Published in "Errors, Bias and Uncertainties in Astronomy", Cambridge University Press, 1990.
Out of print - February 1998

BOOTSTRAP, DATA PERMUTING AND EXTREME VALUE DISTRIBUTIONS: GETTING THE MOST OUT OF SMALL SAMPLES

Suketu P. Bhavsar

Department of Physics and Astronomy, University of Kentucky, Lexington KY 40506-0055, U.S.A.

Abstract. The use of statistical methods on samples of astronomical data are discussed. We discuss the application of bootstrap to estimate the errors in measuring the two-point correlation function for galaxies and the topology of the large-scale structure. We discuss a technique to estimate the statistical significance of filamentary structures in the distribution of galaxies. Extreme value statistical theory is applied to the problem of the nature of the brightest galaxies in rich clusters.

Table of Contents

Next

Next Contents

1. INTRODUCTION

What can a statistician, feet firmly planted on the ground, offer an astronomer whose head is up in the stars? The answer of course is: "quite a lot", which is why we are at this meeting. I shall talk about my own experience, giving an account of how statistics has enriched my astronomical research. Besides formalizing errors and making estimates of uncertainties more rigorous, it has actually opened up new ways to do analysis, providing new insights and approaches to problems which previously had proved elusive.

Small samples are a common occurrence in the field of Astronomy. Data has to be gathered from the laboratory of the universe, over which one has no control. Astronomers have to take what is given, and very often that is very little. In this context, I shall discuss three separate topics which I have been involved in, where statistics has played a major role in furthering our understanding. Please note that this in no way is an attempt at a review, or is representative, of the many uses of statistics in astronomy. It is a description of my personal involvement and excitement at the realization of how much the rich field of statistics has to offer to astronomy. The three applications that I shall talk about here are:

1. The application of bootstrap to estimate the standard errors in measuring the galaxy-galaxy correlation function, and other measures of galaxy clustering.

2. A technique developed to estimate whether large-scale filamentary structures in the universe are statistically significant.

3. An application at extreme value statistical theory to the understanding of the nature of the brightest galaxies in rich clusters.

Next Contents

Next Contents Previous

2. BOOTSRAP

The bootstrap was invented by Efron in 1977 (e.g., see Efron 1979). The method uses the enhanced computing power available in this age to explore statistical properties that are beyond the reach of mathematical analysis. At the heart of this method is the procedure of creating from the available data set, hundreds or thousands of other "pseudo-data" sets. In older methods one assumed or derived some parametric distribution for the data and derived statistical measures. With bootstrap, any property of the original data set can be calculated for all the pseudo-data sets and the resulting variance of a measured quantity calculated over the entire ensemble. We have applied the bootstrap resampling technique (to my knowledge for the first time in astronomy) to illustrate its use in estimating the uncertainties involved in determining the galaxy-galaxy correlation function (Barrow, Bhavsar and Sonoda 1984).

The two-point correlation for galaxies is the factor by which the number of observed galaxy-galaxy pairs exceed the number expected in a homogeneous random distribution, as a function of their separation. It is an average taken over the entire sample and measures the clumping of galaxies in the universe. Usually a galaxy's coordinates in the sky are relatively easy to measure and known accurately. The same cannot be said for a galaxy's distance from us which is a difficult and time consuming measurement. Thus the most common catalogs available are sky surveys, for which the two-point correlation function has been extensively determined; though in the last few years 3-D spatial catalogs of limited extent have become available.

With a sky catalog one measures the angular two-point correlation function. This was pioneered by Totsuji and Kihara (1969), and Peebles (1973) and followed up extensively by the work of Peebles (1980) and his collaborators. They found that the angular correlation function for galaxies is a power law. If the projected distribution has a power law correlation then the spatial two-point correlation is also a power law with a slope that is steeper by -1 than the angular correlation (Limber 1953). The angular two-point correlation for galaxies to magnitude limit 14.0 in the Zwicky catalog (Zwicky et al. 1961-68) is given by

$$\omega(\theta) = \left\langle \frac{0.06}{\theta} \right\rangle^{0.77} \qquad (1)$$

Let me describe in a little more detail the actual determination of the above measurement. The sky positions (galactic longitude and latitude) of all galaxies brighter Than apparent magnitude 14.0 and in the region δ > 0° and b'' > 40° forms our sample. This consists of 1091 galaxies in a region of 1.83 steradian. The angular separation between all pairs of observed galaxies is found and binned to give $N_{oo}(\theta)$, the number of pairs observed to be separated by θ. This is compared to the expectation for the number of pairs that would be obtained if the distribution was homogeneous, by generating a Poisson sample of

1090 galaxies in the region and determining the pair-wise separations between each of the observed 1091 real galaxies and the 1090 Poisson galaxies. This gives us $N_{\rm op}(\theta)$, the number of pairs between observed and Poisson galaxies separated by θ. The angular two-point correlation function is

$$\omega(\theta) = \frac{N_{\rm oo}(\theta) - N_{\rm op}(\theta)}{N_{\rm op}(\theta)} \tag{2}$$

The binning procedure is decided upon before performing the analysis. The data is binned so that there are roughly an equal number of observed-Poisson pairs in each bin. A plot of log $\omega(\theta)$ versus log θ shows the correlation function to be a power law. The slope of the best fitting straight line determined by some well defined statistical measure gives the value of the exponent in equation (1). Our independent determination of this exponent gives a value of 0.77, consistent with earlier determinations. The question is: from the limited data available, how accurately does this describe the correlation function for galaxies in the universe? It is assumed that our sample is a "fair sample" of the universe (Peebles 1980), implying our faith in the assumption that the region of the universe that we sample is not perverse in some way. The *formal errors* on the power law fit to the data do not give us the true *variability* in the slope of the correlation function that may be expected. These formal errors indicate to us a goodness of fit for the best power law that we have found to fit the data; but the question of the statistical accuracy of the value 0.77 representing the slope of the power law for the two-point correlation function for galaxies in the universe, remains uncertain.

