A STATISTICAL APPROACH TO AUTOMATED DETECTION OF MULTI-COMPONENT RADIO SOURCES

A thesis submitted in partial fulfillment of
the requirements for the degree of

MASTER OF SCIENCE

Department of Statistical Sciences

Faculty of Science

UNIVERSITY OF CAPE TOWN

By: Jeremy Stewart Smith

Supervisor: Prof. Russ Taylor

JUNE 2020

PLAGIARISM DECLARATION

I, Jeremy Stewart Smith, know the meaning of plagiarism and declare that all of the work in the document, save for that which is properly acknowledged, is my own.

ACKNOWLEDGMENT

I would like to acknowledge Prof Russ Taylor and Prof Mattia Vaccari for their constant support and advice, and for assisting me to direct this research. I acknowledge Prof Russ Taylor, Dr. Imogen Whittam, Retief Lubbe, Brian Bichanga, Mendrika Rakotomanga and Zafiirah Hosenie for their initial research and work completed on this topic at the JEDI Workshop in Big Data held at the SKA-driven Big Data Challenge in Africa: Science, Innovation and Opportunity Conference in Madagascar, 2018. Student participation at the JEDI workshop was sponsored by Development in Africa with Radio Astronomy (DARA) and the Inter-University Institute for Data Intensive Astronomy (IDIA).

I acknowledge the use of computing facilities of IDIA for this work. IDIA is a partnership of the Universities of Cape Town, of the Western Cape and of Pretoria.

I would also acknowledge Dr Srikrishna Sekhar for all the conversations that we had and all the invaluable support that he provided during my research.

A STATISTICAL APPROACH TO AUTOMATED DETECTION

OF MULTI-COMPONENT RADIO SOURCES

By: Jeremy Stewart Smith, M.Sc.
University of Cape Town
June 2020

Supervisor: Prof.: Russ Taylor

Abstract

Advances in radio astronomy are allowing for deeper and larger observations than ever before. Source counts of future radio surveys are expected to number in the tens of millions. Source finding techniques are used to identify sources in a radio image, however, these techniques identify single distinct sources and are unable to identify multi-component sources, that is to say, where two or more distinct sources belong to the same underlying physical phenomenon, such as a radio galaxy. Identification of such phenomena is an important step in generating catalogues from surveys on which much of the radio astronomy science is based. Historically, identifying multi-component sources was conducted by visual inspection, however, the size of future surveys make manual identification prohibitive. An algorithm to automate this process using statistical techniques is proposed. The algorithm is demonstrated on two radio images. The output of the algorithm is a catalogue where nearest neighbour source pairs are assigned a score. By applying several selection criteria, pairs of sources which are likely to be multi-component sources can be determined. Radio image cutouts are then generated from this selection and may be used as input into radio source classification techniques. Successful identification of multi-component sources using this method is demonstrated.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Dedication

This thesis is dedicated to my partner for his unwavering support and love, without which this thesis would not be complete.

Chapter One

INTRODUCTION

Radio astronomy has seen significant strides in development over the last decade. These developments have resulted in larger and deeper surveys of the sky within the radio spectrum. Current and future projects, such as the Square Kilometre Array (SKA) (Braun, Bourke, et al., 2015)(Braun, Bonaldi, et al., 2019), will further enhance the resolution and dynamic range of observations and allow for large scale surveys to be undertaken.

Different astronomical objects are visible at particular wavelengths due to the emission (or absorption) of electromagnetic radiation due to various physical phenomena. Radio astronomy is used to observe both galactic and extragalactic sources. Galactic sources of radio emissions include the Galactic centre, HI and HII emissions, interstellar medium (ISM), star-forming regions and supernova remnants, among others. Extragalactic sources of radio emissions may include quasars, radio galaxies and starburst or star-forming galaxies, the cosmic microwave background (CMB) and fast radio bursts (FRBs), among others. A number of the ongoing large surveys will focus on the exploration of the extragalactic sky, observing active galactic nuclei (AGN), associated radio galaxies, and star-forming (SF) galaxies. AGNs are particularly important for understanding galaxy evolution and their effects on star formation in host and neighbouring galaxies. AGNs are also highly polarised and are important for investigating galaxy magnetic fields (M. J. Jarvis et al., 2016). Radio galaxies, a type of AGN that are radio loud, are characterised by a central black hole with two

jets extending in opposite directions. The jets often form hot-spots or lobes, which like the central black hole may be visible in the radio spectrum. These jets and lobes often extend beyond the limits of the host galaxy. Depending on the alignment of the black hole and jets to the line of sight of the observer the central black hole and/or one of the jets may not be visible in the radio image. Therefore, these radio galaxies may occur as a single source in the radio image, or two, three or more distinct sources, depending on their visible structure.

An important output from any astronomical observation or survey is a source catalogue or record of sources from the observation. This catalogue provides the basis for much scientific inquiry. Generating such a catalogue requires finding and characterising the sources in a survey image. Such a process includes several steps, including: pre-processing, source finding, and source characterisation, post-processing and cataloguing (Koribalski, 2012). A key challenge for current and future observations is the size of the data. Historically, sources in astronomical images have been identified by eye. Present day techniques employ algorithms to automate the detection of sources in radio images (Hopkins et al., 2015). However, such techniques are primarily used for point sources and their ability to recognised extended sources and diffuse emissions is limited. Furthermore, these techniques do not take into account multi-component sources, where two or more distinct sources are part of the same underlying physical phenomenon, therefore such sources still require manual or visual identification (Ishwara-Chandra et al., 2020)(Wu et al., 2018)(Kozieł-Wierzbowska and Stasińska, 2011). Identification of multi-components sources is crucial for investigations relating to AGNs, galaxy evolution and other science goals (M. J. Jarvis et al., 2016).

Future surveys are expected to produce radio images with sources number in the millions to 10s of millions (Hopkins et al., 2015) (Smolcic et al., 2015). The volume of data will result in manual identification being prohibitive. Automated tools for source identification and source classification are therefore necessary. The aim of this study is to investigate and develop an automated technique to identify multi-component radio sources.

Chapter Two

BACKGROUND

2.1 Radio Astronomy

Radio astronomy is the study of celestial objects and phenomena through the detection of emissions of electromagnetic (EM) radiation at radio frequencies. The radio spectrum includes EM radiation ranging from 1 mm wavelengths to over 10 metre wavelengths, equivalent to a frequency range of 10 MHz to 300 GHz. Radio astronomy observations can probe both thermal and non-thermal processes. Thermal processes are those that arise as a result of excitation due to the thermal properties of a gas, that is to say, the kinetic or internal energy associated with the 'temperature'. Non-thermal processes arise from energetics over and above thermal distributions. Thermal detections include the 21-cm line of neutral hydrogen and other molecular lines from cold gases in the interstellar medium (ISM), free-free emissions from HII regions, and blackbody radiation from the cosmic microwave background. The spectral lines allows for the composition and dynamics of the ISM to be determined and detection of neutral hydrogen is used to determine the rotation dynamics of the galaxy. Non-thermal radiation includes synchrotron radiation, which describes the emissions from charged particles moving at relativistic velocities - that is to say, close to the speed of light - in magnetic fields of the ISM. Synchrotron emissions are related to highly energetic processes such as quasars, supernovae remnants and AGNs (Burke, Graham-Smith, and Wilkinson,

2019).

The diverse range of radio emissions allow for both galactic and extragalactic observations. Furthermore, radio emissions are not obscured by dust as are other emissions such as those at optical wavelengths. This makes deep radio surveys ideal for searching for distant objects that would otherwise be obscured at other wavelengths. For this reason radio astronomy is ideal for observing the centre of our galaxy, and looking beyond our galaxy, at extragalactic regions.

Observations at radio and optical wavelengths are unique in that these observations may be completed using ground-based telescopes. Much of the EM radiation received on Earth is blocked or absorbed by gases and molecules in the Earth's atmosphere. Regions in the spectrum exist at the optical and radio frequencies where EM radiation is able to penetrate the atmosphere, allowing for observations from the Earth's surface to be conducted, while observations at gamma, X-ray, ultraviolet, IR and microwave wavelengths must be conducted at either high altitudes or in space (Burke, Graham-Smith, and Wilkinson, 2019). This allows the unique opportunity for large radio and optical telescopes and radio arrays to be built on the Earth's surface.

The resolution of a telescope is directly proportional to the wavelength at which the observation is being conducted and indirectly proportional to the diameter or collecting area of the telescope. This relationship can be described by Equation 2.1, where θ is the resolution of the telescope, λ is the observed wavelength and D is the diameter of the telescopes collecting area. The intention is to achieve the smallest resolution, allowing for small details to be visible and for small, distant objects to be observed distinctly.

$$\theta = \frac{\lambda}{D} \tag{2.1}$$

The optical spectrum occurs at wavelengths between 400 nm and 800 nm, while the radio spectrum occurs at wavelengths between 1 mm to over 10 m. This means that, for a given

aperture size, an optical telescope can achieve a significantly better resolving power compared to a radio telescope. To achieve a resolving power similar to optical, radio telescopes with large apertures must be built.

To increase resolution, multiple small radio telescopes can be built as radio arrays. These arrays make use of interferometry, whereby the radio telescopes are connected to one another. The telescopes receive the same observed signal but at different times due to their varied positions. The received signals are then combined electronically, with the differences in the phases of the signals known due to prior knowledge of the distance between antenna pairs. Their received signals are combined to synthesize a much larger aperture than any of the individual telescopes. Building a multitude of small telescopes is significantly more cost effective than building a single radio telescope of the equivalent diameter.

An international collaboration has resulted in the initial stages of development of the Square Kilometer Array in order to take advantage of the unique position and scientific benefits of radio astronomy. The primary purpose of this project is to provide significant improvements in resolution, dynamic range and depth of observations, and to provide a flexible observing environment to meet the requirements of a large range of scientific studies. Scientific areas of study include, but are not limited to: terrestrial planet formation; pulsars and black holes; the evolution of cosmic magnetism; galaxy evolution; and the early universe (Carilli and Rawlings, 2004). The SKA is intended to cover frequencies between 70 MHz and 30 GHz and to operate with a large field of view. A collection of coherent ground-based telescopes employ interferometric techniques that will eventually allow for a synthesised aperture equivalent to a million square metres (Dewdney et al., 2009). Such an aperture size will see tremendous improvements to resolution and dynamic range over current telescopes. Astronomical surveys supported by the SKA are expected to produce radio images with source counts in the hundreds of millions. Initial project developments include the Australian SKA Pathfinder (ASKAP), the Murchison Widefield Array (MWA) and the MeerKAT array. The MeerKAT array consists of 64 parabolic dish antennas each with a diameter of 13.5 m

(Brederode et al., 2016). These projects are already conducting observations for scientific studies.

One of the survey projects supported by MeerKAT is the MeerKAT International GHz Tiered Extragalactic Exploration (MIGHTEE) Survey. This large survey project will investigate four extragalactic deep fields, including COSMOS, XMM-LSS, ECDFS and ELAIS-S1, with a total area of 20 deg^2. These fields are well studied and multi-wavelength data is available for the deep fields, allowing for a range of scientific inquiry. The MIGHTEE survey will be conducted over a bandwidth of 900 - 1 670 MHz at a resolution of roughly 6 arsec. The science goals of the MIGHTEE project include: investigating the link between AGN and star formation at high redshifts to determine if AGN activity is responsible for halting star formation in massive galaxies; to understand the interplay between star formation and AGN activity in order to understand galaxy evolution; to investigate the polarization of AGNs pertaining to the relationship between the physical size of the galaxy and the degree of polarization; and investigate evolution of magnetic fields in galaxies; amongst several other goals (M. J. Jarvis et al., 2016). The size of the MIGHTEE survey project and the importance of AGNs in its sciences goals, therefore, necessitate an automated approach to detecting and characterizing multi-component AGNs in the survey data.

2.2 Radio Galaxies

Active galactic nuclei are a type of galaxy whose central nuclei are more luminous than the radiation from the surrounding galaxy as a consequence of black hole accretion processes at their centers. Several different classes of AGNs exist, including Seyferts, radio galaxies, quasars and blazars, amongst others. Radio galaxies and quasars are among the most powerful radio emitters and are characterised by synchrotron radiation. Quasars are compact in comparison to their radio galaxy counterparts, extending less than 1 pc, while radio galaxies can be found to extend up to 3 Mpc. The source of the radio emissions is thought to be the

release of gravitational energy of stellar material that is falling into the central black hole. The release of the energy is seen as the jets, and single or twin radio lobes that characterise these galaxies (Burke, Graham-Smith, and Wilkinson, 2019).

Since the early days of radio astronomy bright radio galaxies have been grouped into two classes, FRI and FRII, named after Fanaroff and Riley who made the distinction (Fanaroff and Riley, 1974)(Urry and Padovani, 1995). FRI galaxies are the less luminous of the two classes. In this class, the core and adjacent portions of the jets tend to be more luminous and the lobes are present but decrease in brightness with distance. FRII galaxies are more radio-loud (luminous in the radio spectrum) than their counterparts and are characterised by a low luminosity core and radio-loud lobes. The difference in the position of the high luminosity areas is thought to be due to whether the particles in the jets maintain relativistic speeds as they exit the host galaxy (Laing and Bridle, 2002). In the case of FRII galaxies, the relativistic velocities are maintained beyond the host galaxy allowing the energetic particles in the jets to interact with the surrounding interstellar or intergalactic medium at a greater distance from the core, forming the characteristic lobes.

A model that describes all AGN classes has been proposed, called the Unified Model of AGN (Burke, Graham-Smith, and Wilkinson, 2019). In this model, a central massive object exists, typically a supermassive black hole. Stellar material and gas from the surrounding ISM falls in toward the central black hole. The angular momentum of this material falling towards the black hole causes the material to form an accretion disc with a steep gradient of angular velocity. A combination of the gravitational potential energy of the surrounding material and friction caused by the differentially rotating disc results in the ejection of energetic material from the nucleus outward along the polar directions. The twin jets are narrow and accurately parallel, and funnel energetic particles and magnetic-field energy outward. A thick torus of inflowing material surrounds the thin accretion disc. The torus material is cooler and opaque and obscures the central region when the system is viewed edge on. If the AGN is viewed from the direction of the polar directions, however, the central

region may be visible (Burke, Graham-Smith, and Wilkinson, 2019). This model has been used to suggest that the differences in the different types of AGNs as well as the number of components visible in the radio image is dependent on the viewing angle at which the AGN is observed.

In the case of radio galaxies the jets extend up to Mpc scales. The jets are continuous with the core and extend outwards culminating in the radio lobes. Bright patches along the jets are caused by turbulent interaction with ISM. Significant interaction with the interstellar and intergalactic medium can be seen in the radio lobes as 'hot spots' which are exceptionally radio loud. In a radio image, these lobes may appear as two or more distinct sources, or depending on the angle at which the radio galaxies is viewed, one or part of a jet may be obscured. In some instances the central black hole is visible in the radio image. A radio galaxy in a radio image, therefore, may manifest itself as multiple distinct components which are part of the same underlying physical phenomenon.

Quasars and radio galaxies are considerably luminous objects, and the most luminous of these objects may therefore be seen at very large redshifts, a term synonymous with distance (and time) in astronomy. The higher the redshift, z, of an observed object, the further back in time one is observing. Radio galaxies, therefore, may potentially provide insights into the large-scale structure of the universe as well as cosmological evolution.

2.3 Multi-wavelength Astronomy

On account of the Earth's atmosphere blocking the reception of large parts of the electromagnetic spectrum, ground-based observations have been primarily optical and radio, while observations at other wavelengths, such as IR, can only be conducted at high altitude, for example, from mountain-tops, stratospheric balloons, earth-orbiting satellites or interplanetary probes (Burke, Graham-Smith, and Wilkinson, 2019). Technological advances have allowed for telescopes to be positioned in orbit around the Earth, such as the Herschel Space

Observatory which undertakes observations at far-infrared and submillimetre wavelengths (Pilbratt et al., 2010). Space-based telescopes have allowed observations at wavelengths that could not be conducted from the Earth's surface.

Galaxies emit EM radiation over a wide frequency range and different physical phenomena are visible at different wavelengths (Mattia Vaccari, 2016). For example, the synchrotron emission of AGN radio jets is visible at radio and X-ray wavelengths, while the accretion disk surrounding the central black hole, or gas and dust of the central galaxy, is visible at infrared wavelengths. Therefore, to understand the full structure and evolution of galaxies it is imperative to draw upon observations conducted at a wide range of frequencies. Furthermore, different statistical techniques of analysis are applicable at different wavelengths and such techniques can be drawn from if information of an object or entire surveys are available at relevant wavelengths. For example, redshifts are often difficult to determine for objects at radio wavelengths, while if optical or infrared data is available for the same object under inquiry, the redshift value of the object can be calculated. This combination of observations and techniques from different frequencies is known as multi-wavelength astronomy.

One of the challenges associated with multi-wavelength astronomy is accurately determining source counterparts across observations at different wavelengths. The resolution of an observation is a function of the observed wavelength, no standardised resolution is prescribed and, therefore, observations at different wavelengths will inherently achieve different resolutions. Furthermore, due to the fact that different physical phenomenon are visible at different wavelengths, the observed angular size of a source may be considerably different across wavelengths. These two factors, among other factors, such as super-positioning, confusion limited observations and differences in dynamic range of an observation (Norris et al., 2009), make accurately determining source counterparts difficult.

Another challenge related to multi-wavelength astronomy is the in-homogeneous observational coverage of the celestial sphere. While certain regions of the sky are commonly observed, such as ELAIS, COSMOS among others, other regions may not be as well ob-

served - such as the plane of the galaxy where dust obscures observations at optical and IR wavelengths.

2.4 Radio Observations

During radio astronomy observations the radio energy from distant sources is measured or detected. The radio telescope detects incoming electromagnetic waves and transforms the radio energy into an electrical signal. The signal includes both the amplitude and phase of the electromagnetic wave. In the simplest case, the amplitude or instantaneous voltage is measured and the phase is discarded. In the case of arrays, the signal is shifted to a lower frequency, and signals from different receivers are cross-correlated. Unlike measuring the instantaneous voltage, cross-correlation of the signals conserves the phase information and is able to magnify signals above background noise, allowing for the detection of faint signals (Burke, Graham-Smith, and Wilkinson, 2019).

The receivers used for radio astronomy are often mounted on parabolic or spherical dishes. In the case of the SKA, MeerKAT (Booth and Jonas, 2012) and the Giant Metrewave Radio Telescope (GMRT) arrays the dishes are parabolic. These dishes act as collecting areas, and are designed to reflect the incoming signals to the receivers mounted either above the dishes or on an offset arm. The measured signal is, therefore, a flux, S, which is the energy, E, that crosses an area, A, lying perpendicular to the direction of the incoming signal, described by Equation 2.2.

$$S = (dE/dt)/A \tag{2.2}$$

Receivers operate at a finite bandwidth and the measured flux is therefore a function of the frequency. The flux per unit bandwidth, S_ν is known as flux density or spectral flux, and the measured flux is the flux density integrated over the receiving bandwidth, as described

by Equation 2.3.

$$S = \int S_\nu \, d\nu \tag{2.3}$$

In radio astronomy, the flux density is measured in units Jansky (Jy), where

$$1 \text{ Jy} = 10^{-26} \, \text{Wm}^{-2} \, \text{Hz}^{-1} \, . \tag{2.4}$$

The received power of an antenna is a function of the flux and the effective area of antenna or dish. The effective area of a dish is direction dependent, calculated as a function of the direction relative to the antenna axis. Signals received in the direction inline with the antenna axis will generally have the maximum response. This response pattern describes the power gain of the antenna, and is characterised by a primary beam, where the maximum response occurs, and the sidelobe response (the response outside of the primary beam) (Burke, Graham-Smith, and Wilkinson, 2019), and is often depicted as a polar diagram. The importance of the primary beam is considered during the source finding step detailed in Section 3.1. The primary beam results in sources toward the centre of the radio image appearing brighter than those sources that occur near the edge of the image, and may be corrected or accounted for during source finding.

2.5 Interferometry

Modern day radio telescopes that employ interferometric techniques achieve significantly higher resolution than previous single dish antennas. A brief description of a radio interferometer follows.

In an interferometric array, multiple telescopes or antennae are positioned at different locations, at different distances from one another. The distance between two antennae is

known as a baseline, and the number of baselines for an array is equal to $N(N-1)/2$, where N is the number of antennae in the radio interferometer. Therefore, for a 64 dish interferometer, such as MeerKAT, there will be 2016 baselines. During an observation, the antennae with track the same position on the sky, accounting for the Earth's rotation. The signals received by an antenna pair are cross-correlated using a correlator, taking into account the time delay. The pairwise array outputs which are amplified and combined to form all possible interferometric combinations has resulted in this technique to be known as *aperture-synthesis*. Using this technique the angular resolution of the array is not a function of the diameter of the individual telescopes, but rather the length of the largest baseline (Burke, Graham-Smith, and Wilkinson, 2019)(Thompson, Moran, Swenson, et al., 2017). This allows for significant improvement to the angular resolution of the array without the costs required to build a large telescope with the equivalent diameter.

The correlator response is a complex value with an amplitude and phase. This response is expressed in the u-,v- plane, which is a coordinate system based on the baseline vector, where u and v are projected east and north respectively, with units in wavelengths (Thompson, Moran, Swenson, et al., 2017). The correlator response expressed in this coordinate system is known as the complex visibility. The Fourier transform of the complex visibility evaluates to the source brightness distribution of the source plane or celestial sphere. A single interferometer observation, that is to say from a single antenna pair, will provide only one value, but a multitude of observations and baselines will result in an approximation of the 2D Fourier transform of the source brightness distribution. In addition to the multitude of baselines, the rotation of the Earth may be used to create additional baselines which allows for further sampling of the u-,v- plane during an observation. The collection of values from the multitude of baselines is known as the visibility data. However, complete sampling of the u-,v- plane is impossible. On account of the incomplete sampling, expression of the source brightness distribution includes a *spectral sensitivity function* component, which is equivalent to a transfer function. The Fourier transfer of this *spectral sensitivity function* is

the response to a point source, that is to say, it is a map that would be generated as a result of the Fourier transform of the visibility data from observing a point source. The resultant map or pattern is known as the *synthesised beam* or *dirty beam* (Burke, Graham-Smith, and Wilkinson, 2019).

In order to produce a science image or map once an observation has been completed, several steps must be undertaken. First the visibility data from the observation must be calibrated. Observations of sources used for calibration are completed during the primary observation. Two classes of calibrators are required, including a phase calibrators and flux calibrators. The source calibrators are used as standards to calibrate the visibility data. The data is then reduced through several steps, including flagging, gridding and calibration. During the flagging step, data that is known to be poor, such as frequency channels that are affected by terrestrial broadcasts, are flagged in order to be ignored during the data reduction process. The visibility data is then gridded in order to complete a fast Fourier transform. Several cycles of flagging, gridding and calibration may be completed during data reduction. The result of this process is a sky map or image of the observation, however, this map is known as the *dirty image* as the Fourier transform of the visibility data to the source brightness distribution still includes the *dirty beam* transfer function. This results in rings and radial lines appearing as artifacts around sources in the map. The form of dirty beam is known from the distribution of the visibility data in u-,v- plane and can be corrected for. An iterative process is used to 'clean' the dirty image. A bright object is selected from the dirty image. The product of the dirty beam with this object is subtracted from the map. This process is repeated for a significant number of objects in the map, with each iteration 'cleaning' the map further. Often a source that was hidden in the sidelobes of the transfer function is revealed after the several iterations (Burke, Graham-Smith, and Wilkinson, 2019). Finally, the result of this process is a map of the sources with much of the effects of the dirty beam removed, and is considered a good approximation of the source brightness distribution. From this image, further analysis may be conducted, crucially including source finding.

Chapter Three

Source Catalogues and Multi-component radio sources

3.1 Source Finding

Source finding is the detection of objects in astronomical images. Source finding is a key component in any astronomical survey, where distinct objects are identified and their properties, such as position, size and intensity, are determined. This process is part of several steps that are used to generate source catalogues on which much of the astronomical science is based. Historically, source finding has been conducted by eye. On account of larger and deeper radio surveys of the sky expected from arrays such as the SKA and other future planned observatories, non-automated identification of sources is prohibitive. Automated algorithms are required to meet the volume and size of the astronomical images, where expected source counts number in the millions (Hopkins et al., 2015) (Smolcic et al., 2015).

Radio astronomy images are comprised of pixels which represent the flux density values of the observed sky. An extra-galactic image produced by radio interferometers may include both extended and compact sources. A radio source will extend across several pixels in the radio image. A compact source will include one or more peak pixels, having the highest flux density value of the pixels in the local region, and an island of pixels surrounding the peak

flux pixel(s) that are considered to be associated with the compact source. A cross-section of a compact source in a radio image will reveal flux values with a Gaussian-like shape. The pixels that are considered to be part of the source will have a flux density value above a certain threshold, usually a function of the background noise.

Several source finding software packages exist, including `PyBDSF` (Mohan and Rafferty, 2015), `Aegean` (P. Hancock et al., 2019), `Blobcat` (Hales et al., 2014), `SExtractor` (Bertin and Arnouts, 1996), among others. These packages largely employ the same techniques with some differences in their applied methods, and mostly focus on the detection of compact sources. In general, the process that is employed includes the following steps. The root-mean-square (rms) or background noise of the image is calculated, often locally using a sliding window technique, a flux detection threshold is chosen as a function of the rms. Pixels above this threshold are considered significant and any contiguous significant pixels are considered islands (Hopkins et al., 2015). At this point, the algorithms tend to diverge and different techniques are employed to determine whether islands include several components or whether they can be considered as a single source. In the case of `PyBDSF`, islands are extended to include pixels above some lower threshold. One or several Gaussians are fitted to the cross-sections of the island drawn through the island's peaks. Early source finding techniques employed a single Gaussian for fitted sources, however, in cases where several sources existed in an island, a single Gaussian would fail to fit the sources and may result in a failed detection (P. J. Hancock et al., 2012). An iterative fitting method was introduced, using multiple Gaussians to fit several peaks within an island. During iterative fitting, a residual threshold is used to determine how well a Gaussian fits a possible source. If the residual is below the threshold, the 'goodness of fit' is considered acceptable and a positive detection occurs. If the fit it poor, the fitting is redone and a second Gaussian component is added. This process is repeated until the fit is considered good or a maximum number of components is reached. Gaussians that overlap significantly are combined and considered single sources. Source properties are then determined and a catalogue of detected sources is

produced. In the case of `PyBDSF` the source position is determined by its centroid, and the position of maximum flux is also recorded, in addition to the maximum or peak flux value and the integrated flux value (the sum of the pixels associated with the source), among other parameters.

The main purpose of most source finding algorithms developed to date is to detect individual compact sources. A number of the source finding algorithms include methods to detect extended or diffuse sources. However, these techniques are challenged when attempting to determine if two or more sources are related to a single physical phenomenon, such as the lobes and/or jets of radio galaxies. Therefore, the sources detected during source finding techniques may be sources of a single physical system, or they may be components of a larger multi-component system.

