Topologies for point distributions

E. Elizalde

Departament de Física Teòrica, Universitat de Barcelona, Diagonal 647, 08028 Barcelona, Spain

(Received 13 November 1986; accepted for publication 13 August 1987)

The concepts of void and cluster for an arbitrary point distribution in a domain $\mathscr{D}$ are defined and characterized by some parameters such as volume, density, number of points belonging to them, shape, etc. After assigning a weight to each void and cluster—which is a function of its characteristics—the concept of distance between two point configurations S_1 and S_2 in $\mathscr{D}$ is introduced, both with and without the help of a lattice in the domain $\mathscr{D}$. This defines a topology for the point distributions in $\mathscr{D}$, which is different for the different characterizations of the voids and clusters.

I. INTRODUCTION

The spatial distributions of points in one, two, or a higher number of dimensions (spatial processes, point processes, spatial patterns, spatial point patterns) constitute a very interesting field of research, not only in pure mathematical statistics but also in its innumerable applications, which range from biometrics to astrophysics and includes such diverse fields as agriculture, econometrics, ecology, traffic problems, and medical sciences. In fact, such point distributions may correspond, for instance, to plants of a given species, to cars along a road, to seeds in a field, to microorganisms in a living body, or to stars—or even galaxies or clusters of galaxies—in our Universe. An important mathematical aspect of spatial processes is the study of the geometrical and topological properties of point distributions.

Although in biometry the study of such distributions has always been very popular during the last years—as can be seen through the large number of research articles and even books which have been issued,[1] in the last couple of months the interest about this field of research has grown very rapidly, due in great part to the remarkable discoveries of de Lapparent, Geller, and Huchra about the spatial distribution of galaxies in our Universe.[2] These authors made an optical red shift survey of all 1099 galaxies brighter than magnitude 15.5 of a thin slice of sky and came to the conclusion that galaxies are concentrated on the surfaces of contiguous bubble like structures with very large typical diameters of about 25 h^{-1} Mpc. The large void in Boötes of 60 h^{-1} Mpc, discovered in 1981 by Kirshner *et al.*,[3] has been thereby proved to be no peculiarity but a very common feature. Too often, the analysis of the point distributions of galaxies, with their voids and clusters, is done simply by looking at pictures and plates with the naked eye, a very primitive procedure which in general is not that bad. Nevertheless, the important discoveries we have just mentioned stress once more the necessity for a more profound understanding of spatial point distributions and, in particular, of the still unsolved problem concerning the construction of a mathematical "measure" for quantifying how far away are two of such point distributions (characterized by the number and magnitude of the voids and clusters, their forms and spatial distribution, etc.).

II. DEFINITIONS OF (SPHERICAL) VOID AND CLUSTER

The very large numbers of points one has to deal with makes it almost a necessity to introduce definitions which are suitable to be treated with a numerical algorithm. This has been pointed out in several previous papers on the subject[4] and will be considered later in detail (Sec. VI). However, I do not think that discrete algorithms alone can solve these problems satisfactorily, and it is much better to play at a time both with discrete and with continuous concepts.

Let S be a set of points in a given domain $\mathscr{D}$ of volume V in d-dimensional Euclidean space. Let N be the number of points in S. For any point $p\in\mathscr{D}$ and any positive real number $r\in\mathbb{R}^+$, the density of points in a ball around p of radius r is given by

$$\rho_p(r) = n_p(r)/V_p(r), \tag{2.1}$$

where $n_p(r)$ is the number of points of S inside the ball, and $V_p(r)$ is the volume of the ball. By definition, there is a void around p of radius bigger than r if the density $\rho_p(r)$ verifies

$$\rho_p(r) < \lambda N/V, \tag{2.2}$$

where $\lambda \leqslant 1$ must be fixed (we may take, for instance, $\lambda = \frac{1}{2}$). The radius of the void around p is defined to be the value r_p such that

$$\rho_p(r_p) = \lambda N/V. \tag{2.3}$$

In this way the density of any void will be the same. Alternatively, one could define the radius of the void as the value of r at which the slope of $\rho_p(r)$ is maximum.