The bootstrap method is a means of estimating the statistical accuracy of the two-point correlation from the single sample of data. A pseudo data set is generated as follows. Each of the 1091 galaxies in the original data is given a label from 1 to 1091. A random number generator picks an integer from 1 to 1091, and includes that galaxy in the pseudo data sample. A galaxy is picked from the original data in this manner through 1091 loops. The pseudo data sample will contain duplicates of some galaxies and will not contain all the galaxies of the original data set. The angular correlation function can be calculated for this new data set in exactly the same manner as the original by fitting it to a power law of the form

$$\omega(\theta) = \left(\frac{\theta_0}{\theta}\right)^{\gamma} . \tag{3}$$

This entire procedure of generating a new data set and determining its angular correlation can be repeated hundreds of times using different sets of random numbers. The samples generated in this way are called bootstrap samples. The frequency distribution for the values of the slope, γ, and the correlation length, θ_0, can be plotted for the ensemble of bootstrap samples to estimate the variance of γ, and θ_0 respectively.

We generated 180 bootstrap samples for the 1091 galaxies in the Zwicky data, determined the two-point correlation function for each as described by equation (2), and fit the power law described by equation

(3) to each correlation function. Figure 1 shows the distribution of γ obtained for these 180 samples. Figure 2 shows the distribution of θ_0.

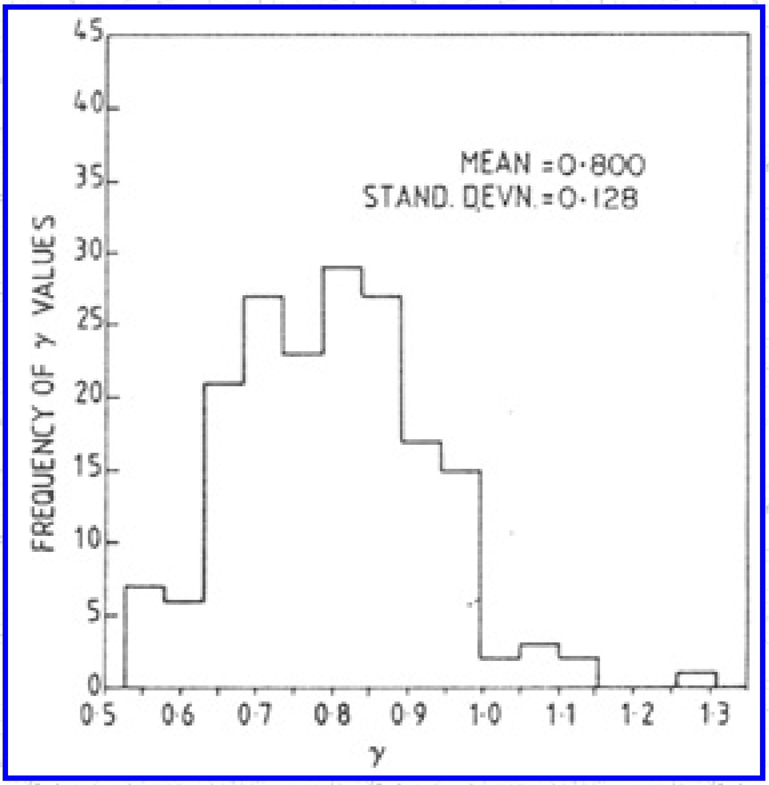

Figure 1.

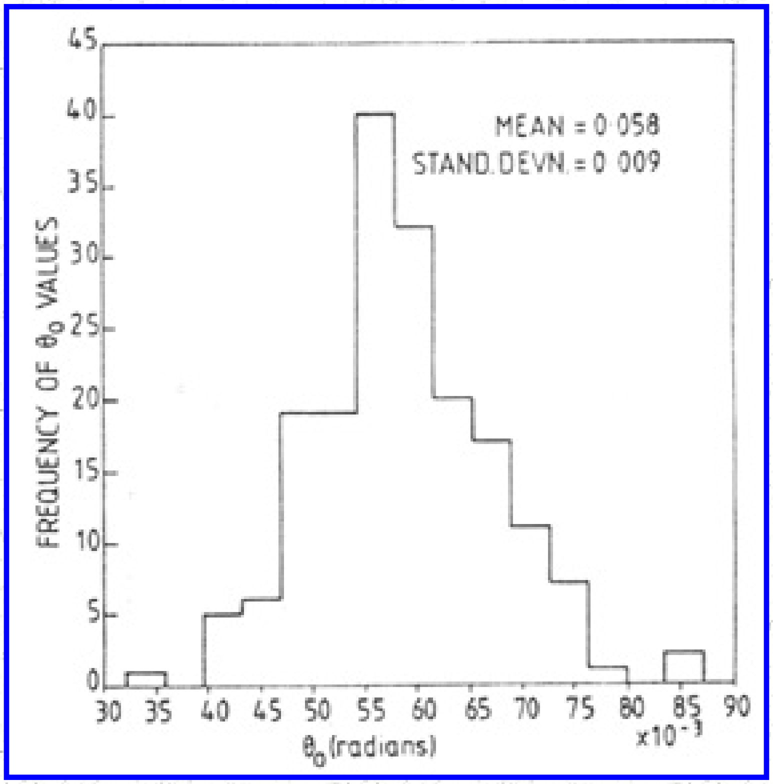

Figure 2.

We can determine the standard deviation in the values of γ and θ_0. These are found to be

$$\sigma(\gamma) = 0.13 \tag{4}$$

$$\sigma(\theta_0) = 0.01 \tag{5}$$

This means that 68 percent of the values of γ lie in an interval whose width is 0.26, or in other words, the bootstrap samples show that 68% of the γ lie in an interval 0.67 to 0.93, and the rest outside this interval, on either side. Similarly 68% of the values of θ_0 lie in an interval 0.05 to 0.07. Both these intervals are much larger than the formal errors that have been assigned to the quantities γ and θ_0 in the literature. It is worth noting that this method does not provide a best estimate for the value of γ or θ_0, but provides a statistical accuracy for these values. The *average* of the bootstrap samples is not an indication of the *true* value any more than the value obtained from the particular set of data is.

In 1915 Sir Ronald Fisher calculated the variance of statistically determined quantities (such as the slope γ here) assuming that the data points on the graph for the two variables were drawn at random from a normal probability distribution. In his time it would have been unthinkable to bootstrap because computation was millions of times slower and expensive. Today's computing power enables us to glean information from the available data without making assumptions about its distribution.

