Box-counting clustering analysis: Corrections for finite sample effects

Stefano Borgani[1,2] and Giuseppe Murante[3]
[1]*Istituto Nazionale di Fisica Nucleare, Sezione di Perugia, c/o Dipartamento de Fisica dell'Università, Via A. Pascoli, I-06100 Perugia, Italy*
[2]*International School for Advanced Studies, Via Beirut 4, I-34014 Trieste, Italy*
[3]*Istituto di Cosmogeofisica del Consiglio Nazionale della Ricerche e di Fisica Generale dell'Università, C.so Fiume 4, Torino, Italy*
(Received 14 May 1993; revised manuscript received 9 March 1994)

We test a method used to correct the box-counting estimate of fractal dimensions that is suitable only when a finite number of points is allowed. This method is based on the existing relations between the moments of cell counts for the sampling point distribution and the moments of point distribution for the underlying fractal structure. After a formal derivation of such relations, we apply them to fractal point distributions having different resolutions, dimensions, and numbers of sampling points. We find that a dilute sampling of a fractal structure often pollutes the scaling behavior, as revealed by a direct box-counting analysis. On the other hand, our correction procedure allows one to enlarge the scaling range up to more than one decade and also to recover the expected fractal dimension value.

PACS number(s): 05.45.+b, 47.53.+n, 98.62.Py

I. INTRODUCTION

Statistical analysis of real data samples based on fractal concepts have been shown to be an extremely useful approach in different physical contexts, in order to properly investigate the scaling properties of a given point distribution. Since the formalization by Mandelbrot [1] of the concept of fractal structure, many applications have been proposed, ranging from the study of the distribution of galaxies in the universe [1–7], to the intermittent behavior in turbulent flows [8,9], to the time series analysis [10–15].

A fundamental quantity that characterizes a fractal structure is the so-called fractal dimension. Mandelbrot defined a fractal as "a mathematical object whose fractal (Hausdorff) dimension is strictly larger than its topological dimension" [1]. Therefore, for a fractal distribution of points embedded in a three-dimensional ambient space, the topological dimension is $D_T=0$, and the fractal dimension D_0 must be $0<D_0\leq 3$. In order to better understand the meaning of the fractal dimension, let us consider a point distribution and suppose to cover it with a set of boxes of size r. For a scale-invariant structure, we expect that in the limit $r\rightarrow 0$ the number of nonempty boxes scales as

$$N_b(r)\propto r^{-D_0} . \tag{1}$$

Here the scaling index D_0 is defined as the box-counting, or capacity, dimension, which in general gives a close estimate of the Hausdorff dimension. According to Eq. (1), D_0 depends only on the number of nonempty boxes, thus dealing only with the geometry of the distribution. However, we are interested in recovering the correct clustering properties (i.e., how many points are in each box), characterized by the behavior of the generalized fractal dimensions (see Sec. II).

Note, however, that while a rigorous fractal extends its self-similarity from arbitrarily small up to arbitrarily large scales, physical systems develop self-similarity only in finite scale ranges. An example of this is represented in the cosmological context by the galaxy distribution. In this case, gravitational force naturally generates fractal behavior on the small scale of nonlinear dynamics [16], while on a large scale the homogeneity of the galaxy distribution is observed. Furthermore, the point distributions one usually deals with often suffer due to intrinsic biases. For instance, if a measurement device is able to detect only signals exceeding a given intensity value, threshold effects destroy self-similarity at large scales, according to the so-called *multiscaling* prescription [17]. Furthermore, while the formal definition of a fractal dimension is given in the limit of infinitesimally small scales, in all practical cases one deals with a finite number of points, so that only a finite scale range can be probed. In fact, at very small scales, all the boxes contain at most only one particle, so that $N_b(r)$ coincides with the total number of points. In this regime, increasing the size of the boxes does not significantly change $N_b(r)$, so that, according to Eq. (1), it is $D_0\simeq 0$, which is just the topological dimension of each single point. For all these reasons, a number of different dimensional estimators have been introduced, which relies on different approximations to the "true" fractal dimension (e.g., Ref. [18], and references therein). Due to such approximations, these methods suffer due to a number of shortcomings which depend on the dimensionality and on the sampling accuracy of the analyzed structure.

