Journal of Statistical Physics, Vol. 81, Nos. 5/6, 1995

Testing for Scaling in Natural Forms and Observables

A. A. Tsonis[1] and J. B. Elsner[2]

Received March 7, 1995

The general procedure of calculating fractal dimensions or other exponents is based on estimating some quantity as a function of scale and on assessing whether or not this function is a power law. This power law manifests itself in a log (quantity) versus log (scale) plot as a linear region (scaling). It has thus become the practice to estimate dimensions by the slope of some linear region in those log–log plots. When we are dealing with exact fractals (the Koch curve, for example) there are no problems. When, however, we are working with natural forms or observables, problems begin to emerge. In such cases the scaling region is subjectively estimated and often is only the result of the generic property of the quantity to increase monotonically or decrease monotonically as the scale goes to zero irrespective of the geometry of the object. Here we discuss these issues and suggest a procedure to deal with them.

KEY WORDS: Self-similarity; self-affinity; fractals; scaling; chaos.

1. INTRODUCTION

Fractal objects,[1] unlike Euclidean objects, possess no characteristic sizes or length scales. They have details on all length scales and as such each small portion when magnified reproduces a large portion. This property is called self-similarity or scaling (scale invariance) and is closely connected to the intuitive notion of dimension. Mathematically, scaling is expressed by a power law of the form $C(\varepsilon) \propto \varepsilon^{\pm A}$ (the sign depends on the statistic), where ε represents the scale, $C(\varepsilon)$ is a statistic obtained at scale ε, and A is related to the fractal dimension, D, which assumes noninteger values. Fractals can be exact or random. Exact fractals are produced by mathematical

[1] Department of Geosciences, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin 53201-413.

[2] Department of Meteorology, Florida State University, Tallahassee, Florida 32306-3034.

822/81/5-6-1

equations (for example, chaotic attractors of dynamical systems) or recursive algorithms (for example, the Koch curve, the Sierpinski carpet, etc.). Exact fractals possess exact self-similarity (i. e., when a small portion is magnified it reproduces exactly a larger portion). Random fractals are products of recursive algorithms plus noise and do not possess exact self-similarity. When a small part is magnified it may not reproduce exactly a larger part, but it reproduces the statistical properties of a larger part. In this case we have that $\langle (C) \rangle \propto \varepsilon^{\pm A}$, where the brackets indicate averages. In both cases scaling extends to infinitely smaller scales.

In cases where the scaling is not uniform (i.e., when shapes are statistically invariant under transformations that scale different coordinates by different amounts), then we do not have self-similarity by self-affinity. This type of scaling often appears in time series. Mathematically this is expressed by $\Delta x(\lambda\, \Delta t) =^{\mathrm{d}} \lambda^H \Delta x(\Delta t)$ for all $\lambda > 0$, where $x(t)$ is the time series and the symbol $=^{\mathrm{d}}$ denotes identity in statistical distributions. This relation dictates that the distribution of increments of x over some time scale $\lambda\, \Delta t$ is identical to the distribution of increments of x over a lag equal to Δt multiplied by λ^H. Therefore, if time is magnified by a factor λ, the x is magnified by a different factor λ^H $(0 < H < 1)$. The quantity H characterizes self-affinity in a similar way to that in which D characterizes self-similarity. The values of $H = 0.5$ corresponds to the trace of a Brownian motion, whereas any value of $H \neq 0.5$ defines a fractional Brownian motion (fBm) that displays infinite long-run correlations (either positive for $H > 0.5$ or negative for $H < 0.5$). Because of the above formulations the general procedure of calculating the fractal dimension or other exponents is based on estimating some quantity C as a function of scale ε and on assessing whether or not this function is a power law. This power law manifests itself in a $\log C(\varepsilon)$ vs. $\log \varepsilon$ plot as a linear region (scaling).

When we are dealing with exact fractals (the Koch curve, for example) or with computer-generated random fractals (Brownian motions, coastlines, etc.), there are no problems. The log–log plots are very linear (on the average in the case of random fractals) and we always recover an expected and *a priori* known result. When, however, we are working with objects or observables from nature whose properties are not *a priori* known, problems begin to emerge. In such cases scaling is assumed and the scaling region is subjectively estimated and often is only the result of the generic property of the quantity to increase monotonically or decrease monotonically as the scale goes to zero irrespective of the geometry of the object.