Next Contents Previous

Next Contents Previous

3. THE STATISTICAL SIGNIFICANCE OF LARGE-SCALE STRUCTURE

3.1. Filaments

How can we quantify the presence of some visually prominent feature in the universe? In particular how "real" are the linear or "filamentary" distributions of galaxies suggested by recent surveys? Their implications for evolution of structure in the universe remains controversial because it is extremely difficult to assess their statistical significance. Though the visual descriptions have been rich in describing form, they are subjective, leading to attributes too vague to model or compare with numerical simulations (Barrow and Bhavsar 1987).

Recently my collaborators and I pioneered the use of a pattern recognition technique - the minimal spanning tree; MST - to identify the existence of filaments (Barrow, Bhavsar and Sonoda 1985 [BBS]) and invented a statistical technique to establish their significance (Bhavsar and Ling 1988a [BL I]; 1988b [BL II]). Though apparent to the eye, this has been the first objective, statistical and quantitative demonstration of their presence (see Nature 334:647).

The famous Lick Observatory galaxy counts (Shane and Wirtanen 1967) are a good example of the problem. The wispy appearance of this sky map (Groth et al. 1977) has evoked a strong image of the universe permeated by a galactic lace-work of knots and filaments. Yet doubts about these large-scale filamentary features remain, and for good reason. The human propensity toward finding patterns in data sets at very low levels has been responsible for some embarrassing astronomical blunders: the most notable example being the saga of the Martian canals [1] (Evans and Maunder 1903; Lowell 1906; Hoyt 1976). To test the validity of controversial visual features an understanding of human visual bias and methods to overcome it are necessary.

There have been several quantitative descriptions of galaxy clustering. Most notably the correlation function approach (Peebles 1980) has been path-breaking in the early studies of structure in the universe. The practical difficulty of measuring high order correlations, however, makes this approach unable to provide information on the profusion of specific shapes of various sizes suggested by recent data. Clustering algorithms and percolation methods have not been very successful at discriminating visual features because they each use only one parameter (respectively, the density enhancement and the percolation radius). This leads to an over simplification of the problem of recognizing patterns. In particular, the difference in visual appearance between the CfA survey and earlier simulations, that were evident to the eye, were only weakly reflected in their percolation properties (Bhavsar and Barrow 1983; 1984). Two essential requirements needed to quantify any particular feature that may be noticed in the data are; first, an objective, well defined, and repeatable method to identify this feature of interest, and second, a means of evaluating the statistical significance of this candidate object.

The Minimal Spanning Tree or MST (Zahn 1971) is a remarkably successful filament tracing algorithm that identifies filaments in galaxy maps (BBS; Ung 1987; BL I). The technique, derived from graph theory, constructs *N* - 1 straight lines (edges) to connect the *N* points of a distribution so as to minimize the sum total length of the edges. This network of edges uniquely connects all (spans) *N* data paints, in a minimal way, without forming any closed circuits (a tree). In order to distil the dominant features of the MST from the noise, the operation of "pruning" is performed. A tree is pruned to level p when all branches with k galaxies, where $k \leq p$, have been removed. Pruning effectively removes noise and keeps prominent features. A further operation, "separating" removes edges whose length exceeds some cut-off, l_c. This gets rid of the unphysical connections in the network, just as the eye distinguishes well separated features. A pruned and separated MST is called a "reduced MST". Figure 3 illustrates the construction of the MST and the processes of pruning and separating on a simple paint set. The reader should refer to BBS for more details.

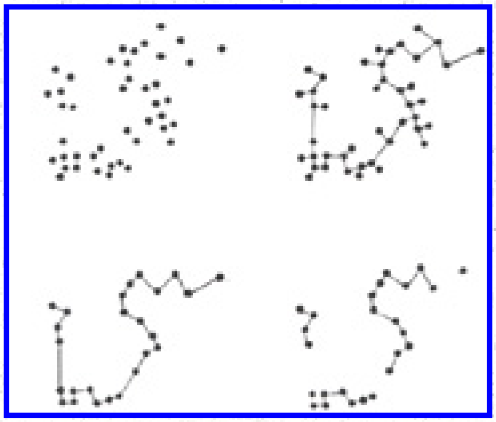

Figure 3.

The reduced MST identifies basic linear patterns in galaxy surveys and can measure them in a quantitative fashion. It is quite a good representation of how the eye-brain forms a global impression of most data sets of random or clustered point distributions. For example the MST does a fairly good job of picking out the constellations from a catalog of bright stars (Hillenbrand 1989; Hillenbrand, Bhavsar and Gott 1989).

How can we determine if the filaments that are identified are statistically significant or not? Are they unique to the distribution of galaxies or does strong small-scale clumping coupled with occasional chance alignments produce the false impression of large-scale linear structure? For this we need to know what kind of filamentary structures occur just by random chance in similarly clumped distributions. One way to do this is to use bootstrap to generate pseudo-samples of data and test them for features. We shall describe the determination of the robustness of another measure, the topology of large-scale clustering, using bootstrap samples later. For the problem of filaments, a nagging doubt is that the small scale dumping is producing the strong *visual* impression of filamentary structure. We can test this by applying the MST to samples that have the same small scale clumping as the data, but no large scale correlations.

The data permuting technique (BL I, BL II) described below achieves this.

Suppose that our eyes (and the eyes of our colleagues) detect features that are of the order of 30 Mpc across. We choose p and l_c such that the reduced MST also identifies these features. If small-scale clustering merely conspires to form large-scale features like these in a statistical way, then rearranging the galaxy distribution on the larger scales (say greater than 10 Mpc for instance), but leaving it intact on the smaller scales (in this case 10 Mpc and smaller), should have no overall effect. The original 30 Mpc features may disassociate but other similar features will come into prominence, and be identified by the reduced MST using exactly the same criteria as before. On the other hand the continued absence of such features after repeated rearrangements of the data would suggest that the original 30 Mpc features were unique to the data and not due to chance illusions.

In practice, this randomizing operation is easily performed by taking a cube of data, dividing it into smaller cubes (cubettes) and randomly shuffling the cubettes within the original volume. We have written a computer program which, given an integer n, divides a data cube of side *L* Mpc into n^3 cubettes of side L/n Mpc, randomly translates and rotates each cubette and reassembles them into a new cube of side *L* Mpc (analogous to manipulating a Rubik's cube, including the inside). Many different random realizations are obtained for each value of *n*, and *n* is varied from $n = 2$ to $n = n_{\max}$, where $n_{\max}$ is determined by the correlation length or some analogous measure. Note that the clustering within any cubette, of length L/n, is unchanged. All the MST techniques can now be used on this "fake" data cube to identify filaments and compare them to the original. The same procedure can also be applied in 2-D to an appropriately chosen shape extracted from 2-D data. The use of the MST ensures that exactly the same criteria will be applied to identify filamentary structures every time. If filamentary features are identified at the same level in these permuted versions of the data as they were in the original data, then we conclude that the original features are spurious, the result of statistical chance. Otherwise, the expected number of spurious filaments and their variance can be obtained for these permuted versions, at different length scales, giving us the statistical significance of the original features.