In this paper we address the problem of treating the effects of undersampling in the determination of the clustering scaling properties, when only a limited number of points is allowed. We discuss a suitable procedure for enlarging the scale range of the detected self-similarity, and for recovering the correct scaling behavior from a poor

point sample, under the assumption that it represents a Poissonian sampling of an underlying fractal structure. A more general treatment of the same problem has already been discussed in Ref. [19]. However, the present analysis will elucidate several aspects concerning the recovering of the background fractal structure in the presence of a limited number of points. Furthermore, our approach is also different from that presented in Ref. [19], and clearly shows the underlying connection between ideal and sampled fractal structures. Our results permit us to check whether the detected absence of self-similarity is intrinsic to the system under analysis or if it is a spurious product of undersampling over an otherwise fractal structure.

II. STATISTICAL BACKGROUND

In order to account for the clustering of a point distribution, let us consider a family $\{\Lambda_i\}_r$ of $N_c(r)$ cells of size r $[i=1,\ldots,N_c(r)]$, which completely covers the fractal structure. Then, if $d\mu(x)$ is defined as a local probability measure over the fractal structure, then the *coarse-grained* probability

$$\chi_i(r)=\int_{\Lambda_i} d\mu(x) \tag{2}$$

gives the fraction of mass, i.e. of the total number of points contained inside the cell volume Λ_i. Therefore, if $p(\chi)$ is the probability density function (PDF) for the χ variable, its moment of order q reads

$$m_q \equiv \langle \chi^q \rangle = \int d\chi\, \chi^q p(\chi)\ . \tag{3}$$

Therefore, the statistics of the fractal can be described by the moment-generating function $M(t)$, which is defined as the Laplace transform of the PDF, according to

$$M(t) \equiv \int d\chi\, p(\chi) e^{t\chi} = \langle e^{t\chi} \rangle\ . \tag{4}$$

In this way, the moments m_q are the coefficients of the McLaurin expansion,

$$M(t)=\sum_{q=0}^{\infty} \frac{m_q}{q!} t^q, \quad m_q = \left. \frac{d^q M(t)}{dt^q} \right|_{t=0}\ . \tag{5}$$

The multifractal spectrum of Renyi dimensions [20] is determined by the scaling of the m_q moments:

$$D_q = \frac{1}{q-1} \lim_{r\to 0} \frac{\log m_q(r)}{\log r}\ . \tag{6}$$

Note that generalized dimensions of positive order q mostly weight the cells having a high probability measure, so that they account for the scaling inside the overdense parts of the distribution. On the other hand, the $q<0$ tail deals with the underdensities. In this sense, the spectrum of Renyi multifractal dimensions gives a comprehensive description of the clustering, other than that of the geometry, of a fractal structure. Under general conditions, it is possible to demonstrate that the shape of the D_q curve is not completely arbitrary, but rather is a nonincreasing function of q. A particularly simple case occurs when D_q is constant and a single scaling index completely describes the statistics. In this case, the structure is called *monofractal*. See Ref. [21] for a technical introduction to multifractals.

Although the D_q dimensional spectrum can be defined for any real q, the m_q moments provided by the McLaurin expansion of $M(t)$ deal only with non-negative integer q's. However, it is possible to show that the $M(t)$ generating function completely specifies the *whole* set of D_q values (e.g., Refs. [22] and [23]). Here we will restrict our attention only to integer $q \geq 2$, so that Eq. (5) can be taken as the starting point for our implementation.

