Rhodes University

Masters Thesis

Automation of source-artefact classification

Author: Makhuduga Lerato Lydia Sebokolodi

Supervisor: Oleg Smirnov

Co-Supervisors: Sphesihle Makhathini & Ian Heywood

A thesis submitted in fulfilment of the requirements for the degree of Master of Science

in Physics

Department of Physics and Electronics

Centre for Radio Astronomy Techniques and Technologies

October 2016

Declaration of Authorship

I, Makhuduga Sebokolodi, declare that this thesis titled, 'Automation of Source-Artefact Classification' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Signed: _____

Date: _____

Abstract

The high sensitivities of modern radio telescopes will enable the detection of very faint astrophysical sources in the distant Universe. However, these high sensitivities also imply that calibration artefacts, which were below the noise for less sensitive instruments, will emerge above the noise and may limit the dynamic range capabilities of these instruments. Detecting faint emission will require detection thresholds close to the noise and this may cause some of the artefacts to be incorrectly detected as real emission. The current approach is to manually remove the artefacts, or set high detection thresholds in order to avoid them. The former will not be possible given the large quantities of data that these instruments will produce, and the latter results in very shallow and incomplete catalogues. This work uses the negative detection method developed by Serra et al. (2012) to distinguish artefacts from astrophysical emission in radio images. We also present a technique that automates the identification of sources subject to severe direction-dependent (DD) effects and thus allows them to be flagged for DD calibration. The negative detection approach is shown to provide high reliability and high completeness catalogues for simulated data, as well as a JVLA observation of the 3C147 field (Mitra et al., 2015). We also show that our technique correctly identifies sources that require DD calibration for datasets from the KAT-7, LOFAR, JVLA and GMRT instruments.

Acknowledgements

In all things, always remember to give thanks...

I would like to give thanks to my supervisors; Oleg, Sphe, Ian and Arun for their patience and guidance. Particularly my co-supervisor Sphe, for not giving up on my writing and coding skills, and also for reminding me to always trust google. Thanks to Roger, Ermias and Ian for allowing me to use their datasets. To my colleagues, thanks for being such as a great team to work with and be around. To those that were brave enough to read my work namely: Mod, Kshitij, Roger, Sbo, Ribijil, Josh, Ulrich, Ermias, Trienko, Saalih, and Jonathan know that I am very thankful. It is through your contributions that this thesis is as is today. And also thanks to the SKA for trusting me with their moola.

To my precious family, I'm very grateful for your support, love and for believing in me. You guys are amazing. Mum, I will forever remember how this journey began, it was through your patience, persistence and love that I made it this far. I thank God for you, and my siblings Tibu, Khekhe and Fido. Koko, my beautiful big mum, thanks for your kind words and teachings. Paps, thanks for calling me a Dr. very early in my career, well the sky is the limit.

Ribijil, thanks for believing in me even when I ceased to and for supporting me throughout. Thank you Kis, Londi, Siv and Sbo for your support. Thanks, Siv, for not giving me a chance for any self-doubt. Monyonyo, my crazy one. Sapi, cool-time, the strongest lady I know, thanks for being a rock when I was tired of *adulting*. Lady D, G,1 Bro Alexido and Mama Drey you guys are the best.

To the Rock Family Church, you are the kindest people I have ever met and thanks for the teachings, guidance and the reminder of what and who God is. I appreciate that you have tolerated my "singing", well you guys are too strong.

Lastly, to the main man, the provider, the healer, the forgiver and the mighty father, thank you, Lord, for giving me unending love, wisdom, patience and the strength when I needed one. You are the Chief Cornerstone.

Contents

List of Figures

List of Tables

Introduction

The objective of this dissertation is to come up with a method to distinguish artefact emission from astrophysical emission in images. Currently, to avoid false detections, we manually have to fine-tune miscellaneous source finding parameters until we find models with better completeness and no (or fewer) artefacts.

The large data rates of the new radio telescopes such as the upcoming Square Kilometre Array (SKA; Dewdney et al., 2013) and its precursors and pathfinders; the Jansky Very Large Array (JVLA in the US; Napier, 2006), Low-Frequency Array (LOFAR in the Netherlands; van Haarlem et al., 2013), Australian Square Kilometre Pathfinder (ASKAP; DeBoer et al., 2009, Johnston et al., 2008), and the MeerKAT in South Africa (Jonas, 2009), require that calibration procedures become highly automated. For example, the sampling rate for the entire LOFAR at 200 MHz is 13 Tbits/s of raw data (van Haarlem et al., 2013). One of the major steps to automating calibration is the ability to generate accurate and complete sky models without the need to inspect the data.

Additionally, these telescopes are extremely sensitive and subject to complex direction-dependent effects (DDEs). The complexities of the DDEs make them difficult to model, therefore resulting in large systematic errors. The high sensitivities cause these systematic errors to emerge significantly above the noise and therefore increases the chance of them being misclassified as astrophysical emission during source finding. Another common practice to avoid artefacts, which is not always successful, is setting source finding thresholds very high. However, even though this may avoid false detections, it normally reduces the completeness of the resulting catalogues, this, in turn, compromises the full scientific potential of these instruments.

To address this problem we developed an algorithm that uses detection features such as peak flux, integrated flux, size, and local noise to classify them into false

or true detections. This algorithm implements and extends the technique known as the Negative detections Method, first developed and implemented in a similar context by Serra et al. (2012). Instead of spectral line data as in Serra, we developed this algorithm for continuum data, in particular, the upcoming continuum surveys such as MIGHTEE with the MeerKAT, Evolutionary Map of the Universe (EMU) with ASKAP, Westerbork Observations of the Deep APERTIF Northern Sky (WOGAN) with the upgraded Westerbork Radio Telescope (WSRT), Giant Metrewave Radio Telescope (GMRT) surveys, JVLA and LOFAR surveys (Norris et al., 2013).

As a result of the first approach, we were able to come up with a technique to identify sources subject to extreme DDEs. This technique is tested on the datasets from the KAT-7 instrument, the JVLA, the GMRT and LOFAR.

The dissertation is structured as follows:

Chapter 1 introduces the reader to basic radio interferometry and the measurement equation. It also describes the concept of calibration using the measurement equation and ways in which artefacts arise in images. This is followed by a brief description of the techniques implemented by source finders, and conditions that lead to misclassification of artefacts as astrophysical emission.

Chapter 2 describes the negative detection method used to distinguish artefacts from astrophysical emission, and further describes the algorithm that we have developed to carry out this task. The analyses are based on the JVLA simulated data and the JVLA observation of the $3C147$ field.

Chapter 3 describes the algorithm that automates the identification of sources that require direction-dependent (DD) calibration solutions. The algorithm is tested on LOFAR, JVLA, KAT-7 and GMRT datasets. This is followed by the general conclusion as well as the future work. Appendix A provides the detailed implementation of the software package we developed in this work.

Chapter 1

Background

This chapter aims to provide the reader with the tools necessary to understand the research conducted in this dissertation. First, we briefly describe the electromagnetic waves associated with astrophysical sources, followed by an introduction to radio telescopes, radio interferometry, and the form of measurements made by an interferometer. Further, we discuss the different types of propagation effects that affect the interferometric measurements and a mathematical formulation known as the Radio Interferometer Measurement Equation (RIME) used to infer and represent these effects. The process of correcting for the propagation effects is known as calibration. Since there exist no perfect calibration, there are always systematic errors induced during the process, which then build up as artefacts in images. When going deeper in an image, these artefacts surfaces significantly above the noise and are usually on the same flux level as the faint astrophysical sources, therefore making it difficult to differentiate. Later in this chapter, we describe source finding techniques and the difficulties encountered in the presence of artefacts. Lastly, we provide a detailed description of our objective, the scope of this thesis and the work done previously in a similar study.

1.1 Electromagnetic Waves

The propagation of the electromagnetic waves (EM) in free space is governed by following Maxwell's wave equation (derived from Maxwell's equations);

$$\frac{\partial^2 \boldsymbol{E}}{\partial t^2} - \frac{1}{\nu^2}\nabla^2 \boldsymbol{E} = 0 \tag{1.1}$$

where

$$\nu = \frac{c}{\sqrt{\mu\varepsilon}} \tag{1.2}$$

and ∇^2 is the Laplacian operator, $\boldsymbol{E}$ is the electric field, μ is the permittivity of the medium and ε_o is permeability of the medium. For propagation in free space; $\mu \to \mu_o$, $\varepsilon \to \varepsilon_o$, and $\nu = c$ (Wilson et al., 2012). A similar wave equation expression holds for the magnetic field, $\boldsymbol{B}$. Equation 1.1 is a homogeneous (i.e free of currents and charges) linear partial differential equation of second order (Wilson et al., 2012), thus, comprises of a large set of solutions, of which some are not solutions to Maxwell's equations (Burke and Graham-Smith, 2002). Thus, some simplifying assumptions must be made. One of the important things to note is that for a solution to satisfy both the wave equation and Maxwell's equations it must satisfy the dispersion relation given by Eq. 1.3.

$$\omega^2 = \frac{|\mathbf{k}|^2 c^2}{\mu\varepsilon} \tag{1.3}$$

where $\mathbf{k}$ is a wave vector, c is speed of light in the vacuum and ω is the angular frequency.

First, if we assume monochromatic harmonic waves then the solution takes the form:

$$\boldsymbol{E}(\boldsymbol{r}, t) = \boldsymbol{E}(\boldsymbol{r}) e^{-\imath\omega t} \tag{1.4}$$

where $\boldsymbol{r}$ is the position vector, $\boldsymbol{E}$ is a complex amplitude and t is time. Since astronomical objects are at large distances from the observer on Earth, incoming waves can be approximated as planar. This implies that $\boldsymbol{E}(\boldsymbol{r})$ varies only in one direction. Note that the direction of $\mathbf{k}$ specifies the direction of the propagation of the waves and is always perpendicular to both the electric field and magnetic field; also these two fields are always perpendicular to each other. If we let the direction of the propagation be along $\mathbf{k} = k\hat{z}$, then the plane wave solution is

$$\boldsymbol{E}(z) = \boldsymbol{E}_o e^{\imath kz}, \tag{1.5}$$

and the full solution becomes

$$\boldsymbol{E}(z, t) = \boldsymbol{E}_o e^{\imath(kz - \omega t)}, \tag{1.6}$$

where $\boldsymbol{E}_o$ is the constant amplitude vector which contains the phase information, $e^{\imath\phi}$, where ϕ is the phase (Burke and Graham-Smith, 2002).

If the direction of the electric field is defined along a unit vector $\hat{\mathbf{a}}$ which lies in the x, y-plane, then Eq. 1.6 can be written as

$$\boldsymbol{E}(z,t) = (\hat{\mathbf{x}}E_x + E_y\hat{\mathbf{y}})e^{\imath(kz-\omega t)}, \tag{1.7}$$

where amplitudes E_x and E_y are $E_x^o e^{\imath\phi_x}$ and $E_y^o e^{\imath\phi_y}$, respectively. Let's consider the physical quantity of the wave by taking the real part of the Eq. 1.7;

$$\begin{aligned}
\mathbf{E}(z,t) &= \Re\Big\{\big(\hat{\mathbf{x}}E_x + E_y\hat{\mathbf{y}}\big)e^{\imath(kz-\omega t)}\Big\}, \\
&= \big(\hat{\mathbf{x}}E_x + \hat{\mathbf{y}}E_y\big)\cos(kz-\omega t), \\
&= \hat{\mathbf{x}}E_x^o\cos(kz-\omega t+\phi_x) + \hat{\mathbf{y}}E_y^o\cos(kz-\omega t+\phi_y).
\end{aligned} \tag{1.8}$$

If the phases are equal, $\phi_x = \phi_y = \phi$, then we have linearly polarised waves of amplitude $|\boldsymbol{E}_o| = \sqrt{(E_x^o)^2 + (E_y^o)^2}$ and polarisation angle $\theta = \tan^{-1}\left(\frac{E_y^o}{E_x^o}\right)$. For a situation whereby the waves are out of phase by $\pm\pi/2$ and $\mathrm{E}_x^o = \mathrm{E}_y^o = \mathrm{E}_o$, the resulting waves will be right or left circularly polarised as shown below;

$$\begin{aligned}
\mathbf{E}(z,t) &= \hat{\mathbf{x}}\mathrm{E}_x^o\cos(kz-\omega t) + \hat{\mathbf{y}}\mathrm{E}_y^o\cos(kz-\omega t+\pi/2) \\
&= \hat{\mathbf{x}}\mathrm{E}_o\cos(kz-\omega t) + \hat{\mathbf{y}}\mathrm{E}_o\sin(kz-\omega t).
\end{aligned} \tag{1.9}$$

For the general values of the phases and amplitudes, the resulting polarisation mode is elliptical, thus making linear and circular polarisation special forms of polarisation. Electromagnetic waves from astronomical sources can be unpolarised or polarised.

1.2 Radio Telescopes

Radio telescopes are instruments designed to measure the EM radiation emitted at frequencies within the radio spectrum. A typical radio telescope consists of an antenna and a receiver. The different antennas include parabolic dish antennas, loop antennas, vertical antennas, dipole antennas, and much more, some still under development. To illustrate how the different parts of the radio telescope work, we consider the MeerKAT parabolic dish antenna shown on Fig 1.1. This

particular antenna layout is called the Offset Gregorian antenna. The incident

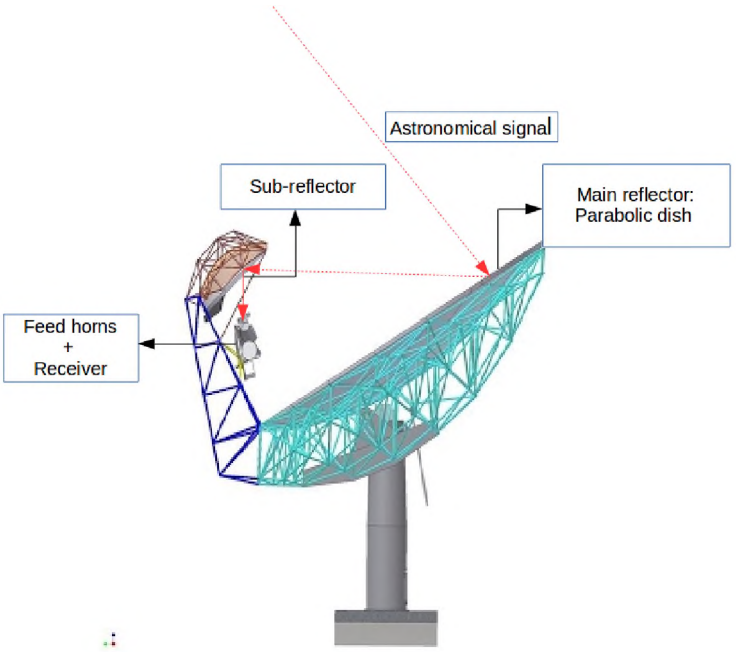


FIGURE 1.1: A Cassegrain radio telescope. Image Credit: SKA SA Public [1]

waves (dotted line) are reflected off the surface of the large parabolic dish (13.5 m) and redirected towards the secondary reflector (3.8 m), and are then focused in the feed horn. Inside the feed horn are the two orthogonal receptor components with each component sensitive to its corresponding polarisation component of the electric field (Hamaker et al., 1996). Thus, feeds are generally designed to measure the linear polarisation (denoted by X, Y) or circular polarisation (denoted as R, L for right and left circular polarisation; Hamaker et al., 1996, Kemball and Martinsek, 2005). For the MeerKAT, in particular, the feed horns are part of the receivers but note that receivers and feeds can be built separately. The purpose of the receivers is to capture the radiation and convert it to electrical voltages and then amplify it.

Telescopes are generally characterised according to two quantities, namely resolution and sensitivity. For a radio telescope, the sensitivity is defined as the faintest flux emission detectable, and is given mathematically by the following radiometer equation (Wrobel and Walker, 1999);

$$S_\nu = \frac{2kT_{\rm sys}}{A_{\rm eff}\sqrt{\Delta t \Delta \nu}}, \tag{1.10}$$

where k is the Boltzmann constant, T_{sys} is system noise temperature, $A_{\rm eff}$ is the dish effective area, Δt is the integration time and $\Delta \nu$ is the bandwidth. The

system noise temperature is given as;

$$T_{sys} = T_A + T_R, \tag{1.11}$$

where T_A is the antenna temperature and T_R is the receiver temperature. The receiver noise temperature is usually reduced by cryogenically cooling the system. The antenna temperature comprises of the following;

$$T_A = T_{CMB} + T_{object} + T_{RFI} + T_{ATM} + \cdots, \tag{1.12}$$

where T_{CMB} is the noise temperature from the cosmic microwave background, T_{object} is the noise temperature from astronomical objects, T_{RFI} is the radio frequency interference (RFI) from man-made radio signals such as cell phones, satellites, etc and T_{ATM} is the atmospheric noise.

The resolution, on the other hand, is defined as the minimum angular separation resolvable by the telescope, and defined mathematically as;

$$\theta \approx \frac{\lambda}{\mathrm{D}}, \tag{1.13}$$

where λ is the wavelength of the radiation and D is the diameter of the telescope. Obtaining high resolution with single dish has been one the challenging aspects to radio astronomy due to its longer wavelengths. Quantitatively, an optical telescope of 10m diameter and 550nm observational wavelength achieves a resolution of ~ 0.01 arcseconds. To obtain the same resolution with a radio telescope operating at 21cm wavelength requires a dish of diameter $\sim$ 4000km, which is not practically possible. Over the years, larger radio telescopes were built, and the current largest telescopes are the 100m Green Bank Telescope (steerable dish, USA), Effelsberg (100m steerable dish, Germany) and a Arecibo (305m non-steerable dish, Puerto Rico; Wijnholds et al., 2010). Large steerable dishes are difficult to build and are most often subject to gravitational pull and thermal expansion which results in a high level of pointing errors, tracking inaccuracy and surface deformation.

1.3 Radio Interferometry

Radio interferometry is a technique used in radio astronomy to combine signals from more than one telescopes into a single signal using the principles of interferometry (Wilson et al., 2012). The resolution in this case is given by λ/B_{max}, where B_{max} is the maximum baseline (where baseline means the projected distances between a pair of telescopes). Baselines can be made as large as possible by placing the telescopes at different locations on Earth or even in space, therefore increasing the resolution. The sensitivity, on the other hand, depends on the total effective area of the dishes as seen in Eq. 1.10. Thus, in order to improve the sensitivity one requires a large number of telescopes, large bandwidths and integration times as well as low noise receivers. For an interferometer, the radiometer equation becomes;

$$\sigma = \frac{\mathrm{SEFD}}{\sqrt{2N(N-1)\Delta\nu\Delta t}}, \tag{1.14}$$

where SEFD is the system equivalent flux density expressed as $2kT_{sys}/A_{eff}$, and N is the number of antennas.

1.3.1 Interferometric Measurements

Figure 1.2 is a simple two-element interferometer consisting of antenna p and q pointing in the direction $\hat{\mathbf{s}}$ towards an astrophysical source. As in section 1.1, consider a quasi-monochromatic source situated in the far-field of the telescope such that planar waves assumption holds. Thus, if $\hat{z}$ denotes the direction of the propagation, then the electric field vector, $\boldsymbol{e}$, in the direction x, y-plane can be represented as column vector of two complex numbers (Clark, 1980, Smirnov, 2011a);

$$\boldsymbol{e} = \begin{pmatrix} e_x \\ e_y \end{pmatrix}. \tag{1.15}$$

Further, let us assume that the instrument is perfect and that the observed sky is small enough that atmospheric variations are negligible. Since the antennas are located at different positions, the signal from an object not situated directly overhead (zenith) reaches the antennas at different times. As shown in Fig. 1.2, the incoming signal reaches antenna q before antenna p, due to the additional travel path length, $c\tau_g$, along the direction p. To ensure that the signals from the antennas meet coherently, we introduce the instrumental delay, τ_i, in the signal

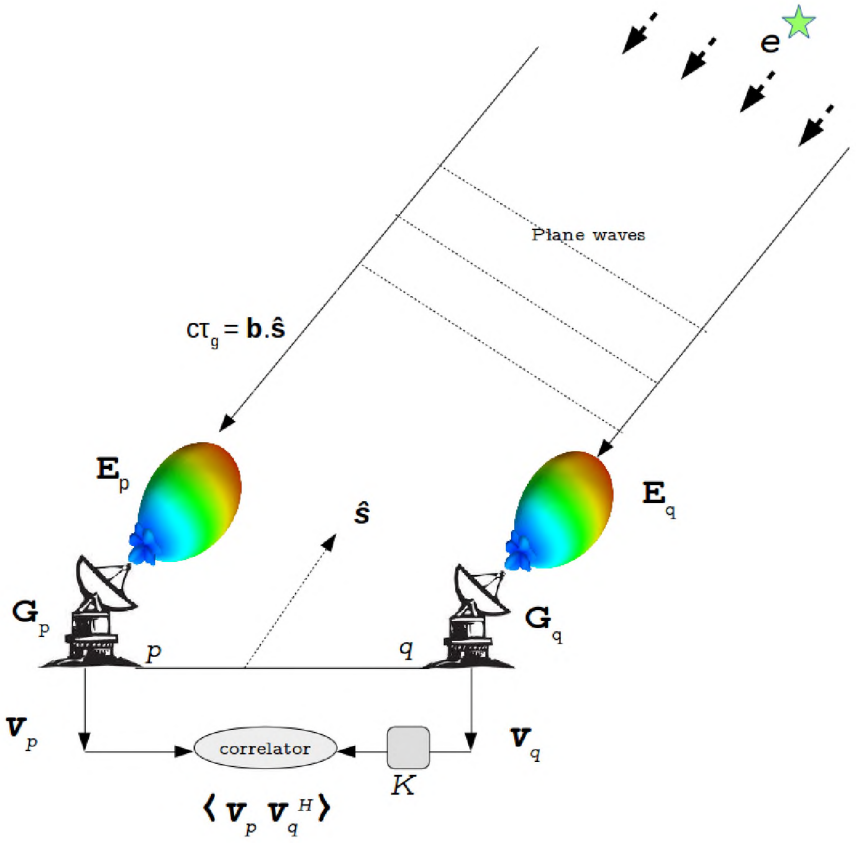


FIGURE 1.2: An interferometer consisting of two antennas, namely antenna p and q. This interferometer is observing a distant astrophysical source (in the far-field) radiating an electric field denoted by a complex vector $\boldsymbol{e}$. $c\tau_g$ is the additional pathlength travelled by signal to antenna p. On its arrival it induces the electrical voltages $\boldsymbol{v}_p$ and $\mathbf{v}_q$ at the antenna feeds. The K, $\boldsymbol{G}$ and $\boldsymbol{E}$ are the phase delay, instrumental gains and primary beam, respectively.

from q to try and match it with the signal from antenna p. If we present the phase delays using a scalar 2×2 *Jones* matrix (Jones, 1941), K, and assume the linear relationship between the phase delays and the electric field (to be explain later), then the induced voltages take the form (Smirnov, 2011a)

$$\boldsymbol{v} = \begin{pmatrix} \boldsymbol{v}_a \\ \boldsymbol{v}_b \end{pmatrix} = K\,\boldsymbol{e}; \quad K = \begin{pmatrix} e^{i\phi} & 0 \\ 0 & e^{i\phi} \end{pmatrix}, \tag{1.16}$$

where the subscript a and b denote feeds polarisation orthogonal components.

This implies the respective voltages in each antenna are $\boldsymbol{v}_p = \begin{pmatrix} \boldsymbol{v}_{pa} \\ \boldsymbol{v}_{pb} \end{pmatrix} = K_p\,\boldsymbol{e}$ and $\boldsymbol{v}_q = \begin{pmatrix} \boldsymbol{v}_{qa} \\ \boldsymbol{v}_{qb} \end{pmatrix} = K_q\,\boldsymbol{e}$, where $K_p = e^{-\imath\omega\tau_g}$ and $K_q = e^{-\imath\omega\tau_i}$. The voltages are then cross-correlated (multiplied and time averaged) to obtain the complex visibilities,

V_{pq} as follows

$$\begin{aligned}\mathsf{V}_{pq} &= \langle \boldsymbol{v}_p\, \boldsymbol{v}_q^H \rangle \\ &= \left\langle K_p\, \boldsymbol{e}\, (K_q \boldsymbol{e})^H \right\rangle \\ &= \left\langle K_p\, (\boldsymbol{e}\, \boldsymbol{e}^H)\, K_q^H \right\rangle \end{aligned} \tag{1.17}$$

where superscript H is a Hermitian transpose. A Hermitian transpose, also known as conjugate transpose, is equivalent to taking the transpose of an m by n matrix consisting of complex entries and then taking its complex conjugate. Thus, an interferometer measures the coherence of the signals (electric field) from two separate antennas (Clark, 1980). Since K is both scalar and diagonal it commutes with everything which means Eq. 1.17 can be written as;

$$\begin{aligned}\mathsf{V}_{pq} &= \left\langle (\boldsymbol{e}\, \boldsymbol{e}^H)\, K_p\, K_q^H \right\rangle \\ &= \left\langle (\boldsymbol{e}\, \boldsymbol{e}^H)\, e^{-\imath\omega(\tau_g - \tau_i)} \right\rangle \\ &= \left\langle (\boldsymbol{e}\, \boldsymbol{e}^H)\, e^{-\imath\omega(\boldsymbol{b}\cdot\hat{s}/c - \tau_i)} \right\rangle, \end{aligned} \tag{1.18}$$

where $\boldsymbol{b}$ is the baseline vector. From here onwards I will be using K as in Eq. 1.17 and not the explicit one in Eq. 1.18. The latter was meant to give a reader a picture of what phase delays are.

Assuming that K is constant across the averaging interval it can be taken out of the brackets in Eq. 1.17, and $\langle \boldsymbol{e}\, \boldsymbol{e}^H \rangle$ is given as (Smirnov, 2011a);

$$\langle \boldsymbol{e}\, \boldsymbol{e}^H \rangle = \begin{pmatrix} \langle e_x e_x^* \rangle & \langle e_y e_x^* \rangle \\ \langle e_x e_y^* \rangle & \langle e_y e_y^* \rangle \end{pmatrix}. \tag{1.19}$$

Assuming linear feeds, Eq. 1.19 can be written in terms of Stokes parameters $(IQUV)$ as follows;

$$\langle \boldsymbol{e}\, \boldsymbol{e}^H \rangle = \begin{pmatrix} I+Q & U+iV \\ U-iV & I-Q \end{pmatrix}. \tag{1.20}$$

where $I = \langle e_x e_x^* \rangle + \langle e_y e_y^* \rangle$ is total intensity, $U = \langle e_x e_y^* \rangle + \langle e_y e_x^* \rangle$ and $Q = \langle e_x e_x^* \rangle + \langle e_y e_y^* \rangle$ specifies linear polarisation, and $V = \langle e_x e_y^* \rangle - \langle e_y e_x^* \rangle$ specifies circular polarisation. This implies that for a telescope with linear feeds, if one measures intensities in U and Q then the incoming electric field is linearly polarised in its origin, or if intensity measurements are found in V then the field is circularly polarised. But since the feeds are not perfect, they tend to experience

polarization leakages whereby the Stokes intensities leak into each other (Asad et al., 2015). However, in this dissertation, we only concern ourselves with Stokes I.

Equation 1.20 is equivalent to the brightness distribution, B (Sault et al., 1996). Thus, Eq. 1.17 becomes;

$$\mathsf{V}_{pq} = K_p \, \mathsf{B} \, K_q^H. \tag{1.21}$$

1.3.2 Radio Interferometer Measurement Equation

The **R**adio **I**nterferometer **M**easurement **E**quation (RIME) is a mathematical formulation used in radio interferometry to model all the distortions endured by a signal throughout its path (Hamaker et al., 1996, Smirnov, 2011a). We briefly describe the RIME with the intention of showing how calibration problems can be addressed using this equation. For an in-depth description and a full derivation refer to Smirnov (2011a) and references therein.

In the previous section, we only considered phase delay effects, but generally, the signal undergoes a series of transformations due to the medium in which it propagates. We can represent all effects along the signal path by the *Jones* matrix $\boldsymbol{J}$ – the phase delays are also contained in this term. Therefore, the visibility equation can be rewritten as follows;

$$\mathsf{V}_{pq} = \boldsymbol{J}_p \, \mathsf{B} \, \boldsymbol{J}_q^H \, , \tag{1.22}$$

where $\boldsymbol{J}_p$ and $\boldsymbol{J}_q$ represents all the propagation effects in the direction of the antenna p and q, respectively.

