A "search for the real sample" method for a blind statistical test: Application to the origin of ultra-high energy cosmic rays

Boris E. Stern[1,2,3], and Juri Poutanen[3]

October 18, 2004

[1]Institute for Nuclear Research, Russian Academy of Sciences,
7a, Prospect 60-letija Oktjabrja, Moscow 117312, Russia
[2]Astro Space Center of Lebedev Physical Institute,
Profsoyuznaya 84/32, Moscow 117997, Russia
[3]Astronomy Division, P.O.Box 3000, 90014 University of Oulu, Finland

Abstract

We suggest a method for statistical tests which does not suffer from *a posteriori* manipulations with tested samples (e.g. cuts optimization) and does not require somewhat obscure procedure of the penalty estimate. The idea of the method is to search the real sample among a large number of random samples representing the null hypothesis as one demonstrating the strongest hypothesized effect. The statistical significance of the effect in this approach is just the inverse of the maximal number of random samples at which the search was successful. We have applied the method to revisit the problem of correlation between the arrival directions of ultra-high energy cosmic rays and BL Lac objects.

1 Introduction

Communications about effects detected at a marginally significant level constitute a considerable fraction of all scientific results. The scientific society usually treats such communication with a great deal of skepticism. Indeed, too many marginally significant effects have not withstand the data accumulation.

The reason for this is quite evident: a 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 more or less statistically independent tests of the same effect. This numerous trials, again, increase the probability to observe a signal in one of the trials by chance.

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

In principle, the researcher can introduce a "penalty factor" accounting for these numerous trials. The most reliable way to estimate the significance is to use random samples, representing the null hypothesis. The investigator can prepare a large array of random samples and do the same estimate of the effect for each of these samples as he does for the real sample in each statistical trial. Then, the significance can be defined as the fraction of samples which demonstrated at *some* trial a more significant "effect" than the real sample at the *most successful* trial.

The above procedure is sufficient if (i) the investigator follows the above procedure precisely; (ii) the investigator does not use the *a posteriori* information on the real sample for the planning of the investigation strategy.

We would like to notice that both conditions are not so easy to satisfy once the investigator studied the real sample and feels which combination of cuts or model parameters will provide the most significant signal. Then he can find the most favorable trial intuitively, avoiding a large number of extra trials which could demonstrate a positive signal for one of random

samples rather than for the real sample. In other terms, the investigator can severely underestimate the "penalty factor" and therefore to overestimate the significance of the effect using *a posteriori* knowledge. We should emphasize that the investigator can introduce such bias not deliberately.

We suggest a simple way how to avoid this "pressure" of the *a posteriori* information. The investigator can get rid of the latter by hiding the real sample inside a large array of random null hypothesis samples. Then the problem has to be inverted: the investigator must find the real sample in the array using the hypothesis rather than to confirm the hypothesis using the real sample. On one hand, this is a kind of a blind test: the investigator does not know where is the real sample. On the other hand, he can feel free to optimize the hypothesis. Just the objective of the optimization will be slightly different: one needs to find the combination of parameters where one of the samples will demonstrate the strongest effect. Then the investigator could make his bet in favor of the dominating sample and the label of the real sample can be disclosed. If the investigator finds the real sample, the significance of the effect is just the inverse of the number of samples in the array.