On the other hand, there exists, by definition, a cluster around p if

$$\rho_p(r) > N/\lambda V. \tag{2.4}$$

The radius of the cluster may be defined to be the value r_p such that

$$\rho_p(r_p) = N/\lambda V. \tag{2.5}$$

As before, one could alternatively define the radius of the cluster as the value of r at which the slope of $-\rho_p(r)$ is maximum.

Until now we have studied only what happens at some given place p. A global analysis has to distinguish between the different voids and clusters, so that we do not count the same point of the distribution twice: as belonging to a void (or cluster) and to another one which intersects the first. Moreover, the preceding definitions are best suited for spherical voids or clusters only and when p is the center of them. These difficulties will be taken care of in the subsequent sections.

III. EFFICIENT SEARCH FOR VOIDS AND CLUSTERS

Let us now introduce a lattice $\mathscr{L}$ of lattice site a in the domain $\mathscr{D}$. For a given set of $r_k \in \mathbb{R}^+$, $k = 1,2,...,m$, select the set S_1 of vertices of the lattice corresponding to the s smallest and to the h highest values of $\rho_{p_i}(r_k)$, for all $k = 1,2,...,m$, for all lattice vertices p_i on $\mathscr{L}$. Improve now the set of points p_i in S_1 to a set S_2 coming from the s smallest and h highest values of $\rho_{q_i}(r_k)$ for all $k = 1,2,...,m$, and for all q_i of the form

$$p_i + \lambda \left[\prod_{k=1}^{m} \rho_{p_i}(r_k) \right]^{-1/2m} \hat{e}, \tag{3.1}$$

where λ is a constant that we can adjust at will (for instance, $\lambda = \frac{1}{2}$ for voids and $\lambda = 2$ for clusters), while $\hat{e}$ sweeps all unitary directions of the form

$$\hat{e}_{i_1} \pm \cdots \pm \hat{e}_{i_l}, \quad 1 \leqslant i_1 < \cdots < i_l \leqslant d, \tag{3.2}$$

where $\hat{e}_i$ is the unitary vector along the i axis of $\mathbb{R}^d$. Notice that, in general, the points q_i are not vertices of the lattice. In fact, the vertices of the lattice serve only as starting points in order to begin the search for the best centers of the voids and clusters.

The procedure is then repeated until it stabilizes. In this way we obtain the positions of the centers of a desired number of the less dense voids and of the more dense clusters in the point distribution S.

IV. WEIGHTS OF THE INDIVIDUAL VOIDS AND CLUSTERS

The weight of a spherical void of radius r and density ρ with the center at the point p is given by the following expression:

$$W_v = k_v [V_p(r)/\rho] = k_v [V_p(r)^2/n], \tag{4.1}$$

where k_v is a constant (independent of the void), $V_p(r)$ the volume of a sphere of radius r in d dimensions, and n is the number of points of S inside the sphere. That this expression is correct can be seen through the following argument. For a given density ρ, increase of W_v in (4.1) is proportional to the volume of the void, while for fixed volume, increase of W_v is proportional to decrease of ρ_i, as it should be by intuition. Alternatively, at fixed n increase of W_v is proportional to the volume and *also* to the decrease of density, i.e., proportional to the volume squared.

The weight of a spherical cluster of radius r and density ρ centered at p is given by

$$W_c = k_c n\rho = k_c [n^2/V_p(r)] = k_c \rho^2 V_p(r), \tag{4.2}$$

where k_c is a constant independent of the cluster. Expression (4.2) can be understood by reasoning as follows. At fixed n, W_c is proportional to increase of ρ (or to decrease of volume). At fixed ρ, W_c is proportional to increase of n (or to increase of volume). Alternatively, with full generality, increase of W_c is due both to increase of n (it would also be true for fixed ρ, i.e., letting volume increase) and to increase of ρ.

