\textbf{\textit{\gamma-ray DBSCAN}: a clustering algorithm applied to \textit{Fermi-LAT} $\gamma$-ray data.}

I: Detection performances with real and simulated data

A. Tramacere\textsuperscript{1} and C. Vecchio\textsuperscript{2}

1 ISDC, University of Geneva, Chemin d’Ecogia 16, Versoix, CH-1290, Switzerland e-mail: andrea.tramacere@unige.ch
2 Politecnico di Milano - Piazza L. da Vinci, 32 - 20133 Milano, Italy

Received; accepted

\textbf{ABSTRACT}

\textbf{Context.} The Density Based Spatial Clustering of Applications with Noise (DBSCAN) is a topometric algorithm used to cluster spatial data that are affected by background noise. For the first time, we propose the use of this method for the detection of sources in $\gamma$-ray astrophysical images obtained from the \textit{Fermi-LAT} data, where each point corresponds to the arrival direction of a photon.

\textbf{Aims.} We investigate the detection performance of the $\gamma$-ray DBSCAN in terms of detection efficiency and rejection of spurious clusters.

\textbf{Methods.} We use a parametric approach, exploring a large volume of the $\gamma$-ray DBSCAN parameter space. By means of simulated data we statistically characterize the $\gamma$-ray DBSCAN, finding signatures that differentiate purely random fields, from fields with sources. We define a significance level for the detected clusters, and we successfully test this significance with our simulated data. We apply the method to real data, and we find an excellent agreement with the results obtained with simulated data.

\textbf{Results.} We find that the $\gamma$-ray DBSCAN can be successfully used in the detection of clusters in $\gamma$-ray data. The significance returned by our algorithm is strongly correlated with that provided by the Maximum Likelihood analysis with standard \textit{Fermi-LAT} software, and can be used to safely remove spurious clusters. The positional accuracy of the reconstructed cluster centroid compares to that returned by standard Maximum Likelihood analysis, allowing to look for astrophysical counterparts in narrow regions, minimizing the chance probability in the counterpart association.

\textbf{Conclusions.} We find that $\gamma$-ray DBSCAN is a powerful tool in the detection of clusters in $\gamma$-ray data, this method can be used both to look for point-like sources, and extended sources, and can be potentially applied to any astrophysical field related with detection of clusters in data. In a companion paper we will present the application of the $\gamma$-ray DBSCAN to the full \textit{Fermi-LAT} sky, discussing the potentiality in the discovery of new sources.

\textbf{Key words.} Gamma rays: general – Methods: statistical – Methods: data analysis

1. Introduction

Modern $\gamma$-ray telescopes operating at energies above the MeV window, provide event-resolved observational data. Each event (after the reconstruction process) is typically described by a tuple (i.e. an ordered list of elements) storing sky coordinates, arrival time, and energy. Detection of discrete sources (either point-like or extended) is performed using various methods. Given the discrete topological nature of $\gamma$-ray images, methods based on cluster search, like the \textit{Minimum Spanning Tree} (MST) \cite{Campana:2007, Campana:2012} have successfully been used.

One of the main advantages of topometric methods, compared to methods using the spatial binning, is to minimize the impact of the poor energy-dependent Point Spread Function (PSF), typical of $\gamma$-ray telescopes, preserving the spatial information of each event. Moreover, these methods are able to detect sources compounded by a small amount of events, but they need to be fine tuned to take into account properly the background. The problem of background rejection is the most penalizing feature of topometric methods, for this reason in this paper, for the first time, we present a method based on the DBSCAN algorithm \cite{Ester:1996}. The DBSCAN is a topometric algorithm used to cluster spatial data that are affected by background noise. Compared to other topometric methods, it has the advantage to embed inside the algorithm itself the discrimination between signal (cluster) and background (noise), according to the local density of events within a typical scanning brush i.e. within a given scanning area.

The aim of the present paper is to show the potentialities of the method, and its statistical characterization when applied to astrophysical $\gamma$-ray data. We apply this method to the detection of point-like sources in the \textit{Fermi-LAT} data. We explore a large volume of the $\gamma$-ray DBSCAN parameter space, by means of simulated data, and we provide a statistical characterization the $\gamma$-ray DBSCAN, finding signatures that differentiate purely random fields, from fields with sources. We define a significance level for the detected clusters, and we successfully test this significance with our simulated data. We apply the method to real \textit{Fermi-LAT} $\gamma$-ray data, and we find an excellent agreement with the results obtained with simulated data.

In a companion paper \cite{Tramacere:2012}, we will apply the method to the \textit{Fermi-LAT} sky, investigating specific issues related to the \textit{Fermi-LAT} response functions, showing the potentiality for the discovery of new sources, in particular of small clusters located at high galactic latitude, or clusters on the galactic plane, affected by a strong background.

The paper is organized as follows. In Sec. 2 we describe the logic of the DBSCAN method, and we present the algorithm...
implemented to analyse γ-ray data, the γ-ray DBSCAN. In Sec. 3 we discuss some caveats regarding the application of the γ-ray DBSCAN algorithm to γ-ray data. In Sec. 4 we study the statistical properties of the γ-ray DBSCAN detection, using a simulated test field with only noise, and five simulated test fields with noise plus point-like sources. In Sec. 5 we evaluate the detection performance of the method in terms of positional accuracy, cluster reconstruction, and rejection of spurious clusters. In Sec. 6 we investigate the significance of the clusters, and describe our algorithmic implementation. In Section 7 we finally use our method with real Fermi-LAT data, investigating the detection performance, and comparing the γ-ray DBSCAN clusters significance, to that returned by the Maximum Likelihood method (sources), conversely low densely populated areas are classified as noise or clusters, and the ID of the cluster the photon belongs to. Each cluster $C_m$, will be described by a tuple storing the position of the centroid with its positional error, the ellipse of the cluster containment, the cluster effective radius ($r_{eff}$), and number of photons in the cluster ($N_p$). The ellipse of the cluster containment, is defined by major and minor semi-axis ($\sigma_x$ and $\sigma_y$, respectively), and the inclination angle ($\alpha_{gal}$) of the major semi-axis w.r.t. the latitude coordinate. (b or DEC). To evaluate the ellipse axis we use the Principal Component Analysis method (PCA) (Jolliffe1986). This method uses the eigenvalue decomposition of the covariance matrix of the two position arrays $x$, and $y$. By definition, the square root of the first eigenvalue will correspond to $\sigma_y$, and the second to $\sigma_x$. The axes represent the two orthogonal directions of maximum variance of the cluster. The effective radius is defined as $r_{eff} = \sqrt{\sigma_x^2 + \sigma_y^2}$. To find the centroid of the cluster and its uncertainty, we use a weighted average of the position of each photon in $C_m$, as follows:

- we define the first order centroid ($C_{ave}$) as the average of the position of each photon: $C_{ave} = (x, y)$. We define the weight array, according to the distance between $p_x \in C_m$ and $p_{ave}$: $w_{ij} = 1/|p_{i,j} - C_{ave}|$. The cluster centroid $C_{ctr}$ will result from average of the position of each cluster point weighted by $w_{ij}$. We have numerically verified that $pos_{err}$ corresponds to a $\approx 95\%$ positional uncertainty.