3.2 Source Classification Techniques

The next step in producing a catalogue of an observation or survey is source classification. Once sources have been identified and characterised in a radio image, the sources must be classified by source type, whether star-forming galaxies (SFGs) and AGNs, or other subcategories. Historically, source classification was conducted manually. Modern day statistical techniques and machine learning techniques are essential on account of the size of current and future astronomical surveys. Machine learning techniques, such as neural networks, have proven to be incredibly successful at image recognition.

A number of machine learning techniques have been investigated as source classification tools. One such study, by Aniyan and Thorat (2017), used a convoluted neural network (CNN) to classify FRI, FRII and radio galaxies with bent-tail morphologies using data from the Faint Images of the Radio Sky at Twenty Centimeters (FIRST) survey. This study required input images, cutouts of extended radio sources, as inputs into the CNN. The CNN achieved a 95 precision score on bent-tail radio galaxies and 91% and 75% precision

on FRI and FRII respectively. Another study, completed by Wu et al. (2018), present the Classifying Radio sources Automatically with Neural networks (CLARAN) classifier. This classifier was demonstrated on the FIRST survey data along with the Radio Zoo Galaxy Data Release 1 catalogue, a catalogue using a crowd-sourced method for classification. The CLARAN classifier implements both source finding and source classification and utilizes a CNN for classification. In this case, CLARAN classifies sources by morphology based on the number of components and the number of peaks associated with the components, such as sources with one component one peak, up to source with one component and seven peaks, two components two peaks and so on. While CLARAN improves classification performance by utilizing IR data, classifying multi-component sources remains a challenge.

Self-organising maps (SOM) or Kohonen maps have also been used for source classification. A study by Polsterer et al. (2016) presents the Parallelized rotation and flipping INvariant Kohonen maps (PINK) software. The study used the Radio Galaxy Zoo catalogue data to generate a SOM of radio galaxies, including both point sources and extended sources. Radio image cutouts where used as inputs into the PINK software, and flipped and rotated versions of the input images were used as inputs to improve results. While SOM technique are unsupervised, areas of the map could be prescribed to a particular morphology and the measure of similarity for each input to its best match could be used to classify the source.

Other techniques involve using multi-wavelength data, astronomical images and catalogues from observations carried out at other wavelengths, such as X-ray and infrared (IR). Sources from the radio images are positionally cross-matched against sources in the multi-wavelength data, and those sources within some threshold distance of each other are assumed to correspond to the same physical system in the two wavelength domains. Identification of AGNs and other galaxy types is then conducted using the information from the multi-wavelength catalogues, or techniques applicable to the other wavelengths, such as X-ray luminosity or Infrared Array Camera (IRAC) colour diagnostics (Ocran et al., 2020).

While these automated techniques are essential to perform classification on large data

sets, accurate cutouts of sources, especially multi-component sources, are required as input for many of the automated techniques. Therefore, a technique where multi-component sources in a radio image are identified may play an integral step in the generation of astronomy catalogues in the future.

3.3 Multi-component Radio Sources

Multi-component sources are a collection of two or more distinct components where the components are part of the same underlying physical phenomenon. An example of such a multi-component sources is a radio galaxy where the lobes, jets and central black hole may appear as distinct components in a radio image. Identification and classification of multi-component sources in a radio image is important when generating a catalogue of a radio survey and aids scientific inquiry in fields of study such as galaxy populations and galaxy evolution, amongst others. Even until recently, multi-component sources have been identified manually through visual inspection (Ishwara-Chandra et al., 2020)(Wu et al., 2018)(Kozieł-Wierzbowska and Stasińska, 2011). However, identification by eye lends itself to bias towards bright, well resolved objects, and multi-component sources where the different components are further apart or one or both is not very bright may be overlooked as the eye does not easily associate them as one object, yet statistically they may be in fact components of the same underlying physical system. Examples of this bias are demonstrated in Figures 4.2 and 4.3 in Section 4.1 Identification by eye will also become prohibitive in the future with larger surveys where source counts number in the tens to hundreds of millions. Source finding techniques which identify sources in radio image are challenged to identify multi-component sources. The aim of this study is therefore to investigate an automated statistical method to identify multi-component sources in radio images in order to overcome the prohibitive nature of visual inspection in the presence of large data volumes, and to provide a more rigorous mathematical approach which will overcome the bias of the eye.

Chapter Four

DATA

4.1 GMRT EN1W

The radio image that will be used during this study is from a survey completed by the GMRT located near Pune in India (Ishwara-Chandra et al., 2020). The survey is a wide-area survey of the European Large-Area ISO Survey-North 1 (ELAIS N1) field. The resultant radio image from the survey is therefore designated by GMRT EN1W and will be referred to as such throughout the rest of this text. The GMRT EN1W image itself is a series of mosaic images. A total of 52 pointings were completed to cover the area of the survey. The GMRT EN1W radio image can be seen in Figure 4.1.

The GMRT EN1W survey was completed at 610 MHz and covers an area of 13 $\deg^2$ with a resolution of 6 arcsec and an rms noise of $\sim$ 40 μJ/beam. This means that components or sources that are greater than 6 arcsec are able to be resolved as separate components or sources, while sources smaller than 6 arcsec may be seen as a single source. As a reference for scale, the angular diameter of the full moon is roughly 31 arcmin.

A total of 6474 sources are recorded (Ishwara-Chandra et al., 2020) to have been detected in the GMRT EN1W image. These sources are primarily compact sources and are associated with normal galaxies, starburst galaxies and AGNs. The AGN sources are generally more radio-bright than their normal and starburst galaxy counterparts and about 10 percent of

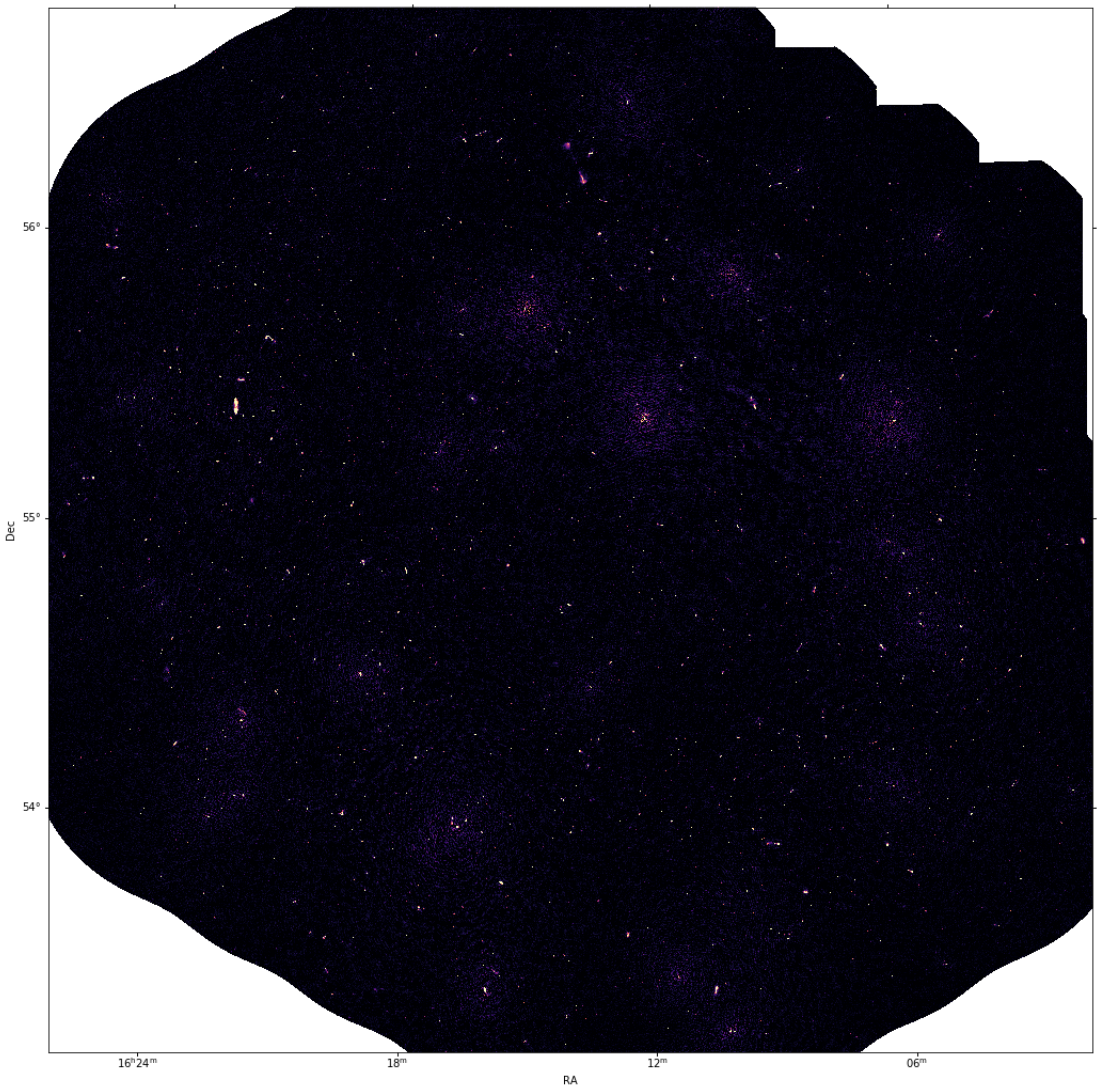


Figure 4.1 GMRT EN1W radio image

the AGN sources will exhibit jets and lobes. By visual inspection, six giant radio galaxies (GRGs) were detected in the GMRT EN1W image, along with several compact doubles.

Figure 4.2 and Figure 4.3 show a cutout from the GMRT EN1W image. In Figure 4.2, the large radio galaxy is clearly visible as the extended bright object. In additional to the radio galaxy, several bright sources that are in close proximity to each other can easily be identified by eye as multi-component sources. These sources are circled in blue in Figure 4.3.

However, there are some source where it is less certain whether or not the sources are single sources or part of the same underlying physical phenomenon. The pairs of sources circled in read in Figure 4.3 demonstrate this uncertainty. These sources may easily be overlooked during visual inspection when trying to identify AGNs.

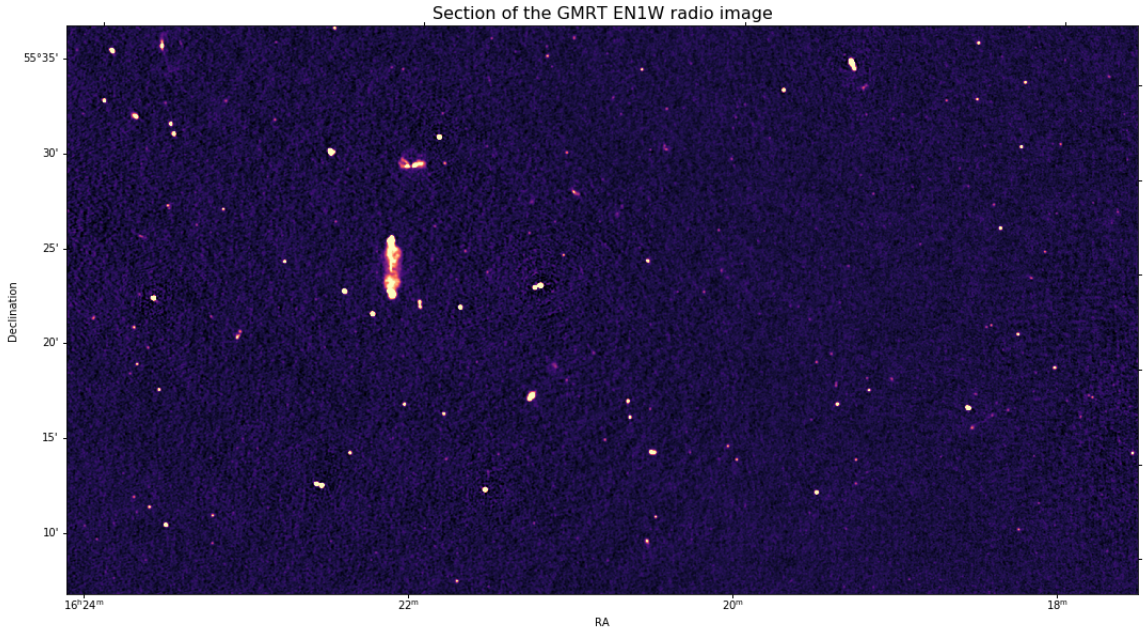


Figure 4.2 Cutout of the GMRT EN1W radio image

4.2 JVLA SDSS Stripe 82

The SDSS (Sloan Digital Sky Survery) Stripe 82 region was observed using the Karl G. Jansky Very Large Array (JVLA) at a frequency range of 1 - 2 GHz with a 1 GHz bandwidth (Heywood et al., 2016). The survey consisted of 1026 pointings using a hexagonal mosaic pattern and observed a region of 100 $\deg^2$. The observation undertaken was particularly sensitive to diffuse and extended radio emissions, making it an excellent candidate for the automated statistical analysis due to the diffuse and extended nature of the AGN radio lobes under investigation. The radio images produced included an eastern and western region,

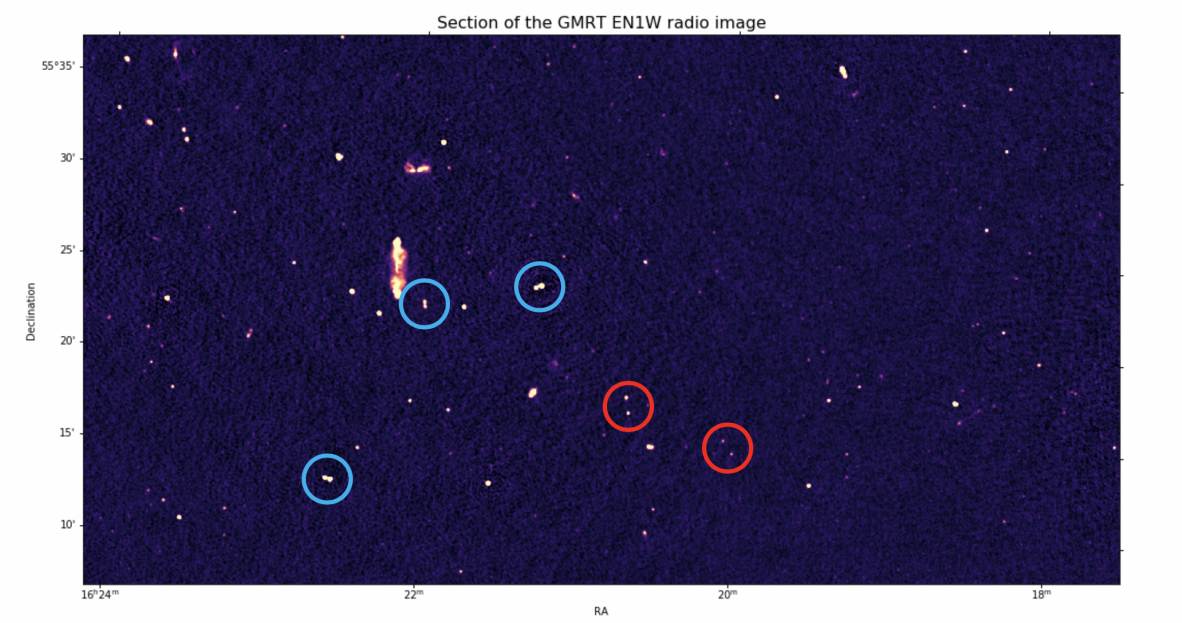


Figure 4.3 Cutout of the GMRT EN1W radio image with possible multi-component sources indicated

the former of which was used during the data analysis. The eastern image extended from roughly 22h00m to 23h21m right ascension (RA) and declination (DEC) of $+1°$ to $-1°$. The resolution of the radio image was 16 $\times$ 10 arcsec, with a 1σ noise level of ~ 88 μJ/beam. The JVLA SDSS Stripe 82 image of the eastern region is included in Figure 4.4. This work focused on the eastern region and from now on JVLA SDSS Stripe 82 will refer to the eastern region.

Source finding (Heywood et al., 2016) using `PyBDSF` (Mohan and Rafferty, 2015) found 4 354 sources in the eastern image using a peak threshold of 5σ and an island threshold of 3σ for the input parameters to the source finding software.

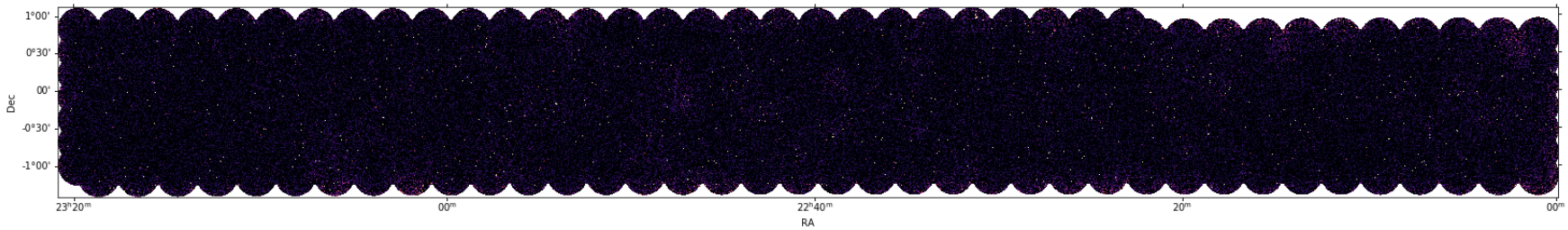


Figure 4.4 JVLA SDSS Stripe 82 radio image

Chapter Five

METHODOLOGY

5.1 Outline of the proposed algorithm

The purpose of the algorithm under investigation is to identify multi-component radio sources. Current techniques rely primarily on identifying the multiple components of a radio source by visual inspection. Visual inspection, however, is easily affected by bias towards certain features. Bright objects, objects that appear to be in alignment or are in close proximity to each other are easily noticeable, while faint sources and sources that are further apart, which may be related to the same underlying physical phenomenon, may be overlooked. The intention is therefore to develop an automated statistical technique that is mathematically rigorous in order to overcome the bias of visual inspection.

The radio sky is homogeneous (J. Condon, 1999), that is to say, galaxy populations are similar in any observed direction. The radio image which results from an observation or survey contains distinct sources. These distinct sources are either single-component (which in turn can be compact or extended) or multi-component sources. There exists some randomness in the distribution of sources throughout the region observed with regards to their position. However, the multi-component sources are structured, and therefore demonstrate statistical properties that lie outside of the random distribution. In the radio image there is a preponderance of faint sources. Their flux distribution follows roughly a log-normal distribu-

tion. Therefore, in a random distribution of sources across an image, a source that exhibits high flux is more likely to be positioned near a faint source, and pairs of bright sources are uncommon. However, with multi-component sources, bright sources do appear in close proximity to one another, for example the radio loud lobes of radio galaxies. Furthermore, the distance between the components of multi-component sources tends to be much smaller than the distances between a random distribution of single sources.

In order to draw on the statistical properties associated with the structured nature of multi-component sources the following method is proposed: compare the distribution of a set of informed parameters of a sample of real radio sources (the real sample), where the sample includes single-component sources and multi-component sources, with a simulated sample that is generated from the flux distribution of the real sample, however, where the position of the sources in the simulated sample are generated using a random uniform distribution. The simulated sample will, therefore, contain only single-component sources and no multi-component sources and will represent the randomness of the distribution of single- and multi-component sources in the radio image. The informed parameters that will be used to generated the distributions include: 1) the separation distance between a source and its nearest neighbour, nn_d; and 2) the product of the flux of a source and that of its nearest neighbour, $S_1 \times S_2$, where S_1 is the target source flux and S_2 is the flux of its nearest neighbour.

The distribution of nn_d for the real sample will include distances that would be expected from a sample of randomly distributed sources as well as those distances that are associated with multi-component sources, while the distribution for nn_d for the simulated sample will only include distances that are expected from a random distribution. Radio galaxies tend to be brighter than their normal and starburst galaxy counterparts. The $S_1 \times S_2$ parameter will therefore accentuate bright and moderately bright pairs of sources, revealing pairs that are statistically significant compared to a uniform distribution of the sources.

It should be noted that not all the multi-component source candidates identified using

our method will be bona-fide multi-component sources. The caveat of a simple model is that not all possible outcomes are accounted for. It is therefore likely that only a portion of the related objects detected through this method are multi-component sources, while a portion may be sources that are gravitationally bound, such as galaxy groups or galaxy clusters. While the nn_d parameter may fail to disassociate the gravitationally bound sources from the multi-component sources, it is expected that the product of the source flux should go some distance in distinguishing these two groups.

5.2 Statistical approach to identify multi-component sources using simulation

The approach taken to identify multi-component sources in a radio image is detailed in the following section. The steps include: determining the flux distribution of the sources in the real sample; generating a simulated set of source fluxes using the empirical distribution technique based on the real sample; generating a simulated sample where source positions have a uniform distribution; generating a catalogue of source and nearest neighbour sources for the real sample and simulated sample respectively, including the distance between the source and their nearest neighbour and the product of these sources' fluxes; generating a distribution of these parameters for the real and simulated samples respectively; applying a comparative function to the distributions in order to assign a probability score to the real sample sources where the score represents the probability that the real sample nearest neighbour pairs are not random. Figure 5.1 depicts the work flow of the steps described. The upper portion of the diagram are steps associated with the real sample data, while the lower portion are steps associated with generating the simulated sample data.

The algorithm described above was implemented in the Python programming language. Python was chosen as it is a common programming language amongst astronomers and has

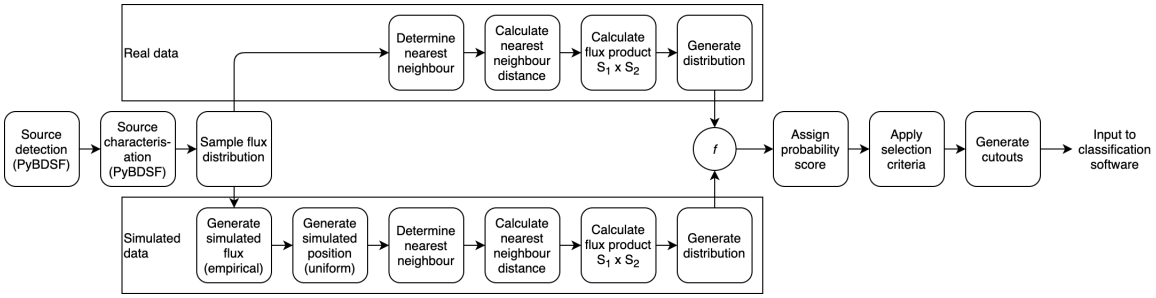


Figure 5.1 Work flow diagram detailing the steps in the proposed algorithm for detecting multi-component sources

a large community of developers and support. In addition, many astronomy and statistical packages are available as part of the Python development environment.

5.2.1 Determine the flux distribution of the radio sources

In order to generate a flux distribution of all the sources in a radio image, the radio sources must first be identified. A source-finding tool was used to detect sources in the radio image and distinguish the sources from background noise and artifacts that may exist in the radio image. Several source-finding tools exist that may be used to complete this process. The Python Blob Detector and Source Finder (`PyBDSF`) (Mohan and Rafferty, 2015) software was used during this study based on its performance in the comparison by Hopkins et al. (2015). A description of the `PyBDSF` tool and its algorithm can be found in Section 3.1. The output of the source-finding tool is a catalogue of sources, including the source id, the position of the source described by its right ascension and declination, the integrated flux or flux density, and peak flux of the source, among other parameters. For a discrete source, such as the majority of sources in the radio survey image, the integrated flux or flux density describes the spectral power received by the detector and is the integration of the spectral brightness over the angle subtended by the source (Burke, Graham-Smith, and Wilkinson, 2019). In this case, `PyBDSF` calculates the integrated flux as the sum of the pixels values of the pixels associated with the source. The peak flux refers to the brightest pixel associated

with a given 'source' in the radio image.

When a radio image is created, the central region of the image is where the primary beam of the telescope is strongest. The primary beam can be thought of as the sensitivity of the detecting instrument as a function of direction. The sensitivity is highest in the direction of the observation and decreases the further away from this direction. Primary beam corrections or weighting needs to be applied to the source fluxes to account for the difference in sensitivity. Often a threshold or weighting will be set for which sources outside of this threshold should not be reliably considered. Several thresholds, including the primary beam correction or weighting thresholds were applied to the output source catalogue of the source-finding tool. These thresholds included a minimum signal-to-noise ratio threshold, a minimum flux density threshold, a maximum primary beam correction threshold and a minimum flux ratio threshold. The signal-to-noise ratio threshold ensures that any source that is detected is significantly stronger than the background noise or rms. The background noise is not uniform across the radio image and it is affected by the primary beam and bright sources resulting from limited dynamic range of the observation. The background noise in the region of each source is calculated using a sliding-window technique and is calculated local to the position of the source (Mohan and Rafferty, 2015) using an image where the detected sources have been extracted (known further as the rms image). The signal-to-noise ratio is then calculated as the ratio of the peak flux to the source rms value. The minimum flux density threshold is calculated as a function of the median rms of the rms image. Any source with a flux density below this threshold was considered too faint to be reliably distinguished from the background noise. The flux ratio test is a test that there is actually a source in the image at the location of the found source. It is the ratio of the source fitted peak to the actual image value at the position of the source. Sources that did not meet these criteria were discarded from the real sample. The weight values of the sensitivity image were then applied to the peak and integrated flux values of the detected sources to account for the primary beam shape and the loss of sensitivity towards the edges

of the radio image. This correction to the flux values was applied using Equation 5.1 to the peak and integrated flux values by multiplication to generate the true sky flux values for the source finding catalogue. This process is considered again when generating the positions of the sources in the simulated sample, as the primary beam correction limit restricts the size of the region where the sources are detected.

$$fluxcorr = \frac{1}{\sqrt{weight}} \tag{5.1}$$

The primary parameters of interest from the catalogue output are the integrated flux and the position of the source. These parameters are pulled from the source-finding output catalogue along with the source identification number. A distribution of the integrated flux is then generated using the `Numpy` histogram function, `hist`. The output of this function is an array containing the histogram bin edges and an array of the distribution value, or in this case, the probability density function of the integrated flux for each bin.

5.2.2 Generate a simulated sample of source flux densities

Radio sources flux density distributions at specific frequencies are detailed in the literature (J. J. Condon, 1988). These distributions are a collection of results from a multitude of radio astronomy surveys spanning a number of decades. It may be possible to generate a sample from these known distributions. However, as these distributions are the statistical results of a large collection of surveys, they do not take into account the limitations of individual telescopes, such as the resolution, signal-to-noise ratio or dynamic range of the observation. These limitations would have to be applied to the distributions if a comparison were to be made between the simulated distribution and the real sample. Applying such a method would therefore require knowledge of these limitations for each observation. It would, therefore, be preferable to use an empirical approach, where the real sample data is

used to generate the flux distributions, in order to allow for this algorithm to be applicable to any radio survey image.

