The Most Beautiful Formulae in GIS

By Nigel Waters

Dr. Waters is currently Professor & Director of the Geographic Information Science, Center of Excellence with the Department of Geography and Geoinformation Science at George Mason University, Fairfax, Virginia, USA; e-Mail: nwaters@gmu.edu. (Oct 2013)

…portions of this white paper by Nigel Waters was first published in GIS World magazine and subsequently compiled for inclusion as a supplemental section in the book Spatial Reasoning for Effective GIS by Joseph K. Berry (Wiley 1995).

(.pdf version posted at **http://www.innovativegis.com/basis/MapAnalysis/Topic30/Topic30.htm/Beautiful_Formulae.pdf**)

One of the most important characteristics of a GIS is that it can be used as a spatial decision support system (SDSS). A GIS can act like an SDSS because it is possible to subject the data stored within a GIS to mathematical manipulations. According to Gail Langran (1989), the ability to analyze data distinguishes first-generation GISs (which could do no more than store, retrieve, and display data) from the second-generation systems that have these powerful processing capabilities.

These mathematical manipulations require a variety of equations and procedures that have their own simplicity, logic, and beauty. The purpose of this white paper is to introduce you to some of the key areas of mathematical manipulation used in a GIS. The appendix stresses the elegance of these formulae and equations, as well as their cleverness and usefulness-a common approach in many expository, mathematical texts (see, for example, Salem et al., 1992; Gardner, 1978; and Dunham, 1991). The idea is that if you have a deep, intuitive understanding of the logic and elegance of the equations then you'll more likely remember and retain that knowledge. In addition, for each formula one or more key references are included that allow readers to dig much deeper into this treasure trove of powerful mathematical tools.

To be included, these formulae meet some of the following criteria:

- They are widely used.
- They are the fundamental building blocks of many key GIS algorithms and analytical procedures.
- They impart powerful ideas (see Papert, 1980, for an extended discussion of this concept).
- They should be representative of and related to other key formulae.

Most vector-based GISs traditionally used a hybrid structure for data storage (Healey, 1991). That means locational information is stored separately from attribute data. Attribute data represent the characteristics of whatever is being mapped. Thus, many GISs, such as ARC/INFO from Environmental Systems Research Institute, Inc. (Esri), Redlands, California (see GIS World Sourcebook 1995 from GIS World, Inc., Fort Collins, Colorado, for a detailed description of GIS packages and the companies that market them), use proprietary software to store the spatial information and a relational database for the attribute information. The proprietary spatial software may be their own, as in the case of Esri, or commercially available computer-aided design (CAD) software. Occasionally some systems use integrated approaches, but that is much less common. I follow the hybrid approach here and divide the algorithms and equations into two groups: spatial and nonspatial.

SPATIAL FORMULAE

1. The Oldest of Them All

Perhaps the most beautiful of all is the oldest. It is the formula for distance between two points and is based on Pythagoras' 'theorem from the sixth century B.C. (see Dunham, 1991, for historical details).

$$D = (| X_1 – X_2 |^2 + | Y_1 - Y2 |^2)^{1/2}$$

The two points are represented by their X and Y coordinates in a Cartesian space. Thus, X1 and Y1 represent the coordinates of the first point, and X2 and Y2 represent the coordinates of the second point. D is the straight line distance between the two points. The vertical bars mean take the absolute value. The 1/2 exponent is, of course, the square root.

The equation is a fundamental building block in many GIS algorithms. Distance is important in many GIS procedures. For example, if a GIS analyst uses a spatial interpolation method, such as a distance weighted moving average (see spatial formula No. 7), to produce an interpolated surface or "drape," he or she probably will interpolate a square grid of points from an irregularly spaced dataset. Each point on the grid will be a distance weighted average of the six nearest points in the original dataset. The above equation is required to determine the six nearest points.

Many other algorithms are made more efficient if the number of possibilities that have to be considered are reduced, and that is achieved often by considering only the closest points. The Very Important Points Algorithm for choosing points to be part of a triangulated irregular network (TIN) model requires distance calculations. The TIN model is a more efficient alternative to the digital elevation model (DEM) for representing a surface (see Unit 39 in the Core Curriculum developed by the National Center for Geographic Information and Analysis (ruccra) by Goodchild and Kemp, 1990; and also Weibel and Heller, 1991).

If the "twos" in the exponents are replaced by R, then R is known as the Minkowski metric. When R=1 we have the so-called city block or Manhattan distance. That formula has been used ingeniously in location-allocation modeling where the analyst uses the GIS to suggest the optimum location for a facility, such as a fire station or a fast food restaurant, and then determines the allocation of residences or potential customers to that facility. The facility users are allocated to the facility to which they live closest, hence the need for distance calculations. Location-allocation modeling is one of the most useful applications of GIS in the public and private sectors. Excellent reviews are provided by Ghosh and Rushton (1987). It is interesting that Esri has seen fit to include state-of-the-art location-allocation modeling procedures only in the most recent revision of ARC/INFO (Version 7). However, the procedure has been incorporated into Caliper Corporation's TransCAD package for several years.

