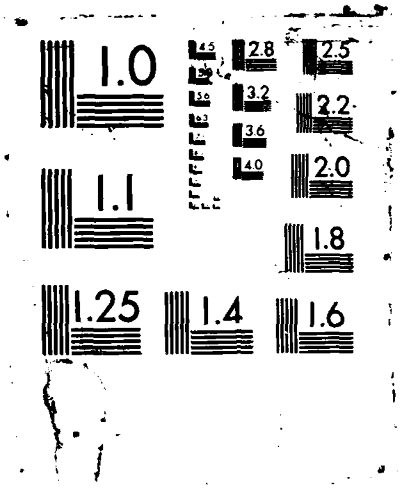
AD-A192 787 AUTOMATIC CORRELATION OF USGS (US GEOLOGICAL SURVEY) 1/1
DIGITAL LINE GRAPH G..(U) BATTELLE MEMORIAL INST
COLUMBUS OH COLUMBUS LABS A M HEARD AUG 86 ETL-0426
UNCLASSIFIED DAAG29-81-D-0100 F/G 8/2 NL


1.0
4.5
5.0
5.6
6.3
2.8
2.5
3.2
2.2
3.6
4.0
2.0
1.1
1.8
1.25
1.4
1.6

ETL-0426

Automatic correlation of USGS digital line graph geographic features to GNIS names data

Andrew M. Heard

Battelle Memorial Institute
Columbus Laboratories
505 King Avenue
Columbus, Ohio 43201-2693

August 1986

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION IS UNLIMITED.

Prepared for
U.S. ARMY CORPS OF ENGINEERS
ENGINEER TOPOGRAPHIC LABORATORIES
FORT BELVOIR, VIRGINIA 22060-5546

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

ADA192787

REPORT DOCUMENTATION PAGE

Form Approved
OMB No 0704-0188
Exp Date Jun 30, 1986

1a. REPORT SECURITY CLASSIFICATION
Unclassified

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE

3. DISTRIBUTION / AVAILABILITY OF REPORT
Approved for public release; distribution is unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)

5. MONITORING ORGANIZATION REPORT NUMBER(S)
ETL-0426

6a. NAME OF PERFORMING ORGANIZATION
Battelle Memorial Institute
Columbus Laboratories

6b. OFFICE SYMBOL *(If applicable)*

7a. NAME OF MONITORING ORGANIZATION
U.S. Army Engineer Topographic Laboratories

6c. ADDRESS *(City, State, and ZIP Code)*
505 King Avenue
Columbus, Ohio 43201-2693

7b. ADDRESS *(City, State, and ZIP Code)*
Fort Belvoir, Virginia 22060-5546

8a. NAME OF FUNDING / SPONSORING ORGANIZATION

8b. OFFICE SYMBOL *(If applicable)*

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
DAAG29-81-D-0100

8c. ADDRESS *(City, State, and ZIP Code)*

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO

11. TITLE *(Include Security Classification)*
Automatic Correlation of USGS Digital Line Graph Geographic Features to GNIS Names Data

12. PERSONAL AUTHOR(S)
Heard, Andrew M.

13a. TYPE OF REPORT
Final

13b. TIME COVERED
FROM ________ TO ________

14. DATE OF REPORT *(Year, Month, Day)*
1986 August

15. PAGE COUNT
32

16. SUPPLEMENTARY NOTATION
This study was subcontracted to Herbert Freeman, 7 Woodview Drive, Cranbury, New Jersey 08512.

17. COSATI CODES

FIELD	GROUP	SUB-GROUP
08	02	

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
Artificial intelligence
Automated cartography
Automated map generation

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

An accelerating interest in automatic name placement has revealed a growing need for the creation of cartographic data bases. An effort has been made to create such a data base from two extant types of digitized USGS map data. Features are defined by geographic data stored as graphic elements in the USGS 1:2,000,000 Digital Line Graph data files. These features are then used as a set onto which names extracted from the Geographic Names Information System data files are mapped. This mapping serves as a cartographic data base, supplying both a label and a graphic description of specific geographic features. Automatic generation of this data base draws nearer the realization of automated map generation.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT
□ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT □ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL
Paul Logan

22b. TELEPHONE *(Include Area Code)*
(202) 355-2777

22c. OFFICE SYMBOL
ETL-TD-MA

83 APR edition may be used until exhausted
All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

Table of Contents

List of Figures

Automatic Correlation of USGS Digital Line Graph Geographic Features to GNIS Names Data

1. Introduction

Much effort in recent years has been spent researching and developing ways to automate map generation systems. Of particular interest and difficulty has been automated name placement. Because of the relative youth of name placement technology, and its still relatively poor quality compared to manual placement, little effort has been made to develop a cartographic data base for use by a name placement expert system. An effort by Umit Basoglu [2] represents one of the few investigations into the possibility of using the digital map files being developed by the U.S. Geological Survey for such a purpose.

The problem of associating labels with geographic features is in some ways as elusive as the task of automated name placement itself. The data used for such association is subject to the prejudices of the cartographer who generated the map data or determined reference points for specific features. A major issue in the emerging science of expert systems is the problem in designing systems to perform tasks done by several experts, each having a personal preference as to how that task should be performed. The U.S. Geological Survey's Digital Line Graph Data (DLG) is digitized by human cartographers following general guidelines. Each cartographer uses his or her best judgment in determining what and how a feature is to be represented. The USGS Geographic Names Information System (GNIS) names data is also digitized in the same manner, with several cartographers following general guidelines. The guidelines for the GNIS digitizors[1] are not necessarily consistent with the guidelines for the DLG digitizors. Thus not only are we presented with the problem of having to interpret the digitizing methodologies of several different individuals for both sets of data, but we are also faced with the problem that the methodologies of different individuals are based on two different standards.

This, of course, is not meant to fault the digitizors; nor is it meant to lay blame on the guidelines they used. Cartography is a very subjective field, one not very easily mastered by today's technology, primarily because conventions in the naming and definition of features often vary widely.

The object of this research, then, is to help determine the feasibility of using digital map data as input to a cartographic name placement system.

[1] A *digitizor* is the term used here to refer to a person doing map digitizing

2. Resources for Development of the Cartographic Data Base

The work for this project was done in the Image Processing Laboratory at Rensselaer Polytechnic Institute. The computer used was a PRIME/750 with a Hewlett-Packard flat-bed plotter. Subsequent plotting was done on CalComp 1012 and 1051 plotters. The nature of the software requires a cartographic expert system to test results; the expert system AUTONAP [1], developed by Dr. John Ahn, was utilized in this capacity. An example of an AUTONAP map made from manually digitized data is shown on the next page.

The United States Geological Survey's Digital Line Graph Data [3] for the northeastern United States (Maine, New York, New Hampshire, Vermont, Massachusetts, Rhode Island, and Connecticut) at a scale of 1:2,000,000 was purchased for use as the geographic feature data base for this project. The data base consists of seven files or overlays, each comprising a disjoint cartographic category: Political Boundaries, Water Bodies, Administrative Boundaries, Roads and Trails, Rivers and Streams, and Cultural Features.

Within each file are four record varieties, the first of which, the header record, contains pertinent projection and boundary information as well as file identification fields. With the exception of the particular file identification, the header records are the same across all seven overlays of a region. The three other records each correspond to the three graphical data types: node, area, and line. The node records each contain the internal file coordinates of the nodes' geographic location. The 1:2,000,000 DLG files only contain point features as degenerate lines; therefore, there is no information related to the node other than an internal file identification number. The area records are represented by their respective internal file identification numbers and geographic reference locations in internal file coordinates. The reference locations need not be located within the bounds of the area. Also contained within the area record is the number of attribute pairs which are used to describe the area.

Every feature type in the DLG data is described in terms of line segments; thus, line records are the most significant data type in the DLG files. Each line record contains the internal identification number of the line segment, the internal file identifications of the nodes which form the line's endpoints, and the internal file identifications of the areas which it bounds to its left and to its right. The record also records how many attributes and intermediate coordinate pairs constitute the line segment. Following each line segment are the intermediate coordinate pairs.

The last information in both the line and area records are the attributes of the features they belong to. It should be noted that a graphic item can belong to several features. For example, a border of a state will usually be a border of a county as well. Because of the separate overlays, situations such as a river being a state boundary are not easily detectable; the river and the state boundary are digitized independently of each other into different overlays and are, therefore, separate graphic elements.

