STATISTICAL MAPPING CAPABILITIES AND POTENTIALS OF SAS

Grady B. Meehan
Institute for Research in Social Science
University of North Carolina
Manning Hall 026A
Chapel Hill, N.C. 27514
U.S.A.

ABSTRACT

SAS, the statistical analysis, data management and graphics system (with SAS/GRAPH) offers several procedures and programming capabilities that provide for the efficient analysis and mapping of geographically based statistical data. Cartographic data is stored in a standard statistical file providing efficiencies in storing and making the data easy to create, document, access, manipulate and transfer between users. Mapping procedures are available to perform map projections, boundary line generalization, the redefinition of geographic regions from existing files and the output of several types of statistical maps to a variety of widely available graphics output devices, including color. Many useful features support computer-assisted statistical mapping and graphing for data display and analysis. The easy exchange of cartographic and statistical data files between researchers at widely separated locations is facilitated, while insuring file compatability. The present and future cartographic potentials of SAS are explored, along with examples of statistical map output.

INTRODUCTION

In the early 1970's, statistical computing began to change, with the introduction of integrated statistical packages. They permitted a user to perform one or more statistical procedures on a single data set stored in a standardized file format. The widespread adoption of such packages by researchers and teachers in social sciences led to the widespread acceptance of statistical computing. Packages such as SPSS, DATATEXT, OSIRIS, SAS and others were used in graduate and undergraduate training. Geographers and others who work with spatial (geocoded) data also employed these packages, but found that maps still had to be produced by manual methods, a slow and tedious chore. On the other hand, computer mapping packages developed and distributed by Harvard University's Laboratory for Computer Graphics and Spatial Analysis reduced the labor intensive map making task, however they lacked the capability to easily transform or manipulate the attribute or cartographic data. In addition, data handling was cumbersome because the mapping packages lacked standard system data files and the data management capabilities which are part of most popular statistical packages.

While many technological advances have lowered costs in computer graphics technology, a growing need for statistical and other thematic maps and graphs combined with software advances have placed statistical cartography in the position of statistical computing five to ten years ago. SAS/GRAPH, a statistical graphics and computer mapping system integrated with SAS, now provides map makers with a tool for the easy use of geographically based data that overcomes the limitations outlined above. The purpose of this paper is to examine some existing mapping capabilities of SAS/GRAPH and potential applications that will benefit the statistical mapping community.

SAS MAPPING FEATURES

SAS is an integrated statistical analysis, reporting and data management software package. In 1980, SAS added SAS/GRAPH, a package for statistical graphing and mapping that could be implemented at existing SAS installations. The combination of features in both packages provides a very powerful tool for those who analyze spatial data and require statistical graphs and maps as output. SAS/GRAPH has four procedures for computer mapping applications (SAS Institute, 1981, 1982a). They are:

1. PROC GREDUCE filters out points contained in the map data set that are not needed for the proper appearance of the map (Douglas and Peucker, 1973). The results are reduced storage requirements and processing costs.
2. PROC GREMOVE deletes internal boundaries of regions to redefine the geographical hierarchy. For example, Census regions are created by removing selected state boundaries from the United States map data set and keeping only those boundaries which make up the external regional boundaries.
3. PROC GPROJECT applies either the Albers equal area, Lambert conformal conic or the gnomonic projection to a map data set containing the unprojected coordinates stored as radians.
4. PROC GMAP produces the map output by using both the map and attribute data sets. Types of plotted output presently includes the choropleth, prism, block and surface maps.

The SAS programming statements used to produce a map or perform utility operations on map data sets are few in number and simple to learn. The ability to analyze, manipulate and manage data requires some understanding of how SAS processes data. This necessitates some training and practice, much the same as if one were using the statistical procedures. Since SAS is a data processing package, one of the major problems it solves for computer mapping is the management of the many, sometimes large data sets required for computer mapping projects.