Distance calculations are important in clustering algorithms, Voronoi or Thiessen polygon algorithms, and nearest neighbor routines in point pattern analysis that are used to determine the adequacy of spatial sampling procedures.

2. The Intersection of Two Lines

$$X_i = - (a_1 – a_2) / (b_1 – b_2)$$
$$Y_i = a_1 + b_1 X_i$$

In the equations above, the b terms represent the slope of the first and second lines, and the a terms represent the intercept with the Y or vertical axis in the Cartesian coordinate system. The

equations for the two straight lines from which these terms are derived are given by the following equations:

$$Y_1 = a_1 + b_1X_1$$
$$Y_2 = a_2 + b_2X_2$$

The initial formulae, which provide the X and Y coordinates of the point of intersection (point X_i,Y_i), are another beauty from the field of coordinate geometry, or as GIS techies say, "COGO." Coordinate geometry is one of the most useful branches of mathematics for GIS design. It is also one of the more accessible, because the rudiments of the subject usually are taught in high school mathematics classes. There are many useful texts in this area. One of the more technical and comprehensive is Hartley (1960).

A related area of mathematics that is equally important is computational geometry, which is concerned with the development of algorithms and procedures that will provide solutions to geometric problems quickly and efficiently on computers. There is now an annual symposium on computational geometry sponsored by the Association for Computing Machinery (ACM) and the special interest groups for Graphics, Automata and Computational Theory of that organization. An advanced text on computational geometry is by Preparati and Shamos (1985). The authors begin their book by noting that the original motivation for geometry among the Greeks and Egyptians was the need to tax lands accurately and fairly and to construct buildings. Of course, these are still important motivations behind today's cadastral and land-related GISs. Algorithm efficiency and speed are vital in GIS, because the operation being considered may have to be performed on tens or hundreds of thousands of lines or polygons. An inefficient algorithm may slow down even the fastest computer.

The formula for the intersection of two lines is the cornerstone of many point-in-polygon and polygon overlay routines. And where would GIS be without them? Point-in-polygon routines are extremely important in GIS operations. Frequently we need to know how many observations fall inside a polygon and how many fall outside. The point-in-polygon operation is straightforward: a line is drawn vertically upward from the point in question. The number of times that line crosses the boundary of the polygon being considered is recorded. If that number is even, the point is outside the polygon; if it is odd, the point must be inside. The algorithm is surprisingly robust and works for polygons that have weird and wonderful shapes' but there are many "special cases" (discussed below) requiring careful consideration.

The polygon overlay operation involves finding the intersecting arcs of two different map coverages (Unit 34 of the NCGIA Core Curriculum). Overlay operations are required for combining the properties of two or more map coverages and, thus, have been important in environmental impact studies since the days when such operations were performed mechanically and manually. Thus, Ian McHarg (1969) used photographic overlay procedures to determine the ideal location for transportation corridors, among other things in his classic text Design with Nature. Polygon overlay is important for buffering and windowing operations in a vector-based GIS. Buffering operations allow the user to put a small polygon around a point, line, or area. For example, a buffer might be placed around a river or a lake to protect such features from environmental damage that might result from logging. A window might be used to take a look at what lies within the window. Thus, the GIS user could find out the population characteristics of the people within the window. Alternatively the window might be used as a moving spatial averaging or filtering device to generate a smoother surface.

Before we leave this topic we should note that the above formulae solve the intersection problem, but only for simple situations. In the real world there are a host of so-called "special cases," which arise with depressing frequency. Thus, the formulae above assume lines of infinite length when we really only want to see if the lines cross somewhere on our map. Vertical lines that have an infinite slope and horizontal lines that have no slope may also cause complications. These and other difficulties are addressed in an article by David Douglas that has been reprinted several times (see, for example, Douglas, 1990).

3. The Gravity Model

$$I_{ij} = k [(p_i) (p_j) / (d_{ij})^2]$$

The gravity model was used to predict traffic flows in early transportation models from the late 1950s. It is developed extensively in the writings of Wlliam Warntz and his students who pioneered the model (see Coffey, 1988, for a review of this work). It states that the amount of interaction, I, between two places, i and j, is equal to some constant, k, multiplied by their population product divided by the distance between them squared. The equation has been used in various forms. Often the population terms are raised to a power, which is determined empirically using least squares regression procedures. The exponent on distance may also be fitted empirically. A full review of these types of models and their applications in the literature may be found in Taylor (1975. The p_i and p_j terms may be used to represent other variables besides population that have an influence on interaction, and the distance term too may represent such surrogates for distance as time and cost.

