

Searching for Small-Scale Anisotropies in Arrival Directions of Ultra-High Energy Cosmic Rays with the Information Dimension

Eli Visbal (Carnegie Mellon University) Advisor: Dr. Stefan Westerhoff

Summer 2005

Abstract

The origin of ultra high energy cosmic rays (UHECRs) is one of today's most interesting mysteries in physics. A direct way to search for the sources of UHECRs is to analyze the distribution of their arrival directions, but all analyses to date show no significant departure from isotropy. Many different types of anisotropies might be expected, for example clusters or lines of arrival directions, or voids with a lack of events in certain regions, and different search methods for each of these deviations from isotropy have been designed. On the other hand, there are methods that test for all deviations from isotropy simultaneously, avoiding any statistical penalty from applying many different tests. One of these methods uses the so-called information dimension of the data set, a quantity analogous to the entropy, to check how evenly distributed a data set is. We compare the effectiveness of identifying small-scale anisotropies with the information dimension test and more specific anisotropy tests, for example the two-point correlation function, using data at energies above 10^{19} eV taken in stereoscopic mode with the High Resolution Fly's Eye (HiRes) experiment.

1 Introduction

Analysis of arrival directions may be the most promising way of identifying sources of UHECRs. The difficulty is that if cosmic rays are charged particles they will be deflected both by galactic and extragalactic magnetic fields. Those with the highest energies would be deflected the least and thus may give strong evidence for a particular source model. It is possible that magnetic fields could cause UHECRs to arrive in a number of ways. The simplest

would be in clusters. Another possibility would be in lines. If a group of particles with different energies is being emitted from the same source those with lower energies would follow a similar path but be deflected more. This could leave a line formation in arrival directions. It is also possible that there are regions of space with fewer sources or magnetic fields deflect particles from entering at particular directions. This would cause voids or lack of events in certain regions. This seems possible because galaxies are distributed with voids.

Three different types of small-scale structure were added to isotropic data, using a simulation which accounts for the exposure of HiRes, and then a test was applied for each to determine the statistical significance with respect to isotropic simulated data. For clusters the two point correlation test was used, for lines the triangle test was developed and used and for voids the void probability function test was designed and used. The same data was then tested using the information dimension. The strength of this test was then compared directly to each of the tests mentioned above for their respective anisotropies. It is useful to search for many anisotropies in one technique because this prevents incurring statistical penalties from applying many tests.

The paper is structured as follows. In Section 2, the information dimension is introduced and discussed as a general test of anisotropy. In Sections 3-5, the tests for each anisotropy type mentioned above are described and used to directly compare their statistical strength with the information dimension. In 6, some of the limitations of the information dimension are described. In Section 7 the information dimension is used to perform an analysis of current HiRes data and conclusions are presented in Section 8.

2 Information Dimension

Fractal dimensionality is a measure of the scaling symmetry of a structure. It can be used to determine how self-consistent data is at different length scales [1]. For analysis of UHECRs it is most useful to use the information dimension case:

$$D_I = - \sum_{i=1}^N \lim_{\epsilon \rightarrow 0^+} \frac{P_i(\epsilon) \log(P_i(\epsilon))}{\log(1/\epsilon)} \quad (1)$$

where $P_i(\epsilon)$ is the probability of finding an event in bin i with an edge size ϵ . The bins should have equal area and shape. Since it is impossible to apply equally shaped and sized bins to a sphere an appropriate approximation must be used. The method used is HEALPix or Hierarchical Equal Area isoLatitude Pixelization [2]. This produces a set of bins on the celestial

sphere which have equal areas and for high pixelizations very similar shapes. For all of the following analysis a pixelization of 3,145,728 was used which is small enough to resolve any possible anisotropy. To assign $P_i(\epsilon)$ all nearby pixels for each event are considered and a contribution which is calculated from a Gaussian function centered on the event with a standard deviation of 0.5° corresponding to the experimental uncertainty of HiRes is added. Each event is given equal weight and all of the contributions are normalized to yield $\sum_{i=1}^N P_i(\epsilon) = 1$. The most useful way to determine ϵ is to approximate each pixel as a square and choose units where the entire sphere has an area of one. This yields $\epsilon = \frac{1}{\sqrt{N}}$. With this choice of units in the case where $P_i(\epsilon)$ is perfectly isotropic, that is $P_i(\epsilon) = \frac{1}{N}$ for all N , $D_I = 2$. This provides an upper limit on the value which can be obtained.

