Paper 163-25

Show Them Where It's At: Data Mine the Earth's Surface for Locational Significance of What's Hidden in Your Data Warehouse

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

Abstract and Introduction

Much of what's in your data warehouse has geographic keys, such as codes for country, state or province, county, or sales or administrative district, as well as postal codes. You can indeed make proximity- or location-based inferences or decisions, or discover and present the locational significance of what's hidden in your data warehouse, with tools you already have. Probably the most widely available SAS® software product in the world, after Base SAS, is SAS/GRAPH®. With PROC GMAP, its map data sets, and auxiliary PROCs and data, you can crea te what I like to call "InfoGeographics", at no added cost.

This paper introduces you to PROC GMAP, if you want to create maps from scratch. To help you get good results more quickly, the paper discusses and demonstrates the author's TOPSPOTS macro, for ready-to-use maximal function and flexibility. TOPSPOTS can work anywhere in the world with any measurement data (which contains, or can be tied to, the SAS geographic keys). Point the macro at your data, identify the geographic area to be mapped, specify your device driver(s), and you are ready to go.

The paper also offers map design guidelines for effective, efficient exploration and presentation of geographic -keyed data. The presentation will include other map examples, for which limits on page-count and time forbid how-to details.

The TOPSPOTS macro was designed and built—using macro-based Software Intelligence—to be widely applicable as is. If needed, a macro-language-fluent SAS user can adapt the macro internals to unique or special situations outside its current scope. The TOPSPOTS macro is too long to publish here, but is available from the author.

Other SAS/GRAPH-based InfoGeographic tools and techniques have been documented elsewhere. See Bessler, LeR., "Map Smart: Design and Build Effective InfoGeographics Using PROC GMAP and Software Intelligence", in *Proceedings of the Twenty-Second Annual SAS Users Group Conference*, SAS Institute Inc. (Cary, N.C.), 1997. Several years ago, the author got started with PROC GMAP with the assistance of Steve Subichin and Gary Plazyk. Some of their work is cited below.

Design First

Information delivery should inform and deliver. The time and attention of a map viewer, and the time of a map creator, are precious resources. One ought strive to provide powerful presentation maps, and map-based or -including, reports that are digestible at a glance.

Technology is no substitute for design. This paper provides design advice, based on experience, observation, and research, to make technology the servant of visual communication. The TOPSPOTS macro was designed for communication.

Design for Communication

Design to inform and influence, not to impress.

Design for Retention

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

Define Your Style— Consistency Accelerates Communication

If you employ a consistent style, your viewers need not "recalibrate" their perceptual and interpretive faculties from map to map, or report to report. And, spared option over-choice and iterative experimentation, you as preparer are more productive.

Just Say "No" to the Designer Drug 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" states hiding those for "low".

Make It "Easy on the Eyes" With Solid Area-Fill

Use of parallel lines or cross-hatching yields an unpleasant image, and can even confuse boundary and area-fill. For some InfoGeographic applications, use of area-fills to encode different levels of response is inapplicable. For how to create dot maps or bubble maps, see Plazyk, G. F., "Using the Annotate Facility with Maps: A Tutorial", in Proceedings of MWSUG '91, MidWest SAS Users Group (Fox Point, Wis.), 1991.

Text Is Essential: Handle It with Care

- **Letters or numbers must be readable**
- Always use black, the most readable color
- If not essential, suppress decimals in numbers
- **Whenever possible, make the title your headline: the main message of your map**

Use Color to Communicate, Not to Decorate

- No response levels/categories (e.g., a dot or bubble map): black & white
- Few levels or categories: gray shades maybe
- Many levels or categories: color necessary

Prof. Jay Neitz (of the Eye Institute of the Medical College of Wisconsin): over 8 percent of American males have some form of color blindness; due to genetic differences, only about one-half percent of American females. **Commonest form of color blindness cannot distinguish red from green.**

In a color-saturated environment, well-designed black and white can be distinctive, "impactful", and memorable. It involves faster, cheaper, more reliable, easier-to-use equipment, and no agonizing over color choice. It is more copyable (there are more, cheaper, faster copiers)—remember: **Good Maps Get Copied.**

Be Careful with the Supposedly Safe Color Gray

- Black area-fill on a map hides shared boundaries.
- Gray shades can be difficult or impossible to distinguish.
- The human eye cannot reliably distinguish more than five shades of gray (or of any other color).
- Sometimes gray shades do not photocopy well.