The model can be extended and manipulated to produce population potential models important in marketing and popularized in GIS packages such as SPANS, now owned by PCI, Incorporated, Richmond Hill, Ontario, Canada. The SPANS package has been used by such firms as Miracle Mart, Ontario, to advise on new supermarket locations. In such studies, areas of high population potential are important indicators of where to place a new store.

4. The Entropy Maximizing Model

$$T_{ij} = (A_i * O_i * B_j * D_j) / e^{(b * cij)}$$

The equation above, in which T_{ij} is the expected or most likely distribution of trips between transportation zone i and transportation zone j, was the basis of the entropy maximizing models developed in the late 1960s. The models provided statistical respectability for the gravity model. 0_i and D_j are the number of workers in the origin zone and number of jobs in the destination zone, respectively; A_i and B_j are weights; e is the irrational number 2.718...; b measures the friction of distance; and c is the cost of travel. Entropy maximizing models have appeared in a variety of forms, including the so-called singly constrained shopping models and the doubly constrained journey-to-work models. They are discussed in great detail in Wilson and Kirkby (1980). In fact, Wilson first introduced these models to geographers, planners, and GIS analysts. A gentler introduction is provided by Gould (1984).

5. Projection Formulae

$$X = R (l\text{-}1_0)$$
$$Y = R \ln \tan [(p / 4) + (f / 2)]$$

Fifth on the list of spatial formulae and those that explicitly incorporate distance are formulae for converting spherical coordinates from Earth's surface into the two-dimensional X and Y coordinates of the Cartesian plane. There are many of these formulae— one or more formulae for each of the dozens of map projections now commonly used in GIS. In the equations above, l and f are longitude and latitude, respectively, and X and Y are the standard Cartesian coordinates. R is the radius of the sphere, and l_0 is the central meridian with all angles measured in radians. Finally, p is the mathematical number pi. This elegant formula (from Snyder, 1987) for the oldest of all standard map projections, Mercator's, has been used to produce maps of the planetary surface of most of Earth's nearest neighbors (Snyder, 1987).

Many of the large, expensive GISs allow users to convert to Cartesian coordinates from any of a large number of projections (in addition to Snyder, 1987, formulae for these projections are discussed in Richardus and Adler, 1972, and Pearson, 1990). There are some excellent standalone packages that carry out vast numbers of these transformations. Perhaps one of the most efficient and effective of these packages is The Geographic Calculator from Blue Marble Geographics, Gardiner, Maine (Waters, 1994a).

6. Trend Surface Models

The formula for a straight line in Cartesian space is also the formula for the common regression line, and that regression equation can be extended easily into two independent dimensions to produce an interpolated trend across space:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + e$$

The equation above is a first-order polynomial with two independent variables. Notation for this model varies, but here I follow that given in Davis (1973, 1986). In the equation, Y is the iependent variable whose value we wish to predict. X_1 and X_2 are the independent variables, and in this model they represent location. The final term, e, is the error or residual term. If the dependent variable is elevation (although it does not have to be) and the two independent variables are eastings and northings, respectively, our trend surface, regression equation provides a linear, interpolated surface portraying the trend in elevation.

From this simple surface we can easily move to higher order surfaces that allow us to model more complex relief. The equation as presented only has the X_1 and X_2 terms raised to their first power, so it produces a surface with no inflections— it is a plane that may or may not be tilted. If we add higher order powers of X_1 and X_2 and terms which represent combinations of their cross-products (e.g., X_1X_2 and $X_1^2 X_2$), we can then create more complex models of our interpolated surface. Indeed, the number of inflexions or bends in the surface is always one less than the highest order of the X_1 and X_2 powers we use in the equation. Complete details of all of these equations and the mathematics behind them, as well as computer programs to carry out the computations, are provided in Davis (1973) and Mather (1976).

To the above equation we can also add another independent variable, X_3, and plunge into the third dimension. So this is a powerful formula indeed. Three-dimensional trend surfaces are difficult to visualize and impossible to draw on a two-dimensional piece of paper or on a flat computer screen. But we can draw them one slice at a time or use some of the visualization techniques discussed in Rosenblum, et al. (1994) to get an impression of how a variable changes over three dimensions.

The error term or residual is important and allows us to see how good our interpolated surface is. Thus, we can use the GIS to produce a map of our model, the trend surface, and we also can use it to produce a map of the residuals or errors to show where the model fails to produce a good fit with reality.

With just one independent variable in the above equation we can model how our dependent variable, for example economic rent, changes over distance, and we have the classic bid rent line— Von Thunen's model flMilson and Kirkby, 1980).

