The Astronomical Journal, 165:66 (15pp), 2023 February https://doi.org/10.3847/1538-3881/aca478

Identifying Diffuse Spatial Structures in High-energy Photon Lists

Minjie Fan[1], Jue Wang[1], Vinay L. Kashyap[2], Thomas C. M. Lee[1], David A. van Dyk[3], and Andreas Zezas[4]
[1] Department of Statistics, University of California, Davis, One Shields Avenue, Davis, CA 95616, USA; tcmlee@ucdavis.edu
[2] Center for Astrophysics | Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138, USA
[3] Statistics Section, Department of Mathematics, Imperial College London, 180 Queen's Gate, London, SW7 2AZ, UK
[4] Physics Department, deUniversity of Crete, P.O. Box 2208, GF-710 03, Heraklion, Crete, Greece
Received 2022 August 15; revised 2022 November 1; accepted 2022 November 17; published 2023 January 23

Abstract

Data from high-energy observations are usually obtained as lists of photon events. A common analysis task for such data is to identify whether diffuse emission exists, and to estimate its surface brightness, even in the presence of point sources that may be superposed. We have developed a novel nonparametric event list segmentation algorithm to divide up the field of view into distinct emission components. We use photon location data directly, without binning them into an image. We first construct a graph from the Voronoi tessellation of the observed photon locations and then grow segments using a new adaptation of seeded region growing that we call *Seeded Region Growing on Graph*, after which the overall method is named `SRGonG`. Starting with a set of seed locations, this results in an oversegmented data set, which `SRGonG` then coalesces using a greedy algorithm where adjacent segments are merged to minimize a model comparison statistic; we use the Bayesian Information Criterion. Using `SRGonG` we are able to identify point-like and diffuse extended sources in the data with equal facility. We validate `SRGonG` using simulations, demonstrating that it is capable of discerning irregularly shaped low-surface-brightness emission structures as well as point-like sources with strengths comparable to that seen in typical X-ray data. We demonstrate `SRGonG`'s use on the Chandra data of the Antennae galaxies and show that it segments the complex structures appropriately.

Unified Astronomy Thesaurus concepts: Spatial point processes (1915); Astrostatistics techniques (1886); Voronoi tessellation (1952); Astronomy data modeling (1859)

1. Introduction

A challenge often encountered in high-energy astronomical analysis is that the images are photon starved and sparse, and contain many "empty" pixels. Unlike photon-rich images encountered at longer wavelengths, complex features in X-ray and γ-ray data are difficult to recognize, characterize, and analyze. Working directly with Poisson-distributed photon counts, while simultaneously separating out the contribution of the background, is a difficult process, especially when trying to detect faint nonuniform emission, or separating faint point sources from larger-scale diffuse emission. Finding the boundaries of extended structures is thus a challenging problem. Such complex structures are common in high-energy astronomical images and include, for example, shock fronts, knots in supernova remnants, regions of diffuse emission in galaxies, point sources embedded in diffuse emission or conglomerates of point sources, entire galaxies or groups/clusters of galaxies, jets, or star-forming regions, which appear to be extended in the X-ray band even with intermediate-resolution ($\lesssim 0\farcs5$) X-ray telescopes.

The analysis of extended X-ray sources is critical for several areas of astrophysics. The spatial scales of extended emission contain information regarding the physical processes that lead to their formation, while their boundaries are often determined by their physical environments. Therefore, identifying the boundary of these regions in a data-driven, rather than a model-driven, fashion is necessary for the scientifically valuable results. In such cases, a primary goal of the researcher is to segment the image into regions with similar properties and to analyze each segment individually.

Multiscale methods like wavelets (Starck et al. 2002) for point-source detection have been efficiently implemented for X-ray images (Freeman et al. 2002), and matched-filter techniques have been successfully used to detect galaxy clusters in ROSAT data (Vikhlinin et al. 1998), but extended structures remain difficult to find and characterize in these low-count Poisson data. Other techniques are generally optimized for high signal-to-noise ratio (S/N) images: they apply adaptive binning, or set S/N thresholds to smoothed images with point sources removed (e.g., Sanders & Fabian 2001; Sanders 2006); adapt methods developed for the analysis of cosmic microwave background images (e.g., Bobin et al. 2016); limit themselves to restrictive assumptions like modeling a combination of point sources ((E)BASCS; Jones et al. 2015; Meyer et al. 2021), or require spectral model similarity across the field of view (Picquenot et al. 2019, 2021). Currently, most astronomical images with complex structures that are processed for public display use some form of flux-nonconserving adaptive smoothing (Ebeling et al. 2006). This approach is inadequate for scientific analysis. Previous efforts at extended-source detection using Voronoi tessellation techniques have been limited by computational cost and the imposition of global thresholding schemes (e.g., `vtpdetect`, Ebeling & Wiedenmann 1993). Methods akin to seeded region growing (see `SrcExtractor`, Bertin & Arnouts 1996; `NoiseChisel`, Akhlaghi & Ichikawa 2015) and machine learning (e.g., `Morpheus`, Hausen & Robertson 2020; `Mask R-CNN`, Farias et al. 2020; `galmask`, Gondhalekar et al. 2022) have been used for the identification of features in optical

Table 1
Glossary of Variables and Notation

Notation	Description
$\overset{\text{iid}}{\sim}$	Independent and identically distributed
$\hat{\zeta}$	Estimate of a generic parameter ζ
$\mathcal{F}$	Field of view, a bounded domain in $\mathbb{R}^2$ that contains the observed photons
n	Number of observed photons
$\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$	Observed location of the n photons, denoting their (sky) coordinates
K	Number of segments within $\mathcal{F}$
$\boldsymbol{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_K\}$	Partition of $\mathcal{F}$, where $\mathcal{S}_k$ is the domain for segmented region k
$\text{Num}(\mathcal{S}_k)$	Number of photons observed in segment $\mathcal{S}_k$
$\text{Area}(\mathcal{S}_k)$	Area of segment $\mathcal{S}_k$
m_{seg}	Number of free parameters per segment
$\lambda(\boldsymbol{x})$	Poisson intensity at location $\boldsymbol{x}$
$\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_K\}$	Collection of the Poisson intensities over each of the segments
$\boldsymbol{\Lambda}$	Total integrated intensity of $\lambda(\boldsymbol{x})$ over $\mathcal{F}$
$\mathcal{V}_i$	Voronoi cell defined by photon i
$\text{Area}(\mathcal{V}_i)$	Area of Voronoi cell $\mathcal{V}_i$
$\widehat{\boldsymbol{\lambda}}_i^{\mathcal{V}_i}$	Voronoi estimator of the Poisson intensity across Voronoi cell i, with $\widehat{\boldsymbol{\lambda}}_i^{\mathcal{V}_i} = 1/\text{Area}(\mathcal{V}_i)$
$f(\boldsymbol{X}, n)$	Joint probability mass/density function of the observed number of photons and their locations
$\mathcal{L}(K, \boldsymbol{S}, \boldsymbol{\lambda} \mid \boldsymbol{X})$	Log-likelihood of the model[a]
$\mathcal{L}_{\text{profile}}(K, \boldsymbol{S} \mid \boldsymbol{X})$	Profile log-likelihood[a] obtained by replacing $\boldsymbol{\lambda}$ in $\mathcal{L}(K, \boldsymbol{S}, \boldsymbol{\lambda} \mid \boldsymbol{x})$ with its estimate $\hat{\boldsymbol{\lambda}}$
$\text{BIC}(K)$	Bayesian Information Criterion as a function of the number of segments, K
m_{grid}	In seed specification, the number of grid points used in a regularly spaced grid
m_{graph}	Number of photons included in initial subgraph for each seed
m_{nn}	Number of nearest neighbors over which local maxima are searched for to specify seeds
m_{R}	Number of strata used during Voronoi-area-stratified sampling to specify seeds
m_{vorthr}	Minimum number of photons required for a Voronoi-area-derived seed to be accepted

Note.
[a] Notation is reversed, by convention, for $\mathcal{L}(b|a)$ compared to conditional probabilities; e.g., $p(a|b)$ represents the probability of a given b.

images of galaxies. However, the Poisson nature and the sparsity of the X-ray data requires statistically better targeted methods.

Here we develop a new method that combines aspects of Voronoi tessellation with region growing by using neighbor similarity clustering. The method can be applied to X-ray data and provides both separation between different structures in a complex image and well-defined apertures to perform photometry. We describe the statistical model that underlies the method in Section 2 and the specific implementation details including the computational methods in Section 3. We carry out several simulations to test the limits of applicability of the algorithm in Section 4 and apply it to Chandra data of the Antennae galaxies in Section 5. We discuss how and when the algorithm may be best used in Section 6 and summarize our work in Section 7.

2. Statistical Methodology

We have developed a method that iteratively aggregates contiguous sets of photons into distinct regions based on similarity of their surface brightness. We employ a likelihood-based method to obtain a piecewise constant estimate of the surface brightness across the image; the likelihood function is derived in Section 2.1. The method starts with the high-resolution segmentation of the spatial distribution of the events based on the Voronoi cells described in Section 2.2 and combines segments by optimizing the Bayesian Information Criterion (BIC) given in Section 2.3. Table 1 provides a glossary of our notation.

2.1. Statistical Model

We consider an event list composed of n photons observed in a bounded domain that defines the field of view, $\mathcal{F} \subset \mathbb{R}^2$. Ignoring instrumental pixelization, we model the set of sky coordinates for the n photons,

$$\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}, \tag{1}$$

via an inhomogeneous Poisson process with intensity function $\lambda(\boldsymbol{x}) \geqslant 0$. The intensity function must be integrable over $\mathcal{F}$, i.e., $\boldsymbol{\Lambda} = \int_{\mathcal{F}} \lambda(\boldsymbol{x}) d\boldsymbol{x}$ must be finite. For simplicity, we assume that the intensity function is piecewise constant. In particular, we assume we can partition $\mathcal{F}$ into K segments, denoted as $\boldsymbol{S} = (\mathcal{S}_1, \ldots, \mathcal{S}_K)$, where the Poisson intensity is constant on each $\mathcal{S}_k$. As $\boldsymbol{S}$ partitions $\mathcal{F}$, the $\mathcal{S}_k$ together cover $\mathcal{F}$ and each pair is disjoint. For a given set of nonnegative intensities, $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_K\}$, we can then express the intensity function as

$$\lambda(\boldsymbol{x}) = \sum_{k=1}^{K} \lambda_k 1_{\mathcal{S}_k}(\boldsymbol{x}), \tag{2}$$

where $1_{\mathcal{S}_k}(\boldsymbol{x})$ is an indicator function that takes a value of 1 if $\boldsymbol{x} \in \mathcal{S}_k$ and is otherwise 0.

A property of the inhomogeneous Poisson process is that the number of photons, $\text{Num}(\mathcal{S}_k)$, recorded in segment $\mathcal{S}_k$ with area $\text{Area}(\mathcal{S}_k)$ follows a Poisson distribution with mean

$$\text{Area}(\mathcal{S}_k) \cdot \lambda_k = \int_{\mathcal{S}_k} \lambda(\boldsymbol{x}) d\boldsymbol{x}, \quad \text{for } k = 1, \ldots, K, \tag{3}$$

with $\lambda_k \geqslant 0$. Likewise, the total photon count is distributed as $n \sim \text{Pois}(\mathbf{\Lambda})$. Another property is that, given n, the sky coordinates, $\boldsymbol{x}_i$ are independent and identically distributed (iid) with (normalized) probability density function $\lambda(\boldsymbol{x})/\mathbf{\Lambda}$ (e.g., Chiu et al. 2013). This means that $\boldsymbol{x}_i$ are distinct—no two photons can have the same recorded coordinates. (The discrete nature of detectors means that occasionally, two photons are recorded with identical coordinates. In this case, we add a very small random scatter, $\sim 10^{-6}$.) Our goal is to estimate the number of segments, K, the segments, $\boldsymbol{S}$, and their respective intensities, $\boldsymbol{\lambda}$.

