A Mathematical Structure for Analyzing Maps

JOSEPH K. BERRY
Yale University
School of Forestry and Environmental Studies
205 Prospect Street
New Haven, Connecticut 06511, USA

ABSTRACT / The growing use of computers in environmental management is profoundly changing data collection procedures, analytic processes, and even the decision-making environment itself. The emerging technology of geographic information systems (GIS) is expanding this revolution to integrate spatial information fully into research, planning, and management of land. In one sense, this technology is similar to conventional map processing involving traditional maps and drafting aids, such as pens, rub-on shading, rulers, planimeters, dot grids, and acetate sheets for light-table overlays. In another sense, these systems provide advanced analytic capabilities, enabling managers to address complex issues in entirely new ways. This report discusses a fundamental approach to computer-assisted map analysis that treats entire maps as variables. The set of analytic procedures for processing mapped data forms a mathematical structure analogous to traditional statistics and algebra. All of the procedures discussed are available for personal computer environments.

The historical use of maps has been for navigation through unfamiliar terrain and seas. Within this context, preparation of maps that accurately locate special features became the primary focus of attention. More recently, analysis of mapped data has become an important part of resource and environmental planning. During the 1960s, manual analytic procedures for overlaying maps were popularized by the work of McHarg (1969) and others. These techniques mark an important turning point in the use of maps—from one emphasizing physical descriptors of geographic space, to one spatially characterizing appropriate land management actions. This movement from descriptive to prescriptive mapping has set the stage for revolutionary concepts of map structure, content, and use.

Spatial analysis involves tremendous volumes of data. Manual cartographic techniques allow manipulation of these data, but they are fundamentally limited by their nonquantitative nature. Traditional statistics, on the other hand, enable quantitative treatment of the data, but, until recently, the sheer magnitude of mapped data made such processing prohibitive. Recognition of this limitation led to "stratified sampling" techniques developed in the early part of this century. These techniques treat spatial considerations at the onset of analysis by dividing geographic space into homogeneous response parcels. Most often, these parcels are manually delineated on an appropriate map, and the "typical" value for each parcel determined. The results of any analysis is then assumed to be uniformly distributed throughout each parcel. The area-weighted average of a set of parcels statistically characterizes the typical response for an extended area. Mathematical modeling of spatial systems has followed a similar approach of spatially aggregating variation in model variables. Most ecosystem models, for example, identify "level variables" and "flow rates" presumed to be typical for vast geographic expanses.

A comprehensive spatial statistics and mathematics, for many years, has been variously conceptualized by both theorist and practitioner. Until modern computers and the computerized map, these concepts were without practical implementation. The computer has provided the means for both efficient handling of voluminous data and the quantitative analysis required by these concepts. From this perspective, the current revolution in mapping is rooted in the digital nature of the computerized map. Increasingly, geographic information system (GIS) technology is being viewed as providing new capabilities for expressing spatial relationships, as well as efficient mapping and management of spatial data. This report discusses the fundamental consideration of the emerging "toolbox" of analytic capabilities.

KEY WORDS: Computers; Digital map analysis; Geographic information systems

Geographic Information Systems

The main purpose of a geographic information system (GIS) is to process spatial information. These systems can be defined as internally referenced, automated, spatial information systems—designed for data mapping, management, and analysis. These systems have an automated linkage between the type of data, termed the thematic attribute, and the whereabouts of that data, termed the locational attribute. This structure can be conceptualized as a stack of "floating maps," with common spatial registration, allowing the user to "look" down and across the stack. The spatial relationships of the data can be examined (that is, inventory queries) or manipulated (that is, analytic processing). The locational information may be organized as a collection of line segments identifying the boundaries of point, linear, and areal features. An alternative organization establishes an imaginary grid pattern over a study area and stores numbers identifying the characteristic at each grid space. Although there are significant practical differences in these data structures, the primary conceptual difference is that the grid format stores information on the interior of areal features, and implies boundaries; whereas, the segment format stores information about boundaries, and implies interiors. Generally, the line segment structure is best for inventory-oriented processing, while the grid structure is best for analysis-oriented processing. The difficulty of line segments in characterizing spatial gradients, such as elevation or housing density, coupled with the frequent necessity to compute interior characteristics limit the use of this data type for many of the advanced analytic operations. As a result, most modern GIS contain programs for converting between the two data structures.