Note that the linearity assumption between the propagation effects $\boldsymbol{J}$ and $\boldsymbol{e}$ (hence B) is valid only in the case where $\boldsymbol{J}$s are independent of the frequency (Hamaker et al., 1996). Real life situations diverge from this assumption, for example, the phase delays themselves are frequency-dependent. However, we usually subdivide the large frequency band into small bands (known as subbands or coherency bandwidth; Hamaker et al., 1996). The idea behind this is that the bandwidth defining the subband is small enough that all transformations occurring within this band are comparable. This similar approach allows us to drop the time dependence of the electric field and only consider it as electric field vector as in the beginning of this section (Hamaker et al., 1996).

These $\boldsymbol{J}$s can be decomposed into multiple propagation effects as shown below

$$\boldsymbol{J} = \boldsymbol{J}_n \, \boldsymbol{J}_{(n-1)} \ldots \boldsymbol{J}_1, \tag{1.23}$$

and the order, n, corresponds to the order in which they act upon the incoming signal. That is, $\boldsymbol{J}_1$ occurred earlier on the signal path while $\boldsymbol{J}_n$ occurred last. Substituting Eq. 1.27 into Eq. 1.22 we find what is known as the onion form of RIME;

$$\mathsf{V}_{pq} = \boldsymbol{J}_{pn} \ldots \Big(\boldsymbol{J}_{p2} \Big(\boldsymbol{J}_{p1} \, \mathsf{B} \, \boldsymbol{J}^H_{q1} \Big) \boldsymbol{J}^H_{q2} \Big) \ldots \boldsymbol{J}^H_{qm} \,, \tag{1.24}$$

where n and m need not be equal. That is to say, the effects acting on the signal in the direction of antenna p and q can be different (Smirnov, 2011a).

For an observation of a single point source situated at the pointing centre (not at zenith), the effects present are those from the instrument itself, and the phase delays. If we represent the instrumental errors by a Jones matrix $\boldsymbol{G}$ –this term incorporates all the effects associated with the instrument such as electronic gains and bandpass gains, then $\boldsymbol{J}$ for antenna p is

$$\boldsymbol{J}_p = \boldsymbol{G}_p \, K_p. \tag{1.25}$$

Equation 1.22 becomes;

$$\begin{aligned} \mathsf{V}_{pq} &= \boldsymbol{G}_p \left(K_p \, \mathsf{B} \, K^H_q \right) \boldsymbol{G}^H_q \\ &= \boldsymbol{G}_p \, \mathsf{X}_{pq} \, \boldsymbol{G}^H_q \,, \end{aligned} \tag{1.26}$$

where X_{pq} $(= K_p \, \mathsf{B} \, K^H_q)$ is the sky coherency or model visibilities. Now, instead of a single source at the pointing centre, consider N number of sources in the observed field. The signal from each of these sources will arrive at different times at the antennas, thus, making K dependent on the direction of sources. The instrumental effects, on the other hand, are the same for all the sources and are generally referred to as direction-independent effects (DIE). Other source-dependent effects are; the primary beam and the ionosphere. The primary beam determines the sensitivity of the instrument and is directional – the maximum sensitivity is at the beam centre (pointing centre) and decreases with the distance from the centre. For longer observations, the instrument tracks sources across the sky, thus making the beam response time-dependent. The ionosphere becomes significant at low frequencies, wide-field observations, and large arrays. LOFAR is an example of an instrument

that is limited by the ionospheric effects. These direction-dependent effects can be collectively represented by a term $\boldsymbol{E}$, then Eq. 1.25 becomes

$$\boldsymbol{J}_{sp} = \boldsymbol{G}_p K_{sp} \boldsymbol{E}_{sp}\,, \tag{1.27}$$

where s is a source in question. The phase delay is usually treated as a separate entity, thus it is not contained in $\boldsymbol{E}$. Thus, for a field with N sources, the model visibilities are

$$\mathsf{V}_{pq} = \boldsymbol{G}_p \Big(\sum_{s}^{N} \boldsymbol{E}_{sp} \mathsf{X}_{spq}\, \boldsymbol{E}_{sq}^{H} \Big) \boldsymbol{G}_q^{H}, \tag{1.28}$$

where X_{spq} is the source coherency defined as $K_{sp}\mathsf{B}_s K_{sq}^{H}$ (Smirnov, 2011a,b,c). Equation 1.28 can be written in terms of the sky coherency as in Eq. 1.26 except that the sky coherency becomes $\sum_s^N \boldsymbol{E}_{sp}\mathsf{X}_{spq}\,\boldsymbol{E}_{sq}^{H}$, which is the sum of coherencies of N sources (Smirnov, 2011b).

1.3.3 Calibration

Calibration is the process of estimating and correcting for instrumental and atmospheric effects embedded in the signal during its propagation towards the interferometer (Grobler et al., 2014). In RIME terms, calibration is a method of finding $\boldsymbol{G}$ and $\boldsymbol{E}$ that minimises the difference between the measured visibilities and the modelled visibilities. Suppose that the interferometer measures the visibilities V_{pq}^{meas} (Grobler et al., 2014), defined in Eq. 1.29.

$$\mathsf{V}_{pq}^{meas} = \check{\boldsymbol{G}}_p \mathsf{X}_{pq} \check{\boldsymbol{G}}_q^{H} + n_{pq} \tag{1.29}$$

where $\check{\boldsymbol{G}}$s are the unknown gains and n_{pq} is the random complex noise with Gaussian distribution of zero mean and standard deviation (root mean square, rms) σ representing the thermal noise (Cornwell and Fomalont, 1999, Kemball and Martinsek, 2005). We try to formulate the modelled visibilities to correct for the present effects. The instrumental gains are always incorporated into the model but depending on the type of observation, the primary beam and the ionosphere can be ignored or included. For example, for the case of small field of views, the primary beam and ionospheric variations are small enough to be ignored. Suppose this was our case, then our model would take the form;

$$\mathsf{V}_{pq}^{mod} = \boldsymbol{G}_p \mathsf{M}_{pq} \boldsymbol{G}_q^{H}. \tag{1.30}$$

where M_{pq} are the visibilities generated from the sky model (Grobler et al., 2014).

Calibration is a χ^2 minimization process where

$$\begin{aligned}\chi^2 &= \sum_{pq} ||\mathsf{V}_{pq}^{meas} - \mathsf{V}_{pq}^{mod}|| \\ &= \sum_{pq} ||\mathsf{V}_{pq}^{meas} - \boldsymbol{G}_p \mathsf{M}_{pq} \boldsymbol{G}_q^H||. \end{aligned} \tag{1.31}$$

Most often, the estimated gains ($\boldsymbol{G}_p$ and $\boldsymbol{G}_q^H$) are considered good enough when

$$||\mathsf{V}_{pq}^{meas} - \mathsf{V}_{pq}^{mod}|| \sim n_{pq} \tag{1.32}$$

But generally there are systematic errors, ϵ_{pq}, that adds to the noise on the right side of Eq. 1.32 which then restricts the depth of the calibration.

1.3.4 Errors During Calibration

Systematic errors result from the imperfect calibration which is usually due to poor knowledge of the instrument and the observed sky. For example, an observation carried out using different antennas, results in beams which may differ in shape and sizes thus making it difficult to model accurately. Also, observation carried out using large arrays (e.g VLBI, LOFAR with international stations) experiences large clock offsets which may lead to significant phase errors. For the latter case, another problematic effect is the ionosphere which may vary significantly across the field of view, thus presenting a more complex structure to model to detail. The direction-dependent effects are generally poorly known due to their complexity such as beam rotations, time-dependent beam shape and extreme ionospheric fluctuations. The accuracy of the sky model generated is also important. Any omitted or poorly modelled source induces errors during calibration. This includes both over- and under-estimation of flux. These systematic errors manifest themselves as calibration artefacts in images (Smirnov, 2011b,c).

1.3.5 Calibration Artefacts

Calibration artefacts can be categorised into two: (1) those that are due to incomplete sky models and (2) those that due to DDEs or incorrect instrumental

modelling. The former are generally very faint and act as real sources in images, an example of these are ghost sources (see Bos (1985), Grobler et al. (2014)). The latter can be very bright and significant in images, and hence more problematic.

Usually, the goal of calibration is to reach the noise and any source below the noise is ignored. But as described above, calibration artefacts add to the noise, thus, defining a new limit to calibration. For less sensitive instruments dominated by DIEs, thermal noise tends to be higher than calibration artefacts. However, for more sensitive instruments also dominated by DIEs, calibration artefacts tend to emerge above the noise level. In the latter, any of the two can be the limiting factor depending on which one is significant. Moreover, for the deepest observations with an extremely sensitive instrument such as the SKA and its precursors, dominated by DDEs, calibration artefacts are expected to emerge significantly above the noise associating themselves with bright sources in the images and therefore, limiting the sensitivity and the dynamic range of the observation. The dynamic range (DR) is defined as the maximum flux in an image divided by the noise in the image (Wilson et al., 2012). Calibration artefacts increases the value of the noise thus affecting the overall DR.

1.3.6 Calibration Methods

Calibration approaches can be loosely grouped into generations as per Noordam and Smirnov (2010): first (1GC), second (2GC) and third (3GC).

1.3.6.1 First-Generation Calibration (1GC)

First generation calibration uses known calibrator sources. These sources have known fluxes, shapes, positions or spectra and are preferably point sources. Usually, the observations are carried in such a way that the observational time is shared between the target field and calibrator field. The calibrator field is used to derive the calibration solutions. Then these solutions are the transferred to the target field (Wilson et al., 2012). The images obtained after 1GC are generally of low dynamic ranges (DRs) and can be improved using 2GC methods (Perley and Smirnov, 2013).

1.3.6.2 Second-Generation Calibration (2GC)

Traditional *self-calibration* (Cornwell and Wilkinson, 1981), in short *selfcal*, is a second-generation calibration. Selfcal allows one to solve for the antenna gains using a model derived from the target field itself. This model represents a subset of the observed sky which mostly consists of the brightest sources in the field derived from 1GC image or taken from pre-existing catalogues. Selfcal assumes that the primary beam effects $\mathbf{E}$ are constant over time and the same across all antennas in a homogeneous array (Smirnov, 2011b). This implies that the beam can be denoted as scalar E. Thus, the primary beam effects can be corrected for by multiplying the source brightness with the inverse E, which we normally refer to as the beam gain (bg). For Selfcal, Eq. 1.28 becomes;

$$\begin{aligned} \mathsf{V}_{pq}^{mod} &= \boldsymbol{G}_p \left(\sum_s^N E_s \, \mathsf{X}_{spq} \, E_s^H \right) \boldsymbol{G}_q^H \\ &= \boldsymbol{G}_p \left(\sum_s^N E_s^2 \, \mathsf{X}_{spq} \right) \boldsymbol{G}_q^H \\ &= \boldsymbol{G}_p \, \mathsf{X}_{pq} \, \boldsymbol{G}_q^H, \end{aligned} \tag{1.33}$$

where $\mathsf{X}_{pq} = \sum_s^N E_s^2 \, \mathsf{X}_{spq}$.

The following are the steps carried out in *selfcal* (Smirnov, 2011b,c):

1. The initial sky model visibilities, M_{pq}, are derived from the 1GC image (Grobler et al., 2014) or taken from previous observations. This model contains some of the brightest sources in the observed field.

2. Using the above model the instrumental gains, $\boldsymbol{G}$, are estimated through minimization.

3. The estimated gains are applied to the measured visibilities; $\mathsf{V}_{pq}^{corr} = \boldsymbol{G}_p^{-1} \mathsf{V}_{pq}^{meas} \boldsymbol{G}_q^{H-1}$, in order to correct for them. This is followed by the primary beam corrections. The term V_{pq}^{corr} is generally referred to as the corrected visibilities.

4. The corrected residuals are obtained by subtracting the initial model from the corrected visibilities, $|\mathsf{V}_{pq}^{corr} - \mathsf{M}_{pq}|$, and then imaged. It should be noted that the residuals contain poorly and/or unmodelled sources, the noise and calibration artefacts.

5. If the above residuals are unsatisfactory, that is if they are not noise-like, the above steps are repeated. The second sky model is created from the above corrected residuals and updated into the initial sky model. The gains are recomputed using the updated sky model and applied for correction.

1.3.6.3 Third-Generation Calibration (3GC)

Third-generation calibration comprises of new techniques and methods, some still under development, aimed at dealing with DDEs. The 2GC methods fall short for the calibration of the observation subject to DDEs; the reason for this that they assume that these effects are constant over time and across all receivers, which is not true particularly for modern instruments. Several 3GC methods already exist (e.g peeling, FACET, SPAM, Sagecal; Noordam, 2004, van Weeren et al., 2016, Intema et al., 2009, Kazemi et al., 2011), but of particular interest to us is the differential gains technique implemented `MeqTrees` package (Noordam and Smirnov, 2010) and first tested on WSRT data presented in Smirnov (2011c). This method corrects for DDEs through Eq. 1.34.

$$\mathrm{V}_{pq} = \mathbf{G}_p \left(\sum_s^N \Delta\mathbf{E}_{sp}\, \mathrm{X}_{spq}\, \Delta\mathbf{E}_{sq}^H \right) \mathbf{G}_q^H, \tag{1.34}$$

where $\Delta\mathbf{E}_{sp}$ represents the differential gains (`dE`). Differential gain solutions are estimated and applied to sources subjected to DDEs in an image. Presently, these sources are identified manually by looking at an image and deciding whether they require these solutions based on the level of artefacts around them. The advantage of this approach is that it requires no physical information about the sources. According to Eq. 1.34 the `dE` are responsible for correcting all the DDEs but this leads to extremely large degrees of freedom making this process computationally expensive. However, if we have parametric model of the primary beam known prior then we can rewrite Eq. 1.34 to be Eq. 1.35 so that `dE` are only responsible for correcting other DDEs except the nominal beam gain. This reduces the amount of computation required to solve for `dE` solutions and also reduces source suppression (Smirnov, 2011b,c).

$$\mathrm{V}_{pq} = \mathbf{G}_p \left(\sum_s^N \Delta\mathbf{E}_{sp}\, E_s\, \mathrm{X}_{spq}\, E_s^q\, \Delta\mathbf{E}_{sq}^H \right) \mathbf{G}_q^H. \tag{1.35}$$

1.3.7 Imaging

Imaging is a process of determining intensity distribution of the sources from the measured visibilities. This is achieved by taking the inverse Fourier transformation of the visibilities as follows;

$$\mathsf{B}(l,m) = \mathrm{FT}[\mathsf{V}(u,v)] \tag{1.36}$$

where l, m are coordinates in the tangent plane of the sky and u, v are the coordinates of the projected baseline of the interferometer expressed in units of wavelength. Each baseline vector u, v corresponds to a single component of the source brightness distribution. As the Earth rotates this baseline vector traces out an arc with each point giving a sample of the measured visibility. The superposition of these arcs (or ellipses) results into what is called u, v coverage of an instrument (Cornwell, 2008, Wilson et al., 2012). The true brightness distribution of source can be recovered if the u, v coverage is completely filled with data points. However, in real life, only discrete data points depending on the array configuration are sampled. This is a result of the finite number of antennas and time available for the observation. Thus the measured visibility is merely the sample of the true visibility V_T

$$\mathsf{V}_M(u,v) = \mathsf{V}_T(u,v)\mathsf{S}(u,v) \tag{1.37}$$

where $\mathsf{S}(u,v)$ represents the sampling function of the u, v. Equation 1.36 then becomes;

$$\mathsf{B}_D(l,m) = \mathrm{FT}[\mathsf{V}_T(u,v)\mathsf{S}(u,v)] \tag{1.38}$$

where $\mathsf{B}_D(l,m)$ is the dirty image. The convolution theorem states that the multiplication of Fourier transforms of two functions in one domain is equal to the convolution of these functions defined in the inverse domain [2];

$$\mathsf{B}_D(l,m) = \mathsf{B}(l,m) * \mathsf{P}(l,m) \tag{1.39}$$

where $\mathsf{P}(l,m) = \mathrm{FT}[\mathsf{S}(u,v)]$ is the dirty beam or the point spread function (PSF) of the instrument. Therefore, to obtain the true source brightness, the dirty beam must be deconvolved from the dirty image. The dirty image is dominated by the PSF sidelobes. These sidelobes result from the gaps in the u,v coverage and the fact that the u, v coverage has a finite extent (Briggs et al., 1999). Deconvolution

[2] `http://www.cv.nrao.edu/course/astr534/FourierTransforms.html`

is technique used to remove these sidelobes (i.e remove the PSF signature) from the resulting image. Popular algorithms that carry out deconvolution are `Högbom` (Högbom, 1974), `Clark` (Clark, 1980) and `Cotton-Schwab` (Schwab, 1984), collectively known as `CLEAN` algorithms. These are best suited for point sources but are less optimum for extended structures in an image. The deconvolution algorithms made to take care of extended sources are `Multi-scale CLEAN` (Cornwell, 2008) and `Maximum Entropy Method` (Weir, 1992, Bontekoe et al., 1994).

1.4 Source Finding Algorithms

Source finding algorithms find and group pixels in an image that are considered to belong to astronomical objects. Most source finders follow some of the following steps: (*i*) background estimation and subtraction, (*ii*) source identification, (*iii*) source characterization and measurement and (*iv*) cataloguing (Hancock et al., 2012, Huynh et al., 2012).

1.4.1 Background Estimation

Any flux contribution in an image which is not astrophysical in origin is considered as background noise. The estimation of this flux is known as background estimation. This task is easier to carry out for a background that comprises of only thermal noise in which a single value can be used to characterise the noise level in the image (Masias et al., 2012). However, this process becomes complex and difficult for backgrounds with large noise variations, artefact structures or classical confusion (Masias et al., 2012). The latter is attributed to deeper observations with a resolution too poor to separate the individual sources thus causing less noisy free space between sources (Wilson et al., 2012). The large noise variations are normally addressed by estimating the background in sub-regions of the image and using the obtained estimations to derive the single optimum background estimate (Hancock et al., 2012, Huynh et al., 2012). Whilst in the presence of artefact structures, the image is first reprocessed using a filter in order to remove these structures before making an estimate (Hancock et al., 2012, Masias et al., 2012).

1.4.2 Source Identification and Characterization

Source identification can be done using two approaches: (*i*) thresholding (commonly used) and (*ii*) the local peak search (not discussed) (Masias et al., 2012). Thresholding is a process that involves grouping neighbouring pixels with intensity, I, above $I_{th} = c\sigma_{background}$ into what is called *islands*, where $c \in \mathbb{N}$ (Hancock et al., 2012, Masias et al., 2012). The local peak search on the other hand, searches for connected pixels with peak values above some I_{peak} and is known to work well for point sources and poorer for complex sources (see Masias et al. (2012) and references therein for further description). The former process is followed by source characterization in which each island is fitted with an appropriate model, for example the elliptical or circular Gaussian are generally used for compact sources. A given island may consist of more than one Gaussian (or model) depending on the threshold used for model fitting (Hancock et al., 2012). Nevertheless, sources are formed by grouping the Gaussian components that fall within an island. This can be further constrained by grouping only those that fall within the size of the point spread function (PSF) of the observation. This is particularly important when bigger islands are formed. For example, the `PyBDSM` (Mohan and Rafferty, 2015) source finder only considers a group of Gaussians to correspond to a single source if no pixel between a pair of Gaussians has a value less than the island threshold, and if the distance between the centre of the Gaussians is less than half the sum of their widths (full width half maximums). At this stage source properties i.e fitted parameters such as total flux, sizes, peak flux, and errors in each measurement are catalogued (Hancock et al., 2012).

1.4.3 Threshold Setting

The purpose of using a threshold is to separate pixels that belong to astronomical objects from those that belong to the background (Hancock et al., 2012, Hopkins et al., 2002, Masias et al., 2012). Background estimation is thus crucial in setting a source finding threshold. The threshold is generally set to be a factor above background noise. For example, `PyBDSM` (Mohan and Rafferty, 2015) performs source identification based on two thresholds; the threshold for forming islands, `thresh_isl`, and the threshold for a model fitting, `thresh_pix`. An island is formed in regions with flux $>$ `mean` $+$ `thresh_isl` $\times$ `rms` and model fitting is

done in islands with peak flux > `mean` + `thresh_pix` × `rms`, where the `rms` is the estimated noise of the image (Mohan and Rafferty, 2015).

There is, however, a statistical way of estimating the threshold using the False Detection Rate (FDR). It was first implemented in a source-finding task by Hopkins et al. (2002). False detection rate evaluates the intensity distribution of all pixels in an image and compares these to an image containing only noise (Hopkins et al., 2002). This method sets a threshold that limits the false detections depending on the statistics of the noise distribution (Huynh et al., 2012, Miller et al., 2001).

1.4.4 Completeness and Reliability

The performance of a source finder is determined by its *completeness* and *reliability* (Hopkins et al., 2015). Completeness is the fraction of sources in an image detected by the source finder tool and is defined by Eq. 1.40.

$$C = \frac{n_r}{n} \tag{1.40}$$

where n_r is number of real sources and n is the total number of sources present in the image (Westerlund et al., 2012).

The reliability refers to the fraction of sources detected by the source finder that are actually real and is given by Eq. 1.41.

$$R = \frac{n_r}{n_t} \tag{1.41}$$

where n_t is the total number of detections both real and spurious (false positives) (Hopkins et al., 2015, Westerlund et al., 2012). The *reliability* can be used in two different ways depending on the context. First, it can be used as a single value to characterise the entire source catalogue (Serra et al., 2012) as defined in Eq. 1.41. It can also be used to characterise the individual sources in the catalogue as in Serra et al. (2012). The latter reliability is obtained from density estimations as will be discussed in Chapter 2.

Ideally, a good source finder should have a high completeness and reliability. However, in reality, there is usually a trade-off between these two quantities. That is, one may choose to detect more sources (real emission) in the image while introducing more background noise (false detections), or one may choose to avoid the

background noise at the expense of overlooking other real emission. A review of the performance of the existing source finders is presented in Hancock et al. (2012), Hopkins et al. (2015), Huynh et al. (2012).

1.5 Problem Statement

Calibration artefacts limit the depth of the source catalogues and images. These artefacts are in some cases brighter than some of the astrophysical emission in an image and thus may be classified incorrectly as real emission during source extraction. The optimum source finding threshold which avoids false emission is set manually. Most of the time this threshold is very high and causes some of the real emission to be overlooked. This leads to shallow and incomplete sky models which impact on the calibration gain estimation. This is typically addressed by iteratively improving the sky model through successive applications of calibration, imaging and source finding i.e multiple *selfcal* iterations. However, this is both computationally expensive and time-consuming.

A second problem is to determine which sources are subject to severe DD effects without manual inspection of the image. These sources need to be identified robustly so that one can perform DD calibration on them.

1.6 Previous Work

There has been work done in distinguishing astrophysical emission from spurious emission. Serra et al. (2012) used the negative detection method for spectral line observations. This work showed that spurious emission, in particular the noise artefacts can be reliably excluded from source catalogues by studying their distributions in a well-known parameter space. The parameters used were the peak flux, integrated flux and the number of voxels. This algorithm is implemented in a source finding tool called `SoFiA` (Serra et al., 2015). They tested this method on an image cube containing noise and 137 neutral hydrogen (HI) input sources and found a total of 303 detections, and out of these 41 were selected as reliable sources. However, 40 of the detections were classified accurately as true emission while a single noise artefact was misidentified as true. Further, they implemented

their method in an image cube containing radio frequency interference (RFI) and found promising results. Thus, they concluded that their method may be able to work for extreme cases consisting of imaging or calibration artefacts and also that the classification can be improved by increasing the parameter space.

1.7 Objectives of the Dissertation

Given the challenges and practical limitations to robust source-artefact discrimination, we aim to develop an algorithm that:

- distinguishes astrophysical sources from the spurious emission in an image, and
- identifies sources that require DD calibration.

An additional goal is to obtain deeper and high fidelity sky models in the initial *selfcal* iterations. That is to say that the sky models must contain a large subset of the actual sources (high completeness) without any false emission (high reliability).

We use a similar approach as Serra et al. (2012) to distinguish spurious emission from astrophysical emission. Following the recommendations in Serra's paper, we extend his work to distinguish astrophysical emission from calibration artefacts in an increased parameter space. However, our algorithm is specifically developed for continuum observations.

Chapter 2

Source-Artefact Discrimination

From the previous chapter, we have learnt about the problem of calibration artefacts in radio images and how they may limit the dynamic range of the modern instruments. This chapter describes the algorithm used to distinguish astrophysical sources from calibration artefacts, as well as its application to simulated data and a JVLA observation. Note that the figures used from section 2.1 to section 2.4.2 are generic examples to illustrate the key concepts of the algorithm. It should also be noted that the algorithm at the time of writing was developed and tested in Stokes I (total intensity) images, and therefore, the images presented in this work are mainly in this intensity. As we mentioned before, this work is focused on continuum data, specifically continuum survey fields (its applicability may extend beyond this) whereby a variety of source morphologies (resolved and unresolved) are present. An unresolved source is a source with an angular size larger than the instrument's angular resolution, and a resolved source has the angular size smaller than the angular resolution. For this work, one does not assume any particular source morphology, thus, the image can comprise of unresolved (point) and resolved (extended) sources.

2.1 Source Extraction

Before going into the details of the algorithm, there are two key terms that we need to explain. These are the positive and negative detections. Figure 2.1 shows an example of an interferometric image. This image is a cut-out from a larger map depicting an off-axis source that is subject to severe DDEs due to the primary beam

rotation. The left side shows the original (or positive) image (restored image) obtained directly from `CLEAN` [1] and the left image is the inverted (or negative) image obtained by multiplying the original image by -1. Source extraction is performed on each of these images independently using the available source finding tools. We refer to the detections obtained from the positive image as *positive detections* and those from the negative image as *negative detections*.

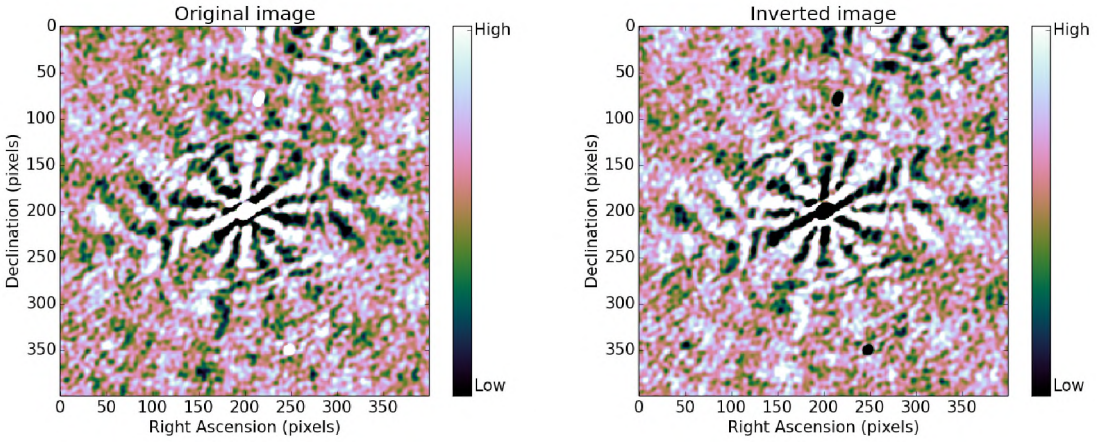


FIGURE 2.1: *Left panel*: the original image obtained using the `CLEAN` algorithm. *Right panel*: the inverted image obtained by multiplying the original image by -1. These images are represented in total intensity (Stokes I).

2.2 Negative Detection Method

We developed an algorithm that implements the negative detection method (NDM) to classify *positive detections* into astrophysical emission and artefact emission. This method hinges on the following assumptions as per Serra et al. (2012):

1. Astrophysical emission has positive total (integrated) flux.
2. There exists a parameter space in which astrophysical emission and artefact emission occupy distinct regions.
3. Instrumental noise is symmetric about zero in the image plane.

[1] The restored image is obtained after deconvolving the dirty beam from the true brightnesses, and adding the resulting point source components reconvolved with an ideal PSF (clean beam) to the residual noise (Högbom, 1974).

For our case, we assume that calibration artefacts are quasi-symmetric about zero in an image plane. The reason for this is that calibration artefacts are generally coupled to the level of the PSF sidelobes. These sidelobes have negative/positive vaguely symmetric structure. The new generation of radio telescopes have very good u, v coverages, even for snapshot observations, which results in lower level of sidelobes, and therefore artefacts. But even though this is the case, the sensitivities of the these instruments make the artefacts significant above the noise. Unlike the noise peaks, calibration artefacts are elongated, and are never isolated, but are always associated with strong sources in the image. The latter is the case because DDEs are multiplicative in the image plane thus, prominent around strong sources (Smirnov, 2011b).

2.2.1 Visualisation of the Assumptions

The purpose of this section is to provide supporting figures for the above assumptions. The detections used were extracted from the JVLA simulated image (see section 2.5 for details on simulations).

