INT. J. GEOGRAPHICAL INFORMATION SYSTEMS, 1994, VOL. 8, NO. 5, 411–427

Research Article

Simulating and mapping spatial complexity using multi-scale techniques

LEE DE COLA
U.S. Geological Survey, 521 National Center, Reston, VA 22092, U.S.A

(*Received 25 January 1993; accepted 9 December 1993*)

Abstract. A central problem in spatial analysis is the mapping of data for complex spatial fields using relatively simple data structures, such as those of a conventional GIS. This complexity can be measured using such indices as multi-scale variance, which reflects spatial autocorrelation, and multi-fractal dimension, which characterizes the values of fields. These indices are computed for three spatial processes: Gaussian noise, a simple mathematical function, and data for a random walk. Fractal analysis is then used to produce a vegetation map of the central region of California based on a satellite image. This analysis suggests that real world data lie on a continuum between the simple and the random, and that a major GIS challenge is the scientific representation and understanding of rapidly changing multi-scale fields.

1. Introduction

A major cartographic challenge has traditionally been the provision of spatial information for the location of resources, settlement of the land, and transportation across the surface of the Earth (Stegner 1962). But now that most of the Earth's features have been mapped and much of its land exploited, scientists are turning to the problem of understanding phenomena at a much wider range of physical and temporal scales. In effect, having mapped the important *objects* (the things that are important to locate in space), we turn our attention to *fields* (the complex and everchanging stuff of physical reality) (Goodchild and Gopal 1990). Because human welfare is now strongly dependent upon understanding and adapting rapidly to evolving transformations in the environment, a prime cartographic and analytical challenge today is to characterize dynamic spatial patterns. Topography and land cover, whether physical (e.g., soils), natural (e.g., vegetation), or human-induced (e.g., urbanization), can be regarded as kinds of biogeophysical and socioeconomic fields whose representation requires new analytical and cartographic techniques.

The objective of this paper is to contribute to the theory and practice of mapping these fields. Section 2 develops a formal model that demonstrates how spatial data represented at multiple scales can be regarded as an abstract field. Section 3 develops two measures of spatial complexity: multi-scale variance and multi-fractal dimension. Section 4 illustrates the statistical behaviour of these measures for three simulated datasets that span a continuum of spatial complexity from the simple to the random. Section 5 uses multi-fractal dimension in the mapping of vegetation in central California. The paper concludes with a few speculations about the nature of complexity.

2. Data and measurement

The discussion that follows is a formal and rather detailed presentation of the problem of characterizing the spatial complexity of fields. The thoroughness of this discussion is useful in clarifying several key ideas surrounding the definition of the fractal dimension of regions in space. The interested reader may find the symbol table at the end of the article useful, while the impatient reader may skim this section and concentrate on the more visual presentation of § 4.

We begin with the raw material for spatial analysis, which are data based on empirical measurements. Let the integer $E = 2$ represent the dimensionality of physical space, so that $x \in \mathscr{R}^2$ will be the location of a sampled data point and $A = \{x\}$ will be a set of data points in physical space. Although $\mathscr{R}^2$ is a continuous space, resources are finite, so the problem of sampling within a limited region arises.

Because it is simply not possible to collect data for all time and from everywhere in the universe, A must be bounded, so that its diameter $|A|$ (the length of the line segment determining its largest axis) must be finite. In fact, there may be strong constraints on the volume of $A = \text{vol}(A) \leqslant |A|^2$. (Note that although the term 'volume' refers in this case to the area of the minimum rectangle bounding A, the term can be generalized to a space of any dimension.) Although a more rigorous treatment of fractal dimension is presented below, table 1 can serve as an example of the kinds of datasets that are typically used to sample a field (Laurini and Thompson 1992).

In addition, because it is physically impossible to measure a phenomenon at every point within physical space $\mathscr{R}^2$, it is necessary to sample in some way. Consider therefore a scale index $\lambda = 0, \ldots, L$ representing decreasing degrees of sampling, so that the structure $\mathbf{A} = \{A_\lambda: \lambda = 0, \ldots, L\}$ includes the highest resolution sample $A_0 = A$ and its generalizations up to A_L, some aggregate location representing all the data points (for example a centroid). (The use of the bold face $\mathbf{A}$ represents a hierarchical data structure composed of a number of sampled layers A_λ). Theoretically it is possible for $\lambda < 0$ and even for $\lambda \to -\infty$, which in the limit represents infinitely small (and therefore infinitely expensive) sampling; hence the sample density can always be increased.

In the most general case the density of sampling (which in the case of regular spatial sampling is termed 'resolution') may be increased regularly, so that the dimensionality of the set $D(A)$ approaches $E = 2$. Such approaches provide a hierarchical data structure spanning a range from the smallest to the largest resolution level. A regular such structure $\mathbf{A}$ would be a pyramid (figure 1 shows an $L = 5$-level pyramid) (Falconer 1990: 33), but a sampling scheme more sensitive to the nature of the data would be a quadtree (Samet and Webber 1988). It is not always acknowledged that any given dataset represents measurements made at a single resolution level. The specificity of A_λ conveys the fact that data are always a sample and that information from each level of $\mathbf{A}$ can be used to tell us something about the phenomenon (Arbia 1990, Strahler *et al.* 1986). Certainly fractal description of such characteristics as irregularity, self-similarity, scaling exponents, and dimension requires a hierarchy.

So far we have set forth the problem of sampling the physical space in which a phenomenon will be measured without saying anything about the empirical nature of the phenomenon itself. Consider the nature of a field as succinctly described by Einstein (1950): 'The physical reality of space is represented by a field whose components are continuous functions of four independent variables—the coordinates of space and time.' Assume that the phenomenon is such a field and let M measurements be made at each point x, so that $\mathscr{R}^M$ is a state space. Now consider just one of the measurements

Table 1. Dimensions of sampling schemes.

