BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION REPORT SERIES
SRD Research Report Number: Census/SRD/RR-87/26

Statistical Analysis of Map Differences

by

Alan Saalfeld
Statistical Research Division
Bureau of the Census

Recommended by: Tom O'Reagan

Report completed: September, 1986

Report issued: October 1, 1987

Statistical Analysis of Map Differences

Alan Saalfeld, Bureau of the Census

ABSTRACT

Digital map data lends itself to computerized statistical analysis much like any other computer-readable data file. Comparative data analysis is possible when two files are present and linkages can be established between some of the feature records of the two files. Modern digital map files contain various measures of spatial relations, including adjacency relations, network patterns, and measures of position and distance. This paper examines those measures and their interactions and proposes some statistical tools for automating feature matching procedures.

INTRODUCTION

Conflation is the consolidation or merging of two map representations of the same region into a third composite conflated map. Recently the Bureau of the Census has begun consolidating or conflating pairs of digital map files of the same region in order to measure and improve the quality of the Bureau's digital maps. A second set of digital maps for the entire country is being provided by the United States Geological Survey for the Bureau of the Census to use with its own DIME files for comparative updating of both sets of maps.

In the past, measures of similarity and differences of maps, primarily of paper maps, were not quantitative or even fully quantifiable; and this limitation made the comparative analysis of maps quite subjective and non-numerical. Often differences and discrepancies were merely enumerated or listed; and there was no readily understood measure of map similarity or sameness. The digital map file, on the other hand, is considerably more amenable to numerical analysis, and its format invites computerization of that analysis.

Now it is not only feasible and informative to quantify and analyze map similarities and differences; it is also useful to find and use concrete numerical measures of similarity to establish statistical evidence that two maps, or parts of the two maps, are the same.

This paper presents some initial attempts at quantifying map similarities and differences. The paper outlines approaches to analysis of those differences and similarities; it does not contain extensive empirical justification for those approaches. This last shortcoming is due primarily to the limitations of available data. Although the Bureau of the Census will eventually have to process a large number of map pairs (over 5,500 7.5 minute quadrangles), only three digital data pairs were made accessible to the research unit; and all three of those digital map pairs required considerable editing before they could be compared. Problems such as duplicate data records and special data formats had to be addressed before map feature comparisons could begin. Effective data classification methods are still evolving as a result of initial comparison attempts.

A HEURISTIC APPROACH TO AUTOMATED FEATURE MATCHING

A cartographer, in order to compile two maps of the same region and produce a third new map, uses numerous visual clues and cues to match features of one map to features of the other; and, convinced of a match, he extracts a single feature from the two maps. After a cartographer has matched features on the two maps, a statistical analysis of the numerical properties of the matched and unmatched features may be performed. The resulting analysis yields information on the numerical characteristics of the cartographer's matching operation or matching algorithm. The resulting analysis, in turn, may be used to drive an automatic statistical matching procedure, which can then replicate the cartographer's results and, thus, more fully automate the map conflation process.

Because of the need for uniform processing and also because of the large number of map files to be processed, the final production system for computerized matching and merging of two map files should be as fully automated as possible. Nevertheless, in order to assess various rules for matching, a semi-automatic system has been implemented. A computer operator uses the semi-automatic system to simulate the decision-making of a cartographer who has been asked to match as many features of the two maps as possible. A computer operator uses a color graphics workstation to select and view matches of street intersections and matches of street segments on the pair of maps.

The system is semi-automatic in that it has been programmed to test matching criteria and to prompt the operator. Thus the operator needs only to verify or affirm proposed matches. The possible matches are proposed by the machine based on various pre-programmed criteria selected by the operator. Those criteria involve positional and relational characteristics of the map features being matched. Color screen displays have facilitated decision-making procedures; and interactive rubber-sheeting algorithms have realigned the maps, thereby permitting very effective immediate visual verification of matching decisions. The most valuable element of the color graphics/alignment approach has been the ease and accuracy of assessing whether or not a match was made correctly. An iterative matching/alignment procedure brings matchable pairs closer and closer together and moves pairs which do not match farther and farther apart. Our initial analysis shows that these relative distances are significantly different and useful in determining matches.

