

Automated Map Insetting in SMIMS

Ming Liu and David Alexander
Geography Division
U.S. Bureau of the Census

ABSTRACT

The Single MIM-based Integrated Mapping System (SMIMS), an automated mapping software system developed and maintained at the Census Bureau, has produced over 16 million maps to support the 2000 decennial Census. In addition to producing all the different type of maps to support field operations, it is used for other high-volume mapping project, such as data exchange programs with state and local governments. This paper and presentation will explore the different strategies that we have employed to automate inset selection and creation from the point of view both of cartographic design and of algorithmic design and mathematical techniques employed. It will include recent research into the application of spatial autocorrelations and other spatial statistics to the problem of the recognition and isolation of dense areas of feature networks as candidates for inset creation.

INTRODUCTION

The Census Bureau's Geography Division has among its responsibilities the continuing development and maintenance of the Bureau's master geographic database and the creation of a wide variety of cartographic products using this data.

The database was developed in the 1980's to support the 1990 Decennial Census and is known as the TIGER (for Topologically Integrated Geographic Encoding and Reference) system.

The cartographic products created based on TIGER include both maps that are prepared to disseminate Census data and information about Census geography to the public and maps that are used internally by Bureau personnel to help data collection operations and in data exchange programs with state and local governments and other partners.

The changes in the organization and conduct of the Decennial Census between 1990 and 2000 placed much greater demands on Geography Division's automated mapping capability. The plans for Census 2000 called for a large increase in the number of different map types in order to tailor the maps to different field procedures in different areas. They also called for being able to produce different maps of the same areas at

different stages of the field operations. The need for higher volume and tight deadlines in turn made it necessary to create a more decentralized system for the production and distribution of maps.

In the mid-1990's, staff in Geography Division started developing a new mapping system to satisfy all the high-volume map production needs of the 2000 Census as well as to produce maps for other programs which need large volumes of maps created. In order to produce large volumes of maps under tight deadlines with limited staff, this system, needed to be completely automated. In order to produce all the different map types needed for the Decennial and other programs and provide maximum flexibility to Census planners in meeting their cartographic needs, it was necessary for this system to allow for the specification of a wide range of cartographic designs without the need for software rewrites. In order to enable decentralization of processing and rapid distribution of maps, it was necessary for the output format to be something which will allow us to print all the different types of maps using widely available printers and plotters.

The automated mapping system that was developed to meet these challenges was called the Single MIM-based Integrated Mapping System (SMIMS). It consists of a Perl script which calls in sequence 44 software modules written in C. Each of the modules carries out a specific task in the creation of a particular map, either in the configuration of the map, e.g. scale determination, sheet definition, or inset determination, or in the creation of a layer of symbology for one sheet, e.g. linear features, political boundaries or areal or point landmarks.

The flexibility to create different types of maps without rewriting the software comes from the use of map definition tables. Each time SMIMS makes a map; it reads all the information needed to make the right kind of map from a set of files, which contain the complete specification of the map's design. These definition tables include such information as the size of the sheets, whether they are to be color or black-and-white, whether they are to be single or multi-sheeted and whether insets are to be allowed. When the modules are run to determine scale and the location and numbers of insets, definition tables are read which contain parameters that relate feature densities to scale and set thresholds for dense areas to become insets. When modules are being run to symbolize features, the definition tables are read to determine how features of different types are to be symbolized.

The development of SMIMS was closely bound to the development of the Map Image Metafile (MIM), a complete hardware-independent map description format. Every SMIMS module, which placed any symbol on a map, did so by creating output in MIM format. The final output of a SMIMS run was a set of MIMs, one per sheet, which could be converted to any printer or plotter language. This made it possible for SMIMS to have

a single output mode which could be distributed to field offices with Postscript printers or HP plotters and which could be put out on any device without rewriting SMIMS for only the trouble of writing a new converter routine. This solved some problems that came out of the 1990 mapping efforts in which the output of the mapping software was too tightly bound to a particular output device.

Having provided a general introduction to the cartographic challenges that SMIMS has addressed, our attention will now be focused on the particular challenges of automating the delineation of inset areas for a wide variety of types of maps. The problem has common features across the different map types: how to show the densest areas of a map's subject area at a sufficiently detailed scale without being forced to map all the less-dense areas at that same detailed scale, which would produce an impractically large number of sheets. SMIMS has developed multiple methods of choosing inset areas and placed them at the disposal of the cartographers who design our maps, but has continued to look for better ways that can come closer to capturing the way a human cartographer could choose insets, given the luxury of being able to consider each entity to be mapped.