One might ask at this point: What determines a filament? Does it depend arbitrarily on the choice of l_c and p? Actually our procedure provides a working definition of a filament. Filaments can be identified at all levels, depending on the values of the parameters we choose to reduce the MST. If a tree is not pruned enough the small filaments persist even after the permutations. For some minimum choice of the pruning level the features are unique to the original distribution, showing the statistical presence of linear structures, and providing a measure of what to call a filament. Our experiments show that for some distributions this point is never reached because there are no filaments present. Remember, our motivation is to objectively define a visual feature, then check for its statistical significance. That a linear feature identified at a particular p and l_c is unique to the data, in fact defines for us "a filament."

Samples to be analyzed have to be chosen with care. For the above procedure to be of value it is imperative that the data cube be a complete sample and free from obscuration. Since this is stressed and elaborated on in detail in BL I we shall not say any more here. We used the CfA redshift catalog (Huchra et al 1983), in the region $\delta > 0°$ and $b'' > 40°$, with complete red-shifts down to apparent magnitude 14.5

and corrected for Virgo in-fall. We volume-limited this truncated cone by applying an absolute magnitude cutoff at -18.5 mag. Inside this volume-limited truncated cone, which contains 489 galaxies, we inscribed the largest possible cube, measuring 41.6 Mpc (H_0 = 50 km/sec/Mpc) on each side. The data in this cube is complete and practically tree of galactic obscuration. For details of the positioning of the cube and its dimensions the reader should refer to BL I. Figure 4a shows the data cube described above, its reduced MST and the end-to-end lengths and total lengths of the identified filaments. This cube was permuted/reassembled many times for each value of *n*, from *n* = 2 to *n* = 10. *We did not find filaments at the level of the original data in any of these realizations*. In fact they are present at a significantly lower level. Figure 4b shows one of these realizations for *n* = 5. Again, for more details and figures, refer to BL I.

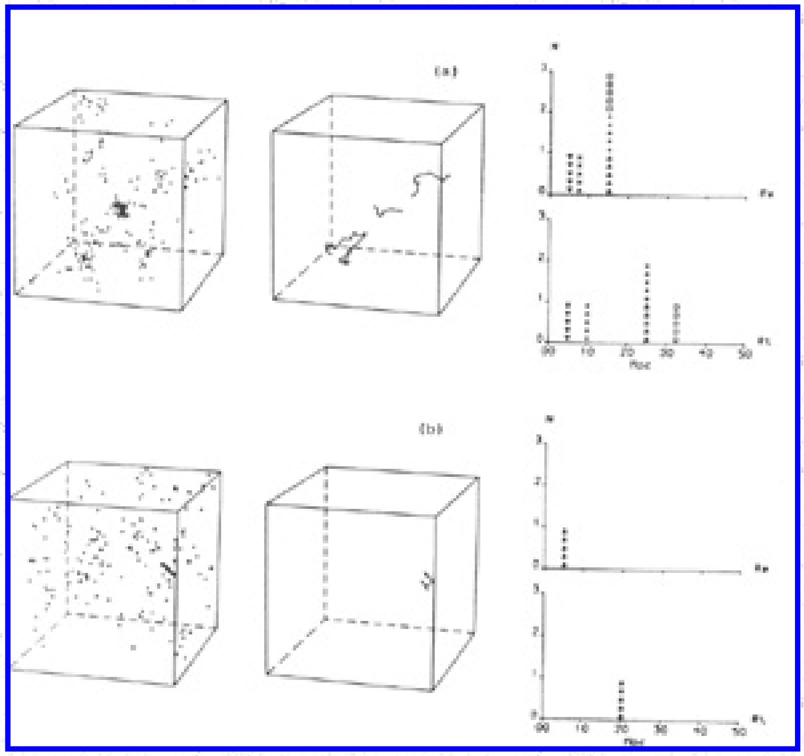

Figure 4.

We have also shown (Bhavsar and Ling 1988b, hereafter referred to as BL II) how easily our eye-brain is fooled into seeing large-scale filamentary features where none exist. The MST on the other hand identifies real filaments in toy models (see figure 1a and 1b in BL I) but is not fooled by spurious ones. This is shown using a construction now called Glass patterns (Glass 1969), produced by taking a random distribution of dots (figure 5a) which is then superimposed on itself with a slight rotation (figure 5b); figure 5c is the pruned and separated MST of the pattern in figure 5b.

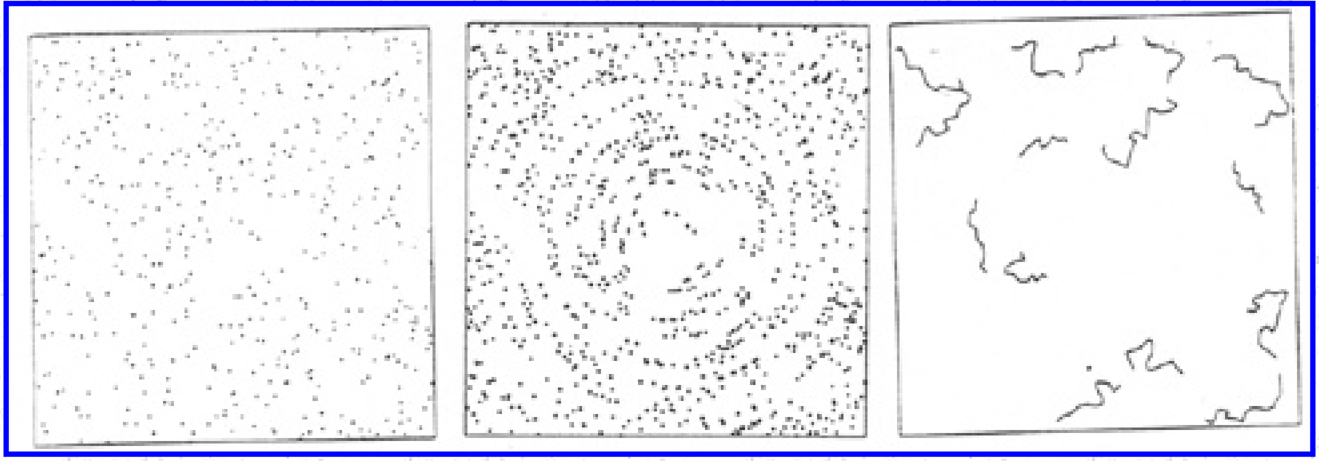

Figure 5.