Other Design and Construction Problems

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. For 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"—see the author's "Map Smart" paper cited above. The TOPSPOTS macro does not support annotation, but the author has in mind an enhancement to provide an alternative.

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.), 1993.

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. 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.

Automated Rationale-Based Response Range Assignment was developed to optimally and flexibly handle the cases of: four response ranges; five response-change ranges; and N cluster-based response ranges. (See "Map Smart".) Such maps are shown in the presentation, but cannot be included here. Here our focus is on the "Top Spots" rationale.

TOPSPOTS Macro and Figures 1 & 2

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, the most memorable and most readily assimilated 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 a large majority, or even 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 99% of the total response. For the data depicted in Figure 1, the Top 10 countries (out of 54) account for 63% of the total response.

The TOPSPOTS macro does 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 just enough of the highest response areas to account for a specified percent of the total response; and
(c) highlight all areas with a response at or over a minimum. If the response variable being mapped is not additive, Option (b) is inapplicable. E.g., population is additive, but population density is non-additive. For Options (a) and (b), the macro user can also specify a minimum. I.e., highlight the Top areas, but only those at or over the minimum.

The macro is maximally informative. It dynamically builds explanatory subtitles and footnotes. It states the selection criterion. If a minimum was specified, it states this auxiliary criterion. If the response variable is additive: (1) it states how many areas qualified, and what subtotal and what percent of the total response they account for; and (2) it states the total count of areas and their total response. When exception situations alter the significance of the map, they are noted.

For each area-fill block, the legend can optionally show either the actual range of responses, default text, or custom text. When the legend displays ranges instead of text, it presents "trimmed" ranges with the actual data bounds, emphasizing the inter-range separation. Traditional ranges are "contiguous", not maximally informative. The user can specify any standard SAS format for the legend range bound entries, if the macro's algorithm gives unsatisfactory results.

If the user does not request a separate category for zero response, the map is always two-color. Otherwise, the map will be three-color, if any zero response is found. The user can specify that missing values be set to zero. If the user requests it, the presence of any missing values will be footnoted; and if they have been set to zero, that will be part of the footnote.

Future Enhancements. Between press time and presentation time, the author will solidify a statement of planned enhancements, and may actually have implemented some.

How To Use PROC GMAP

For wider applicability, the discussion here is in the context of Version 6, specifically Release 6.12, which was used for development. For more information, consult *SAS/GRAPH Software: Reference, Version 6, First Edition, Volumes 1 and 2,* SAS Institute Inc. (Cary, N.C.), 1990. There are several related manuals and Technical Reports, too numerous to list. For their titles, please see the Publications Catalog at the SAS web site.

The remarks below are based on the above manual. They are not meant to be complete. Options or statements omitted are unused here, and, in some cases, actually "anti-recommended".

```
PROC  GMAP  MAP=map-data-set
            DATA=response -data-set  ALL;
```

The map data set contains boundary coordinates for the geographic unit areas. It is typically one supplied by SAS Institute, but need not be. The response data set is user-supplied information to be charted in map form. For map and response data to work together, both must contain ID variables, or "keys", for the geographic unit areas. The "ALL" option assures the map will contain every unit area in the map data set, even if an area's keys are absent in the response data, or the response value is missing for the unit area.

ID variable(s);

The ID statement specifies the unit area identification variable(s) (e.g., STATE for MAPS.US, COUNTRY for MAPS.AFRICA). In some cases, there may be multiple ID variables.

CHORO response-variable / DISCRETE
LEGEND=LEGENDn COUTLINE=BLACK;

The CHORO statement requests a choropleth map. Such a map uses area-fill to distinguish the response value/range.

DISCRETE means that a numeric response value is being treated as a discrete variable, not a continuous variable. This is a bit confusing. The "raw" response variable in an InfoGeographic may actually be, and usually is, continuous. However, usually, and particularly in the context of the TOPSPOTS macro, it gets converted to a smaller set of discrete values by use of a FORMAT statement. The TOPSPOTS macro automatically and dynamically generates the FORMAT statement for you, depending on what option you select for the macro, and reflecting the specific content of the response data set.