3. Caveat on the application to γ-ray data

The application of clustering methods, such as the γ-ray DBSCAN, leads to deal with practical difficulties, related mainly to the instrument PSF, and to gradient and/or structures in the background. In order to deal with these issues, without biasing the detection results, it’s recommended to apply some criteria that we discuss in the following. As first, we comment on the PSF impact. The PSF imposes a limit on the capability of an instrument to resolve sources separated by a distance smaller then the PSF size. Sources with sizes below the PSF classified as point-like, otherwise are classified as extended. A further complication is that the PSF often depends on the energy; in the case of Fermi-LAT, the 68% containment angle of the reconstructed incoming photon direction, for normal incidence photons, has a typical size of a couple of degrees at 100 MeV (Fermi-LAT Collaboration 2012), and scales down to few tens of degree above the GeV energies. The size of the PSF is strongly connected to the size $e$ of the γ-ray DBSCAN scanning brush. Indeed, if $e$ is much smaller than the PSF size, it might occur the risk to loose clusters characterized by small $N_p$, or to fragment a cluster with large $N_p$ in smaller fake satellite clusters. We stress that the formation of satellite clusters is a very rare event, thanks to our recursive DBSCAN implementation, that is explained in Sec. 6. On the contrary, if

\footnote{http://www.slac.stanford.edu/exp/glast/groups/canda/lat_performance.html}
A. Tramacere and C. Vecchio: γ-ray DBSCAN application to Fermi-LAT γ-ray data.

Fig. 1. Photon map for the sky test field 1, with the result of the γ-ray DBSCAN detection for $K = 5$ and $\varepsilon = 0.17$ deg. The blue crosses refer to the simulated sources, the green boxes to 51 detected true clusters, and the red boxes to the 2 fake ones. The black dots represent the background events, the remaining colors indicate cluster events.

A careful and self-consistent analysis of the effects of the energy dependence of the PSF, and in general of issues related to the Fermi-LAT response function, is beyond the scope of this paper, where we focus mostly on a statistical characterization of the method. These subjects will be investigated in the companion paper (Tramacere 2012).

A second relevant issue, is the inhomogeneity of the background, that affects both the choice of $\varepsilon$ and $K$. If the background is homogeneous over the entire field, the optimal choice of a single pair of values of $\varepsilon$ and $K$, guarantees a safe rejection of the background. Indeed, values of $\varepsilon$ and $K$, such that the average density of photons within $\varepsilon$ is significantly larger the the average density of the background photons, make rare the chance to grow a cluster from a background fluctuation. Unfortunately, the γ-ray sky shows strong gradients of background, in particular at low galactic latitudes. To solve this issue, one could think to adapt the value of $\varepsilon$ and $K$ according to a local value of the background photon density. Since $\varepsilon$ has a strong constraint imposed by the PSF, one should tune mostly the value of $K$. The drawback is that as we increase the value of $K$ to compensate for the background, we decrease the capability to detect cluster with small $N_p$. To overcome this difficulty, we adopt an alternative solution. We use a unique pair of values of $\varepsilon$ and $K$, for each field, where $\varepsilon$ is mostly constrained by the PSF, and $K$, by the field average background, ad we take into account the background inhomogeneities by defining a significance level of the cluster, according to the signal to noise ratio (Li & Ma 1983), evaluated from the local background. This is explained in detail in Section 6. The capability to reject clusters according to a low significance level, allows to relax the constrain on $\varepsilon$ and $K$, increasing the number of clusters detected, hence increasing the detection ratio, and at the same time allows to reject spurious sources, due to the significance threshold. Anyhow, to

$\varepsilon$ is much larger w.r.t the PSF, it is likely to build extended clusters contaminated by the background, or by close sources.
avoid that the background is so high, that the fluctuations in the background events, can lead to densities comparable to those of weak sources, it’s recommended to apply a cut in energy, to make this possibility rare. In order to optimize the ratio between background and clusters events, in the following we use a threshold of 3 GeV, that mitigates the possible bias due to the background fluctuations.

4. Statistical properties of the $\gamma$-ray DBSCAN clusters

4.1. The test fields

In this section we study the statistical properties of the clusters, looking for signatures that characterize random Poissonian fields, and fields with point-like sources. To accomplish this task we compare results obtained for a test field with only noise (random test field), and the five test fields with noise plus point-like sources (sky test fields 1-5).

As sky test fields we use the same fields used in the Campana et al. (2012). Each of these five sky fields covers a broad sky region, with a galactic longitude extension of $80^\circ < l < 170^\circ$ and a galactic latitude extension of $40^\circ < b < 65^\circ$. The $\gamma$-ray background has been simulated using the standard gtobssim tool, developed by the Fermi-LAT collaboration, simulating both the Galactic and isotropic components for a 2-year long period, using a threshold energy of 3 GeV, for a total amount of 9322 photons. To this photon list we added 70 simulated sources: for each source, the number of photons was chosen from a probability distribution given by a power-law, with exponent 2, centered at the source location. Five simulated test fields have been generated, adding the simulated sources to the diffuse background. The only difference in the five realizations is the source location, randomly chosen to have different brightness contrast between sources and the background. The random test field covers the same area of the sky test fields, and a number of events equal to the sky test field-1 (background and sources), for a total amount of 11044 events.

In Fig. 1 we show the photon map for the sky test field 1, and the result of the $\gamma$-ray DBSCAN detection for $K = 5$ and $\varepsilon = 0.17$ deg. We detect 51 true clusters, and only 2 fake ones. A cluster is defined as a circle centered on the centroid cluster, with a radius equal to 2 $\sigma_{\text{psf}}$. We call fake, the remaining clusters. In Fig. 2 we show a close-up of two true clusters. The black ellipses correspond to the ellipses of the positional error, and the purple and orange thick points represent the cluster points, while the black thick dots represent the background.

4.2. Test strategy

We now investigate the statistical properties of the $\gamma$-ray DBSCAN clusters, in particular signatures that distinguish purely random fields from fields with point-like sources, and their dependence on $K$ and $\varepsilon$. To investigate systematically a broad volume of the parameter space, we use a parametric approach. We set the range of $\varepsilon$ in $[0.1, 0.50]$ deg, with a step of 0.01 deg., and the range of $K$ in $[2, 15]$, with a step of 1. The total amount of detection trials for each test field is 574. We collect the statistics of the trials, and we investigate the distribution of $r_{\text{eff}}$ and $N_p$, and their connection with $\varepsilon$ and $K$, respectively.

