Source Finding in Crowded Fields

W. D. Cotton (NRAO), J. J. Condon (NRAO), A. Matthews (NRAO),T. Mauch (SARAO) April 3, 2019

***Abstract*—The problems of finding and categorizing sources in crowded interferometer images are examined and a scheme for avoiding many of them is discussed. Analysis centers on a seriously confusion–limited high Galactic latitude image obtained using the MeerKAT array at 1.3 GHz with a thermal noise of $\approx$ 570 μJy/beam and a resolution of 7.6". Limiting the region of the image being used to "islands" above some threshold is needed to keep fits of nearby sources from interacting. Constraining the maximum angular size of components is critical to obtaining accurate peak values in fitted components. A simulation with properties of the observed field and with a realistic distribution of flux densities is used to evaluate the reliability of associating features in the image with real sources. Associations of peaks below about 30 times the thermal noise or 15 times the RMS "rumble" in confusion limited areas are found to be problematic.**

***Index Terms*—source finding, crowded fields**

I. Introduction

IDENTIFYING and categorizing discrete sources in images of the sky is a critical part of extracting the astrophysical information. This process is simpler for extragalactic than galactic objects as they tend to be relatively spatially–confined and are associated with an individual galaxy or group of galaxies. These objects can be characterized by fitting components of some basis function, e.g. elliptical Gaussians, to the image; these components are then the basis of the catalog derived from the image. Extracting a list of discrete sources simplifies the association of individual objects with similar lists derived from other instruments and is used for much astrophysical interpretation. Details depend on the type of instrumentation involved and the following is explicitly for images made from radio interferometer arrays.

When images are sparse well separated by areas dominated by noise, identifying sources is fairly straightforward and the fitting of the basis functions is generally not unduly affected by other sources. In the case where the field is crowded, adjacent sources may overlap the problem is more complex. In the limit of a "confusion limited" image, blends of weaker sources are difficult to distinguish from real sources and even the "noise" may be dominated by the contributions of sources too weak to distinguish individually. This memo discusses source finding in crowded fields as implemented in the program FndSou in the Obit package [1] [1].

II. Basic Source Finding

In the current context, source finding in an image consists of the following:

1) Identify continuous"islands" of emission above some threshold representative of significant emission. These islands can be defined by rectangular boxes in the image enclosing all pixels in the island. In sparse radio images where sources are located many beams apart, the majority of islands enclose a single source. In crowded fields, these "islands" may themselves contain multiple smaller islands. Since these rectangular boxes may contain more than one island, they are referred to as "regions" in the following.
2) Identify local maxima in each region representing potential sources or parts of a source.
3) Further define islands in a region by blanking (indicating pixels with no value) any pixel not contained in the island surrounding any peak in the region. This process may combine previously disjoint islands in a given region but not change the size or location of the region. This is necessary to minimize the effects of the separate sources on the fitting of each other. The "waterline" (blanking threshold) in this process need not be the same as in the original definition of the islands but is constant across the image. A lower value will allow more pixels in valid sources to be used in the fitting.
4) Fit a set of basis functions (e.g. elliptical Gaussians) with initial locations at pixel peaks in the region and subject to constraints:
   - The size of fitted components can be constrained to be no smaller than the restoring beam used in the imaging. This allows the fitted components to always be deconvolved from the restoring beam giving an estimate of the true size. For example, fitted Gaussians smaller than the restoring beam correspond to source with imaginary sizes.
   - An upper limit of fitted component sizes can be applied, especially if *a priori* knowledge supports this. The sizes of weak sources are poorly constrained by the image and a maximum allowed size of the peak flux density may be useful. For very weak sources an upper limit equal to the restoring beam size may be reasonable.
   - Positions of the centers of components can be constrained to be inside the region.
   - If a single component fit is deemed inadequate on the basis if the residuals to the fit, a two component fit can be attempted and accepted if it gives a better fit and the components are more separated than a given fraction (e.g. half) of the restoring beam size.
5) Valid fitted components may be selected based on a number of criteria:
   - Peak flux density. To get a (nearly) complete of

National Radio Astronomy Observatory, 520 Edgemont Rd., Charlottesville, VA, 22903 USA email: bcotton@nrao.edu
South African Radio Astronomy Observatory, 2 Fir St., Observatory, South Africa

[1]http://www.cv.nrao.edu/∼bcotton/Obit.html

sources to a given flux density level, the initial waterline level used to defining islands needs to be below the desired minimum.
- Minimum SNR of the fitted peak. Low significance fits can be excluded.
- Minimum island/region size. Imaging artifacts are frequently narrow in size.

