Mon. Not. R. Astron. Soc. **272**, 231–240 (1995)

Quantifying the geometry of large-scale structure

Russell C. Pearson and Peter Coles
Astronomy Unit, School of Mathematical Sciences, Queen Mary and Westfield College, Mile End Road, London E1 4NS

Accepted 1994 September 5. Received 1994 July 14

ABSTRACT
We investigate a method for quantifying the geometrical properties of large-scale structure in the galaxy distribution. We use a graph-theoretical construction known as the Minimal Spanning Tree (MST) to delineate the main features of the structure. We then extract quantitative measures of the shape of the MST known as *structure functions*. By using simple models, we show that these measures can quantify the tendency of clustering to occur preferentially in filaments, sheets or isolated clumps, and that they are relatively insensitive to the addition of an unclustered background of points and to the effects of redshift-space distortion. We then apply the method to a set of simulations of the ultra-large-scale structure traced by rich clusters of galaxies. These pose a difficult test for our method, because of the low sampling density of clusters. Nevertheless, the method does reveal significant differences between various models of clustering. We also compare our results for the structure functions with those obtained using a statistic proposed by Vishniac.

Key words: methods: statistical – galaxies: clustering – galaxies: formation – cosmology: theory – large-scale structure of Universe.

1 INTRODUCTION

One of the most important outstanding tasks in cosmology is to understand the origin of large-scale structure in the distribution of galaxies that we observe in galaxy redshift surveys. This task involves an interplay between theory and observation which is necessarily statistical, and the first step in the confrontation of theory with observations is the extraction of statistical descriptors of the correlations displayed by galaxies. Much can be learned about the clustering of galaxies using relatively simple statistical measures, such as the two point correlation function (Peebles 1980), power-spectrum or counts-in-cells analysis. On the other hand, as the available galaxy catalogues get larger and larger, and theoretical models of structure formation become more and more sophisticated, attention will become increasingly focused upon statistical descriptors that are more sensitive to pattern (i.e. geometry) than the crude measures mentioned above.

One particular question one might like to ask of the galaxy distribution is whether galaxies form preferentially in isolated clumps, on filaments, on sheets or even on a mixture of these. The possible existence of 'filaments' (one-dimensional structures) in the galaxy distribution has been discussed for many years, though the early claims were based on a visual analysis of the data (Jôeveer, Einasto & Tago 1978). One must be mindful of the 'Canals on Mars' fiasco, and be sure to employ objective methods to avoid the eye's readiness to ascribe pattern where none exists in reality (Barrow & Bhavsar 1987). More recent galaxy catalogues have been claimed to show that the typical geometry of superclusters is not filamentary, but sheet-like (de Lapparent, Geller & Huchra 1986, 1989; Wegner et al. 1990). It has also been suggested that the visual texture of galaxy clustering resembles a network of bubbles, with large void regions surrounded by these sheets (de Lapparent et al. 1986). In 'standard' models of structure formation, wherein structures at the present epoch form by the action of a gravitational instability from small primordial density perturbations, the formation of 'pancakes' (two-dimensional collapsed structures) is expected to be commonplace; there are also good reasons to expect some filaments to appear (Shandarin & Zel'dovich 1989; Sahni & Coles 1994). The relative numbers and sizes of these different geometrical features depend in a complicated way on the initial fluctuation spectrum and normalization, so a precise quantification of the geometrical properties of observed clustering may place important constraints on structure formation models.

Given the importance of this problem, it is not surprising that many ideas have been put forward as to how one might quantify the geometry of the clustering pattern in an objective way. Some of these have been successful, others less so. Among the methods proposed* have been percolation analysis (Zel'dovich, Einasto & Shandarin 1982), alignment statis-

* Note that we do not include in this list the *genus* statistic proposed by, among others, Gott and collaborators (Gott, Melott & Dickinson 1986; Melott 1990) because it is a *topological*, rather than *geometrical*, measure: a filament is topologically equivalent to a pancake.

tics (Kuhn & Uson 1982), minimal spanning trees (Barrow, Bhavsar & Sonoda 1985; Bhavsar & Ling 1988a,b), dimensionality and scaling (Jones et al. 1988; Martinez et al. 1990), quadrupole statistics (Fry 1985; Vishniac 1986), ridge-finding (Moody, Turner & Gott 1983) and structure functions (Babul & Starkman 1992); these, and other, methods are reviewed by Coles (1992).

In this paper we propose and test a method for measuring the geometrical properties of galaxy clustering which is, in fact, a 'hybrid' of two methods drawn from the above list. Our method involves constructing the Minimal Spanning Tree (MST; we define this below) around the set of points, separating and pruning the MST (these operations are defined below), and then computing structure functions (again defined below) of the resulting structure(s). Babul & Starkman (1992) have already discussed the use of structure functions in clustering studies, but they applied them directly to the point distribution of galaxies. Previous experience with geometrical descriptors, such as the Vishniac quadrupole statistic (Vishniac 1986), has demonstrated that they are liable to be confused if the density of points is low, or if there is an unclustered background of points around the filaments. Our proposed method allows one to select candidate structures from the background objectively, which in turn reduces the susceptibility of the structure functions to statistical 'noise'.

2 SPANNING TREES, MOMENTS OF INERTIA AND STRUCTURE FUNCTIONS

In this section, we outline the technical definitions of the various constructions we use. We shall give only brief outlines; the interested reader is referred to the references for more detailed descriptions. For the MST, see Gower & Ross (1969), Ore (1962) or Zahn (1971); structure functions are described in some detail by Babul & Starkman (1992).

2.1 The Minimal Spanning Tree (MST)

The MST is a construction borrowed from graph theory (Ore 1962). In the context of this study, our data set can be thought of as a set of points in three-dimensional space. We follow the definitions of Barrow et al. (1985). In general terms, a graph defined on this set will be a collection of *nodes* (galaxies) and *edges* (straight lines joining galaxies). The number of edges emerging from a node is called the *degree* of the node. A sequence of edges joining nodes is called a *path*; a part of a path that is closed is called a *circuit*, and if there is at least one path between any two nodes then the graph is *connected*. A connected graph containing no circuits is called a *tree*; if a tree contains all the nodes of the set, it is called a *spanning tree*. The MST of the set is the spanning tree that has the smallest total length, i.e. the minimal set of edge-lengths. If no two edge-lengths are equal, then the tree is unique.

There is a simple algorithm for constructing the MST of a set of points, which is sometimes called the *Greedy Algorithm* and which is described in some detail by Barrow et al. (1985).

