Quick, Look! A New Method for Artifact Detection

by

Suhasini Satish Rao

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

Master of Science

Department of Physics
University of Alberta

Abstract

In the era of all-sky surveys, wide-area radio surveys like the Very Large Array Sky Survey (VLASS) place strong emphasis on the identification and classification of candidate transient events. To aid the transient searches, automatically generated VLASS Quick-Look (QL) images are designed and produced to deal with high data volumes and enable rapid followup observations. But, the incomplete and snapshot sampling of the sky in the *uv*-plane during VLASS observations combined with imaging choices made when generating QL images often lead to residual linear artifacts, particularly around brighter sources. While well-established techniques (like `clean`) maximizes the information from this partial sampling, such techniques can still be imperfect. Therefore, the need for automatic image-quality classification and assurance is more important than ever to ensure rapid data quality assessment and for enabling the best science. In this thesis, I present a new technique to identify these linear streaks around sources detected in the VLASS Epoch 1 QL images by extending the results of a line detection technique called the Hough Transform. After robustly quantifying the identified streaks, their effects on the sources/components that they are overlapping are removed. The resulting artifacts-subtracted components' brightnesses are then used to distinguish real astrophysical sources from imaging artifacts. Finally, I discuss the use of this streak detection method as an additional quality assessment step during interferometric image reconstruction.

Preface

This thesis is an original work by Suhasini Satish Rao. All the data used for in this thesis is publicly available. The code to predict the angles of the sidelobe streaks and to scale the Hough Transform 1D background profile was developed by Dr. Gregory R. Sivakoff.

“Y’all, astronomy is kind of interesting.”

-Gina Linetti

Acknowledgements

I would like to thank my supervisor, Gregory Sivakoff for his continued guidance, motivation and support. I am forever grateful to him for putting up with my frequent visits to his office and answering all of my questions with at-most patience and interest. I would also like to thank Coleman Dean and Steven Fahlman for giving me the “juice” on my plots and figures. Finally, I would like to thank my family for being my constant pillars of support.

Table of Contents

List of Tables

List of Figures

List of Algorithms

Abbreviations

2MASS The Two Micron All Sky Survey.

ALMA Atacama Large Millimeter/submillimeter Array.

ATCA Australia Telescope Compact Array.

BDP Basic Data Droducts.

CASA Common Astronomy Software Applications.

CHT Circle Hough Transform.

CIRADA Canadian Initiative for Radio Astronomy Data Analysis.

Dec Declination.

ESO European Southern Observatory.

EVLA Expanded Karl G. Jansky Very Large Array.

FIRST Faint Images of the Radio Sky at Twenty-centimeters.

FWHM Full Width at Half Maximum.

GRB Gamma Ray Burst.

HT Hough Transform.

ISM Inter-Stellar Medium.

LOFAR Low Frequency Array.

LoTSS LOFAR Two-metre Sky Survey.

MAD Median Absolute Deviation.

MGPS Molonglo Galactic Plane Survey.

NRAO National Radio Astronomy Observatory.

OTFM On-The-Fly Mosaic.

PSF Point Spread Function.

PyBDSF Python Blob Detector and Source Finder.

QL VLASS Quick-Look.

QLIP Quick-Look Imaging Pipeline.

R.A Right Ascension.

RHT Rolling Hough Transform.

RMS Root Mean Square.

RMSD Root Mean Square Deviation.

SDSS The Sloan Digital Sky Survey.

SNR Signal-to-Noise Ratio.

SOM Self-Organizing Maps.

VLA (Karl G. Jansky) Very Large Array.

VLASS Very Large Array Sky Survey.

WISE Wide-field Infrared Survey Explorer.

Chapter 1

Introduction

1.1 Scientific Motivation

Over the course of history, people have seen and recorded hints about the dynamic nature of the Universe. For example, ancient astronomical records from China, Japan, Iraq etc., describe the sightings of the supernova of 1054 as a “guest star” (Brecher, 1978; Ho et al., 1972), the remnant of which is now famously known as the Crab Nebula. With the advancements in technology and the invention of telescopes, astronomers not only focused on recording rare astrophysical sightings, but also on studying the objects and processes producing such events. The 1054 “guest star” was later (independently) re-discovered as a nebula by John Bevis in 1731 (Ashworth, 1981) and was even catalogued by Charles Messier as the first Messier object or M1 (Messier, 1781). The physical nature and origin of the nebula was finally confirmed by Jan Oort in 1942 as a remnant of a supernova (Mayall & Oort, 1942). Eventually, using the 300 ft Green Bank radio antenna, David H. Staelin and Edward C. Reifenstein III reported the observations of the pulsar associated with the Crab Nebula (Howard et al., 1968), only a year after the first discovery of a pulsar by Jocelyn Bell in 1967. Since then, nebulae and pulsars have been a study of interest for astronomers not just out of curiosity but also to understand the origins of such extreme conditions that give rise to pulsations and energetic particles. Such discoveries of extreme objects raised many questions about the range of dynamic objects present in the Universe and the diverse physical conditions that lead to such scenarios.

Similar to the Crab nebula and pulsar, various other astronomical objects that show extreme variations in the morphological, photometric, or spectroscopic properties — termed as transients — have been studied by astronomers for several decades. Many of these transients trace the dynamic and violent Universe — such as supernovae, relativistic jets from supermassive black-holes, neutron star mergers, pulsars and fast radio bursts. Based on the mechanism producing these transients, the spectrum of these objects peak (and in some cases, only emit) at specific electromagnetic frequencies. To discover all types of transients and cover the entire transient phase space, the sky should be studied at all frequencies **and at all times!**

Compared to the entire electromagnetic spectrum (especially the optical part of the spectrum) the emphasis on studying radio transients (which motivates the work in this thesis) is because some objects like radio galaxies, radio pulsars, and fast radio bursts, are only visible at radio frequencies, while other objects like quasars stand out strongly at radio frequencies. Moreover, the discovery of other phenomena in more familiar objects like low frequency bursts on the surface of the Sun (Reber, 1944) and Jupiter (Franklin, 1959), were only possible due to radio observations since they are only emitted at radio frequencies. The range of radio transients mentioned above (and many more), their peak emission frequencies, and their luminosities are shown in Figure 1.1 (from Rowlinson et al., 2022). The transient studies at radio frequencies will thus provide a unique opportunity to study the nature of objects that are elusive under traditional optical observations.

In many transient detection cases, targeted followup observations are conducted after first detections of a source. Although this technique is well suited to study the specific (proposed) targets at higher resolution (and hence in more detail), it often does not allow for a population study of the different types of dynamic objects present in the radio sky. Additionally, even for a single type of transient, targeted observations help in understanding the object itself, but not how frequent the transient behaviours

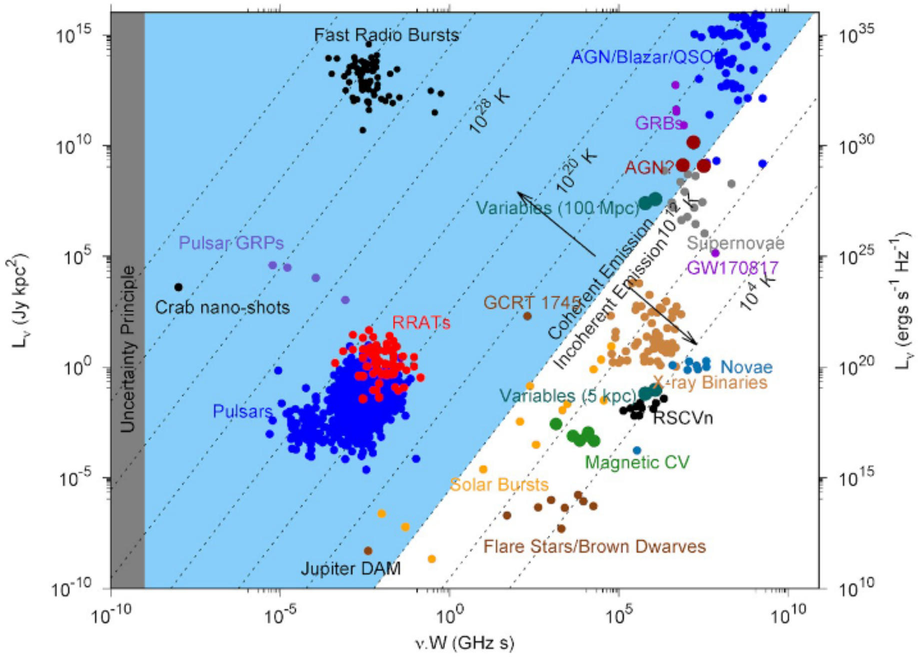


Figure 1.1: The range of variable radio sources with the luminosity plotted against frequency (adapted by Rowlinson et al. (2022) from an original plot in Pietka et al. (2015)). Here L_ν is the specific luminosity, ν is the frequency of radiation and W is the timescale of the emission.

are. Alternatively, wide (all) sky surveys allow for this required population analysis. Moreover, the detection of shorter timescale transients (that evolve in the order of days) — like GRBs and flares — can trigger deeper follow ups at other frequencies that can help characterize the transient sources.

Usually in the optical and infrared domains, panoramic sky surveys are conducted to analyze the types of objects populating each spectral regime. For example, the Sloan Digital Sky Survey (SDSS) is an optical/infrared imaging and spectroscopic redshift survey of the entire sky at high resolutions (1.43″ in r-band) and has currently catalogued the photometry of more than 10^9 astrophysical sources (as of data release 16; Ahumada et al., 2020). The Two Micron All Sky Survey (2MASS) is an all-sky infrared survey (with 4″ spatial resolution in all three bands) that produced a catalogue of more than 2×10^6 astrophysical sources (Skrutskie et al., 2006), with a greater ability to identify objects behind large extinction columns from Galactic dust.

The Gaia observatory, tasked with charting the three dimensional map of the Milky Way, has catalogued over 10^9 stars including their proper motions and distances (Gaia Collaboration et al., 2016).

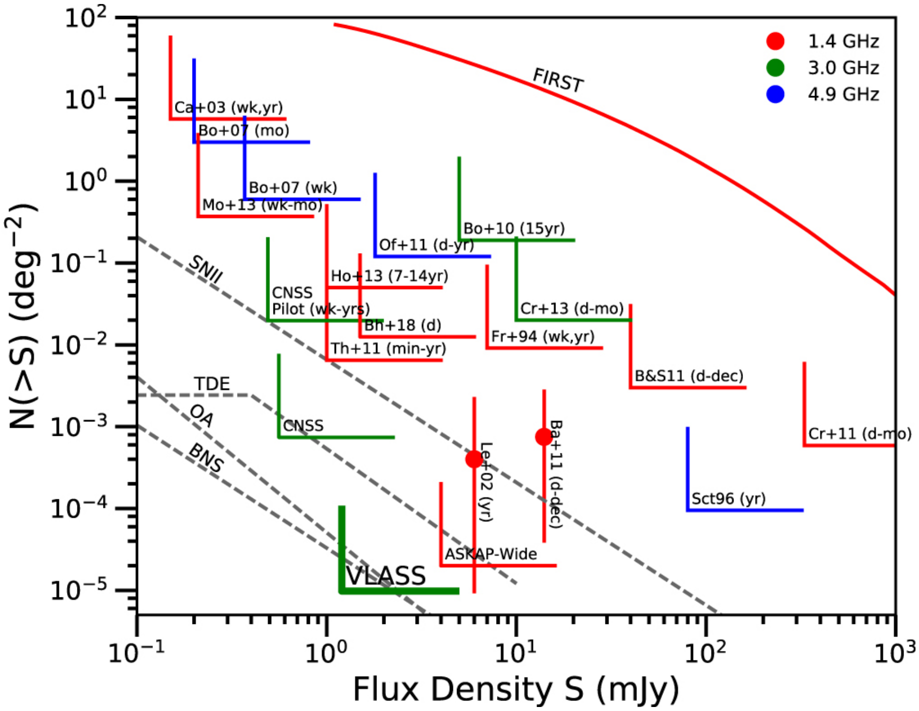


Figure 1.2: The sensitivities of various surveys (at different frequencies indicated by colour) and their constraints or measurements on the angular density of transient sources. The *dashed* lines indicate the expected projected angular density of various extra-galactic radio transients (from Lacy et al., 2020).

Similarly, with the advent of interferometry, sky surveys at radio frequencies with resolution comparable to optical observations/surveys are made possible. For example, the Faint Images of the Radio Sky at Twenty-centimeters (FIRST) was a radio survey of the northern sky conducted using the Very Large Array at $5''$ resolution and has catalogued almost a million radio sources. In the same way, the Very Large Array Sky Survey (VLASS) was proposed as an all-sky, relatively shallow, survey with transient searches as one of its key science goals. Compared to the FIRST survey, VLASS is more sensitive to binary neutron star mergers, Gamma Ray Bursts (GRBs) and Orphan GRB Afterglows, and Tidal Disruption Events, as shown in the areal density plot of extra-galactic radio transients in Figure 1.2 (from Lacy et al., 2020)). This makes VLASS an ideal survey for studying radio transients.

Here we provide a few examples of transient studies already made possible by VLASS.

- Law et al. (2018) discovered a decade-long extra-galactic transient, called FIRST J141918.9+394036, by comparison of archival data from the FIRST survey at 1.4 GHz with VLASS Epoch 1 data at 2–4 GHz. Follow up radio observations, optical spectroscopy, and the absence of a counterpart GRB suggests that the transient could be an Orphan GRB Afterglow, an off-axis late afterglow of a GRB seen at lower frequencies. The slow evolution of the radio afterglow in such an origin would confirm some of the predicted theories (*e.g.*, Sironi & Giannios, 2013) about the non-relativistic bulk motion of shocks in jets.

- A study by Nyland et al. (2020) found a sample of quasars that have transitioned from radio-quiet to radio-loud within a decade (with ∼2–5 times increase in the flux densities from L-band, 1.4 GHz, to S-band, 2–4 GHz) between the observations made by the FIRST survey and VLASS Epoch 1. Follow up radio observations that study their variability and spectral energy distribution shapes suggested that the emergence of young jets driven by accretion onto supermassive-black holes led to the highly increased radio loudness. The discovery of strongly changing quasar behaviour on human timescales suggests there may be more short-term AGN activity than previously believed. Such activity could have strong implications for how AGN feedback could drive galaxy evolution at high redshifts.

- Dong et al. (2021) discovered a transient, VT J121001+495647, with a peak radio luminosity of 1.5×10^{29} erg s^{-1} Hz^{-1} (unusually high for a non-nuclear transient) that is consistent with a merger-triggered core collapse supernova using VLASS Epoch 1 data. In this type of system, during the common envelope phase, the core of the companion star is tidally disrupted by the compact object, causing the star to undergo a premature merger-driven supernova. The discovery of VT J121001+495647 provided the first observational evidence of a merger triggered supernova.

1.2 The Very Large Array and its Sky Survey

The Karl G. Jansky Very Large Array (VLA) is a radio interferometric array located in New Mexico, United States of America and is operated by the National Radio Astronomy Observatory (NRAO). The VLA comprises 27 antennas of 25 m diameter each, positioned along 3 equilateral arms of 21 km with 9 antennas per arm. Each of the 27 antennas can be moved along the arms to different locations to arrange the telescope in different configurations. The 4 standard configurations: A, B, C and D (from the largest to the smallest) have maximum baseline lengths of 36, 11, 3.4 and 1 km respectively. The larger configurations provides higher resolution, while the smaller configurations provide better surface brightness sensitivity. Thus, along with changes in the frequency bands, the VLA provides a wide range in resolution and sensitivity. There are other intermediate hybrid configurations, namely BnA, CnB and DnC, where the northern arm is extended to the larger configuration while the lower eastern and western arms are kept in one of the more compact configurations. A depiction of the VLA in the BnA configuration is shown in Figure 1.3.

VLASS is a 3 epoch all-sky radio survey observed using the upgraded VLA known as the expanded Karl G. Jansky Very Large Array (EVLA; Lacy et al., 2020). The survey covers the entire sky with Declination (δ) above $-40°$ — a 33,885 deg^2 area corresponding to 82% of the entire sky. During VLASS scheduling, the antennas in VLA are in either the B or BnA configuration. When observing at lower declinations, the extended northern arm in the BnA configuration produces projected north-south baselines similar to the east-west baselines of the other two (southern) arms. Therefore, the BnA configuration is mostly used to map the southern regions to obtain a *uv*-space coverage similar to the B configuration observations in northern regions.

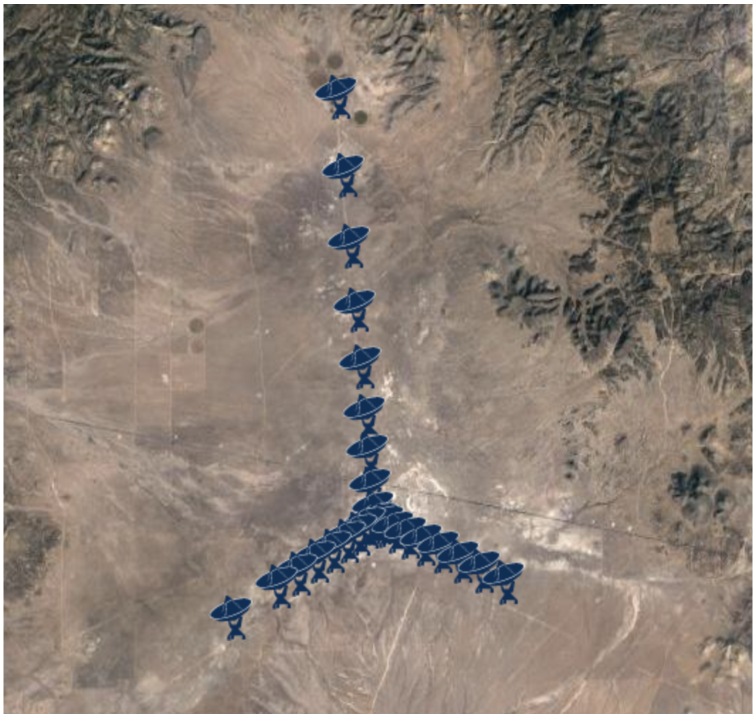


Figure 1.3: A pictorial representation of the VLA antennas in the BnA configuration taken from https://public.nrao.edu/vla-configurations/.

The survey maps the radio sky in the 2–4 GHz bandwidth (S-band) in 2 MHz channels, along with full calibrated polarization parameterization in Stokes I, Q and U. The survey has an imaging resolution of $2''.5$ with a 1σ sensitivity of 120 μJy/bm per epoch and a projected root mean square (RMS) depth of 70 μJy/bm when all three epochs are combined.

1.2.1 Transient-Focused Design and Implementation

The three epochs of VLASS are separated by approximately 32 months to observe and measure the variable and transient sources (Lacy et al., 2020). Each single epoch is divided into two parts, with half the sky (e.g., Epoch 1.1) observed about 16 months after the other half (e.g., Epoch 1.2) . The first two epochs were completed by 2019 July and 2022 June, respectively, with the first half of the third epoch completed in June 2023.

As the field of view of VLASS is significantly larger than the primary beam response of a single VLA antenna, the observations are done using a mosaicing technique. VLASS utilizes a type of mosaicing called the On-The-Fly Mosaics (OTFM) mode where the telescope “scans” the sky with the position of the antennas continuously moving with respect to the sky. Given the time it takes for the telescopes to slew and settle at a new position, a more traditional point-stare-repoint (discrete) mosaic method will have significantly more overhead than OTFM mode observations. This is because the cumulative VLASS depth in a single epoch for a given part of the sky is about the same as a 5-second pointed VLA observation.

The data collected in OTFM mode is then stitched together to allow for the high time and high angular resolution observations required for shallow surveys and transient searches. This OTFM mode was tested in the S-band Stripe 82 program of 13B-370, which is a precursor of VLASS (Mooley et al., 2016).

The VLA telescopes in the OTFM mode scan a strip along Right Ascension (R.A.) at a constant Declination (Dec) at a scan speed of 3.31 arcmin/s (Lacy et al., 2020). During the scan, the phase center of the array is discretely sampled every 0.9 seconds to minimize the smearing of the response. With this, VLASS covers 23.83 deg^2 per hour. To facilitate this scanning and mosaicing, the sky is divided into 899 Tiles of 10° × 4°, each observed for approximately 2 hours each. The tiles are ordered into 32 strips of Dec called “tiers”, laid out from South to North. Sets of 2–4 of these tiles are grouped together to form a scheduling block of 4–8 hours of observations.

1.2.2 Science and Data Products

VLASS was designed with 4 science goals in mind (Lacy et al., 2020):

1. Searching for transients;

2. Tracing the evolution of galaxies;

3. Measuring the magnetic field; and

4. Imaging and surveying the dust obscured regions in the Galactic plane

To enable these science goals, VLASS produces or will produce seven Basic Data Products (BDP): Raw Visibility Data; Calibrated Data; Quick Look Images; Single Epoch Images, Single Epoch Component Catalogues; Cumulative VLASS Images; and a Cumulative VLASS Component Catalogue. The planned release of these products ranges from immediately after observations to 12 months after observations.

1.2.3 VLASS Quick Look Images

The Quick-Look (QL) images were mostly designed to aid the search for radio transients (Lacy et al., 2020) and enable rapid identification of observations that needed to be retaken within a given epoch. These images are released within 2 weeks of the observations and are available[1] at https://archive-new.nrao.edu/vlass/quicklook/. Here, we discuss how those images are generated by an automated pipeline.

First bad data is flagged to remove problems caused by issues like recorded data acquisition errors and known or observed radio frequency interference. Then, the VLASS data that remained un-flagged are calibrated. Corrections to the raw data's amplitude and phase are made based on the intensity of previously known flux and complex gain calibration sources. These calibrated data are then used to produce the QL images using the automated Quick-Look Imaging Pipeline, which utilizes the `tclean` task in the Common Astronomy Software Applications (CASA) software package[2]. The `tclean` task reconstructs images from the visibility space, i.e., the sky as seen by the interferometer. For the gridder parameter in `tclean`, a *mosaic* convolution function is used to combine the measure visibilities (from the different OTFM scans) as the telescopes slew across a single 4° × 10° region (i.e., a "tile"). With this, 2° × 2° mosaic images are generated; from which the central 1° × 1° segments called "Subtiles" (QL images) are extracted. These QL Subtile images have

[1] The processed VLASS images are also ingested into the Canadian Astronomy Data Centre https://www.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/en/vlass/

[2] CASA is a data processing software primarily developed for radio telescopes, especially the VLA and the Atacama Large Millimeter/submillimeter Array (ALMA).

a pixel size of $1''$. The edge trimming ensures the corners of the images are correctly de-convolved. This whole process is performed in parallel to keep up with the rate of observation.

Another task called `imdev` in CASA is used to produce "RMS" images that estimate the noise level, i.e. the RMS of the QL images as a function of position. These RMS images are also included as a part of the QL images' BDPs.

Since these QL images are mainly used for quality assurance and transient searches they are produced on a short timescale (∼two weeks). Fast production of these images using relatively simpler algorithms results in residual artifacts around bright sources (that will likely be better minimized in VLASS Single Epoch images). The cause of such artifacts is given below.

1.3 Interferometers and the Creation of Artifacts

To understand the origin of the artifacts seen in VLASS QL images, interferometry and the synthesis imaging technique used to image the sky needs to be understood.

