Bureau of the Census
Statistical Research Division Report Series
SRD Research Report Number: Census/SRD/RR-85/11

COMPARISON AND CONSOLIDATION OF DIGITAL DATABASES USING INTERACTIVE COMPUTER GRAPHICS

by

Alan Saalfeld
Statistical Research Division
Bureau of the Census

Recommended by: Tom O'Reagan

Report completed: November 7, 1985

Report issued: November 7, 1985

Comparison and Consolidation of Digital Databases Using Interactive Computer Graphics

Alan Saalfeld
Statistical Research Division
Bureau of the Census
Washington, DC 20233
(301) 763-7530

ABSTRACT

Over the past decade a considerable amount of interactive graphics software has been developed to maintain digital map files. Maintenance generally involves updating or changing feature representations, and the software used for this purpose usually focuses on manual adjustment and screen image verification of a change to a local region of a single map. Powerful interactive color graphics tools have made more sophisticated types of editing possible.

Map compilation, or conflation, may be accomplished by simultaneously manipulating the graphic images of two different digital maps of the same region in order to recognize matched features and differences more easily. Fast computer programs for rubber-sheeting one or both of the maps permit a user of a computer graphics workstation to align the maps in stages through successive approximation methods. After the maps have been aligned, the user may readily identify similar features on the two maps by overlaying one feature on the other. Features of one map which have no counterpart on the other map stand out; they become highlighted in isolation when different colors are used for the two maps.

With some limitations, the computer may be programmed to recognize matches. Mathematical relations of geometric position may be used to test for matches; and, when the tests are satisfied, corresponding features may be flagged automatically as matches. Similarly, features which have no counterpart on the other map may be identified by mathematical tests which measure their isolation; and they may be flagged automatically as unmatchable features. Displaying the matched, unmatched, and unmatchable features in different colors as they are identified permits further operator interaction in the compilation process.

Introduction: Motivation for the Conflation Experiment

Computerized map matching theory, or map compilation, or map ***conflation***, began at the Census Bureau in 1984 as an experiment. The Census Bureau's interest in implementing some type of conflation system arose when the United States Geological Survey agreed to provide raster-scanned line files from their 1:100,000 scale cartographic data base to be compared and combined with the Census Bureau's own DIME files. The Census Bureau's Statistical Research Division began developing and implementing new triangulation and rubber-sheeting techniques on high-resolution color graphics terminals to initiate the experiment.

Scope and Objectives of the Joint USGS/Census Project

Information transfer between files is an important goal. The Geological Survey has made their digital files available to the Census Bureau so that the Bureau will tag the features (assign attributes, such as name and address) in exchange for the use of the files. Adding names and address ranges requires a preliminary step of matching enormous numbers of features in a ***consistent*** and ***comprehensive*** manner. That preliminary mass matching step is one of the products of the present conflation package being developed and used at the Census Bureau.

The Census Bureau also wants to compare the two digital representations of the same area in order to find discrepancies that will permit the Bureau to focus on possible recent housing development and on map drawing errors of one file or the other.

Alignment of the maps is the key tool to recognizing matches and differences of the two maps; and it is accomplished through rubber-sheeting. Rubber-sheeting involves first moving some of the Census map's points to align them exactly with corresponding points on the USGS map. Then the remaining Census map points are moved in a proportional fashion so that the relative positions of all of the Census map's points are not changed. The Census map is shifted and stretched to fit the USGS map.

During the rubber-sheeting of the Census map, the USGS map does not move. One conceivably could move both maps some average distance to bring them together at points that can be matched; however, the matching of features would not be improved and both maps would lose their original shapes.

The Census Bureau's specific objectives for the conflation system may be summarized as follows:

1. Adjust coordinates on the Census Bureau's Geographic Base File by exactly adopting the USGS coordinates wherever possible.

2. Establish match flags and pointers between Census map files and USGS map files for all line and point features that can be paired on the two files.

3. Establish non-match flags for those line and point features which appear on one map and not the other. Classify non-matches into categories for future resolution.

4. Create maps which highlight matches and non-matches in a manner that facilitates secondary verification of controversial features.