HARDWARE UTILIZATION REQUIREMENTS

While SAS is presently running on IBM 370 compatible mainframes under several different operating systems, a version for Data General "super" minicomputers (32 bit), has been announced by the SAS Institute (SAS Communications, 1982b, p. 3). The somewhat limited choice of mainframe computers, however doesn't apply to the choice of graphic output devices. At the present time, many different models of interactive monochrome and color CRTs and hard copy plotting devices from more than a dozen manufacturers are directly supported by SAS/GRAPH. In addition, a "universal device driver" will interface those graphic devices not directly supported. Program directed options available to the programmer within SAS/GRAPH resolve hardware differences and also take advantage of special features that are built into certain graphics terminals and plotters. For example, if a user works with a Tektronix 4027 color crt, SAS/GRAPH has a procedure that will enable the terminal's function keys to help streamline a terminal session. Thus, different hardware characteristics are resolved by the software for each specific device. Printer produced maps resembling SYMAP are not available in SAS, but some attempts to program them have been made by individual users (Spitznagel, 1980).

SAS MAP DATA SETS

One of the most powerful features of SAS is the ability to read and store nearly any type of machine-readable data using any of a variety of input formats. A single SAS data set can store over 1000 variables and an unlimited number of observations. Map data sets require several variables, including a geographic code variable, horizontal and vertical coordinates, and segment identifiers to accomodate any case in which a single region is made up of more than one polygon. Hierarchical files are also accomodated by SAS.

When creating a SAS map data set, variable names are stored with the data set as are user comments. The later are a valuable documentation feature, especially for storing a description of the data. For example, map data sets might contain statements describing the coordinate system, the source, scale and projection of the source map, the name of the digitizing orgainization and any other information required for internal documentation. Any subsequent user of the map data set can print the internal documentation by using PROC CONTENTS. In addition, automatic documentation such as the type of storage device, names and sizes of files, the time and date of creation, the names and data formats of all variables are also printed. Another useful feature is the ability to store geographical area names up to 200 characters long as a single variable.

The importance of complete map file documentation has been a subject addressed in the cartographic literature. Information stored internally with the data will not be missing or hard to find as is sometimes the case with separate printed documentation. Universities, governments and private industry concerned with maintaining data archives are concerned with data integrity that includes adequate documentation for cataloging machine-readable data files. SAS facilitates the efficient management of large data bases through internal data set documentation. Since map data sets are expensive to create and maintain, efforts should be made to protect this investement with proper documentation.

Good documentation also facilitates the transfer of data between widely separated installations. Problems associated with tape file transfers are minimized with SAS data sets, often saving much time and effort when tape based files are transferred from one place to another.

USING THE MAPPING PROCEDURES

The SAS Institute supplies four map data sets (cartograpghic data bases) with SAS/GRAPH. They are:

1. United States by state (unprojected)
2. United States by county (unprojected)
3. United States by State (projected and reduced)

4. Canada by province and census district (projected and reduced).

Only the reduced and projected United States map data set is in a form ready to be used in a mapping procedure. The other data sets must be projected and/or reduced before they can be used.

Preparing SAS map data sets for a mapping project is easy and may be accomplished within the same job that produces the map. In normal practice, map data sets are prepared once and stored for later use with the mapping procedure for the sake of economy. If, for example, a map were needed that required the data contained in the unprojected "states" map data set, several preprocessing steps would be necessary. Our example will assume we want to create a map of the four census regions for the contiguous United States.

The first step would consist of removing the state boundaries internal to the census regions using PROC GREMOVE. To accomplish this, we create a data set with cross references from each of the 48 states to its corresponding census region. This data set is then merged with the map data set so that each of the 15,506 observations contains the appropriate region code. This is easily accomplished with several SAS statements that invoke the data management programming steps of SAS. Next, the data is processed by PROC GREMOVE, resulting in the creation of a data set containing the desired regional boundaries. A map of the census regions is shown in Figure 1. If it were necessary to reduce