Interferometry was developed to gain better resolution that cannot be obtained with single antenna telescopes. For example, with an angular resolution θ given by

$$\theta = 1.22\frac{\lambda}{D} \tag{1.1}$$

where λ is the wavelength (in m) and D is diameter of an antenna (in m), a single dish antenna should have a diameter of $\sim 25\,\mathrm{km}$ to achieve $1''$ resolution at $3\,\mathrm{GHz}$. The difficultly in building and maintaining such large telescope antennas led to synthesizing large apertures by using multiple significantly smaller antennas separated by a distance D instead of one large antenna (although here $\theta = \frac{\lambda}{D}$ is the more relevant equation).

Each pair of antennas in an interferometer observes a source of brightness $I(\mathbf{s})$, where $\mathbf{s}$ is the unit vector in the direction of what the pair of antennas is observing, as a complex visibility $V_\nu(\mathbf{b})$,

$$V_\nu(\mathbf{b}) = \int_S A_\nu(\mathbf{s}) I_\nu(\mathbf{s}) e^{-2\pi i \nu \frac{\mathbf{b}\cdot(\mathbf{s}-\mathbf{s}_o)}{c}} d\Omega \tag{1.2}$$

where A_ν is the primary beam sensitivity, S is the surface area of the source, $\mathbf{b}$ is the baseline length, $d\Omega$ is the solid angle and $\mathbf{s}_o$ is the unit vector in the direction of the phase center of the antennas. As Equation (1.2) is in the form of a 2D Fourier transform, the brightness of the source can be recovered by taking the inverse Fourier transform of the complex visibility. Combining the obtained brightness of the source through each pair of antennas provides a reasonable estimate of the true brightness of the source.

Since each pair of antennas observes the source as a complex visibility V_ν, which is a function of its baseline $\mathbf{b}$, V_ν can be represented in terms of the baseline geometry as shown in Figure 1.4 (adapted from Thompson et al., 2017). Each baseline vector $\mathbf{b}$ (the separation between each pair of antennas) has 3 components (u, v, w) towards the East, North and the direction of the source respectively (measured in number of wavelengths). Similarly the source in the sky plane is represented by (l, m, n). If the source is sufficiently small, i.e., l and m are small, the nw component of the $\mathbf{b}\cdot\mathbf{s}$ term in Equation (1.2) approaches 0. With this the visibility V can be written in terms of (u, v) as,

$$V_\nu(u, v) = \int_S A_\nu(l, m) I_\nu e^{-2\pi i(ul+vm)} dl dm \tag{1.3}$$

Therefore, the sky (in this example, only a single source) is observed as a set of visibilities V in the (u, v) plane. More antennas leads to more points in the uv-plane, resulting in a clearer observation of the source. But, due to the finite number of antennas in any interferometer, the uv-plane is sparsely sampled. If the uv-plane is thought of as observing a point source with all the pair of antennas, the Fourier transform of the uv-plane produces the synthesized "dirty" beam of the interferometer, i.e., how the interferometer sees the point source. An example of the uv-plane of VLA in B configuration with a phase center at R.A. $\sim 289.50^\circ$ and

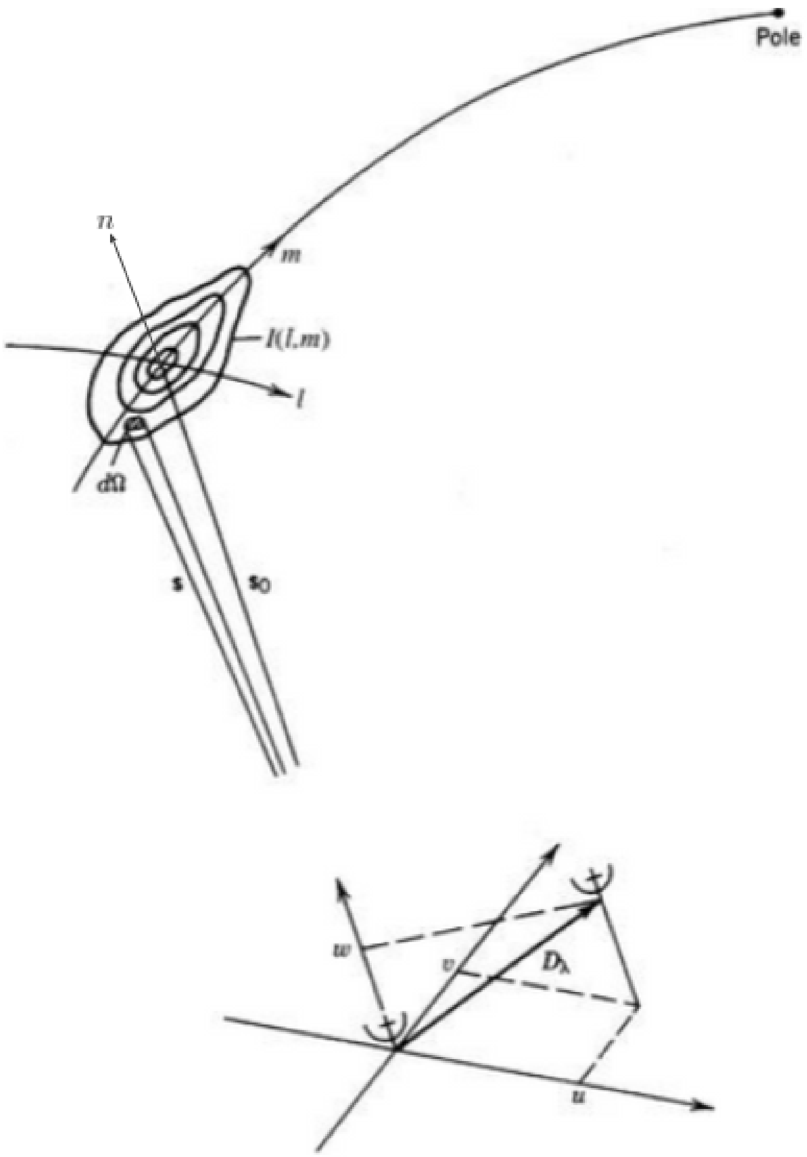


Figure 1.4: This figure (adapted from Thompson et al., 2017) shows the geometric representation of $(u,\, v)$ plane and the $(l,\, m)$ plane.

Dec ∼ −39.50° and its resulting dirty beam is shown in Figure 1.5. As seen in the example, the dirty beam of VLA has a characteristic hexagonal spokes pattern produced due to the lining of the antennas along the three linear arms. The angles of the spokes in the dirty beam depend on the projected orientation of each arm relative (on the *uv*-plane) to the location of the source in the sky. Unlike in typical synthesis imaging, where the Earth's rotation allows an interferometry to better sample the *uv*-plane, VLASS data are essentially equivalent to a single snapshot in *uv*-plane with

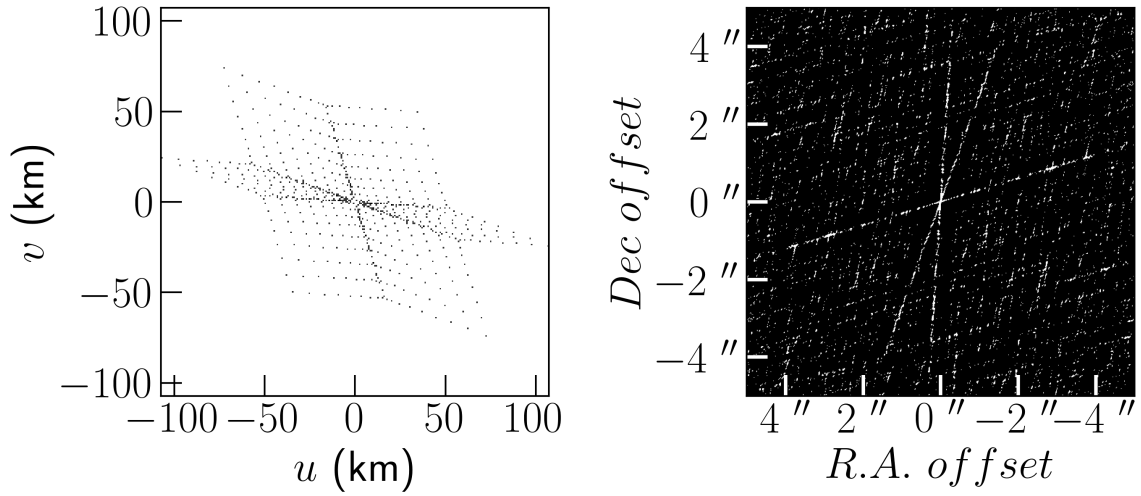


Figure 1.5: An example of a uv-plane and the corresponding dirty beam resulting from the uv-plane.

up to $N(N-1)/2$ baselines, where N is the number of antenna with unflagged data ($N = 27$ if no antenna has completely flagged data).

Since the sky is imaged as the sky brightness I_ν convolved with the dirty beam, most of the linear artifacts seen around sources are aligned with the spokes in the dirty beam. Additional linear artifacts can also arise due to the incomplete sampling of the interference pattern caused by two sources near each other, or even two components of a multi component astrophysical source. Well established techniques (like `clean`; Högbom, 1974) de-convolve the Fourier transform of the complex visibility V_ν i.e. dirty image, to reconstruct the sky brightness by maximizing the information from the partially sampled uv-plane. However, these techniques can sometimes over or under "clean" (de-convolve) the image resulting in linear structures/streaks particularly around bright sources. Two examples of these linear artifacts around bright sources are shown in Figure 1.6. When these linear streaks overlap with an astrophysical source or one of its components, the streak can enhance or suppress the measured flux leading to incorrect results. Furthermore, automated source identification algorithms can mistake local maxima on these streaks from noise as real astrophysical sources.

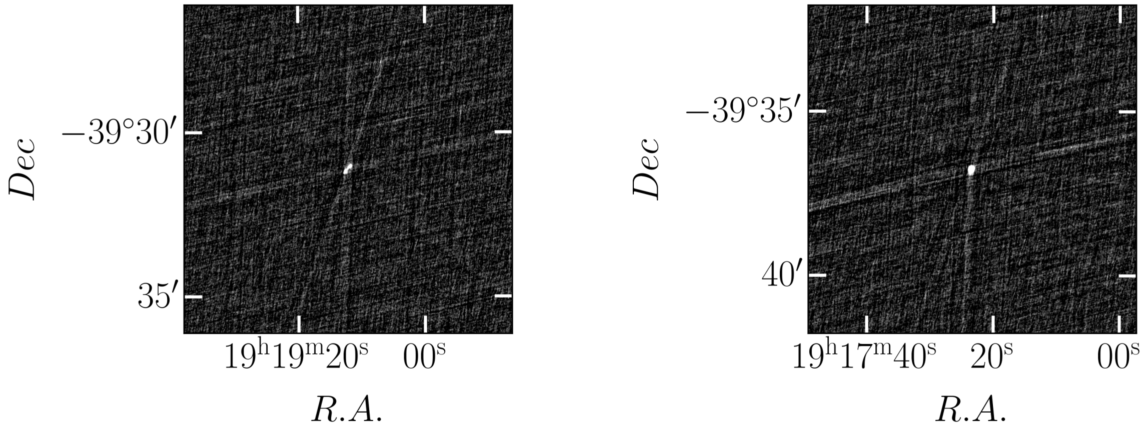


Figure 1.6: Cutouts of VLASS QL images around bright sources shows linear artifacts mostly at the angles seen in the dirty beam overlapping the source. Streaks seen at other angles are from the interference pattern caused by two nearby components/-sources.

1.4 CIRADA Quick-Look Components Catalogue

The Canadian Initiative for Radio Astronomy Data Analysis (CIRADA)[3] produced a catalogue of radio sources/components[4] using the QL images (Gordon et al., 2021). The authors of Gordon et al. (2021) published an updated Version 2 catalogue to the CIRADA website, which will be the catalogue that we use throughout this thesis. The components were extracted by applying a source detection algorithm from the Python Blob Detector and Source Finder (PyBDSF)[5] (Mohan & Rafferty, 2015) called `process_image` on 35,825 Subtiles to detect and fit ellipses to flux regions of 3σ above the image average. The identified regions for which an ellipse could not be fitted were also included in the catalogue for completeness. Since VLASS QL images have overlapping regions, there are some sources that are identified in more than one Subtile. To indicate the presence of such duplicates or poor detections, flags were included in the catalogue. The catalogue comprises of sources that are simple single, or at most, close-double/triple components. More complex sources were not included

[3] https://cirada.ca/

[4] At the time of this catalogue creation, the Epoch 1 QL images were not reprocessed with the updated QL processing pipeline and the catalogue was generated with an older version of the images.

[5] See: https://github.com/lofar-astron/PyBDSF

in the catalogue.

The catalogue consists of 3 tables: Component Table; Host ID Table; and Subtile Information Table that provide basic information of the parameters fit by PyBDSF. The three tables are linked to one another via (primary) key columns. The Host ID Table and the Subtile Information Table are not used for this work; here we only focus on the Component Table.

Component Table

The component table has a total of 3,381,277 radio components identified in 35,825 Subtiles. Here complex sources with more than one component are listed as different components rather than a single source. The key columns of this table used for this thesis work are given in Table 1.1.

For the identified components, parameters like the sky coordinates, image plane coordinates, component size like the full width at half maximum (FWHM) of the semi major and minor axis, source flux density, total flux along with their uncertainties are provided in the table. Other parameters related to components detection like *Quality_flag* (to indicate spurious or poor detection), *Duplicate_flag* (to indicate whether the same component is found in any other adjacent QL image) and *NN_Dist* (distance to the component's nearest neighbouring identified component) are also included in the table.

The table also contains a quality assurance metric called *Peak_to_ring*. It is the ratio of the peak flux of the component to the peak flux in an annular region of radii $5''$ and $10''$ around the component. For isolated components (with no significant overlapping streaks), the peak flux in an annular region is just the maximum value of the background noise. But, for components with overlapping streaks, the peak flux in an annular region is contributed by the flux density of the overlapping streak(s), which tend to have a higher peak flux density than the background noise level. Therefore, low peak-to-ring flux density ratios can indicate false positive components that were

Table 1.1: Key columns from the CIRADA catalogue's Component Table (reproduced from Version 2 of Gordon et al. (2021)).

Column Name	Column Description	Units
RA	Right ascension of component	[deg]
DEC	Declination of component	[deg]
Peak_flux	Peak flux of component	[mJy/beam]
E_Peak_flux	Error in component peak flux	[mJy/beam]
Maj_img_plane	Component major axis size in the image plane (FWHM)	[arcsec]
Min_img_plane	Component minor axis size in the image plane (FWHM)	[arcsec]
PA_img_plane	Component position angle in the image plane, east of north	[deg]
Tile	VLASS tile this component is located in	
Subtile	VLASS subtile this component is located in	
Peak_to_ring	Ratio of the Peak flux to the maximum flux in annulus of r, R = 5, 10" centred on component RA, Dec.	
P_sidelobe	The probability that the component is a false-positive detection caused by a sidelobe. Determined using a SOM and manual classification	
NN_dist	Angular distance to nearest other component in the Component Table	[arcsec]

fit to imaging artifacts. Yet, such low values can also arise from nearby components. Therefore the *Quality_flag* is increased by 1 only if the *Peak_to_Ring* value is < 2 and the nearest identified component is $> 20''$ away. Although this metric captures fairly isolated false-positives, they misidentified spurious sources caused by linear streaks around bright sources. Conversely, some components in multi-component sources are misclassified as false-positives with this metric. The Version 2 of the catalogue includes 1,291 sources that do not have a validly measured *Peak_to_ring* value (-99). From the entire sample, 291,870 sources have a quality flag that indicate they are likely artifacts (8.63%). In addition, there are an additional 386,706 isolated components with a *Peak_to_Ring* value of 2–3 (11.44%).

An additional metric, *P_sidelobe*, that indicates the probability of an identified component being an artifact caused by sidelobes was later added to the catalogue by Vantyghem et al. (2021). This probability is only calculated for components with *Peak_to_Ring* < 3, and some sources in this range are identified as artifacts in this metric, but not the *Peak_to_ring* method. This is why we report the number of isolated components (mentioned above) with a *Peak_to_Ring* value of 2–3, and will consider the full *Peak_to_Ring* < 3 range as potential artifacts in Section 4.3. Vantyghem et al. (2021) set the threshold to 3 with the assumption that most components with higher values are likely to be real astrophysical sources.

To calculate *P_sidelobe*, an unsupervised machine learning technique called the Self-Organizing Maps (SOM) was used to examine component morphologies. For this, a 10×10 neuron grid SOM was trained using 40,000 images containing 1,080,393 components in total (Vantyghem et al., 2021). 100 random components that best matched with that particular neuron were visually inspected by a group of astronomers. Here best matches refers to the neuron with the most similar morphology compared to the component (i.e., the least distant neuron). Based on the fraction of artifacts that were associated with each neuron during inspection, a probability is assigned to that neuron. The process yields a *P_sidelobe* for each neuron. All the components in the catalogue with *Peak_to_Ring* < 3 were mapped to the best matching neuron and inherited their associated *P_sidelobe*. The *Quality_flag* of components with a *P_sidelobe* ≥ 0.10 were incremented by 8 to ensure the catalogue has less false-positive components.

Adopting the same nearest neighbour difference of $> 20''$ as above to indicate a component's level of isolation, the SOM method identifies: 134,731 components as isolated sources likely to be artifacts; 148,822 components as non-isolated sources likely to be artifacts; 544,222 components as non-isolated sources unlikely to be artifacts; 328,613 components as non-isolated sources unlikely to be artifacts; and leaves 2,224,886 components unevaluated. In total, this method finds 8.39% of the total component list to likely be artifacts.

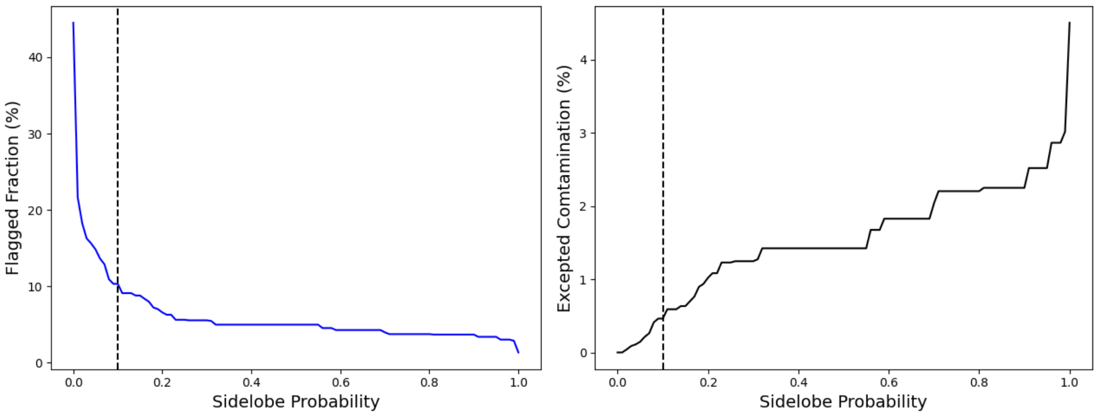


Figure 1.7: This figure (from Vantyghem et al., 2021) shows the percentage of flagged components and estimated contamination percentage as a function of the probability of sidelobe calculated using SOM. Here the probability threshold is set to 0.10 which flags about 10% (*left panel*) of the components and leaves around 0.46% of the false-positive component undetected (*right panel*).

Figure 1.7 (from Vantyghem et al., 2021) shows the effect of setting the sidelobe probability threshold to 0.10 on the contamination percentage and number of spurious components flagged. At the 0.10 threshold only 0.46% of the components in the catalogue are likely to be false-positive components after flagging (which flags about 10% of the components in the catalogue as false-positives). Increasing this suggested threshold to a higher value leads to more tolerance towards likely spurious components in the catalogue.

1.5 Thesis Goals

In this work, we present a new empirical method of detecting linear artifacts present in VLASS QL images using the Hough Transform. The applications of the identified artifacts in characterizing and classifying the already identified sources in the QL images is also discussed. Furthermore, we review the use of the new method as a quality assurance step during image reconstruction. Finally, we compare the new method with two previous techniques of classifying the sources.

Chapter 2

Hough Transform

The Hough Transform (HT) is a pattern recognition technique invented to identify straight lines in images. Since its inception (Hough, 1962), the algorithm has been advanced to detect other features and shapes that have a 2D analytical form — like circles and ellipses. This technique is based on a voting mechanism where each set of parameters that describe a feature in the image space is voted for in the parameter space. Local maxima in this parameter space represents the features' presence in the image space.

In contrast to the above mentioned *classical* HT, a *generalized* HT has also been developed to recognize shapes that do not have a 2D analytical form (Ballard, 1981). Here we will only focus on the *classical* HT.

2.1 History

The *classical* HT was originally developed to detect complex lines in photographs with applications in tracing the path of subatomic particles in viewing fields and pictures of bubble chambers (Hough, 1962). In this technique an image is transformed from an (x, y) space to the feature's parameter space (in this case, those of straight lines).

For example, a straight line given in the form

$$y = mx + b \tag{2.1}$$

is characterized by the slope m and the intercept b. Therefore the parameter space

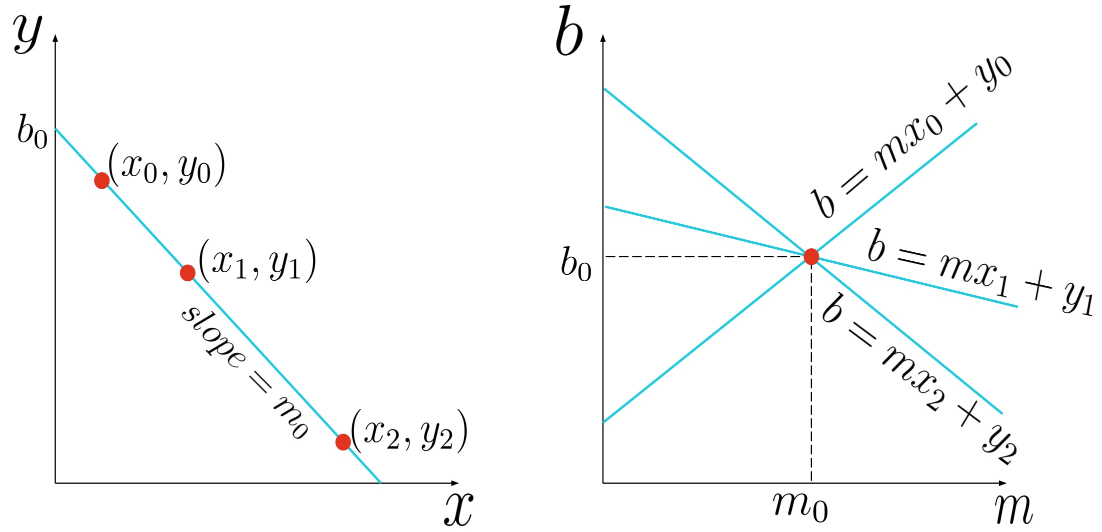


Figure 2.1: Mapping of points on a line in the image space (*left panel*) to lines in the *classical* HT (m, b) parameter space (*right panel*). The peak in the parameter space describes the (m, b) of the line in the image.

is an (m, b) space. For every straight line in the image, the *classical* HT's voting procedure votes for a (m, b) pair in the parameter space. Consequently, every straight line in the image is transformed to a point in the parameter space.