INTERMAP AND INTRAMAP DIFFERENCES

This study of map similarities and differences focuses on street intersections and their configuration and position. In order to reduce the data storage and computational requirements,

the intersection patterns were classified by type, the types were assigned numerical code values, and the distributions of various types of intersections in the plane were examined. The coding of types attempted to reflect similarities of types through similar code values; however, the coding scheme requires additional review, especially with respect to nearness of one type to another. As one would expect, the intersections are not clustered in space, but are generally fairly evenly distributed in the plane.

The average distance from any intersection to its nearest neighbor intersection on the same map is large compared to the average amount of distortion between maps; and this fact makes an iterative alignment approach very effective.

The following figures show that a good initial alignment can bring nearly all matchable pairs into proximity in such a way that the proximity relation almost becomes a necessary (but not sufficient) condition for matchability.

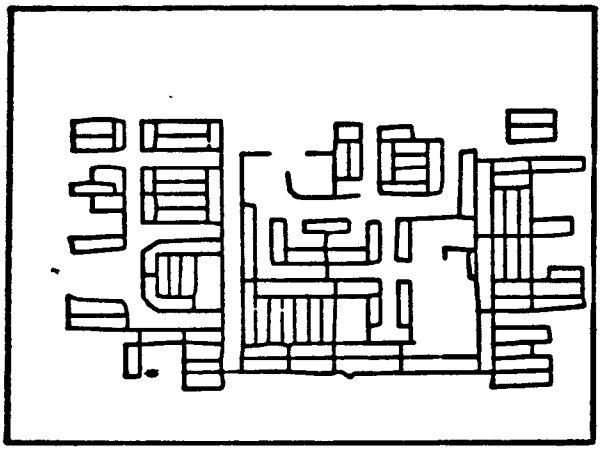

Figure 1A. USGS Map of Part of Fort Myers, Florida

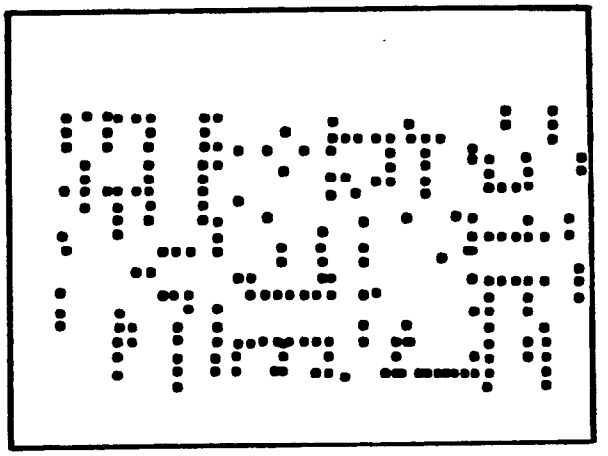

Figure 1B. Street Intersection Configuration for Same Map

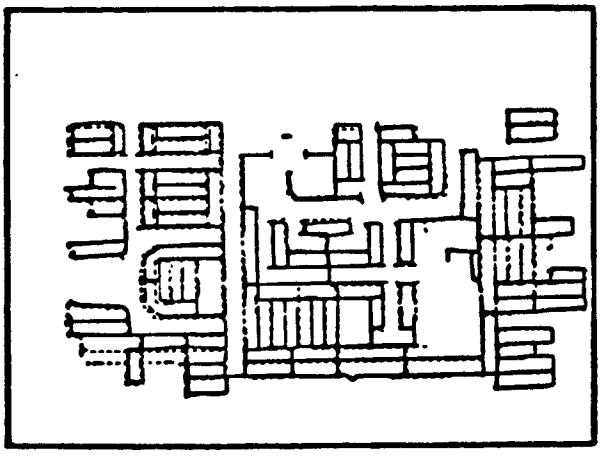

Figure 2A. Overlay of Two Maps of Fort Myers, Florida

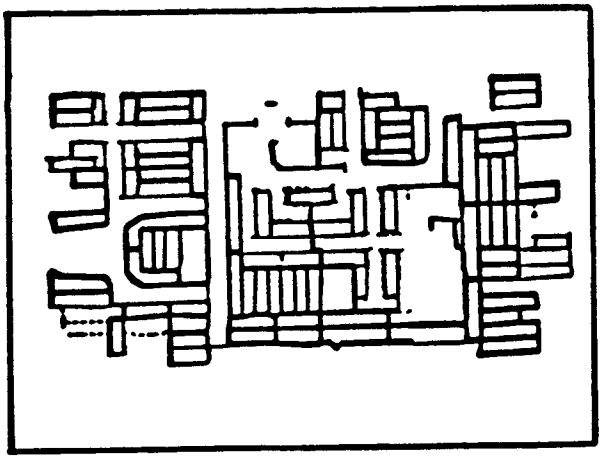

Figure 2B. Matched and Aligned Maps of the Same Area

for one particular test map of Fort Myers, FL, the efficacy of an alignment approach to matching can be illustrated in the following summary statistical diagrams. In general, two types of false classifications may occur. A map feature may be labelled incorrectly as having a match when indeed it does not (false positive); or a feature may be judged incorrectly not to match any feature when it has a true match (false negative). Because an iterative matching procedure identifies new matches at each stage, and, in general, does not flag non-matches as such until the final stage, false negatives may be corrected at an intermediate stage. False positive errors are less desirable and less managable than false negatives because at no point in the iteration procedure is there an automatic un-match capability for correcting false positives.