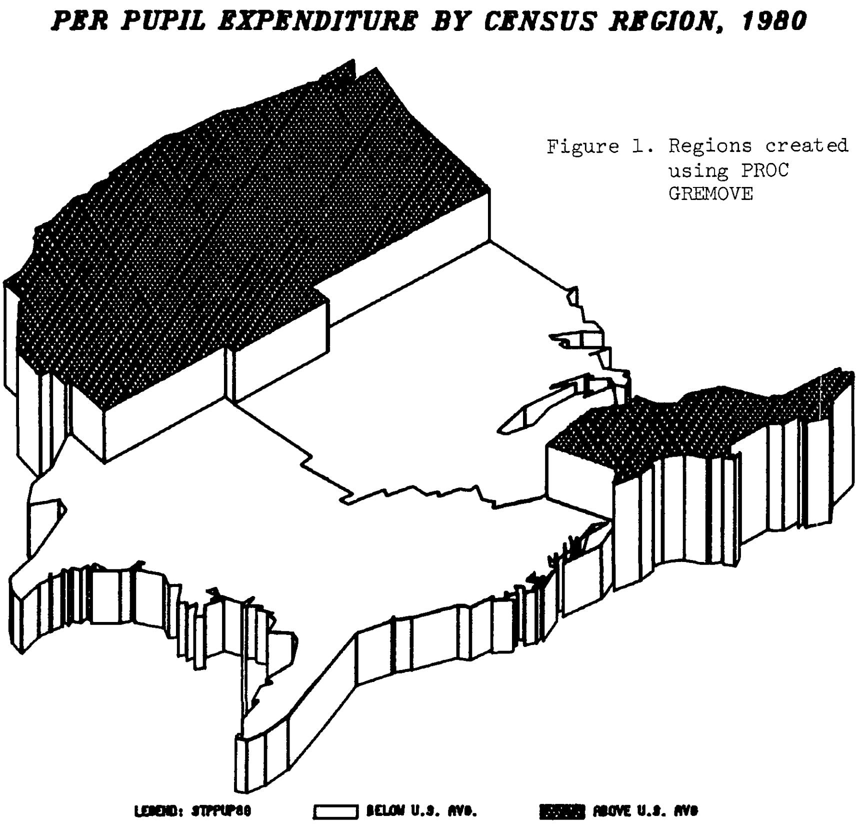

Figure 1. Regions created using PROC GREMOVE

the physical size of the final output file or limit the maximum number of coordinates per polygon, then PROC GREDUCE would have been used prior to creating the final map data set. PROC GREDUCE employs a line filtering algorithm to reduce the number of coordinates in a digitized line. For this particular figure, however, the projected and reduced United States map data set was used.

Next, the data set containing the census regions is processed by PROC GPROJECT, the final step necessary prior to mapping when an unprojected map data set is selected. For the United States, the default Albers equal area map projection is appropriate. If one of the three available map projections is not suitable, then SAS programming statements could be employed to compute the desired projection for the particular data set. The ease of using programming statements within SAS by a user makes it a very flexible package for performing any type of transformation on map (or attribute) data and mapping the result.

Statistical maps are produced by selecting an attribute data variable and matching it with the appropriate geographical code stored in a map data set. The link between map data and attribute data is the geographic code contained in each data set. The mapped variable is specified by name, the geocode variable name is identified, the map type specified and the number of symbolism classes (if appropriate) is chosen. Since the attribute data controls which polygons are plotted on the map, subsets of large data sets need not be specially created. For example, if an attribute data set contained only data for the states of the southeastern United States, then only they would be plotted from a map data set containing all forty-eight states.

The plotter output space has title space at the top, footnote space at the bottom and map/graph space between the two. SAS/GRAPH scales a map to fit in the plotter space remaining after titles and footnotes are plotted. If a series of maps are being produced, for example, then a constant number of title and footnote lines on each map will result in each one being at the same scale. If necessary, dummy title and/or footnote lines should be inserted for consistancy.

The choropleth map is frequently selected to display statistical data and it is available in PROC GMAP. Here an example of a choropleth map of the southeastern states is shown (Figure 2). The newest types of maps available in SAS/GRAPH are the prism and block maps which were released in the latest version of the package (SAS Institute, 1982a). The prism map example shows the 1980 populations for twelve south-eastern states (Figure 3).