V. THE SET OF ALL POINT DISTRIBUTIONS AS A METRIC SPACE

Once the spherical voids and clusters have been constructed by the procedures described in Secs. II and III, and just before their individual weights (Sec. IV) are calculated, one has to look for superpositions of them which may result in nonspherical voids and clusters. The idea is very simple: to consider as a unique void (resp. cluster) the union of all of them which are connected by a chain of intersections (Fig. 1). In this way, the numbers of voids and clusters, s and h, respectively, diminish and, at the same time, they are no more spherical but acquire a form of the type depicted in Fig. 2. It is now immediate to modify the formulas (4.1) and (4.2) accordingly: in both cases the volume $V_p(r)$ of a d-dimensional sphere of radius r centered at p must be substituted by the volume of the void or cluster considered. Of course, the density ρ and the number of points n will also correspond now to the whole, nonspherical void or cluster.

Once all the voids and all the clusters have been constructed, the remaining region of the domain $\mathscr{D}$ is filled up with a (under ideal conditions) sensibly uniform distribution of points of S with a density almost equal to $\rho_0 = N/V$. In practice this must be checked *a posteriori* and if it were not true, the free parameters introduced in the definitions and construction of the voids and clusters (s, h,...) ought to be changed accordingly. For instance, if the density of the remaining region were smaller than ρ_0, then the number s of voids should be increased. On the other hand, if the homogeneity of the remaining region were not very good then both s and h ought to be augmented.

Let us now consider the plane (V,ρ) and the points (V_i,ρ_i) in it, where the index i goes through all the different voids and clusters, with one value of the index corresponding to the intermediate, remaining region. Introduce a regular lattice in this plane and denote the different cells by (V_j,ρ_j), $j\in\mathscr{J}$. Define now the function $f(V_j,\rho_j)$ which as-

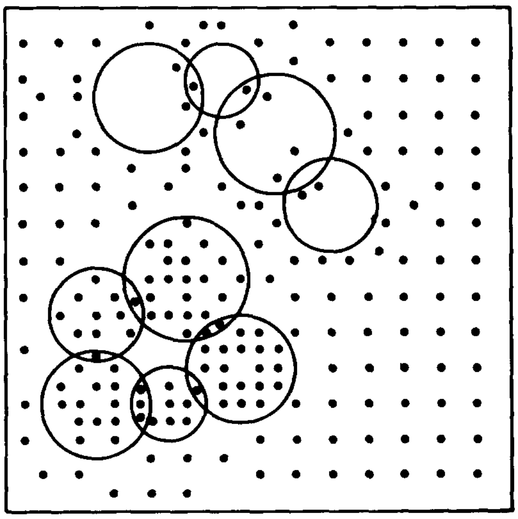

FIG. 1. Examples for distributions of spherical voids and clusters. They can superpose in a variety of ways.

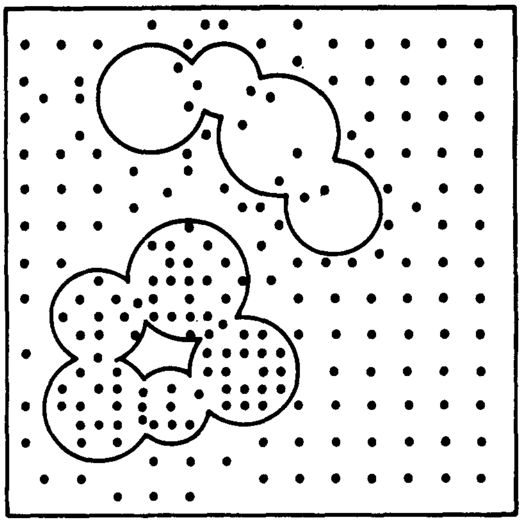

FIG. 2. Voids and clusters constructed from the distributions of Fig. 1 in order to avoid overcounting.