High energy astrophysics gives a number of instructive examples of searches for marginally significant effects. Indeed, there are many detection of particles or transient gamma-ray events which sources are unknown. This stimulates intensive searches of various correlations between different classes of events and objects. For example, there is a number of works reporting detections of correlation between locations of gamma-ray bursts (or their subsamples) and various objects: galaxy clusters (Kolatt & Piran 1996), galaxy plane (Belli,1997) the local galactic arm (Komberg, Kurt & Tikhomirova 1997). We will not discuss these results and will concentrate on ultra-high energy cosmic rays (UHECRs) and searches for their hypothetic sources. A claim of significant autocorrelation in the arrival directions of UHERCs detected by the Akeno Giant Air Shower Array (AGASA) (Hayashida et al. 1996; Takeda et al. 1999) motivated searches of cross-correlations between UHECRs and various astrophysical objects. Particularly there were reported statistically significant cross-correlation signals between UHECRs and BL Lac objects (Tinyakov & Tkachev 2001, hereafter TT01), between UHECRs (more precisely - clusters in UHECRs) and the super-galactic 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 abundance of reported cross-correlation signals with quite different objects rises a doubt that any of these signals is real. Below we concentrate only on UHECRs – BL Lacs correlation: our main objective is the illustration of the method rather than a search for the correlation.

2 Procedure

2.1 Catalogs

We used the AGASA sample of UHECRs with 58 events above 4×10^{19} eV and a catalog of Véron-Cetty & Véron (2003) containing 876 BL Lac objects. We do not combine the AGASA sample with the data from other experiments because other samples are smaller and problems associated with the non-uniform structure of a joint sample would overweight the statistical gain. The BL Lac catalog has been cut in declination at -10^o and was subject to various brightness cuts. We also tried a sub-catalog of *confirmed* BL Lacs which includes 491 objects. Actually we do not know which catalog, the entire Véron-Cetty & Véron (2003) BL Lac catalog or its confirmed sub-catalog is more relevant. We believe that, e.g., a radio-bright source suspected to be a BL Lac, even if not confirmed, is a relevant object for the test due to its radio brightness, because the latter is a signature of particle acceleration. On the other hand, TT01 used a confirmed sub-catalog and it would be interesting to try the same approach too.

2.2 Null hypothesis and random samples

Null hypothesis in our case is just the isotropic distribution of arrival directions of UHECRs 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 avoiding possible uncertainties in the latitude exposure function: to sample the right ascension uniformly keeping the actually observed declination for each event. We, nevertheless, have dispersed the declinations of UHECRs by $\pm 3^o$ around their real values in order to destroy possible small-scale latitude correlations, if the latter exist. Such small dispersion does not distort a much wider exposure function.

When performing the test we have distributed roles: one of the coauthors

acts as an "investigator", another plays a role of "examiner". Examiner has prepared an array of 99 random samples as described above and inserted the real sample into the array keeping the sequential real sample number in secret from the investigator.

2.3 Measure for the correlation signal

We used usual two-point correlation function counting the number n of UHECRs within angle δ from any BL Lac of a given catalog. Then we compare this number with expectation $n_{\rm e}$ for the null hypothesis:

$$n_{\rm e} = N_{\rm BL} N_{\rm U} \frac{1-\cos\delta}{1-\cos(-10^o)}, \tag{1}$$

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

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

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) use a plynomial 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 $4\cdot10^{19}$ eV) in a form of histogram in $\cos\theta$ with the bin width 0.1 since this is the simplest option that can be easiely reproduced by any researcer.

Factor F depends on the BL Lac catalog and therefore on cuts. According to our estimates with Eq. (2) F is close to 1 for radio-bright objects and ~ 1.2 for optically-bright objects (probably due to anisotropy caused by galactic absorption). We introduce the measure of the signal, p (which depends on δ and cuts in the BL Lac catalog), as the probability to sample n or more hits from the Poisson distribution at expectation $Fn_{\rm e}$.

Note that for autocorrelated samples the distribution of n is not Poisson, therefore our measure is, in principle, not exact. It was verified with Monte-Carlo simulations using a large numbe of random UHECR samples and we find that maximal disagreement between Poisson and Monte-Carlo probability for BL Lac subsamples is by factor 2. Finaly, we use Poisson probability for preliminary estimate and recalculate the probability for leading samples summarized in table 2 with Monte-Carlo simulation.