The implementation of this algorithm is designed to account for pixelation of images. Each non-zero pixel (x_i, y_i) in the image can be thought to have infinite lines passing through the pixel. The (m, b) pair of each of these lines are voted in the parameter space that traces out a line. Therefore each point in the image is represented as a line in the parameter space as shown in Figure 2.1. Since the voting mechanism accumulates a count for each pair of parameters that is a potential line candidate, the local maxima of accumulated counts in the parameter space indicates the slopes and intercepts (m_i, b_i) of the most probable lines in the image space.

Although an ingenious method, the original *classical* HT has a few limitations. Such limitations predominantly occur because both the slope and intercept in the

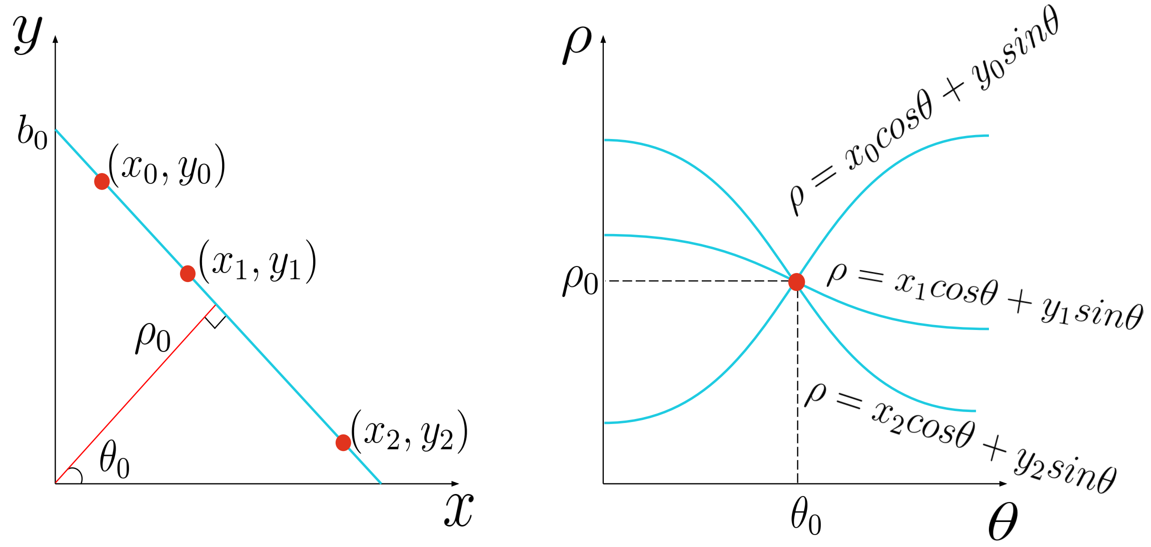

y
b_0
(x_0, y_0)
(x_1, y_1)
ρ_0
(x_2, y_2)
θ_0
x
ρ
$\rho = x_0 cos\theta + y_0 sin\theta$
ρ_0
$\rho = x_1 cos\theta + y_1 sin\theta$
$\rho = x_2 cos\theta + y_2 sin\theta$
θ_0
θ

Figure 2.2: Mapping of points on a line in the image space (*left panel*) to lines in the HT (θ, ρ) parameter space (*right panel*). The peak in the parameter space describes the (θ, ρ) of the line in the image.

parameter space are unbounded. For example, identifying vertical lines using this method is not possible as their slopes are undefined.

2.2 Existing Algorithm

Duda & Hart (1972) improved the *classical* HT by moving from slope-intercept parameter space to an angle-radius parameter space. In this *enhanced classical* HT approach (henceforth referred to without the prefixes *enhanced* or *classical*), every straight line is represented by the normal of the line passing through the origin. Accordingly, every straight line in the image is characterized by the normal line of the polar form

$$\rho = x \cos\theta + y \sin\theta \tag{2.2}$$

where ρ is the length of the normal from the origin and θ is the angle the normal makes from the X-axis. This characterization of the straight line leads to a (θ, ρ)

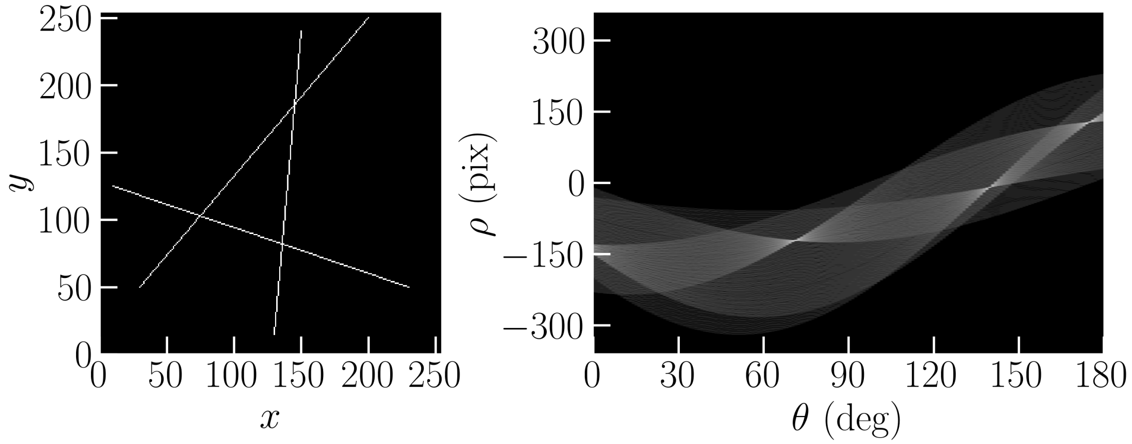


Figure 2.3: *Left:* An example of the HT of a Boolean image with three clear lines. *Right:* HT parameter space of the image with axes corresponding to the position angle θ and the distance to a given line ρ. The three maxima are the (θ, ρ) of the three lines in the image.

parameter space. Again, straight lines in this parameterization are represented as points in a parameter space, here (θ, ρ).

During implementation every non-zero pixel in the image is denoted by a cosine curve in the parameter space since the infinite possible lines that could pass through a pixel given by Equation (2.1) trace a sinusoidal curve. Therefore any line, a collection of co-linear points, produces a set of cosine curves, as shown in Figure 2.2, that intersect at a specific (θ, ρ) pair. An example of HT of image with three intersecting lines is shown in Figure 2.3. In this example and in all further references to HT, we consider θ as position angles (PA). Due to the voting procedure involved, the resulting local maxima in the parameter space (shown in the second panel of both Figure 2.2 and Figure 2.3) strongly indicate the presence of straight lines in the image space at the corresponding θ and ρ. This technique overcomes the previously mentioned limitations of identifying vertical lines since θ and ρ are both bounded in this parameterization.

2.3 Key Characteristics

There are a few properties that are intrinsic to the HT that can affect the detection of lines:

- The HT is tolerant towards imperfections in the features it is designed to detect. For example, the HT can detect lines even if there are gaps or missing pixels in a line. As a result, different line segments are detected as a single line instead of separate distinct lines.

- The binning of θ and ρ determines whether a straight line is picked by the algorithm or not. If the HT has high resolution binning, the votes accumulated by a pair of parameters might be too low to be detected as a peak amidst the background noise.

- The algorithm can be fooled by apparent lines in a noisy image. For example, in an binary image with random noise, straight lines at 45° and 135° can be detected. This is because longer lines are made up of more pixels and hence accumulate more counts. Therefore diagonals (the longest lines that can be drawn on an image) account for the highest peaks even if there are no real lines in the image at those angles.

2.4 Previous Uses in Astronomy

The HT and its variants have been extensively used in astronomical data analysis. A variation of the HT was used by Ragazzoni & Barbieri (1994) to determine the period and period variations of an eclipsing binary star GW Cep by studying their light curve. Ballester (1994) used the HT for the automated wavelength calibration of long-slit spectra at the European Southern Observatory (ESO) New Technology Telescope. Fridman (2010) demonstrated the use of the HT to detect dispersed transients by identifying straight lines in the time-frequency plane of a LOFAR observation to trace the transient's track before de-dispersion. Hollitt & Johnston-Hollitt (2012) illustrated the use of the *circle* HT to detect and characterize arc-like sources such as supernova remnants, narrow-tailed radio galaxies, and radio relics in radio observations from Molonglo Galactic Plane Survey (MGPS) and Australia Telescope Compact Array (ATCA). A variant of the HT, called the *rolling* HT (RHT), was used by Clark et al. (2014) to study the atomic hydrogen (HI) maps by identifying local linear fiber-like structures. They were able to show that the magnetic field lines and the HI fibers were aligned in the high Galactic latitude inter-stellar medium (ISM), therefore indicating that the fine magnetic field structure can be traced by using RHT. Zuo & Chen (2020) also provided a proof of concept for the detection of radio bursts by detecting straight lines in the time-frequency plane using the HT.

Chapter 3

New Methods to Detect and Quantify Linear Artifacts

Our development of new methods to detect and quantify linear artifacts are the center of this thesis. These methods include both methods that can be generalized to a wide range of cases both in astronomy and beyond, as those that are more specifically suited for images generated by the VLA.

3.1 New Detection Methods

3.1.1 The Need to Extend the Hough Transform

Even though VLASS QL images mostly contain linear artifacts, there are a few reasons why the standard HT cannot be directly used to detect these artifacts:

- The HT does not consider the strengths of pixels; instead it transforms any non-zero pixel to a cosine curve in the parameter space. With most pixels in the VLASS QL images being non-zero, applying HT directly on the image weighs all the pixels equally and produces a uniform parameter space without any distinct maxima from the streaks. Although this is a standard issue in the use of the HT on generalized images, basic thresholding on the input image did not considerably improve results. An example is shown in Figure 3.1. The absence of distinct peaks results in low global peak detection threshold, which

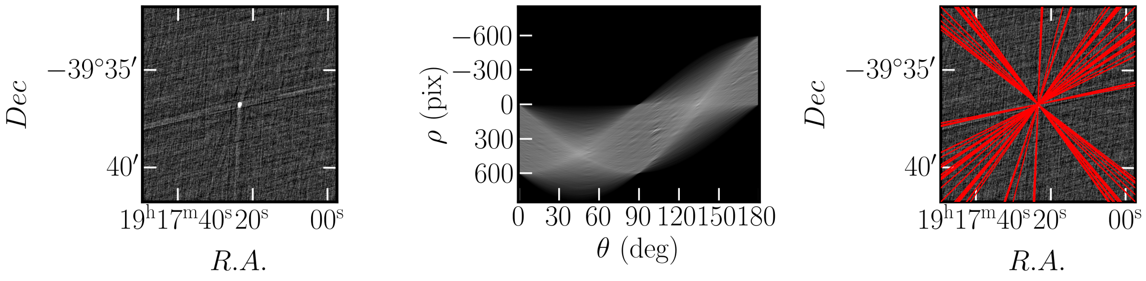


Figure 3.1: *Left:* Cut out of a VLASS QL image around a component with overlapping artifacts. *Center:* Hough Transform of a Boolean-converted version of the *left* image, where the Boolean image was produced by clipping values below 0. *Right:* The lines identified by HT around the component at a detection threshold of 0.6 times the parameter space maxima.

in turn leads to an increase in the detections of false positive streaks around bright sources and false negative streaks around relatively dimmer sources.

- The HT can be ineffective in identifying thick lines and does not intrinsically measure the widths of lines. As illustrated in the rightmost panel of Figure 3.1, the algorithm detected a thick line (at ∼110° position angle) as multiple single-pixel wide lines at slightly different angles.

- Edge detection algorithms, often used as a pre-processing step prior to HT (to increase the chances of true positive detection), cannot be used for streak detection since VLASS QL images are particularly noisy. Additionally, in the process of finding the edge of any object/feature in the image, edge detection algorithms significantly reduces the possibility of regaining other statistical information like the width of the lines.

Therefore a statistically-motivated pre-processing procedure and extension of the results of HT specifically designed to take advantage of intrinsic patterns present in VLASS QL images is necessary for better artifacts identification. While some of the techniques discussed in this chapter can be applied to a broad set of noisy images, other techniques are designed to take advantage of the properties specifically present in QL images due to the snapshot nature of the observations performed using VLA.

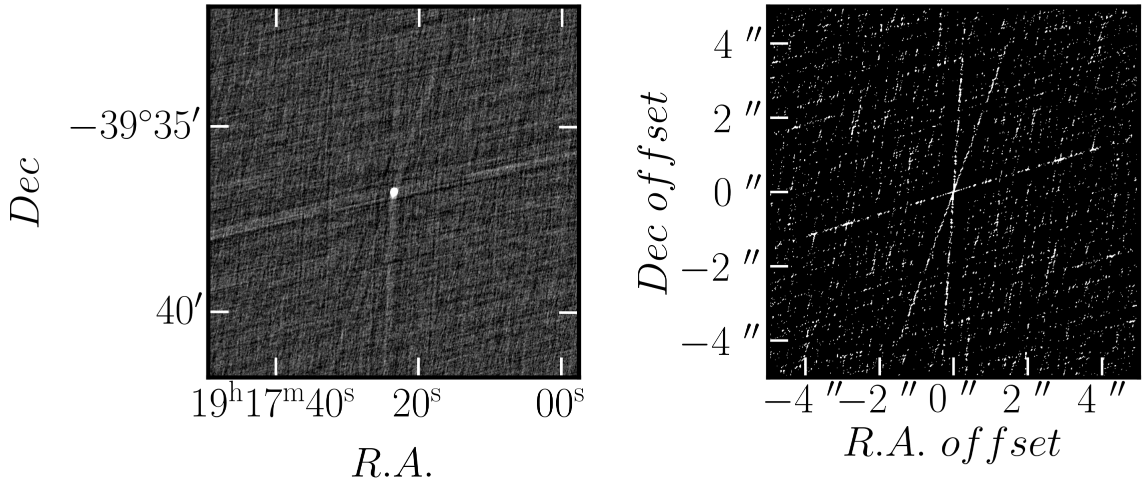


Figure 3.2: *Left*: Cutout of a VLASS QL image of a component used throughout this Chapter to illustrate the extended HT procedures. *Right*: The Point Spread Function (PSF, also known as the dirty beam) of image shown in the *left* panel. The artifacts seen in the image are at a similar angles to the sidelobe streaks seen in the dirty beam image. Artifacts like those seen at other angles, which happen to be lighter in this image, can be from signal interference between nearby components or artifacts from another source that coincidentally overlap the targeted component.

The large number of parallel lines seen in VLASS QL images are usually at the same angles as that of the "hex" line pattern seen in the Point Spread Function (PSF; or dirty beam) of the image (see Figure 3.2). As previously mentioned in Section 1.3, streaks seen at other angles are usually caused by interference pattern from nearby sources. The parallel lines seen in the images are mostly due to the effects of phase errors, the linear arms of the VLA, and the incomplete image reconstruction; and therefore are inherent to all VLA images. Since a significant fraction of artifacts are sidelobe streaks resulting from incomplete *uv*-sampling, these streaks are usually found at the same angles as that of the parallel lines in images. The new method developed here takes advantage of the presence of these parallel lines to efficiently identify and characterize the linear artifacts in the image by increasing the signal-to-noise ratio of these artifacts during detection.

Since our scientific goals centre around detected sources in VLASS QL images, some of the techniques are developed such that they need not be blindly applied to the whole image. The subset of techniques presented here that focus on streak detection in a specific portion of an image can still be applied beyond VLASS QL images, but are currently tuned for use on VLASS QL images.

3.1.2 Identifying Linear Streaks

The new method of linear artifacts identification is applied to VLASS QL images specifically around components already identified and catalogued by Gordon et al. (2021). Basic information about the potential radio sources (or components of radio sources) like sky coordinates, image plane coordinates, total and peak flux of the component, semi-major and semi-minor axis of the ellipse fitted to the component, etc., in the catalogue is used in the new artifact identification and analysis.

The steps of the new algorithm are discussed individually in the following sections. Throughout these steps, the example of a component at R.A. 261.584° and Dec 32.322° in the VLASS QL Subtile VLASS1.1.T19t22.J172805+323000 is used to illustrate and demonstrate the procedures. The source code for our new artifact identification and component classification (discussed in Chapter 4) algorithm is available at https://github.com/SiniSRao/VLASS-Artifact-Identification.

Clipping & Masking

As mentioned in Section 3.1.1, the direct HT of VLASS QL images do not identify prominent linear artifacts without significant false positive detections. Therefore, as a part of the pre-processing stage, the images are converted to Boolean images by clipping at the 90^{th} percentile confidence interval that would be consistent with noise. This is done by converting all pixels with value > than 1.64σ to 1 while the rest are converted to 0. This threshold was chosen after careful visual inspection of the remnant noise level in the image after clipping. The estimation of the threshold level is made robust by calculating the standard deviation equivalent σ from the median absolute deviation (MAD) of the image ($\sigma_{\text{MAD}} = 1.4826\,\text{MAD}$), which is more resilient

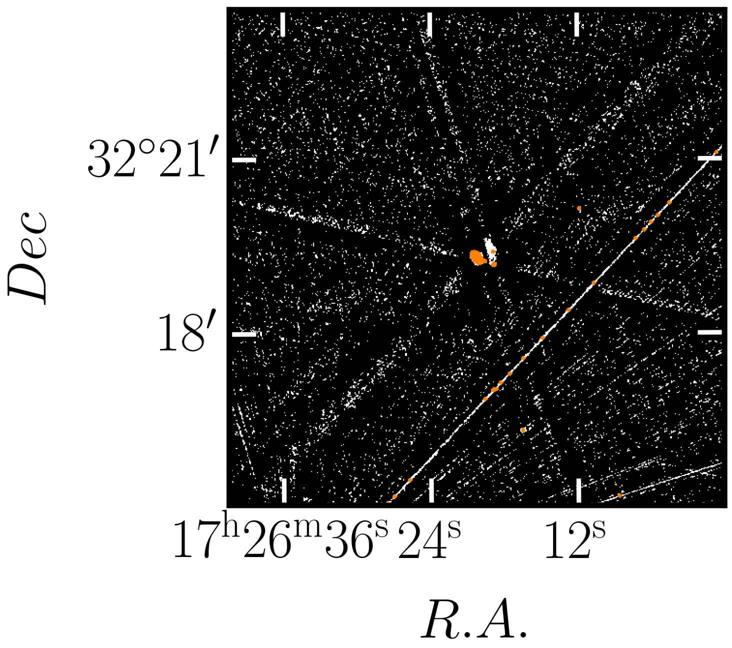


Figure 3.3: An example (cutout of a) VLASS QL image clipped to show outliers (in *white*) above the 90^{th} percentile positive threshold. The outlier-robust threshold is calculated before masking out regions associated with catalogue-identified components in the image region (orange regions). Since component signal in these masked regions can affect artifact detection, the masked regions are excluded from future analysis steps.

to outliers. Since a VLASS QL image that is devoid of signal is expected to have zero intensity i.e., without noise, the noise measurement is applied against this null hypothesis, resulting in a threshold ($T = 1.6449\,\sigma_{\mathrm{MAD}}$) that is applied separately to both positive and negative outliers.

For an image pixel p_i, a clipped image pixel q_i given by,

$$q_{\mathrm{pos},i} = \begin{cases} 1, & \text{if } p_i > T \\ 0, & \text{otherwise} \end{cases} \qquad \text{and} \qquad q_{\mathrm{neg},i} = \begin{cases} 1, & \text{if } p_i < -T \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

VLASS QL images have both positive and negative linear artifacts. The negative linear artifacts can arise due to excessive `clean`-ing of the images during QL image generation and de-convolution. Therefore, to detect both types of streaks, the image is clipped for both positive and negative pixels to get a corresponding positive-space image (q_{pos}) and negative-space image (q_{neg}), respectively. For the rest of the chapter the positive-space case is used for illustrations and examples.

Next, all components identified by Gordon et al. (2021) in the image are masked, even if they are classified as potential artifacts by the *Peak_to_Ring* or the SOM metrics. This ensures that only signal from the streaks and the background (noise) is considered. Figure 3.3 shows an example of a clipped and masked VLASS QL image around a component. The masked regions are highlighted in orange. Pixels in the resulting clipped and masked image, either $q_{\mathrm{pos,msk},i}$ or $q_{\mathrm{neg,msk},i}$ are passed to the next step; hereafter refer to as q_{msk}.

Source & Background Region Extraction

Applying a global HT to the entire image can lead to failed artifact detections around fainter sources. To prevent this, local annular regions around components are considered. As mentioned in Section 1.4, the *Peak_to_Ring* metric used an annular region of radii $5''$ - $10''$ to catch false positive components. Efficient identification of linear streaks in the same annular region around components is not achievable in the new method as the the signal-to-noise ratio (SNR) of the streaks is not enhanced in such small regions. Therefore an annular region of radii 5–50 times the semi-major axis (Full Width at Half Maximum, FWHM along the major axis) of the component is used in the new method.

The annular region around every component extracted from the clipped and masked image (q_{msk}) is considered separately as a source region, src. Additionally, a background region, bkg, is extracted by setting all the pixels in the same region to 1.

For the extraction of these annular regions for each component, a mask, which we label the annularMask, is generated centered at the component. The source and background regions, src_i and bkg_i respectively, are extracted by masking out all the

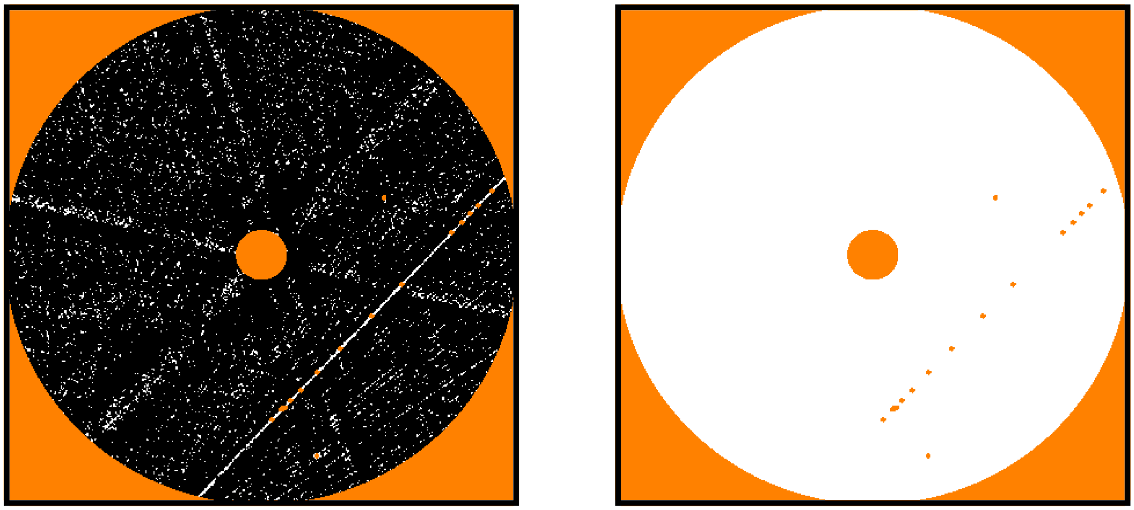


Figure 3.4: Annular source (*left panel*) and background (*right panel*) regions around the component under consideration. The orange regions marked inside the annular regions are the masks applied to the components in the catalogue.

pixels outside the annular region, annularMask.

$$\mathrm{src}_i = \begin{cases} q_{msk,i}, & \text{if } \mathrm{src}_i \in \text{annularMask} \\ 0, & \text{otherwise} \end{cases}$$

$$\mathrm{bkg}_i = \begin{cases} 1, & \text{if } \mathrm{bkg}_i \in \text{annularMask} \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

The inclusion of the aforementioned background region is to make sure that the pattern intrinsic to the shape of any given region of interest (as mentioned in Section 3.1.1) in the HT does not influence the detection of the linear artifacts. Figure 3.4 shows an example of the source and background regions extracted around a component.