There are additional considerations for source finding in crowded fields. As noted above, there may be, and frequently are, multiple individual peaks and/or islands in a given region. This has to be taken into account in the initial definitions of components in a region. Furthermore, all image features represented in the pixels going into a given fit need to be included in the fit. For instance, if a fitting region includes the shoulder of a nearby source but not its peak, the fitting will not attempt a separate component for the nearby source but will try to incorporate it in the fitting of other components. In order to avoid this only pixels in the islands of components being fitted should be included.

Fitted sizes of weaker sources may be made larger by the fitting process attempting to accommodate the even weaker, nearby sources. Blanking pixels below an appropriate level helps reduce this problem as does appropriate upper limits on fitted component sizes.

In confusion limited portions of images, the RMS of brightness variations may be dominated by the contributions of numerous weak source and not be describable as a Gaussian, or even a zero mean process. This renders the interpretation of the significance of a component based on its measured "SNR" of questionable value. This problem is further complicated by the potential presence of many blends of weaker source which are difficult to distinguish from real sources. This problem becomes acute for sources a few times the RMS "rumble". This is examined further in Section V-A.

III. Implementation: Obit/FndSou

A source finding program as outlined in the previous section in implemented in the Obit package as the program FndSou. This program reads an image and finds, and fits Gaussians to sources in a specified portion of the image. The results are saved in an AIPS MF (model fit) table and optionally an AIPS VL table (source catalog) as well as a text file. Least squares fitting with constraints uses an adaptation of the dvdmin routine in the Obit ObitImageFit class. FndSou is an adaptation of the AIPS task VSAD which was adapted from task SAD ("Search And Destroy").

Features controlling the selection of islands and constraints on fitting are described in the following:

- **NGauss:** The maximum number of regions allowed.
- **CutOff:** The flux density "waterline" for the initial definition of islands.
- **Blank:** The flux density "waterline" for the blanking of non-island pixels in a region just prior to fitting.
- **Retry:** Constant part of test on maximum absolute residual for attempting breaking single Gaussians into two. The test level is $\sqrt{(\mathrm{Retry}^2 + (\mathrm{Gain} * \mathrm{peak})^2}$ where peak is the peak value of the single Gaussian fit.
- **Gain:** See **Retry**.
- **doMult:** Allow multiple sources in an island?
- **doWidth:** Fit for component sizes?
- **Parms:** General fitting constraints.
  - Parms[0]: The minimum acceptable peak flux density. Components with fitted values less than this are rejected.
  - Parms[1]: Maximum size of components (arc-sec) applied as a constraint in the fitting.
  - Parms[2]: Maximum distance of a component centroid outside of its enclosing region, applied as a constraint in the fitting.
  - Parms[3]: If > 1.0 then only fit a size larger than the CLEAN restoring beam. This only applies if the restoring beam is round.
  - Parms[4]: Minimum width in pixels of an island.
  - Parms[5]: If > 0 then the maximum allowable False Detection Rate [2] [3].
  - Parms[6]: The maximum allowable fraction of a region allowed to be blanked.
- **sizeLim:** Peak flux density variable limits on maximum component size. If sizeLim[0]>0, these override Parms[1]. These specify a constant maximum size below a given peak flux level and above another and a linear ramp between the two.
  - sizeLim[0]: Max. size (cells) for sources with peak brightness $<$ sizeLim[2].
  - sizeLim[1]: Max. size (cells) for sources with peak brightness $>$ sizeLim[3].
  - sizeLim[2]: Lower flux limit for linear ramp.
  - sizeLim[3]: Upper flux limit for linear ramp.

IV. MeerKAT Deep 2 examples

Examples of the use of Obit/FndSou in a crowded field used a confusion limited image of the "DEEP_2" field (RA 04 13 26.4, Dec -80) made with the MeerKAT array with a center frequency of 1.28 GHz. This field was chosen to be particularly devoid of strong sources allowing a very sensitive image with a minimum of artifacts. This image is strongly confusion limited, the RMS "rumble" near the center is ≈ 1.5 μJy/beam while the thermal noise in the outer parts of the image has an RMS ≈ 0.57 μJy/beam.

