AUTOMATED MAP INSET DETERMINATION

Frederick Roland Broome
Constance Beard
April A. Martinez
Geography Division
U.S. Bureau of the Census
Washington, D.C. 20233

ABSTRACT

Cartographers typically use their judgement to determine the need for an inset by visually inspecting a map and subjectively identifying the areas which are too dense for adequate feature or text placement. To determine insets an automated system must emulate the human decision process. This paper considers eight different approaches for automating map inset determination. It describes the development of a system for Census Bureau needs that is based on the most efficient of the eight methods.

INTRODUCTION

There are many situations where a map design must provide adequate space for text placement within areas delineated by map features. This is particularly true for most census field operations maps. These maps require space for placing census block numbers and sometimes space for marking residential structure locations along a street. Since maps used for census field operations are constrained by sheet size and number of sheets, the sizes of the areas for text placement on portions of the maps can become too small and insets must be produced.

Standard, good cartographic procedure calls for each map to be designed, examined, and on sheets where the text cannot be adequately placed, either the base map is enlarged or insets prepared . The field maps for the 1990 Decennial Census of Population and Housing of the United States will be prepared by computer programs run in a non-interactive, batch mode. Hundreds of thousands of maps, each different must be produced within a few months. There will be no opportunity to prepare "trial" maps either as hard copy plots or on a graphics terminal screen. If cartographic principles of map design are to be applied, they must become part of the computer programs used to generate the map plots. The map volume and short production schedule necessitates an efficient method of identifying when and where an inset is needed. This paper describes the research, development and implementation of an automated inset determination procedure.

DISCUSSION

Cartographers typically use their judgement to determine the need for an inset. They visually inspect the map and subjectively identify the limits of areas which are "too crowded", "too small", or "too dense." It was apparent that to determine insets an automated system must emulate the human decision process. The human process appears to arrive at a decision by answering the following three questions.

What features are considered when determining density?
What constitutes "too dense?"
What are the bounds of dense areas to determine inset limits?

Operationally there are two ways to execute automated inset determination: by preprocessing the files and storing the information or at the time of production. Each method has advantages and disadvantages. Extraction at time of map production means that the effort is expended for one scale (the scale of the map being produced) and that the extraction must be done every time the map is generated. This method has lower requirements for storing and updating of control information than the extraction in advance method. Extraction in advance allows for tailoring the system for use at many scales is only redone when the TIGER File partition from which the extract has been made is modified.

Since the time it takes the computer to determine the insets depends upon the efficiency of the algorithms, various algorithms were developed and tested. The algorithms developed were specific to the Census Bureau's TIGER File structure. A different file structure will probably yield different efficiencies, but the procedures for inset determination will be similar.

The TIGER File structure is based upon entities known as zero-cells, one-cells and two-cells. A full description of these entities and related files is provided in other papers presented at this conference*. For purposes of discussion, the zero-cells can be considered as the intersection points of features and, therefore, the endpoints of one-cells. One-cells are the lines connecting zero-cells and bounding two-cells. All the other points along the feature between the endpoints are stored in an auxiliary file. The two-cells are the smallest areas bounded by one-cells. Aggregates of two-cells make up higher level geography such as census blocks.

Eight different algorithms were proposed and tried. The eight approaches tested all aspects of the spatial data available from the TIGER file. Several other methods were proposed and immediately discarded as computationally too expensive. For example, computing the average length of all the one-cells necessitates use of the distance formula.

The underlying principle is based upon feature counts within a grid cell. A grid with a known cell size is determined for an area to be mapped. The TIGER File is then read and the number of occurrences for each item within a cell is calculated. The resulting matrix of values is then smoothed by summing the counts for groups of nine cells and recording the results as the value for the center cell of the 3 X 3 group. The smoothing operation removes local irregularity due to the use of a single coordinate to represent a linear and/or areal feature.

---

*See papers by Kinnear, Knott, and Meixler in these proceedings for discussions of the TIGER System.

Algorithms Selected for Consideration

The following algorithms were selected for testing:

1. One-cell, midpoint method. The coordinates of the endpoints of the one-cells are added together and divided by two to get a midpoint.

2. One-cell, average of all points method. The coordinates of all the points along the one-cell and the endpoints are added and the results divided by the count to get an average.

3. One-cell, endpoint method. The zero-cells are used.

4. Two-cell, envelope midpoint method. The maximum and minimum coordinates of the two-cells are added and divided by two to get a midpoint.

5. Two-cell, weighted area centroid method. The area and geographic centroid of the two-cell is determined and the value at the centroid is the area.

6. Census block, envelope midpoint method. The maximum and minimum coordinates for each census block is determined by aggregating the two-cells which constitute the block. The sum is divided by two and the resulting coordinates are used.

7. Census block, two-cell average centroid method. The two-cell average centroid is derived by adding all the maximum and minimum coordinates of the two-cells and dividing by the count to get an average.

8. Census block, weighted area centroid method. The area and geographic centroid of the block is determined and the value at the centroid is the area.