HT & its 1D profiles

The SNR of the signal from linear artifacts can be increased by utilizing the signal from parallel streaks seen in the images. To do so, the HT of the entire clipped and masked image ($\mathrm{HT}_{q_{\mathrm{msk}}}$) along with the HT of the source and the background regions are generated ($\mathrm{HT}_{\mathrm{src}}$ and $\mathrm{HT}_{\mathrm{bkg}}$, respectively). Figure 3.5 shows in highlights the

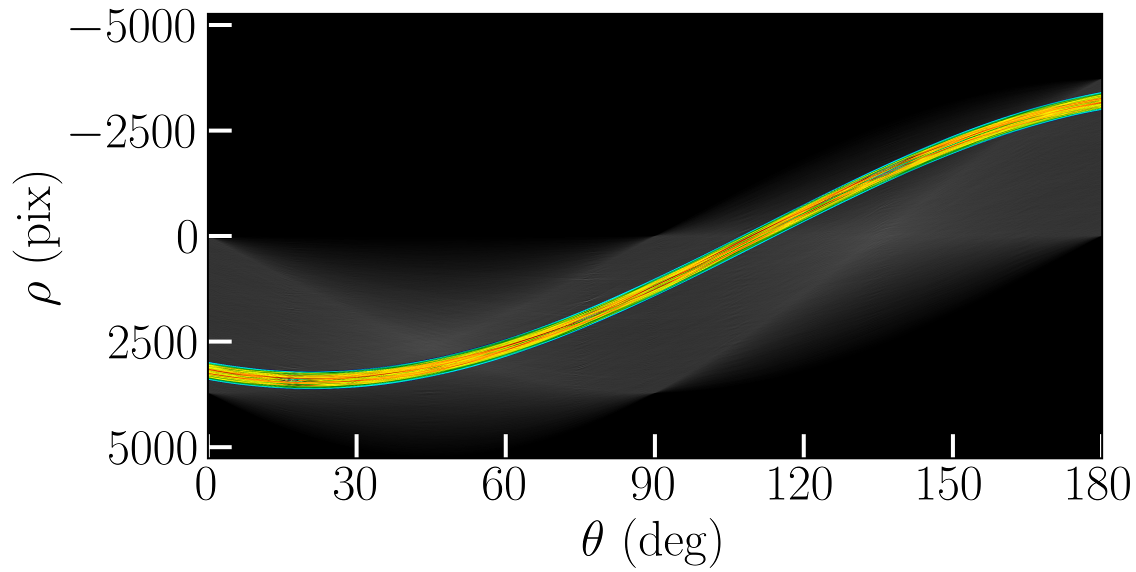


Figure 3.5: Hough Transform parameter space of the annular source region (in grey) overlaid on the Hough Transform parameter space of the full clipped & masked image (in colour).

HT of the source (and background) region against the HT of the clipped and masked image. Here the resolution of the HT parameter θ is fixed at 0.2° for better accuracy.

Next, the average of the clipped image HT weighted by the HT of the source region along the distance parameter is generated:

$$\begin{aligned}\text{Profile}_{\text{src}} &= \frac{\sum_\rho \text{HT}_{\text{src}}(\theta_i,\rho)\cdot \text{HT}_{q_{\text{msk}}}(\theta_i,\rho)}{\sum_\rho \text{HT}_{\text{src}}(\theta,\rho)} \\ \text{Profile}_{\text{bkg}} &= \frac{\sum_\rho \text{HT}_{\text{bkg}}(\theta_i,\rho)\cdot \text{HT}_{q_{\text{msk}}}(\theta_i,\rho)}{\sum_\rho \text{HT}_{\text{bkg}}(\theta,\rho)}\end{aligned} \tag{3.3}$$

This produces one-dimensional profiles (against position angles) of the source region HT. The same procedure is repeated for the background region HT. Weighing the HT of the clipped image with the HT of the source region increases the signal strength of the annular region and simultaneously suppresses the signal from other regions.

Figure 3.6 shows the 1D source and background profiles. Here the y-axis represents the weighted average counts of pixels at every tested position angle; and hence increases the signal-to-noise ratio of the streaks. The profiles can be considered analogous to the signal strength at each angle. But, the peak at 45°, the peak at 135°, and the general trend seen in the 1D profiles are direct results of the HT algorithm itself, and are not related to linear artifacts in the image.

Background Scaling & Subtraction

Assuming the source region consists of signal from the streaks and random noise, during 1D profile generation procedure discussed in Section 3.1.2, the source region emphasizes/enhances both the signal and noise (present in the image) in the 1D profile. On the contrary, the background region (where all unmasked pixels are set to 1) uniformly weighs the HT of the clipped image (i.e. no weight is added). Although the general trend is seen in both profiles, the background 1D profile has fewer counts compared to the source 1D profile, as seen in Figure 3.6, because of the uniform weights.

To account for the noise in the source region and its weight on the 1D profile, the background 1D profile is scaled up, appropriately incorporating the size of image noise in the source profile. A uniform noise throughout the image is assumed and (thus) applied to the component in question as well. Algorithm 1 shows the iterative process used for noise extraction and addition to the background profile to generate a scaled background profile. Figure 3.7 shows the HT profile of the source and the HT profile of the background profile after scaling.

Peak Detection

To make the streak detection independent of the effects of HT, the scaled background 1D profile is subtracted from the source 1D profile. The highly pixelated image produces a noisy background-subtracted 1D profile. The profile is smoothed to remove the effects of this noise. The resulting smoothed background-subtracted profile is therefore used as a robust statistical basis for streak identification as it is independent

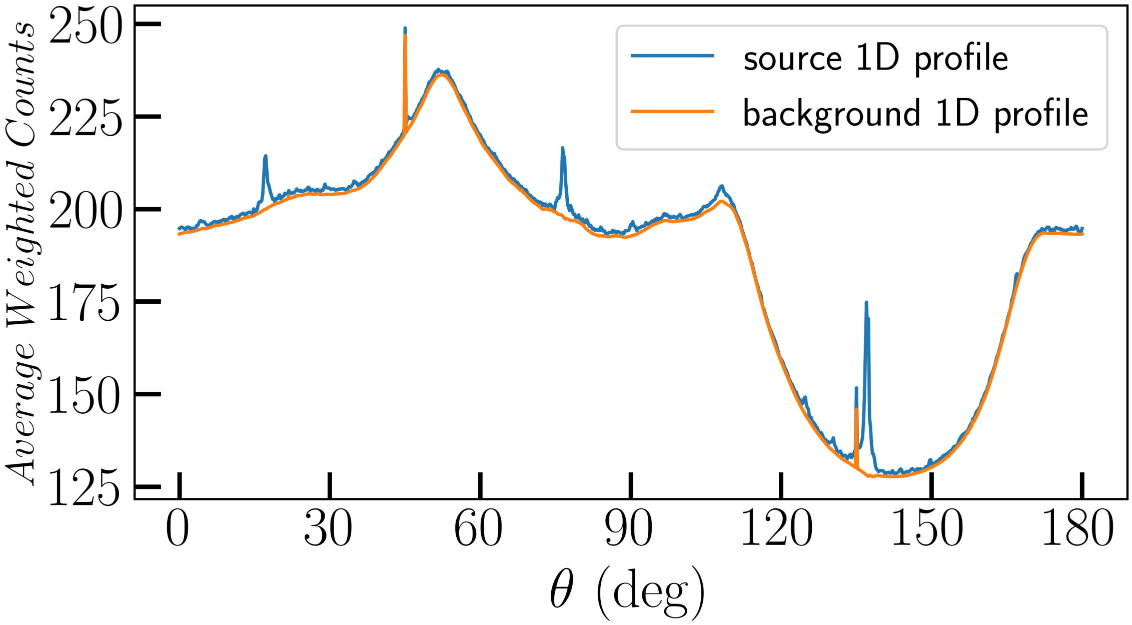


Figure 3.6: Source and background 1D profile of the weighted HT parameter space against position angle θ. The background 1D profile is (slightly) lower than the source 1D profile due to the absence of signal and noise of the component. The peaks seen at 45° and 135° are not from the signal in the image but from the nature of the HT itself.

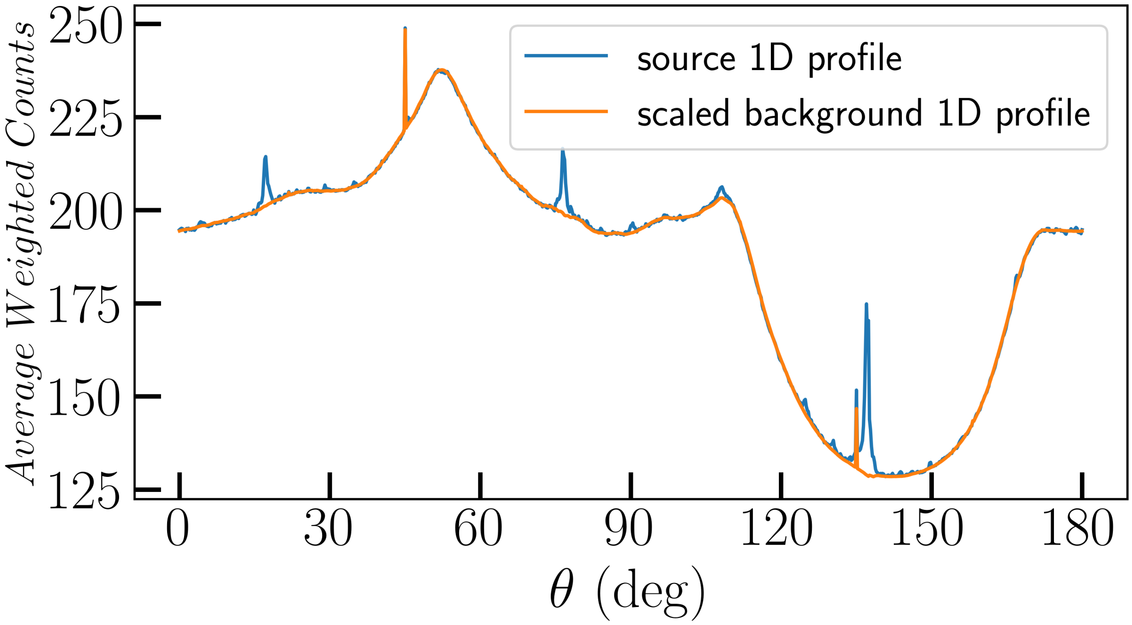


Figure 3.7: Source 1D profile with the scaled background 1D profile of the Hough Transform against position angle θ. The background profile is scaled up to the source profile by adding the mean noise from the source 1D profile.

Algorithm 1 Background 1D Profile Scaling

Given:

$Profile_{src}$ = the 1D profile of the HT of the source region

$Profile_{bkg}$ = the 1D profile of the HT of the background region

RMS() = function to calculate the robust root mean square of given values

AVG() = function to calculate the mean of given values

MDN() = function to calculate the median of given values

```
  $iter \leftarrow 0$
  $noiseLevel \leftarrow 0$
  $Profile_{scaledbkg} \leftarrow Profile_{bkg}$
  **while** $iter < 3$ **do**
    $Profile_{src-bkg} \leftarrow Profile_{src} - Profile_{scaledbkg}$
    $noiseRegime \leftarrow Profile_{src-bkg}$
    $condition \leftarrow True$
    **while** $condition = \text{True}$ **do**
      $thresh \leftarrow 3 \times \text{RMS}(Profile_{src-bkg})$
      $noiseRegime \leftarrow Profile_{src-bkg} < thresh$
      $newNoiseLevel \leftarrow \text{AVG}(noiseRegime)$
      **if** $newNoiseLevel = noiseLevel$ **then**
        $condition \leftarrow \text{False}$
      **else**
        $noiseLevel \leftarrow newNoiseLevel$
      **end if**
    **end while**
    $Profile_{scaledbkg} \leftarrow Profile_{scaledbkg} \times \left(1 + \frac{noiseLevel}{\text{MDN}(Profile_{scaledbkg})}\right)$
    $iter \leftarrow iter + 1$
  **end while**
```

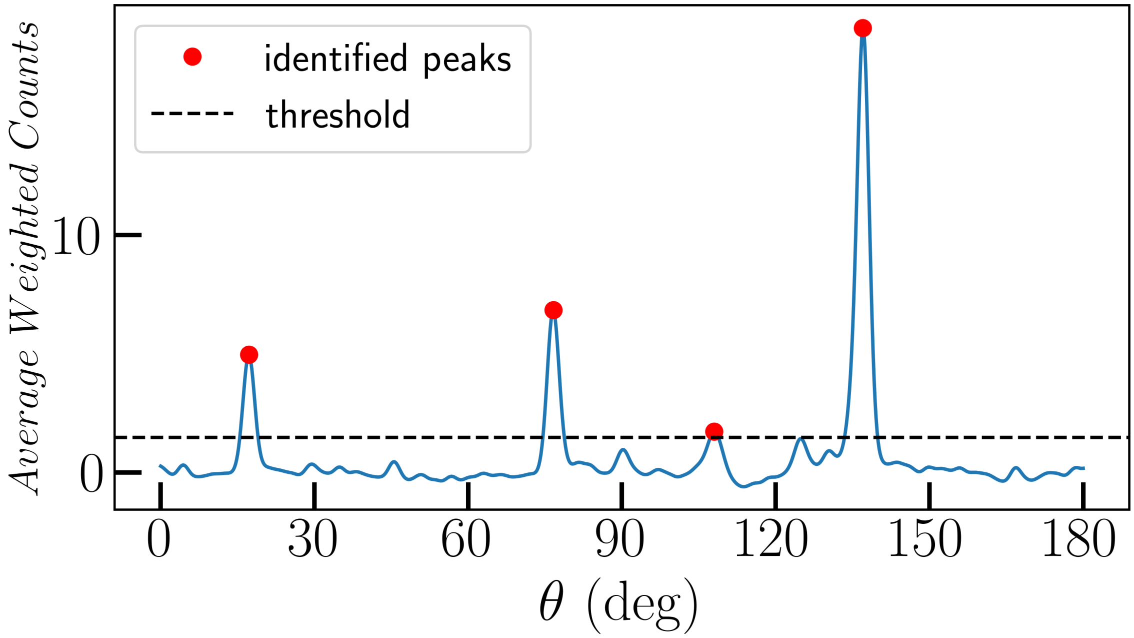


Figure 3.8: Smoothed background-subtracted source 1D profile against position angle θ. All peaks detected above the 5σ horizontal black dotted threshold line are shown as red dots. The position angles at these peaks represent the position angles at which linear artifacts are present around the component.

of all image noise and intrinsic patterns that arise from the HT.

All peaks in the smoothed profile that are $\geq 5\sigma$ threshold are identified as angles of the linear streaks around the component in question. As mentioned in Section 3.1.2, the threshold is made robust by using the MAD. This conservative threshold of 5σ is used to ensure the detection of “apparent” streaks, arising from the pixelated input image, is avoided. Figure 3.8 shows the peaks detected for the previously selected example data.

In cases where there are vertical streaks in the image, peaks are usually identified at both 0° and 180°. To avoid this duplication, the smoothed background-subtracted 1D profile is rolled onto itself during peak detection.

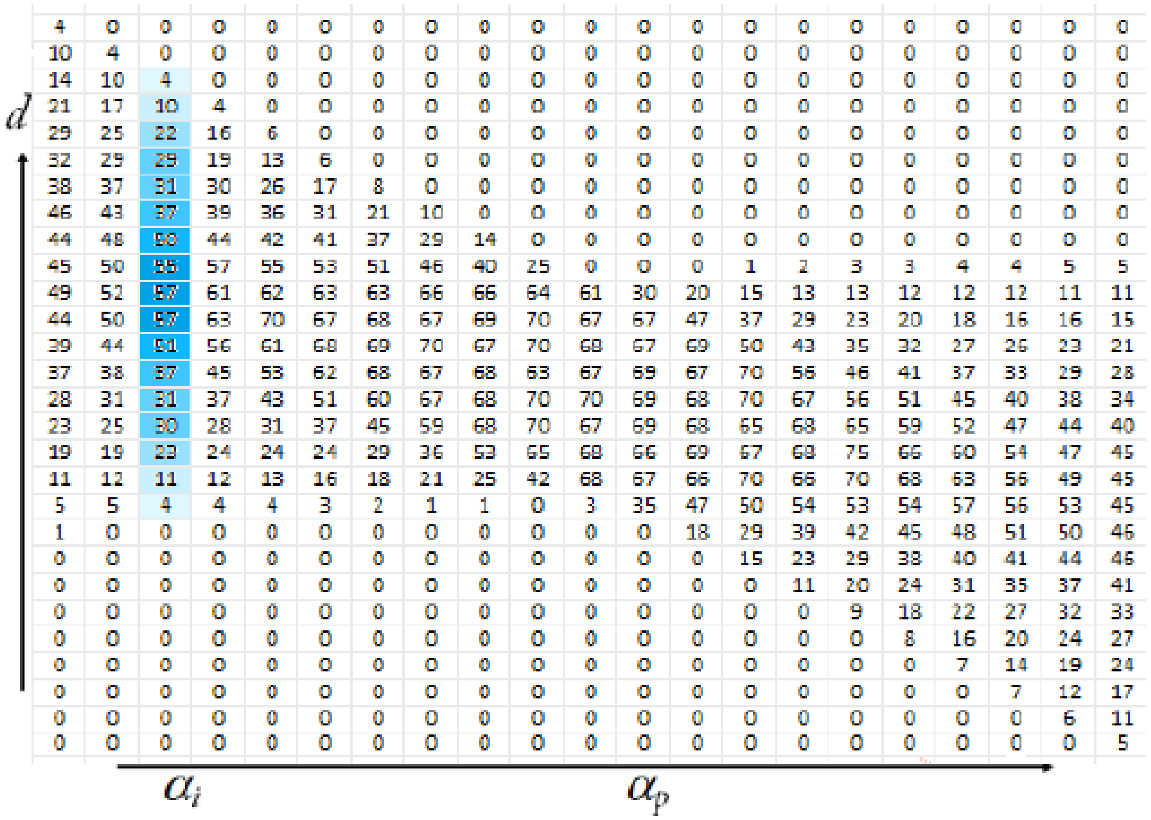
3.2 New Characterization Methods

The linear artifact identification algorithm only extracts the position angles of the streaks in the source region. The quantitative and qualitative properties of the identified linear artifacts are required to classify and characterize the component around which the streaks are found. To do so, the width and surface brightness of the streaks are extracted.

3.2.1 Width Characterization

In the HT parameter space, any of the identified position angle's ($\theta_{\text{peak},i}$) columns show the distribution of the strength of parallel lines at varying distances (along the ρ axis of HT parameter space) from the origin of the image. This distance column can have multiple peaks and/or plateaus indicating the presence of ”strong” parallel lines around the component. Figure 3.9 from Xu et al. (2014) illustrates an example of HT around a position angle θ_i. The gradient of the peaks and/or plateaus represents the spread of the signals at those distances (i.e., the width of the parallel streaks).

Figure 3.10 shows the distance column at an identified position angle $\theta_{\text{peak},i} = 108^\circ$ around the component considered. The FWHM of peaks in the distance column are taken as the widths of these parallel streaks. Again, a robust standard deviation i.e., MAD is used to identify the peaks 3σ above the median of the column. Here, two peaks are detected. Annular regions can sometimes have non-central bright streaks (which do not overlap the component). These streaks can raise the median level of the signal and affect the estimation of the robust standard deviation. To counter for such scenarios, a lower threshold of 3σ (compared to the previously chosen 5σ) is chosen to ensure that the signal from the streaks that are actually overlapping the streaks are also identified.



4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	10	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	17	10	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	25	22	16	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	29	29	19	13	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	37	31	30	26	17	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
46	43	37	39	36	31	21	10	0	0	0	0	0	0	0	0	0	0	0	0	0
44	48	50	44	42	41	37	29	14	0	0	0	0	0	0	0	0	0	0	0	0
45	50	55	57	55	53	51	46	40	25	0	0	0	1	2	3	3	4	4	5	5
49	52	57	61	62	63	63	66	66	64	61	30	20	15	13	13	12	12	12	11	11
44	50	57	63	70	67	68	67	69	70	67	67	47	37	29	23	20	18	16	16	15
39	44	51	56	61	68	69	70	67	70	68	67	69	50	43	35	32	27	26	23	21
37	38	37	45	53	62	68	67	68	63	67	69	67	70	56	46	41	37	33	29	28
28	31	31	37	43	51	60	67	68	70	70	69	68	70	67	56	51	45	40	38	34
23	25	30	28	31	37	45	59	68	70	67	69	68	65	68	65	59	52	47	44	40
19	19	23	24	24	24	29	36	53	65	68	66	69	67	68	75	66	60	54	47	45
11	12	11	12	13	16	18	21	25	42	68	67	66	70	66	70	68	63	56	49	45
5	5	4	4	4	3	2	1	1	0	3	35	47	50	54	53	54	57	56	53	45
1	0	0	0	0	0	0	0	0	0	0	0	18	29	39	42	45	48	51	50	46
0	0	0	0	0	0	0	0	0	0	0	0	0	15	23	29	38	40	41	44	46
0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	20	24	31	35	37	41
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	18	22	27	32	33
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	16	20	24	27
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	14	19	24
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	12	17
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	11
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5

Figure 3.9: This example figure of a Hough Transform parameter space is taken from Xu et al. (2014). Here d is the distance parameter ρ and α is the position angle parameter θ. The cell values are the counts at each (α, d). The highlighted blue gradient shows the values peaking at location of the streak/line and decreasing away from the peak.

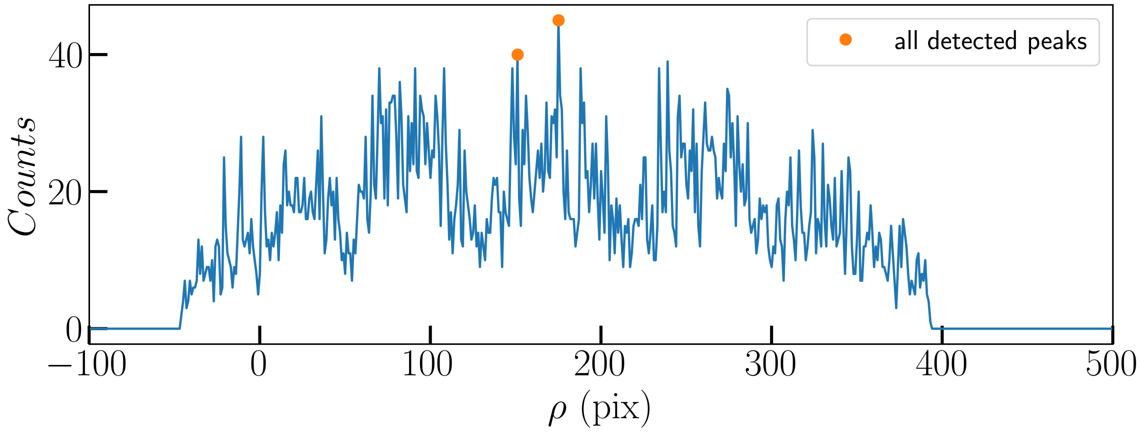


Figure 3.10: A single column in the Hough Transform parameter space at one of the position angles of the identified streak at $\theta_{\text{peak},i} = 108°$ against the distance parameter ρ. The two peaks identified indicate the distances to the potential streaks from the image origin.