For example, if $C(\varepsilon)$ is the number of boxes of size ε needed to cover the object, then $C(\varepsilon)$ increases as ε decreases regardless of the geometry of the object. This problem is further accentuated by the fact that strictly speaking exponents related to scaling are meaningful only for $\varepsilon \to 0$, which

in applications is either overlooked or limited by data quality and sample size. For that reason the statistical significance of scaling itself has to be tested before it can be attributed to a self-similar or self-affine fractal set. This issue has never been considered in studies establishing the fractal or chaotic character of natural objects and observables. As we will see next, this is not a trivial problem; certain admissions have to be made, and more work is needed in order to design a general procedure to test for scaling.

2. TESTING FOR SELF-SIMILARITY

Let us consider the case of the coast of Great Britain. It has been suggested[2,3] that coastlines are random fractals with their length L as measured with a yardstick of size r, $L(r)$, scaling according to $L(r) \propto r^{1-D}$, where D for the coast of Great Britain is ~ 1.25. We considered the coastline with a resolution of 1 km and we applied box counting. Accordingly, Fig. 1 shows the logarithm of number of squares of size r, $N(r)$, that include a piece of the coastline as a function of the logarithm of r. The least-squares fit over $0 < \log r < 2.5$ gives a slope (an estimate of D) of about -1.24, in accordance with the previously claimed value. At this point we can claim that the coastline of Great Britain is fractal with a dimension of 1.24. However, all we did was to assume that $\log N(r)$ is a linear function of $\log r$ over the above range of scales and to subsequently estimate the least squares slope. Thus we assumed fractality or self-similarity before we are able to prove it. What we should have first asked is whether or not a linear model provides the best fit for the data in Fig. 1.

Figure 2 shows $d \log N(r)/d \log r$ as a function of $\log r$. We choose to show the first derivative because it is equal to the dimension D,

$$N(r) = Ar^{-D} \Rightarrow \log N(r) = \log A - D \log r \Rightarrow d \log N(r)/d \log r = -D$$

In such a figure scaling manifests itself as a plateau (zero slope). We observe the following: (1) At very small scales the data points tend to minus one. (2) At very large scales the data tend to a value of minus two. Those two features are artifacts of the algorithms, used to estimate dimensions and can be explained fully. Due to limits in the resolution as r approaches the resolution of the data, the coastline becomes more and more linear and thus $D \to 1$. For large r most if not all of the squares used in the boxcounting have a high chance to include a piece of the coastline and thus $D \to 2$. Due to those artifacts if the structure is scaling, a plateau will be observed in between very small and very large scales. In our case we do not observe (visually) a clear plateau, as the data show a slight trend over the whole range of scales. It is thus imperative to use statistical tests before a scaling

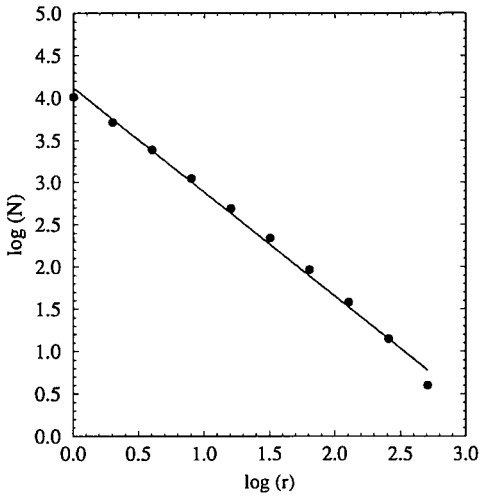

Fig. 1. Log–log plot of the number of squares of sizes r, $N(r)$, that contain a piece of the coast of Great Britain as a function of the size r. The straight line is a least squares fit with slope -1.24 over the indicated range of scales.