Sampling scheme	Fractal dimension
Dust	$D \approx 0$
Transect	$D \approx 1$
Raster	$D \approx 2$
Lattice	$D \approx 3$

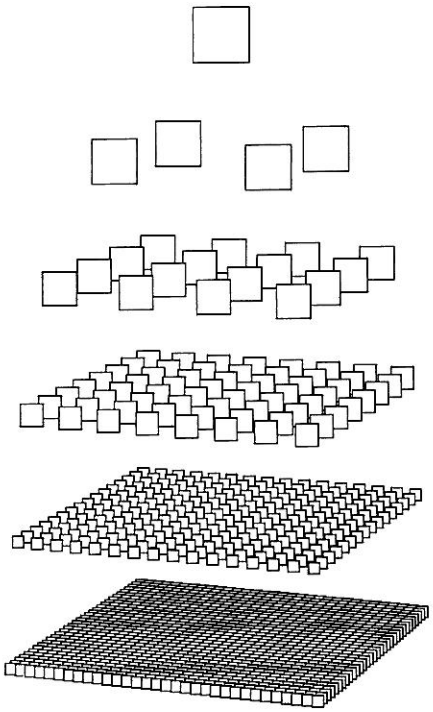

Figure 1. A 5-level data pyramid.

made in this state space and let there be a field measured by this state φ: $\mathscr{R}^2 \rightarrow \mathscr{R}$ at a given time and place (we shall neglect time in what follows). Values of $\varphi(x)$ are represented by measurements $f(A)$ subject to some unknown error $\varepsilon = f - \varphi$ reflecting such problems as sensitivity, reliability, calibration, measurement resolution, categorical resolution (fineness of classes) etc. (Goodchild and Gopal 1990). Let us however assume that these measurements are unbiased (expectation $\mathbf{E}(\varepsilon) = 0$)) and simply define the data set $f(A_\lambda) = \{f(x): x \in A_\lambda\}$.

3. Characteristics of spatial complexity

Analysis is a data-centred activity that focuses on extracting statistical information or parameters from the data (Cressie 1992). Multi-scale analysis involves examining the behaviour of $f(A_\lambda)$. It obviously matters whether we examine $f(A_\lambda)$ by making measurements at distinct scale levels, or whether we examine generalizations of the measurements $(f(A))_\lambda$, by sampling, averaging, or filtering the original data

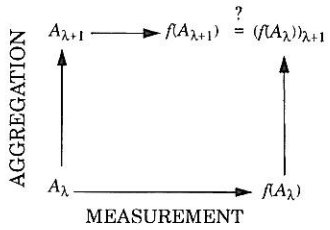

Figure 2. Measurement and aggregation.

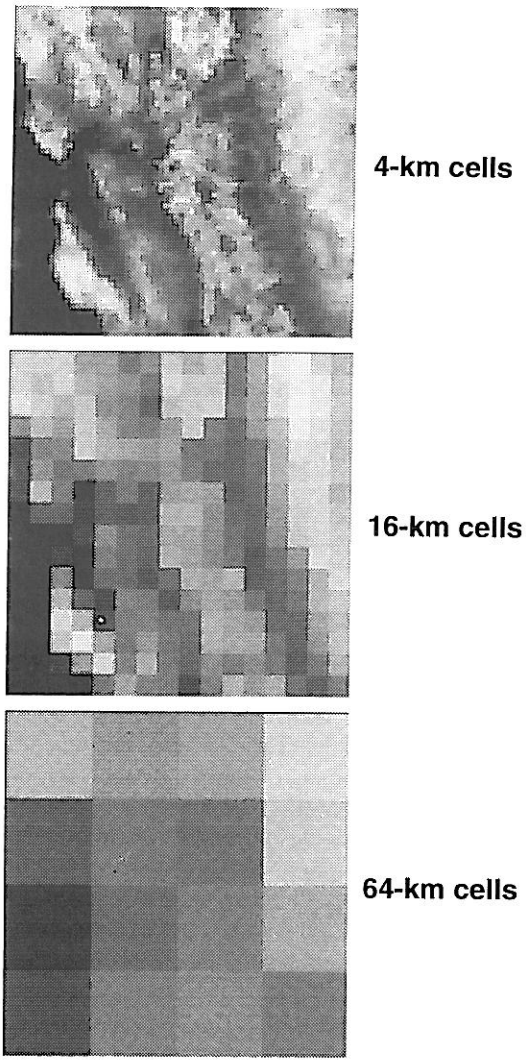

Figure 3. Three levels of an 8-level pyramid image of Central California.

(Gaydos 1992) (see figure 2). Because measurement and aggregation may not be commutative, such considerations are relevant in comparing two satellite sensors of differing resolutions, and in determining, for example, how aggregated (sampled, decimated, averaged, or integrated) measurements made from a 10-metre sensor might differ from raw measurements made from a 20 metre sensor. (Although the problem of constructing multiple-resolution sensors is one of great theoretical as well as practical interest (Basseville *et al.* 1992), this paper examines generalizations $(f(A))_\lambda$ of the data.)

To make the discussion simpler let A be a raster of $L \times L$ cells, where L is an integral power of 2, so that aggregation is straightforward; and let f be restricted to integral values in order to make the discussion consistent with an examination of typically encountered remotely-sensed data. From among the many possible techniques of data aggregation I have chosen a simple one that averages every 2 by 2 nonoverlapping window in the raster. This scheme creates a power-2 image pyramid, such as that of figure 3, which shows three levels from an 8-level satellite image pyramid to be analysed below.

For each aggregation level λ then we have a variance σ^2_λ. Multiscale description of variance examines the behaviour of σ^2_λ with varying λ. In general variance declines with aggregation, and in particular the degree to which it does reflects lack of spatial autocorrelation, so that if $\sigma^2_{\lambda+1}/\sigma^2_\lambda \approx 1$ the spatial structure of the data tends to retain its coherence at varying scale levels (Arbia 1990). Perhaps the simplest way to characterize the multiscale behaviour of variance is

$$\log \sigma^2_\lambda = a + b\lambda \tag{1}$$

whose linear form assumes that variance is scaling with exponent $b < 0$; the coefficient a is the predicted variance at level $\lambda = 0$. Although this expression does not always ideally characterize multi-scale variance, the parameter b will be used below as a summary measure of spatial autocorrelation (Cressie 1992), with small values of b representing low spatial autocorrelation.

A more detailed approach to spatial structure than autocorrelation is fractal analysis, which examines the geometry of the spatial behaviour of f at multiple scales. Let $\{B_k\}$ be a collection of nonoverlapping but possibly disjoint sets such that $\cup_k B_k \supseteq f(A)$. For example, the set

$$B_k = \{y : k \leqslant y\} \tag{2}$$

is simply the half-space bounded below by the plane $y = k$. Let f^{-1} be the inverse of the data function f, corresponding to points in the physical space $\mathscr{R}^2$ that correspond to values in the state space $\mathscr{R}^1$. Then let a segmentation of the physical space be

$$f^{-1}(B_k) = \{x : f(x) \in B_k\} \tag{3}$$

which represents the 'image' of B_k, namely those locations in the physical space corresponding to the state set B_k. Figure 4 illustrates the basic ideas of a state space, an interval, and the segmentation of the physical space into regions, which moreover are quite likely to be disconnected. These regions are spatial and therefore geometric, and as such they may be examined at a specific level λ: $F_{k,\lambda} = f^{-1}(B_k)$, so that strictly speaking there is another hierarchy $\mathbf{F}_k = \{F_{k,\lambda} : \lambda = 0, \ldots, L\}$ corresponding to the forms of each image. At any given scale these regions are geometric objects with sizes and shapes. Their size distributions can be examined nonhierarchically with such models as the lognormal (Aitchison and Brown 1957) or random split (Hagget *et al.* 1977) processes. The Pareto model uses a rank-size hierarchy to examine sizes (De Cola