4.3. Statistics of $r_{\text{eff}}$ and connection with $\varepsilon$

We start by investigating the distribution of the log($r_{\text{eff}}$) values, in the case of the random, and of the sky test field-1. The distribution for the detections collected over the full $K-\varepsilon$ parameter space (top left panel of Fig. 4), shows a symmetric shape well fitted by a Gaussian distribution (log-normal w.r.t. $r_{\text{eff}}$), with the mean value of $\log_{10}(r_{\text{eff}}) \simeq -0.45$ (corresponding to $r_{\text{eff}} \approx 0.3$ deg) and a dispersion of $\sigma_{\log_{10}(r_{\text{eff}})} \approx 0.23$. The log-normal distribution provides a reasonable description of the empirical distributions also for individual pairs of $(K, \varepsilon)$ values. An example is given in panel c of Fig. 3 for the case $K = 3$, $\varepsilon = 0.3$ deg., where the best fit values are $\log_{10}(r_{\text{eff}}) \approx -0.51$, and $\sigma_{\log_{10}(r_{\text{eff}})} \approx 0.16$. We now investigate the empirical distribution of log($r_{\text{eff}}$) for fields with point-like sources. In the right panel of Fig. 3, we show the case of the sky test field-1. The distributions of log($r_{\text{eff}}$) are still described by a normal. In the case of fake clusters (red dashed line), the best fit values of the mean ($\log_{10}(r_{\text{eff}}) \approx -0.46$) and of the dispersion ($\sigma_{\log_{10}(r_{\text{eff}})} \approx 0.24$), are very similar to those found in the case of the random test field. On the contrary, the true cluster distribution (blue hatched histogram) is peaking around the value of log($r_{\text{eff}}$) $\approx -0.67$ deg, corresponding to $r_{\text{eff}} \approx 0.21$ deg., very close to the value of the dispersion $\sigma_{\log_{10}} = 0.20$ deg., used to simulate the sources. Since the simulation parameter $\sigma_{\log_{10}}$ reproduces the effect of the instrumental PSF, we observe that for non-random fields, the typical size of the reconstructed clusters is constrained by the PSF, suggesting the empirical rule to set the value of $\varepsilon$ of the order of the PSF size.

To investigate more accurately the connection between $\varepsilon$ and the PSF, we analyse the statistical properties of the quantity $r_{\text{eff}}/\varepsilon$ as a function of $\varepsilon$. For each value of $\varepsilon$, we determine the median, and the two-sided 1-$\sigma$ confidence level (CL) interval around the median, of the $r_{\text{eff}}/\varepsilon$ distributions. In the left panel of Fig. 4 we plot the $r_{\text{eff}}/\varepsilon$ median (blue solid circles) and 1-$\sigma$ CL region, as a function of $\varepsilon$, for the random field. We note that the $r_{\text{eff}}/\varepsilon$ trend is slightly increasing with $\varepsilon$, and that the 1-$\sigma$ CL region is consistent with the case $r_{\text{eff}}/\varepsilon = 1$, but the upper boundary shows a systematic increase, compared to the lower boundary, for $\varepsilon < 0.30$ deg. The trend for the case of the true clusters in sky test field 1 (right panel Fig. 4), shows a different behaviour. The median of $r_{\text{eff}}/\varepsilon$ (red solid circles) is slightly decreasing with $\varepsilon$, showing that, for true clusters, $r_{\text{eff}}$ is not sensitive to the size of $\varepsilon$, being mostly constrained by the simulated PSF size. As expected, for the case of fake clusters (blue dashed line), the trend is almost identical to that of the clusters in the random field.

4.4. Statistics of $N_p$ and connection with $K$

We now investigate the statistics of the distribution of the number of photons per cluster. In the case of random fields, we expect that the number of photons in a cluster attends a Poisson distribution. Indeed, for a generic two-dimensional Poisson pro-
A. Tramacere and C. Vecchio: γ-ray DBSCAN application to Fermi-LAT γ-ray data.

Fig. 3. Panel a: distribution of the values of $\log_{10} r_{\text{eff}}$ for the random field case, for the full parameter space (black line) and fit by means of Gaussian distribution (blue line). Panel b: the same as in the top panel, for the case of $K = 3$ and $\varepsilon = 0.3$ deg. Panel c: distribution of $\log_{10} r_{\text{eff}}$ for the case of the sky test field 1, for fake clusters (red solid line), and true clusters (blue solid line, hatched histogram). The dashed lines represent a Gaussian best fit.

Fig. 4: Left panel: the $r_{\text{eff}}/\varepsilon$ statistical distribution as a function of $\varepsilon$, for the random field case. The blue solid circles represent the median, and the grey shaded area represents the 1-σ confidence level region, for each value of $\varepsilon$. Right panel: the same as in the bottom left panel, for the case of the sky test field 1. The red solid circles represent the median of the sky field 1, true clusters (red solid line), and true clusters (blue solid line, hatched histogram) the dashed line shows the 1-σ confidence level region, for the case of the fake clusters.

cess, the probability to observe a number of events ($N(S) = j$) enclosed by a surface $S$ is given by:

$$P(N(S) = j) = \frac{(\lambda |S|)^j \exp(-\lambda |S|)}{j!},$$  

(2)

where $\lambda$ is the average spatial density. Translating $S$ in terms of $\varepsilon^2$, we can rewrite:

$$P(N(\varepsilon^2) = j) = \frac{(|\varepsilon^2|)^j \exp(-|\varepsilon|^2)}{j!},$$  

(3)

from which follows that, given the value of $K$ and $\varepsilon$, the probability to find a cluster as function of $K$ and $\varepsilon$ will be given by

$$P_{\text{clus}}(\varepsilon, K) = P(N(\varepsilon^2) > K) = 1 - \sum_{j=0}^{K} \frac{(|\varepsilon|^2)^j \exp(-|\varepsilon|^2)}{j!},$$  

(4)

namely the Poissonian survival function. Anyhow, due to the logic of the DBSCAN clustering process, the Poisson statistics can’t be extended from $\varepsilon$ to $r_{\text{eff}}$, for any value of $\varepsilon$. Indeed, a cluster is not a simple collection of points enclosed within a surface $S$, this holds only within the $\varepsilon$-sized circle, namely the seed of the cluster ($C'$). If we consider the annulus defined between $\varepsilon$ and the cluster radius $r_{\text{clus}}$, not all the points in the annulus will be cluster member, but only those that are at least density reachable. This implies that we expect a deviation from the Poissonian statistics, when $r_{\text{eff}}$ is significantly larger than $\varepsilon$, i.e. $\varepsilon \geq 0.3$ deg (according to the analysis presented in the previous section). This expected deviation from the Poissonian statistics, is confirmed by the plots in the left panels of Fig 5. In panel a we show the distribution of $N_p$ for the case $K = 2$ and $\varepsilon = 0.20$ deg. We note that the Poissonian distribution (Eq 3) gives a reasonable description of the empirical distribution. On the contrary, for the case of $\varepsilon = 0.30$ deg. (panel b), we observe that the Poisson distribution shows larger deviations, in particular for $K > 6$. When we take into account the $N_p$ distribution for the full parameter space (panel c), we note the Poissonian distribution is failing in providing a reasonable description of the empirical distribution, whilst a log-normal one gives a good fit.