3.2.2 Component & Streak Matching

Applying the HT on the annular regions around a component causes the signal from streaks that do not overlap with the component to be included in the parameter space. The spatial information of the recognized streaks are lost since the new linear artifacts detection method collapses the HT to a 1D position-angle profile of the input image to boost the signal strength of the streaks. This implies that the new algorithm is designed so that it could identify streaks that have no effect on the component since they are offset from the component. Since a brighter streak can be next to the component without overlapping with the component, choosing the highest peak for the width measurement is not an option.

To ensure that only streaks that overlap with the component are identified, a two step verification is applied when measuring the width of the streaks.

1. For every streak identified around a component, only a section of the distance column around the component's ρ_{comp} is considered during the width measurement. ρ_{comp} is calculated using Equation (2.2) with x and y as the image coordinates of the component and θ as the identified angle under considera-

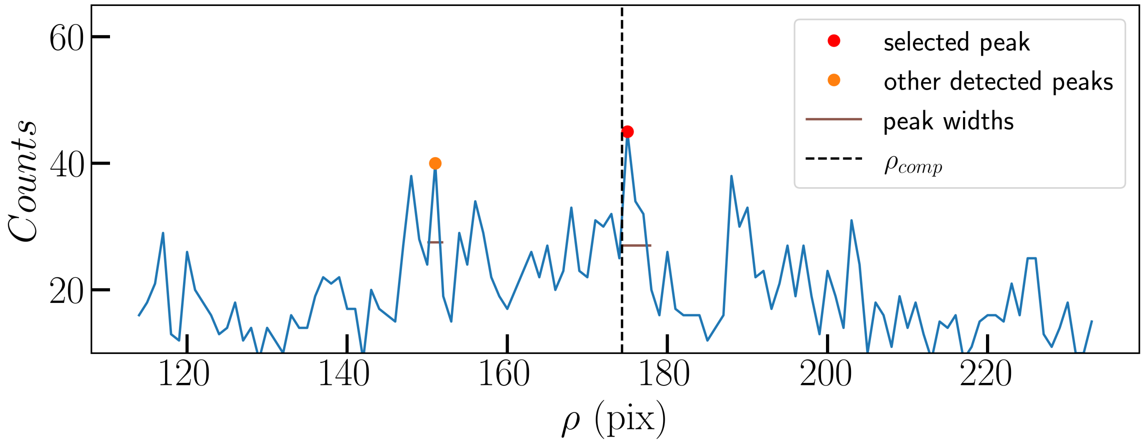


Figure 3.11: The Hough Transform parameter space at the position angle of an identified streak at $\theta_{\mathrm{peak},i} = 108°$ against the distance parameter ρ. The orange peak is a potential streak detected at an offset from the component and has no effects on the component whereas the red peak represents a streak overlapping with the component. The brown horizontal lines at the peaks shows the width of these peaks.

tion. Through trial and error, the size of subsection of the distance column considered is chosen as 120 cells/pixels centred around ρ_{comp}. This ensures that the threshold calculated for peak detection during width calculation is accurate but at the same time also removes signal from streaks that are offset from the component yet are in the annular region.

2. Of all the peaks and their respective widths identified in the subsection of the distance column, the highest peak that has a FWHM (i.e., width) that overlaps or grazes the component's ellipse is chosen. This step makes sure that only the streaks that are affecting the component are considered.

Figure 3.11 illustrates the subsection region centered around ρ_{comp} in the distance column. The peak shown in red is chosen as the width of the streak as it is seen to be going over the ρ_{comp} marked with the dotted black line. Other potential peaks that were detected are marked in orange. Algorithm 2 describes the verification steps in detail. Streaks that are detected in the linear artifacts detection method but found to not overlap with the component during the width characterization step are filtered out.

Algorithm 2 Streak Filtering

```
Given:
$\rho_{comp}$ = distance of the component from image origin.
$distCol$ = $\rho$ column of HT at an position angle $\theta_i$ of streak identified
  for every streak $i$ identified around a component do
    $srcCenteredDistCol \leftarrow distCol[\rho_{\text{comp}} - 60 : \rho_{\text{comp}} + 60]$
    identify peaks $\rho_i$ in $srcCenteredDistCol$ and measure their widths $wid_i$
    for every peak $\rho_i$ in $srcCenteredDistCol$ do
      if $\rho_i$ - $wid_i/2$ < ellipse of component < $\rho_i$ + $wid_i/2$ then
        accept streak at $\theta_{\text{comp}}$ with width $w_i$
      else
        reject identified streak at $\theta_{\text{comp}}$
      end if
    end for
  end for
```

3.2.3 Quantifying the Streak Surface Brightness

Assessing the strength of streaks (i.e., the surface brightness of the streaks) is essential to analyze how much of a component's flux density is due to overlapping streaks. This helps in the accurate measurement of the component's flux density.

The values of the pixels in VLASS QL images are the flux density surface brightness given in Jy/bm (Jansky per beam). To calculate the surface brightness of the identified streaks, the image is first converted to units of Jy/pix using

$$\text{Jy/pix} = \frac{\text{Jy/bm}}{\Omega_b}, \tag{3.4}$$

where Ω_b is the area of a Gaussian synthesized beam (known as the beam volume) given by

$$\Omega_b = \frac{\pi}{4\ln 2} \times \Theta_{\text{maj}} \times \Theta_{\text{min}}, \tag{3.5}$$

where Θ_{maj} and Θ_{min} are the major and minor axis of the synthesis beam, respectively. The conversion is to make sure that the surface brightness of the streaks are calculated over the area of the streak; and not over the beam per area of the streak.

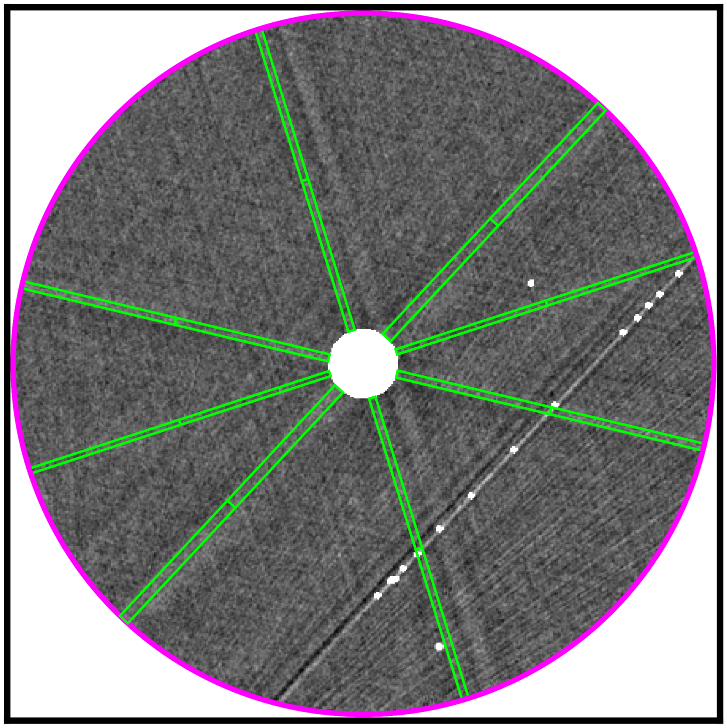


Figure 3.12: The four sections of all the identified streaks around a component is shown in green. For each of the identified streaks, aperture photometry is performed to extract a robust surface brightness of the streak to determine the strength and the effects of the artifacts on the component.

The surface brightness of the streak is calculated by performing aperture photometry on the Jy/pix values from the VLASS QL image. The width of the streak is taken as the width calculated during the streak width characterization and the length of the streak is taken as twice the radius of the annular region. Sometimes, linear artifacts are not uniformly bright around the component. To measure a more accurate surface brightness of the streak, the streak region is divided into 4 sections. An example of a streak divided into 4 sections is shown in Figure 3.12. Aperture photometry is performed separately on each of the 4 sections to measure their surface brightnesses. The median of the surface brightnesses of the sections is taken as the surface brightness of the streak.

A part of the signal from the streak region can be due to background noise that is not associated with the streak. Therefore the RMS of the annular region that is not affected by streaks is measured to quantify this noise. The measured background RMS is used to calculate the signal-to-noise ratio of the streak and therefore the extent of effects the streak has on the flux density of the component.

Chapter 4

Application of New Methods to VLASS Epoch 1

A random subset of VLASS QL images from Epoch 1 was used to develop the new artifact detection technique and tune the selection of parameters (like the size of the annular region considered), However, this subset only consisted of ≈ 1000 VLASS QL images and their components i.e., about 1.5% of the Epoch 1 data. In this chapter we apply the new technique to the entire VLASS Epoch 1 data-set[1]. We first discuss the steps taken to confirm the streak detection. Then we detail how the confirmed streaks were used to re-evaluate the properties of components in VLASS QL 1 images. Finally, we compare our method of identifying artifacts to the two methods already in the component catalogue.

4.1 Streak Confirmation

The new artifacts detection algorithm identifies ∼1.4 streaks around each component on an average. While 17% (574,193) of components have no linear artifacts overlapping them, multiple streaks including the sidelobe streaks are detected around 83% (2,807,084) of components (including a majority of the brightest components). The new method is also very sensitive to relatively low surface brightness. Therefore, the absence of above-detection-threshold streaks around some sources could be due to

[1] The component catalogue for VLASS QL images in Epoch 2, which contains critical input data for that epoch, has not undergone peer review at the time of this thesis

accurate de-convolution resulting in almost no residuals.

Gordon et al. (2021) initially identified about 8% of components potentially having artifacts based on *Peak_to_Ring* < 2 and nearest neighbour $> 20''$ criterion. Therefore, we selected a subset of the detected streaks ($\approx$ 1%) to perform visual inspection, verify the presence of streaks, and manually check the impact on the components they overlap. With these inspections, we developed and performed additional analysis (for the full sample of streaks) to quantify the candidate streaks' properties and verify that the detected streaks are valid and statistically well understood for future analysis that removes their effect on the overlapping component.

As previously mentioned in Section 3.2.3, each identified streak is divided into four sections, and the total flux density and angular extent of each quadrant is measured. If a quadrant has N beams, the total flux density of each quadrant, f_{q_i}, is:

$$f_{q_i} \pm \delta f_{q_i} = f_{\text{beam},i} \cdot N \pm \sqrt{N} \cdot \delta f_{\text{beam},i}, \tag{4.1}$$

where f_{q_i} is the total flux density of the i-th quadrant and $f_{\text{beam},i}$ is the flux density of the beams in the i-th quadrant.

The uncertainty in the flux density for a single beam ($\delta f_{\text{beam},i}$) can be approximated as the RMS of the background local annular region since the noise level in the local region is assumed to be uniform:

$$\delta f_{\text{beam},i} = \text{RMS}_{\text{local}} \tag{4.2}$$

Therefore, the surface brightness of each quadrant is,

$$\begin{aligned} \mu_{q_i} \pm \delta\mu_{q_i} &= \frac{f_{q_i}}{A_{q_i}} \pm \frac{\sqrt{N} \cdot \text{RMS}_{\text{local}}}{A_{q_i}} \\ &= \frac{f_{q_i}}{A_{q_i}} \pm \frac{\text{RMS}_{\text{local}}}{\sqrt{N} \cdot \Omega_{\text{beam}}}, \end{aligned} \tag{4.3}$$

where A_{q_i} is the area of each quadrant in square pixels and Ω_{beam} is the synthesized beam size in square pixels. With the surface brightness of the sections already calculated in the streak detection algorithm, the uncertainty of the surface brightness of the sections are calculated using Equation (4.3).

The streaks are split into sections to verify whether the streaks have uniform surface brightness in the considered local annular region. Furthermore, splitting of a streak makes sure that the strength of the streak is not biased by the surface brightness of any particular quadrant(s). Here, we have chosen to divide each streak into four sections to have the minimum number of sections needed for unbiased estimation of the surface brightness, while maintaining symmetry around the component. Next, the inverse-variance weighting of the surface brightness of the sections is taken as the surface brightness of the whole streak, $\mu_{\rm streak}$.

$$\mu_{\rm streak} \pm \delta\mu_{\rm streak} = \frac{\sum \frac{\mu_{q_i}}{\delta{\mu_{q_i}}^2}}{\sum \frac{1}{\delta{\mu_{q_i}}^2}} \pm \left[\frac{1}{\sum \frac{1}{\delta{\mu_{q_i}}^2}}\right]^{\frac{1}{2}}. \qquad (4.4)$$

Masking the components from the Gordon et al. (2021) catalogue can reduce the considered area of the any of the four sections, thus changing the error in surface brightness of each section. The inverse-variance weighting rectifies the effect these different uncertainties have on the surface brightness of the streak.

Here the surface brightness of the sections are calculated as flux density per unit pixel. Then the surface brightness is converted to flux density per beam

$$\mu_{\rm streak}[\mathrm{Jy/bm}] = \mu_{\rm streak}[\mathrm{Jy/pix}] \times \ \Omega_{\rm beam}[\mathrm{pix/beam}] \qquad (4.5)$$

to match the units with the units of intensity given for each pixel in the QL images. Here $\mu_{\rm streak}$ is taken as the uniform surface brightness of the entire streak, at least in the considered local annular region.

A histogram of the SNR of the surface brightness of all the streaks identified in VLASS Epoch 1 QL images is shown in Figure 4.1. The peak of the histogram (shown as a solid back line) represents the approximate completeness limit of the new algorithm (i.e., not all streaks below an SNR of 30 have been detected by the new algorithm). This could be due to high or non-uniform noise levels around components, which can lead to peaks in 1D profile (discussed in Section 3.1.2) to be below

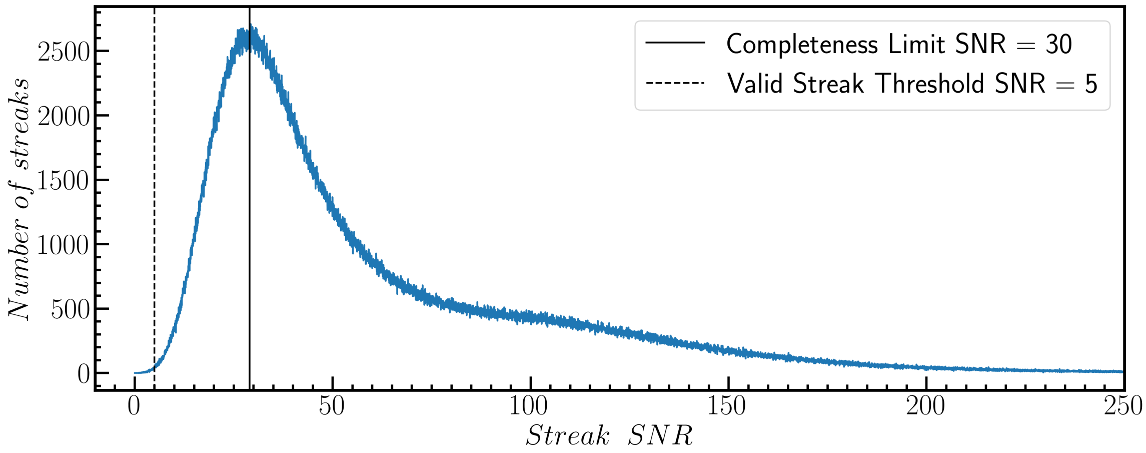


Figure 4.1: Histogram of the SNR level of the all the identified streaks in VLASS Epoch 1 QL images. The histogram peaks at SNR of ∼30, which indicates the likely completeness limit of our new artifact-identification method (i.e. some streaks below SNR of 30 are not identified by the new method). However, streaks with SNR below 30 are less likely to strongly affect component characterization.

the set threshold. In addition, we note that since a typical quadrant size is $\approx$ 25 beams, the uncertainty in the average streak surface brightness is usually known at the $\mathrm{RMS}_{\mathrm{local}}/(4 \cdot 25)^{1/2}$ level. Thus, missing streaks with SNR below 30 will have little effect on the peak flux density of components, especially for components above 1 mJy/beam (recall a typical RMS is $\sim 0.15\,\mathrm{mJy/beam}$). Therefore, although the components' peak flux density might be slightly (almost negligibly) overestimated, the absence of these missing streaks will not lead to mis-classification of components as imaging artifacts.

4.1.1 Streak Selection Based on Application

The identified streaks can be used: as an additional quality assurance step while the QL images are generated and; for component brightness estimation and classification. Based on the use case, a subset of the identified streaks is considered for further analysis.

Quality Assurance

The detection of the streaks indicates inadequate `clean`-ing on an image as a whole. Although all streaks cannot be perfectly removed from the images, a few additional checks can be added to ensure that images are `clean`-ed as much as possible. While QL-image generation will not be changed for the final epoch of VLASS, most of VLASS Single Epoch continuum images are still being generated. The quality assurance modes and steps we discuss below could be integrated into that process, or used more generally for snapshot VLA observations.

There are two modes one can envision for quality assurance: in one mode, source detection is run and the full extended HT analysis described in this thesis is performed; this mode is considered below. In the second mode, the extended HT analysis described in this thesis is adjusted for component-blind analysis. This could include treating the entire image as the "src" region and forgoing the component and streak matching step. The second mode will be more computationally expensive, but may provide superior artifact detection.

During image generation, the following quantities can be checked after a suitable round of `clean`-ing .

- The total number of streaks identified in each QL image. Based on a set threshold of acceptable number of streaks in an image, to drive decisions on the need for more iterations of de-convolution.

- The change in the total number of streaks identified in each QL image. If this difference is negligible even after 3 consecutive iterations, the de-convolution step can be stopped.

- The median surface brightness of the identified streaks in an image. Again, based on a set acceptable threshold for the surface brightness of the streaks, more iterations of `clean`-ing can be performed.

- The change in the total surface brightness of all streaks between each iteration of `clean`. If this difference is negligible even after 3 consecutive iterations, the de-convolution step can be stopped.

When considering the numbers of streaks, one would also have to set a surface brightness density threshold (based on the RMS of the image) above which to "count" a streak. This criterion could be optional when considering the surface brightness of streaks. It is to be noted that above steps should only be considered as suggestions. Fine tuning of these steps is beyond the scope of this thesis.

Component Brightness Estimation and Classification

To check whether the component is a real source or not, and to correctly estimate the flux densities of those real sources, a more stringent streak selection criterion is needed.This is because streaks with larger surface brightness uncertainties can negatively affect the estimation (with good accuracy) of the peak flux density of just the component. Furthermore, this selection ensures that only streaks that are significantly affecting the components are considered during components' brightness estimation. Therefore only streaks with SNR ≥ 5 are considered (i.e. not all the identified streaks will be considered for this analysis) for the calculation of the collective effect of the streaks on a component. We consider streaks that fail this criteria as both too poorly quantified and to likely have little effect on components. With this threshold, of the total of 4,733,249 streaks identified around 3,381,277 components, only 1,548 streaks are below the 5σ limit. These low-SNR streaks are found in only 1,500 Subtiles. Figure 4.2 shows an example of identified streaks that are below the threshold.

4.2 Component Re-evaluation

Once the streaks are identified and quantified, the effect of the overlapping streaks on the components is removed. Based on the corrected flux density of the component,

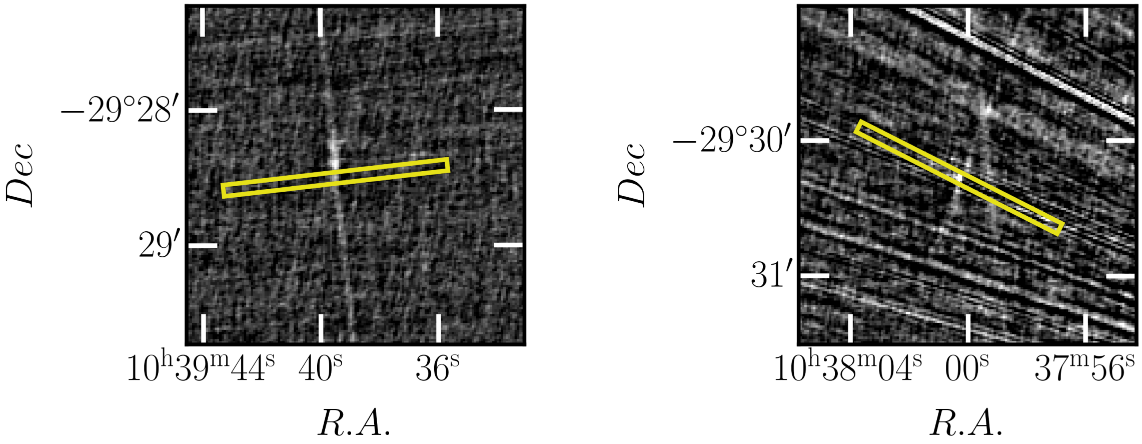


Figure 4.2: Examples of streaks identified around sources/components that have a signal-to-noise ratio of $\leq 5\sigma$. These detections are either results of a nearby streak in the local annular region that does not overlap the component (*left panel*), or around very bright sources that produce lots of streaks (*right panel*). Streaks below 5σ limit are not considered in re-evaluating component properties under the new method.

the components are then classified as either likely being a real astrophysical source or an imaging artifact.

4.2.1 Removing Effects of Streaks' Surface Brightnesses

To remove the effects of the streaks on any component, the components' peak flux densities already available in the Gordon et al. (2021) component catalogue are considered. The reasons for considering the peak flux density instead of the total flux density of the components are threefold:

- Transient sources, the major scientific motivation for this thesis, must be unresolved point sources (due to light-travel time considerations) and are statistically unlikely to be embedded in an extended source. In (noise-free) point sources, if the peak flux density is X mJy/beam, then the total flux density is X mJy.

- The streaks are assumed to have uniform surface brightness, which in these cases are the calculated surface brightness flux densities. Therefore the calculated surface brightness flux density of a streak is also its peak flux density.

- The Gordon et al. (2021) catalogue does not have the total flux calculated for all the identified components. This is mainly because PyBDSF was not able to fit Gaussians to the components and hence the dimensional parameters like the semi-major and minor axes were not estimated.

For a component with at least one overlapping streak, the peak flux density can be visualized as shown in Figure 4.3. The flux density Gaussian curve of an unresolved component is increased based on the (average) flux density of each streak that is overlapping the component. This increases the observed peak flux density of the component. Conversely, the peak flux density of components that have no identified overlapping streaks or have streaks that are not overlapping the brightness region of the component are unaffected.

To determine the unaffected peak flux density of a component, the flux density of the overlapping streak(s) is removed from the previous estimations. The algorithm describing the subtraction is shown in Algorithm 3. As previously mentioned in Section 4.1, only streaks that have a SNR ≥ 5 are considered for component peak brightness correction. Of these valid streaks, only streaks that are overlapping more than half of the component (i.e., the center of the component + 1 pixel) are considered as streaks that are affecting the component. Here the additional pixel is added as a safety measure to make sure that the centre of the component (i.e., the peak flux density pixel of the component), is in fact fully covered by the streak. As seen in Figure 4.3, these are the streaks that increase the previously estimated peak flux density of a component.

