Anisotropy studies with multiscale autocorrelation function

M. De Domenico\textsuperscript{ab}, H. Lyberis\textsuperscript{cd}

\textsuperscript{a}Laboratorio sui Sistemi Complessi, Scuola Superiore di Catania, Catania
\textsuperscript{b}Istituto Nazionale di Fisica Nucleare, Sez. di Catania, Catania (Italy)
\textsuperscript{c}CNRS/IN2P3 - IPN Orsay, Paris, France
\textsuperscript{d}Dipartimento di Fisica, Università di Torino, Torino, Italy

We present a novel method, based on a multiscale approach, for detecting anisotropy signatures in the arrival direction distribution of the highest energy cosmic rays. This method is catalog independent, i.e. it does not depend on the choice of a particular catalog of candidate sources, and it provides a good discrimination power even in presence of contaminating isotropic background. We present applications to simulated data sets of events corresponding to plausible scenarios for events detected, in the last decades, by world-wide surface detector-based observatories for charged particles.

1. Introduction

In many field involving data analysis, the search for anisotropy has played a crucial role. Many estimators, namely two-point correlation functions \[1–4\], have been proposed and widely used to search for clustering of objects and to measure deviation from isotropy of angular distributions. A non-differential variant of such an estimators is widely used for small data set of points \[5–9\]. These methods apply to angular coordinates of objects as well to distributions of arrival directions of events: in this work we will indifferentely refer to both as arrival direction distributions of events.

Recently, new estimators have been introduced to study the anisotropy signature of sky’s arrival direction distributions: the modified two-point Rayleigh \[10\], and shape-strength method derived from a principle component analysis of triplets of events \[11\]. Such a test statistics have been recently applied to both P. Auger data and to synthetic maps of events, the latter generated by sampling the Veron-Cetty & Veron catalog \[12\] of nearby candidate active galactic nuclei (AGN) within 75 Mpc \((z \leq 0.018)\), showing an higher discrimination power than other estimators \[13\].

Within the present work, we introduce a new fast and simple method for anisotropy analysis, which makes use of a multiscale approach and depends on one parameter only, namely the angular scale of the intrinsic anisotropy. The main advantage of our estimator is the possibility to analytically treat the results: the analytical approach drastically reduces computation time and makes available the possibility of applications to very large data sets of events. We test the method on several simulated isotropic and anisotropic arrival direction distributions (mock maps) and perform an extensive analysis of its statistical features under both the null and the alternative hypotheses. However, it is worth remarking that the scope of applicability of our method is not limited to UHECR physics, and it is valid for any distribution of angular coordinates of objects.

2. Multiscale Autocorrelation Function

Let \(S\) be a region of a spherical surface and let \(P_i (\phi, \theta) (i = 1, 2, ..., n)\) be a set of points locating \(n\) arrival directions on \(S\), defining a sky. The sky \(S\) is partitioned within a grid of \(N\) equal-area (and almost-equal shape) disjoint boxes \(B_k (k = 1, 2, ..., N)\) as described in Ref. \[14\]. Let \(\Omega\)
be the solid angle covered by \( S \), whereas each box \( B_k \) covers the solid angle

\[
\Omega_k = \frac{1}{N} \int_{\theta_{\min}}^{\theta_{\max}} \int_{\phi_{\min}}^{\phi_{\max}} d\cos\theta d\phi = 2\pi (1 - \cos \Theta)
\]

where \( 2\Theta \) is the apex angle of a cone covering the same solid angle: \( N, \Theta \) and \( \Omega \) are deeply related quantities that define a scale.

Let \( \psi_k(\Theta) \) be the density of points in the data set that fall into the box \( B_k \): the function \( A(\Theta) \) that quantifies the deviation of data from an isotropic distribution at the scale \( \Theta \), is chosen to be the Kullback-Leibler divergence \(^{13, 16}\)

\[
A(\Theta) = \sum_{k=1}^{N} \psi_k(\Theta) \log \frac{\psi_k(\Theta)}{\bar{\psi}_k(\Theta)} \tag{1}
\]

where \( \bar{\psi}_k(\Theta) \), generally a function of the domain meshing, is the expected density of points isotropically distributed on \( S \) falling into the box \( B_k \). The Kullback-Leibler divergence is an information theoretic measure widely used in hypothesis testing and model selection criteria, statistical mechanics, quantum mechanics, medical and ecological studies (see \(^{17}\) and Ref. therein). This measure quantifies the error in selecting the density \( \psi(\Theta) \) to approximate the density \( \bar{\psi}(\Theta) \): it can be shown that minimizing the Kullback-Leibler divergence is equivalent to maximum likelihood estimation \(^{17}\). It is straightforward to show that \( A(\Theta) \) is minimum for an isotropic distribution of points, or, in general, when \( \psi(\Theta) \sim \bar{\psi}(\Theta) \), i.e. if the model is correct.