Regardless of the data storage structure, all GIS contain hardware and software for data input, storage, processing, and display of digital maps. The processing functions of these systems can be grouped into four broad categories:

- Computer mapping
- Spatial data base management
- Spatial statistics
- Cartographic modeling

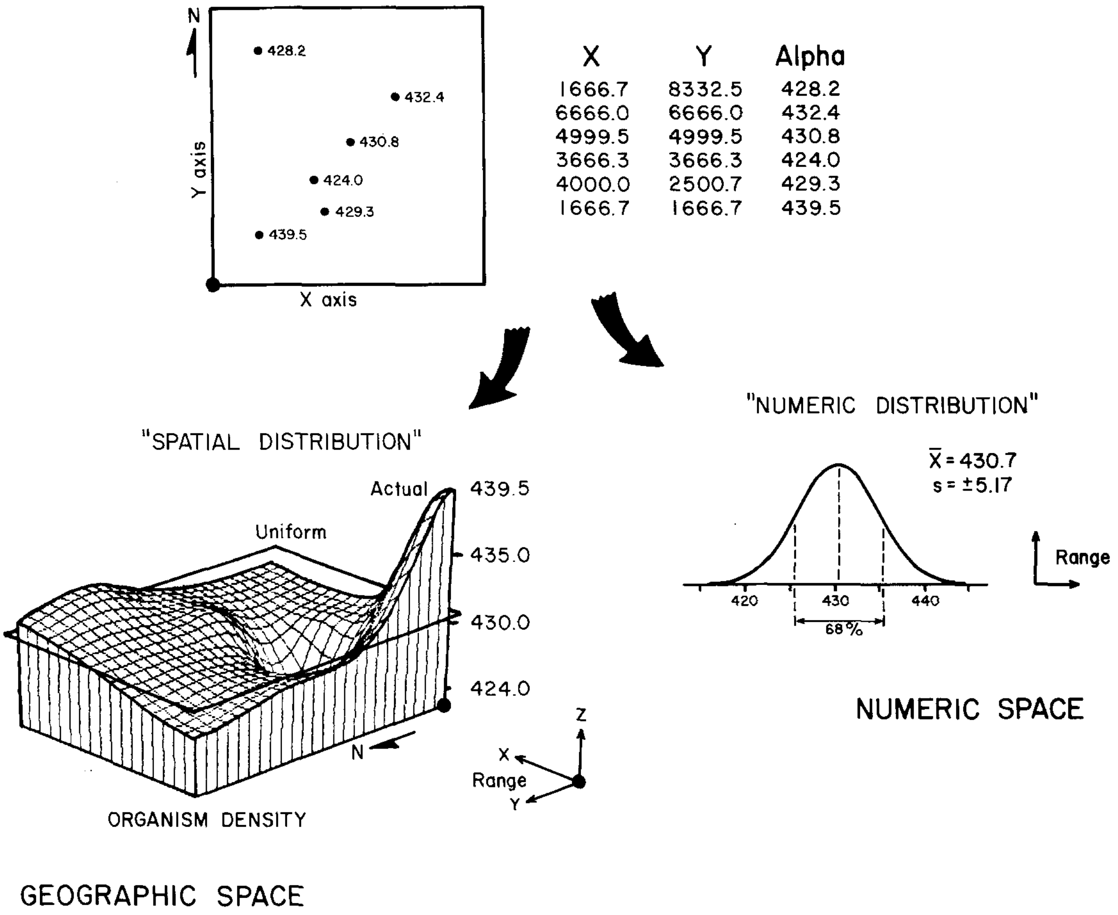
Most GIS contain some capabilities from each of these categories. An inventory-oriented system will emphasize mapping and management functions, whereas, an analysis-oriented system will focus on statistics and modeling functions.

Computer mapping, also termed automated cartography, involves the preparation of map products. The focus of these operations is on the input and display of computerized maps. Spatial data base management, on the other hand, focuses on the storage component of GIS technology. Like nonspatial data base systems, these procedures efficiently organize and search large sets of data for frequency statistics and/or coincidence among variables. For example, a spatial data base may be searched to generate a map of areas of silty-loam soil, moderate slope, and ponderosa pine forest cover. These mapping and data base capabilities have proven to be the backbone of current GIS applications. Once a data base is compiled, they allow rapid updating and examining of mapped data. However, other than the significant advantage of speed, these capabilities are similar to those of manual techniques. The remainder of this report investigates the emerging analytic concepts of spatial statistics and cartographic modeling.

Spatial Statistics

Spatial statistics seeks to characterize the geographic pattern or distribution of mapped data. This approach differs from traditional statistics as it describes the spatial variation in the data, rather than distilling the data for typical responses that are assumed to be uniformly distributed in space. For example, consider the hypothetical data presented in Figure 1. The large square at the top is a map of sample locations for a portion of a lake. The tabular data to the right identify both the location and microorganism density determined from laboratory analysis of the surface water samples. Traditional statistics would analyze the data by fitting a numerical distribution to the data to determine the typical response. Such density functions as standard normal, binomial, and Poisson could be tried, and the best-fitting functional form chosen. The lower-right inset characterizes the fitting of a "standard normal curve" indicating an average organism density of 430.7 ± 5.17 units. These parameters describing the central tendency of the data in numerical space are assumed to be uniformly distributed in geographic space. As shown in the lower-left plot, this assumption implies a horizontal plane at 430.7 units over the entire area. The geographic distribution of the standard deviation would form two planes (analogous to "error bars") above and below the average. This traditional approach concentrates on characterizing the typical thematic response, and disregards the locational information in the sampled data.

