BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION REPORT SERIES
SRD Research Report Number: Census/SRD/RR-87/27

Classifying and Comparing Spatial Relations of Computerized Maps for Feature Matching Applications

by

Alan Saalfeld
Statistical Research Division
Bureau of the Census

Recommended by: Tom O'Reagan

Report completed: September, 1986

Report issued: October 1, 1987

Revised and updated version of STATISTICAL ANALYSIS OF MAP DIFFERENCES

This version to be published in JOS

CLASSIFYING AND COMPARING SPATIAL RELATIONS OF COMPUTERIZED MAPS FOR FEATURE MATCHING APPLICATIONS

Alan Saalfeld
Statistical Research Division
Bureau of the Census
Washington, DC 20233

ABSTRACT

Modern computerized maps either contain digital information on spatial relations, such as adjacency relations, shape, network patterns, and measures of position and distance of features, or they permit derivation of that information from the feature data that they do contain. Such spatial attributes lend themselves to computerized statistical analysis much like any other data. Comparative data analysis of spatial relations is possible when two map files are known to cover the same area. In this case, spatial characteristics alone may be used to established linkages between many of the feature records of the two files. This paper presents examples of some spatial measures of distance and local configuration that were used to develop an automated feature matching system at the Bureau of the Census. For a particular sample pair of maps, global summaries and spatial depictions of distance and configuration measures are presented; and some additional uses for the measures are suggested.

KEY WORDS

Computerized maps, map distortion, automated cartography, feature matching, record linkage, configuration, conflation, spider function.

1. INTRODUCTION

1.1. Background.

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 (computerized) 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 (USGS) for the Bureau of the Census to use with its own metropolitan map files for comparative updating of both sets of maps. The second set of USGS digital maps was created by mechanically scanning line drawings of road and water networks and thus contains only spatial information about line segments and their intersections. It does not contain any name or attribute information. Thus, only comparisons involving line segments, their locations and locations of their intersections, and derived spatial measures are possible. All of the work in this paper, therefore, treats a map as nothing more than a plane line graph or network.

In the past, measures of similarity and differences of linear features 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 noted or listed; and there was no readily understood measure of map similarity or sameness. The digital map file, on the other hand, is by its very nature considerably more amenable to numerical analysis, and its format invites computerization of that analysis.

1.2. Scope of this paper.

Now it is not only feasible and informative to quantify and analyze individual linear feature similarities and differences; it is also useful to try to develop concrete numerical measures and graphic displays of regional and global similarity in order to establish statistical evidence that two maps, or specific significant regions within the two maps, are piece by piece the same. A challenging open problem is to find a local or regional numerical signature that can be used to block or group together feature records in order to limit or localize a search for matches. Finding such a blocking algorithm, typically a critical component to any record linkage system, is currently the principal obstacle to fully automating the feature record matching subsystem of the conflation system.

The aim of this paper is to present some initial attempts at quantifying map similarities and differences when the maps consist exclusively of spatial information. The paper outlines approaches to analysis of those differences and similarities; it does not contain extensive empirical justification for those approaches. This last constraint is due in part to the limited available data. Although the Bureau of the Census will eventually have to conflate over 5,500 map pairs (each map covering approximately 50 square miles), only three such map pairs were made available for this research. While the results of the initial quantification measures appear very encouraging in the few examples to which they have been applied, it is necessary to note that the measures themselves are only a few of many possible measures; and the observations based on three map sets illustrate the potential for, rather than prove, the measures' effectiveness.

2. CONFLATION AND 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 develop a rule-based system and to drive an automatic statistical matching procedure, which can then replicate the cartographer's results and, thus, automate the map conflation process. Due to the need for uniform processing and 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.

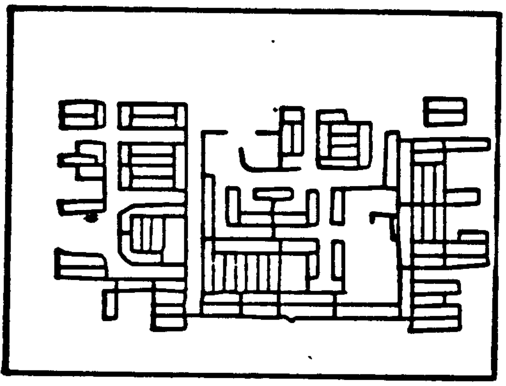
At the Bureau of the Census, in order to assess various rules for matching, a semi-automatic interactive color graphics prototype conflation system has been implemented on a Tektronix 4125B Workstation (Lynch and Saalfeld, 1985). A computer operator uses the system to manipulate map images and to classify street intersections and street segments as matches or non-matches.