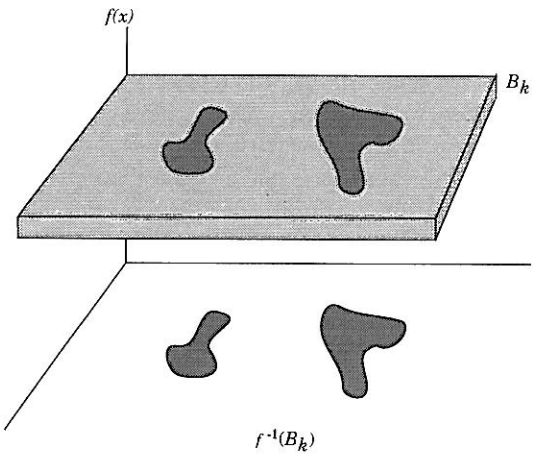

Figure 4. Regions in a state space mapped on the physical space domain of a function by an interval of its range.

1989 a), and the allometric model explains the sizes of regions in terms of other variables (Lee 1989).

Examination of the shapes of these regions is generally a more difficult problem. For data represented at a single scale, area-perimeter analysis (Lovejoy 1982, De Cola 1989 b) is one technique. An approach to this problem for multiple scales can be found in De Cola (1991), in which A is assumed to be a raster and $F_{k,\lambda}$ to be regions of countable cells with determinate boundaries. In this case let $\delta(F_{k,\lambda})$ measure the length of the boundary of the $F_{k,\lambda}$, which in general will vary with scale level λ. The fractal characterization of $F_{k,\lambda}$ reflects the tendency of δ to decrease with scale level (decreased resolution):

$$D(F_{k,\lambda}) = 3 + \frac{\partial \delta(F_{k,\lambda})}{\partial \lambda}, \tag{4}$$

where the constant term is the sum of the physical space ($E = 2$) and the state space ($M = 1$) (Falconer 1990). This expression implies that data (or underlying phenomena) that reveal significantly more detail at higher resolution have a low fractal dimension, as opposed to a structure that is less affected by resolution. This chain of reasoning develops $D(F_{k,\lambda})$ as a scaling exponent that is a description of f and therefore is a characterization of the spatial form of ϕ. Equation (4) indicates that fractal dimension may depend on λ, and in fact, breaks in scale can occur, although estimates of D based on the scaling property do typically have quite high goodness of fit measures (large R^2). So if we assume that $\delta(F_{k,\lambda})$ is perfectly scaling than D will not be a function of λ and D_k therefore becomes a straightforward *multifractal* that characterizes each possible value k of f, rather than all values simultaneously.

Given the dependence of $D(F_{k,\lambda})$ upon so many factors—the data pyramid **A**, the measurements f, the interval B_k, and the scale level λ—it is little wonder that fractal analysis may give a range of answers to the question 'what is the dimension of a phenomenon?' Two points are emphasized. First, geometric sets whose measure (usually area) increases with resolution have a lower fractal dimension than those sets whose measure does not (see the examples below). Corresponding to our intuition, sets that are 'solid' or coherent have high fractal dimensions. Secondly, there is a direct and

inverse relationship between the fractal dimension of a region and that of its boundary: disks have $D = 2$ and they are bounded by curves of dimension $D = 1$. As the D of a set decreases that of its boundary increases, so that we may be tempted colloquially to atribute the dimension of a region's boundary—which gives it shape—to that of the region itself. Nevertheless, as with the concept of spatial autocorrelation, the underlying idea of fractal dimension is both powerful and straightforward: spatial phenomena that are little affected by the resolution of the data structure generally have a high dimension, which is why the negative scaling exponent to be derived from (4) will diminish D. In what follows, moreover, I shall assume that this expression reflects perfect scaling, so that $D(F_{k,\lambda}) = D(F_k)$ is not a function of λ

To review the argument so far, we begin with a hierarchical pyramid $\mathbf{A} \subset \mathscr{R}^2$ in physical space pervaded by a field ϕ for which measurements in a state space $f \in \mathscr{R}$ are made at a set of points x. We next create a set of regions $\{B_k\}$ in the state space of f and invert the function to see what kind of regions $F_{k,\lambda}$ are created at a given scale in the physical space of the function. Finally, we examine the multi-scale behaviour of the variance of the data as well as its multi-fractal geometric form to make inferences about the complexity of the field.

4. Simulation

In general, a field may be regarded as the physical manifestation of some spatial process. Topography reflects such processes as tectonics and weathering, land cover is determined by topography and climate, urbanization responds to accessibility and demographics, and so forth. Measures of complexity should therefore reflect underlying processes, but rather than leaping to an examination of the behaviour of these measures for empirical data, it is useful to explore how they characterize more abstract spatial processes. The usual approach to this problem is simulation, which provides datasets based on controllable factors as well as random events.

The ideas of the previous sections are illustrated using datasets based on three simulation spatial processes as well as one set of empirical measurements. The processes are represented as spatial forms that result from simulations run on a $256^2 = 65\,536$-cell raster, so that the phenomenon are systematicallly sampled on a grid A_0 of $(2^8)^2$ cells that can be aggregated in a pyramid of levels $\lambda = 0,\ldots 8$ (see figure 3). Moroever, each of the processes is confined to the range $f(x) \in [0,\ldots,255]$, although this restriction is merely for convenience. These processes are analysed in four ways; visual appearance, histogram, multi-scale variance σ^2_λ, and multi-fractal dimension $D(F_k)$ corresponding to isarithmic thresholds $B_k = \{y: k \leqslant y\}$.

The first simulated data set is a random noise image, specifically for the Gaussian function, $\phi(x) \sim N(\mu, \sigma)$ shown in figure 5 (*a*). In the present case the 65 536 cell values are normally distributed with mean $\mu = 16$ and variance $\sigma^2 = 16$ (see the histogram of figure 6 (*a*)). The variance of the Gaussian process is nearly perfectly scaling, declining with the data volume by a factor of 4 with each aggregation, as shown in figure 7 (*a*). In this case equation (1) yields the estimate $b = -2$. Such multi-scale variance indicates a complete lack of spatial autocorrelation. The few values, near μ, for which it is possible to estimate D from equation (4), give $D \sim 0$ (figure 8 (*a*)); they are essentially individiual locations (points) whose values are uncorrelated with those of their neighbours (De Cola 1989a). Another way of interpreting this spatial pattern is that it is virtually all edge, surrounding extremely complex regions of no coherence.