The log-normal trend of $N_p$ is consistent with the log-normal trend of the distribution of $r_{\text{eff}}$. Since the number of photons in
a cluster will be approximatively \( N_p \propto \lambda r^2_{eff} \), we can write the PDF of \( N_p \):

\[
f(N_p) \propto f(r_{eff}^2)\lambda.
\]

(5)

To evaluate the distribution of \( r_{eff}^2 \) we can use the standard theory of the transformation of Random Variables (RV) (Papoulis1965). It can be easily proved that, given a RV \( X \) having a log-normal distribution,

\[
f(X) = \frac{1}{x} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln(x) - \mu)^2}{2\sigma^2}\right).
\]

(6)

the RV \( Y = X^2 \), will follow a log-normal distribution given by:

\[
f(Y) = \frac{1}{2Y} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln(\sqrt{Y}) - 2\mu)^2}{4\sigma^2}\right).
\]

(7)

Indeed, our \( r_{eff}^2 \) distribution, for the random field (panel d, Fig. 5), is fitted by a log-normal distribution peaking at \( \approx 0.03 \) deg\(^2\). Hence, according to Eq. 6 we expect that also \( f(N_p) \) will follow a log-normal distribution, when \( N_p \) is not ruled by a Poissonian statistics.

We verify, that the same statistical trends, describe the real sky fields. The panels e and f in Fig. 5 show the statistical distribution of \( N_p \) for the sky test field 1 case. In agreement with the analysis concerning the random test field, we see that the fake clusters (\( \varepsilon = 0.30 \) deg., panel e in Fig. 5), are described by a Poissonian statistic, whilst, the true clusters (panel f in Fig. 5), are better described by a log-normal distribution (red dashed line), compared to a Poissonian one (solid black line). We also observe that the log-normal law, describes reasonably the empirical distribution, only for values of \( N_p \leq 50 \), whilst shows significant deviation in the tail, consistent with the statistics of our simulated sources population.

To complete this statistical characterization, we investigate the distribution of the number of detected clusters, as a function of the threshold \( K \). According to Eq. 4 we expect that the number of detected cluster, for a random field, follows a Poisson survival distribution. The plot a of Fig. 6 confirms our hypothesis, indeed the Poisson survival function provides a reasonable description of the empirical distribution. The same holds for the case of fake clusters of the sky test field 1 (plot c Fig. 5). On the contrary, in the case of true clusters (panel d Fig. 6), the Poisson survival distribution is not able to reproduce the observed trend, consistently with the non-poissonian statistic of the simulated clusters. The panels b and e of Fig. 6 show the \( 1 - \sigma \) CL region...
Fig. 6. Panel b: the \( N_p \) statistical distribution as a function of \( K \), for the random field case. The blue solid circles represent the median, and the grey shaded area represents the 1-\( \sigma \) confidence level region around the median, for each value of \( K \). The dashed black line represents the \( N_p = K + 1 \) law. Panel e: same as in panel b, for the sky test field 1 case. Panel a: number of detected clusters for the random test field case (blue solid points), as a function of \( K \), and best fit by means of a Poissonian survival function (red empty bars). Panel c: number of detected cluster for the sky test field 1 case (black solid points), for the case of fake clusters, as a function of \( K \), and best fit by means of a Poissonian survival function (red empty bars). Panel d: number of detected cluster for the sky test field 1 case (red solid boxes), for the case of true clusters, as a function of \( K \), and best fit by means of a Poissonian survival function (black empty boxes).

5. Testing the detection performance with simulated gamma-ray data

In this section we investigate the detection performance of the gamma-ray DBSCAN. As first point, we study the dependency of the detection efficiency on \( K \) and \( \varepsilon \), and their impact on the spurious ratio, and on the detection efficiency. Then, we investigate the capability of the algorithm to reconstruct the simulated clusters, and the positional accuracy of the reconstructed centroids. We test the detection performance of the gamma-ray DBSCAN using as benchmark the five sky test fields used in the previous section, and exploring the same parameter space.

5.1. Detection efficiency and spurious ratio as a function of \( K \) and \( \varepsilon \)

To investigate the detection performance of the gamma-ray DBSCAN, we run, for each of the five sky test fields, and for each pair of values \( K, \varepsilon \), a gamma-ray DBSCAN detection. For each detection run, we build a cluster catalog. Starting from the cluster catalog, we build the corresponding candidate catalog. The candidate catalog is a list of sources built by taking into account two possible biases, the confusion, and the multiple association, in detail:

- a cluster is defined true, i.e. with a possible counterpart, if the position of the simulated source falls within a circle centered on the cluster centroid, with a radius equal to 2 \( \sigma \text{pos}_{src} \).
- Two or more true clusters are defined confused, if they have the same counterpart.
- A true cluster has a multiple association, if has more than one counterpart.

We stress that, the number of confused clusters is negligible, indeed the average number of confused clusters per run is about 0.08, and no confused clusters are found for \( K > 4 \), and that the average number of multiple associations per run is about 0.2.

The final candidate catalog will count a number of candidate sources \( N_{src} \), each identified by a unique \( SRCID \). The number of spurious sources will be \( N_{fake} = N_{src} - N_{true} \). In order to characterize the performance, we define the following parameters:

- the detection efficiency:

\[
D_{eff} = \begin{cases} 
\frac{N_{true} - N_{fake}}{N_{sim}(N_{true}, > K)}, & \text{if } (N_{true} - N_{fake}) \leq N_{sim}(N_{p, sim}, > K) \\
1.0, & \text{if } (N_{true} - N_{fake}) > N_{sim}(N_{p, sim}, > K)
\end{cases}
\]  

(8)

where \( N_{sim}(N_{p, sim}, > K) \) is the number of simulated sources with a number of simulated events larger than \( K \).

- the true detection ratio \( D_{true} = N_{true}/N_{src} \).
The $D_{\text{fake}}$ parameter shows the fraction of simulated clusters, above the threshold $N_{p,\text{sim}} = K$, detected by the method, net of the fake ones. Hence, does not provide an indication of the spurious contamination. For this reason we have introduced the $Q$ parameter, which rescale the $D_{\text{eff}}$ according to the ratio between fake clusters, and found clusters $N_{\text{src}}$. We remind that, according to the $D_{\text{eff}}$ definition in Eq. (8), it’s possible to obtain values of $D_{\text{eff}} > 1.0$. Assume to have a simulated cluster such that, for a given $K$ and $\varepsilon$, the corresponding seed cluster has a size $N^* = N_{p,\text{sim}} = K$. In the case of no background events within the circle of radius $\varepsilon$, this cluster will be rejected. If we have one or more background events contained within the circle of radius $\varepsilon$, i.e. $N^* > K$, the cluster will be detected. For this reason, in such a case, we report a value of $D_{\text{eff}} = 1.0$. The same applies to $Q$.