Thus far, we have not discussed the sources or background. If the field of view includes multiple separated sources, we expect the piecewise constant intensity function to capture the intensity peaks associated with point and extended sources. Between these sources (or around a single source) is the background region. If the background intensity is constant across $\mathcal{F}$ and the source region(s) is/are isolated within the field of view, we expect a single large segment representing the background to encircle the source regions and to extend to the boundary of $\mathcal{F}$. If the background intensity varies slowly, we might find several large segments that together comprise the background region. In any case, there are segments associated with background and with sources. We do not attempt to classify the segments in this regard. Of course, if there is a single large low-intensity segment encircling smaller higher-intensity segments, it is easy enough to identify the background with the large low-intensity segment.

We are particularly interested in the case where small-scale point-like sources lie within a larger extended source as this is a challenging task for existing methods. We do not distinguish between extended or point sources, and in fact ignore the effect of telescope's point-spread function (PSF), assuming that it is small compared to the size of the $\mathcal{F}$. Our method is agnostic by design to the sizes of individual structures and is thus capable of isolating sources at all scales.

To derive the likelihood function, recall $n \sim \text{Pois}(\mathbf{\Lambda})$ and given n the $\boldsymbol{x}_i \overset{\text{iid}}{\sim} \lambda(\boldsymbol{x})/\mathbf{\Lambda}$. Thus, their joint probability mass/density function under the inhomogeneous Poisson process is

$$\begin{aligned} f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, n) &= f(\boldsymbol{X}|n) \cdot f(n) \\ &= \frac{1}{\mathbf{\Lambda}^n} \prod_{i=1}^{n} \lambda(\boldsymbol{x}_i) \cdot \frac{\exp(-\mathbf{\Lambda})\mathbf{\Lambda}^n}{n!} \\ &= \frac{\exp(-\mathbf{\Lambda})}{n!} \prod_{i=1}^{n} \lambda(\boldsymbol{x}_i) \end{aligned} \tag{4}$$

and their log-likelihood function is given by

$$\begin{aligned} &\mathcal{L}(K, \boldsymbol{S}, \boldsymbol{\lambda}|\boldsymbol{X}, n) = \log f(\boldsymbol{X}, n) \\ &= \sum_{i=1}^{n} \log \lambda(\boldsymbol{x}_i) - \int_{\mathcal{F}} \lambda(\boldsymbol{x}) d\boldsymbol{x} - \log n!, \end{aligned} \tag{5}$$

where we write out $\mathbf{\Lambda} = \int_{\mathcal{F}} \lambda(\boldsymbol{x}) d\boldsymbol{x}$. Replacing $\lambda(\boldsymbol{x})$ by the piecewise constant expression given in Equation (2), we have

$$\begin{aligned} &\mathcal{L}(K, \boldsymbol{S}, \boldsymbol{\lambda}|\boldsymbol{X}, n) \\ &= \sum_{\substack{k=1 \\ \text{Num}(\mathcal{S}_k)\neq 0}}^{K} \text{Num}(\mathcal{S}_k) \log \lambda_k - \sum_{k=1}^{K} \text{Area}(\mathcal{S}_k)\lambda_k - \log n!. \end{aligned} \tag{6}$$

Recall that $\text{Num}(\mathcal{S}_k)$ and $\text{Area}(\mathcal{S}_k)$ denote the number of photons in $\mathcal{S}_k$ and the area of $\mathcal{S}_k$, respectively. (When $\text{Num}(\mathcal{S}_k) = 0$, the summand

$$\text{Num}(\mathcal{S}_k) \log \lambda_k$$

is excluded from the first sum in Equation (6).)

We aim to maximize $\mathcal{L}$ as a function of K, $\boldsymbol{S}$, and $\boldsymbol{\lambda}$ to obtain their maximum likelihood estimates. For fixed K and $\boldsymbol{S}$, $\mathcal{L}$ is maximized as a function of $\boldsymbol{\lambda}$ by

$$\widehat{\lambda}_k = \text{Num}(\mathcal{S}_k)/\text{Area}(\mathcal{S}_k) \text{ for } k = 1, \ldots, K. \tag{7}$$

Plugging $\widehat{\lambda}_k$ into $\mathcal{L}$, we obtain the profile log-likelihood of K and $\boldsymbol{S}$, i.e.,

$$\begin{aligned} &\mathcal{L}_{\text{profile}}(K, \boldsymbol{S}|\boldsymbol{X}, n) \\ &= \sum_{\substack{k=1 \\ \text{Num}(\mathcal{S}_k)\neq 0}}^{K} \text{Num}(\mathcal{S}_k) \log\left(\frac{\text{Num}(\mathcal{S}_k)}{\text{Area}(\mathcal{S}_k)}\right) - n - \log n!. \end{aligned} \tag{8}$$

The same profile log-likelihood can be derived by modeling the data as a mixture of uniform distributions. Allard & Fraley (1997) considered a special case with a uniform background with a contiguous extended source superposed.

To estimate $\boldsymbol{S}$, we first deploy a (greedy) algorithm that finds an optimal segmentation, $\widehat{\boldsymbol{S}}(K) = \arg\max \mathcal{L}_{\text{profile}}(K, \boldsymbol{S})$, for fixed K, as described in Sections 2.2 and refined in Section 3. In Section 2.3, we introduce a penalized version of $\mathcal{L}_{\text{profile}}$ that we maximize over K to obtain final estimates of the number of segments, $\widehat{K}$, and thereby of the segments themselves, $\widehat{\boldsymbol{S}}(\widehat{K})$.

2.2. Estimating $\boldsymbol{S}$ via Voronoi Tessellation

Obtaining estimates of the segments, $\boldsymbol{S}$, requires us to constrain the set of possible partitions. For any fixed K, for example, we can make $\mathcal{L}_{\text{profile}}$ arbitrarily large by including a segment that is small enough to contain exactly one photon and shrinking the segment's area toward zero (as $\text{Area}(\mathcal{S}_k)$ appears in the denominator of Equation (8)). Similarly, any $\mathcal{S}_k$ with $\text{Num}(\mathcal{S}_k) = 0$ can have arbitrary shape as it does not contribute to the profile likelihood. As we cannot estimate the intensity function at a higher resolution than that of the data, we only consider candidate segments that include at least one photon.

One way to constrain $\boldsymbol{S}$ is to only consider candidate segments that consist of the Voronoi cells derived from the Voronoi tessellation of the data, or the union of several Voronoi cells. The Voronoi tessellation of the observed photons uniquely partitions $\mathcal{F}$ into n convex cells, denoted as $\mathcal{V}_i$, $i = 1, \ldots, n$, such that cell $\mathcal{V}_i$ contains exactly one photon, say $\boldsymbol{x}_i$, and consists of all locations in $\mathcal{F}$ closer to photon $\boldsymbol{x}_i$ than to any other photon. These cells are called Voronoi cells, and $\boldsymbol{x}_i$ is called the nucleus of $\mathcal{V}_i$. Figure 1 gives an example of the Voronoi tessellation of 50 photons drawn from a normal distribution truncated to the unit square. The photon locations are plotted in the left panel and the Voronoi cells in the middle panel. (We discuss the graph in the right panel in Section 3.1.)

To avoid unclosed Voronoi cells near the border of the field of view, we restrict the tessellation to Voronoi cells whose vertices are all in $\mathcal{F}$. Based on the Voronoi tessellation, Barr & Schoenberg (2010) introduced the Voronoi estimator $\widehat{\boldsymbol{\lambda}}_i^{\mathcal{V}_i}(\boldsymbol{x}) = 1/\text{Area}(\mathcal{V}_i)$ for any location $\boldsymbol{x} \in \mathcal{V}_i$. They show that under certain conditions, the Voronoi estimator is approximately unbiased for the Poisson intensity $\lambda(\boldsymbol{x})$, and its

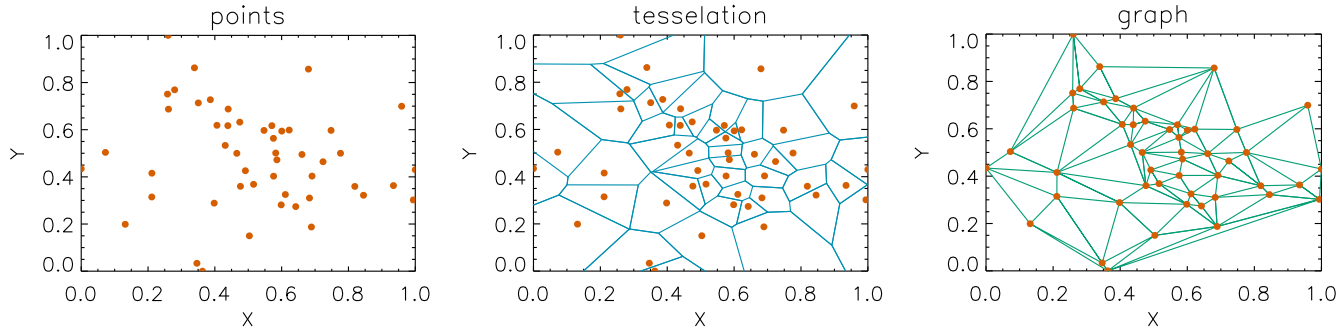

Figure 1. Illustration of tessellation and triangulation of points. Left panel: 50 points drawn randomly from a multivariate normal $N(0.5,\ 0.2^2\mathcal{I})$ are plotted as red dots, where $\mathcal{I}$ is a unit matrix. Middle panel: The Voronoi cells generated by these points are shown as cyan polygons surrounding each point. Right panel: The corresponding graph of Delaunay triangulation is shown as green line segments connecting adjacent points.

sampling distribution is approximately the inverse Gamma distribution.[5]

The algorithm that we propose to combine the Voronoi cells to form the segments (by approximately maximizing $\mathcal{L}_{\text{profile}}$ for each fixed K) is detailed in Section 3. When K is fixed in advance, this algorithm can be used to estimate $\boldsymbol{S}$. To fit K we use the method in Section 2.3.

2.3. Estimating K via the Bayesian Information Criterion

Unfortunately, the number of sources, K, cannot be reasonably estimated by maximizing the profile likelihood, because the likelihood in Equation (8) increases with K and is thus maximized by $K = n$, i.e., with the full set of Voronoi cells.[6] We avoid such overfitting by adding a term to Equation (8) that suitably penalizes model complexity. In particular, we use the so-called BIC, which has been shown to produce statistically consistent results for many model selection problems. For the current problem, the BIC is defined as

$$\text{BIC}(K) = -2\mathcal{L}_{\text{profile}}(K, \widehat{\boldsymbol{S}}(K)|\boldsymbol{X}, n) + Km_{\text{seg}} \log n, \qquad (9)$$

where m_{seg} is the number of free/independent parameters per segment; thus $m_{\text{seg}}K$ is the total number of free parameters in the model.[7] The BIC estimate for K is given by

$$\widehat{K} = \arg\min \text{BIC}(K). \qquad (10)$$

For fixed K, optimizing the BIC is equivalent to optimizing $\mathcal{L}_{\text{profile}}$; thus this estimate of $\boldsymbol{S}$ is equivalent to that described in Section 2.2.

Unfortunately, because we are using a nonparametric model, m_{seg} is not well-defined. Following Aue & Lee (2011), we set m_{seg} by approximating the model by a parametric one using an assumed specific shape for the segments. When the segments are close to ellipse-shaped, for example, we set $m_{\text{seg}} = 6$ to account for the coordinates of the center, lengths of the two axes, orientation, and intensity. When the segments are close to circular, m_{seg} is reduced to 4. Another possibility is to simply set $m_{\text{seg}} = 1$ and the number of model parameters to the number of segments, which to some extent reflects the overall model complexity (Magnussen et al. 2006), but ignores the shapes of the segments. While these parametric approximations allow us to assign a reasonable value to m_{seg}, the model itself remains nonparametric.

The BIC is closely related to the "fitness function" used in the Bayesian block method (Scargle et al. 2013), where model complexity is penalized via a geometric prior on the number of sources, i.e., $p(K) = P_0\gamma^K$, with γ being a tuning parameter and P_0 a normalization constant. Setting $\gamma = 1/n^{Km_{\text{seg}}}$ makes the fitness function equivalent to the penalty term in the BIC.