The result is astonishing, the reader is urged to make transparencies of figure 5a and experiment. The visual impression is that of broken circular concentric rings of dots. Actually the dots have no auto-correlations, beyond the local pairings, but the eye-brain is fooled because it misinterprets a low level of global coherence (a result of the rotation) for large-scale linear features.

The MST is not fooled into identifying the perceived filaments. When the data permuting technique described above is used it shows that the filaments that the MST does identify are not statistically significant (see BL II for details.) Figure 6 shows the results of using the data permuting technique on figure 5b for n = 3, 5 and 10. Not only is the MST not fooled by the spurious filaments in figure 5b, but the ones it does identify are shown to be not statistically significant.

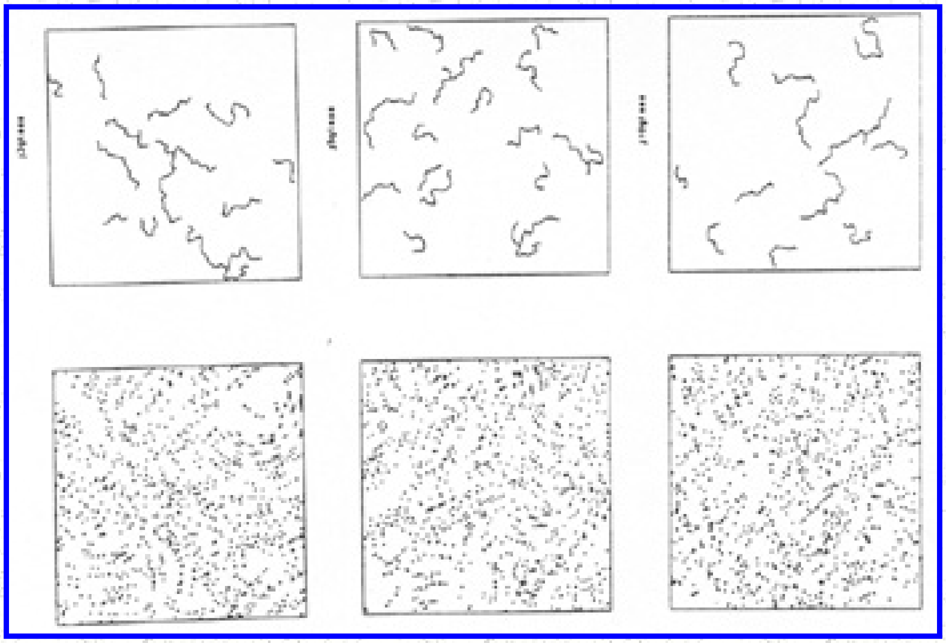

Figure 6.

Questions about the type of initial conditions and mechanisms in the early universe that make filamentary structures remain unanswered. Can they result simply from gravitational instability on initial random fluctuations in matter, or do they require specific initial conditions, or even non-gravitational mechanisms? We propose to use the demonstrated success of the MST techniques to answer this question by analyzing filamentary structure among numerical simulations and comparing them with data.

[1] A more recent controversy about a visual feature on Mars, photographed in July 1976 by the Viking orbiter, is that of a mile-long rock resembling a humanoid face (Carlotto 1988)! Back.

Next Contents Previous

Next Contents Previous

3.2. Topology

An analysis of the present day topology of the large-scale structure in the universe (Weinberg, Gott and Melott 1987) is directly related to the topology of the initial density fluctuations. The series of papers by Gott and his collaborators detail the way a measure of the topology - the genus of contours of a smoothed density distribution of galaxies - can be studied to determine the type of initial density fluctuations that existed in the early epochs of the universe.

The available data is sparse, and the relation between the genus and the threshold density has to be made from this available sample. The process involved in obtaining this relation involves smoothing the data, drawing the contour and then determining the genus per unit volume.

Figure 7, taken from Weinberg, Gott and Melott (1987) shows the contours for the CfA data (Huchra et al. 1983), drawn at various thresholds, v, of volume fractions, containing respectively 7%, 16%, 24% and 50% of the low dshow the mean values obtained for the bootstrap models. As mentioned the bootstrap is a procedure for determining the variance and not the mean values. No significance should be attached to the position of the squares.

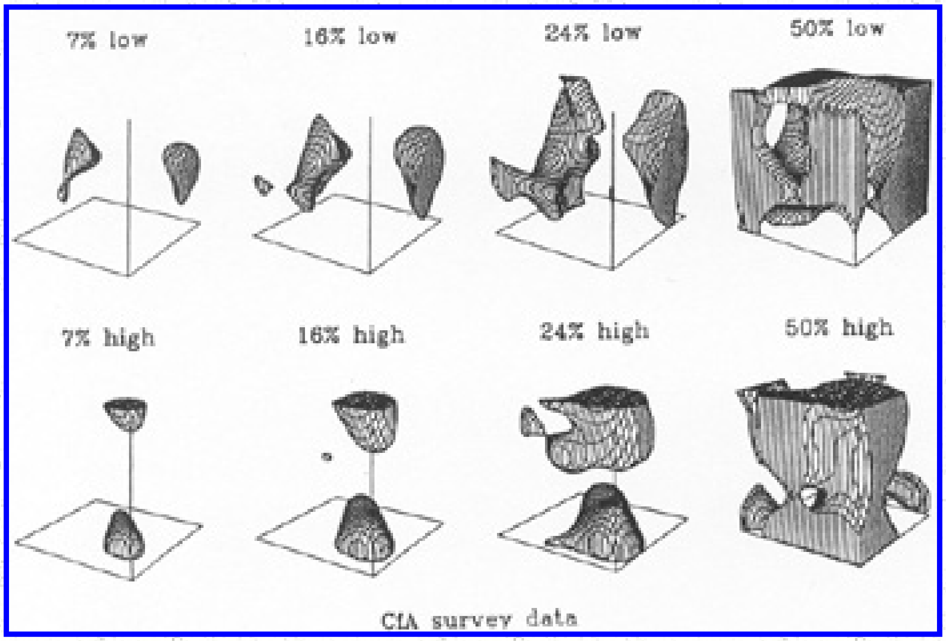

Figure 7.