III. CORRECTING THE BOX-COUNTING ALGORITHM

Although the probability moments m_q and the corresponding generating function $M(t)$ characterize the fractal as a mathematical structure, in practical estimates of fractal dimensions one usually deals with a finite number of points, which represents a sampling of the underlying "true" structure. In this context, the box-counting method represents a classical approach to estimate the fractal dimension, and is based on the definition (6) of Renyi indices. For a distribution of a total number N_t of points, we define the box-counting partition function

$$Z(r,q)=N_c(r)\frac{\langle N^q \rangle_r}{N_t^q}\ , \tag{7}$$

where $N_c(r)$ is as before the total number of boxes, and $\langle N^q \rangle_r$ is the qth order moment for the count of points within boxes of size r. According to Eq. (6), for a fractal distribution, we expect that

$$Z(r,q) \propto r^{(q-1)D_q}\ , \tag{8}$$

and the multifractal dimension spectrum is recovered from a log-log linear regression of the partition function in the scale range where a pure power-law shape is detected.

However, it is clear that sampling a fractal structure with a finite number of points allows only an approximate estimate of the "true" dimension. Therefore, a suitable prescription should be devised to recover it from the scaling analysis of a limited data set. To this aim, let us consider the PDF for a Poisson point distribution (e.g., Ref. [24]):

$$p(\chi)=\sum_{N=0}^{\infty} \frac{1}{N!} \bar{\chi} e^{-\bar{\chi}} \delta_D(\chi - N)\ . \tag{9}$$

In the above expression, $\bar{\chi}$ is the average value of the coarse-grained probability χ, while the Dirac δ-function accounts for the discrete nature of the distribution and constrains it to take only non-negative integer values. According to Eq. (4), the corresponding moment-generating function becomes

$$M(t)=\exp[\bar{\chi}(e^t-1)] \tag{10}$$

and the expression $M(t)=e^{t\bar{\chi}}$, expected for a uniform continuous field, is recovered after substituting $t \to e^t - 1$. Therefore, the effect of the discrete sampling is accounted for by this change of variable in the functional dependence of the moment-generating function. Accordingly,

under the assumption that a given point distribution represents a Poissonian sampling of an underlying structure, it is straightforward to recognize (see, e.g., Ref. [25]) that the generating function of the moments of the discrete sampling, $M_{\rm discr}(t)$, is connected to that $[M(t)]$ of the underlying structure according to

$$M_{\rm discr}(t)=M(e^t-1) . \tag{11}$$

Equation (11) represents the basic relation which connects the "true" statistics to those of the sampled structure. In fact, by successively differentiating the above relation the $\langle N^q\rangle$ moments can be expressed order by order in term of the m_q moments. At the first four integer q values, it is

$$\langle N\rangle=m_1,\quad \langle N^2\rangle=m_1+m_2 ,$$
$$\langle N^3\rangle=m_1+3m_2+m_3,\quad \langle N^4\rangle=m_1+7m_2+6m_3+m_4 , \tag{12}$$

and more complicated expressions follow at higher orders. Following Eqs. (12), it is easy to obtain recursively the values of the "true" moments m_q from the measured $\langle N^q\rangle$, so that the correct fractal dimension can be recovered (see Eq. (8) in Ref. [19]).

We note that the same set of Eqs. (12) can also be applied to correct the scaling detected by the correlation-integral method. In this case, the $\langle N^q\rangle_r$ quantities represent the moments of counts of neighbors within the distance r from a point belonging to the set. Its application in the cosmological context to the analysis of the distribution of galaxy clusters [26] has shown it to provide reliable dimension estimates.

Grassberger [19] devised a general prescription to correct dimensional estimates based on box-counting and correlation-integral methods for any real q value. However, our method should be considered as an alternative approach, to derive corrections to dimensional estimates due to finite statistics. In Sec. IV we will test in detail the reliability and robustness of this method, when applied to different fractal point distributions, generated by using different algorithms, and having different dimensions, sampling rate, and resolution.

IV. RESULTS

In order to test the reliability of the correction procedure described in Sec. III, we apply it to *a priori* known fractal point distributions, generated by means of the β-model algorithm [27,28], as well as of the Henon map [29].