It is possible to perform various operations on the tree, with the aim of eliminating extraneous features. The two most useful such operations are *pruning* and *separating*; they are both illustrated by Fig. 1. Define a k-branch to be a path of k edges connecting a node of degree 1 to a node of degree exceeding 2, in such a way that all intervening nodes are of degree 2. An MST is *pruned* to a level p if all k-branches with $k \leqslant p$ have been removed. A *separated* tree is obtained by removing any edges with length exceeding some cut-off, l. It is useful to define l in terms of the mean length of edges in

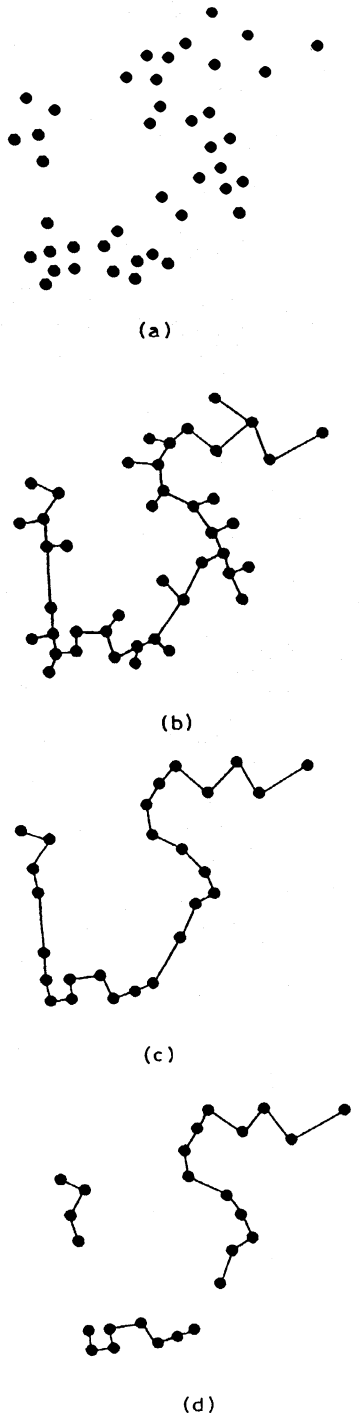

Figure 1. Illustration of the MST and various operations on it: (a) shows a schematic set of points in two dimensions; (b) shows the MST of the set; (c) shows an example of pruning - all nodes of degree 1 connected to nodes of degree 2 or more have been removed; (d) shows the effect of separation - all edges exceeding a certain critical length have been removed.

the MST, $\langle l \rangle$. For example, one might choose $l = 2\langle l \rangle$. We shall discuss optimal choices of p and l later. Pruning removes 'foliage' on the MST that is not essential to the main structure of the pattern; separation tends to disconnect 'accidental' linkages. The tree remaining after these operations may consist of several disjoint pieces, which can be analysed further. Pruning is most useful for the specific task of identifying filaments in the galaxy distribution (Barrow et al. 1985). We have in mind the less specific problem of identifying, and classifying, the dominant structures whether they be filamentary, sheet-like or even spherical. We do not therefore implement pruning on our MST constructions. On the other hand, separation is a valuable operation for eliminating the effect of 'chance' linkages and is therefore good for enhancing a weak structure superimposed upon an unclustered background.

The great advantage of the MST is that it seems to mimic the activity of the eye in emphasizing filaments embedded in background noise. Running the MST over a data set therefore produces a set of candidate connected structures, which should register with a high signal-to-noise ratio in any quantitative statistic. Although its use has been mainly confined to studies of possible filamentary structure, it can be used to connect points with any intrinsic geometry. Nevertheless, the MST is not itself a statistical descriptor: we still need a quantitative way to extract information from the tree. Various methods have been suggested, such as the frequency distribution of edge-lengths (Barrow et al. 1985; Barrow & Ling 1988a,b). We want to find a more geometrical way to describe the tree, so we have turned to quantities that have been previously applied only to point sets.

2.2 Moments of inertia and structure functions

Suppose we have a set of N points, located at position vectors $\boldsymbol{r}^{(k)}$, where $k = 1 \ldots N$. In this case, the set of points will be the set of nodes of each piece of a pruned and separated tree. One can define a local moment of inertia tensor about the centre of mass of each piece by

$$I_{ij} = M_{ij} - M_i M_j, \tag{1}$$

where

$$M_i = \frac{1}{N} \sum_{k \in \text{ MST}} r_i^{(k)}, \tag{2}$$

and

$$M_{ij} = \frac{1}{N} \sum_{k \in \text{ MST}} r_i^{(k)} r_j^{(k)}. \tag{3}$$

The sums are taken over all nodes in the tree; notice that each node is assigned equal weight in this procedure.

Following Babul & Starkman (1992), we define the parameters v and μ:

$$v = (I_2/I_1)^{1/2}; \tag{4}$$

$$\mu = (I_3/I_1)^{1/2}, \tag{5}$$

where I_1, I_2 and I_3 are the eigenvalues of the tensor I_{ij}, sorted into decreasing order of magnitude. Structure functions S_1, S_2 and S_3 are then defined as follows:

$$S_1 = \sin\left[\frac{\pi}{2}(1 - v)^p\right] \tag{6}$$

$$S_2 = \sin\left[\frac{\pi}{2} a(\mu, v)\right] \tag{7}$$

$$S_3 = \sin\left(\frac{\pi\mu}{2}\right). \tag{8}$$

In equations (6)–(8), we have $p = \log 3 / \log 1.5$, $a(\mu, v)$ is given by

$$\frac{v^2}{a^2} - \frac{\mu^2}{a^2(1 - \alpha a^{1/3} + \beta a^{2/3})} = 1, \tag{9}$$

and the parameters α and β are

$$\alpha = 1.9; \tag{10}$$

$$\beta = -\left(\frac{7}{8}\right) 9^{1/3} + \alpha 3^{1/3}. \tag{11}$$

These somewhat complicated definitions are designed to ensure the following properties: $0 \leqslant S_i \leqslant 1$ for all i; a perfect sphere has $(S_1, S_2, S_3) = (0, 0, 1)$; a flat sheet has $(S_1, S_2, S_3) = (0, 1, 0)$; a filament has $(S_1, S_2, S_3) = (1, 0, 0)$. The functions are constructed so as to fall rapidly away from unity as the structure moves away from the specific shape described by a particular S_i. Full details are given by Babul & Starkman (1992).

