The Astrophysical Journal, 623:L33–L36, 2005 April 10

BLIND SEARCH FOR THE REAL SAMPLE: APPLICATION TO THE ORIGIN OF ULTRA–HIGH-ENERGY COSMIC RAYS

Boris E. Stern[1,2] and Juri Poutanen[3]
Astronomy Division, Department of Physical Sciences, P.O. Box 3000, FI-90014 University of Oulu, Finland;
stern@bes.asc.rssi.ru, juri.poutanen@oulu.fi
Received 2004 November 2; accepted 2005 March 7; published 2005 March 18

ABSTRACT

We suggest a method for statistical tests that does not suffer from a posteriori manipulations of tested samples (e.g., cuts optimization) and does not require the somewhat obscure procedure of applying a penalty estimate. The idea of the method is to hide the real sample (before it has been studied) among a large number of artificial samples, drawn from a random distribution expressing the null hypothesis, and then to search for it as the one demonstrating the strongest hypothesized effect. The statistical significance of the effect in this approach is the inverse of the maximum number of random samples for which the search is successful. We have applied the method to revisit the correlation between the arrival directions of ultra–high-energy cosmic rays and BL Lacertae objects. No significant correlation is found.

Subject headings: catalogs — cosmic rays — methods: data analysis — methods: statistical

1. INTRODUCTION

Communications about effects detected at marginally significant levels constitute a considerable fraction of all scientific results. The scientific community usually treats such reports with skepticism. Indeed, too many marginally significant effects have not withstood the accumulation of additional data.

High-energy astrophysics provides a number of instructive examples of searches for marginally significant effects. Indeed, there are many detection of particles or transient gamma-ray events whose sources (i.e., objects known from observations at other wavelengths) are unknown. This stimulates intensive searches for various correlations between different classes of events and objects. For example, there are a number of works reporting the detection of correlations between the locations of gamma-ray bursts (or subsamples thereof) and various objects: galaxy clusters (Kolatt & Piran 1996), the Galactic plane (Belli 1997), and the local Galactic arm (Komberg et al. 1997). None of these results has been confirmed. Another, similar, area is that of ultra–high-energy cosmic rays (UHECRs) and searches for their hypothetical sources. A claim of significant autocorrelation in the arrival directions of UHECRs detected by the Akeno Giant Air Shower Array (AGASA) (Hayashida et al. 1996; Takeda et al. 1999) motivated searches for cross-correlations between UHECRs and various astrophysical objects. In particular, there were reported statistically significant cross-correlation signals between UHECRs and BL Lacertae objects (Tinyakov & Tkachev 2001, hereafter TT01; but see Evans et al. 2003), the supergalactic plane (Uchihori et al. 2000), radio-loud compact quasars (Virmani et al. 2002), highly luminous bulge-dominated galaxies (presumably, nearby quasar remnants; Torres et al. 2002), and Seyfert galaxies (Uryson 2004).

The reason for the abundance of detected correlations is quite evident: the number of various possible effects, which have been searched for with statistical methods, is large, and it is not surprising that some of them demonstrate a marginally significant signal just by chance. The situation is even worse because, typically, a probed effect is somewhat uncertain and the researcher tries different versions of the hypothesis, varying parameters and applying various cuts to the data samples. This means that the researcher performs a number of tests of the same effect that are neither independent nor completely dependent. These numerous trials, again, increase the probability of observing a signal in one of the trials by chance, and an analysis of this kind of bias is difficult. We illustrate this problem in § 4 and Figure 1.

Does this mean that one should reject the possibility of manipulating data samples with cuts and parameters? A blind statistical test in which all cuts and parameters have been set and motivated a priori is good style. But there are many situations in which such an a priori definition of a test is very problematic, and the investigator sometimes really needs the ability to vary the testing procedure to see what will happen.

In principle, a researcher can account for these numerous trials using random samples, representing the null hypothesis. Often this is done in the following way (see, e.g., TT01): The investigator prepares a large array of N random samples, d^i, and performs the same estimate of the effect for each of these samples as was done for the real sample, d^0, in each statistical trial. Let a statistic associated with the confidence of the effect (e.g., $1-p$, where p is the probability of obtaining the result from the null hypothesis) be $S^i_j = F(d^i, C_j)$, where C_j is a set of cuts and/or analysis parameters from the universe C of all cuts and parameters. First one finds the maximum for the real sample, $S^0_{\max} = \max\{F(d^0, C_j)\}$, which is reached at some $j = j_0$. Then one performs a similar search for the random samples' $S^i_{\max} = \max\{F(d^i, C_j)\}$. The significance can be defined as the fraction of random samples satisfying the condition $S^i_{\max} > S^0_{\max}$. This value differs from the straightforward estimate (uncorrected for the numerous trials) of the significance by the "penalty factor" $N(S^i_{\max} > S^0_{\max})/N(S^i_0 > S^0_{\max})$, where $S^i_0 = F(d^i, C_{j_0})$ and $N(\bullet)$ is the number of samples satisfying the condition $(\bullet)$. This procedure is sufficient if the investigator (1) follows the above procedure precisely and (2) does not use the a posteriori information from the real sample in planning the investigation strategy.