In Fig. 7 we summarize the detection runs for the case of sky test field 1, for the full parameters space with $K > 2$. The panel a shows the isolevel map of the fake clusters detection ratio. The gradient in the isolevel map is quite sharp, and roughly half of the parameter space shows no fake clusters (white isolevel line).

The $D_{\text{false}}$ parameter shows the fraction of simulated clusters, above the threshold $N_{p,\text{sim}} = K$, detected by the method, net of the fake ones. Hence, does not provide an indication of the spurious contamination. For this reason we have introduced the $Q$ parameter, which rescale the $D_{\text{eff}}$ according to the ratio between fake clusters, and found clusters $N_{\text{src}}$. We remind that, according to the $D_{\text{eff}}$ definition in Eq. (8), it’s possible to obtain values of $D_{\text{eff}} > 1.0$. Assume to have a simulated cluster such that, for a given $K$ and $\varepsilon$, the corresponding seed cluster has a size $N^* = N_{p,\text{sim}} = K$. In the case of no background events within the circle of radius $\varepsilon$, this cluster will be rejected. If we have one or more background events contained within the circle of radius $\varepsilon$, i.e. $N^* > K$, the cluster will be detected. For this reason, in such a case, we report a value of $D_{\text{eff}} = 1.0$. The same applies to $Q$.

In Fig. 7 we summarize the detection runs for the case of sky test field 1, for the full parameters space with $K > 2$. The panel a shows the isolevel map of the fake clusters detection ratio. The gradient in the isolevel map is quite sharp, and roughly half of the parameter space shows no fake clusters (white isolevel line). To have a better understanding of the impact of fake clusters, it’s interesting to compare the $D_{\text{false}}$ isolevel map to the $D_{\text{true}}$ isolevel map (panel b Fig. 7). Also in this case the map shows a sharp gradient, and the region with $D_{\text{true}} > 0.95$ overlaps the $D_{\text{false}} = 0$ region. These two maps, clearly show the region of the parameter space where the algorithm has the best performance, but the $D_{\text{true}}$ and $D_{\text{false}}$ ratios do not provide information on the ratio between the number of true detected clusters and the number of simulated clusters. At this regard more information are provided by the $D_{\text{eff}}$ isolevel map (panel c, Fig. 7). To focus on the “effective” volume of the parameter space, we hide by a white area the region where $D_{\text{eff}} < 0$. We note that the isolevel lines $D_{\text{eff}} = 0$ and the isomap lines in the maximum gradient.
area show a positive correlation between $K$ and $\epsilon$, meaning that an increased value of $\epsilon$ requires an increased value of $K$, to have better background rejection. To evaluate better the trade-off between $D_{\text{true}}$ and $D_{\text{fake}}$, we plot in the panel d of Fig. 7 the isolevel map of $Q$. This plot shows that the area corresponding to $Q > 0.95$, is consistent with that found in the case of $D_{\text{eff}}$. In Tab. 1 we report the $D_{\text{eff}}$ values obtained for all the five sky fields, for detections with a number of fake sources $\leq 6$. We note that the average values of true clusters ranges between 44 and 51, with the fake ones ranging between 1 and 3, and an average $D_{\text{eff}}$ between 0.96 and 1.0. This is a very promising result.

5.2. Cluster reconstruction, and positional accuracy

The positional accuracy of the toptometric methods, is probably the most important feature of this class of algorithms. In Sec. [we have described our weighting method to reconstruct the centroid of the cluster.

The panel a of Fig. 8 shows by red solid boxes the mean positional error of the clusters centroid and the standard deviation (vertical error bar) vs. $N_p$, for the true clusters of the sky test field 1 with $\epsilon \leq 30$ deg. The clusters are binned in $N_p$, with the bin width indicated by the horizontal error bar. As expected, the uncertainty on the reconstructed cluster centroid is $\sigma_{\text{pos.err}} \approx \sigma_{\text{sim}}/\sqrt{N_p}$ (solid red line). The solid black circles represent the corresponding trend for the separation between the simulated cluster position and the reconstructed cluster centroid. For $N_p \geq 30$, the separation is below $2^\prime$. In the panel b of Fig. 8 we plot the histogram of the distribution of the angular separation between the position of the simulated source, and the position of the cluster centroid. For the three cases of $\epsilon = 0.10$ deg., $\epsilon = 0.15$ deg., and $\epsilon = 0.20$ deg., the positional error is below the $1.5^\prime$ for the 68% of the sample.

Besides positional accuracy, is also important to understand the capability of the $\gamma$-ray DBSCAN to reconstruct the simulated cluster in terms of number of photons. Indeed, this information gives an idea of the average number of background photons contaminating the reconstructed cluster. In the top left panel of Fig. 9, we show the scatter plot of $N_p$ vs. the number of simulated events ($N_p$ sim.). The solid points represent the average value of $N_p$, for a given value of $N_p$ sim., and the error bar, corresponds to the standard deviation. The solid green line represents the case $N_p = N_p$ sim., and the dashed upper and lower lines, represent $N_p = N_p$ sim. ± 10 sim., respectively. Both for the cases of $\epsilon = 0.15$ deg., and $\epsilon = 0.20$ deg., the scatter is bounded by the dashed lines, showing that the largest excess in the $N_p$
Table 1. Summary of the detections obtained for all the five sky fields, for detections with a number of fake sources \(\leq 6\). \(N_{\text{sim}} > K\), is the number of simulated sources with a number of simulated events larger than \(K\), the number separated by the / symbol, indicates the full number of simulated sources.