It is also interesting to compare these structure functions with another descriptor extracted from the moment of inertia tensor: the Vishniac quadrupole statistic (Fry 1985; Vishniac 1986). In terms of M_i and M_{ij}, this is

$$Q = \frac{2M_{ij}(M_{ij} - M_i M_j) - M_{ii}(M_{jj} - M_j M_j)}{(M_{ii})^2}, \tag{12}$$

where summation over all repeated indices is implied. This statistic is also designed to have specific properties: $Q = 0$ for a random field, and $Q = 1$ for a straight filament. It is, however, a rather cruder descriptor than the S_i; we make some remarks upon this in the Conclusions.

2.3 Our algorithm

The general-purpose method we propose for identifying relevant geometrical properties of a clustered pattern is simply to construct the MST, apply separation (with a variable separation length, l), and then look at statistics of the structure functions of each piece of the resulting tree. Of course, one could also look at other properties of the MST and some of these have indeed been looked at previously (Barrow et al. 1985). We feel, however, that the structure functions S_1, S_2 and S_3 constitute a good compromise between simplicity and sophistication: they do not seek to encode all the information about the distribution in a single number (we have three numbers for each disjoint piece of tree at each value of the separation length), but the amount of information extracted from a graph for a reasonable data set is quite manageable.

3 STOCHASTIC MODELS

In this section, we discuss the preliminary testing of our proposed algorithm using simple static Monte Carlo simulations of various types of galaxy clustering. These are not intended to be realistic simulations of galaxy clustering in any particular scenario; we merely seek an indication of the performance of the method on simple stochastic 'toy' models with known clustering geometry. We have a number of variables describing

the algorithm, the simulations and the results. We first have the separation length $f\bar{x}$, where f is a dimensionless parameter giving the separation length in units of $\bar{x}$, the mean edge length of the (unseparated) tree. For a given value of f, and a given data set, the resulting MST will fall into a number of disjoint pieces each containing a different number of nodes, N. For each of these pieces, and at each value of f, we have a set of three structure functions (S_1, S_2, S_3).

3.1 Poisson random field

The simplest conceivable clustering model is one in which there is no clustering at all, i.e. a random (Poisson) distribution of points. We have used simulations of such a distribution to get a feel for the behaviour of our algorithm in the presence of an unclustered background. As one might expect, the S_i depend very sensitively upon the number of nodes in the tree for trees obtained from a random distribution. For low values of N, S_1 comes out consistently around unity while S_2 and S_3 are around zero. Indeed, for N around 50, typical values are $S_1 \sim 0.75$, indicating a fairly strong filamentary pattern. As N increases, the typical value of S_1 falls in such a way that $S_1 < 0.2$ for $N > 150-200$. The value of S_2 remains uniformly low for all N. On the other hand, S_3 increases as N increases in such a way that, by the time N reaches 300, the value of S_3 is close to unity.

This is the behaviour we would expect for a random field. For low N, even a random pattern appears 'filamentary' because of the inherently one-dimensional nature of the connections making up the MST. Similar results are noted by Barrow et al. (1985). For larger N, the MST is large enough so that branches head off in all directions, forming a more spherical structure. The persistently low value of S_2 merely indicates that there is no intrinsic two-dimensional pattern in the simulations: the MST construction itself is one-dimensional on small scales, whereas the large-scale structure is homogeneously filling three-dimensional space. The method we are using does therefore behave sensibly on these, the most simple stochastic point processes.

3.2 Filaments embedded in a random field

The next model upon which we have tested our technique is known filamentary structures embedded within a random (Poisson) distribution of points. We choose a simple method for generating filaments. Inside the simulation box we choose two points at random. Points are then spread randomly along the line joining the two points. The number of points in the filament is simply proportional to its length. Filaments generated in this way therefore all have the same mean density of points along them.

We began by testing our algorithm on simulations containing just filaments. The results were as expected, with consistently high values of S_1 being obtained with S_2 and S_3 being low, for all reasonable values of the separation parameter f. This is not a very severe test, however, and realistic situations are better modelled by filaments superimposed on a random background. To model this, we ran a series of numerical experiments with random (Poisson) backgrounds of points with different random intensity (mean number-density). We give an example of the sort of results obtained in Fig. 2.

The figure shows the effect of varying the separation parameter upon S_1 and S_3 for filaments generated by the above algorithm in a fixed background random field of points. For small values of f, we see clearly one-dimensional trees ($S_1 \simeq 1$, $S_3 \simeq 0$), whereas for larger f the effectiveness of the algorithm deteriorates. The reason for this is that strong separation removes virtually all chance linkages between filament and background, while many of these remain if f is large. If one did not separate or prune the tree at all, it would appear very much like a spherical 'bush'. At high values of f, only small linear branches are cut off while the remaining long tree segments are clearly not filamentary in character. One therefore needs to separate the tree with quite a small f, of order unity or less, for our algorithm to work effectively in this situation.

3.3 Spherical clusters embedded in a random field

Now let us turn to simulations of isolated spherical clusters. Again, we adopt a very simple prescription for these simulations. Cluster centres are placed at random in the simulation box. We then select a radius for the cluster, and an overdensity. We then spread a uniform distribution of points inside the cluster radius. The number of points is chosen so that there is a known *overdensity* in the cluster compared to a random background of points with which, as with the filaments, we surround the cluster. The size and density of each cluster can be varied by increasing/decreasing the cluster radius or the number of points within the cluster. We have performed a large number of numerical experiments using this prescription.

The MST construction has been used in the past primarily to look for filaments, so we first needed to understand how our algorithm performs at picking out simple, spherical clusters in the absence of any background distribution. In fact, it does very well. We found, for any reasonable choice of simulation parameters, that $S_1 \simeq 0$ and $S_3 \simeq 1$. Note that *pruning* of the tree would be a very bad idea for this kind of structure, as the somewhat 'bushy' trees one obtains would fall into a large number of smaller pieces upon pruning. Separation is less of a problem in this respect, as long as the parameter f is kept within reasonable bounds. Obviously, our algorithm would be expected to work better for denser clusters, and this is indeed the case. At a fixed overdensity, and a fixed background random intensity, the algorithm prefers clusters of larger radius because these contain more points.

