Paper 192-26

At Home and Abroad with PROC GMAP: A Global Atlas of InfoGeographic Design for Effective Visual Communication

LeRoy Bessler, Bessler Consulting & Research, bessler@execpc.com

Abstract and Introduction

Maps of the spatial distribution of geo-coded data are useful in business, government, not-for-profits, education, and research. This paper and poster are based on over a decade of experience solving a wide variety of problems in creating maps with SAS/GRAPH® to explore the locational significance of the contents of the data warehouse. This continuing mission to maximize "The Power to Show" in PROC GMAP has confronted and surmounted diverse obstacles in communication, classification, and annotation with maps—which are probably the information-richest and most interesting data visualization images you can create. This paper recommends principles for InfoGeographic design. The companion poster provides my design examples from around the world, as well as the USA and individual states.

The associated poster will present twenty or more example maps, each in 8-1/2 inch by 11 inch full-page format. Since the Proceedings page-count limit (ten) cannot accommodate *all* of the maps, I decided to display almost all of them only on the poster, and to make all available upon request as an emailed zip file. This paper contains *only one* illustration.

This paper emphasizes design principles and discussion of problems and solutions—solutions as depicted in the poster example maps—not coding details. If you have questions, comments, suggestions, see me at SUGI, or refer to the Author Information section for telephone number, etc.

Maps Have "The Power to Show"

"Of all the contrivances hitherto devised for the benefit of geography, the map is the most effective. In the extent and variety of its resources, in rapidity of utterance, in the copiousness and completeness of the information it communicates, in precision, conciseness, perspicuity, in the hold it has upon the memory, in vividness of imagery and power of expression, in convenience of reference, in portability, in the happy combination of so many and such useful qualities, a map has no rival. Everything we say or do has reference to place, and wherever place is concerned a map deserves welcome. There is scarcely one department of knowledge, physical or moral, beyond the sphere of its usefulness; to geography it is indispensable. Modern technology has advanced the process of making maps considerably, and a map still has no rival in its usefulness."

G.B. Geenough
Presidential Address to The Royal Society
London, 1840

Design for Communication and Retention

Information delivery should inform and deliver. The time and attention of a map viewer, and the time and effort of a map creator, are precious resources. One ought strive to provide maps that interpreted quickly and easily.

Design to inform and influence, not to impress. A powerful image sticks in the memory, and is easy to "replay". But special effects are for movies. A map is already inherently a complex image. Good design and interesting data can stand on their own. Communication, not decoration, is the objective.

Just Say "No" to 3D

Use the informative two-dimensional CHOROPLETH map. The 3D alternatives—SURFACE, PRISM, and BLOCK maps—are picturesque, but impractical. SURFACE maps are too vague for real communication. PRISM and BLOCK maps have responses for some "high" areas hiding those for "low".

Use Solid Area-Fill

Use V=MSOLID or V=MEMPTY for PATTERN statements. Parallel lines or cross-hatching yield an unpleasant image, and can confuse boundary and area-fill. For dot or bubble maps, use of area-fills to encode different levels of response is unnecessary.

Text Is For Reading

Letters or numbers must be readable. On light background, always use black. If not essential, avoid decimals in numbers.

Use Color to Communicate, Not Decorate

If there are no response levels/categories (e.g., a bubble or dot map), black and white are enough. For few levels or categories, gray shades may suffice. But, for many levels or categories, color is essential.

Prof. Jay Neitz (of the Eye Institute of the Medical College of Wisconsin): in the USA over 8 percent of males have some form of color blindness; due to genetic differences, only about one-half percent of females. The commonest form of color blindness here cannot distinguish red and green, which are widely used in combination.

In a color-saturated environment, well-designed use of black, white, and gray shades can be distinctive, "impactful", and memorable. Such work involves equipment that is faster, more reliable, easier-to-use, and less expensive (for replenishable supplies at least), and requires no agonizing over color choice. It is more copyable (there are more, cheaper, faster black-and-white copiers)—remember: *Good Maps Get Copied.*

Be Careful with the "Safe" Color Gray

Gray shades are often difficult to use.