3. Algorithm for Combining Voronoi Cells into Segments

3.1. SRGonG: Seeded Region Growing on Graph

Using the Voronoi cells as building blocks, we start by proposing an algorithm to estimate the segments in $\boldsymbol{S}$ for fixed K. The first step is to identify pairs or groups of Voronoi cells that can potentially be combined. We accomplish this via the dual graph of the Voronoi tessellation, known as the Delaunay triangulation. This graph's vertices are the centers of the Voronoi cells, i.e., the photons, and its edges connect pairs of the adjacent Voronoi cells. The right panel of Figure 1 depicts the graph derived from the Voronoi cells in the middle panel. We assign vertex $\boldsymbol{x}_i$ the value of the Voronoi estimator, denoted by $\widehat{\boldsymbol{\lambda}}_i^{\mathcal{V}_i}$, i.e., an estimate of the intensity in Voronoi cell $\mathcal{V}_i$. Using the graph constructed by the Delaunay triangulation, the problem of estimating the $\mathcal{S}_k$ can be naturally translated to the problem of graph segmentation, i.e., partitioning the graph into subgraphs such that the Voronoi cells therein form a single segment, $\mathcal{S}_k$. Thus, each subgraph/$\mathcal{S}_k$ is formed by a set of vertices/photons connected by a collection of edges in the full graph. As we assume that the intensity function is a piecewise constant, all the vertices in each subgraph should share similar values of $\widehat{\boldsymbol{\lambda}}_i^{\mathcal{V}_i}$.

Unfortunately, finding $\boldsymbol{S}$ to maximize Equation (8) for fixed K remains an intractable combinatorial optimization problem even when confined to combinations of Voronoi cells. A distinct advantage of representing the problem as graph segmentation is that we implicitly impose an additional constraint that each $\mathcal{S}_k$ is a subgraph. In this way, traditional image segmentation[8] methods can be adapted and used to segment the graph. In particular, we

[5] The probability density function of an inverse Gamma distribution is $\frac{b^a}{\Gamma(a)}x^{-a-1}\exp\left(-\frac{b}{x}\right)$, where $x > 0$, a and b are the shape and rate parameters, respectively, and $\Gamma(\cdot)$ denotes the Gamma function.

[6] As n is fixed, maximizing $\mathcal{L}_{\text{profile}}$ is equivalent to maximizing the sum in Equation (8), which can be written as $\sum_{i=1}^{n}\log\hat{\lambda}(\boldsymbol{x}_i)$, where $\hat{\lambda}(\boldsymbol{x}_i)$ is the local optimizer, $\text{Num}(\mathcal{S})/\text{Area}(\mathcal{S})$, for the segment containing $\boldsymbol{x}_i$. Increasing the number of segments allows for better local optimization of local fluctuations and thus increases $\mathcal{L}_{\text{profile}}$. Of course, with too many segments, better fitting of local fluctuations amounts to fitting noise, i.e., overfitting the data.

[7] As the shape of the final segment is determined by the first $K-1$ segments, a more precise formulation of the total number of parameters in the model is $m_{\text{seg}}(K-1)+1$, where the intensity parameter of the final segment is accounted for by the "+1." The difference between this more precise formulation and the one used in Equation (9) is $(1-m_{\text{seg}})n$, which does not depend on any of the unknown parameters and thus does not affect estimation.

[8] Image segmentation is the process of separating an image into a number of regions such that each region is composed of connected pixels with similar characteristics, such as similar pixel values.

propose the seeded region growing (SRG) on graph (SRGonG) method, which is similar to the original SRG method used for images except that the concept of "neighbors" is determined by the edges of the graph instead of neighboring pixels. The original SRG is proposed in Adams & Bischof (1994) and is extended to several variants to deal with more complicated cases in Fan & Lee (2014). SRGonG starts by identifying, either manually or automatically, a set of initial seeds from the graph. Each seed can be a single vertex/photon or a seeding subgraph, i.e., a set of connected vertices/photons.

For the moment, we present a simplified version of SRGonG that requires a perfect set of seeds, i.e., a set with exactly one seed in each $\hat{\mathcal{S}}_k$. Recall that K is fixed; thus initially we assume K seeds. The details of seed specification in more realistic settings are described in Section 3.2 and the full version of SRGonG (which requires an extra step to merge segments and estimate K) is detailed in Section 3.3.

SRGonG grows the seeds into subgraphs by successively adding neighboring vertices to them. In particular, at each iteration, the method selects a pair that consists of a growing subgraph, S, and one of its unassigned neighboring vertices, i, such that

$$\delta(i, S) = |\log \widehat{\boldsymbol{\lambda}}_i^{\mathcal{V}_i} - \log\{\mathrm{Num}(\mathcal{S})/\mathrm{Area}(\mathcal{S})\}| \tag{11}$$

is minimized. This criterion compares the logarithm of the estimated intensities of the subgraph and the neighboring vertex because $\mathcal{L}_{\mathrm{profile}}$, which we aim to optimize, combines the segment-specific intensity estimates on the log scale. The vertex in the pair with the smallest difference is added to the corresponding subgraph. This process finishes when all the vertices of the full graph are assigned to exactly one subgraph. The Voronoi cells contained in the subgraphs give the final segmentation of $\mathcal{F}$, i.e., $\widehat{\boldsymbol{S}}(K)$, for prespecified K.

In practice, we save the index, i, of the neighboring Voronoi cell that minimizes $\delta(i, S)$ for each growing subgraph, $\mathcal{S}_k$, at each iteration. This reduces the time complexity of the method to be linear in terms of the number of photons.

3.2. Seed Specification

As SRGonG begins by building out regions starting from a specified set of seeds, the number and location of the seeds are important considerations. As discussed in Section 3.1, we would ideally have exactly one seed within each $\hat{\mathcal{S}}_k$. Unfortunately, this is not feasible in practice. A brute-force solution is to overspecify the seed set to the extreme, by setting every photon location to be a seed, and devising an algorithm to merge the seeds into segments. However, merging such a large seed set would be challenging in terms of both the computational speed and statistical accuracy (see discussion in Section 6.1). If the field being analyzed is known to have a large number of point sources, or if the scientific question requires focusing on point sources, then running a source detection algorithm first to find all such sources and specifying all of them as seeds will be helpful. Here we describe three generic strategies to specify smaller initial seed sets. These strategies still overspecify the set in that they use a larger number of seeds than the expected number of segments (but less so than setting each photon to be a seed). Thus, after growing the seeds into subgraphs as described in Section 3.1, we require a method to merge the resulting subgraphs into segments; we describe our merging algorithm in Section 3.3.

Regular grid: This method starts by overlaying a regular grid of m_{grid} points onto the field of view, $\mathcal{F}$. For each grid point, we specify a seeding subgraph composed of the m_{graph} photons closest to the grid point (in terms of the Euclidean distance). We typically set m_{grid} and m_{graph} so that their product is much smaller than n to enable the seeded regions to grow. Conflicts in the allocation of photons to seeding subgraphs (e.g., when a single photon is among the m_{graph} closest to two or more grid points) are broken by the order of assignment. The number, m_{graph}, of the photons assigned to each seeding subgraph can be increased to stabilize the initial estimates of the growing subgraph intensities, especially when the contrast (i.e., the ratio between the intensities of an extended source and the background) is low. In practice, there is no universally best choice for m_{grid} and m_{graph}, as their optimal values depend on factors including the number of observed photons n and the complexity and number of true astronomical sources. To ensure a sufficient number of photons for the seeding subgraphs, we require

$$m_{\mathrm{graph}} \leqslant \frac{n}{m_{\mathrm{grid}}}.$$

As it is possible for a seed to fall on the boundary between two distinguishable segments and adversely affect subsequent processing, we propose an additional *seed-rejection* step. In particular, we compare the range of the Voronoi areas for each photon $\mathrm{Area}(\mathcal{V}_k)|_{\{k=1,\ldots,m_{\mathrm{graph}}\}}$ of a seeding subgraph of size m_{graph} with the expected empirical 2σ confidence interval for a homogeneous distribution of photons (Møller 1994, Chapter 4.2), i.e.,

$$\frac{1}{\widehat{\boldsymbol{\lambda}}_s} \pm 2 \times \frac{0.53}{\widehat{\boldsymbol{\lambda}}_s} \tag{12}$$

(Møller 1994, Chapter 4.2), where

$$\frac{1}{\widehat{\boldsymbol{\lambda}}_s} = \frac{1}{m_{\mathrm{graph}}} \sum_{k=1}^{m_{\mathrm{graph}}} \mathrm{Area}(\mathcal{V}_k) \tag{13}$$

is the average of Voronoi areas for the photons in the seeding subgraph. Thus, if the actual range of Voronoi areas $\mathrm{Area}(\mathcal{V}_k)$ exceeds the expected empirical confidence interval, the seeding subgraph is rejected.

Grid supplemented by local maxima: If the regular grid used to generate the seeds is too sparse, some image structures may not be captured in the segmentation. For example, if there is not a grid point sufficiently near a point or extended source, the source may be merged into the background or another source. One remedy is to include additional seeds near the likely locations of sources. Sources induce an elevated intensity over small spatial scales. Thus, locations of high photon density are likely associated with sources. We propose to identify vertices that are local maxima of the graph constructed by the Delaunay triangulation, in the sense that the vertex value (i.e., $\widehat{\lambda}_i^{\mathcal{V}_i}$) is greater than or equal to that of its closest k vertices (including itself), where closeness is measured by the Euclidean distance. For each local maxima we find in this way, we include a seeding subgraph composed of its closest m_{graph} vertices (including itself).

Voronoi-area-stratified sampling: More complex schemes, designed to locate seeds over a broader range of surface

brightness values, can also be devised. Methods such as Otsu's thresholding (Otsu 1979) can also be used to specify seeds or seeding subgraphs for point-like or localized extended sources. As we discuss in Section 6.1, the grid supplemented by the local maxima is adequate to identify structures that exist at a large variety of scales in astronomical data. Here, as an example case, we describe a third method, which selects seeds via stratified sampling of the photons, with strata determined by the areas of the Voronoi cells, Area($\mathcal{V}_i$). In particular, the photons are divided into m_{R} strata bounded by equally spaced quantiles of the distribution of Area($\mathcal{V}_i$). The number of strata depends on the sample size, but we typically use $m_{\mathrm{R}} \approx 10$–$20$. Clumps of near-neighbor photons within each stratum are put together (see the Appendix) into a set of labeled groups such that spatially nearby photons within a given stratum are all assigned the same label. If a given label is assigned to fewer than m_{vorthr} photons (typically $m_{\mathrm{vorthr}} = 5$), then the photons with this label are discarded for the purpose of seed specification; otherwise the central photon[9] among those with each label is set as a seed. Subgraphs are then constructed for each retained seed photon in the same manner as described in Section 3.1.

Algorithm 1. Seeded Region Growing on Graph (SRGonG)

Data: Coordinates of observed photons $\boldsymbol{x}_i = (x_{1i}, x_{2i})$, $i = 1, \ldots, n$ in field of view $\mathcal{F}$.
Result: Piecewise constant estimate of intensity function with a segmentation of $\mathcal{F}$ into regions of constant intensity, $\widehat{\boldsymbol{S}} = (\hat{\mathcal{S}}_0, \ldots, \hat{\mathcal{S}}_{\widehat{K}})$.
begin
1 Use Voronoi tessellation to obtain a graph whose vertices are the observed photons with the Voronoi estimators $\widehat{\lambda}_i^{\mathcal{V}_i}$ as their values.
2 Using a method in Section 3.2, specify the initial seeds for subgraph growing.
3 Grow seeds into subgraphs that oversegment the entire graph:
while *there are unassigned vertices* **do**
Select a pair of a growing subgraph S and one of its neighboring vertices i such that
$\delta(i, S) = |\log \widehat{\lambda}_i^{\mathcal{V}_i} - \log\{\mathrm{Num}(\mathcal{S})/\mathrm{Area}(\mathcal{S})\}|$ is minimized.
Add the vertex in the pair with the smallest difference to the corresponding subgraph.
4 Greedily merge subgraphs by minimizing the BIC at each merger to obtain a nested sequence of segmentations.
5 Finally, set $\widehat{K}$ and $\widehat{\boldsymbol{S}}$ to the values of the nesting level with the smallest BIC. $\widehat{K}$ is the final segmentation of $\mathcal{F}$.