Figure 2.2 shows an example of the distribution of the *positive detections* in the parameter space defined by three different source parameters (or features); X, Y and Z. For clarification purposes, note that X, Y, Z refers to some abstract example source parameters and not conventional spatial coordinates. Specifically, X, Y and Z are the local variance, peak flux and total flux (see section 2.4.2) respectively. Since this is simulated data, we have good knowledge of what is real (shown in blue) and what is false (shown in red).
From these plots, it can be seen that the distribution of the *positive detections* can be categorized into two distinct classes of distributions; a class that describes the spurious emission and another that describes astrophysical emission. The separation is more apparent in the parameter spaces defined by X & Y and X & Z. The parameter space defined by Y and Z shows one of the challenges of classification when the distributions are not distinct across the parameter space. The key point is that for some well-chosen features, it is possible to separate the astrophysical emission from the artefact emission.

Figure 2.3 shows the distribution of *positive detections* and *negative detections*. The *negative detections* appear to occupy a more similar parameter space as the

spurious emission from the positive image. This shows that the distribution of the spurious emission in the positive and negative image is quasi-symmetric.

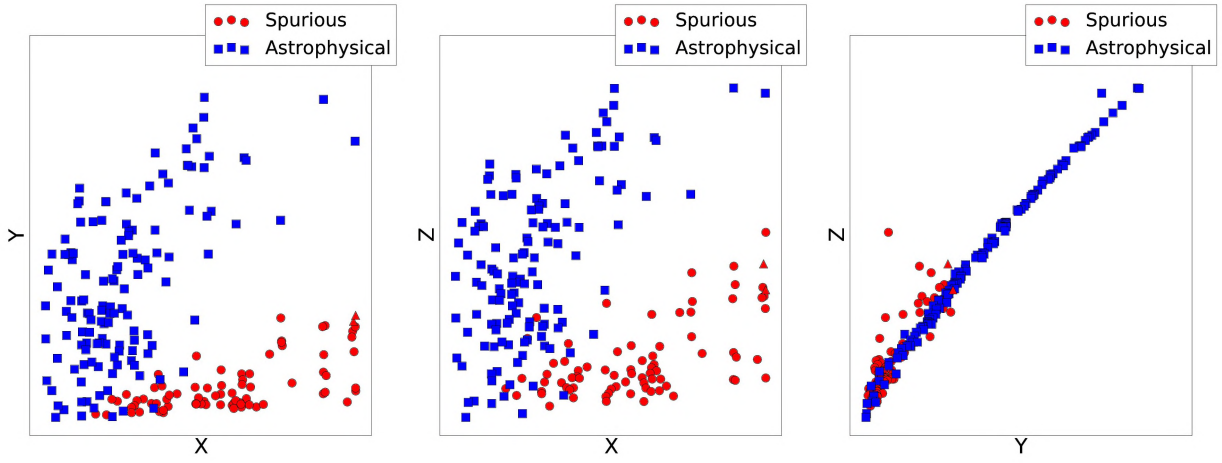


FIGURE 2.2: The distribution of positive detections. These detections were obtained from the JVLA simulated image. It can be seen that the distributions of the astrophysical emission and spurious emission form two distinct classes. The distinction is more apparent in the parameter spaces defined by X & Z and X & Y. The parameter space defined by Y and Z shows that in some cases the distributions may become indistinct.

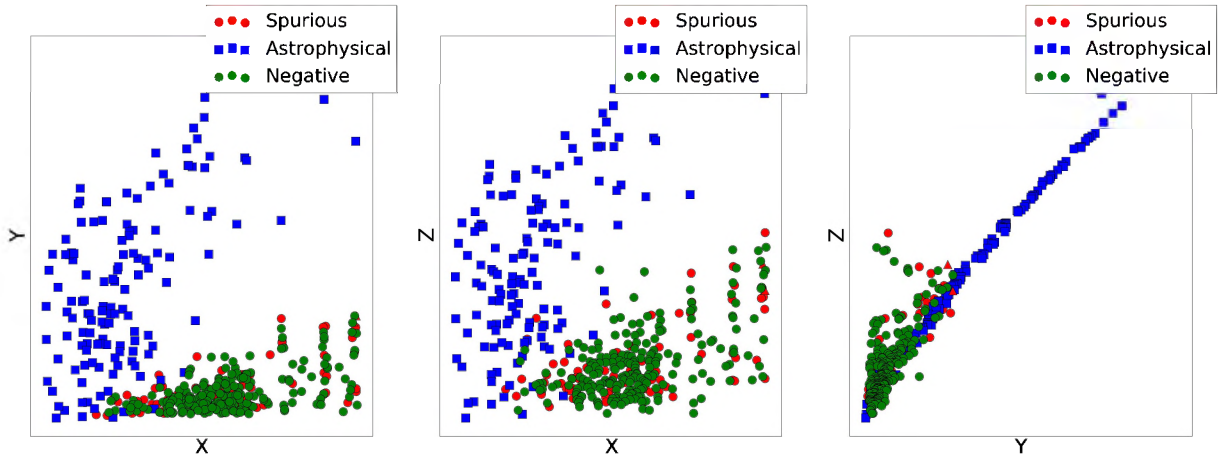


FIGURE 2.3: The distribution of *positive detections* and *negative detections*. The negative detections occupy similar regions of the parameter space as the spurious emission.

2.3 Density Estimations

From the stated assumptions, we let astrophysical emission and spurious emission be defined by Class 1 (C_1) and Class 2 (C_2), respectively. Our goal is to use

probability measures to quantify and hence classify positive detections as either belonging to C_1 or C_2. This will be achieved through density estimations. There are parametric and non-parametric ways of estimating density functions. Parametric approaches assume the data to be drawn from some specific distribution and estimate the defining parameters of this distribution using the data. For example, if a normal distribution is assumed, the main task will be to estimate the mean and variance of this distribution from the data. Although this may simplify the problem and be less computationally intensive, this approach poses a subjective model and therefore not optimum for complicated data structures (Duong, 2004). Non-parametric, on the other hand, allows the data itself to decide on the underlying distribution. Examples of non-parametric density estimators are the histogram and kernel density estimation (KDE, Duong, 2004). In this work, we are using KDE to construct the density functions.

2.3.1 Kernel Density Estimation

Given a random sample $\{X_1, X_2, ..., X_n\}$ drawn from a distribution with an unknown density $f(x)$, the kernel estimator is defined as

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big), \tag{2.1}$$

where n is sample size, $K(\cdot)$ is a kernel function (also known as the probability density function) which satisfies the following conditions

$$\int_{\infty}^{\infty} K(x)dx = 1 \tag{2.2}$$

and $K(x) \geq 0$ for all $x \in \mathbb{R}$, and h is a bandwidth also known as the smoothing parameter which determines the width of the kernel (Duong, 2004, Hasen, 2009).

Figure 2.4 shows a univariate kernel density estimate obtained from the data, $\{-1, -0.8, -0.6, 0.5, 1.2\}$, using a Gaussian kernel (see Eq. 2.3) with a bandwidth of 0.3517 (Duong, 2004).

$$F(x, h) = \frac{1}{\sqrt{2\pi h}} e^{-\frac{1}{2}\left(\frac{x - X_i}{h}\right)^2} \tag{2.3}$$

where $i = 1, ..., n$ and X_i is the mean (data point).

The kernel $n^{-1}h^{-1}K\left(\frac{x-X_i}{h}\right)$ is centred and evaluated at each data point (X_i) shown by the Gaussian dotted lines. The final kernel density estimator is obtained by summing the contributions from each data point and is represented with a solid line. This implies that the estimate, $\hat{f}(x)$, becomes larger if observations are closer to X and is smaller if there are fewer observations (Hasen, 2009).

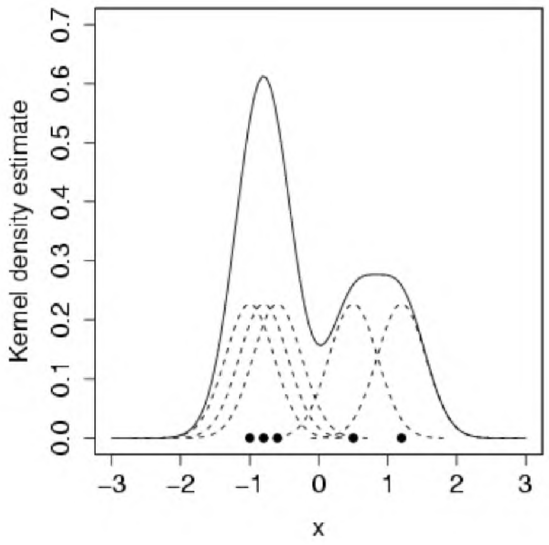


FIGURE 2.4: Univariate kernel density estimation. The dotted normal lines are the individual kernels centred at each data point. The solid line is the density estimate obtained by taking a sum of the dotted lines kernels. Image Credit Tarn Duong Duong (2004)

For a general case of a multi-dimensional data $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{id})^T$; $i \in \mathbb{N}$, the kernel density estimator is defined as

$$\hat{f}(\boldsymbol{x}) = \frac{1}{n|\boldsymbol{H}|^{\frac{1}{2}}} \sum_{i=1}^{n} K\left(\frac{\boldsymbol{x} - \boldsymbol{X}_i}{\boldsymbol{H}^{\frac{1}{2}}}\right), \tag{2.4}$$

where $K(\cdot)$ is the d-dimensional kernel function, $\boldsymbol{H}$ is a $d \times d$ bandwidth matrix and $\boldsymbol{x} \in \mathbb{R}^d$ (Duong, 2004).

Figure 2.5 illustrates the kernel density estimator of a 2-dimensional data using a 2-dimension Gaussian kernel. The left diagram depicts the individual kernels centred about each data point and the right depicts the estimated density obtained by summing up kernel contributions. Kernel density estimation works well for data with dimension less than 6. Higher dimension causes instabilities in the estimate (Duong, 2004, Scott, 1992).

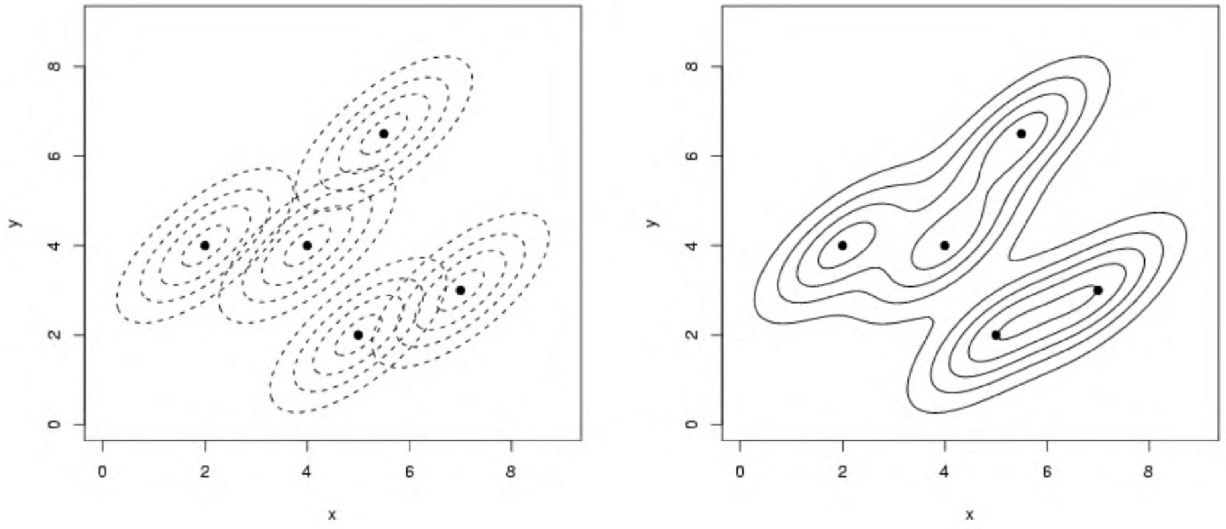


FIGURE 2.5: Bivariate kernel density estimation. *Left* are the kernels centred at each data point. *Right* is the density estimate. Image Credit Tarn Duong Duong (2004)

2.3.1.1 Bandwidth Selection

The resulting kernel estimator is highly dependent on the choice of the smoothing parameter $\boldsymbol{H}$ (Duong, 2004, Hasen, 2009). A broad $\boldsymbol{H}$ results in over-smoothing and hides the important features within the data (Hasen, 2009). This increases the estimation bias. Meanwhile, a narrower $\boldsymbol{H}$ under-smooth the data and thus, increases the estimation variance (Hasen, 2009). A bandwidth can be chosen in such a way as to minimise the error between estimation bias and estimation variance. The measure of the error is referred to as the mean squared error (MSE), defined as (Hasen, 2009):

$$\begin{aligned} \text{MSE} &= E\Big(\hat{f}(x) - f(x)\Big)^2 \\ &= \text{bias}(\hat{f}(x))^2 + \text{Var}(\hat{f}(x)), \end{aligned} \tag{2.5}$$

where E is the expectation value, estimation bias is defined as $\text{bias}(\hat{f}(x)) = E\hat{f}(x) - f(x)$ and estimation variance is $\text{Var} = E[\hat{f}(x) - E\hat{f}(x)]$ (Duong, 2004). Scott (1992), Bowman (1997) showed that for a Gaussian kernel with normally distributed data, the optimal bandwidth that minimizes the MSE is

$$h_j = \sigma_j \Big(\frac{4}{(d+2)\,m}\Big)^{1/(d+4)}, \tag{2.6}$$

where $j = 1, 2, ..., d$; $h_j \in \boldsymbol{H}$, $\boldsymbol{H}$ is a diagonal matrix with diagonal entries h_j and σ_j is the standard deviation of jth random variable (Zhang et al., 2004). Figure 2.6 shows a univariate example of over- and under- smoothing, and bandwidth derived from MSE.

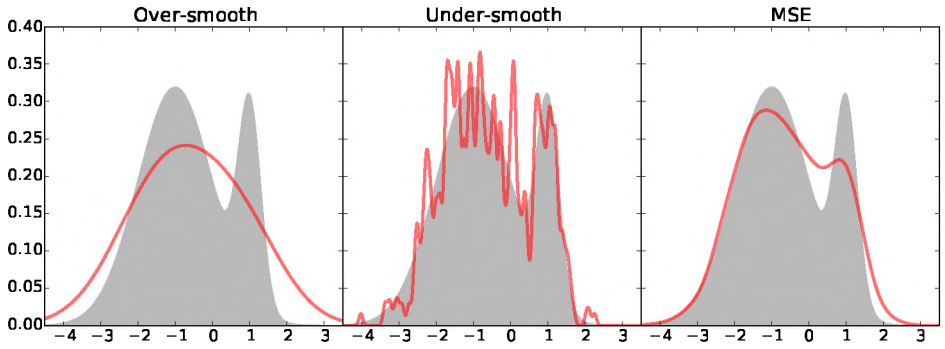


FIGURE 2.6: Plots derived using three different bandwidths. *Left* shows the over-smoothing ($h = 1.0$), *middle* plot is the under-smoothing ($h = 0.05$) and *right* is the plot for the bandwidth derived using MSE approximation.

2.3.1.2 Density Function Implementation and Reliability Estimations

We used Gaussian kernels to estimate the densities for the positive and negative detections with bandwidth entries derived from Eq. 2.6. Figure 2.7 is an example of the estimated density for the positive detections $\hat{f}_P(\boldsymbol{x})$ (red contours) and negative detections $\hat{f}_N(\boldsymbol{x})$ (blue contours) where $\boldsymbol{x} \in \mathbb{R}^2$. The plotted data points represent the positive detections. Note that from now onwards I will be referring to random variables $\boldsymbol{X}$ as features or parameters. Finally, our goal is to evaluate the probability of each data point in each density field. For example, if we consider a data point described by features (x, y) we compute $\hat{f}_P(x, y)$ and $\hat{f}_N(x, y)$ which is the probability evaluated in the positive and negative density field, respectively. Depending on the resulting probability value we can thereafter classify it as either belong to C_1 or C_2. For instance, since the density $\hat{f}_P$ consists of contributions from astrophysical and artefact emission thus, if $\hat{f}_P(x, y) \sim \hat{f}_N(x, y)$ then the detection in question is likely to be an artefact.

To properly quantify the above probabilities, we need to introduce a measure known as the reliability (Serra et al., 2012), which is distinct from the R parameter in Eq. 1.41. If you recall, R was initially defined as a reliability that categorises the entire catalogue but in this case the reliability will be used to quantify each

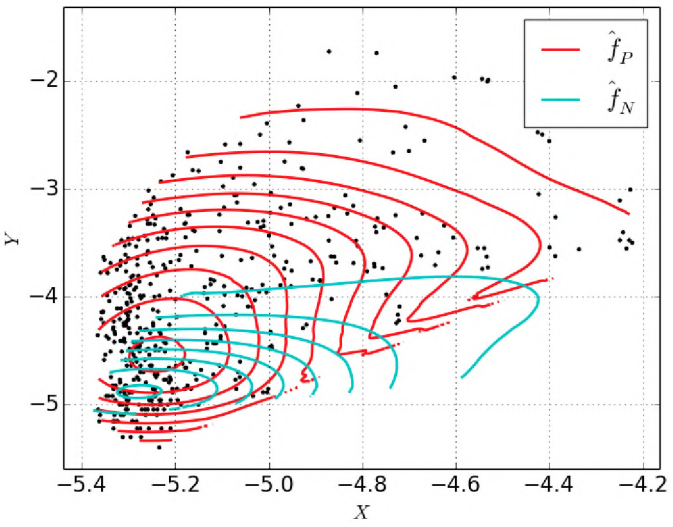


FIGURE 2.7: The density field estimated using Kernel Density Estimation (KDE). The $\hat{f}_P$ represents the density of the positive detections and $\hat{f}_N$ is the negative density field estimated using the negative detections. The dotted points represent the positive detections. This parameter space is defined by the example parameters X and Y.

detection in a catalogue. The reliability of a positive detection with features $(x, y, ...)$ is defined as

$$R = \frac{\hat{f}_P(x, y, ...) - \hat{f}_N(x, y, ...)}{\hat{f}_P(x, y, ...)}. \tag{2.7}$$

Let P be a number density given as $P = \hat{f}_P(x, y, ...)/n$ and $N = \hat{f}_N(x, y)/m$, where n and m are length of data points of the positive and negative detections respectively. The reliability equation (Eq. 2.8) becomes

$$R = \frac{P - N}{P}. \tag{2.8}$$

For the earlier example, when we had $\hat{f}_P(x, y) \sim \hat{f}_N(x, y)$, this implied that $R \sim 0$, while $\hat{f}_P(x, y) > \hat{f}_N(x, y)$ implies $R > 0$.

2.4 The Algorithm

We developed a software tool called `sourcery` which computes these reliability estimates. This tool mainly adds some measurement of fidelity to the sources

extracted by the source finder. The implementation of `sourcery` is described in Appendix A. This tool implements the following algorithm;

1. Takes in the image in FITS format.
2. Smooths the image.
3. Extracts sources from the image.
4. Optional: computes extra source parameters and assigns them to each detection.
5. Computes the reliabilities.
6. Selects sources that require DD calibration solutions (described in chapter 3).

2.4.1 Smoothing

A key motivation for smoothing in `sourcery` is to de-noise the data. This is essentially filtering out the noise or large background variations, thus giving emphasis to astrophysical emission as well as some of the spurious emission (calibration artefacts).

We smooth the data using a Gaussian kernel at different scales in order to compensate for differing structures in the image. The small scales emphasise point sources and large scales emphasise extended emission. Currently, the employed scales are [0.1b, 2b, 5b, 10b, 20b, 40b] where b is the PSF size. We smooth the data at each kernel size and mask the pixels below a user specified threshold in each smoothed image. The masks are then added up to form a single mask image. The resultant mask is then applied to the actual image. Figure 2.8, 2.9 and 2.10 show the unmasked (original image) and masked data at different thresholds to illustrate the above approach. Note that a masked image is used in `sourcery` during source finding to form islands while the actual image is used to do the model fitting. This is so that we avoid incorrect parameter estimations due to smoothing and masking.

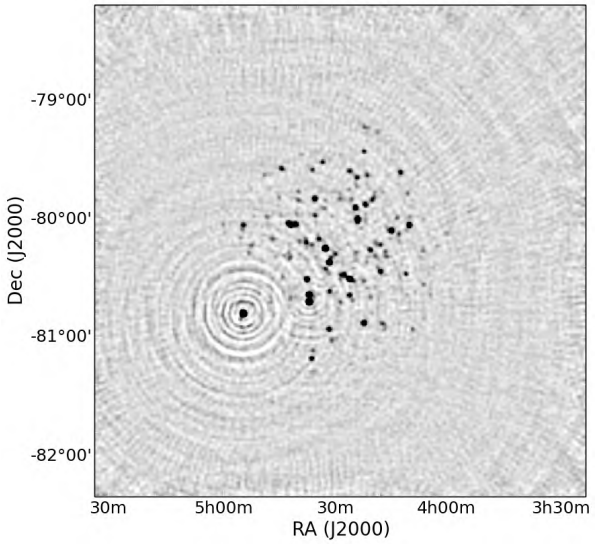


FIGURE 2.8: KAT-7 image of observed data.

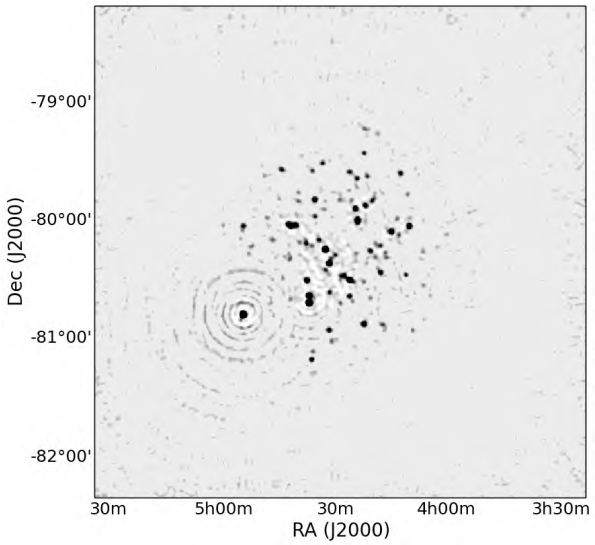


FIGURE 2.9: KAT-7 as above. Smoothing was performed and pixels with values less than $1.6 \times \sigma_{rms}$ (image noise) were masked.

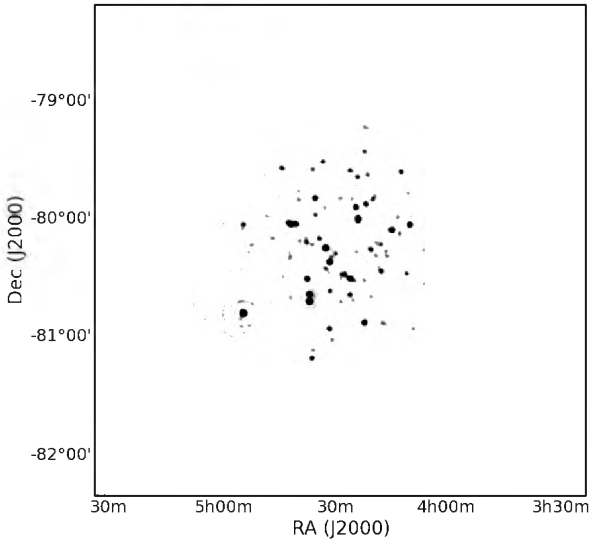


FIGURE 2.10: KAT-7 as above. Masking pixels with values less than $5 \times \sigma_{rms}$ (image noise).

2.4.2 `sourcery` Source Parameters

The following source parameters are used in `sourcery` to estimate the density and then do the classification: source area, source integrated flux, source peak flux, correlation factor, local variance and number of nearby sources.

1. The source area is computed from the major and minor axis derived by a source finder. For point sources, these axis will be zero thus, since we are working in logarithmic space, we consider the PSF size as the minimum instead. The source area is computed in units of arcseconds.

2. The integrated flux and peak flux are taken as they are from source catalogue. These are in units of Jansky (Jy).

3. The local variance of a source is obtained by taking the standard deviation of a subset of pixels centred about the source in question. The size of the subset of pixels is chosen in terms of the PSF sizes, where the default value is 10 PSFs. The choice is arbitrary. The local variance of a source surrounded by calibration artefacts should, in general, be larger than that of a source in a noise limited region. Another factor that results in high local variance is that of poorly modelled sources. To avoid biases from nearby sources or the smearing of poorly modelled sources, we compute the local variance using pixels in the inverted image only.

4. Correlation gives a measure of how two quantitative variables vary together (Gertman, Retrieved on 12 December 2015). The correlation factor is obtained by correlating a given portion of the Point Spread Function (PSF) image with the local region centred on the source. Note that this portion must be a good representation of the PSF. Thus, correlating a main-lobe down to at least 5 side-lobes seemed reasonable. However, this is a user-controllable option with the default value 5.

5. The last parameter is the number of neighbouring sources (`Source Near`). This is obtained by counting the number of sources in a circle centred at a source and having a radius of $c \times b$, where b is the PSF size and c is an optional parameter with default value 5. Sources situated in a vicinity dominated by calibration artefacts should have a larger number of neighbours (detections) while those that are situated in a noise-limited regions This parameter contributions is as that of the local variance.

Turning back to the PSF correlation, since this parameter is not as intuitive, we further investigated how the PSF correlates with sources of different sizes. We simulated a JVLA observation with 11 sources of equal brightness (1Jy) but different sizes ranging from 4.5 arc-seconds to 50 arcseconds and thermal noise. Figure 2.11 shows the PSF correlation as a function of source size. From this plot, it can be observed that the correlation is higher for sources with similar sizes to the PSF and decreases for highly extended sources. In addition to this, we evaluated how the PSF correlates with calibration artefacts and the noise. We used the detections from the $3C147$ image (see section 2.8) to observe this. We found that the noise correlates well with the PSF, similarly to extended sources. Calibration artefacts, on the other hand, correlate poorly with the PSF with values of order 10^{-4}. Thus, the correlation is expected to be optimum for the case of calibration artefacts and astrophysical emission. Thus, we can then assume that any detection with higher R but poor correlation can be excluded. We will investigate the latter at the result sections.

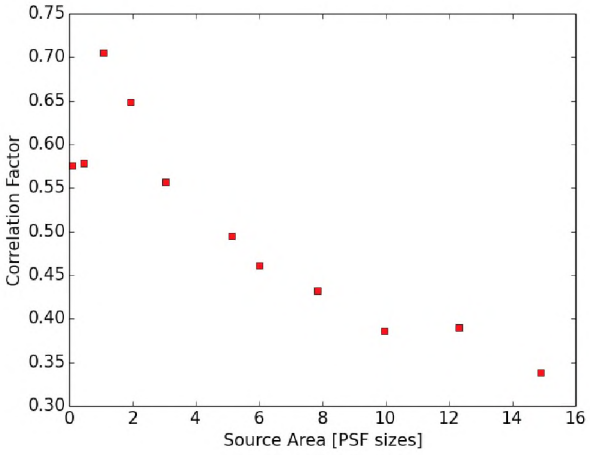


FIGURE 2.11: Correlation of the PSF and simulated sources. These are taken from a noise + sources only JVLA simulation. The sources have the brightness of 1Jy and different sizes ranging from $5'' \times 4.5''$ to $50'' \times 45''$.

2.5 The Experimental Setup

The following sections describe the simulations that were carried out in order to test the performance of the above algorithm.

2.5.1 Telescope Configurations

We simulated a JVLA observation using the *C*-configuration which consists of 26 antennas and a maximum baseline of 3.4km. The configuration for the observation is shown in Table 2.1.

TABLE 2.1: Simulation Configurations.

Parameter	Value
Central frequency	1.4 GHz
Frequency band	L-band
Channel width	3 MHz
Number of Channels	40
Total observational bandwidth	120 MHz
Field of view	2 degrees
Phase centre coordinates	Right ascension (RA) = 85 degrees Declination (Dec) = 49.5 degrees
Integration time	5 seconds
Synthesis time	2 hours

2.5.2 Sky Models

We generated artificial sources with random positions and sky brightnesses. The positions were uniformly random and fluxes followed a power-law distribution as shown in Fig. 2.12. The radio astronomical sources are known to have flux densities and positions that follow the aforementioned distributions (Felli and Spencer, 1989).

2.5.3 The Tools

We used `MeqTrees` (Noordam and Smirnov, 2010), for simulations. MeqTrees provides options to perform realistic simulations which include ionospheric effects, the parallactic rotation for alt-az mounts, primary beam effects such as pointing errors and sky rotation, feed errors and gain errors (G Jones).