Spatial statistics, on the other hand, incorporates locational information in mapping the variation in thematic values. Analogous to traditional statistics, a density function is fitted to the data. In this instance, the distribution is characterized in geographic space, rather than numerical space. To conceptualize this process, visualize a pillar at each sample location rising above the lake in Figure 1 to the height of its thematic value (that is, measured organism density). The lower-left inset shows a continuous surface that responds to the peaks and valleys implied by the pillars of the sampled data. This surface was fitted by "inverse-distance-squared weighted averaging" of the sample values using an inexpensive personal computer software package (Golden 1985). The distribution shows con-



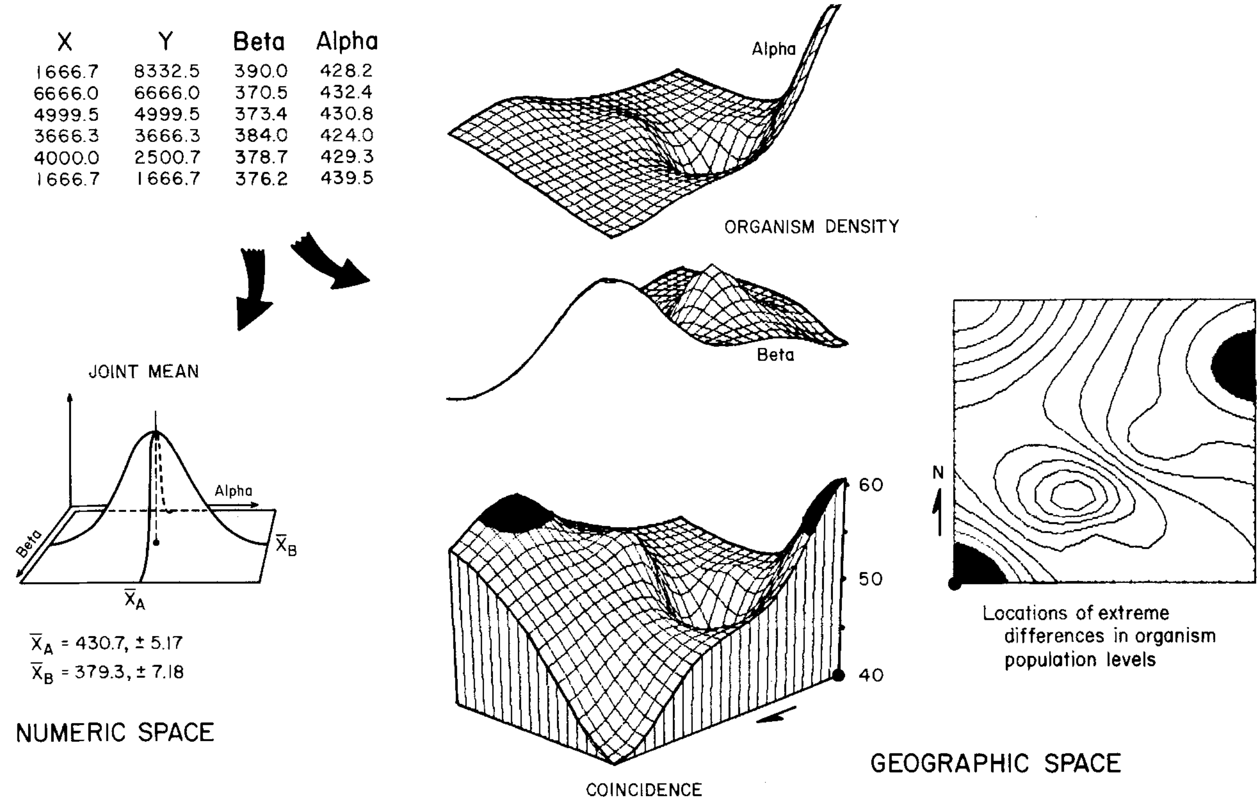
X	Y	Alpha
1666.7	8332.5	428.2
6666.0	6666.0	432.4
4999.5	4999.5	430.8
3666.3	3666.3	424.0
4000.0	2500.7	429.3
1666.7	1666.7	439.5

Figure 1. Spatially characterizing data variation. Traditional statistics identifies the typical response and assumes this estimate to be distributed uniformly in space. Spatial statistics seeks to characterize the geographic distribution, or pattern, of mapped data.

siderable variation in the data in the southwest portion, tending to conflict with the assumption that variation is uniformly distributed in space. Analogous to traditional statistics, other surface fitting techniques, such as Kriging, spline, or polynomial functions, could be tried and the best-fitting functional form chosen. Analysis of the residuals, from comparisons of the various surfaces with the measured values, provides an assessment of fit.

