COMMUNITY ECOLOGY 14(2): 144-152, 2013
1585-8553/$20.00 © Akadémiai Kiadó, Budapest
DOI: 10.1556/ComEc.14.2013.2.4

Sample size requirements for fractal dimension estimation

N. C. Kenkel

Department of Biological Sciences, University of Manitoba, Winnipeg, Canada. Email: kenkel@cc.umanitoba.ca

Keywords: Box counting, Point pattern, Probability theory, Resolution, Scale.

Abstract: Over the past several decades, fractal geometry has found widespread application in the theoretical and experimental sciences to describe the patterns and processes of nature. The defining features of a fractal object (or process) are self-similarity and scale-invariance; that is, the same pattern of complexity is present regardless of scale. These features imply that fractal objects have an infinite level of detail, and therefore require an infinite sample size for their proper characterization. In practice, operational algorithms for measuring the fractal dimension D of natural objects necessarily utilize a finite sample size of points (or equivalently, finite resolution of a path, boundary trace or other image). This gives rise to a paradox in empirical dimension estimation: the object whose fractal dimension is to be estimated must first be approximated as a finite sample in Euclidean embedding space (e.g., points on a plane). This finite sample is then used to obtain an approximation of the true (but unknown) fractal dimension. While many researchers have recognized the problem of estimating fractal dimension from a finite sample, none have addressed the theoretical relationship between sample size and the reliability of dimension estimates based on box counting. In this paper, a theoretical probability-based model is developed to examine this relationship. Using the model, it is demonstrated that very large sample sizes – typically, one to many orders of magnitude greater than those used in most empirical studies – are required for reliable dimension estimation. The required sample size increases exponentially with D, and a 10^D increase in sampling effort is required for each decadal (order of magnitude) increase in the scaling range over which dimension is reliably estimated. It is also shown that dimension estimates are unreliable for box counts exceeding one-tenth the sample size.

Introduction

Natural phenomena characteristically show high levels of structural and organizational complexity over a broad range of spatial and temporal scales (Schroeder 1991, Falconer 2013). This scale-invariant complexity limits the use of classic Euclidean geometry in quantifying natural patterns and processes (Kenkel and Walker 1996). The alternative geometry of fractals (Mandelbrot 1967) often provides a more realistic and meaningful descriptor of nature, as evidenced by the wide-ranging application of fractal methodology in ecology (Seuront 2010, Bez and Bertrand 2011) and other disciplines (Falconer 2003, Brewer and Di Girolamo 2006, Sun et al. 2006, Lopes and Betrouni 2009).

The defining features of a fractal are self-similarity and scale-invariance; fractals possess an infinite (and unresolvable) level of detail (Theiler 1990). The basic requirement of a fractal, therefore, is that structural "irregularities" are preserved at all – including the finest – scales (Hall 1995). In practice, estimating the fractal dimension of real-world objects (e.g., point patterns, curves, or surfaces) requires recording the object in a Euclidean embedding space at finite resolution (sample size); that is, one begins with a finite description of the fractal (Theiler 1990, Taylor and Taylor 1991, Hall 1995). Resolution may be limited by the image itself (e.g., digital photograph), the capacity of recording equipment (e.g., scanner, stylus or filter), or sampling limitations (e.g., mapped tree locations, and animal movement paths). As a consequence, the more detailed properties of self-similar objects or processes – that is, the very features that define a fractal – are "smoothed" or destroyed during the recording process (Hall 1995). The process of data recording substantially degrades detailed fractal properties, since higher-magnification features of the original process are smoothed to a greater or lesser extent (depending on the level of resolution). Since higher-magnification properties are destroyed, so too is self-similarity. However, provided that the smoothing is not too severe (i.e., image resolution is high, or sample size is large), lower-magnification features are faithfully preserved and can be used to estimate the fractal dimension provided the data are examined at relatively coarse scales (Hall 1995). An ideal methodological approach provides an estimate of the dimension of the underlying fractal set, not of the finite description of that set (Theiler 1990).

Rigorous theoretical definitions of dimension are impractical for direct numerical estimation but provide the basis for developing operational algorithms for estimating the dimension of finite sets (Theiler 1990). The Hausdorff dimension of a point pattern embedded in a *P*-dimensional Euclidean space is defined by the limit function:

$$D_H = \lim_{r \to \infty} \log [1/N(r)] / \log r \qquad [1]$$

The upper limit of D_H, variously known as the box-count, capacity or fractal dimension, is obtained by covering the embedding space (e.g., two-dimensional surface) with a fixed-size grid. For a specified grid of boxes of size (side

length) r, the number of boxes N(r) containing at least one point is determined. The scaling:

$$N(r) \sim r^{-D}$$

defines the fractal dimension D. In empirical applications only a finite sample size M is available so that the limit $r \rightarrow 0$ cannot be taken. Intuitively, one could apply equation [1] using the smallest r available, but this approach is impractical since limit convergence is logarithmically slow (Theiler 1990). A practical alternative is to plot log N(r) versus log r, estimating D as the negative slope of the log-log plot over a defined range of r:

$$D = -\Delta[\log N(r)] / \Delta[\log r]$$

This approach introduces a new challenge: over what range of r should the slope (i.e., the estimate of D) be determined? This is not an issue when sample size $M \rightarrow \infty$, since the log-log plot of an infinite set is linear over all r. However, in empirical applications the sample size M is necessarily finite; consequently, the range over which the log-log plot is strictly linear is limited at both ends of the scaling range (Theiler 1990, Bernston and Stoll 1997, Halley et al. 2004). At coarse scales (large box size r), the proportion of occupied grid boxes rises as r increases, a consequence of the fact that finite empirical patterns (images) are necessarily of limited spatial extent and therefore space filling at coarse scales (Taylor and Taylor 1991, Halley et al. 2004). Above a certain threshold value of r all grid boxes are occupied, and the log-log plot therefore curves as the dimension of the embedding space is approached (Ramsey and Yuan 1990). In practice this problem is alleviated easily by excluding from consideration all values of r that are space filling, e.g., when $N(r) = (1/r)^2$ given a two-dimensional embedding space.