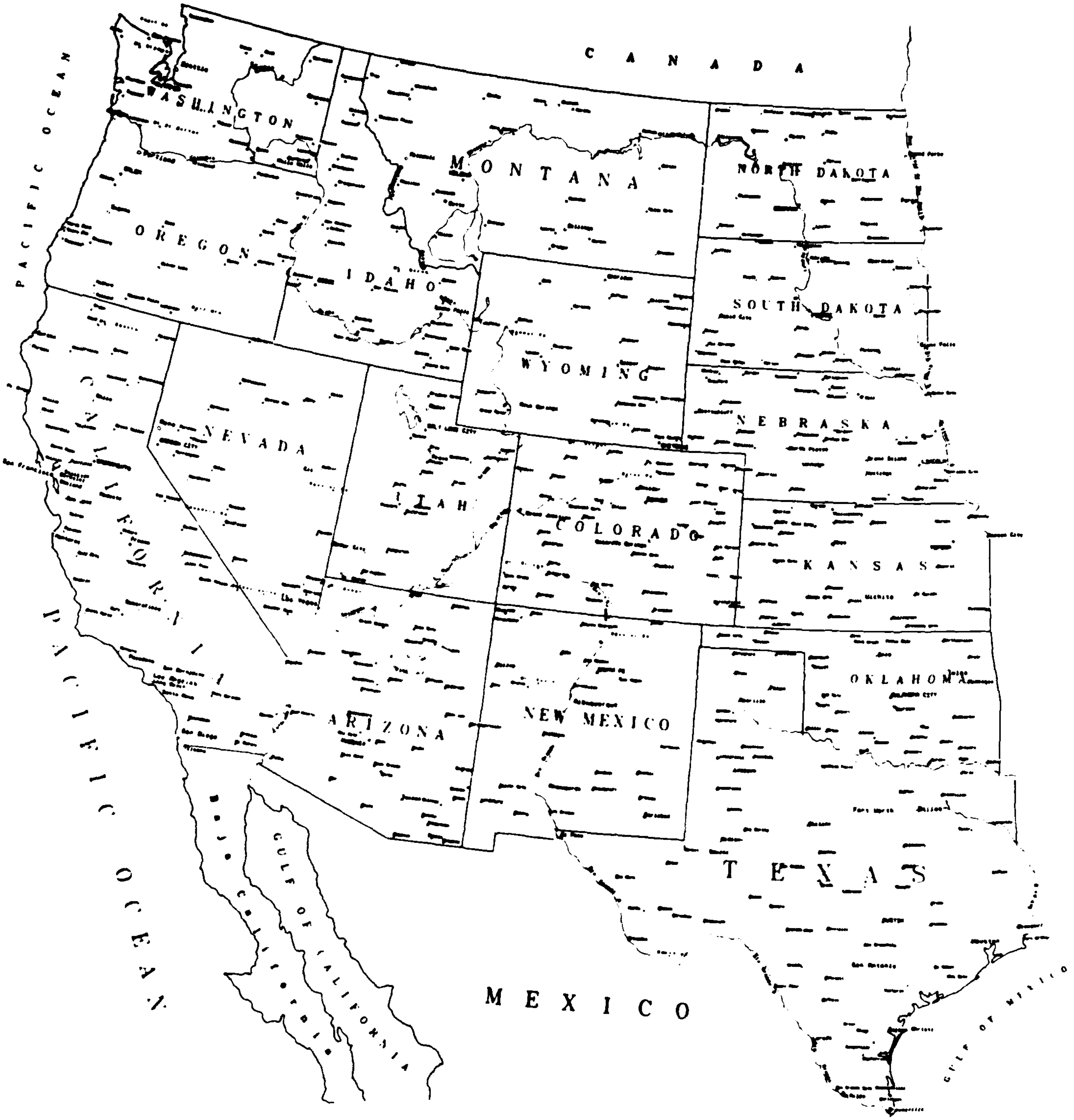


Figure 1: An AUTONAP Map Using Manually Digitized Data

There may be some difficulty determining some attributes. For example, a branch of the Hudson River near its source may be a small stream; yet if it should be determined that the branch should be the extension of the river towards its source, then it will be given the attributes of being a major river.

The Geographic Names Information System (GNIS) Names File [4, 5] is an alphabetic listing of all geographic names in a state. The state used in this study was New York. Associated with each name in the GNIS file is a primary reference coordinate, state and county FIPS code, and a feature classification code called a generic. (Examples of generics are 'river' for rivers and streams, and 'ppl' for populated places.) Some names may be given a secondary reference point. For names which belong to areal features, the primary reference point will be the approximate center of the area (an 'eye-balled' estimate by the person doing the digitizing). The exceptions to this are populated-places reference points which may be the center of the original town, or some important landmark such as a city hall. If the primary coordinate is not within the map boundary, the secondary reference point is used to indicate the part of the feature within the boundary. There are always two reference points for linear features, one being its mouth. The secondary coordinate is simply any place else on the feature within the map. Associated with rivers is a source coordinate which is always the determined source of the river. The way in which this coordinate is digitized is somewhat subjective. The digitizor is asked to follow the river as far as he can take it, letting the shortest drain be the source of the river. Other factors may influence the decision of the digitizor; for example, a branch of 'Big River', the branch having its source near a town called 'Big River', would be a reasonable choice for a source of the river 'Big River'.

3. Inherent Factors Inhibiting Correlation of Names to Geographic Features

Automated association of names to features is made difficult by the problematic issues inherent in the correlating process and by the problems posed by the particular data set being used. From points and lines the automated process must create and classify features. However, many ambiguities develop in this feature definition process. For example, if a river forks, has one river split into two, has one river split off from another, or, since, as in this case, the data may not give the direction of flow, have two rivers combined into one? The same sort of ambiguity can be seen in the naming process. If there is a lake on an island in another lake, does a name which points inside both lakes belong to the smaller or the larger lake?

The large amount of cartographic data does not explicitly give the answers to questions like these. Decisions must be based on assumptions and extrapolations from a limited amount of information. In particular, the USGS data, while not resolving the issues, does help in answering both the preceding questions. The DLG data has an extensive assortment of attributes which may help in differentiating between dissimilar intersecting features, but will not help if the attributes for all the features are the same. A stated practice of GNIS digitizors is to place areal reference points outside of interior features whenever possible. This may not help if scaling differences make the difference in interior coordinates negligible.

The definition of a feature is very much a *function of scale.* An area feature at one scale may be a point feature at another; a mountain range on a large-scale map may be just a peak on a small-scale map, or non-existent on an even smaller-scale map. The GNIS data is digitized, for the most part, from 7.5 minute maps, a much larger scale than the 1:2,000,000-scale maps of the DLG data. Thus, the GNIS digitizor trying to determine a source coordinate for a river may face an entirely different set of choices than those faced by his DLG counterpart.

The DLG data was created as a graphic data base to serve the purposes of a visual representation of geographic features. To that end it has few (if any) shortcomings. The problem in deciphering points and lines into a data base of independent geographic features made up and chosen from their constituent graphic elements is the major problem in the extraction of these features. A precise definition of a feature must be available in order to address it as an entity. Areas must be bounded by a closed lattice of bounding line segments. Line features must be described as a continuous set of coordinates. Until the entire feature can be identified, any labelling will be dubious since it will be based on an approximation of the feature in both location and orientation.

For the most part, the DLG area features are closed and bounded; there are exceptions when the feature encounters the map boundary. The biggest problem encountered in the extraction of area features, however, was the determination of a state's effective boundaries. The information contained within the political boundaries file does not show the geographic boundaries of a state or county when the boundary encounters a major water body, such as the ocean or a Great Lake. Instead, the political boundary extends into the water to show what is called the "seaward extent" of the state and county. Much of the time, when there is no obvious choice such as a national boundary line, this seaward extent is simply an arbitrary boundary set by the digitizing cartographer to close the area. The state and county political boundary data is shown in Figure 2. The graph in 2(a) shows both state and county boundary data, the second graph in 2(b) shows only state boundary data, and the graph in 2(c) shows only the state boundaries on land. This data is, of course, unacceptable as an areal description for the purposes of name placement; aesthetically effective name placement must be based on the geographic limits of the area. The shoreline data necessary to close the effective boundary of the the states in the third graph are not digitized in the political boundaries overlay.

Line-feature extraction is the most difficult. Typically, a line feature is made up of several segments which must necessarily be joined together in order attain a precise definition of the feature as an entity. A line segment does not inherently carry the attributes that will disassociate it from a similar line segment. For example, the line features dealt with in this study were rivers. Rivers merge, divide, and bend in an unpredictable manner. When a river forks, the graphic representation will be a line segment encountering a node from which there are two other, possibly similar, line segments extending. The dilemma, obviously, is choosing which forking line segment belongs to the incompletely extracted line feature. A line feature's attributes may also change from one line segment to another. For example, the Hudson River's southern extension is classified as a double-line river (meaning it is wide enough to represent its shore lines); the northern extension of the Hudson River is attributed as a 300+ kilometer, perennial river charted as a single line. The change in attribute makes it difficult to determine whether the end of the feature has been encountered or whether, as in the case of the Hudson, there is an extension of the feature under another attribute type. To further aggravate this dilemma is the fact that the line segments that switch the feature attributes from perennial single line to a double line river are not connected. As a matter of fact, for the Hudson River in particular, there are two perennial line segments which should but do not actually connect. This is probably not an intentional occurrence on the part of the digitizor; information received from the USGS seemed to indicate that line segments belonging to a feature would in fact be connected. Regardless, encountering a line segment which does not join with another line segment does not necessarily indicate that an endpoint of the linear feature has been found.

