**γ-ray DBSCAN: a clustering algorithm applied to Fermi-LAT γ-ray data**

I. Detection performances with real and simulated data

A. Tramacere¹ and C. Vecchio²

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

Received 30 July 2012 / Accepted 30 September 2012

**ABSTRACT**

**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 this method to detect sources in γ-ray astrophysical images obtained from the Fermi-LAT data, where each point corresponds to the arrival direction of a photon.

**Aims.** We investigate the detection performance of the γ-ray DBSCAN in terms of detection efficiency and rejection of spurious clusters.

**Methods.** We used a parametric approach, exploring a large volume of the γ-ray DBSCAN parameter space. By means of simulated data we statistically characterized the γ-ray DBSCAN, finding signatures that distinguish purely random fields from fields with sources. We defined a significance level for the detected clusters and successfully tested this significance with our simulated data. We applied the method to real data and found an excellent agreement with the results obtained with simulated data.

**Results.** We find that the γ-ray DBSCAN can be successfully used in detecting clusters in γ-ray data. The significance returned by our algorithm is strongly correlated with that provided by the maximum likelihood analysis with standard 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 one to look for astrophysical counterparts in narrow regions, which minimizes the chance probability in the counterpart association.

**Conclusions.** We found that γ-ray DBSCAN is a powerful tool for detecting of clusters in γ-ray data. It can be used to look for both point-like sources and extended sources, and can be potentially applied to any astrophysical field related to detecting clusters in data. In a companion paper we will present the application of the γ-ray DBSCAN to the full Fermi-LAT sky, discussing the potential of the algorithm to discover new sources.

**Key words.** gamma rays: general – methods: statistical – methods: data analysis – methods: numerical

---

1. **Introduction**

Modern γ-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. Discrete sources (either point-like or extended) are detected with various methods. Given the discrete topological nature of γ-ray images, methods based on cluster search, such as the minimum spanning three (MST; Campana et al. 2007, 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 γ-ray telescopes, and to preserve 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 the background properly into account. The problem of background rejection is the most penalizing feature of topometric methods. Therefore we here present for the first time a method based on the density based spatial clustering of applications with noise (DBSCAN) algorithm (Ester et al. 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 the discrimination between signal (cluster) and background (noise) inside the algorithm itself, 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 potential of the method, and its statistical characterization when applied to astrophysical γ-ray data. We applied this method to detect point-like sources in the Fermi-LAT data. We explored a large volume of the γ-ray DBSCAN parameter space by means of simulated data, and we provide a statistical characterization of the γ-ray DBSCAN, finding signatures that distinguish purely random fields from fields with sources. We defined a significance level for the detected clusters, and we successfully tested this significance with our simulated data. We applied the method to real Fermi-LAT γ-ray data and we found an excellent agreement with the results obtained with simulated data.

In a companion paper (Tramacere, in prep.), we will apply the method to the Fermi-LAT sky, investigating specific questions related to the Fermi-LAT response functions, showing the potential to the discovery new sources, in particular of small clusters located at high galactic latitude, or clusters on the Galactic plane that are affected by a strong background.
The paper is organized as follows. In Sect. 2 we describe the logic of the DBSCAN method and present the algorithm implemented to analyze γ-ray data, the γ-ray DBSCAN. In Sect. 3 we discuss some caveats regarding the application of the γ-ray DBSCAN algorithm to γ-ray data. In Sect. 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 Sect. 5 we evaluate the simulated test field with only noise, and five simulated test fields with statistical properties of the γ-ray DBSCAN clusters significance to that returned by the maximum likelihood method in terms of positional accuracy, cluster reconstruction, and rejection of spurious clusters. In Sect. 6 we investigate the significance of the clusters and describe our algorithmic implementation. In Sect. 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 using a weighted average of the position of each photon in $C_m$, as follows:

1. In a list of photons $D$, where each element $p_i$ is a tuple storing positional sky coordinates, $\rho(p_i, p_j)$ is the angular distance between two photons $p_i$ and $p_j$.
2. We iterate over the full photon list $D$. A seed cluster $C^*_m$ is built when a minimum number of photons $K + 1$ is enclosed within a circle of radius $\varepsilon$ centered on $p_i$.
3. For each photon $p_j \in C^*_m$, we build the photon list $C^*_j$ by collecting all photons $p_k$ that meet the condition $\rho(p_j, p_k) < \varepsilon$, and $p_k \notin C^*_m$.
4. For each photon $p_j \in C^*_m$, if the number of photons enclosed within a circle of radius $\varepsilon$ centered on $p_j$ is $\leq K$ and $p_j \notin C^*_m$, then $p_j$ will be attached to the final photon list of the cluster without a recursive search for more neighbors, these points are defined density-reachable.
5. For each photon $p_j \in C^*_m$, if the number of photons enclosed within a circle of radius $\varepsilon$ centered on $p_j$ is $> K$ and $p_j \notin C^*_m$, $p_j$ is attached to the $C^*_m$, and step 3 is repeated recursively.
6. When both conditions at step 4 and 5 are false, the cluster $C^*_m$ is built by joining the density-reachable events to those in the $C^*_m$ and in the $C^*_m$ lists.
7. The process starts again from step 1 searching for new clusters, skipping the events already flagged as noise or clusters, until all events in $D$ are flagged as cluster, or noise, or density-reachable events.
8. At the end of the process the full photon list will be partitioned as follows:

$$D = D_{\text{clus}} \cup D_{\text{noise}} = (p_i \in \bigcup_{m} C^*_m) \cup (p_i \notin \bigcup_m C^*_m)$$

$$\emptyset = D_{\text{clus}} \cap D_{\text{noise}}.$$

In this way high-density areas are classified as clusters (sources), conversely low-density areas are classified as noise (background). The recursive call of step 3 is not implemented in the original DBSCAN algorithm and represents a novelty. This new feature allows us to reconstruct clusters significantly larger than the resolution radius, which makes it unlikely that a single cluster is fragmented into small satellite clusters. Moreover, it allows the possibility to reconstruct extended structures, in particular extended sources, or filamentary structures in the background.

After the clustering process, each photon in $D$ will be described by a tuple storing the photon position (both in galactic and celestial coordinates), the photon class type (noise or cluster), 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 ($r_{\text{eff}}$), 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 $\sigma_y$) and the inclination angle ($\sigma_{\text{dphi}}$) of the major semi-axis w.r.t. the latitudinal coordinate ($b$ or Dec). To evaluate the ellipse axis we use the principal component analysis method (PCA; Jolliffe 1986). 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 $\sigma_x$ and the second to $\sigma_y$. The axes represent the two orthogonal directions of maximum variance of the cluster. The effective radius is defined as $r_{\text{eff}} = \sqrt{\sigma_x^2 + \sigma_y^2}$. To find the centroid of the cluster and its uncertainty, we used a weighted average of the position of each photon in $C_m$, as follows:

- We define the first order centroid ($C_{\text{ave}}$) as the average of the position of each cluster photon: $C_{\text{ave}} = \langle (x), (y) \rangle$.
- We define the weight array, according to the distance between $p_k \in C_m$ and $C_{\text{ave}}$: $w_k = 1/\rho(p_k, C_{\text{ave}})$.
- The cluster centroid $C_{\text{cen}}$ will result from the average of the position of each cluster point weighted by $w_k$.
- The centroid position uncertainty ($pos_{\text{err}}$) is determined by propagating the error on the weighted average of $C_{\text{cen}}$. We have numerically verified that $pos_{\text{err}}$ 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 practical difficulties that are mostly related to the instrument PSF and to gradient and/or structures in the background. To deal with these problems without biasing the detection results, we recommend to apply some criteria that we discuss in the following.

We first comment on the PSF impact. The PSF imposes a limit on the capability of an instrument to resolve sources separated by a distance smaller than the PSF size. Sources with sizes smaller than the PSF are classified as point-like, otherwise they are classified as extended. Another 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 size of about 5 degrees at 100 MeV (Ackermann et al. 2012), and scales down to a few tenths of degree above the GeV energies. The size of the PSF is strongly connected to the size $\varepsilon$ of the γ-ray DBSCAN scanning brush. Indeed, if $\varepsilon$ 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 false satellite clusters. We stress that the formation of satellite clusters is a very rare event, thanks to our recursive DBSCAN implementation, which we explain in Sect. 2. In contrast, if $\varepsilon$ is much larger than

---

the PSF, it is likely to build extended clusters contaminated by the background or by close sources.

A careful and self-consistent analysis of the effects of the energy dependence of the PSF, and in general of problems 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 topics will be investigated in the companion paper (Tramacere, in prep.).

A second relevant problem, is the inhomogeneity of the background, which affects the choice of both $e$ and $K$. If the background is homogeneous over the entire field, the best choice of a single pair of values of $e$ and $K$ guarantees a safe rejection of the background. Indeed, values of $e$ and $K$ such that the average density of photons within $e$ is significantly higher than the average density of the background photons, make it unlikely that a cluster grows from a background fluctuation. Unfortunately, the γ-ray sky shows strong gradients of background, in particular at low galactic latitudes. To solve this problem, one could think to adapt the value of $e$ and $K$ according to a local value of the background photon density. Since $e$ has a strong constraint imposed by the PSF, one should tune mainly 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 clusters with small $N_p$. To overcome this difficulty, we adopted an alternative solution. We used a unique pair of values of $e$ and $K$ for each field, where $e$ is mostly constrained by the PSF, and $K$ by the field average background, and we take into account the background inhomogeneities by defining a significance level of the cluster, according to the signal-to-noise ratio ($S/N$; Li & Ma 1983), evaluated from the local background. This is explained in detail in Sect. 6. The capability to reject clusters according to a low significance level allows one to relax the constrain on $e$ and $K$, increasing the number of clusters detected, hence increasing the detection ratio, and at the same time allows one to reject spurious sources, because of the significance threshold. To 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. To optimize the ratio between background and cluster events, we use a threshold energy of 3 GeV in the following, that mitigates the possible bias caused by the background fluctuations.

