ABSTRACT

Most SAS/GRAPH™ mapping is done at the state and county levels of geography using cartographic data bases (boundary coordinates) supplied by the SAS Institute. However, SAS/GRAPH has a completely general mapping capability limited only by the kind of data and cartographic databases supplied to it. This paper briefly outlines the history of cartographic database development as the basis for thematic mapping and describes options for SAS users desiring to map data at various levels of geography.

INTRODUCTION

Displaying data in thematic map form is a powerful tool for assisting an analyst in identifying spatial patterns that may be present in the data. The spatial patterns may reveal insights which would be obscured in a tabular data presentation: a cluster of homes all insured by one company which would make the company vulnerable to large losses in the event of a spreading fire, tornado or flood; a pocket of profitable accounts, which might be better served by a branch office or expanded by a targeted marketing campaign; systematically shifting demands for emergency services at different times of the day or days of the week, which could be anticipated by changing deployment of emergency services.

Historically, the power of visual examination of data patterns has been hampered by the tedious nature of map preparation. For example, consider a study involving plotting every call for emergency service as a point symbol on a series of maps: one for each hour of the day, a series for each day of the week, a breakout by season or weather condition, another by type of complaint, and yet another showing lengthy response time. Although such an analysis could reveal potentially life-saving insights into deployment patterns of emergency vehicles, the number of maps required makes the study impossible by traditional cartographic methods.

However, it's now possible to make hundreds of computer maps detailed to reveal subtle phenomena using computer mapping programs. And one such program — probably the most widely used at present — is the mapping feature of SAS/GRAPH. SAS/GRAPH provides the basic functionality found in most commercially available mapping software packages, and in addition it has an often overlooked feature which can be used to produce incidence maps — a powerful tool little used by geographers.

Getting a map out of a computer — especially with a program like SAS/GRAPH — is easy, once the map is in the computer in the first place. The need to "digitize" or computerize the base map for any study is often overlooked. Creation of digitized base maps — a prerequisite for any computer thematic mapping — is the subject of this paper.

"CARTOGRAPHIC DATA BASES" (CDB)

There are five ingredients necessary for mapping data by computer. Four are obvious: the computer itself, the data to be mapped, a plotter or other kind of graphic display device, and software to generate the map display. The fifth element is
not so obvious. It is a special kind of data file which describes the relationships between the location of the various data events in the absolute geographic language of latitude and longitude. This database, which really is just a digitally encoded map, is called a cartographic database or CDB.

Just as paper maps can define boundaries of states, counties, towns, census tracts, city blocks, land use zones or ownership parcels, cartographic data bases can also represent these boundaries. SAS/GRAPH comes with state and county boundaries, but you need ZIP Code boundaries to map data by ZIP, or block boundaries to map data by block. Once a CBD is made for any set of boundaries, it can be used repeatedly to make many different maps of different data variables grouped and classified in a wide variety of ways.

COMPONENTS OF A CDB

The CDBs used by SAS/GRAPH are relatively simple and typical. Here is a sample CBD for mapping data for the states of Colorado and Wyoming:

<table>
<thead>
<tr>
<th>State Name</th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colorado</td>
<td>37.0</td>
<td>102.1</td>
</tr>
<tr>
<td>Colorado</td>
<td>37.0</td>
<td>109.1</td>
</tr>
<tr>
<td>Colorado</td>
<td>41.0</td>
<td>109.1</td>
</tr>
<tr>
<td>Colorado</td>
<td>44.0</td>
<td>102.1</td>
</tr>
<tr>
<td>Wyoming</td>
<td>41.0</td>
<td>104.1</td>
</tr>
<tr>
<td>Wyoming</td>
<td>41.0</td>
<td>111.1</td>
</tr>
<tr>
<td>Wyoming</td>
<td>45.0</td>
<td>111.1</td>
</tr>
<tr>
<td>Wyoming</td>
<td>45.0</td>
<td>104.1</td>
</tr>
</tbody>
</table>

A little research with a map that has latitude and longitude marks will reveal that the SAS/GRAPH CBD for these states consists only of the four corner point coordinate readings for each state, arranged in order around the boundary and identified by the state name. This is sufficient information to allow SAS/GRAPH to map any data for the two states.

This is a typical CBD. There are a few complications; obviously, states with more complex boundaries will have more latitude/longitude pairs to describe the shape. The number of points determines the fidelity of the final thematic map. States like Hawaii, Michigan and Virginia require special treatment since their boundaries consist of several polygons, each of which must be described to SAS/GRAPH. Finally, some data analysis zones are completely embedded in others and a special coding convention described in the SAS literature is required to define the embedded portions.

We made the CBD for Colorado and Wyoming with a ruler. Since there were only eight points to digitize, this wasn’t too bad a job. However, the danger of manual digitizing is that it is easy to make errors and hard to detect them. For example, in our Colorado/Wyoming CBD, the last latitude measurement for Colorado should be 41.0 instead of 44.0, a difference that is not obvious in the printed listing, but which would grossly distort the map. The greater the number of data zones, and the more complex their boundaries are, the harder it is to make a usable CBD with manual coordinate measurement.

USING A DIGITIZER

If one is available, it is preferable to make a CBD using a digitizer: a graphical input device capable of making precise coordinate measurements. Although there are many types of digitizers, they all share three things in common: a working surface on which the user can affix the map to be digitized, a pointing device or cursor which has one or more switches to signal when a measurement is desired, and an electronics unit to generate the measurement and convert it into a data value for storage or immediate computer processing. While early digitizers simply punched out an IBM card every time the button was pushed, virtually all modern units operate under the control of an interactive digitizing program that gives the user immediate graphical feedback as the digitizing progresses.