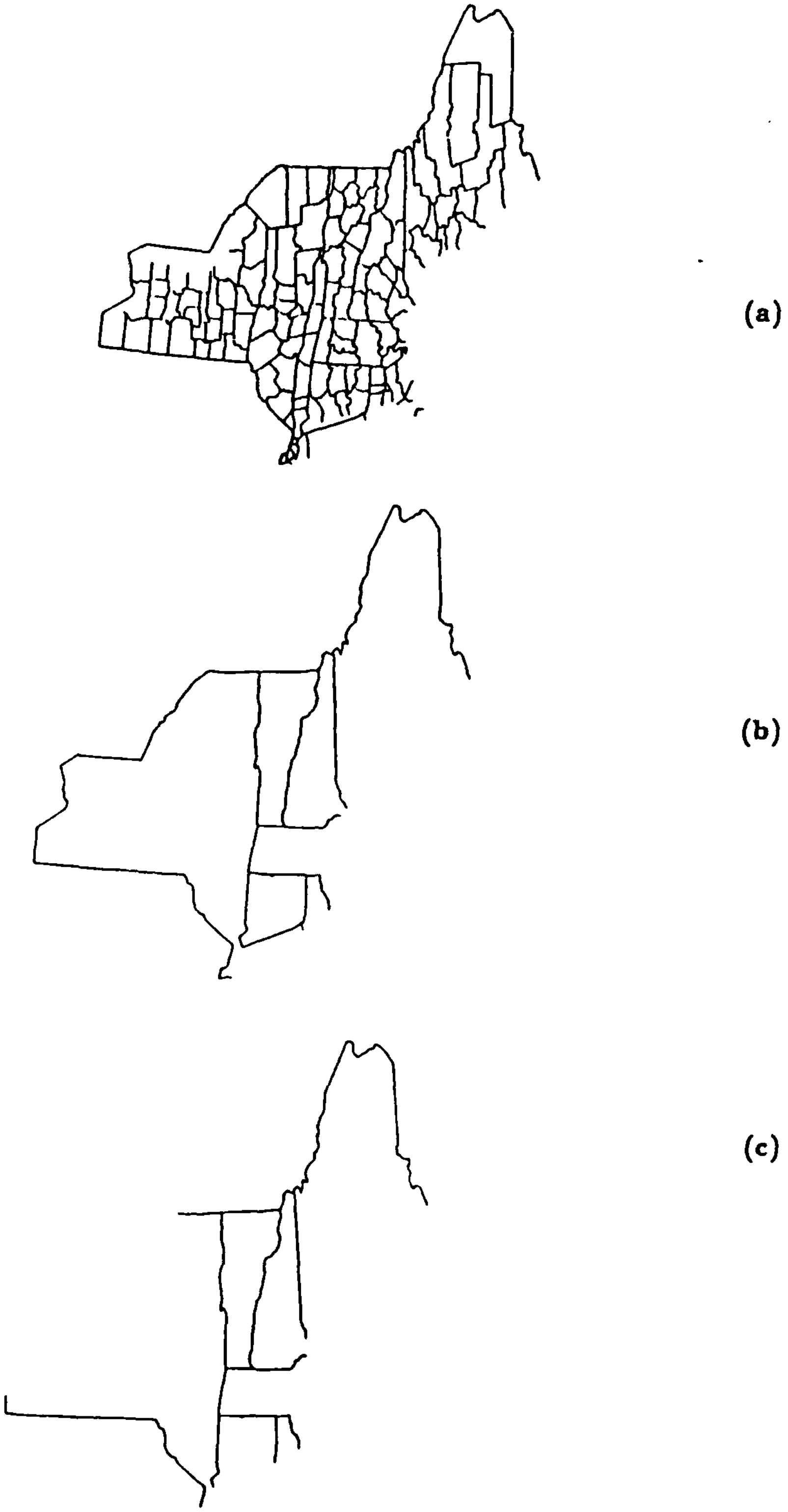


Figure 2: State and County Political Boundary Data
(a) state and county boundaries, (b) only state boundary data,
and (c) only state boundaries on land

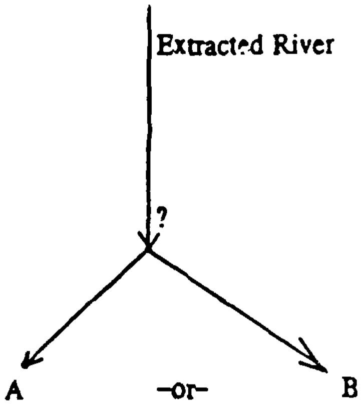


Figure 3: The Forking River Dilemma

4. Software Data Structures

The initial task of the correlating software developed in this study is to read in the DLG data. Each data record is read and stored in its own FORTRAN "COMMON BLOCK", regardless of its attributes. (Certain inherent relationships between dissimilar attributes make it important to store all the data and not just data of the type the user requested.) The information in the header record is translated into the parameters used in translating latitude/longitude coordinates into Albers Equal Area Conic Projection.

The node data structure contains the file information about each node (internal identification number and geographic location) as well as a record of which lines intersect each node. The area data structure contains the identification number of the area, a reference location, and all the areas' attributes. The line data structure contains the node identification numbers for each line segment's two endpoints, the area identification number of the area that it bounds to its left and to its right, and the attributes of each line segment. The intermediate points comprising each line segment are stored in a direct-access, binary file; the record numbers corresponding to the first record of each line segment's intermediate points are also stored in the line data structure.

All area components which have attributes requested by the user to be extracted and named are stored as an extractable area feature in a separate data structure. This separate data structure contains the name if found for each feature, plus a weight measuring the viability of the name. Each area feature also points to a list of pointers which point to the line segments which make up its border. As a line feature is extracted (from line segments which are the same as or are compatible with the attributes that the user requested be extracted and named), a record is initialized in an extractable line feature data structure where the line-feature name (if found) can be stored together with a viability weight. The feature is stored as a sequential linked list of pointers to the line segments which comprise the feature. The state boundaries are treated as belonging to a special area feature since they are extracted and named in a different manner and, therefore, are stored in a separate data structure. Its structure is similar to that of the extractable line data structure.

The data structure of any correlating software has to be extremely extensive to handle the enormous amount of cartographic data. Cross-referencing and searching data constitutes a large amount of the correlating process. The data structure developed for this study is based on a multiple grid structure in order to optimize the search and cross-reference time. For each unique data type, there is a grid which covers the entire area of the map. From each cell of each grid extends a linked list of pointers to a specific data element. There are four such grids used in this study:

1. A grid to search the node data

2. A grid to search shoreline data (special case to extract state boundaries)

3. A grid to store nodes of state boundaries which have only one intersecting state land boundary line

4. A grid to search area data

Figure 4 depicts the general organization of the grid-oriented data structure.

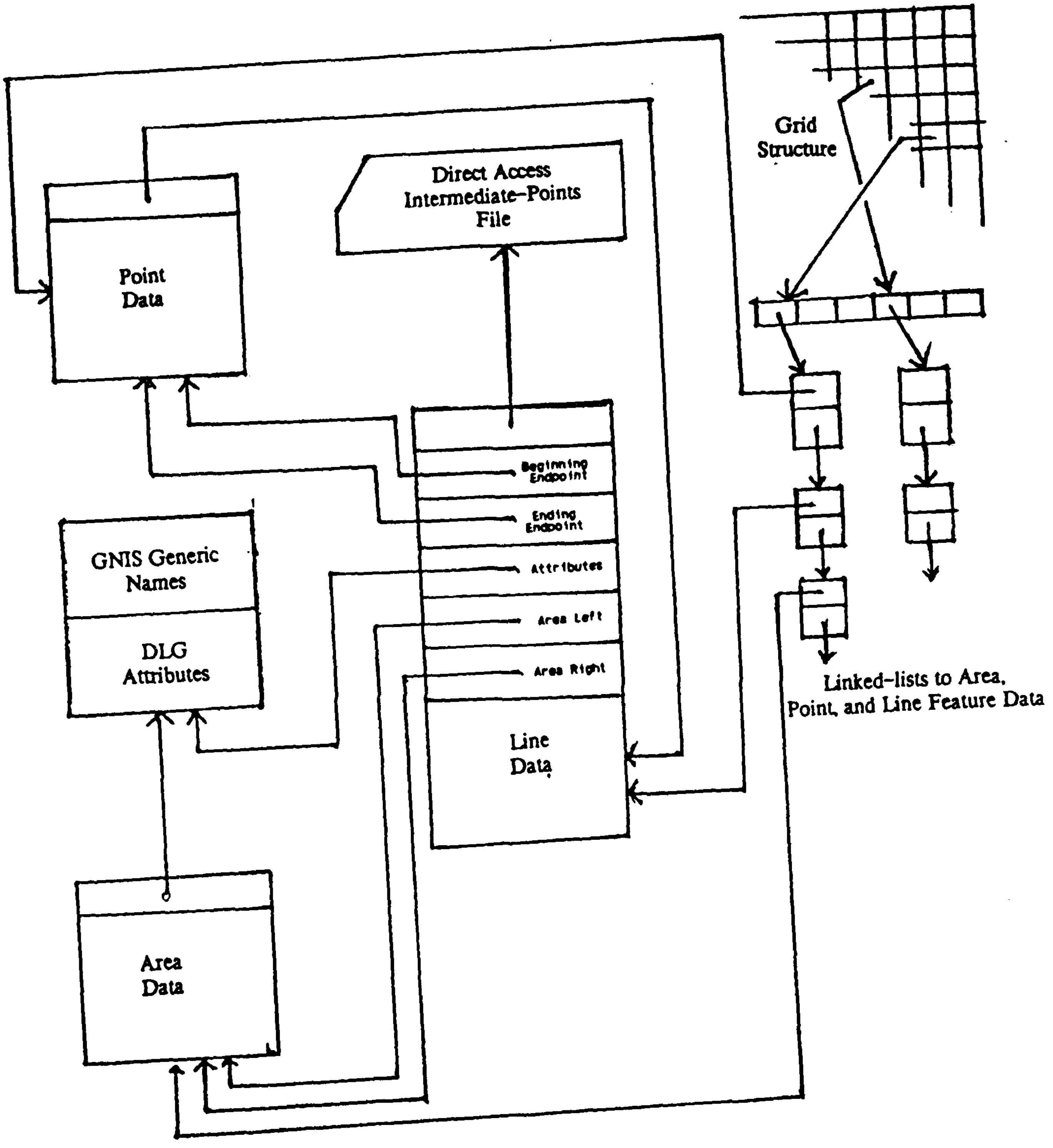


Figure 4: Diagram of the Internal Software Data Structure.