4. Statistical properties of the γ-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 as in 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 γ-ray background was simulated using the standard gtobssim\footnote{http://fermi.gsfc.nasa.gov/ssc/data/analysis/scitools/help/gtobssim.txt} tool, developed by the Fermi-LAT collaboration, simulating both the Galactic and isotropic components for a two-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 from a minimum value of 4 up to 40 photons, joined to a constant tail up to 240 photons. The number of the sources is similar to that reported in the Fermi-LAT Second Source Catalog (Nolan et al. 2012, 2FGL hereafter), in the same region of the sky. The source events are spatially distributed with a bivariate Gaussian probability density function (PDF) with $\sigma_y^{\text{sim}} = \sigma_y^{\text{sim}} = 0.2$ deg. centered at the source location. Five simulated test fields were 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 as the sky test fields and a number of events equal to the sky test field-1 (background and sources) for a total amount of 11 044 events.

In Fig. 1 we show the photon map for the sky test field 1 and the result of the γ-ray DBSCAN detection for $K = 5$ and $e = 0.17$ deg. We detect 51 true clusters, and only 2 fake ones. A cluster is defined as true if the position of the simulated source falls within a circle centered on the cluster centroid, with a radius equal to 20 pixels. We call the remaining clusters fake. In Fig. 2, we show a close-up of two true clusters. The black ellipses correspond to the ellipses of the cluster containment, and the purple and orange thick points represent the cluster points, while the black thick dots represent the background.

4.2. Test strategy

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

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

We started by investigating the distribution of the $\log(r_{\text{eff}})$ values for the random and the sky test field 1. The distribution for the detections collected over the full $K$-$e$ parameter space (top left panel of Fig. 3) shows a symmetric shape well fitted by a Gaussian distribution (log-normal w.r.t. $r_{\text{eff}}$), with the mean value of $\langle \log(r_{\text{eff}}) \rangle \simeq -0.45$ (corresponding to $(r_{\text{eff}}) \simeq 0.3$ deg) and a dispersion of $\sigma_{\log(r_{\text{eff}})} \simeq 0.23$. The log-normal distribution provides a reasonable description of the empirical distributions also for individual pairs of $(K,e)$ values. An example is given in panel c of Fig. 3 for $K = 3$, $e = 0.3$ deg, where the best fit values are $(\log(r_{\text{eff}})) \simeq -0.51$, and $\sigma_{\log(r_{\text{eff}})} \simeq 0.16$. 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 a normal. For fake clusters (red dashed line) the best fit values of the mean $(\langle \log(r_{\text{eff}}) \rangle \simeq -0.46)$ and of the dispersion $(\sigma_{\log(r_{\text{eff}})} \simeq 0.24)$ are very similar to those found for the random test field. In contrast, the true cluster distribution (blue hatched histogram) is peaking around the value of
Fig. 1. Photon map for the sky test field 1, with the result of the $\gamma$-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.

Fig. 2. Close-up of two true clusters reported in Fig. 1. The ellipses correspond to the ellipse of the cluster containment. The purple and orange points represent the cluster points, the black dots represent the background events, the blue crosses the position of the simulated sources, and green boxes the position of the cluster centroid.

$\sigma_{\text{sim}}$ reproduces the effect of the instrumental PSF, we observe that the typical size of the reconstructed clusters for non-random fields, is constrained by the PSF, suggesting that the empirical rule is to set the value of $\varepsilon$ of about the PSF size.