Initial attempts to generate a simulated sample of the flux values from the real sample set used the Monte Carlo method. The Monte Carlo method uses repeated random sampling of a probability distribution function in order to generate a sample of numerical values. This method, therefore, requires a continuous function to be applied to the flux distributions. Inspection of the flux distribution lead to the selection of two candidate functions, including the exponential and log normal function, that would best describe the flux distribution form. From the `Scipy optimize` package suite, the `curvefit` function was used to determine the exponential and log normal parameters respectively that would result in the best fit curve to the flux distribution function. It was found that the log normal function was the preferred candidate probability distribution function. Random sampling of the log normal function was undertaken to produce a sample of numerical values that could be used as a simulated data set that represented the flux distribution of the real sample. However, the deviation from the true distribution was found to be too great and alternative methods were investigated.

Subsequently, it was decided to generate the simulated sample of source fluxes using an empirical distribution method (Barton and Schruben, 1993). From the probability density function of the integrated flux of the real sample, a cumulative distribution function was constructed using the `Numpy cumsum` function. As its name describes, this function is used to generate the cumulative sum of the probability density function values, a discrete cumulative distribution function pertaining to the original bin values of the discrete probability density function. A random sample, using a uniform distribution between 0 and 1, are then generated using the `Numpy random.uniform` function. Applying these sample values to the cumulative distribution function results in a source flux value per sample value. Instead of a single flux value per bin, the source flux value was calculated by interpolation, using a gradient determined by the bin edges and the source flux range per bin. Using this method a large sample of source flux densities could be generated for the simulated data set that would

closely adhere to the distribution of the real data.

5.2.3 Sample the simulated set and apply source positions in a uniform distribution

Using the empirical distribution method described above, a large simulated data set was generated from the flux density distribution of the real sample. The size of this simulated data set could be adjusted depending on the size of the real sample. A simulated set that was hundreds of times the size of the real sample was created. The large simulated set was then sampled for a simulated sample, having the same size as the real sample. By sampling the simulated set repeatedly the comparison between the real and simulated sample could be completed a large number of times. The average of the results would then be closer to the expected value, as opposed to the result of a single trial.

In order to determine the nearest neighbour distances and the flux product for each source and its nearest neighbor, the sources of the simulated sample were given a randomly generated position, described by a right ascension value and declination value. These values can be equated to an X and Y position value in a two dimensional space. The position values were first generated in the 'pixel space' based on the dimensions of the radio image in pixels and then converted to right ascension and declination using the coordinate data associated with the radio image `FITS` file and the `WCS` function from Python `astropy.wcs` package which converts pixel values to world coordinate systems, such as right ascension and declination. The primary beam correction threshold or weighting threshold that was used during the source identification step were taken into account when generating the position values. The primary beam correction values are assigned per pixel based on the distance from the observing centre. Therefore the primary beam correction threshold limits the region where the real samples are detected. Even though the radio image describes a 13 square degree area, the primary beam correction threshold limits the sample area to a subsection

of the total image area. If this limit was not taken into account, the simulated sample, which contains the same number of data points as the real sample, would have randomly generated positions in an area greater than that of the real sample. This would result in much larger distances between nearest neighbours, and the simulated sample would not accurately represent the uniform distribution that is assumed to exist in the real sample. Therefore, the same primary beam correction threshold was applied to the positions generated for the simulated sample. A uniform distribution was used to generate both the right ascension and declination values. The ranges of the uniform distribution were set as the bounds of the radio image, described by the maximum and minimum right ascension and declination values respectively. However, only positions that fell above the primary beam correction threshold were accepted. Furthermore, a minimum flux threshold was applied to each source using the sensitivity image weights. As the sources in the simulated set are based on the true sky integrated flux (after applying the weights to the real sources), the weights were applied to the simulated sources based on their randomly generated position, so that sources that were placed at the edge of the GMRT EN1W region were reduced in flux value to reflect the primary beam shape. If reducing a source's flux values caused the flux to fall below the flux threshold the position of the source was recalculated. These positions were generated until all the data points of the simulated sample were assigned a position.

5.2.4 Generate a catalogue of nearest neighbours

At this point in the pipeline, a real sample and simulated sample are available. Both samples have the same number of sources, and each source is described by an integrated flux value and a position. The next step in the algorithm is to determine the nearest neighbour for each source in the respective samples.

K-nearest-neighbour (knn) algorithms are useful to determine the closest data points to a particular point in a sample space, where k is the number of nearest points that are of

interest. A naive approach to the *knn* algorithm is a brute-force method, where the distance, d, from each point of enquiry, n, is calculated for every data point in the sample space. This naive approach results in an $O(nd + kn)$ runtime. In the case of our samples, where $d = n$ and $k = 1$, the expected runtime using a naive approach would therefore have an $O(n^2)$ runtime (Cislak and Grabowski, 2014). This is an important consideration when dealing with the scale of data that is expected from future large radio surveys, where the data points may number in the millions. Two alternative *knn* algorithms were considered, including the ball-tree *knn* algorithm and the k-d tree *knn* algorithm.

The ball-tree algorithm implements a hierarchical binary tree structure in order to map out the sample space. The sample space is first divided into two hyper-spheres, where each point is associated with the hyper-sphere to whose centroid it is closest to. Each hypersphere is then sub-divided into two hyper-spheres, and each sample is again associated with the sub-sphere whose centroid is closest. This process is repeated until a certain tree depth is reached. Points that fall within a ball are expected then to be closest to points inside the ball. Points that lie close to the boundary of a ball, however, may be closer to data points in a neighbouring ball. The initial setup of the ball-tree binary tree requires considerable memory and computation, as the distance of each data point must be calculated to determine the hyper-sphere in which it falls, but once the structure is set up, the searches for the nearest neighbours to any point in the sample space is quick (Cislak and Grabowski, 2014). The ball-tree algorithm is considered particularly efficient in multi-dimensional spaces, where the number of sample attributes are high (Rajani, McArdle, and Dhillon, 2015).

The k-d tree nearest neighbour algorithm is similar to ball-tree algorithm, in that it also implements a hierarchical binary tree structure. The sample space data is considered for the attribute with the highest variance. The sample space is then divided in two at the median point of this attribute, known as the splitting hyperplane. The two subspaces are then considered, and each subspace is divided at the median point of the attribute with the greatest variance within the subspace, without considering the previous attribute. This

process is repeated until there are either one or two data points in each sub-space. The first median point is inserted as the root node, and subsequent median points are inserted as the sub-nodes. This results in a balanced binary-tree structure with $log(n)$ depth (Cislak and Grabowski, 2014). To find the nearest neighbour to a given point, the binary-tree is searched recursively. Starting from the root node, it traverses the tree, going left or right depending on whether the point under enquiry is less than or greater than the value at the node. At each node, it calculates the distance between the point and the node value, and stores the node with the current smallest distance. Points that lie close to the splitting plane may be close to data points on the other side of the splitting hyperplane. A hypersphere centered on the point of enquiry whose diameter is equal to the smallest distance is determined. If the hypersphere includes a neighbouring area to the splitting hyperplane, this area must also be searched for nearest neighbour points. The k-d tree nearest neighbour algorithm is considered to perform poorly with high dimensionality. High dimensional data results in more data points having to be searched to find the nearest neighbour, resulting in a similar performance to the brute-force method.

In the case of the astronomical data that is under investigation, the sample space is positional data in a two-dimensional space, including the right ascension and declination value for each radio source. On account of the low dimensionality of the data the k-d tree algorithm is preferred to the ball algorithm. The k-d tree nearest neighbour algorithm is available in the `SciPy` suite in the `spatial` package as the `cKDTree` function, as well as in `AstroPy` suite in the `SkyCoords` package through the `match_sky_coords` function. The `cKDTree` was chosen to be used as this function returns a binary tree object that can be queried, and therefore easily allows for the nearest neighbour search to be expanded from $k = 1$ to higher values of k.

The `cKDTree` package implements the k-d tree using a sliding midpoint split as described by Maneewongvatana and Mount (2002) rather than a median or midpoint split when determining the position of the hyperplane. The sliding window technique overcomes an issue

where if points are all clustered at one end of the sample space, the standard splitting of the plane with greatest variance may result in elongated subspaces with poor aspect ratio, or if midpoint splitting is utilized, subspaces without data points may occur, where all the data points occur on one side of the split, resulting in an elongated tree. In the first instance, the hypersphere used to determine which neighbouring subspaces should be queried will cross many additional boundaries due to the elongated nature of the subspaces, while in the second instance, an elongated tree results in longer search times when completing a query. With the sliding-midpoint method, the midpoint of a subspace is first determined for the split. If all the data points lie to one side of the splitting hyperplane, the position of the hyperplane is adjusted until it intersects with at least one point. This method ensures that no trivial subspace exists, where no data points are located.

Once the nearest neighbour for each source was determined the index, position and flux of the nearest neighbour and distance in arcsec between the source and its nearest neighbour was recorded for each source and included in the real sample catalogue.

5.2.5 Comparison of the real sample and the simulated sample

At this point the nearest neighbouring source for each data point in the real sample and simulated sample respectively has been determined. The intention of the analysis is to identify multi-component sources in the real sample and consequently in the radio image. Radio galaxies are known to be multi-component sources. These sources are radio loud and are often considerably more luminous than the surrounding galaxies. A distribution of the flux density of galaxies in a radio survey image of the sky reveals that faint sources dominate and that bright sources are rare. Therefore, the occurrence of two large, bright sources in close proximity to one another is particularly rare. In the real sample, it is known that radio galaxies exist and therefore two bright sources in close proximity may identify a radio galaxy. While in the simulated random sample, no multi-component radio galaxies exist

and similarly two bright sources in close proximity to one another will be considerably rare. Using this knowledge, the analysis of the differences between the real and simulated samples will be conducted using two parameters: 1) the distance between the nearest neighbour pairs, nn_d, and 2) their flux product, $S_1 \times S_2$. The flux product will be large for the case where both sources in a pair are bright. A two-dimensional discrete distribution for each of the samples was constructed using the parameters nn_d and $S_1 \times S_2$. For each sample the distributions had k^2 bins, where each bin was length Δx and Δy for the parameters nn_d and $S_1 \times S_2$ respectively, and the bins were indexed (x_i, y_j). In order to conduct the comparison between the distributions of the real and simulated sample, the Function 5.2 is derived to compare the number of events per bin of the discrete distributions. Here, $n_{sim_{(x,y)}}$ is the number of events that occur in bin (x_i, y_j) of the simulated sample distribution, and $n_{real_{(x,y)}}$ is the number of events that occur in bin (x_i, y_j) of the real sample distribution. The function results in a value assigned to each bin of the sample space.

$$f_{x,y} = \begin{cases} 1 - \dfrac{n_{sim_{(x,y)}}}{n_{real_{(x,y)}}} & n_{sim_{(x,y)}} \leq n_{real_{(x,y)}} \\ 0 & n_{sim_{(x,y)}} > n_{real_{(x,y)}} \end{cases} \tag{5.2}$$

For bins where the number of events in the real sample is equal to the number of events in the simulated sample the function will result in a value of zero. For bins where the number of events in the real sample exceed the number of events in the simulated sample, the function will result in a value between 0 and 1, where the larger the ratio of real events to simulated events per bin, the closer the resultant value is to 1. A value of 0.5 indicates that there are twice as many events in the real sample bin than in the equivalent bin in the simulated sample distribution. A value of 1 indicates that no event occurred in the simulated sample bin, while one or more events occurred in the real sample bin equivalent. In bins where more events occurred in the simulated sample than in the real sample, that is to say, where the score would be negative, the score was set to zero. The score can be interpreted as the

probability for each object in bin (x_i, y_j).

When constructing the distributions of the simulated sample, a large sample set, much greater than the size of the real sample, was generated. This large set was then sampled to create a simulated sample that was equal in size to the real sample. The two-dimensional discrete distributions for this simulated sample was then constructed and Function 5.2 was applied to this simulated sample and the real sample distributions. This process of sampling from a larger sample set was repeated, each time applying Function 5.2. Bin edges of the distributions were kept constant. The output values of the function were recorded for each bin. These values were then averaged to produce a final score per bin and the standard deviations for each bin was calculated. The bin index for each source in the real sample was determined and the score associated with the bin was assigned to this data point. A catalogue of sources in the radio image was then output with the scores and standard deviation of the scores for each source. As the score tends towards a value of one, it represents a likely multi-component source that is more and more statistically significant when compared to the random sample. That is to say, high scores indicate a relationship between the flux and distance of the source and its nearest neighbour that lies outside a random distribution of sources.

In order to identify radio galaxies a filter could be applied to the catalogue which selects sources where the product of the flux of a source and its nearest neighbour is above some threshold value. High scores associated with sources when this filter is applied will then identify likely multi-component source candidates, for example, radio galaxies. High scores at other extremes, such as at low nearest neighbour flux product values, may indicate faint radio galaxies that may require further investigation.

5.3 Extension of detection using multiwavelength data

The GMRT EN1W field is one of the most observed regions of the sky (Ocran et al., 2020). Data from several observations of this region exist across different wavelengths. Images and catalogues from Spitzer Adaptation of the Red-sequence Cluster Survey (SPARCS), SWIRE and SDSS, amongst others, are available, which include observations at optical and IR wavelengths. The correlation between optical and radio flux is not very certain, the radio beam is often much larger than the optical point spread function, and the source density of optical sources is much greater than that of radio sources. For these reasons, the identification of radio sources with their optical counterparts is very ambiguous. However, in the infrared, some of these limitations can be overcome. While the jets and lobes associated with radio galaxies are not visible, the central galaxy may be visible. On account of this, the SWIRE IRAC infrared images and catalogues were used to extend the detection method for radio galaxies.

The proposed method is described as follows. The detection of radio galaxies at radio wavelengths should identify the radio lobes and nearest neighbour in the automated detection method. Using the output catalogue of the automated detection method and the IRAC catalogue, matches between the radio sources and sources in the IR catalogue are determined, with a maximum distance of 5 arcsec between matched sources, due to the synthesised beam size of the radio image. Sources from the radio catalogue are grouped into four groups, including: group A, where both the source and it's nearest neighbour have a match in the IR catalogue; group Ba, where neither the source nor its nearest neighbour has a match in the IR catalogue, and both sources are positioned within the area observed by the IRAC EN1 observations; Bb, where neither the source nor its nearest neighbour have a match in the IR catalogue, and both sources are positioned outside the area observed by the IRAC EN1 observation (that is to say, no IRAC image or catalogue data for these sources are available); and group C, where either the source or it's nearest neighbour has a match in the

IR catalogue. Of these four groups, it would be unlikely that group A would represent radio galaxy components as both source and nearest neighbour have IR matches, group Ba is of considerable interest as neither source or its nearest neighbour has an IR match and therefore could potentially represent the radio galaxy lobe components. Group C is also considered interesting as it's possible that an IR source could be superpositioned at the location of one of the components, however, it's unlikely that such a superposition would occur at both components, in the case of group A. Group Bb is not considered as no data is available.

Once these groups are determined, a search for any IR sources positioned between the location of the radio source and its nearest neighbour is carried out. A rectangular-shaped area is determined, with a width equal to that of the major axis of the synthesised beam of the radio image, extending along a line drawn between the radio source and it's nearest neighbour, while excluding these radio sources' positions. If one or more IR sources are found within this rectangular area, the probability of finding a source within this area must be determined. A low probability would suggest that finding a source within the area between the two radio sources is highly unlikely and therefore significant.

The IRAC observations include 16 fields that together encompass much of the EN1W field, however a portion of the GMRT EN1W field has no corresponding IRAC data. A total of 573843 sources were included in the IRAC catalogue. In order to determine the probability of finding an IR source in the area between the two radio sources, the IR catalogue was searched for all sources that occurred in each of the 16 regions. The central four fields (fields 6,7,10,11) of the 16 IRAC fields were used to calculate the probability of finding one or more source as a function of area. A box was then determined, centred at a random position within the four central regions, where the area of the box would fall entirely within this region. The position of the box centre was determined by generating a random value from a uniform distribution within the bounds of the fields and confirming that the bounds of the box fell within the region of the four fields. The number of sources within the box was then counted. This process was repeated 50000$\times$ and the number of events, where more than one

source was found within the box, was recorded. The probability of finding more than one source was determined by taking the average of these values. The box size was then varied and the process was repeated. The result of this process was an empirically determined probability distribution of finding one or more source as a function of area.

The groups Ba and C were then examined for IR sources that fell within the region between the primary source and its nearest neighbour. A rectangular area was determined between the two sources, centred on a line drawn between the two sources, with a width equal to the length of the major axis of the synthesised beam. In each case, the number of sources that were found within this area was recorded as well as the area in square degrees. The area was then compared to the empirically generated probability distributions and a probability of finding one or more IR sources within the area between the radio sources was determined and recorded. Instances with a lower probability were considered to more likely represent occurrences of a radio galaxy, while a high probability of finding an IR source between the radio sources suggest that such a source could have occurred as a random event.

A catalogue was generated of the radio source pairs of groups Ba and C, and included the area between the sources in square degrees, the number of IR sources that were found in this area, and the probability of finding one or more IR sources in this region as a function of area. From this catalogue, where one or more IR source were found in the region between the radio sources, cutout images were created for visual inspection.

Chapter Six

Application of the algorithm to GMRT EN1W radio image

The automated statistical method was applied to the GMRT EN1W radio image. The radio image data included a Flexible Image Transport System (`FITS`) image file and a FITS weights file. The `FITS` file is a common format for astronomy images and typically contains a header and data. The header contains information about the observation, such as which telescope or array undertook the observation, the date of the observation, the coordinates of the image centre, the frequency at which the observation took place, details about the synthesised beam, and the dimensions of the pixels in terms of an astronomical coordinate system, such as right ascension and declination. The data is commonly a $n \times m$ matrix, where each value indicates the brightness, or other parameter, associated with the pixel of the image. In the case of GMRT EN1W image, the unit of the pixel value is μJy/beam. The `FITS` weights file is used for primary beam correction. As mentioned in Section 5.2.1, a sensitivity or weights file is applied to the radio image to correct for the fact that the sensitivity of the detector is highest towards the centre of the image.

The GMRT EN1W image was observed at 610 MHz. The image centre was 16h14m00.1162s RA and +54:59:59 DEC. The dimensions of the image were `13000x13000 pixels`, with a total area of 13 deg^2. The synthesised beam major and minor axis were calculated to be 5.3

arcsec respectively.

Source finding on the GMRT EN1W image was conducted using `PyBDSF`. The input parameters for the `bdsf.process_image` function that were used are described. The rms box size was set to 200 with a step size of 50 (`rms_box=(200,50)`). The step size is the number of pixels the box is moved before the rms is calculated again. An adaptive box to calculate the local rms was used (`adaptive_rms_box=True`), where the size of the box used to calculate the rms is reduced for bright sources in order to account for strong artifacts that are common around bright sources. The rms box size near bright sources was set to a size of `30 pixels` with a step size of 11 (`rms_box_bright=(30,11)`). The threshold for identifying source peaks was set to 5.0 (`thresh_pix=5.0`) and would find all sources peaks 5σ above the rms value. The rms value here is local to the source and calculated using the rms box. The threshold for determining the island pixels that would be included when fitting was set as 4.0 (`thresh_isl=4.0`), where island pixels 4σ above the rms mean would be included. `PyBDSF` includes wavelet decomposition module that is useful for determining extended sources, however, this module was disabled (`atrous_do=False`) during source finding.

A total of 5570 sources were detected in the GMRT EN1W image using `PyBDSF`. The thresholds to exclude unreliable sources were applied to the output source list of the `PyBDSF` source finding, included a minimum signal-to-noise ratio threshold of 4.0, a minimum flux density threshold of 0.499 μJy, a maximum primary beam correction threshold of 20 (calculated as the inverse of the primary beam limit of 0.05), and a minimum flux ratio threshold of 0.25. 23 sources were rejected based on these limitations, resulting in a catalogue of 5547 sources. The weight values of the sensitivity image were then applied to the peak and integrated flux values of the detected sources using 5.1 to generate the true sky flux values for the final catalogue. This catalogue is the real sample for the remainder of the statistical approach.

For the real sample catalogue, the minimum peak flux was found to be 0.16 mJy and the

maximum peak flux was 664.80 mJy. The minimum integrated flux (the sum of fluxes of the pixels associated with a source) was found to be 0.24 mJy and the maximum integrated flux was 1701.97 mJy. The mean and median integrated flux was recorded as 8.26 mJy and 1.07 mJy respectively. Figure 6.1 shows the distribution of the integrated fluxes for the real sample. As mentioned, there is a preponderance of faint sources, with a 95th percentile integrated flux value of 27.37 mJy.

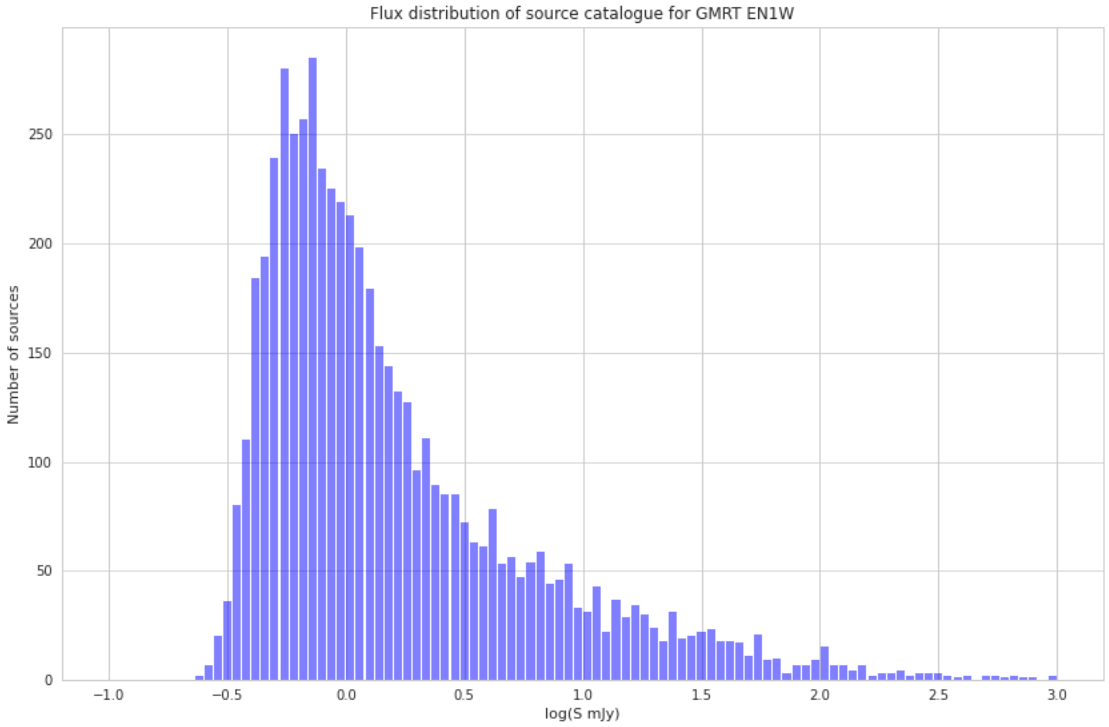


Figure 6.1 A Distribution of the integrated flux of sources in the real sample for GMRT EN1W

From the real sample, the integrated fluxes for the simulated sample are generated using the empirical method. Figure 6.2 and Figure 6.3 show comparisons between the distributions of the integrated flux for real sample (blue) and the simulated sample (red), with 5 547 and 500 000 samples respectively. From the distributions it can be seen that the variance reduces

with increased sample size. The variance for the sample size of 5 547 sources was calculated as 2 538.14, while the variance of the larger sample size was calculated to be 2 428.23.

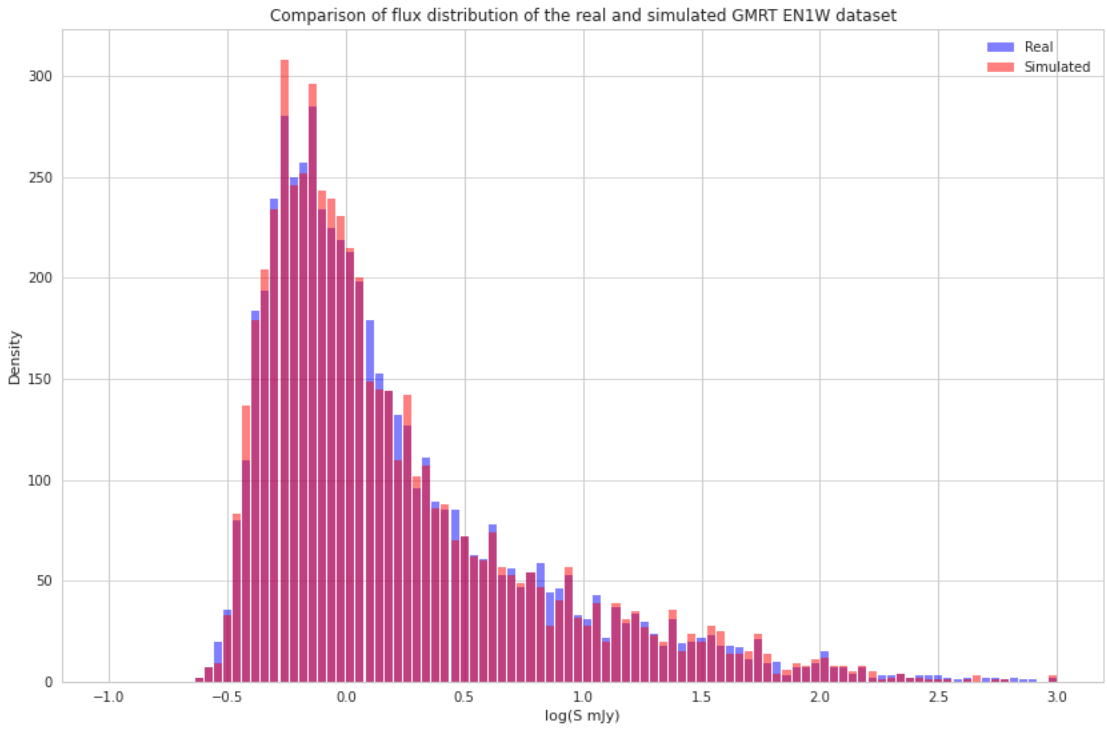


Figure 6.2 A Distribution of the integrated flux of 5 547 sources from the real (blue) and simulated (red) data set for GMRT EN1W region

Table 6.1 shows a comparison of the mean, median and standard deviation of the real and simulated data sets, where the full simulated data set is 500 000 data points, and the small simulated sample is a sample (the same number of values as the real sample) of the full data set. It can be seen that the values of the simulated sets lie close to those of the real data, where the full simulated data set is more closely correlated, due to the number of data points samples, than the simulated data sample. This is also evident in Figure 6.2 and Figure 6.3. As expected, there is greater variance in the smaller simulated sample than the full simulated data set. The average correlation between the real data and multiple samples

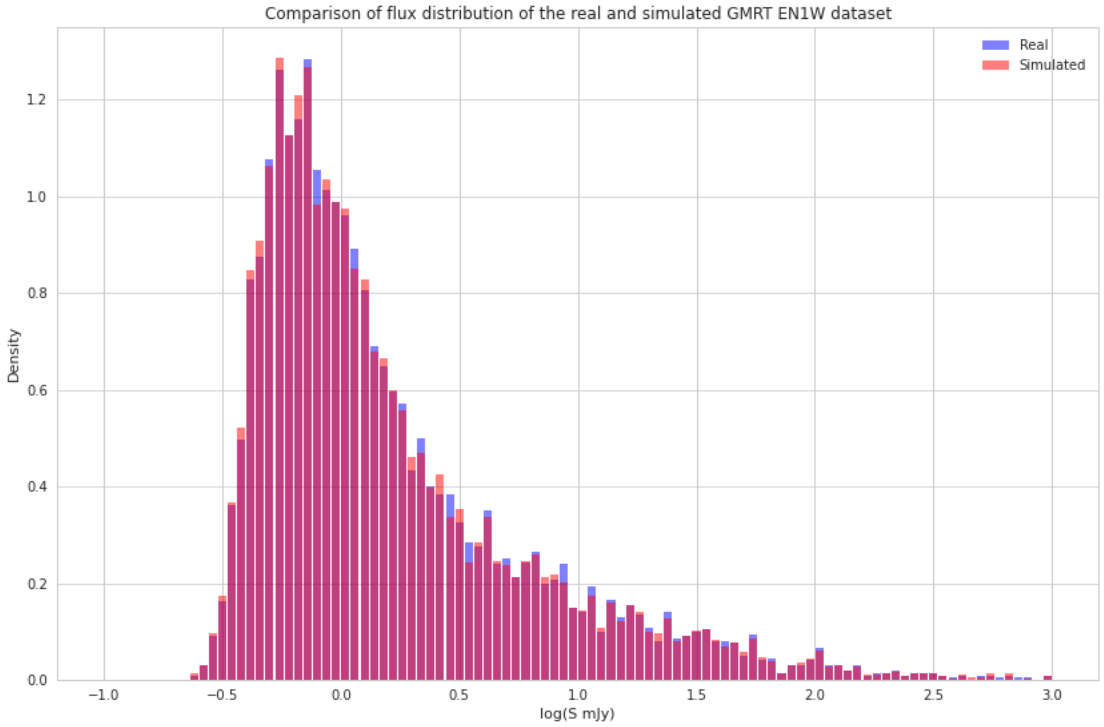


Figure 6.3 A Distribution of the integrated flux of the real (blue) data and 500 000 sources from the simulated (red) sample for GMRT EN1W region

of the simulated data set was found to be 0.983.