The final example is the block map, consisting of graduated vertical bars placed at the centroid of each polygon. The map of Greensboro's 1960-1970 population change is a block map (Figure 4). The map data set containing the census tracts is a modified version of the Urban Atlas map files developed by the U.S. Census Bureau (Schweitzer, 1973) and distributed separately by the SAS Institute as a map data set. The 1970 tract boundaries are available for over 200 cities of the United States. The Urban Atlas data set is much easier to use for mapping as a SAS data set than the original version. Although the 1980 tract boundaries are not available, the 1970 tract boundaries, still have value as an educational tool.

SUMMARY AND CONCLUSIONS

SAS facilitates statistical mapping. Features such as the standard file format for the storage of both attribute and map data, internal documentation, data transformation, manipulation and management give the map maker many capabilities combined with the flexibility of SAS. Map making requires a few simple SAS statements to create any of several types of statistical maps. The possibility of creating, storing and archiving map data sets will permit SAS users to carry out large projects that necessitate the use of large attribute and/or map data sets. SAS is well supported with frequent updating, the addition of new procedures and a large, active world-wide user group that convenes annual meetings and publishes proceedings.

PERCENT BELOW POVERTY, 1979

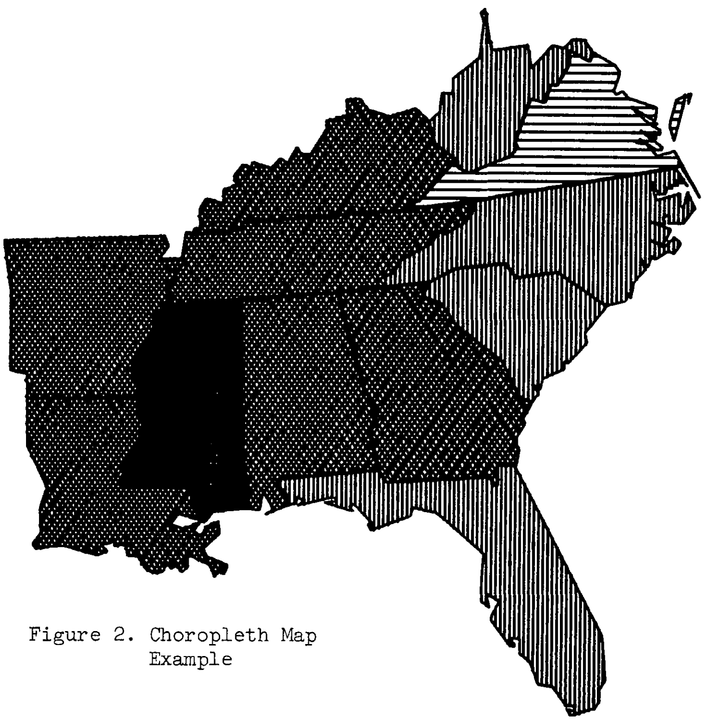

Figure 2. Choropleth Map Example

SOUTHEASTERN U.S. POPULATION, 1980

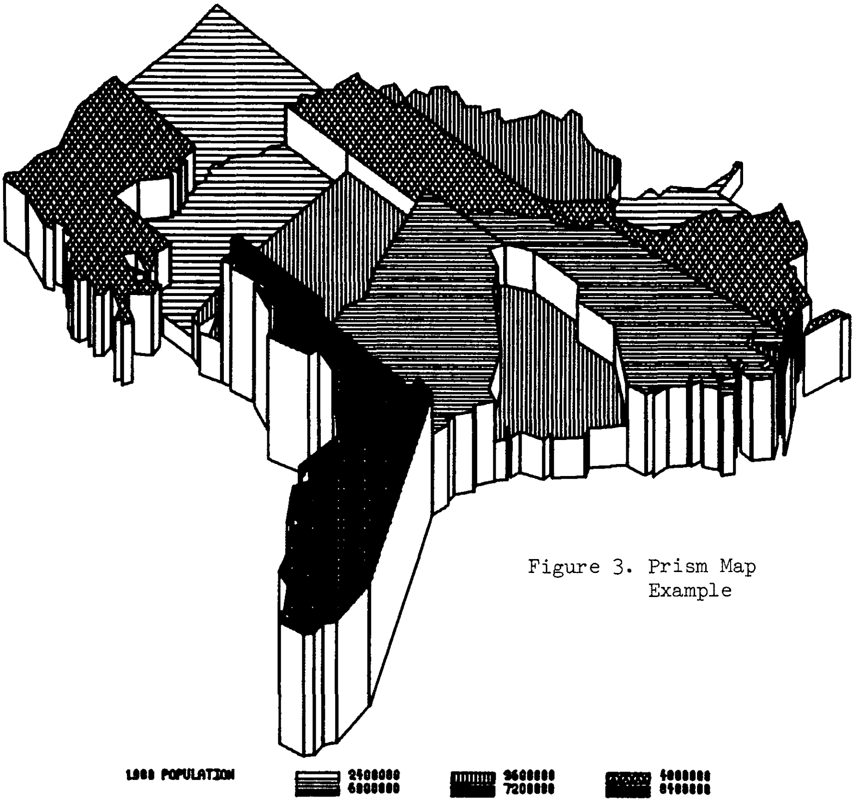

Figure 3. Prism Map Example

REFERENCES

Douglas, David H. and T.K. Peucker 1973, Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or its Caricature: The Canadian Cartographer, Vol. X, pp. 112-122

SAS Institute 1981, SAS/GRAPH User's Guide, 1981 Edition, SAS Institute, Inc., Cary, North Carolina

SAS Institute 1982a, SAS/GRAPH Enhancements and Updates for 79.6, SAS Technical Report P-119, SAS Institute, Inc., Cary, North Carolina

SAS Institute 1982b, SAS Institute Picks Data General: SAS Communications, Vol. VIII, p. 3

Schweitzer, Richard H., Jr. 1973, Mapping Urban America with Automated Cartography: paper presented at the Fall Convention of the American Congress on Surveying and Mapping, Orlando, Florida

Spitznagel, Edward L., Jr. 1980, Shaded Map Reports: Proceedings of the Fifth Annual SAS User's Group International Conference, SAS Institute, Inc., Cary, North Carolina, pp. 475-481

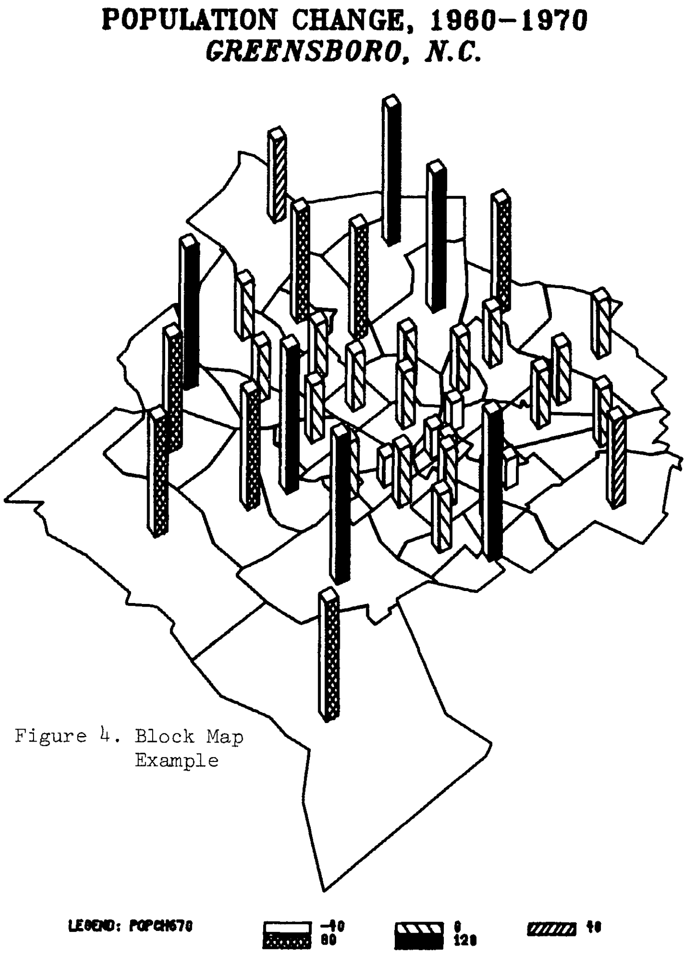

Figure 4. Block Map Example