The distance between potential matches after initial alignment is an excellent tool for controlling both types of errors. Distance may be used for predicting both types of error and reducing one or the other: in the Fort Myers map, for example, if the threshold for matching is set at 20 meters, (that is, no matches are accepted unless the candidate pair are within 20 meters), then the measured probability of making a false negative error is 11%, and the probability of a false positive is 1%. By decreasing the threshold, false positives may be reduced further; however the increase in false negatives will require additional iterations of the file processing; and the threshold may even need to be relaxed in the final iterations in order to detect all matches.

The cumulative relative histograms below summarize the distance from a matchable point to its match and the distance from a matchable point to the nearest non-match on the other map.

Table 1 shows the summary statistics for the distances between nearest neighbors and for the distances between corresponding points on the two Fort Myers maps.

Frequency Distribution of Distances

Distance Range (in meters)	Number within Distance of Matching Point	Number of Points within Distance of Nearest Non-matching Point
0 to 5	162	-
5 to 10	359	-
10 to 15	272	4
15 to 20	132	8
20 to 25	70	14
25 to 30	19	25
30 to 40	13	54
40 to 50	3	90
50 to 60	2	227
60 to 70	-	302
70 to 80	1	134
80 to 100	-	86
100 to 200	1	82
200 to 400	-	8
400 and above	-	-

	Mean distance	Range	Std Dev.
To Matching Point	11.45	112.25	7.75
To Nearest Nonmatch	66.68	278.89	28.55

Table 1. Comparison of Distances to Matches and Nonmatches (After Initial Alignment)

The distance measures above show that nearest neighbor pairs are excellent candidates for matching if other match criteria tests are also met.