A more pernicious effect arises at finer scales (small box size r). Here, the problem is that the number of occupied boxes N(r) is necessarily bounded by M; that is, the number of non-empty boxes cannot exceed the number of points (Theiler 1990, Taylor and Taylor 1991). At a certain threshold value of r the number of occupied boxes is "saturated" at $N(r) = M$, giving a log-log slope of zero at the finest scales (Liebovitch and Toth 1989). In practice the effect is more insidious, however, since the log-log slope declines very gradually with decreasing r as the saturation limit $N(r) = M$ is approached. The result is a concave downward log-log plot at finer scales, and a corresponding reduction in the slope (i.e., fractal dimension) estimate (Liebovitch and Toth 1989, Bernston and Stoll 1997, Foroutan-pour et al. 1999, Halley et al. 2004, Agterberg 2013). Note that the saturation limit of zero slope (i.e., $D \rightarrow 0$ as $r \rightarrow 0$) is strictly correct since the topological dimension of a finite point set is zero. However, from an analytical perspective this is a trivial result, since it is the dimension of the underlying point pattern that is of interest, not the topological dimension of a finite set (Ramsey and Yuan 1990, Pruess 1995).

Unfortunately, no rigorous theory-based method exists for determining the scaling range over which the log N(r) versus log r plot is strictly linear (i.e., the range that ensures unbiased dimension estimation). The naïve approach (as defined by Gneiting et al. 2012) involves determining the log-log plot slope over the entire scale range, but this produces biased – generally underestimated – dimension estimates (e.g., Pruess 1995, Gonzato et al. 1998, Gneiting et al. 2012). To alleviate this bias, Liebovitch and Toth (1989) suggest excluding the smallest scales at which $N(r) > M/5$, although they provide no theoretical justification for this recommendation. More recent recommendations are often deliberately vague (e.g., Halley et al. 2004), encouraging an empirical "choice by eye" approach to determining the appropriate scale range (Ramsey and Yuan 1990, Foroutan-pour et al. 1999).

The problem of finite sample size (or equivalently, finite image resolution) is a "serious limitation" to empirical dimension estimation (Theiler 1990), making "dimension calculations … very much a large-numbers game" (Ramsey and Yuan 1990). In practical applications, a finite sample requires that N(r) in equation [1] be estimated as <N(r)>, the observed number of boxes containing at least one of M points. In general <N(r)> will underestimate N(r), although for very large sample sizes it provides a reasonably good approximation since, for given r:

$$N(r) = \lim_{M \rightarrow \infty} \langle N(r) \rangle$$

Unfortunately, this limit converges very slowly, implying that a large sample size M is required to obtain a reasonably reliable and accurate, empirically derived estimate of fractal dimension (Theiler 1990). In fact, finite sample size (or equivalently, finite resolution) is the most serious limitation to dimension estimation, a problem that is eliminated only at the $M \rightarrow \infty$ limit (Pruess 1995).

Initial research on empirical fractal dimension estimation, undertaken by experimental physicists and chemists, employed data sets containing tens of thousands of points (e.g., Grassberger and Procaccia 1983). However, later applications in other fields (including ecology, biology, meteorology and geology) often made do with much smaller data sets: "indeed, very small, with numbers of observations ranging from less than a thousand to less than two hundred … such data sets are miniscule" (Ramsay and Yuan 1990). In ecology and other disciplines the serious limitations of using such "miniscule" data sets to estimate fractal dimension are not always appreciated, or are downplayed (e.g., Kallimanis et al. 2002). Numerous empirical studies have demonstrated that small data sets lead to estimation problems that seriously compromise the reliability of dimension estimates (e.g., Ramsey and Yuan 1990, Taylor and Taylor 1991, Pruess 1995). These estimation problems are largely alleviated using very large data sets, but currently there is no consensus as to how 'large' a data set needs to be (Ramsey and Yuan 1990).

While is it known that finite sample size strictly limits the linearity (unbiased dimension estimation) range of a log <N(r)> versus log r plot, the exact relationship between sam-

ple size M and range of log-log plot linearity (in units of decades or orders of magnitude of r) has not been systematically investigated. A rigorous theory-based approach for determining the minimal sample size required to achieve a reliable estimate of fractal dimension (log-log slope) over a specified scaling range is clearly needed, given that a scaling range of two to three decades (i.e., 100 to 1000-fold range in r) is often recommended as a minimal requirement for "proving" fractal self-similarity (e.g., Malcai et al. 1997, Avnir et al. 1998, Gonzato et al. 1998, Ciccotti and Mulargia 2002, Falconer 2003, Halley et al. 2004). Currently, the theoretical sample sizes necessary to achieve this minimal requirement are unknown.

In this study, I develop a probability-based box counting model and use it to determine the theoretical sample sizes required to achieve linearity (i.e., unbiased dimension estimation) of log <N(r)> versus log r plots over a specified scale range. The model is used to determine, for a given sample size M and fractal dimension D, the value r_m at which the log-log plot begins to deviate from linearity as the saturation limit <N(r)> = M is approached. Given r_m, the sample size required to obtain log-log plot linearity over a specified scale range (in units of decades of r) is easily determined. The model is based on box counting of fractal point patterns, but the results are directly applicable to other operational algorithms for estimating D (e.g., variograms and power spectra; Seuront 2010) and other fractal image types (e.g., lines and pixels). The paper also discusses briefly the merits of representing (and analyzing) fractal images as point patterns, rather than as traced or pixelated lines or areas as commonly practiced.

Box-counting probability model

Continuous or complete data (M = ∞)

Random point pattern. Consider an infinite number of random points (uniform random distribution) embedded on the one-dimensional unit interval [0,1]. In the box-counting method, the unit interval is divided into r^{-1} equal-size segments or "boxes" of length r. Following Hamburger et al. (1996), define N(r) = number of boxes of size r that include at least one point. For r^{-1} boxes, the expectation value is:

$$<N(r)> = 1/r$$

The log <N(r)> versus log r plot is strictly linear over all values of r, with slope:

$$\Delta[\log <N(r)>] / \Delta[\log r] = -1$$

over the entire scale range $0 < r \leq 1$.