If \( A_{\text{data}}(\Theta) \) and \( A_{\text{iso}}(\Theta) \) refer, respectively, to the data and to an isotropic realization with the same number of events, we define multiscale autocorrelation function (MAF) the estimator

\[
s(\Theta) = \frac{|A_{\text{data}}(\Theta) - \langle A_{\text{iso}}(\Theta) \rangle|}{\sigma_{A_{\text{iso}}}(\Theta)} \tag{2}
\]

where \( \langle A_{\text{iso}}(\Theta) \rangle \) and \( \sigma_{A_{\text{iso}}}(\Theta) \) are the sample mean and the sample standard deviation, respectively, estimated on several isotropic realizations of the data. If \( \mathcal{H}_0 \) denotes the null hypothesis of an underlying isotropic distribution for the data, the chance probability at the angular scale \( \Theta \), properly penalized because of the scan on \( \Theta \), is the probability

\[
p(\Theta) = \Pr (s_{\text{iso}}(\Theta') \geq s_{\text{data}}(\Theta)|\mathcal{H}_0, \forall \Theta' \in \mathcal{P}) \tag{3}
\]

obtained from the fraction of null models giving a multiscale autocorrelation, at any angular scale \( \Theta' \) in the parameter space \( \mathcal{P} \), greater or equal than that of data at the scale \( \Theta \). The null hypothesis is eventually rejected in favor of the alternative \( \mathcal{H}_1 = -\mathcal{H}_0 \) — being \( - \) the negation operator — at the angular scale \( \Theta \), with probability \( 1 - p(\Theta) \).

Under the null hypothesis \( \mathcal{H}_0 \), the estimator \( s(\Theta) \) follows a half-gaussian distribution, independently on the value of the angular scale \( \Theta \) and on the number of events on \( S \)^{17}.

### 3. Dynamical Boxing

The simplest definition of the boxing algorithm, as shortly described in Sec. \( \textbf{2} \) involves the fixed grid introduced in Ref. \(^{14}\), where each box only embodies the relative number of events falling in it. Unfortunately, such a static binning approach could not reveal an existing cluster. Indeed, the fixed grid may cut a cluster of points within one or more edges, causing a further loss of information at the angular scale under investigation. To overcome this possible loss of information, we introduced the dynamical binning, a type of smoothing of the grid by applying it on the data \(^{17}\).

The smoothing, adopted in our study, deals with a new counting procedure for the estimation of the density \( \psi_k(\Theta) \). However, such a density depends on the observatory’s exposure, generally a function on the celestial sphere depending on both the latitude of the experiment and the maximum zenith of detection, quantifying the effective time-integrated detection area for the flux of particles from each observable sky position. The relative exposure \( \omega \) is the dimensionless function corresponding to the exposure normalized to its maximum value. For a single full-time operating detector, i.e. with constant exposure in right ascension, fully efficient for particles arriving with zenith angles smaller than \( \theta_{\max} \), it still depends on the declination \( \delta \)^{18}. The proposed method takes into account the effect of non-uniform ex-
posure on the arrival direction distribution of objects, by weighting the angular region around a given direction with the local value of the exposure. Our numerical studies show that such a dynamical binning approach recovers the correct information on the amount of clustering in the data [17].

4. Interpretation of MAF

Any catalog-independent method provides information about the angular scale $\Theta^*$ where the significance is minimum. In the case of a simple two-point method, such an angular scale is quite difficult to interpret and topologically different configurations of events lead to the same significance. In the case of the modified two-point Rayleigh method, the estimation of the significance includes another set of parameters, independent from the angular distribution, as described in Ref. [10]: parameters are sensitive to the orientation of the pairs and therefore to skies showing preferential directions and filamentary structure of points. It follows that $\Theta^*$ is the most significant angular size for the mix of these informations, still linked to the pair configuration. In the case of the shape-strength method, the estimation of two parameters, namely the shape and strength, is performed: both can be interpreted, respectively, in terms of size and elongation of the triangles defined by a triplet of points. It follows that all information is recovered from the configuration of triplets.