As for the β model, in order to generate a fractal embedded in a three-dimensional ambient space, let us start with a parent cube with side L_0 and break it into 2^3 parent subcubes having side $L_1=L_0/2$. After n breaking iterations, we generated a total number $M=2^{3n}$ of small cubes, each having size $L_n=L_0/2^n$. Let us assign to each cube a probability p to survive after each breaking step, and continue the cascading. Therefore, the number of active cubes after n iterations is a random variable, having mean $\langle m\rangle=pM$. In the limit of an infinite number of iterations, it is easy to show that the distribution of survived object is a monofractal structure, with dimension

$$D_0=\lim_{k\to\infty}\frac{\log\langle m\rangle^k}{\log 2^k}=\frac{\log\langle m\rangle}{\log 2}=\frac{\log pM}{\log 2} , \tag{13}$$

k being the iteration order. From the above formula, for $p=1$ all the cubes survive and $D_0=3$, as expected for a homogeneous, space-filling distribution. In general, lower and lower p values correspond to more and more clustered fractals, having a progressively lower dimension. A further interesting possibility occurs when the probability p depends on the scale L (i.e., on the cascading iteration). In this case, the resulting structure displays different scaling properties at different scale ranges, or no scaling at all. From an operative point of view, the limit of infinite iteractions *cannot* be achieved for a real distribution generated by means of a computer algorithm; usually one stops after a given number of iterations and associates a point with each active survived object. This simple model has been shown to be quite useful for modeling the fractal properties of the galaxy distribution [28].

According to the prescription of the β model, we generated different fractal point distributions. The distribution (a) has $N_p=128\,000$ points, obtained with three homogeneity iterations with $p=1$, and eight fractal iterations with $p=\frac{1}{4}$ corresponding to a fractal dimension $D=1$. Therefore, the scale at which the distribution becomes homogeneous is $\frac{1}{8}$ of the whole box size, while the smallest resolution scale is $1/2^{11}$ of L_0. The distribution (b) has the same dimension and scaling regimes of the first one, except that we stop the cascading at the eight iteration. The resulting number of points is $N_p=32\,000$. Taking fewer iteration steps clearly decreases the small-scale resolution with which we generate the fractal distribution, and allows us to check the effect of this on the efficiency of the Poissonian corrections, introduced in Sec. III. In order also to investigate the effect of taking different fractal dimensions, we also generated a distribution (c), with one filling iteration and eight iteration with $p=\frac{1}{2}$ that forces $D=2$. The resulting total number of points is $N_p=115\,000$. For each of these structures, with the corrected and noncorrected box-counting algorithms we analyzed several randomly selected subsamples with varying number of points: $N_1=32\,000$, $N_2=5000$, and $N_3=500$.

We also generated a point distribution obtained from a Henon recursive map, with parameters $a=1.4$ and $b=0.3$, iterated 100 000 times, extracting $N_4=1000$ points with a Poissonian sampling.

In Fig. 1 we plot the results of the analysis for $q=2$, 3, and 4 (from left to right) for the distribution (a) with N_1 points, as well as for the distribution (b) with the same number of points. Note that in the first case the distribution represents a Poissonian sampling of the whole structure. The lower part of each panel shows the corresponding moments of cell count, and the upper part the resulting local dimension obtained by a log-log linear regression on the moment slope. With this kind of plot the flattening of the local dimension corresponds to the detection of a scaling (fractal) range. Filled and open circles