The human eye cannot reliably distinguish more than five to seven shades of gray (or of any single hue). I have seen conflicting numbers for this limit. Your own experience will find *your* limit.

Gray shades sometimes do not copy well.

Design Problems and Solutions

The SAS/GRAPH map is a chart type particularly vulnerable to detrimental defaults. Without specifying response ranges, you get results—often sub-optimal, if not unacceptable—based on a default algorithm. An example of this problem is shown in Figure 5, the only illustration included here from the poster.

Even if defaults are tolerable, it is better to make a deliberate choice of ranges, based on a rationale. In principle, that requires knowledge of the data distribution. Before creating a map, one can inspect the data. However, that is inconvenient, time-consuming, and laborious, and can result in an arbitrary decision anyhow. Rationale-based ranges create a talking point for the map. Software defaults or arbitrary breakpoints cannot provide concept-based defendability.

I first got involved with the classification problem in 1992. My blatantly disappointing result when taking PROC GMAP defaults while working on a trade area analysis problem has been reported and displayed often, in a series of prior papers.

I have developed a series of Automated Rationale-Based Response Range Assignment macros, to optimally and flexibly handle the cases of: four response ranges; five signed response ranges; N cluster-based response ranges; and highlighting response range ("Top Spots"). See the Bibliography.

One can assign response ranges based on percentiles. I particularly like to use four ranges, with breaks at percentiles 20, 50, and 80. (Since this solution uses four response ranges, and therefore four colors, I like to call it—with tongue in cheek and with unabashed grandiosity—"solving The Four Color Map Problem". There is a famous classic problem in mathematics which requires a proof that four colors are sufficient to color any map so that no two adjacent areas are the same color.)

There is nothing sacred about percentiles 20 and 80, and the macro I developed accepts

other user preferences. You might prefer percentiles 10 and 90. (One could also develop a macro to use mean and some factor(s) times standard deviation to provide four or six response ranges, if desired. The important thing is to use *some* rationale, not defaults or arbitrary ranges.)

Why *four* ranges? Well I admit to a fascination with the charm and the power of the median. That automatically begets an even number of ranges. Six might sometimes be better than four, and certainly would be easy to automate.

When a response measure is potentially influenced by the magnitude of the unit area (e.g., populations in a non-hostile environment), it is important to consider normalizing the data over land area. That eliminates any undue emphasis that might be due merely to size of the territory. Though you may wish to map the distribution of population by magnitude, you might also want to map the density. Obviously, there are many other applications where such normalization is useful.

Cluster analysis is another tool I have used to assign response range boundaries. Cluster Analysis is the ultimate in fairness: you decide how many ranges you want, and let a mathematical algorithm find that many most-clustered groupings of responses. One can object that the number of clusters is arbitrary, but at least the boundaries are unequivocally "neutral" and immune to human intervention or judgment. Sometimes you will get some (or what might be deemed as too many) single-value clusters. Now, that might be judged as a failure and not a good way to present the data, but another interpreter might view that as eminently meaningful and helpful. A limitation on the desirability of cluster analysis may be that it can tempt one to call for so many clusters that, when using only shades of gray or one hue, area fills are impossible to distinguish.

The easiest way to "Show Them Where It's At" is to "Show Them What's Important". Even if you want or need to show them the geo-based data with one or more other maps, with a different analytical or presentation purpose, a readily assimilated and very memorable image is a map which highlights the Top N areas, i.e., the N areas with the highest response.

Typically, a small subset of observations account for most, or almost all, of the total response. A Top 10 or Top N Report (i.e., a concise report) usually suffices, often accounting for 80% to 90% or more of the total response. E.g., in Europe, the Top 10 areas (out of 48) account for 74.6% of the total population of Europe.

The macro I developed to create a Top N map can do much more than just put more conspicuous area-fill on the high-response areas. It supports three options: (a) highlight the Top N areas; (b) highlight all areas with a response at or over a minimum; and (c) highlight just enough of the highest response areas to account for a specified percent of the total response. If the response variable being mapped is not additive, Option (c) is inapplicable. E.g., population is additive, but population density is non-additive. For Options (a) and (c), the macro user can also specify a minimum. I.e., highlight the Top areas, but only those at or over the minimum. My SUGI 25 paper focused on the TOPSPOTS macro. (See the Bibliography.)