Next, consider M = ∞ random points embedded on a two-dimensional unit square [0,1] x [0,1]. The box-counting method divides the unit square into a grid of r^{-2} equal-sized "boxes" of side length r. The expectation value is:

$$<N(r)> = 1/r^2$$

The log <N(r)> versus log r plot has slope:

$$\Delta[\log <N(r)>] / \Delta[\log r] = -2$$

over the entire scale range $0 < r \leq 1$.

Note that these derivations are easily extended to cases of M = ∞ random points embedded in three or higher dimensions.

Fractal point pattern. The above derivations can be generalized to include any self-similar fractal set of dimension D. The corresponding expectation value is:

$$<N(r)> = 1/r^D \quad [2]$$

Over the entire scale range $0 < r \leq 1$, the log <N(r)> versus log r has slope:

$$\Delta[\log <N(r)>] / \Delta[\log r] = -D \quad [3]$$

For example, consider Cantor "dust" (D = log(2)/log(3) = 0.6309), a self-similar fractal set on the line (D ≤ 1) obtained by deleting at each iteration the middle third of all remaining portions of the unit interval [0,1]. It follows that box count values are N(1/3) = 2, N(1/9) = 4, and so forth (Schroeder 1991). Equation [2] provides the correct expectation values, e.g.:

$$<N(1/3)> = 1/r^D = [(1/3)^{0.6309}]^{-1} = 2$$

$$<N(1/9)> = 1/r^D = [(1/9)^{0.6309}]^{-1} = 4$$

Equation [3] gives the correct slope of the log-log plot:

$$\Delta[\log <N(r)>] / \Delta[\log r] = [\log 2 - \log 4]/[\log (1/3) - \log (1/9)] = -0.6309$$

Next, consider the two-dimensional Cantor set (D = log (4)/log(3) = 1.2619), a fractal set on the unit square (D ≤ 2). In this case N(r) = 4 at box size r = 1/3 (i.e., 3x3 grid), N(1/9) = 16, and so forth. Equation [2] yields the correct expectation values, e.g.:

$$<N(1/3)> = 1/r^D = [(1/3)^{1.2619}]^{-1} = 4$$

As a third example, consider the Sierpinski carpet (D = log[8]/log[3] = 1.8928). This fractal set is obtained by dividing the [0,1] x [0,1] unit square into 9 sub-squares (3 x 3 grid), removing the central sub-square. The process is then repeated for each of the eight remaining sub-squares, and continued iteratively (Schroeder 1991). Box-count values are therefore N(1/3) = 8, N(1/9) = 64, and so forth. Once again, equation [2] gives the correct expectation values, e.g.:

$$<N(1/9)> = 1/r^D = [(1/9)^{1.8928}]^{-1} = 64$$

Equation [2] calculates box-count expectation values <N(r)>, where $0 < r \leq 1$, for any self-similar fractal set of known dimension D.

Discrete or sampled data (M < ∞)

Random point pattern. Consider now the discrete case, in which a sample of M random points (unit random distribution) is embedded on the one-dimensional unit interval [0,1]. The expectation value for N(r) is:

$$<N(r)> = p / r$$

where p = probability that a box of size r includes at least one point. After placing the first random point on the unit interval, the probability that a given box does not include that point is simply:

$$q_1 = (1\text{-}r)$$

After placing all M points, the probability that a given box does not include a point is:

$$q = (1\text{-}r)^M$$

Since the probability that a box does include a point is $p = 1 - q$, it follows that:

$$\langle N(r) \rangle = [1 - (1 - r)^M] / r$$

Using the same definitions as above, for M points embedded on a two-dimensional unit square [0,1] x [0,1]:

$$q = (1\text{-}r^2)^M$$

Giving:

$$\langle N(r) \rangle = [1 - (1 - r^2)^M] / r^2$$

These derivations are easily extended to cases of M random points embedded in three or higher dimensions.

Fractal point pattern. The above is generalizable to any sampled self-similar fractal set of dimension D. The corresponding expectation value is:

$$\langle N(r) \rangle = [1 - (1 - r^D)^M] / r^D \quad [4]$$

Equation [4] allows investigation of the relationship between box size r and expected box count $\langle N(r) \rangle$ for a set of M points of fractal dimension D. For point patterns of finite sample size, the expected box count has a fixed upper limit of $\langle N(r) \rangle = M$. Thus, the log $\langle N(r) \rangle$ versus log r plot has slope:

$$\Delta[\log \langle N(r) \rangle] / \Delta[\log r] = -D$$

over a limited range of r. With decreasing box size r (that is, as the saturation limit $\langle N(r) \rangle = M$ is approached), a value r_m is reached at which the log-log plot begins to deviate from strict linearity (i.e., when slope < D). It follows that the slope of the log-log plot at values of $r < r_m$ is a biased estimator (underestimate) of D.

Methods

Equation [4] was used to obtain box count expectation values $\langle N(r) \rangle$ as a function of r. A total of 28 simulations were obtained, using four sample sizes ($M = 10^3$, 10^4, 10^5 and 10^6) for each of seven fractal dimensions (D = 0.8, 1.0, 1.2, 1.4, 1.6, 1.8 and 2.0). For each M-D combination, a log $\langle N(r) \rangle$ versus log r plot was obtained by incrementally decreasing r and determining $\langle N(r) \rangle$ from equation [4]. For each incremental change in r, the local slope (i.e., local estimate of D) was computed as:

$$\Delta[\log \langle N(r) \rangle] / \Delta[\log r]$$

When $M = \infty$, the log-log plot has local slope = -D over all values of r. For finite sample size M, as r is incrementally decreased the local slope at some point will begin to deviate from D:

$$|\Delta[\log \langle N(r) \rangle] / \Delta[\log r]| < D$$

For a given sample size M, define r_m as the smallest box size at which the absolute value of the local slope equals D. Since estimates of D are typically expressed to the third or even fourth decimal place (e.g., Gonzato et al. 1998, Foroutanpour et al. 1999, Pérez-Rodríguez et al. 2013), r_m was obtained by decreasing r incrementally until the local slope deviated from D to the third decimal place:

$$\Delta[\log \langle N(r) \rangle] / \Delta[\log r] + D = 0.001$$

The value r_m is the smallest scale value at which the log-log plot is linear with slope = -D, and therefore the lower bound of the scale range that correctly estimates D.