5. Transfer name and address attribute information from GBF/DIME records to matched USGS records.

Even with the new coordinates, the Census maps will maintain their topological integrity. The rubber-sheeting process will not alter relative positions of map features. If a Census map requires correction of topological relations, that adjustment must occur elsewhere.

Rubber-sheeting of the Census maps is accomplished by computationally fast piecewise-linear local transformations defined on a triangulation of the map space. Coordinates of 0-cells are recomputed and 1-cells are redrawn. Rubber-sheeting is applied iteratively to increase the number of map points aligned at each step. Each rubber-sheeting step leaves Census map features in the same relative positions in the topological sense.

Evolution of the Computerized Alignment System

The iterative process of map alignment was initially designed to be highly manual because of the anticipated difficulty in deciding what to match. An operator was expected to apply complex decision rules to a large variety of situations. The initial design called for real-time operator interaction with a moving color graphics screen display. That design produced a video game type of situation. The similarity to video games was further exploited by adding system features which rewarded or reinforced correct choices by the operator via the screen display. The operator was challenged to move quickly and accurately and could see concrete results of moves made.

Recent experience with real maps points to the feasibility of a more fully automated approach to matching, and such an approach is underway. The situations and decisions facing an operator are not as complex in general as was first imagined. The initial fears that a fully automated rubber-sheeting system would completely distort the maps have been allayed. A stepwise iterative alignment procedure now matches a few points based on very strict criteria on the first pass, then relaxes match criteria (and adds criteria dependent on prior matches) with each successive iteration of the matching operation. A first-pass match of a few points distributed throughout the map gives, in general, a very good initial alignment of the maps, and after an initial alignment of maps is accomplished, future distortions may be avoided entirely by simply not allowing large point movements.

The Conflation Project has fostered the development of several new mathematical algorithms and computerized techniques. These new tools are grouped into (1) feature match routines, (2) triangulation routines, and (3) rubber-sheeting routines. The routines are described below after an initial system design section explains the interactions of the routines in the full conflation system.

Fundamental Structure of the Conflation System

The Census Bureau's Conflation System employs an iterative approach to map matching. That iterative approach may be outlined as follows:

One begins by defining a sequence of test criteria to be applied successively to determine feature matches:

MATCH-CRITERIA(1), MATCH-CRITERIA(2),...,MATCH-CRITERIA(K),

The programs implementing these criteria receive, as input, a pair of maps, possibly already partially matched; and they return, as output, sets of pairs of matched features and lists of unmatchable features for each map. The conflation system matching operation at the Bureau of the Census can be diagrammed succinctly as the following *K*-loop:

```
REPEAT FOR J = 1 TO K

    APPLY MATCH-CRITERIA(J)
      (RETURNS (MATCHES(J), UNMATCHABLES(J))
    TRIANGULATE ON MATCHES(1)...MATCHES(J)
    RUBBER-SHEET ON TRIANGULATION

END LOOP
```

Feature Matching Routines and Algorithms

The match criteria ***MATCH-CRITERIA(J)*** need not all be different. Each rubber-sheeting operation will improve total alignment of the maps; and match criteria based on good alignment may fail in one iteration and succeed in the next. Some match criteria in use now are based on previous successful matches; and these match criteria will undoubtedly give different results when applied at different iteration points. One simple iteration rule is to apply a set of match criteria until they give no new matches before moving on to the next different set of match criteria.

All of the match criteria developed and tested at the Bureau of the Census contain a component based on nearness or proximity of features (see reference [4]). Nearness may refer to usual Euclidean distance or to pseudo-distances along a graph or network. Successively improved alignment through rubber-sheeting will progressively adjust Euclidean distances without changing topological pseudo-distances of the graph or network.

Feature matching initially refers to 0-cell matching, (or intersection matching, or point matching). Intersection matching is the easiest to define and the easiest to implement in a fashion that permits topologically consistent extension to the whole space. Matched points serve as vertices for triangulations of the space; and the triangulations permit each triangular region of one map to be associated with a corresponding region of the other map.