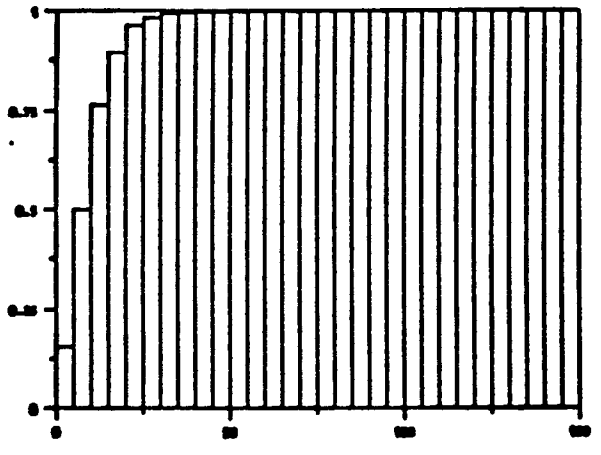

Figure 3A. Fraction of Matchable Points whose Matching Point is Within the Indicated Distance of the Point

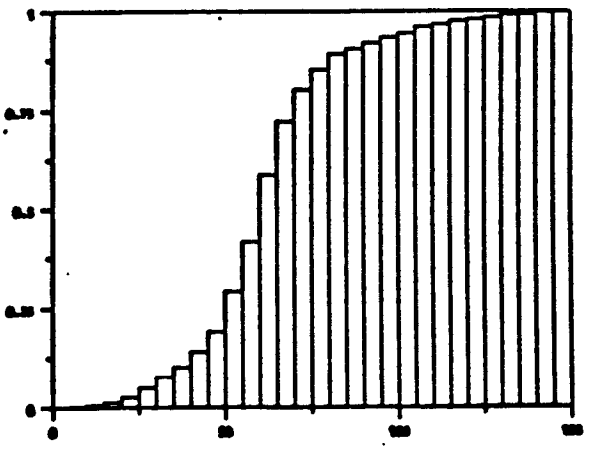

Figure 3B. Fraction of Matchable Points whose Nearest Non-Matching Point is Within the Indicated Distance of the Point

The remainder of the analysis in this paper focuses on other tests which utilize the following integer measures of local configuration:

(1) *The Index of a 0-cell (intersection)*. The number of streets emanating from an intersection is called the *index* of the 0-cell or intersection. The index provides a good measure on which to match intersections if it is unique or locally unique (e.g. the only intersection in the neighborhood with seven streets coming into it.)

(2) *The Spider Function of a 0-cell*. The street pattern at a 0-cell (that is, the emanating 1-cells) has infinitely many possibilities for street 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 street pattern is assumed to have at most one street in each of the eight sectors (more than one street in any sector will not be reflected in the spider function representation, but will reduce the chances for making a match. However, it will not lead to false matches). The eight sectors in counter-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 street 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 intersection 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 Bureau of the Census has developed several simple measures of nearness of configuration.

The representation of the spider function value as a hexadecimal integer has additional nice properties. It is a two-digit number; and each digit describes the street directional behavior in a four-sector band constituting a semi-circular region. A digit in the second position describes the same configuration as the same digit would in the first position except for a rotation of 180°.

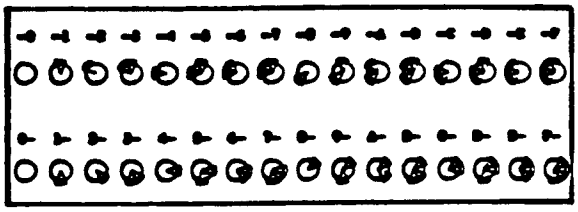

Figure 4. Hexadecimal and Sector Patterns for Spider Function

A frequency distribution of spider function values for a map may be organized in a sixteen-by-sixteen table whose columns correspond to second (units) digit values and whose rows correspond to first (or sixteens) digit possibilities. In a highly urbanized area, for example, the frequency of the hexadecimal number 55, representing the north-east-south-west intersections, would be very large, and could help distinguish between urban and other areas. More generally, the frequency table establishes a kind of signature for the street network; and parts of the table, such as the diagonal, have special meaning. (The principal diagonal of the table includes those intersections all of whose streets continue straight through the intersection.)

Two tables (one for the USGS map and one for the DIME map) showing the distribution of spider function values for all map intersections for the 25 square mile Fort Myers area are given below.