In Algorithm 3, this selection is implemented by only considering the streaks that are less than their width + 1 pixel distance away from the center of the component. The centres of the streak obtained during the artifact's detection (x_{streak}, y_{streak}), and the component (x_{comp}, y_{comp}) are converted to distance from the origin of the image

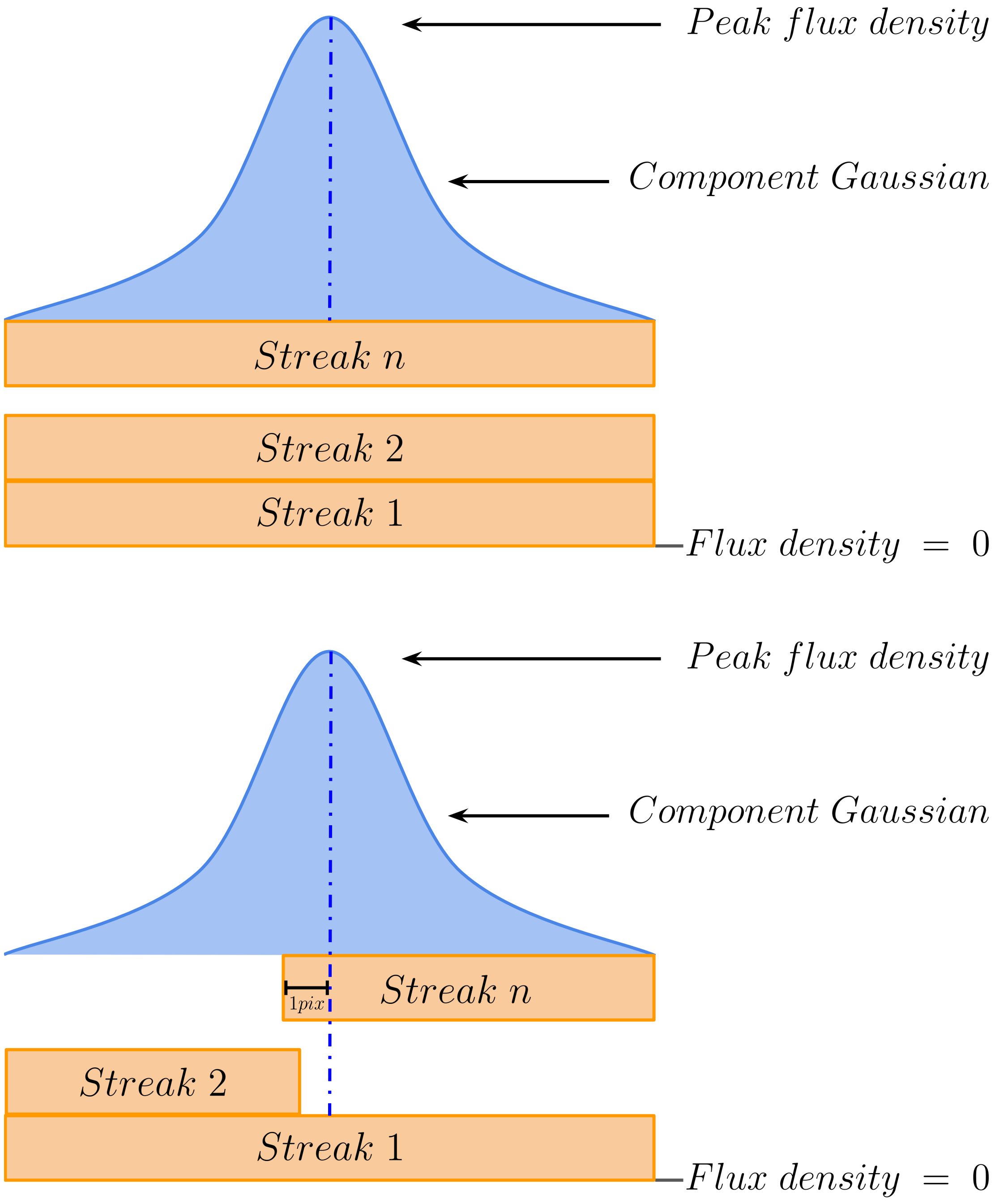


Figure 4.3: Graphical 1D representations of the flux density around a source/component. PyBDSF fits a Gaussian to the component shown by the blue curve. *Upper:* shows the enhancement of the detected peak flux density of the component by n streaks. *Lower:* shows the enhancement of the peak flux density of the component by the 1^{st} and n^{th} streak, whereas the the 2^{nd} streak is not affecting the peak flux density of the component.

Algorithm 3 Total Peak Flux Density of Valid Streaks

```
Given:
$S_{streak,i}$, $w_{streak,i}$ = Surface brightness and width of each overlapping streak.
$\rho_{streak,i}$, $\rho_{comp}$ = HT distance equivalent of the center of the streak and component
    from origin of the image at the streak angle $\theta_{streak}$.
$S_{allStreaks}$ = Sum of surface brightness of all selected streaks.
  for all components in catalogue do
    $S_{allStreaks} \leftarrow 0$
    $\delta S_{allStreaks} \leftarrow 0$
    if $\frac{S_{streak,i}}{\delta S_{streak,i}} \geq 5$ then
      if $|\rho_{streak,i} - \rho_{comp}| < w_{streak,i}+1$ then
        $S_{allStreaks} \leftarrow S_{allStreaks} + S_{streak,i}$
        $\delta S_{allStreaks} \leftarrow \sqrt{S^2_{allStreaks} + S^2_{streak,i}}$
      end if
    end if
  end for
```

ρ_{streak} and ρ_{comp} respectively, at the considered streak angle θ_{streak}.

$$\begin{aligned}\rho_{\text{streak}} &= x_{\text{streak}} \cdot \cos\theta_{\text{streak}} + y_{\text{streak}} \cdot \cos\theta_{\text{streak}} \\ \rho_{\text{comp}} &= x_{\text{comp}} \cdot \cos\theta_{\text{streak}} + y_{\text{comp}} \cdot \cos\theta_{\text{streak}}\end{aligned} \tag{4.6}$$

This distance-equivalent value is considered instead of the Euclidean distance to maintain consistency with the HT's distance parameter ρ. Finally, the flux densities of all these selected streaks are summed to obtain the total effect of the streaks on the component $\mu_{\text{allStreaks}}$. The detected peak flux density of the component, f_{comp}, is then corrected by:

$$\begin{aligned}f_{\text{corr,comp}} &= f_{\text{comp}} - \mu_{\text{allStreaks}}[\text{Jy/bm}] \text{ and} \\ \delta f_{\text{corr,comp}} &= \sqrt{\delta f_{\text{comp}}{}^2 + \delta\mu_{\text{allStreaks}}{}^2},\end{aligned} \tag{4.7}$$

subtracting the streak surface brightness (in Jy/bm). The corrected value $f_{\text{corr,comp}}$ is the estimated peak flux density from only the component.

4.2.2 Component Classification

Through visual inspection, we noted that many potential imaging artifacts are local maxima on individual streaks (likely sidelobe streaks from nearby a component/-source) or at the intersection of two or more streaks, most of which are originating from nearby components. Such examples of potential imaging artifacts are shown in Figure 4.4. Additionally, if the component is a real source in the *uv*-plane, incomplete cleaning of a point source convolved with the dirty beam would result in hexagonal pattern residuals (three streaks) around the component. Unless the images are perfectly de-convolved leaving no residuals, which is usually unlikely for QL images, most real astrophysical sources will have traces of sidelobe streaks around them. The absence of overlapping sidelobe streaks or the comparable flux densities of these streaks to the peak flux density of the component can both suggest that a component is probably an imaging artifact. Inferring from this, the classification of the components is based on the signal-to-noise ratio of the peak flux density of component before and after streak subtraction, the number of streaks overlapping the component, the number of sidelobe-predicted streaks overlapping the component, and the strength of the peak flux density of the component compared to the brightness of all the overlapping streaks combined.

The classification follows the flowchart/decision tree shown in Figure 4.5. The components are given a binary flag based on the results at each decision node (shown as rounded rectangles). The set of binary flags collected at all the decision nodes determines the classification category (shown as oval leaf nodes) of the component. The procedure for classification is as follows:

1. The components for which the peak flux density SNR ≤ 3 are not classified (flag $= 0$; *no classification*). This is because such components do not have enough signal strength to know whether they are valid components and hence we cannot further classify them accurately. The remaining components with SNR > 3 are

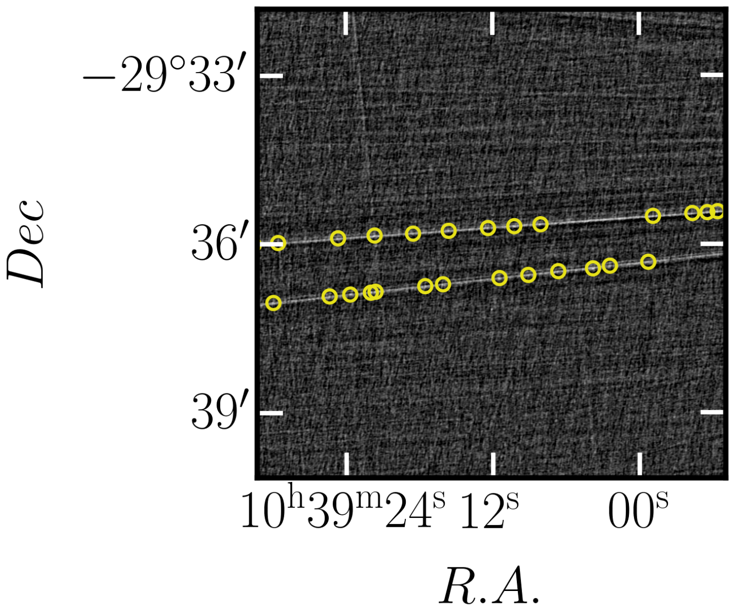


Figure 4.4: Cutout of the VLASS QL image highlighting false-positive imaging artifacts detected as components/sources by PyBDSF. Many of these false-positives are local maxima on a prominent streak originating from a nearby bright source.

analyzed for classification (flag = 1).

In the Gordon et al. (2021) catalogue, some components do not have an estimate of the error in peak flux density, δf_{comp}. To rectify this, we assume that the error is the target RMS noise level per VLASS epoch (i.e., 120 μJy/bm). The same 3σ threshold is considered for these components (flag = 1), with low SNR components being given no classification (flag = 0; again with the classification label as *no classification*).

2. For components that do not have any streaks (after the streak selection described in Section 4.2.1), the previous estimation of peak flux density remains the same. Since our classification is based on the identified streaks, such components are omitted from our classification process (flag = 0) and are classified as *unknown*. Components with streaks as further analyzed (flag = 1).

3. After the peak flux density correction of the components, if the SNR of the corrected peak flux density < 3, it implies that the previously estimated peak

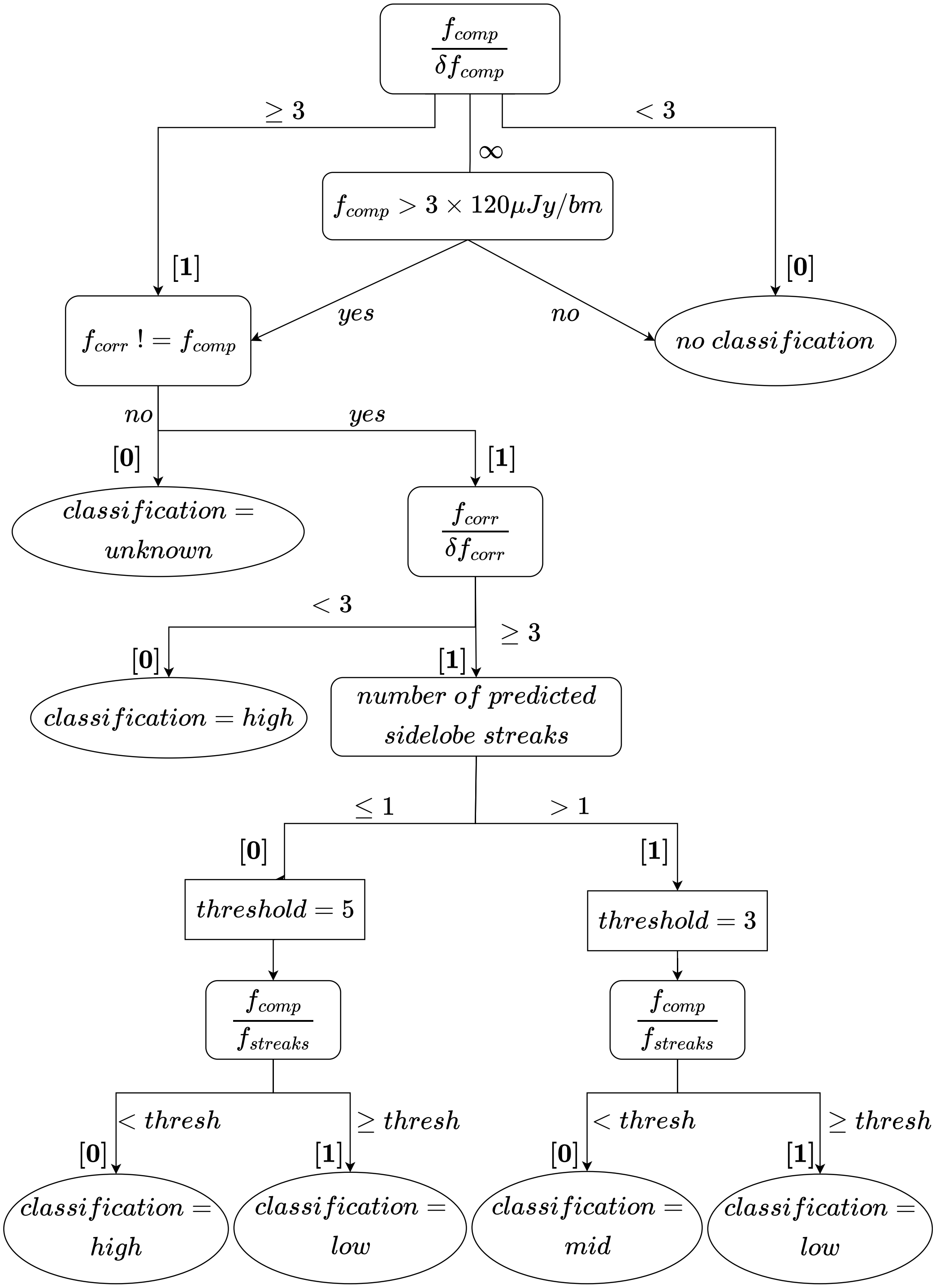


Figure 4.5: Component classification flowchart. At each decision node the components are given a binary flag. Based on the flags, the components are given a classification tag of either no classification and unknown, or one of the three probability ratings that a component arises from a linear streak artifact: low, mid and high.

flux density mostly arose from the overlapping streak. The low signal strength from only the component (i.e., after subtraction), suggests that the component is likely to be an artifact (flag = 0). Such components are given the artifact probability rating of *high*. If the SNR ≥ 3, the strength of the signal is mostly from the component itself and therefore such components are considered for further analysis (flag = 1).

4. As previously mentioned in Section 4.2.2, the presence of sidelobe streaks around components suggests that it is a real astrophysical source. Therefore components are separated into two classes based on the number of overlapping streaks that are at the angles of the predicted sidelobe streaks.

   Around the components, the angles of the detected sidelobe streaks might be at slightly different angles from that of the predicted sidelobes because of various reasons like interference of signal from nearby component(s), overlapping of streaks from other sources at different angles, etc.. Therefore, while counting the number of sidelobe streaks, streaks that are within 5° of any of the three predicted sidelobe streaks are considered as being "predicted".

   If the number of predicted sidelobe streaks ≤ 1, the threshold for further SNR comparison and analysis is set to 5 (flag = 0) otherwise the threshold is set to 3 (flag = 1). The change in threshold for these two classes is to make sure that the components with no sidelobe streaks are in fact real astrophysical sources instead of a local maxima on a single bright streak.

5. Finally the peak flux density of the component is compared with the total flux density of all affecting streaks (SNR). If the SNR is $\geq$ the threshold, then the probability of such components being imaging artifacts is low and hence are given a artifact probability rating of *low* (flag = 1).

   For components that have a threshold of 5 i.e., 1 or less overlapping sidelobe streaks, if the SNR is below the threshold, the components are given an artifact probability rating of *high* (flag = 0) indicating that the components are likely imaging artifacts. Whereas for components with a threshold of 3 i.e. more than 1 overlapping sidelobe streaks, if the SNR is below the threshold, the components are given an artifact probability rating of *mid* (flag = 0) implying that even though the SNR is low, the component might be a real astrophysical source. Such components can be omitted or included in the selection based on the user's requirements.

In total, the components are classified into 5 categories (with binary decision flags given in brackets):

- ***no classification [0]:*** Components that have a SNR < 3. These components are not given a probability rating.
- ***unknown [10]:*** Components for which no overlapping streaks were detected. No probability rating is given since the classification is based on the identified streaks.
- ***high [110] or [11100]:*** Components that are most likely to be imaging artifacts.
- ***low [11111] or [11101]:*** Components that are most likely to be real astrophysical sources.
- ***mid [11110]:*** Components that have the probability of being an artifact between the *low* and *high* categories.

Out of the 3,381,277 components from Gordon et al. (2021), we only identify 988 components with no classification flag (0.029% of all components). We find that only 616,497 components contain no overlapping streaks (18.23% of all components); this strongly suggests that the incomplete identification of streaks with surface brightness

SNR of 5–30 is not a strong concern. We identify 850,718 components (25.16% of all components) as having a "high" likelihood of arising from linear streak artifacts. Of these highly likely imaging artifact components, 73,133 have [110] flag (2.16% of all components) and 777,585 have [11100] flag (23.00% of all components). Since we only identify 47,025 components (1.39% of all components) as having a "mid" probability of arising from linear streak artifacts, it makes relatively little difference whether users include or exclude such components in their analysis. Finally, we find that 1,866,049 components (55.19% of all components) have a "low" probability of arising from linear streak artifacts. 646,559 of these components have the [11111] flag (19.12% of all components) while 1,219,490 of them have the [11101] flag (36.07%).

The five categories of the components in a $f_{\rm comp}$ – $f_{\rm corr,comp}$ plane are shown in Figure 4.6. The diagonal black line represents the ideal scenario where no streaks are found in the images and therefore the peak flux density of the components before and after streak subtraction stays the same. In the not ideal scenario, (i.e., in QL images), the diagonal line also traces the components around which no above-the-threshold streaks were identified (i.e., components for which the new method fails to identifies components due to the reasons discussed in Section 4.1). The horizontal and vertical black lines represents an SNR of 3 (threshold above which the components are considered for the 3 artifact probability rating). As the components deviate from the diagonal line (i.e., as the overlapping streaks' strength increases with respect to the measure peak flux density of the components), components are more likely to be classified as "high" or "mid". The few outliers of "unknown" (shown in blue) (i.e., the components for which overlapping streaks are not detected) lie below the diagonal line due to their missing peak flux density uncertainty. For such components, the local RMS level $\mathrm{RMS}_{\rm local}$ is substituted as their peak flux density uncertainty, which changes their SNR. The outliers seen in the "low" components population are mostly components that have been heavily effected by a streaks or streaks; but some components for which the overlapping streaks' combined strength is under-estimated

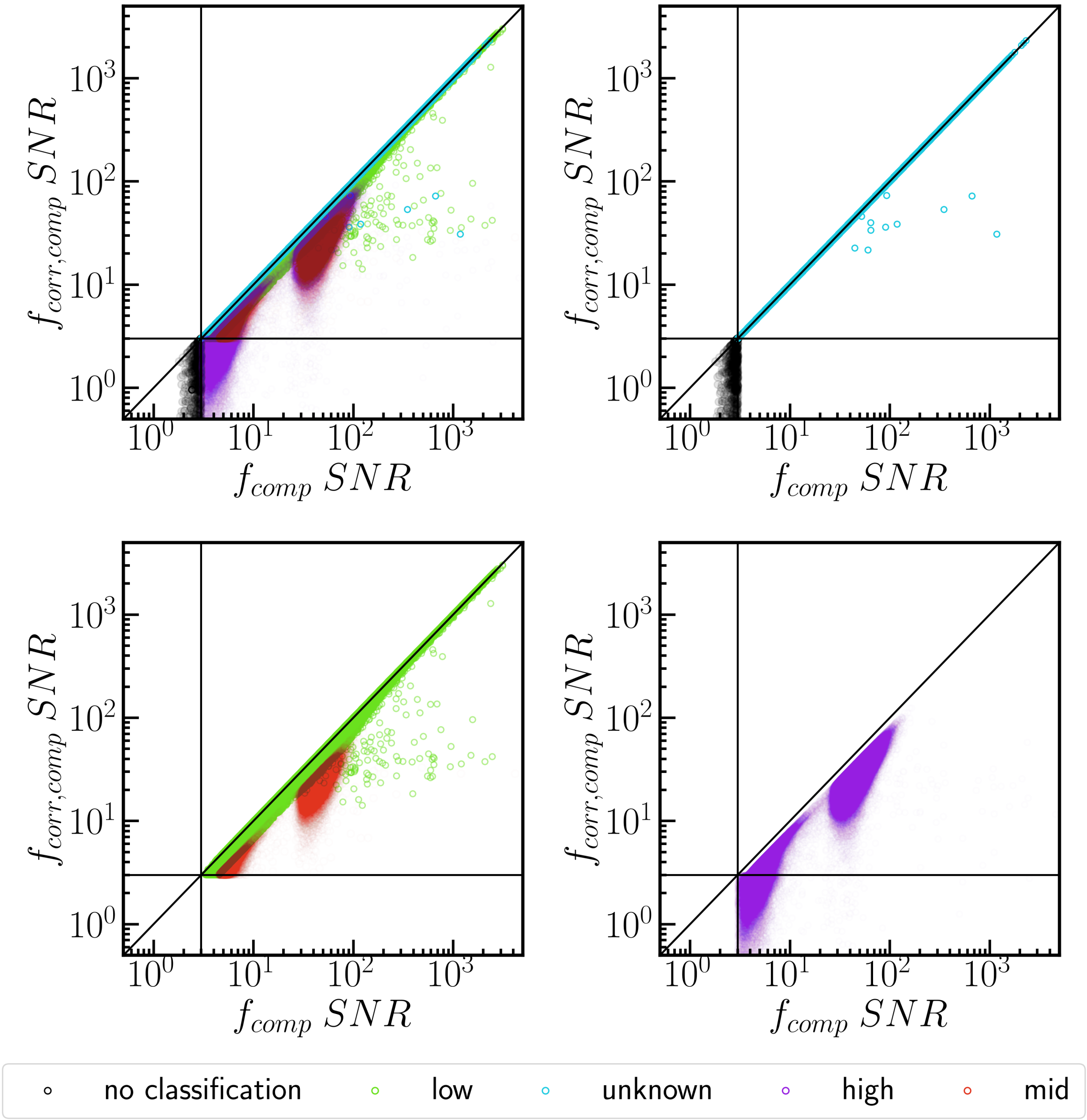


Figure 4.6: Comparison of the SNR of the 5 categories of components before and after streak subtraction. *top left panel* shows all 5 categories; *top right panel* shows all "unknown" and "no classification" categories; *bottom left panel* shows all "low" and "mid" categories; and *bottom right panel* shows the "high" category components. The vertical and horizontal black lines represents an SNR level of 3; and the diagonal black line represents the region where the SNR did not change after streak subtraction. The two population seen in the components rated "high" and "mid" (around f_{comp} SNR of 10^1 and 10^2) is due to the two sets of population (based on their peak flux density) present in catalogue.