Matched 1-cells arise as induced matches when the corresponding end points of the 1-cells have been matched as 0-cell matches. 2-cell matches are induced by matches along their boundaries, namely the 1-cells and 0-cells.

The 0-cells are matched using three primary tools:

(1) ***Nearest Neighbor Pairings***. The 0-cells on one map are compared with the 0-cells of the other map after an initial alignment of the maps is made; and nearest neighbor pairs are identified. A pair of points is a nearest neighbor pair if the points are from different maps and each is the closest point from the other map to its paired point. Nearest neighbor pairs are found by applying a nearest point algorithm to create a sequence of points which alternates from one map to the other. Such a sequence stabilizes rapidly in a nearest neighbor pair. The nearest point algorithm uses a list ordering of the points called a Peano-key ordering which is based on interlacing the digits of a binary representation of the vertical and horizontal coordinates. The Peano-key list

ordering is accessed through a B-tree because the list is updated throughout the conflation process.

Nearest neighbor pairs are candidates for matching if other match criteria tests are also met. Several of the other tests utilize the following integer measure of local configuration:

(2) ***The Spider Function of a 0-cell***. The ray pattern at a 0-cell (that is, the emanating 1-cells) has infinitely many possibilities of rays and ray directions. In order to simplify the possibilities, the number of directions was reduced to 16 and later 8 sectors. The eight sectors finally decided upon correspond to 45° pie slices in the principal directions of north, northeast, east, southeast, south, southwest, west, and northwest. The ray pattern is assumed to have at most one ray in each of the eight sectors (more than one ray in any sector will alter the spider function representation and reduce the chances for making a match--it will not lead to false matches). The eight sectors in clockwise order are assigned consecutive bit positions (from right to left) in an 8-bit binary number, and the bit for a given sector is turned on if and only if there is a ray in that sector. The resulting number has been descriptively named the spider function of the 0-cell. With this function, an integer between 0 and 2^8-1 describes the ray pattern of the 0-cell. The binary number 01010101 (which is the decimal 85) represents the typical 4-street north-south-east-west intersection, for example. Intersection patterns which differ by a power of two are "close" in one of two geometric senses: either one pattern is missing a single street, but agrees everywhere else; or else one street is shifted, off by a single sector. By comparing the ***index*** of a 0-cell as well as the spider function (the index is the number of emanating rays), the Bureau developed several simple measures of nearness of configuration.

(3) ***Dependent Matching Routines***. Dependent matching routines are rules for applying relaxed matching criteria to 0-cells which are adjacent to already matched 0-cells. These routines use the 1-cell network of the map to precipitate additional matches. For that reason, it is important to store the topological relations of the map in order to facilitate network or graph traversals.

After each matching phase, matched points are moved into perfect alignment; and other points of the map are assigned to triangles whose vertices are the matched points. All of the unmatched points are then moved according to the movements of the vertices of the triangles which contain them.

Triangulation Routines and Algorithms

For the conflation system developed at the Bureau of the Census, the rubber-sheeting transformations are defined on triangular subregions of each map; and the triangular subregions must cover each map and not overlap each other (except to share an edge). A polygonal region such as a map area may be subdivided into non-overlapping triangles. Such a subdivision into triangles is called a triangulation of the region. One type of triangulation, called the ***Delaunay triangulation***, exhibits special properties which make it particularly useful for the conflation system. A good presentation of fundamental properties of the Delaunay triangulation may be found in the referenced paper by Lee and Schachter [2].

The following tools were developed to deal with triangulation requirements of conflation:

(1) ***The Triangle-based Point Directory***. Since all 0-cells and other point features need to be transformed according to the triangle that contains them at any particular iteration, and since the triangles are changing with each iteration in a local fashion (in other words, not all of the triangles change), one may keep track of triangle containment by maintaining a triangle-based point directory and updating it with each triangulation update. A triangle-based point directory eliminates the need for determining the containing triangle of each point using some point-in-polygon test. As the number of triangles increases, the rubber-sheeting routines do not require additional computer time to search for the containing triangle if the triangle-based point directory is used.