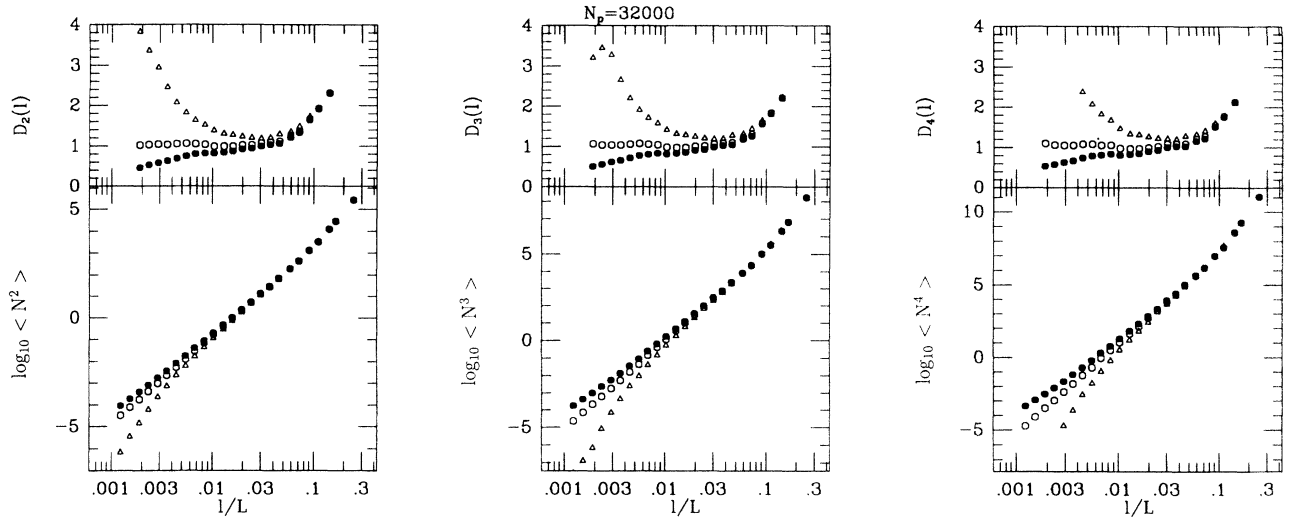

FIG. 1. The box-counting fractal analysis for the (a) and (b) scale-dependent fractal distributions (see text) with $N_p = 32\,000$ points. From left to right we report the results for $q = 2$, 3, and 4. The lower panels are for the cell-count moments, and the upper panels are for the corresponding local dimension, obtained from a five-point log-log linear regression on the $\langle N^q \rangle$ shape. Filled circles are for the noncorrected analysis of the (a) distribution, the open circles for the corrected analysis of the same distribution, and open triangles are for the corrected analysis of lower resolution (b) distribution. The reliability of the correction when applied to a high-resolution structure, and its failure when trying to correct a low-resolution distribution, are apparent.

are for the (a) distribution and represent noncorrected and corrected values, respectively. The open triangles are for the corrected box-counting analysis of the (b) distribution. From this plot, the reliability of introducing the Poissonian corrections in the box-counting estimate of the generalized fractal dimensions in the high-resolution structure is apparent; after correcting, the scale range in which the local dimension remains flat is increased by more than a decade. On the other hand, the effects of correction destroy the scaling for lower resolution structure. This is not surprising, since in this case the distribution does not represent a Poissonian sampling of a richer structure. Therefore, correcting for Poissonian undersampling amounts to working out the noise, which is present below the resolution scale, with a subsequent increase of the local dimension to the $D = 3$ value. The noncorrected analysis of distribution (b) would of course show discreteness at small scales, but it is "structural," since it is related to the limited number of iterations performed in the β model.

In Fig. 2 we show the same analysis for the distribution (a), but for the subsample with $N_2 = 5000$ points. The correction still recovers the scaling and the correct fractal dimension, especially for $q = 2$. On the contrary, the noncorrected analysis shows evidence of discreteness effect that lowers the measured dimension value and com-