Figure 2 depicts multivariate statistical analysis. If the level of another microorganism was also determined for each water sample, its joint occurrence with the previously described organism could be assessed. In traditional statistics, this involves fitting of a density function in multivariate statistical space. For the example, a standard normal surface was fitted, with its "joint mean" and "covariance matrix" describing the typical paired occurrence. Generally speaking, the second organism occurrs less often (379.3 vs 430.7) with a negative correlation (−.78) between the two populations. The right portion of Figure 2 depicts an analysis of the spatial distributions of the two populations. For this analysis the two maps of population density are compared (that is, one subtracted from the other) to generate a coincidence surface. A planimetric map of the surface, registered to the original map of the lake, is on the right side of the Figure. The darkened areas locate areas of large differences (that is, more than average difference plus average standard deviation) in the two organism populations. The combined information provided by traditional and spatial statistics complements each other for decision making. Characterizing the typical response develops a general impression of the data. The mapping of the variance in the data refines this impression by locating areas of abnormal response. In the example, the approximately 50 units difference in average population densities may not warrant overall concern. However, the pockets of larger differences may direct further investigation, or special management action, within those areas of the lake.

In both statistical approaches, the nature of the



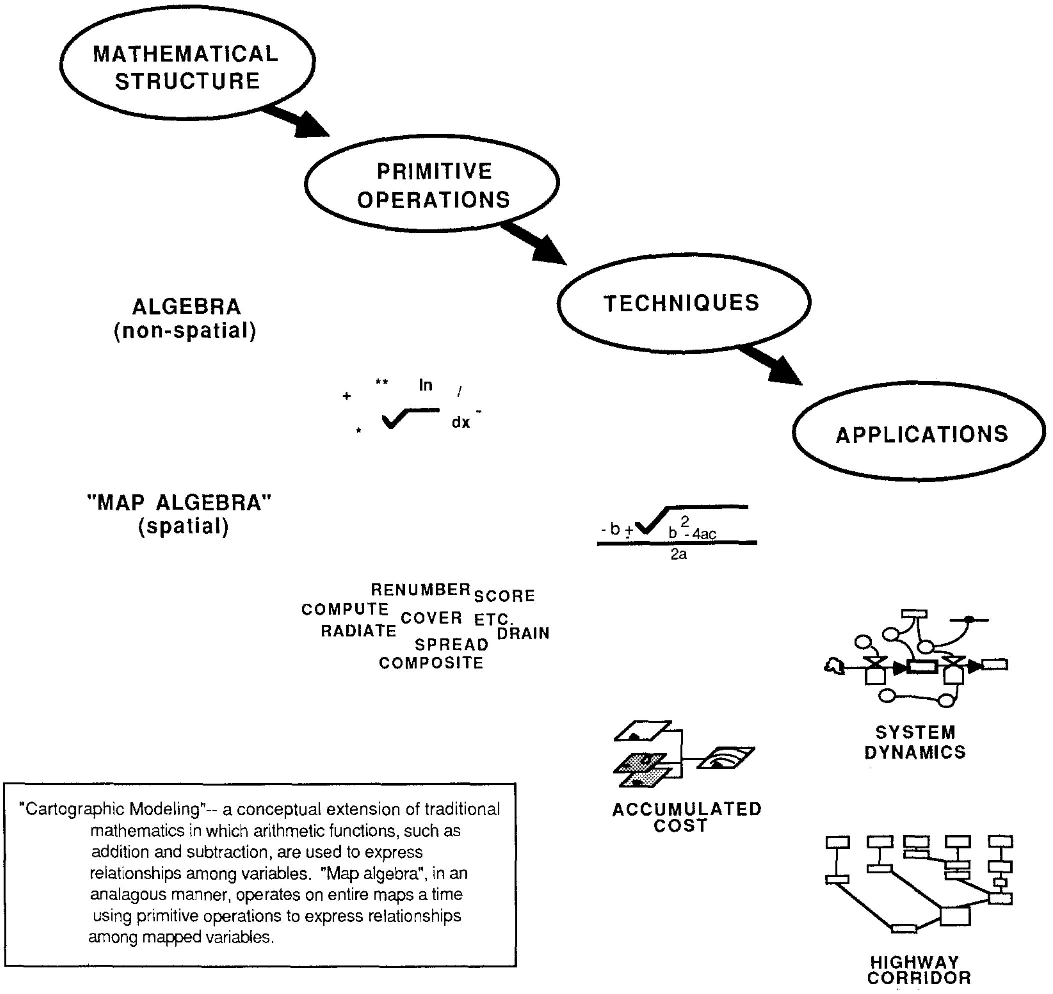
X	Y	Beta	Alpha
1666.7	8332.5	390.0	428.2
6666.0	6666.0	370.5	432.4
4999.5	4999.5	373.4	430.8
3666.3	3666.3	384.0	424.0
4000.0	2500.7	378.7	429.3
1666.7	1666.7	376.2	439.5