(2) ***The Delaunay Triangulation Iterative Building Routines: Add-a-Vertex***. The Delaunay triangulation on a set of vertices may be built by adding one vertex at a time and re-triangulating after each addition. Each re-triangulation involves changing a few of the triangles in the neighborhood of the added point based on the following result (see reference [5]):

If a new vertex is added to a Delaunay triangulation, then the set of new edges in the updated Delaunay triangulation is precisely the set of segments linking the new vertex to all three vertices of every triangle whose circumcircle contains the new vertex.

(3) ***Triangulation Extendability Results***. In order for the rubber-sheeting routines to be applied, the one-to-one correspondence of matched point pairs must be able to be extended to a one-to-one transformation of the entire triangulated spaces. Tests for extendability of the one-to-one transformations are applied; and when the tests fail, the vertex set must be modified.

(4) ***The Delaunay Triangulation Modification Routines: Delete-a-Vertex***. The Delaunay triangulation may require modification in the form of vertex deletion due to non-extendability of the corresponding triangulation in the other space. A delete-a-vertex routine (see [7]) provides a triangulation update based on re-triangulating the star-shaped polygon containing the vertex to be deleted.

After a map is triangulated and the image of the triangles is tested for one-to-one-ness, a transformation is applied which moves every point of the map into a new position and aligns those points which have been flagged as matches.

Rubber-Sheeting Routines and Algorithms

In their paper on rubber-sheeting, Marvin White and Pat Griffin [1] describe a ***local affine transformation*** of maps which sends triangles into triangles in such a way that shared edges of neighboring triangles are transformed in a consistent manner. The Census Bureau has adopted the rubber-sheeting transformation of White and Griffin, applied it to their own Delaunay triangulations, and expressed it in terms of simplicial coordinates for greater ease of manipulation and speed of computation.

The primary tool developed for rubber-sheeting is a computational algorithm for finding ***simplicial coordinates*** using linear equations for triangle edges (see Saalfeld [6]). The same computation serves to test for triangle containment of a point, since a point lies in a triangle if and only if its simplicial coordinates are between 0 and 1.

Local affine transformations preserve linearity and parallelism on each triangle; however, these continuous transformations are not differentiable at the edges of triangles. Because the rubber-sheeting transformations are applied only to discrete points (0-cells and shape points), however, there is no need to preserve differentiability everywhere. After the 0-cells and shape points have been transformed, the connecting lines or 1-cells are redrawn to complete the map image.

Other non-affine rubber-sheeting methods have been considered and rejected because of greater computational complexity and topological inconsistencies.

Conclusion

A prototype system for compiling two digital maps interactively at a graphics terminal is working well and promises new opportunities for merging and comparing two digital source maps. Current research is focusing on refinements of the system such as local data access methods which allow better data handling for local rubber-sheeting operations. The experiment at the Bureau of the Census has shown that the mathematical algorithms and the computer technology are sufficiently developed to offer a good solution to the problem of automated map conflation.

References

1. Griffin, P., and M. White, 1985, "Piecewise Linear Rubber-sheet Map Transformations," The American Cartographer.

2. Lee, D.T., and B.J. Schachter, 1980, "Two Algorithms for Constructing a Delaunay Triangulation," International Journal of Computer and Information Sciences, 9(3), 219-242.

3. Lynch, M. P., and A. Saalfeld, 1985, "Conflation: Automated Map Compilation--A Video Game Approach," AUTOCARTO 7 Proceedings.

4. Rosen, B., and A. Saalfeld, 1985, "Matching Criteria for Automatic Alignment," AUTOCARTO 7 Proceedings.

5. Saalfeld, A., 1985, "Direct Computation of Delaunay Triangulations," Internal Census Bureau Document.

6. Saalfeld, A., 1985, "A Fast Rubber-Sheeting Transformation Using Simplicial Coordinates," The American Cartographer.

7. Saalfeld, A., 1985, "Updating Delaunay Triangulations After Point Removal," Internal Census Bureau Document.

8. U. S. Geological Survey/Bureau of the Census, 1983, "Memo of Understanding for the Development of a National 1:100,000 Scale Digital Cartographic Data Base," Washington, DC.