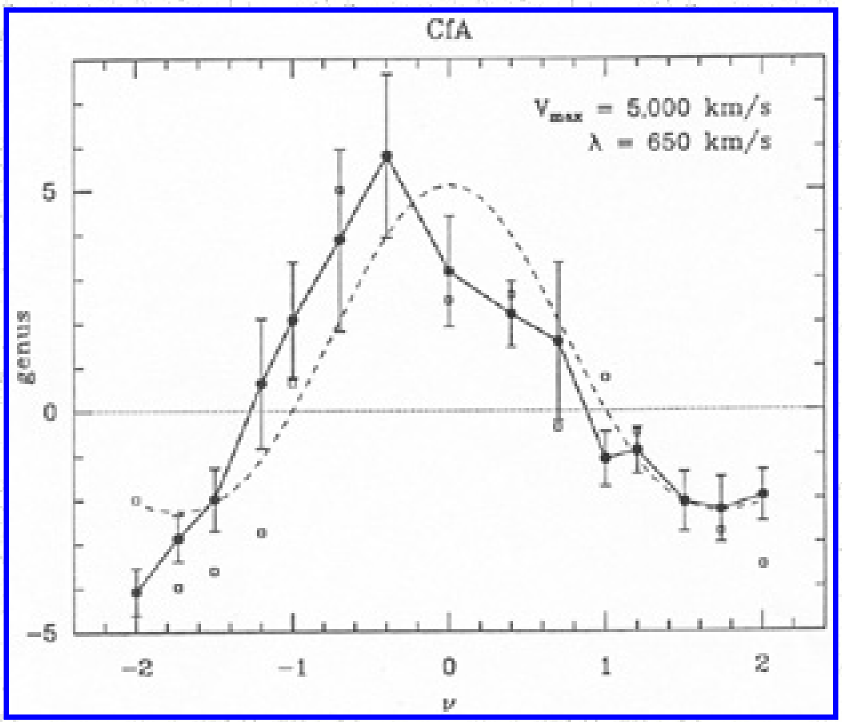

Figure 8.

Next Contents Previous

Next Contents Previous

4. EXTREME VALUE THEORY AND BRIGHTEST GALAXIES

This is an application of the statistical theory of extreme values to astronomy. Once again the name of the illustrious statistician Sir Ronald Fisher comes up. for it was he who developed the theory of extreme values. The problem has to do with the remarkably small dispersion ($\approx$ 0.35 mag) in the magnitudes, M_1, of the first brightest galaxies in clusters. This has made them indispensable as "standard candles", and they have, for this reason, become the most powerful yardsticks in observational cosmology. There has existed, however disagreement as to the nature of these objects. The two opposing viewpoints have been: that they are a class of "special" objects (Peach 1969; Sandage 1976; Tremaine and Richstone 1977); and at the other extreme, that they are "statistical", representing the tail end of the luminosity function for galaxies (Peebles 1968, Geller and Peebles 1976).

In 1928, in a classic paper, R.A. Fisher and L.H.C. Tippett had derived the general asymptotic form that a distribution of extreme sample values should take - independent of the parent distribution from which they are drawn! This work was later expanded upon by Gumbel (1958).

From extreme value theory, for clusters of similar size, the probability density distribution of M_1 for the "statistical" galaxies is given (Bhavsar and Barrow 1985) by the distribution in equation 6, which is often referred to as the first Fisher-Tippett asymptote, or Gumbel distribution.

$$P_{\rm Gum} = a \exp[a(M_1 - M_0) - e^{a(M_1 - M_0)}], \qquad (6)$$

where a is the parameter which measures the steepness of the luminosity function at the tail end. M_0 is the mode of the distribution and is a measure of the cluster mass, but with only a logarithmical dependence on the cluster mass.

We can compare the above distribution with the data to answer the question: "Are the first-ranked galaxies statistical?" Figure 9 shows the maximum likelihood fit of equation (6) to the magnitudes of the 93 first-ranked galaxies from a homogeneous sample of richness 0 and 1 Abell clusters. This excellent data for M_1 was obtained by Hoessel, Gunn and Thuan (1980, [HGT]). The fit is very bad. A Kolmogorov-Smirnov test also rejects the statistical hypothesis with 99% confidence. If all the galaxies are special, it is not possible to have a theoretical expression for their distribution in magnitude. Though we can make a simple argument that if they are formed from normal galaxies as a standard "mold", we expect a Gaussian distribution for their magnitudes or luminosities. This is in fact what is assumed for their distribution by most observers, as seen from the available literature. This possibility is explored also, and figure 9 shows a Gaussian with the same mean and variance as the data compared with the data for both cases where the magnitudes or the luminosities have a Gaussian distribution. The case where the luminosities are distributed as a Gaussian is called expnormal (in analogy to lognormal) for the magnitudes. Neither, it can be seen from figure 9 is acceptable.

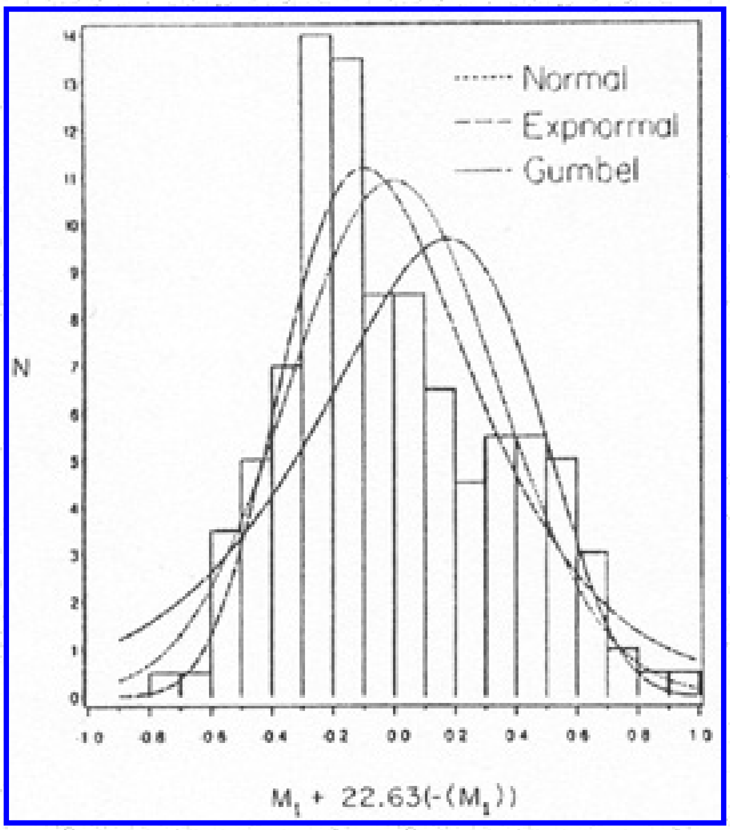

Figure 9.