LEGENDn (n in the range 1 to 99) identifies a LEGEND statement that associates samples of area-fill with descriptions of the response ranges used for response value classification. The TOPSPOTS macro creates a customized LEGEND1 statement.

COUTLINE=BLACK assures the boundary lines are BLACK, and, therefore, maximally conspicuous—an eminently reasonable design goal in a chart that delineates territory.

Trademarks

SAS/GRAPH and SAS are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® denotes USA registration.

Author Information and Related Work

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

LeRoy Bessler is a SAS consultant, with interests in macro-based Software-Intelligent Application Development, visual communication, graphic design, information visualization, color, and InfoGeographics. An internationally recognized expert on SAS/GRAPH and graphic design, and an award winner for papers on graphic design and visual communication, Dr. Bessler is writing a book titled "Chart Smart: Design Guide and Solution Toolkit for SAS Graphs, Tables, and Maps That Inform and Influence".

See his companion SUGI 25 paper: "Show Them What's Important: Solutions for a Finite Workday in an Era of Information Overload".

Invocation of TOPSPOTS Macro for Figures 1 & 2

```
 /* first set GOPTIONS & get population data */

%TOPSPOTS(DATA=POPDATA,
          MAPDATA=AFRICA,
          IDVAR=ID,
          AREASNAM=Countries,
          CHOROVAR=POP,
          MAPTYPE=NUM,
          N=10,
          FMT=COMMA11.,
          LEGTITLE=Population,
TTLTEXT=Distribution of Population in Africa,
          FTNTTXT=
SAS/GRAPH Africa map does not display six island
countries:,
          FTNT2TXT=
%STR(Mauritius, Reunion, Comoros, Seychelles,)
%STR(Sao Tome & Principe, Cape Verde),
          LEGSHFTH=-14.4 PCT,
          HEADROOM=0.25 IN)
RUN;

%TOPSPOTS(DATA=POPDATA,
          MAPDATA=AFRICA,
          IDVAR=ID,
          AREASNAM=Countries,
          CHOROVAR=POPDENS,
          CHVARTYP=NONADD,
          CHVARMIN=50,
          MAPTYPE=MIN,
          LEGTITLE=Population Per Sq Mi,
          LEGFMT=9., /* "over-allow" */
          TTLTEXT=Population Density in Africa,
          FTNTTXT=
SAS/GRAPH Africa map does not display six island
countries:,
          FTNT2TXT=
%STR(Mauritius, Reunion, Comoros, Seychelles,)
%STR(Sao Tome & Principe, Cape Verde),
          LEGSHFTH=-21.7 PCT,
          EXPLJ=C,
          HEADROOM=0.25 IN)
RUN;
```

NOTES:

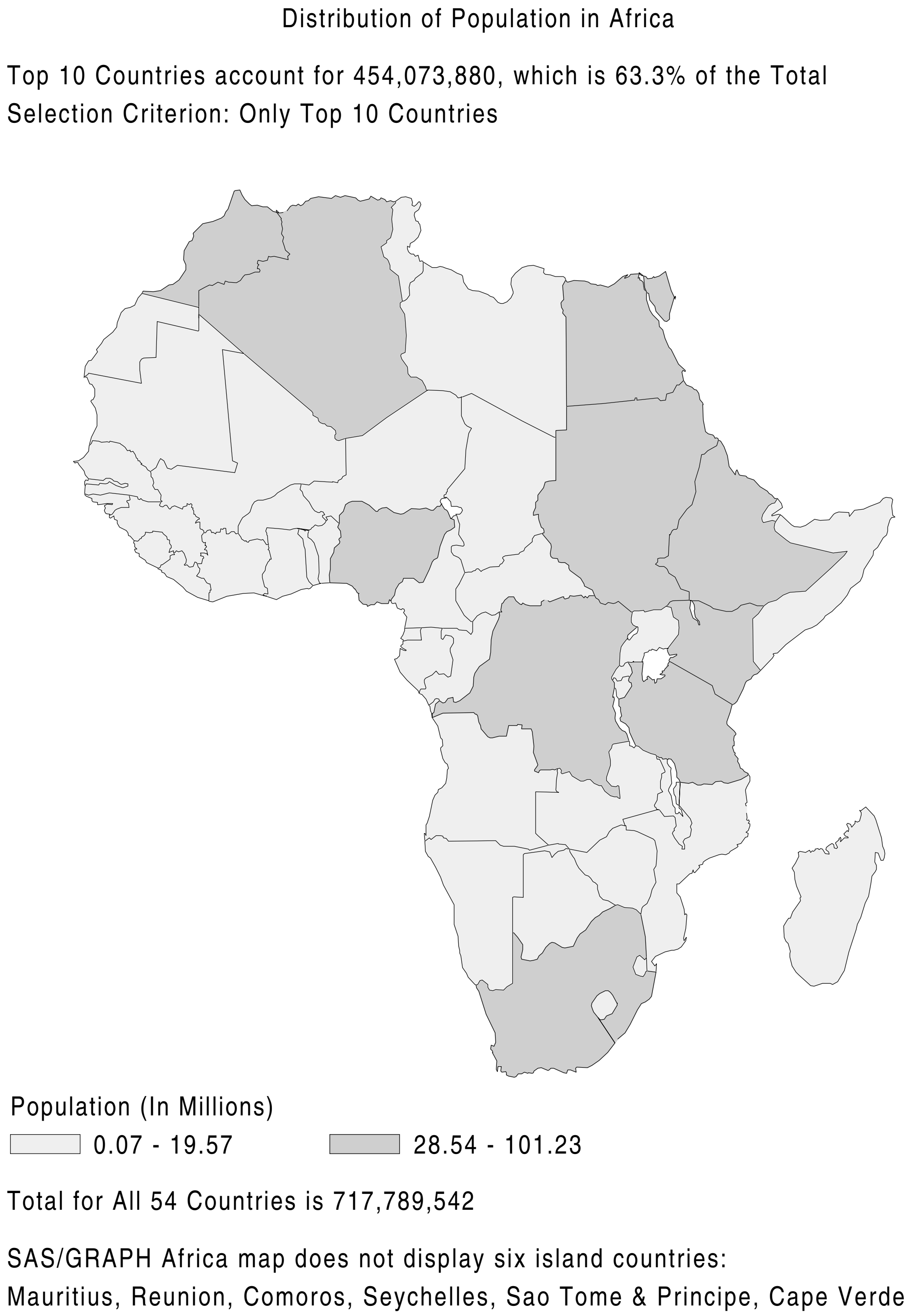
For the six small island countries, the SAS/GRAPH Africa map data set contains only points. So, polygons cannot be drawn to create visible areas.

Population data used is from "The World Factbook 1995", published by the US Central Intelligence Agency .