9. White, M., 1981, "The Theory of Geographical Data Conflation," Internal Census Bureau draft document.

Appendix 1: Conflation Map Facts for Non-Cartographers

United States Geological Survey Maps

The USGS is providing 4 different overlays of each of the 57,000 $7^{1}/_{2}$ minute quadrangles which comprise the nearly 2,000 1:100,000 map sheets for 100% coverage of the land area of the U.S. (over 3,000,000 square miles). The 4 overlays of a single region will be integrated into a single map by the Census Bureau's Geography Division.

The USGS map files were produced by machine-scanning (raster scanning followed by vector conversion) of hand-detailed line maps. The hand-detailed line maps were drawn at the 1:100,000 scale. At that scale, 1 inch equals approximately 0.8 miles. Scale in this case provides an indication of the degree of detail present on a digitized map, since a digitized map can be redrawn at any scale.

Each 1:100,000 map sheet (1° longitude by $^{1}/_{2}$° latitude, each covering approximately 1600 sq.miles) consists of 32 $7^{1}/_{2}$ minute quadrangles.

In metropolitan areas, each $7^{1}/_{2}$ minute quadrangle is divided into two Metropolitan Map Sheets (MMS's).

For different latitudes, the length of the angular units of latitude and longitude vary. For latitudes in the continental U.S., each minute of latitude or longitude (one sixtieth of a degree) is (on the average) approximately one mile. Equating 1 mile to 1 minute is a good rule-of-thumb approximation. A $7^{1}/_{2}$ minute quad (short for quadrangle) is a square of approximately $7^{1}/_{2}$ miles to the side.

A $7^{1}/_{2}$ minute quad contains approximately 50 square miles. A MMS contains approximately 25 square miles.

Census GBF/DIME Coverage

Census GBF/DIME coverage exists in 278 Metropolitan Statistical Areas (MSA's). GBF/DIME coverage corresponds to approximately 5% of the land area and 60% of the population of the U.S.

GBF/DIME coverage in those 278 metropolitan areas is not complete. GBF/DIME coverage in those areas intersects, but does not necessarily cover, approximately 5,500 MMS's.

The average number (mean) of MMS's per MSA is just under 20 and the median is 12. The range of MMS's is 2 to 98.

The GBF/DIME files contain one hand-digitized record per 1-cell or line feature segment. The range of MSA file size is from 3,000 to 226,000 records.

Appendix 2: Processing Considerations

Total processing requirements

With a conservatively low estimate of 2,000 segment records per MMS, total GBF/DIME coverage involves 11,000,000 1-cell records. For a typical urban configuration, this coverage requires 6,000,000 0-cells (segment intersections), and 5,000,000 2-cells (usually city blocks). Matching at the 0-cell level therefore requires 6,000,000 classifications as matches or non-matches be made.

Manual system limitations

A manual system averaging one classification per operator per minute would require 100,000 person-hours or 12,500 person-days or 60 person-years to process the maps.

An area of 1 mile square is the largest area that an operator can review comfortably on the screen at one time. For a larger area, the operator cannot easily distinguish the individual 0-cells which lie close to each other. GBF/DIME coverage includes approximately 140,000 square miles.

Semi-automatic system gains

A semi-automatic system capable of making 90% of the decisions automatically in batch, thereby providing excellent alignment to facilitate the final 10% of the classifications, would reduce the manual processing phase to an estimated 10,000 person-hours or 6 person-years. Gains in consistency of case-handling and error-free processing, while not quantifiable here, are an expected product of automating the major part of the map conflation.

Semi-automatic processing, followed by expert review, will be used to correct errors by resubmitting jobs for processing with additional manually entered constraints which prevent repeating errors made in the first job run. In this way, batch processing to insure consistent case-handling for the large majority of decision-making will be maintained.

Additional Considerations

Exact alignment of maps embeds the GBF/DIME coverage area of a map in the USGS MMS, showing clearly where GBF/DIME coverage extension is required. A coverage extension is planned as an activity to be contracted out to some mapping firm. If coverage extension is done without alignment, there is a greater likelihood of mismatching, double counting, and omissions.