5. Point-Feature Correlation

Of the three data types, point-feature correlation is perhaps the easiest to achieve. The process may simply involve finding the best match between reference points from the DLG and GNIS data, and much of the time, it may simply involve translating the latitude/longitude coordinates from the GNIS data to Albers Equal Area and incorporating the feature into the extracted points data structure, thereby circumventing any need to refer to the DLG data. But because of the large number of point-feature names (particularly populated places -- cities and towns), it becomes necessary to implement a selection criterion. A selection criterion is also necessary in order to differentiate between different classes of feature types. Unfortunately, the data used in this study does not provide a viable selection criterion for populated-places point features. Originally, because there are three different data types for cities and towns in the DLG data, it was thought that a selection criterion would be provided. When the data was examined, however, it was found that there is insufficient data stored under the three data types to constitute a viable selection criterion. It is felt that the GNIS populated-places file, which contains population figures from the 1980 census surveys, would provide a good selection criterion based on population. The effort by Umit Basoglu [2] outlines a more credible approach to the point selection process for populated places.

The GNIS Names files do not differentiate between various type of populated places. A subdivision having one inhabitant is as significant as a major metropolitan center such as New York City. To compensate for this shortcoming, the most significant cities and towns from the GNIS names data were manually chosen based on their representation on the National Atlas 1:2,000,000 scale map of the northeastern United States. Having created, then, an albeit limited but sufficient names data base for cities and towns, a selection criterion is possible based on the populated-places class chosen for each figure during manual selection. Figure 5 is an AUTONAP plot exclusively of cities' and towns' names selected by the correlating software from the manually created populated-places names data base.

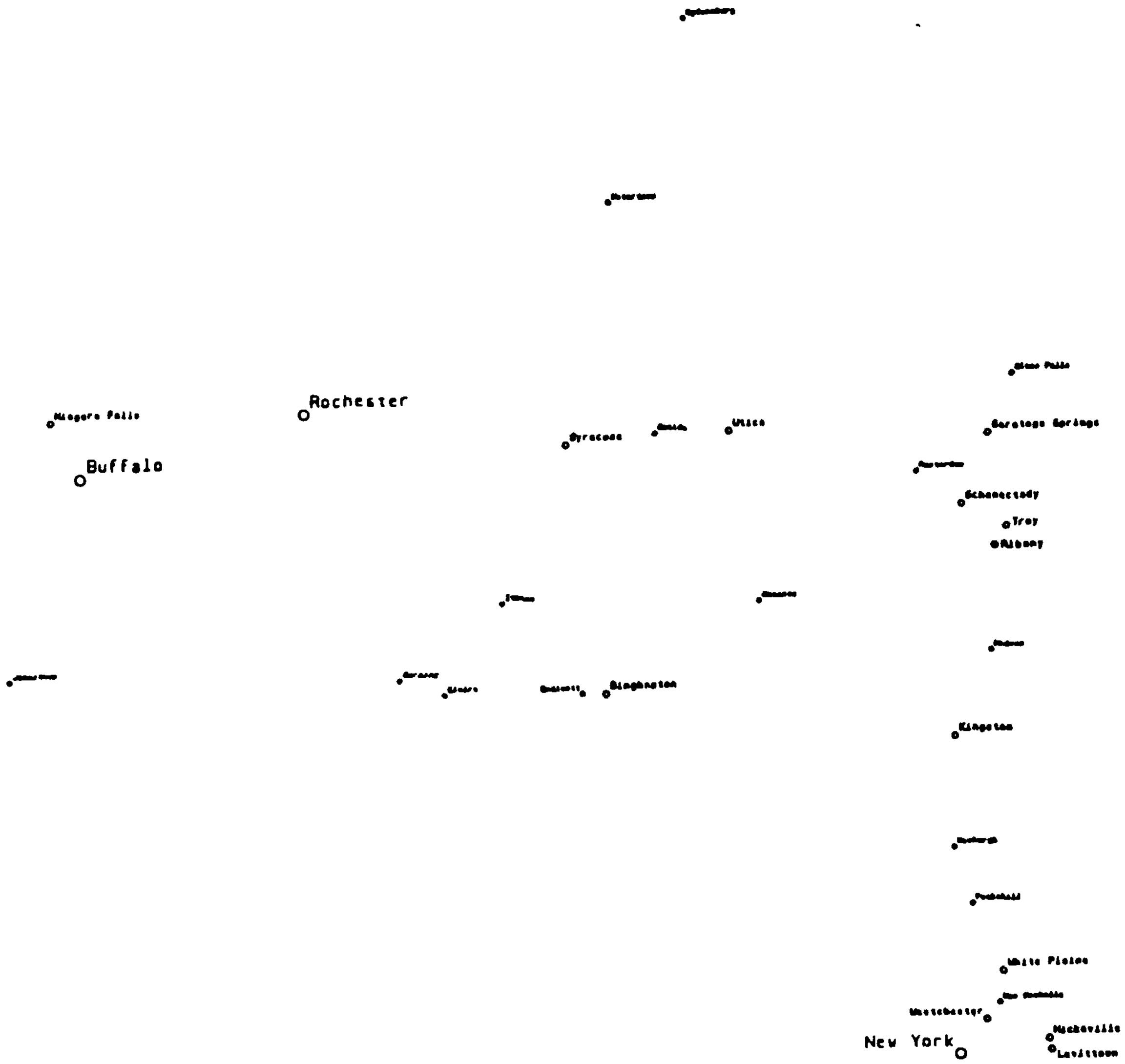


Figure 5: Populated-Places Names Placement

6. Line-Feature Extraction and Names Correlation

6.1 Selection Criterion For Linear Features

The selection algorithm for linear features is straight-forward. The user selects the attribute codes which represent the types of data he or she wishes to extract and have named (the codes are listed in [3]). Subsequently, all segments which match a code and all line segments which are attached and extend the feature to which they belong are extracted.

6.2 Heuristics For Feature Extraction

The problem of feature extraction is a particularly crucial and difficult one for linear features. The amount of information associated with a linear feature name is limited; beyond the generic feature type, there remains only the geographic primary and source coordinates. To obtain the greatest likelihood of successful names-to-features correlation, as much information must be determined about a feature as possible. Finding a segment which matches a chosen attribute is straight-forward. Determining which segment (if any) extends that segment as a part of the desired feature presents some difficulties. A human may have some a priori notion about such choices; an objective of this study is to quantify these pertinent notions in order to provide a credible method for automatically making such choices.

A major problem is the forking river dilemma (see Figure 3). Consistent rules in nature are hard to come by. A river tends to flow in many directions until it reaches its mouth. However, when a river forks, or a tributary enters it, the extension of the river seems consistently to be more in line with the generalized path of the already extracted river than does the segment which joins or leaves it. Hence, when a choice has to be made between two or more river segments, the generalized slopes of the segments are calculated and compared against the generalized slope of the partially extracted linear feature. The generalized slope is simply the slope calculated from the two endpoints of a line. The segment which best matches the partial feature's slope is chosen. Fortunately, the forking-river dilemma does not occur often due to the relatively large number of classifying attributes about linear features; many intersecting line segments are eliminated by virtue of their incompatible attributes. The centerline of a double-lined river, however, has many identically attributed branches coming off of itself, and, thus, provides a good test base. In this case, the heuristic works without flaw in eliminating the branches and choosing the correct centerline.

Another inhibiting factor in linear-feature extraction is the discontinuity between segments of a linear feature. When an ending node of a line segment is encountered which does not have an intersecting line segment to continue it, surrounding nodes have to be searched in order to be sure that the actual ending node has been found. Furthermore, this search must be sure not to include surrounding nodes of segments which do not actually belong to the feature being extracted.