The system is semi-automatic in that it has been programmed to simulate the feature-matching detection of a cartographer by applying various matching criteria and then prompting the operator with its findings. The criteria involve position and configuration characteristics of the map features being compared. After the computer locates a likely match based on the matching criteria, the operator needs only to verify or reject the proposed match. The use of color to distinguish between the two maps and to distinguish features that have already been classified as matches or non-matches has also facilitated operator decision-making procedures. After matches have been confirmed, fast rubber-sheeting algorithms are used to align the maps, thereby permitting very effective immediate visual verification of matching decisions. The most valuable element of the color graphics/image alignment approach has been the ease and accuracy of assessing whether or not a match was made correctly. The currently used matching/alignment procedure is iterative: with each iteration, it brings more and more matched feature pairs into exact alignment, moves matchable pairs closer and closer together, and moves pairs which do not match farther and farther apart (Saalfeld, 1985).

3. DIFFERENCES WITHIN AND BETWEEN MAPS

3.1. Measures of feature position or location.

This study of map similarities and differences focuses on street intersections and their configuration and location. Intersection locations are stored by their coordinates; and as one would expect, the intersections are not clustered in space, but are fairly evenly distributed in the plane, as shown in Figure 1B.



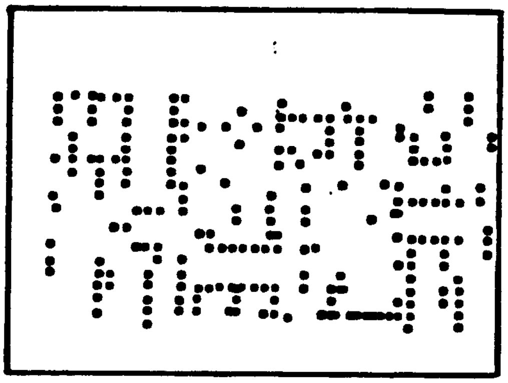


Figure 1A. USGS Map of Part of Fort Myers, FL

Figure 1B. Street Intersection Point Distribution for Same Map

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

Distortion is most easily analyzed through overlay techniques. Alignment may be achieved through elementary transformations called rubber-sheeting functions that relocate key points of one or both maps on top of corresponding points and move other points of the maps proportionately. The transformations used in the Census Bureau system are piecewise linear homeomorphic (PLH) functions defined on a triangulation of the map space or spaces. (Griffin and White, 1985).

Others have used smooth functions such as bivariate quintics, again defined on triangulations, (Lupien and Moreland, 1987) for their rubber-sheeting alignment.

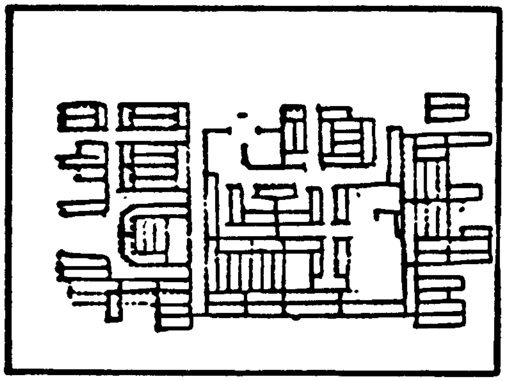


Figure 2A. Overlay of Two Map Sections of Fort Myers, FL

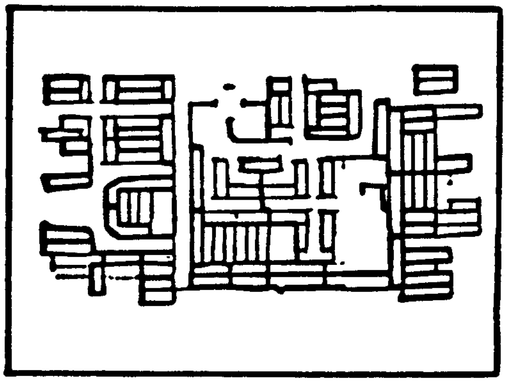


Figure 2B. Matched and Aligned Sections of the Same Area

Figures 2A and 2B suggest that a good initial alignment achieved with PLH transformations can bring nearly all matchable pairs into proximity in such a way that being a nearest street intersection on the other map almost becomes a necessary (but not sufficient) condition for intersection matchability.