This simulated data with 500 000 was kept as the simulated sample. The position values for the simulated sample were generated based on the right ascension and declination values of the original GMRT EN1W image and were restricted to fall within the same region as the real sample when applying the primary beam correction threshold from the `FITS weights file`.

Figure 6.4 and Figure 6.5 show the plots of the positional data associated with each source in the real sample and a sampling of the simulated sample data respectively. The larger simulated sample was sampled using the same size as the real sample in order to provide a reasonable visual comparison.

	Real	Simulated (full)	Simulated (sample)
mean (mJy)	8.26	8.13	7.88
median (mJy)	1.07	1.06	1.02
σ *(mJy)*	49.40	48.26	50.65
min (mJy)	0.24	0.24	0.24
max (mJy)	1701.97	1686.27	1685.47

Table 6.1 Quantitative comparison of the integrated flux of the real and simulated GMRT EN1W data sets

The nearest neighbours for each source in the real sample were determined. The distance between each source and its nearest neighbour, nn_d, was calculated, as well as the product of their integrated flux values, $S_1 \times S_2$. This process was conducted on the simulated sample. However, the simulated sample was again sampled using a size equivalent to the real sample, in order to maintain a comparable nearest neighbour distance. An increase in the number of data points in the sample would significantly reduce the nearest neighbour distances, as more sources would exist in the same field size.

Figure 6.6 shows a comparison between the distribution of the nearest neighbour distances in the real sample (blue) compared to the distribution of the simulated sample (red). Portions of the distributions that overlap are visible as purple. To generate the simulated sample distribution, the simulated sample was sampled several hundred times using the aforementioned process to calculate nn_d and $S_1 \times S_2$. These results were then aggregated to create the distribution. A total of 500 000 data points were determined. The difference in size of the real and simulated sample should be therefore noted as the variance in the larger simulated sample would be expected to be greatly reduced in comparison to the real sample.

The mean and median nn_d for the real sample were found to be 75.05 and 66.60 arcsec respectively, while for the simulated sample the mean and median nn_d were 82.68 and 77.33

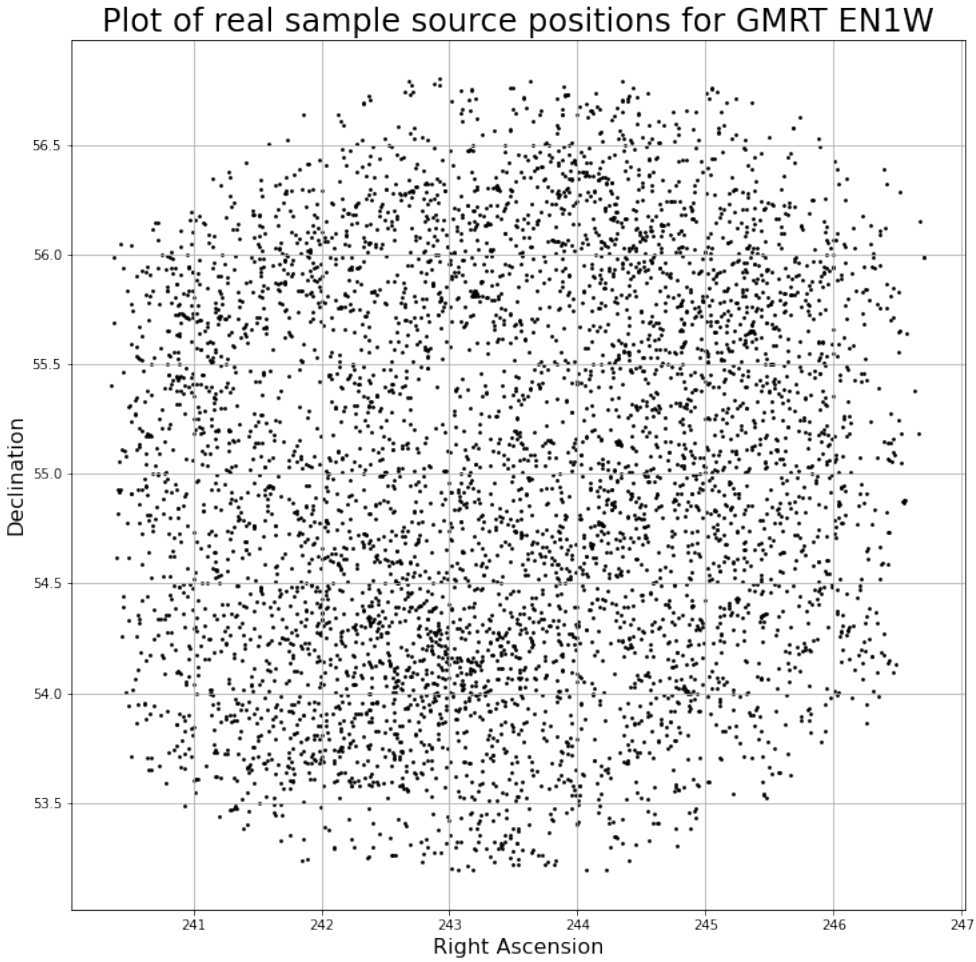


Figure 6.4 A plot of sources positions of sources from the real GMRT EN1W data set

arcsec. The minimum nn_d of the real and simulated data was found to be 6.24 and 6.00 arsec respectively. From Figure 6.6, it can be seen that the distribution of the simulated data takes on a Poisson-like distribution, while the distribution of the real sample has a bimodal or multimodal distribution indicating a significant portion of data points with smaller values of nn_d in the real sample compared to the simulated sample. This is aligns with the mean and median separation distance values of the real sample being lower than the simulated

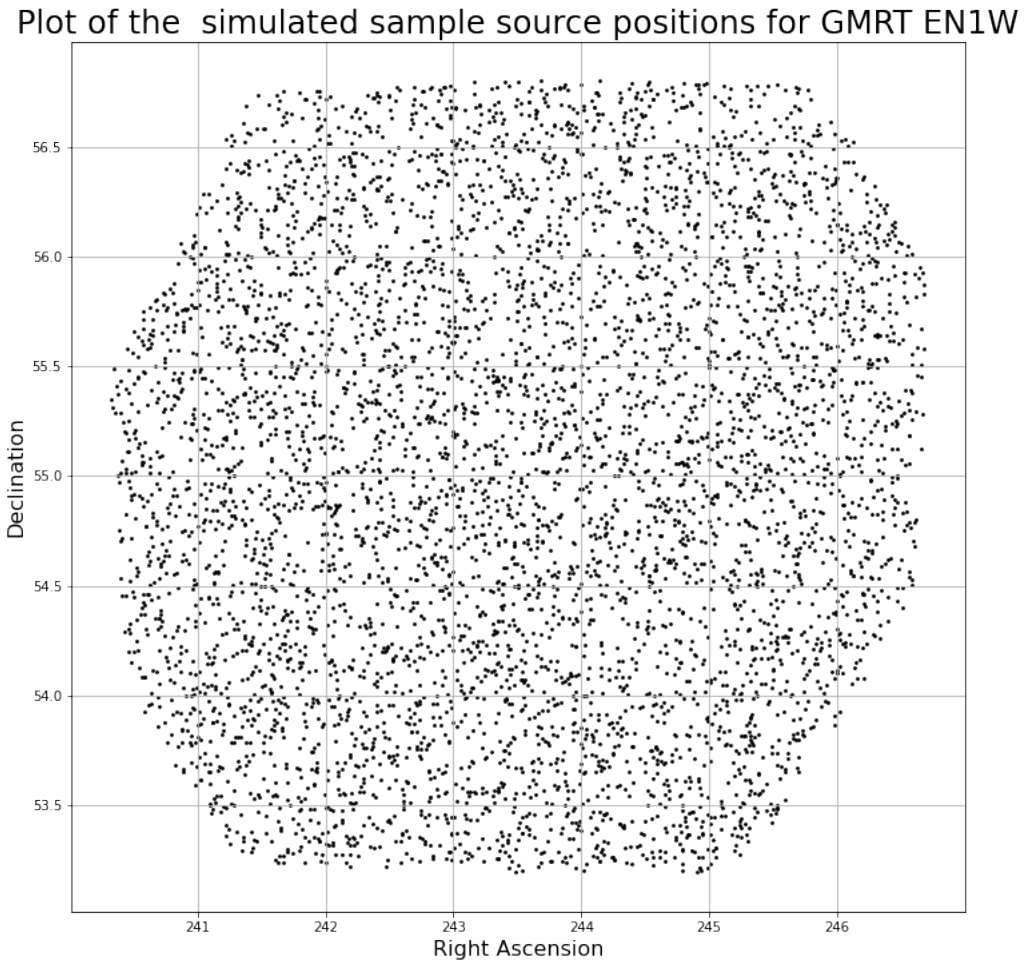


Figure 6.5 A plot of sources positions from the simulated GMRT EN1W data set

same. This suggests that there are more sources that are in close proximity to each other in the real sample than would be found if sources were uniformly spread throughout the area. This elevated occurrence of small nearest neighbour distances suggests relationships between these sources. It should be noted that the data in the two samples have not been standardized at this point, that is to say, converting the data to a z-score and, therefore, a normal distribution has not been completed. Usually the standardization technique allows

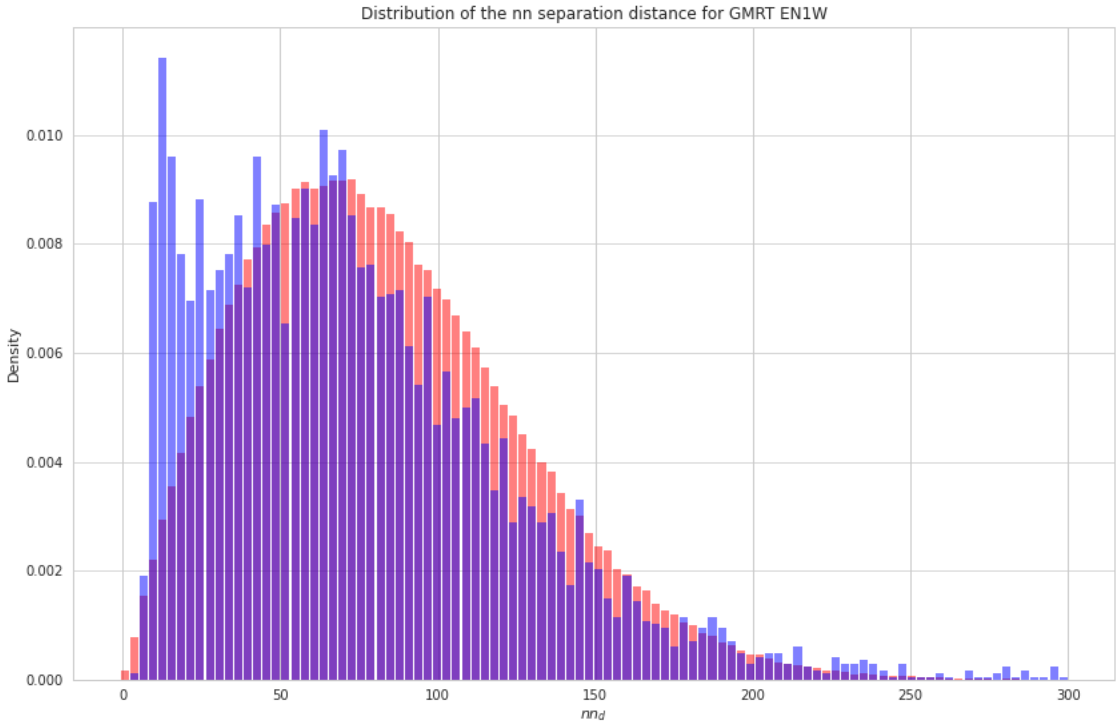


Figure 6.6 Distribution of nn_d for the real sample (blue) and the simulated sample (red) for GMRT EN1W

for comparisons of different distributions - even non-normal distributions can be converted to normal distribution by converting the scores to a z-score, and the resultant normal distributions may be compared. However, when converting the real data into a z-score it was found that the distribution narrows considerably more than the simulated data due to the multimodal structure of the real sample. The real sample's mean is less than the mean of the simulated sample, and its standard deviation is larger, resulting in a narrower z-score normal distribution. In Figure 6.6 and the distributions to follow, the real sample and simulated sample distributions have been separately normalized so that the area under the distribution is 1. However, when the real and simulated sample distributions are compared using Function 5.2 to generated the probability scores, the distributions are normalized by the value

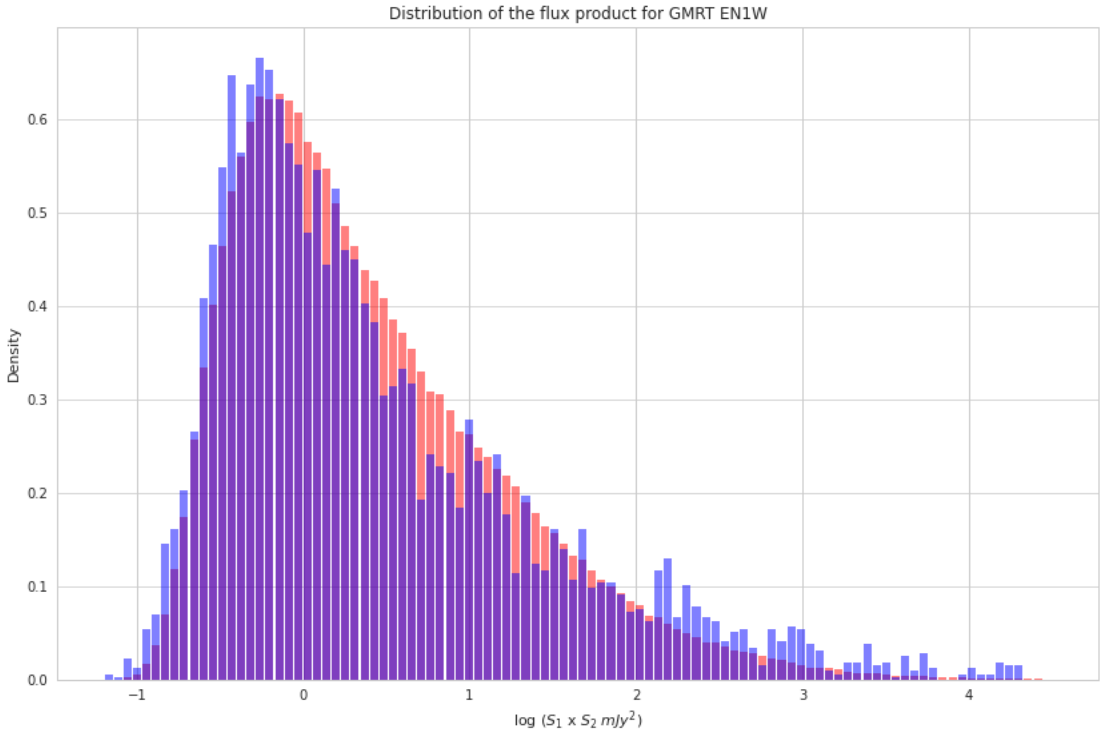


Figure 6.7 Distribution of $S_1 \times S_2$ for the real sample (blue) and the simulated sample (red) for GMRT EN1W

of the bin at the position of the bin with maximum value in the simulated two-dimensional distribution.

Figure 6.7 shows a similar comparison between the distribution of the flux product of the nearest neighbours identified in the real sample (blue) and the simulated sample (red). Portions of the distributions that overlap are visible as purple. The same process as described above was used to generate these distributions, therefore the real sample in this instance has 5 547 data points, while the simulated sample has 500 000 data points. The mean and median $S_1 \times S_2$ of the real sample were found to be 458.84 and 1.49 mJy^2 respectively, while for the simulated data set these values were found to be 62.93 and 1.66 mJy^2. The higher average $S_1 \times S_2$ value in the real sample can be attributed to the higher occurrence of radio

loud nearest neighbour sources that are components of a single physical system.

While the distributions of the real and simulated sample in Figure 6.7 match closely, there is some divergence in the real sample at low flux product values and at high flux product values. This divergence would suggest that there is a population in the real sample of higher flux product values than in the simulated sample. In the case of the high flux products values, the divergence in the real sample from the simulated sample distribution, where two nearest neighbour sources with high flux values are unlikely to occur due to the distribution of population source fluxes, may indicate occurrences of radio galaxies.

Plotting the above distributions as a single two-dimensional distribution further reveals the divergence between the real sample distributions and those of the simulated sample. Figure 6.8 and Figure 6.9 shows the two-dimensional distribution of real sample and the distribution of the simulated sample respectively, with the distribution of nearest neighbour distances on the horizontal axis and flux product on the vertical axis.

From Figure 6.8, it can be seen that the real sample distribution extends further to the top left of the distribution, indicating a high occurrence of high flux product values at small nearest neighbour separation distances compared to the simulated sample distribution. This indicates bright nearest neighbour pairs that cannot be paired by random association. There also appears to be a higher occurrence of high flux product values at nearest neighbour separation distances of 75 arcsec, however at a lower occurrence and lower flux product values than the first divergence mentioned. Figure 6.10 shows an overlay of the real and simulated sample two-dimensional distributions.

From the overlay, what stands out beyond the other two divergences between the real and simulated sample distributions that were mentioned, is the area of the distribution of the real sample at both low flux product values and low nearest neighbour separation distances. A number of data points exist in this area that did not occur in the simulated sample, suggesting a significant population of low flux density sources that are in close proximity to one another

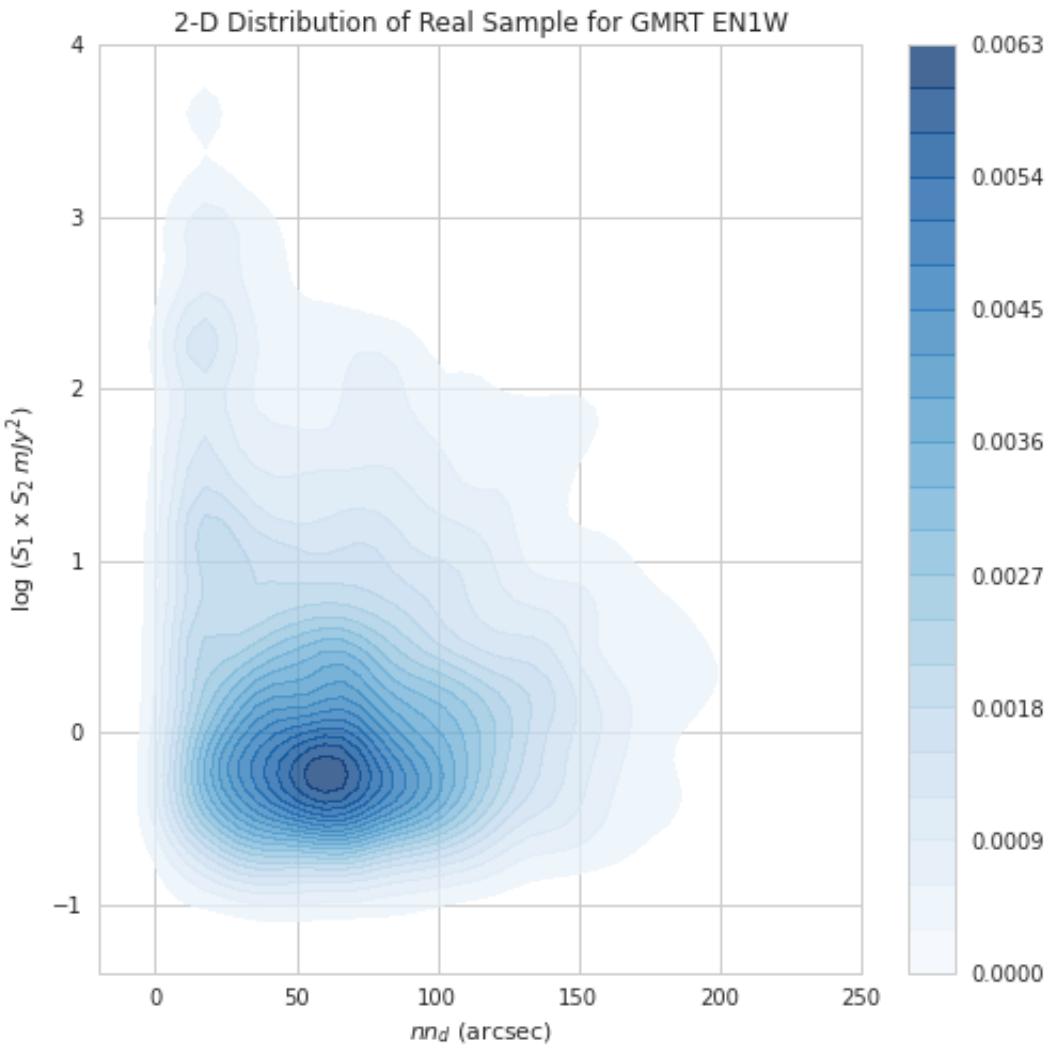


Figure 6.8 2-D distribution of nn_d by $S_1 \times S_2$ for the real sample for GMRT EN1W

The Function 5.2, which calculates the inverse of the ratio of the simulated samples to real samples bin count, was applied to the two distributions. As mentioned in the description of the method, the simulated sample was sampled using the size of the real sample, so that the number of points in each group were equal when applying the function. This was repeated a number of times and the average of the function result and standard deviation for each

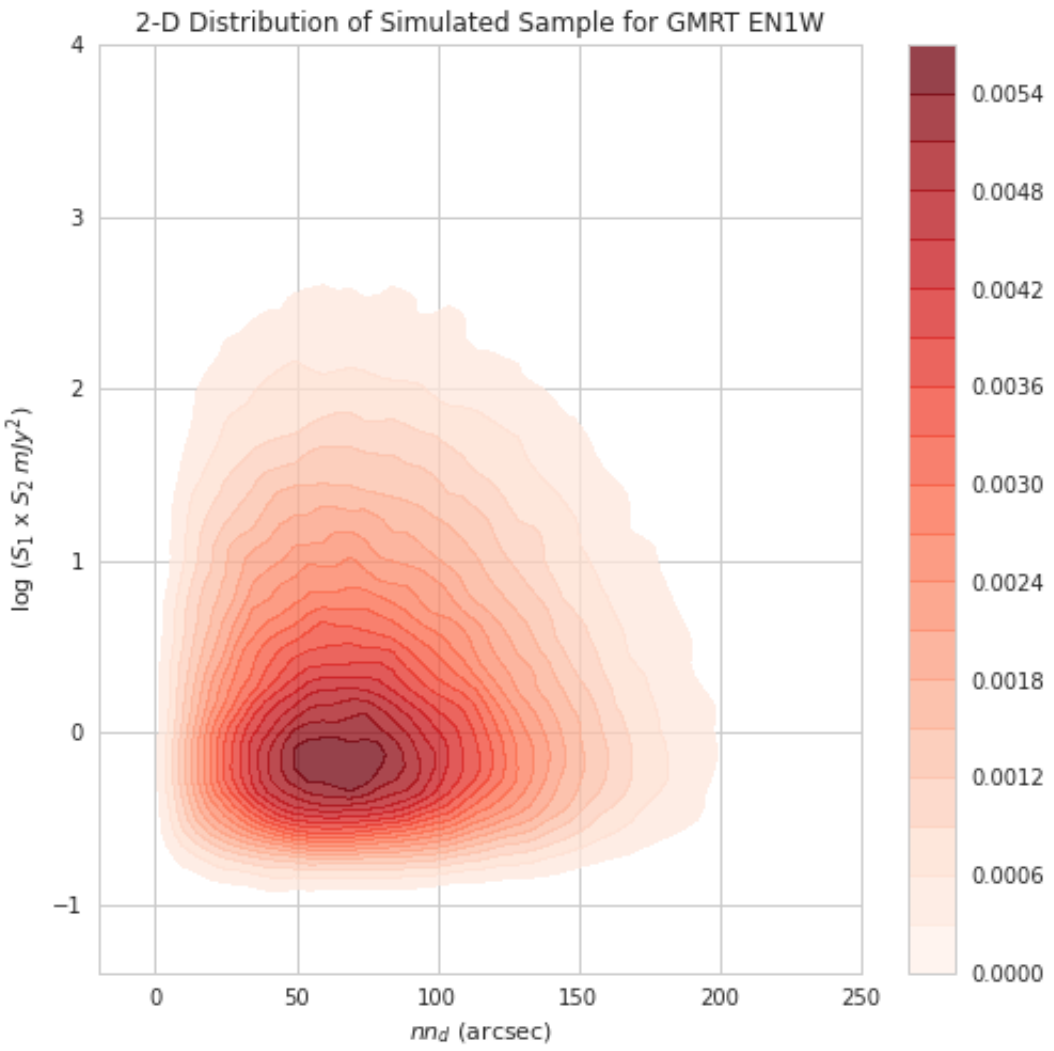


Figure 6.9 2-D distribution of nn_d by $S_1 \times S_2$ for the simulated sample for GMRT EN1W

bin were determined. The average value of Function 5.2 per bin was set as the bin score. All data points that fell into a particular bin were assigned the score associated with that bin. Figure 6.11 and Figure 6.12 show 'heatmap' representations of the bin scores and the standard deviation of the bin score respectively.