Results

A representative set of log $\langle N(r) \rangle$ versus log r plots at $M = 10^4$ for dimensions D = 0.8, 1.0, 1.2, 1.4, 1.6, 1.8 and 2.0, and corresponding r_m values, are shown in Figure 1. For a given sample size, r_m increases with D (e.g., in Figure 1 $r_m = 0.001$ when D = 1, but $r_m = 0.0316$ when D = 2). As expected, r_m declines with increasing sample size M for a given fractal dimension (Figure 2). For example, at D = 1.2 the value $r_m = 0.021544$ when $M = 10^3$, versus $r_m = 0.000464$ when $M = 10^5$. Results from the 28 simulations revealed that M, D and r_m are related through a simple power law:

$$r_m = [0.1\ M]^{-1/D}$$

or

$$M = 10\ r_m^{-D} \quad [5]$$

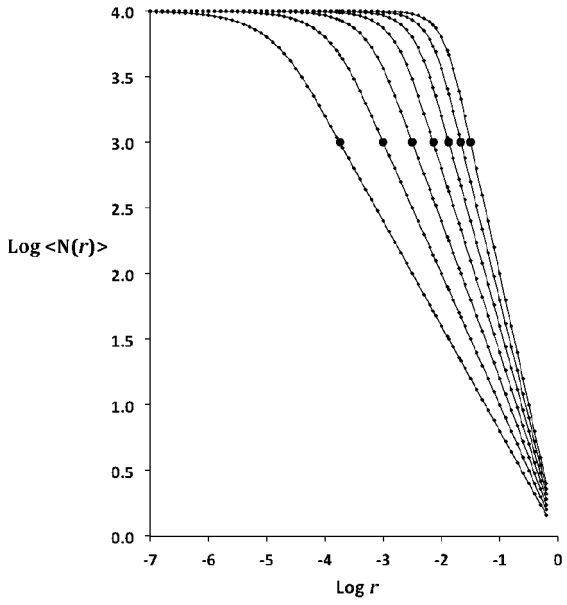

Figure 1. Theoretical log $\langle N(r) \rangle$ versus log r plots at sample size $M = 10^4$ for D = 0.8, 1.0, 1.2, 1.4, 1.6 1.8 and 2.0 (left to right). The filled circles indicate r_m, the point at which the log-log plot begins to deviate from slope = -D as the limit log $\langle N(r) \rangle$ = log M = 4.0 is approached. Logarithms are base 10.

Using this power law relationship, r_m (the smallest usable box size for estimating D from a log-log plot) can be determined for a sample of M points at any specified fractal dimension D. For example, given $M = 10^3$:

D	Smallest Box Size (r_m)	Maximum Grid Size (r_m^{-1} x r_m^{-1})
1.25	0.025119	~ 40 x 40 boxes
1.5	0.046416	~ 22 x 22 boxes
2.0	0.1	10 x 10 boxes

This result is not especially useful in empirical studies, i.e., when D is not known *a priori* and is to be estimated. A more useful result is obtained by rearranging equation [5] as:

$$r_m^D = 10/M$$

Substituting r_m^D into equation [4] gives:

$$\langle N(r_m)\rangle = [1 - (1 - 10/M)^M] / (10/M)$$

The $(1 - 10/M)^M$ term is vanishingly small (~ 4.5×10^{-5} for large M), so this simplifies to:

$$\langle N(r_m)\rangle = M/10$$

This relationship is independent of D and therefore very useful in empirical studies. Simply stated, the log-log plot deviates from linearity (slope = -D) for all box sizes that produce a box-count $\langle N(r)\rangle$ that exceeds one-tenth the sample size M (see Figures 1 and 2). This is a simple and straightforward rule for determining the lower bound of r when estimating fractal dimension from a log-log plot.

Equation [5] can also be used to determine, for a given fractal dimension D, the proportional increase in sample size M necessary to achieve a decadal (order of magnitude) increase in linearity of the log-log plot. A decadal increase in linearity implies a ten-fold decline in r_m:

$$M' = 10\,(r_m/10)^{-D}$$

Thus, the required proportional increase in sample size (from M to M') is given by the ratio:

$$M'/M = (r_m/10)^{-D}/(r_m)^{-D} = [r_m/(r_m/10)]^D = 10^D \qquad [6]$$

This power law relationship indicates that the proportional increase in M necessary to achieve a decadal increase in log-log plot linearity increases exponentially with D. Thus, when D = 1 a ten-fold increase in sample size is required, but D = 2 calls for a hundred-fold increase in sample size.