It is also interesting to compare D_I with:

$$S = -k \sum_r p_r \log p_r \quad (2)$$

the formula for entropy. D_I is analogous in a very true sense. It is a measure of how evenly distributed a data set is. Very evenly distributed is analogous to a high entropy, while more “clumpy” or structured data is analogous to low entropy.

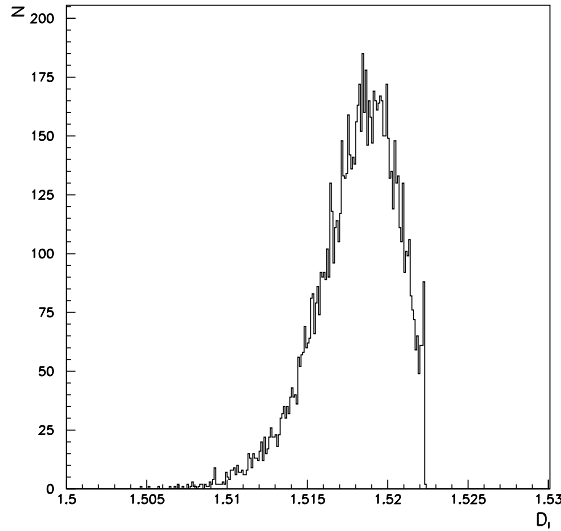


Figure 1: Distribution of D_I values with Isotropic Data for 55 Events

For our particular study sets of 55 and 271 data points were used. These numbers were chosen because they correspond to the number of events in

the data sets above 30 EeV and 10 EeV respectively. To test a potentially anisotropic set the information dimension was applied to this data and the D_I value was compared to that of many isotropic simulated data sets. The fraction of these data sets which had equal or lower D_I values gave the statistical significance. All statistical significances were determined with 10000 data sets. A distribution of isotropic D_I values is shown in Fig. 1. It has a sharp cutoff on the right side which corresponds to as evenly distributed as possible for that number of points. This happens when all points are far enough away from each other that there is no overlap in contributions of $P_i(\epsilon)$ from different events.

3 Clustering

Data sets were produced with both 2-point and 3-point clusters. Events were placed with a random Gaussian function with a standard deviation of 0.5° . They were first tested with the two point correlation technique. In this technique the distance between every pair of points is examined and those below a certain threshold are counted. This number can then be compared to isotropic simulated data to give a statistical significance. We used a threshold of 2° corresponding to 4 standard deviations of the placement smearing.

The information dimension described above was then performed on the same data and the statistical significance of each test was compared. Looking at Fig. 2 one can see that in the large data set the information dimension test is not as good as the two point correlation. In the small data set it is equally effective in the case of 3-point clusters and better in the case of 2-point clusters. Both are effective methods of discerning points, especially in a small data set.

4 Lines

As described above, it seems possible for magnetic fields to deflect UHECRs into lines across our sky. Data with artificial lines of 3 events 4° long and lines of 4 events 6° long was created. All the points were again placed with the experimental uncertainties of HiRes.