INSETTING

The current implementation of automated inseting performs selection and creation from the point of view of cartographic design. Specialized capabilities include creating insets of whole geographic areas regardless of density; inseting by densities of features such as housing units, linear features, and polygons, looking for areas where the density within arbitrary windows exceeds a specified threshold; and inseting by geography and density where areas of dense features are identified and expanded to include whole geographic entities. These are among the options that can be specified in the map design by means of the cartographers specifying keywords and parameter values in the inseting definition table for the particular mapping project.

If the cartographer desires insets based directly on the density of some type of feature, the THRESHOLD key word is used to specify a minimum number of features that a cell must have in order to qualify as a potential inset. Other parameters allow the cartographer to set a smallest distance between inset windows allowed before inset windows are combined, a minimum size a potential inset must be in order to qualify as an actual inset, and a maximum distance between two potential insets for merging.

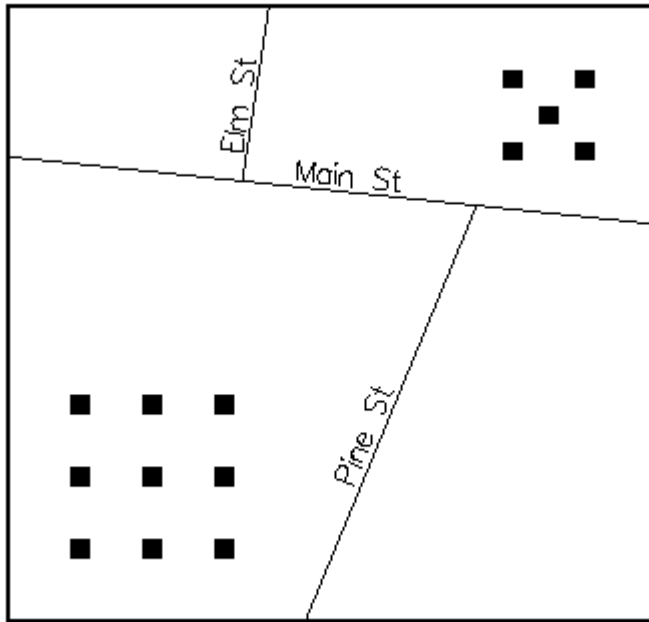
One problem that has been observed with the method that has been used to identify dense areas for inseting is that it relies on an arbitrary partitioning of the area to be mapped into cells to calculate densities. It is not unusual for an area with a dense

concentration of features to be split between two or sometimes even four different cells. This dilutes the density observed in each of the cells and may prevent any of the cells from being recognized as having a density above the threshold. Cartographers can try to compensate for these occurrences by lowering the threshold, but this leads to situations where unnecessary insets are made.

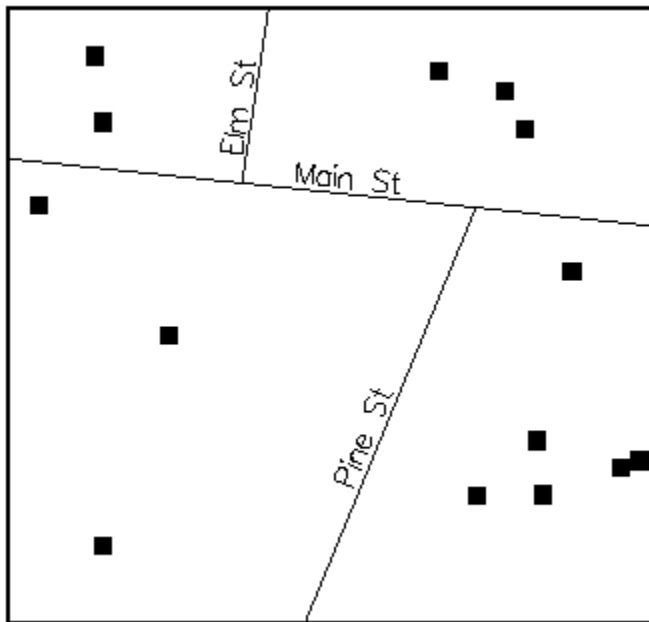
The next implementation of insetting will employ more sophisticated algorithmic design and mathematical techniques. It will include recent research into the application of spatial autocorrelations. This method will be provided to the cartographers as another option in addition to the currently available density methods. Instead of simply using THRESHOLD to identify arbitrarily defined cells with high densities, we can calculate some type of autocorrelation statistic that will provide information about the tendency of map features to bunch together into clusters.

One option for this implementation is to calculate a cross-product statistic, such as Moran's I , which is based upon the distances between every pair of map features. Such a statistic can be calculated and used to distinguish, given a degree of confidence between occurrences of positive autocorrelation, where features tend to bunch together, negative autocorrelation, where they tend to space themselves apart from each other, and random distribution. The principal concern of an insetting algorithm is to respond to high degrees of autocorrelation by the creation of insets in the areas where the map features are bunched together and so threshold values of positive autocorrelation will have to be developed by cartographers which provide a useful test for determining when a map may benefit from the creation of insets.