All nodes within a certain distance threshold are examined to see whether they are intersected by line segments with compatible attributes. All such line segments are called candidates for extension. If a candidate is found, it is examined to see that it has not already been extracted (meaning another feature has already claimed it). If the candidate has not already been extracted, then the node is examined to see that there is no other line segment intersecting that node which is compatible with the candidate. If this third line segment is found, then it is most likely that it and the candidate are extensions of each other and not of the feature. In summation, all neighboring nodes of a potential linear feature endpoint are searched for the following:

1. A candidate for extension of the line feature intersects the node.

2. The candidate has not already been extracted.

3. There are no other lines intersecting the node which are compatible with the candidate.

If such a node is found, then the line is extended with the candidate of the node. This heuristic seems to work with relatively no flaw. A problem does occur when the river passes in and out of the map boundary. The Susquehanna River is an example where the algorithm defines two separate features instead of one because of the gap created by the missing section of the river outside the map boundary. It is virtually impossible for an automated algorithm to state with certainty that the two sections are part of one feature.

6.3 Heuristics For Linear Names Correlation

The GNIS data, as stated previously, holds little in the way of information concerning a geographic description of a river. The primary (mouth) coordinate of a feature should be relatively close to the coordinate determined from the DLG data, but the method through which the source coordinate is chosen for both files makes it a weak factor in determining which name belongs to which feature.

Rivers in the DLG data are classified as perennial single lines, intermittent single lines, or double lines. All single lines are sub-classified by length. Although the straight distance between the source and mouth of a river is not necessarily a precise measure of the length of a river, it certainly is an indication. One heuristic, then, in determining a match between feature and name is to match a threshold based on the smallest sub-classification of river length selected for extraction against the distance between the source and mouth coordinates of a name from the GNIS file. There are also clues within the name of the feature. Most of the features stored under the generic 'river' are actually small, insignificant streams. It was discovered that a lot of time was being spent trying to match these insignificant features against the larger rivers from DLG data. Many of these insignificant features contained more descriptive generics within their names. In trying to match the names of more significant rivers, then, it became prudent to eliminate all the features whose names contained generics such as 'Brook', 'Stream', or 'Creek'. Most other insignificant features, whose names do not yield any informative generics, would be eliminated by the distance test.

If a line segment defined by the two endpoints of a linear feature has the same or similar linear equation as the line segment defined by the source and mouth coordinates of a name, then the likelihood of a match is good. The closer the two equations are, the greater the likelihood of their match. This determination provides the basis for computing a weight to associate with each potential match. Not reflected in this, however, is the fact that the primary mouth coordinate of the name is more likely to match a feature endpoint than the source coordinate. Hence, the weight is determined by considering the distances from the endpoints of a line to the corresponding source or mouth coordinate. Since the linear equation of the line segments are only a function of the two endpoints, all the characteristics of the linear equation match are necessarily incorporated, but an extra characteristic is added in that the distance between the mouth coordinate and a feature endpoint may be given the greater influence in computing the weight.

If a name is not eliminated by the generic and distance tests, then the grid cell in which the primary coordinate of the name lies and neighboring grid cells are searched for an endpoint to a line feature compatible with the GNIS file generic of the name. (In this study, the generic was always 'river'.) If such a node and feature is found, a weight is computed based on twice the inverse distance between the node and the mouth coordinate plus only one times the inverse distance between the other endpoint of the feature and the source coordinate of the name. If the weight indicates a better match than any previous match against the the feature (the weight is larger than a previous weight) and is also the largest weight found against the name, then the name is matched at least temporarily with the feature.

The results indicate that this heuristic provides an adequate matching algorithm for linear features. Wrong associations between features and names were made when a feature passed in and out of the map boundary and back in again. Figure 6 shows an example of such a matching. The river labelled Cowanesque is actually a part of the Susquehanna. As one would expect, the part of the Susquehanna which has either the source or the mouth on the map is correctly matched. The Genesee River has its mouth outside of the map boundary, and so is incorrectly labelled as the Tioga. The Housatonic River is never in New York state, and therefore is not in the GNIS file. The mouth of the Titicus is apparently near either the source or the mouth of the Housatonic.

One may note that the secondary coordinate would have been a useful determinate in matching names to linear features. Because it is guaranteed to be within the map boundaries, the secondary coordinate could only increase the certainty of a match and would help eliminate the problem where incorrect matches are made against features whose source or mouth coordinates fall outside of the map boundary. Unfortunately, however, the secondary coordinates were not included with the GNIS data that we received.

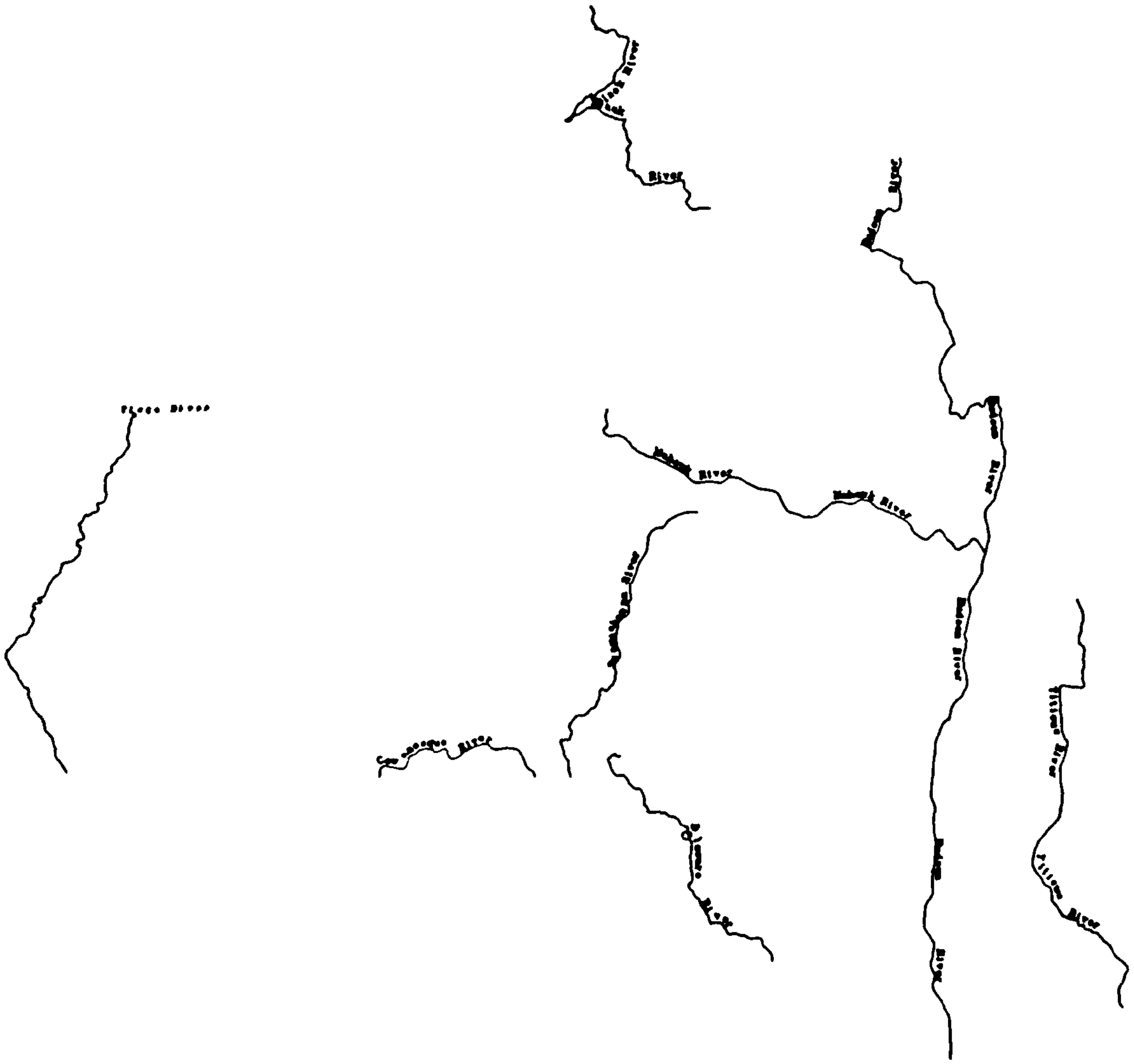


Figure 6: Linear Names Match and Placement

7. Area-Feature Extraction and Names Correlation

7.1 Area-Feature Selection Criterion

Like the line features, area-feature selection is straight-forward. Two area features were used in this study: states and lakes. State features are chosen by the user who supplies the FIPS code for the state. The state feature chosen determines the boundary for the extracted map. Other area features are chosen by the user by supplying the attribute codes [3] of the boundary lines for the areas which he or she may want extracted and labelled. Currently, the software is only equipped to match lake features to names, but any feature type may be chosen for extraction.

7.2 Heuristics For The Extraction Of Areal Features

7.2.1 Extracting The State's Effective Boundary

The political boundaries overlay, as stated before, does not provide shoreline data and, therefore, does not provide sufficient data for determining the effective state boundary necessary for name placement. The water bodies overlay, which contains shoreline data, is then necessary for the extraction of the boundaries of states which border an ocean or a Great Lake. All shoreline line segments are stored in a grid structure. If a shoreline passes through a grid cell, it is referenced in the linked list corresponding to the cell. Another grid is set up to store all the nodes which have only one intersecting state land-boundary segment.

The line data structure is searched until a line segment is encountered which has the attribute of being the boundary of the state to be extracted. The boundary of the state is traversed counter-clockwise, extending the boundary with connecting state boundaries until the initial line segment is encountered or until no connecting state boundaries can be found. Figure 7 is the extracted boundary for Vermont which has no water body boundary.