Bird’s Eye View of the TOPSPOTS Macro (If defaults and required parameters suffice, just ignore the other options.)

```
%MACRO TOPSPOTS(DATA=,                     /* response data set, REQD */
 MAPDATA=,                                      /* map data set, REQD */
 IDVAR=ID,                 /* unit area identification variable, REQD */
 AREASNAM=,            /* description of unit areas (e.g., Countries) */
 CNTRYSEL=, /* condition to select COUNTRY, if applicable to map data */
 REGNSEL=,  /* condition to select REGION,  if applicable to map data */
 IDSEL=,                                     /* condition to select ID */
 CHOROVAR=,                                /* response variable, REQD */
 CHVARTYP=ADD,   /* response var is ADD(itive) or NONADD(itive), REQD */
 MAPTYPE=NUM,                /* NUM for Top N, alternatives PCT & MIN */
                     /* MAPTYPE=PCT not applicable if CHVARTYP=NONADD */
 N=10,                                  /* Top 10 map, if MAPTYPE=NUM */
 PCT=80,  /* show areas for top 80% of total response, if MAPTYPE=PCT */
 CHVARMIN=,               /* minimum response value to qualify as TOP */
 SEP_ZERO=YES,  /* make 0 separate response range, NO to put in OTHER */
 MISSZERO=NO,   /* do not assign 0 to missing values, YES if to do so */
 MISSNOTE=YES,            /* display a footnote if any missing values */
 SHOTOTAL=YES,  /* display Total response for all areas in a footnote */
                    /* SHOTOTAL=YES not applicable if CHVARTYP=NONADD */
 FMT=BEST32.,     /* optional format for displayed non-Legend numbers */
 LEGTXT=NO,     /* range bounds for LEGEND entries, YES for text desc */
 LEG_NONE=,                 /* LEGEND entry desc for response value 0 */
 LEG_OTH=,           /* LEGEND entry desc for Other response category */
 LEGTITLE=,    /* LEGEND desc, ignored if LEGTXT=YES & there is OTHER */
 LEGFMT=,     /* if LEGTXT=NO, use this format, not default algorithm */
 TTLTEXT=,                                          /* text for title */
 FTNTTXT=,                         /* text for first  custom footnote */
 FTNT2TXT=,                        /* text for second custom footnote */
 PREVIEW=NO,              /* display on Monitor, with option to Print */
 DISPDRVR=WIN,                                 /* DEVICE= for Monitor */
 PRTDRVR=WINPRTG,               /* DEVICE= or TARGETDEVICE= for Print */
 ORIENT=BEST, /* let macro decide; alternatives PORTRAIT VS LANDSCAPE */
 LANVSPOR=1.29,   /* if map width/height ratio GE this, use LANDSCAPE */
 FCOLNONE=GRAYF0,              /* area-fill COLOR= for category None  */
 FCOL_OTH=GRAYD0,              /* area-fill COLOR= for category Other */
 FCOL_TOP=GRAYB0,              /* area-fill COLOR= for category Top   */
 BORDER=NO,                   /* YES to get a border around the image */
 LEGMODE=RESERVE,      /* REQD, other alternatives are SHARE, PROTECT */
  /* SHARE with LEGPOS=INSIDE can produce overlay of legend and map   */
  /* RESERVE requires LEGPOS=OUTSIDE; PROTECT is not recommended      */
 LEGPOS=(BOTTOM CENTER OUTSIDE),             /* Legend POSITION, REQD */
  /* Y-pos: BOTTOM, MIDDLE, or TOP; X-pos: LEFT, CENTER, or RIGHT;    */
  /* position the Legend INSIDE or OUTSIDE the area used for the map  */
 LEGSHFTH=,  /* optional right (+) or left (-) PCT OFFSET from LEGPOS */
 LEGSHFTV=,  /* optional up    (+) or down (-) PCT OFFSET from LEGPOS */
 TTLF=NONE,                        /* NONE if default font for TITLEs */
 TTLH=1,                                     /* height for TITLE text */
 TTLJ=C,       /* CENTER the TITLE text, alternative L(eft) & R(ight) */
 TTLSHFT=,    /* indent +KK PCT, if TTLJ=L; indent -KK PCT, if TTLJ=R */
 EXPLF=NONE,        /* NONE if default font for explanatory subtitles */
 EXPLH=1,                                 /* height for subtitle text */
 EXPLJ=L,     /* Justify L(eft) subtitle text, alt C(enter) & R(ight) */
 EXPLSHFT=+10 PCT,   /* indent +10 PCT for EXPLJ=L; do not if EXPLJ=C */
 TOTALJ=L, /* Justify L(eft) for Tot footnote, alt C(enter) & R(ight) */
 TOTSHFT=+10 PCT,  /* indent +10 PCT for TOTALJ=L; do not if TOTALJ=C */
 FTNTF=NONE,             /* NONE if default font for custom footnotes */
 FTNTH=1,                              /* height for custom footnotes */
 FTNTJ=L,  /* Justify L(eft) custom footnotes, alt C(enter) & R(ight) */
 FTNTSHFT=+10 PCT,   /* indent +10 PCT for FTNTJ=L; do not if FTNTJ=C */
 HEADROOM=, /* add space betw titles & map (IN or CM or PCT or CELLS) */
 TMRGNADD=,         /* add a Top    Margin (IN or CM or PCT or CELLS) */
 BMRGNADD=);        /* add a Bottom Margin (IN or CM or PCT or CELLS) */
 /* macro internal code not shown */ %MEND TOPSPOTS;
```

Figure 1. *Note: The map program did use the population of the island countries not displayed.*