A. Island Selection

Figure 1 shows a portion of the DEEP_2 image displaying the island/regions and marking the positions of fitted sources. The island/regions are initially set at a level of 5 μJy/beam and the final island definitions used 4μJy/beam. The maximum source size was flux density dependent and varied from 1 to 2 times the restoring beam size (7.6").

B. Source Size Limit

One way of imposing prior knowledge on the fitting process is to set an upper limit of the component sizes. We expect that the vast majority of real source in the DEEP_2 field will be at most marginally resolved at 7.6" resolution. Figure 2 shows

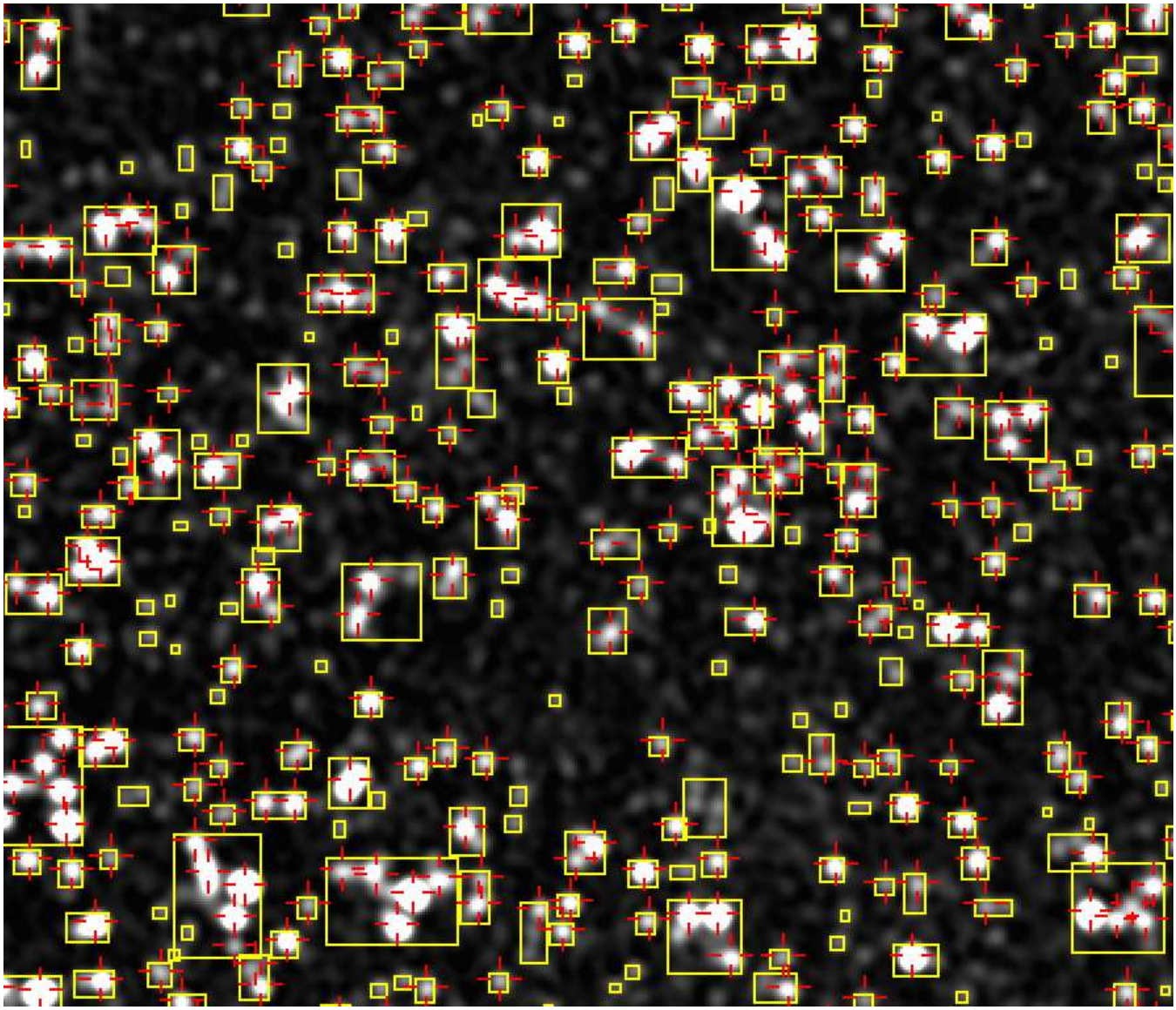

Fig. 1. Portion of the DEEP_2 field in gray-scale with regions overploted in yellow and positions of fitted sources with peaks > 10 μJy/beam are marked with open red crosses. The gray-scale range displayed is -1 to +20 μJy/beam.

the effect of size limits of 30" and 12.5"; the former giving an implausible result for the sources near the center.

Sizes of weaker sources are poorly constrained by the image in the best of circumstances and this is aggravated by the presence of many even weaker sources. On the other hand, source with large SNRs can have their sizes relatively well constrained by the image. A peak flux density variable upper limit on component size is defined by parameter sizeLim and is applied as a fitting constraint to both axes of the Gaussians fitted.

In the DEEP_2 image there are a number of well resolved AGNs which will be poorly represented by a collection of Gaussians. Other than these, most sources in the image are expected to be marginally resolved at most and an upper limit on the size of twice the psf (restoring beam) size appears to give plausible results, even for most stronger sources.

V. Simulations

Finding and fitting components to an image is merely a way to reduce the image to a smaller number of parameters and can shed limited light on the correspondence between features in the image and astronomical objects. This is especially true of crowded fields in which many of the weaker features are merely blends of even weaker objects. This lends the second meaning of "confusion" to confusion limited images. In real images, the ground truth is rarely known and we must resort to simulations.

In order to evaluate the association of various fitted features in the DEEP_2 image with actual galaxies; a simulation was created. This simulation matched the resolution and thermal noise of the DEEP_2 image and used a randomly spatially distributed collection of point sources with a distribution of flux densities given by the best current log(N)-log(S) curve at this frequency [4]. Simulated components were added to