signs to each cell (V_j,ρ_j) the number of clusters plus the number of voids which belong to this cell, each of them multiplied by the corresponding weight (4.2) and (4.1), respectively, i.e.,

$$f(V_j,\rho_j) = n_c(V_j,\rho_j)W_c(V_j,\rho_j) + n_v(V_j,\rho_j)W_v(V_j,\rho_j), \quad j\in\mathscr{J}. \qquad (5.1)$$

Here the intermediate region is to be counted as an additional void or cluster depending on its density ρ being $\rho\leqslant\rho_0$ or $\rho>\rho_0$, respectively. Notice that $f(V_j,\rho_j)$ is different from zero only in a finite number of cells (V_j,ρ_j). The following step is to construct with these points the minimal triangulated surface with vertices at these points. Thus, we get a continuous function $f(V,\rho)$ defined on the plane (V,ρ). Now, given another point configuration on the domain $\mathscr{D}$, we define the distance between these two configurations S_1 and S_2 by

$$d(S_1,S_2)^2 = \iint [f_1(V,\rho) - f_2(V,\rho)]^2\,dV\,d\rho, \qquad (5.2)$$

where f_1 and f_2 are the functions corresponding to the point configurations S_1 and S_2, respectively.

The problem we are dealing with is not so standard. No wonder, therefore, that definition (5.2) is not a *usual* measure of the configuration space. However, it is important to observe that d, as given by (5.2), can be easily implemented to yield a true distance by the usual mathematical procedures. Let us be completely rigorous.

The set which is going to turn into a metric space is $\mathscr{S}$ = set of all finite point distributions in the domain $\mathscr{D}$. The "distance" defined by (5.2) is actually only a semidistance. In fact, it satisfies (i) $d(S_1,S_2)=0$, (ii) $d(S_2,S_1)=d(S_1,S_2)$, and (iii) $d(S_1,S_2)\leqslant d(S_1,S_3)+d(S_3,S_2)$, for any $S_1,S_2,S_3\in\mathscr{S}$. All we have to do is to define the coset $\bar{\mathscr{S}}=\mathscr{S}/\sim$, where $S_1\sim S_2$ iff $d(S_1,S_2)=0$, in order to obtain a metric space $\bar{\mathscr{S}}$ with the distance $\bar{d}$ given by

$$\bar{d}(\bar{S}_1,\bar{S}_2) = d(S_1,S_2),\ S_1\in\bar{S}_1,\ S_2\in\bar{S}_2. \qquad (5.3)$$

In fact, this is a consistent definition for, let us consider two other configurations $S_1'\in\bar{S}_1$ and $S_2'\in\bar{S}_2$. Then, we have

$$d(S_1,S_2)\leqslant d(S_1,S_1') + d(S_1',S_2') + d(S_2',S_2).$$

But $d(S_1,S_1')=0$ and $d(S_2,S_2')=0$, so that $d(S_1,S_2)\leqslant d(S_1',S_2')$. Moreover,

$$d(S_1',S_2')\leqslant d(S_1',S_1) + d(S_1,S_2) + d(S_2,S_2'),$$

and we get $d(S_1',S_2')\leqslant d(S_1,S_2)$. Therefore, $d(S_1',S_2') = d(S_1,S_2)$. Equation (5.3) defines a distance $\bar{d}$ on $\bar{\mathscr{S}}$. In fact, $\bar{d}$ satisfies the axioms (i)–(iii) above and, moveover, the additional one, (i′) $\bar{d}(\bar{S}_1,\bar{S}_2)=0$ implies $\bar{S}_1=\bar{S}_2$. This is immediate from (5.3) and from the definition of the coset $\bar{\mathscr{S}}$.