When no connecting boundary is found, then the grid area where the dead-end node resides is searched for a nearby shoreline. When a shoreline is found, its nearest intermediate point is taken as the intersection of the political boundary and the shoreline. Since the separate overlays were also digitized separately, the intersection between shoreline and political boundary is almost never exact, but is always close enough to ensure certainty.

At this point, a question arises as to which direction to continue the traversal of the shoreline in extracting the state boundary. Because the boundary is being traversed in a counter-clockwise direction, the state must be to the left of the bounding shoreline. Hence, because a shoreline should never cross over itself, the direction of traversal is chosen which describes the left-most turn from the intersecting political segment.

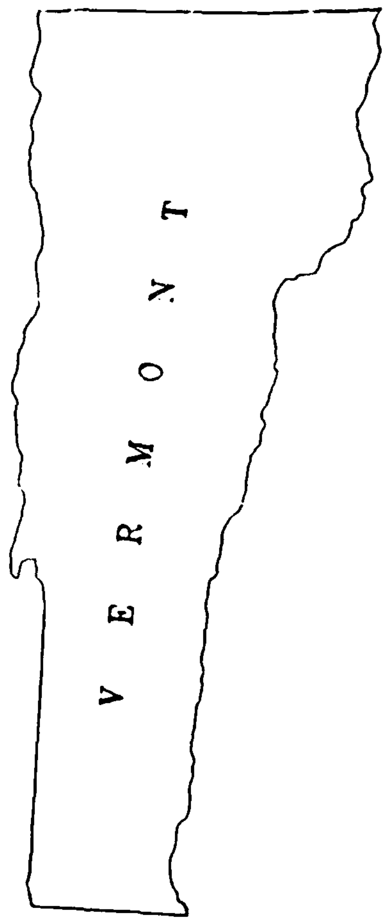


Figure 7: Extracted State Boundary With No Water Body Border.

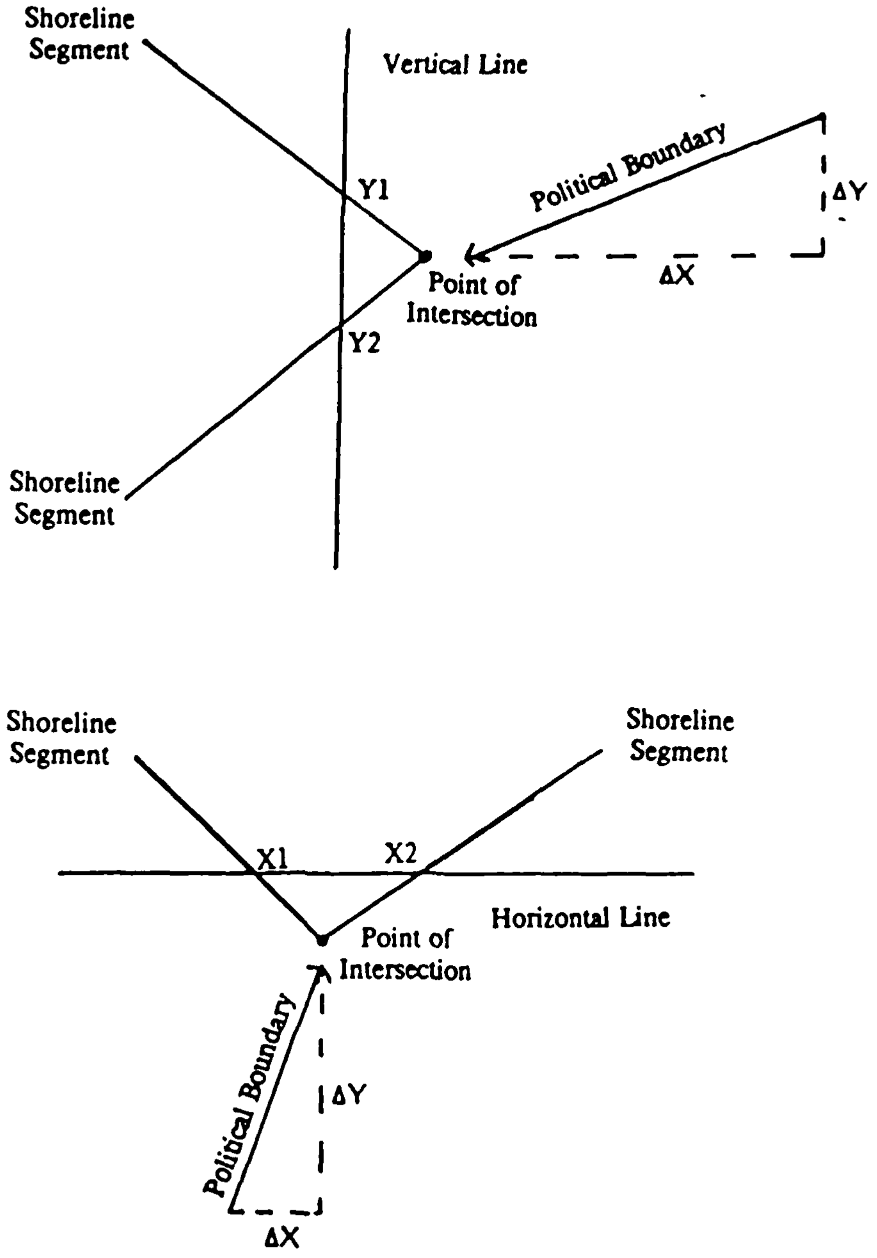


Figure 8: Computing the Best Left Turn

As shown in Figure 8, If the change in the Y-direction is greater than the change in the change in the X-direction, then a horizontal line is drawn just beyond the point of intersection (which can be thought of as the vertex of an angle between the two segments of the shoreline). If the change in the X-direction is greater, then a vertical line is drawn. Depending on whether the change is negative or positive, a correct choice as to the left-most turn can be made by comparing the corresponding X or Y values where the vertical or horizontal lines intersect the two segments. There are special cases where one or both the line segments do not intersect the horizontal or vertical line; but in general :

1. If $|DX| > |DY|$ & $DX < 0$ The line segment whose Y-value at its intersection with the vertical line is less is chosen.

2. If $|DX| > |DY|$ & $DX > 0$ The line segment whose Y-value at its intersection with the vertical line is greater is chosen.

3. If $|DX| < |DY|$ & $DY > 0$ The line segment whose X-value at its intersection with the horizontal line is less is chosen.

4. If $|DX| < |DY|$ & $DX < 0$ The line segment whose X-value at its intersection with the horizontal line is greater is chosen.

The shoreline is traversed in the chosen direction, checking at each intermediate point for a nearby node which contains the continuation of the political boundary of the state. When this node is encountered, traversal is continued in the normal fashion.

For New York, the effective boundary consists of two separate intersections of the political and shoreline data, making it a particularly good test for the extraction algorithm. The result is shown in Figure 9.

7.2.2 Extraction of Other Areas

Other area boundaries are more straight-forward. When an area boundary segment is encountered, it is extended in much the same way as the state's political boundary was extended. The area border is traversed in a counter-clockwise direction, keeping the area being extracted to the left of the direction of traversal. The node at the endpoint of the boundary is scanned for a segment with the same relevant attribute as the previous segment and with the area being extracted to its left (in the direction of the traversal -- it may actually be to the right, depending on the direction in which the line was digitized). As each area is extracted, it is stored in a linked list attached to all the grids it spans.