In the specific case of MAF, the angular scale $\Theta^*$, where the significance is minimum, turns to be the significative clustering scale: it is the scale at which occurs a greater accumulation of points respect to that one occurring by chance, with no regard for a particular configuration of points, e.g. doublets or triplets. Fig. 1 shows the chance probability, as estimated by MAF, for three different point source mock skies contaminated by $80\%$ isotropic background. In each sky, $20\%$ of events are normally distributed, with dispersion $\rho$, around a single source, and $80\%$ of events are isotropically distributed.

5. Statistical analysis of MAF

In this section, we investigate the statistical features of MAF by inspecting its behavior under both the null or the alternative hypothesis. In particular, we estimate the significance $\alpha$ (or Type I error), i.e. the probability to wrongly reject the null hypothesis when it is actually true, and the power $1 - \beta$ (where $\beta$ is known as Type II error), i.e. the probability to accept the alternative hypothesis when it is in fact true.

Null hypothesis. We generate isotropic maps of $10^5$ skies, by varying the number of events from 20 to 500: for each sky in each map, we estimate the MAF for several values of the angular scale $\Theta$. Hence, we choose the value of $\Theta = \Theta^*$ where the chance probability is minimum, as the most significant clustering scale: $\tilde{p}(\Theta^*) = \arg \min_{\Theta} p(\Theta)$, properly penalized because of the scan on the parameter $\Theta$, according to the definition in Eq. [3]. Indipendently on the number of events in the

![Figure 1. MAF: chance probability versus the angular scale for three mock skies of 60 events. In each sky, 20% of events are normally distributed, with dispersion $\rho$, around a single source, and 80% of events are isotropically distributed.](image)
mock map, we find an excellent flat distribution of probabilities $\tilde{p}(\Theta^*)$, as expected for analyses under the null hypothesis $H_0$, i.e. the MAF is not biased against $H_0$.

Because of the definitions in Eq. (1), (2) and for the central limit theorem, an half-normal distribution is expected for the estimator $s(\Theta)$. We find an excellent agreement between the distribution for Monte Carlo realizations and the expected one. It follows that the (unpenalized) probability to obtain by chance a value of the MAF, greater or equal than a given value $s_0$, is just $1 - \text{erf} \left( \frac{s_0}{\sqrt{2}} \right)$, being erf the standard error function, independently on the angular scale $\Theta$ [17]. Although this nice feature of the MAF estimator, generally the distribution of $s_{\text{max}} = \max \{s(\Theta)\}$ is of interest for applications, because of the required penalization due to the scan over the parameter $\Theta$. Hence, it is important to identify the distribution of the penalized probability $p(\Theta)$, if any. Intriguingly, our numerical studies show that such a distribution exists and it corresponds to one of the limiting densities in the extreme value theory (see [17] and Ref. therein). In particular, the probability density of maxima is known as the Gumbel distribution [19, 20]:

$$g(x) = \frac{1}{\sigma} \exp \left[ - \frac{x - \mu}{\sigma} - \exp \left( \frac{x - \mu}{\sigma} \right) \right]$$  \hspace{1cm} (4)

where $\mu$ and $\sigma$ are the location and shape parameters, respectively.

In Fig. 2 are shown the probability densities of $s_{\text{max}}$ for $n = 40, 60, 80, 100$ and 500 events: independently on $n$, each density is in excellent agreement with the Gumbel distribution of extreme values, for the parameters $\mu = 1.743 \pm 0.002$ and $\sigma = 0.470 \pm 0.0022 \times 10^{-5}$.