3.3. Subgraph Merging

Using one of the seed sets of Section 3.2 to grow subgraphs as described in Section 3.1 leads to an oversegmented graph as the number of seeds is invariably more than the predetermined K or the $\widehat{K}$ that optimizes the BIC. To merge the subgraphs into segments, we propose a subgraph merging method that aims to minimize the BIC. Similar ideas were used by Lee (2000) and Peng et al. (2011) in image segmentation. In particular, the subgraph merging method starts by computing the BIC for the oversegmented graph and then iteratively selects two neighboring subgraphs, merges them, and recomputes the BIC. The two merged subgraphs are selected so that their merger gives the largest decrease (or the smallest increase) in the BIC among all possible merges (of neighboring subgraphs). In this sense, this is a greedy algorithm. We continue merging the subgraph until all subgraphs are merged into the entire graph, except when K is fixed, in which case we stop when K segments remain. In this way, we obtain a sequence of nested segmentations, $\{\widehat{\boldsymbol{S}}(K), K = 1, \ldots, n\}$, each with a BIC value. Finally, we set $\widehat{K}$ and $\widehat{\boldsymbol{S}}$ to the values of the nested level with the smallest BIC.

At each iteration, we use an updating formula to speed the computation of the BIC. Consider the graph segmentation associated with K and the graph segmentation after merging two of the subgraphs of $\boldsymbol{S}(K)$ and label the merged subgraphs i and j, with $1 \leqslant i < j \leqslant K$. This merger decreases the BIC by

$$\begin{aligned}\Delta\mathrm{BIC}_{K,i,j} &= \mathrm{BIC}(K) - \mathrm{BIC}(K-1)\\ &= 2\,\mathrm{Num}(\mathcal{S}_i)\log\frac{\mathrm{Num}(\mathcal{S}_{i\cup j})\mathrm{Area}(\mathcal{S}_i)}{\mathrm{Area}(\mathcal{S}_{i\cup j})\mathrm{Num}(\mathcal{S}_i)}\\ &\quad + 2\,\mathrm{Num}(\mathcal{S}_j)\log\frac{\mathrm{Num}(\mathcal{S}_{i\cup j})\mathrm{Area}(\mathcal{S}_j)}{\mathrm{Area}(\mathcal{S}_{i\cup j})\mathrm{Num}(\mathcal{S}_j)}\\ &\quad + m_{\mathrm{seg}}\log n, \end{aligned} \tag{14}$$

where $i \cup j$ denotes the union of photons that belong to subgraphs i and j. The complete procedure for SRGonG is summarized in Algorithm 1.

4. Simulation Studies

Our simulation study is conducted assuming a hypothetical instrument that produces fields of view, $\mathcal{F}$, with two-dimensional coordinates on the unit square. The simulations are designed to assess the performance of SRGonG when applied to fields of view of point-like sources embedded in extended sources of different shapes, while varying the exposure time (or equivalently, the overall counts in the field) and the contrast between the different components. In all our simulation settings, "point-like sources" are circular sources of radius 0.025 and extended sources are of area ≈0.2 relative to the $\mathcal{F}$. We consider three "true images": (a) four point-like sources embedded within a circular extended source of radius 0.25 (covering an area 0.196 of the unit square), (b) three point-like sources embedded within a polygonal zigzag-shaped extended source comprising five squares of size 0.2×0.2 (total area of 0.2), and (c) three point-like sources embedded within an arc-shaped extended source (a half-annular shape with an inner radius of 0.2, an outer radius of 0.4, and a total area of 0.189). We consider these three settings because point-like sources embedded within a complex extended source are commonly observed in astrophysical fields of view, as illustrated in Section 5, and the extended sources mimic typical astronomical shapes. Furthermore, the variety of shapes and the contrasts considered are a stringent test of the algorithm. Letting β denote the exposure time (in arbitrary units) and σ denote the contrast level between the different components, for each simulated $\mathcal{F}$ we generated $\beta\sigma$ counts for each point-like source, $10\beta\sigma$ counts for the extended source, and 1000β counts in expectation for the background, with the photons corresponding to each component distributed uniformly over the area allocated to it. In Figure 2, we have adopted $\beta = 1$ and $\sigma = 30$, so in all cases, the point-like sources

[9] Consider the set, L, of photons with a given label. For each photon $l \in L$, we calculate $d_l = \sqrt{\mathrm{Area}(\mathcal{V}_l)} + \sum_{k\in L} d_{lk}$, where d_{lk} is the Euclidean distance between photons l and k. That photon in L with the smallest $\{d_l\}$ is flagged as the central photon among the photons in L. This measure of centrality is better than computing a centroid as it ensures that the seed is guaranteed to be included inside the labeled region even when the region shape is complex, and that the seed is unambiguously assigned to one of the photons.

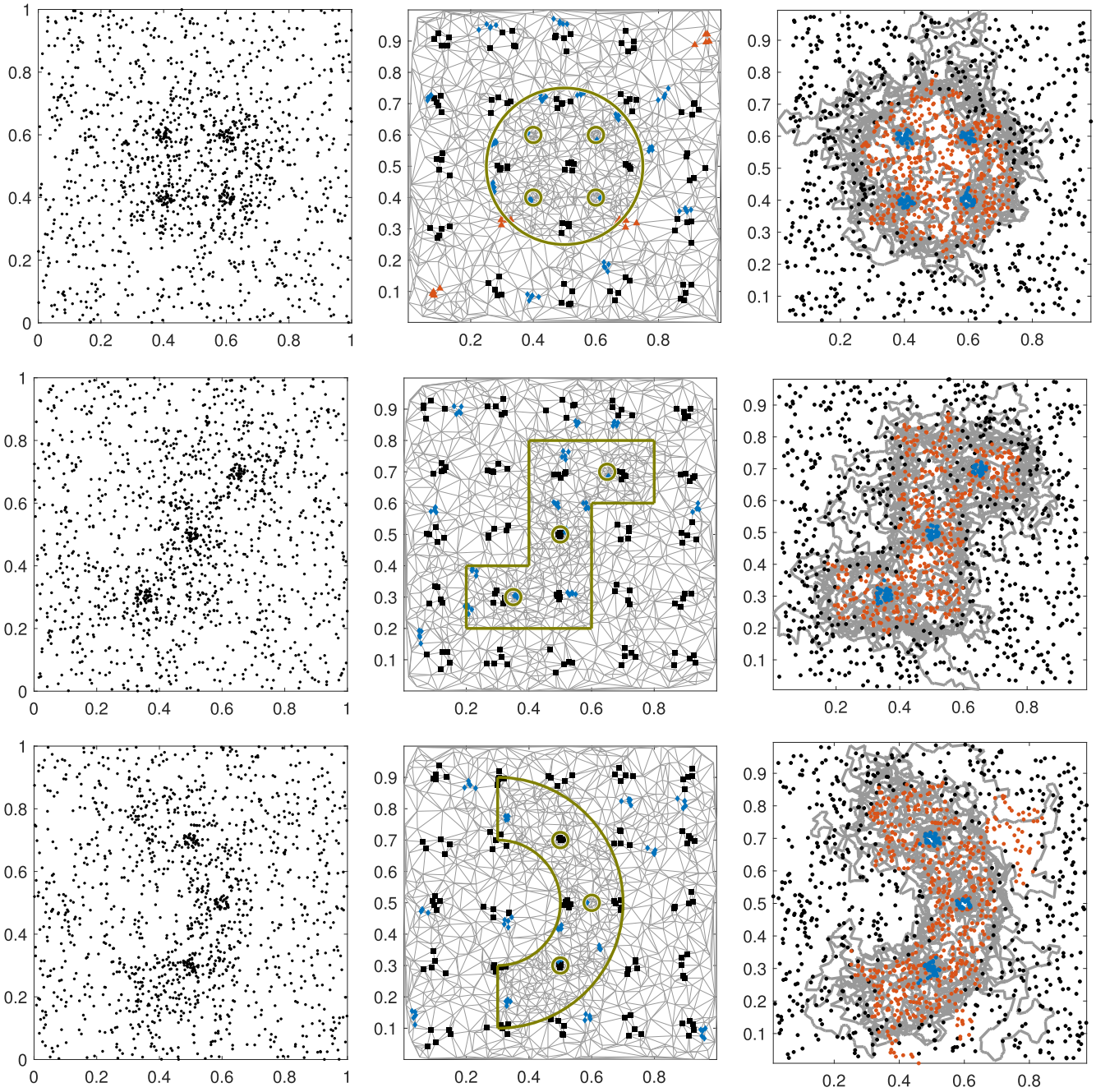

Figure 2. Simulation studies with various shapes of the extended source (with $\beta = 1$ and $\sigma = 30$; see Section 4). Each row corresponds to one of the three image shapes (circular, top row; polygonal zigzag, middle row; semi-annular, bottom row) considered. Left column: simulated photon locations for one simulation instance. Middle column: The graph for the exemplar (first) simulation of the left column is shown (gray lines), and the true segment boundary is overlaid (solid green lines). The initial seeds are marked for the 5×5 regular grid (black squares), local maxima (blue diamonds), and rejected seeds (red triangles). Right column: The segment boundaries from 10 additional simulations (gray solid lines) are overlaid on the photons from the first simulation, which are marked as red for the extended source and blue for the point-like sources.

have 30 photons, the extended shapes have 300 photons, and the background has $\sim$Poisson(1000) photons, all distributed uniformly over their allocated areas, with $\approx$200 background counts under the area of the extended source. The contrast in surface brightness between the extended source and the background is thus $\approx 1.5\times$, which is sufficiently large on the scale of the extended sources that the presence of the extended sources are clearly recognizable. However, it is clear from inspection of Figure 2 that local fluctuations can be sufficiently large as to make estimating the boundary of the extended sources challenging.

The photons randomly generated for each of the settings are shown for one case in the left column of Figure 2. The middle column shows the shapes of the extended sources overlaid on the corresponding Delauney triangulation for each of the photons, as well as the seeds chosen for that case. The right column shows the segmentation, with points colored blue and the extended source colored red, for the simulation in the left column; superposed in gray lines is the result of segmentations from 10 additional simulations. The superpositions of the segment boundary lines over the expected lines of the shapes of both the point-like and the extended sources show that, while

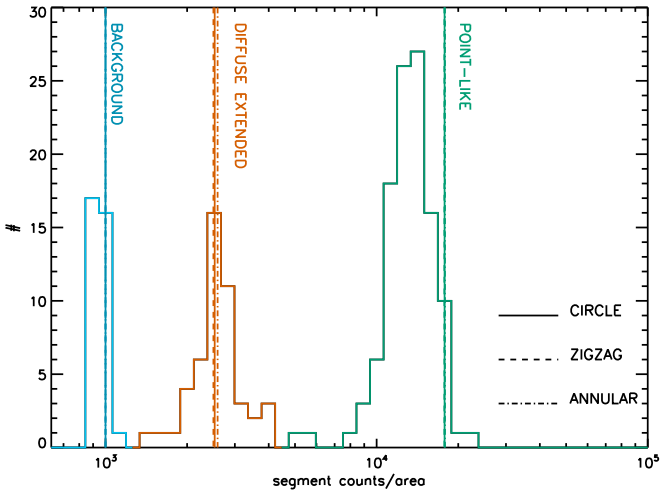

Figure 3. Demonstration of recovery of simulation parameters in segmentation. The brightness of each segment recovered by SRGonG for all of the simulations shown in Figure 2 is shown as the black histogram. The expected brightness for each of the components (background, extended, point-like) and the shapes of the extended sources (circle, polygonal zigzag, and semi-annular) are marked as vertical lines with different line styles, as labeled. The background and extended emission brightnesses are recovered well, but because of the large contrast, the brightness of point-like objects shows a bias (see text).

fluctuations are present in individual simulations, on average the boundaries are picked out well. A detailed examination of the locations of the boundaries and their uncertainties requires modeling the boundaries, and we defer discussion to follow-up work (J. Wang et al. 2022, in preparation). Here, we demonstrate that the components are well recovered in all cases. We show the distribution of the segment brightnesses found for all the simulations for all three cases in Figure 3: the components are clearly separated, with uncertainties of ≈10%–15% on the expected brightness in each component. We find that the brightness of the point-like component suffers from a bias because of the tendency of the segment areas to preferentially encroach on the much larger area of the surrounding extended source, thus causing a downward shift in the estimated brightness.