Among the simplest spatial processes are those that can be completely characterized as differentiable, smooth mathematical functions. The second simulated process is a

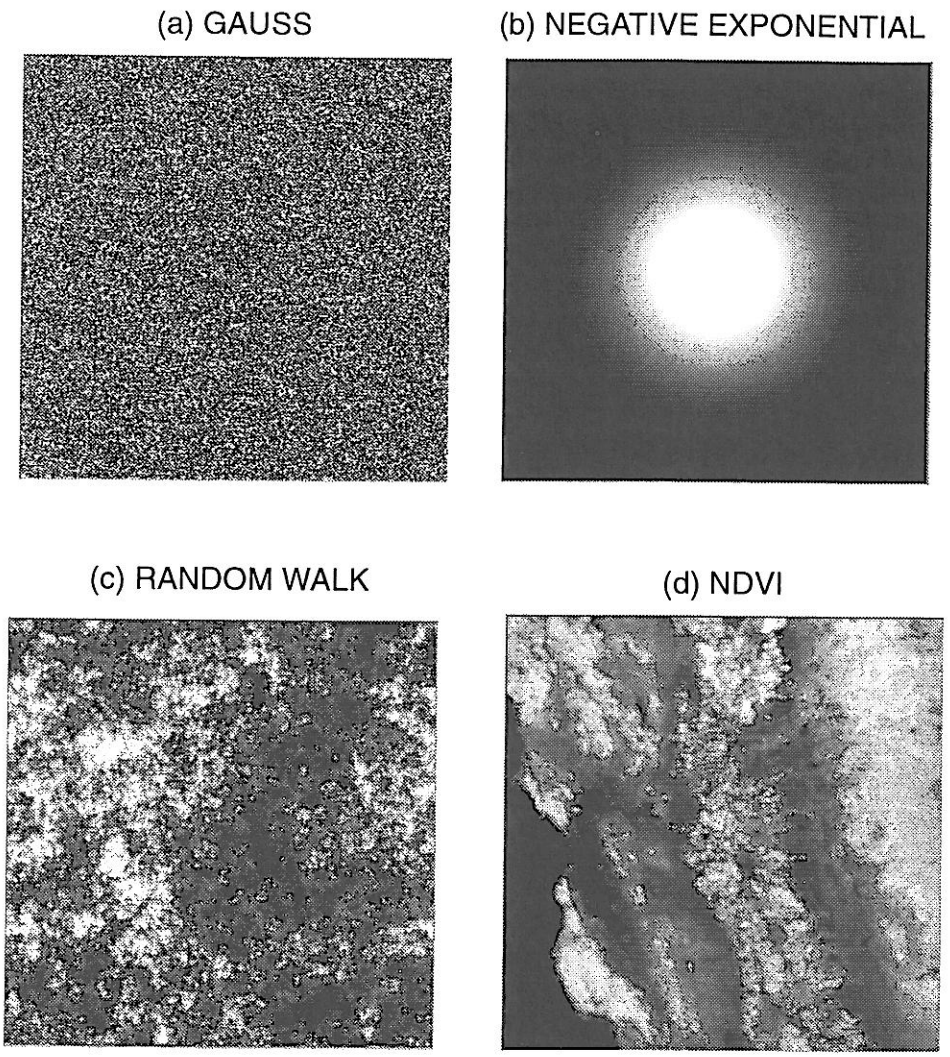

Figure 5. Three simulated processes and a satellite image.

negative exponential function $\phi(x) \propto e^{-|x-c|^2}$ where the values of ϕ of a cell x decline with its distance from c, the centre of the raster. Figure 5 (b) shows data sampled from this process, which resembles a round hill with a summit of $\phi = 100$ smoothly sloping to a plane of value $\phi = 0$. Figure 6 (b) is the histogram, which is strongly skewed because most of the values are concentrated around the value 0. The multiscale variance σ_λ^2 for these data, shown in figure 7 (b) is basically insensitive to aggregation because $f^{-1}(B_k)$ is simply a set of concentric disks whose shape at each aggregation level is unchanged. Even after 6 aggregations (yielding just 16 values) variance has been reduced by only 30 per cent. The extremely high level of spatial autocorrelation in these data is illustrated by the fact that, $b = -0{\cdot}06$ (note that this coefficient is estimated only for the data at $\lambda = 0, \ldots 6$ and not to the outlier at level 7, lest we overstretch the notion of regression. Figure 8 (b) shows that the dimensions of these disks estimated for values from 0 to 100, are all nearly 2, corresponding to simple Euclidean figures. The spatial