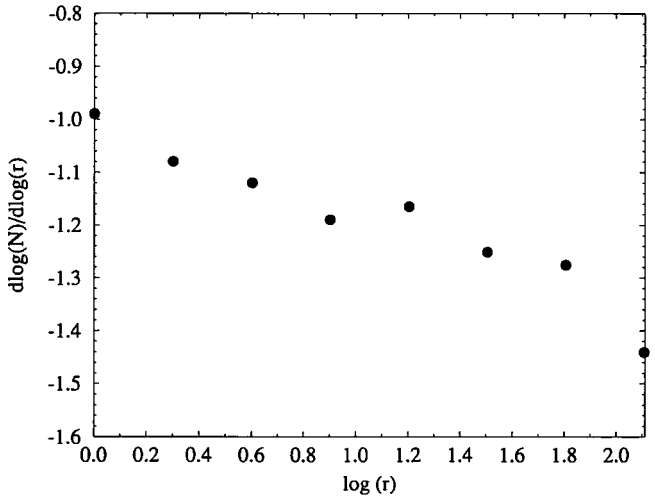

Fig. 2. The first derivative of the function in Fig. 1 $[d \log N(r)/d \log r]$ versus $\log r$. Since the derivative provides to the fractal dimension, plateaus in such figures indicate scaling (see text for details).

region (a region over which the data in Fig. 2 have a slope which is not statistically significantly different than zero) is accepted. Note that the above-described artifacts are also present in figures such as Fig. 1, but due to the compression properties of the logarithmic function they are "suppressed." Therefore, one must question the validity of fitting a straight line over the whole range of scales in such figures. We thus are faced with the question as to what is the appropriate procedure to establish scaling. Here is where the problem becomes nontrivial. The scaling region is a function of the sample size. The larger the size, the wider the region. For a given set and sample size the width and location of the scaling region (if there is one) are not known a priori. We thus cannot consider a certain range of scales between very small and very large scales and test whether the points in that range exhibit a slope that is not statistically significantly different than zero. One could assume a sliding "window" of a varying width $\Delta \log(r)$ and test whether or not over the included range of scales in that window the slope is not significantly different than zero. This process requires enough points to ensure accurate statistics. This is not a problem, as the box-counting operation can be repeated several times, each time starting with a different size square. The problems, however, with this approach are that often (1) zero slopes result from highly nonlinear functions and (2) nonzero slopes may result from points belonging to a scaling random fractal structure (see below for examples).

Next we propose a way to test for "alleged" scaling that is devoid of such problems. The philosophy behind it is as follows: In order to decide that the data points in Fig. 2 are a manifestation of scaling and not the result of a random nonfractal structure, we should test the hypothesis that the data in Fig. 2 come from a population of random fractal boundaries with an average $D = 1.24$ over the same range of scales. In order to do this, it is necessary to have a model that generates such fractal boundaries. In the case of coastlines several algorithms can be employed to produce boundaries with a desired dimension. We employed the successive random midpoint displacement technique[4] in order to generate 10,000 random fractal boundaries with the same number of points as the Great Britain data, which on the average have a dimension equal to 1.24. The results are summarized in Fig. 3. The dots correspond to the data from the coastline of Great Britain (from Fig. 2). For each simulation and for a given r we have a D. Thus, from the 10,000 simulations we can estimate average D, standard deviation, and frequency distribution of D. The solid line is the average D as a function of $\log r$. The bounds indicate the 5%–95% interval of the frequency distribution of D. As in Fig. 2, we observe the following artifacts: (1) The bounds are narrower for small scales, and (2) the solid line is rather flat, showing a plateau at -1.24 for the range of scales