Exploratory studies of distortion (ibid.) have displayed as elevation the displacement in each coordinate direction between maps to produce distortion surfaces such as the following:

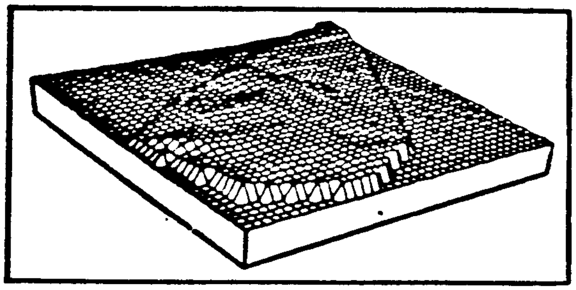


Figure 3A. 50 Link Distortion Surface for X Coordinate*

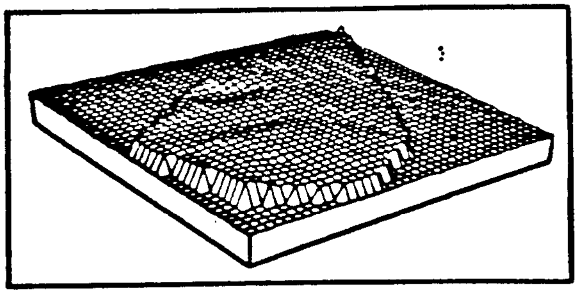


Figure 3B. 50 Link Distortion Surface for Y Coordinate*

*Figures 3A and 3B reproduced with authors' permission

Available rubber-sheeting techniques have no difficulty aligning maps of different scales and orientation. Distortion surfaces measure the amount of movement required for that alignment. The mean slope of each distortion surface, for example, would reflect the overall scale difference in each of the respective coordinates. Orientation change has a similarly predictable and detectable effect on the distortion surfaces.

In general error theory, 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 pairing (false negative). In matching theory a third type of error occurs when a feature is judged correctly to have a match, but the wrong correspondent is judged to be the matching element. This type of error is called mismatch. The iterative matching procedure used with the conflation system identifies new matches at each stage and does not label non-matches as such until the final stage. False negatives are a residual and do not present a problem at an intermediate iteration. False positive errors and mismatches are less desirable and less managable than false negatives because they may precipitate additional errors at subsequent iterations, and at no point in the iteration procedure is there an un-matching capability for correcting false positives and mismatches.

The Euclidean distance between potential matches after initial alignment is an excellent measure for controlling both mismatch and false negative errors. For one particular test map of Fort Myers, FL, Table 1 shows the distribution of instances of distance ranges from matchable points on the Census map to their matched or paired points on the USGS map (column 2), and from the same matchable points on the Census map to their nearest non-matching neighbors on the USGS map (column 3).

The initial alignment used to produce Table 1 was accomplished through hardware and software image manipulation of Census and USGS maps. First the Census map was subdivided into 32 equal sized rectangular pieces. Each rectangular piece could be moved anywhere on the screen by the operator. Using the entire USGS map as background, the operator positioned each small Census rectangle in order to produce the best possible visual alignment near each small rectangle's centroid. The movement that had been required to so position the 32 centroids was recorded and averaged locally (using PLH functions on a triangulation of the Census map) in order to rubber-sheet the Census map and re-compute all of its coordinates (Saalfeld, 1985).

The cumulative relative histograms which follow and summarize Table 1 support the idea that, after initial map alignment, nearest neighbor pairs are excellent candidates for matching.