Generally speaking, however, the algorithm is less successful at identifying spherical objects than filaments. This is not surprising, given the intrinsically one-dimensional local structure of the MST. Fig. 3 shows an example which is for clusters with an overdensity of $\delta\rho/\rho = 1000$ and a number of points per cluster of around 800; there is not a clear signal with $S_1 \simeq 0$ and $S_3 \simeq 1$ for any length of tree. The reason for this is that, at this overdensity level and separation parameter, points outside the actual cluster have a significantly high probability of being contained in the MST, producing a somewhat elongated tree. Rather than adjusting f in an ad hoc fashion for each cluster, we can improve the performance by introducing another node deletion operation. It is advantageous in this situation to remove a node from the tree if the number-density of other nodes around the node is less than some specified threshold. The number-density is calculated by simply counting the number of nodes within some specified distance of the given node. This is not the same as simple separation,

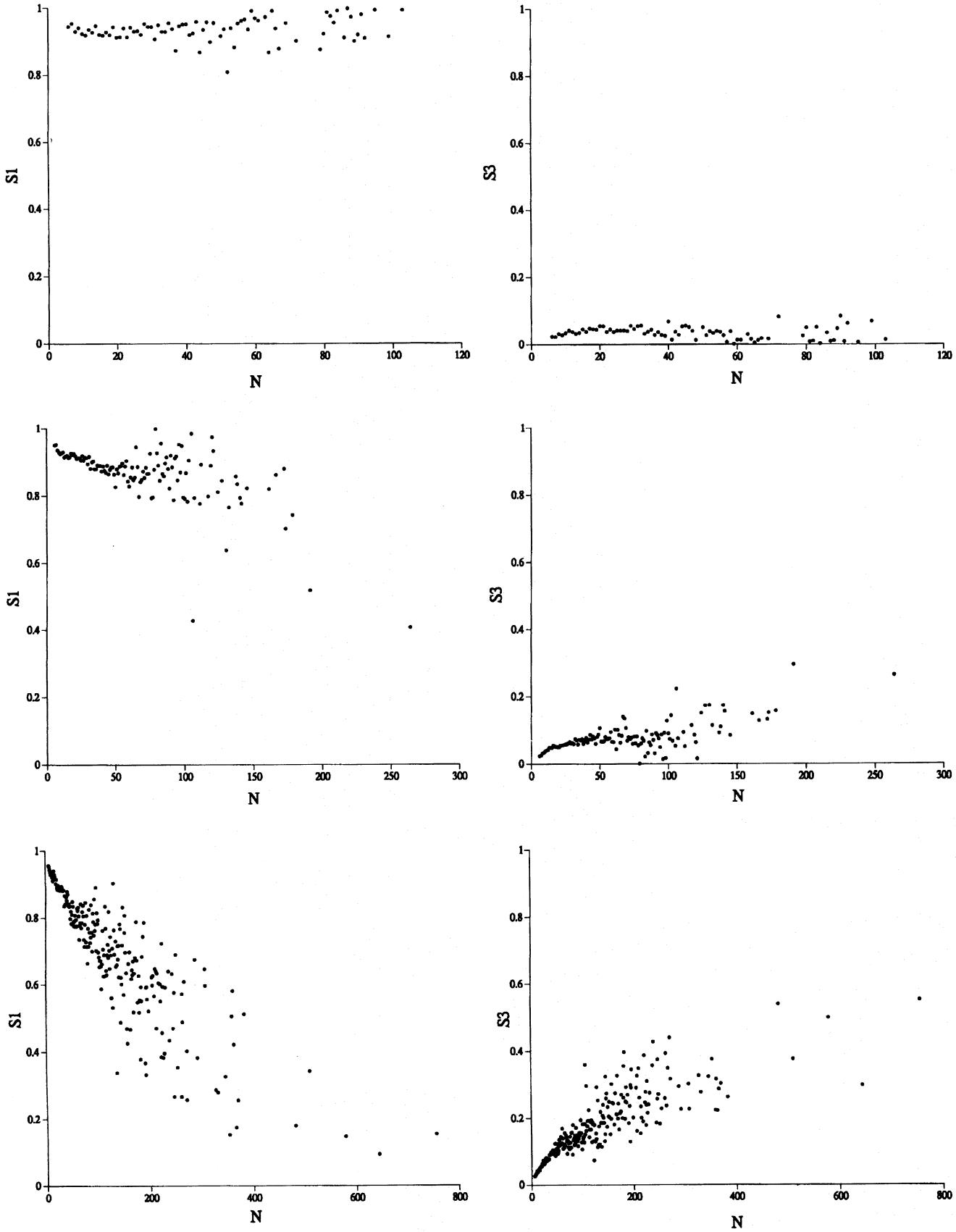

Figure 2. The effect of varying the separation parameter f upon the values of S_1 and S_3 for the random filament models described in the text; the graphs show results for $f = 0.75$ (top), 1.25 (middle) and 1.75 (bottom).