3 Search for the best-correlating sample and its results

Optimizing cuts in all existing parameters of objects 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 most relevant cuts. The most relevant parameters are the apparent radio- or non-thermal optical brightness of objects (represented in the catalog by their observed radio flux density measured in Jy and the visual magnitude V). We assume that the main fraction of optical BL Lac emission is nonthermal while actually a component of the luminosity can come from thermal emission of accretion disk. Apparent brightness is preferable to intrinsic luminosity because the flux of hypothetical UHECRs scales with distance in the same way as the radio flux or even faster, if we deal with charged particles which are deflected by intergalactic magnetic field.

The non-thermal optical and radio emission are indicators of particle acceleration in an object. Both fluxes generally correlate with each other, however, there exists a set of radio-bright and optically-dim objects, maybe due to the absorption in optics. Theoretically, it is not clear which kind of emission is a better indicator of particle acceleration to ultrahigh energies and we tried optimization in both radio brightness and optical brightness cuts. To avoid "over-optimization" of random samples in two-dimensional scan, we performed two separate scans:

1. We optimized cut $C_{\rm r}$ in the radio flux at 6 GHz within the limits 0.01 Jy $< C_{\rm r} <$ 2 Jy, varying it with the step 0.1 in decimal logarithm. No cuts in optical brightness was applied. This scan is marked with letter R in Table 1.

2. We optimized cut $C_{\rm o}$ in visual magnitude within the range from $V = 12$ to $V = 24$ with the step $\Delta V = 0.5$. No cuts in radio flux was applied and we excluded objects with no data on their radio brightness. This scan is marked with letter O in Table 1.

The proper correlation angle δ is somewhat uncertain. The most significant correlation should not certainly appear at a correlation angle equal to 1σ experimental error (the latter depends on the particle energy). If UHECRs are charged, then the correlation could appear at δ corresponding to a typical angle of particle deflection. We optimized δ between $1\overset{\circ}{.}5$ and 5^o with the step $0\overset{\circ}{.}5$. Cases of the most significant correlation are summarized in Table 1. We also tried a scan over the intrinsic radio luminosity. The strongest effect gave sample #11: $p = 4 \times 10^{-4}$ with 25 intrinsically brightest BL Lac objects and $\delta = 3\overset{\circ}{.}$.

With these results at hand, the investigator had to make a bet concerning the real sample. First of all, it is clear that the effect of correlation is marginally significant at best since the difference in the significance levels between the best and the second best samples is moderate (factor of 4). All best samples (except #4) have a reasonable value of optimal δ (2^o and 3^o), which is close to the angular resolution of AGASA of $2\overset{\circ}{.}3$. Finally, the "investigator" used sample #11 as the first bet. The second option was sample #90.

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

$$n_{\rm e} = \frac{N_{\rm U}(N_{\rm U}-1)}{2}\frac{1-\cos\delta}{1-\cos(-10^o)}, \quad F = \frac{\Sigma_{i=1}^{N_{\rm U}}\,\xi(\theta_i)}{N_{\rm U}\langle\xi\rangle}, \tag{3}$$

where factor $F = 1.4$.

Now, sample #67 showed maximum significance of $p = 0.5 \times 10^{-3}$ at $\delta = 2\overset{\circ}{.}5$ (8 hits). The second sample showing strong autocorrelation was #30 with $p = 1.7 \times 10^{-3}$ at $\delta = 2\overset{\circ}{.}1$. The bet of the investigator was #67.

The real sample number was #67. Therefore the test at 99% confidence level was unsuccessful for UHECRs–BL Lacs correlation and successful for UHECRs autocorrelation. Then we checked sample #67 for the cross-correlation with BL Lacs by varying $C_{\rm r}$ and have not found any significant signal.

Table 1: Samples demonstrating the most significant correlation with BL Lacs.