Table 1. Distribution of Distances from Matchable Points to their Matches and Nearest Nonmatches (After Initial PLH Alignment*)

Distance Range (meters)	Number of Matchable Points whose Matching Pair is within range	Number of Matchable Points whose Nearest Non-match is within range
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	-	-
All distances	1034	1034

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

*PLH alignment uses 32 local alignments and 66 triangles

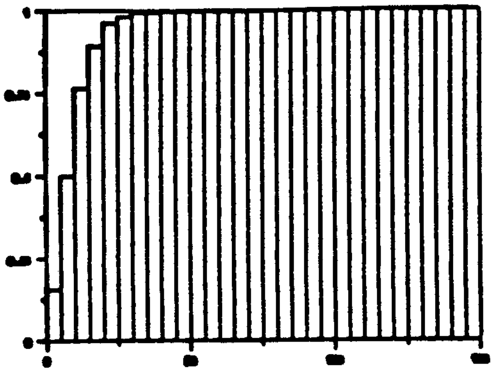


Figure 4A. Fraction of Matchable Points whose Matching Point is Within the Indicated Distance of the Point

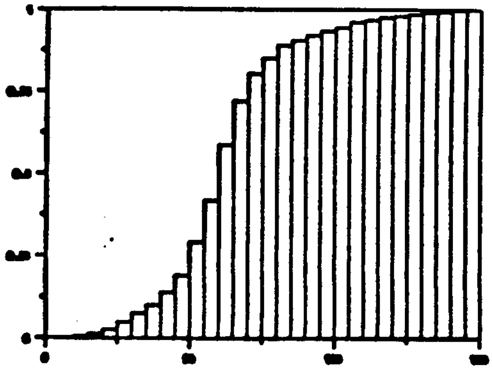


Figure 4B. Fraction of Matchable Points whose Nearest Non-Matching Point is Within the Indicated Distance of the Point

Nearness alone will not suffice for matching. Nonetheless, distance tolerances may be used for estimating both mismatch and false negative error types and reducing one type 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 of each other), then the measured probability of omiting a match (false negative) is 11%, and the probability of mismatching a matchable point is 1%. By decreasing the threshold, mismatches 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.

3.2. Measures of configuration.

The remainder of this paper focuses on other match criteria tests to supplement nearest neighbor tests. In order to facilitate quantitative comparisons of intersection patterns, the configurations are assigned numerical summary values and are grouped according to those values. The coding scheme reflects similarities of patterns through the assignment of nearly equal summary values when the intersection configurations themselves are nearly identical (Rosen and Saalfeld, 1985). These additional criteria utilize the following numerical measures of local configuration:

3.2.1. The Degree of an intersection. The number of streets emanating from an intersection is called the *degree* of the intersection. The degree 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.)

3.2.2. The Spider Function of an intersection. The street pattern at an intersection (that is, the emanating rays) has infinitely many possibilities for street directions. In order to simplify the possibilities, the number of directions was reduced to 8 sectors. The eight sectors correspond to 45° pie slices centered upon the principal directions of north, northeast, east, southeast, south, southwest, west, and northwest. 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 named descriptively the spider function of the intersection. With this function, an integer between 1 and 2^8-1 describes the street pattern of the intersection. The binary number 01010101 (which is the decimal 85 and hexadecimal 55) represents the typical 4-street north-south-east-west intersection, for example. The street pattern is assumed to have at most one street in each of the eight sectors. (If more than one street occurs in any sector, the spider function may be given a special value 0 or it may simply ignore the extra street. Limited experience suggests that ignoring the extra street will not adversely affect our matching procedure since (1) two street in the same sector are very rare, and (2) matching is allowed if street configurations are only similar--e.g. "off by one"--and not identical.) Intersection patterns which differ by a power of two are usually "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 *degree* of a intersection 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 (base 16) integer has additional nice properties:

(1) The spider function value is always a two-digit number.

(2) Each digit describes the street directional behavior in a four-sector band constituting a semi-circular region.

(3) A digit K in the second (units) position describes the same configuration as the same digit K would describe in the first (sixteens) position except for a rotation of 180° (see Figure 5).

(4) The configuration with hexadecimal digits NM is the 180° rotation of the configuration with hexadecimal representation MN, the number with digits M and N transposed.

(5) Numbers with repeated digits KK and only those numbers have all streets continuing straight through the intersection.

