Management", in *Proceedings of the Seventeenth Annual SAS Users Group International Conference*, SAS Institute Inc. (Cary, N.C.), 1989), it was emphasized that, typically, a small subset of the observations account for a large majority, or even almost all, of the total response. A Top 10 or Top NN Report (i.e., some one-page data report) usually suffices, often accounting for 80% or 99% of the total response. In this data set, the Top 10 states account for 66.4% of attendance. (Over the last five years, 63.1%.) With 50 states and DC, the Top 10 states are by definition above the 90th percentile.

Regardless of the specific choices, it is natural to break up the total range based on percentiles. One can use, e.g., the 25th percentile, median (the 50th percentile), and 80th percentile. The resulting four ranges may be characterized as Very Low, Below Median, Above Median, and Very High. One might use more ranges if there were many more geographic areas to differentiate (e.g., all counties in the USA), but would always include the median as a break point. (Other mapmakers might use mean and standard deviation to develop ranges.)

Percentile-based ranges create a talking point for the map, as opposed to relying on software defaults. The solution uses what has been called "Software Intelligence" (SI). For other examples of SI, see Bessler's paper "Software Intelligence: Applications That Customize Themselves", in *Proceedings of the Eighteenth Annual SAS Users Group International Conference*, SAS Institute Inc. (Cary, N.C.), 1993.

The program in Exhibit 4 shows how to let software automatically solve the Four Color Map Problem. Give the program data, it sets four ranges automatically. You need only pick four colors, and supply title text. This technique can be called Automatic Rationale-based Response Range Assignment (ARbRRA). ARbRRA is implemented with the PROC UNIVARIATE step, the following DATA _NULL_ step, the PROC FORMAT step, and the FORMAT statement in the PROC GMAP step.

If desired, the legend could display the actual numeric ranges, instead of the text strings used--modify the PROC FORMAT VALUE statement by replacing the right-hand side of each assignment with the left-hand side enclosed in double quotes.

Before closing discussion of response range assignment, we must acknowledge there are many mapping applications where ARbRRA is not needed. One may wish the map to show whether the change in a measurement over time within each unit geographic area has been positive, zero, or negative. Or to show whether the response is above or below some critical value. Such examples fall into the category of situations where the map communication objective directly defines the response ranges. Whether or not those situations are exceptional, rather than the norm, ARbRRA is a powerful concept/tool, and a productivity aid, for geographic data exploration and presentation.

Annotation Without Blanking Unacceptable

If the map in Exhibit 3 were done without the white boxes (blanking) behind the annotation, the text and data in many states would be nearly or actually unreadable. This is due to the use of grey-shade area fill. Blanking is still missing from SASGRAPH. Behind map annotation we need a rectangle of white space within the solid color (or grey shade) in the geographic area. The only area fill that might be acceptable without blanking would be very light shades (e.g., light pink, light yellow, light green, light blue, etc.). But not all devices can render light shades of color.

Furthermore, not all publications will accept color illustrations. Finally, a very interesting, informative map is likely to be one from which someone will wish to make copies. Though color copiers are becoming increasingly available, they still are not as widespread, cheap, and fast as black-and-white copiers. Therefore, annotated grey-shade maps are usually the most practical solution.
Exhibit 3: Having It All - Maximal, Optimized Annotation, and Blanking

This map (Exhibit 3, done with the program in Exhibit 4) tries to do just about anything one could think of doing with a statistical map.

Of course, if one also wants, say, Percent of Whole, a four-line of annotation could be done with a more complicated alternative to the ANNOVALU DATA step, or Percent of Whole could replace State abbreviation in the demonstrated three-line annotation.

Four-line annotation would require use of a smaller font.


Annotation with response value has obvious benefit. And since not everyone knows each state name just by shape and relative location on the map, it is "nice to have" that identification. The provision of Rank based on response value reflects Bessler's graphic design postulate that most people prefer to be able to quickly identify what's most important.

How to provide a "must have", i.e., the white boxes, was reported by Subichin in "Enhanced Useability for Annotation on SASGRAPH Maps", in W/ASAS Proceedings, Volume 5, June issue, W/ASAS Inc. (South Milwaukee, Wis.), 1993.

The program in Exhibit 4 is an extension of one presented in Subichin's paper (op. cit.). Included here, however, are recommended adjustments to the state-center coordinates from MAPS.USCENTER.

Also, the resulting map here incorporates a star to highlight the conference location. (Instead, the city marker could have been the city name, or "SUGI 14".) See DATA steps CITYSTAR and ANNODATA for how to do this. In Exhibit 1 is a program to list the GMAP US cities. For a city not listed, pick the nearest city that is listed, and adjust coordinates. Annotating cities was first demonstrated in the manual for Version 5 of SASGRAPH. It is included here for completeness—to provide the reader a single, reusable, adaptable program that includes all the techniques likely to be needed for professional-grade statistical mapping.