Equation [5] can also be used to determine the sample size M required to achieve log-log plot linearity over a single decade (order of magnitude) of r. A finite image, when measured at the coarsest scales, has trivially the same dimension as the space in which it is embedded (Theiler 1990). Therefore, the largest values of r (i.e., large box sizes) should not be used when estimating D from a log-log plot (Halley et al. 2004). Following Liebovitch and Toth (1989), an upper value of $r_s = 0.25$ ($\log r_s = -0.6$) is used here; this corresponds to a 4 x 4 grid of boxes. A decade (order of magnitude) of log-log plot linearity therefore requires that $r_m = 0.025$ ($\log r_m = -1.6$, or 40 x 40 box grid). For a given fractal dimension D, substituting $r_m = 0.025$ into equation [5] indicates how large M must be to ensure a single decade (order of magnitude) of log-log plot linearity, e.g.:

D = 1	$M = 10(0.025)^{-1.0} = 400$
D = 1. 5	$M = 10(0.025)^{-1.5} = 2{,}530$
D = 2.0	$M = 10(0.025)^{-2.0} = 16{,}000$

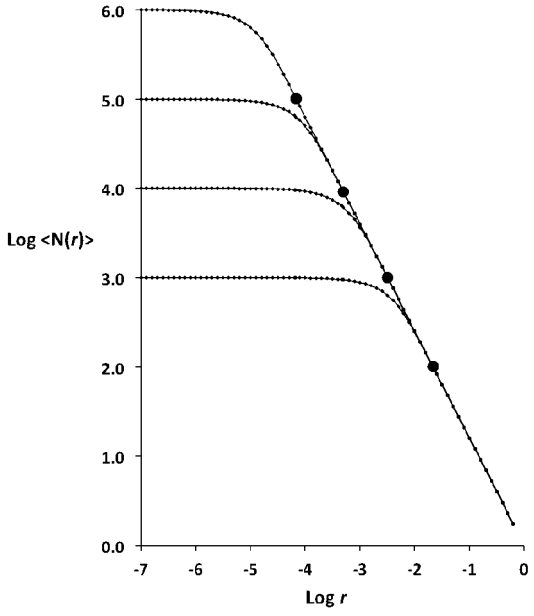

Figure 2. Theoretical log $\langle N(r)\rangle$ versus log r plots for D = 1.2 at sample sizes $M = 10^6$, 10^5, 10^4 and 10^3 (top to bottom). The filled circles indicate r_m, the point at which the log-log plot begins to deviate from slope = -D as the limit log $\langle N(r)\rangle$ = log M is approached. Logarithms are base 10.

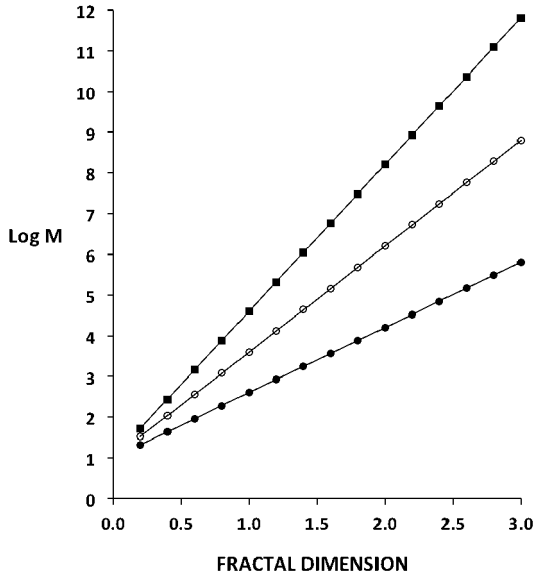

Figure 3. Sample size (M) requirements for a scaling range of one (filled circle), two (open circle) and three (filled square) decades, as a function of fractal dimension D. Note that sample size is on a logarithmic scale (base 10).

Since equation [5] is a power-law relationship, the required sample size M increases exponentially with D. The sample size required to achieve two (or more) decades of log-log plot linearity is easily obtained using equation [6], e.g.,

$$D = 1.5 \qquad M = 2{,}530(10)^{1.5} = 80{,}000$$

The required sample sizes to achieve one, two and three decades of log-log plot linearity, as a function of D, are summarized in Figure 3. It is notable that very large sample sizes are required to achieve two or more decades of linearity. For example, a space-filling fractal (D = 2) requires a sample size of 16,000 points to achieve one decade of linearity, but two decades requires 1.6 million points and three decades 160 million points. Required sample sizes for smaller values of D, while less daunting, are still considerable; for example, at D = 1.25 (the estimated fractal dimension of the west coast of Britain, Mandelbrot 1967) well over 300,000 points are needed to ensure three decades of linearity.

Discussion

Fractal dimension estimates have been determined for a great many natural patterns and processes in ecology (Seuront 2010, Bez and Bertrand 2011) and other disciplines (Malcai et al. 1997, Falconer 2003), but very few studies have attempted to develop theory-based protocols for obtaining reliable dimension estimates (Brewer and Di Girolamo 2006). The importance of sample size – in particular the necessity of large samples (high resolution images) for reliable dimension estimation – is recognized and acknowledged (Theiler 1990, Hall 1995), but specific guidelines for determining the required sample sizes are lacking (Halley et al. 2004). While empirical investigations have provided valuable insights into the importance of sample size, they cannot quantify the relationship between sample size and dimension estimate reliability (Ramsey and Yuan 1990). The present study uses a theoretical probability-based model to determine specific sample size requirements for reliable dimension estimation. The major findings are that the sample sizes required are very large – indeed, much larger than those used in most empirical studies – and that sample size determines the scaling range available for reliable dimension estimation.

Determining the scale range over which the slope (i.e., fractal dimension estimate) is calculated is a major challenge to box count dimension estimation (Ramsey and Yuan 1990, Halley et al. 2004). Ideally, dimension estimates should be obtained over the broadest scale range possible, with particular emphasis given to the finest scales (Gneiting et al. 2012). However, in empirical studies finite sample size imposes a strict limit on the lower bound of the scaling range (Theiler 1990). Established methods for choosing this lower bound include the "choice by eye" approach (Ramsey and Yuan 1990, Foroutan-pour et al. 1999, Halley et al. 2004, Agterberg 2013) and recommendations based on experience rather than theory (Liebovitch and Toth 1989, Theiler 1990). In this study, a probability model was developed to determine the theoretical lower bound r_m of the scaling range. The box count corresponding to this lower bound is $N(r_m) = M/10$, emphasizing that the lower bound is a function of sample size. This relation also provides an objective and easily implemented "rule of thumb": scaling values $r < r_m$ (i.e., r values with box counts $N(r) > M/10$) must not be used for slope determination, as they produce biased (underestimated) dimension estimates. It is notable that this theoretically derived lower bound is more conservative than the empirically derived $N(r) > M/5$ suggested by Liebovitch and Toth (1989).