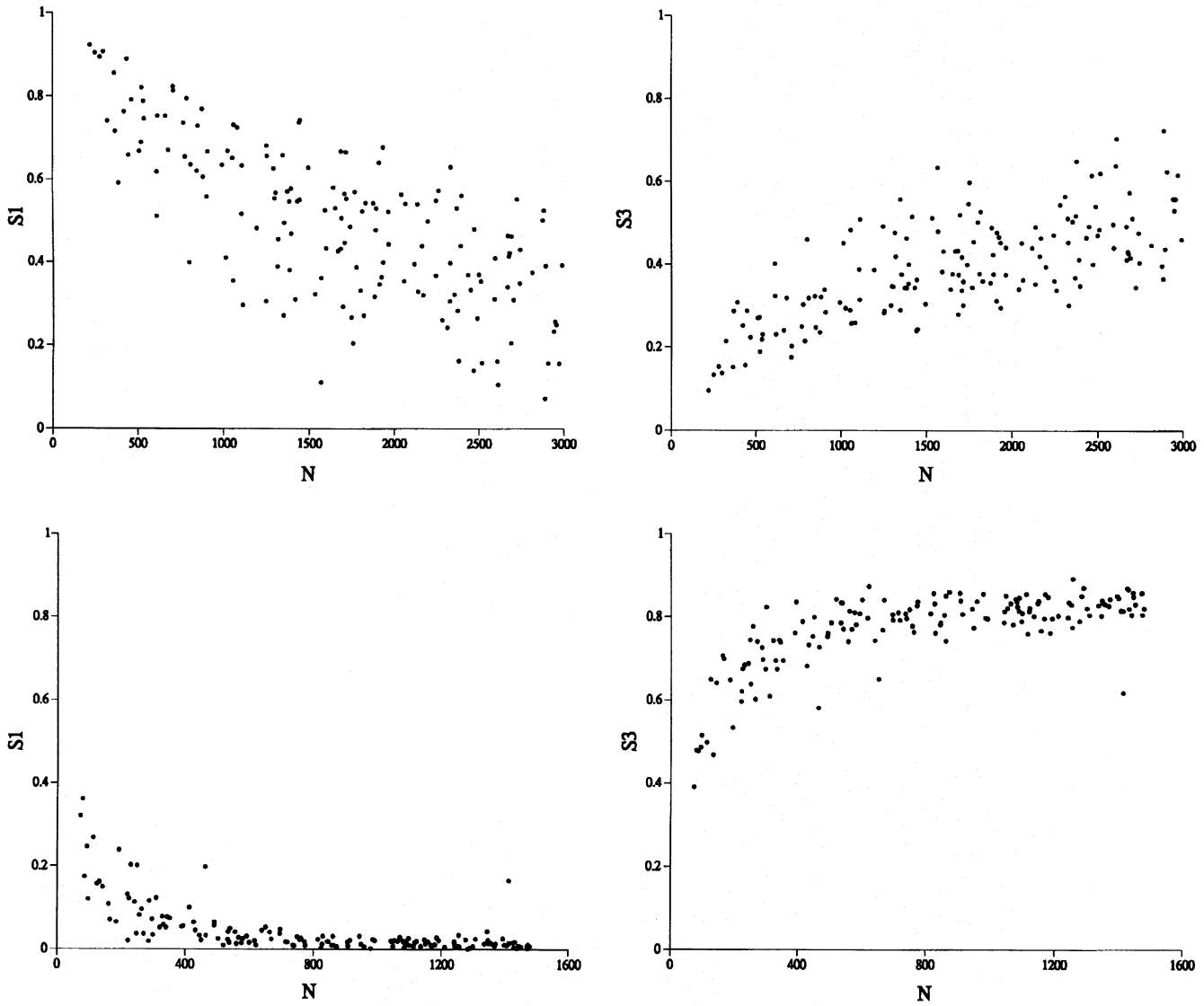

Figure 3. The behaviour of S_1 and S_3 for simulated spherical clusters, as described in the text. The upper pair of plots shows the simple version of the algorithm which does not display a definite detection of isolated spherical clusters (which would be $S_1 \simeq 0$ and $S_3 \simeq 1$). The addition of an extra objective node-deleting criterion in the lower pair of plots does, however, improve the performance of the algorithm in this direction.

but it does very effectively remove straggly branches from the edge of a spherical cluster. The bottom half of Fig. 3 shows what a dramatic improvement can be achieved by adopting this simple new deletion procedure. This shows that there is a very effective way of correcting for the one-dimensional 'bias' intrinsic to the MST which should be used if one requires sensitivity to three-dimensional clustering.

3.4 Redshift-space distortions

The clusters we generated in the previous section are constructed in real space. In most analyses of galaxy clustering we actually have a survey of galaxy positions in redshift space, rather than real space. The redshift of a galaxy will be due both to the Hubble expansion and to the radial component of the galaxy's peculiar motion. Since clusters will have relatively large velocity dispersions, they can appear significantly elongated along the line-of-sight even if they are spherical in real space: the well-known 'Fingers of God' effect. We therefore decided to run a series of simple experiments to see if redshift space distortions have any effect upon the structure functions. Again, the simulation technique is very simple. We take the clusters to be generated as before, but assign a random peculiar motion to each galaxy in the cluster. To get something approximating to reality, we adopt cluster parameters similar to those of the Coma cluster, i.e. a recession velocity of 5000 km s^{-1}, and a radius of around 7 Mpc. For realistic velocity dispersions, the effect upon S_1 and S_3 is not dramatic, although there is a clear trend for S_3 to decrease as one increases the radial velocity dispersion up to around 500 km s^{-1}. The redshift-space effect also makes it harder to pick out the cluster from the noise. In any case, this effect is potentially important so, when applying our method to the testing of a specific galaxy formation model against real data, one must be careful to use simulated data in redshift space.

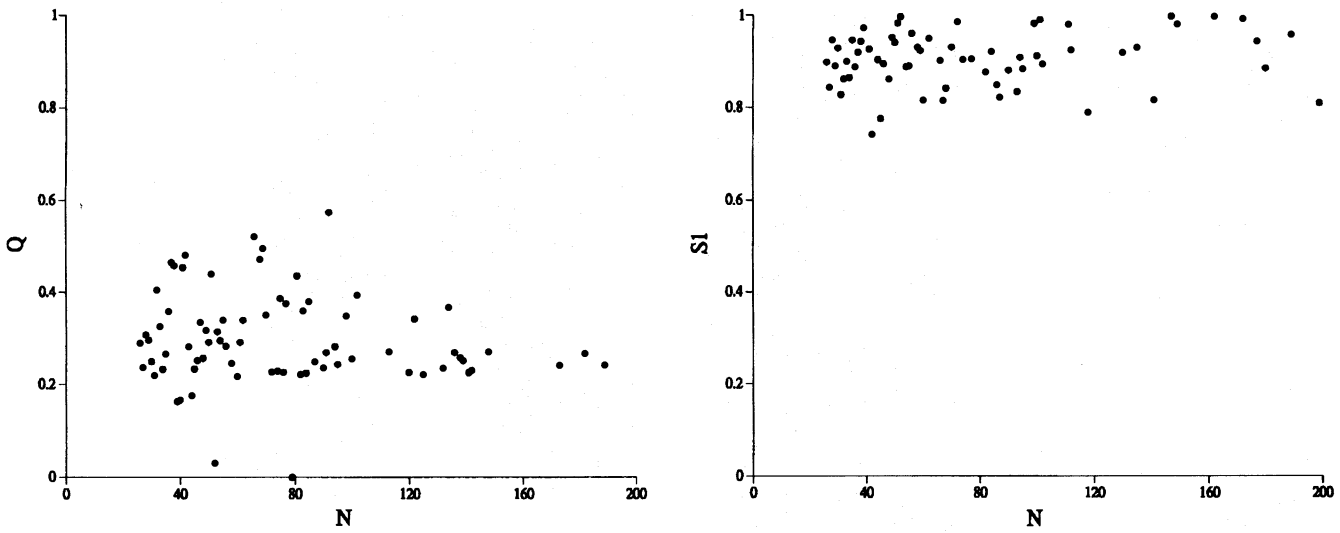

Figure 4. A comparison of the structure function S_1 and the Vishniac Q parameter for filamentary models. The former shows a much clearer detection of filaments ($S_1 \simeq 1$) than the latter ($Q \simeq 0.3$).