Of course, one should realize that DATA step ANNODATA could be simplified. E.g., for two-line annotation, use POSITION values for 1/2 cell above and 1/2 cell below, and consider making the boxes smaller.

Notes

SAS and SASGRAPH are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. * denotes USA registration.

All SAS code in this paper was tested, but can only be presented "as is". Any adopted by you must be tested by you.

Author Contact Information

LeRoy Bessler
P.O. Box 17725
Fox Point, WI 53217-0725, USA
Telephone: 414-351-6748

Steven J. Subichin
P.O. Box 482
Milwaukee, WI 53201-0482, USA
Telephone: 414-931-2271

DATA US CITIES;
SET MAPS.USCITY;
STATENAME + FNAME(STATE);
IF CAPITAL = "N" THEN CAPITAL = ",
RUN;
PROC PRINT DATA=USCITIES UNIFORM LABEL;
PANEL STATE;
BY STATE STATENAME;
VAR CITY X CAPITAL POP;
LABEL STATE='ST';
LABEL STATENAME='STATE';
LABEL CAPITAL='CAP';
TITLE 'SAS/GRA PH MAPS.USCITY FILES';
RUN;

Exhibit 1: Program to List GMAP US Cities

Exhibit 2.

SUGI 14 Attendance, By State

Defaults Unacceptable

VALUE | 120 | 240 | 360 | 480 | 600 | 720

882
Exhibit 3. Map of San Francisco SUGI 14 Attendance – Showing State, Count, & Rank

Legend:
- Very Low
- Below Median
- Above Median
- Very High

Map of the United States showing the attendance distribution by state, with colors indicating the level of attendance.

States are color-coded to indicate their attendance level, with a key at the bottom of the map providing the legend:
- Very Low
- Below Median
- Above Median
- Very High

The map includes major state abbreviations and numbers indicating the count and rank for each state.
DATA RESPONSE; /* input data
INFILE ddbname;
INPUT STATEA88 $ VALUE;
STATE = STFIPS(STATEA88); /* convert two alphabetic characters to numeric ID */
RUN;
PROC UNIVARIATE DATA=RESPONSE NOPRINT PCTLDDEF=2; /* get min, max, percentiles */
OUTPUT OUT=STATS MIN=MIN MAX=MAX PCTLPTS=PCTL PCTLPRE=PCTL;
RUN;
DATA _NULL_; /* boundaries for response ranges as global variables */
SET STATS;
GLOBAL MIN B1 B2 B3 B4 B5 B6 MAX;
CALL SYMPUT('MIN',TRIM(LEFT(MIN 100)));
CALL SYMPUT('B1',TRIM(LEFT(PCTL20 100)));
CALL SYMPUT('B2',TRIM(LEFT(PCTL20 + 1) ));
CALL SYMPUT('B3',TRIM(LEFT(PCTL50 100)));
CALL SYMPUT('B4',TRIM(LEFT(PCTL50 + 1) ));
CALL SYMPUT('B5',TRIM(LEFT(PCTL80 100)));
CALL SYMPUT('B6',TRIM(LEFT(PCTL80 + 1) ));
CALL SYMPUT('MAX',TRIM(LEFT(MAX 100)));
RUN;
PROC FORMAT; /* build format to be used by PROC GMAP for ranges & legend text */
VALUE VALUEFMT &MIN - &B1 = 'Very Low'
&B2 - &B3 = 'Below Median'
&B4 - &B5 = 'Above Median'
&B6 - &MAX = 'Very High';
RUN;
PROC SORT DATA=RESPONSE OUT=RANKED; BY DESCENDING VALUE;
RUN;
DATA RANK;
SET RANKED;
RANK = _N_;
RUN;
PROC SORT DATA=RANK OUT=RANK; BY STATE;
RUN;
PROC SORT DATA=RESPONSE OUT=TOANNO; BY STATE;
RUN;
PROC SORT DATA=MAPS.USCENTER OUT=CENTERS; /* get coords. of the state centers */
BY STATE DESCENDING OCEAN; /* if OCEAN = 'V', must precede OCEAN & 'N' */
RUN;
DATA CHANGES; /* changes empirically selected by LeRoy Bessler */
INFILE CARDS; /* to be applied to maps.uscenter coords as of SAS Release 6.08 */
INPUT STATE OCEAN $ CHANGEX CHANGEY;
CARDS;
01 N -0.0037 +0.0000 AL
02 N +0.0050 +0.0000 AK
04 N -0.0050 +0.0000 AZ
05 N +0.0037 +0.0000 AR
08 N +0.0000 +0.0000 CO
09 Y -0.0140 -0.0050 CT
10 Y -0.0025 -0.0060 DE
11 N -0.0100 -0.0200 DC
12 N +0.0040 -0.0060 FL
15 N +0.0050 +0.0250 HI
16 N -0.0050 +0.0000 ID
17 N -0.0015 +0.0050 IL
18 N -0.0013 +0.0000 IN
19 N +0.0000 -0.0000 IA
20 N +0.0000 -0.0000 KS
21 N +0.0000 +0.0035 KY
22 N -0.0015 +0.0000 LA
23 N -0.0025 +0.0060 ME
24 Y +0.0250 -0.0200 MD
25 Y -0.0050 +0.0015 MA
26 N +0.0010 -0.0030 MI
28 N -0.0013 +0.0000 MS
29 N -0.0037 +0.0000 MO
30 N +0.0070 -0.0005 MT
33 Y -0.0010 +0.0150 NH
34 Y +0.0275 -0.0200 NJ
36 N +0.0025 +0.0015 NY
37 N +0.0000 +0.0035 NC
38 N +0.0000 +0.0020 ND
RUN;
PROC SORT DATA=CHANGES OUT=CHANGES; BY STATE DESCENDING OCEAN; /* SORT */
RUN;
DATA CENTERS; /* shift of coordinates for states needing it */
MERGE CENTERS CHANGES(IN=CHANGE); BY STATE DESCENDING OCEAN;
IF CHANGE THEN DO;
  X = X + CHANGEX;
  Y = Y + CHANGEY;
END;
RUN;
DATA MERGED; /* merge rank & response value with coords of the state center */
MERGE RANK TOANNO(IN=INRESP) CENTERS; BY STATE;
IF INRESP;
RUN;
DATA ANNO DATA; /* put all text at center of each state, except coastal states */
SET MERGED;
RETAIN FLAG;
IF _N_ = 1 THEN FLAG = 0; /* this annotation record is a label */
FUNCTION = 'LABEL'; TEXT = FIPSTATE(STATE); IF OCEAN = 'N' OR FIPSTATE(STATE) = 'VT' THEN POSITION = '8'; ELSE POSITION = '9'; OUTPUT;
FUNCTION = 'DRAW'; SIZE = 1.00; /* The block of statements at left */
END;
IF OCEAN = 'Y' THEN DO;
FUNCTION = 'MOVE'; IF FIPSTATE(STATE) NE 'VT' THEN X = X - .005; ELSE Y = Y - .0020;
END;
RUN;
OUTPUT; /* */ FLAG = 1; /* */ END; /* */