Figure 2. Assessing coincidence among mapped data. Maps characterizing spatial variation among two or more variables can be compared and locations of unusual coincidence identified.

data is crucial. For spatial analyses similar to the example, the thematic values must be numerically "robust" and the locational attribute "isopleth." These terms identify map variables that form gradients in both numeric and geographic space. By contrast, a variable may contain "nonrobust" values (for instance, arbitary numbers associated with soil types) that are discontinuous, or "choropleth," in geographic space (for instance, abrupt land use boundaries). A complete discussion of the considerations dealing with the various types of data in spatial statistics is beyond the introductory scope of this article. Both Davis (1973) and Ripley (1981) offer good treatise of this subject.

A Mathematical Structure for Map Analysis

Just as a spatial statistics may be identified, a spatial mathematics is also emerging. This approach uses sequential processing of mathematical primitives to perform a wide variety of complex map analyses (Berry 1985). By controlling the order in which the operations are executed, and using a common database to store the intermediate results, a mathematical-like processing structure is developed. This "map algebra" (Figure 3) is similar to traditional algebra in which primitive operations, such as addition, subtraction, and exponentiation, are logically sequenced for specified variables to form equations; however, in map algebra the variables represent entire maps consisting of numerous values. Most of the traditional mathematical capabilities, plus an extensive set of advanced map-processing primitives, comprise this map analysis "toolbox." As with matrix algebra (a mathematics operating on groups of numbers defining variables), new primtives emerge that are based on the nature of the data. Matrix algebra's transposition, inversion, and diagonalization are examples of extended operations. Within map analysis, the spatial coincidence and juxtapositioning of values among and within maps create new operators, such as proximity, spatial coincidence, and optimal paths. These operators can be accessed through general purpose map analysis packages, similar to the numerous matrix algebra packages. The map analysis package (MAP) (Tomlin 1983) is an example of a comprehensive general purpose system that has been acquired by over 350 computer centers throughout the world. A commercial implementation of this system, the Professional Map Analysis Package (pMAP) (SIS 1986), has recently been released for personal computers. All of the processing discussed in



"Cartographic Modeling"-- a conceptual extension of traditional mathematics in which arithmetic functions, such as addition and subtraction, are used to express relationships among variables. "Map algebra", in an analagous manner, operates on entire maps a time using primitive operations to express relationships among mapped variables.

Figure 3. Cartographic modeling. In mathematics, primitive operators, such as addition and subtraction, are used to express relationships among variables. "Map algebra," in an analogous manner, operates on entire maps at a time using primitive operators to express relationships among mapped variables.

this report was done with this inexpensive pMAP system in a standard IBM PC environment. The logical sequencing of map processing involves:

- Retrieval of one or more maps from the database
- Processing those data as specified by the user
- Creation of a new map containing the processing results
- Storage of the new map for subsequent processing

Each new map derived as processing continues is spatially registered to the other maps in the database. The values comprising the derived maps are a function of the statistical or mathematical summary of the values on the "input maps."

This cyclical processing provides an extremely flexible structure similar to "evaluating nested parentheticals" in traditional algebra. Within this structure, a mathematician first defines the values for each dependent variable and then solves the equation by performing the primitive mathematical operations on those numbers in the order prescribed by the equation. For example, the simple equation

$$A = (B + C)/D$$

identifies that the variables B and C are first defined, and then added, with the sum being stored as an intermediate solution. The intermediate value, in turn, is divided by the variable D to calculate the solution variable A. This same mathematical structure provides the framework for computer-assisted map analysis.

Within this processing structure, four fundamental classes of map analysis operations may be identified. These include:

- *Reclassifying maps*—involving the reassignment of the values of an existing map as a function of the initial value, position, size, shape, or contiguity of the spatial configuration associated with each category.
- *Overlaying maps*–resulting in the creation of a new map where the value assigned to every location is computed as a function of the independent values associated with that location on two or more existing maps.
- *Measuring distance and connectivity*—involving the creation of a new map expressing the distance and route between locations as simple Euclidean length, or as a function of absolute or relative barriers.
- *Characterizing and summarizing neighborhoods*—resulting in the creation of a new map based on the consideration of values within the general vicinity of target locations.