Other important software packages used were `simms` [2] to create measurement sets (MS) by taking observational configurations shown in Table 2.1, `tigger` [3] for

[2] `https://github.com/SpheMakh/simms`
[3] `https://github.com/ska-sa/tigger`

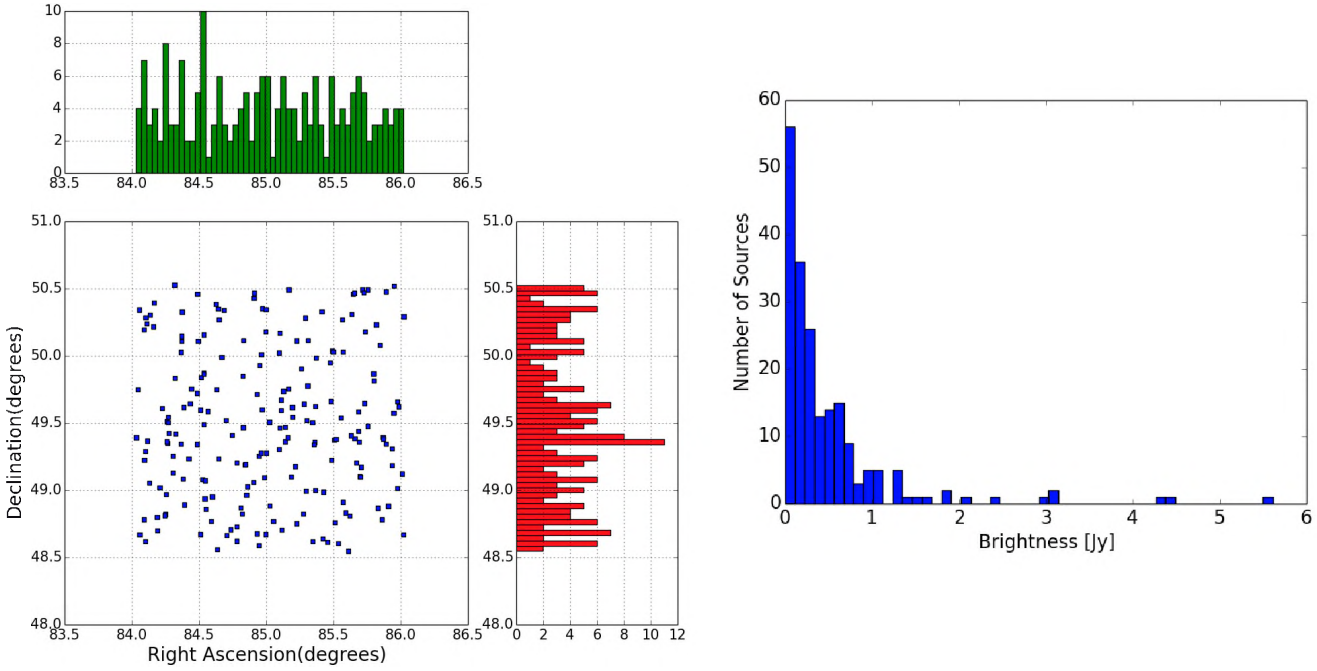


FIGURE 2.12: *Left panel*: is the distribution of the source positions (red plot is the declination and green plot is the right ascension). *Right panel*: represent the distribution of source' brightnesses.

viewing, reading in and modifying the sky models, and `wsclean` (Offringa et al., 2014) for imaging. These were incorporated into a pipeline using a framework known as `pyxis` [4]. This pipeline simulates random skies, calibrates the visibilities, then images the corrected data and finally runs `sourcery` on the resulting image.

2.5.4 Visibility Simulation

For corrupting the visibilities, we included noise according to the radiometer formula presented in Eq. 1.14. The SEFD and the angular resolution of the JVLA for this configuration are represented in Table 2.1 and they are 420 Jy and 14 arc-seconds, respectively.

Since we are interested particularly in artefacts induced by primary beam rotation, we introduced the DDEs using the measured beams called holographic beams. These beams represent more realistic beams of the JVLA.

[4] `https://github.com/ska-sa/pyxis/`

2.5.5 Visibility Calibration

We corrected for the above induced primary beam effects using primary beam models derived from `cassbeam`. Cassbeam models beams for the Cassegrain antennas such as the VLA antennas, using the geometrical ray tracing method. This technique is described in detail in: `https://github.com/ratt-ru/cassbeam/blob/master/doc/cassbeam.tex`.

2.6 Algorithm Test Procedure

The *input model*, Q, used for simulations is compared with the output catalogue from `sourcery`, S. The catalogue S contains the reliability estimates. The positions of the sources in S are cross-matched with those in Q. This dissertation uses four classes of detections to evaluate the performance of the algorithm and these are dependent on the choice of the reliability threshold, R_{thr}. The four classes are:

True detections, T, are the detections with $R > R_{thr}$, and exist in both Q and S. These detections are classified correctly as astrophysical sources.

False positives, FP, are the detections with $R > R_{thr}$ but exist in S and not in Q. These are artefacts classified as sources.

False negatives, FN, are detections appearing in both S and Q but have $R < R_{thr}$. These are true sources falsely identified as artefacts.

False detections (true negatives), F, are detections with $R < R_{thr}$ and that exist in S and not in Q. These detections are classified correctly as artefacts.

2.7 Simulated Data Classification Test

Figure 2.13 shows the image used to test the algorithm that does source-artefact discrimination. This image provides a good test case as it is dominated by calibration artefacts. We extracted emission from this image using the `PyBDSM` (Mohan and Rafferty, 2015) source finder. And the corresponding source finding parameters for island `thresh_isl` and model fitting `thresh_pix` were set to 1 and 2,

respectively. The value of 2σ was used to create a mask. Thus, any emission in the smoothed images with flux less than $2\times$ the noise were masked (see section 2.4.1).

2.7.1 Increase of the Parameter Space

Here we want to investigate whether the increase in the number of parameters has improved the classification. The initial reliability thresholds of 60% and 80% are chosen arbitrarily and gave the detections presented using confusion matrices shown in Table 2.2.

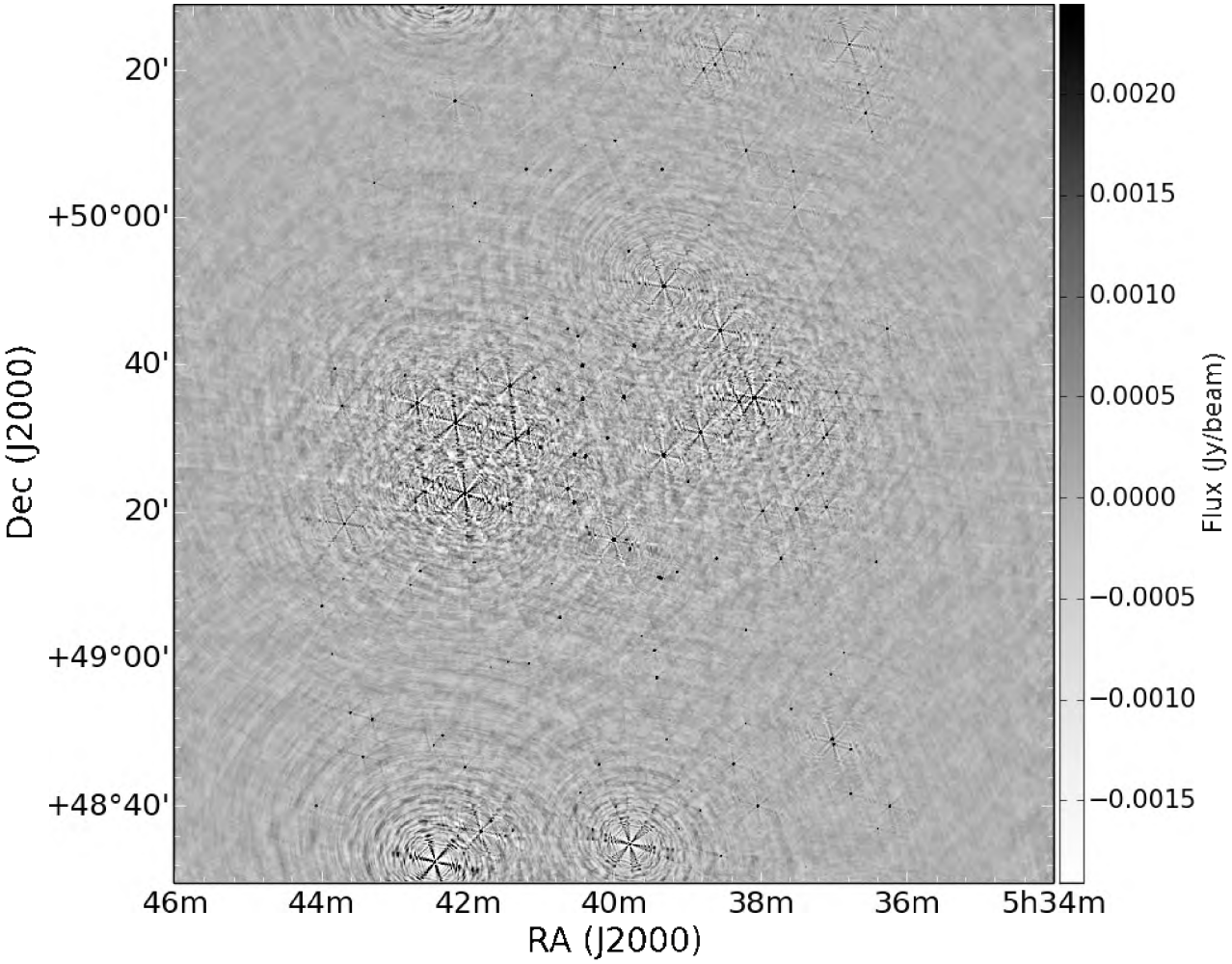


FIGURE 2.13: Simulated JVLA image: 2 hours synthesis time and 120MHz observational bandwidth.

Figure 2.14 contains the plots showing the distribution of the detections obtained using $R_{thr} = 60\%$, as presented in Table 2.2.

TABLE 2.2: Four classes of detections. *Left*: the reliability threshold 60%. *Right*: the reliability threshold 80%

		Predictions	
		True	False
Actual	True	158	32
	False	3	322

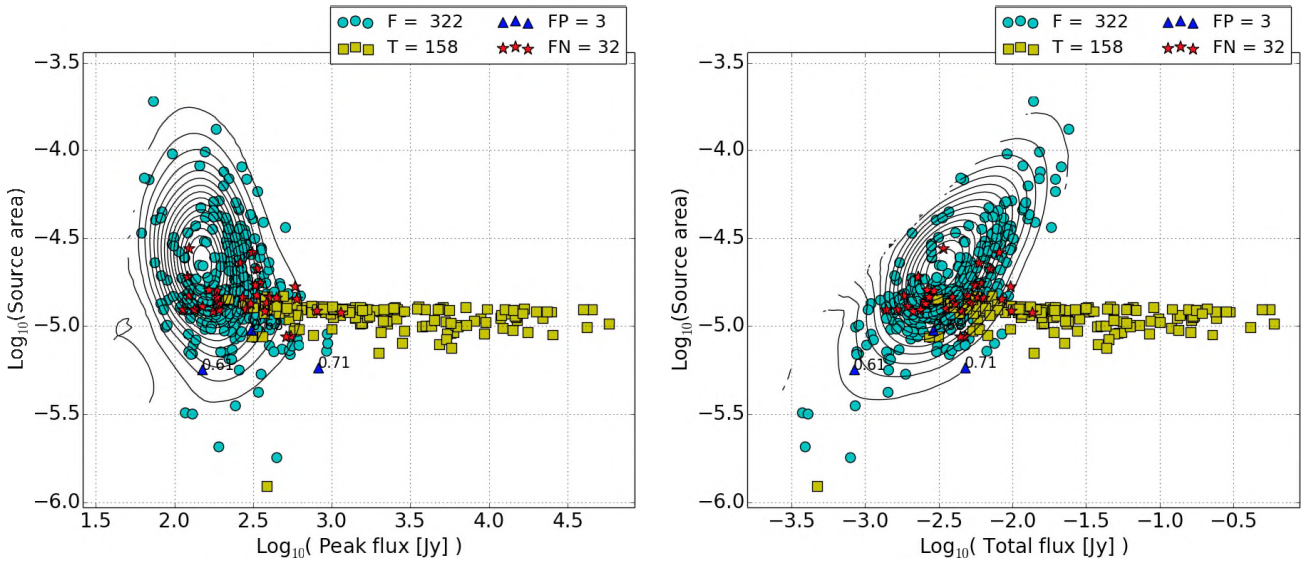
		Predictions	
		True	False
Actual	True	146	44
	False	0	325

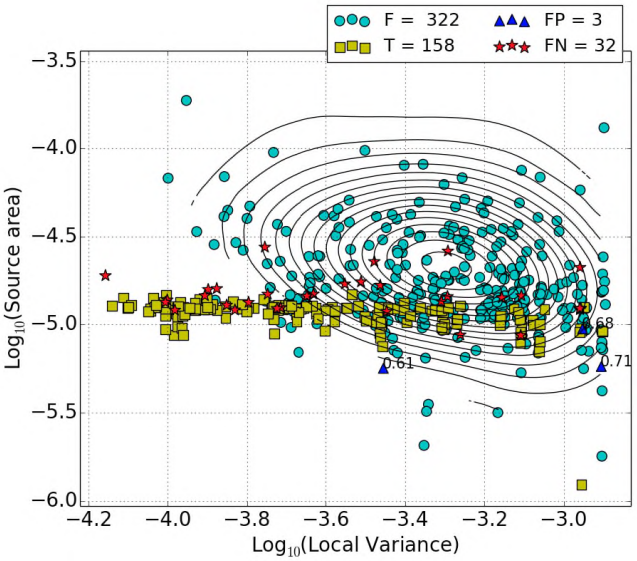
Table 2.3 shows the detections found when using the parameter space defined by the source area, peak flux and total flux which is equivalent to the parameter space in Serra et al. (2012). The only difference is that, in spectral line data the source size is defined as "source volume" (voxel) whereas in continuum data is defined in terms of surface area.

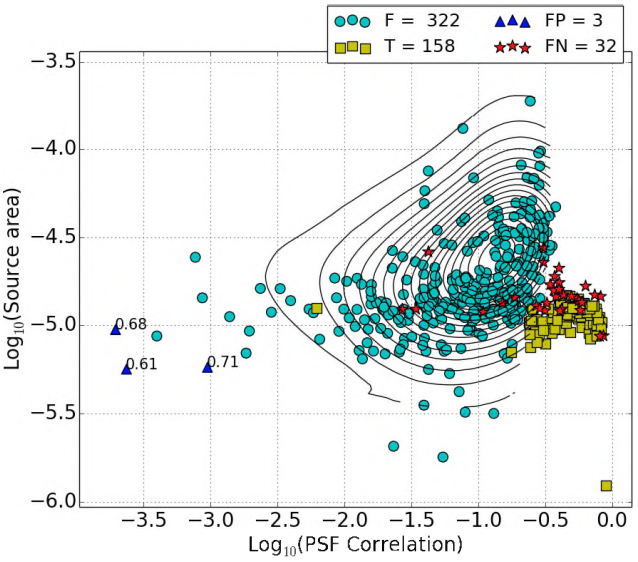
TABLE 2.3: The detections found using the parameter space defined by source area, peak flux and total flux. *Left*: reliability threshold is 60%. *Right*: the reliability threshold is 80%.

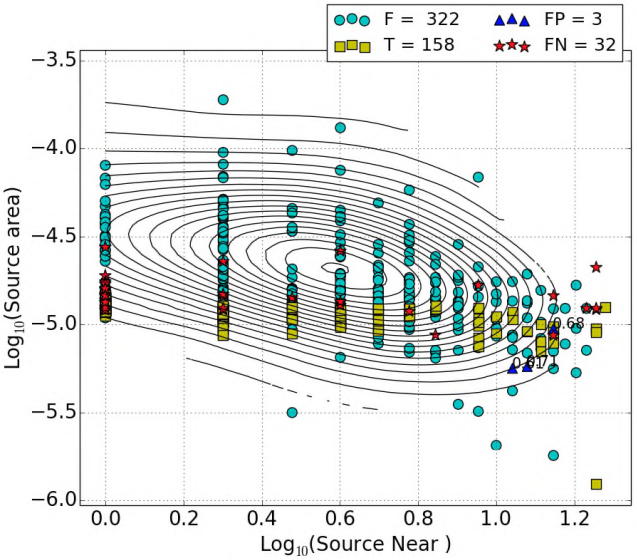
		Predictions	
		True	False
Actual	True	128	62
	False	3	322

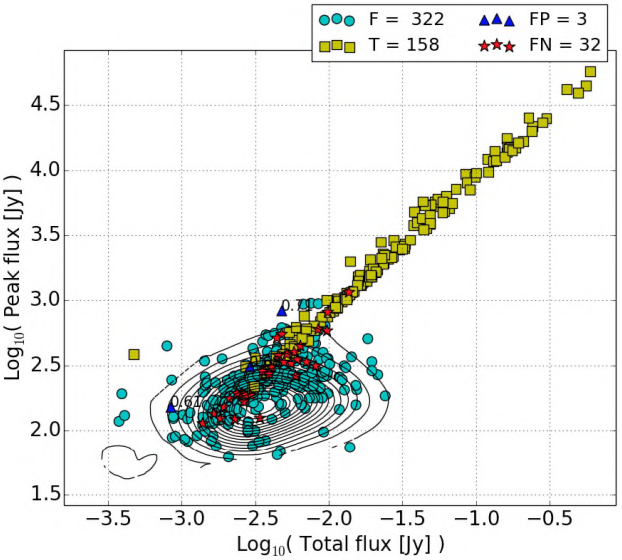
		Predictions	
		True	False
Actual	True	115	75
	False	0	325

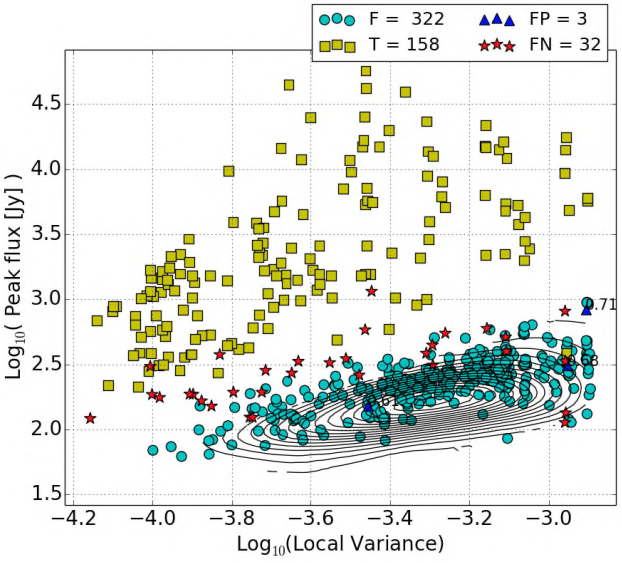


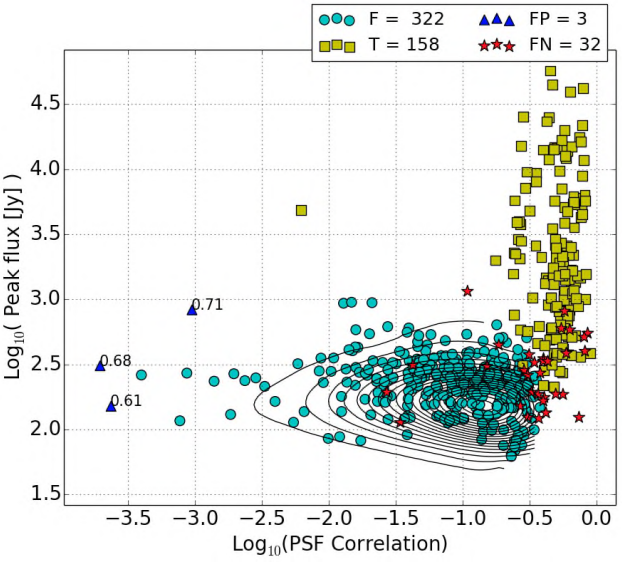

F = 322
T = 158
FP = 3
FN = 32
Log10(Source area)
Log10(Local Variance)
0.61
0.68
0.71

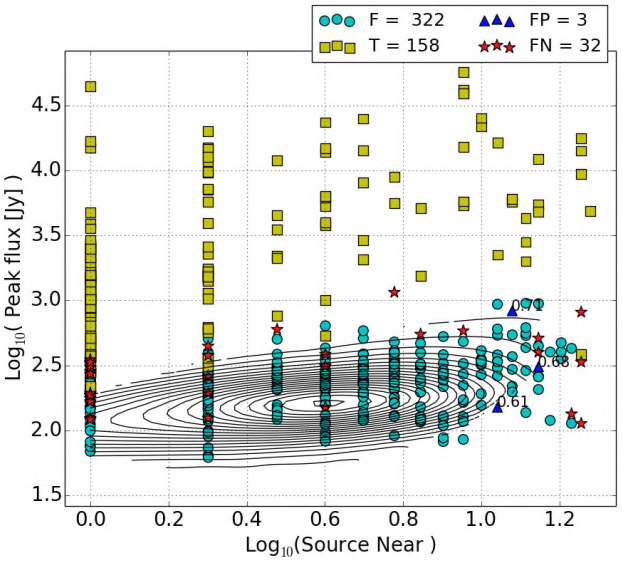

F = 322
T = 158
FP = 3
FN = 32
Log10(Source area)
Log10(PSF Correlation)
0.61
0.68
0.71

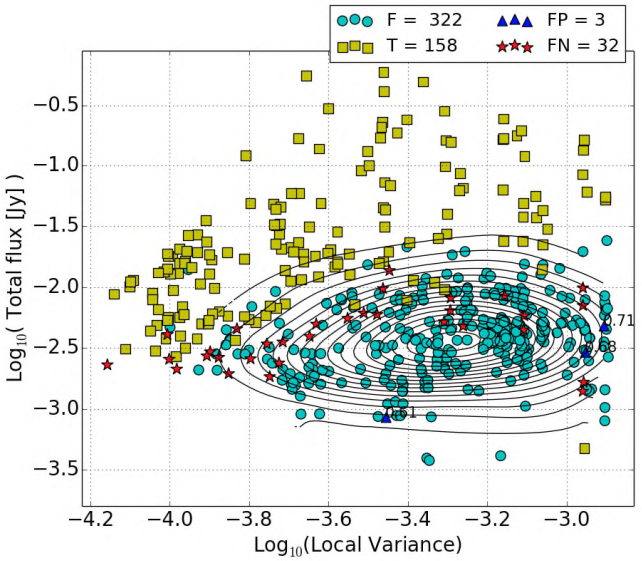

F = 322
T = 158
FP = 3
FN = 32
Log10(Source area)
Log10(Source Near)

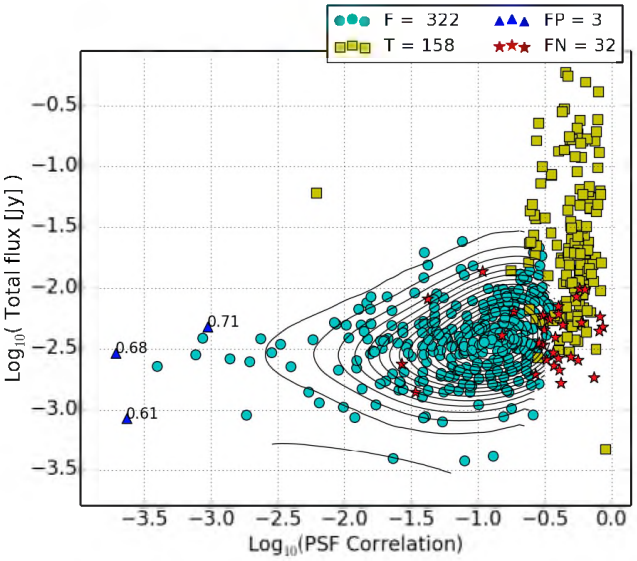

F = 322
T = 158
FP = 3
FN = 32
Log10(Peak flux [Jy])
Log10(Total flux [Jy])

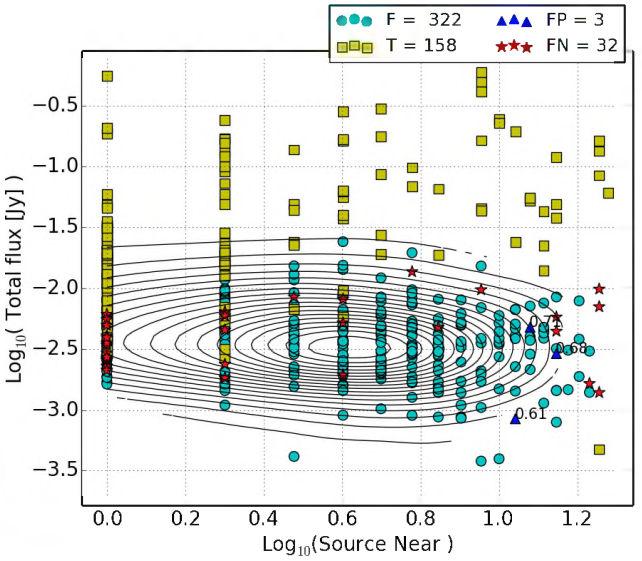

F = 322
T = 158
FP = 3
FN = 32
Log10(Peak flux [Jy])
Log10(Local Variance)
0.71
0.68

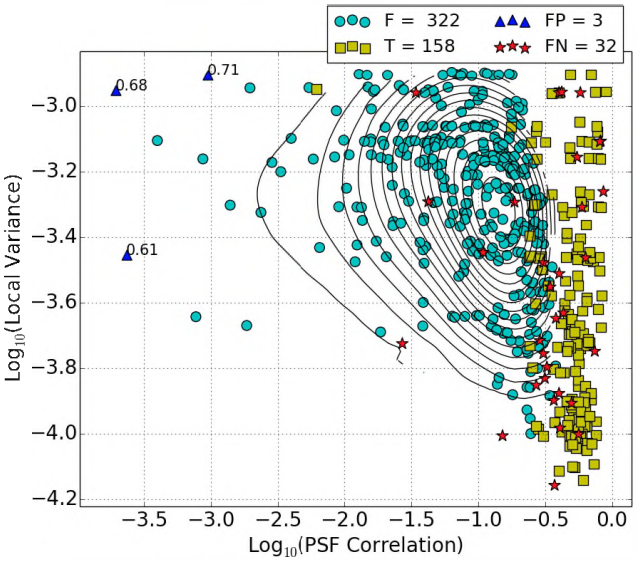

F = 322
T = 158
FP = 3
FN = 32
Log10(Peak flux [Jy])
Log10(PSF Correlation)
0.71
0.68
0.61

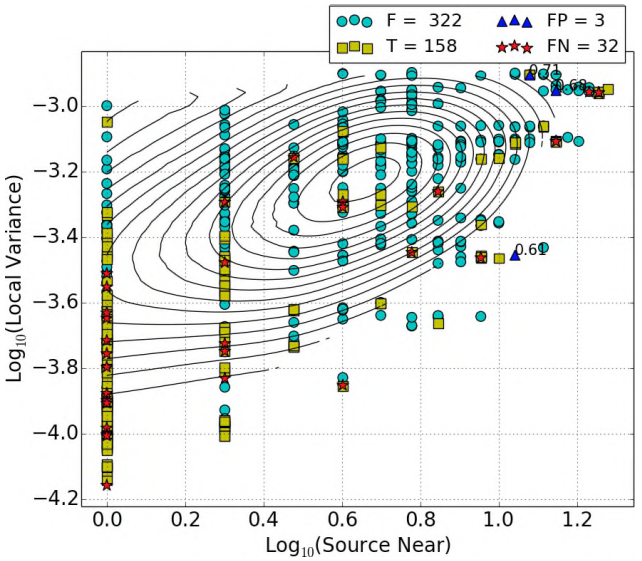

F = 322
T = 158
FP = 3
FN = 32
Log10(Peak flux [Jy])
Log10(Source Near)
0.71
0.68
0.61


F = 322
T = 158
FP = 3
FN = 32
Log10(Total flux [Jy])
Log10(Local Variance)
0.71
0.68
0.61


F = 322
T = 158
FP = 3
FN = 32
Log10(Total flux [Jy])
Log10(PSF Correlation)
0.71
0.68
0.61


F = 322
T = 158
FP = 3
FN = 32
Log10(Total flux [Jy])
Log10(Source Near)
0.71
0.68
0.61


F = 322
T = 158
FP = 3
FN = 32
Log10(Local Variance)
Log10(PSF Correlation)
0.71
0.68
0.61


F = 322
T = 158
FP = 3
FN = 32
Log10(Local Variance)
Log10(Source Near)
0.71
0.68
0.61

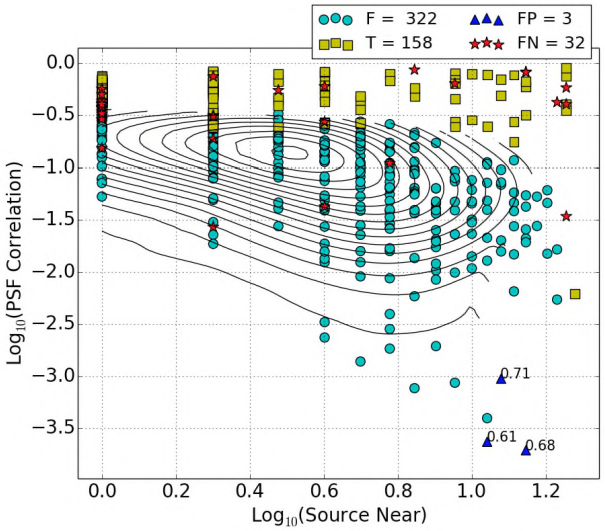


FIGURE 2.14: The distributions of the detections. T are the true detections, FP are the false positives, FN are the false negatives and F are false detections. Source area is in units of arcseconds. The labelled points represent the reliabilities of the FP detections. The contours represent the negative density field derived from the negative detections.