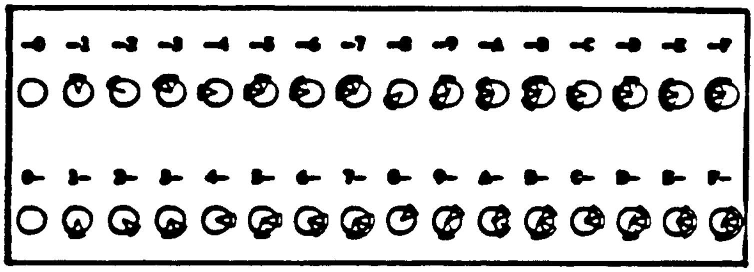


Figure 5. Hexadecimal and Sector Patterns for Spider Function.

3.3. Summary Statistics on Global Configuration.

3.3.1. Spider Function Tables.

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 the hexadecimal representation. 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 is comprised precisely of those intersections all of whose streets continue straight through the intersection.)

Two tables (one for the USGS map and one for the Census map) showing the distribution of spider function values for all map intersections for the 25 square mile Fort Myers area are given below. Such tables can orient an initial exploratory data analysis of intersection patterns of the area. After viewing the tables, one may display, 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, as is illustrated below.

	-0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-A	-B	-C	-D	-E	-F
0-	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-
1-	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-	[illegible]	[illegible]	[illegible]	-	[illegible]	-	-	-
2-	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-	[illegible]	-
3-	[illegible]	[illegible]	[illegible]	-	[illegible]	[illegible]	-	-	[illegible]	-	-	-	-	-	-	-
4-	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-	[illegible]	[illegible]	[illegible]	-	[illegible]	[illegible]	-	-
5-	[illegible]	[illegible]	[illegible]	-	[illegible]	[illegible]	[illegible]	-	[illegible]	[illegible]	-	-	-	-	-	-
6-	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-	-	-	[illegible]	-	[illegible]	-	-	-
7-	[illegible]	-	-	[illegible]	-	-	[illegible]	-	-	-	-	-	-	-	-	-
8-	[illegible]	-	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-	[illegible]	[illegible]	[illegible]	-	[illegible]	-	-	-
9-	-	[illegible]	[illegible]	-	[illegible]	-	-	-	[illegible]	[illegible]	[illegible]	-	-	-	-	-
A-	[illegible]	[illegible]	[illegible]	-	[illegible]	-	[illegible]	-	[illegible]	[illegible]	[illegible]	-	-	-	-	-
B-	[illegible]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C-	[illegible]	[illegible]	[illegible]	-	[illegible]	-	-	-	[illegible]	-	[illegible]	-	-	-	-	-
D-	[illegible]	[illegible]	[illegible]	-	[illegible]	-	-	-	[illegible]	-	[illegible]	-	-	-	-	-
E-	[illegible]	-	-	-	-	-	-	-	-	-	[illegible]	-	-	-	-	-
F-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 2A. Spider function distribution for USGS intersections.

	-0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-A	-B	-C	-D	-E	-F
0-	-	[illegible]	[illegible]	-	[illegible]	[illegible]	-	-	[illegible]	[illegible]	[illegible]	-	-	[illegible]	-	-
1-	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-	-	[illegible]	[illegible]	-	-	[illegible]	-	-
2-	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-	-	[illegible]
3-	-	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-	-	-	-	-	-	-	-	-	-
4-	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-	-
5-	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-	[illegible]	[illegible]	[illegible]	-	-	-	-	-
6-	[illegible]	[illegible]	[illegible]	-	[illegible]	[illegible]	[illegible]	-	[illegible]	-	[illegible]	-	[illegible]	-	-	-
7-	-	[illegible]	[illegible]	[illegible]	-	-	[illegible]	-	-	-	-	-	-	-	-	-
8-	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-	[illegible]	[illegible]	[illegible]	-	[illegible]	-	-	-
9-	[illegible]	[illegible]	[illegible]	-	[illegible]	[illegible]	-	-	[illegible]	[illegible]	[illegible]	-	-	-	-	-
A-	[illegible]	[illegible]	[illegible]	-	[illegible]	[illegible]	-	-	[illegible]	[illegible]	[illegible]	-	[illegible]	-	-	-
B-	[illegible]	-	[illegible]	-	-	-	-	-	-	-	-	-	-	-	-	-
C-	[illegible]	-	[illegible]	-	[illegible]	-	-	-	[illegible]	-	-	-	[illegible]	-	-	-
D-	[illegible]	-	-	-	[illegible]	-	-	-	-	-	-	-	-	-	-	-
E-	[illegible]	-	-	-	[illegible]	-	-	-	-	-	[illegible]	-	-	-	-	-
F-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 2B. Spider function distribution for Census intersections.