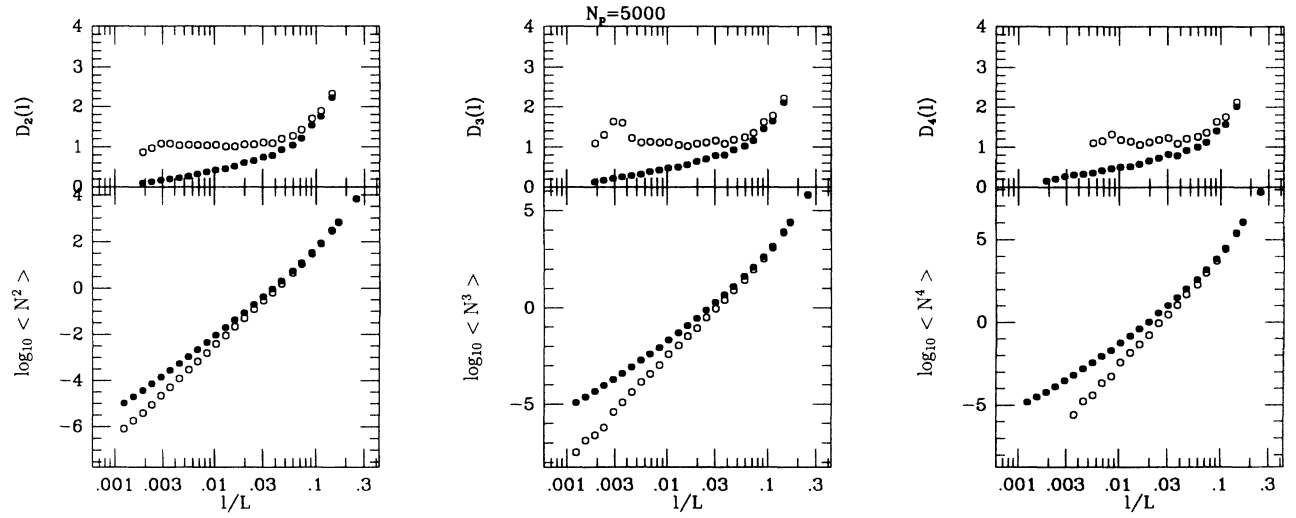

FIG. 2. The same as in Fig. 1 but only for the (a) distribution with $N_p = 5000$ points. Note the reliability of the correction to recover the correct scaling, despite the fact that poor statistics completely pollute the scaling in the uncorrected analysis.

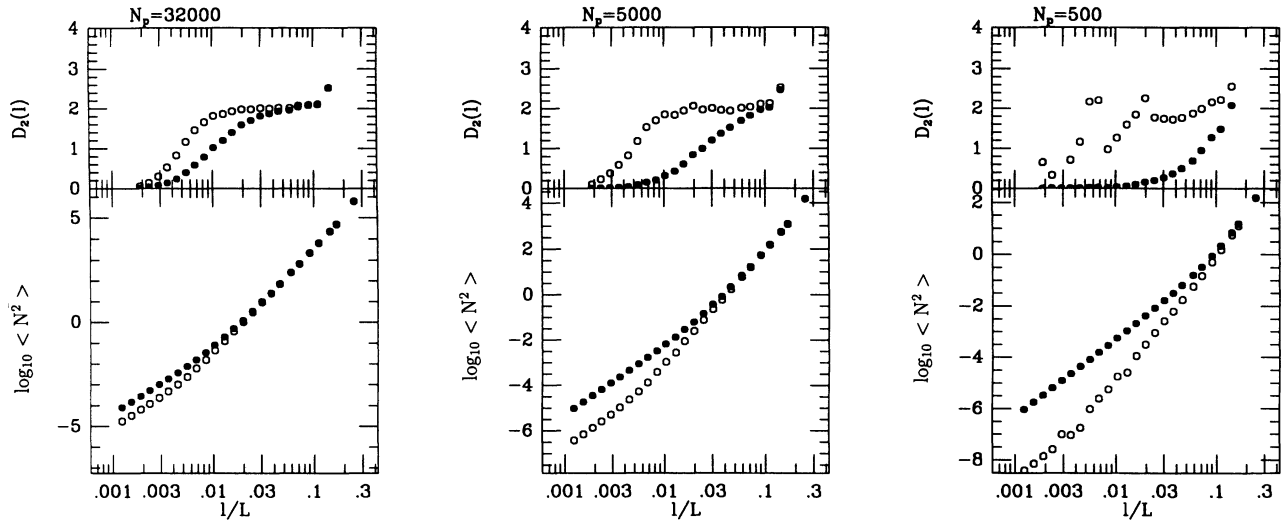

FIG. 3. The effect of taking randomly selected subsamples with different dilution factors on the (c) distribution. Only the results for $q=2$ are shown. Again the correction recovers the scaling behavior over a one decade scale range. Only for the poorest sample is the local dimension affected by noise, although the moment $\langle N^2 \rangle$ recovers an overall correct shape.

pletely masquerades any evidence of scaling behavior. Once more, this result supports the reliability of subtracting the discreteness contributions from the measured cell count moments according to Eqs. (12), especially for diluted distributions, which otherwise do not show any evidence of fractality.