Summing up, $\bar{\mathscr{S}}$ is a metric space endowed with the distance $\bar{d}$. This constructive procedure is very well known to mathematicians, in fact, it is the standard way to proceed. This allows one to be a little loosely in the notation and speak of the metric space $\mathscr{S}$ and of the distance d, as given by (5.2). The alternative definitions of distance which will follow have to be compared with (5.2). Actually all of them ought to be submitted to the same procedure as given above in order that they become true distances $\bar{d}$.

A metric space is readily made into a topological space, the topology being provided by the distance, much as in the standard example of the metric space $\mathbb{R}^n$. The neighborhoods of the basis of this topology are open balls of the form $B_p(\bar{S}) = \{\bar{S}'\in\bar{\mathscr{S}}\,|\,\bar{d}(\bar{S},\bar{S}')<p\}$, p being any rational number $p\in\mathbb{Q}$. Being again a little loosely with the notation we may say that the set $\mathscr{S}$ of all finite point distributions in $\mathscr{D}$ is a topological space, the topology being given through the distance d in (5.2).

This is by no means the only possibility to define a distance between two point configurations. But the definition which has just been given above is quite a sensible one. An example of a different, more simple definition is the following. Consider the weights (4.1) and (4.2) and place them at the negative and positive semiaxis x, respectively (Fig. 3). Then discretize this axis by considering intervals of a given length l. For each interval of the x axis, on the y axis set the number of voids (resp. clusters) with a value of W_v (resp. W_c) which belongs to this interval. Now consider the segment-wise curve constructed with the resulting points (Fig. 3). Let us call this curve $g(x)$. The distance between two point configurations S_1 and S_2 can then by defined by

$$d(S_1,S_2)^2 = \int [g_1(x) - g_2(x)]^2\,dx. \qquad (5.4)$$

Notice, however, that on taking the weights from the beginning we have implicitly introduced in this last case an equivalence relation among voids (and among clusters). In some cases this can actually be convenient in order to simplify the problem from the beginning, but in other situations a finer definition such as the first one will have to be adopted.

VI. POINT DISTRIBUTIONS IN A DOMAIN WITH A LATTICE

In order to treat all the preceding questions in a way better suited for numerical manipulations, one can carry all these definitions to a lattice $\mathscr{L}$ of certain site a on the do-

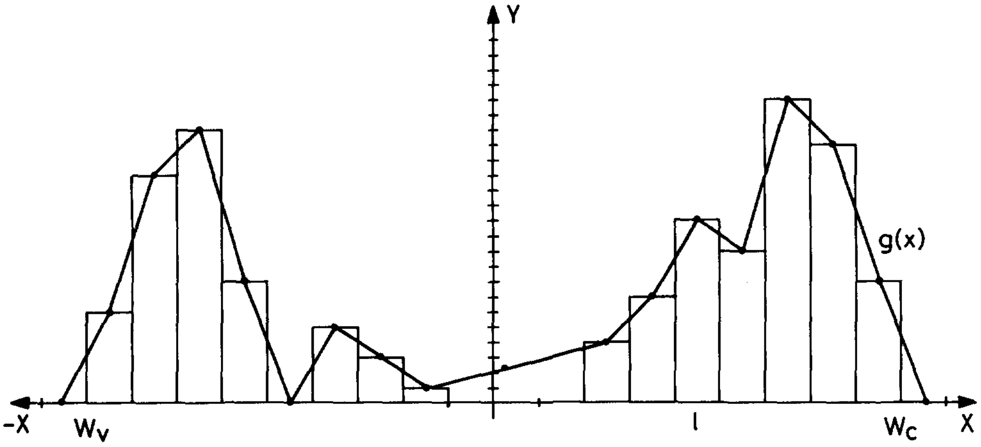

FIG. 3. Segment-wise curve $g(x)$ constructed using a discretization of the weights corresponding to voids (negative x axis) and clusters (positive x axis). In the y axis the numbers of voids and clusters that fall into each interval of x are represented.

main $\mathscr{D}$. That is, a discretization of the methods which have been elaborated above has to be developed.