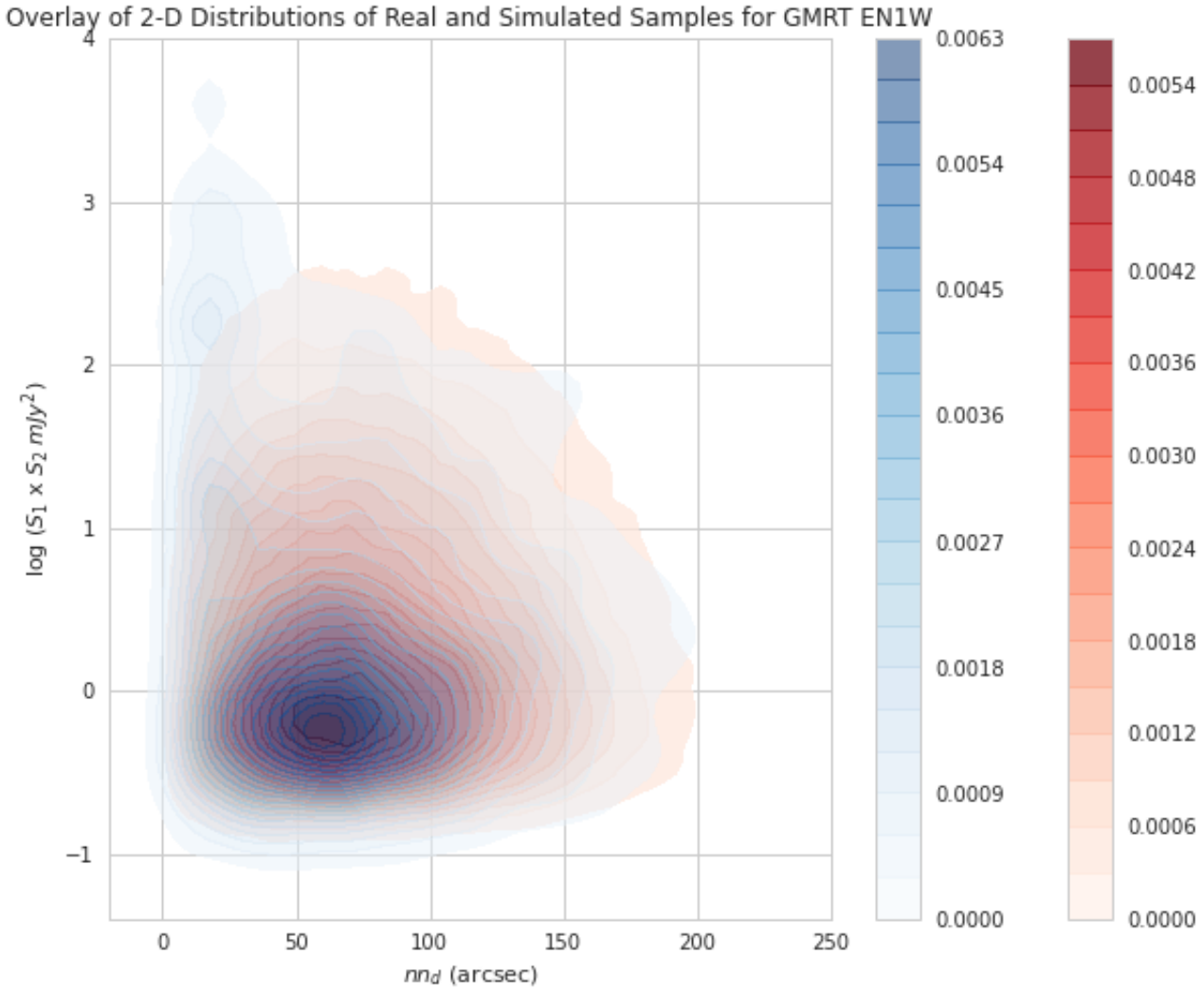


Figure 6.10 2-D distribution overlay of nn_d by $S_1 \times S_2$ for the real (blue) and simulated (red) samples for GMRT EN1W

A significant number of bins with a score > 0.8 can be seen at high flux product and low flux product values where the separation distance of nearest neighbours is low. Bins around the centroid of the distribution resulted in a low score, as much of the distributions overlap in this region. Further away from the centroid, at higher flux products and at both small and

Figure 6.11 2-D distribution of the output score by nn_d and $S_1 \times S_2$ for GMRT EN1W

large separation distances scores > 0.6 can be seen. It should be noted that the score does not infer any probability density information in so far it has no indication to the number of points per bin. The standard deviation of the score for each bin can be seen in Figure 6.12. Bins towards the centroid of the sample distribution have a standard deviation between 0.2 and 0.3. The centroid area is then ringed by a region of bins with higher standard deviation. Beyond this ring the standard deviation value for the bins drops rapidly, and outlying bins tend to have a standard deviation value of zero. The standard deviation value is interesting as it allows us to infer some information about the number of points per bin, where the score does not. Towards the centroid of the distribution, the standard deviation is mid-range, between 0.2 and 0.3. It is known that the majority of data points lie in this region. A low standard deviation value here suggest that there are many points per bin and that the variance in the bin count and score value was small. The standard deviation then increase in

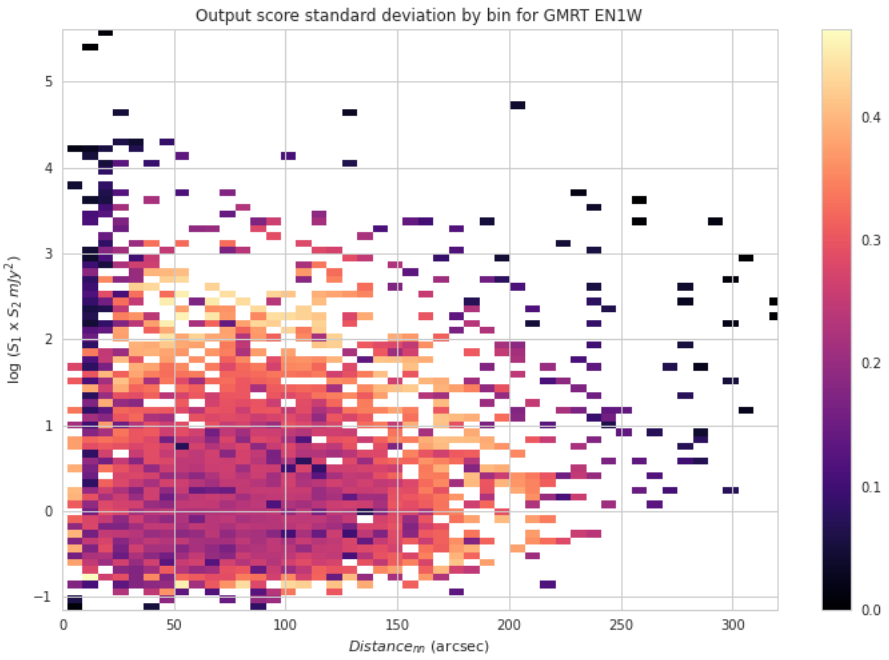


Figure 6.12 2-D distribution of the standard deviation of the output score by bin for GMRT EN1W

the ring around the centroid due to the fact that there are fewer points in this area than in the centroid, however, a number of points must occur in both the real and sample distribution in the area, due to the high standard deviation. Where the standard deviation is zero and the score is one outside of this ring, one can infer that no samples existed in the simulated sample in these bins, while one or more data points in the real sample occurred in these bins in this region. The score and standard deviation values were captured and included in the output catalogue.

The intention of the method is to detect multi-component sources, specifically radio galaxies. Radio galaxies that are radio loud and are detected through this method as nearest neighbours would result in high flux product values. By filtering the sources by flux product, score value and standard deviation it may be possible to extract the radio galaxies directly from the catalogue. Figure 6.13 shows the two dimensional distribution of the real sample

with a box (red) outlining an area of interest for where radio galaxies are presumed to be located in the 2D distribution.

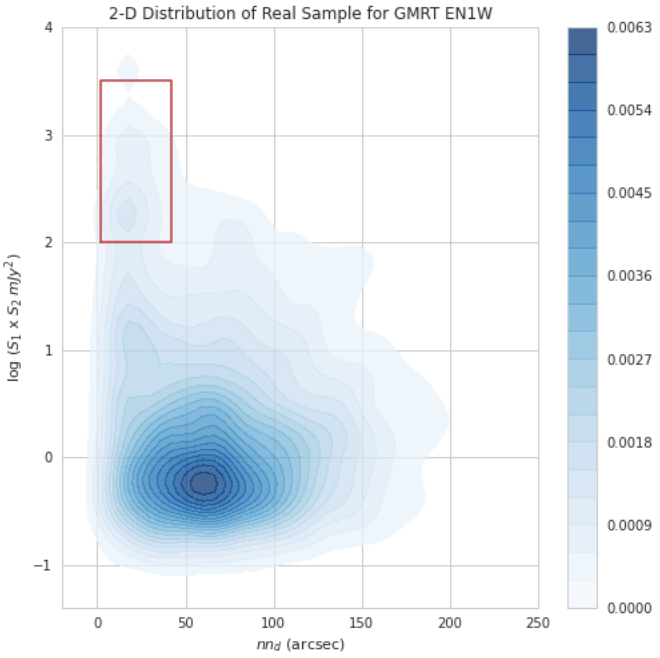
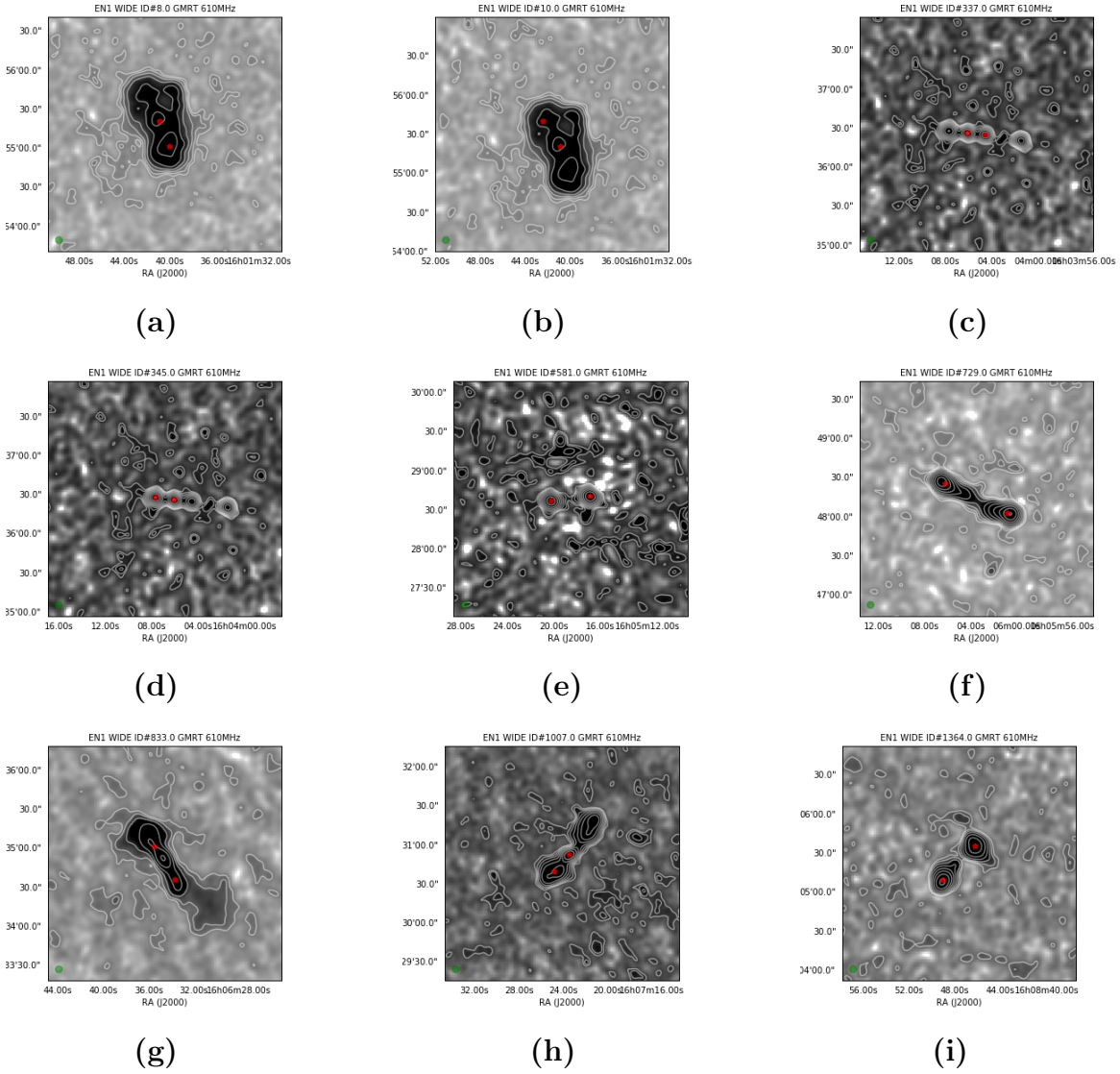


Figure 6.13 2-D distribution of nn_d and $S_1 \times S_2$ for the real sample for GMRT EN1W, with box indicating area of interest

The output catalogue of the algorithm was filtered using the following parameters and values: $nn_d < 100$; $log(S_1 \times S_2) > 2$; *score value* > 0.9, *score* $\sigma < 0.15$. A total of 93 unique nearest neighbour pairs met these criteria. Figures 6.14, 6.15 and 6.16 show the cutout images of several of these nearest neighbour pairs. The background image of the cutouts is the data from the GMRT EN1W radio image with contours of the flux values indicated in white. The position of the detected sources is indicated by the red stars. The synthesized beam size is shown in green in the bottom left corner of each image.



(a) **(b)** **(c)**

(d) **(e)** **(f)**

(g) **(h)** **(i)**

Figure 6.14 Cutout images of the nearest neighbour pairs meeting the following criteria: $nn_d < 100$; $log(S_1 \times S_2) > 2$; *probability score* > 0.9, *score* $\sigma < 0.15$. The background image is from the GMRT EN1W radio image data, with the contours of flux values indicated in white. The position of the source pairs indicated by the red stars.

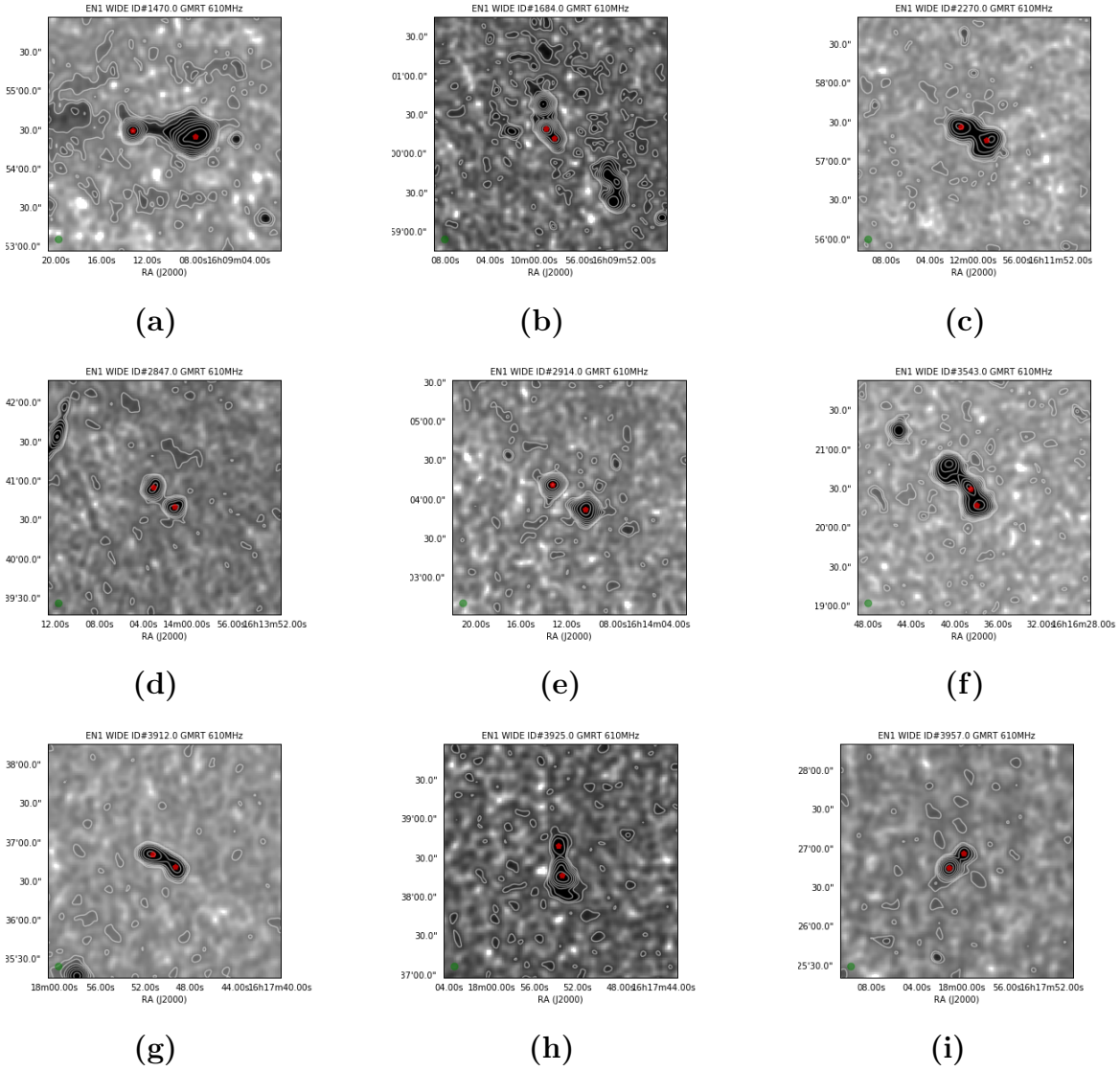


Figure 6.15 Cutout images of the nearest neighbour pairs meeting the following criteria: $nn_d < 100$; $log(S_1 \times S_2) > 2$; $probability\ score > 0.9$, $score\ \sigma < 0.15$. The background image is from the GMRT EN1W radio image data, with the contours of flux values indicated in white. The position of the source pairs indicated by the red stars.

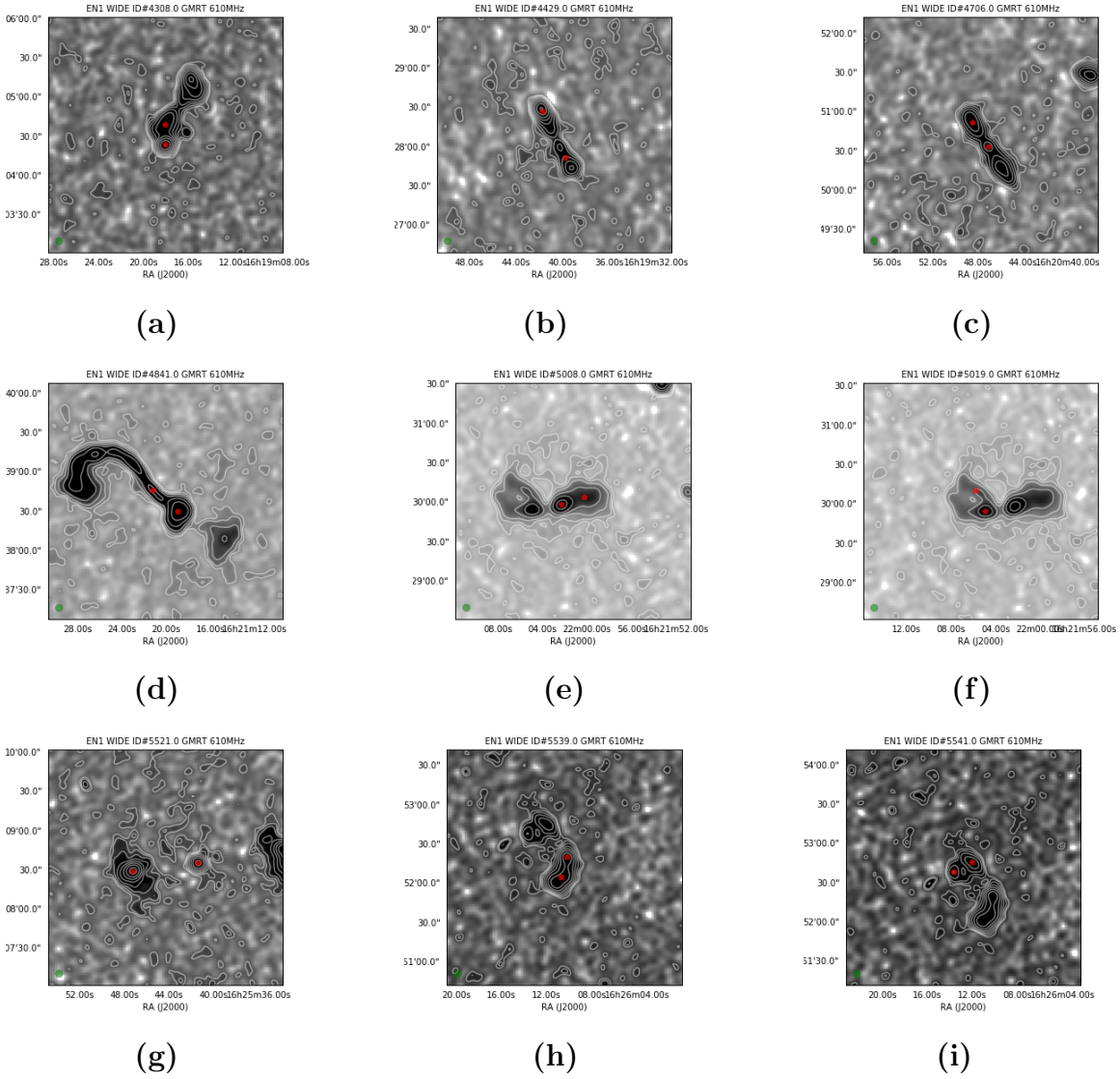


Figure 6.16 Cutout images of the nearest neighbour pairs meeting the following criteria: $nn_d < 100$; $log(S_1 \times S_2) > 2$; *probability score* > 0.9, *score* $\sigma < 0.15$. The background image is from the GMRT EN1W radio image data, with the contours of flux values indicated in white. The position of the source pairs indicated by the red stars.

Chapter Seven

Application of the algorithm to Stripe 82 radio image

The algoritm was applied to the JVLA SDSS Stripe 82 radio image using the method as described in Chapter 5, Sections 5.2.1 - 5.2.5. During the following chapter, the real sample and simulated sample will refer to those samples generated during the statistical method in association with the Stripe 82 radio image. The Stripe 82 image that was available had no sensitivity or weight information for primary bream correction, therefore, the primary beam threshold used during the source finding step of the statistical method was not completed. In addition, due to the absence of the sensitivity information, the boundaries for determining the position of simulated sources were calculated based on the ranges of the RA and DEC values of the real sample. This was considered appropriate due to the shape of the region of the Stripe 82 image, which was mosaic of hexagonal regions that were joined together to form a rectangular-shaped field, rather than the 'circular' shape of the GMRT EN1W image due to the primary beam shape.

The JVLA SDSS Stripe 82 radio image was `6168 x 49092 pixels` in dimension with a synthesized beam of 6 arcsec. A total of 4 393 sources were detected in the radio image using the `PyBDSF` software. The mean and median rms values of the radio image were found to be 114.01 μJy and 98.89 μJy respectively. Two sources, with peak flux $<$ 98.89 μJy or

snr < 4.0, were rejected from the source finding catalogue, based on the threshold criteria described in Section 5.2.1, leaving a final 4391 sources for the output catalogue.

The minimum integrated flux from the source catalogue was found to be 0.44 mJy and the maximum integrated flux was found to be 3 492.66 mJy with a 95th percentile of 160.808 mJy. Figure 7.1 shows the integrated flux distribution of the 4 391 sources detected during source finding. Similar to the integrated flux distribution of sources detected in the GMRT EN1W radio image, the distribution in Figure 7.1 shows a dominance of low flux sources.