[1] Institute for Nuclear Research, Russian Academy of Sciences, Prospekt 60-letiya Oktyabrya 7a, 117312 Moscow, Russia.
[2] Astro Space Center, P. N. Lebedev Physical Institute, ulitsa Profsoyuznaya 84/32, 117997 Moscow, Russia.
[3] Corresponding Fellow, NORDITA, Copenhagen.

We note that satisfying both these conditions is not so easy once the investigator has studied the real sample and feels which combination of cuts or model parameters will provide the most significant signal. Then he or she can find the most favorable trial intuitively, avoiding a large number of unfavorable ones. In other words, the investigator can introduce a bias in the choice of C_j and overestimate the significance of the effect by using a posteriori knowledge. We should emphasize that the investigator can introduce such a bias quite accidentally. This is a serious disadvantage of the approach. This kind of bias is difficult to trace, and we consider such a method to be insufficiently credible.

In this Letter, we suggest a new approach that provides a simple way of avoiding this "pressure" from the a posteriori information. The investigator can hide the real sample inside a large array of random null-hypothesis samples *prior to any data analysis*. Now one has a single array of N samples d^i. One of them is real (the investigator does not know which), and the others are random. The problem is thus inverted: instead of confirming the hypothesis using the real sample, the investigator must find the real sample in the array using the hypothesis that the verified effect exists, that is, find $i_{\max}$ corresponding to the maximum value of $S_j^i = F(d^i, C_j)$. This is a blind test: the investigator does not know where the real sample is and can feel free to perform numerous trials. If the investigator finds the real sample, the significance of the effect is just the inverse of the number of samples in the array.

An alternative to our approach is the cross-validation method, in which the search for an effect is carried out on a *fraction* of the data sample. Therefore, it is less sensitive. Below we demonstrate our method by applying it to the problem of the UHECR–BL Lac correlation.

2. PROCEDURE

2.1. *Catalogs*

We used the AGASA sample of UHECRs, with 58 events above 4×10^{19} eV (Hayashida et al. 2000), and the catalog from Véron-Cetty & Véron (2003) containing 876 BL Lac objects. We do not combine the AGASA sample with the data from other experiments, because the other samples are smaller and problems associated with the nonuniform structure of a joint sample would outweigh the statistical gain. The BL Lac catalog has been cut in declination at $-10°$ and was subject to various brightness cuts. We also tried a subcatalog of *confirmed* BL Lac objects that includes 491 objects. Actually, it is not clear which catalog is more relevant (TT01 used a confirmed subcatalog), and therefore we try both variants.

2.2. *Null Hypothesis and Random Samples*

The null hypothesis in our case is just an isotropic distribution of UHECR arrival directions convolved with the AGASA exposure function. The latter is a function of declination and does not depend on right ascension. This provides a simple way to prepare random, null-hypothesis samples that avoid possible uncertainties in the latitude exposure function, by sampling the right ascension uniformly while keeping the actually observed declination for each event. We nevertheless have dispersed the declinations of the UHECRs by $\pm 3°$ around their real values in order to destroy any possible small-scale latitude correlation. Such a small dispersion will not distort the much wider exposure function.

When performing the test, we have distributed roles: one of us acts as an "investigator," while the other plays the role of "examiner." The examiner prepared an array of 99 random samples as described above and inserted the real sample into the array, keeping the sequential real sample number secret from the investigator. He did not participate in the data analysis until the investigator made his final choice.