One must start by counting the number of points of the distribution S in $\mathscr{D}$ that fall into each of the elementary cells of the lattice. The voids will consist of all those cells whose number of points does not reach a value n_1 fixed in advance, while the clusters are made up of cells with a number of points above a second number n_2 also fixed in advance. Of course, one must have

$$n_1 < N/N_c < n_2, \tag{6.1}$$

where $N_c = V/a^d$ is the number of cells of the lattice $\mathscr{L}$. (To begin with, we consider all cells equal and obviate the small modifications in these definitions which had to be made for cells touching the border of the domain $\mathscr{D}$.) In this way, extended voids and clusters made up of cells will arise, in general. A huge void (cluster) will consist of several contiguous cells with a small (big) number of points. Figure 2 will be almost the same, only that the curved contour will be substituted by a segment-wise one, with segments of longitude proportional to a. Formulas (4.1) and (4.2) will remain unchanged: only $V_p(r)$ will be substituted by the volume V_v or V_c of the void or cluster under consideration (a volume always proportional to a^d, the volume of an elementary cell).

Notice that this procedure is less time consuming than the former one when it is carried out in practice. However, it is not so sensible to detect the voids and clusters with precision. In fact, once the lattice $\mathscr{L}$ has been fixed, a given cell can participate at the same time of a void and of a cluster so that the total number of points in it may compensate (Fig. 4), thus hiding this fact completely. Clearly, everything becomes better as a is made smaller (continuum limit). However, with a (discrete) point distribution this cannot be done indefinitely: for a small enough every cell contains at most one point only and for such small cells the whole procedure ceases to be of much use (this was the difficulty with the topology of discrete point distributions in the first place).

Once the weights (4.1) and (4.2) have been adapted to the lattice voids and clusters, the definitions (5.2)–(5.4) for the distance between two point configurations S_1 and S_2 go through immediately. Thus, we complete the treatment of the lattice case and define a topology for point configurations on the domain $\mathscr{D}$. One could think, in principle, that in order to proceed in accordance with the discretization of $\mathscr{D}$, the plane (V,ρ) ought also to be discretized, i.e., divided into rectangles of sides a_V and a_ρ, and the minimal triangulated surfaces constructed using the vertices corresponding to the centers of these cells. However, it must be pointed out that this last discretization is completely independent from the one of the domain $\mathscr{D}$.

Finally, notice that in our definition only the volume and the density (we may substitute one of these by the number of points inside) of the void or cluster have been taken into account in the definition of the distance $d(S_1,S_2)$. A more elaborate definition should also include other parameters such as some characterizing the shape of the void or cluster (for instance, a combination of the diameters along each of the axes, as the sum or the product of these diameters). The function $f(V,\rho)$ given in (5.1) and the distance (5.2) have to be redefined accordingly. That is (we drop the subindex j for convenience)

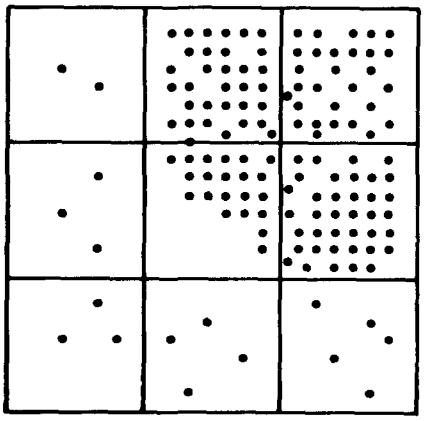

FIG. 4. Some cells (here the one in the middle) of a lattice in $\mathscr{D}$ may participate both from some void and from some cluster. They may compensate and give a deceptive mean density approximately equal to $\rho_0 = N/V$.