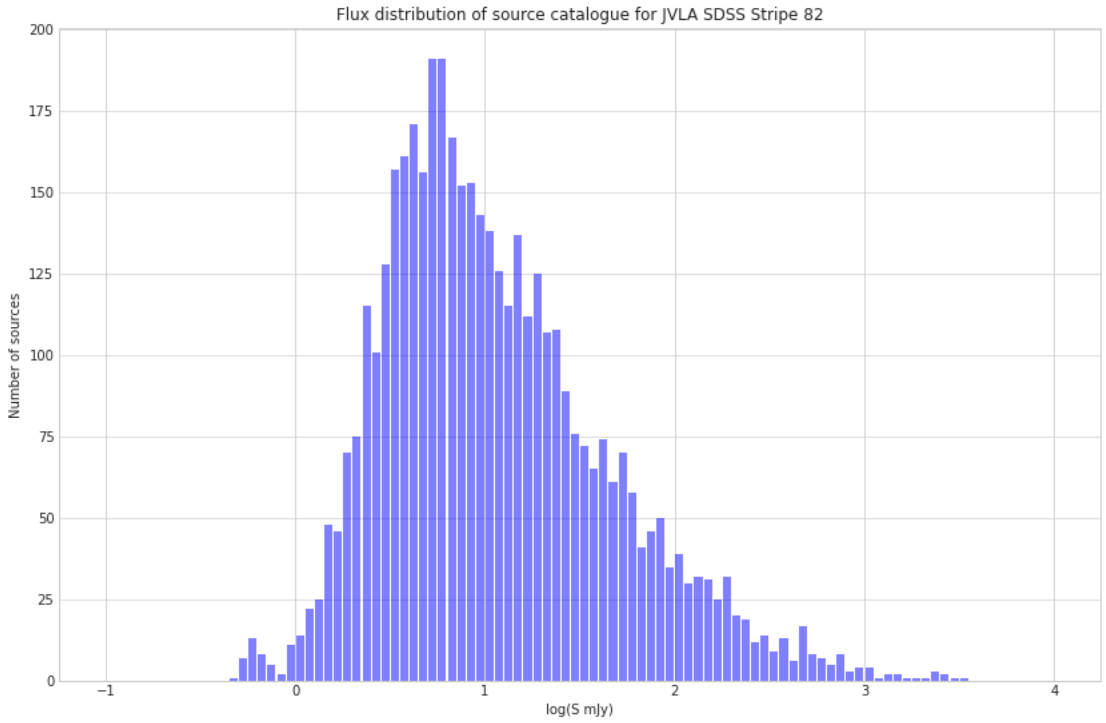


Figure 7.1 A Distribution of the integrated flux of sources in the real sample for JVLA SDSS Stripe 82

The simulated set, with 500 000 data points, was generated using the empirical distribution method from the flux distribution of the real sample. Figure 7.2 and 7.3 show the distribution of the integrated flux of the real sample (blue) compared to the simulated sample

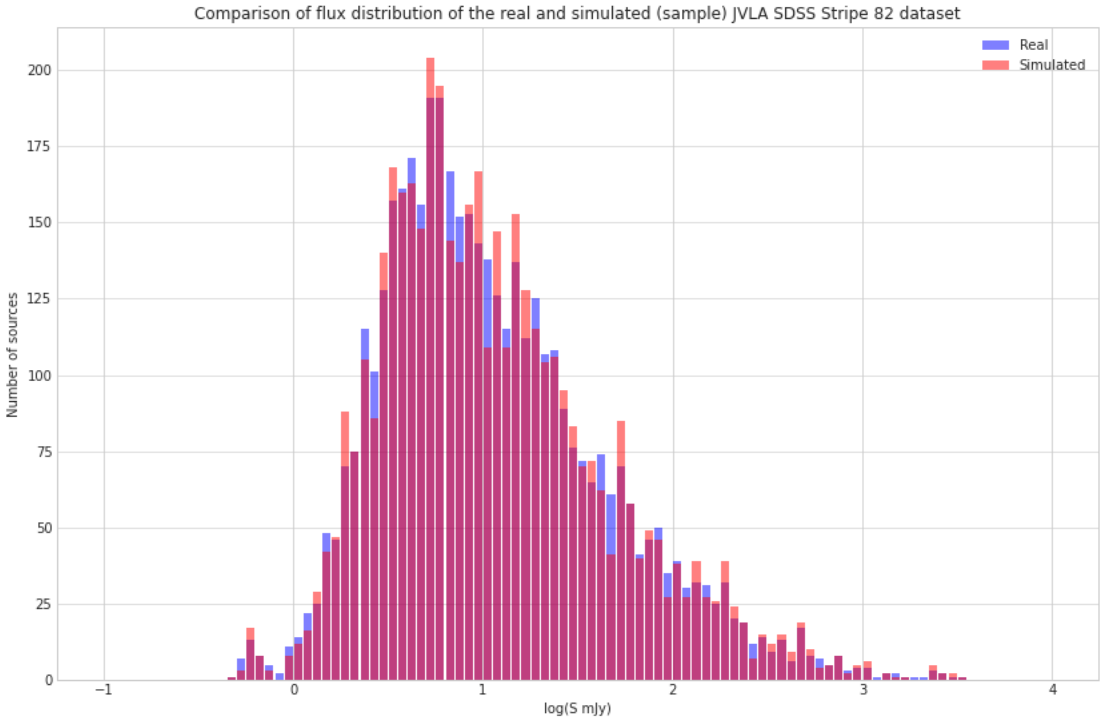


Figure 7.2 A Distribution of the integrated flux of 4 391 sources from the real (blue) and simulated (red) data set for JVLA SDSS Stripe 82 region

(red). Figure 7.2 demonstrates this comparison on a sample (4 391 values) of the simulated data set, while Figure 7.3 shows the comparison as a density with all 500 000 data points of the simulated data set. Table 7.1 shows a comparison of the mean, median and standard deviation of the real and simulated data sets, where the full simulated data set is 500 000 data points, and the small simulated sample is a sample (the same number of values as the real sample) of the full data set. Again, it can be seen that the values of the simulated sets lie close to those of the real data. The average correlation between the real data and many random samples of the simulated data set was found to be 0.993.

Positions were assigned to samples of the simulated data set where each sample was the size of the real sample. Figure 7.4 and Figure 7.5 show the source positions for the real

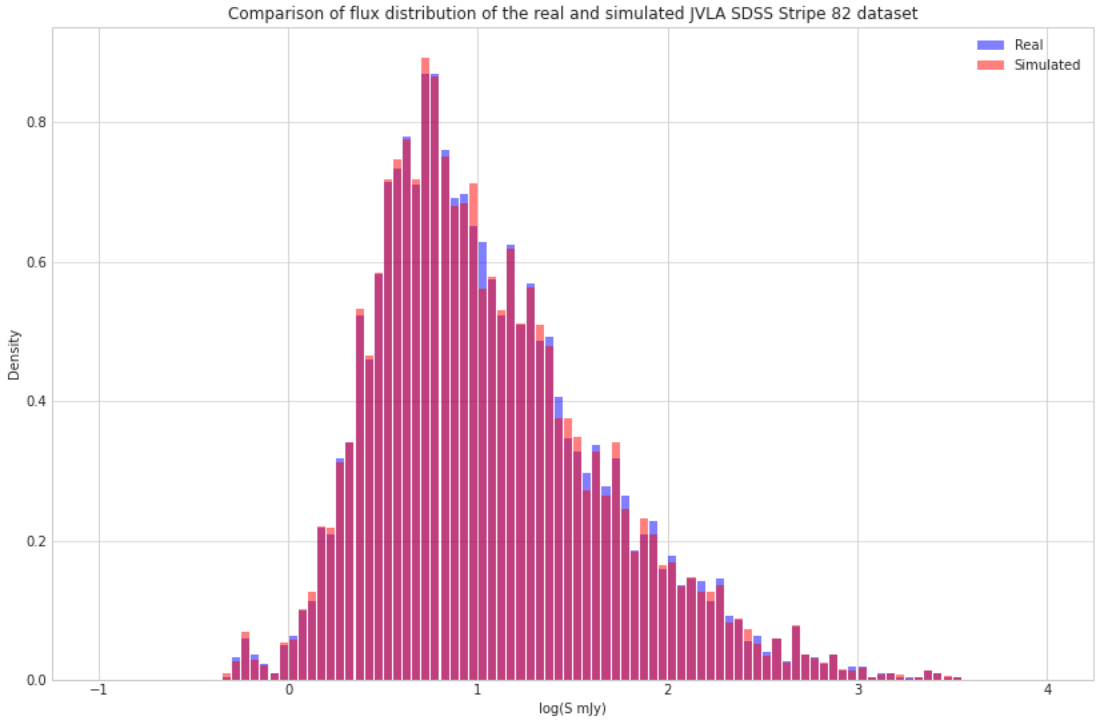


Figure 7.3 A Distribution of the integrated flux of 500 000 sources from the simulated sample for JVLA SDSS Stripe 82 region

sample and a sample of the simulated data set respectively.

The nearest neighbours for each of the sources in the real sample were determined, as well as for sources in samples of the simulated data set, where again the sample size for each sample was equated to the real sample size. The separation distance, nn_d, and flux product, $S_1 \times S_2$, for each nearest neighbour pair was determined. The mean and median nn_d for the real sample were found to be 173.30 and 160.17 arcsec respectively, while for the simulated sample the mean and median nn_d were 195.59 and 182.87 arcsec. Similar to the GMRT EN1W data, we expect that the real sample will tend to have smaller separation distances between sources due to the known multi-component systems that exist in the real sample, and not in the simulated sample. Figure 7.6 shows the distribution of nn_d for the real

	Real	Simulated (full)	Simulated (sample)
mean (mJy)	41.88	41.53	40.51
median (mJy)	8.98	8.89	8.56
σ *(mJy)*	153.49	153.03	147.83
min (mJy)	0.45	0.45	0.45
max (mJy)	3492.66	3461.00	2933.55

Table 7.1 Quantitative comparison of the integrated flux of the real and simulated JVLA SDSS Stripe 82 data sets

sample (blue) and simulated sample (red). From Figure 7.6 it can be seen that a considerable number of nearest neighbour pairs in the real sample show a small separation distance in comparison to the simulated sample. Note that the same normalization techniques that were applied to the sample distributions when the algorithm was applied to the GMRT EN1W radio image were applied to the JVLA SDSS Stripe 82 data.

Figure 7.7 show the distribution of the $S_1 \times S_2$ for the real sample (blue) and simulated sample (red). The distributions seem to match closely, however, it can be seen that there exists a higher occurrence of high flux product values in the real sample when compared to the simulated sample. The mean and median $S_1 \times S_2$ of the real sample were found to be 2729.5 ± 52.06 and 104.96 ± 11.42 mJy^2 respectively, while for the simulated data set these values were found to be 1666.72 and 104.83 mJy^2.

The two-dimensional distributions of nn_d and $S_1 \times S_2$ for the real and simulated samples were constructed for comparison. Figure 7.8 shows the two-dimensional distribution for the real sample. From Figure 7.8 the characteristic 'bulges' can be seen at high flux product values corresponding to low ($\sim$ 25 arcsec) and medium ($\sim$ 180 arsec) separation distances.

Figure 7.9 shows the two-dimensional distribution for the simulated sample. The distribution is significantly more smooth than the real sample distribution and lacks the 'bulges'

that are visible in the real sample distribution.

Figure 7.10 shows an overlay of the two-dimensional distribution of the real sample (blue) and the simulated sample (red). Areas of the real sample can be seen to extend beyond the distribution of the simulated sample.

Finally the score value per bin was determined by applying Function 5.2 to the real and simulated sample distributions. Samples of the simulated data set, where the sample size was equal to the size of the real sample, were repeatedly applied to the real sample using Function 5.2. This process was repeated several thousand times and the average score and standard deviation of the score as a result of Function 5.2 was recorded for each bin. Figure 7.11 shows the score value by bin of the two-dimensional distribution of $S_1 \times S_2$ and nn_d. Lower scores can be seen toward the centroid of the distribution, while scores tending towards a value of 1 can be seen at the outer edges of the distribution. A significant region of high scores can be seen at low nearest neighbour separation distances, along the left side of the distribution.

Figure 7.12 shows the standard deviation of the score value per bin of the two-dimensional distribution. A high standard deviation can be seen near the centroid of the distribution. A band of low standard deviation values, similar to that of the high score values in Figure 7.11, can be seen along the left side of the distribution, and to a lesser degree around the outer regions of distribution.

An output catalogue was produced containing information about the primary source, its nearest neighbour, the flux and position values of each source of the pair, the separation distance, flux product, score and score standard deviation for the Stripe 82 radio image. Figure 7.13 shows the two-dimensional distribution of the real sample with a box to indicate the region of interest. Source pairs in this region are likely to be bright radio galaxies as this region is characterised by high flux product values. This output catalogue was filtered using the following parameters and values: $20 < nn_d < 100$; $log(S_1 \times S_2) > 3$; $score > 0.9$, $score\ \sigma < 0.15$. A total of 91 unique nearest neighbour pairs met these criteria. Figures

7.14, 7.15 and 7.16 show the cutout images of several of these nearest neighbour pairs. The background image of the cutouts is the data from the JVLA SDSS Stripe 82 radio image with contours of the flux values indicated in white. The position of the detected sources is indicated by the red stars. The synthesized beam size is shown in green in the bottom left corner of each image.

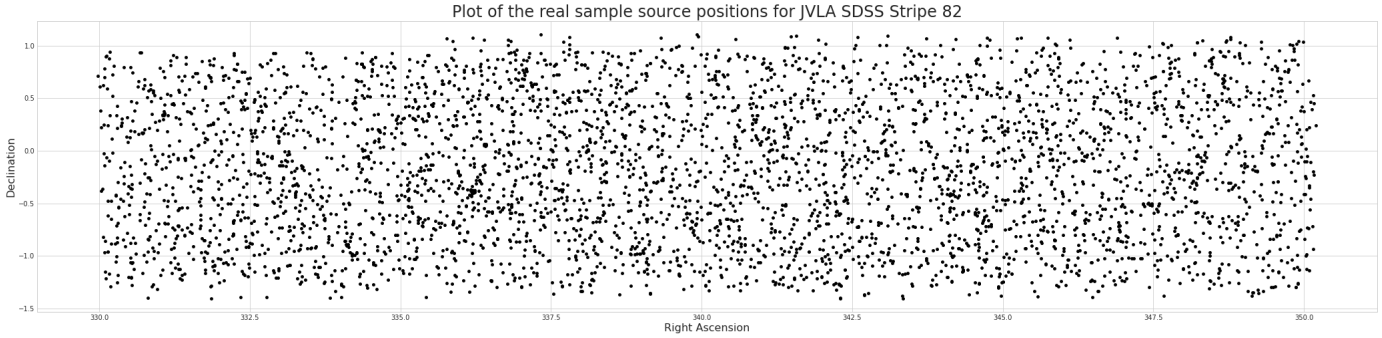


Figure 7.4 A plot of sources positions of sources from the real sample for JVLA SDSS Stripe 82

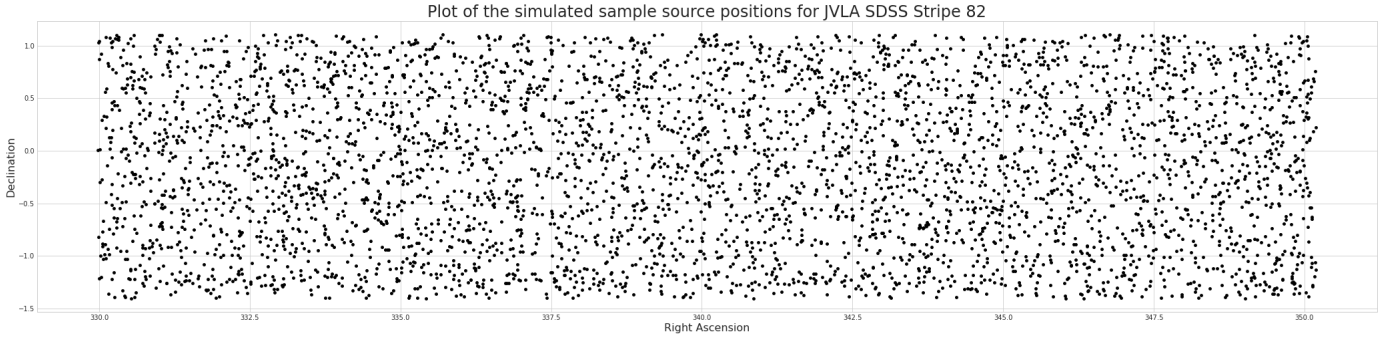


Figure 7.5 A plot of sources positions of sources from a sample pf the simulated data set for JVLA SDSS Stripe 82

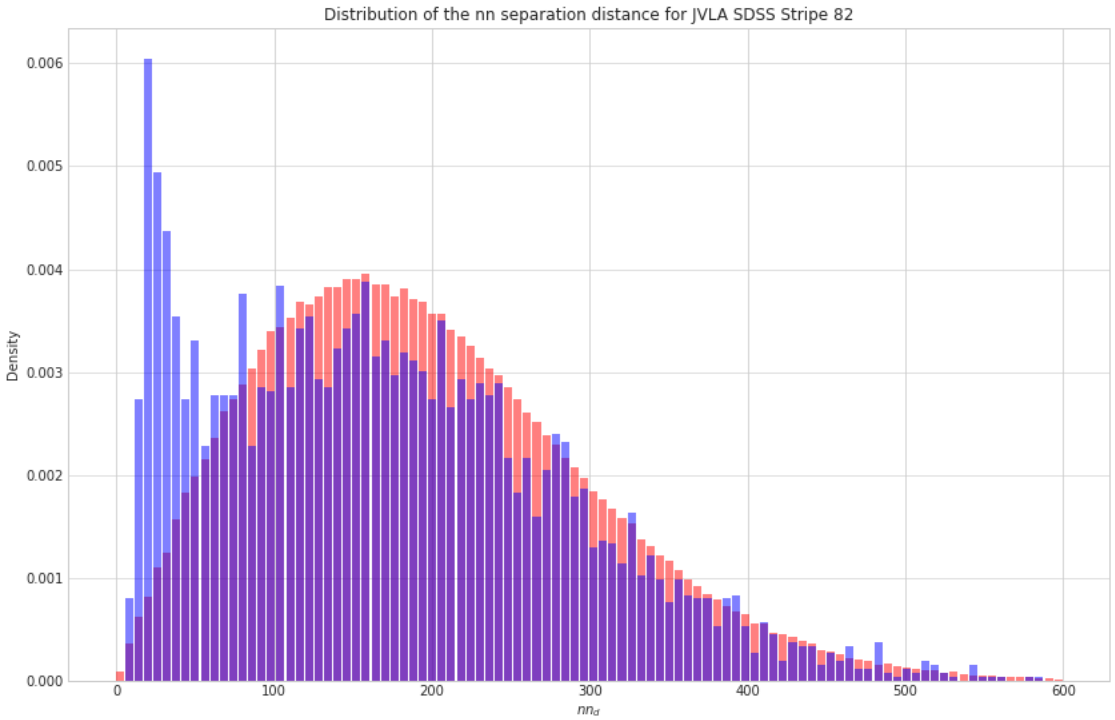


Figure 7.6 Distribution of nn_d for the real sample (blue) and the simulated sample (red) for JVLA SDSS Stripe 82

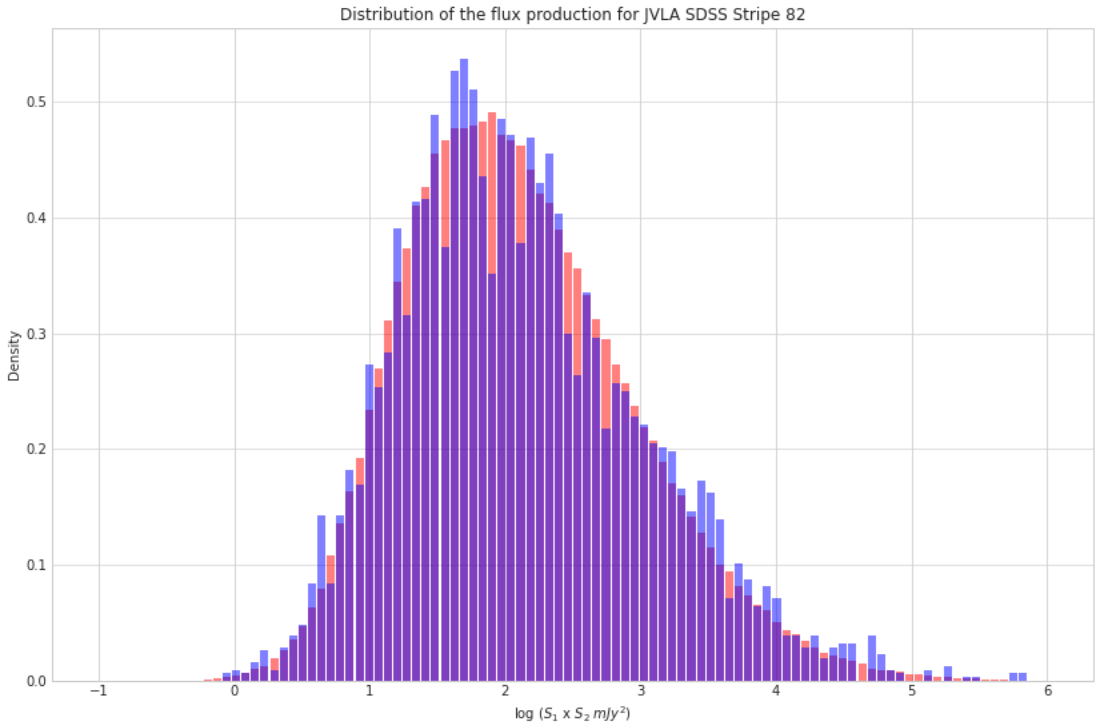


Figure 7.7 Distribution of $S_1 \times S_2$ for the real sample (blue) and the simulated sample (red) for JVLA SDSS Stripe 82

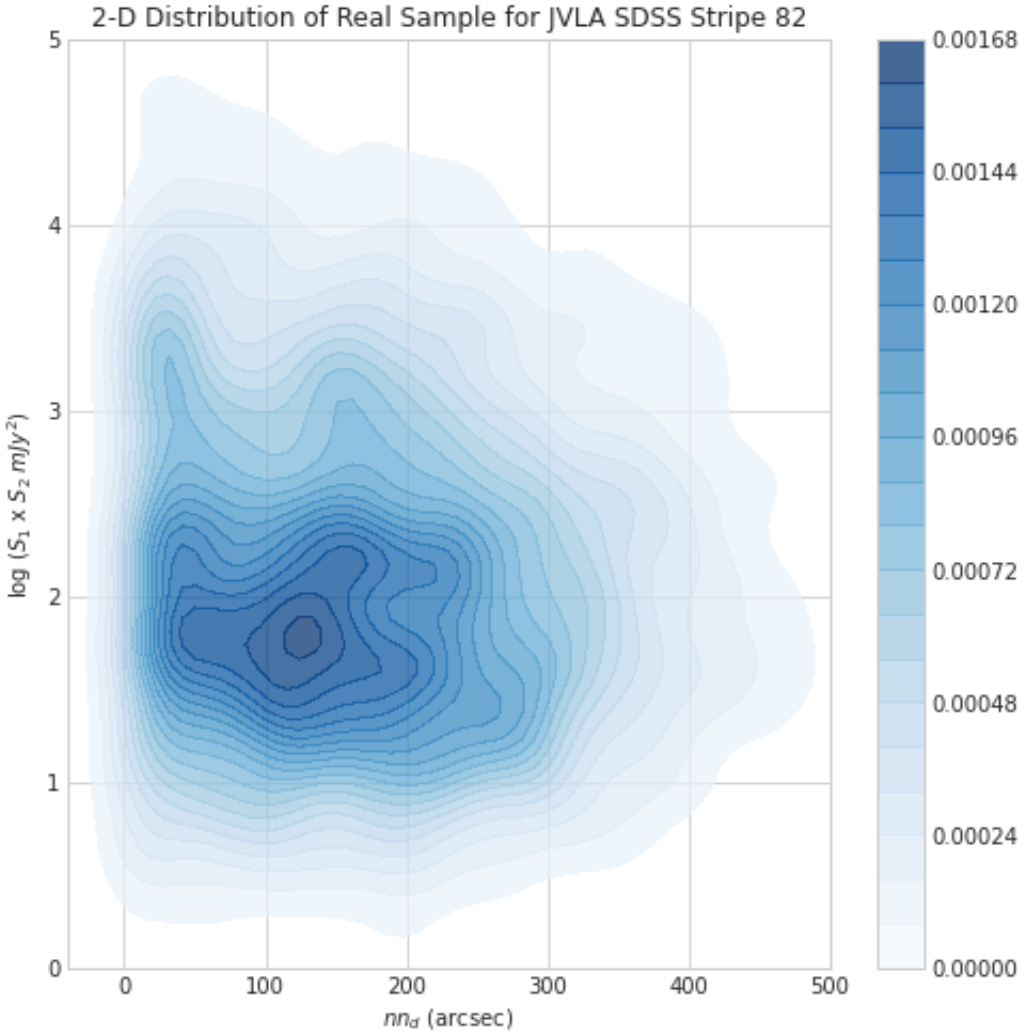


Figure 7.8 2-D distribution of nn_d by $S_1 \times S_2$ for the real sample for JVLA SDSS Stripe 82

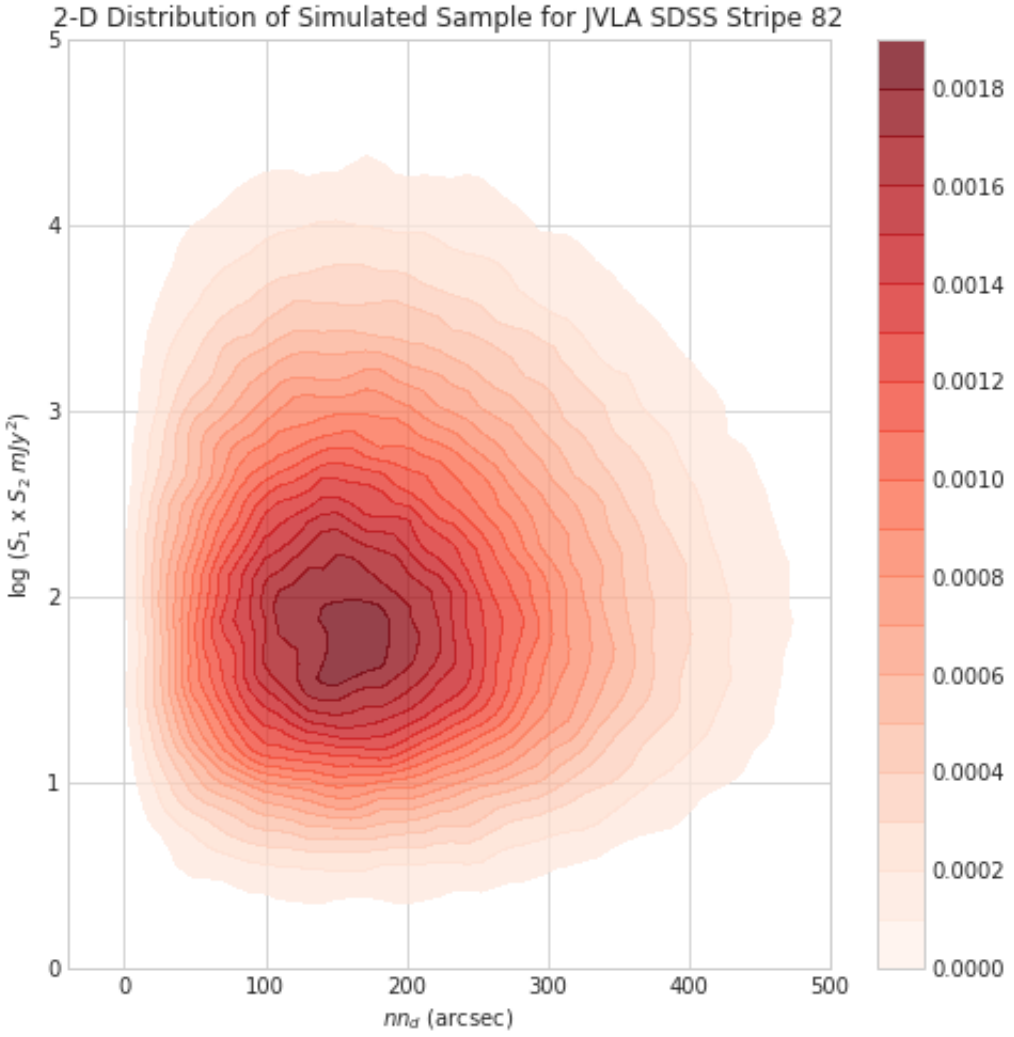


Figure 7.9 2-D distribution of nn_d by $S_1 \times S_2$ for the simulated sample for JVLA SDSS Stripe 82