RUN; DATA CITYSTAR; /* city marker annotation */ SET MAPS.USCITY; IF CITY = 'San Francisco'; /* get coordinates for San Francisco */ XSYS = '2'; /* do not change */ YSYS = '2'; /* do not change */ WHEN = 'A'; /* do not change */ POSITION = '5'; /* adjust, if desired, for limited repositioning */ SIZE = 1.00; /* adjust size, if needed */ STYLE = 'MARKER'; /* font comprised of markers (see also MARKERE) */ COLOR = 'BLACK'; /* change color, if desired */ TEXT = 'V'; /* this marker is a star */ X = X - 0.010; /* for this city, shift left for better visibility */

RUN; DATA ANNODATA; /* append city marker to other annotation data */ SET ANNODATA CITYSTAR;

RUN; PROC FREQ DATA=MAPS.US; /* count the points on each segment of the state */ TABLES STATE.SEGMENT / OUT=COUNTS NOPRINT;

RUN; PROC SORT DATA=COUNTS; /* order state & segment ID by decreasing point-counts */ BY STATE DESCENDING COUNT;

RUN; DATA BIGSEG; /* output 1 record per state with segment ID of biggest segment */ SET COUNTS; BY STATE DESCENDING COUNT;

RUN; DATA DELOCEAN; /* won't need boxes for Hawaii & in ocean */ SET CENTERS; IF FIPSTATE(STATE) = 'HI' THEN DELETE;

RUN; DATA INNER; /* keep state, segment ID, & coordinates only for states with box */ MERGE BIGSEG DELOCEAN(IN=NOOCEAN); BY STATE;

DATA BOXES(KEEP=STATE SEGMENT X Y); /* make boxes for inner states */ SET INNER; RX=X; /* save X & Y coordinates */ RV=Y;

RUN; /* different application may need boxes of different size & shape */ /* if so, BOXES DATA step still needs X=, & Y=, and to close the polygon */ DATA US_WBOX; /* interleave map and box coordinates */ SET MAPS.US BOXES; BY STATE SEGMENT;

RUN; PATTERN1 V=MSOLID C=GRAY;

PATTERN2 V=MSOLID C=BLUE;

PATTERN3 V=MSOLID C=GREEN;

PATTERN4 V=MSOLID C=DAB;

LEGEND1 LABEL=NONE VALUE=(F=SIMPLEX) SHAPE=BAR(3.4,0.8) DOWN=l;

TITLE.; PROC GMAP MAP=US_WBOX DATA=RESPONSE ANNO=ANNODATA ALL; /* map with blanking */ /* using MAP=MAPS.US will turn off blanking */ FORMAT VALUE VALUEFHT.; ID STATE; CHORO VALUE / COUTLINE=BLACK DISCRETE LEGEND=LEGEND1;

RUN; QUIT;

Exhibit 4: Program for Exhibit 3

886