To test for lines the triangle test was developed. In this technique the triangle areas defined by each triplet in a data set are measured. To identify lines two thresholds are introduced. First, all triangles with a side longer than would be expected in a line are excluded, 8° was chosen. Then all triangles with area below what is consistent with smearing due to uncertainty are

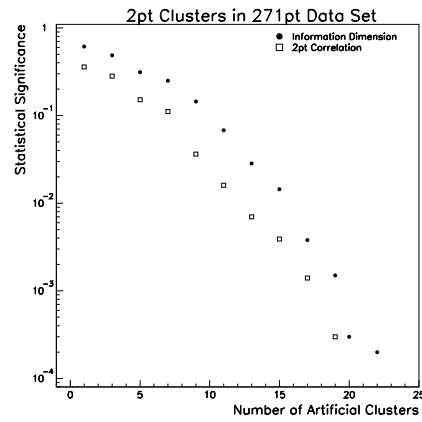
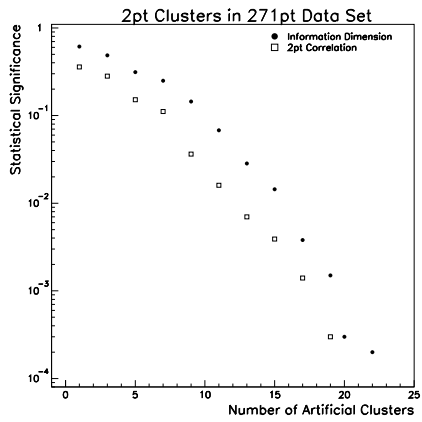
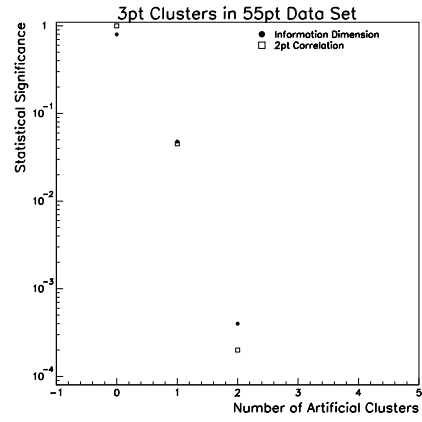
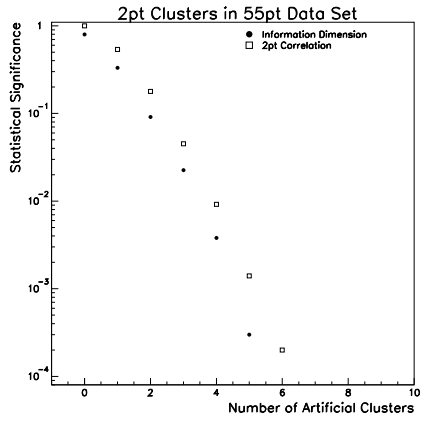


Figure 2: Statistical Significance of Tests on Clustering

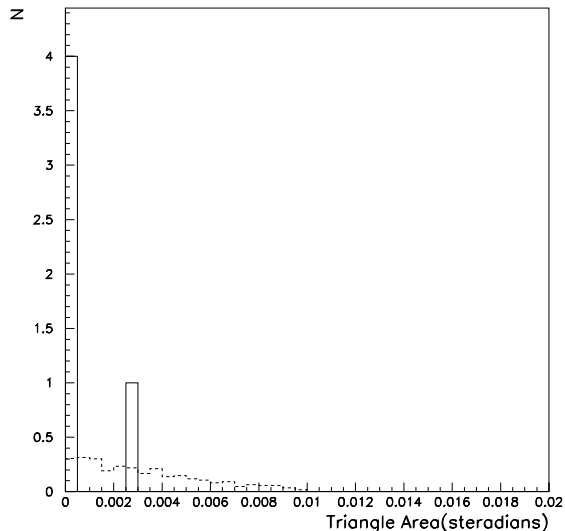


Figure 3: Comparison of Data with Four Artificial Lines (solid line) and Average Isotropic Data (dashed line)

counted. The area used for this count was 5×10^{-4} steradians. This count is compared to isotropic simulated data to determine a statistical significance.