In our simulation design, in addition to varying the three "true images," we also vary the exposure time with β taking values 0.5, 1, and 2, and the contrast with σ taking values 10, 20, and 30. We simulate 500 fields of view under each of the 27 resulting simulation settings.[10] Each of the 13,500 simulated fields of view is analyzed with SRGonG, with initial seeds specified following the "grid supplemented by local maxima" method of Section 3.2. The regular grid used for seed specification is 5×5, with a seed size of $m_{\rm graph} = 5$, and a neighborhood size of $k = 50$ for finding the local maxima. As the sources we are considering are simple, we set the BIC parameter to be $m_{\rm seg} = 4$; when more complicated shapes are expected, larger values of $m_{\rm seg}$ should be used.

The second and third columns of Figure 2 show the initial seed specifications and segmentation results for the first of the 500 fields of view generated with $\beta = 1$ and $\sigma = 30$. All point-like sources are clearly identified. The fitted boundaries of the extended sources are generally quite good, except for some mild leakage for the arc-shaped extended source. In general, we expect sources with longer perimeters per unit area[11] to be more challenging. This is because photons nearer the boundary of a segment are more likely to be misclassified than are those nearer the middle. Thus, more irregularly shaped sources, such as the arc-shaped source in this simulation, are more challenging, including SRGonG.

Furthermore, several of the initial seeds placed by the regular grid happen to fall near the boundary of the arc-shaped source, which can also jeopardize the performance of seed-based methods.

We use a clustering verification metric, specifically the *adjusted Rand index* (ARI; Hubert & Arabie 1985), to assess the quality of the SRGonG segmentations. The Rand index (RI; Rand 1971) quantifies how well a given segmentation matches the ground truth segmentation. In particular, each pair of photons is classified as either (a) being in the same fitted segment *and* in the same ground truth segment, (b) being in different fitted segments *and* in different ground truth segments, or (c) not being in class (a) or (b). (For the ground truth, the segments are the background, extended source, and each point-like source.) The RI is defined to be the number of photons pairs in class (a) or (b), relative to the total number of photon pairs. Thus, a perfect match to the ground truth results in RI = 1. The ARI corrects the RI such that accidental overlaps of segments due to chance are accounted for, yielding values in the range $-1 < \mathrm{ARI} < +1$.

Figure 4 summarizes the ARI and the fitted value for the number of segments, $\widehat{K}$ for the 500 replicates under each of the 27 simulation setting. For each of the three true images, as expected, the SRGonG segmentation improves as either the exposure time or the contrast between the brightness of the components increases. This is seen in the progression from the top left panel to the bottom right panel of the right column of plots in Figure 4: the method fails to identify the embedded point sources when there are ≈5 counts in each source, but correctly identifies all components in ⩾70% of the cases (80% for the circular and polygonal cases) when there are 60 counts in each point source. Similarly, the ARI increases to close to one (i.e., perfect agreement between ground truth and SRGonG segmentation, where the fitted number of segments equals the true number of sources) as β and σ increase.

5. Application to Antennae Galaxies

The Chandra observations of the Antennae galaxies provide a good test case for the application of SRGonG. The X-ray data (see Figure 5, top left panel) show complex structures. In particular, the data reveal several point sources and extended regions, with several clumps of diffuse emission of different extent and surface brightness, along with a population of unresolved point-like sources superposed. Some of the point sources lie within the extended sources (e.g., in the extended region at the bottom of the image), and some of the extended sources are entangled with each other.

As a conservative scenario we used the first Chandra observation of the Antennae galaxies obtained on 1999 December 1st (OBSID 315; Fabbiano et al. 2001). The

[10] The number of source counts was held fixed in all simulations, while the number of background counts was generated as a Poisson with mean 1000β in order to explore the effect of background fluctuations. Thus, the total counts in a given data set are $\beta\sigma + 10\beta\sigma + \mathrm{Poisson}(1000\beta)$.

[11] A standard measure of shape irregularity is the "perimeter index" of Angel et al. (2010), which is defined to be the perimeter of a circle of area equal to that of the shape divided by the actual perimeter of the shape.

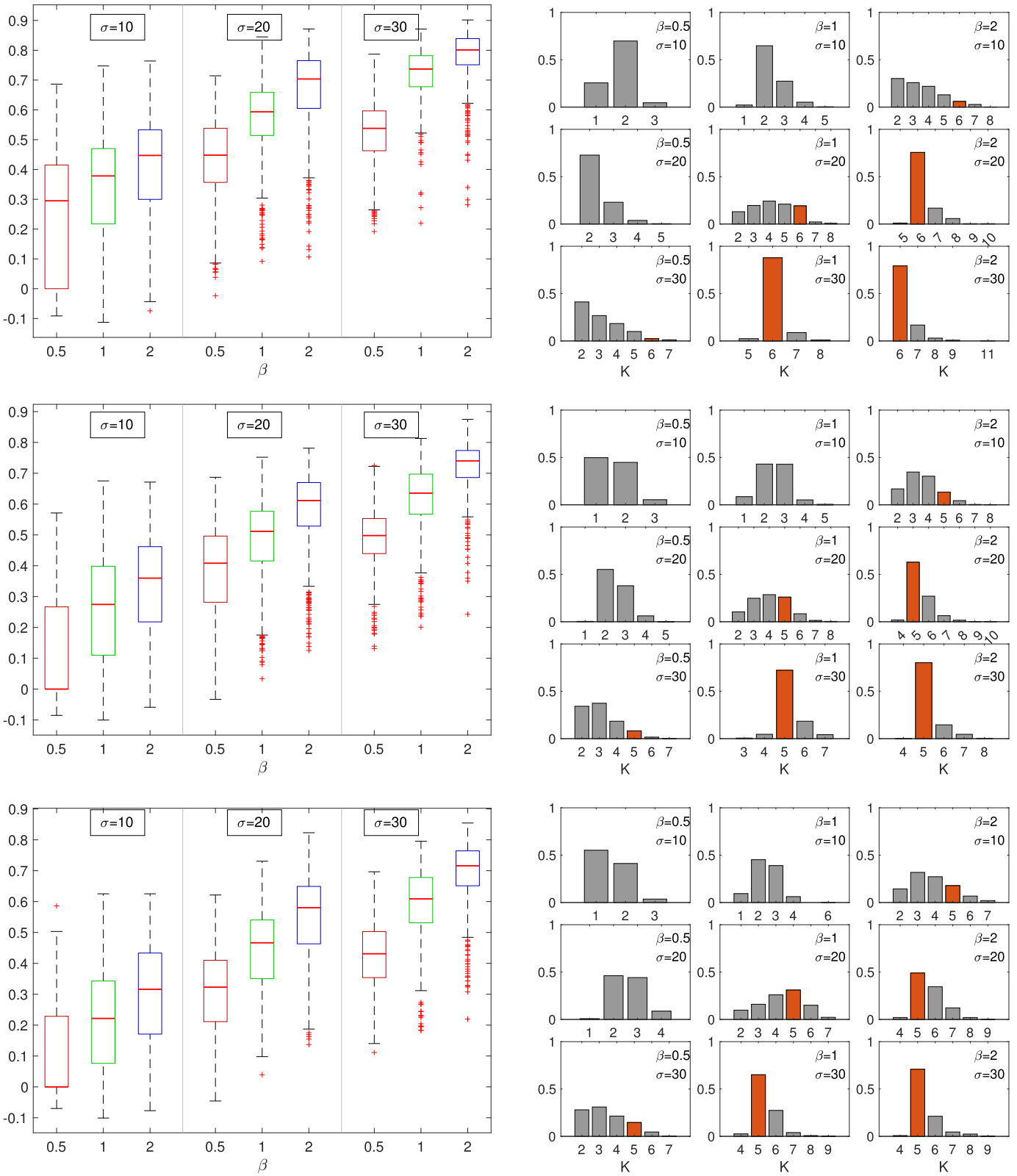

Figure 4. Simulation studies for the various shapes of the extended source, different exposure times, β, and different contrast ratios, σ. Each point source has $\beta \cdot \sigma$ counts in a circle of radius 0.025; each diffuse extended source has $10 \cdot \beta \cdot \sigma$ counts spread uniformly over an area $\approx$0.2; and there are $1000 \cdot \beta$ counts in the background, spread uniformly across $\mathcal{F}$. The three rows correspond to the three extended-source shapes in the rows of Figure 2. Left column: box plots of the adjusted Rand indices for 500 replicates under each simulation setting. Right column: histograms of the fitted number of sources $\widehat{K}$ for the 500 replicates under each simulation setting, where the true number of sources is highlighted in red.

observation was performed with the ACIS-S detector for a total exposure of 72 ks. We process and screen the data (e.g., initial calibrations, removal of strong background flares, selection of good grades) as in Zezas et al. (2022; *CIAO* v3.2, CALDB v2.11). Again as a conservative scenario, we use the full data set without any screening for events of very low or

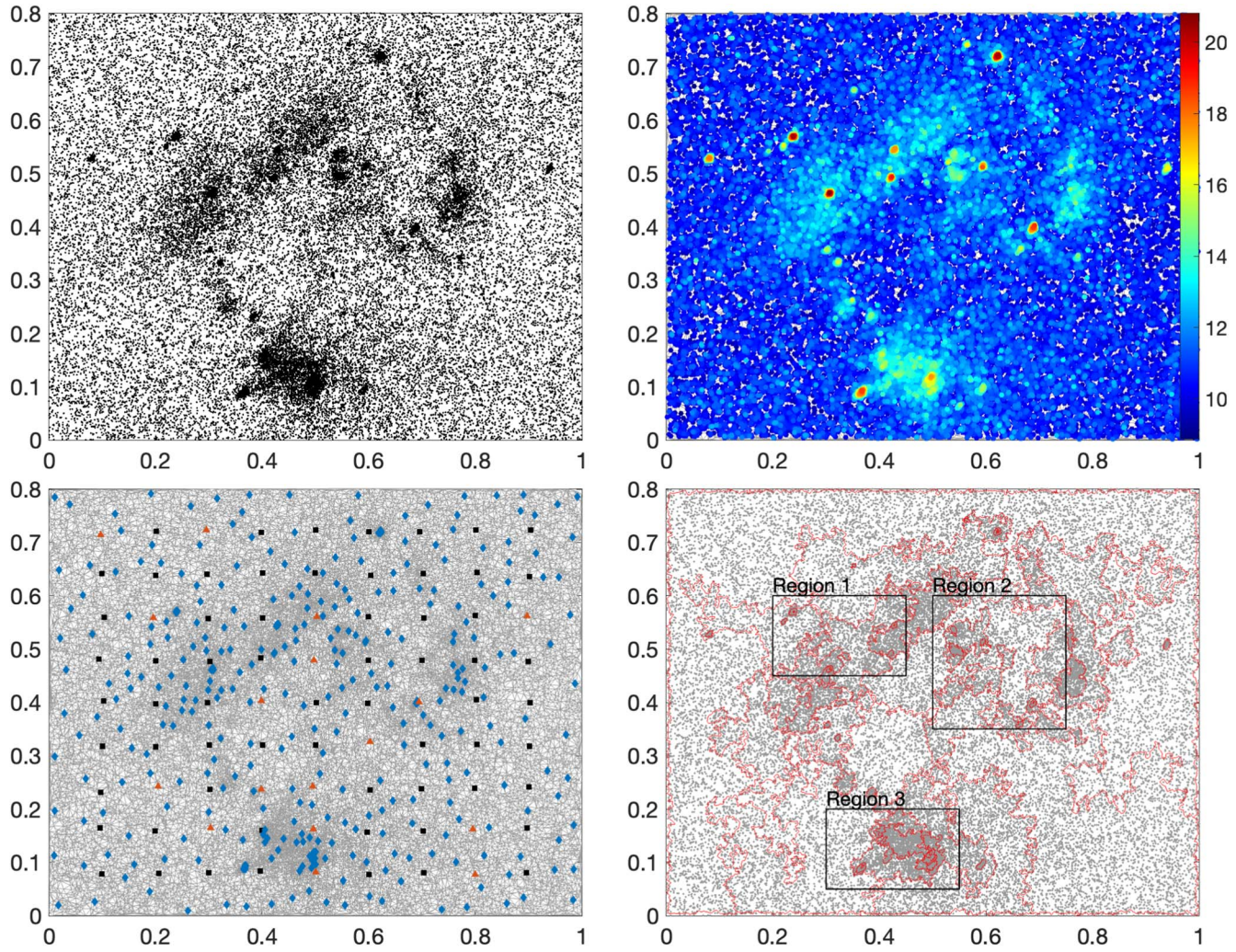

Figure 5. Representations of Antennae data. Top left: scatter plot of the photons from six Chandra observations of the Antennae galaxies carried out between 2001 December and 2002 November. The x-axis is normalized to range from 0 to 1, and the y-axis is normalized accordingly, by the same ratio, such that the data are depicted in the form that is input to SRGonG. Top right: The Delaunay triangulation of the photons is shown, with each vertex marked by a point that is color coded by the brightness determined as the reciprocal of the Voronoi area (see scale on the right). Bottom left: The seeding subgraphs for SRGonG via a regular grid, marked with black squares, supplemented by the local maxima marked with blue diamonds. Seeds discarded due to a strong indication of being on the boundary of two segments are marked as red triangles. Each subgraph is represented by one photon in the middle of the seeding subgraph. Bottom right: The SRGonG segmentation of the Antennae galaxies data, with the red curves depicting the boundaries of the fitted segments. Three smaller fields of view are highlighted by black boxes, labeled Regions 1–3, and magnified to show details in Figure 7.