The theoretical model developed here cannot be used to determine the upper bound of the scaling range; in this study, an upper bound of $r_s = 0.25$ (as suggested by Liebovitch and Toth 1989) was used to determine the sample size requirements summarized in Figure 3. It is well known that the dimension estimate approaches (or equals) the embedding dimension at large values of r (Theiler 1990, Halley et al. 2004). In practice, all values of r returning the embedding dimension (i.e., values of r at which all boxes are filled) should be excluded. For higher fractal dimensions this may result in an upper bound $r_s < 0.25$, increasing the sample size requirements reported in this study. For example, if the upper bound is lowered to $r_s = 0.125$, the sample size range necessary to ensure a single decade of linear scaling (assuming $1 \le D \le 2$) rises from 400 – 16,000 to 800 – 64,000.

This study used point data to examine the relationship between sample size and the reliability of dimension estimates; here, sample size is equal to the number of points. Examples of fractal point patterns include strange attractors (Grassberger and Procaccia 1983, Liebovitch and Toth 1989, Ramsay and Yuan 1990) and earthquake epicenters (Ogata and Katsura 1991); ecological examples include animal locations (Hagen et al. 2001) and spatial patterns of forest trees (Cheng and Agterberg 1995). Similarly, for time series and spatial transect data the sample size is simply the number of discrete measurements (e.g., Brewer and Di Girolamo 2006, Gneiting et al. 2012). However, determining sample size for other data types can be more problematic. Many empirical studies employ data consisting of lines, or line networks; examples include landscape features such as coastlines (Mandelbrot 1967) and geological fracture networks (Pruess 1995, Gonzato et al. 1998). Typically, line data are obtained by converting a map or photographic image to a digitized "line" of pixels. Under ideal conditions, and provided that the digitized "line" is one pixel wide (Gonzato et al. 2000), the sample size is equal to the total number of pixels. However, digitized "lines" often contain segments of three or more linearly adjacent pixels (Ciccotti and Mulargia 2002), indicating that fine image details have been smoothed; this may reflect limited resolution of the original image, or loss of detail during the digitization process (Hall 1995). In such cases, the effective sample size is somewhat less than the total number of pixels. A conservative estimate of effective sample size $M = P/k$ is suggested, where P is the total number of pixels and k is the mean length of linearly adjacent pixel segments on the digitized "line" (see also Pruess 1995). In cases where a "line" is created by joining discrete point observations using linear segments (e.g., branched biological networks, Panico

and Sterling (1995); animal movement pathways, Nams (2006); spatial and temporal series, Gneiting et al. (2012)), the sample size equals the number of individual points, not the number of pixels or other elements used to denote the linear segments. In fact, such data are better analyzed as point patterns, i.e., with the joining line segments (which are Euclidean, not fractal) removed. Another common data type is the "spot" image, obtained by digitally converting a high-resolution photograph to a black-white (filled-unfilled) pixelated image of "spots" (e.g., Pérez-Rodríguez et al. 2013). The fractal properties of the spots (black regions) themselves, or the spot boundaries, can be examined. However, only the boundary pixels (i.e., the black to white edges) are relevant to fractal analysis (Foroutan-pour et al. 1999, Halley et al. 2004), so the sample size is equal to the total number of boundary (edge) pixels. However, as with line data a more conservative effective sample size may have to be calculated if the edges contain segments of three or more linearly adjacent pixels.

It is apparent that sample size and image resolution are strongly linked (Huang et al. 1994). Images processed for fractal analysis often have limited resolution (e.g., digital photographs) or contain simplified, smoothed edges (e.g., mapped features such as geological faults, topographic lines, lakeshores, forest edges, etc.). Limited resolution implies that an image is Euclidean (non-fractal) at the finest scales (Hall 1995), which are paradoxically the scales of greatest interest in fractal analysis (Bez and Bertrand 2011, Gneiting et al. 2012). When detailed features are absent from an original image, attempts to increase the sample size by increasing image processing resolution are fruitless; the effective sample size is constrained by image resolution (Huang et al. 1994, Pruess 1995). The only way to increase the effective sample size is to begin with a more detailed (higher resolution) image.

Malcai et al. (1997) suggest that a minimal scaling range of two to three decades is necessary to demonstrate fractal self-similarity (see also Gonzato et al. 1998, Ciccotti and Mulargia 2002, Falconer 2003, Halley et al. 2004). However, the theoretical model developed here indicates that a 2-3 decade scaling range may be difficult if not impossible to achieve in practice, since the required sample sizes are enormous. For example, a fractal embedded in two dimensions requires a sample size of 4,000 to 1.6 million (assuming $1 \leq D \leq 2$) to ensure two decades of log-log plot linearity; for three decades, the range is 40,000 to 160 million. For fractal surfaces (i.e., embedded in three dimensions, and assuming $2 \leq D \leq 3$) the sample sizes are astronomical: 16,000 to 640,000 for just a single decade of scaling, rising to 1.6 – 640 million for two decades. Given these numbers, the most researchers can reasonably hope for is one or two decades of linear scaling, but even then sample sizes much larger than those currently used in most empirical studies are required (cf. Kallimanis et al. 2002).

It has been suggested that natural objects have a limited range of fractal self-similarity (e.g., Panico and Sterling 1995), and that this is reflected in the limited scaling ranges observed in empirical studies (Avnir et al. 1998, Halley et al. 2004). However, the results presented here suggest that a limited scaling range is more likely a sampling artifact related to the difficulty or impossibility of obtaining a sufficiently large sample size to achieve more than one or two decades of linear scaling. This study has shown that an additional decade of scaling range requires a 10^D increase in sampling effort, i.e., a 10 to 1000-fold increase in sample size (assuming $1 \leq D \leq 3$). This exponential increase in sampling effort makes it extremely difficult, if not impossible, to achieve more than two or three decades of linear scaling (cf. Malcai et al. 1997). Consider, for example, the cerebral cortex of the human brain, which has an estimated surface fractal dimension D = 2.80 (Kiselev et al. 2003). Using the theory-based equations developed here, a sample size (i.e., number of individual point measures on the brain surface) of over 193 million would be required to achieve just two decades of linear scaling.