The information dimension was applied to the data, the sensitivity of the two tests is compared in Fig. 4. It is important to note that the triangle test has an advantage for the longer 4-point lines while the information dimension is better for the shorter 3-point lines. This is because the information dimension is looking at how evenly distributed the data sets are and 4-point lines produce a lot of small triangles, but not as high density of events.

5 Voids

Voids are another interesting anisotropy to examine. Data with voids was created by simulating isotropic data with the same isotropic simulation, but excluding circular areas on the sphere with radii of 5, 10 and 15 degrees. Voids were only applied to the large data set. We attempted to identify voids using a void probability function test. In this test the probability of finding no other points within a certain radius at a random point as a function of the radius is determined. Fig. 5 shows how this function looks for an isotropic data set and the same data set with voids added. The statistical significance would be determined by adding up the values of the function

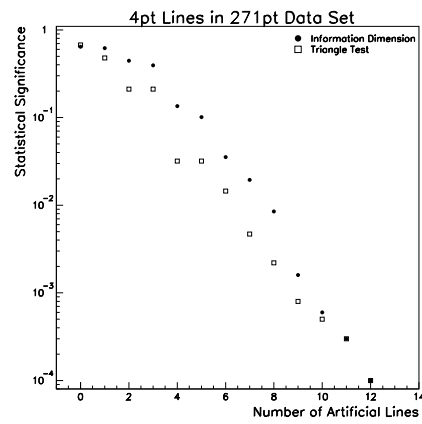
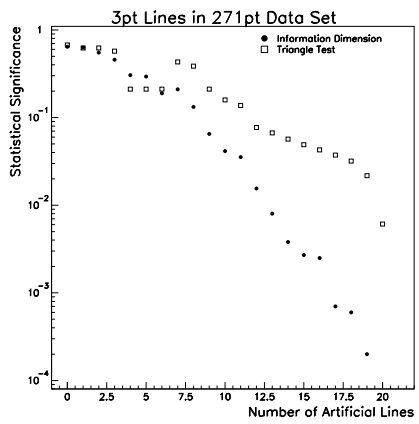
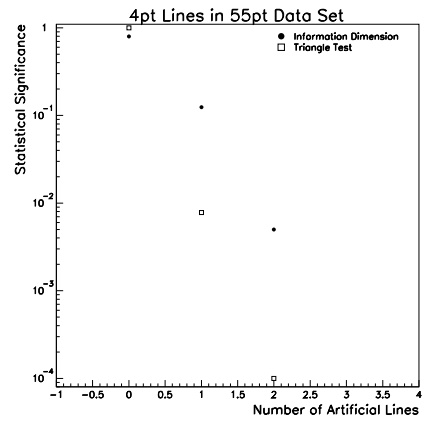
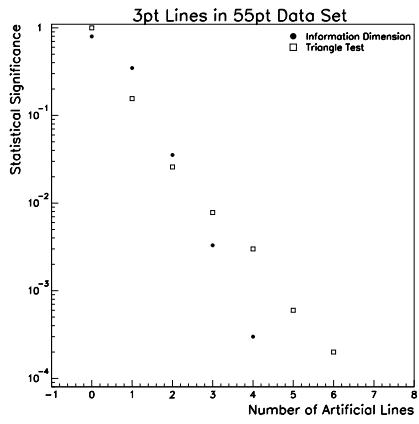