To investigate the connection between $\varepsilon$ and the PSF more accurately, we analyzed the statistical properties of the quantity $r_{\text{eff}}/\varepsilon$ as a function of $\varepsilon$. For each value of $\varepsilon$, we determined 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 $r_{\text{eff}}/\varepsilon = 1$, but the upper boundary shows a systematic increase compared to the lower boundary for $\varepsilon \gtrsim 0.30$ deg. The trend for the true clusters in sky test field 1 (right panel Fig. 4), shows a different behavior. 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 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. For random fields we expect that the number of photons in a cluster follows a Poisson distribution. Indeed, for a generic two-dimensional Poisson process, 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

$$P(N(\varepsilon^2) = j) = \frac{(\lambda |\varepsilon|^2)^j \exp\left(-|\varepsilon|^{2}\right)}{j!}, \quad (3)$$

from which it 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

$$P_{\text{clus}}(\varepsilon, K) = P(N(\varepsilon^2) > K) = 1 - \sum_{j=0}^{\infty} \frac{(\lambda |\varepsilon|^2)^j \exp\left(-|\varepsilon|^{2}\right)}{j!}, \quad (4)$$

that is the Poissonian survival function. Owing to the logic of the DBSCAN clustering process, the Poisson statistics cannot 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, the seed of the cluster ($C^*$). If we consider the annulus defined between $\varepsilon$ and the cluster radius $r_{\text{clus}}$, not all points in the annulus will be cluster members, but only those that are at least density reachable. This implies that we expect a deviation from the Poisson statistics, when $r_{\text{eff}}$ is significantly larger than $\varepsilon$, i.e. $\varepsilon \gtrsim 0.3$ 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 $K = 2$ and $\varepsilon = 0.20$ deg. We note that the Poisson distribution shows stronger deviations, in particular for $K > 6$. When we take into account the $N_p$ distribution for the full parameter space (panel c), the Poissonian distribution is failing to provide a reasonable description of the empirical distribution, whilst a log-normal distribution gives a good fit.

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

$$f(N_p) \propto f(r_{\text{eff}}^2) \lambda. \quad (5)$$

To evaluate the distribution of $r_{\text{eff}}^2$, we can use the standard transformation theory of random variables (RV; Papoulis 1965). It can be easily proved that for an RV $X$ with a log-normal distribution,

$$f(X) = \frac{1}{X \sqrt{2\pi\sigma^2}} \exp\left(-\frac{\ln(X)^2 - \mu}{2\sigma^2}\right), \quad (6)$$

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 $K = 3$ and $\varepsilon = 0.3$ deg. Panel c) distribution of $\log_{10} r_{\text{eff}}$ for the sky test field 1, for fake clusters (red solid line), and true clusters (blue solid line, hatched histogram). The dashed lines represent the 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 gray 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 sky test field 1. The red solid circles represent the median of the true clusters case, and the grey area the 1-σ confidence level region. The dashed line shows the 1-σ confidence level region, for the fake clusters.
the RV \( Y = X^2 \) will follow a log-normal distribution given by

\[
f(Y) = \frac{1}{2Y\sqrt{2\pi\sigma}} \exp\left( -\frac{\ln(Y) - 2Y}{4\sigma^2} \right). \tag{7}\]

Indeed, our \( r_{\text{eff}}^2 \) 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. (5) we expect that also \( f(N_p) \) will follow a log-normal distribution, when \( N_p \) is not ruled by a Poissonian statistics.

We verified that the same statistical trends describe the real sky fields. Panels e and f in Fig. 5 show the statistical distribution of \( N_p \) for the sky test field 1. In agreement with the analysis for 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) than by a Poissonian distribution (solid black line). We also observe that the log-normal law reasonably describes the empirical distribution only for values of \( N_p \leq 50 \), but shows significant deviation in the tail, consistent with the statistics of our simulated source population.

To complete this statistical characterization, we investigated 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 clusters for a random field follows a Poisson survival distribution. 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 fake clusters of the sky test field 1 (plot c Fig. 6). In contrast, for true clusters (panel d Fig. 6), the Poisson survival distribution is unable to reproduce the observed trend, consistently with the non-Poissonian statistic of the simulated clusters. Panels b and e of Fig. 6s show the 1–\( \sigma \) CL region for \( N_p \) as a function of \( K \). Both for random and sky field true clusters the lower boundary of the region is constrained by the equation \( y = K + 1 \), which is consistent with the \( \gamma \)-ray DBSCAN logic. In contrast, the upper boundary shows a different behavior. For the random field, the upper boundary deviates from the lower boundary, which is compatible with the fluctuations of the events around the \( \varepsilon \) circle, and ranges from about 8 to about 16. In contrast, for the true sky field the upper boundary is constrained by the statistics of the number of events in the simulated sources, and ranges from about 60 to 100.

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

In this section we investigate the detection performance of the \( \gamma \)-ray DBSCAN. We first 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
We ran for each of the five sky test fields used in the previous section, exploring the same parameter space.

mark the five sky test fields used in the previous section, explored detection performance of the γ-rays. We test the positional accuracy of the reconstructed centroids. We test the detection performance of the γ-ray DBSCAN, using as benchmark the five sky test fields used in the previous section, exploring the same parameter space.