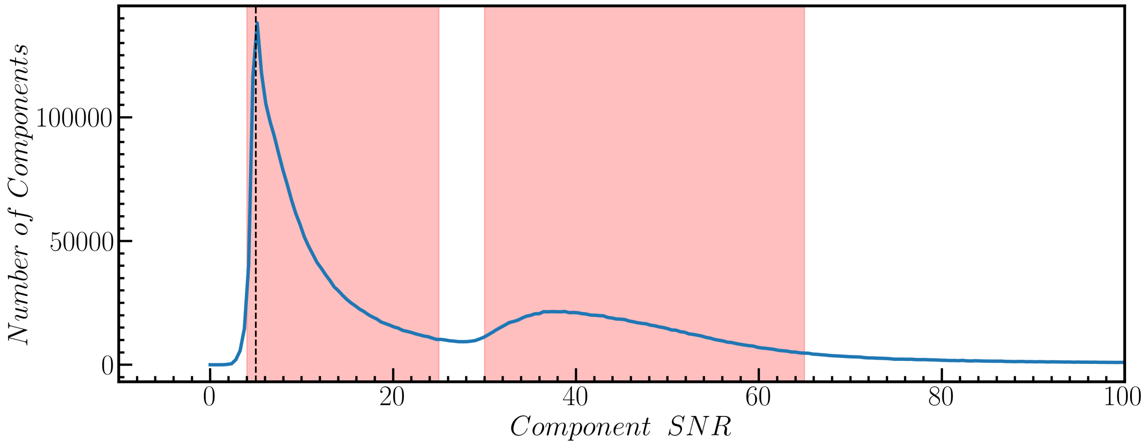


Figure 4.7: Histogram of the SNR level of all identified components in VLASS Epoch 1 QL images. The two highlighted regions represent the two population groups present in the catalogue.

can also be found in these outliers. All these outliers have a [11101] flag, while the rest of the "low" population have both [11101] and [11111] flags. Finally, two population seen in the "high" and "mid" categories are due to the presence of two sets of components with their peak flux density's SNR peaking at ∼5 and ∼40 (even at SNR $\sim$ 100 there are approximately 1000 components).

This second population is seen in Figure 4.7, showing the histogram of the SNR of all components identified in VLASS Epoch 1 QL images. The histogram of the SNR of "low", "mid" and "high" components before and after correcting for the overlapping streaks is shown in Figure 4.8. The change in the peak of the "high" and "mid" components' SNR seen at ∼40 before correction (to ∼23 after correction) indicates that large corrections are done for this population. Therefore, we believe that the population of VLASS components at SNR ∼40 includes a large population of artifact components that lie on sidelobe streaks from unrelated, very bright components. Yet, the presence of the bump/peak after correction (at SNR of ∼23) indicates that more accurate streak subtraction needed.

On the other hand, the population of VLASS components at SNR ∼5 include both the expected higher number of astrophysical components at low flux densities and the

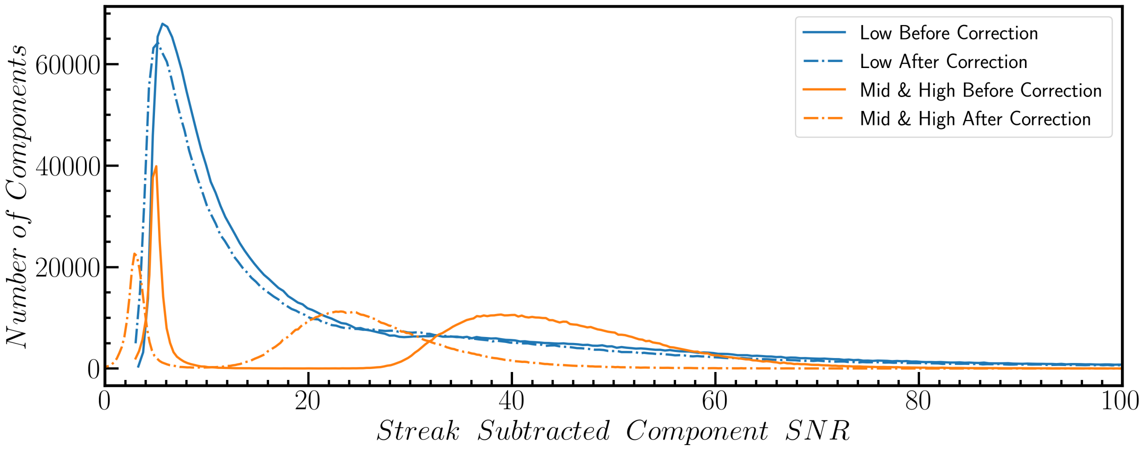


Figure 4.8: Histogram of the SNR level of the components classified as “low”, “mid” and “high” before and after subtracting the effects of the overlapping streaks. This shows that there is no clear secondary peak in the components with a “low” probability of arising from a streak artifact. Moreover, components with a “mid” or “high” probability of arising from a streak artifact show adecrease in the secondary peak SNR (to ∼23 from ∼40) of the “mid” and “high” component population after correction. While this indicates an improvement, there is still an unexpected population of components (and thus potential artifact components) at moderate SNR.

greater ease for low surface brightness streaks to cause artifacts. The two population groups seen in the “high” components are not caused separately by the two flags, [11100] and [110], given to this category. Both the “high” populations groups seen above $f_{\mathrm{corr,comp}}$ of 3 have the [11100] flag, while the population below $f_{\mathrm{corr,comp}}$ of 3 have the [110] flag.

4.3 Comparison of New Methods with VLASS QL Catalogue

Our two-fold “high” likelihood classification of a component arising from artifact streaks brackets past estimates of artifacts contamination in the Gordon et al. (2021) catalogue. We only identify $\sim 2\%$ of all components with our most strict [110] flag identifying artifacts, compared to the $\sim 8\%$ of the most strict flags of Gordon et al. (2021). Excluding only this flag, a larger fraction of the components Gordon et al. (2021) may be astrophysical sources with recoverable peak flux densities than previously believed.

However, our "looser" [11100] flag identified 23% components that are likely to be artifacts. This flag largely arises from components where the streaks make up a sizeable fraction (at least 20%) of the original peak flux density estimated for that component. While these components may include real sources, our ability to interpret their peak flux density will depend critically on how well our streak subtraction techniques work. This is a topic for future consideration beyond this thesis. For now, we conservatively label these sources as having a high likelihood of arising from a linear artifact.

As discussed in Section 1.4, the first version of the Gordon et al. (2021) catalogue flags potential spurious detections originating from sidelobe streaks using the *Peak_to_ring* metric. The left panel of Figure 4.9 shows the comparison of these *Peak_to_ring* with the new method's classification with respect to the measured flux density of the components. As mentioned in (Gordon et al., 2021), most of the potential artifacts are below *Peak_to_ring* of 3 as seen in the left panel. Only 597 components that are flagged as highly likely to be artifacts are seen above *Peak_to_ring* of 3.

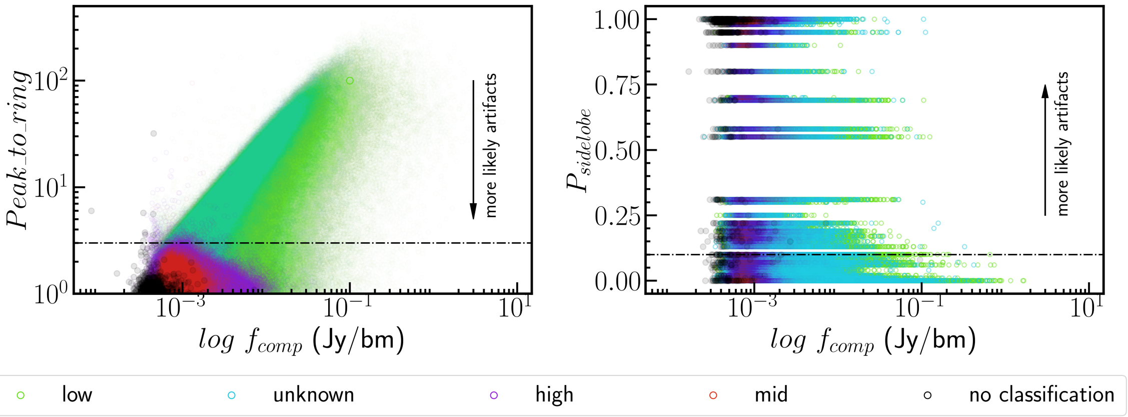


Figure 4.9: *Left:* The spread of each classification category in the *log* f_{comp} vs *Peak_to_ring* plane. A major fraction of the components classified as *high* or *mid* are seen below *Peak_to_ring* of 3 (shown as *dash-dotted* line). *Right:* The spread of each classification category in the *log* f_{comp} vs $P_{sidelobe}$ plane. A large fraction of the components classified as *high* or *mid* are seen above $P_{sidelobe}$ of 0.1 (shown as *dash-dotted* line).

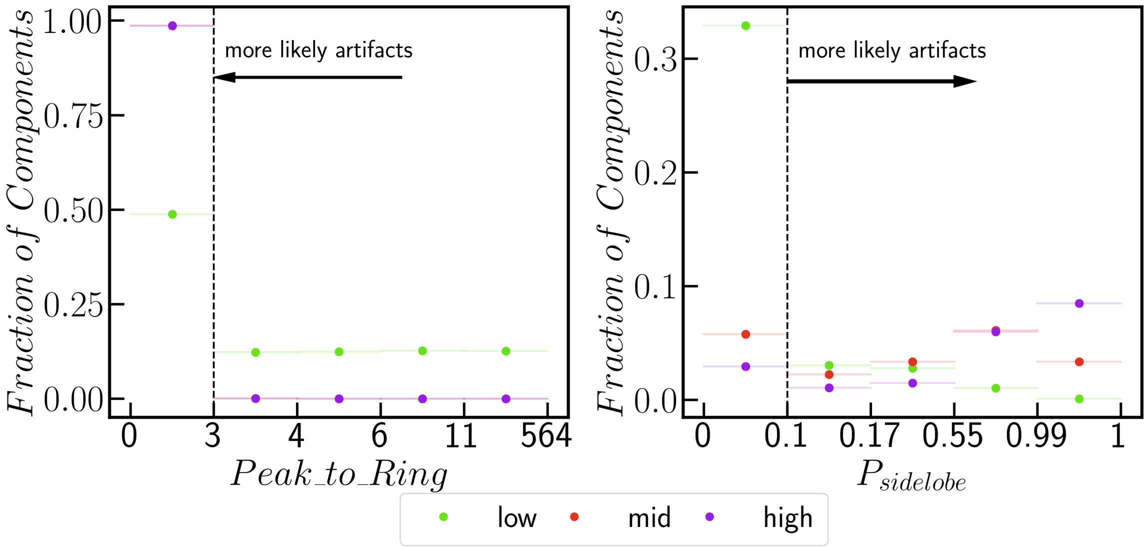


Figure 4.10: New classification component fraction comparison with *Peak_to_ring* and SOM $P_{sidelobe}$. Each of the 4 *Peak_to_ring* and $\mathrm{P_{sidelobe}}$ bins above the threshold (3 and 0.1 respectively) contains equal number of components. All components below the *Peak_to_ring* and $\mathrm{P_{sidelobe}}$ thresholds are binned together. The negligible number of "high" and "mid" components seen above *Peak_to_ring* of 3 and increase in the number of "high" and "low" components with the increase in $\mathrm{P_{sidelobe}}$ suggests a good agreement between the new classification and the two metrics.

The updated version of the catalogue has the SOM's $\mathrm{P_{sidelobe}}$ metric that is based on *Peak_to_ring*. As discussed in Vantyghem et al. (2021), only components that have *Peak_to_ring* < 3 are considered for the SOM sidelobe probability prediction. Of these components some (especially components that have a peak flux density of ≤ 1 mJy/bm) are classified as "high" or "mid" in our classification method as seen in the right panel of Figure 4.9. The vertical spread of the "high" and "mid" at different $\mathrm{P_{sidelobe}}$ is currently not yet understood and therefore further investigation is necessary. In general, the decrease in number of components classified as "low" with the increase in $\mathrm{P_{sidelobe}}$ indicates an agreement between the two methods.

The agreement between our classification and the two metrics is highlighted in Figure 4.10 showing the fraction of components in each category found in bins with each number of components. A negligible fraction of "high" and "mid" category components have *Peak_to_ring* > 3, which Gordon et al. (2021) associate with sources that are unlikely to be artifacts. Here we note that both the “high” and “mid” fractions are similar in all the bins and they overlap each other in the *left* panel of Figure 4.10. Moreover, since they both have similarly very low fractions at high *Peak_to_ring*, and the figure was made by evenly binning the *Peak_to_ring* > 3 into 4 populations, the fraction of “low” components in each bin above the *Peak_to_ring* of 3 are all approximately the same (by design). In the *right* panel of Figure 4.10, we evenly bin the P_sidelobe > 0.1 into 4 populations. Here, the fraction of components categorized as “high” and “mid” increase with the increase in $\mathrm{P_{sidelobe}}$, while the fraction of “low” components in each bin decreases with the increase in $\mathrm{P_{sidelobe}}$.

Although there is a general agreement between our new classification and the two older metrics, there are components where the newer and older methods disagree on artifact identification. The number of components in each category from our method is cross referenced with *Peak_to_ring* and $\mathrm{P_{sidelobe}}$ values in Tables 4.1–4.2. In both tables, for components below the threshold (3 for *Peak_to_ring* for and 0.1 for $\mathrm{P_{sidelobe}}$), an additional comparison of whether the components are isolated ($> 20''$) or not is added. This is added to understand the conditions where our new classification succeeds and fails in correctly identifying the components.

In Table 4.1, the components that are classified as “low” but have a *Peak_to_ring* < 3 and the components that are classified as “high” but have a *Peak_to_ring* ≥ 3 are the regions of disagreement. Similarly, in Table 4.2, the components that are classified as “low” but have a $\mathrm{P_{sidelobe}} \geq 0.1$ and the components that are classified as “high” but have a $\mathrm{P_{sidelobe}} < 3$ are the regions of disagreement. Here we must note that components that are given a *Peak_to_ring* of -99 and $\mathrm{P_{sidelobe}}$ of -1 are ignored in their respective tables. A random sample of components in the above mentioned four disagreement regions are shown in Figures 4.11–4.14. These examples show that while many components are accurately classified in our new method, some components are misclassified as well; this also appears true for the older *Peak_to_ring* and SOM methods. Further visual inspection and analysis is required to quantify the fraction

Table 4.1: Numbers of components by classification, cross referenced with *Peak_to_ring* (PTR) and nearest neighbour properties

	PTR < 3, not isolated[a]	PTR < 3, isolated[a,b]	PTR ≥ 3	Total (PTR ≥ 0)
No Classification	590	381	8	979
Unknown	122,300	146,672	338,961	607,933
Low	493,921	416,154	932,830	1,843,905
Mid	40,672	5,725	43	46,440
High	738,105	101,604	597	840,306
Total	1,447,661	618,463	1,272,439	3,338,563

NOTE.—The *Peak_to_ring* (PTR) and nearest neighbour properties are from Gordon et al. (2021). This table excludes components with poorly measured Peak_to_ring values (PTR=-99).

[a] Gordon et al. (2021) consider these components to likely be artifacts.

[b] Following Gordon et al. (2021), we label components with NN_dist $\geq 20''$ as isolated.

of misidentified components and identify the origin behind the wrong classification in our new method.

Finally, since the goal of this thesis is to aid transient searches by correctly estimating the brightness of the identified components and thus distinguish real astrophysical sources from spurious detections, here we give two examples of how our new method correctly estimates and classifies such components. Figure 4.15 shows a real astrophysical source (R.A. 293.933° and Dec -6.981°) with overlapping streaks seen in both Epoch 1 and Epoch 2. This source is correctly given a probability rating of "low", but has a *Peak_to_ring* of 2.53 and $\mathrm{P_{sidelobe}}$ of 0.7. Figure 4.16 shows a example of an artifact (at R.A. 94.525° and Dec 62.453°) correctly flagged as a spurious detection ("high") but has a $\mathrm{P_{sidelobe}}$ of 0.02 and *Peak_to_ring* of 1.38.

Table 4.2: Numbers of components by classification, cross referenced with P_{sidelobe} and nearest neighbour properties

	$P_{\text{sidelobe}} < 0.1$, not isolated[a]	$P_{\text{sidelobe}} < 0.1$, isolated[a,b]	$P_{\text{sidelobe}} \geq 0.1$	Total ($P_{\text{sidelobe}} \geq 0$)
No Classification	104	33	814	951
Unknown	51,448	136,853	28,543	216,844
Low	237,602	376,441	130,598	744,641
Mid	1,031	1,687	7,103	9,821
High	10,344	14,728	145,201	170,273
Total	1,546,230	1,480,174	312,159	3,338,563

NOTE.—The P_{sidelobe} and nearest neighbour properties are from Vantyghem et al. (2021) and Gordon et al. (2021) respectively. This table excludes components with poorly measured P_{sidelobe} values (P_{sidelobe}=-1).

[a] Vantyghem et al. (2021) consider these components to likely be artifacts.

[b] Following Gordon et al. (2021), we label components with NN_dist $\geq 20''$ as isolated.

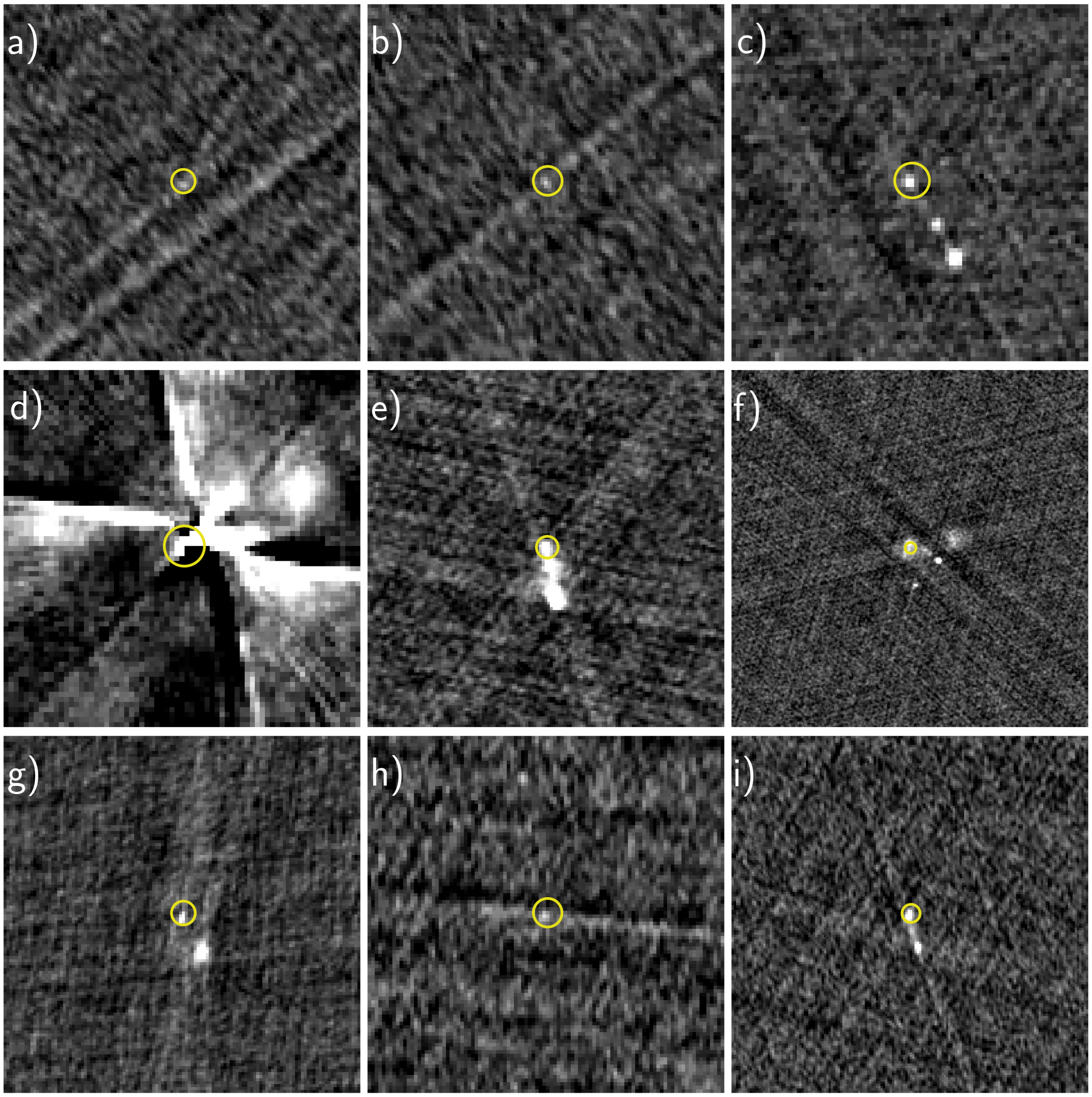


Figure 4.11: Sample of components that we classify as having a "low" probability of arising from a streak artifact, but that the older *Peak_to_ring* method classifies as an artifact. Panels *a* and *d* are likely artifacts and thus our method appears to have failed to accurately classify them. We believe that we correctly classified the rest of the components as potential real astrophysical sources/ components, whereas the older method misclassified these sources as likely artifacts.

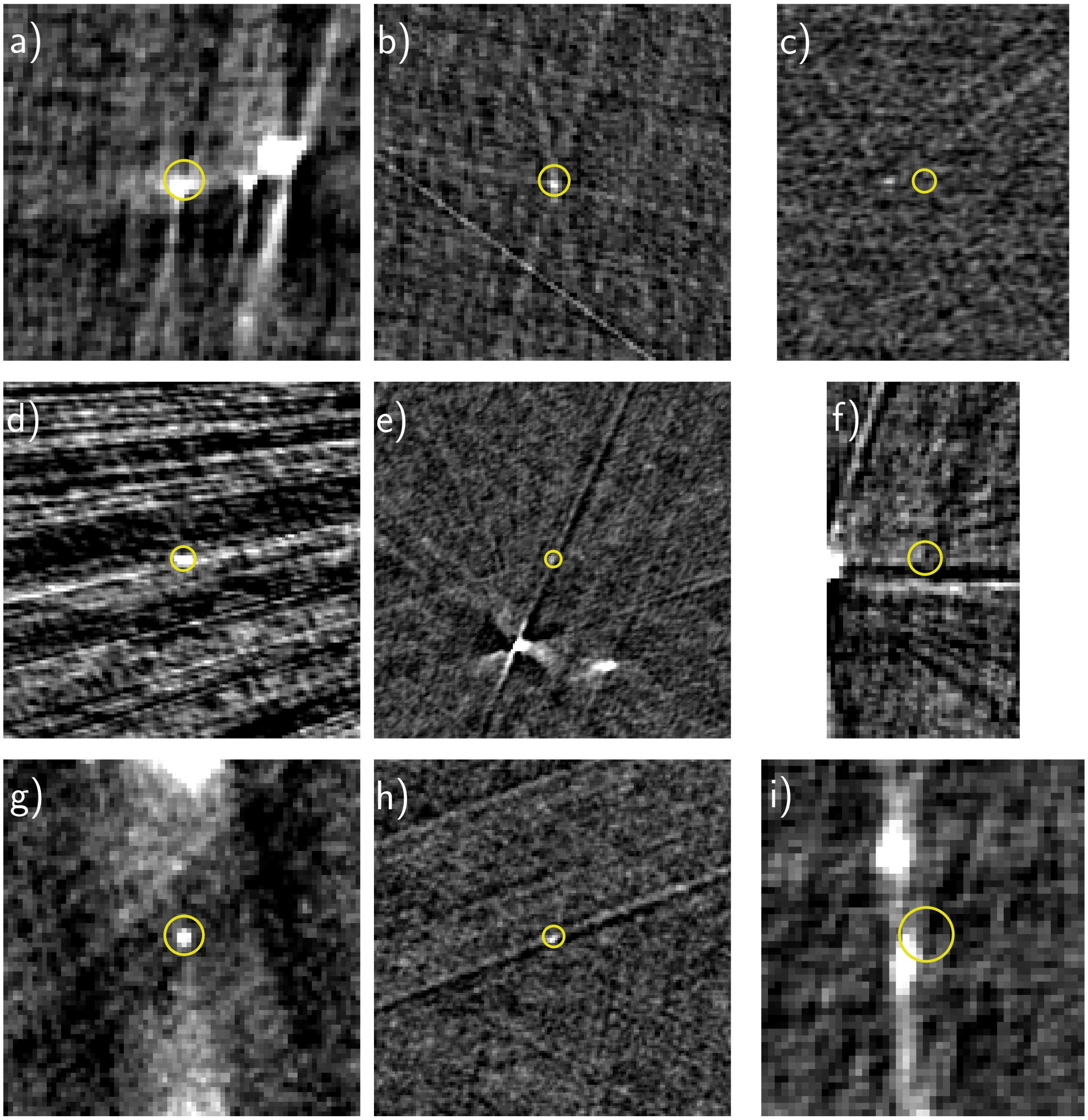


Figure 4.12: Sample of components that we classify as having a “high” probability of arising from a streak artifact and that the older *Peak_to_ring* method classifies as an artifact. Panels *a*, *b*, *g* and, *h* appear to be wrongly classified as likely artifacts in the new method, while the rest of the panels appear appropriately classified as artifacts.