This result does not necessarily imply that all brightest galaxies are special. It does demand though, that not all first ranked galaxies in clusters are statistical. The question as to their nature, implied by their distribution in magnitudes, was recently addressed by this author (Bhavsar 1989).

It was shown that a "two population model" (Bhavsar 1989) in which the first brightest galaxies are drawn from two distinct populations of objects - a class of special galaxies and a class of extremes of a statistical distribution - is needed. Such a model can explain the details of the distribution of M_1 very well. Parameters determined purely on the basis of a statistical fit of this model with the observed distribution of the magnitudes of the brightest galaxies, are exceptionally consistent with their physically determined and observed values from other sources.

The probability density distribution of the magnitudes of the special galaxies is assumed to be a Gaussian. This is the most general expression for the distribution of either the magnitudes or luminosities of these galaxies, if they arise from some process which creates a standard mold with a small scatter, arising because of intrinsic variations as well as experimental uncertainty in measurement. The distribution of M_1 for the special galaxies is given by

$$P_{\rm sp}(M_1) = \frac{1}{\sigma_{\rm sp}\sqrt{2\pi}} \exp\left\{\frac{(M_1 - M_{\rm sp})^2}{2\sigma_{\rm sp}^2}\right\}. \qquad (7)$$

If a fraction, *d*, of the rich clusters have a special galaxy which competes with the normal brightest galaxy for first-rank, then the distribution that results is derived in Bhavsar (1989). This expression, *f* (M_1), which describes the distribution of M_1 for the first-ranked galaxies in rich clusters, for the two population model is given by equation (8)

$$f(M_1) = d[P_{\rm sp}(M_1)\int_{M_1}^{\infty} P_{\rm Gum}(M')dM' + P_{\rm Gum}(M_1)\int_{M_1}^{\infty} P_{\rm sp}(M')dM'] + (1-d)P_{\rm Gum}(M_1).. \qquad (8)$$

The first term in the above expression gives the probability density distribution of special galaxies with the condition that the brightest normal galaxy in that cluster is always fainter. The second term gives the probability density distribution of first-ranked normal galaxies, in clusters containing a special galaxy, but the special galaxy is always fainter. The last term gives the probability density of normal galaxies in clusters that do not have a special galaxy. Equation (8) is our model's predicted distribution of M_1 for the brightest galaxies in rich clusters. The parameters in this model are: i) $\sigma_{\rm sp}$ - the standard deviation in the magnitude distribution of the special galaxies; ii) $M_{\rm sp}$ - the mean of the absolute magnitude of the special galaxies; iii) *a* - the measure of the steepness of the luminosity function of galaxies at the tail end; iv) $M_{\rm ex}$ - the mean of the absolute magnitude of the statistical extremes given by $M_{\rm ex} = M_0$ - .577/*a*, we shall instead use the parameter $b = M_{\rm sp} - M_{\rm ex}$, the difference in the means of the magnitudes of special galaxies and statistical extremes; and v) *d* - the fraction of clusters that have a special galaxy.

We have chosen the maximum-likelihood method, being the most bias free, and therefore best suited, to determine the values of the parameters. There are five independent parameters and 93 values of data. We maximize the likelihood function, defined by

$$\pounds = \prod_{i=1}^{93} f(M_1[i]), \qquad (9)$$

where the function *f* (M_1[*i*]) is the value of *f* (M_1) defined in equation (8), evaluated at each of the 93 values of M_1 respectively for *i* = 1 to 93. The values of the parameters that maximize £ give the maximum-likelihood fit of the model to the data. The parameters, thus determined, have the following values:

$$\sigma_{\rm sp} = 0.21 \ {\rm mag} \qquad (10)$$

$$M_{\rm sp} = -22.80 \ {\rm mag} \qquad (11)$$

$$b = M_{\rm sp} - M_{\rm ex} = -0.53 \text{ mag} \tag{12}$$

$$a = 4.33 \tag{13}$$

$$d = 0.65 \tag{14}$$

Figure 10 compares the data to the model [equation (8)] evaluated for the parameter values determined above. The fit is very good. Note that the fit is calculated using all 93 independent observations, and not tailored to fit this particular histogram.

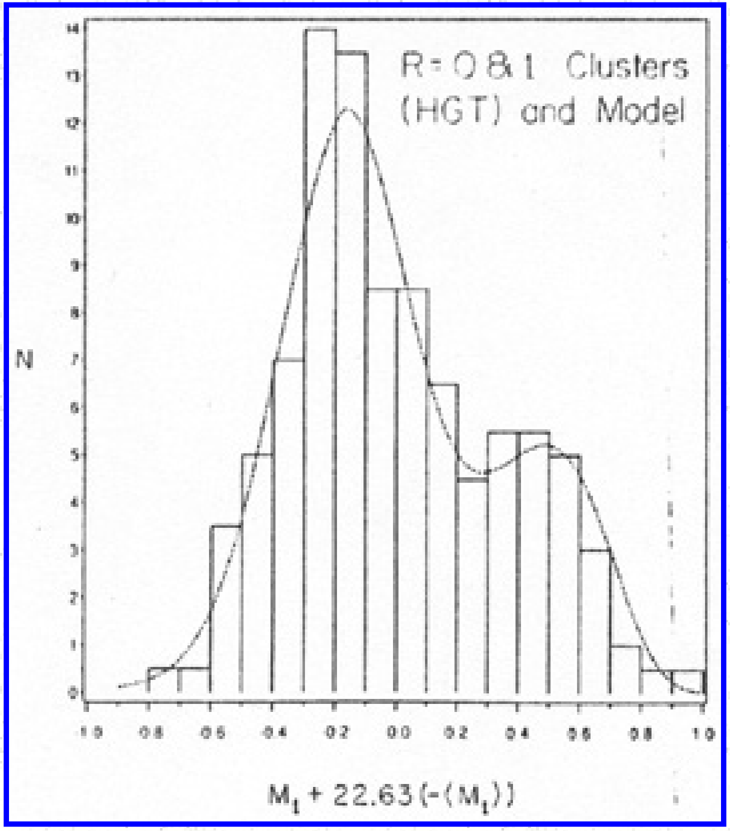

Figure 10.