2.3. *Measure for the Correlation Signal*

We used the usual two-point correlation function, counting the number n of UHECRs within an angle δ of any BL Lac object in a given catalog. Then we compare this number with the expectation n_e for the null hypothesis:

$$n_e = N_{\rm BL} N_U \frac{1 - \cos\delta}{1 - \cos(-10°)}, \tag{1}$$

where $N_{\rm BL}$ is the number of BL Lac objects in the catalog, $N_U = 58$ is the number of UHECRs, and $-10°$ is the declination cut on the BL Lac objects. Note that this expectation implies an isotropic distribution for at least one sample. This is not the case, because the AGASA sample has a latitude anisotropy and the BL Lac catalog is anisotropic with respective to the Galactic plane (selection effect) and large-scale cosmological structure. A more accurate estimate differs from that given by equation (1) by a factor

$$F = \frac{\sum_{i=1}^{N_{\rm BL}} \xi(\theta_i)}{N_{\rm BL}\langle\xi\rangle}, \tag{2}$$

where $\xi(\theta)$ is the AGASA exposure function. The exposure function depends on particle energy and is hardly known better than one can extract from the latitude distribution of detected UHECRs. Takeda et al. (1999) used a polynomial fit to the observed latitude distribution of events above 10^{19} eV. We prefer to use the observed distribution of the available AGASA sample (above 4×10^{19} eV) in the form of a histogram in $\cos\theta$ with a bin width of 0.1, since this is the simplest option that can be easily reproduced.

The factor F depends on the BL Lac catalog and therefore on the cuts. According to our estimates with equation (2), F is close to 1 for radio-bright objects and $\sim$1.2 for optically bright objects (probably because of the anisotropy caused by Galactic absorption). We introduce a measure of the signal, p (which depends on δ and cuts in the BL Lac catalog), as the probability of sampling n or more hits from the Poisson distribution at expectation Fn_e.

Note that for autocorrelated samples the distribution of n is not Poissonian, and therefore this measure is not exact. In order to correct this probability for the actual autocorrelated distribution of BL Lac objects on the sky, we perform Monte Carlo simulations using a large number of random UHECR samples. The maximal disagreement between the Poisson and Monte Carlo probabilities is by a factor of 2. Thus, we use the Poisson probability p for preliminary estimates and recalculate the probability for the leading samples (given in Table 1) with Monte Carlo simulations.

3. SEARCH FOR THE BEST-CORRELATED SAMPLE AND RESULTS

By optimizing cuts in all existing parameters, we can fit a BL Lac catalog to any set of locations in the sky so that it will demonstrate a highly significant correlation (see § 4). Therefore, if our objective is to find the real sample, we have to try the most relevant cuts. The apparent radio or optical brightness of objects (represented in the catalog by their observed radio

TABLE 1
Samples of UHECRs Demonstrating the Most Significant Correlation with BL Lacertae Objects

Brightness Cut	ID[a]	N_{obj}[b]	C_r or C_o[c] (Jy or V)	δ[d] (deg)	p[e] ($\times 10^4$)
All Quasars					
Radio	90	256	0.04	2	2.6
	40	139	0.16	3	3.2
	11	35	0.79	2	5
Optical	90	153	17.5	2	3.12
Confirmed BL Lac Objects					
Radio	11	6	0.79	2	1.1
	4	197	0.02	1.5	3.47
	90	6	0.79	3	8
Optical	4	118	18	1.5	1.15

[a] Identification number of the sample giving the strongest correlation signal.
[b] Number of objects passing the cut.
[c] Optimal cut in 6 GHz radio flux or visual magnitude.
[d] Optimal correlation angle.
[e] Significance level.

flux density measured in janskys and the visual magnitude V) seems to be a good indicator of particle acceleration to ultrahigh energies. To avoid "overoptimization" of random samples in a two-dimensional scan, we performed two separate passes:

1. We optimized cut C_r in the 6 GHz radio flux within the limits 0.01 Jy $< C_r <$ 2 Jy, varying it by steps of 0.1 in decimal logarithm. No cuts in optical brightness were applied.
2. We optimized cut C_o in visual magnitude within the range from $V = 12$ to $V = 24$ with a step of $\Delta V = 0.5$. No cuts in radio flux were applied, and we excluded objects with no radio brightness data.

The proper correlation angle δ is somewhat uncertain. The most significant correlation certainly should not appear at a correlation angle equal to the 1 σ experimental error (the latter depends on the particle energy). If UHECRs are charged, then the correlation could appear at a δ corresponding to the typical angle of particle deflection. We optimized δ between 1°.5 and 5° with a step of 0°.5. The samples that yielded the most significant correlations are listed in Table 1. In addition, we also tried a scan over the intrinsic radio luminosity as was done in TT01. The strongest effect was exhibited by sample 11: $p = 4 \times 10^{-4}$ with the 25 intrinsically brightest BL Lac objects and $\delta = 3°$.

With these results in hand, the investigator had to make a choice concerning the real sample. All the best samples (except No. 4) had a reasonable value for the optimal δ (2° and 3°), which is close to the AGASA angular resolution of 2°.3. Finally, the investigator took sample 11 as the first choice. The second option was sample 90.