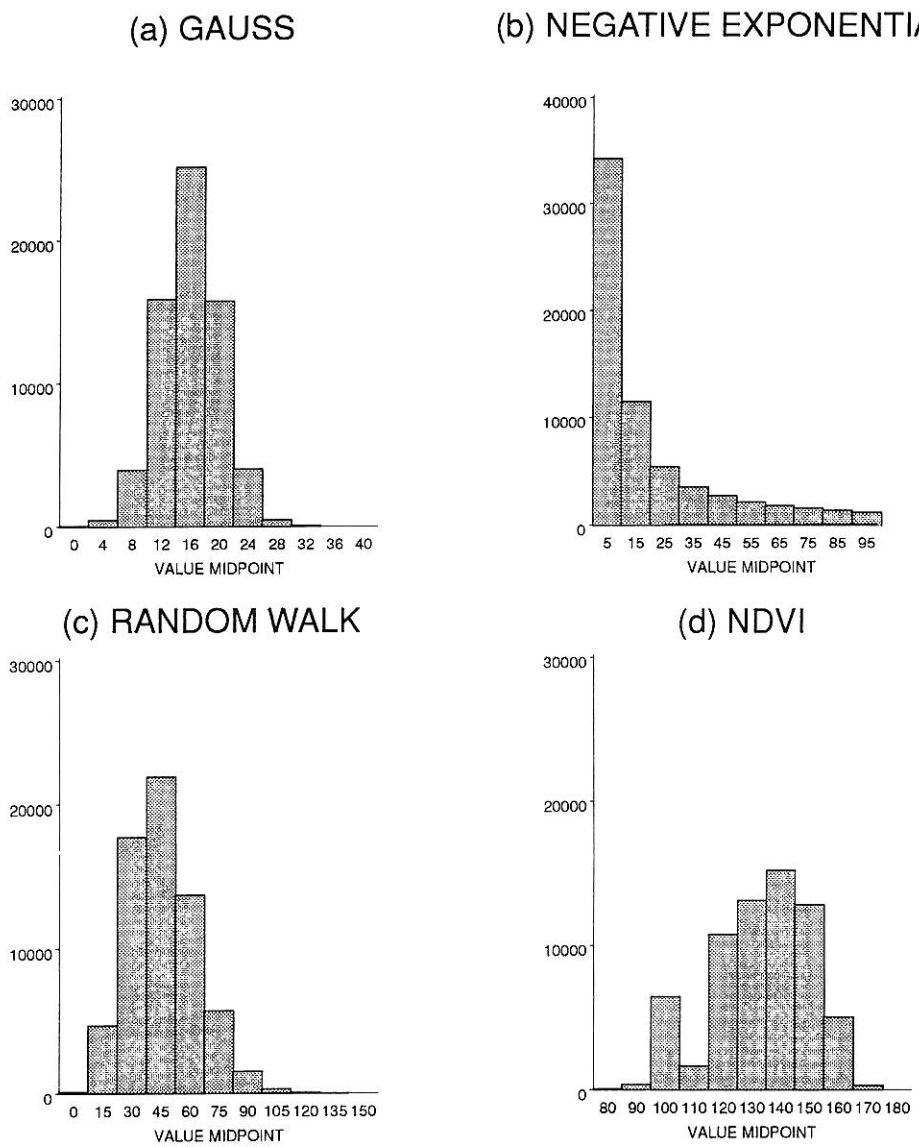

Figure 6. Histograms of the data shown in figure 5.

process that gives rise to these data is very artificial and lies on the simple end of a complexity continuum; it can be thought of as corresponding to a phenomenon that could be cartographically represented as a set of concentric circles.

The above two processes were chosen in part because they bound the spatial complexity continuum, from the simple to the completely random. The Gaussian image is noise whose inverse regions F_k are simply scattered points, and the negative exponential process is just a smooth 'hill' with circular inverse regions of $D = 2$. Their spatial structure is obvious—the value of ϕ at each location has either nothing or a great deal to do with that of its neighbours. These processes do not invite further examination.

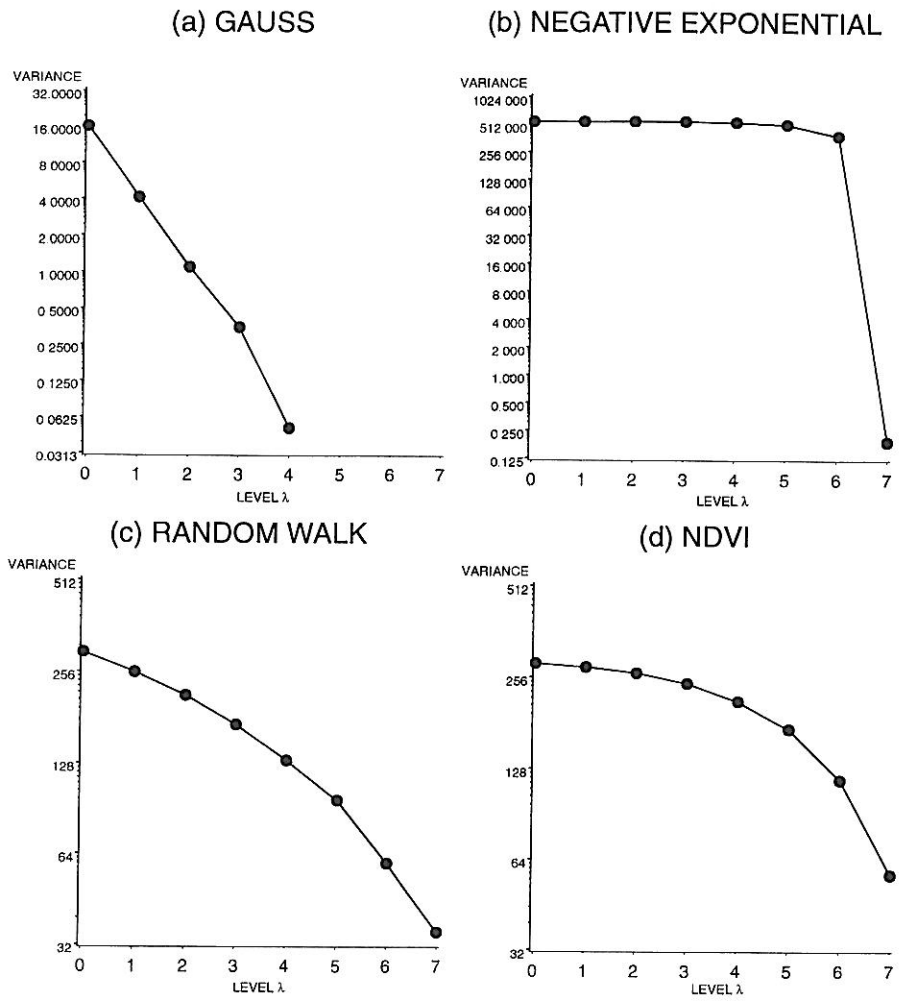

Figure 7. Multi-scale variance σ_λ^2 of simulated and satellite image data.

More interesting is a third process, shown in figure 5 (*c*), a 3 000 000-step random walk on the 65 536 cells of the raster with a toroidal topology (Lam and De Cola 1993, chapter 14). Its structure is quite complex, although this is not suggested by its histogram in figure 6 (*c*). This image is similar to data encountered in remote sensing and digital elevation modelling. This complexity is illustrated by the behaviour of the multiscale variance shown in figure 7 (*c*), which declines less steeply with aggregation than does the Gaussian process but more so than does the negative exponential process. Although variance is not strictly scaling (as it would be if figure 7 (*c*) were linear), equation (1) yields a value of $b = -0{\cdot}47$, which is between that of the Gaussian ($b \approx -2$) and the negative exponential ($b \approx 0$) processes.