high energies, which are dominated by the background. The final data set we use consists of ≈50,700 events within a $\sim 3\farcm45 \times 3\farcm45$ region around the galaxy (screening for events in the generally used 0.5–8.0 keV band would result in a reduction of ∼43% in the total number of counts).

Figure 5 shows different depictions of these data, with the coordinates scaled linearly to the range [0, 1], as is assumed by our implementation of SRGonG, and processed to show the resulting Voronoi tessellation.

We apply SRGonG to these data in order to obtain statistically meaningful nonparametric segmentations of the different clumps of diffuse emission, as well as to separate diffuse and point-like emission sources. We apply the Voronoi tessellation to the photons and construct the graph of Delaunay triangulation (see top right panel of Figure 5). We specify the initial seeds for SRGonG via a regular grid supplemented by the local maxima (see Section 3.2; shown in bottom left panel of Figure 5). We start with a regular 9×9 grid (i.e., $m_{\rm grid} = 81$), the initial estimates of which are stabilized by assigning the $m_{\rm graph} = 20$ nearest photons to each seed; these cover the large-scale variations in the data. The local maxima are determined over a neighborhood size of $k = 100$. The 419 seeds that result from this process provide a sufficiently large number to ensure that there is generally at least one seed in each point-like or extended source or the background. As we expect the segments of the extended sources to be more irregularly shaped in the real data than in the simulation, we choose a larger value of the BIC parameter, $m_{\rm seg} = 6$ (see Equation (9)); this corresponds to assuming that each segment has the complexity of an ellipse.

The results of SRGonG are shown in the bottom right panel of Figure 5 (the regions outlined in black are discussed in more detail in Section 6.1), showing the boundaries of the fitted segments as thin red lines around the black dots depicting the photons. SRGonG correctly segments areas with similar surface brightness such that photons that correspond to these diffuse components are grouped together. The photons that belong to each of these segments can be trivially collected together for

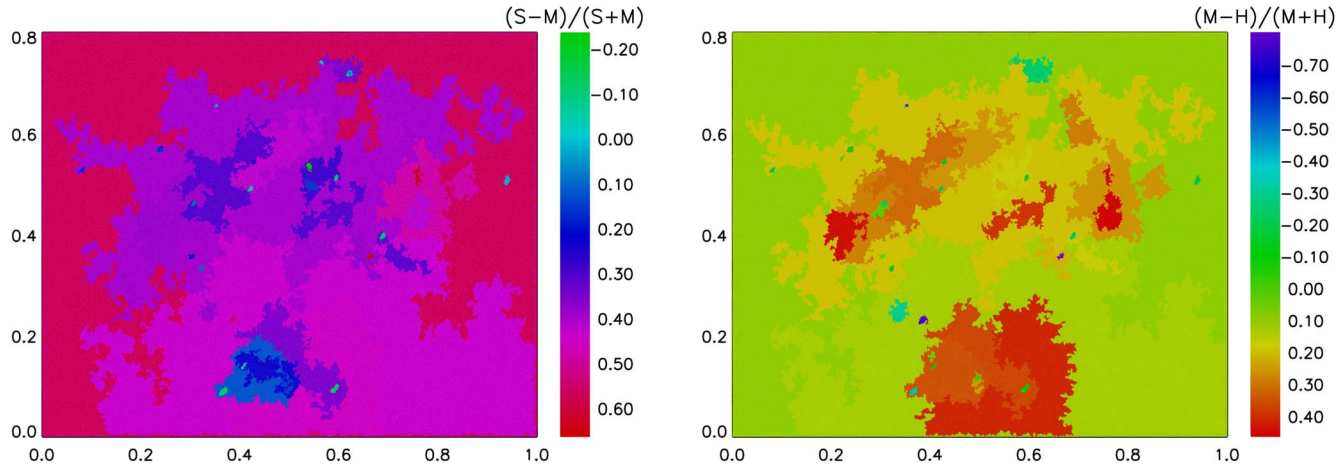

Figure 6. Fractional hardness ratios of counts in each of the SRGonG-determined segments of the Antennae data. The images show $\mathrm{HR}_{SM} = \frac{S-M}{S+M}$ (left) and $\mathrm{HR}_{MH} = \frac{M-H}{M+H}$ (right), where S, M, H are counts in the passbands 0.5–0.9 keV, 0.9–1.2 keV, and 1.2–2 keV, respectively. In both figures, redder colors indicate softer spectra. The background is rendered in reddish pink in HR_{SM} and light green in HR_{MH}.

further analysis, depending on the scientific question being explored. For instance, in Figure 6, we show segment-wise maps of the fractional hardness ratios $\mathrm{HR}_{SM} = (S-M)/(S+M)$ and $\mathrm{HR}_{MH} = (M-H)/(M+H)$, where S, M, H are counts in PI channels [35:61] (≈0.5:0.9 keV), [62:82] (≈0.9:1.2 keV), and [83:135] (≈1.2:2 keV), respectively. Notice that the maps clearly demonstrate that the diffuse emission in the Antennae generally have softer spectra than the point sources. Maps such as these can be used to identify the extent of dust lanes in the Antennae system; e.g., the segments at ∼(0.4, 0.22), ∼(0.3, 0.25), and ∼(0.6, 0.75), which are characterized by harder spectra than the surrounding segments, a characteristic of increased absorption (see Zezas et al. 2006). Furthermore, notice that the southern region (around Region 3 in the bottom left panel of Figure 5) is surrounded by a halo of relatively soft X-ray emission, in agreement with the spectral analysis of Baldi et al. (2006) who find emission from soft ∼0.6 keV thermal emitting gas.

6. Discussion

6.1. Performance on the Antennae Data

Here we discuss the quality of the SRGonG segmentation of the Antennae in greater detail. To begin with, we note that SRGonG successfully identifies a number of point-like sources, characterized by the presence of a large number of photons within a small space. Several of these point-like sources are superposed on extended diffuse emission and surrounded by complex structures. Furthermore, unlike the case usually with methods that use piecewise constant models, the point-like sources are invariably defined by single segments and not several concentric rings that approximate the typical profile of the PSF where intensity increases from the wings inward to rise to a peak at the core. However, such cases are not entirely absent: see Figure 7, specifically the sources at ∼(0.23, 0.57), ∼(0.43, 0.54), and ∼(0.31, 0.46) in Region 1; ∼(0.67, 0.4) in Region 2; and ∼(0.49, 0.12) and ∼(0.37, 0.07) in Region 3.

Further note that the segment boundaries are not smooth because of the boundary being formed by the outermost Voronoi cells. The photons that comprise the boundary are also subject to stochasticity, due both to PSF-induced statistical variations in photon arrival locations, as well as the greedy merging process. Visual inspection of the results suggests that at the lowest surface brightness levels, fluctuations in the counts could result in oversegmentation of what is usually considered to be the background (e.g., the two large extended regions along the left side of the bottom edge of $\mathcal{F}$). Nevertheless, the expanded views of the inset regions in Figure 7 show that the segmentation correctly separates diffuse emission structures at different spatial scales and surface brightness levels. In particular, transitions in the spatial density of photons across the boundaries are clearly discernible by eye, such as those between segments B ↔ C, B ↔ D, D ↔ E, D ↔ F, E ↔ F in Region 1; between segment B and segments A, C, D, F, G, I, J in Region 2; and segment A and segments B, C, D, G, and H as well as C ↔ F and C ↔ D in Region 3. Some transitions are too subtle to be visually recognizable (e.g., A ↔ B in Region 1, B ↔ E and B ↔ H in Region 2, and D ↔ E in Region 3) but are required due to the computed contrasts in the counts per unit area. Conversely, the brightness transitions across B ↔ C ↔D in Region 1, B ↔ C ↔K and B ↔ I ↔L in Region 2, and A ↔ B ↔ C, F ↔ C ↔ D, and D ↔ A ↔ G, H are apt demonstrations of the capability of SRGonG to perform at the level of human visual acuity. Parametric modeling to capture the spatial variations in such structures would be much more difficult than the segmentations achieved here.

An important factor in obtaining a reliable segmentation is the initial seed specification. It is worth establishing that the scheme we propose generates a useful segmentation and does not miss features. For this, we compare the SRGonG method against a brute-force segmentation where every photon in the data set is taken to be a seed, and the corresponding Voronoi cells are merged using the BIC as described in Section 3.3. This brute-force scheme is similar to Scargle's (Scargle 2002) method (but using the BIC instead of Bayes factors) in that it eschews the SRG on graph step developed and described in Section 3.1.

In Figure 8, we compare SRGonG-based segmentation (left panel) against brute-force segmentation (right panel). Although at first glance the two segmentations look similar, a closer inspection reveals crucial differences. While the quality of the identification of point-like sources does not differ significantly, there are significant differences in the diffuse emission regions that strongly favor the SRGonG segmentation. Notice that