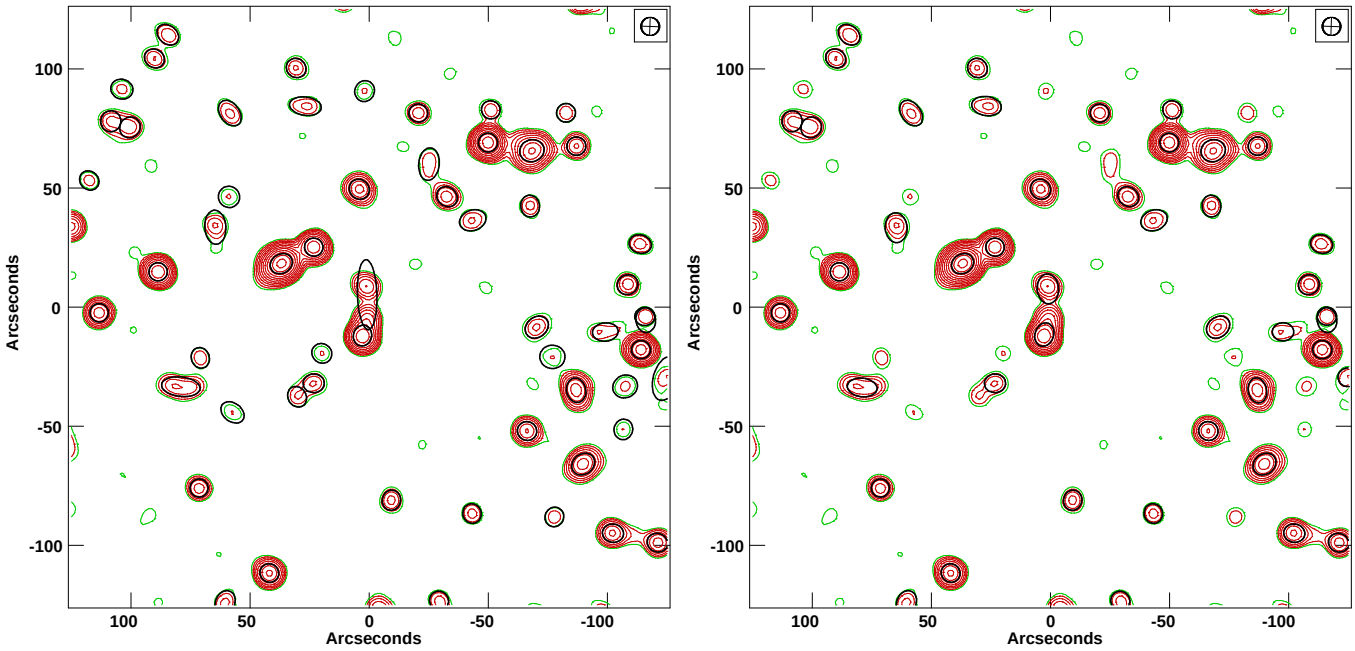

Fig. 2. Contour plot of a region of the DEEP_2 image showing the effect of upper limits on component sizes of the weak sources near the center of the plots. **Left:** Allowed upper limit of 30",
Right: Allowed upper limit of 12.5",
Fitted Gaussians are shown as black ellipses , the lowest, green, contour is at 4 μJy/beam and contours at higher levels are separated by a factor of $\sqrt{2}$ and are shown in red. Fitted components as faint at 7 μJy/beam are included. The restoring beam size is shown in the upper right.

the image convolved with the restoring Gaussian down to the CLEAN limit used in the DEEP_2 image and the actual psf ("Dirty beam") below that level. A portion of this image is shown in Figure 3. The comparison with Figure 1 immediately shows striking similarities. Like Figure 1, many of the weaker features in Figure 3 appear resolved whereas the simulation used exclusively point sources.

A. Reality of Fits to the Simulated image

In the case of the simulated image, the ground truth is known. A comparison of flux densities is given in Figure 4. Program FndSou was run on the simulated image allowing only point components and accepted results above 7 μJy. The derived catalog was then matched against the list of components contributing to the image.

In Figure 4 **Left:** an attempt was made to match a single component in the simulation with each fitted feature. As there are many weaker components in the simulation, constraints on position and flux density matches were required. Even with this, many of the fitted components were not matched and the large scatter in the plot, even to relatively high flux densities indicates that many of the fitted components are blends of weaker components.

Figure 4 **Right:** trys to compensate for this by summing simulation components within 1/2 of a beam-width of each fitted component. This plot still has considerable scatter as the summed components will not add their full flux density to the position of the fitted peak.

A traditional measure of the effects of confusion is the average number of beam areas per source above some level. These measures are summarized in Table I. A traditional cutoff

TABLE I
SYNTHESIZED BEAMS PER SOURCE

S_p μJy/bm	Beams per Source
11	16
17	25
27	40
43	66

Notes: At a peak flux density level of "S_p" and above there are "Beams per Source" synthesized beam areas per source on the average in the simulation.

for a confusion limited images is a source density at which there are more than 25 beam areas per source. By this measure sources fainter than 17 μJy are unreliable. This is supported by Figure 4.

VI. DISCUSSION

There are a number of critical steps in identifying and categorizing discrete objects in a crowded field. The first is identifying a set of pixels in the image containing connected "islands" of emission which should be fitted as a collection of components. These islands are a set of pixels above some "waterline" connected by adjacent pixels; all above this waterline. Initial locations of components are then identified with the positionss of one or more peaks in the island. This is straightforward for images derived from interferometers as the noise has a correlation function the same shape as the image "Dirty" psf.

Once the set of pixels to be used in a given least squares fit is identified, the actual fitting is best done with constraints on