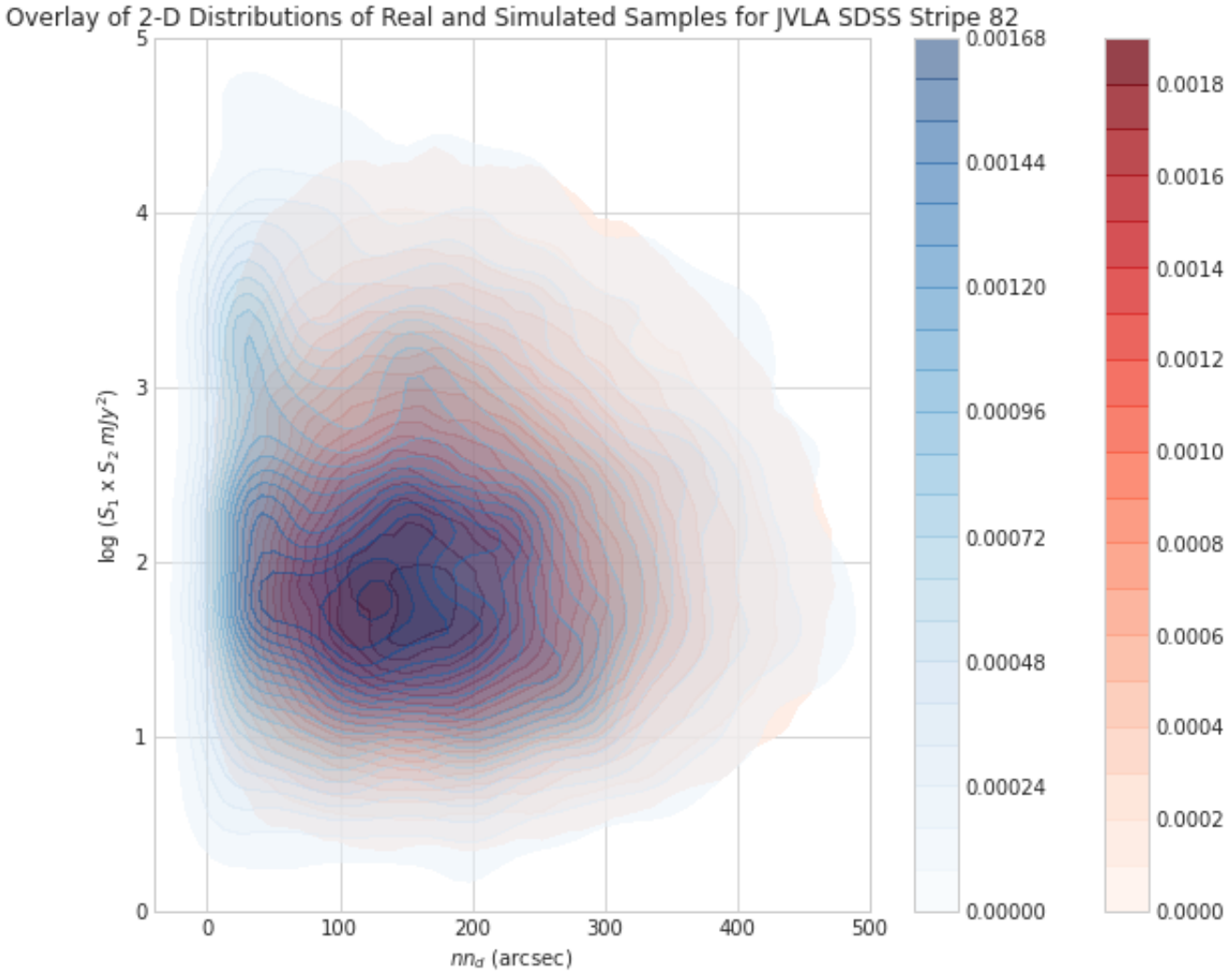


Figure 7.10 2-D distribution overlay of $S_1 \times S_2$ by nn_d for the real (blue) and simulated samples (red) for JVLA SDSS Stripe 82

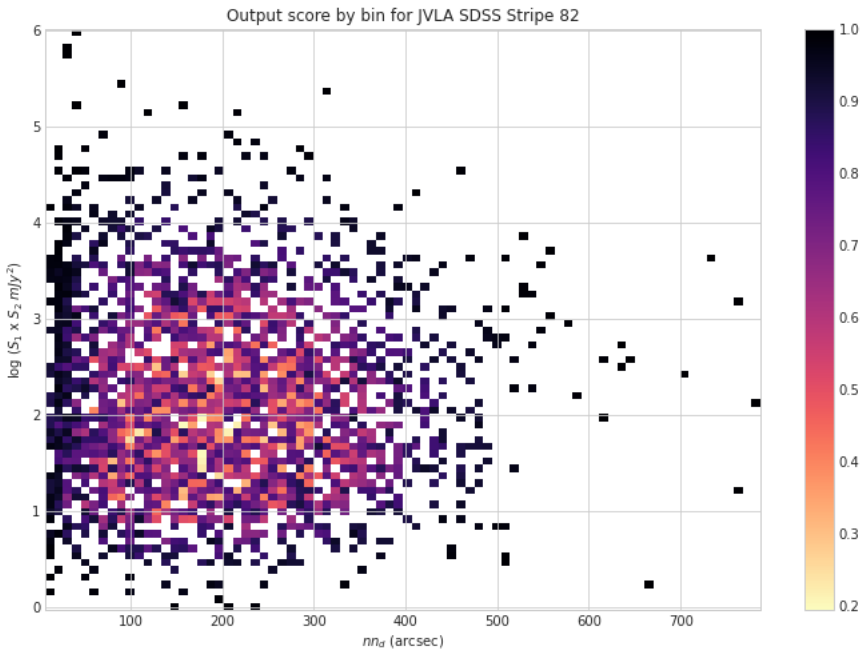


Figure 7.11 2-D distribution of the output score by nn_d and $S_1 \times S_2$ for JVLA SDSS Stripe 82

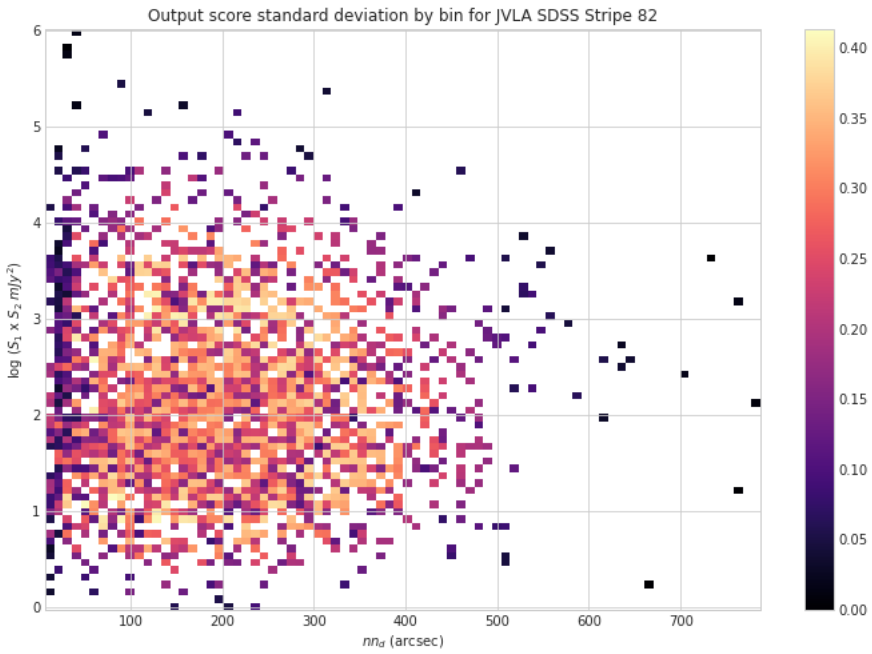


Figure 7.12 2-D distribution of the standard deviation of the output score by bin for JVLA SDSS Stripe 82

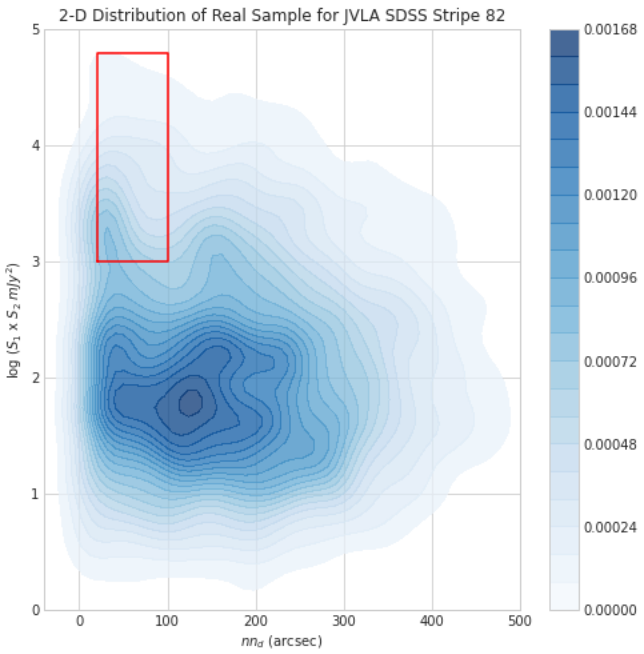


Figure 7.13 2-D distribution of nn_d and $S_1 \times S_2$ for the real sample for JVLA SDSS Stripe 82 , with box indicating region of interest

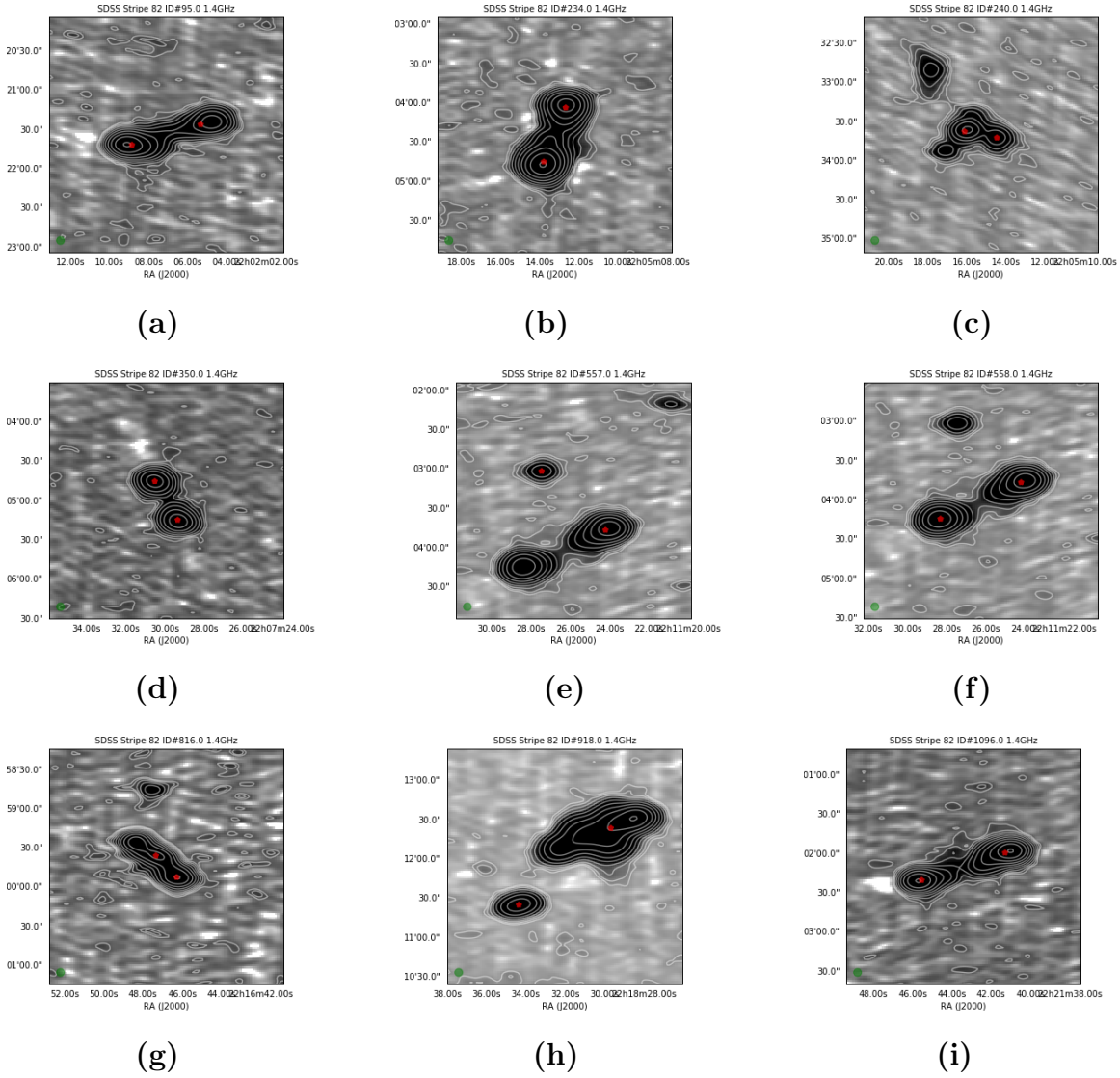


Figure 7.14 Cutout images of the nearest neighbour pairs meeting the following criteria: $20 < nn_d < 100$; $log(S_1 \times S_2) > 3$; *probability score* > 0.9, *score* $\sigma < 0.15$. The background image is from the GMRT EN1W radio image data, with the contours of flux values indicated in white. The position of the source pairs indicated by the red stars.

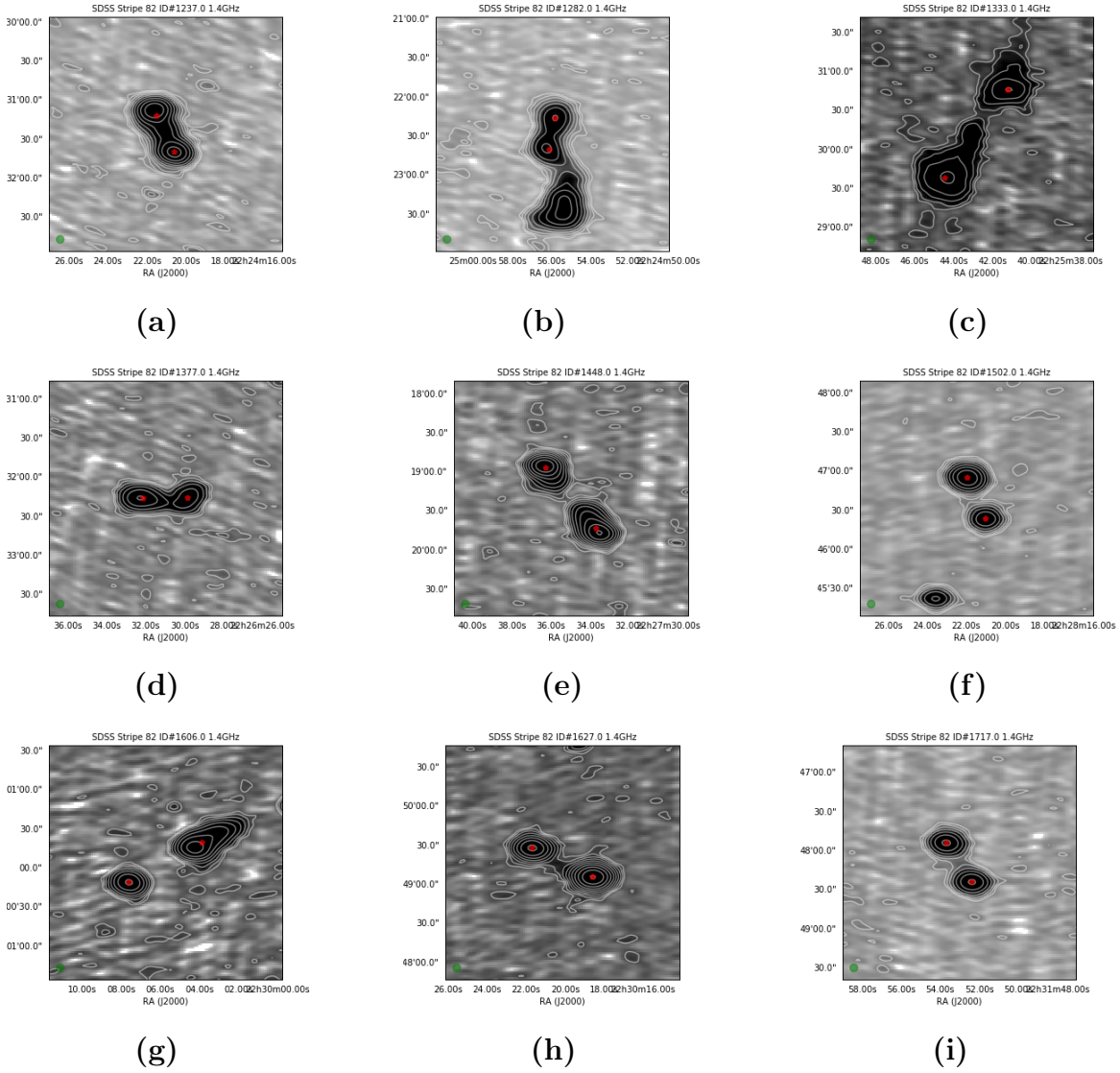


Figure 7.15 Cutout images of the nearest neighbour pairs meeting the following criteria: $20 < nn_d < 100$; $log(S_1 \times S_2) > 3$; *probability score* > 0.9, *score* $\sigma < 0.15$. The background image is from the GMRT EN1W radio image data, with the contours of flux values indicated in white. The position of the sources pairs indicated by the red stars.

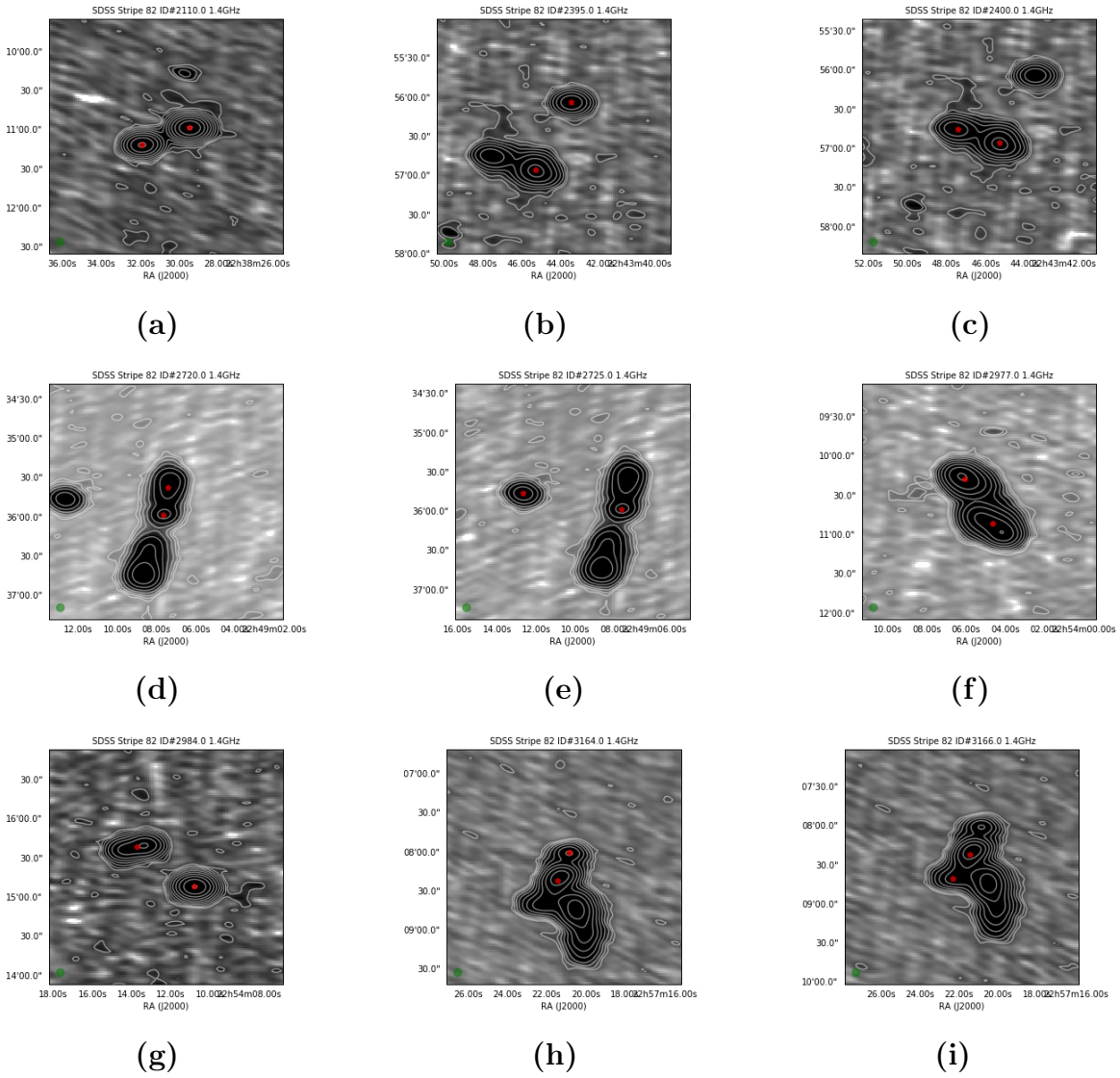


Figure 7.16 Cutout images of the nearest neighbour pairs meeting the following criteria: $20 < nn_d < 100$; $log(S_1 \times S_2) > 3$; $probabilityscore > 0.9$, $score\ \sigma < 0.15$. The background image is from the JVLA SDSS Stripe 82 radio image data, with the contours of flux values indicated in white. The position of the sources pairs indicated by the red stars.

Chapter Eight

Application of source matching with multi-wavelength data

The method described in Section 5.3 was applied to the results of the algorithm from the GMRT EN1W radio image. This method matches the source pairs detected in the radio image with potential matches in an IR image of the same region, and attempts to find a central source between the two radio sources, applying an empirically generated statistical probability of finding such a source as a function of area. A catalogue of detected IR sources from the Spitzer Wide-area InfraRed Extragalactic (SWIRE) survey of the EN1 field was used for cross-matching the radio sources. When cross-matching the sources the following is expected: if the radio source pairs are jets or lobes, neither radio source will have an IR counterpart; if the radio source pairs are a combination of a jet or lobe and the core galaxy or central black hole, there will be an IR counterpart for the core, but not for the jet or lobe; if IR counterparts exist for both sources in the radio source pair then the pair does not represent components of a radio galaxy.

The nearest neighbours catalogue for the GMRT EN1W was the input catalogue for this multi-wavelength cross-matching method. The input catalogue was prepared by removing any duplicate nearest neighbour pairs from the catalogue, that is to say, for any nearest neighbour pair where the primary source occurred as the nearest neighbour for its nearest

Group	No. of nn pairs	Ave. probability score
A	2728	0.517
Ba	58	0.775
C	452	0.608

Table 8.1 Number of source pairs found per group in the GMRT EN1W real sample

neighbour when considered as the primary source, the second occurance of the pair was removed from the input data. A total of 1714 duplicate pairs were removed from the input catalogue, leaving 3833 nearest neighbour pairs. The remaining unique nearest neighbour pairs were cross-matched with the sources from the SWIRE catalogue using the k-d nearest neighbour algorithm using the position values (right ascension and declination) of the radio sources and the IR sources. A separation threshold matching the major axis of the synthesised beam, in this case of 5.3 arcsec, was used to determine positive matches between the radio sources and the IR sources. Table 8.1 shows a break down of the number of sources that were found per group, along with the average probability score associated with the sources pairs that were assigned to each group.

Group A includes radio sources where both the primary source and it's nearest neighbour have positive matches from the IR catalogue. Group Ba includes radio sources where neither the primary source nor the nearest neighbour source have matches in the IR catalogue, however, a maximum separation threshold was applied to the radio and IR sources, indicating that the radio sources do fall within the SWIRE survey area. Group Bb includes radio sources with the same criterion as group Ba, however, the radio sources fell outside of the SWIRE survey region for EN1 and therefore the method of searching for a source between the primary source and its nearest neighbour would not be applicable. Group C includes radio sources where only one, the primary source or the nearest neighbour source, has a positive IR source match.

Sources in groups A and Bb are considered to be not applicable to this method and are discarded. Group A, with IR counterparts on each component, are distinct unrelated sources. Groups Ba and C exhibit behaviour that is expected from radio galaxies, two lobes would have no IR counterpart, while a core and a jet would have an IR counterpart on one. Sources in groups Ba and C are considered further. The SWIRE catalogue was searched for sources that were positioned between the nearest neighbour pairs, in a rectangular-shaped area positioned along a line drawn between the two sources, with a width equal to the synthesised beam major axis. If one or more sources were detected in this box, the area of the box was compared to a pre-generated empirical probability distributions that describe the probability of finding one or more sources as a function of area in the EN1 region.

The overall average probability score of sources pairs identified in the GMRT EN1W radio image is 0.545. Group A demonstrated a average probability score that fell below the sample average, while groups Ba, and C all showed average probability scores that were above the sample average, in particular group Ba which demonstrated an average probability score of 0.775. The average score of group Ba reinforces the prediction that sources pairs in this group are likely to be radio galaxies and should be investigated.

Of the 58 source pairs in group Ba, 22 were found to have one IR source positioned between the radio source pairs, while 29 source pairs were found to have more that one IR sources between the radio source positions. Similarly, for group C, of the 452 radio source pairs, 145 pairs were found to have one IR source positioned in the region between the radio source pairs, while 286 pairs were found to have more than one IR sources.

Figures 8.1 and 8.2 show examples of the radio source pairs in group Ba where one IR source is detected between the radio sources of the pairs. The background images in the figures are from the SWIRE 3.6μ IRAC observation. The flux values of the GMRT EN1W radio image can be seen as the green contours. The radio source pairs are indicated by the red stars. The region that was searched for IR sources is indicated by the white box. The probability of finding one or more source within this region is indicated by the

probability value in the figure title, $P[X > 1]$. As can be seen from the probability value, this value is a function of the area between the sources. As the width of the box is kept constant, this probability is therefore a function of the distance between the radio sources identified as nearest neighbour pairs, previously referred to as nn_d. A larger region indicates a greater probability of finding one or more IR source between the radio sources. While this dependency draws parallels to the parameters used in the statistical method, the distances between the sources may not be sufficient to infer a likelihood that the source pairs are radio galaxies, that is to say, based on a low probability of finding one or more IR source in the search region. From the Figure 8.1, while not confirmed as radio galaxies, one can see that the potential radio galaxies reveal a range of nn_d values.

Figure 8.2 shows further cutouts of samples from the radio pairs of group Ba. Figure 8.2(a) demonstrates where this method is less effective as an IR sources is found between the position of the two radio sources, while it is likey that these sources do not represent a radio galaxy. However, the probability of finding an IR source in the region between the radio sources is quite high, $P[X > 1]$ equal to 0.87. A high probability score here is less likely to be associated with a radio galaxy. Figures 8.2(a) 8.2(b) and 8.2(c) demonstrate radio pairs that have very small separation distances, but where an IR source is located in the region between the radio source pairs.