These major groupings can be further classified as to the basic approaches used by the various processing algorithms. This mathematical structure forms a conceptual framework that is easily adapted to modeling the spatial relationships in both physical and abstract systems. Detailed discussion of the various analytic procedures is beyond the introductory scope of this article. Reference to the papers by Berry and Tomlin (1982a and b) provides comprehensive discussion and examples of each of the classes of operations.

Cartographic Modeling

The cyclical processing structure of map analysis enables primitive spatial operations to be combined to form equations, or "cartographic models," in a familiar and intuitive manner. For example, in traditional algebra, an equation for the percent change in value of a parcel of land may be expressed as

$$\%\text{ change} = \frac{(\text{new value} - \text{old value})}{\text{old value}} \times 100$$

In a similar manner, a map of percent change in market values for an entire town may be expressed, in pMAP command language, as

```
COMPUTE NEWVALUE.MAP MINUS OLDVALUE.MAP
TIMES 100 DIVIDEDBY OLDVALUE.MAP
FOR PERCENTCHANGE.MAP
```

Within this model, data for current and past land values are collected and encoded as computerized maps. These data are evaluated, as shown above, to form a solution map of percent change. The simple model might be extended to provide coincidence statistics, such as,

```
CROSSTABULATE ZONING.MAP WITH
PERCENTCHANGE.MAP
```

for a table summarizing the spatial relationship between type of zoning and change in land value (that is, which zones have experienced the greatest decline in market value?). The basic model might also be extended to include geographic searches, such as,

```
RENUMBER PERCENTCHANGE.MAP FOR
BIGCHANGES.MAP
ASSIGNING 0 TO -20 THRU 20
```

for a map isolating those areas which have experienced more than a +20% or −20% change in market value.

Another simple model is outlined in Figure 4. It creates a map characterizing the number of houses and their proximity to major roads, given maps of housing locations (HOUSING) and the road network (ROADS). The model incorporates reclassifying, distance measuring, and overlaying operations. An extension to the model (not shown) uses a neighborhood operation to characterize each map location as to its general housing density within a tenth of a mile radius, as well as its proximity to roads. Incorporation of this modification requires the addition of one line of code and slight editing of two others. A working-copy display of the results and tabular summary is shown at the bottom of Figure 4. The pMAP command sequence for this analysis is shown in the upper portion.

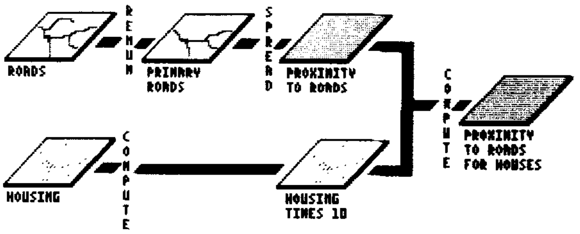
The procedure used in the example may be generalized to identify "type–distance" combinations among any set of features within a mapped area. Consider the following generalization.

```
     onemap   →   X
 anothermap   →   Y
      XDIST   =   distance fn(X)
          Z   =   (Y * 10) + XDIST
compositemap  ←   Z
```