7. Distance Weighted Averages

Distance weighted averages are like spatial moving averages. They are important in surface interpolation routines and in attempts to smooth a surface and to increase the signal-to-noise ratio of whatever is being portrayed. The equation is simple and straightforward in its unadorned version, but the beauty of the formula is that it is possible to include all sorts of ad hoc adjustments.

$$Y_k = SUM (Y_i / D_{ik}) / SUM (1 / D_{ik})$$

This equation simply states that the estimated value at Y_k is given by taking each of the i points and dividing by the distance between each of the i points and point k, which is the point being estimated. These values are summed for however many points are used (4, 6, and 8 points are common). The sum then is divided by the sum of the reciprocals of all the distances. Chapter 3 in Harbaugh, Doveton and Davis (1977) discusses the effect of using different numbers of points, different distance weighting procedures and other variations on this generic version of the equation. The equation has been used extensively in the SURFACE II and III series of spatial interpolation programs (Sampson, 1978, 1994).

8. Afflne and Curvilinear Transformations

Closely related to the ideas expressed in spatial formula number 5 above and to the equation used in spatial formula number 6 above is the idea of an affine transformation. Transformations are required when we wish to register different sets of coordinates for information from the same mapped area, but the coordinates come from a variety of maps that use different projections or coordinate grids. Such transformations come in two types: affine transformations that keep parallel lines parallel and curvilinear transformations in which that is not necessarily the case.

Affine transformations include translations (where the origin of the system is moved), scalings (where the scale of the map is changed), rotaions (where the map is rotated around the origin), and reflections (where a mirror image of the map is produced). These transformations are the bread-and-butter of a host of algorithms in computer graphics. Foley and Van Dam (1934) provide an exhaustive account together with computer pseudo-code for these operations. A gentler introduction, along with BASIC programs and code, is given by Myers (1982). The four operations are defined in the following equations:

Translation:	*Scaling:*	*Rotation:*	*Reflection (about the X axis):*
U = X - a	U=c X	U = X cos(alpha) + Y sin(alpha)	U = X
V = Y - b	V=d Y	V = -X sin(alpha) + Y cos(alpha)	V = f - Y

Where,

- U and V are the new transformed coordinates,
- X and Y are the original coordinates,
- a and b represent the number of units the Y and X axes are shifted,
- c and d represent the scale change for the X and Y coordinates,
- alpha is the angle of rotation measured anticlockwise, and
- f is the maximum value on the Y axis.

The transformations have several specific applications. For example, a reflection is needed to convert from a Cartesian coordinate system (in which the origin of the system is in the bottom left-hand corner and Y increases as we move up the map) to a raster display system, such as a line printer or a video monitor (in which the origin is in the top left-hand corner and Y increases as we move down the map). A series of these transformations (a translation, two rotations, and a reflection, respectively) also are needed to transform three-dimensional world coordinates into two-dimensional screen or map coordinates (see Myers, 1982).

One method that combines these transformations is to use two multiple regression equations of the following form:

$$U = b_1 + b_2X + b_3Y$$
$$V = b_4 + b_5X + b_6Y$$

These equations are of the same form as the first-order polynomial used in the trend surface. U and V and X and Y are defined as before, but b_1-b_6 are constants and coefficients that could be fitted using a multiple regression program, one run for each equation. Thus, a GIS analyst would have a series of X and Y coordinates and enter them as independent variables into the multiple regression program. In the first run the U coordinates from the second map would represent the dependent variable. Thus, the multiple regression program would yield the b_1,b_2, and b_3 coefficients for the first equation. A second run of the program using the V coordinates from the second map would allow the remaining b_4, b_5, and b_6 coefficients to be found. Advanced treatments of affine geometry may be found in many textbooks (see, for example, Snapper and Troyer, 1971).

Curvilinear transformations involve the use of higher order polynomials such as the following:

$$U = b_1 + b_2X + b_3Y + b_4X^2 + b_5Y^2 + b_6XY$$
$$V = b_7 + b_8X + b_9Y + b_{10}X^2 + b_{11}Y^2 + b_{12}XY$$

This is a second-order polynomial, but more complex models might be used. Again a good discussion is found in Davis (1986). The whole topic of transformations is discussed in Goodchild (1984).

9. Estimating Slope and Aspect

A local trend surface equation can be calculated using a 2 by 2 or 3 by 3 window in a digital elevation model (DEM) from a raster GIS. Once the equation has been determined (either by the usual least squares method or using the simplified formulae found in Unit 38 of the NCGIA Core Curriculum) and put in the form shown in spatial formula Number 6 above, we can calculate the landscape's useful properties, such as slope and aspect. Using the same notation

as in the trend surface equation above, these properties would be found by the following equations:

$$\text{Slope} = (b_1^2 + b_2^2)^{1/2}$$
$$\text{Aspect} = \tan^{-1} (b_2 / b_1)$$

The first equation states that slope is equal to the square root of the sum of the squares of the b_1 and b_2 coefficients from the trend surface equation. These properties represent just two from a large number of geo-morphometric properties that can be calculated from the DEM. The best discussion of this is found in articles by Evans (1990) and Pike (1988).