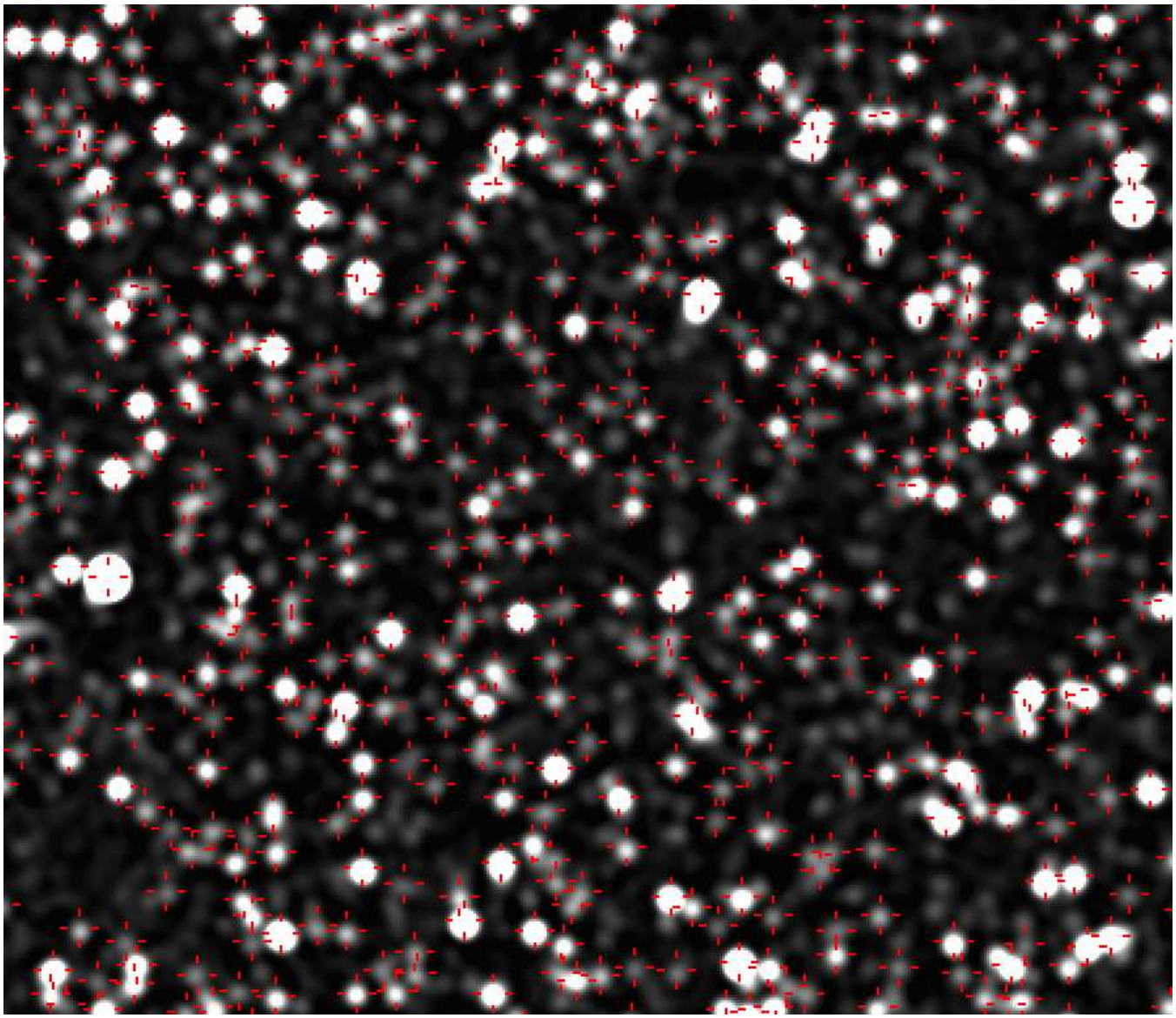

Fig. 3. Portion of the simulation matching DEEP_2 in gray-scale displaying a similar spatial size and stretch as Figure 1. Open red crosses indicate the locations of fitted components. All components contributing to this simulation are points.

the parameter values. We note that the most critical constraint on fitting is the allowed size of components. This is especially so in crowded fields where weak sources are surrounded by even weaker sources and the fitting will typically enlarge the fitted size to reduce the residuals from the weaker sources. In the case of the DEEP_2 image used for these tests, the vast majority of the sources are expected to be unresolved [5]. Our best results are obtained from requiring the fitted size of the faintest components be that of the restoring beam and allowing up to twice that for sources brighter than 100 μJy with a linear ramp in allowable size.