Figure 4: Statistical Significance of Tests on Lines

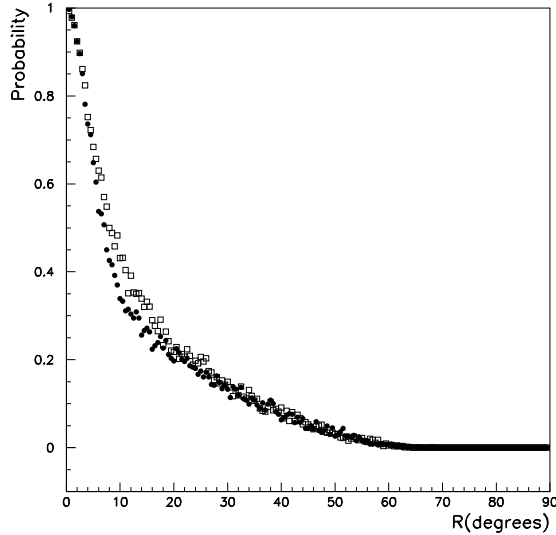


Figure 5: Void Probability Function of isotropic data (dots) and the same data with 9 15-degree voids added (squares)

and comparing this to isotropic data. It turns out that this test does a very poor job of identifying voids because random fluctuations in isotropy cause changes in the function that are just as large as artificially added voids. The test can only distinguish very large voids. It may be possible that the test would improve with more trials in determining the function but this was not computationally practical and seems unlikely. Although this method did not work it is an example of the type of method which could be developed to specifically find voids.

The information dimension does a better job of identifying voids. A few large voids are very easy to distinguish from isotropy with this test. Even with small voids good statistical significances can be achieved with many voids. This can be seen clearly in Fig. 6.

6 Limitations

There is one very important limitation of the information dimension technique which must be realized. It can only find anisotropies which are on the same distance scale as the standard deviation used in assigning $P_i(\epsilon)$ or smaller. In our analysis the uncertainty of HiRes was used. This however ignores any magnetic deflections. To best be able to identify anisotropies