3.5 The Vishniac statistic

As we mentioned in Section 2.2, there are some similarities between the structure functions S_i and a parameter suggested by Vishniac (1986), given by equation (11). The Vishniac statistic Q has in the past only been applied to 'raw' galaxy counts, without the amplifying effect of the MST to pick out structures. It is worthwhile asking instead whether Q is better or worse than the S_i at quantifying the structure of the MST. The answer can be seen very easily from Fig. 4 where we compare the Vishniac statistic Q and S_1 for the same simulated filamentary data sets, and the same separation parameter $f = 0.75$. Notice that Q does not give an unambiguous detection of filaments (which would be $Q \simeq 1$): $Q \simeq 0.2 - 0.6$ is the best we get. On the other hand, $S_1 \simeq 0.7 - 1.0$ for the same data. The structure function S_1 is therefore *much* better at picking out filaments than is Vishniac's Q. This result is not a consequence of the particular simulation parameters we have used: different background intensities yield similar results.

4 MORE REALISTIC MODELS

Having evaluated the performance of our method on simple stochastic models with known clustering properties, we must now see whether it provides a useful statistical description of more realistic data sets. At the present time there is no clear front runner for a standard model of large-scale structure so, rather than try to explore the entire parameter space of clustering models, we decided to look at one particular set of simulations which we thought would pose a severe challenge to our method. Our plan is to study the geometry of superclustering, i.e. the structures traced out by clusters of galaxies, rather than galaxies themselves. This is expected to be difficult, for many reasons. First, the spatial number-density of clusters of galaxies is very low. This means that any filamentary or sheet-like structures will be sampled very sparsely, and shot-noise is likely to be a problem. Recall, however, that our technique of *separating* the MST is designed to cope with this problem. Secondly, filaments and sheets are thought to arise in the non-linear phase of the evolution of clustering and therefore might be expected to exist on only relatively small scales in hierarchical clustering scenarios. It is not clear, in these models, whether one expects a significant number of filaments a priori because the mass distribution on cluster scales may still be evolving linearly. Since we do not know whether we should see filaments, sheets or even quasi-spherical superclusters, we decided to look at all three structure functions S_i and merely ask the question: can we tell the difference between different cluster distributions using this diagnostic?

The simulations we use for this test were performed by Borgani, Coles & Moscardini (1994); they are generated in a box $640h^{-1}$ Mpc on a side, using the Zel'dovich approximation. We have selected a subset of these simulations: standard CDM (SCDM); open CDM (OCDM; with Ω_0=0.2); a mixture of hot and cold DM (CHDM). For further details, see Borgani et al. (1994).

Some measure of the difficulty of the task we have set can be judged by looking at Fig. 5, which shows a slice through each of the simulations together with a random distribution of points having the same number-density. As we have explained above, although the clusters are strongly correlated, the level of geometric pattern one can see is very slight. Can our algorithm capture enough of the pattern to provide a useful discriminant between these simulations?

We found for all these simulations that S_1 and S_2 have values very close to zero, indicating that there are indeed few filamentary or sheet-like structures in these data sets. At first sight, therefore, the results look very similar for the models and the random data set. Looking a little closer, however, we see some important differences. We ran our algorithm using different values of the separation cut-off f, and compared the MST properties with those of a random distribution. As we have explained, at each value of f, the MST consists of a number of disjoint pieces. In Fig. 6 we show the distribution of number of pieces, as a function of f. Note that, for large f, the MST consists of only one piece. One can see that this distribution does indeed carry information which can discriminate between the models and the random data: the