The situation of signed data offers an opportunity for a straightforward and intuitive rationale. My solution was really what might be called a two-sided TOPSPOTS rationale. It's a five-color map, which uses white (or V=MEMPTY area-fill) for zero values, and dark and light shades of any two contrasting colors for the positive and negative values. For example, the Top 10 positive-valued areas could be dark, bright, or normal blue, the remaining positives light blue, the Bottom 10 negative-valued areas could be dark, bright, or normal red, and the remaining negatives light red. There's nothing magic about blue and red. The important thing is to avoid green and red, as noted above in the discussion of color blindness.

My range assignment macro for five signed response ranges—which like all my others is built with what I like to call "Software Intelligence"—is capable of adapting to whatever is the situation of the data (e.g., no

positives, no negatives, not more than N positives, not more than N negatives, no zeros, and any combination of those, where N is the number of "top areas" that the user chooses to highlight). I don't recall the exact number, but I believe there are more than thirty possible test cases for which the associated macro had to be validated.

A chart can both show relative magnitude, and supply detail. Presentations or reports that deliver both image (impact) and numbers (precision) are memorable, quickly and easily comprehended, and both influencing and reliable for decisions.

How to best annotate the geographic unit areas of a map with a variety of numeric and text items—in effect making it a "spatial table"—has been previously reported for North American audiences by me most recently in the "Map Smart" paper (see Bibliography).

It should be noted that the "founding paper" on effective annotation of maps was by S. J. Subichin: "Enhanced Useability for Annotation on SAS/GRAPH Maps", in WISAS Proceedings, Volume 5, June Issue, WISAS Inc. (Fox Point, Wis., USA), 1993. This work showed how to create boxes of white background for text inside solid-filled areas.

Essential to annotation of a map is a data set of map coordinates for the centers of the geographic unit areas. Though SAS® Institute supplies centroids for MAPS.US, those coordinates did require some adjustment, first by Subichin, and further by me. However, there is no such data set for other countries or continents. When preparing maps for the SEUGI 18 Conference in Dublin, I developed such coordinates for the counties of the Republic of Ireland.

It remains to be seen for what other maps I will be able to develop centroids between paper submission deadline (22 January 2001) and SUGI 26 Conference poster exhibition time. I certainly expect to develop the centroids for the counties of one or more states in the USA, among them California (the SUGI host state).

This is not a trivial task. E.g., last year (2000), I deferred trying to develop usable centroids for the countries of Europe. In the USA, the very small states and the District of Columbia (the seat of federal government) are fortunately located near the Atlantic Ocean. For them the solution is to locate the annotation in the ocean and draw arrows back to the pertinent state (or D.C.). However, Europe includes, among other problems, small areas imbedded in the continental land mass (e.g., Andorra, Liechtenstein, and Luxembourg). Coastal or near-coastal areas such as Monaco and San Marino can probably be annotated from out in the Mediterranean or Adriatic Sea, respectively.

When annotation is infeasible, the solution is either a companion table to show the details for small areas and large, or a dynamically drillable interactive map. Drillable maps are outside the scope of this paper, and do not provide a retainable hardcopy document—the mouse-triggered pop-up text is not printable. When very compact annotation is feasible, a two-character key can be used to tie areas to entries in a companion table.

Bubble maps and dot maps do not require any response range assignment.

Bubble maps indicate the magnitude of a response at a location by the area of a circle centered at the centroid of that geographic unit area. One can create more than a simple bubble map, by annotating the unit areas with area name and area response. The bubbles provide an easy means to compare magnitudes with a visual indicator, and the details are available. An annotated bubble map has the precision decision-support power of annotation, and enables *the ultimate in visual comparison of magnitude of response* because the scale is absolute and continuous, rather than clumping responses into a few ranges.