2.7.2 Discussion of the Classification

The following are the analyses based on the above tables:

- It is observed that at $R_{thr} = 60\%$ three of the artefacts were classified as real, and were eliminated by setting a higher reliability threshold of 80%. However, a high threshold rendered twelve of the T detections to be classified incorrectly as artefacts.
- Table 2.2 contains a larger number of T detections than Table 2.3 at the respective reliability thresholds. In particular, at R_{thr} of 80% where there are no artefacts, the completeness increased from 60.5% to 76.8%. This shows that the increase in the number of well-defined parameters improves the completeness of the final models.

By studying the distributions of T, FN, FP and F detections across different parameter spaces, we observe that:

- The detections are indistinguishable in the parameter spaces defined by the combination of the source area, local variance and the number of neighbours. These parameter combinations contribute negatively to source-artefact discrimination.

- A clear separation of the detections is observed in the parameter spaces defined by the local variance combined with other parameters, except the ones mentioned above.

- Using the source area, the PSF correlation and the number of neighbours combined with the peak flux and total flux, distinguishes the detections as well as the above parameter combinations, except that this classifies more real emission as artefacts.

Based on the above plots, it can be seen that the misclassified artefacts occupy unique positions in the parameter spaces defined by the PSF correlation. In Section 2.4.2 we showed that artefacts correlate poorly with the PSF. It is for this reason that we can choose to discard the detections with low correlation values. The correlation values of the three false detections shown in Table 2.2 are 2.35×10^{-4}, 1.95×10^{-4} and 9.48×10^{-4}. These correlation values are very low that these detections appear as outliers in the parameter space thus, giving rise to high reliabilities. After discarding poor correlation sources ($cf < 0.0999$, chosen arbitrarily), we obtained the minimum reliability of 10% which resulted in a total of 179 true detections, 11 false negatives, 0 false positives and the remaining detections are false.

2.7.3 `Sourcery` as a complementary tool to `PyBDSM`

The development of sourcery was intended to complement PyBDSM or other source finders, in general, by removing (or reducing the number of) falsely detected artefacts from the catalogues, while still retaining good enough completeness. To evaluate if this is actually the case, I simulated a total of 60 JVLA images with 4 hours synthesis time, a single channel of width 128 MHz, a pointing centre at 0 degrees right ascension and 40 degrees declination, and a field of view of 2 degrees. I generated the positions and fluxes of the sources, and corrupted and calibrated the visibilities similarly to the above simulations. Furthermore, I ran PyBDSM on each image using the default parameters (`thresh_isl`=3 and `thresh_pix`=5). Similarly, I used the same thresholds for `sourcery` except for the negative image where I've used `thresh_isl`=2 and `thresh_pix`=4. Figure 2.15 shows PyBDSM detections; T are the true detections, FP are the artefacts detected as real (false positives) and T at FP=0 are the true detections obtained after removing FP

detections. PyBDSM detects sources based on their flux values. For example, if an image contains true sources with flux values between (2 , 0.5) Jy and artefacts with fluxes between (0.8, 0.5) Jy, then detecting sources down to (0.8, 0.5) Jy will also introduce artefacts. Therefore to obtain the blue bars in Fig. 2.15, I discarded all the detections with fluxes below the maximum flux of the FP detections.

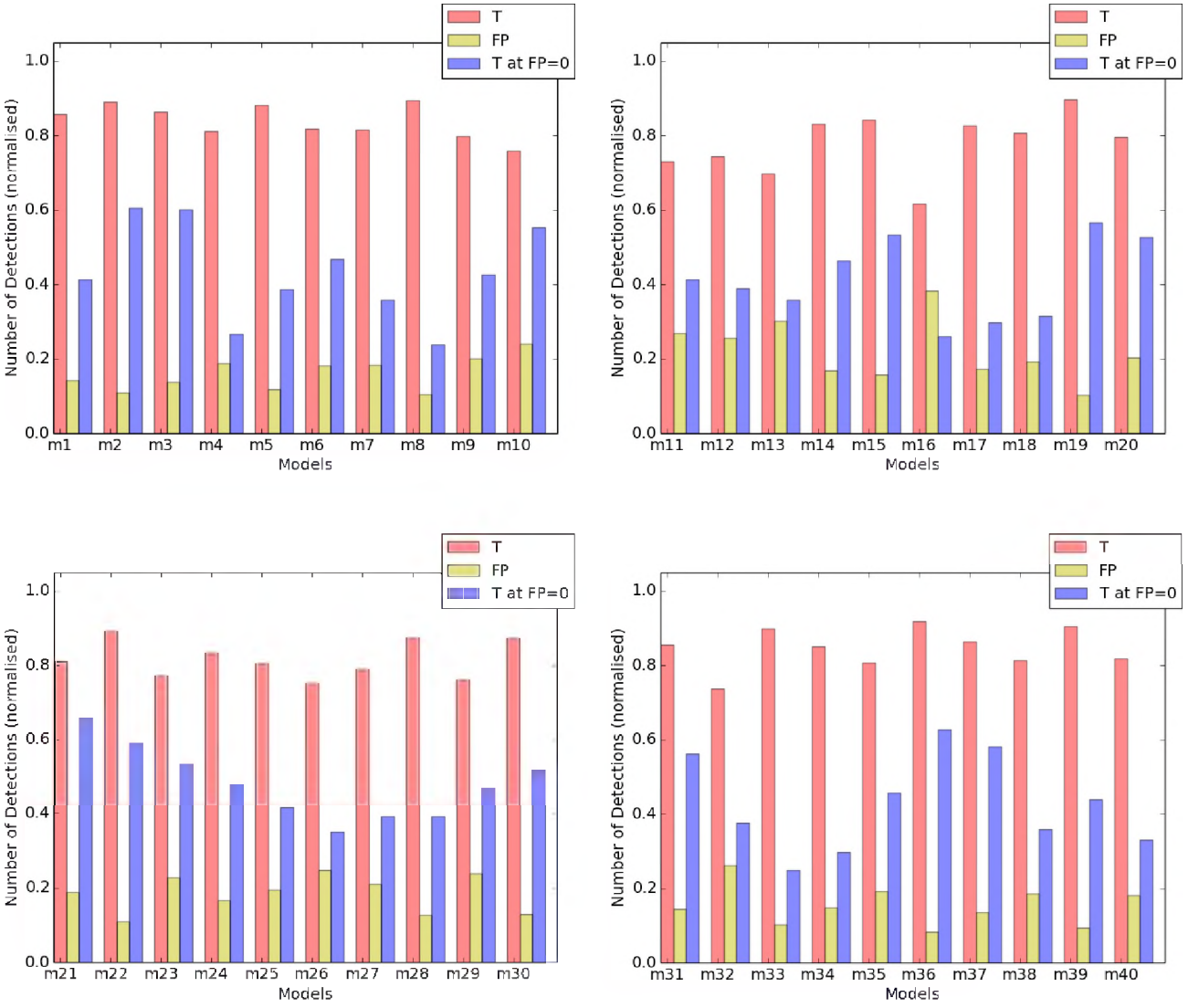


Figure 2.16 shows the detections obtained by `PyBDSM` and `sourcery`. I chose the reliability threshold of 50% and removed sources with $cf < 0.0999$. From this plots, one can see that although we don't recover all the sources with `sourcery` we still get models with better completeness than `pybdsm` at FP=0. Sources which are closer to the noise level are misclassified as false and including them means introducing noise peaks. Our parameter space is optimized to deal with calibration artefacts and not noise peaks. The false positive detections which are shown by black bars in Fig 2.16 are very bright noise peaks which can be eliminated by

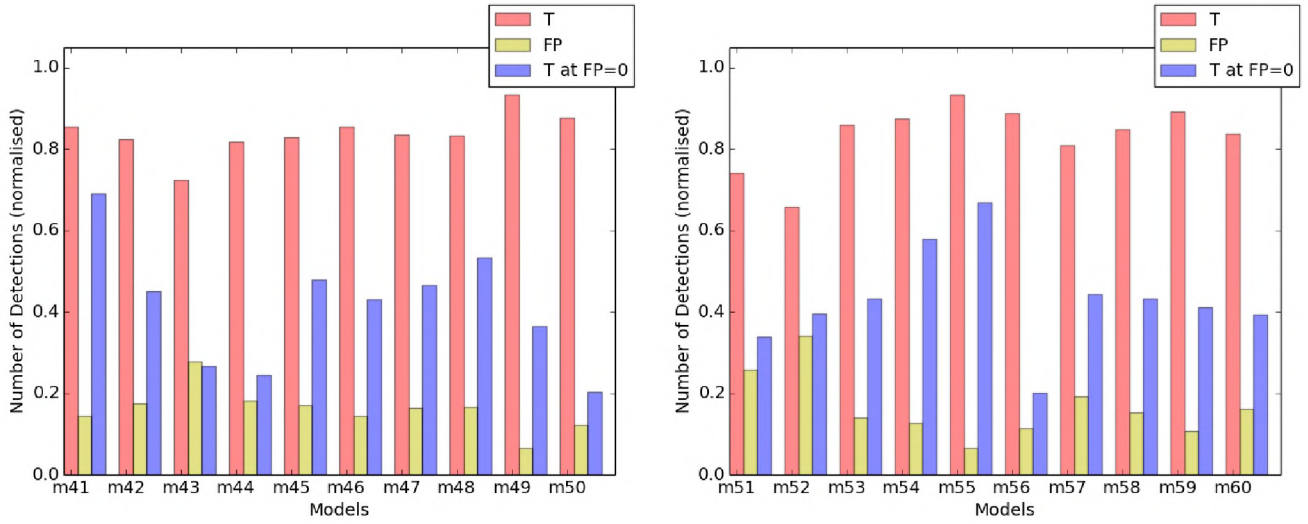
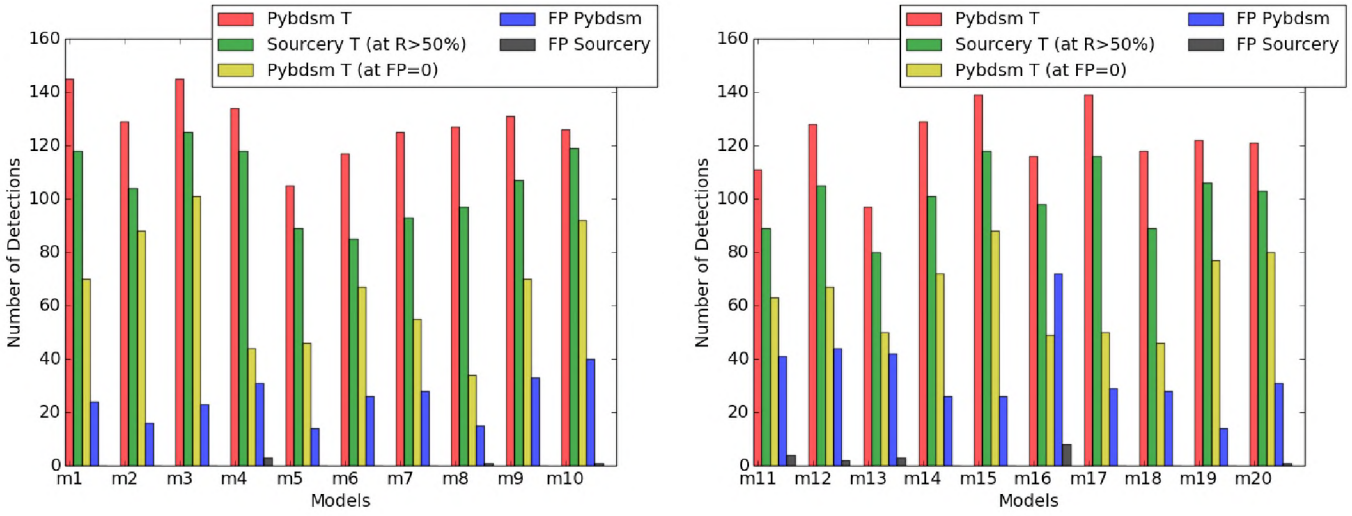


FIGURE 2.15: Detections with `PyBDSM` on 60 JVLA simulated images. T are the true detections, FP are the false positives and T at FP=0 are the true detections obtained after removing the FP.

increasing the reliability thresholds or by finding parameters that can deal with noise peaks. In a nutshell, `sourcery` improves the reliability of the catalogues and completeness.



2.8 Test on Real Data

We tested `sourcery` on the $3C147$ field image observed using JVLA (Perley and Smirnov, 2013). The object $3C147$ (J054+4951) is very compact ($\sim$ 700mas) and is also the strongest radio source in its field with brightness of 22.82Jy at the L-band (Perley and Smirnov, 2013). The observational configuration is presented in

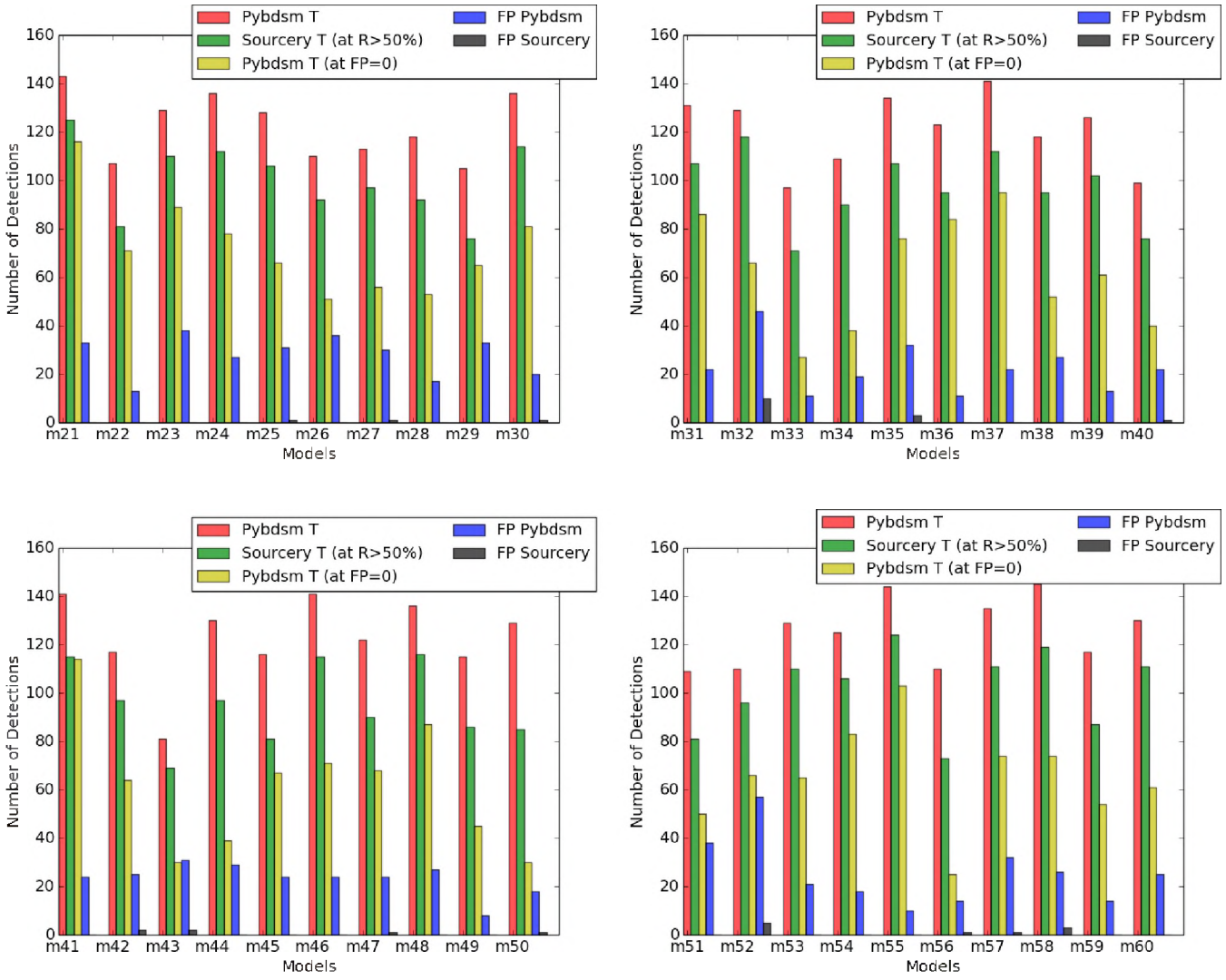


FIGURE 2.16: Detections obtained from both `PyBDSM` and `Sourcery`.

Table 2.4. The reliability threshold of 80% was used based on our experience with the simulations discussed above, and we further inspected (visually) the resulting catalogue to ensure that none of the spurious emission was included. However, we do know that with real data one may never be entirely certain as to whether a source is real or not. Thus, for this comparison we ensured that 80% reliability only included sources that were obvious. For comparison purposes, we have also masked all the detections that are less than 20 arc-minutes (size determined by eye) from the phase centre due to significant subtraction artefacts in the vicinity of $3C147$ itself. Interestingly enough, two of the artefacts were removed through the use of cf_0. This allowed the lowest R_{thr} to become 60%.

We compared catalogues obtained from images created at different stages of the

TABLE 2.4: Observational Configuration for the $3C147$ Field.

Parameter	Value
Antenna configuration	C-Config
Central Frequency	1.5 GHz
Channel width	3 MHz
No. of Channels	64
Total bandwidth	192 MHz
Frequency Range	1260 -1520 MHz
Synthesis Time	8 Hours
Integration time	1 sec
Polarisation	4
Calibrator source	J0555 + 3948 (every 10 minutes)

$3C147$ calibration pipeline. This calibration pipeline was initially created by Perley and Smirnov (2013) and then later modified by Mitra et al. (2015) who introduced primary beam corrections via `cassbeam` (Brisken, 2003). Below is the summary of what it entails (Perley and Smirnov, 2013):

1. A prior full-polarisation model for $3C147$ is used for $\boldsymbol{G}$ calibration.
2. Source subtraction is performed to remove $3C147$. The residuals are imaged and deconvolved, this is followed by source finding and the resulting detections are corrected for the power beam and updated into the initial model.
3. The updated model is used to perform $\boldsymbol{E}$, dE and $\boldsymbol{G}$ calibration. The new residuals are imaged and deconvolved, and source finding is performed on the resulting residuals. The acquired model is beam corrected and added to the existing model.
4. Step 3 is repeated one more time to improve $\boldsymbol{G}$, $\boldsymbol{E}$ and dE solutions using the above updated model. At the end the residuals are imaged and deconvolved.
5. The final sky model is restored into the residual image.

This pipeline corrects for DDEs in two ways: (1) first it applies the primary beam corrections using the available parametric models (Brisken, 2003) followed by (2) the differential gains (to selected sources, Smirnov, 2011c) to correct for the remaining DD effects.

2.8.1 Model Comparisons

We applied `sourcery` to images obtained at step 2 and 3. Let us refer to the resulting catalogues as S_{ini} and S_{fin}, respectively. We call the models that were obtained by Mitra et al. (2015) in the respective steps A_{ini} and A_{fin}. Note that we combined multiple components that fell within the PSF size at 20% deviation into a single source component so that there is a one to one match when cross-matching different source catalogues [5].

Figure 2.17 shows the initial models S_{ini} and A_{ini} and Fig. 2.18 depicts the final models S_{fin} and A_{fin}. The model S_{ini} had 118 sources and A_{ini} had 27 sources, and it was observed that all the sources in A_{ini} were in S_{ini}. On the other hand, the model S_{fin} had 192 detections and A_{fin} had 168 detections. When comparing the final calibration models we found that 39 sources in S_{fin} were not found in A_{fin} while 14 sources in A_{fin} were not found in S_{fin}.

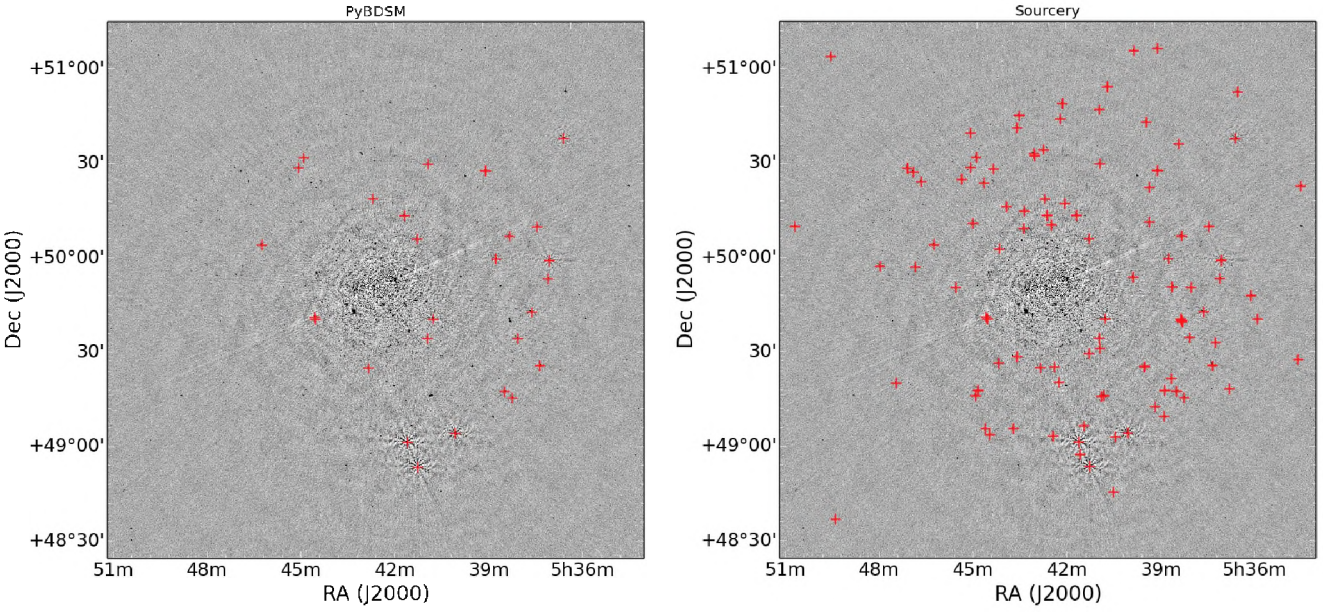


FIGURE 2.17: JVLA 3*C*147 field. *Left*: the initial model obtained when using a `PyBDSM` (A_{ini}) and *Right*: the initial model, S_{ini}, obtained using `sourcery`. The sources in the `sourcery` model have reliabilities $> 60\%$. The A_{ini} model contains 27 sources and has a total of 47 sources when including those within $20'$ from the phase centre. The S_{ini} model contains 118 sources and has a total of 165 sources when the central region sources are included.

[5] `PyBDSM` does provide an option to obtain a catalogue in the form of a source list, with neighbouring components similarly merged, however, for purposes of calibration the Gaussian component list provides a more accurate model.

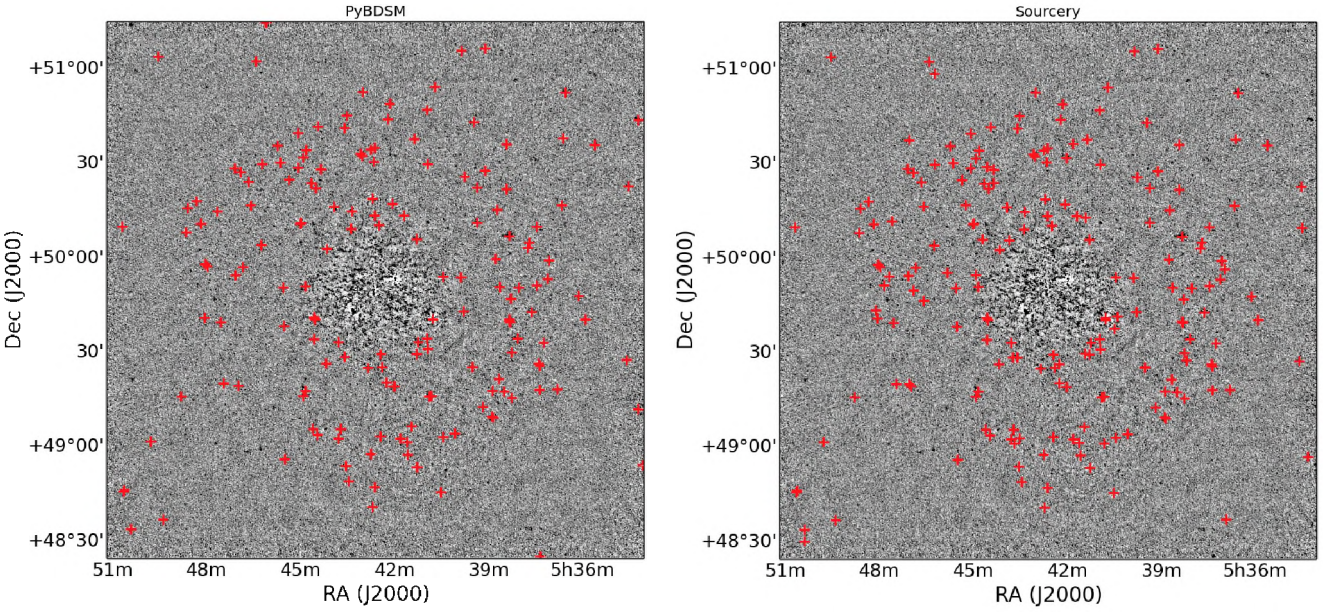


FIGURE 2.18: JVLA $3C147$ field. *Left*: the final model obtained when using `PyBDSM` (A_{fin}) and *Right*: the final model, S_{fin}, obtained using `sourcery`. The sources in the `sourcery` model have reliabilities $> 80\%$. The A_{fin} model contains 168 sources and has a total of 305 sources when including those within $20'$ from the phase centre. The S_{fin} model contains 192 sources and has a total of 467 sources when the central region sources are included.

2.8.2 $3C147$ Calibration Pipeline Implementation

In addition to checking the completeness of the catalogues, we further incorporated our algorithm into the $3C147$ calibration pipeline. Previously, Mitra et al. (2015) had obtained a dynamic range (DR) of 3.27 million. They found the aforementioned DR using 3 calibration steps. We find a DR of 3.118 million in 2 calibration steps and 3.145 million in 3 calibration steps when replacing the source finding steps of the pipeline with `sourcery`.

There are three approaches by which we estimated the noise: (1) by taking the standard deviation of the pixels in the noise-only regions, (2) by computing the mean absolute deviation (MAD) of the pixels in an image and (3) computing the negative noise. The negative noise is obtained by taking the standard deviation of the pixels in the inverted image. Figure 2.19 shows the noise estimate for the three methods. The negative noise and noise-only regions provided better estimates of the noise but we chose to use the former to compute the above DRs. Figure 2.20 shows the difference in the final corrected residual images obtained when using `PyBDSM` and `sourcery`.

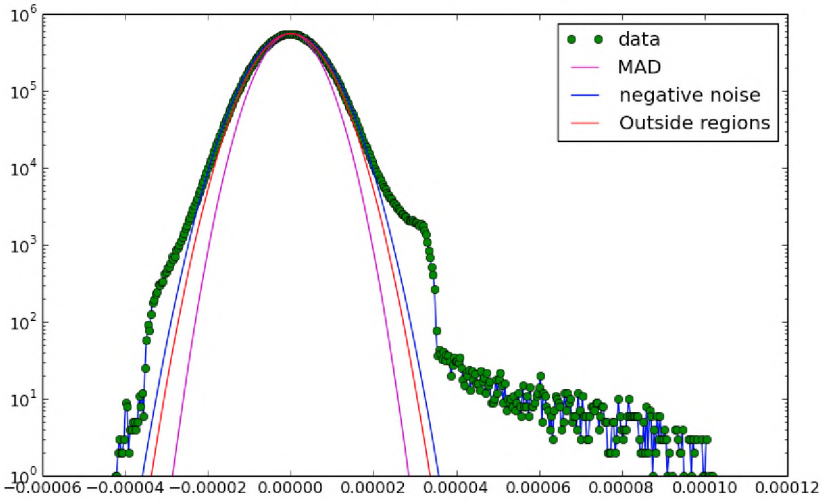


FIGURE 2.19: Three methods to estimate the noise in an image: median absolute deviation (MAD), standard deviation of the noise-only regions (outside regions) and the negative noise. The histogram is the $3C147$ image data (final residuals).