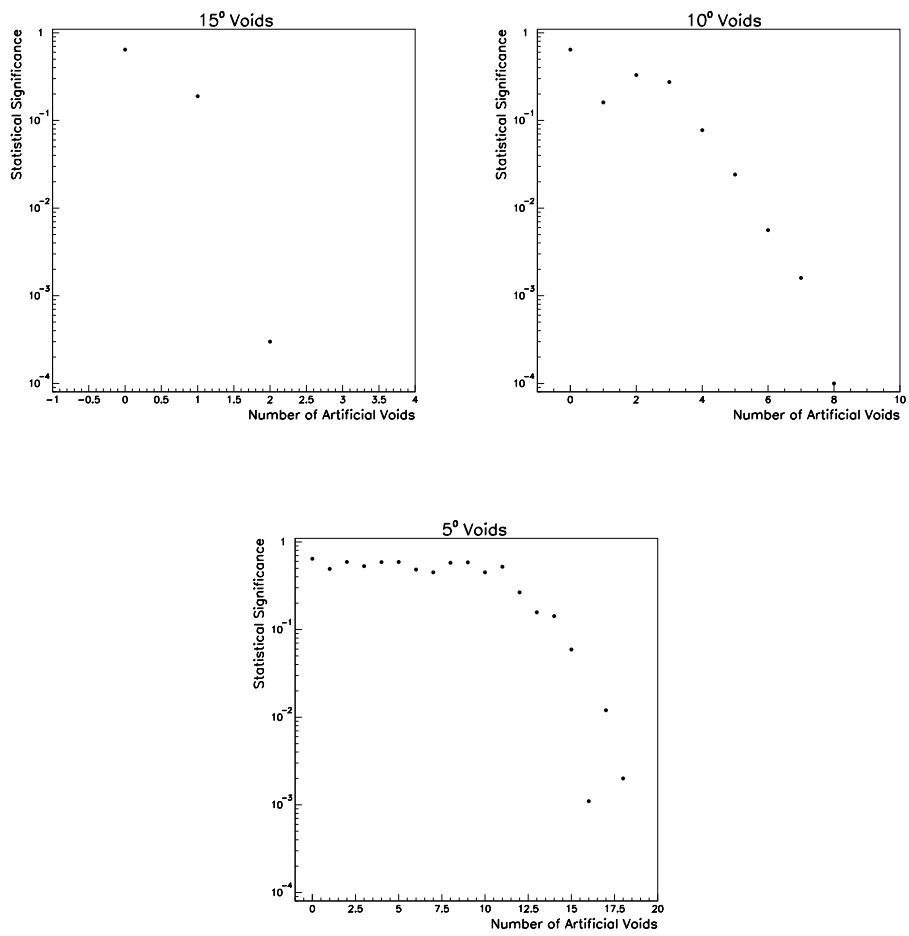


Figure 6: Statistical Significance of D_I on Voids

caused by deflection, such as lines, it would be best to assign $P_i(\epsilon)$ using a Gaussian with a standard deviation that accounts for deflection. A different distance scale was not chosen because we wanted to first evaluate this test without any arbitrary thresholds.

This limitation can be seen clearly if a much lower uncertainty for HiRes is considered. Attempting to find lines with the triangle test is much more successful, but finding lines with the information dimension becomes impossible unless you increase the standard deviation in assigning $P_i(\epsilon)$. This is because in data sets without definite point sources and no magnetic deflection, points contribute only to very nearby pixels in assigning $P_i(\epsilon)$. This will result in the upper limit discussed earlier which can be seen in Fig 1.

7 HiRes Analysis

The information dimension was used to analyze current HiRes data. All energy cutoffs for events over 10 EeV were considered. All 271 events above 10 EeV were tested, then the 270 most energetic, then the 269 most energetic and so forth. Looking at Fig. 7 one can see that none of these cuts is statistically significant. Another option would be to perform the same test across

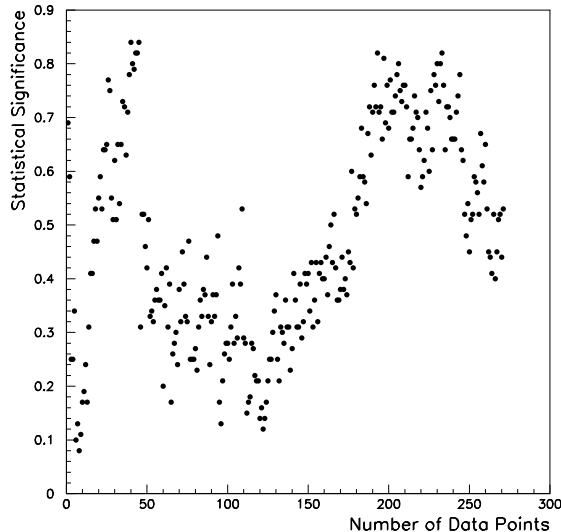


Figure 7: Statistical Significance of All Possible Energy Cuts Above 10 EeV on HiRes Data

a wide range of standard deviations in assigning $P_i(\epsilon)$. This would search

the data for potential anisotropies at larger distance scales. It is important to note that in any scan such as these if a strong statistical significance is found appropriate statistical penalties must be applied to account for the large number of tests performed.

8 Conclusions

The information dimension test's ability to discern three types of small-scale anisotropies in the arrival direction of UHECRs has been investigated. This test provides two major advantages. In one test it looks for many different types of anisotropies without *a priori* thresholds beyond adjusting for the distance scale. It does a good job of identifying these anisotropies relative to other more specific tests which can be employed.

9 Acknowledgments

I would like to thank my advisor Stefan Westerhoff for guidance in this summer project. I would also like to thank Segev Benzvi, Brian Connolly, Chad Finley, and Andrew O'Neill for helping as well. Finally, I would like to express my gratitude towards the National Science Foundation for making this valuable experience possible.

References

- [1] Stokes, B., Jui, C., Matthews, J., "Using Fractal Dimensionality in the Search for Source Models of Ultra-High Energy Cosmic Rays," *Astropart.Phys.* 21 (2004) 95.
- [2] Gorski K.M., Hivon E., Banday A.J., Wandelt B.D., Hansen F.K., Reinecke M., Bartelman M., "HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere," *ApJ.* 622 (2005) 759.