I first learned how to create bubble maps and dot maps from Gary Plazyk. See Plazyk, G. F., "Using the Annotate Facility with Maps: A Tutorial", in Proceedings of MWSUG '91, MidWest SAS Users Group (Fox Point, Wis., USA), 1991.

Dot maps can show the location of every instance of a unit response.

They are feasible and appropriate when you have either precise coordinates of specific locations, or centroid coordinates for very small areas, such as postal code or census district. Also, they can best be used to present the geographic distribution of "binary information", i.e., whether or not the measured condition is present or absent at a point. Positive is represented by the dot, Negative by absence of the dot.

Bibliography (Partial)

My InfoGeographic tools and techniques have been documented elsewhere.

See, e.g., my papers "Map Smart: Design and Build Effective InfoGeographics Using PROC GMAP and Software Intelligence", in *Proceedings of the Twenty-Second Annual SAS Users Group International Conference*, SAS Institute Inc. (Cary, N.C., USA), 1997, and "Show Them Where It's At: Data Mine the Earth's Surface for Locational Significance of What's Hidden in Your Data Warehouse", in *Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference*, SAS Institute Inc. (Cary, N.C., USA), 2000, as well as a paper with the same title ("Show Them Where It's At . . .") but with much broader content (presented at the 2000 SAS European Users Group International Conference in Dublin), in *2000 SEUGI 18 Proceedings*, SAS Institute Inc. (Heidelberg, Germany), 2000.

Author Information and Related Work

LeRoy Bessler, Ph.D.
Bessler Consulting & Research
PO Box 96
Milwaukee, WI 53201-0096, USA
bessler@execpc.com
1-414-351-6748

Since 1987, at various venues in the USA and overseas, Dr. LeRoy Bessler has been sharing his ideas and experiences in the areas of macro-based Software-Intelligent Application Development, graphic design, visual communication, color, information visualization, and InfoGeographics. The content for paper and poster are extracted from a book in preparation, which will offer guidelines, examples, and macros for best practices with SAS/GRAPH software.

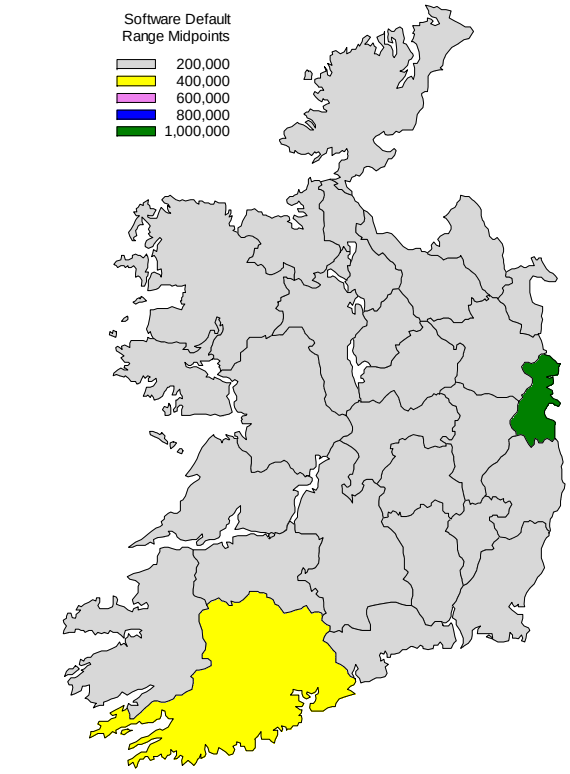
At SUGI 26, he is also presenting "The Power to Show: Using Information and Image to Reveal and Persuade".

More information related to these SUGI presentations may be available on request.

Trademarks

Only Map Shown Here from the Poster Atlas—why you must get beyond defaults:

Figure 5: Default Ranges for Population Map of Republic of Ireland--Clearly Unacceptable. Highest and second lowest ranges are each allocated to 1 county, while lowest range (0-300,000) is allocated to remaining 24 counties. "Midpoint" means, e.g., that 400,000 is for range 300,001-500,000.



Complete Range: 25,032 (County Leitrim) to 1,056,666 (County Dublin)

Design + SAS/GRAPH = The Power to Show