2.8.3 Discussion

When comparing the models we observed that:

- Our initial model has better completeness compared to the initial model found by `PyBDSM`.
- The final model with `sourcery` has a higher completeness than the final model constructed with `PyBDSM`. However, there are sources that were entirely missed in one model but are in the other model. For instance, although S_{fin} was more complete than A_{fin}, there were 14 sources in A_{fin} that it overlooked. Some of the sources that were overlooked by `sourcery` were due to source exclusion that takes place when computing the local variance and PSF correlation. When computing these parameters a region around a source is selected and if this region happens to go beyond the boundary of the image a source is flagged from the catalogue and excluded from the reliability estimations. Out of the 14 sources, 6 fell into this category. One way to deal with this problem is to increase the image size.

Typically, a complete calibration model implies better estimations of the instrumental gains. Therefore, one would expect a far better DR with `sourcery` than

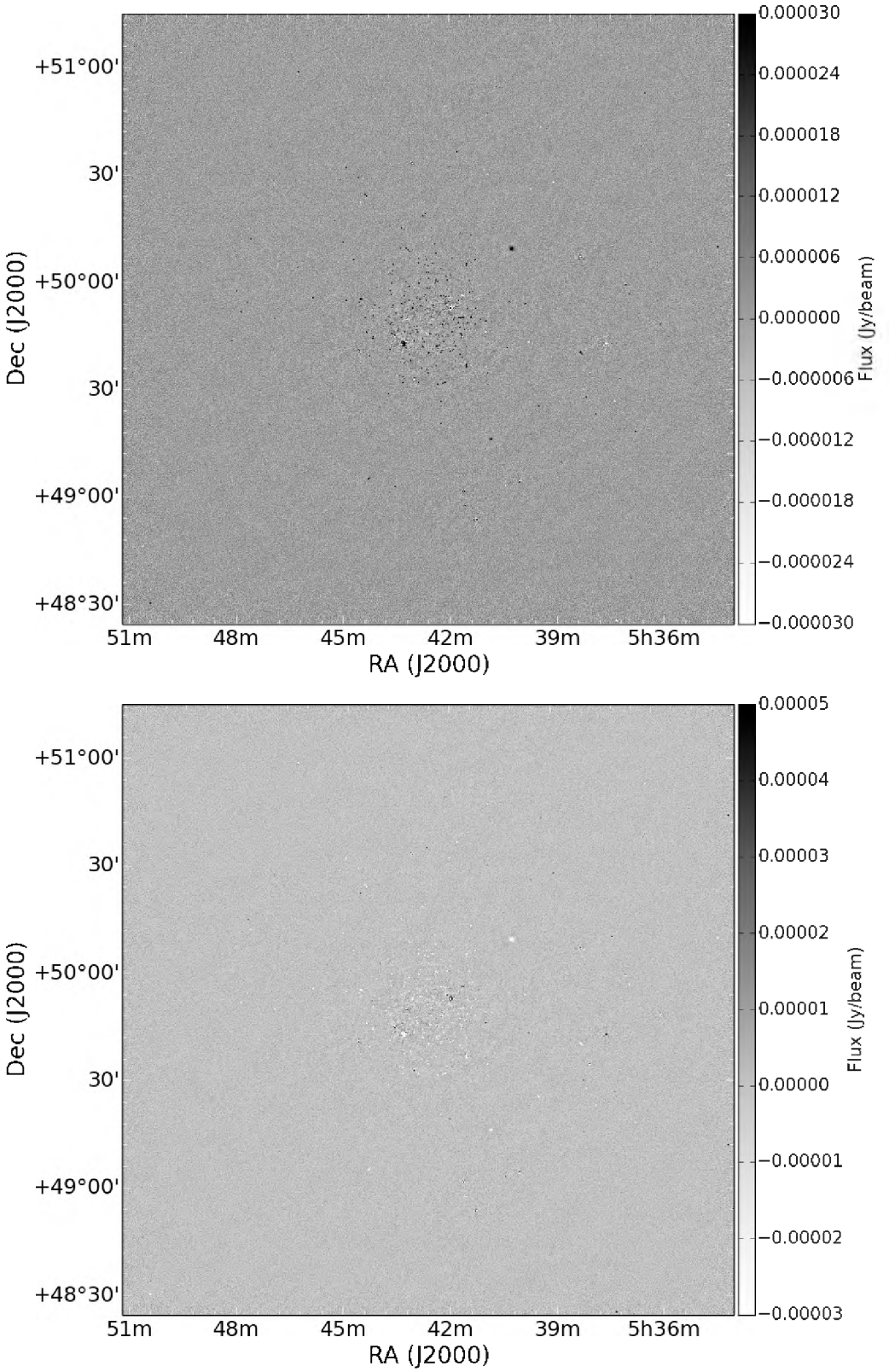


FIGURE 2.20: JVLA 3*C*147 calibrated field. *Top panel*: `PyBDSM` residual image minus the `sourcery` residual and *bottom panel:* the reverse subtraction.

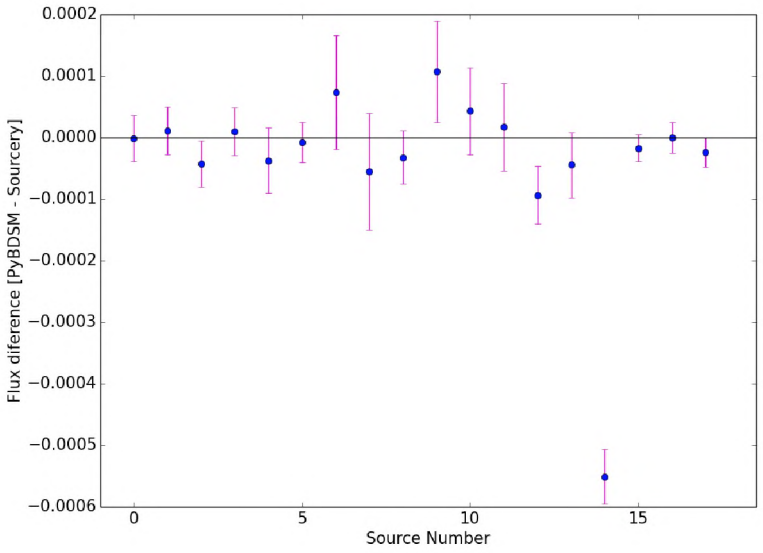


FIGURE 2.21: The differences in the modelled fluxes. The `PyBDSM` model fluxes minus the `sourcery` fluxes. Only 18 sources from both models were used. These are the sources with a single Gaussian component. The Y-axis is in units of Jansky and the error-bars represent the measurement errors found during source extraction.

`PyBDSM`. Instead, the DR with `PyBDSM` was slightly better than `sourcery`. By looking at the difference in the residuals it can be seen that `sourcery` model had accounted for more sources in the image than `PyBDSM`. Figure 2.21 shows the difference in the fluxes for some sources (using only the sources with single components matching in both models – this was to avoid dealing with multiple components) in `sourcery` and `PyBDSM` model. It can be seen that the flux difference of the sources is not significant and thus cannot be a responsible for lower DR.

When visually inspecting the residual (clean) images, we observed some regions are over- or under-subtracted. The degree of the subtraction couldn't be determined by eye, though we suspect that this may have some responsibility for the lower dynamic range. Poor subtraction is amongst many challenges in calibration. Source finders do not necessarily provide good models for all the detections. In some cases, the sources fluxes or shapes are underestimated or overestimated. Note that the completeness of the parameters and poor subtraction are out of the scope for this work, but are worth discussing given their importance to future calibration. For example, consider an image composed of Gaussian sources and suppose we happen to fit an appropriate model to each of them. If somehow, we have overestimated or underestimated the fluxes or shapes, the resulting images will be

comprised of over- or under-subtracted regions. In reality, images are composed of a wide range of source morphologies that are not Gaussian. Source finders including `PyBDSM` do offer an option to fit shapelets or wavelets to extended emissions, but these may only reduce the effect of poor subtraction. An alternative approach to improve source subtraction is to improve source modelling. We propose the use of optimisation methods such as Markov Chain Monte Carlo (MCMC) or Genetic Algorithms (GA) in which a variety of Gaussians, shapelets or wavelets are fit to a source and the best fit is obtained. An even better approach would be to develop algorithms that do the modelling in the visibilities.

The lower DR at the last step could also be due to incorrect parameter estimations due to inaccuracies in the images. Sourcery finds most of its sources at the early stages of the pipeline when the image is less accurate. Thus, the incorrectly modelled sources are carried over to the last calibration model. We reduced the dE smoothing both in time and frequency by 50% so that the inaccuracies in the model do not play a huge role in the calibration. This has slightly improved the DRs. We can improve our initial models by re-running `sourcery` on the final images. Thus, the final and complete catalogues would be found from rerunning `sourcery` after the final step.

The advantage of using `sourcery` in addition to `PyBDSM` is that it offers reliable models with better completeness at the early stages of the calibration, thus making for more efficient pipelines.

Chapter 3

The Identification of Sources that require DD Calibration

3.1 The Algorithm

Conventionally, sources that are subject to DD effects in an image are identified by eye. This work attempts to automate this process. The approach used here results from the knowledge acquired in the previous section. Hence, this task also uses the source parameters to classify the sources of interest. Sources that require DD calibration usually have the following properties:

1. They are the brighter sources in an image. Calibration artefacts are present around all the sources in an image, however, they are more prominent around bright sources and are usually hidden in the noise for the fainter sources. As explained in the previous chapter, calibration artefacts are multiplicative errors in the image domain and thus, become more significant around bright sources.

2. Because these sources are surrounded by calibration artefacts, they tend to have high local variance. The two scenarios that result in high local variance were discussed in Section 2.4.2.

3. As seen in Chapter 2, calibration artefacts correlate poorly with the PSF, unlike astrophysical sources of any size. Thus, the PSF correlation can be

used to eliminate calibration artefacts, which may have high local variance, from the classification.

4. Due to the quasi-symmetric behaviour of calibration artefacts, it can be observed that these sources have a large number of negative detections in their immediate proximity.

Based on the stated characteristics, we came up with the algorithm described in Fig. 3.1 to identify sources that require DD calibration. The thresholds used in each step are user specified and are defined as follows:

- A high SNR source has SNR greater than $c \times \text{SNR}_{min}$, where c is a user specified factor and SNR_{min} is the minimum SNR in the source catalogue.
- A high local variance source has the local variance $> d \times \sigma$, where d is a user specified factor and σ is the noise level of the image.
- The correlation can be any value between [0, 1], high values imply high levels of correlation. A source with high correlation has a correlation factor greater than cf, where cf is a user specified value, $cf \in [0, 1]$.
- The number of negative detections can be any integer value, i. Thus, a source with a high number negative neighbours has $i > M$, where M is a user specified number.

We made the PSF correlation optional so that the DD tagging does not depend on the availability of the PSF image. However, the inclusion of the PSF correlation may improve the classification.

3.1.1 Testing of the Algorithm

The algorithm was directly tested on four different real datasets: (*i*) the KAT-7, (*ii*) the JVLA, (*iii*) the Giant Metrewave Radio Telescope (GMRT) and (*iv*) the Low-Frequency Array (LOFAR). The algorithm classification was compared to that made by eye. Table 3.1 shows the thresholds used in each dataset. We have used default values except for the number of negative detections in the LOFAR image which we set higher to constrain the evaluation due to the unavailable PSF image and measurement set (MS).

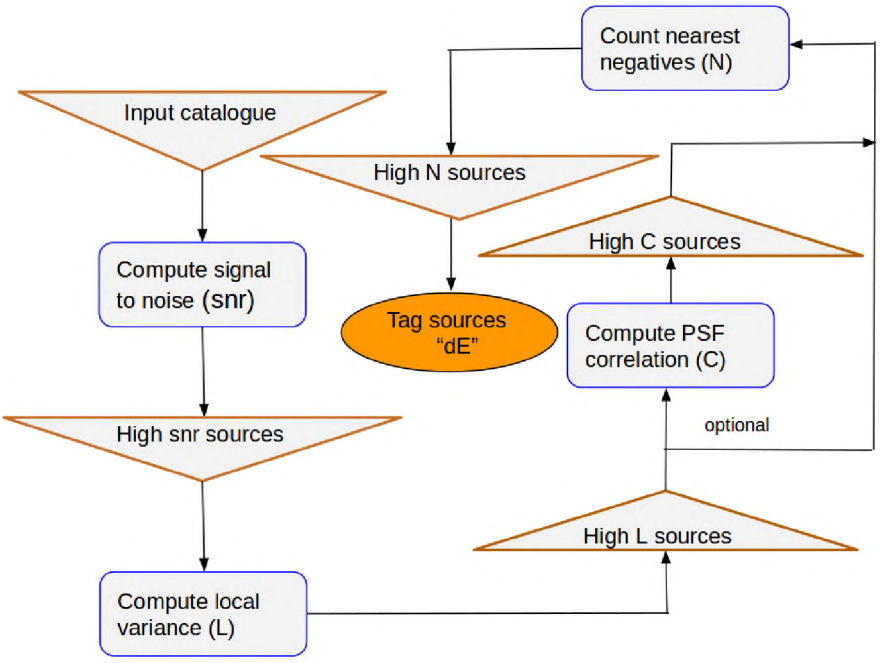


FIGURE 3.1: The algorithm that identifies sources that require DD calibration.

TABLE 3.1: Thresholds used for carrying out the source tagging for sources that require DD calibration.

Thresholds	KAT-7	JVLA	GMRT	LOFAR
SNR	40	40	40	40
Local variance	0.8	0.8	0.8	0.8
Correlation factor	0.4	0.4	0.4	None
No. of nearby negative detections	5	5	5	10

The source finding thresholds are similar to those used in Chapter 1, section 2.7. The island and model fitting thresholds were set to 1 and 2, respectively. The masking threshold was set to 2 for both the original image and the inverted image.

3.2 Classification Results

Figure 3.2 shows a KAT-7 field. The square highlights a source that was selected as requiring DD calibration.

Figure 3.3 shows the 3C147 field observed with the JVLA telescope (Mitra et al., 2015, Perley and Smirnov, 2013). The circled sources are those identified by our

algorithm. Figure 3.4 shows the zoomed images of the sources circled in Fig. 3.3.

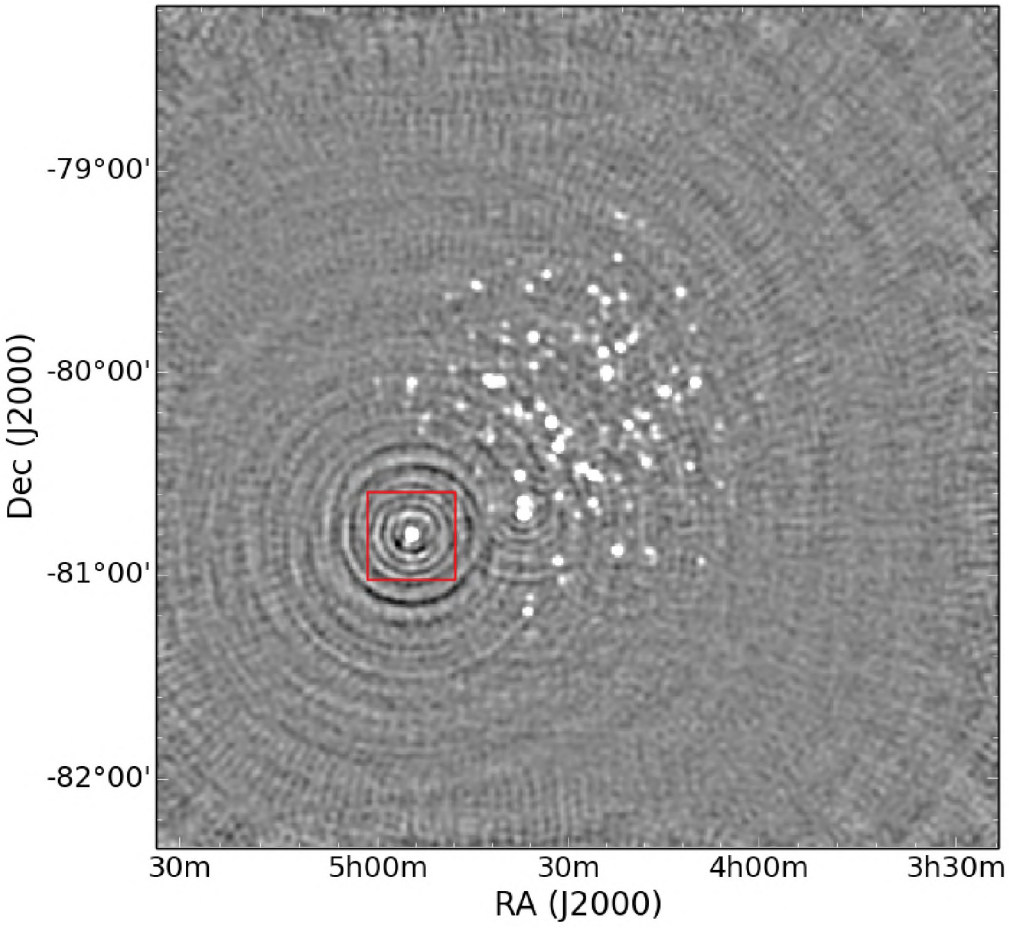


FIGURE 3.2: KAT-7 field. The indicated source was identified requiring DD calibration. Image credit: Ermias Kassaye (Rhodes University)

Figure 3.5 shows a GMRT observation. The circled sources shown in this image are those that require DD calibration solution as identified by the algorithm. Figure 3.6 contains the zoomed images of these sources.

Figure 3.7 shows a LOFAR observation. Sources inside the squares are those that require DD calibration as identified by our algorithm. The top image represent the classification obtained using the settings in Table 3.1 and the bottom image shows the classification after setting the number of nearby negative detections to 20. Figure 3.8 shows the zoomed images of these sources.

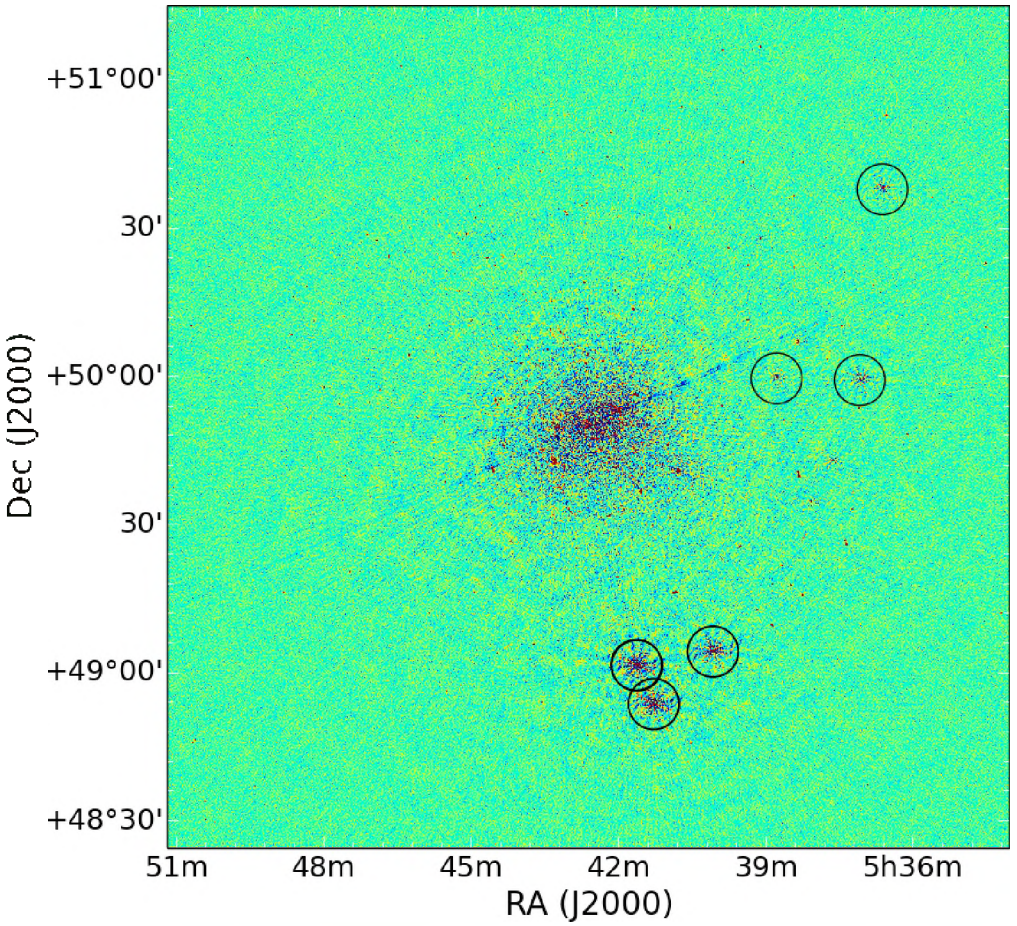
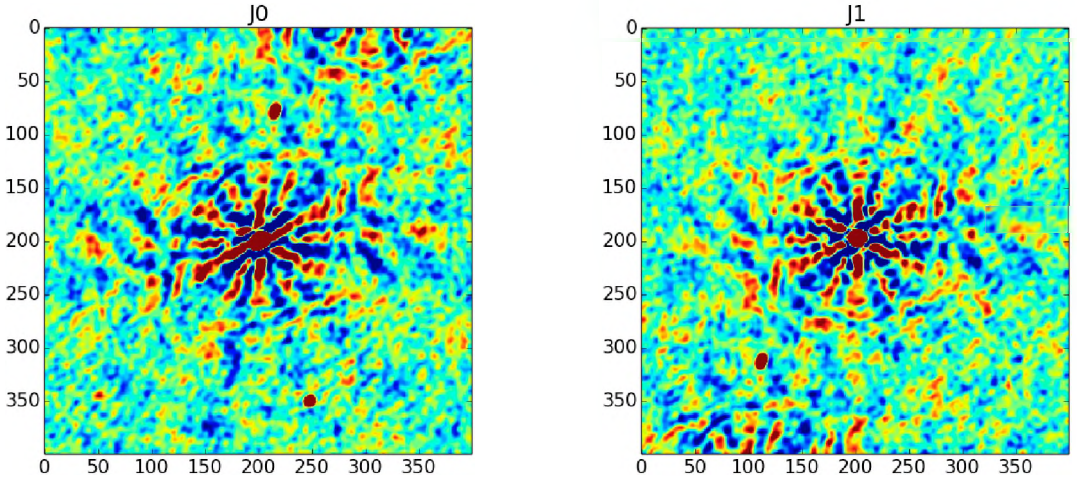


FIGURE 3.3: JVLA field. The circled sources are those that require DD calibration solutions as identified by our algorithm. Image credit: Modhurita Mitra (Rhodes University).



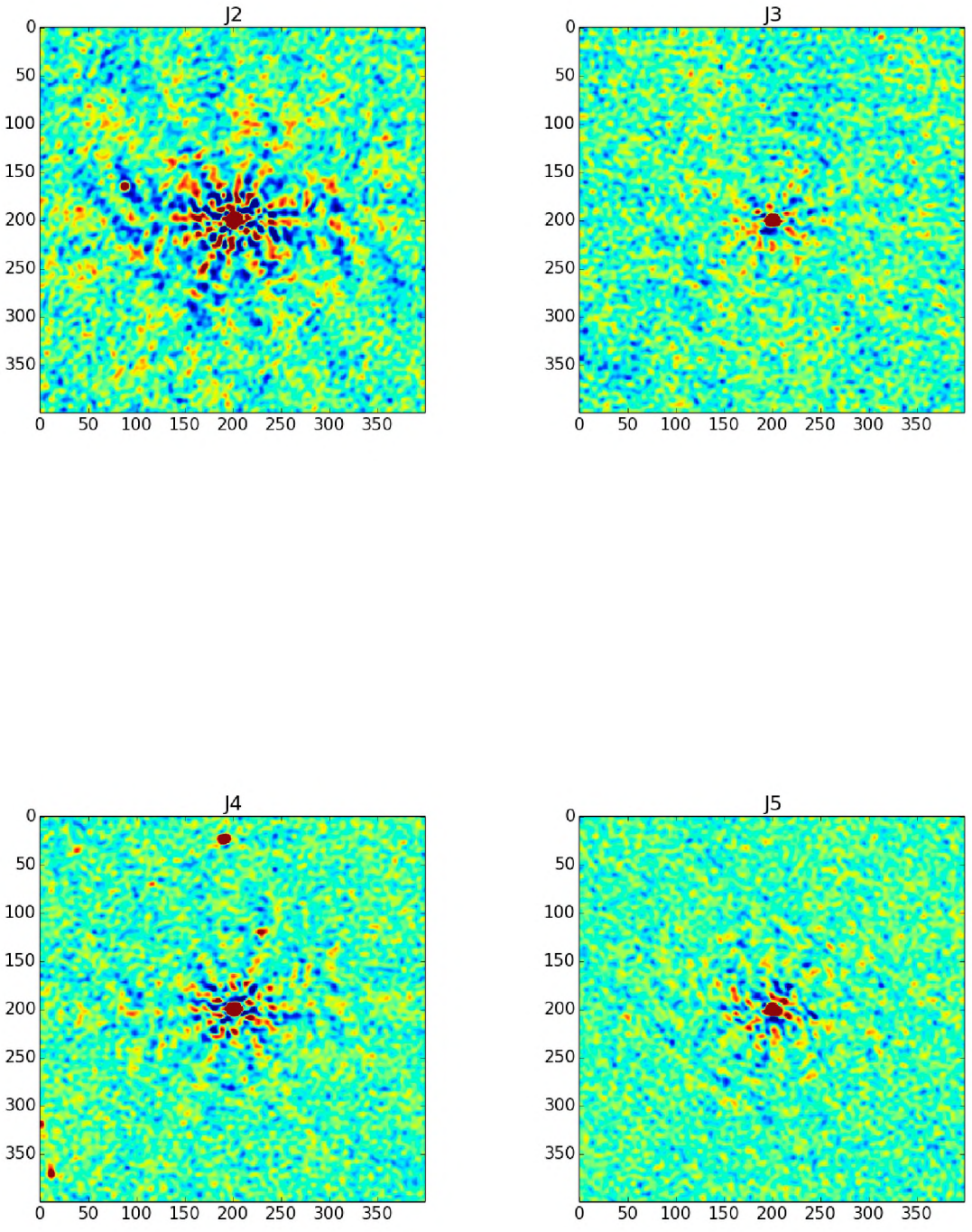


FIGURE 3.4: JVLA observation. Zoomed images of sources requiring DD calibration as identified by our algorithm.