Particularly interesting is the behaviour of $D(F_k)$ for the random walk process. At the extreme values (near $f = 0$ and $f = 100$) $D \approx 0$ because at the valleys and peaks the process is basically point-like. But for mid-range values the regions F_k corresponding to $k \sim \mathbf{E}(f) = 45$ are extremely complicated. For example, the number of cells at $\lambda = 0$ for which $f \geqslant 45$ is 32 382 (about half of the image), and the

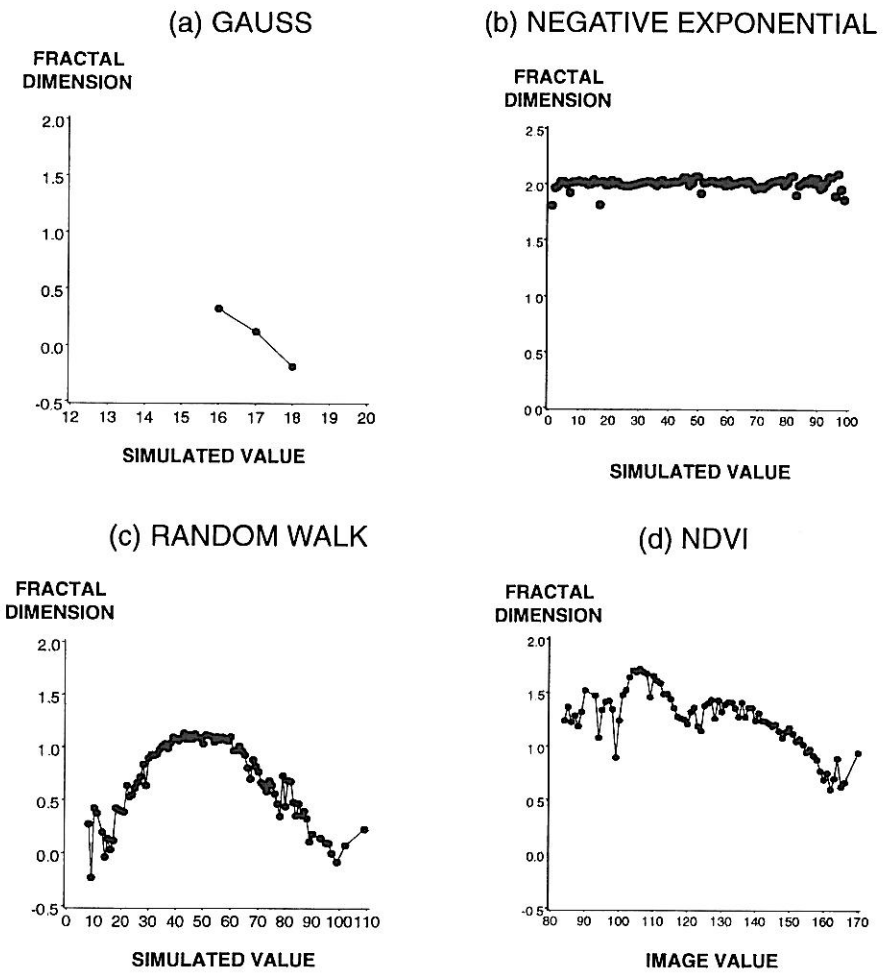

Figure 8. Multifractal dimension $D(F_k)$ of simulated and satellite image data.

number of edges between $f^{-1}(B_{45})$ and its complement is 22 905 out of a possible total number of unlike pairs of about 130 000. Clearly the regions bounded by the $f = 45$ isoline are extremely complicated, as reflected by the fact that $D(F_{45}) = 1{\cdot}12$. Does this mean that these regions are nearly linear? On the contrary, the regions bounded by the isarithm have a dimension approaching 1, which makes them nearly maximally complex, midway betwee the point-like $D \approx 0$ regions of the Gaussian process and the polygonal $D \approx 2$ regions of the negative exponential. Unfortunately for cartography, most real spatial data manifest this complexity, which makes them difficult to map.

5. Mapping vegetation

The final dataset to be analysed is not a simulation, but a satellite image of central California centred around the San Jose–San Francisco–Sacramento megalopolis (figures 3, 5 (*d*) and 9). The data, measured on an $L = 8$ or 256^2 raster of 1-km cells, are normalized difference vegetation index (NDVI) measurements from the U.S. Geological survey's EROS data Center (Jenson 1991, Loveland *et al.* 1991). The values of this index are computed as follows: NDVI = (IR − Visible)/(IR + Visible), where IR

Figure 9. Location of the central California NDVI image shown in figures 5(*d*) and 10.

is the infrared channel and Visible is the red channel from the National Oceanic and Atmospheric Administration's Advanced Very High Resolution Radiometer (AVHRR) sensor. The equation produces NDVI values in the range -1 to $+1$, where negative values generally represent clouds, snow, water, and other non-vegetated surfaces, while positive values represent vegetated surfaces. The values are then rescaled to a range [0, 255] and geometrically registered (EROS Data Center 1991, De Cola 1920. The histogram of this image, shown above in figure 6(*d*), is bimodal, the lower maximum reflecting the large number of low-brightness water pixels of the Pacific Ocean and the bays. The multiscale variance of this complex image is similar to that of the random walk, with equation (1) yielding $b = -0{\cdot}29$, which indicates that multi-scale variance does not decline as rapidly with scale as does the random walk data, and that spatial autocorrelation is fairly high.

Multifractal analysis of this image presents an interesting pattern, midway between the polygonal simplicity of the negative exponential and the raster complexity of the random walk (see figure 8(*d*)). The multifractal dimension $D(F_k)$ is quite high and appears to have three distinct intervals: disconnected values of D for $k < 100$, and two parabolic regions, $100 \leqslant k \leqslant 120$ and $k > 120$. The interval in which $D(F_k)$ is rising to a maximum of 1·72 corresponds to the darkest regions of the image, mainly the Pacific Ocean and the San Francisco and San Pablo Bays. On the other hand, D_k declines rapidly for the higher values of the image, corresponding to scattered regions of coniferous forest (Loveland *et al.* 1991) in the Sierra and Coast Ranges. These multi-fractal dimensions, it will be recalled, apply to thresholded intervals in the data, where the interval in the domain of f is open-ended (see equation (2)).

It is clear so far that the random walk and NDVI data (unlike the Gaussian and negative exponential) are truly multifractal, because D varies strongly with k.

The Gaussian process is inherently point-like, and the negative exponential hill is a surface, neither of which is particularly complex. Yet even the successful statistical parameterization of regions in the NDVI data does not mean that they can be easily rendered cartographically; this is the heart of the vector-raster dichotomy. The production of maps from images uses techniques designed to increase $D(F_k)$ based on such cartographic criteria as smoothness or interpretability, yet generalization almost always conflicts with such statistical criteria as error minimization (Morrison 1971, Lam 1983). A critical mapping problem, then, is that while we wish the width of B_k to be small enough to correspond to a meaningfully specific range of states, we also want $D(F_k)$ to be large enough to represent a visually interpretable region in the physical space—but these objectives conflict. A good map of a phenomenon should illustrate it using a few broad intervals that are represented on the ground by few large regions (Robinson *et al.* 1984). The specific criteria for a visually interesting map of a complex biogeophysical phenomenon are that:

— there should be relatively few intervals;
— the intervals should comprise a relatively large part of the map;
— the intervals should cover the range of values of the mapped variable; and
— the regions representing the intervals should be spatially simple.