**Alternative hypothesis.** In order to investigate the behavior of MAF under the alternative hypothesis of an underlying anisotropic distribution of objects, we generate anisotropic maps of $10^4$ skies, by varying the number of events from 20 to 100. In general, the anisotropy of a sky depends on several factors: for instance, in the case of cosmic rays, it depends on the distribution of sources, on magnetic fields and on propagation effects as energy loss or the GZK cutoff [21, 22] (and Ref. therein). Thus, a more complicated approach is required for the Monte Carlo realization of the maps. In order to estimate the power of MAF, we build reasonable anisotropic maps reflecting in part the real-world scenario, keeping in mind that our purpose is to build an anisotropic set of events for statistical analysis and not to generate events mimicking real data sets with the best available approximation. We proceed as follows:

1. **Catalog of candidate sources.** Although several models for production mechanisms of UHECR are available [21, 22, 23–30], it is generally accepted that the candidate sources are extragalactic and trace the distribution of luminous matter on large scales [31]. In particular, it has been shown that correlation with
possible high redshift sources is unlikely [32], whereas compact sources are favored [33, 34]; the recent result reported by the P. Auger Collaboration experimentally supports the latter claim, showing an high correlation between the observed data and the distribution of nearby active galactic nuclei (AGN) [35, 36]. For these reasons, we use the Palermo Swift-BAT hard X-ray catalogue of AGN with known redshift within 200 Mpc ($z \leq 0.047$) [37], as the reference catalog of candidate sources providing the most complete and uniform all-sky hard X-ray survey up to date.

2. **Source effects.** Events, from each source in the reference catalog, are generated by weighting for the source flux and for the expected geometrical flux attenuation. Hence, the number of events coming from a source is proportional to its flux and to the factor $z^{-2}$; because of the small scales and the high energy of cosmic rays involved in anisotropy studies ($E \geq 4.0 \times 10^{19}$ EeV), we assume a flat universe with zero cosmological constant ($\Omega = 1, \Lambda = 0$) and non-evolving source. Indeed, we naively take into account the possible deflections of the particles, due to the random component of the magnetic field, by producing arrival directions gaussianly-distributed with dispersion $\rho$ around the source. It is worth remarking that such a dispersion is strictly related to both the injection energy and the mass of the particle, as well as other physical quantities [21].

3. **Background.** We take into account the possibility for a contaminating isotropic background of the anisotropy signal, by generating a number of events isotropically distributed, corresponding to a fraction $f_{\text{iso}}$ of the whole data set.

4. **Detection effects.** As previously explained, the number of events detected by a single fully efficient and full-time operating surface detector, depends on its own relative exposure. In order to take into account such a detection effect, we generate the events according to the relative exposure of the single detector.

However, for a more realistic distribution of events, several more constraints, in general based on further assumptions or debated models, are required: the mass of the particle, the injection spectrum of the source, the intervening magnetic field, to cite some of the most important. In our study of the MAF discrimination power, we fix $\rho = 3^\circ$, as the mean angular deviation of UHECR in the galactic and extra-galactic magnetic field, and a background fraction $f_{\text{iso}} = 0.3$.

In order to produce a likely map of UHECR, we choose to generate events distributed in the whole sky, according to the number of events collected by surface detectors in the last decades. In particular, we consider events with energy $E \geq 4.0 \times 10^{19}$ EeV and error on the arrival direction smaller than $5^\circ$, as detected at the Sidney University Giant Airshower Recorder (SUGAR), Akeno Giant Air Shower Array (AGASA), Havera Park, Volcano Ranch, Yakutsk and P. Auger Observatory (we refer to [17] for a comprehensive list of references for the data released by each experiment). However, the fluxes of particles as measured by those experiments do not agree each other in the absolute fluxes, and a rescaling is needed [38]. By assuming that the spectrum reported by the HiRes Collaboration [39] corresponds to the correct energy scale, the rescaling, based on some specific characteristics of the UHECR spectrum, fixes the energy shift factors $\lambda$ for the other experiments [6, 38]. Positions, maximum zenith angles $\theta_{\max}$, exposures and energy shift factors are reported in Tab. I, for each experiment, as well as the number of detected events with rescaled energy $E' \geq 4.0 \times 10^{19}$ EeV ($E' = \lambda E$). In Fig. 3 is shown the relative geometrical exposure of each single detector listed in Tab. I, as well as the joint exposure of all experiments. For reference, in Fig. 4a is shown the all-sky data set of 102 detected events with rescaled energy $E'$, superimposed on the distribution of AGN within 200 Mpc from the reference catalog, whereas in Fig. 4b is shown the mock map of simulated events according to phys-
ical constraints previously described.