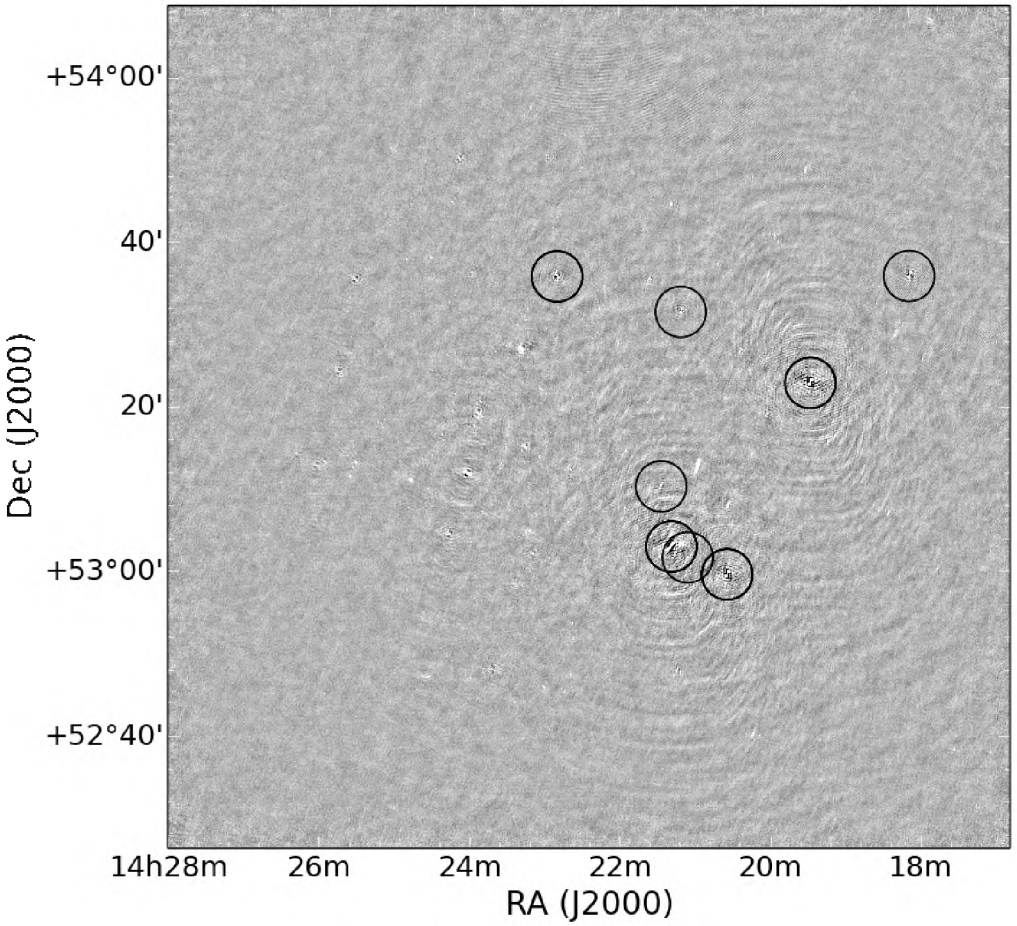
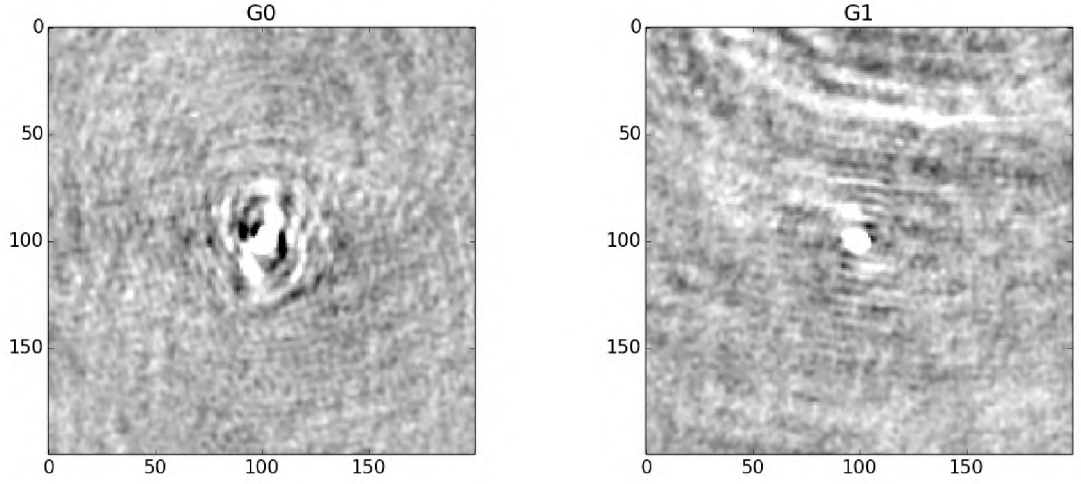


FIGURE 3.5: GMRT field. The sources shown are those selected by our algorithm as sources that require DD calibration solutions. Image credit: Roger Deane (Rhodes University).



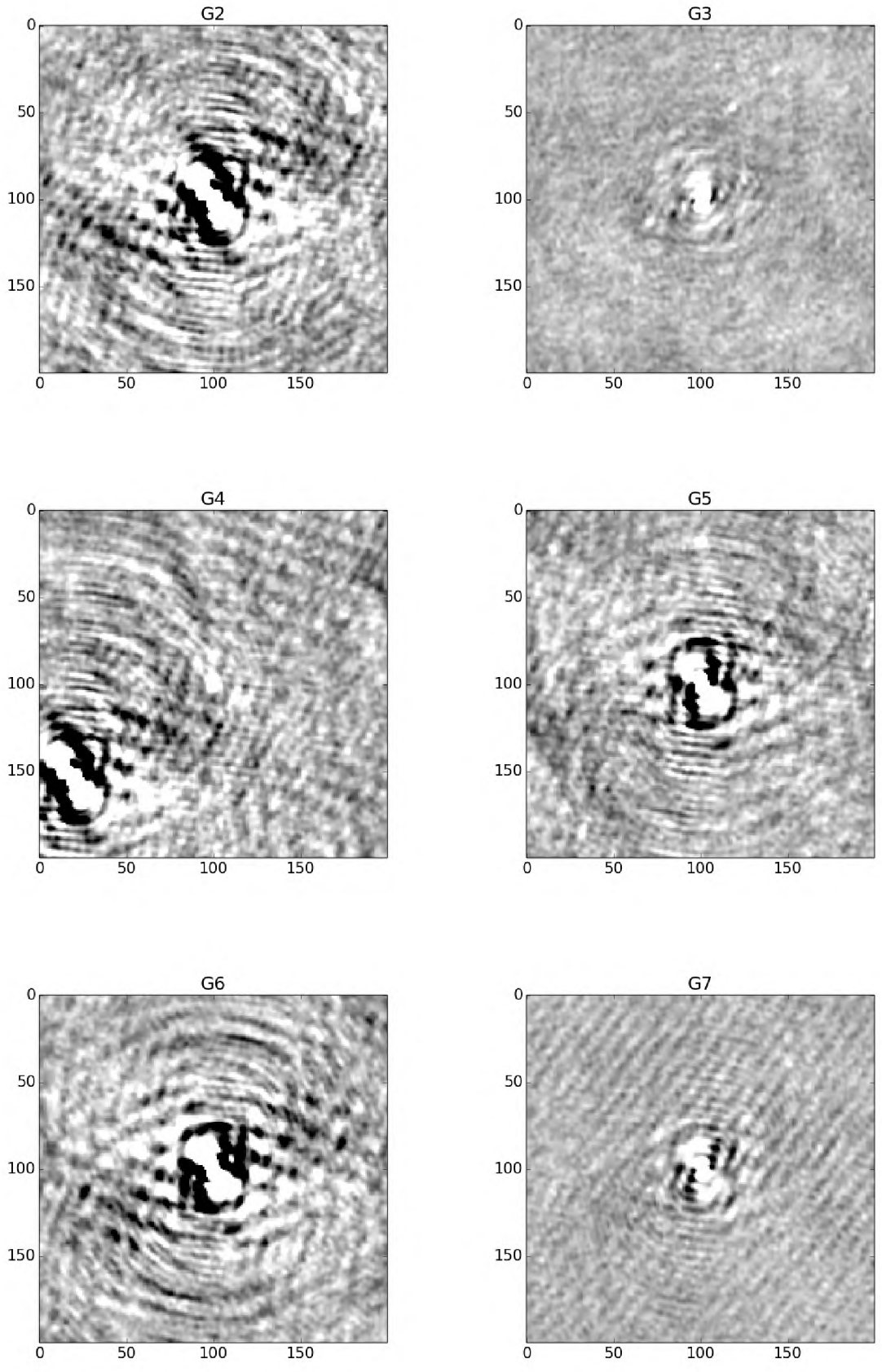


FIGURE 3.6: Sources that require DD calibration in the GMRT data.

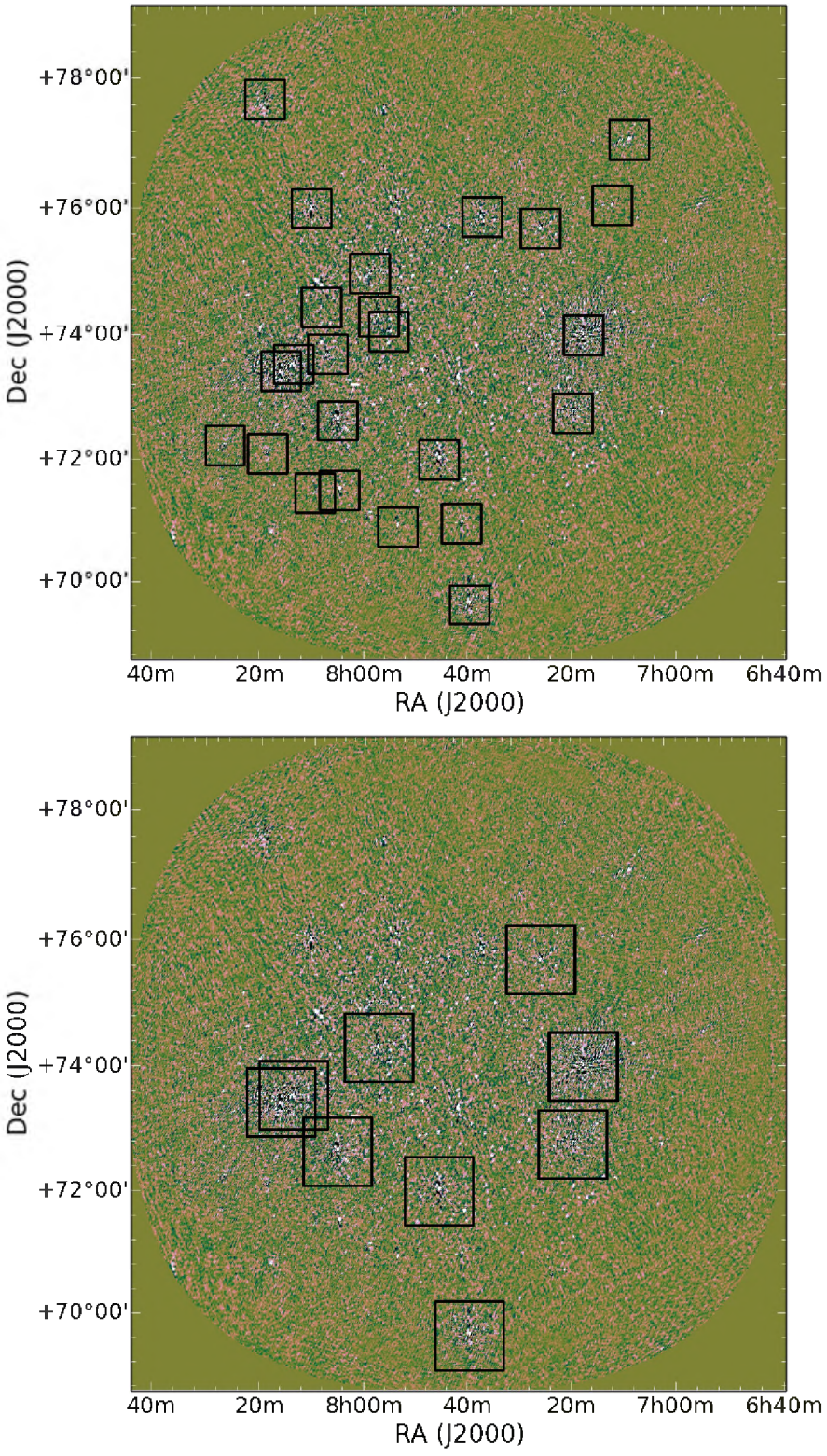
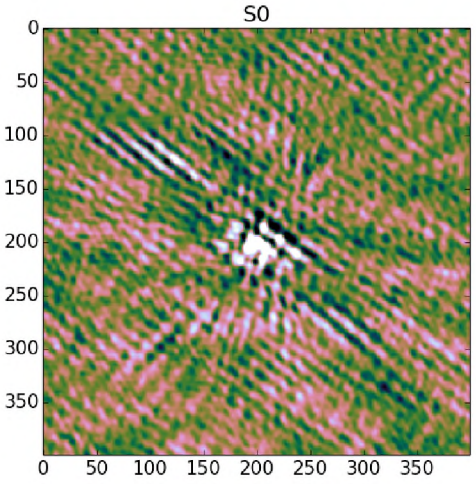
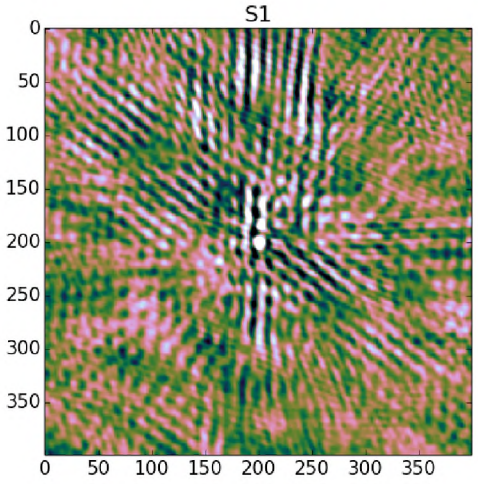
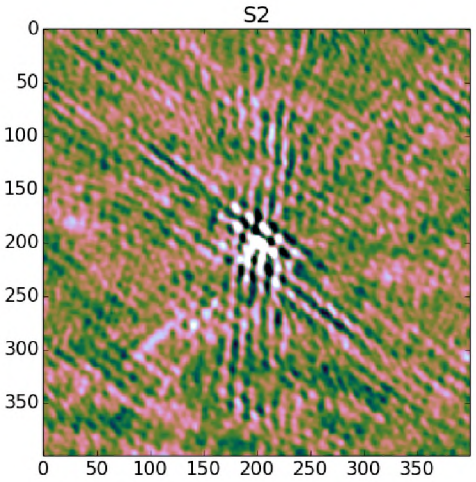
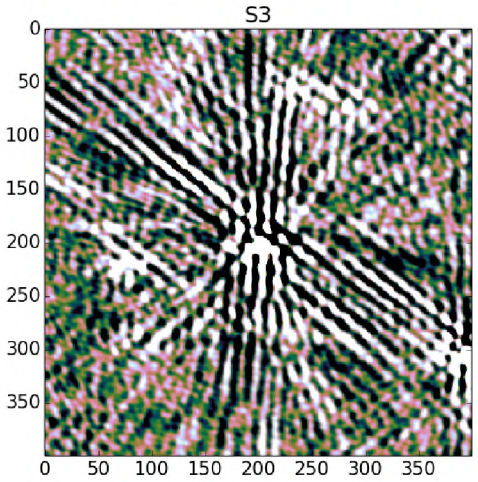


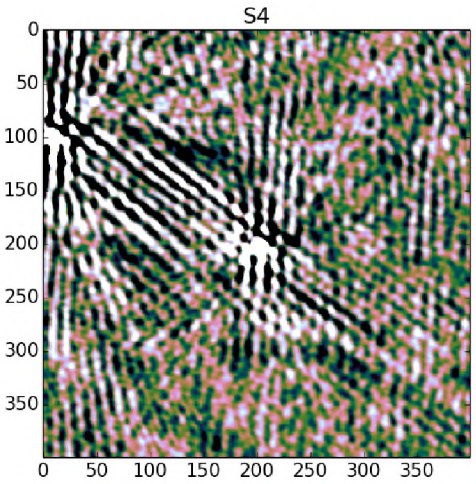
FIGURE 3.7: LOFAR observation. *Top panel:* Squares centred about a source indicates a source or a region that require DD calibration solutions as classified by our algorithm, for parameter setting described in Table 3.1. *Bottom panel:* is the classification when the number of nearby negative detections is set to 20. Image credit: Ian Heywood (CSIRO) [1]

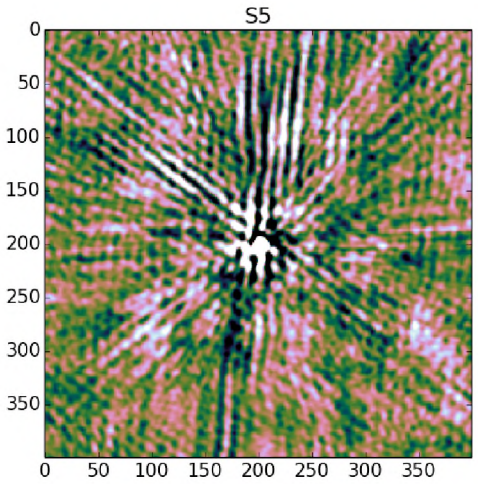

S0
0
50
100
150
200
250
300
350
0 50 100 150 200 250 300 350

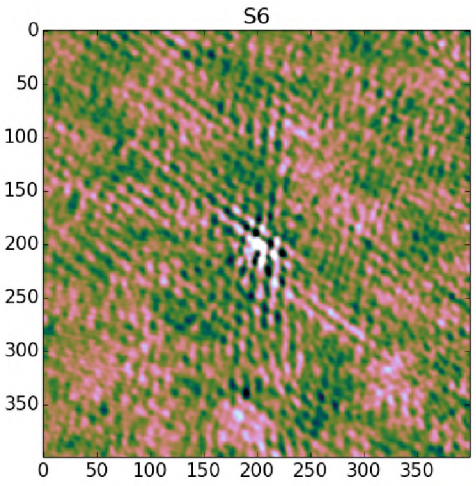

S1
0
50
100
150
200
250
300
350
0 50 100 150 200 250 300 350

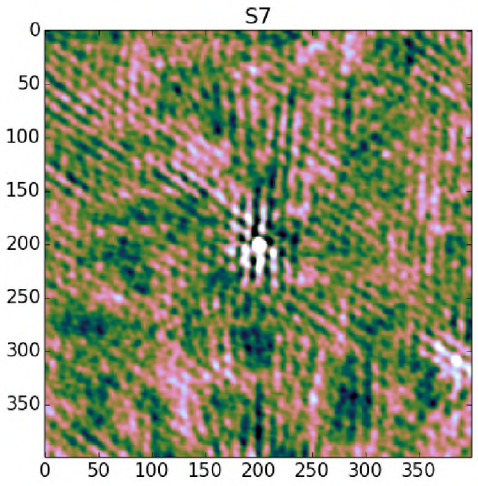

S2
0
50
100
150
200
250
300
350
0 50 100 150 200 250 300 350

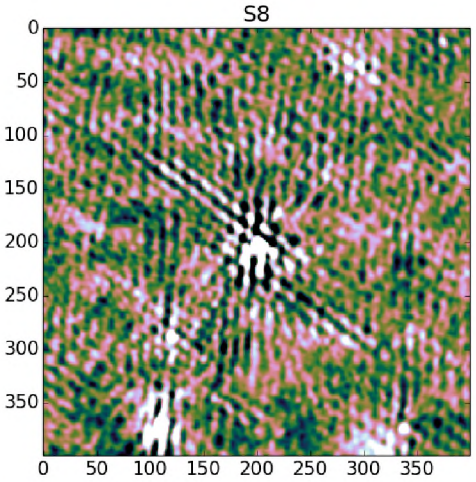

S3
0
50
100
150
200
250
300
350
0 50 100 150 200 250 300 350

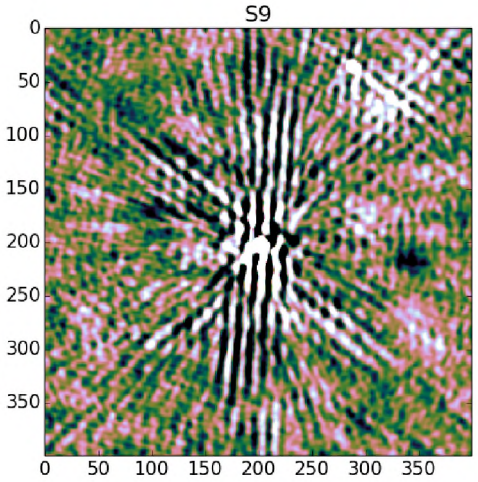

S4
0
50
100
150
200
250
300
350
0 50 100 150 200 250 300 350

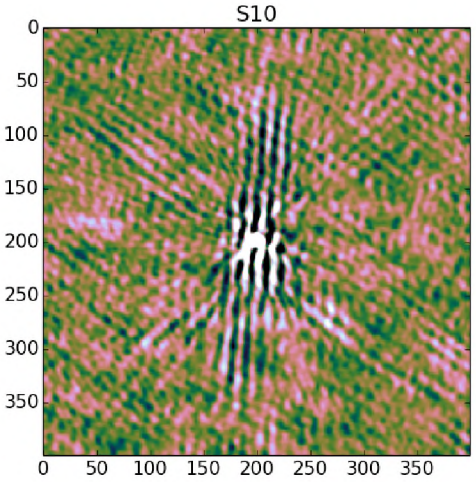

S5
0
50
100
150
200
250
300
350
0 50 100 150 200 250 300 350

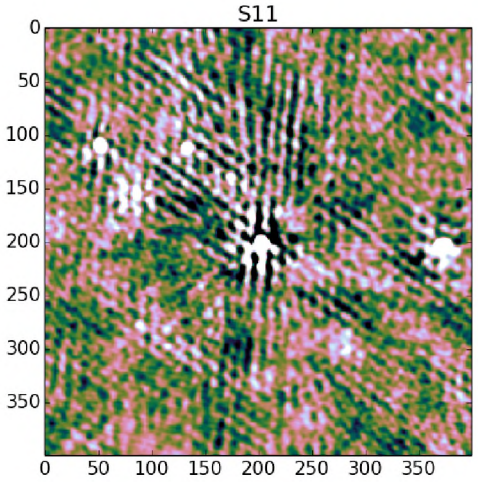

S6
0
50
100
150
200
250
300
350
0 50 100 150 200 250 300 350

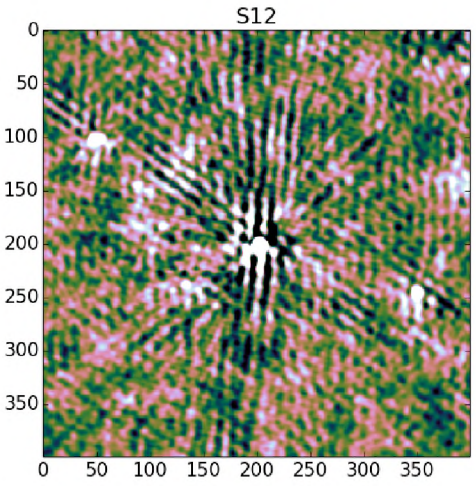

S7
0
50
100
150
200
250
300
350
0 50 100 150 200 250 300 350

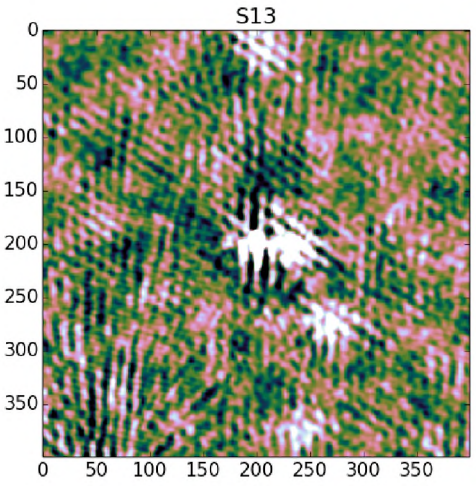

S8
0
50
100
150
200
250
300
350
0 50 100 150 200 250 300 350

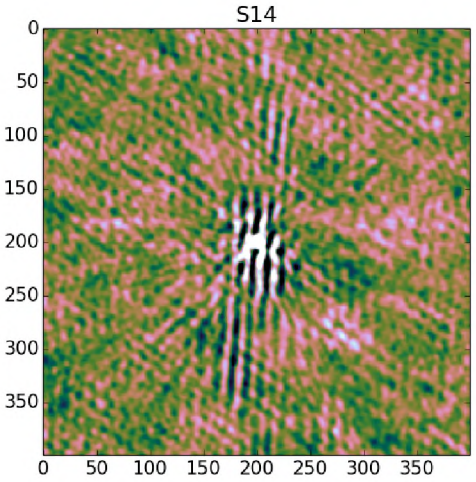

S9
0
50
100
150
200
250
300
350
0 50 100 150 200 250 300 350

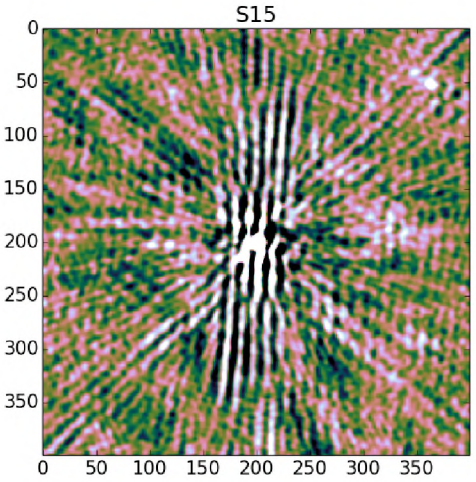

S10
0
50
100
150
200
250
300
350
0 50 100 150 200 250 300 350

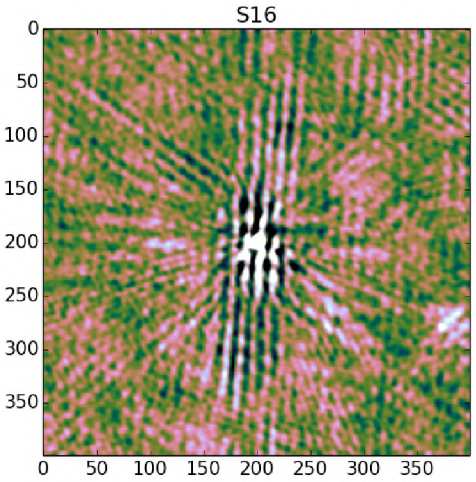

S11
0
50
100
150
200
250
300
350
0 50 100 150 200 250 300 350

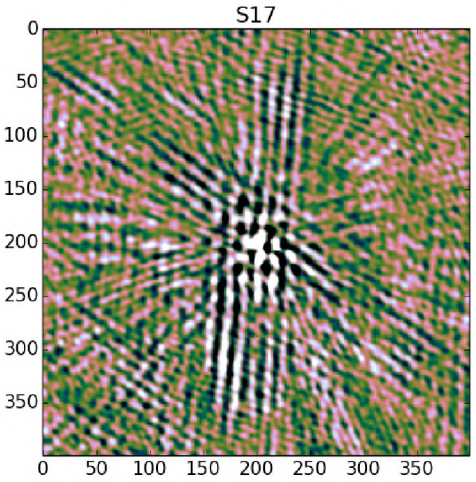

S12
0
50
100
150
200
250
300
350
0 50 100 150 200 250 300 350


S13
0
50
100
150
200
250
300
350
0 50 100 150 200 250 300 350


S14
0
50
100
150
200
250
300
350
0 50 100 150 200 250 300 350


S15
0
50
100
150
200
250
300
350
0 50 100 150 200 250 300 350


S16
0
50
100
150
200
250
300
350
0 50 100 150 200 250 300 350


S17
0
50
100
150
200
250
300
350
0 50 100 150 200 250 300 350

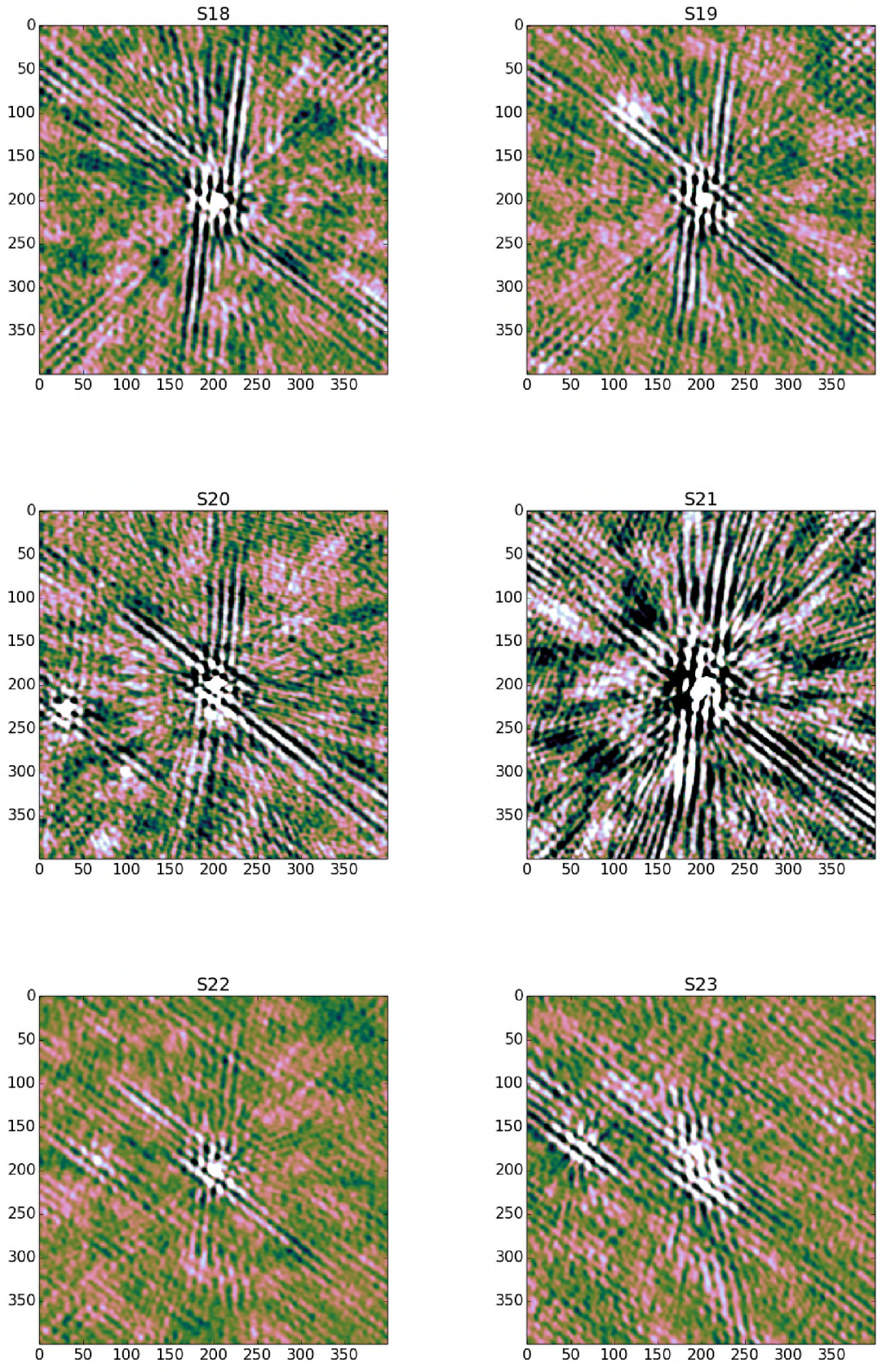


FIGURE 3.8: The sources that require DD calibration in the LOFAR image.