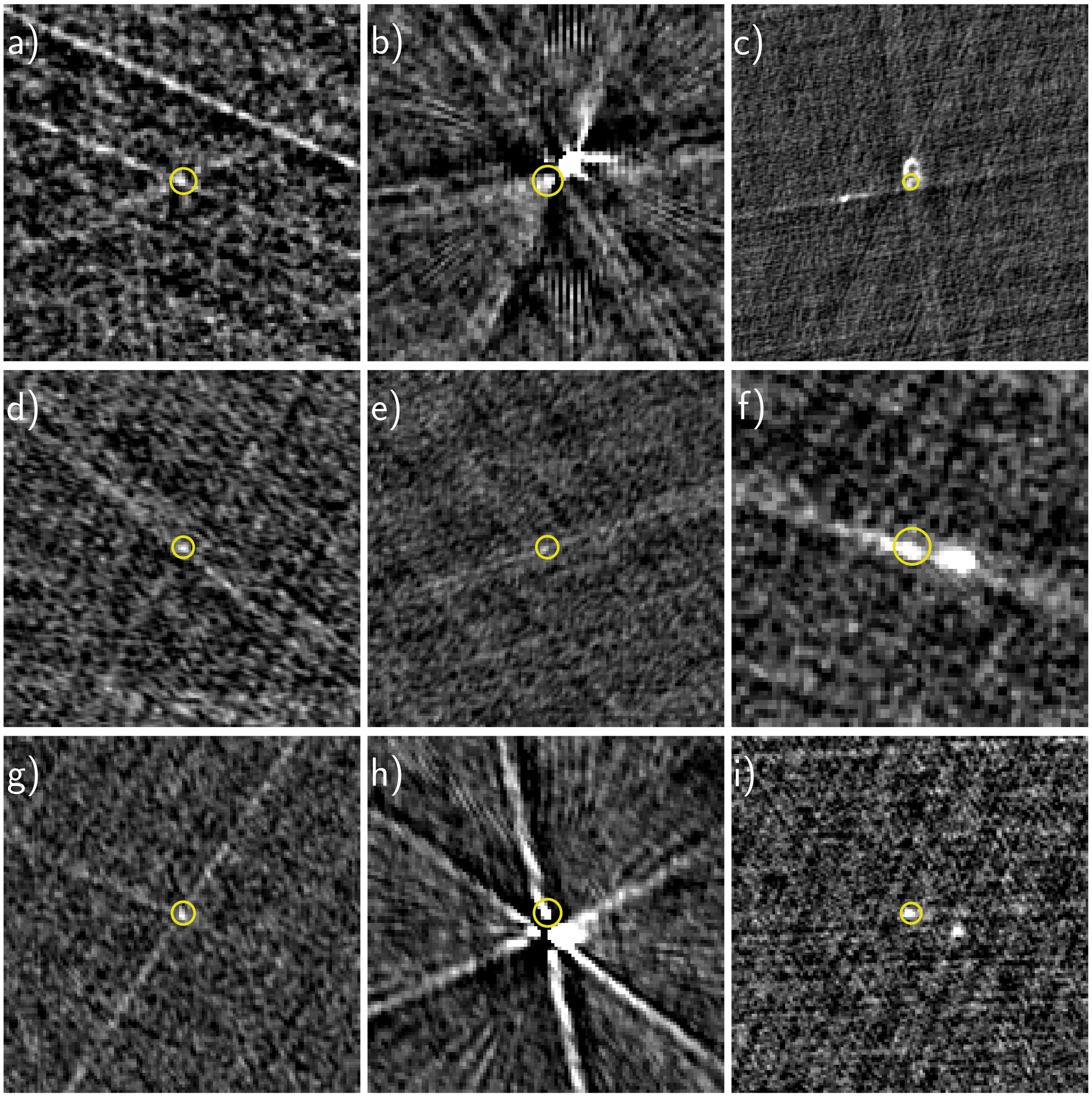


Figure 4.13: Sample of components that we classify as having a "low" probability of arising from a streak artifact and that the older SOM method classifies as an artifact. Panels a, d g and, i appear to be appropriately classified as likely astrophysical sources in the new method, but incorrectly classified in the $\mathrm{P_{sidelobe}}$ method. Panels b, e, and h are misclassified in the new classification method, but correctly classified in the $\mathrm{P_{sidelobe}}$ method. The remaining two panels c and f are slightly ambiguous and require further visual inspection to verify the classification.

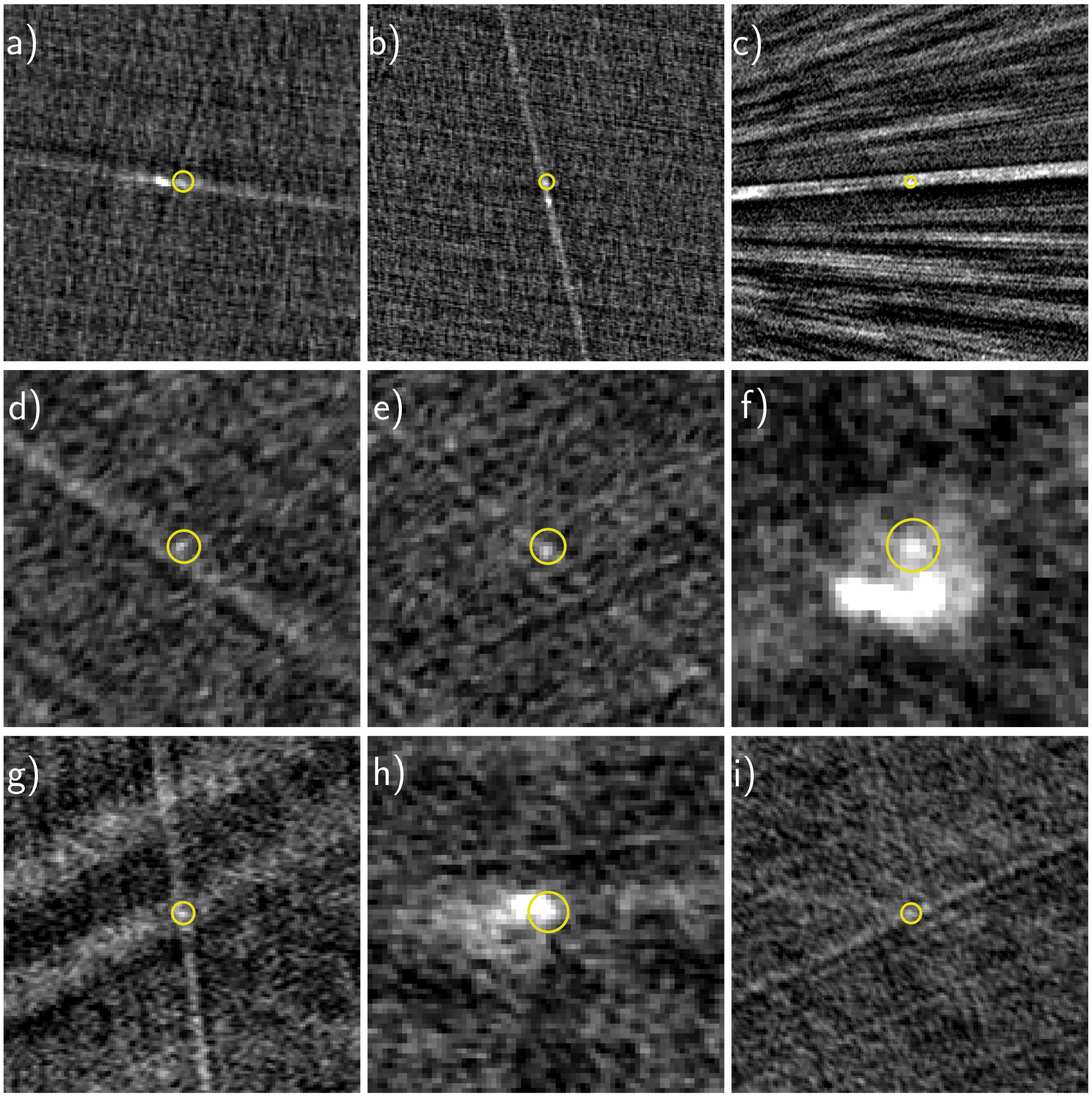


Figure 4.14: Sample of components that we classify as having a “high” probability of arising from a streak artifact, but that that the older SOM method classifies as likely being astrophysical. Panels a, c, g and, i appear correctly classified as artifacts in the new method, but incorrectly in the SOM method. Panels b, d and, e are potential mis-classifications by the new method that the SOM method identified as likely real sources. Our method definitely fails to correctly classify components in panels f and h.

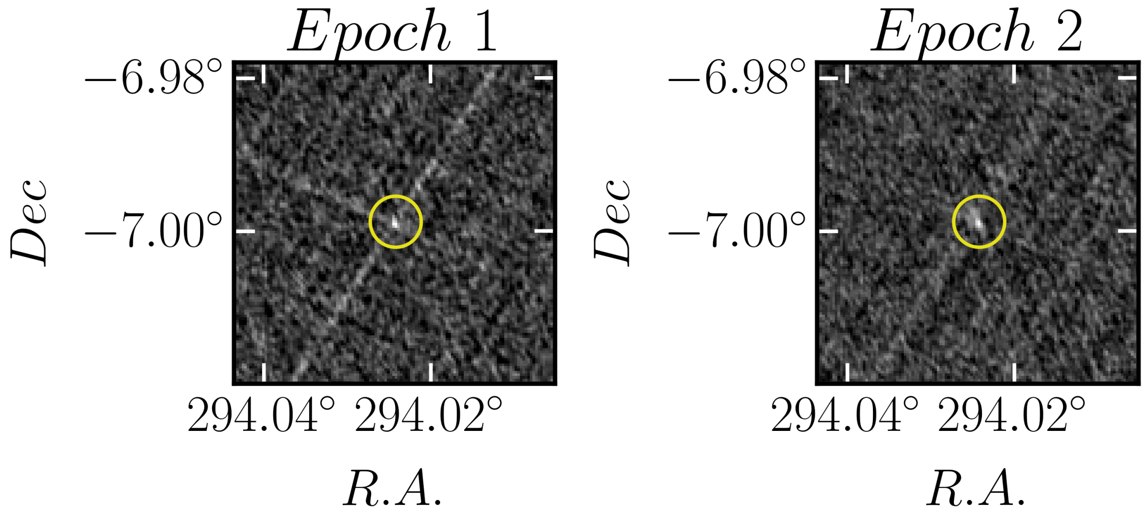


Figure 4.15: Epoch 1 and 2 QL image cutouts of a real source correctly estimated as having a "low" probability in Epoch 1 of being an artifact by our method. However, since the *Peak_to_ring* was 2.53, the SOM method found a $\mathrm{P}_{\mathrm{sidelobe}}$ of 0.7.

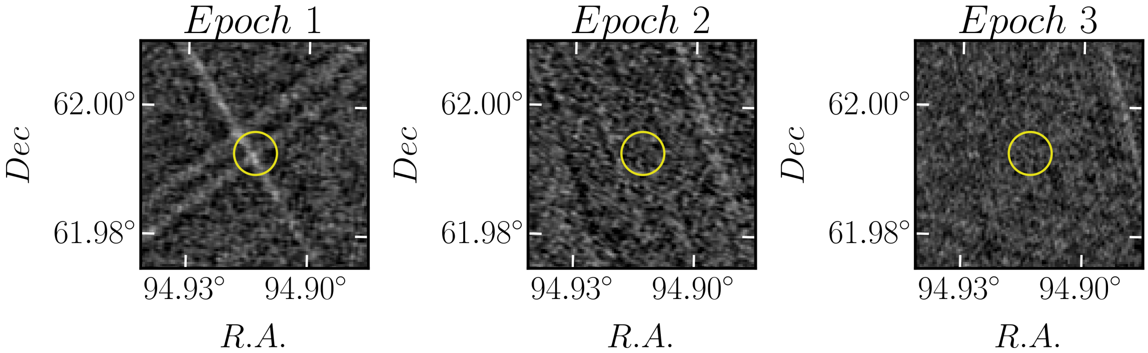


Figure 4.16: Epoch 1–3 QL image cutouts of a spurious detection that our new method correctly classifies as having a "high" likelihood of being an artifact. While the older *Peak_to_ring* method also found it likely to be an artifact, the SOM method found a $\mathrm{P}_{\mathrm{sidelobe}}$ of 0.02.

Chapter 5

Conclusion

5.1 Summary

In this thesis we developed a new artifact identification method that identifies linear streaks overlapping previously identified components across all VLASS Epoch 1 Quick-Look images with no user input. A visual inspection of over 500 images was performed as one of the verification steps for this method.

In this technique that follows component detection, the input VLASS QL images are pre-processed before passing through a line detection technique called the Hough Transform; the images are clipped and detected sources are masked to obtain Boolean images. A local region-based analysis is performed for each component to decrease the false-positive identification of overlapping linear streaks. To identify linear streaks the results of the Hough Transform are extended to adapt for the high noise level relative to signals present in the QL images as compared to typical images outside of astronomy.

The identified streaks around the components are quantified by measuring the width and average surface brightness of the detected streaks. During the quantification process, only streaks that are at least partially overlapping the ellipse of a component are considered.

Next, the effect of overlapping streaks are removed by subtracting the total surface brightness of all overlapping streaks above a SNR of 5σ. The components are then classified into three major categories: *high*, *mid*, and *low* based on the probability of them being imaging artifacts. Components that have a SNR of less than 3 are not classified and flagged as *no classification.* Since this classification is based on overlapping linear streaks, components for which no linear streaks were found are also not classified in this method and are flagged as *unknown.* Finally, the new artifact detection and classification methods are compared with the previously developed *Peak_to_ring* and SOM metrics.

This new technique has applications in any photometric techniques that will test if detected components are local maxima in linear streaks, real astrophysical sources on top of related or unrelated streaks, or astrophysical sources that are uncontaminated by streaks. This technique is especially well suited to be extended to other VLA projects, particularly those where time constraints or large data volumes prevent optimal image reconstruction.

5.2 Key Results

Our new artifact detection algorithm was applied on all 3,381,277 components in the Gordon et al. (2021) catalogue. Of these, a total of 4,733,249 overlapping streaks were identified around 2,807,084 components. No overlapping streaks were identified for the remaining 574,193 components. Of the remaining components, 988 components did not have a peak flux density SNR ≥ 3 and hence were removed from the classification. The rest of the components were given a “low”, “mid” or “high” probability of being potential imaging artifacts. A total of 850,718 components were classified as “high”; 47,025 components were classified as “mid”; and 1,866,049 components were classified as “low”. Two different flags lead to a “high” classification. In particular, the most strict flag ([110]) identifies 73,133 components (2.16% of all components) as likely being artifacts with streak-corrected peak flux densities that do not reach above the 3-σ level. The “looser” flag ([111100]) identifies 777,585 components (23.00% of all components) as likely being an artifact that lead to a streak-corrected peak flux densities that have streak surface brightnesses that are at least 20% of the original peak flux density. In particular, the fact that our strict flag identifies fewer artifacts than past methods suggests that our new method may lead to a greater recovery of astrophysical components compared to past methods, if future tests prove that our streak removal process is robust.

5.3 Limitations of New Methods

Although the new method is capable of identifying artifacts that are missed by standard approaches, there are a few limitations in the new method.

- The new method is based on the already identified components in the CIRADA QL catalogue. Although the new detection algorithm is capable of identifying streaks around any arbitrary component or coordinate, the pre-processing step of masking surrounding components is dependent on the the catalogue. For example, some components in catalogue are not identified or their dimensions are not estimated accurately. The lack of, or incorrect, masking of such components affects the measured signal strength in the local region; this sometimes leads to false-positive streak detections.

- In the new technique, apart from the streaks and the sources/components, a uniform noise level is assumed in all QL images. Non-uniform noise level around some components can result in improper background profile subtraction of the 1D HT profile. This issue can lead to an overestimation or underestimation of the noise level of the streaks. This either leads to false-positive streak detections or missing obvious streaks around components.

- PyBDSF was unable to fit a Gaussian to few real components/sources in the (Gordon et al., 2021) catalogue. For a few such components, the peak flux densities are severely underestimated. The streak subtraction of such components leads to inaccurate corrected peak flux densities and consequently, to being wrongly classified as artifacts by the new method.

5.4 Future Works

In this method, the components are preliminarily classified into three categories based on the probability of them being imaging artifacts. In the future, we intend to add a more statistically based probabilistic classification. This will include an extensive review of a large number of classified components. Next, we plan on adding the host information present in the CIRADA catalogue as a factor that affects the classification.

Many of the mis-classifications in our method is due to the lack of an accurate Gaussian fit to the identified components. To counter this and for the completeness of the new methods, we consider adding a source detection step prior to the artifact detection. Finally, as of now, the artifacts detection and classification is only performed on VLASS Epoch 1 images. We plan on expanding this to Epoch 2 and 3 images as well.

Bibliography

Ahumada, R., Allende Prieto, C., Almeida, A., et al. 2020, ApJS, 249, 3, doi: 10.3847/1538-4365/ab929e

Ashworth, William B., J. 1981, Proceedings of the American Philosophical Society, 125, 52

Ballard, D. H. 1981, Pattern Recognition, 13, 111, doi: 10.1016/0031-3203(81)90009-1

Ballester, P. 1994, in Astronomical Society of the Pacific Conference Series, Vol. 61, Astronomical Data Analysis Software and Systems III, ed. D. R. Crabtree, R. J. Hanisch, & J. Barnes, 319

Brecher, K. 1978, Nature, 273, 728, doi: 10.1038/273728a0

Clark, S. E., Peek, J. E. G., & Putman, M. E. 2014, ApJ, 789, 82, doi: 10.1088/0004-637X/789/1/82

Dong, D. Z., Hallinan, G., Nakar, E., et al. 2021, Science, 373, 1125, doi: 10.1126/science.abg6037

Duda, R. O., & Hart, P. E. 1972, Commun. ACM, 15, 11, doi: 10.1145/361237.361242

Franklin, K. L. 1959, AJ, 64, 37, doi: 10.1086/107852

Fridman, P. A. 2010, MNRAS, 409, 808, doi: 10.1111/j.1365-2966.2010.17346.x

Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2016, aap, 595, A2, doi: 10.1051/0004-6361/201629512

Gordon, Y., O'Dea, C., Rudnick, L., et al. 2021, https://cirada.ca/catalogues, 24. https://cirada.ca/catalogues

Gordon, Y. A., Boyce, M. M., O'Dea, C. P., et al. 2021, ApJS, 255, 30, doi: 10.3847/1538-4365/ac05c0

Ho, P.-Y., Paar, F. W., & Parsons, P. W. 1972, Vistas in Astronomy, 13, 1, doi: 10.1016/0083-6656(72)90002-5

Högbom, J. A. 1974, A&AS, 15, 417

Hollitt, C., & Johnston-Hollitt, M. 2012, PASA, 29, 309, doi: 10.1071/AS11051

Hough, P. V. C. 1962, U.S. Patent No. 3,69,654, U.S. Patent and Trademark Office

Howard, W. E., Staelin, D. H., & Reifenstein, E. C. 1968, IAUC, 2110, 2

Lacy, M., Baum, S. A., Chandler, C. J., et al. 2020, PASP, 132, 035001, doi: 10.1088/1538-3873/ab63eb

Law, C. J., Gaensler, B. M., Metzger, B. D., Ofek, E. O., & Sironi, L. 2018, ApJ, 866, L22, doi: 10.3847/2041-8213/aae5f3

Mayall, N. U., & Oort, J. H. 1942, PASP, 54, 95, doi: 10.1086/125410

Messier, C. 1781, Catalogue des Nébuleuses et des Amas d'Étoiles (Catalog of Nebulae and Star Clusters), Connoissance des Temps ou des Mouvements Célestes, for 1784, p. 227-267

Mohan, N., & Rafferty, D. 2015, PyBDSF: Python Blob Detection and Source Finder, Astrophysics Source Code Library, record ascl:1502.007. http://ascl.net/1502.007

Mooley, K. P., Hallinan, G., Bourke, S., et al. 2016, ApJ, 818, 105, doi: 10.3847/0004-637X/818/2/105

Nyland, K., Dong, D. Z., Patil, P., et al. 2020, ApJ, 905, 74, doi: 10.3847/1538-4357/abc341

Pietka, M., Fender, R. P., & Keane, E. F. 2015, MNRAS, 446, 3687, doi: 10.1093/mnras/stu2335

Ragazzoni, R., & Barbieri, C. 1994, PASP, 106, 683, doi: 10.1086/133429

Reber, G. 1944, ApJ, 100, 279, doi: 10.1086/144668

Rowlinson, A., Meijn, J., Bright, J., et al. 2022, MNRAS, 517, 2894, doi: 10.1093/mnras/stac2460

Sironi, L., & Giannios, D. 2013, ApJ, 778, 107, doi: 10.1088/0004-637X/778/2/107

Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, aj, 131, 1163, doi: 10.1086/498708

Thompson, A. R., Moran, J. M., & Swenson, George W., J. 2017, Interferometry and Synthesis in Radio Astronomy, 3rd Edition (Springer), doi: 10.1007/978-3-319-44431-4

Vantyghem, A., O'Dea, C., Rudnick, L., et al. 2021, CIRADA website, 16

Xu, Z., Shin, B.-S., & Klette, R. 2014, in Advanced Information Systems Engineering, ed. C. Salinesi, M. C. Norrie, & O. Pastor, Vol. 7908 (Springer Berlin Heidelberg), 190–201, doi: 10.1007/978-3-319-09955-2_16

Zuo, S., & Chen, X. 2020, MNRAS, 494, 1994, doi: 10.1093/mnras/staa891