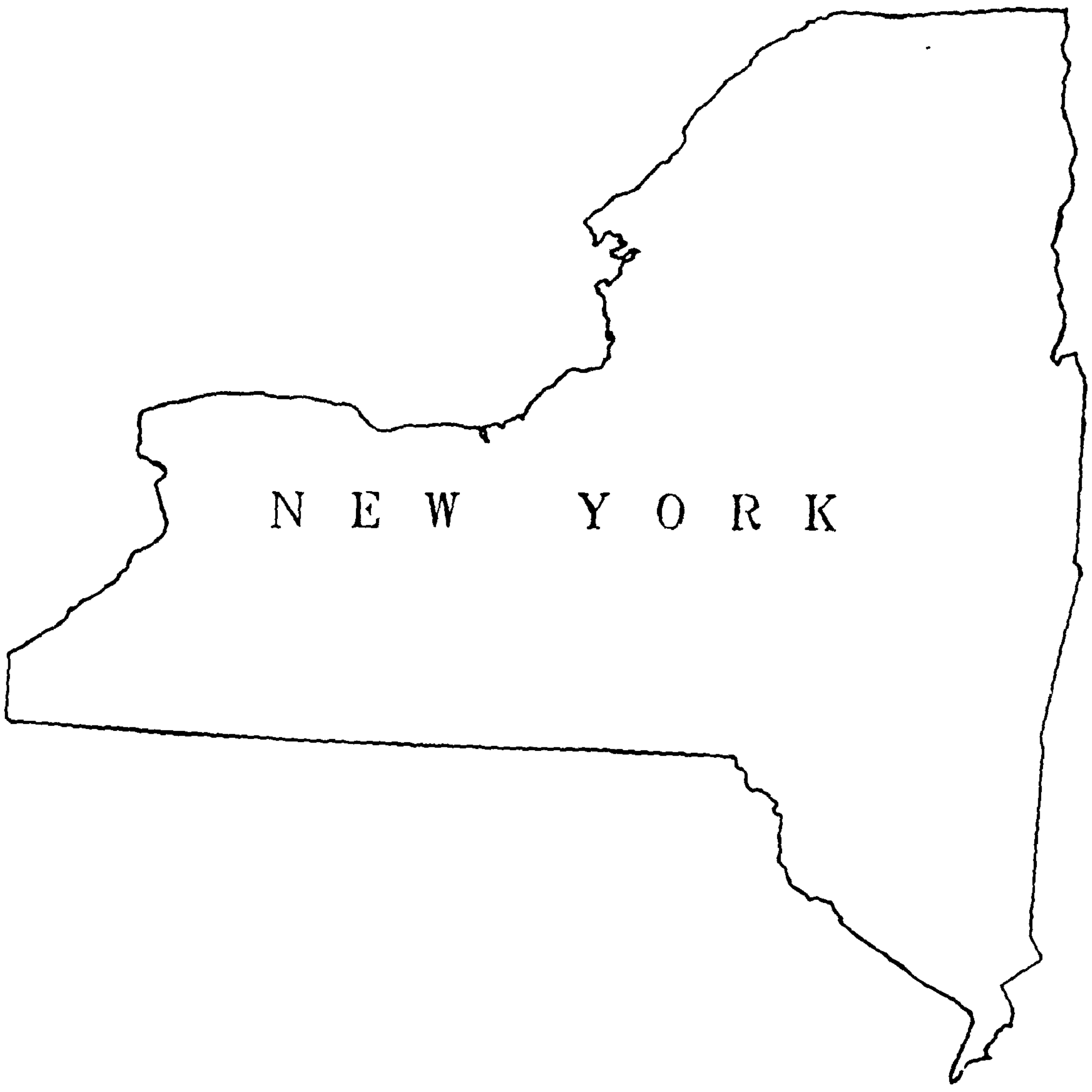


Figure 9: New York State's Effective Boundary With Two Separate Shorelines

7.3 Heuristics For Areal Names Correlation

Because state names are determined from the FIPS code recorded by an attribute attached to the boundary segments, there is no uncertainty about their names. Other area features, on the other hand, must be correlated with their names extracted from the GNIS data.

The features used in this study were lakes. The reference points associated with each name are purported to be the approximate centers of the areas to which they belong. The reference point from the DLG data, on the other hand, may not even lie in the interior of its area. The grid area in which the name's reference point lies is searched for an extracted area feature of compatible type with the generic of the name. An algorithm is implemented to test whether a name's reference point is contained within the area defined by the line segments of the area's border. The specific algorithm, exemplified in Figure 10, is one which extends a horizontal line to the right of the reference point. If the line extended to the right intersects the boundary of the area an odd number of times, then the point is contained within the area. The special cases where the horizontal line coincides with the border are handled by only considering the points on the border immediately before and and after the region of coincidence. Although not specifically handled by the software developed in this study (AUTONAP cannot handle areas with both an internal and external border such as lakes with islands), this algorithm is particularly well suited for excluding occurrences of areas contained within the external boundary of a larger, similarly attributed area.

If it is determined that a reference point is contained within an area, then a weight measuring the viability of the match is assigned the value of 1. Normally, this is enough to ensure a match, but due to either differences in the scaling or simply in the digitization process, the reference point may not lie in the DLG area. Also, it could be that more than one lake name's reference point will lie in the DLG area. To compensate for this, the inverse distance between the DLG reference point and the GNIS reference point is added to the weight. The match with the heaviest weight will be preserved.

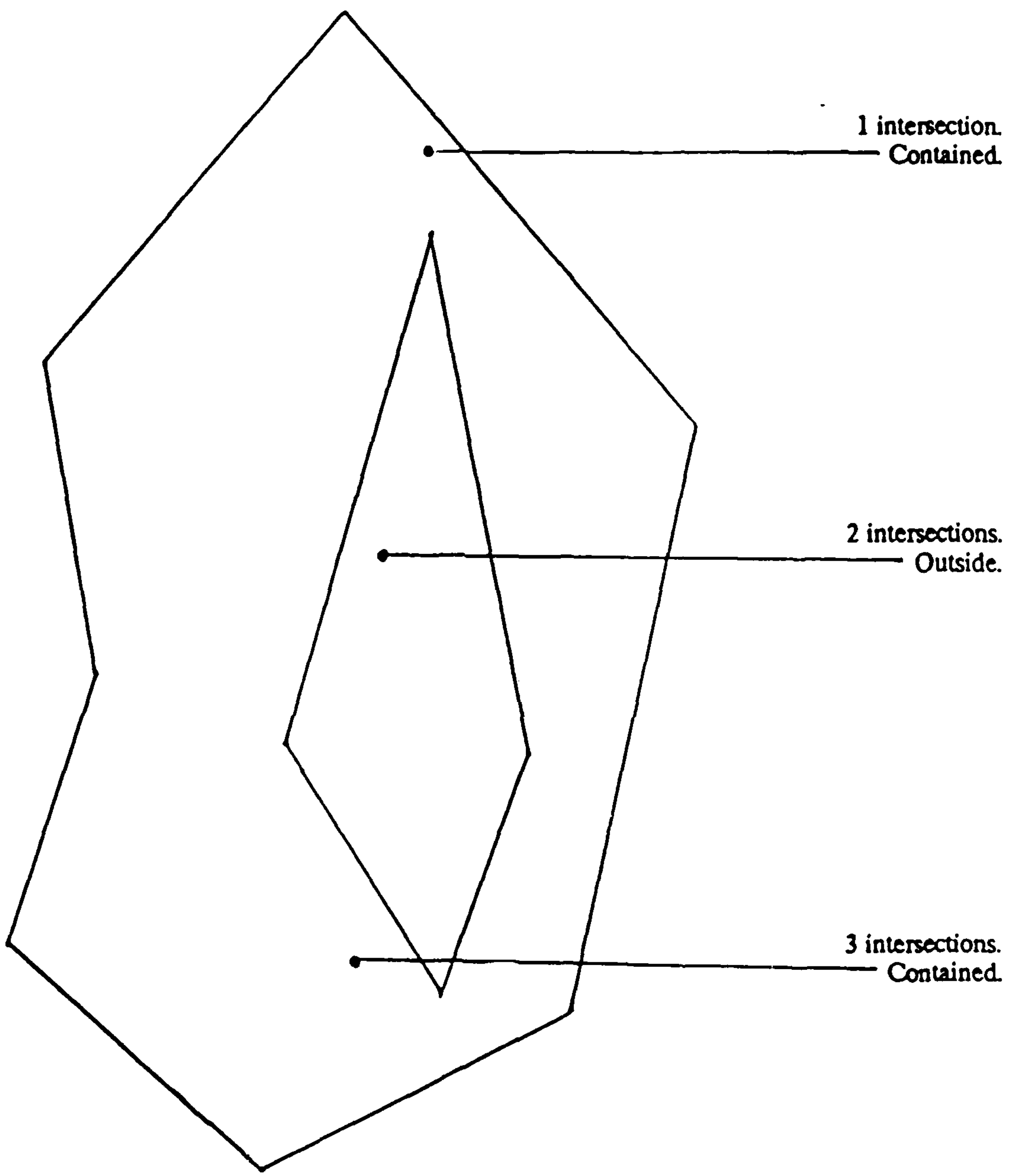


Figure 10: Odd Number Of Intersections Containment Algorithm

Figure 11 shows two of the lake features matched with names. The tests against this algorithm prove to be quite encouraging. One discrepancy, however, does occur. Some lakes are not referenced in the GNIS Names file under the generic 'lake'; they are instead referenced under the generic 'tank'. There is no documentation concerning this generic but it was assumed that 'tank' was synonymous to 'reservoir'. In any case, this seems to be an unreasonable determination. Cayuga Lake, one of the largest of the Finger Lakes in New York, is labelled a 'tank' although it certainly is a natural lake. Cayuga is the unlabelled lake next to Seneca in Figure 11. The problem arises from the fact that the 'tank' features' reference points are digitized outside the boundary of the feature to which they refer, possibly at the point where a dam is thought to be. This makes it virtually impossible to determine a proper match. The fact that there are so many occurrences of lakes categorized as 'tanks' makes this a severe problem. Lakes whose names are categorized as 'lakes', are, without failure, properly matched.

Figure 12 is a map with point, line, and area features generated automatically by AUTONAP with data comprised from AUTOCOR, the software developed for this study. Because point, line, and area feature correlation represent three independent processes, there are no extra difficulties imposed by having the three types of data correlated and merged together into one cartographic data file.