As an illustration of exploratory analysis that could be applied to the above tables, notice that the total number of intersections for the USGS map is far greater than the total for the Census map. This difference is due to the greater extent or coverage of the USGS map. The Census map merely covers a subregion of the USGS map. Nevertheless, cell percentages are very similar, indicating that the distribution of intersections by configuration types is the same. Moreover, the anomaly of having fewer "55" or north-south-east-west intersections than any type of "T" intersection: 15, 51, 45, and 54, is apparent in both tables. The prevalence of "T" intersections in the Fort Myers area is due to frequent water inlets that result in numerous natural road barriers. It is indeed a signature or identifying characteristic for the area.

Since the occurrences are linked to spatial position, the tables shown above could further be decomposed according to subareas or subregions of the map. Although the total number of entries would decrease, the entries present would then reflect more accurately local characteristics of the chosen subarea of the street network.

3.3.2. Spider Displays as Point Patterns. After the spider function tables are compiled, one may choose to display as point patterns only those intersections whose occurrences in the spider function tables are judged extraordinary. One may look at rare occurrences such as the unique "6C" intersection appearing on both maps; or one may draw all "15 T" intersections to try to determine why they are so frequent. The second option is illustrated in the figures below as a filtering operation. In the first set of figures the entire range of spider function values in a subregion are plotted in their intersection locations. In the other sets only those intersections with particular spider function values are plotted.

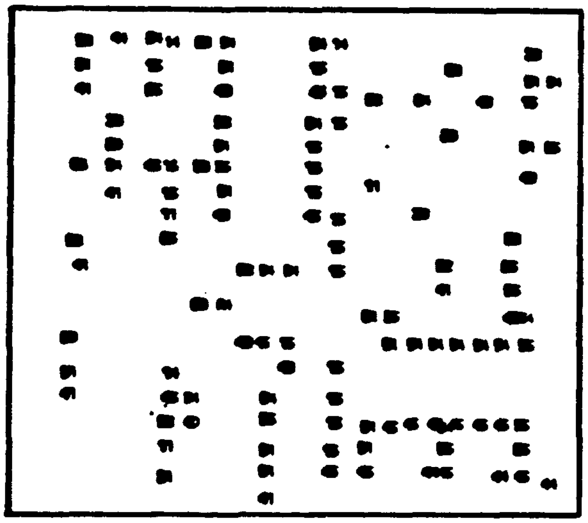


Figure 6A. All Spider Function Values in a Subarea of USGS Map

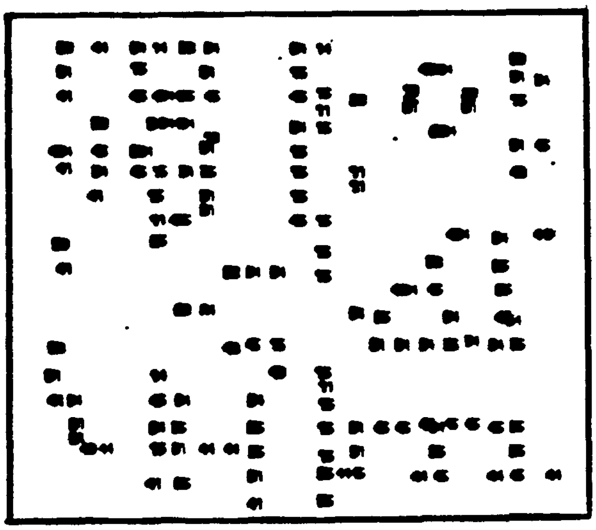


Figure 6B. Spider Values in Same Subarea of Census Map

By looking only at "T" intersections in the area, Figures 7A and 7B, (and using knowledge that each "T" value corresponds to a single direction failing to "go through"--for instance "15" does not go through to the west or left), one may almost visualize the barriers (in this case known to be water inlets). A vertical string of "15's" just to the right of a vertical string of "51's" clearly flank one such inlet! A horizontal string of "54's" sitting above a similar string of "45's" clearly flank a horizontal inlet.