Up to now we have used equation (4) to compute fractal dimensions for threshold intervals based on each **individual** value in the data: $B_k = \{y\colon k \leqslant y,\ k \in [\min, \max]\}$ see (2)), which approach has given a fractal dimension for each value. Now let us turn to the somewhat more complicated problem of examining *all* intervals

$$b_{k,l} = \{y\colon k \leqslant y \leqslant l,\ (k,l) \in [\min, \max]\} \tag{5}$$

First, it should be noted that because the pair (k, l) may be chosen independently of one another the number of such intervals is roughly equal to $(\max - \min)^2/2$, which can be quite large; for example, if an 8-bit integer extends over its full dynamic range then the number of intervals that can be examined is $2^{15} = 32\,768$, and looking at all of the corresponding fractal dimensions $D_{k,l}$ can be very time consuming. Note further that examining the parallelepipeds (boxcars) for a multi-band image is even more daunting because such intervals in the state space multiply with its dimensions—for example, a 3-band image would require the analysis of over 10^{13} 3-dimensional intervals. But in the case of the central California data, $\{\min, \max\} = [82, 174]$, so that the number of intervals to examine is only 4324.

In general an effective cartographic rendition of a field generates intervals (k, l) that represent relatively narrow range of $B_{k,l}$ corresponding to regions that both contain a large number of cells (the area of $F_{k,l}$) and have relatively large fractal dimension, so that the regions should be relatively simple and easy to represent as few polygons. One way to seek such intervals is to examine pairs (k, l) corresponding to particularly high values of $D_{k,l}$. It should be possible to select such pairs using the criteria listed above, but a preliminary approach based on judgement rather than rules uses the information in table 2 to select six intervals from the 4324 pairs.

This process of interval selection would be improved with the addition of supervised classification based on spectral information, as well as the automation of the spatial analysis. The technique does however give us additional spatial information, beyond simple rules based on the statistical distribution of the measurements of a field, in order to help create effective maps. This technique was used to generate the interpolated contours shown in figure 10. This map is based on the level $\lambda = 5$ generalization of the

Table 2. Intervals selected to represent NDVI regions.

Interval	Interval limits		Interval width	Cells	Cells per interval width	$D_{k,l}$
	Lower: k	Upper: l				
1	82	97	16	649	41	1·332
2	98	101	4	5730	1432	1·501
3	102	131	30	21568	719	1·301
4	132	135	4	5870	1467	1·299
5	136	152	17	24256	1427	1·152
6	153	174	22	7456	339	1·012

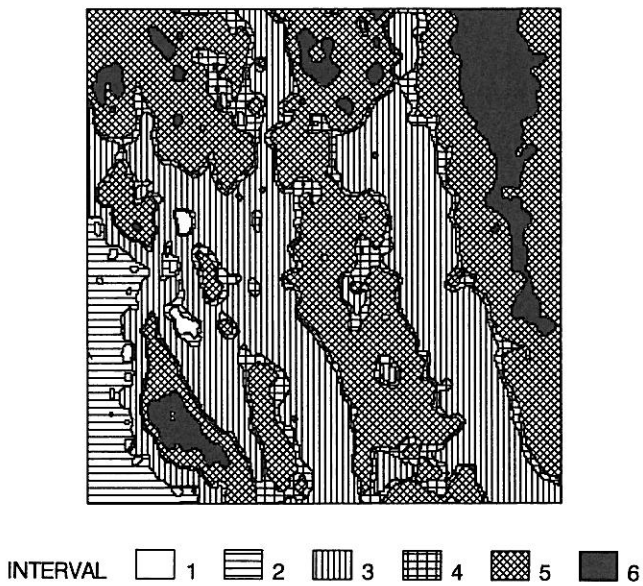

Figure 10. Level $\lambda = 5$ NDVI data for the central California region classified into six intervals and interpolated to a 256^2 raster.

data (an 8 by 8 array of generalized values) that has been interpolated to a level $\lambda = 0$ (256 by 256) array using bilinear interpolation (SAS Institute 1990). This map therefore represents a semiautomated segmentation of the original based on a regionalization of adjacent values into six intervals generating relatively large-dimensional (spatially simple) regions. In this sense, the technique has produced a satellite-measured vegetation map that can be used to make thematic interpretations of the phenomenon and can obviously help in the determination of land cover. First, we clearly see the relatively low values associated with the Pacific Ocean, as well as two large regions of very low NDVI in San Pablo and San Francisco Bays noted above (these may be regions of sediment or pollution). Somewhat more complex but relatively coherent are two predominantly agricultural regions (intervals 3 and 5) separated by a relatively simple but small transition transition zone (interval 4). At the other end of the continuum

are the forested regions (interval 6) largely associated with the western slopes of the Sierra Nevada and Coast Ranges.

6. Conclusion

The extraction of information from data is rarely a straightforward process, particularly as the amounts of data and the complexity of the phenomenon they measure increases. The cartographic challenge of summarizing, characterizing, and visualizing information has never been greater (Méaille and Wald 1990). Some kinds of data are quite easy to characterize. For example, Gaussian noise is fully described by its mean and standard deviation, while the spatial structure of simple polygons describing low-order mathematical functions is immediately obvious. It is phenomena like random walks (the results of chaotic processes or random events with memory)—and especially biogeophysical and socioeconomic fields—that really interest us. These processes may be thought of as forming a continuum, as shown in figure 11. Although the discussion here has been confined to systems unrelated in time, I suggest an evolutionary scenario in which simple processes evolve through complex patterns that eventually degrade into random noise (Nicolis and Prigogine 1989). It is the middle stages that are difficult to describe.

This discussion demonstrates that spatial phenomena, whether simulated or empirical, can sometimes be characterized as fields that give rise to multiscale data whose spatial complexity may be not only summarized by multiscale variance but also specifically measured the multifractal dimension. The latter measure can be used to make choices for the mapping of appropriate intervals in the range of the data from a satellite image. This technique facilitates the creation of a map (Vasiliev *et al.* 1990) from an image (Besag 1987), and we have used the fractal dimensions of regions bounded by isolines to select those regions that are relatively large (comprise a significant proportion of the image) and relatively simple (have a high fractal dimension).

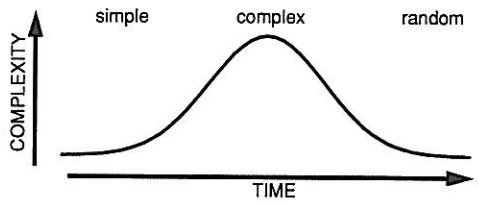

Figure 11. System complexity.

Table 3. Data and information.