	-0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-A	-B	-C	-D	-E	-F
0-	0	26	7	2	20	29	2	-	8	2	1	1	8	6	1	-
1-	26	23	1	29	23	209	11	-	3	26	2	-	2	-	-	-
2-	0	2	4	6	1	0	20	-	3	4	9	1	1	-	1	-
3-	1	13	6	-	20	2	-	-	1	-	-	-	-	-	-	-
4-	20	22	4	14	23	215	12	-	2	20	7	-	6	4	-	-
5-	40	272	20	-	204	225	2	-	20	2	-	-	-	-	-	-
6-	6	20	4	1	4	2	4	-	-	-	8	-	1	-	-	-
7-	1	-	-	1	-	-	2	-	-	-	-	-	-	-	-	-
8-	8	-	1	2	2	13	2	-	2	20	21	-	2	-	-	-
9-	-	29	2	-	17	-	-	-	21	2	2	-	-	-	-	-
A-	2	2	20	-	20	-	2	-	29	2	9	-	-	-	-	-
B-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C-	6	1	2	-	11	-	-	-	6	-	2	-	-	-	-	-
D-	8	1	1	-	2	-	-	-	1	-	1	-	-	-	-	-
E-	1	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-
F-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 2A. USGS Map Spider Function Distribution

	-0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-A	-B	-C	-D	-E	-F
0-	-	90	22	-	80	24	-	-	21	1	1	-	-	6	-	-
1-	78	23	4	20	0	209	4	-	-	21	1	-	-	2	-	-
2-	14	2	0	4	4	12	7	-	1	1	6	2	2	-	-	1
3-	-	0	2	1	6	2	-	-	-	-	-	-	-	-	-	-
4-	77	20	5	9	21	204	13	1	5	22	5	1	24	1	-	-
5-	10	170	20	2	204	124	4	-	13	5	1	-	-	-	-	-
6-	2	6	2	-	20	7	2	-	2	-	5	-	1	-	-	-
7-	-	2	2	1	-	-	1	-	-	-	-	-	-	-	-	-
8-	12	1	2	1	2	0	2	-	2	1	17	-	9	-	-	-
9-	2	13	20	-	24	2	-	-	20	1	2	-	-	-	-	-
A-	2	5	0	-	0	1	-	-	25	2	12	-	4	-	-	-
B-	4	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-
C-	2	-	5	-	24	-	-	-	5	-	-	-	1	-	-	-
D-	7	-	-	-	2	-	-	-	-	-	-	-	-	-	-	-
E-	1	-	-	-	1	-	-	-	-	-	1	-	-	-	-	-
F-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 2B. Census Map Spider Function Distribution

Tables can orient an initial exploratory data analysis of intersection patterns of the area. One may also display, for example, in the plane, all of those intersection points having a particular spider value (or a range of related spider values) and then proceed to apply pattern recognition techniques to the pattern.

Since the patterns themselves are linked to spatial position, the tables shown above could further be decomposed according to subareas or subregions. Although the total number of entries would decrease, the fewer entries would then reflect more accurately the neighborhood or local characteristics of the street network.

The figures shown below illustrate the spatial distribution of spider function values. In the first set of figures the entire range of values are plotted in their intersection locations. In the other sets only those intersections with particular function values are plotted.

Although condensing the network information at an intersection to a single number inevitably causes some loss of information, the resulting patterns lend themselves to many standard pattern recognition and analysis techniques. A number of references are listed in the bibliography for such spatial analysis techniques. The pattern differences need to be viewed not only in terms of statistical error measurements, but also in terms of geometric relations of similarity shared by subsets of the spider function values (see examples in reference by Rosen and Saalfeld).