Conclusions and recommendations

Box counting remains the most widely used computational algorithm for estimating the fractal dimension of natural patterns and processes (Falconer 2003, 2013). A number of other algorithms for estimating D have been developed (see Seuront 2010), but all depend on the availability of large amounts of data at sufficient spatial or temporal resolution (Ramsey and Yuan 1990, Bez and Bertrand 2011, Gneiting et al. 2012). The following recommendations are specific to the box counting approach, but the most important recommendation – that researchers should strive to obtain the largest sample size possible – is broadly relevant and applicable to all estimation algorithms.

- The formulae developed in this paper should be used to determine minimal sample size requirements for reliable dimension estimation in empirical investigations. As an absolute minimum, a sample size sufficient to achieve a scaling range of at least one decade is strongly recommended. For images in a two-dimensional embedding space, a minimal sample size of 2,500 or greater is required for a scaling range of one decade when $D < 1.5$; for higher-dimension fractals ($1.5 < D \leq 2$), a sample size of 5,000 – 15,000 or greater is required. It must be emphasized that sample sizes less than these minimal recommended values will result in an unreliable, highly biased dimension estimates irrespective of the computational algorithm used.

- The lower bound of the scaling range for reliable estimation of the box count dimension is r_m, where the box count $N(r_m) = M/10$. Therefore, scaling values $r < r_m$ (i.e., $N(r) > M/10$) should never be used in calculating the log-log slope (fractal dimension estimate). At the upper end of the scaling range, as a minimum all r for which $N(r) = (1/r)^2$ (i.e., all boxes filled, assuming a two-dimensional embedding dimension) should be excluded.

- Whenever possible, point or coordinate data (i.e., topological dimension of zero) should be used to estimate fractal dimension. Most data sets in ecology and other disciplines can be represented as point coordinates, rather than as lines or pixels. For example, landscape features (e.g., coastlines, forest edges) are typically represented as "lines" (topological dimension of one), obtained by tracing and digitizing a map or aerial photograph. Image processing will necessarily smooth the image, resulting in a "line" representation that has both fractal and Euclidean (non-fractal) features (Hall 1995). An alternative approach involves using a stylus to record point coordinates at regular intervals along the entire mapped image. This point-coordinate approach, while much more tedious, avoids smoothing the image (provided that the interval between adjacent point coordinates matches the resolution of the map or photograph). There are two additional benefits: the sample size M is clearly defined, and the box-count has a fixed upper limit of $N(r) = M$. By contrast, contour lines (and line networks) have no fixed upper box-count limit since they have a topological dimension of one at finer scales. Thus, the log-log plot of a linear feature scales as D at intermediate scales, but as one-dimensional at finer scales. The transition from D to one dimension may be quite subtle and difficult to detect, particularly when D is small (see examples in Gonzato et al. 1998, Gneiting et al. 2012). Conversely, log-log plots based on point data transition from D to zero dimension, making deviation from log-log linearity much easier to detect (see Figures 1 and 2). Conversion to point-coordinate data can also be applied to "spot" images, by using a stylus to record point coordinates at regular intervals along all edges of the image. Perhaps the greatest advantage of using point coordinate data, however, is that problems of "apparent fractality" (Hamburger et al. 1996, Buczkowski et al. 1998, Halley et al. 2004) and "physical cutoffs" (Ciccotti and Mulargia 2002) associated with "line" and "spot" data are entirely avoided.

- The sample size used to obtain a dimension estimate, and the scaling range (upper and lower bounds) used to determine the fractal dimension, should always be reported. Also, as many box sizes as possible should be utilized. Many studies only provide box-counts at $r = 1/2^k$ where k is a non-negative integer (Liebovitch and Toth 1989, Gneiting et al. 2012), but additional values of r will provide greater statistical power (Gonzato et al. 1998). As a general rule, at least ten box count values (i.e., regression analysis data points) should be obtained over the scaling range r_m to r_s.

- Exercise caution when using regression analysis for dimension estimation. Significance tests based on R^2 are strictly invalid since box-count values are not independent (Halley et al. 2004), although R^2 remains a useful descriptor. The log-log plot should always be examined for evidence of curvature (a concave downward trend) over the scaling range. The most common method for examining deviation from linearity is the analysis of residuals, but tests on residuals lack statistical power unless the number of regression data points is large (Gonzato et al. 1998). A more powerful exploratory approach is to include a quadratic term in the regression equation of the log-log plot. A statistically significant quadratic (non-linear) term indicates curvature. If curvature is detected, one or two data points (from one or both ends of log-log plot) are removed and the regression equation is fitted again; this process is repeated until the quadratic term is no longer significant. Note that it is unlikely that a log-log plot will be concave downward if the sample size is sufficiently large and r_m is used as the lower bound of the scaling range.

Perusal of the contemporary literature indicates that applications of fractal geometry in ecology and other disciplines are severely compromised by the lack of a sufficiently rigorous approach to obtaining reliable dimension estimates. Specifically, sample sizes are often wholly inadequate, and the methodologies used to obtain dimension estimates lack a theoretical framework. This study provides statistically rigorous guidelines for determining required sample sizes in empirical studies, as well as a rigorous theory-based methodological approach to obtaining unbiased box count dimension estimates from log-log plots.

Acknowledgements. This paper is dedicated to Professor L. Orlóci, who first piqued my interest in fractal geometry. I thank Dr. D. Walker for inspiring conversations on some of the topics touched upon here.

References

Agterberg, F.P. 2013. Fractals and spatial statistics of point patterns. *J. Earth Sci.* 24: 1-11.

Avnir, D., O. Biham, D. Lidar and O. Malcai. 1998. Is the geometry of nature fractal? *Science* 279: 39-40.