5.1. Detection efficiency and spurious ratio as a function of K and ε

To investigate the detection performance of the γ-ray DBSCAN, we ran for each of the five sky test fields and for each pair of values K, ε, a γ-ray DBSCAN detection. For each detection run, we built a cluster catalog. Starting from this, we built the corresponding candidate catalog. This 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 as 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 2σ\text{src}.
- Two or more true clusters are defined as confused if they have the same counterpart.
- A true cluster has a multiple association if it 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. Moreover, 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}. To characterize the performance, we define the following parameters:

- the detection efficiency:

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

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

- the true detection ratio \( D_{\text{true}} = N_{\text{true}}/N_{\text{src}} \)

- the spurious detection ratio \( D_{\text{false}} = N_{\text{false}}/N_{\text{src}} \)

- the overall detection quality factor (Q), which takes into account the tradeoff between \( D_{\text{eff}} \) and \( D_{\text{false}} \), defined as

\[ Q = D_{\text{eff}} \left( 1 - \frac{N_{\text{false}}}{N_{\text{src}}} \right). \]
reason, we then report a value of $D_{\text{eff}} = 1.0$. The same applies to $Q$.

In Fig. 7 we summarize the detection runs for sky test field 1 for the full parameters space with $K > 2$. 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 $D_{\text{fake}}$ isolevel map to the $D_{\text{true}}$ isolevel map (panel b Fig. 7). The map shows also in this case a sharp gradient, and the region with $D_{\text{true}} > 0.95$ overlaps the $D_{\text{fake}} = 0$ 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 $D_{\text{fake}}$ ratios do not provide information on the ratio between the number of true detected clusters and the number of simulated clusters. For this point more information is provided by the $D_{\text{eff}}$ isolevel map (panel c, Fig. 7). To focus on the “effective” volume of the parameter space, we hide the region where $D_{\text{eff}} < 0$ with a white area. 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 $\varepsilon$, meaning that an increased value of $\varepsilon$ requires an increased value of $K$ to obtain a better background rejection. To evaluate the trade-off between $D_{\text{true}}$ and $D_{\text{fake}}$, we plot in 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 for $D_{\text{eff}}$. In Table 1 we report the $D_{\text{eff}}$ values obtained for all five sky fields, for detections with a number of fake sources $\leq 6$. 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 topometric methods, is probably the most important feature of this class of algorithms. In Sect. 2, we have described our weighting method to reconstruct the centroid of the cluster.

Panel a of Fig. 8 shows with red solid boxes the mean positional error of the cluster centroids and the standard deviation (vertical error bar) vs. $N_p$, for the true clusters of sky test field 1 with $\varepsilon \leq 30$ 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
Table 1. Summary of the detections obtained for all the five sky fields for detections with a number of fake sources \(\leq 6\).

<table>
<thead>
<tr>
<th>Sky field</th>
<th>(N_{\text{sim}}^{\text{cut}} &gt; K)</th>
<th>True</th>
<th>Fake</th>
<th>(\epsilon)</th>
<th>(D_{\text{off}})</th>
<th>(Q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sky field 1</td>
<td>44/70</td>
<td>47</td>
<td>1</td>
<td>6</td>
<td>0.19</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>38/70</td>
<td>48</td>
<td>2</td>
<td>8</td>
<td>0.27</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>53/70</td>
<td>50</td>
<td>3</td>
<td>5</td>
<td>0.17</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>62/70</td>
<td>51</td>
<td>5</td>
<td>4</td>
<td>0.14</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>53/70</td>
<td>53</td>
<td>6</td>
<td>5</td>
<td>0.18</td>
<td>0.89</td>
</tr>
<tr>
<td>Sky field 2</td>
<td>50.00</td>
<td>49.80</td>
<td>3.40</td>
<td>5.60</td>
<td>0.19</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>62/70</td>
<td>47</td>
<td>0</td>
<td>4</td>
<td>0.11</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>44/70</td>
<td>49</td>
<td>2</td>
<td>6</td>
<td>0.20</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>53/70</td>
<td>51</td>
<td>3</td>
<td>5</td>
<td>0.17</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>62/70</td>
<td>52</td>
<td>6</td>
<td>4</td>
<td>0.14</td>
<td>0.74</td>
</tr>
<tr>
<td>Sky field 3</td>
<td>55.25</td>
<td>49.75</td>
<td>2.75</td>
<td>4.75</td>
<td>0.15</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>41/70</td>
<td>42</td>
<td>0</td>
<td>7</td>
<td>0.22</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>44/70</td>
<td>45</td>
<td>2</td>
<td>6</td>
<td>0.20</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>53/70</td>
<td>46</td>
<td>3</td>
<td>5</td>
<td>0.16</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>62/70</td>
<td>48</td>
<td>5</td>
<td>4</td>
<td>0.14</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>62/70</td>
<td>50</td>
<td>6</td>
<td>4</td>
<td>0.15</td>
<td>0.71</td>
</tr>
<tr>
<td>Sky field 4</td>
<td>52.40</td>
<td>46.20</td>
<td>3.20</td>
<td>5.20</td>
<td>0.17</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>53/70</td>
<td>47</td>
<td>1</td>
<td>5</td>
<td>0.16</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>53/70</td>
<td>50</td>
<td>5</td>
<td>5</td>
<td>0.18</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>62/70</td>
<td>52</td>
<td>6</td>
<td>4</td>
<td>0.14</td>
<td>0.74</td>
</tr>
<tr>
<td>Sky field 5</td>
<td>56.00</td>
<td>49.67</td>
<td>4.00</td>
<td>4.67</td>
<td>0.16</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>44/70</td>
<td>47</td>
<td>1</td>
<td>6</td>
<td>0.19</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>44/70</td>
<td>50</td>
<td>2</td>
<td>6</td>
<td>0.20</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>44/70</td>
<td>53</td>
<td>3</td>
<td>6</td>
<td>0.21</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>44/70</td>
<td>55</td>
<td>5</td>
<td>6</td>
<td>0.22</td>
<td>1.00</td>
</tr>
<tr>
<td>Average</td>
<td>44.00</td>
<td>51.25</td>
<td>2.75</td>
<td>6.00</td>
<td>0.20</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Notes. \(N_{\text{sim}}^{\text{cut}} > K\) 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.