| Sky Field | \(N_{\text{in}}\) | True | Fake | \(K\) | \(\varepsilon\) | \(D_{\text{avr}}\) | \(Q\) |
|-----------|---------------------|------|------|------|---------|---------|------|
| sky field 1 | 44/70 | 47 | 1 | 6 | 0.19 | 1.00 | 1.00 |
| 38/70 | 48 | 2 | 8 | 0.27 | 1.00 | 1.00 |
| 53/70 | 50 | 3 | 5 | 0.17 | 0.89 | 0.84 |
| 62/70 | 51 | 5 | 4 | 0.14 | 0.74 | 0.68 |
| 53/70 | 53 | 6 | 5 | 0.18 | 0.89 | 0.80 |
| average | 50.00 | 49.80 | 3.40 | 5.60 | 0.19 | 0.90 | 0.86 |
| sky field 2 | 62/70 | 47 | 0 | 4 | 0.11 | 0.76 | 0.76 |
| 44/70 | 49 | 2 | 6 | 0.20 | 1.00 | 1.00 |
| 53/70 | 51 | 3 | 5 | 0.17 | 0.91 | 0.86 |
| 62/70 | 52 | 6 | 4 | 0.14 | 0.74 | 0.67 |
| average | 55.25 | 49.75 | 2.75 | 4.75 | 0.15 | 0.85 | 0.82 |
| sky field 3 | 41/70 | 42 | 0 | 7 | 0.22 | 1.00 | 1.00 |
| 44/70 | 45 | 2 | 6 | 0.20 | 0.98 | 0.94 |
| 53/70 | 46 | 3 | 5 | 0.16 | 0.81 | 0.76 |
| 62/70 | 48 | 5 | 4 | 0.14 | 0.69 | 0.63 |
| 62/70 | 50 | 6 | 4 | 0.15 | 0.71 | 0.63 |
| average | 52.40 | 46.20 | 3.20 | 5.20 | 0.17 | 0.84 | 0.79 |
| sky field 4 | 53/70 | 47 | 1 | 3 | 0.16 | 0.87 | 0.85 |
| 53/70 | 50 | 5 | 5 | 0.18 | 0.85 | 0.77 |
| 62/70 | 52 | 6 | 4 | 0.14 | 0.74 | 0.67 |
| average | 56.00 | 49.67 | 4.00 | 4.67 | 0.16 | 0.82 | 0.76 |
| sky field 5 | 44/70 | 47 | 1 | 6 | 0.19 | 1.00 | 1.00 |
| 44/70 | 50 | 2 | 6 | 0.20 | 1.00 | 1.00 |
| 44/70 | 53 | 3 | 6 | 0.21 | 1.00 | 1.00 |
| 44/70 | 55 | 5 | 6 | 0.22 | 1.00 | 1.00 |
| average | 44.00 | 51.25 | 2.75 | 6.00 | 0.20 | 1.00 | 1.00 |

6. Cluster significance, background inhomogeneities, and rejection of spurious clusters

Even though, we have identified the region of the \(K\)-\(\varepsilon\) parameter space, where the detection efficiency is larger, and the probability to detect fake cluster is lower, in the application to real data, it’s mandatory to provide a significance level, expressing the probability of a cluster being not originated in a background fluctuation. We propose a method derived from the \cite{LiMa1983} approach, based on the evaluation of the signal to noise (S/N) ratio. A significance method based on the S/N ratio fits well the the \(\gamma\)-ray DBSCAN implementation, because the algorithm directly provides a partition of the photon list in cluster and noise events. Hence, for each cluster we can evaluate easily the S/N ratio, knowing the exact nature of each event. The procedure to evaluate the significance is summarized by the following items:

1. for each cluster, we define an annular region, with an inner radius \(r_{in}\), and an external radius \(r_{out}\).
2. \(r_{in}\) is set to an initial value of \(r_{in} = 2r_{eff}\), and is adaptively increased with a step of \(r_{in}/10\), for a maximum of 10 trials, until at least the 95\% of the cluster events are enclosed within \(r_{in}\).
3. \(r_{out}\) is set to 3\(r_{in}\).
4. We count all the cluster events \(N_{\text{in}}\) and all the background events \(N_{\text{bkg}}\), enclosed within the circle with radius \(r_{in}\) and centered on the cluster centroid.
5. We determine the \(N_{\text{bkg}}\) background level, rescaling the number of background events in \(r_{in} < r < r_{out}\), to a circle with radius \(r_{in}\).
6. To evaluate possible gradients in the background, we select a region enough far from the cluster to sample properly the background level, and enough close to the cluster, to measure a local background level. At this regard, we define the radius \(r_{ave}^{\text{eff}} = (r_{out} + r_{in})/2\), and we evaluate the average background level \(N_{\text{bkg}}^{\text{eff}}\) in a circle of radius \(r_{ave}^{\text{eff}}\), centered on each point \(r_{ave}^{\text{eff}} < r < r_{out}\).
7. If no background points are found in \(r_{ave}^{\text{eff}} < r < r_{out}\), we set \(N_{\text{bkg}}^{\text{eff}} = N_{\text{in}}^{\text{eff}}\).
8. By comparing \(N_{\text{bkg}}^{\text{eff}}\) to \(N_{\text{in}}^{\text{eff}}\), we evaluate the fraction of noise already resolved by the \(\gamma\)-ray DBSCAN, and we evaluate the effective background level \(N_{\text{bkg}}^{\text{eff}}\), by correcting \(N_{\text{bkg}}^{\text{eff}}\) for \(N_{\text{in}}^{\text{eff}}\).
9. We evaluate the significance according to the Likelihood Ratio Test (LRT) method proposed by \cite{LiMa1983}:

\[
S_{\text{cls}} = \sqrt{2 \left( \frac{N_{\text{in}}^{\text{eff}} \ln \left( \frac{2N_{\text{in}}^{\text{eff}}}{N_{\text{in}}^{\text{eff}} + N_{\text{bkg}}^{\text{eff}}} \right) + N_{\text{bkg}}^{\text{eff}} \ln \left( \frac{2N_{\text{bkg}}^{\text{eff}}}{N_{\text{in}}^{\text{eff}} + N_{\text{bkg}}^{\text{eff}}} \right) \right)}
\]

Under the hypothesis that a cluster is due to a background fluctuation, the variable \(S_{\text{cls}}\) is expected to follow a chi square distribution, with one degree of freedom (\(\chi^2\)). In the left panel of Fig. 10, we plot the distribution of \(S_{\text{cls}}\) for the fake clusters in the sky test field 1 (blue histogram), compared to a \(\chi^2\) distribution. The empirical distribution, is well described by the expected \(\chi^2\) distribution, proving that the value of \(S_{\text{cls}}\), can be used as the "significance" of the detected cluster. A very illustrative example of the power of \(S_{\text{cls}}\) in rejecting fake clusters, is given by the plot in the right panel in Fig. 10, where we plot the \(D_{\text{fake}}\) ratio isolavel map, applying the selection \(S_{\text{cls}} > 4\). The fake ratio is 0 for the parameter space with \(\varepsilon < 0.25\) deg. For 0.25 deg. \(\leq \varepsilon < 0.35\) deg., there are fluctuations showing \(D_{\text{fake}} \leq 0.05\). Only for \(\varepsilon \geq 0.40\) deg. and \(K \leq 8\), the fake ratio shows a significant increase, but we stress that in this region of the parameter space, \(\varepsilon\) is more then double of the PSF size, hence this is a region of the parameter space that should not be used in the detection with real data.

7. Application to real Fermi-LAT data

The last step in our investigation of the \(\gamma\)-ray DBSCAN, is the application to real Fermi-LAT \(\gamma\)-ray data. We select the same region of the sky used for the simulated test field (80° < \(l < 170°\), and 40° < \(b < 65°\)), and we extract all the photons with energy

---

A. Tramacere and C. Vecchio: \(\gamma\)-ray DBSCAN application to Fermi-LAT \(\gamma\)-ray data.
A. Tramacere and C. Vecchio: γ-ray DBSCAN application to Fermi-LAT γ-ray data.