A program was developed for each method and run using a TIGER 87 file partition from the 1987 Test Census. These programs extracted the features from the TIGER File, computed the centroid in latitude/longitude (and the area if required), and stored the results in intermediate files. The points calculated by these programs were analyzed for determination of dense areas. The two methods based upon area calculation were discarded immediately as computationally inefficient when their dot patterns and computer times were compared to the other methods.

Census field operations are primarily concerned with roads as access to the population and as statistical boundaries. For this, our concern is to identify dense areas of road features. The one-cell, endpoint method was discarded because of the difficulties in avoiding multiple counting of the endpoints.

The census block, two-cell average centroid method was discarded because most blocks consist of one two-cell. Where blocks consist of two or more, two-cells, the difference of the centroid from the whole block midpoint position is insignificant. Thus, the extra computation was deemed inefficient.

The four methods retained for the second developmental phase were: the one-cell midpoint; the one-cell, average of all points; the two-cell envelope midpoint and the census block envelope midpoint. The intermediate files for the four methods were processed through a program which produced a grid cell count matrix, i.e. the number of feature centroids that fell within each grid cell. Each grid cell was .25 inches on a side. The number of grid cells along each axis is determined by converting the differences between the minimum and maximum longitude and latitude into inches at map scale and then dividing by the grid cell side size. The map can never be greater than 36 x 36" because of plotter paper size and map design constraints. Consequently the grid never was limited to no more then 144 cells to a side at the selected cell size. The four count matrices were printed and their patterns analyzed.

After the second phase of development, the one-cell average of all points was discarded because there were no apparent differences between the results of this method and those of the one-cell, midpoint method. Then the census block envelope midpoint method was dropped because of the many computations and file accesses required to determine the block envelope. Also the fact that text placement must avoid conflict even with the other two-cell boundary features within the block. Together, these made this method inefficient when compared to the remaining methods.

The one-cell, midpoint is computationally the easiest of the linear feature methods. The two-cell, midpoint method is the easiest of the areal methods, and it is particularly easy when using the TIGER File since the centroid coordinates of the two-cell are stored as part of the base file information. The execution time for both methods is directly proportional to the number of features processed. Thus, the areal method is faster. The execution times are also low because the programs are accessing elementary units of the TIGER File and only performing additions when required, a binary shift for division by two.

Calibrating the Inset Determination Algorithm

The next phase in the development was to introduce human cartographic expertise. Ten professional cartographers, with an average experience of five years in census map making were asked to examine the same maps that the computer programs were producing and to mark on overlays the extent of the insets needed, if any. Their only guidance was that the maps were to show census block numbers. The cartographers marked the overlays without discussion among themselves and without knowledge of what the previous cartographers had marked.

The results showed a surprisingly close match between the cartographers. The variances averaged less than one-fourth inch at map scale. This variance was within the grid cell size and considered quite good for the intended purpose.

Next a smoothed count matrix was plotted out at the scale of the map used by the cartographers. The overlays were placed over the plotted count matrix. A visual examination revealed that the human cartographers placed their inset boundaries so that they enclosed grid cell clusters with counts above about one-half the difference between the maximum and minimum cell counts. This not only worked for large dense areas, but it identified smaller dense areas. Significantly timeframe, it avoided picking up extremely small clusters of one or two census blocks where an inset would be inappropriate.

The numeric value that was developed from the cartographer's efforts to delimit insets was shown to be a function of grid cell size, map scale, and size of text to be plotted. The grid cell size and size of the text to be plotted remained the same throughout development. Only the map scale varied.

Final System Description

The automated map inset determination system developed is based upon the two-cell envelope midpoint method and uses the human expert derived factors. It also processes a map at a time for insets rather than preprocessing a whole TIGER File partition. The preprocessing approach may be reevaluated when TIGER Files become available in sufficient quantity to compare the results. The two-cell envelope midpoint method was selected because the major problem in current field map design is census block number text placement.

The system operates in the following steps:

1. For a whole TIGER File partition, compute the latitude and longitude midpoint of all two-cells that are bounded by at least one one-cell that is a road and store the results in an intermediate file.

2. For all the two-cells within a given map sheet image area convert the latitude and longitude midpoints from the intermediate file into inches at map scale and store these in a temporary file.

3. Process the temporary file through the program that creates the grid cell count matrix and smooths it. Then use the smoothed matrix to determines the need for and limits of insets. For each inset determined, a record is put out to an inset control file. The control file record contains the inset latitude/longitude windows for that map.

Future Research

The current system was developed within a short timeframe imposed by production schedules. While the system performs the task for which it was designed, many questions remain. Will the expert derived numeric value continue to be adequate when the TIGER Files for more geographic areas are available? Is the relationship between map scale, size of text to be plotted, grid cell size and the numeric value linear or does some other relationship exist? Is there a more efficient programming sequence? What is the effect of other map features on the need for an inset and how is it detected? These and other questions are currently being researched, but many areas for study remain.