10. Fractals and Determining the Fractal Dimension

Fractals stem from a branch of mathematics that was popularized by Mandelbrot (1977, 1982). A popular treatment of Mandelbrot and his work is provided by Gleick (1987). Fractals have many uses, including shape measurements, evaluating the degree of convolution of two-dimensional shapes (hence they help with error measurements), evaluating the roughness of surfaces, and other generalization such as line operations. They can be used for data compression, and Barnsley (Chapter 3, 1988) discusses their relationship to affine transformations (see spatial formula number 8 above). Geographical treatments are contained in Goodchild and Mark (1987) and Batty and Longley (1994).

Here I provide a simple equation for determining the fractal dimension of a line. Other methods are discussed in Goodchild and Mark (1987).

$$D = (\log (n_2 / n_1)) / (\log (s_1 / s_2))$$

D is the fractal dimension; s1 is the step size of a pair of dividers used to measure the line the first time; s2 is the step size of a pair of dividers used to measure the line the second time; n1 is the number of steps used in the first measuring attempt; and n2 is the number of steps used in the second measuring attempt. The formula relies on the fact that the line will appear to become longer as the step size is decreased (equivalent to an increase in the scale of the map), because now we use a more accurate measuring device. The length increase is measured by the fractal dimension.

ATTRIBUTE-RELATED FORMULAE

1. The Normal Distribution or Bell Curve

The normal distribution describes many types of objects, including the characteristics of people, naturally occurring phenomena, and errors. The normal distribution is based on the principle that extreme values are rare. Thus, short and tall people are uncommon, whereas people of an average height are common. The bottom line is that most of us are pretty average. The same is true of animal and plants. And it is true of unbiased and nonsystematic errors as well. As a result, small errors, for example in digitizing, tend to be more common than large errors, which is why we use statistics to describe our errors.

For example, the root mean square is another name for the standard deviation, and in a normal distribution 68 percent of the values will be between the value of the mean (or average) minus one standard deviation and the value of the mean plus one standard deviation. So by quoting

the root mean square for our errors on an ARC/INFO tic coverage, for example, we can get a good idea of how serious the errors are.

The normal distribution also has been shown to provide a good model for the rate at which people and organizations adopt innovations. That idea figures prominently in the work of Rogers (1962). The point is that initially it is difficult to persuade individuals to adopt something new because no one wants to take the risk. But there are a few people who are true innovators. They adopt right away. And then comes a larger group, the early adopters, followed by yet a larger group, the so called early majority. These groups are followed by the late majority and finally the laggards, the latter being a small group of die-hards who need to be convinced before they adopt. The groups are divided on the basis of the standard deviation and the normal distribution.

To illustrate, suppose it takes the average person 12 years to adopt and the standard deviation of the distribution is four years. The innovators will adopt before four years have elapsed since the product's introduction. The early adopters will adopt between four years and eight years, and the early majority will adopt between eight and 12 years after the introduction. The late majority adopt between 12 and 16 years after introduction. Finally, the laggards adopt after the product has been on the market 16 years. Because we can use mathematical tables to determine the percentage between the mean and one or more standard deviations from the mean, we know exactly how many people fall into the five categories: 2.5 percent, 13.5 percent, 34 percent, 34 percent, and 16 percent, respectively. Now what does that have to do with GIS? Jeffrey Lane and David Hartgen (1989) used the model to illustrate the rate at which state transportation departments adopt GIS and related technology.

2. Standardizing Data

$$Zi = (X_i - \text{Mean}) / \text{Standard Deviation}$$

This simple formula allows us to standardize the values of different variables. Thus, variables that have different variances because they are measured in different units are reduced to the same measuring stick. The formula is important in cluster analysis in which a measure of dissimilarity is based on Euclidean distances (see spatial formula number 1). If we did not standardize the variables, those with larger variances would exert a greater influence on the clustering process. The formula is also useful when our variables are distributed normally, because the formula will convert all our values to standard normal deviates and then we can use tables of standard normal deviates to determine the probability of finding values between certain ranges, among other applications. Many test statistics follow a normal distribution, so if we convert an observation to one of these test statistics and we know the mean and standard deviation of the sampling distribution of the test statistic, we can again convert it to a standard normal deviate and work out the probability of observing that particular value. That is the basis of many inferential statistical tests (see Siegel and Castellan, 1988, for a more detailed discussion).

3. The Chi-Square Statistic

The chi-square statistic is used to compare observed with expected distributions. It is ideal as a one-sample, goodness-of-fit test, because we can use it to see how our data are distributed. Thus, we can tell if a variable follows a normal or a poisson or some other distribution and, in the first two instances, that will allow us to take advantage of all the benefits described under attribute formulas numbers 1and 8. The formula for chi-square is as follows:

$$X^2 = \text{SUM} \ [(f_o - f_e)^2 / f_e]$$

Here f_o is the observed frequency, and f_e is the expected frequency. A thorough discussion of this statistic is given in Williams (1984). Another useful introduction is provided by Siegel and Castellan (1988). The use of formal hypothesis testing, and the scientific method is an approach that has been recommended by Waters (1994b) and Wellar and Wilson (1993).