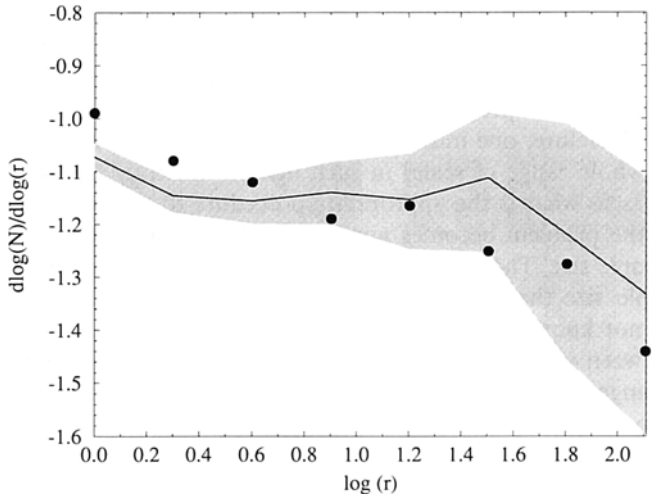

Fig. 3. Same as Fig. 2, but showing the results from 10,000 simulations of random fractal coastlines having an average dimension of 1.24. The solid line indicates the average dimension as a function of log r and the shaded area shows the bounds of the 5% – 95% limits of the observed frequency distribution of D as a function of log r. The dots are as in Fig. 2 (see Fig. 2 for details).

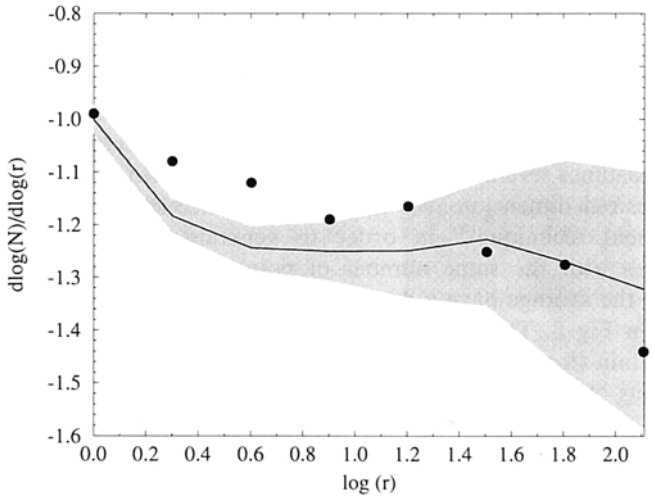

Fig. 4. Same as Fig. 3, but for 10,000 simulations of random fractal coastlines having an average dimension of 1.14.

$0.3 < \log r < 1.5$ (as would be expected from a random fractal with $D = 1.24$); it tends to smaller values for $\log r < 0.3$ and to greater values for $\log r > 1.5$. As before, those features are expected and can be fully explained. Due to limits in resolution, as r approaches the resolution of the data, the structure becomes more and more linear anf thus $D \to 1$. For large r most if not all of the squares used in the box-counting algorithm have a high chance to include a piece of the coastline and thus $D \to 2$. This also results in unreliable statistics and consequently the bounds are wider for larger scales than for smaller scales where the statistics are much more reliable.

From Fig. 3, for a resolution of about 1 km we have thus established the features of scaling of random fractal boundaries with an average dimension of 1.24. These bounds can now be used as the population against which the Great Britain data (black dots) can be compared. Everything outside those limits will indicate departure from the model with $D = 1.24$. Apparently, only the points corresponding to larger scales ($\log r > 1.0$) can be considered as significant at a 95% confidence level. Therefore a $D = 1.24$ is not appropriate. A $D = 1.14$ (Fig. 4) appears to be the most appropriate model, but again it does not explain the small scales, which still deviate from the model. It may be that scaling breaks at very small scales, indicating different mechanisms at larger and smaller scales, which by itself could be a very significant result. Note that (for $D = 1.14$), even though over intermediate scales the black dots exhibit no obvious plateau they nevertheless fall within the 5%–95% limits of the distribution and therefore are consistent with a scaling model of $D = 1.14$

Alternatively, it may be that Fig. 1 exhibits two scaling regimes, one for smaller scales and one for larger scales. In this case, however, we were not able to produce a bidimensional model that will include all the black dots. If none of the above procedures is conclusive, it may very well be that the $\log N(r)$ is a nonlinear function of $\log r$. This could be an indication either that there is no scaling or we are faced with a new type of scaling far more complicated than simple scaling. For simplicity we will call this "nonlinear" scaling.