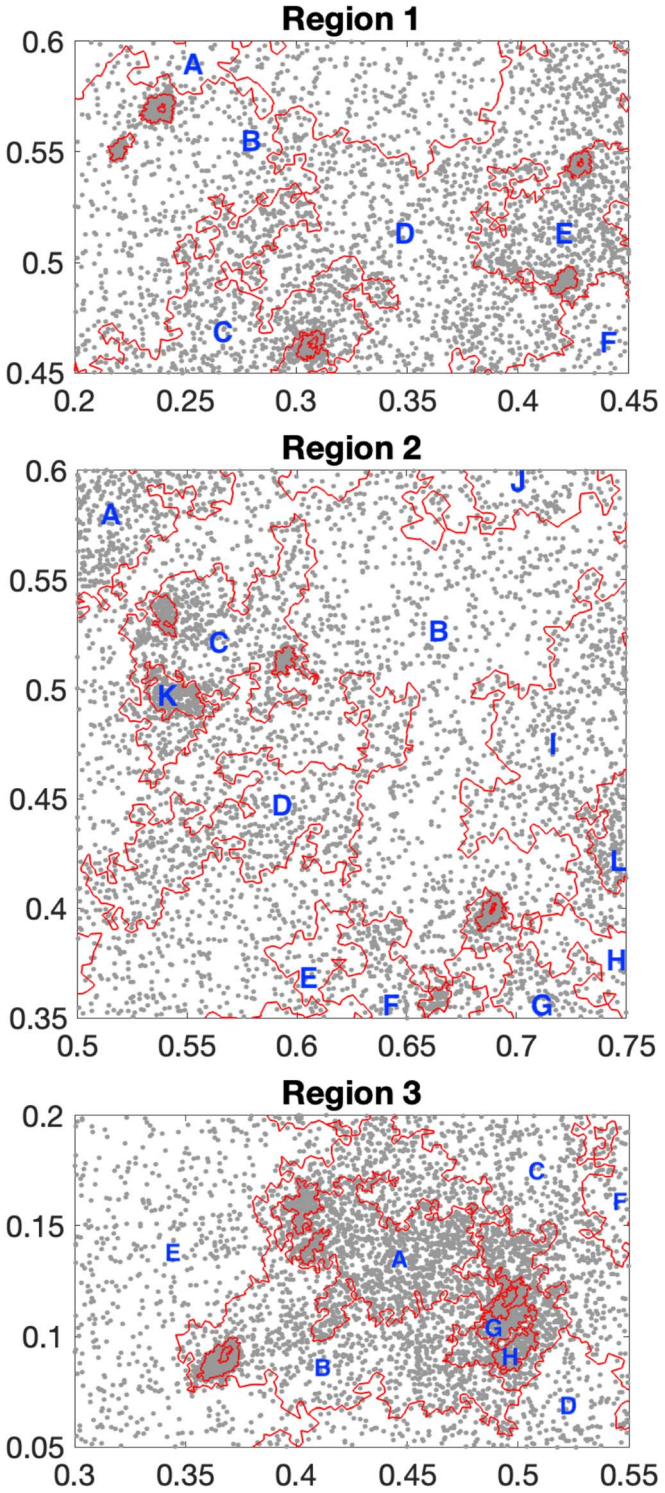

Figure 7. Magnified views of the three regions of the Antennae galaxies highlighted in the bottom right panel of Figure 5. The gray dots represent individual photons, and the thin red curves represent the SRGonG segmentation (see Section 5). Some of the larger segments are labeled with blue letters (see text).

segment C of Region 1 from SRGonG (top panel of Figure 7) is missing in the brute-force segmentation, and is effectively subsumed into segment D, which in turn also subsumes segment B. These changes are prima facie unsupported by the visible variations in the surface density of photons. Similarly, we see that segment E is incorrectly extended, and a different segment extends down into the middle of segment D. Such cases are also seen in Region 2 (middle panel of Figure 7), where all of the complexity found as segments C, D, and E are lost in the brute-force segmentation; and in Region 3 (bottom panel of Figure 7) where the clear separation of segments A and B is lost in the brute-force segmentation, as is the point-like source at ∼(0.4, 0.15). In summary, clear variations in the surface brightness are recovered in the SRGonG segmentation, unlike in the brute-force method. Using a smaller set of seeds *improves* the robustness of the segmentation by avoiding the chaotic development of early merging steps; errors in early stages accumulate because of the greedy merging process. We thus conclude that SRGonG is superior because of, and not despite, the much smaller, but perceptively selected number of seeds used to carry out the segmentation.[12]

Although SRGonG is not designed as a point-source detection method, it is instructive to see how it behaves in the case of point-like sources. In Figure 9 we show point-like sources can be identified in SRGonG (left panel) compared to a wavelet-based method that is optimized to detect point sources (right panel, wavdetect; Freeman et al. 2002). Based on the typical size of the Chandra PSF, we isolate all SRGonG segments that cover an area of comparable or smaller size to the PSF[13] and show them in the left panel. (We emphasize that this is not a method to *detect* point-like sources; while SRGonG segments with larger areas than the PSF size can be flagged as extended, regions with small areas cannot be definitively flagged as point sources, as such segments can occur due to the layered segmentation of extended sources or even due to statistical fluctuations in the surface brightness of diffuse emission.) In the right panel, we show all wavdetect-detected sources, superposed on a counts image of the same field. As wavdetect is optimized to find point sources in a variety of scales, it may also detect more diffuse sources. Sources that are identified as extended based on visual inspection and/or comparison with the PSF profile (e.g., lack of a core, or PSF fitting for sources with more than 100 counts; Zezas et al. 2002) are marked by red circles, while point-like sources are marked by cyan circles. We note that the SRGonG segmentation invariably marks as "point-like" (based on the segment area criterion) the point sources that are confirmed by the inspection process of Zezas et al. (2002) and does not find the extended sources identified by wavdetect. The latter are instead components of larger diffuse emission segments. In this respect, although the SRGonG is not a point-source detection algorithm, screening of the identified segments based on the segment area, Area($\mathcal{S}_k$), can be used to distinguish extended regions from point-like sources.

[12] Just as Markov Chain Monte Carlo techniques rely on running multiple chains and verifying consistency to gain confidence in the analysis results (Gelman & Rubin 1992), we recommend that analyses that use SRGonG also consider the sensitivity of the results to the adopted seed set. The schemes that we recommend in Section 3.2 are adequate to handle most scenarios encountered in astronomy, but are nonetheless characterized by several runtime-specified parameters ($m_{\rm grid}$, $m_{\rm graph}$, $m_{\rm nn}$, $m_{\rm R}$, $m_{\rm vorthr}$). Work to formalize this process via bootstrap analysis is ongoing (J. Wu 2022, private communication).

[13] We choose segments identified by areas Area($\mathcal{S}_k$) $\leqslant 0.0003$ in normalized coordinates, which corresponds to a circular area of radius $\approx 1\farcs6$ on the sky, comparable to the extent of the Chandra PSF.

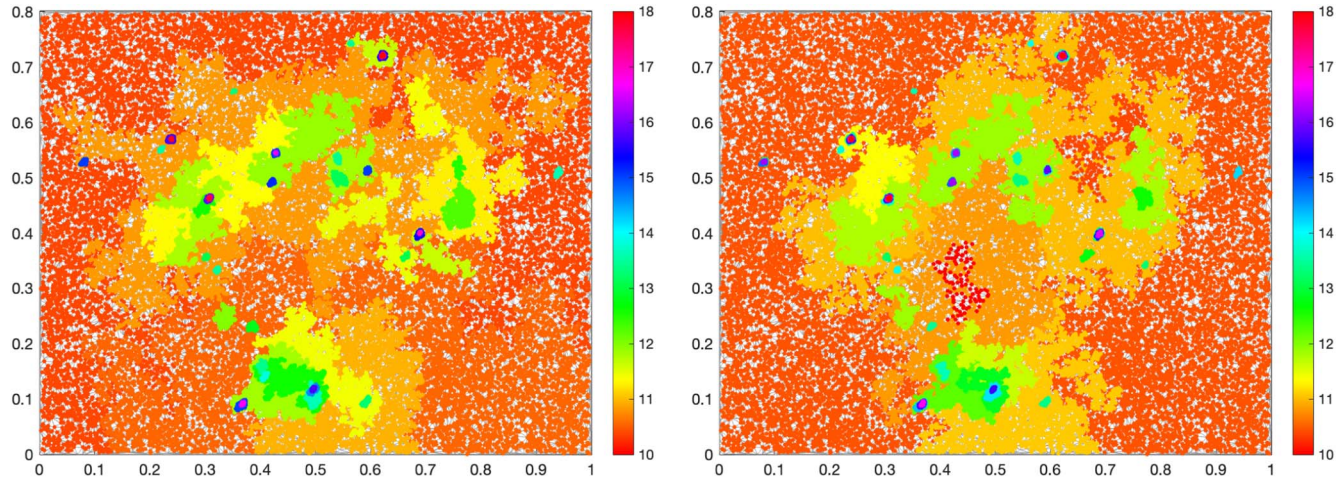

Figure 8. Segmentation of the photon list shown in Figure 5. Photons are color coded by the estimated intensity of their segment. Left panel: 55 fitted segments obtained with SRGonG with seeds specified via a grid supplemented by local maxima. Right panel: 61 fitted segments obtained for brute-force segmentation, where all photons are assumed to be seeds. The differences between these two segmentations are described in Section 6.1.

6.2. Advantages and Limitations of SRGonG

Our simulations (Section 4) and analysis of the Chandra Antennae data set (Sections 5 and 6.1) illustrate SRGonG's strength in identifying sources at many different scales. The method allows the identification of extended diffuse structures in X-ray data regardless of their shape, i.e., no assumptions are made about the morphology or the homogeneity of the sources. Note that, while the blurring due to the shape of the PSF is not explicitly modeled, this has negligible effect on any source structure at scales larger than the size scale of the PSF. Thus, we expect that useful results can be obtained even when the PSF varies across the field of view, which can happen due to several reasons (the quality of the telescope optics can degrade away from the aim point; or fields are observed that have a large diversity of soft and hard sources, each with significantly different PSF shapes and sizes; or when complex combinations of data sets, such as multiple observations carried out at different angular offsets, are combined). At spatial scales larger than the PSF size, we expect that results are not reliant on the specific characteristics of the PSF.[14]

Even when the photons are sparsely distributed, e.g., when the observation is dominated by diffuse structures at low surface brightness, and blurring due to the shape of the PSF is not included, point sources that may exist in the field of view can be identified due to the increased concentration of photons at their locations. However, note that SRGonG is designed to identify large-scale extended regions, so the focus and trade-offs are different. Thus, weak point sources with low contrast against the surrounding diffuse emission are likely to be subsumed into the diffuse regions. However, because these are by definition weak, they are unlikely to contribute significantly to the brightness (or hardness) of the diffuse component. This situation is effectively similar to the situation where the detection sensitivity of a telescope is insufficient to resolve apparently diffuse emission into its point-source population. An additional issue to consider is the bias in the point-source brightness demonstrated in Figure 3. This bias arises as area fluctuations from small point sources are naturally bounded at zero, but can extend into the area of the diffuse emission, leading to a skewed error distribution. So point-source intensities found by SRGonG should not be used directly but must be reestimated using appropriate techniques (e.g., Primini & Kashyap 2014). However, note that the bias demonstrated in Figure 3 comes from point sources that do not have PSF wings; in real sources where point sources are sharply peaked due to the PSF, the area distribution bias works to the advantage of SRGonG, incorporating more of the PSF wings into the point source and reducing the resulting contamination of the diffuse emission by strong point sources.

Unlike adaptive smoothing, source detection, or contouring methods, SRGonG does not set S/N thresholds or rely on thresholds of source significance to determine the presence or extent of contiguous regions. Thus, even regions that may be characterized by low surface brightness tend not to be oversegmented. Conversely, as the uncertainty in the estimated brightness is dependent on the number of photons that fall within the segment, small variations in adjacent regions can be more easily distinguished when the areas of the segments are sufficiently large.

Also of note is that SRGonG works directly on photon lists, the most basic form of high-energy X-ray and γ-ray data sets, and the Poisson nature of the data is explicitly accounted for during the merging process. While this can have detrimental effects on the running time when the size of the data set is large,[15] using the data at the highest available resolution avoids the requirements to define artificial binning sizes.

Of greater concern is the dependence of SRGonG results on the distribution of the initial seeds, especially for fields with low contrast. This may result in unstable behavior because of fluctuations in the local minima in the spatial intensity distribution, leading to both false segmentation and false merging. We caution that while the schemes we describe in Section 3.2 are generally adequate and perform well (see, e.g., Section 6.1), as is typical with seeded-region-growing methods,

[14] Note that PSF size information *may* be incorporated into the analysis, e.g., by requiring that any segment that is found to have a smaller area than that of the PSF be subsumed into a surrounding or adjacent segment. We do not use such a criterion in this work, though such a strategy is demonstrated in Figure 9.

[15] For illustration, the analysis of the Antennae data set, with 50,700 photons and 491 seeds, takes ≈240 s on a 2021 epoch 14″ MacBook Pro with an Apple Silicon M1 SOC.