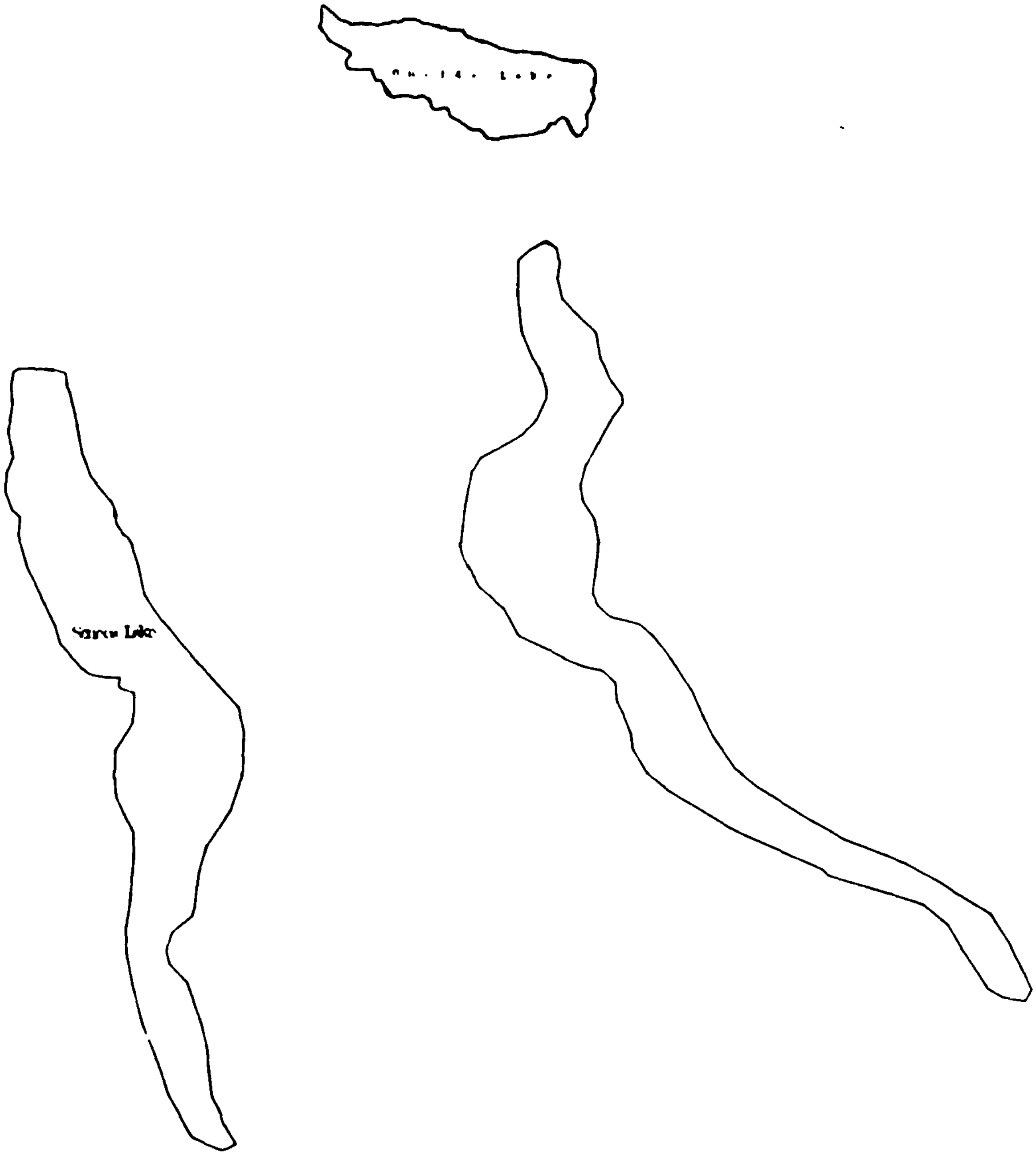


Figure 11: Names Correlated and Matched With Lakes

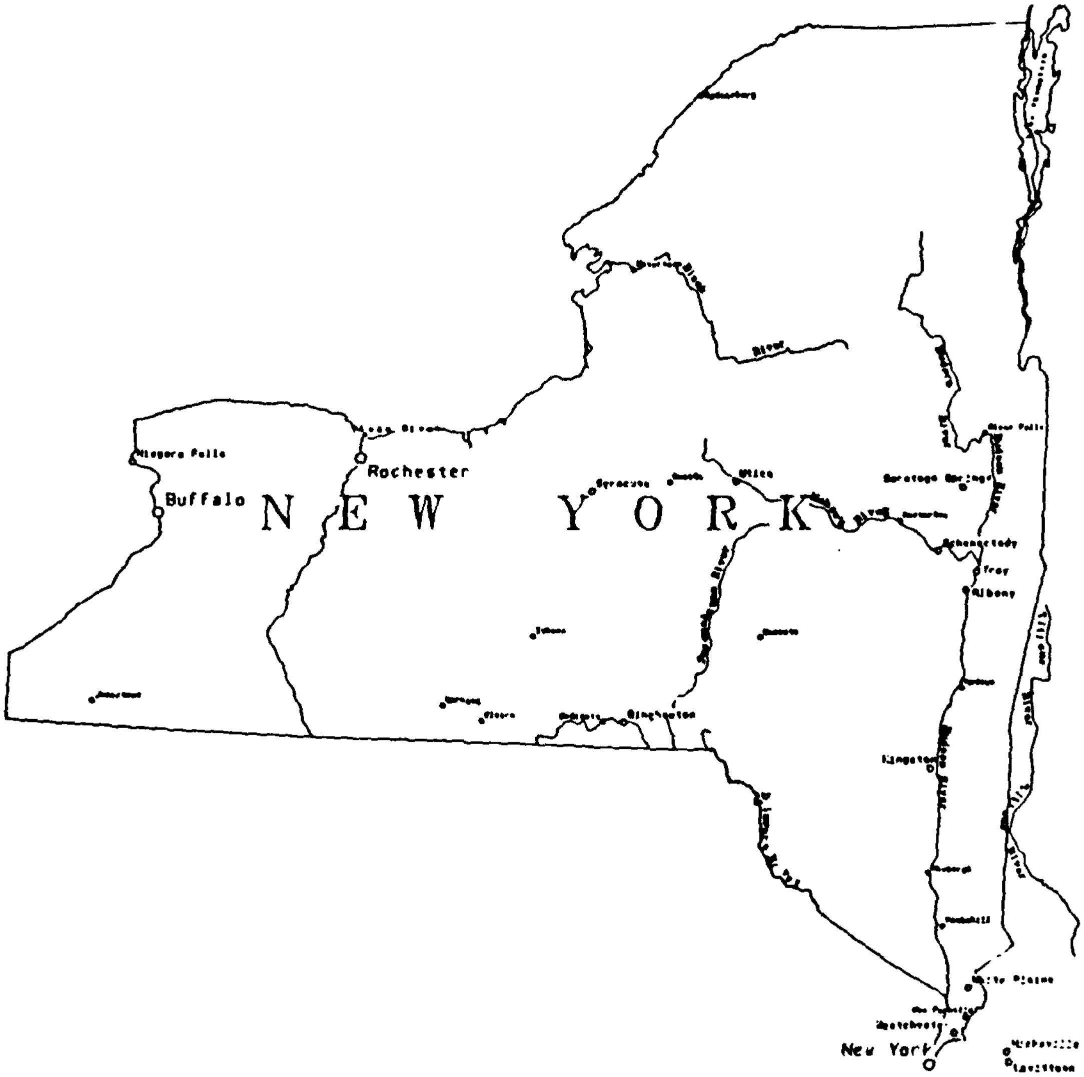


Figure 12: Point, Line, and Area Features Correlated Together

8. Conclusion

The research described in this paper is for the most part based on empirical testing methods. Several versions of each algorithm were designed before a final one was chosen. Each version has its own advantages and disadvantages; such is the nature of automated cartography. The results are quite promising. The successes within this research indicate that the GNIS and DLG data files can be merged together with some reasonable credibility. Particularly, area features, when categorized properly, showed a good percentage of proper matches. Linear features were more difficult, but showed good results when the entire feature is contained in and can be extracted from the DLG data.

The failures in this research indicate room for improvement in both the algorithms and the data. Many of the failures can be attributed to lack of data integrity, such as lakes being called 'tanks'. Other inadequacies in the data, such as disconnected line segments of the same feature, were overcome. A key element that limited the performance was the fact that the project tried to link 1:2,000,000 DLG data with 1:24,000 names data. There is little doubt that many of the difficulties encountered resulted from this wide disparity in scale.

As the integrity of the data and the algorithms that use them are improved, incorrect matches will decrease and correct matches will become more common. It is then a conclusion of this study that the automatic generation of cartographic data from GNIS and DLG data is an attainable prospect. [4, 5]

References

1. J. Kwangho Ahn. "Automatic Name Placement System". *IPLab IPL-TR-063* , (May 1984).

2. Basoglu, Umit. "A New Approach to Automated Name Placement". *Fifth ISymp on CAC Crystal City, Virginia* , (August, 1982), .

3. Domaratz, Hallman, Schmidt, Calkins. "Digital Line Graphs From 1:2,000,000-Scale Maps". *USGS Digital Cartographic Data Standards* , (1983).

4. Payne, Roger L. "Geographic Names Information System Users Guide". *Department of the Interior, U.S. Geological Survey, National Mapping Division* , (1984), 84-551.

5. Payne, Roger L. "Geographic Names Information System". *USGS Digital Cartographic Data Standards* (1983), .