The second task is to test for autocorrelation of the UHECR arrival directions. This was performed with the same array of random samples, before the investigator was informed about the results of his choices in the first test. The autocorrelation signal is estimated in a similar way as described above for the cross-correlation signal:

$$n_e = \frac{N_U(N_U - 1)}{2}\frac{1 - \cos\delta}{1 - \cos(-10°)}, \quad F = \frac{\sum_{i=1}^{N_U} \xi(\theta_i)}{N_U\langle\xi\rangle}, \quad (3)$$

where the factor $F = 1.4$.

Now, sample 67 showed the maximum signal, $p = 0.5 \times 10^{-3}$ at $\delta = 2°.5$ (eight hits). The second sample showing a strong autocorrelation was No. 30, with $p = 1.7 \times 10^{-3}$ at $\delta = 2°.1$. The choice of the investigator was sample 67.

The real observed sample of UHECRs was No. 67. Therefore, the test at the 99% confidence level was unsuccessful for a UHECR–BL Lac correlation and successful for a UHECR autocorrelation. We then checked sample 67 for a cross-correlation with BL Lac objects by varying C_r and did not find any significant signal.

4. INTERPRETATION OF THE RESULTS

We can confirm that the autocorrelation signal in the AGASA sample with the given energy threshold has a significance of at least 10^{-2}. To determine the level of significance, we would have to vary the size of the random array and find the limit at which we are able to find the real sample. This objective is beyond the scope of this work. According to the correlation signal in the second-best sample, the significance is probably around 3×10^{-3}, in agreement with Finley & Westerhoff (2004). One should note, however, that this result refers to a specific sample with an energy cut of 4×10^{19} eV (see Finley & Westerhoff 2004 for a discussion). To estimate the significance of the real autocorrelation, one has to perform the same procedure with an untruncated sample of UHECRs, varying the energy cut over a reasonable range.

Our negative result on the cross-correlation with BL Lac objects does not mean that we have found a quantitative disagreement with the results of TT01. They found a positive signal with another catalog of confirmed BL Lac objects. Their cuts were $z > 0.1$ or unknown, $C_r = 0.17$ Jy, and $C_o = 18$ mag. With these cuts, a positive signal still exists at $p = 1.9 \times 10^{-2}$ and $\delta = 2°.5$ (with factor $F = 1.24$; see eq. [2]), and the real sample (No. 67) is the second most significant among the 99 random samples (having a similar significance as three other samples, including No. 11).

We have just demonstrated that using the most straightforward assumptions, blindly, one can hardly find the correlation signal. Regarding more specific cuts, such as those in TT01, one encounters the problem of interpreting the signal—whether it is real or just a consequence of cuts optimization (see also Evans et al. 2003, 2004). The claim that a given cut was motivated independently rather than being optimized is not convincing unless the motivation has been established a priori.

Now let us demonstrate how multiple-cuts optimization can actually mimic a significant signal. In this demonstration, we use 10^4 random UHECR samples prepared as described in § 2.2 and the BL Lac catalog with cuts optimized for each random sample. Figure 1 shows the fraction of random samples η that demonstrated a "significance of correlation" higher than p, after cuts optimization. If we fix all the cuts (curve 1), then there is an approximate agreement between η and p. If we optimize one cut, C_r, then we obtain η a few times greater than p (actually, the ratio η/p can be interpreted as the penalty factor discussed above). With two-cut optimization, adding a scan over visual magnitude, the ratio η/p reaches almost 2 orders of magnitude and one out of five samples demonstrates $p < 0.01$. If we add an optimization for the correlation angle δ, then every third random sample demonstrates a "significance" of 10^{-2}, every 10th gives $p < 10^{-3}$, and one out of 1000 gives $p = 10^{-6}$!

5. SUMMARY

We have presented a method of blind search for a hypothesized effect in which various trials with different subsamples