3. TESTING FOR SELF-AFFINITY

Next we proceed with testing for self-affinity in sequences. Our example involves DNA sequences. DNA sequences are strings of the bases (nucleotides) A, T, C, G. Bases C and T are pyrimidines, and bases A and G are purines. For each sequence a DNA walk may be defined as follows: Starting with the first base, the walker steps to the right if the base is a pyrimidine and to the left if it is a purine. It has been suggested and

debated[5–7] that such walks correspond to traces that display scaling properties appropriate to fBms for intron-containing sequences (noncoding sequences) or to pure random walks for intronless sequences (coding sequences).

In one-dimensional random walks the displacement after l steps $y(l)$ is given by

$$y(l) = \sum_{i=1}^{l} u(i)$$

where $u(i) = +1$ or -1 if the ith step is to the right or to the left, respectively. A walk may be an uncorrelated walk, where the direction of each step is independent of the previous steps, or it may be a correlated walk, where the direction of each step depends on the past motion. In any case a suitable statistical quantity that characterizes a walk is the root mean square fluctuation $F(l)$ about the average displacement,[8]

$$F^2(l) = \overline{[\Delta y(l)]^2} - \overline{[\Delta y(l)]}^2$$

where $y(l) = y(l_0 + 1) - y(l_0)$ and the bars indicate an average over all positions l_0 in the walk.

As mentioned earlier, a scaling process $y(l)$ satisfies the relationship $y(l) =^{\mathrm{d}} \sigma^{-1} y(\lambda l)$, where $=^{\mathrm{d}}$ indicates equality in distribution and σ, $\lambda > 0$. Consequently, any moment of order k, μ'_k, satisfies the relation $\mu'_k(l) = \sigma^{-k} \mu'_k(\lambda l)$. The general solution to the last equation is $\mu'_k = l^{kb} \chi(\log l / \log \lambda)$ with $b = \log \sigma / \log \lambda$ and where χ is a periodic function of period one superimposed on the power law.[9] Note that since the usual power law $\mu'_k(\varepsilon) = A l^{kb}$ is a particular solution, depending on the scaling process, these oscillations may or may not present or significant. In any case if we consider the definition of $F(l)$, we expect that if $y(l)$ is scaling, then

$$F(l) \propto l^H \tag{1}$$

where $H = 2b$. The parameter H is called the scaling exponent. A value of $H = 0.5$ corresponds to a purely random walk and a value of $H \neq 0.5$ corresponds to walks that display infinitely long-run correlations (positive if $H > 0.5$ and negative if $H < 0.5$). It thus follows that scaling exists if there exists a parameter H that satisfies Eq. (1). Note that all self-affine processes of a given H exhibit spectra of the form $1/f^{2H+1}$, where f is the frequency.[10, 11]

Figure 5 shows $\log F(l)$ vs. $\log l$ for the noncoding sequence of β-cardiac myosin heavy-chain gene. This sequence is 28,438 bases long. As expected by the defenition of $F(l)$, $\log F(l)$ is a monotonically increasing function of $\log l$. A linear model here appears to be very convincing and a

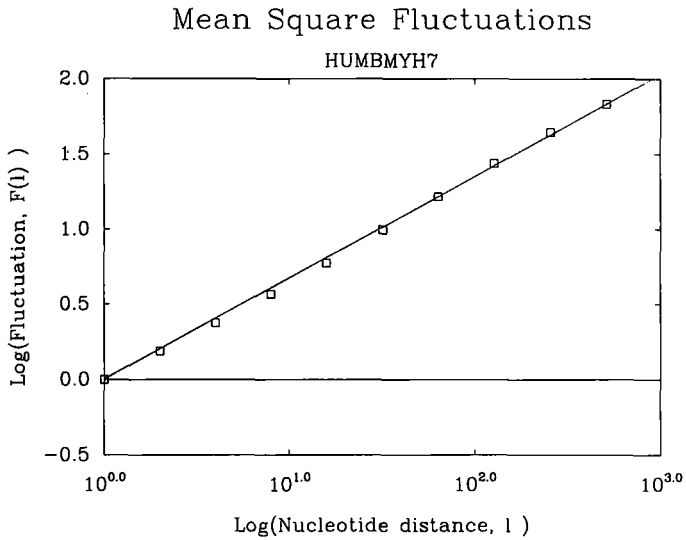

Fig. 5. Log–log plot of the root mean square fluctuation $F(l)$ of a walk generated by the noncoding sequence of β-cardiac myosin heavy-chain gene as a function of the nucleotide distance l. The straight line of slope 0.67 is a least squares fit over the indicated range of scales.