Figure 3 reports the results for the $D_2=2$ distributions with $N=32\,000$, 5000, and 500 points, for $q=2$ only. The correct value of $D=2$ and an extension of the scaling regime are again obtained. Note that for $N=5000$ points no evidence of a scaling regime is shown without the use of this correction. The lack of such a statistic becomes dramatic when only 500 points are analyzed; in this case, no scaling at all is shown without the correction, while after correcting some marginal evidence indicates that the local dimension takes the correct values, despite the large dispersion due to the limited statistics.

In Fig. 4 we report results for the analysis of the distribution sampled from the Hénon map. Also in this case, the recovery of the scaling regime is confirmed, despite the low sampling rate of the underlying fractal structure.

On the basis of these results, we conclude that the procedure described in Sec. III to correct the box-counting fractal dimensions for Poissonian sampling is a rather reliable procedure in order to verify whether the absence of scaling is intrinsic to the distribution or is only an effect of poor statistics. Using a fractal model (the β model) with known fractal properties, an extension of the scaling regime ranging from half to one decade has been always achieved, while also recovering the expected value of the local dimension. A Hénon undersampled distribution has shown the same improvement in detecting the scaling re-

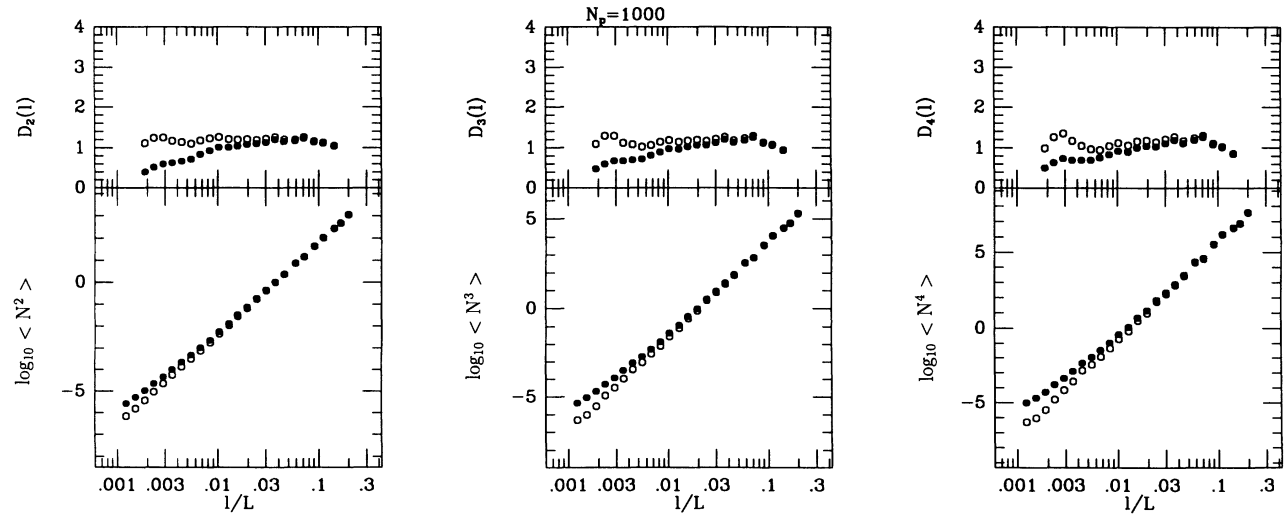

FIG. 4. The same as in Fig. 2, for distribution (d). The correction is reliable in this case, as well. Note that the sampling rate is remarkably low ($\frac{1}{100}$).

gime. Even in the presence of a rather heavy undersampling, precise hints about the behavior of the fractal nature of the point distribution are achieved. We point out that this method is also able to discriminate between a distribution which represents a Poissonian sampling of an underlying highly resolved fractal structure, and one which samples a low-resolution structure. In the second case, applying the correction at scales near to that of the limiting resolution leads only to the detection of noise, which translates into an increase of the small-scale local dimension.