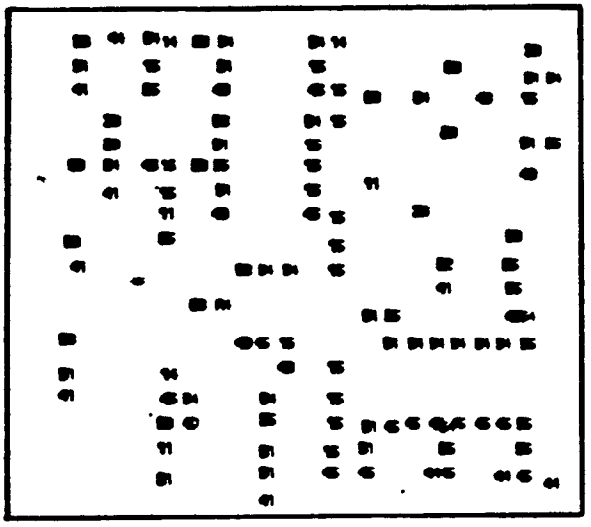

Figure 5A. All Spider Function Values in a Region of USGS Map

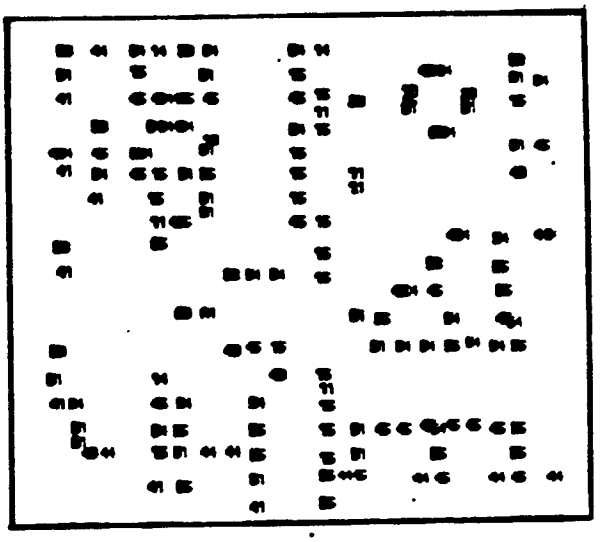

Figure 5B. Spider Function Values in Same Region of DIME Map

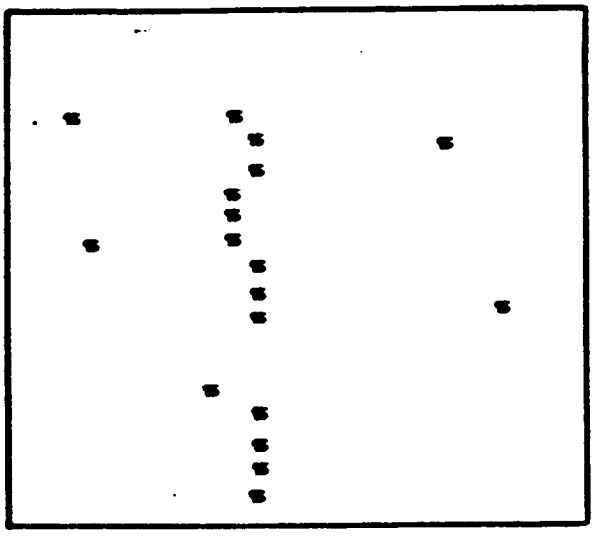

Figure 6A. USGS Intersections with value 15

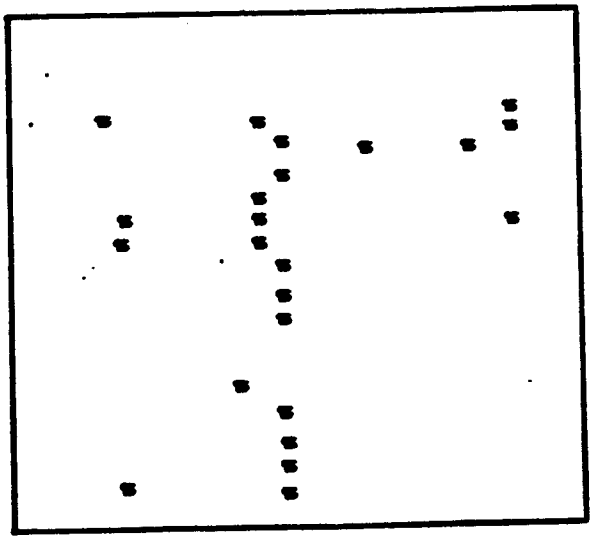

Figure 6B. Census Intersections with Value 15

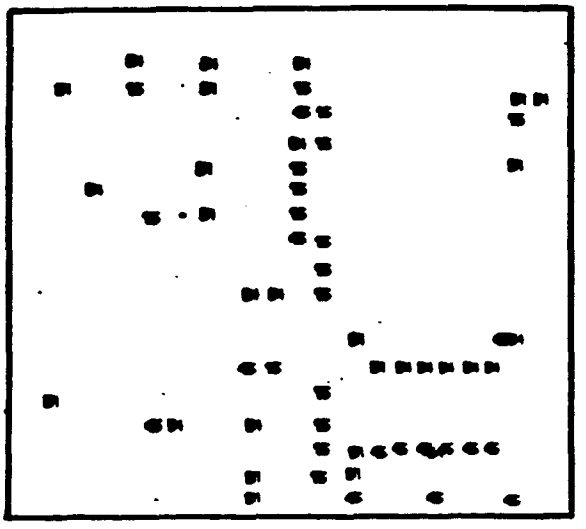

Figure 7A. USGS Intersections with Hexadecimal Values 15, 51, 45, and 54 (T's)

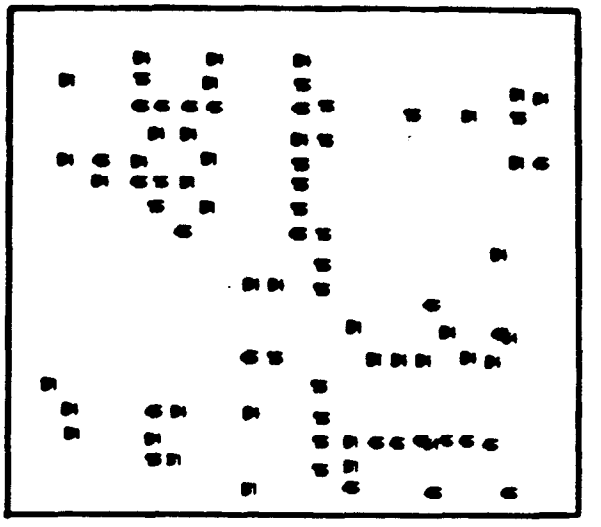

Figure 7B. Census Intersections with Hexadecimal Values 15, 51, 45, and 54

CONCLUSIONS

An analysis of distances between matching and nonmatching map features indicates that nearness measures can and should play a key role in automated map matching routines. A further link between computer cartography and spatial statistical analysis is provided by an integer-valued function defined on map intersection points. Preliminary exploratory work to study properties of this function has begun with limited data resources; and the approach used in that work has been outlined here. The next stage in the research will involve the application of pattern recognition techniques to attempt fully automated map matching.

ACKNOWLEDGMENT

The author wishes to cite the outstanding computer graphics support provided throughout this research work by Maureen Lynch of the Statistical Research Division of the Census Bureau.

REFERENCES

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

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

Pavlidis, T., 1982, Algorithms for Graphics and Image Processing, Computer Science Press, Rockville, MD.

Ripley, B.D., 1981, Spatial Statistics, John Wiley, New York, NY.

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

Saalfeld, A., 1985, "Comparison and Consolidation of Digital Databases Using Interactive Computer Graphics," Census/SRD/Research Report Number 85-11, Washington, DC.

Serra, J., 1982, Image Analysis and Mathematical Morphology, Academic Press, New York, NY.

Tou, J.T., and R.C. Gonzalez, 1974, Pattern Recognition Principles, Addison-Wesley, Reading, MA.

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.

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