R or O	ID	$N_{\rm obj}$	$C_{\rm r}$ or $C_{\rm o}$ (Jy or V)	δ (deg)	$p \times 10^4$
			All quasars		
R	90	256	0.04	2	2.6
R	40	139	0.16	3	3.2
R	11	35	0.79	2	5
O	90	153	17.5	2	3.12
			Confirmed BL Lacs		
R	11	6	0.79	2	1.1
R	4	197	0.02	1.5	3.47
R	90	6	0.79	3	8
O	4	118	18	1.5	1.15

Two catalogs of quasars and active nuclei are considered: all objects from the Véron-Cetty & Véron (2003) catalog and a sub-catalog containing only confirmed BL Lac objects. R or O represent the applied cut (in radio or optical brightness), ID is the identification number of a sample giving the strongest correlation signal, $N_{\rm obj}$ is the number of objects passing the cut, $C_{\rm r}$ or $C_{\rm o}$ is the optimal cut in radio flux at 6 GHz or visual magnitude for a given sample, δ is the optimal correlation angle, p is the significance level. Only cases with $p < 10^{-3}$ are presented.

4 Interpretation of the results

We can confirm that the autocorrelation signal in AGASA sample with the given energy threshold has a significance of at least 10^{-2}. To find the significance level we would have to vary the size of the random array and to find the limit when we are able to find the real sample. This objective is beyond the scope of this work. Probably, according to the correlation signal in the second best sample, the significance is around 3×10^{-3} in agreement with Finley & Westerhoff (2003). One should notice, however, that this result refers to a specific sample with the energy cut of 4×10^{19} eV (see Finley & Westerhoff 2003, for the discussion). To estimate the significance of real autocorrelation one has to perform the same procedure with an untruncated sample of UHECRs varying the energy cut in a reasonable range.

Our negative result on cross-correlation with BL Lacs does not mean of course that we have found a disagreement with the results of TT01. They have found a positive signal with another catalog of confirmed BL Lac. Their cuts were: $z > 0.1$ or unknown, $C_{\rm r} = 0.17$ Jy, $C_{\rm o} = 18^{\rm m}$. At these cuts the positive signal still exists at the significance of 1.9×10^{-2} and $\delta = 2\overset{\circ}{.}5$ (with factor $F = 1.24$, see Eq. 2) and the real sample #67 is the second significant among 99 random samples together with three other of the same significance including sample #11. Note that some discrepancy between results of different authors at the fixed set of cuts can appear from different treatment of the AGASA exposure function which affects factor F. If one sets $F = 1$, then $p = 1.0 \times 10^{-2}$.

We just demonstrated that using the most straightforward assumptions, blindly, one can hardly find the correlation signal. Regarding more specific cuts, like in TT01, one meets a problem of interpretation of the signal whether it is real or is just a consequence of cuts optimization. The claim, that a given cut was motivated independently rather than optimized, is not convincing unless the motivation has been done *a priori*.

Now let us demonstrate how the multiple cuts optimization can actually mimic a significant signal. In this demonstration we use 2000 random UHECRs samples prepared as described in Sect. 2.2 and the BL Lac catalog with cuts, optimized for each random sample. Fig. 1 shows the fraction of random samples η which 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, $C_{\rm r}$, then we obtain η a few times greater than p (actually, the ratio η/p can be inter-