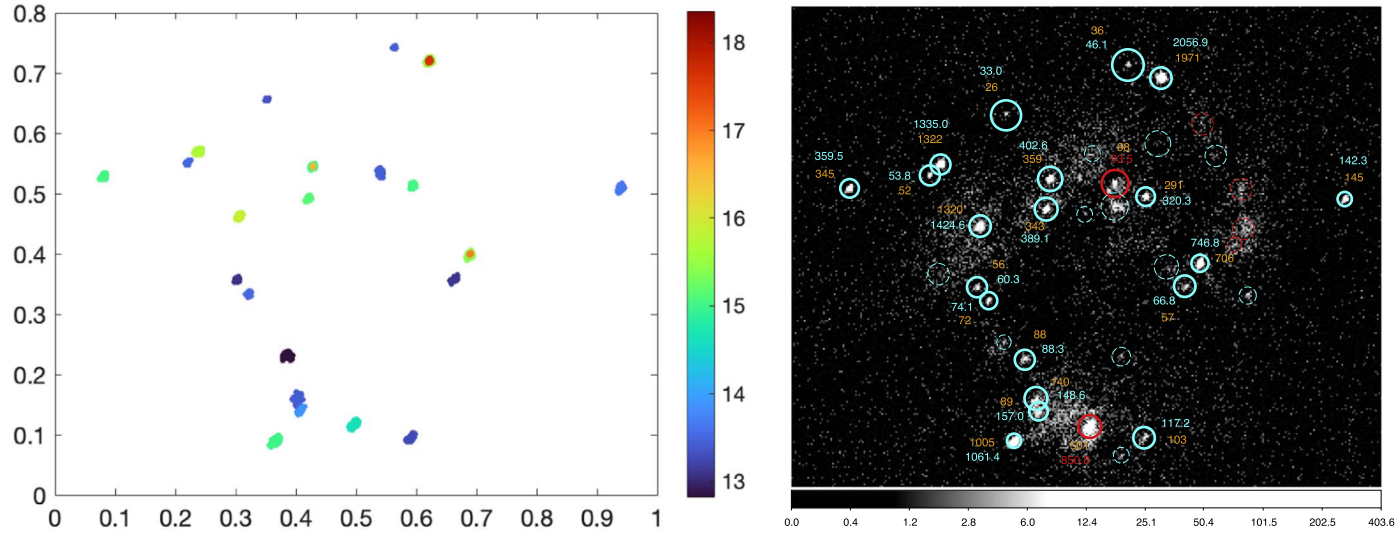

Figure 9. Identifying point-like sources in the event list data in Figure 5. Left panel: "Point-like" `SRGonG` segments selected to have a segment area $\leqslant 0.0003$ (31 total segments; 22 after concentric segments are merged). Right panel: sources detected with `wavdetect` (Zezas et al. 2002) overlaid on a full-band image of the Chandra data (OBSID 315). The size of the field is $2'.8 \times 2'.3$. Red circles mark the detected sources that were shown to be inconsistent with Chandra's PSF according to the analysis of Zezas et al. (2002), while cyan circles mark the remaining sources (some of which may also be extended). The size of the circles corresponds to the optimal `wavdetect` scale. Thicker solid circles indicate sources that are also identified by `SRGonG`; the remaining sources are indicated by the thinner dashed circles. The cyan and orange numbers give the source counts measured by the customized aperture photometry of Zezas et al. (2002) and the `SRGonG` method, respectively (the `SRGonG`-based counts are not background subtracted). Note that most strong point-like sources are identified as such by the `SRGonG` method, and in general there is good agreement in the estimated source intensities.

the sensitivity of the segmentation to the adopted seed structure must always be checked.

Future extensions of this method will include quantification of the uncertainty of the segmentation resulting from the stochastic nature of the data, which would be quantified in terms of uncertainty on the number of segments, the outline of the segments, and the corresponding source flux within each segment. Other avenues to explore include different merging procedures as substitutes for the greedy merge to address the oversegmentation. The goal of such alternative merging options would be to search more possible final segmentations and make the algorithm more robust to seed initialization. Yet another potential extension is to perform the analysis in three dimensions, incorporating photon energy information. Currently spectral information can only be used by running the code on passband filtered data.

7. Summary

We have developed an algorithm that provides a piecewise constant segmentation of a photon event list that approximates the spatial structure present in the data. Point-wise surface brightnesses are initially estimated as the inverse of their Voronoi cell areas and cells with similar brightnesses are grouped together to grow segments. The seeds needed to grow the segments can be initialized as regular grids, additionally supplemented with local maxima, or set using more complex processes by stratified sampling of Voronoi cell areas. The process begins with a deliberate oversegmentation, and neighboring segments are sequentially merged by maximizing the BIC change. The resulting (greedy) segmentation generates apertures on the sky plane that can be used to collect photons and carry out further analysis in a way that removes manual intervention in selecting regions of interest. We have explored this method via both simulations and application to a complex Chandra data set, and find that it consistently provides a good description of both point-like and extended diffuse regions of arbitrary shapes.

We note that this is not a source detection method, but a robust method for the definition of source regions, especially for extended sources. In this way, it can be used to perform photometry or spectroscopy on arbitrarily shaped extended sources.

This method provides several advantages over other commonly used methods for the analysis of extended sources in high-energy photon data. Namely, it allows the identification of sources at different scales even when they are embedded within each other without imposing any restrictive assumptions on the spatial distribution of the source photons or the source intensity.

This work was conducted under the auspices of the CHASC International Astrostatistics Center. CHASC is supported by NSF DMS-18-11308, DMS-18-11083, DMS-18-11661, DMS-21-13615, DMS-21-13397, and DMS-21-13605; by the UK Engineering and Physical Sciences Research Council [EP/W015080/1]; and by NASA 18-APRA18-0019. We thank our CHASC colleagues for many helpful discussions, especially Jilei Yang for his valuable comments on an earlier draft. M.F., J.W., and T.C.M.L. acknowledge further support from NSF through CCF-19-34568, DMS-18-11405 and DMS-19-16125. V. L.K. further acknowledges support from NASA contract to the Chandra X-ray Center NAS8-03060. D.v.D. and A.Z. were also supported in part by a Marie Skodowska-Curie RISE (H2020-MSCA-RISE-2015-691164, H2020-MSCA-RISE-2019-873089) grants provided by the European Commission.

Facility: Chandra (ACIS).

Software: CIAO (https://cxc.harvard.edu/ciao/; Fruscione et al. 2006); *PINTofALE* (Kashyap & Drake 2000); *Matlab* (https://www.mathworks.com/products/matlab.html); `SRGonG` (https://github.com/jujWang96/Astro_sim).

Appendix
Nearest -neighbor Labeling

Here we describe the heuristic by which selected photons are collected into groups characterized by their proximity (used in the Voronoi-area-stratified sampling scheme for seed specification; see Section 3.2). The photons considered in a given stratum are defined by a small range of Voronoi areas, or analogously, are located at similar contour levels if an image were constructed from the photons. Thus, they are likely to be sparsely distributed, but with clumps of photons surrounding higher-intensity regions. The goal here is to group the clumped photons that are near each other, without breaking up rings or other complex shapes. We emphasize that this heuristic is a quick but approximate preprocessing method to pick seeds for the full-fledged SRGonG algorithm. We expect this heuristic to be useful in situations where the astronomical data set is characterized by sparsely distributed structures with a large dynamic range in surface brightness.

We first determine an average characteristic length scale for the ensemble of photons included in stratum Υ, as

$$L_{\Upsilon} = 2\sqrt{\frac{1}{2}\left[\max_{i\in\Upsilon}\{\mathrm{Area}(\mathcal{V}_i)\} + \min_{i\in\Upsilon}\{\mathrm{Area}(\mathcal{V}_i)\}\right]}.$$

This ensures that the length scale is typical of stratum Υ. We begin with an arbitrary photon from Υ, assigning it a unique group label, and recursively assign this group label to any neighbor, i.e., any photon in stratum Υ located within a Euclidean distance of L_{Υ} from any photon assigned to this group. The recursive labeling ends when no new neighbors are present, and we move to another arbitrary as yet unlabeled photon in Υ, assign it a different label, and repeat the process. We continue this labeling until all photons in Υ are assigned labels. For a case where the photons are placed uniformly on a regular grid, this results in all the photons being aggregated into one clump with one label. If there are multiple clumps separated by $>L_{\Upsilon}$, each clump will be assigned a separate label. We eventually discard all clumps with fewer than m_{vorthr} photons and do not use them to set a seed. The entire process is repeated for each of the m_{R} strata.

ORCID iDs

Vinay L. Kashyap https://orcid.org/0000-0002-3869-7996
Thomas C. M. Lee https://orcid.org/0000-0001-7067-405X
David A. van Dyk https://orcid.org/0000-0002-0816-331X
Andreas Zezas https://orcid.org/0000-0001-8952-676X

References

Adams, R., & Bischof, L. 1994, ITPAM, 16, 641
Akhlaghi, M., & Ichikawa, T. 2015, ApJS, 220, 1
Allard, D., & Fraley, C. 1997, JASA, 92, 1485
Angel, S., Parent, J., & Civco, D. L. 2010, TCG, 54, 441
Aue, A., & Lee, T. C. M. 2011, Ann. Stat., 39, 2912
Baldi, A., Raymond, J. C., Fabbiano, G., et al. 2006, ApJS, 162, 113
Barr, C. D., & Schoenberg, F. P. 2010, Biometrika, 97, 977
Bertin, E., & Arnouts, S. 1996, A&AS, 393, 117
Bobin, J., Sureau, F., & Starck, J.-L. 2016, A&A, 591, A50
Chiu, S. N., Stoyan, D., Kendall, W. S., & Mecke, J. 2013, Stochastic Geometry and its Applications (New York: Wiley), doi:10.1002/SERIES1345
Ebeling, H., White, D. A., & Rangarajan, F. V. N. 2006, MNRAS, 368, 65
Ebeling, H., & Wiedenmann, G. 1993, PhRvE, 47, 704
Fabbiano, G., Zezas, A., & Murray, S. S. 2001, ApJ, 554, 1035
Fan, M., & Lee, T. C. M. 2014, IET Image Process., 9, 478
Farias, H., Ortiz, D., Damke, G., Jaque Arancibia, M., & Solar, M. 2020, A&C, 33, 100420
Freeman, P. E., Kashyap, V., Rosner, R., & Lamb, D. Q. 2002, ApJS, 138, 185
Fruscione, A., McDowell, J. C., Allen, G. E., et al. 2006, Proc. SPIE, 6270, 62701V
Gelman, A., & Rubin, D. B. 1992, StaSc, 7, 457
Gondhalekar, Y., de Souza, R. S., & Chies-Santos, A. L. 2022, RNAAS, 6, 128
Hausen, R., & Robertson, B. E. 2020, ApJS, 248, 20
Hubert, L., & Arabie, P. 1985, J. Classif., 2, 193
Jones, D. E., Kashyap, V. L., & van Dyk, D. A. 2015, ApJ, 808, 137
Kashyap, V., & Drake, J. J. 2000, BASI, 28, 475
Lee, T. C. M. 2000, JASA, 95, 259
Magnussen, S., Allard, D., & Wulder, M. A. 2006, Scand. J. For. Res., 21, 239
Meyer, A. D., van Dyk, D. A., Kashyap, V. L., et al. 2021, MNRAS, 506, 6160
Møller, J. 1994, Lectures on Random Voronoi Tessellations, Vol. 87 (New York: Springer)
Otsu, N. 1979, ITSMC, 9, 62
Peng, B., Zhang, L., Zhang, D., et al. 2011, ITIP, 20, 3592
Picquenot, A., Acero, F., Bobin, J., et al. 2019, A&A, 627, A139
Picquenot, A., Acero, F., Holland-Ashford, T., Lopez, L. A., & Bobin, J. 2021, A&A, 646, A82
Primini, F. A., & Kashyap, V. L. 2014, ApJ, 796, 24
Rand, W. M. 1971, JASA, 66, 846
Sanders, J. S. 2006, MNRAS, 371, 829
Sanders, J. S., & Fabian, A. C. 2001, MNRAS, 325, 178
Scargle, J. D. 2002, in AIP Conf. Proc. 617, Bayesian Inference and Maximum Entropy Methods in Science and Engineering (Melville, NY: AIP), 163
Scargle, J. D., Norris, J. P., Jackson, B., & Chiang, J. 2013, ApJ, 764, 167
Starck, J. L., Pantin, E., & Murtagh, F. 2002, PASP, 114, 1051
Vikhlinin, A., McNamara, B. R., Forman, W., et al. 1998, ApJ, 502, 558
Zezas, A., Fabbiano, G., Baldi, A., et al. 2006, ApJS, 166, 211
Zezas, A., Fabbiano, G., Rots, A. H., & Murray, S. S. 2002, ApJS, 142, 239