| Experiment     | φ₀       | θ_max | λ      | #Ev. |
|----------------|----------|-------|--------|------|
| Volcano R.     | 35.15°N  | 70°   | 1.000  | 6    |
| Yakutsk        | 61.60°N  | 60°   | 0.625  | 20   |
| H. Park        | 53.97°N  | 74°   | 1.000  | 7    |
| AGASA          | 35.78°N  | 45°   | 0.750  | 29   |
| SUGAR          | 30.43°S  | 70°   | 0.500  | 13   |
| P. Auger       | 35.20°S  | 60°   | 1.200  | 27   |

Table 1

Surface detectors: positions, maximum zenith angles θ_max, exposures, energy shift factors and number of detected events with rescaled energy \( E' \geq 4.0 \times 10^{19} \text{ EeV} \).

Figure 3. Relative geometrical exposure of each single detector listed in Tab. I (lines and points), and the joint exposure of all experiments (solid line).

Figure 5. MAF power vs. the number of events sampled from anisotropic mock maps generated as described in the text, for values of the significance corresponding to \( \alpha = 0.1\% \) and \( \alpha = 1\% \).

In Fig. we show the power \( 1 - \beta \) vs. the number of anisotropically distributed events, generated as described above. A sky is labelled as anisotropic if, for a fixed value of the significance \( \alpha \), the penalized chance probability as defined in Eq. 3 is lesser or equal than \( \alpha \), i.e. if the condition \( \hat{p}(\Theta^*) = \arg \min_\Theta \hat{p}(\Theta) \leq \alpha \) holds for some angular scale \( \Theta^* \). In Fig. 4 is shown the power for two values of the significance threshold, namely \( \alpha = 0.1\% \) and \( \alpha = 1\% \), estimated through the analytical approach. For applications, a power of 90\% is generally required: under this threshold the method could miss to detect an existing anisotropy signal. In the case of the MAF, and for the considered anisotropic mock map, the power increases with the number of events \( n \) and it is able to detect the anisotropic signal for \( n \geq 60 \), with significance \( \alpha = 1\% \). However, by decreasing the significance for the statistical test, the power requires a greater number of events to reach the 90\% threshold, as expected: our test clearly shows that the MAF provides an excellent discrimination power for \( n \geq 80 \). Indeed, we verified the agreement between the analytical and the Montecarlo estimations of the discrimination power.

6. Discussion and conclusion

We introduced a new statistical test, based on a multiscale approach, for detecting an anisotropy signal in the arrival direction distribution of UHECR, that makes use of an information theoretical measure of similarity, namely the Kullback-Leibler divergence, and of the extreme value theory. Within the present work we showed that our procedure is suitable for the analysis of both small and large data sets of events, by applying it on several Montecarlo realizations of isotropic and anisotropic synthetic data sets, cor-
responding to plausible scenarios in the physics of highest energy cosmic rays. In fact, for small data sets as well as for larger ones, the method is able to recover the information about the most significant angular scale of clustering in the data, even in presence of strong isotropic contamination.

The advantages of our approach over other methods are multiples. First, the method allows an analytical description of quantities involved in the estimation of the amount of anisotropy signal in the data, avoiding thousands of Montecarlo realizations needed for the penalizing procedure of results and drastically reducing the computation time. Second, the method allows the detection of a physical observable, namely the clustering scale, in the case of a point source. In the case of multiple sources, the information is about the most significant clustering scale(s), according to source distribution. Third, the method is unbiased against the null hypothesis and it provides an high discrimination power even in presence of strong contaminating isotropic background, for both small and large data sets. Although in this work we referred to UHECR physics for our applications, it is worth remarking that the method is suitable for the detection of the anisotropy signal in each data set involving a distribution of angular coordinates on the sphere, and it can be adapted to non-spherical spaces by properly re-defining the dynamical boxing algorithm.

Acknowledgments. Authors thank the P. Auger Collaboration for generous comments and fruitful discussions and the “Fondo per il potenziamento per la ricerca in informatica”, (Dipartimento di Matematica e Informatica, Università degli Studi di Palermo) for having kindly made available the computing resources required for the analyses presented within the present work.

References

[1] M. Davis and P.J.E. Peebles, Ap. J. 267, 465 (1983).

[2] S.D. Landy and A.S. Szalay, Ap. J. 412, 64 (1993).