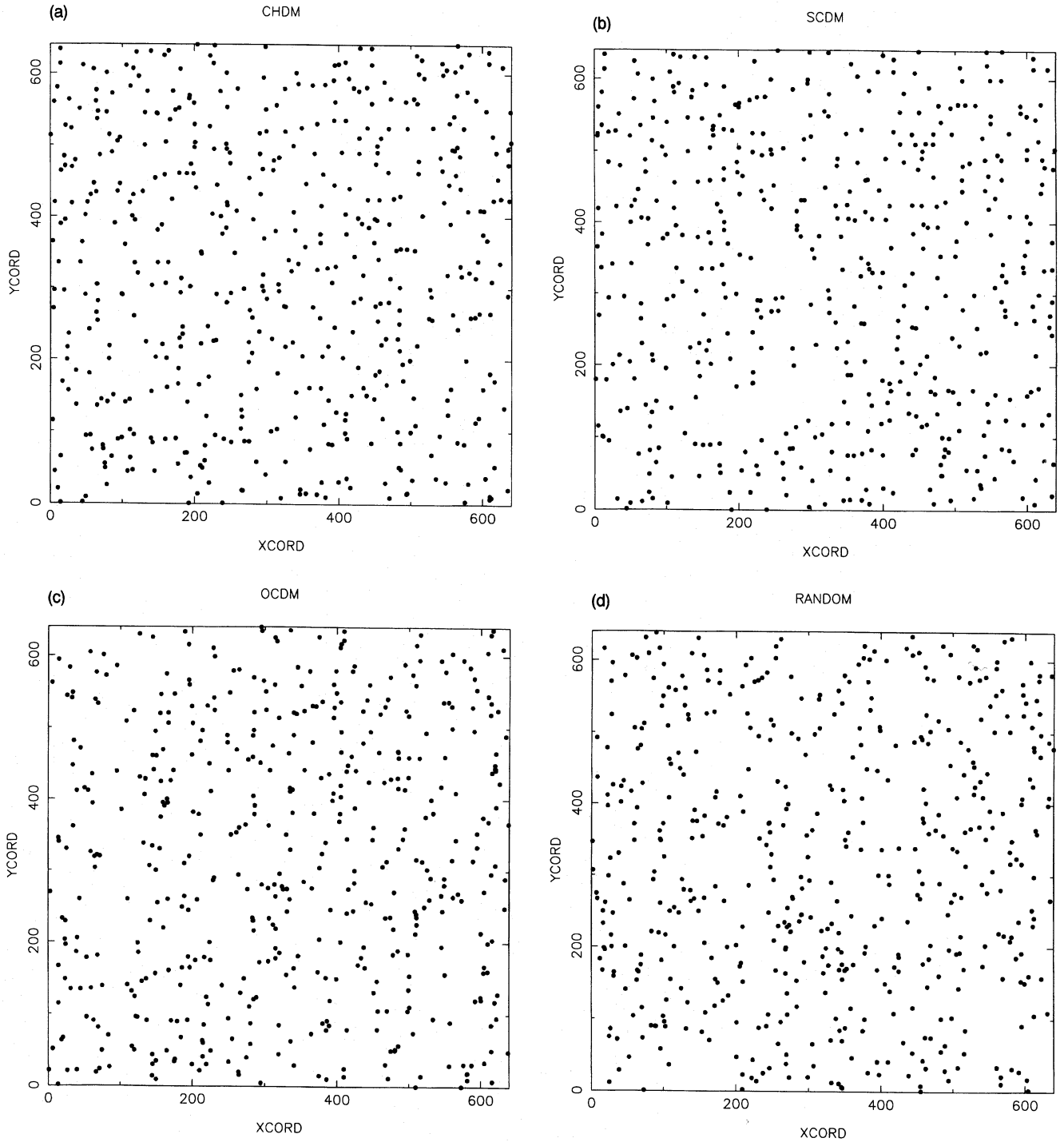

Figure 5. Simulated samples of rich clusters, as described in Borgani et al. (1994): (a) cold and hot dark matter (CHDM); (b) standard CDM (SCDM); (c) open CDM (OCDM); (d) a random (Poisson) set of points with the same number-density. All plots have a mean inter-cluster separation corresponding to that of APM clusters, i.e. $35h^{-1}$ Mpc. The boxes are $640h^{-1}$ Mpc on a side and are 'slices' through three-dimensional simulations.

CHDM and SCDM data show up as significantly different from the random set at the 95 per cent level (with χ^2 of 50.96 and 73.84 respectively, for $n = 4$). Somewhat surprisingly, the OCDM model is not significantly different from the random distribution ($\chi^2 = 1.83$ for $n = 4$). This is surprising because the OCDM model has a much longer correlation length than CHDM or SCDM and might therefore be expected to be less like the random distribution than the other two. Remember, however, that the MST construction is quite independent of the correlation function, and this expectation is therefore hard to justify with any rigour.

Now we turn to the structure functions. In a similar vein to the above calculations, we have calculated the structure functions of each of the pieces of the MST. The figure we

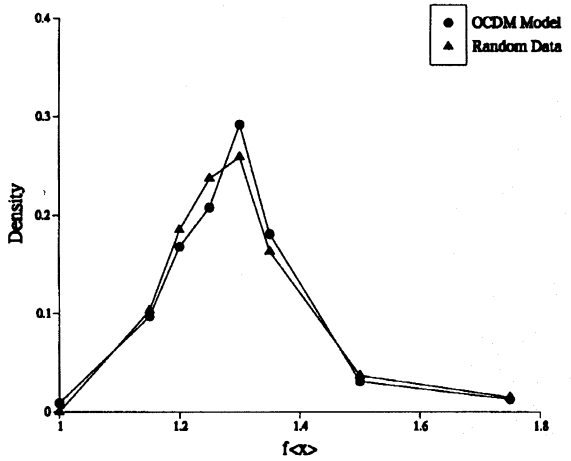

Figure 6. The number-density of trees as a function of f for the OCDM, CHDM and SCDM models. Corresponding results for a random simulation are also plotted for comparison.

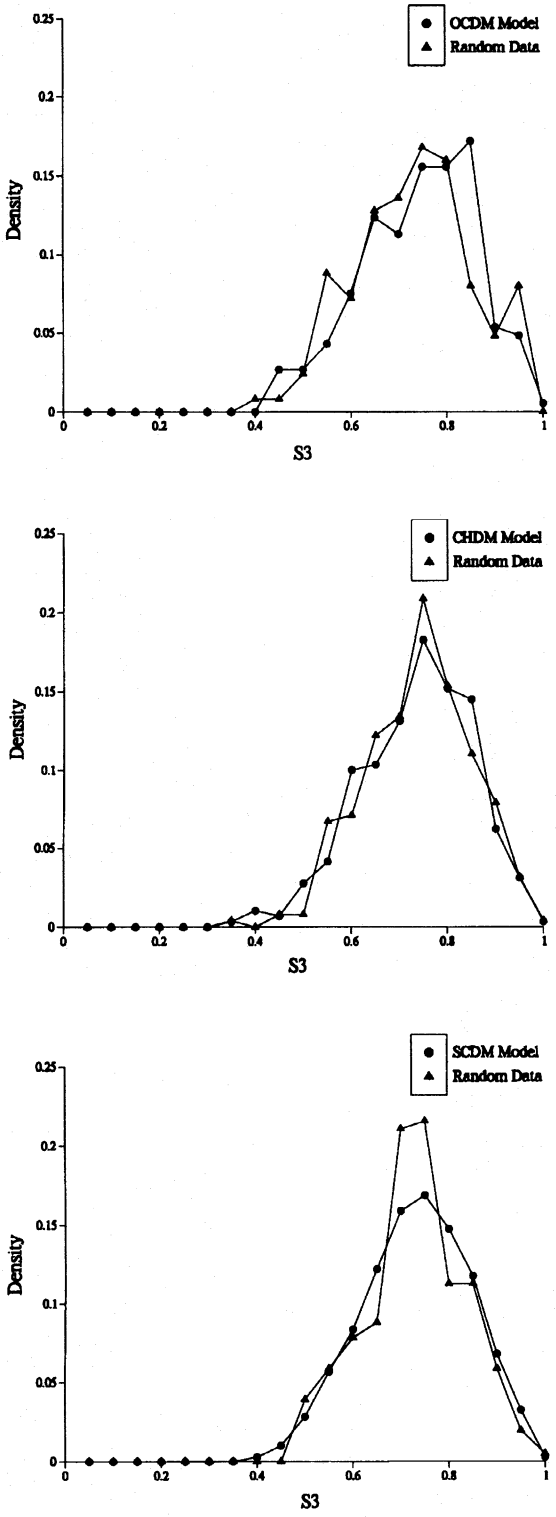

Figure 7. Distribution of S_3 values (integrated over f) corresponding to Fig. 6. Results for a random simulation are plotted for comparison.

have shown is actually summed over all values f shown in Fig. 6, in order to project out the f-dependence. As we explained above, we concentrate upon S_3, since S_1 and S_2 do not seem to carry useful information for these data sets. The distributions of S_3 are shown in Fig. 7; the OCDM and SCDM data are significantly different from the random set at the 95 per cent level (with χ^2 of 29.3 and 41.3 for 11 and 10 degrees of freedom, respectively); the CHDM is consistent ($\chi^2 = 12.3$ for 9 degrees of freedom). Notice that the OCDM model is significantly non-random, when seen in terms of this description.