least squares fit in the range $1 < \log l < 3$ results in a line with a slope ≈ 0.67. Thus it was claimed that in this case $F(l) \propto l^{0.67}$, indicating that the sequence displays infinitely long-run positive correlations. Again, here the procedure assumes that the best fit is a straight line and thus scaling was assumed before it was proven. The question remains: Are the data in Fig. 5 consistent with population of fBms having $H = 0.67$?

We generated 1000 fBms with $H = 0.67$ and length 28,438. These sequences were generated by inverting spectra of the form $f^{-(2H+1)}$. Even though other approaches to generate fBms exist, this approach is widely used and is considered the purest interpretation of fractional Brownian motion.[12] From each one of theses sequences we obtained a $d \log F(l)/d \log l$ vs $\log l$ graph. The solid line in Fig. 6 shows the average $d \log F(l)/d \log l$ (i.e., H) vs. $\log l$ plot, which, as expected, displays a plateau at $H \approx 0.67$. The bounds show again the 5%–95% interval of the frequency distribution of H. If a reported scaling with $H = 0.67$ is to be significant at the 95% significance level, the plot corresponding to the gene, $d \log F(l)/d \log l$ vs. $\log l$ (dots), should fall within the bounds. Note that over the whole range of scales a least squares fit would result in a slope close to zero (despite the fact of an overall nonlinearity) and in a value of $H = 0.67$. But the data show no plateau at $H = 0.67$, as almost all

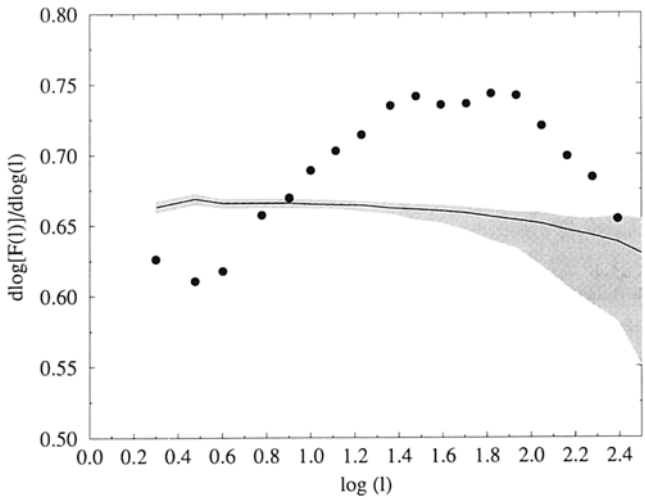

Fig. 6. For any random walk one expects the root man square fluctuation $F(l)$ about the average of the displacement to scale with l according to a power law $F(l) \propto l^H$. The exponent H is the slope of a log $F(l)$ vs. log l plot. If a scaling region clearly exists in such a plot, then we should be able in a $d \log F(l)/d \log l$ vs. log l plot to observe a plateau over a significantly wide range of scales. The dots show $d \log F(l)/d \log l$ vs. log l for the entire noncoding sequence of β-cardiac myosin heavy-chain gene. This figure was produced as follows. First the function $F(l)$ was obtained for $l = 1,..., 250$. Then the slope $= [\log F(l) - \log F(l-1)][\log l - \log(l-1)]^{-1}$ was calculated for only those l's that are powers of 1.3 [$l =$ integer $\cdot (1.3^n)$, $n = 1, 22$]. This arrangement gives a good representation of the function over the range of scales involved. The average of all the values corresponding to dots or a least squares fit of log $F(l)$ on log l over the interval $2^0 < l < 2^9$ yields a slope of around 0.67, which is the value that was initially claimed.[5] The solid line shows the average $d \log F(l)/d \log l$ as a function of log l based on a sample of 1000 fBms with $H = 0.67$. The shaded bounds indicate the 5%–95% percentiles of the frequency distribution. From this figure it is concluded that the reported scaling with $H = 0.67$ (indicating long-range correlations) is not significant at the 95% confidence level (see text for details).