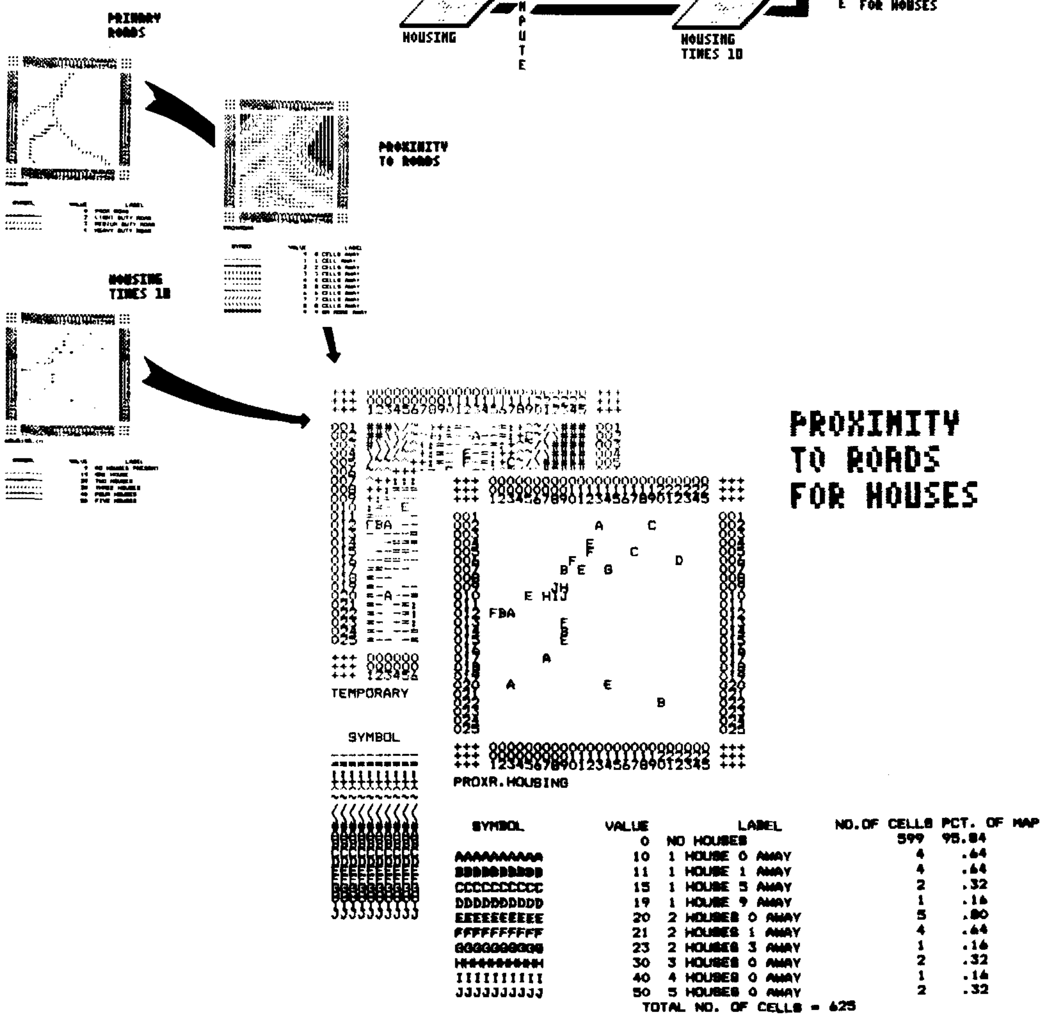
This "macro" technique may be stored as a command file and accessed at anytime. For example, the following command sequence can be entered to assess the covertype (Y) and distance to the nearest house (X) for all map locations.

COMMAND SEQUENCE

```
RENUMBER ROADS FOR PROADS ASSIGNING 0 TO 1
SPREAD PROADS TO 9 FOR PROXROAD
COMPUTE HOUSING TIMES 10 FOR HOUSING.10
COMPUTE HOUSING.10 PLUS PROXROAD FOR TEMPORARY
RENUMBER TEMPORARY FOR PROXR.HOUSING /
  ASSIGNING 0 TO 1 THRU 9
DISPLAY PROXR.HOUSING
```



OUTPUT



SYMBOL	VALUE	LABEL	NO.OF CELLS	PCT. OF MAP
	0	NO HOUSES	599	95.84
AAAAAAAAAA	10	1 HOUSE 0 AWAY	4	.64
BBBBBBBBBB	11	1 HOUSE 1 AWAY	4	.64
CCCCCCCCCC	15	1 HOUSE 5 AWAY	2	.32
DDDDDDDDDD	19	1 HOUSE 9 AWAY	1	.16
EEEEEEEEEE	20	2 HOUSES 0 AWAY	5	.80
FFFFFFFFFF	21	2 HOUSES 1 AWAY	4	.64
GGGGGGGGGG	23	2 HOUSES 3 AWAY	1	.16
HHHHHHHHHH	30	3 HOUSES 0 AWAY	2	.32
IIIIIIIIII	40	4 HOUSES 0 AWAY	1	.16
JJJJJJJJJJ	50	5 HOUSES 0 AWAY	2	.32

TOTAL NO. OF CELLS = 625

Figure 4. A simple cartographic model. This analysis combines the information on housing locations (HOUSING) and the road network (ROADS) to derive a map characterizing the number of houses and their proximity to the nearest major road.

```
    Keyboard
ASSOCIATE X WITH
  HOUSING
ASSOCIATE Y WITH
  COVERTYPE
READ c:macro.cmd
                                 Stored file
                         SPREAD X TO 9 FOR PROX.X
                         COMPUTE Y TIMES 10 FOR Y.10
                         COMPUTE PROX.X PLUS Y.10 /
                           FOR Z
DISPLAY Z
```

The ASSOCIATE command temporarily defines specified maps as generalized variables. The READ command transfers input control to the designated file, and the stored set of generalized commands are processed as if they were being entered from the keyboard. When the model is finished executing, input control is returned to the keyboard for user interactive processing.