Although we practically take them for granted, digitizers have not always been available. When computer mapping was in its infancy in the 1960’s, many researchers created CBDs by hand, measuring each coordinate point with a ruler. In fact, the first widely used computer mapping program, SYMAP, came with a special ruler graduated in the SYMAP coordinate system. Digitizing cartographic databases was such an obstacle that few agencies used computer mappings. Two organizations, the US Census Bureau and the Tri-State Transportation Commission (New York), actually built their own digitizers for in-house computer mapping projects.

A-107 PROCEDURES

The “Bible” for digitizing a CDB for SAS/GRAPH is SAS Technical Report A-107 "Creating your own SAS/GRAPH Map Data Sets with a Digitizer". This report is a must for anyone desiring to digitize a CDB for SAS/GRAPH. It is clear and concise, so we will not repeat a lot of its content here. Basically, A-107 presents several SAS programs which help the user to detect coding errors in the digitizing process and a way to permit one-time digitizing of points, even if they are to be used twice in describing the boundaries of two adjacent polygons.

A-107 also addresses issues of coordinate transformation and how to digitize a study area that is defined on two or more map sheets. A-107 lets researchers go beyond the limits of manually digitizing with a ruler, and vastly expands the usefulness and flexibility of SAS/GRAPH.

However, even with A-107, some jobs are unwieldy, for example digitizing the boundaries of 45,000 Census tracts or 35,000 cities and towns. Projects of this size require the ability to "sew" together coordinate readings taken from thousands of mapsheets, and they demand an interactive environment for operator efficiency.
THIESSEN POLYGONS

It is possible to make zone boundaries computationally, and skip the manual effort required to trace the boundary of each data collection zone. If a centroid coordinate is known for each polygon, it's possible to calculate a "Thiessen polygon" as a surrogate boundary by splitting the distance between adjacent centroids. The algorithm for making so-called "proximal" boundaries is well known, and has been applied to centroid coordinates supplied by the Census Bureau in the "MARF" (Master Area Reference File). Unfortunately, proximal boundaries end up looking like rather abstract honeycombs, and are often hard to relate to the actual boundaries they're meant to simulate.

OTHER SOURCES

Fortunately for mappers, various organizations have made commonly-used Cartographic Data Bases available as off-the-shelf products. The first to do so was the Census Bureau which released two widely-used CBDBs in the mid-1970's: nationwide county boundaries, which until recently have been distributed with SAS/GRAPH, and the boundaries of about 35,000 Census Tracts defined for the 1970 Census. Availability of these two CBDBs kicked off the mini-boom in thematic mapping that has been building for the past decade.

Troubles with processing the 1980 Census prevented the Census Bureau from upgrading its Tract boundary coordinates to represent the 45,000 tracts defined for the 1980 Census. All of these boundaries are available from commercial sources, however, as well as block, block-group and enumeration district boundaries. Other commercially available CBDBs include both true and proximal ZIP boundaries (about 40,000 zones) and the 35,000 MCDs and CCDs (cities and towns) in the USA.

In a typical large-scale digitizing operation — in this case, making a nationwide CBDB of true 5-digit ZIP boundaries — the operations are analogous to those described in A-107. There is a map-preparation step, where information on the boundaries to be digitized is brought together from various sources onto the base map. The digitizing itself is highly automated. The operator designates a ZIP to be digitized, then traces the boundary clockwise, making a coordinate reading at critical points determined by the digitizing standards. When the ZIP boundary is closed — and verified by the monitoring software — the operator enters the next 5-digit ZIP Code to trace. If the new ZIP shares a boundary with any others already digitized — and this can be verified by inspecting the screen display — then only the new portion of the boundary must be traced. The control program automatically hooks the new string of boundary points in to any previous digitizing. When an entire map sheet is completed, another is done, and so forth until the whole study area is completed. The final steps include sewing all the digitizing together across map-sheet borders, computer edits and visual checking of plots, and reformatting to accommodate a variety of desired distribution formats for various applications.

As part of the reformatting operation, the software makes land area and centroid calculations, and centroids are used for distance and direction computations.

INCIDENCE MAPPING

All of the mapping discussed involves displaying data that have been aggregated to a zone such as census tract or county. Often, this method of analysis averages out important details in the data, and another kind of display is more appropriate. Probably the most famous incidence map was made by Dr John Snow in 1854 depicting the location of cholera deaths in central London. As the story goes, analysis of the map led Snow to remove the handle of the Broad Street water pump, which stopped the epidemic, which by then had taken 500 lives.
Although cholera is no longer the threat it was a century ago, a latter-day Dr. Snow could use SAS/GRAPH to map incidence data by computer. The latest release of SAS/GRAPH has a feature called "Annotate Data Sets" which is intended for adding cosmetic annotation to a map display in the form of points, lines or polygons which SAS/GRAPH overlays on the map.

There is no rule that the annotate data set must be cosmetic; it could represent the primary data being displayed -- the location of a death by cholera in the example above.

Preparing an annotate data set can be done manually, by measuring an across and up coordinate reading for each point. As in the case of preparing a CBD with the A-107 procedures, access to a digitizer would speed up the process. However, as is the case with boundary digitizing a large number of points -- especially if they are only identified by a street address or intersection and are not plotted on a map -- can make the data preparation tedious and unreliable.

There are automated methods for putting coordinate measurements on data records identified by street addresses. In the public domain are the Admatch program and GBF/DIME files -- both available from the Census Bureau -- which can code addresses to coordinates in the IBM mainframe environment. Commercially available packages can do the same coding in an IBM-PC environment with the advantage of interactive correction of non-matched addresses.

This largely unutilized capability of SAS/GRAPH is potentially the most valuable automated mapping technique on the market since it bypasses the information-destroying steps of data aggregation and classifying that are required for mapping using the ubiquitous choropleth method.