The reality of "sources" fitted to features in the image was evaluated using a simulation with the properties expected for the actual image and constructed from a set of point components with a distribution of flux densities given by the best available log(N)-log(S) curve. This set of components was then convolved with the restoring and dirty psfs of the observed image taking into account the limited depth of the CLEAN and then Gaussian distributed noise was added. Components in this image were then located and fitted by a set of point sources and compared with the list used in the simulation. Figure 4 gives this comparison. The left-hand panel shows that below about 20 μJy (35 $\times$ the thermal RMS of 0.57 μJy or about 15 $\times$ the RMS confusion "rumble") association of fitted components with individual components in the simulation is very problematic. Even above this level, at least to 100 μJy, blends with weaker components disturb the derived flux densities. Table I shows that 17 μJy is the traditional cutoff for confusion. Caution should be exercised in the interpretation of fainter features in crowded images.

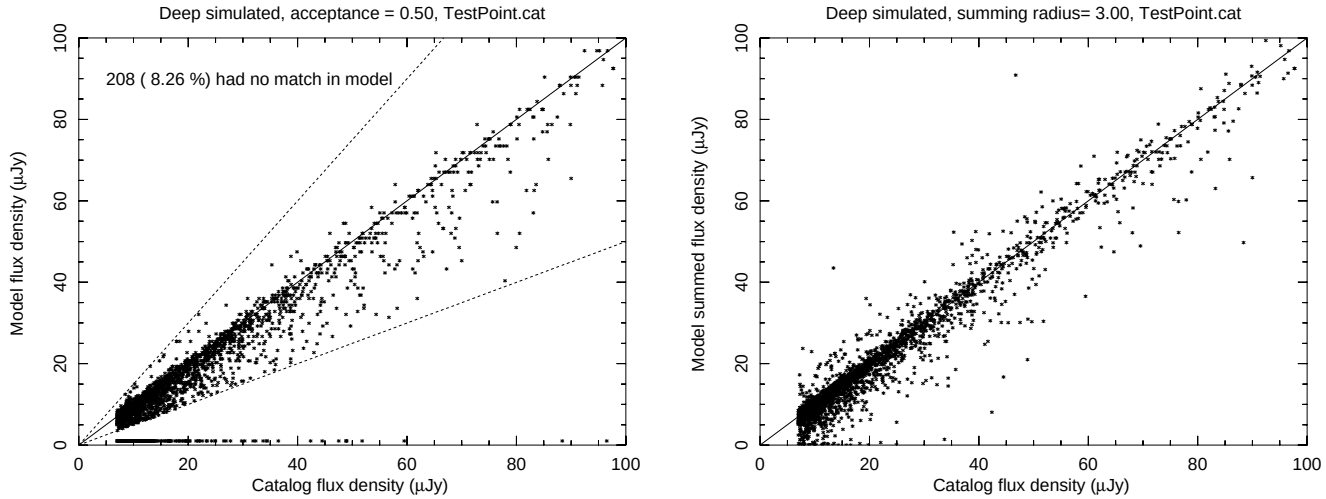

Fig. 4. Comparison of the fitted peak flux densities with the list of simulated components.
Left: Flux-Flux plot using the simulation component ("Model") closest to the fitted ("Catalog") component but within 1/2 of a beam and within a factor of 2 of the flux density. The solid line is at unit ratio and the dashed lines show the region in which a simulation component could match the fitted component. Fitted components with no matches are shown at the bottom.
Right: Flux-Flux plot using the sum of simulation component flux densities within 1/2 of the beam of the positions of fitted components.

REFERENCES

[1] W. D. Cotton, "Obit: A Development Environment for Astronomical Algorithms," *PASP*, vol. 120, pp. 439–448, 2008.
[2] W. D. Cotton and W. Peters, "False Detection Rate of Source Finding," *Obit Development Memo Series*, vol. 25, pp. 1–4, 2011.
[3] W. D. Cotton, "False Detection Rate of Source Finding Revisited," *Obit Development Memo Series*, vol. 43, pp. 1–5, 2016.
[4] J. J. Condon, W. D. Cotton, E. B. Fomalont, K. I. Kellermann, N. Miller, R. A. Perley, D. Scott, T. Vernstrom, and J. V. Wall, "'Resolving the Radio Source Background: Deeper Understanding Through Confusion," *ApJ*, vol. 758, p. 23, 2012.
[5] W. D. Cotton, J. J. Condon, K. I. Kellermann, M. Lacy, R. A. Perley, A. M. Matthews, T. Vernstrom, D. Scott, and J. V. Wall, "The Angular Size Distribution of μJy Radio Sources," *ApJ*, vol. 856, p. 67, 2018.