points are outside the 5%–95% interval of the control model with $H = 0.67$. The figure suggests that a small scaling region exists at about $H = 0.73$. This might be important, but the conclusion here is that no statistically significant scaling or long-range correlations exist in the β-cardiac myosin heavy-chain gene.

Inverting $f^{-(2H+1)}$ spectra is by now routine and thus when it comes to testing for "alleged" self-affinity the procedure proposed above is sufficient. When, however, we wish to test for an "alleged" self-similarity, the task may not be as easy. In this case a model might not be available. We may not even know what the data look like (for example, when we try to estimate dimensions of attractors from observables and we search for

scaling regions in data embedded in some high dimension). In this case the only procedure would be the "sliding-window" approach in Section 2, which however, has drawbacks. The difficulty with testing self-similarity could be overcome if we could assume that different families of random fractals having on the average the same dimension have the same scaling limits. For example, any fractal with a dimension 1.24 has the limits shown by the bounds in Fig. 3 provided that the resolution is the same and that their scales have been normalized between zero and one. Computer simulations, however, using various sets of the same dimension do not support this. Each family of fractals seems to have its own limits.

4. CONCLUSIONS

We have presented an investigation into an issue that has been overlooked in studies establishing fractals and chaos in natural forms and observables. Even though problems still exist, we have suggested some ways to deal with the issue. Even though testing for scaling may be in many cases a nontrivial problem, it is evident that we cannot keep on avoiding testing for scaling, as proper testing can potentially reveal important properties of the system in question, such as limited scaling, multiple scaling, or even "nonlinear" scaling. As such, it may enrich our understanding of the character and processes involved in the system. If we assume that a scaling regime represents a rule that dictates the properties of an object over the corresponding scales, then a nonlinear $\log N(r)$ function will indicate that there exist many rules for many scales and thus we are dealing with a far more complicated problem than a simple scaling will indicate. Similarly, multiple scaling (say, two distinct scaling regions) will suggest that two major processes are involved each one at a different range of scales. Whatever the case, testing for scaling is necessary, for it can solidly establish the existence or the absence of an alleged scaling with a high degree of confidence, which in turn might provide useful insights into how the different scales actually are related in a given physical problem.

REFERENCES

1. B. B. Mandelbrot, *The Fractal Geometry of Nature* (Freeman, New York, 1983).
2. B. B. Mandelbrot, *Science* **156**:636 (1987).
3. J. Feder, *Fractals* (Plenum Press, New York, 1988).
4. H.-O. Peitgen, H. Jürgens, and D. Saupe, *Chaos and Fractals: New Frontiers of Science* (Springer-Verlag, New York, 1992).
5. C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, *Nature* **356**:168 (1992).

6. A. A. Tsonis, J. B. Elsner, and P. A. Tsonis, *Biochem. Biophys. Res. Commun.* **197**:1288 (1993).
7. S. Karlin and V. Brendel, *Science* **259**:77–680 (1993).
8. E. Montroll and M. F. Shlesinger, In *Nonequilibrium Phenomena II, From Stochastics to Hydrodynamics*, J. L. Lebowitz and E. W. Montroll, eds. (North-Holland, Amsterdam, 1984), pp. 1–121.
9. A. A. Tsonis, G. N. Triantafyllou, and R. Picard, *Appl. Math. Lett.* 7:19 (1994).
10. A. R. Osborne and A. Provenzale, *Physica D* **35**:357 (1989).
11. A. A. Tsonis, *Chaos: Theory to Applications* (Plenum Press, New York, 1992).
12. D. Saupe, In *The Science of Fractal Images*, H.-O. Peitgen and D. Saupe, eds. (Springer-Verlag, New York, 1988).