4. The Kappa Statistic

Remote sensing and GIS often are lumped together. Remote sensing is a cheap, effective way to acquire large amounts of data over extensive parts of the world covered by harsh or inaccessible terrain. The journal Photogrammetric Engineering and Remote Sensing has a special section of each issue devoted to GIS and a complete issue devoted to GIS annually. The fact that remotely sensed data arrive as a grid of pixels makes them an ideal data source for a raster GIS.

One of the most common activities in remote sensing is to classify the pixels into groups that represent, for example, vegetation classes. Then we can compare the classified pixels to the vegetation that actually occurs on the ground in a number of test sites. That information usually is represented in a matrix in which each row of the matrix represents the true classes and the columns represent the predicted classes that result from the use of a procedure such as a discriminant equation (see attribute formula number 4). In the matrix the diagonal elements represent a successful classification, i.e., where the predicted value is the same as the true value. Off-diagonal elements indicate a failure of the model. The GIS analyst wants to know how good the model is, how much better than a random allocation of pixels to the elements in the matrix. Even a random allocation would have some success. So to see how much better the model is we use the kappa statistic:

$$\text{Kappa} = (d - q) / (N - q)$$

Here d is the number of cases in the diagonal cells, N is the total number of cases recorded in the matrix, and q is the number of cases expected in the diagonal cells (for a given matrix, q can be found by taking each row total multiplying by each column total and then dividing by N, an operation carried out for each row, and then all the row values are summed).

Alternative formulae and associated significance tests and references for this useful, elegant statistic can be found in Siegel and Castellan (1988). One of the most important properties to know about these statistics is their range. Ideally a statistic such as kappa should have the same range for matrices of any size, and it simplifies things if this range is from 0 to 1. Kappa conforms to these norms. A matrix that shows a classification which is no better than chance will have a kappa value of 0, while a matrix with a perfect classification in which all the entries are on the diagonal will have a value of 1, because d and N will be the same number.

5. Color Matching Systems

Color is important in GIS. It assists in visualization (Rosenblum, et al., 1994, and spatial formula number 6), and almost every GIS text includes the obligatory section of color images-usually increasing the cost of the book by $5 to $10! A detailed introduction to the use of color and computer systems is provided by Durrett (1987) and Waters (1989). One of the problems with the use of color is that there are numerous ways of describing color. Usually a three-

coordinate system is used. Jain (i989) provides many formulae for moving from one coordinate system to another. These formulae are usually matrix equations of the following format:

$$[O] = [TM]\ [I]$$

Here [O] is the 3-by-l output vector of three coordinates describing the characteristics of the color, [I] is the 3-by-1 input vector describing the color characteristics in the original system, and [TM] is a 3-by-3 transformation matrix. A specific transformation matrix is required for each type of transformation.

6. Optimization Models

$$\text{Maximize } Z = aX_1 + bX_2 \textit{ subject to constraints}$$

Here we have the basis for all the programming models in their various guises, including linear, integer, and dynamic programming, and the so-called transportation models. They are also part of most location-allocation models, which now are incorporated into many of the most important GISs (see discussion of spatial formula number 1). All of these models are important if we wish to use our GIS as an SDSS. Extensive discussion of optimization models is provided in Wilson and Kirkby (1980). See Densham (1991) for an SDSS discussion.

7. The Rank Size Rule

$$Pi = P1 / i$$

This is a powerful formula with all sorts of statistical implications and applications. A hint of the treasure trove awaiting the investigation of this formula is given by Simon (1991). Basically, the formula says big things are less common than little things, and it describes log-normal distributions that are common in the data we store in GISs. The formula has been used extensively in urban studies in which the size distribution of urban centers is expected to follow this distribution, and it also has been used in hydrocarbon exploration studies in which hydrocarbon reservoirs also are expected to follow the distribution. In the former version of the model, Pi would represent the population of the town at rank i in the hierarchy, and Pl is the population of the largest town in the system. Thus, the formula states that the second town in the urban system is half as big as the first, the third town is a third the size of the first, etc.

8. The Poisson Distribution

$$p(n) = (e^{-d})\ (d^{n})\ /\ n!$$

Here d represents the density of the process, and p(n) is the probability of observing n points in a given interval of either time or space. The formula has been used to evaluate the distribution of samples of points to see if the GIS provides adequate coverage when the data are spaced irregularly. It also can be used to generate data for traffic simulation models in which vehicles arrive independently of each other, as well as many other situations.

9. Classification and the F Statistic

$$F = \text{Within class variance} / \text{Between class variance}$$

So simple, so elegant, and yet it is the basis of most attempts to classify data in a GIS. And classification is so fundamental to GIS (Davis, 1986) use in resource management using remotely sensed data. Classification raises interesting philosophical and methodological issues. The former are detailed in Lakoff (1987), and the latter are addressed in Mather (1976) and Everitt (1993). Procedures that depend on this statistic include cluster analysis, trend surface analysis, regression analysis, and discriminant analysis.