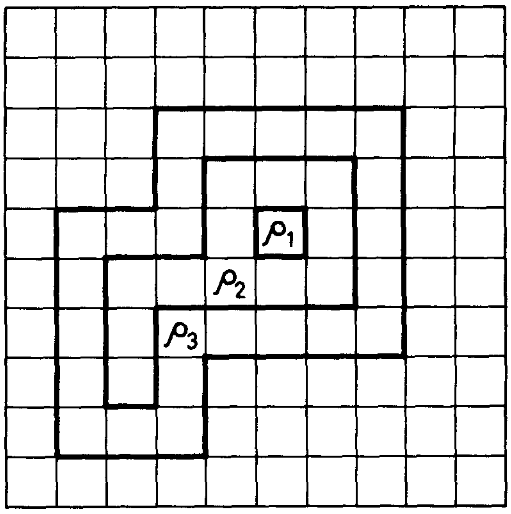

FIG. 5. A huge void or cluster (the points have not been depicted) displaying the gradient of density: ρ_1 is the density of the most inner cell; ρ_2 that of the surrounding crown of cells (here 12 cells); ρ_3 is the density of the most exterior crown (here 24 cells).

$$f(V,\rho,k) = n_c(V,\rho,k)\,W_c(V,\rho,k) - n_v(V,\rho,k)\,W_v(V,\rho,k), \tag{6.2}$$

where k is the new parameter, and

$$d(S_1,S_2)^2 = \iiint [f_1(V,\rho,k) - f_2(V,\rho,k)]^2\,dV\,d\rho\,dk, \tag{6.3}$$

respectively. In an analogous way, we may introduce other parameters, such as the gradient of density for large voids or clusters, as one proceeds from inside to the border (Fig. 5). We may define, for instance,

$$h = [(\rho_2 - \rho_1)^2 + (\rho_3 - \rho_2)^2 + \cdots]^{1/2} \tag{6.4}$$

and include h besides k as a new parameter. All these parameters and others one may think of improve the definition of the topology (6.3) and may be introduced at ease into our formalism in the way we have just shown.

VII. OUTLOOK

The procedures introduced here for the first time (to our knowledge) are currently being applied to the point distributions that correspond to the analysis of galaxies of de Lapparent *et al.*[2] and also to other related results. Moreover, simulation methods are being developed with the purpose of checking the reliability of the distance between point distributions as defined here compared with the only one which is presently available, namely, the "distance" that our naked eye would *grosso modo* assign to them. The investigation is in progress. Its partial results are pretty good and will be published elsewhere with a detailed account of the analysis involved.

ACKNOWLEDGMENTS

Useful conversations with E. Gaztañaga are gratefully acknowledged.

This paper has been partially supported by the CAICYT.

[1] M. S. Bartlett, *The Statistical Analysis of Spatial Pattern* (Chapman and Hall, London, 1975); A. D. Cliff and J. K. Ord, *Spatial Processes: Models and Applications* (Pion, London, 1981); D. R. Cox and V. Isham, *Point Processes* (Chapman and Hall, London, 1980); P. J. Diggle, *Statistical Analysis of Spatial Point Patterns* (Academic, London, 1983); B. D. Ripley, *Spatial Statistics* (Wiley, New York, 1981); A. Rogers, *Statistical Analysis of Spatial Dispersion* (Pion, London, 1974).

[2] V. de Lapparent, M. J. Geller, and J. P. Huchra, Astrophys. J. Lett. **302**, L1 (1986).

[3] R. P. Kirshner, A. Oemler, Jr., P. L. Schechter, and S. A. Schectman, Astrophys. J. Lett. **248**, L57 (1981).

[4] F. R. Bouchet and M. Lachièze-Rey, Astrophys. J. Lett. **302**, L37 (1986); J. E. Moody, E. L. Turner, and J. R. Gott, III, Astrophys. J. **273**, 16 (1983); S. Otto, H. D. Politzer, and M. B. Wise, Phys. Rev. Lett. **56**, 1878 (1986); H. D. Politzer and J. P. Preskill, *ibid.* **56**, 99 (1986); S. Otto, H. D. Politzer, J. Preskill, and M. B. Wise, Astrophys. J. **304**, 62 (1986).