\[
p_{\text{pos}} \approx \frac{\sigma_{\text{sim}}}{\sqrt{N_p}}\] (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\'. In panel b of Fig. 8 we plot the distribution histogram 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', for 68% of the sample.

In addition to 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 (\(N_{\text{sim}}\)). The solid points represent the average value of \(N_p\), for a given value of \(N_{\text{sim}}\), and...
Fig. 10. Left panel: the distribution (blue line) of the square of the significance for the fake clusters in 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 $D_{\text{fake}} = 0.0$.

the error bar corresponds to the standard deviation. The solid green line represents $N_p = N_{p \text{ sim.}}$, and the dashed upper and lower lines represent $N_p = N_{p \text{ sim.}} \pm 10$ sim. For $\varepsilon = 0.15$ deg and $\varepsilon = 0.20$ deg, the scatter is bounded by the dashed lines, showing that the highest excess in the $N_p$ is about 10 photons, independently of $N_p$ sim. For $\varepsilon = 0.15$ deg, the number of reconstructed photons systematically underestimates the simulated number, whilst the $\varepsilon = 0.20$ deg case does not show this bias. It is possible to appreciate this effect better in the bottom left panel of Fig. 9, where we show the fractional reconstruction error ($N_p - N_{p \text{ sim.}})/N_{p \text{ sim.}}$ vs. $N_p$ sim. The solid green line represents the 0 error, and the dashed lines represent the ±20% boundaries.

The bias on $N_p$ for $\varepsilon = 0.15$ deg again shows the strong correlation between $\varepsilon$ and the PSF radius. When $\varepsilon$ is smaller than the $\sigma^{\text{sim.}}$ (that in our simulations reproduces the PSF effect), the number of reconstructed events $N_p$ is systematically smaller than $N_p$ sim., in contrast, when the $\varepsilon$ radius matches the PSF radius size ($\varepsilon = 0.20$ deg), the bias disappears.

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 higher and the probability to detect fake clusters is lower, in the application to real data it is mandatory to provide a significance level that expresses the probability that a cluster is not originated in a background fluctuation. We propose a method derived from the Li & Ma (1983) approach, based on the evaluation of the S/N. A significance method based on the S/N fits the $\gamma$-ray DBSCAN implementation well, because the algorithm directly provides a partition of the photon list in cluster and noise events. Hence, for each cluster we can easily evaluate the S/N, 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 $r_{\text{in}}$ and an external radius $r_{\text{out}}$.

2. $r_{\text{in}}$ is set to an initial value of $r_{\text{in}} = 2r_{\text{eff}}$, and is adaptively increased with a step of $r_{\text{in}}/10$ for a maximum of 10 trials until at least the 95% of the cluster events are enclosed within $r_{\text{in}}$.

3. $r_{\text{out}}$ is set to $3r_{\text{in}}$.

4. We count all cluster events $N_{\text{src}}^{\text{sim.}}$ and the background events $N_{\text{bkg}}^{\text{sim.}}$ that are enclosed within the circle with radius $r_{\text{in}}$ and are centered on the cluster centroid.

5. We determine the $N_{\text{bkg}}^{\text{sim.}}$ background level, rescaling the number of background events in $r_{\text{in}} < r < r_{\text{out}}$, to a circle with radius $r_{\text{in}}$.