Distribution of Population in Africa

Top 10 Countries account for 454,073,880, which is 63.3% of the Total

Selection Criterion: Only Top 10 Countries

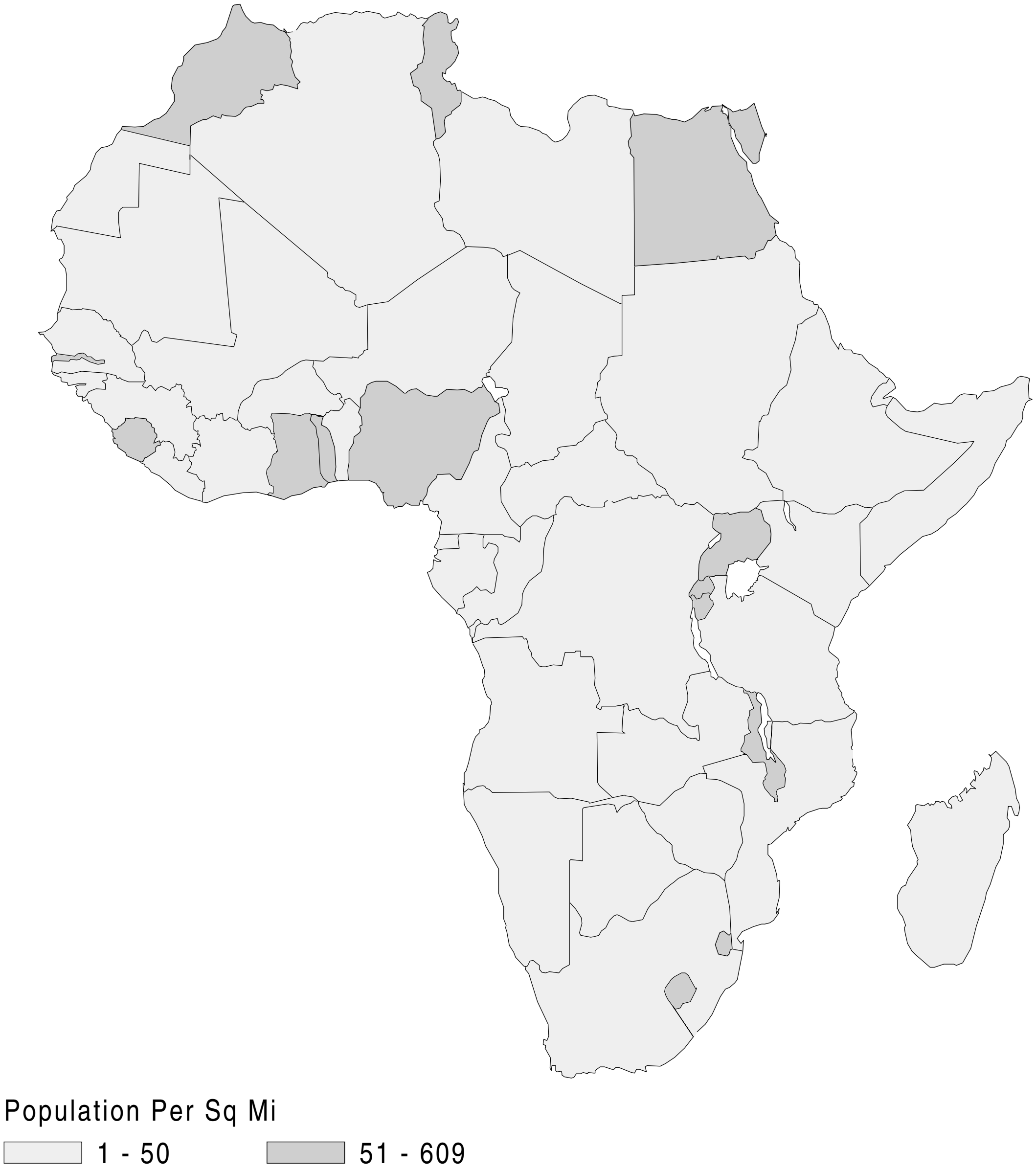
Total for All 54 Countries is 717,789,542

SAS/GRAPH Africa map does not display six island countries:
Mauritius, Reunion, Comoros, Seychelles, Sao Tome & Principe, Cape Verde

Figure 2. *Note: The map program did use the population density of the island countries not displayed.*

Population Density in Africa

Selection Criterion: Only Countries with value not less than 50



SAS/GRAPH Africa map does not display six island countries:
Mauritius, Reunion, Comoros, Seychelles, Sao Tome & Principe, Cape Verde