Characterization	Data	Information
Representation	Image	Map
Features	Fields	Objects
Discipline	Science	Management
Queries	What is at x?	Where is y?
Everyday experience	Stuff	Things

This argument can itself be generalized to explore the tension that often exists between the realms of data and information, as shown in table 3. Spatial data frequently are in the form of images that need to be represented as maps containing information. In many cases these data represent complex and rapidly changing multidimensional fields that often must be generalized as objects, for example within a GIS. While scientists are usually concerned with ways of characterizing fields across space, asking what is or might be the value of the field at a given location (Openshaw *et al.* 1990), managers frequently wish to know where a given object is, or can or should be located (Guptill 1990). Yet even in everyday experience we recognize what might be crudely called the distinction between the complex stuff of everyday experience that we all recognize but find it quite difficult to describe, and a preoccupation with the management of things in space. These tensions are a central theme of geography. Indeed, a fruitful dialogue between science and management crucially depends upon a keen sensitivity to the 'geography of data' in our efforts to understand the data of geography.

List of symbols

E integer representing the dimension of physical space
$\mathscr{R}$ the real line ($\mathscr{R}^2$ = the plane, etc.)
A a set of data points in physical space
x a point in physical space where measurements are made or a field exists
λ integer index representing scale or aggregation level
L maximum level of aggregation
$\mathbf{A}$ the data structure $\{A_0, \ldots, A_1\}$
M integer representing the dimension of the state space
$\phi(x)$ a field value at point x
$f(x)$ measurement at point x
ε error between measurement and field value
$\mathbf{E}$ expected value
σ^2 variance
a predicted variance
b scaling exponent of multiscale variance
B subset of state space
k a value in the state space
f^{-1} inverse mapping from state to physical space
F subset (region) of physical space
$\delta(F)$ length of a boundary of a region
D fractal dimension of a set or space
N Gaussian (normal) probability function

References

AITCHISON, J., and BROWN, J. A. C., 1957, *The Lognormal Distribution* (Cambridge: Cambridge University Press).

ARBIA, G., 1990, *Spatial data configuration in statistical analysis of regional and related problems* (Hingham, MA: Klewer).

BESAG, J., 1986, On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society,* **B48,** 259–302.

BASSEVILLE, M., BEINSTE, B.,CHOU, K. C., GOLDEN, S. A., NIKONKHAH, R., and WILLSKY, A. S., 1992, Modeling and estimation of multiresolution stochastic processes *I.E.E.E. Transactions of Information Theory,* **38,** 766–784.

CRESSIE, N., 1992, *Statistics of spatial data* (New York: Wiley).

DE COLA, L., 1989 a Pareto and fractal description of regions from a binomial lattice. *Geographical Analysis,* **21,** 74–81.

DE COLA, L., 1989 b, Fractal analysis of a classified Landsat scene. *Photogrammetric Engineering and Remote Sensing,* **55,** 601–610.

DE COLA, L., 1991, Fractal analysis of multiscale spatial autocorrelation among point data. *Environment and Planning. A,* **23,** 545–556.

DE COLA, L., 1992, Multiscale interaction between topography and vegetation in Colorado. In *Proceedings of Resource Technology 92* (Bethesda: ASPRS–ACM).

EINSTEIN, A., 1950, On the generalized theory of gravitation. *Scientific American,* **182,** 13–17.

EROS DATA CENTER, 1991, Conterminous U.S. AVHRR companion disc, CD/ROM AVHRR-9107 Eros Data Centre, U.S. Geological Survey, U.S.A.

FALCONER, K. J., 1990, *Fractal Geometry: Mathematical Foundations and Applications* (New York: Wiley).

GAYDOS, L. J., 1992, Scale dependent measurement of vegetation in urban areas, USGS, Western Mapping Center, Menlo Park, CA.

GOODCHILD M., and GOPAL, S., 1990, *The accuracy of spatial databases* (London: Taylor & Francis).

GUPTILL, S. C., 1990, Multiple representations of geographic entities through space and time. In *Proceedings of the 4th International Symposium on Spatial Data Handling, Zürich* (Zürich: International Union), vol. 2, pp. 859–868.

HAGGETT, P. A. D., CLIFF, A., and Frey, A., 1977, Locational analysis in human geography (London: Edward Arnold).

JENSEN, S. K., 1991, Applications of hydrologic information automatically extracted from digital elevation models. *Hydrological Processes,* **5,** 31–44.

LAM, N., 1983, Spatial interpolation methods: a review. *The American Cartographer,* **10,** 129–149.

LAM, N., and DE COLA, L., 1993, *Fractals in geography* (Reading MA: Prentice-Hall).

LAURINI, R., and THOMPSON, O., 1992, *Fundamentals of Spatial Information Systems* (San Diego: Academic Press).

LEE, Y., 1989, An allometric analysis of the US urban system: 1960–80. *Environment and Planning A,* **21,** 463–476.

LOVEJOY, S., 1982, Area-perimeter relation for rain and cloud areas. *Science,* **216,** 185–187.

LOVELAND, T. R., MERCHANT, J. W., OHLEN, D. O., and BROWN, J. F., 1991, Development of a land cover characteristics data base for the conterminous U.S. *Photogrammetric Engineering and Remote Sensing,* **57,** 1453–1464.

MÉAILLE, R., and WALD, L., 1990, Using geographical information system and satellite imagery within a numerical simulation of regional urban growth. *International Journal of Geographical Information Systems,* **4,** 445–456.

MORRISON, J. L., 1971, Method-produced error in isarithmic mapping. Monograph, Department of Geography, University of Wisconsin.

NICOLIS, G., and PRIGOGINE, T., *Exploring Complexity: an Introduction* (New York: Freeman).

OPENSHAW, S., CROSS, A., and CHARLTON, M., 1990, Building a prototype geographical correlates exploration machine. *International Journal of Geographical Information Systems,* **4,** 297–311.

ROBINSON, A. H., SALE, R. D., MORRISON, J. H., and MUEHRCKE, P. C., 1984, *Principles of Cartography,* 5th edn (New York: Wiley).

SAMET, H., and WEBBER, R. E., 1988, Hierarchical data structures and algorithms for computer graphics Part I: fundamentals. *I.E.E.E. Computer Graphics and Applications,* **8,** 48–68.

SAS INSTITUTE, INC., 1990, *SAS/GRAPH Software: Reference,* Vol. 2 (Cary NC: SAS Institute, Inc).

STEGNER, W., 1962, *Beyond the hundredth meridian* (New York: Scribner's).

STRAHLER, A. H., WOODCOCK, C. E., and SMITH, J. A., 1986, On the nature of models in remote sensing. *Photogrammetric Engineering and Remote Sensing,* **20,** 121–139.

VASILIEV, I., FREUNDSCHUH, S., MARK, D. M., THEISEN, G, G. D., and MCAVOY, J., 1990, What is a map? *The Cartographic Journal,* **27,** 119–123.