[3] A.J.S. Hamilton, Ap. J. 417, 19 (1993).

[4] P.J.E. Peebles, *The Large Scale Structure of the Universe* (Princeton Univ. Press, 1980).
[5] M. Kachelrieß and D. Semikoz, Astrop. Phys. 23, 486 (2005).
[6] M. Kachelrieß and D.V. Semikoz, Astrop. Phys. 26, 10 (2006).
[7] A. Cuoco, G. Miele, and PD Serpico, Physical Review D 74, 123008 (2006).
[8] A. Cuoco, S. Hannestad, T. Haugbølle, M. Kachelrieß, and PD Serpico, The Astrophysical Journal 676, 807 (2008).
[9] A. Cuoco, S. Hannestad, T. Haugbølle, M. Kachelrieß, and P.D. Serpico, Astrop. J. 702, 825 (2009).
[10] M. Ave, L. Cazon, J. Cronin, J.R.T. de Mello Neto, A.V. Olinto, V. Pavlidou, P. Privitera, B.B. Siffert, F. Schmidt, and T. Venters, J. Cosm. Astrop. Phys. 2009, 023 (2009).
[11] J.D. Hague, B.R. Becker, M.S. Gold and J.A.J. Matthews, J. Phys. G: Nucl. Part. Phys. 36, 115203 (2009).
[12] Veron-Cetty M.P., Veron P., Astron. Astroph. 455, 773 (2006).
[13] de Mello Neto, J.R.T. et al, Proc. 31st ICRC, Lodz (2009).
[14] BT Stokes, CCH Jui, and JN Matthews, Astroparticle Physics 21, 95 (2004).
[15] S. Kullback and RA Leibler, The Annals of Mathematical Statistics 22, 79 (1951).
[16] S. Kullback, The American Statistician 41, 340 (1987).
[17] M. De Domenico, H. Lyberis, A. Insolia, and M. Scuderi, Arxiv preprint arXiv:1001.1666 (2010).
[18] P. Sommers, Astrop. Phys. 14, 271 (2001).
[19] E.J. Gumbel, Statistical theory of extreme values and some practical applications: A series of lectures (National Bureau of Standards Washington DC, 1954).
[20] E.J. Gumbel, Statistics of extremes (Dover Pub., 2004).
[21] M. Nagano and A.A. Watson, Rev. Mod. Phys. 72, 689 (2000).
[22] P. Bhattacharjee and G. Sigl, Phys. Rep. 327, 109 (2000).
[23] A.M. Hillas, Ann. Rev. Astr. Astrop. 22, 425 (1984).
[24] C.T. Hill, D.N. Schramm, and T.P. Walker, Phys. Rev. D 36, 1007 (1987).
[25] V. Berezinsky and A. Vilenkin, Phys. Rev. Lett. 79, 5202 (1997).
[26] V. Berezinsky, M. Kachelrieß, and A. Vilenkin, Phys. Rev. Lett. 79, 4302 (1997).
[27] A. Venkatesan, M.C. Miller, and A.V. Olinto, Astrop. J. 484, 323 (1997).
[28] B.T. Gare and P.L. Biermann, Phys. Rev. Lett. 81, 3579 (1998).
[29] D. Fargion, B. Mele, and A. Salis, Astrop. J. 517, 725 (1999).
[30] J. Arons, Astrop. J. 589, 871 (2003).
[31] E. Waxman, K.B. Fisher, and T. Piran, Astrop. J. 483, 1 (1997).
[32] G. Sigl, D.F. Torres, L.A. Anchordoqui, and G.E. Romero, Phys. Rev. D 63, 81302 (2001).
[33] Z. Fodor and S.D. Katz, Phys. Rev. D 63, 23002 (2000).
[34] P.G. Tinyakov and I.I. Tkachev, JETP Lett. 74, 1 (2001).
[35] Abraham, J. et al, Science 318, 938 (2007).
[36] Abraham, J. et al, Astrop. Phys. 29, 188 (2008).
[37] Cusumano, G. et al, Astron. Astroph. 510, (2010).
[38] V. Berezinsky, Nucl. Phys. B - Proc. Supp. 188, 227 (2009).
[39] Abbasi, R.U. et al, Phys. Rev. Lett. 100, 101101 (2008).