A further detailed statistical analysis by Bhavsar and Cohen (1989) of alternatives to the assumed Gaussian distribution of the magnitudes of special galaxies, along with a study of the confidence limits of the parameters determined by maximum-likelihood and other statistical methods has determined a "best" model. It turns out that the models in which the luminosities have a Gaussian distribution work marginally better. This may have been expected, luminosity being the physical quantity. This model requires 73% of the richness 0 and 1 clusters to have a special galaxy which is on average half a magnitude brighter than the average brightest normal galaxy. As a result, about 66% of all first-ranked galaxies in richness 0 and 1 clusters are special. This is because in 7% of the clusters, though a special galaxy is present, it is not the brightest.

Although it is generally appreciated that some of the brightest galaxies in rich dusters are a morphologically distinct class of objects (eg cD galaxies); we have approached the problem from the viewpoint of the statistics of their distribution in M_1, and conclude that indeed some of the brightest galaxies in rich clusters are a special class of objects, distinct from the brightest normal galaxies. Further we have been able to model the distribution of these galaxies. We have presented statistical evidence that the magnitudes of first-ranked galaxies in rich clusters are best explained if they consist of two distinct populations of objects; a population of special galaxies having a Gaussian distribution of magnitudes with a small dispersion (0.21 mag), and a population of extremes of a statistical luminosity function. The best fit model requires that 73% of the clusters have a special galaxy that is on average 0.5 magnitudes brighter than the brightest normal galaxy. The model also requires the luminosity function of galaxies in clusters to be much steeper at the very tail end, than conventionally described.

Next Contents Previous

Contents Previous

REFERENCES

1. Barrow, J.D., and Bhavsar, S.P. 1987, Ouart.J.R.A.S., 28, 109 (BB).
2. Barrow, J.D., Bhavsar, S.P., and Sonoda, D.H. 1984, M.N.R.A.S., 210, 19p
3. Barrow, J.D., Bhavsar, S.P., and Sonoda, D.H. 1985, M.N.R.A.S., 216, 17 (BBS).
4. Bhavsar, S.P. 1989, Ap. J. 338, 718.
5. Bhavsar, S.P., and Barrow. J.D. 1985, M.N.R.A.S. 213, 857 (BB).
6. Bhavsar, S.P., and Barrow, J.D. 1983, M.N.R.A.S., 205, 61p
7. Bhavsar, S.P., and Barrow, J.D. 1984, in Clusters and Groups of Galaxies eds. F. Mardirossian, M. Giuricin, and M. Mezzetti (Dordrecht: Reidel) p. 415.
8. Bhavsar, S.P., and Ling, E.N. 1988, Ap. J. (Letters), 331, L63 (BL I).
9. Bhavsar, S.P. 1988, Pub. Ast. Soc. Pac. 100, 1314 (BL II).
10. Carlotto, M.J. 1988, Appl. Optics, 27, 1926.
11. Efron, B. 1979, Ann. Stat., 7, 1.
12. Evans, J.E., and Maunder, E.W. 1903, M.N.R.A.S., 63, 488.
13. Fisher, R.A., and Tippett, L.H.C. 1928, Proc. Cambridge Phil. Soc. 24, 180.
14. Geller, M.J., and Peebles, P.J.E. 1976, Ap. J. 206, 939.
15. Glass, L. 1969, Nature, 223, 578.
16. Gott, J.R. et al. 1989, ap. J. 340, 625.
17. Groth, E.J, Peebles, P.J.E., Seldner, M., and Soneira, R.M. 1977, Sci. Am.
18. Gumbel, E.J. 1966, Statistics of Extremes, (New York: Columbia Univ. Press).
19. Hillenbrand, L. 1989, B.A. thesis, (Princeton University).
20. Hillenbrand, L., Bhavsar, S.P., Gott, J.R. in preparation
21. Hoessel, J.G., Gunn, J.E., and Thuan, T.X. 1980, Ap. J. 241, 486 (HGT). 237, 76 (May).
22. Hoyt, W.G. 1976, Lowell and Mars, (Tucson: University of Arizona Press).
23. Huchra, J., Davis, M., Lantham, D.W., and Tonry, J. 1983, Ag. J. Suppl., 52, 89.
24. Ling, E.N. 1987, Ph. D. thesis, (University of Sussex).
25. Limber, D.N. 1953, Ap. J., 117, 134
26. Lowell, P. 1906, Mars and Its Canals, (New York: The Macmillan Company).
27. Peach, J.V. 1969, Nature, 223, 1140.
28. Peebles, P.J.E. 1968, Ap. J. 153, 13
29. Peebles, P.J.E. 1973, Ap. J. 185, 413.
30. Peebles, P.J.E. 1980, The Large Scale Structure of the Universe (Princeton, New Jersey: Princeton University Press)
31. Peebles, P.J.E., and Groth, E.J. 1975, Ap. J., 196, 1.
32. Sandage, A. 1976, Ap. J. 205, 6.
33. Shane, C.D., and Wirtanen, C.A. 1967, Pub. Lick Cbs., Vol. 22, Part 1.

34. Totsuji, H., and Kihara, T. 1969, Pub. Ast. Soc. Japan, 21, 221.
35. Tremaine, S.D., and Richstone, D.O. 1977, Ap. J. 212, 311.
36. Weinberg, D.H., Gott, J.R., and Melott, A.L. 1987, Ap. J. 321, 2.
37. Zahn, C.T. 1971, IEEE Trans. Coma., C20, 68.
38. Zwicky, F., Herzog, E., Wild, P., Karpowicz, M., and Kowal, C.T. 1961-68, Catalogue of Galaxies and Clusters of Galaxies. (Cal. Inst. of Tech. Pasadena)

Contents Previous