ACKNOWLEDGMENTS

The authors are grateful to Dr. A. Provenzale for suggestions and for a critical reading of the manuscript. G.M. wishes to acknowledge Professor C. Castagnoli for support and encouragement.

[1] B. B. Mandelbrot, *The Fractal Geometry of Nature* (Freeman, San Francisco, 1982).
[2] G. Efstathiou, S. M. Fall, and G. Hogan, Mon. Not. R. Astron. Soc. **189**, 203 (1979).
[3] B. J. T. Jones, V. J. Martinez, E. Saar, and J. Einasto, Astrophys. J. Lett. **332**, L1 (1988).
[4] V. Martinez and B. J. T. Jones, Mon Not. R. Astron. Soc. **242**, 517 (1990).
[5] V. Martinez, B. J. T. Jones, R. Dominguez-Tenreiro, and R. van de Weygaert, Astrophys. J. **357**, 50 (1990).
[6] L. Pietronero, Physica A **144**, 257 (1987).
[7] A. Provenzale, in *Applying Fractals in Astronomy,* edited by A. Heck and J. Perdang (North-Holland, Amsterdam, 1992).
[8] K. R. Sreenivazan and C. Meneveau, J. Fluid. Mech. **173**, 357 (1986).
[9] K. R. Sreenivazan, R. Ramshankar, and C. Meneveau, Proc. R. Soc. London Ser. A **421**, 79 (1989).
[10] P. Grassberger and I. Procaccia, Physica D **9**, 189 (1983).
[11] J. P. Eckmann and D. Rouelle, Rev. Mod. Phys. **57**, 617 (1985).
[12] A. Provenzale, L. A. Smith, R. Vio, and G. Murante, Physica D **58**, 31 (1992).
[13] L. A. Smith, Phys. Lett. A **133**, 283 (1988).
[14] J. Theiler, Phys. Rev. A **34**, 2427 (1986).
[15] J. Theiler, J. Opt. Soc. Am. A **7**, 1055 (1990).
[16] R. Valdarnini, S. Borgani, and A. Provenzale, Astrophys. J. **394**, 422 (1992).
[17] M. H. Jensen, G. Paladin, and A. Vulpiani, Phys. Rev. Lett. **67**, 208 (1991).
[18] S. Borgani, G. Murante, A. Provenzale, and R. Valdarnini, Phys. Rev. E **47**, 3879 (1993).
[19] P. Grassberger, Phys. Lett. A **128**, 369 (1988).
[20] A. Renyi, *Probability Theory* (North Holland, Amsterdam, 1970).
[21] G. Paladin and A. Vulpiani, Phys. Rep. **156**, 147 (1987).
[22] R. Balian and R. Schaeffer, Astron. Astrophys. **226**, 373 (1989).
[23] S. Borgani, Mon. Not. R. Astron. Soc. **260**, 537 (1993).
[24] M. Abramowitz and I. A. Stegun, *Handbook of Probability Functions,* 10th ed. (Dover, New York, 1972).
[25] D. Layzer, Astrophys. J. **61**, 383 (1956).
[26] S. Borgani, V. J. Martinez, M. A. Perez, and R. Valdarnini, Astrophys. J. (to be published).
[27] U. Frish, P. L. Sulem, and M. Nelkin, J. Fluid Mech. **87**, 719 (1978).
[28] C. Castagnoli and A. Provenzale, Astron. Astrophys. **246**, 634 (1991).
[29] M. Henon, Commun. Math. Phys. **50**, 69 (1976).