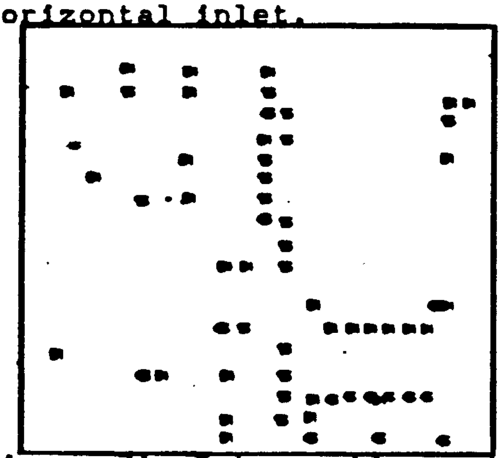


Figure 7A. Intersections of USGS Map with Hexadecimal Values {15,51,45,54} (T's)

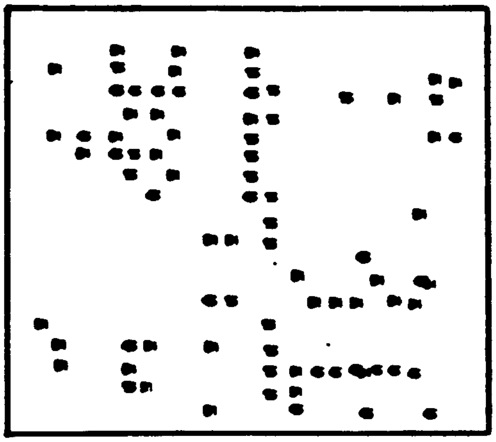


Figure 7B. Intersections of Census Map with Hexadecimal Values {15,51,45,54} (T's)

A second filtering operation to reduce one's view to only a single class of intersections ("15's") produces a set of figures even more amenable to standard pattern recognition techniques.

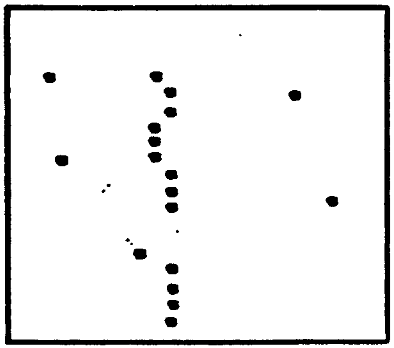


Figure 8A. Intersections of USGS Map with Value = 15

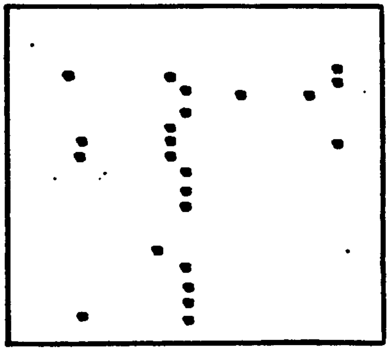


Figure 8B. Intersections of Census Map with Value = 15

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. The pattern distributions need to be viewed not only in terms of statistical error measurements, but also in terms of geometric relations of similarity and dependence shared by subsets of the spider function values. Two spider function values represent similar intersections patterns and, hence, are "close," for instance, if one value is twice the other or if their difference is a power of two.* Likewise, values occurring at opposite ends of the same line segment must exhibit clear geometric dependence reflected in one of their digits. Only exploratory work has been undertaken to study geometric implications of spider function value distributions (Rosen and Saalfeld, 1985).

4. 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 and illustrated here. The next stage in the research will involve the application of image analysis and pattern recognition techniques to attempt fully automated map matching.

5. ACKNOWLEDGMENTS

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.

6. REFERENCES

Griffin, P., and M. White, 1985, "Piecewise Linear Rubber-Sheet Map Transformations," The American Cartographer, v. 12:2, pp 123-131.

Lupien, A., and W. Moreland, 1987, "A General Approach to Map Conflation," AUTOCARTO 8 Proceedings, Baltimore, MD, pp 630-639.

Lynch, M. P., and A. Saalfeld, 1985, "Conflation: Automated Map Compilation--A Video Game Approach," AUTOCARTO 7 Proceedings, Washington, DC, pp 342-352.

Rosen, B., and A. Saalfeld, 1985, "Matching Criteria for Automatic Alignment," AUTOCARTO 7 Proceedings, Washington, DC, pp 456-462.

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