10. The Discriminant Equation

Discriminant analysis is used extensively in remote sensing and GIS. For example, an analyst might classify pixels from a digital image into vegetation classes using discriminant functions. Although part of the general linear model, which assumes the data are distributed normally, the technique is relatively robust for departures from such assumptions and is used frequently. Excellent accounts of its application are found in Mather (1976), Davis (1986), and Tabachnik and Fidell (1989). The equation is similar to a multiple regression equation and is given by the following formula:

$$DS = b_0 + b_1X_1 + b_2X_2 + b_nX_n$$

Here DS is the discriminant score, b_0 is a constant, and b_1 to b_n are coefficients that are fitted empirically according to some criterion, such as maximizing the minimum difference among classes. The equation can be written in a standardized form in which case the constant term drops out and the b_1 to b_n terms are replaced by standardized coefficients that indicate the relative importance of the variables in contributing to the discriminant function.

Conclusion

The formulae chosen here were selected for their elegance, simplicity, and generality, as well as their wide applicability and power. They represent a host of approaches and analytical procedures that have proved fruitful in constructing GISs and in subsequent GIS analysis. There are many others, but these few provide a glimpse into the power of a GIS as a tool for spatial reasoning and spatial decision support.

C

This appendix originated in a column that appeared in G1S WORLD's November 1994 issue. That original, much abbreviated treatment asked readers for their input concerning what they felt to be formulae and equations deserving of inclusion in a more lengthy treatment. Many readers responded to me by E-mail, and I am truly grateful to them for their suggestions. These readers included Lee De Cola, U.S. Geological Survey, who suggested the Normal or Gaussian distribution. I am still collecting ideas. If you have suggestions, please E-mail them to me at nwaters@gmu.edu.

References

Barnsley, M. 1988. *Fractals Everywhere*. Academic Press, New York.

Batty, M. and Longley, P. 1994. *Fractal Cities*. Academic Press, New York.

Coffey, W. J., ed. 1988. *Geographical Systems and Systems of Geography: Essays in Honor of William Wamtz*, Department of Geography, University of Western Ontario, London, Ontario.

Dunham, W. 1991. *Journey Through Genius; The Great Theorems of Mathematics*. Penguin Books Ltd., Harmondsworth, England.

Davis, J. C. 1973. Statistics and Data Analysis in Geology (First Edition), John Wiley, New York. (This edition of the book contains Fortran code for carrying out spatial and statistical operations of interest to the GIS analyst.)

Davis, J. C. 1986. *Statistics and Data Analysis in Geology* (Second Edition). John Wiley, New York.

Densham, P. J. 1991. “Spatial Decision Support Systems," Chapter 26, pp.403-412 in Maguire et al., op. cit.

Douglas, D. 1990. "It Makes Me So Cross," Chapter 21, pp. 303-307 in Peuquet and Marble, op. cit.

Evans, L S. 1990. "General Geomorphometry," pp. 44-56 in A. Goudie (ed.) *Geomorphological Techniques*, Unwin Hyman, London.

Everitt, B. S. 1993. *Cluster Analysis* (Third Edition). Arnold, London.

Foley, J. and Van Dam, A. 1984. *Fundamentals of Interactive Computer Graphics*. Addison-Wesley, Reading, Mass.

Gardner, M.1978. Aha! Insight. Scientific American Inc., W.H. Freeman and Company, San Francisco.

Ghosh, A. and Rushton, G. 1987. *Spatial Analysis and Location-Allocation Models*. Van Nostrand Reinhold, New York.

Gleick, J. 1987. *Chaos: Making a New Science*. Penguin, Harmondsworth, England.

Goodchild, M. F. 1984. "Geocoding and Geosampling." pp. 3153 in *Spatial Statistics and Models*, G. L. Gaile and C. J. Wiltmott, eds., Reidel Publishing Co., Dordrecht, Holland.

Goodchild, M. F. and Kemp, K. K. eds. 1990. *NCGIA Core Curriculum*. National Center for Geographic Information and Analysis, University of California, Santa Barbara.

Goodchild, M. F. and Mark, D. M. 1987. "The Fractal Nature of Geographic Phenomena." *Annals*, Association of American Geographers, vol. 77, pp. 265-278

Gould, P. 1984. *The Geographer at Work*. Routledge, New York.

Harbaugh, J. W., Doveton, J. H. and Davis, J. C. 1977. *Probability Methods in Oil Exploration*. John Wiley, New York.

Hartley, E. M. 1960. *Cartesian Geometry and the Plane*. Cambridge University Press, Cambridge, England.

Healey, R. G. 1991. "Database Management Systems." Chapter 18, pp.251-267 in Maguire et al., op. cit.