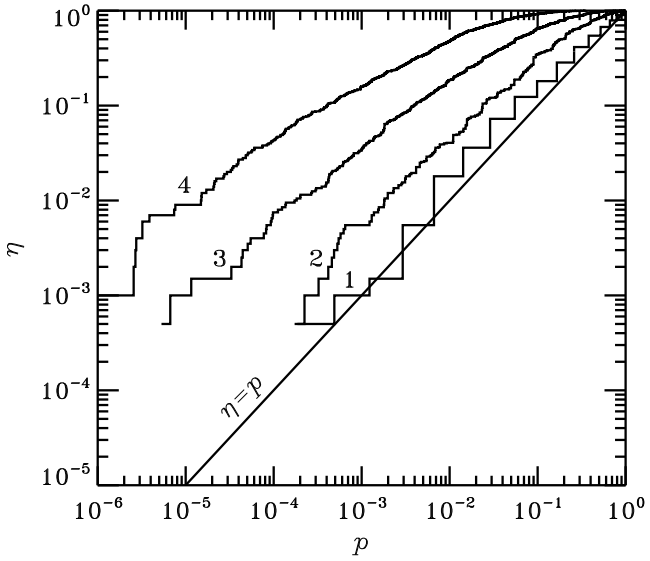

Figure 1: The fraction of 2000 simulated random UHECRs samples, η, demonstrating a higher significance level for the "correlation signal" with the BL Lac catalog of Véron-Cetty and Véron (2003) (876 objects) than a given value p for different cut optimization. From lower to higher curves: 1 – no cuts optimization with $C_{\rm r} = 0.2$ Jy, $\delta = 2.^\circ5$, no cuts in optical brightness; 2 – optimization in $C_{\rm r}$ with $\delta = 2.^\circ5$ and no cuts in optical brightness; 3 – optimization in both $C_{\rm r}$ and $C_{\rm o}$ with $\delta = 2.^\circ5$; 4 – $C_{\rm r}$, $C_{\rm o}$ and δ optimization.

preted as the penalty factor discussed above). With two cuts optimization, adding a scan over visual magnitude, the ratio η/p reaches almost two orders of magnitude and one out of 5 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 tenth gives $p < 10^{-3}$, and one out of thousand gives $p = 10^{-6}$!

5 Summary

We presented a method of a blind search for a hypothetic effect where various trials of the researcher with different subsamples or model parameters do not affect the stated significance level. In most cases a manipulation with data sets leads to the overestimate of the significance. The method still allow

such manipulation which is unavoidable if one does not know exactly which fraction of data should display the effect most prominentely. We believe that a tradition to use this method when possible would dramatically reduce the number of unconfirmed claimes of marginally significant effects.

The method is useful when: (i) there is a clear null hypothesis and a way to prepare random samples representing it; (ii) there exists a convenient measure of the statistical significance of the effect; (iii) the effect is uncertain in some respects, otherwise a test with the blind *a priory* formulation (i.e. it is *a priory* 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 like to 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. Actually if the objective is to find the significance level of the effect, one should make several trials with different array size starting from a larger array, then reducing its size until the real sample is found. The examiner should not disclose the real sample after unsuccessful trial.

An effect detected with this method is credible because it ensures a researcher against not deliberate overestimation of the significance. The only possible source of errors that can mimic a positive result is a wrong null hypothesis distinguishing random samples from the real sample. In the case considered in this paper, this could be for example a wrong exposure function of 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 at the successful search.

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

References

[1] Belli B. M., 1997, ApJ, 479, L31

[2] Finley C. B., Westerhoff S., 2004, Astroparticle Physics, 21, 359

[3] Hayashida N. et al. 1996, PhRevLett, 77, 1000

[4] Kolatt T., Piran T., 1996, ApJ, 467, L41

[5] Komberg B. V., Kurt V. G., Tikhomirova Ya. Yu., 1997, Ap&SS, 252, 465

[6] Takeda M. et al., 1999, ApJ, 522, 225

[7] Tinyakov P. G., Tkachev I. I., 2001, JETP Lett, 74, 445 (TT01)

[8] Torres D. F., Boldt E., Hamilton T., Loewenstein M., 2002, PhRvD, 66, 23001

[9] Uchihori Y., Nagano M., Takeda M., Teshima M., Lloyd-Evans J., Watson A. A, 2000, Astroparticle Physics, 13, 151

[10] Uryson A. V., 2004, Astronomy Reports, 48, 81

[11] Véron-Cetty M.-P., Véron P., 2003, A&A, 412, 399

[12] Virmani A., Bhattacharya S., Jain P., Razzaque S., Ralston J. P., Mckay D. W., 2002, Astroparticle Physics, 17, 489

[13] Yoshiguchi H., Nagataki S., Sato K., 2004, astro-ph/0404411