3.3 Classification on Simulated Data

In this section, I intend to give a comparison between a classification made by eye and that made by `sourcery` on JVLA simulated data. I have chosen 20 images from Chapter 2 at random. The confusion matrices are shown below. Note that, I refer to the classification made by eye as actual and the one made by `sourcery` as the predictions. I ran `sourcery` using similar settings as in Table 3.1.

TABLE 3.2: Actual classification made by eye compared the predictions made by `sourcery`.

		Predictions	
		True	False
Actual	True	4	0
	False	0	215

		Predictions	
		True	False
Actual	True	3	1
	False	1	297

		Predictions	
		True	False
Actual	True	2	0
	False	2	253

		Predictions	
		True	False
Actual	True	1	0
	False	3	157

		Predictions	
		True	False
Actual	True	3	1
	False	1	294

		Predictions	
		True	False
Actual	True	4	2
	False	2	192

		Predictions	
		True	False
Actual	True	6	1
	False	2	239

		Predictions	
		True	False
Actual	True	1	0
	False	2	224

		Predictions	
		True	False
Actual	True	4	2
	False	0	260

		Predictions	
		True	False
Actual	True	3	1
	False	0	291

		Predictions	
		True	False
Actual	True	3	0
	False	0	229

		Predictions	
		True	False
Actual	True	3	0
	False	0	194

		Predictions	
		True	False
Actual	True	2	2
	False	0	270

		Predictions	
		True	False
Actual	True	3	1
	False	1	237

		Predictions	
		True	False
Actual	True	4	0
	False	1	213

		Predictions	
		True	False
Actual	True	2	0
	False	1	205

		Predictions	
		True	False
Actual	True	4	0
	False	3	244

		Predictions	
		True	False
Actual	True	3	1
	False	0	247

		Predictions	
		True	False
Actual	True	3	0
	False	2	227

		Predictions	
		True	False
Actual	True	3	0
	False	1	229

To quantify the performance of the classification I have used three terms named accuracy, miss rate (also known as false negative rate or FNR) and fall out rate (known as false positive rate or FPR). These terms are defined mathematically as;

$$\text{accuracy} = \frac{\text{T} - \text{F}}{\text{T} + \text{F} + \text{FP} + \text{FN}}, \quad \text{FNR} = \frac{FN}{FN + T}$$
$$\text{and FPR} = \frac{FP}{FP + F}.$$

The miss rate defines a rate at which the classifier misclassifies a true detection as false, while the fall out gives the rate at which false detections are classified as true. For this classification, we obtained the average accuracy, miss rate and fall out rate of 99.27%, 12.79% and 0.498%, respectively.

3.4 Discussion

With the exception of sources $G1$ and $G4$ in the GMRT dataset, which were falsely identified due to their proximity to a source dominated by DD effects, all the other sources in all the dataset were properly classified. This is also the case for the misclassification for simulated data.

However, it can be seen from the top and bottom image of Fig. 3.7 that the classification is highly dependent on the choice of the number of negative detections threshold, M, with a higher threshold seemingly identifying the sources that are subject to extreme DD effects. Thus, the choice of M and other thresholds could significantly affect the outcome of the classification. An example of this is the choice of source finding thresholds. The lower thresholds will result in a large number of detections and thus may misclassify sources similar to $G1$ and $G4$ in the GMRT dataset, while the higher detection thresholds will result in fewer detections and therefore may overlook other sources of interest. This suggests that our current approach of deriving the thresholds, which is arbitrary, needs to be refined so that it overcomes situations similar to the one described above.

This algorithm has been shown to work well on all the datasets from different telescopes. This is because none of our assumptions depend on any specific telescope configurations.

The algorithm has automated the identification of sources that require DD calibration. Thus it has eliminated the need to halt the calibration pipeline and visually inspect the image to select these sources.

Summary

This dissertation addresses two of the automation problems associated with classification. First, we developed an algorithm, based on prior work by Serra et al. (2012), that distinguishes spurious emission from astrophysical emission. We then implemented and tested a software tool called `sourcery` based on this algorithm. This was followed by an algorithm that identifies sources that are subject to extreme DD effects and thus require DD calibration.

For the first algorithm, we implemented the negative detection method which uses a probability measure known as the reliability to quantify and classify the above emissions into two distinct classes. Our focus was particularly on calibration artefacts, thus, we reduced or removed (ideally) the noise peaks from the classification by smoothing the data and masking the peaks that were above a certain threshold. For the second algorithm, we used detection properties such as brightness, local variance, PSF correlation, and the number of nearby negative peaks, to identify astrophysical emission that produces calibration artefacts in images. The latter algorithm has been shown to work well for different datasets taken by KAT-7, GMRT, JVLA and LOFAR instruments.

The increase in the number of parameters has improved the classification just as Serra et al. (2012) initially postulated. The extra parameters are the local variance, PSF correlation factor and the number of nearby sources (the previously used parameters were source area, peak flux and total flux). It was observed with simulated data that some of the parameter combinations contributed negatively to the classification. However, the increase in the number of parameters has allowed the approach to become robust in such a way that it reduced the rate of misclassification, thus increasing the completeness. The high completeness was also evident on the 3C147 field data and also for simulated data. Our algorithm obtained sky models that were better than those found using `PyBDSM` alone. For

instance, our final model for the $3C147$ field had a much higher completeness compared to `PyBDSM` final model with only 14 of which 6 were missed due to edge effects, i.e would have been found with a larger image size. Meanwhile, `PyBDSM` missed a total of 39 sources.

Using `sourcery` in the $3C147$ calibration pipeline, we arrive at a final DR of 3.145M, which is slightly less than the 3.27M obtained previously with `PyBDSM`. This is not a significant reduction, but still needs to be explained. Note that the `PyBDSM`-based pipeline contains multiple source detection steps (and is thus much slower), but effectively this means that fainter sources have their properties measured at a later stage of the pipeline, when the images are correct. The single-step `sourcery`-based pipeline detects most of the sources in one step, which means their measured parameters are potentially less accurate (since the image at this stage is less correct). We hypothesise that this what led to a slightly reduced DR. Note that to improve catalogue correctness, one could re-run the source finder on the final images. We also suggest the use of optimisation techniques such as MCMC and GA to perform source modelling. In any case, our algorithm has shown to provide high-completeness catalogues, and has also automated the identification of sources that requires DD calibration. Our algorithm working together with `PyBDSM` will be useful for pipelined calibration of future continuum observations.

In the future, one of the most important things would be to find a robust way of determining the thresholds. Particularly, a threshold for the number of negative peaks, thought to depend on the sensitivity of the telescope and the choice of source finding thresholds. Another threshold is that of the reliability. The latter needs to be tested on additional datasets other than that of the JVLA.

Ian Heywood proposed an alternative approach to source-artefact discrimination, which we did not explore in this work. His approach hinges on the fact that calibration artefacts are to some extent due to the primary beam which scales with frequency, they too should scale and move in frequency. That is, as the beam gets smaller and narrower at higher frequencies, calibration artefacts should also move inward towards their parent astrophysical source [2] and vice versa. Meanwhile, the positions of the astrophysical sources are independent of the beam. Except that their signal-to-noise ratio are dependent on the frequency of the observation. Thus, their different behaviours in frequency can be used to distinguish them.

[2] An astrophysical source responsible for them (calibration artefacts).

This method can be potentially very powerful for wide-band multi-frequency observations.

Appendix A

Sourcery Implementation

The algorithms described in this thesis are implemented by the `sourcery` tool (`https://github.com/radio-astro/sourcery`). Sourcery is an open-source project hosted on github. The tool is freely available, and is already distributed as part of the `radio-astro` packaging scheme developed by Gijs Molenaar (UvA/Rhodes/SKA SA), and thus can be installed via pip.

Sourcery is written in Python. The entirety of the source code was developed by the author (with minor inputs from supervisors) as part of the present work, and should be considered an effective part of the submitted thesis (release 1.2.6 at time of writing).

Installation

There are two ways to install `sourcery`:
1. Direct build:

```
$ git clone https://github.com/radio-astro/sourcery
$ cd sourcery
$ pip install
```

2. Using `Pip`:

```
pip install sourcery
```

Other Software Dependencies

To be able to run `sourcery` you need the following software packages:

`numpy`, `pylab`, `matplotlib`, `pyfits`, `scipy`, `tigger`, `pybdsm` (LOFAR software) and `astLib`.
These are available in the standard Ubuntu repository and/or `https://launchpad.net/~radio-astro/+archive/ubuntu/main`.

Requirements

The files needed to run `sourcery` are:
1. An image of your observation in FITS format (Non-optional)
2. PSF image in FITS format (Optional)

Commands to Run Sourcery

The following are the ways in which you can run `sourcery`:
1. You can specify everything on the terminal, for example:

```
$ sourcery -i image.fits -p psf_image.fits -od sunshine -pref
   datalsm -ppix=2 --pisl=3
```

The above command runs `sourcery` on the image called *image.fits* (NB: if the image is not in your current working directory you need to specify a correct path to your image). We have specified a PSF image (*psf_image.fits*) using *-p* and the out directory using *-od*. The latter ensures that all the files produced are directed into a directory called *sunshine*. You can also specify a prefix e.g in the above example all the generated files will have *datalsm* as a prefix. Furthermore, *-ppix* and *-pisl* are the source finding detection thresholds to be used for the positive image. To see other options in `sourcery` make use of the help command line:

```
$ sourcery -h
```

2. You can use `Json` **config** file (shown below) using the command:

```
$ sourcery -jc config.json


 {
   "imagename": "image.fits",
   "psfname" : "psf_image.fits",
   "prefix" : "datalsm",
   "outdir" : "sunshine",
   "reliability" : {
      "sourcefinder_name" : "pybdsm",
      "makeplots" : true,
      "do_psf_corr" : true,
      "do_local_var" : false,
      "do_nearsources": false,
      "psf_corr_region" : 5,
      "local_var_region" : 10,
      "rel_excl_src" : null,
      "pos_smooth" : 2,
      "neg_smooth" : 2,
      "loglevel" : 0,
      "thresh_isl" : 20,
      "thresh_pix" : 25,
      "neg_thresh_isl" : 2,
      "neg_thresh_pix" : 3
   },
   "dd_tagging" : {
      "enable": true,
      "snr_thresh" : 40,
      "local_thresh" : 0.8,
      "high_corr_thresh" : 0.4,
      "negdetec_region" : 10,
      "negatives_thresh" : 4,
      "phasecenter_excl_radius" : 100,
      "loglevel" : 0
   },
   "source_finder_opts": {
   }
```

Definitions of the Parameters in the Config File

The reliability estimations, *reliability*:

• **makeplots** (boolean)- make reliability plots to show the distribution of the detections across the different parameter spaces.

• **do_psf_corr** (boolean) - add PSF correlation as a parameter.

• **do_local_var** (boolean) - add the local variance.

• **do_nearsources** (boolean) - add the number of nearby sources.

• **psf_corr_region** (float) - the size of the region to compute the correlation, in PSF sizes.

• **local_var_region** (float) - the size of the region to compute the local variance, in PSF sizes.

• **rel_excl_src** (list) - a list of regions (containing the detections) to remove from the reliability estimations, e.g ["0.0,35,0.5"] will remove sources within a radius 0.5° from RA=0.0°, DEC=35°.

• **pos_smooth** (float) - a threshold to create a mask, e.g if 2 then all the pixels in smoothed image with value less than 2× rms are masked (rms is the noise of an image).

• **neg_smooth** (float) - similar to *pos_smooth* but applied to the inverted image.

• **thresh_isl** and **thresh_pix** (float) - the detection thresholds for forming island and model fitting (for the positive image).

• **neg_thresh_isl** and **neg_thresh_pix** (float) - similar to above thresholds but applied to the inverted image.

The selection of sources that require DD calibration, *dd_tagging*:

• **snr_thresh**, **local_thresh**, **high_corr_thresh** and **negatives_thresh** (float) - are the thresholds for high signal-to-noise ratio, local variance, correlation factor and the number of nearby negative detections, respectively.

• **negdetec_region** (float)- the size of the region to count the number of nearby negative detections, in PSF sizes.

• **phasecenter_excl_radius** - the radius w.r.t to the phase centre to exclude from *dd_tagging*.

3. You can use both the terminal and **config** file. This is particularly, more efficient

when working with multiple images. First you need to replace the following lines in the **config** file as shown below (and the rest can be left unchanged) and then run the command that follows:

```
{
    "imagename": "<input FITS image>",
    "psfname" : null,
    ... }
```

```
$ sourcery -i image.fits -p psf_image.fits -jc config.json
```

Provided you have more than one image say *image0.fits*, *image1.fits* and *image2.fits* and *psf_image0.fits*, *psf_image1.fits* and *psf_image2.fits* then you can specify these as:

```
$ sourcery -i image0.fits,image1.fits,image2.fits -p psf_image0.fits,
   psf_image1.fits,psf_image2.fits -jc config.json
```

If the PSF image is common to all of the images, then you can specify it once and it will be used for all the images, otherwise the number PSFs must equal that of images provided and must be ordered accordingly. For a common PSF the command to use is:

```
$ sourcery -i image0.fits,image1.fits,image2.fits -p psf_image.fits
   -jc config.json
```

Parameters

The local variance, PSF correlation and the number of nearby sources are optional. And whether to include them into the reliability estimations is up to the user. On the terminal you can add them using *-apsf, -alv, -dn* (see the help command above) or in the **config** file set their options to true. Take note that the parameters source area, peak and total flux are not optional.

Bibliography

K. M. B. Asad, L. V. E. Koopmans, V. Jelić, et al. Polarization leakage in epoch of reionization windows - I. Low Frequency Array observations of the 3C196 field. *MNRAS*, 451:3709–3727, August 2015. doi: 10.1093/mnras/stv1107.

A. Bos. *On instrumental effects in spectral line synthesis observations.* PhD thesis, Ph. D. thesis, University of Leiden (1985), 1985.

A. Bowman, A. W & Azzalini. *Applied Smoothing Techniques for Data Analysis.* London: Oxford University Press, 1997.

W Brisken. Using grasp8 to study the vla beam, evla mem 58. May 2003. URL `http://www.aoc.nrao.edu/evla/geninfo/memoseries/evlamemo58.pdf`.

C. Carilli. Square Kilometre Array key science: a progressive retrospective. *Advancing Astrophysics with the Square Kilometre Array (AASKA14)*, art. 171, April 2015.

C. L. Carilli and S. Rawlings. Motivation, key science projects, standards and assumptions. *Nucleic Acids Research*, 48:979–984, December 2004. doi: 10.1016/j.newar.2004.09.001.

S. Wijnholds, S. van der Tol, R. Nijboer, and A.-J. van der Veen. Calibration challenges for future radio telescopes. *IEEE Signal Processing Magazine*, 27: 30–42, January 2010. doi: 10.1109/MSP.2009.934853.

B. G Clark. Coherence in radio astronomy. In *Synthesis Imaging in Radio Astronomy II, ASP Conference Series*, 2009. doi: 10.1007/978-3-540-85122-6.

J. J. Condon, W. D. Cotton, E. W. Greisen, et al. The NRAO VLA Sky Survey. *AJ*, 115:1693–1716, May 1998. doi: 10.1086/300337.

T. Cornwell and E. B. Fomalont. Self-calibration. In *Synthesis Imaging in Radio Astronomy II, ASP Conference Series*, volume 180, pages 187–199, 1999.

T. J. Cornwell and P. N. Wilkinson. A new method for making maps with unstable radio interferometers. *MNRAS*, 196:1067–1086, September 1981.

B. F Burke and F Graham-Smith. *An Introduction to Radio Astronomy*. Cambridge University Press, second edition, 2002.

D. S Briggs, D. R Schwab, and R. S Sramek. Imaging. In *Synthesis Imaging in Radio Astronomy II, ASP Conference Series*, pages 151–173, 1999.

N. Weir. A Multi-Channel Method of Maximum Entropy Image Restoration. In D. M. Worrall, C. Biemesderfer, and J. Barnes, editors, *Astronomical Data Analysis Software and Systems I*, volume 25 of *Astronomical Society of the Pacific Conference Series*, page 186, 1992.

T. R. Bontekoe, E. Koper, and D. J. M. Kester. Pyramid maximum entropy images of IRAS survey data. *A&A*, 284:1037–1053, April 1994.

T. J. Cornwell. Multiscale CLEAN Deconvolution of Radio Synthesis Images. *IEEE Journal of Selected Topics in Signal Processing*, 2:793–801, November 2008. doi: 10.1109/JSTSP.2008.2006388.

F. R. Schwab. Relaxing the isoplanatism assumption in self-calibration; applications to low-frequency radio interferometry. *AJ*, 89:1076–1081, July 1984. doi: 10.1086/113605.

B. G. Clark. An efficient implementation of the algorithm 'CLEAN'. *A&A*, 89: 377, September 1980.

M. P. van Haarlem, M. W. Wise, A. W. Gunst, et al. LOFAR: The LOw-Frequency ARray. *A&A*, 556:A2, August 2013. doi: 10.1051/0004-6361/201220873.

D. R. DeBoer, R. G. Gough, J. D. Bunton, et al. Australian SKA Pathfinder: A High-Dynamic Range Wide-Field of View Survey Telescope. *IEEE Proceedings*, 97:1507–1521, August 2009. doi: 10.1109/JPROC.2009.2016516.

P. E. Dewdney, W. Turner, R. Millenaar, et al. Ska1 system baseline design. 12: 23–24, 2013. doi: SKA-TEL-SKO-DD-001.

T. Duong. May 2001. URL `http://www.mvstat.net/tduong/research/seminars/seminar-2001-05.pdf`.

T. Duong. Bandwidth selectors for multivariate kernel density estimation. October 2004. URL `http://www.mvstat.net/tduong/research/publications/duong-2005-thesis.pdf`.

M. Felli and R. E. Spencer. *Very Long Baseline Interferometry Techniques and Applications*. Kluwer Academic Publishers, 1989. doi: 10.1007/978-94-009-2428-4.

B. B Gertman. *Stat Primer Version 6.4*. Retrieved on 12 December 2015. URL `http://www.sjsu.edu/faculty/gerstman/StatPrimer/correlation.pdf`. Part B Section 14 Correlation.

T. L. Grobler, C. D. Nunhokee, O. M. Smirnov, A. J. van Zyl, and A. G. de Bruyn. Calibration artefacts in radio interferometry - I. Ghost sources in Westerbork Synthesis Radio Telescope data. *MNRAS*, 439:4030–4047, April 2014. doi: 10.1093/mnras/stu268.

J. P. Hamaker, J. D. Bregman, and R. J. Sault. Understanding radio polarimetry. I. Mathematical foundations. *AAPS*, 117:137–147, May 1996.

P. J. Hancock, T. Murphy, B. M. Gaensler, A. Hopkins, and J. R. Curran. Compact continuum source finding for next generation radio surveys. *MNRAS*, 422:1812–1824, May 2012. doi: 10.1111/j.1365-2966.2012.20768.x.

B. E Hasen. *Lecture Notes on Non-Parametrics*. 2009. URL `http://www.ssc.wisc.edu/~bhansen/718/NonParametrics1.pdf`.

G. H. Heald, R. F. Pizzo, E. Orrú, et al. The LOFAR Multifrequency Snapshot Sky Survey (MSSS). I. Survey description and first results. *A&A*, 582:A123, October 2015. doi: 10.1051/0004-6361/201425210.

J. A. Högbom. Aperture Synthesis with a Non-Regular Distribution of Interferometer Baselines. *AAPS*, 15:417, June 1974.

A. M. Hopkins, C. J. Miller, A. J. Connolly, et al. A New Source Detection Algorithm Using the False-Discovery Rate. *AJ*, 123:1086–1094, February 2002. doi: 10.1086/338316.

A. M. Hopkins, M. T. Whiting, N. Seymour, et al. The ASKAP/EMU Source Finding Data Challenge. *PASA*, 32:e037, October 2015. doi: 10.1017/pasa.2015.37.

N. Hurley-Walker, J. Morgan, R. B. Wayth, et al. The Murchison Widefield Array Commissioning Survey: A Low-Frequency Catalogue of 14 110 Compact Radio Sources over 6 100 Square Degrees. *PASA*, 31:e045, November 2014. doi: 10.1017/pasa.2014.40.

M. T. Huynh, A. Hopkins, R. Norris, et al. The Completeness and Reliability of Threshold and False-discovery Rate Source Extraction Algorithms for Compact Continuum Sources. *PASA*, 29:229–243, December 2012. doi: 10.1071/AS11026.

D. C. Jacobs, J. E. Aguirre, A. R. Parsons, et al. New 145 MHz Source Measurements by PAPER in the Southern Sky. *ApJL*, 734:L34, June 2011. doi: 10.1088/2041-8205/734/2/L34.

S. Johnston, R. Taylor, M. Bailes, et al. Science with ASKAP. The Australian square-kilometre-array pathfinder. *Experimental Astronomy*, 22:151–273, December 2008. doi: 10.1007/s10686-008-9124-7.

J. L. Jonas. MeerKAT - The South African Array With Composite Dishes and Wide-Band Single Pixel Feeds. *IEEE Proceedings*, 97:1522–1530, August 2009. doi: 10.1109/JPROC.2009.2020713.

R. C. Jones. *A new calculus for the treatment of optical systems, I. Description and Discussion of the Calculus.* Journal of the Optical Society of America, 1941.

A. Kemball and A. Martinsek. Bootstrap Resampling as a Tool for Radio Interferometric Imaging Fidelity Assessment. *AJ*, 129:1760–1775, March 2005. doi: 10.1086/428006.

M. Masias, J. Freixenet, X. Lladó, and M. Peracaula. A review of source detection approaches in astronomical images. *MNRAS*, 422:1674–1689, May 2012. doi: 10.1111/j.1365-2966.2012.20742.x.

C. J. Miller, C. Genovese, R. C. Nichol, et al. Controlling the False-Discovery Rate in Astrophysical Data Analysis. *AJ*, 122:3492–3505, December 2001. doi: 10.1086/324109.

M. Mitra, S. Makhathini, G. Foster, O. Smirnov, and R. Perley. Incorporation of antenna primary beam patterns in radio-interferometric data reduction to produce wide-field, high-dynamic-range images. In *Electromagnetics in Advanced Applications (ICEAA), 2015 International Conference on*, pages 494–497, Sept 2015. doi: 10.1109/ICEAA.2015.7297163.

N. Mohan and D. Rafferty. PyBDSM: Python Blob Detection and Source Measurement. Astrophysics Source Code Library, February 2015.

P. J. Napier. The EVLA Project: Ten Times More Capability for the VLA. In D. C. Backer, J. M. Moran, and J. L. Turner, editors, *Revealing the Molecular Universe: One Antenna is Never Enough*, volume 356 of *Astronomical Society of the Pacific Conference Series*, page 65, December 2006.

J. E. Noordam and O. M. Smirnov. The MeqTrees software system and its use for third-generation calibration of radio interferometers. *A&A*, 524:A61, December 2010. doi: 10.1051/0004-6361/201015013.

J. E. Noordam. LOFAR calibration challenges. In J. M. Oschmann, Jr., editor, *Ground-based Telescopes*, volume 5489 of *Proc. SPIE*, pages 817–825, October 2004. doi: 10.1117/12.544262.

R. J. van Weeren, W. L. Williams, M. J. Hardcastle, et al. LOFAR Facet Calibration. *ApJS*, 223:2, March 2016. doi: 10.3847/0067-0049/223/1/2.

H. T. Intema, S. van der Tol, W. D. Cotton, A. S. Cohen, I. M. van Bemmel, and H. J. A. Röttgering. Ionospheric calibration of low frequency radio interferometric observations using the peeling scheme. I. Method description and first results. *A&A*, 501:1185–1205, July 2009. doi: 10.1051/0004-6361/200811094.

S. Kazemi, S. Yatawatta, S. Zaroubi, P. Lampropoulos, A. G. de Bruyn, L. V. E. Koopmans, and J. Noordam. Radio interferometric calibration using the SAGE algorithm. *MNRAS*, 414:1656–1666, June 2011. doi: 10.1111/j.1365-2966.2011.18506.x.

R. P. Norris, J. Afonso, D. Bacon, et al. Radio Continuum Surveys with Square Kilometre Array Pathfinders. *PASA*, 30:e020, March 2013. doi: 10.1017/pas.2012.020.

A. R. Offringa, B. McKinley, N. Hurley-Walker, et al. WSCLEAN: an implementation of a fast, generic wide-field imager for radio astronomy. *MNRAS*, 444:606–619, October 2014. doi: 10.1093/mnras/stu1368.

R. Perley and O. Smirnov. High dynamic range imaging. In *The Radio Universe Ger's Feest, 2013 Gerfeest Conference in honor of Ger De Bruyn - Astron*, Nov 2013. URL `http://www.astron.nl/gerfeest/presentations/perley.pdf`.

R. J. Sault, J. P. Hamaker, and J. D. Bregman. Understanding radio polarimetry. II. Instrumental calibration of an interferometer array. *AAPS*, 117:149–159, May 1996.

D.W Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization.* New York: John Wiley, 1992.

P. Serra, R. Jurek, and L. Flöer. Using Negative Detections to Estimate Source-Finder Reliability. *PASA*, 29:296–300, February 2012. doi: 10.1071/AS11065.

P. Serra, T. Westmeier, N. Giese, et al. SOFIA: a flexible source finder for 3D spectral line data. *MNRAS*, 448:1922–1929, April 2015. doi: 10.1093/mnras/stv079.

O. M. Smirnov. Revisiting the radio interferometer measurement equation. I. A full-sky Jones formalism. *A&A*, 527:A106, March 2011a. doi: 10.1051/0004-6361/201016082.

O. M. Smirnov. Revisiting the radio interferometer measurement equation. II. Calibration and direction-dependent effects. *A&A*, 527:A107, March 2011b. doi: 10.1051/0004-6361/201116434.

O. M. Smirnov. Revisiting the radio interferometer measurement equation. III. Addressing direction-dependent effects in 21 cm WSRT observations of 3C 147. *A&A*, 527:A108, March 2011c. doi: 10.1051/0004-6361/201116435.

R. A Thompson. Fundamentals to radio interferometry. In *Synthesis Imaging in Radio Astronomy II, ASP Conference Series*, pages 9–36, 1999.

S. J. Tingay, R. Goeke, J. D. Bowman, et al. The Murchison Widefield Array: The Square Kilometre Array Precursor at Low Radio Frequencies. *PASA*, 30: e007, January 2013. doi: 10.1017/pasa.2012.007.

M. P. van Haarlem, M. W. Wise, A. W. Gunst, et al. LOFAR: The LOw-Frequency ARray. *A&A*, 556:A2, August 2013. doi: 10.1051/0004-6361/201220873.

S. Westerlund, C. Harris, and T. Westmeier. Assessing the Accuracy of Radio Astronomy Source-Finding Algorithms. *PASA*, 29:301–308, February 2012. doi: 10.1071/AS11049.

T. L Wilson, S Rohlfs, and Hüttemeister. *Tools of Radio Astronomy.* fifth edition, December 2012.

J. M Wrobel and R. C Walker. Sensitivity. In *Synthesis Imaging in Radio Astronomy II, ASP Conference Series*, pages 171–186, 1999.

Xibin Zhang, Maxell. L King, and Rob. J Hyndman. Bandwidwith selection for multivariate kernel density estimation using mcmc. July 2004. URL `http://repec.org/esAUSM04/up.1603.1077410300.pdf`.