Only one of the structure functions, S_3, appears powerful in the analysis of this kind of data set, but it is reasonable to expect that the distributions of S_2 and S_1 would provide useful information for data sets with a stronger degree of intrinsic filamentary or sheet-like pattern, as we have shown above. In conjunction with the distribution of sizes of MST segments as a function of f, these parameters seem to provide a useful potential discriminant. Notice also that our method is clearly providing information that is in some sense orthogonal to the correlation functions: there is no simple relationship between the results we have obtained, and the properties of the two-point correlation functions of these models discovered by Borgani et al. (1994).

5 CONCLUSIONS

We have suggested a new way to analyse the geometrical properties of galaxy clustering data. The method involves constructing the MST of the data set, separating the tree with a variable cut-off parameter f, and then looking at the structure functions S_i of each disjoint piece of MST. We suggest that pruning should not be adopted in this approach, so as not to prejudice the algorithm against non-filamentary structures such as sheets or blobs.

We have shown that the combination of the MST with structure functions is very adept at picking our geometrical structures from spatial point patterns, even in the presence of an unclustered background component. Both the MST and the structure functions have been suggested before in the literature, but the combination of the two is a new idea; it seems most promising. We have also tried to apply a statistic suggested by Vishniac (1986) in conjunction with the MST, and found it less successful.

Because we did not only want to look at any particular geometry of clustering (i.e. filaments or sheets), we have spurned the idea of 'pruning' the MST. Our method can pick out isolated filaments, sheets or simple isolated blobs. The structure function then reveals the geometry. On the other hand, separation of the MST is a good way to enhance the clustering pattern by picking out connected structures from an unclustered background. By looking at all different values of f, the separation parameter, we avoid having to use an ad hoc value. Indeed, the behaviour of MST properties as a function of f yields more information than one would obtain by fixing f a priori. The MST does have a kind of 'bias' towards one-dimensional or filamentary geometries simply because of the way it is constructed. This is reflected by the fact that our algorithm is very good at picking out filaments from noise, but its success at picking out spherical clusters is more mixed. Nevertheless, the algorithm extracts information about a generic clustering pattern that can be used to compare data sets with theoretical predictions.

In a general situation, we suggest the best approach is to construct the MST, separate with a variable f, then construct the distribution functions of the number of MST fragments as a function of f and the distribution functions of S_3 values. This avoids trying to encode all the geometrical properties of clustering in a very small number of parameters, but reduces the amount of data to handle by a substantial amount. Obviously, there are other ways of using the information present in the MST: pruning the tree, and looking at the behaviour as a function of the pruning parameter, is one example.

Our method reveals differences between models of cluster clustering, even when the structures seem to be so sparsely sampled that little intrinsic gemometry can be seen visually. We expect the algorithm to work even better on more obviously 'filamentary' or 'sheet-like' structures, or with data sets having a higher density of points (such as a bright galaxy survey, rather than a rich cluster survey). This has yet to be proven, however. We shall return to this in a more comprehensive study of this and other ways of analysing galaxy clustering on relatively small scales. We shall also be applying this method, and others, to a larger ensemble of simulations of superclustering and to data on superclustering scales, with the aim of seeing which of the available models (if any) fits *all* the cluster clustering data. In the meantime, however, we recommend this method as a potentially valuable addition to the 'toolkit' of methods available for the analysis of clustering in point patterns.

ACKNOWLEDGMENTS

RCP receives a PPARC postgraduate studentship; PC is a PPARC Advanced Research Fellow. Both are grateful to Stefano Borgani and Lauro Moscardini for permission to use the cluster simulations from Borgani et al. (1994) in this work.

REFERENCES

Babul A., Starkman G.D., 1992, ApJ, 401, 28
Barrow J.D., Bhavsar S.P., 1987, QJRAS, 28, 109
Barrow J.D., Bhavsar S.P., Sonoda D.H., 1985, MNRAS, 216, 17
Bhavsar S.P., Ling E.N., 1988a, ApJ, 331, L63
Bhavsar S.P., Ling E.N., 1988b, PASP, 100, 1314
Borgani S., Coles P., Moscardini L., 1994, MNRAS, 271, 223
Coles P., 1992, in Feigelson E.D., Babu G.J., eds, Statistical Challenges in Modern Astronomy. Springer, New York, pp. 57-81
de Lapparent V., Geller M.J., Huchra J.P., 1986, ApJ, 302, L1
de Lapparent V., Geller M.J., Huchra J.P., 1989, ApJ, 343, 1
Fry J.N., 1985, ApJ, 289, 10
Gott J.R., Melott A.L., Dickinson M., 1986, ApJ, 306, 341
Gower J.C., Ross G.J.S., 1969, Appl. Stat., 18, 54
Jôeveer M., Einasto J., Tago E., 1978, MNRAS, 185, 357
Jones B.J.T., Martinez V.J., Saar E., Einasto J., 1988, ApJ, 332, L1
Kuhn J.R., Uson J.M., 1982, ApJ, 263, L47
Martinez V.J., Jones B.J.T., Dominguez-Tenreiro R., van de Weygaert R., 1990, ApJ, 357, 50
Melott A.L., 1990, Phys. Rep., 193, 1
Moody J.E., Turner E., Gott J.R., 1983, ApJ, 273, 16
Ore O., 1962, Am. Math. Soc. Colloq. Publ., 38
Peebles P.J.E., 1980, The Large-scale Structure of the Universe. Princeton Univ. Press, Princeton
Sahni V., Coles P., 1994, Phys. Rep., in press
Shandarin S.F., Zel'dovich Ya.B., 1989, Rev. Mod. Phys., 61, 185
Vishniac E.T., 1986, in Kolb E.W., Turner M.S., Lindley D., Olive K., Seckel D., eds, Inner Space/ Outer Space. University of Chicago Press, Chicago, pp. 190-193
Wegner G., Thorstensen J.R., Kurtz M.J., Geller M.J., Huchra J.P., 1990, AJ, 100, 1405
Zahn C.T., 1971, IEEE Trans. Comput., C20, 68
Zel'dovich Ya.B., Einasto J., Shandarin S.F., 1982, Nat, 300, 407

This paper has been produced using the Royal Astronomical Society/Blackwell Science LaTeX style file.