Bernston, G.M. and P. Stoll. 1997. Correcting for finite spatial scales of self-similarity when calculating the fractal dimensions of real-world structures. *Proc. Royal Soc. London B* 264: 1531-1537.

Bez, N. and S. Bertrand. 2011. The duality of fractals: roughness and self-similarity. *Theor. Ecol.* 4: 371-383.

Brewer, J. and L. Di Girolamo. 2006. Limitations of fractal dimension estimation algorithms with implications for cloud studies. *Atmos. Res.* 82: 433-454.

Buczkowski, S., P. Hildgen and L. Cartilier. 1998. Measurements of fractal dimension by box-counting: a critical analysis of data scatter. *Physica A* 252: 23-34.

Cheng, Q. and F.P. Agterberg. 1995. Multifractal modeling and spatial point processes. *Math. Geol.* 27: 831-845.

Ciccotti, M. and F. Mulargia. 2002. Pernicious effect of physical cutoffs in fractal analysis. *Phys. Rev. E* 65: 037201.

Falconer, K.J. 2003. *Fractal Geometry: Mathematical Foundations and Applications*. Wiley, Chichester.

Falconer, K.J. 2013. *Fractals: A Very Short Introduction*. Oxford Univ. Press, Oxford.

Foroutan-pour, K., P. Dutilleul and D.L. Smith. 1999. Advances in the implementation of the box-counting method of fractal dimension estimation. *Appl. Math. Comput.* 105: 195-210.

Gneiting, T., H. Sevcikova and D.B. Percival. 2012. Estimators of fractal dimension: assessing the roughness of time series and spatial data. *Statist. Sci.* 27: 247-277.

Gonzato, G., F. Mulargia and W. Marzocchi. 1998. Practical application of fractal analysis: problems and solutions. *Geophys. J. Int*. 132: 275-282.

Gonzato, G., F. Mulargia and M. Ciccotti. 2000. Measuring the fractal dimensions of ideal and actual objects: implications for application in geology and geophysics. *Geophys. J. Int.* 142: 108-116.

Grassberger, P. and I. Procaccia. 1983. Characterization of strange attractors. *Phys. Rev. Letters* 50: 346-349.

Hagen, C.A., N.C. Kenkel, D.J. Walker, R.K. Baydack and C.E. Braun. 2001. Fractal-based spatial analysis of radio telemetry data. In: Millspaugh, J.J., J.M. Marzluff and R. Kneward (eds.), *Radio Tracking and Animal Populations.* Academic Press, San Diego. pp. 167-187.

Hall, P. 1995. On the effect of measuring a self-similar process. *SIAM J. Appl. Math.* 55: 800-808.

Halley, J.M., S. Harley, A.S. Kallimanis, W.E. Kunin, J.J. Lennon and P. Sgardelis. 2004. Uses and abuses of fractal methodology in ecology. *Ecol. Letters* 7: 254-271.

Hamburger, D., O. Biham and D. Avnir. 1996. Apparent fractality emerging from models of random distributions. *Phys. Rev. E* 53: 3342-3358.

Huang, Q., J.R. Lorch and R.C. Dubes. 1994. Can the fractal dimension of images be measured? *Pattern Recog.* 27: 339-349.

Kallimanis, A.S., S.P. Sgardelis and J.M. Halley. 2002. Accuracy of fractal dimension estimates for small samples of ecological distributions. *Land. Ecol.* 17: 281-297.

Kenkel, N.C. and D.J. Walker. 1996. Fractals in the biological sciences. *Coenoses* 11: 77-100.

Kiselev, V.G., K.R. Hahn and D.P. Auer. 2003. Is the brain cortex a fractal? *NeuroImage* 20: 1765-1774.

Liebovitch, L.S. and T. Toth. 1989. A fast algorithm to determine fractal dimensions by box counting. *Phys. Letters A* 141: 386-390.

Lopes, R. and N. Betrouni. 2009. Fractal and multifractal analysis: a review. *Med. Image Anal*. 13: 634-649.

Malcai, O., D.A. Lidar and O. Biham. 1997. Scaling range and cutoffs in empirical fractals. *Phys. Rev. E* 56: 2817-2828.

Mandelbrot, B. 1967. How long is the coast of Britain? Statistical self-similarity and fractional dimension. *Science* 156: 636-638.

Nams, V. 2006. Improving accuracy and precision in estimating fractal dimensions of animal movement paths. *Acta Biotheor.* 54: 1-11.

Ogata, Y. and K. Katsura. 1991. Maximum likelihood estimates of the fractal dimension for random spatial patterns. *Biometrika* 78: 463-474.

Panico, J. and P. Sterling. 1995. Retinal neurons and vessels are not fractal but space-filling. *J. Compar. Neurol.* 361: 479-490.

Pérez-Rodríguez, L., R. Jovani and F. Mougeot. 2013. Fractal geometry of a complex plumage trait reveals bird's quality. *Proc. Royal Soc. B* 280: 20122783.

Pruess, S.A. 1995. Some remarks on the numerical estimation of fractal dimension. In: Barton, C.C. and P.R. La Pointe (eds.), *Fractals in the Earth Sciences*. Plenum, New York. pp. 65-75.

Ramsey, J.B. and H.-J. Yuan. 1990. The statistical properties of dimension calculations using small data sets. *Nonlinearity* 3: 155-176.

Schroeder, M.R. 1991. *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise.* Freeman, New York.

Seuront, L. 2010. *Fractals and Multifractals in Ecology and Aquatic Science*. CRC Press, Boca Raton.

Sun, W., G. Xu, P. Gong and S. Liang. 2006. Fractal analysis of remotely sensed images: a review of methods and applications. *Int. J. Remote Sens.* 27: 4963-4990.

Taylor, C.C. and S.J. Taylor. 1991. Estimating the dimension of a fractal. *J. Royal Stat. Soc. B* 53: 353-364.

Theiler, J. 1990. Estimating fractal dimension. *J. Opt. Soc. Amer. A* 7: 1055-1073.

Received April 26, 2013
Revised July 8, 2013
Accepted July 11, 2013