Jain, A. K. 1989. *Fundamentals of Digital Image Processing*, Prentice Hall, Englewood Cliffs. New Jersey.

Lakoff, G. 1987 . *Women, Fire and Dangerous Things*. The University of Chicago Press, Chicago.

Lane, J. S. and Hartgen, D. T. 1989. "Factors Affecting the Adoption of Information Systems in State Departments of Transportation." Paper 18, *Transportation Publication Series*, University of North Carolina at Charlotte, Charlotte, N.C.

Langran, G. 1989. "A Review of Temporal Database Research and Its Uses in GIS Applications." *International Journal of Geographical Information Systems*, vol. 3, 2l5-232.

Maguire, D. J., Goodchild, M. F. and Rhind, D. W. 1991. *Geographical Information Systems*. Longman Scientific and Technical. London.

Mandelbrot, B. M.1977. *Fractals: Form, Chance and Dimension*.W. H. Freeman and Co., San Francisco.

Mandelbrot, B. M.1982. *The Fractal Geometry of Nature*. W. H. Freeman and Co., San Francisco.

Mather, P. M. 1976. *Computational Methods for Multivariate Analysis in Physical Geography*. John Wiley, New York.

McHarg, I. L.1969. *Design with Nature*. Doubleday, New York.

Myers, R. E.1982. *Microcomputer Graphics*. Addison-Wesley, Reading, Mass.

Papert, S. 1980. *Mindstorms: Children, Computers and Powerful Ideas*. Basic Book, New York.

Pearson, F. 1990. *Map Projections: Theory and Applications*. CRC Press Inc., Boca Raton, FIa.

Peuquet, D. J. and Marble, D.E, 1990. *Introductory Readings in Geographic Information Systems*, Taylor and Francis, London.

Pike, R. J. 1988. "The Geometric Signature: Quantifying Landslide Susceptible Terrain Types from Digital Elevation Models." *Mathematical Geology*, vol. 20, pp. 491-511.

Preparata, F. P. and Shamos, M. I. 1985. *Computational Geometry: An Introduction*. Springer-Verlag, New York.

Richardus, P. and Adler, R.K.1972. *Map Projections for Geodisists, Cartographers and Geographers*.

Rogers, E. M.1962. *Diffusion of Innovations*, Free Press, New York.

Rosenblum, L., Earnshaw, R.A., Encarnacao, J., Hagen, H., Kaufman, A., Klimenko, S., Nielson, G., post, F. and Thalmann, D.1994. *Scientific Visualization: Advances and Challenges*. Academic Press, New York.

Salem, L., Testard, F. and Salem, C. 1992. *The Most Beautiful Mathematical Formulas*. John Wiley, New York.

Sampson, R. J. 1978. *Surface II Graphics System*. Series on Spatial Analysis, #1, Kansas Geological Survey, Lawrence, Kan.

Sampson, R. J. 1994. *Surface III Manual*. Surface III Office, Kansas Geological Survey, Lawrence, Kan.

Siegel, S. and Castellan, N. J. 1988. *Nonparametric Statistics for the Behavioral Sciences* (Second Edition). McGraw-Hill, New York.

Simon, H. 1991. *Models of My Life*. Basic Books, New York.

Snapper, E. and Troyer, R. J.1971. *Metric Affine Geometry*. Academic Press, New York.

Snyder, J. P. 1987. *Map Projections—A Working Manual*. U.S. Geological Survey Professional Paper 1,395. U.S. Government Printing Office, Washington, D.C.

Tabachnik, B. G. and Fidell, L. S. 1989. *Using Multivariate Statistics*. Harper and Row, New York.

Taylor, P. J. 1975. "Distance Decay in Spatial Interactions." *Concepts and Techniques in Modem Geography*, #2, Geo Abstracts Ltd., Norwich, England.

Waters, N.M. 1989. “Geography, Microcomputers and the Use of Color." *The Operational Geographer*, vol. 7 #3, pp. 33-38.

Waters, N. M. 1994a. "The Geographic Calculator." *GIS W0RLD*, vol.6, # 5, p. 64.

Waters, N. M. 1994b. "Statistics: How Much Should the GIS Analyst Know?" *GIS WORLD*, vol. 6, # 3, p. 62.

Weibel, R. and Heller, M. 1991. "Digital Terrain Modeling." Chapter 19, pp. 269-297, in Maguire et al., op. cit.

Wellar, B. and Wilson, P. 1993. "Contributions of GIS Concepts and Capabilities to Scientific Inquiry: Initial Findings." Pp. 753-767 in the *Proceedings of GIS/LIS '93*, Association of American Geographers, Washington, D.C.

Williams, R. B. G. 1984. *Introduction to Statistics for Geographers and Earth Scientists*. MacMillan Publishers Ltd., London.

Wilson, A. G. and Kirkby, M. l. 1980. *Mathematics for Geographers and Planners*. Clarendon Press, Oxford.