**Fig. 10.** Left panel: The distribution (blue line) of the square of the significance, for the fake clusters in the sky test field 1, for the full $K,\varepsilon$ parameter space, compared to a $\chi^2$ distribution with one degree of freedom. Right panel: the spurious ratio $D_{\text{fake}}$ for $S_{\text{cls}} > 4.0$, the white line shows the isolevel $D_{\text{fake}} = 0.0$.

**Fig. 11.** Aitoff projection of the Fermi sky region. The purple boxes represent the γ-ray DBSCAN sources ($K = 8, \varepsilon = 0.21$ deg.). The green crosses are the 2FGL sources with $TS > 16$, the red one those with $TS \leq 16$. There are no fake sources, and the γ-ray DBSCAN finds all the sources with $TS > 16$, except only one, enclosed by the red circle, and with the center positioned at the edge of the field.

$E > 3$ GeV. The photons are collected for the same time span of the 2FGL catalog. We repeat the detection test performed in the case of simulated data (see Sec. 5 and Sec. 6), restricting the parameter space to $2 \leq K \leq 10$, and $0.10 \leq \varepsilon \leq 0.30$ deg.

To properly understand the detection performance, we need to take into account that the 2FGL catalog has been built using photons with an energy threshold of 100 MeV, whilst we use a value of 3 GeV. A possibility is to select sources with a reported flux larger than zero, in the 3-10 GeV band flux column of the 2FGL. This flux-based selection, is not the best way to study the detection performance of the γ-ray DBSCAN, indeed the flux does not contain a unambiguous relation with the significance of the detection, for that energy threshold. A more reliable criterion is to select the sources according to the significance reported in the 2FGL. The 2FGL detection significance is given by the $\sqrt{TS}$. The $TS$ is the test statistic defined as
Fig. 13. $D_{\text{fake}}$ (left panels) and $D_{\text{eff}}$ (right panels), for the real sky detections, using the 2FGL$_{TS>16}$ catalog. Panels $a,b$: no cut on $S_{\text{cls}}$ applied. Panels $c,d$: $S_{\text{cls}} > 2.0$ Panels $e,f$: $S_{\text{cls}} > 4.0$
**Fig. 12.** The scatter plot the positional error of the γ-ray DBSCAN clusters vs. the positional error of the corresponding associated 2FGL$_{TS>16}$ sources. For each 2FGL$_{TS>16}$ source associated to one or more γ-ray DBSCAN clusters, we plot the error on the position of the reconstructed cluster centroid and its standard deviation (represented by the error bar). The dashed red line represents a linear best fit, with a slope of ≃ 0.99, and an intercept of ≃ 9.53.

**Fig. 14.** Left panel: scatter plot of $S_{cls}$ vs. $\sqrt{TS}$. For each source in our 2FGL$_{TS>16}$ list, associated to one or more γ-ray DBSCAN cluster, we plot the $\sqrt{TS}$ in the 3-10 GeV band, v.s. the average values of $S_{cls}$ and its standard deviation (represented by the error bar). Right panel: the distribution (blue line) of the square of the significance, for the fake clusters in the Fermi-LAT real sky, for the full $K,\varepsilon$ parameter space, compared to a $\chi^2$ distribution with one degree of freedom.

$TS = 2(\log L(\text{source}) - \log L(\text{no source}))$, where $L$ is the likelihood of the data given the model with or without a source present at a given position on the sky (Nolan et al. 2012). We apply a selection according to $\sqrt{TS} > 4$, and we refer to the corresponding source list (counting 35 sources) as 2FGL$_{TS>16}$.

An example of the application of the γ-ray DBSCAN to real Fermi-LAT data is given in Fig. 11, where we report an Aitoff projection in galactic coordinates of the analysed γ-ray sky region. The red crosses represent the 2FGL sources with $TS < 16$ in the 3-10 GeV band, and the green crosses represent those with $TS \geq 16$. The purple boxes represent the γ-ray DBSCAN sources found for $K = 8$, $\varepsilon = 0.21$ deg. For this choice of parameters, we find no fake sources, and we find all the sources with with $TS > 16$, except only one, enclosed by the red circle, and positioned at the edge of the sky region, with a galactic latitude $l = 64.85$ deg. In Tab. 2 we summarize the detection performance, for detections with a number of fake sources ≤ 4. We note that values of true clusters ranges between 35 and 34, out of the 35 present in the 2FGL$_{TS>16}$. The fake ones range between 1 and 4, and we obtain an average detection efficiency of $D_{eff} = 0.94$.

In Fig. 12 we compare the localization performance of the γ-ray DBSCAN algorithm with that returned by the likelihood analysis implemented in the Fermi Science Tools. For each source in our 2FGL$_{TS>16}$ list, associated to one or more γ-ray DBSCAN clusters, we plot the the error on the position of the reconstructed cluster centroid and its standard deviation (represented by the error bar) vs. the 95% positional uncertainty reported in the 2FGL. We evaluate the 2FGL 95% positional uncertainty as $\sqrt{\sigma_{95,\text{min}}^2 + \sigma_{95,\text{max}}^2}$, where $\sigma_{95,\text{min}}$ and $\sigma_{95,\text{max}}$ are the semimajor and semiminor axes of the 95% confidence source location region, respectively. The dashed red line represents a linear best fit, with a slope of ≃ 0.99, and an intercept of ≃ 9.53, showing that the error on the position of the reconstructed cluster centroid, performed with a threshold of 3 GeV, is of the same order of the 95% positional uncertainty reported in the 2FGL catalog, performed above 100 MeV.

To test the reliability of the significance $S_{cls}$ to reject spurious sources, in Fig. 13 we plot the $D_{fake}$ and $D_{eff}$, based on the 2FGL$_{TS>16}$ catalog. The panels a and b, correspond to the case of no selection on $S_{cls}$. Both the $D_{fake}$ and the $D_{eff}$ trends are very similar to the case of the simulated sky. If we apply a significance cut of $S_{cls} > 2.0$ (panels c,d), we observe that the number of spurious ratio is $D_{fake} \leq 0.05$ for almost half of the parameter space (region to the right of the purple line). The more severe cut of $S_{cls} > 4.0$ (panels d,e), removes all the fake clusters except two, for $\varepsilon \leq 0.15$ deg. Only for $\varepsilon \geq 0.25$ deg., the $D_{fake}$ ratio shows a significant increase, ranging from 0.05 up to ≃ 0.1. In agreement with our analysis on simulated data, the
region of the parameter space where $\epsilon$ is comparable to the PSF size, gives the better performance.