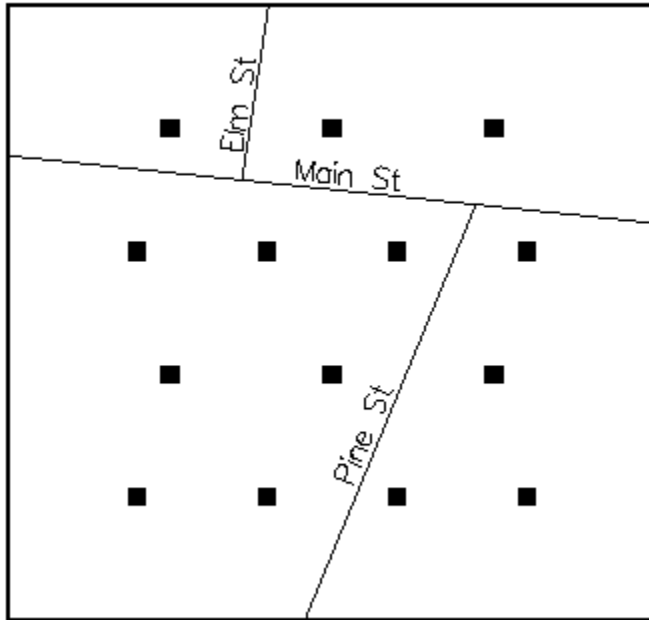
The three maps of different sets of point features below illustrate the phenomena that spatial autocorrelation measures can be used to classify.



This map shows positive autocorrelation, as the point features are clustered together in two clusters, in the lower-left and upper right corners.



This map shows a random distribution of the same number of point features. An autocorrelation statistic should make it possible to conclude, with some measure of confidence, that these point features have neither a tendency to cluster together nor to separate themselves, but are located randomly and independently.



This third map shows the same number of point features displaying negative spatial autocorrelation. The distribution of distances between pairs of features is skewed toward larger values than can be expected from random, independent distribution. Instead of features tending to be located close together, they tend to be located as far apart as possible. This represents a situation which may or may not occur in practice, but which requires no inseting remedy if it is detected, since mapping such an area at a uniform scale without insets will result in an efficient map.

The Moran I statistic, a measure of spatial autocorrelation, can be employed to estimate the correlation of the locations of different types of map features. The Moran's I coefficient is computed as

$$I = \frac{N}{S_0} \frac{\sum_{i=1}^N \sum_{j=1}^N W(i, j)(z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^N (z_i - \bar{z})^2}$$

with

$$S_0 = \sum_{i=1}^N \sum_{j=1}^N W(i, j).$$

One possible application is in the measurement of the densities of polygons to be displayed on the map. In such an application, N would be the number of polygons, the z_i would be the areas of polygon i , \bar{z} is the average of the $\{z_i\}$ over the N polygons. The numerator is a measure of covariance among $\{z_i\}$ and the denominator is a measure of variance. S_0 indicates the summations of the elements $\{W(i, j)\}$ in the weighting matrix W .

Another possible application might be in the measurement of the densities of point features, such as housing units. This problem can be transformed into the equivalent of the problem involving the densities of polygons by the creation of Voronoi polygons surrounding each of the point features. The areas of these Voronoi polygons can then be substituted for the areas of the polygons to be mapped and the rest of the calculation proceeds the same way.

The matrix W is the heart of this statistic. We can choose $W(i, j)$ to be any simple function of the distance between the centroid of polygon i and the centroid of polygon j . Therefore we are able to consider the problem of the recognition and isolation of dense areas of feature networks as candidates for inset creation. Moran's I varies from -1 to 1 . If $I > 0$, it means spatial clustering; random if I closes to 0 and uniform $I < 0$. Use of this method would be specified in the inseting definition table with the key word AUTOCOR, set to a threshold value between 0 and -1 .

CONCLUSION AND OUTLOOK

Comparing the by density/THRESHOLD approach to the by density/AUTOCOR approach for inseting, we note that the former has been used for many years and can meet the needs of most mapping projects, including the 2000 Decennial Census. However, we still experience feature congestion sometimes. Using the autocorrelation approach should resolve those few problems. Since there are numerous spatial autocorrelation methods being used in GIS and a great amount of calculation could be involved, it is very important for us to choose the right method and implement a practical algorithm for it. Moran's I has shown promising results, though we are still researching it. Nevertheless, we expect it to be fully implemented very soon.