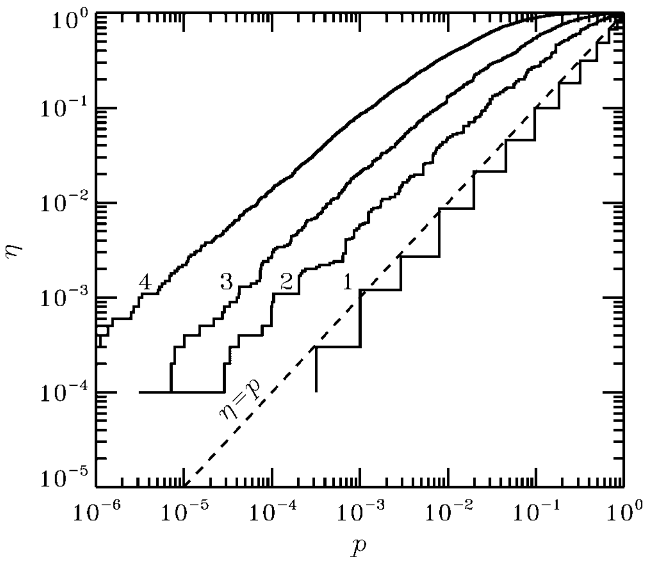

FIG. 1.—Fraction of 10^4 simulated random UHECR samples, η, demonstrating a higher significance level for the "correlation signal" with the BL Lac catalog of Véron-Cetty & Véron (2003; 876 objects) than a given value p for different optimized cuts. From lowest to highest curve, (1) no cuts optimization, with $C_r = 0.2$ Jy, $\delta = 2^\circ.5$, and no cuts in optical brightness; (2) optimization in C_r with $\delta = 2^\circ.5$ and no cuts in optical brightness; (3) optimization in both C_r and C_o with $\delta = 2^\circ.5$; (4) optimization in C_r, C_o, and δ.

or model parameters do not affect the stated significance level. We believe that a tradition of using this method, when possible, would dramatically reduce the number of unconfirmed claims of marginally significant effects. The method is especially useful when (1) there is a clear null hypothesis and a way to prepare random samples representing it, (2) there exists a convenient measure of the statistical significance of the effect, and (3) the effect is uncertain in some respects—otherwise, a test with the blind a priori formulation (i.e., it is a priori clear which data should be used and how the effect should look) is sufficient. Such problems as searches for cross-correlation between two classes of astrophysical objects usually satisfy all three conditions. We would emphasize that the proposed method is, in principle, applicable in any field of science.

In this work we performed a demonstration for only one size of the array of random samples. To find the level of significance of the effect, one should make several trials with different array sizes, starting from a larger one and then reducing its size until the real sample is found. The examiner should not disclose the real sample after an unsuccessful trial.

An effect detected with this method is credible because the method insures the researcher against unintentional overestimation of its significance. The only possible source of error that can mimic a positive result is a wrong null hypothesis distinguishing random samples from the real sample. In the case considered in this Letter, this could be, for example, a wrong exposure function for the UHECR detector. Otherwise, a positive result would have an explicit meaning: the chance that the effect does not exist is the inverse of the size of array of samples in the successful search.

As an application of the proposed method, we analyzed a possible correlation between UHECRs and BL Lac objects. We found no significant correlation, but we cannot claim, of course, that no such correlation exists.

We are grateful to P. Tinyakov and I. Tkachev for useful discussions, and to the anonymous referee for numerous useful suggestions. The work is supported by RFBR grant 04-02-16987, Academy of Finland grants 80750 and 102181, the Jenny and Antti Wihuri Foundation, the Vilho, Yrjö, and Kalle Väisälä Foundation, and the NORDITA Nordic project in High Energy Astrophysics.

REFERENCES

Belli, B. M. 1997, ApJ, 479, L31
Evans, N. W., Ferrer, F., & Sarkar, S. 2003, Phys. Rev. D, 67(10), 103005
———. 2004, Phys. Rev. D, 69(12), 128302
Finley, C. B., & Westerhoff, S. 2004, Astropart. Phys., 21, 359
Hayashida, N., et al. 1996, Phys. Rev. Lett., 77, 1000
———. 2000, preprint (astro-ph/0008102)
Kolatt, T., & Piran, T. 1996, ApJ, 467, L41
Komberg, B. V., Kurt, V. G., & Tikhomirova, Ya. Yu. 1997, Ap&SS, 252, 465
Takeda, M., et al. 1999, ApJ, 522, 225
Tinyakov, P. G., & Tkachev, I. I. 2001, JETP Lett., 74, 445 (TT01)
Torres, D. F., Boldt, E., Hamilton, T., & Loewenstein, M. 2002, Phys. Rev. D, 66(2), 023001
Uchihori, Y., Nagano, M., Takeda, M., Teshima, M., Lloyd-Evans, J., & Watson, A. A. 2000, Astropart. Phys., 13, 151
Uryson, A. V. 2004, Astron. Rep., 48, 81
Véron-Cetty, M.-P., & Véron, P. 2003, A&A, 412, 399
Virmani, A., Bhattacharya, S., Jain, P., Razzaque, S., Ralston, J. P., & McKay, D. W. 2002, Astropart. Phys., 17, 489