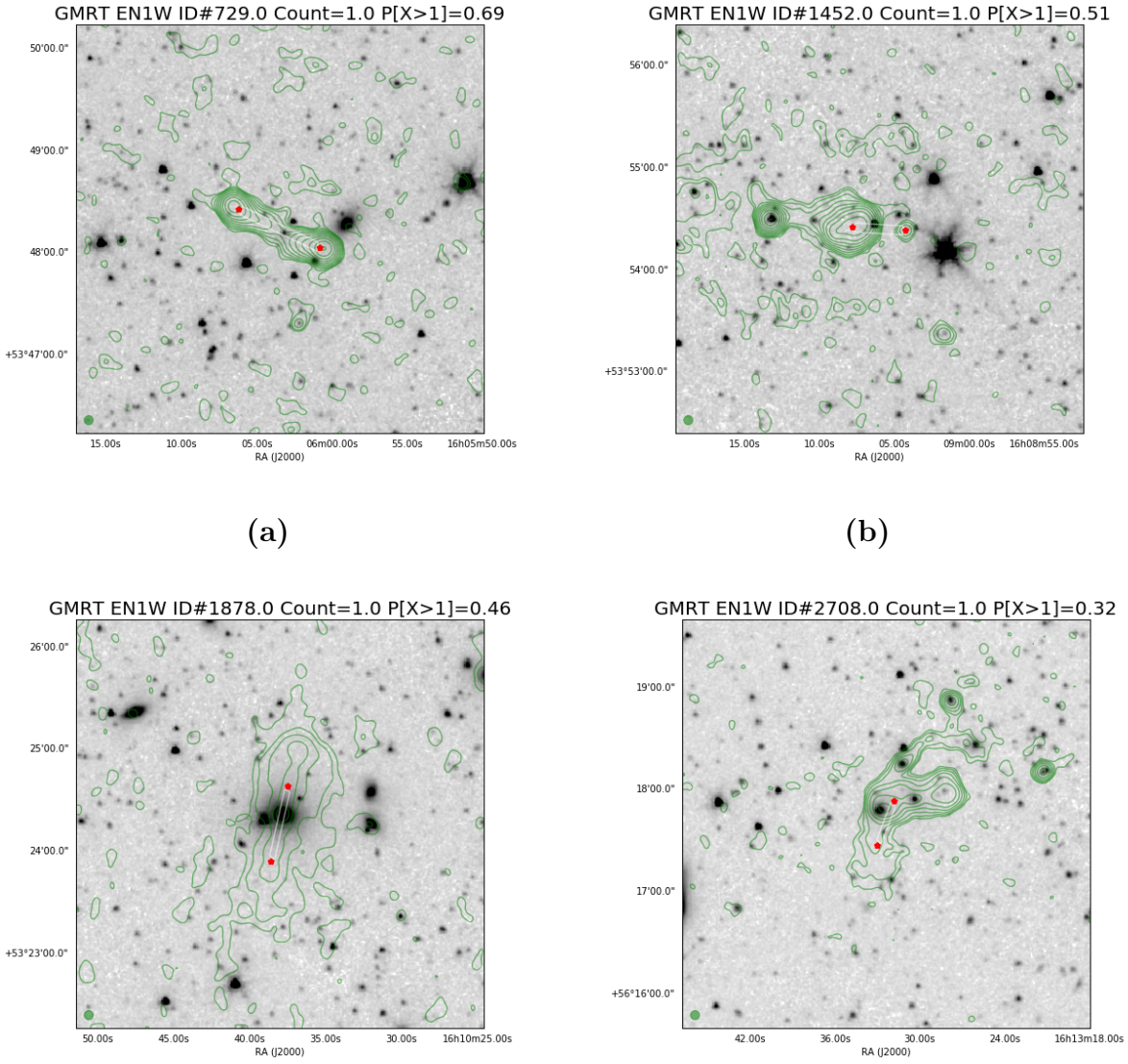


Figure 8.1 A sample of the cutout images of the nearest neighbour pairs from group *Ba*. The background image is from the 3.6μ IRAC image from the SWIRE catalogue. The contours of flux values of the GMRT EN1W radio image are indicated in green. The position of the source pairs indicated by the red stars. The search box for IR sources indicated by the white box.

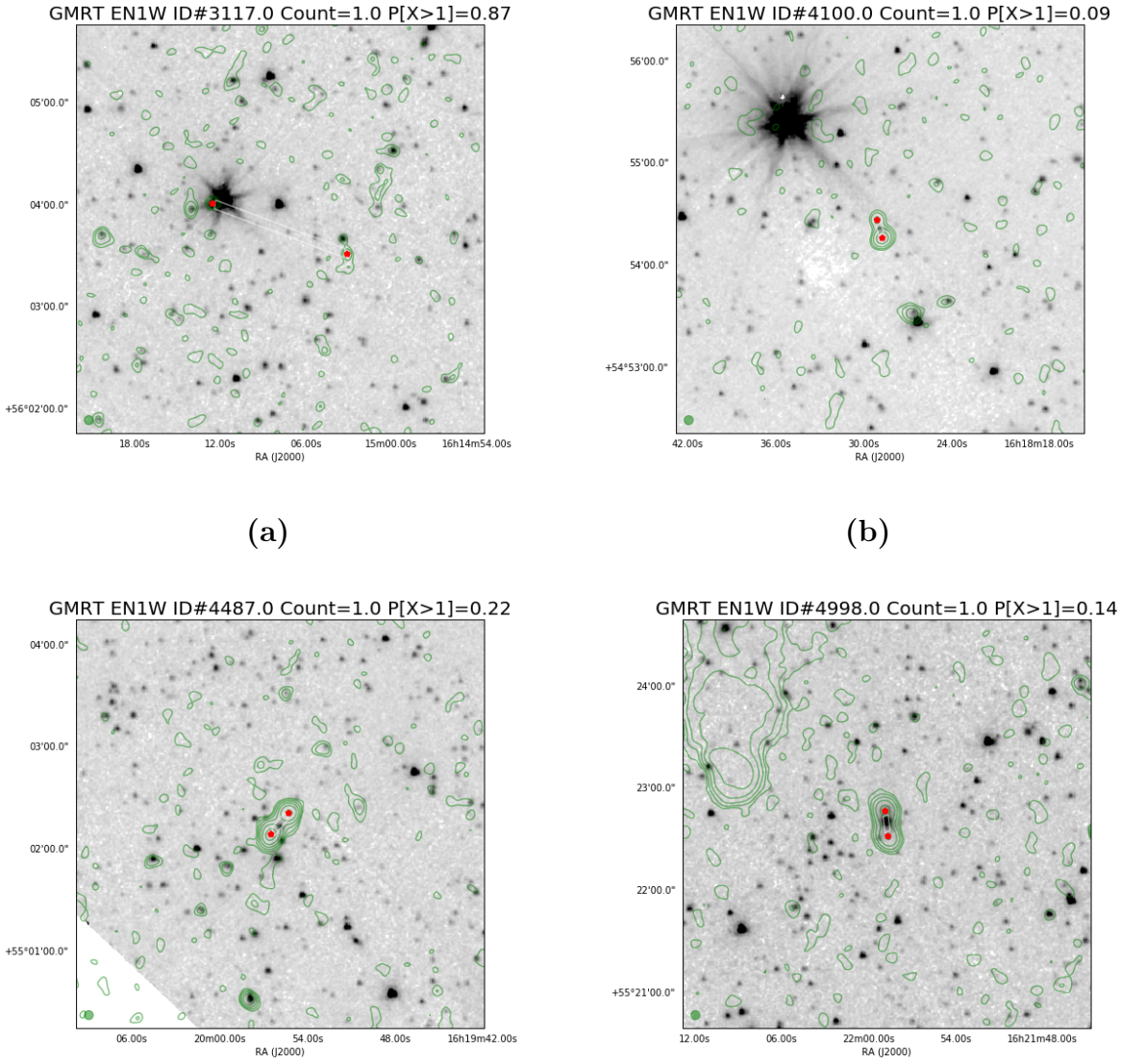


Figure 8.2 Additional samples of the cutout images of the nearest neighbour pairs from group *Ba*. The background image is from the 3.6μ IRAC image from the SWIRE catalogue. The contours of flux values of the GMRT EN1W radio image are indicated in green. The position of the source pairs indicated by the red stars. The search box for IR sources indicated by the white box.

Chapter Nine

DISCUSSION

From the results it can be seen that the distributions of the flux product and nearest neighbour distance for both the GMRT and Stripe 82 sample data that were produced were similar in shape, that is to say, both distributions showed a divergence from the respectively simulated samples at low nearest neighbour distances and high flux product values. However, from Figure 6.8, a second divergence in the GMRT distribution can be seen at high flux product values and at around 75 arcsec nn_d, while in Figure 7.8, the divergence in the JVLA SDSS Stripe 82 data is centred at roughly 175 arcsec nn_d.

It should be noted that integrated flux values from the GMRT sample data range between 0.24 mJy and 1701.97 mJy with a mean value of 8.25 mJy and a 75 percentile at 2.77 mJy, while the JVLA SDSS Stripe 82 sample integrated flux values ranged from 0.44 mJy and 3492.66 mJy with a mean value of 41.88 mJy and a 75 percentile value of 24.91 mJy. This shows that the GMRT EN1W radio image has a higher number of faint sources compared to the Stripe 82 radio image. From Figure 6.1 and Figure 7.1 it can be seen that there is a higher preponderance of high flux sources in the JVLA SDSS Stripe 82 sample than in the GMRT EN1W sample. The differences in flux distributions between the samples can be attributed to the differences in sensitivity of the observations and the frequencies at which the observations were conducted. Since the GMRT image is deeper than the JVLA image, the GMRT catalogue will contain fainter objects. Furthermore, the mean nn_d value

for the GMRT EN1W sample was found to be 75.03 arcsec, while for the JVLA SDSS Stripe 82 sample data the mean nn_d value was found to be 173.30 arcsec. The JVLA SDSS Stripe 82 observation covered a significantly large square degree area than the GMRT EN1W observation. In addition, the total source count of the JVLA SDSS stripe 82 radio image was found to be lower than that of the GMRT EN1W radio image. This would suggest that the number density in the JVLA SDSS Stripe 82 image is lower. The sparse distribution of sources in the JVLA SDSS Stripe 82 image would account for the higher separation distances found. While these differences do show divergences in the distributions of the flux product and nearest neighbour distances between the two observation samples, they do highlight the fact that the algorithm for identifying multi-component sources is applicable to observations spanning a wide range of depths and frequencies.

In Figures 6.10 and 7.10 it can be seen that the centre of the two two-dimensional distributions that are overlaid are not positioned at the same location in the respective figures. The respective distributions are normalized individually in these figures, however, this indicates that the data should be standardized. It was found that application of the standardization method, where the data is converted to a z-score caused the distribution of the real sample to narrow significantly more than that of simulated sample, as the real sample has a bimodal or multimodal distribution. To improve the accuracy of the probability score that results from the algorithm the random component of the distribution of the real sample data should be standardized to the simulated distribution. The same issue applies to implementing the algorithm on the JVLA SDSS Stripe 82 radio image.

The algorithm currently is implemented to search for nearest neighbour pairs, but many multi-component sources may have more than two components. The cutout images, from the GMRT EN1W data, in Figure 6.14 to 6.16 show possible multi-component sources that have a high probability of being radio galaxies. From Figure 6.14a and Figure 6.14b, and Figure 6.14c and Figure 6.14d, it can be seen that the primary source and its nearest neighbour are not always reciprocal, that is to say, the nearest neighbour, when viewed as the primary may

not associated the same source as its nearest neighbour. From the example cutouts mentioned, 'chaining' of nearest neighbours for different primary sources can be seen which may be identifying multiple components of the same physical phenomenon. The same 'chaining' can be seen in Figure 6.16e and Figure 6.16f, and again in Figure 6.16h and Figure 6.16i. A similar occurrence can be seen in Figure 7.16h and Figure 7.16i in the JVLA SDSS Stripe 82 cutouts. Combining these chains of sources that are identified as nearest neighbours for bright sources may be useful when expanding the statistical technique beyond two components. However, it should be noted that not all chained sources may be associated with the same physical phenomenon, such as in Figure 7.16d and Figure 7.16e, it is uncertain whether the source to the left of the cutout image is part of the multi-component source that can be seen positioned on the right.

While nearest neighbours with a high flux product value are more likely to be representative of radio galaxies, it is possible that a faint source positioned close to a bright primary source is selected as its nearest neighbour rather than a bright source that lies close to the primary source, however, further than the faint source. In observations with high dynamic range and radio images with a high occurrence of faint sources, such as the GMRT EN1W radio image, this scenario is more likely to occur, while in images with low dynamic range, bright sources are more likely to obscure faint sources. This could result in failed detection of a radio galaxy if some flux product threshold value is used as a selection criteria for output cutouts that are likely radio galaxy candidates as the flux product of such a scenario may result in the nearest neighbour pair falling below this threshold. This may be overcome to some degree with the chaining of sources mentioned above, however, such a feature would not be an absolute solution.

Selection criteria, using threshold values for the flux product, nearest neighbour distances and the score value, were used to select nearest neighbour pairs that would likely represent multi-component sources, such as radio galaxies. While the criteria were crudely selected, the purpose of the selection was to demonstrate the cutout images and how easily radio

galaxies and other multi-component sources could be separated out from point sources. While not all sources selected this way are radio galaxies, the cutouts produced may be used as inputs into classification techniques in order to recognise radio galaxies or other galaxy subtypes. Using this process the vast majority of point sources could be excluded from input into source classification tools, thereby significantly reducing processing time for the classification methods. Significant improvements to the selection criteria are likely required in order to avoid failed detections of multi-component sources.

The application of source matching with multi-wavelength data was used to search for sources in infrared observation data that were positioned between the nearest neighbour radio sources of the GMRT EN1W data sample. From Figure 8.1 it can be seen that this technique may be useful to identify potential radio galaxies, however, as the results from Figure 8.2 indicate, not all instances of a single IR source positioned between the two radio source pairs will be a positive match of a radio galaxy. One issue with this proposed method is perhaps the use of the probability of finding one or more source as a function of area. The area that was used had a set width, based on the major axis of the synthesised beam of the radio image. On account of this, the probability is actually a function of the nn_d value of the radio source pairs. This parameter may not be enough to infer some likelihood of identifying a radio galaxy. Alternatively, if the area was determined in a staggered fashion, for example, by looking at a small area in the central region between the two radio sources for a positive occurrence of an IR source, and then increasing this area in the event that no IR source is found, this changes the probability of finding one or more IR sources to a function that does not depend on the nn_d value.

A further issue with this multi-wavelength technique is that it assumes that the nearest neighbour sources that are detected during the statistical method are jets or lobes of the radio galaxy which is not necessarily the case. Depending on the morphology of the radio galaxy, the nearest neighbours detected may include a lobe and the central black hole, in which case the multi-wavelength matching technique would fail to detect an IR source between these

nearest neighbour sources. Alternative, if the nearest neighbour pairs determined by the statistical technique identify a radio galaxy lobe and another source that is not part of the multi-component system the multi-wavelength technique may detect an IR source leading to a false positive. These two scenarios therefore reduce the usefulness of this approach.

Chapter Ten

CONCLUSION

The distributions of the flux product value and separation distance of nearest neighbour sources in the GMRT EN1W and JVLA SDSS Stripe 82 radio image were compared to simulated data constructed from the respective radio images. A comparison between the real and simulated data using the above parameters was used to generate a score that would indicate a statistical divergence from the random distribution of source pairs in the simulated data. It was demonstrated that selection criteria, using the probability score values, flux product values and separation distance, could be used to filter the nearest neighbour samples for multi-component sources that have a high probability of representing radio galaxies. A sample of cutout images of those sources that met the criteria was produced. These cutout images could be used as input images into source classification techniques in order to automate the detection and classification of radio galaxies and other multi-component sources in radio images, assisting in the automated generation of source catalogues for radio images.

Future work that may expand and improve this research include:

1. Improve the standardization and normalization techniques that are applied to the data so that the the random component in the distribution of the real sample is more closely approximating the distribution of the simulated sample.

2. Expanding the statistical technique to detect multi-component sources with more than

two components. This would also further reduce input data to automated classification techniques as sources that have more than two components would form a single input rather than the current output of the statistical technique that sees multiple cutouts for a single physical phenomenon where the components are greater that two.

3. Improve the process of detecting multi-component sources. Currently, the nearest neighbour to the primary source is chosen as the candidate multiple-component. This may select a faint source when a bright source is in close proximity to the primary source and is more likely to be a component of a radio galaxy. A potential solution to this may be to implement a function where source flux values are weighted as a function of distance from the primary source using an exponential weight function. This way, a significantly bright source in close proximity to the primary source may be selected as the nearest neighbour candidate over a closer faint source. Alternative solutions or weighting functions should be investigated.

4. Accurately determine the centre of multi-component sources. This process may be important in order to centre the sources in cutout images for input into classification techniques.

5. Investigate the high occurrence of nearest neighbour pairs with low flux product values and low nearest neighbour distances.

6. Improve the steps when determining the region for searching for an IR source between the radio source pairs in the multi-wavelength technique so that the probability of finding one or more IR sources is not a function of the nearest neighbour distance.

REFERENCES

Aniyan, AK and Kshitij Thorat (2017). "Classifying radio galaxies with the convolutional neural network". In: *The Astrophysical Journal Supplement Series* 230.2, p. 20.

Barton, Russell R and Lee W Schruben (1993). "Uniform and bootstrap resampling of empirical distributions". In: *Proceedings of 1993 Winter Simulation Conference-(WSC'93)*. IEEE, pp. 503–508.

Bertin, E. and S. Arnouts (1996). "SExtractor: Software for source extraction." In: 117, pp. 393–404. DOI: 10.1051/aas:1996164.

Booth, RS and JL Jonas (2012). "An overview of the MeerKAT project". In: *African Skies* 16, p. 101.

Braun, Robert, Anna Bonaldi, et al. (2019). "Anticipated Performance of the Square Kilometre Array–Phase 1 (SKA1)". In: *arXiv preprint arXiv:1912.12699*.

Braun, Robert, Tyler L Bourke, et al. (2015). "Advancing astrophysics with the square kilometre array". In: *Advancing Astrophysics with the Square Kilometre Array*. Vol. 215. SISSA Medialab, p. 174.

Brederode, Leonardus R et al. (2016). "MeerKAT: a project status report". In: *Ground-based and Airborne Telescopes VI*. Vol. 9906. International Society for Optics and Photonics, p. 990625.

Burke, Bernard F, Francis Graham-Smith, and Peter N Wilkinson (2019). *An introduction to radio astronomy*. Cambridge University Press.

Carilli, C.L. and S. Rawlings (2004). "Science with the Square Kilometer Array: Motivation, Key Science Projects, Standards and Assumptions". In: *New Astronomy Reviews* 48.11-12, pp. 979–984. ISSN: 1387-6473. DOI: 10.1016/j.newar.2004.09.001. URL: http://dx.doi.org/10.1016/j.newar.2004.09.001.

Cislak, Aleksander and Szymon Grabowski (2014). "Experimental evaluation of selected tree structures for exact and approximate k-nearest neighbor classification". In: *2014 Federated Conference on Computer Science and Information Systems*. IEEE, pp. 93–100.

Condon, James J (1988). "Radio sources and cosmology". In: *Galactic and Extragalactic Radio Astronomy*. Springer, pp. 641–678.

Condon, JJ (1999). "Very large radio surveys of the sky". In: *Proceedings of the National Academy of Sciences* 96.9, pp. 4756–4758.

Dewdney, Peter et al. (2009). "The square kilometre array". In: *Proceedings of the Institute of Electrical and Electronics Engineers IEEE* 97.8, pp. 1482–1496.

Fanaroff, Bernard L and Julia M Riley (1974). "The morphology of extragalactic radio sources of high and low luminosity". In: *Monthly Notices of the Royal Astronomical Society* 167.1, 31P–36P.

Hales, C. A. et al. (2014). "ATLAS 1.4 GHz Data Release 2 - I. Observations of the CDF-S and ELAIS-S1 fields and methods for constructing differential number counts". In: 441.3, pp. 2555–2592. DOI: 10.1093/mnras/stu576. arXiv: 1403.5307 [astro-ph.GA].

Hancock, Paul J et al. (2012). "Compact continuum source finding for next generation radio surveys". In: *Monthly Notices of the Royal Astronomical Society* 422.2, pp. 1812–1824.

Hancock, Paul et al. (2019). *PaulHancock/Aegean: Aegean2.1 - full python3 support*. Version v2.1.0. DOI: 10.5281/zenodo.3474072. URL: https://doi.org/10.5281/zenodo.3474072.

Heywood, I et al. (2016). "A deep/wide 1–2 GHz snapshot survey of SDSS Stripe 82 using the Karl G. Jansky Very Large Array in a compact hybrid configuration". In: *Monthly Notices of the Royal Astronomical Society* 460.4, pp. 4433–4452.

Hopkins, Andrew M et al. (2015). "The ASKAP/EMU Source Finding Data Challenge". In: *Publications of the Astronomical Society of Australia* 32.

Ishwara-Chandra, C H et al. (2020). "A wide-area GMRT 610-MHz survey of ELAIS N1 field". eng. In: *Monthly notices of the Royal Astronomical Society* 497.4, pp. 5383–5394. ISSN: 0035-8711.

Jarvis, Matt J et al. (2016). "The MeerKAT international GHz tiered extragalactic exploration (MIGHTEE) survey". In: *Proceedings of Science*.

Koribalski, Bärbel S (2012). "Source Finding and Visualisation". In: *Publications of the Astronomical Society of Australia* 29.3, pp. 213–213.

Kozieł-Wierzbowska, Dorota and G Stasińska (2011). "FR II radio galaxies in the Sloan Digital Sky Survey: observational facts". In: *Monthly Notices of the Royal Astronomical Society* 415.2, pp. 1013–1026.

Laing, RA and AH Bridle (2002). "Relativistic models and the jet velocity field in the radio galaxy 3C 31". In: *Monthly Notices of the Royal Astronomical Society* 336.1, pp. 328–352.

Maneewongvatana, Songrit and David M Mount (2002). "Analysis of approximate nearest neighbor searching with clustered point sets". In: *Data Structures, Near Neighbor Searches, and Methodology* 59, pp. 105–123.

Mohan, Niruj and David Rafferty (2015). "Pybdsf: Python blob detection and source finder". In: *Astrophysics Source Code Library*.

Norris, Ray P et al. (2009). "ASKAP-EMU: Overcoming the challenges of wide deep continuum surveys". In: *arXiv preprint arXiv:0909.3666*.

Ocran, EF et al. (2020). "Deep GMRT 610 MHz observations of the ELAIS N1 field: catalogue and source counts". In: *Monthly Notices of the Royal Astronomical Society* 491.1, pp. 1127–1145.

Pilbratt, GL et al. (2010). "Herschel Space Observatory-An ESA facility for far-infrared and submillimetre astronomy". In: *Astronomy & Astrophysics* 518, p. L1.

Polsterer, Kai et al. (2016). "Parallelized rotation and flipping INvariant Kohonen maps (PINK) on GPUs". In: European Symposium on Artificial Neural Networks.

Rajani, Nazneen, Kate McArdle, and Inderjit S Dhillon (2015). "Parallel k nearest neighbor graph construction using tree-based data structures". In: *1st High Performance Graph Mining workshop, Sydney, 10 August 2015*.

Smolcic, Vernesa et al. (2015). "Exploring AGN Activity over Cosmic Time with the SKA". In: *Advancing Astrophysics with the Square Kilometre Array*. Vol. 215. SISSA Medialab, p. 069.

Thompson, Anthony Richard, James M Moran, George Warner Swenson, et al. (2017). *Interferometry and synthesis in radio astronomy*. Springer.

Urry, C Megan and Paolo Padovani (1995). "Unified schemes for radio-loud active galactic nuclei". In: *Publications of the Astronomical Society of the Pacific* 107.715, p. 803.

Vaccari, Mattia (2016). "HELP-ing Extragalactic Surveys: The Herschel Extragalactic Legacy Project and The Coming of Age of Multi-Wavelength Astrophysics". In: *arXiv preprint arXiv:1605.01215*.

Wu, Chen et al. (2018). "Radio Galaxy Zoo: CLARAN–a deep learning classifier for radio morphologies". In: *Monthly Notices of the Royal Astronomical Society* 482.1, pp. 1211–1230.

APPENDIX

Appendix A

target_id	neighbour_id	target_i_flux	target_ra	target_dec	neighbour_i_flux	neighbour_ra	neighbour_dec
1	2	12.883	329.980	0.712	3.566	330.007	0.612
2	8	3.566	330.007	0.612	10.039	330.070	0.624
3	10	19.941	330.023	0.385	367.374	330.077	0.249
4	6	10.720	330.030	-0.379	22.785	330.054	-0.361
5	10	46.929	330.037	0.224	367.374	330.077	0.249
6	4	22.785	330.054	-0.361	10.721	330.030	-0.379
7	16	11.577	330.062	-0.822	13.424	330.103	-0.856
8	2	10.039	330.070	0.624	3.566	330.007	0.612
9	19	63.793	330.071	-0.027	15.293	330.116	-0.064
10	5	367.374	330.077	0.249	46.929	330.037	0.224
11	14	12.124	330.087	0.781	3.701	330.091	0.698
12	24	99.859	330.089	-0.966	6.367	330.142	-0.961
13	19	2.144	330.090	-0.140	15.293	330.116	-0.064
14	8	3.700	330.091	0.698	10.039	330.070	0.624
15	18	236.784	330.094	-1.149	18.605	330.113	-1.211

Table A.1 Sample output catalogue from the algorithm applied to the SDSS Stripe 82 radio image 1/2

target_id	neighbour_id	flux_product	flux_product_log	separation_dist	probability_score	probability_score_std
1	2	45.939	1.662	370.998	0.871	0.156
2	8	35.799	1.554	233.334	0.484	0.346
3	10	7325.878	3.865	528.676	0.997	0.033
4	6	244.270	2.388	109.529	0.695	0.271
5	10	17240.605	4.237	167.651	0.951	0.133
6	4	244.270	2.388	109.529	0.695	0.271
7	16	155.409	2.191	194.226	0.564	0.313
8	2	35.799	1.554	233.334	0.484	0.346
9	19	975.588	2.989	208.932	0.624	0.318
10	5	17240.605	4.237	167.651	0.951	0.133
11	14	44.866	1.652	298.071	0.714	0.266
12	24	635.815	2.803	192.354	0.596	0.335
13	19	32.787	1.516	289.696	0.881	0.132
14	8	37.150	1.570	275.698	0.798	0.194
15	18	4405.339	3.644	232.531	0.864	0.212

Table A.2 Sample output catalogue from the algorithm applied to the SDSS Stripe 82 radio image 2/2