Generalized Structure for Map Analysis

The development of a generalized analytic structure for map processing is similar to those of many other nonspatial systems. For example, the popular dBASE III package contains less than 20 analytic operations, yet they may be flexibly combined to create "models" for such diverse applications as address lists, inventory control, and commitment accounting. Once developed, these logical sequences of dBASE sentences can be "fixed" into menus for easy end-user operations. A flexible analytic structure provides for dynamic simulation as well as database management. For example, the Multiplan "spreadsheet" package allows users to define the interrelationships among variables. By specifying a logical sequence of interrelationships and variables, a financial model of a company's production process may be established. By changing specific values of the model, the impact of uncertainty in fiscal assumptions can be simulated. The advent of database management and spreadsheet packages has revolutionized the handling of nonspatial data.

Computer-assisted map analysis promises a similar revolution for handling spatial data. For example, a model for siting a new highway could be developed. The analysis would likely consider economic and social concerns (for example, proximity to high housing density, visual exposure to houses), as well as purely engineering ones (for example, steep slopes, water bodies). The combined expression of both physical and nonphysical concerns, in a quantified spatial context, is a major benefit. However, the ability to simulate various scenarios (for instance, steepness is twice as important as visual exposure; proximity to housing four times as important as all other considerations) provides an opportunity to fully integrate spatial information into the decision-making process. By noting how often and where the optimal route changes as successive runs are made, information on the unique sensitivity to siting a highway in a particular locale is described.

In addition to flexibility, there are several other advantages in developing a generalized analytic structure for map analysis. The systematic rigor of a mathematical approach forces both theorist and user to consider carefully the nature of the data being processed. It also provides a comprehensive format for instruction which is independent of specific disciplines or applications (Berry 1986). In addition, the flowchart of processing succinctly describes the components of an analysis. This communication enables decision makers to understand more fully the analytic process, and actually comment on model weightings, incomplete considerations, or erroneous assumptions. These comments, in most cases, can be easily incorporated and new results generated in a timely manner. From a decision-maker's point of view, traditional manual techniques of map analysis are separate from the decision itself. They require considerable time to perform and many of the considerations are subjective in their evaluation. From this perspective, the decision maker attempts to interpret results, bounded by the often vague assumptions and system expression of the technician. Computer-assisted map analysis, on the other hand, encourages the involvement of the decision maker in the analytic process. From this perspective, spatial information becomes an active and integral part of the decision process.

Conclusion

Geographic information system (GIS) technology is revolutionizing how maps are handled. Since the 1970s, an ever-increasing portion of mapped data is being collected and processed in digital format. Currently, this processing emphasizes computer mapping and database management capabilities. These techniques allow users to update maps quickly, generate descriptive statistics, make geographic searches for areas of specified coincidence, and display the results as a variety of colorful and useful products. Newly developing capabilities extend this revolution to how mapped data are analyzed. These techniques provide an analytic "toolbox" for expressing the spatial interrelationships of maps. Analogous to traditional algebra and statistics, primitive operations are logically sequenced on variables to form spatial models; however, the variables are represented as entire maps. This quantitative approach to map analysis is changing basic concepts of map structure, content, and use. From this perspective, maps move from images describing the location of features to mapped data quantifying a physical or abstract system in prescriptive terms. This radical change in map analysis has promoted a more complete integration of spatial information into the decision-making process.

Acknowledgments

The professional map analysis package (pMAP) described in this article was developed and written by the author and Dr. Kenneth L. Reed of Spatial Information Systems, Inc. Several of the concepts and algorithms are based on the widely distributed mainframe map analysis package (MAP) developed by C. Dana Tomlin (1983) and distributed by Yale University.

Literature Cited

Berry, J. K. 1985. Computer-assisted map analysis: fundamental techniques. Pages 369–386 *in* Proceedings of 6th annual conference, National Computer Graphics Association, Dallas, Texas, April.

Berry, J. K. 1986. The academic map analysis package (aMAP) instructions materials. Yale University, School of Forestry and Environmental Studies, New Haven, Connecticut.

Berry, J. K., and C. D. Tomlin. 1982b. Cartographic modeling: computer-assisted analysis of spatially defined neighborhoods. Pages 227–241 *in* Conference on energy resource management, American Planning Association, Vienna, Virgina, October.

Davis, J. C. 1973. Statistics and data analysis in geology. John Wiley and Sons, New York, pp 170–411.

Golden Software. 1985. Golden Software Graphics System Information Manual. Golden Software, Golden, Colorado.

McHarg, I. L. 1969. Design with nature. Doubleday/Natural History Press, Doubleday and Company, Garden City, New York, 197 pp.

Ripley, B. D. 1981. Spatial statistics. Series in Probability and Mathematical Statistics. John Wiley and Sons, New York, 241 pp.

SIS. 1986. The professional map analysis package (pMAP) user's manual and reference. Spatial Information Systems, Omaha, Nebraska, 228 pp.

Tomlin, C. D. 1983. Digital cartographic modeling techniques in environmental planning. PhD dissertation, Yale University, School of Forestry and Environmental Planning, New Haven, Connecticut, 298 pp.