To have a further confirmation about the robustness of our significance, we plot in the right panel of Fig. 14 $S_{cls}$ vs. $\sqrt{T S}$. For each source in our 2FGL $T S > 16$ list, associated to one or more $\gamma$-ray DBSCAN cluster, we plot the $\sqrt{T S}$ in the 3-10 GeV band, vs. the average value of $S_{cls}$ and its standard deviation (represented by the error bar). The average value of $S_{cls}$ and its standard deviation are evaluated from the list of all the cluster associated to the same 2FGL source. The solid blue boxes represent the full $K, \epsilon$ parameter space case, and the red solid circles represent the $\epsilon = 0.10$ deg. case. The dashed black line represents a linear best fit. The slope of the linear fit is $\approx 0.5$. The strong correlation in the scatter plots ($r \approx 0.98$, for both data sets), proves that our significance implementation is consistent with the $\sqrt{T S}$ reported in the 2FGL, and the slope of the linear fit suggests that $S_{cls} \approx 0.5 \sqrt{T S}$.

### Table 2. Summary of the detection performance for the real Fermi-LAT field, for detections with a number of fake sources $\leq 4$.

| 2FGL $T S \geq 16$ True Fake $K$ $\epsilon$ $D_{eff}$ $Q$ |
|-----------------|-----------------|---------|-------|-------|-------|
| Fermi-sky       | 35              | 34      | 8     | 0.21  | 0.97  | 0.97  |
|                 | 35              | 34      | 1     | 7     | 0.20  | 0.94  | 0.92  |
|                 | 35              | 35      | 2     | 7     | 0.21  | 0.94  | 0.89  |
|                 | 35              | 35      | 4     | 6     | 0.19  | 0.89  | 0.79  |
| average         | 35              | 34.50   | 1.75  | 7.00  | 0.20  | 0.94  | 0.89  |

8. Conclusions

For the first time, we have used the DBSCAN for the detection of sources in $\gamma$-ray astrophysical images. We have implemented a new version of the DBSCAN, the $\gamma$-ray DBSCAN, that is optimized for the application to $\gamma$-ray astrophysical images, with relevant background noise. Our $\gamma$-ray DBSCAN, presents the novelty of recursive call of the DBSCAN algorithm, that allows an excellent reconstruction of the cluster, with an effective background rejection. We have tested the algorithm with a sample of simulated $\gamma$-ray Fermi-LAT fields, to give a statistical characterization of the method, and to benchmark the detection performance. The results, with the simulated $\gamma$-ray data, are summarized by the following items:

- The radius of $\gamma$-ray DBSCAN scanning brush $\epsilon$, has a strong correlation with the instrumental PSF radius. We find that the typical size of the reconstructed true cluster is of the order of the simulated PSF size $\sigma_{sim}$, and that the precision of the reconstructed centroid is of the order of $\sigma_{sim}/\sqrt{N_p}$.
- The number of reconstructed events $N_p$ is ruled by the Poissonian statistics in the random fields, and for the fake clusters. On the contrary, for true clusters, the statistics of $N_p$ is ruled by that of the simulated sources.
- The fractional error on the reconstructed events number is of the order of 20% for $N_p, sim, \leq 50$, and is negligible for larger values, with best performance obtained when $\epsilon = \sigma_{sim}$.
- We have investigated the detection performance, for a wide range of the $K, \epsilon$ parameter space, and we have identified the region with the best performance in terms of detection efficiency, and spurious ratio.
- We have implemented an algorithm for the estimate of the Signal to Noise (S/N) ratio, able to deal with local background inhomogeneities and nearby sources contamination, and we have successfully used the S/N estimate to determine the significance of the clusters, using the definition in Li & Ma (1983).
- Our cluster significance, $S_{cls}$, for random clusters, follows the $\chi^2(1)$ statistics, and can be used to reject spurious sources. The chance to find spurious sources for $S_{cls} > 4$, is negligible. This means, that our $S_{cls}$ is a robust a reliable tool to reject spurious sources, and that $\chi^2(1)$ statistics can be used to evaluate the probability of a cluster to be spurious.

We have successfully applied the $\gamma$-ray DBSCAN to real Fermi-LAT data. We have found an excellent agreement with results from the simulated fields. We tested our detection performance using as catalog, the 2FGL sourced with a $\sqrt{T S} > 4$ cut. The results, with the real Fermi-LAT $\gamma$-ray data, are summarized by the following items:

- the error on the position of the reconstructed cluster centroid, performed with a threshold of 3 GeV, is of the same order of the 95% positional uncertainty reported in the 2FGL, performed above 100 MeV.
- We tested the $\gamma$-ray DBSCAN significance, finding that it is strongly correlated with the $T S$ provided in the 2FGL. The significance cut, allows to remove safely spurious clusters.
- The detection efficiency with real data is excellent, we are able to find all the 35 sources with $\sqrt{T S} > 4$.
- When working with $\epsilon$ of the order of the instrumental PSF size, we obtain the best performance, in terms of spurious rejection, and detection efficiency.

In general, we find that the $\gamma$-ray DBSCAN is a very powerful detection method to find clusters in $\gamma$-ray images, corresponding to real sources. It has the great advantage to deal self-consistently with gradient in the background, providing an effective rejection of spurious clusters. Our implementation of the detection significance, in addition to the algorithm to evaluate local fluctuations in the background, allows to apply statistically significant selection, making even more effective the rejection of spurious sources.

In a companion paper [Tramacere 2012], we will apply the method to the Fermi-LAT sky, showing the potentiality for the discovery of new sources, in particular of small clusters located at high galactic latitude, or cluster on the galactic plane, affected by a strong background. We will also investigate how to plug the energy dependence of the PSF into the $\gamma$-ray DBSCAN algorithm, and how to improve the detection performance taking into account other Fermi-LAT calibration properties.

We remark that, since the $\gamma$-ray DBSCAN provides also density maps, it can potentially be used in the detection of large scale structures in the galactic $\gamma$-ray background, providing patterns to compare to the interstellar gas distribution. We also stress, that the application of this method are not limited to $\gamma$-ray images, but can be potentially used for any application related to the detection of spatial, and/or spatio/temporal clusters.

Acknowledgements. We are grateful to Enrico Massaro, Riccardo Campa, and Enrico Bernieri, for helpful comments, and for providing us the simulated test fields. We are grateful to Gino Tosti, for helpful comments. We thank the anonymous referee for providing us with constructive comments and useful suggestions.

References

Campa, R., Massaro, E., Bernieri, E., Tinebra, F. & Tosti, G. 2012, submitted Campa, R., Massaro, E., Gasparrini, D., Cutini, S. & Tramacere, A. 2007, Monthly Notices of the Royal Astronomical Society, 383, 1166
Ester, M., Kriegel, H., Sander, J., & Xu, X. 1996, In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining
Fermi-LAT Collaboration. 2012, eprint arXiv, 1206, 1896
Jolliffe, I. T. 1986, Principal component analysis
Li, T.-P. & Ma, Y.-Q. 1983, Astrophysical Journal, 272, 317
Nolan, P. L., Abdo, A. A., Ackermann, M., et al. 2012, The Astrophysical Journal Supplement, 199, 31
Papoulis, A. 1965, Probability, Random Variables and Stochastic Processes
Tramacere, A. 2012, In prep.