6. To evaluate possible gradients in the background, we select a region far enough from the cluster to properly sample the background level, and close enough to the cluster to measure a local background level. For this, we define the radius $r_{\text{out}}^{\text{local}} = (r_{\text{out}} + r_{\text{in}})/2$, and evaluate the average background level ($N_{\text{bkg}}^{\text{local}}$) in a circle of radius $r_{\text{in}}$, centered on each point in $r_{\text{out}}^{\text{local}} < r < r_{\text{out}}$.

7. If no background points are found in $r_{\text{out}}^{\text{local}} < r < r_{\text{out}}$, we set $N_{\text{bkg}}^{\text{local}} = N_{\text{bkg}}^{\text{in}}$.

8. By comparing $N_{\text{bkg}}^{\text{local}}$ to $N_{\text{src}}^{\text{sim.}}$, we evaluate the fraction of noise already resolved by the $\gamma$-ray DBSCAN and evaluate the effective background level $N_{\text{bkg}}^{\text{eff.}}$, by correcting $N_{\text{bkg}}^{\text{local}}$ for $N_{\text{bkg}}^{\text{in}}$.

9. We evaluate the significance according to the Likelihood Ratio Test (LRT) method proposed by Li & Ma (1983):

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

Assuming that a cluster is due to a background fluctuation, the variable $S_{\text{cls}}^2$ is expected to follow a chi square distribution, with one degree of freedom ($\chi^2(1)$). In the left panel of Fig. 10, we plot the distribution of $S_{\text{cls}}^2$ for the fake clusters in sky test field 1 (blue histogram), compared to a $\chi^2(1)$ distribution. The empirical distribution is well described by the expected $\chi^2(1)$ distribution, proving that the value of $S_{\text{cls}}$ can be used as the “significance” of the detected cluster. A highly illustrating 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 isolevel map, applying the selection $S_{\text{cls}} > 4.0$. The fake ratio is 0 for the parameter space with $\epsilon \leq 0.25$ deg. For $0.25 \, \text{deg} \leq \epsilon \leq 0.35$ deg, there are fluctuations showing $D_{\text{fake}} \leq 0.05$. The fake ratio shows a significant increase only for $\epsilon \geq 0.40$ deg and $K \leq 8$, but we stress that in this region of the parameter space $\epsilon$ is more than twice 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 selected the same region of the sky as was used for the simulated test field (80° < $l$ < 170°, and 40° < $b$ < 65°) and extracted all the photons with energy $E > 3$ GeV. The photons are collected for the same time span of the 2FGL catalog. We repeated the detection test performed with simulated data (see Sects. 5 and 6), restricting the parameter space to $2 \leq K \leq 10$ and $0.10 \leq \epsilon \leq 0.30$ 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 used a value of 3 GeV. A possibility is to select sources with a reported flux higher 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 $\gamma$-ray DBSCAN, indeed, the flux does not contain an 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 $\sqrt{T S}$. The $T S$ is the test statistic defined as $T S = 2(\log L(\text{source}) - \log L(\text{no source}))$, 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 applied a selection according to $\sqrt{T S} > 4$ and refer to the corresponding source list (counting 35 sources) as 2FGL$_{TS>16}$.

An example of the application of the $\gamma$-ray DBSCAN to real Fermi-LAT data is given in Fig. 11, where we report an Aitoff projection in galactic coordinates of the analyzed $\gamma$-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 $\gamma$-ray DBSCAN sources found for $K = 8$, $\epsilon = 0.21$ deg. For this choice of parameters, we find no fake sources, and we find all sources with $TS > 16$, except for one that is enclosed by the red circle and is positioned at the edge of the sky region, with a galactic latitude $l = 64.85$ deg. In Table 2 we summarize the detection performance for detections with a number of fake sources $\leq 4$. Values of true clusters range between 1 and 16. The fake ones range between 1 and 4, and we obtain an average detection efficiency of $D_{\text{eff}} = 0.94$.

In Fig. 12 we compare the localization performance of the $\gamma$-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 $\gamma$-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. The dashed red line represents a linear best fit with a slope of $\approx 0.99$ and an intercept of $\approx 9.53$, 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

Fig. 11. Aitoff projection of the Fermi sky region. The purple boxes represent the $\gamma$-ray DBSCAN sources ($K = 8, \epsilon = 0.21$ deg). The green crosses are the 2FGL sources with $TS > 16$, the red crosses those with $TS \leq 16$. There are no fake sources, and the $\gamma$-ray DBSCAN finds all the sources with $TS > 16$, except one, enclosed by the red circle, whose center is positioned at the edge of the field.
Table 2. Summary of the detection performance for the real Fermi-LAT field for detections with a number of fake sources $\leq 4$.

<table>
<thead>
<tr>
<th>2FGL$_{TS\geq16}$ True</th>
<th>Fake</th>
<th>$K$</th>
<th>$\epsilon$</th>
<th>$D_{\text{eff}}$</th>
<th>$Q$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fermi-sky</td>
<td>35</td>
<td>34</td>
<td>0</td>
<td>8</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>35</td>
<td>34</td>
<td>1</td>
<td>7</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>35</td>
<td>35</td>
<td>2</td>
<td>7</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>35</td>
<td>35</td>
<td>4</td>
<td>6</td>
<td>0.19</td>
</tr>
<tr>
<td>Average</td>
<td>35</td>
<td>34.5</td>
<td>1.75</td>
<td>7.00</td>
<td>0.20</td>
</tr>
</tbody>
</table>

95% positional uncertainty reported in the 2FGL catalog, performed above 100 MeV.

To test the reliability of the significance $S_{\text{cls}}$ to reject spurious sources, we plot the $D_{\text{fake}}$ and $D_{\text{eff}}$ based on the 2FGL$_{TS\geq16}$ catalog in Fig. 13. Panels a and b correspond to no selection on $S_{\text{cls}}$. Both the $D_{\text{fake}}$ and the $D_{\text{eff}}$ trends are very similar to the case of the simulated sky. If we apply a significance cut of $S_{\text{cls}} > 2$ (panels c,d), we observe that the spurious ratio is $D_{\text{fake}} \leq 0.05$ for almost half of the parameter space (region to the right of the purple line). The more severe cut of $S_{\text{cls}} > 4.0$ (panels d and e), removes all fake clusters for $\epsilon \leq 0.20$ deg, except two for $\epsilon \leq 0.15$ deg. Only for $\epsilon \geq 0.25$ deg, the $D_{\text{fake}}$ ratio shows a significant increase, ranging from 0.05 up to $\approx 0.1$. 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 an additional confirmation about the robustness of our significance we plot in the right panel of Fig. 14, $S_{\text{cls}}$ vs. $\sqrt{T S}$. For each source in our 2FGL$_{TS\geq16}$ 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 $S_{\text{cls}}$ and its standard deviation (represented by the error bar). The average value of $S_{\text{cls}}$ and its standard deviation are evaluated from the list of all clusters 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 ($r = 0.98$, 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 $S_{\text{cls}} \approx 0.5 \sqrt{T S}$.

8. Conclusions

For the first time, we have used the DBSCAN to detect sources in $\gamma$-ray astrophysical images. We implemented a new version of the DBSCAN, the $\gamma$-ray DBSCAN, which 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, which allows an excellent reconstruction of the cluster with an effective background rejection. We 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 below.

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

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

- The error on the position of the reconstructed cluster centroid, performed with a threshold of 3 GeV, is on the same order as 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 one to safely remove spurious clusters.
- The detection efficiency with real data is excellent, we are able to find all 35 sources with $\sqrt{T S} > 4$. 

Fig. 12. Scatter plot of the positional error of the $\gamma$-ray DBSCAN clusters vs. the positional error of the corresponding associated 2FGL$_{TS\geq16}$ sources. For each 2FGL$_{TS\geq16}$ source associated to one or more $\gamma$-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 $\approx 0.99$ and an intercept of $\approx 9.53$. 

A138, page 12 of 14
Fig. 13. $D_{fake}$ (left panels) and $D_{eff}$ (right panels) for the real sky detections using the 2FGL-16 catalog. Panels a), b) no cut on $S_{cls}$ applied. Panels c), d) $S_{cls} > 2.0$. Panels e), f) $S_{cls} > 4.0$. 
When working with ε on 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 γ-ray DBSCAN is a very powerful detection method to find clusters in γ-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 one to apply a statistically significant selection, making the rejection of spurious sources even more effective.

In a companion paper (Tramacere, in prep.), we will apply the method to the Fermi-LAT sky, showing the potential to discover new sources, in particular small clusters located at high galactic latitude, or clusters on the Galactic plane that are affected by a strong background. We will also investigate how to include the energy dependence of the PSF into the γ-ray DBSCAN algorithm, and how to improve the detection performance taking into account other Fermi-LAT calibration properties.

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

Acknowledgements. We are grateful to E. Massaro, R. Campana and E. Bernieri for helpful comments and for providing us the simulated test fields. We are grateful to G. Tosti for helpful comments. We thank the anonymous referee for providing us with constructive comments and useful suggestions.

References
Campana, R., Massaro, E., Behroozi, E., Tinebra, F., & Tosti, G. 2012, submitted
Jolliffe, I. T. 1986, Principal component analysis
Papoulis, A. 1965, Probability, Random Variables and Stochastic Processes