UNIVERSITY OF CAPE TOWN

A DISSERTATION PRESENTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF ASTRONOMY

The MeerKAT Radio Frequency Interference Environment

Author:
Isaac Sihlangu

Supervised by:
Professor Bruce Bassett
Dr Nadeem Oozeer
Professor Russ Taylor

November 2019

Abstract

Radio signals from astronomical sources are extremely weak and easily distorted/-corrupted or overwhelmed by man-made radio signals such as cellphones, satellites, aircraft and telescope electronics. These Radio Frequency Interference (RFI) are increasingly threatening radio observatories due to our increasingly technological world. To detect and mitigate RFI, observatories need to understand their RFI environment, what contributes to it and how it is changing. While there are few dedicated RFI monitoring systems on the MeerKAT site, the most sensitive RFI detector is the MeerKAT array itself. In this thesis we use approximately 1500 hours of MeerKAT observations to create a multi-dimensional view of the RFI at the MeerKAT site.

Here we investigate a probabilistic approach to characterise the RFI environment around the MeerKAT radio telescope. In order to achieve our goal, we propose the MeerKAT Historical Probability of RFI (KATHPRFI) framework. We produced the high level requirements of the KATHPRFI framework driven by the needs of the MeerKAT users. The design approach and the design decision of the framework is presented that cover both the software and hardware constraints. The KATHPRFI produces a 5-dimensional array of the RFI probability as measured by the MeerKAT telescope during the commissioning phase (May 2018 - December 2018) for each observation file.

From the 5-D array, we extracted various statistics and characterised the RFI environment around MeerKAT site. We found that there is a correlation between RFI occupancy and the time of the day which is most probably related to human activities. Furthermore, we found a correlation between the time of the day and flights passing over a region of site. Our results showed that the highest probability of RFI points towards a region including nearby towns. The results obtained are consistent with the argument that the major RFI sources for MeerKAT site are the Global Positioning System (GPS) satellite, flight Distance Measurement Equipment (DME) and the Global System for Mobile Communications (GSM). Our data also showed that the RFI occupancy decreases with an increase of baseline length, this is a result of moving RFI sources with respect to the static sky. Therefore, the phase of the RFI changes rapidly on long baselines compared to short baselines.

As a result when a correlation is carried out the RFI amplitude will vanish less on short baselines compared to the long baselines.

Our results provide the first highly detailed view of the MeerKAT RFI environment allowing us to track the historical evolution of the RFI, both on average, and as a function of frequency, baseline and direction. With historical baselines known, one can also provide alerts about sudden changes. This could be due to new sources of RFI or stem from any outliers in the data, which could signal telescope or correlator issues. Hence the KATHPRFI framework also provides a window into the operational health of the telescope.

Contents

List of Figures

Acknowledgements

I would like to convey my sincere gratitude to my parents for always supporting and encouraging me to reach for the stars. This work would not be possible if it was not for their prayers and consistent hard-working spirit that they have taught me. I am dedicating this thesis to them.

To my friends, Tumelo, Albert, Simphiwe, Abu and the craziest of them all Daniel, thank you guys for being there when I needed you the most. Special thanks to my girlfriend, Tshireletso for not adding more stress during the process of researching and writing this thesis but instead she provided me with unfailing support and continuous encouragement throughout.

Dr T. Mauch, you are one of the best people that I have ever met. Thank you for your continuous support and guidance, moreover, I would like to pass my humble gratitude to you for reading this thesis several times and providing honest and constructive feedback.

To my manager, Mr Khutso Ngoasheng, I cannot express how I feel with what you have done in my life. Thank you for believing in me, for all the advice, the knowledge and wisdom that you have shared me since I joined the SARAO. May the Lord helps to get the desires of your life.

To my supervisors, Professor Bruce Bassett and Dr Nadeem Oozeer, I could not have asked for awesome supervisors like you. Thank you for your encouragement, honest constructive feedback and above all your mentorship. I have grown a lot to be a better researcher through your guidance. I would also like to thank SARAO data science team for lending a shoulder to lean on, more especially Chris Finlay.

Finally, I would like to thank the one above us, the Almighty God for awarding me with such an opportunity.

Declaration

This thesis is an account of research undertaken between June 2018 and November 2019 at The Department of Astronomy, Faculty of Science, University of Cape Town.

I, Sihlangu Isaac, hereby declare that the work presented in this thesis to the best of my knowledge is my own, except where otherwise clearly indicated in the customary manner. The material in this thesis has not been previously submitted, as a whole or in part, to any university for award of academic degree.

Sihlangu Isaac
November, 2019

Chapter 1

Introduction

1.1 Astronomy

Astronomy is one of the oldest sciences, with origins spanning back to the earliest human civilisations (Krupp 2003). For thousands of years, people have been observing the sky with the naked eye, as is noticeable in the naming of many constellations, notably the largely Greek names used today such as Orion, Leo and Centaurus. Before the invention of the compass, ancient sailors used the position of the stars to find their way across the seas (Nielbock 2017). Humankind's inquisitiveness about the cosmos led to the invention of the telescope. Galileo Galilee contributed massively to the development and the building of telescopes (e.g. Fig. 1.1) which he used for astronomical observations. Ever since more sensitive instruments have been built to improve the depth and the precision of astronomical measurements (Walker 1996).

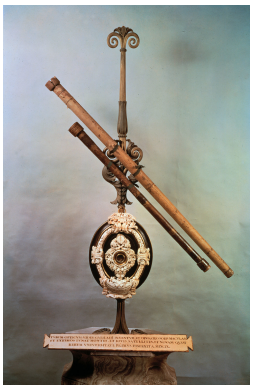


Figure 1.1: *Two of Galileo's first telescopes (Source:* https://www.mpg.de/7913340/Galileo_Galilei_telescope*.*

Celestial sources emit Electromagnetic (EM) radiation that is classified according

to wavelength into radio, microwave, infrared, visible, ultraviolet, X-rays and gamma rays as depicted in Figure 1.2. By studying this EM radiation, scientists deduced that the Universe is filled with an enormous number of planets, stars, galaxies and other astronomical objects. Properties such as the chemical composition, temperature, density, mass, distance and relative motion of these objects can be worked out by analysing the radiation they emit.

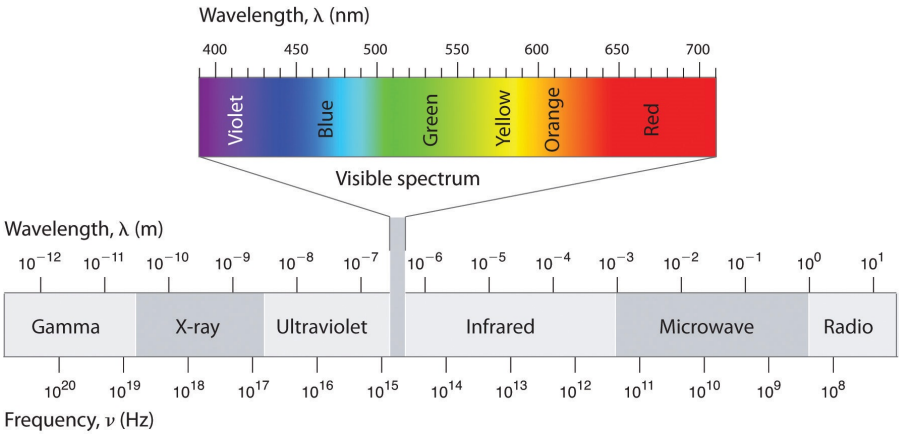


Figure 1.2: *The electromagnetic spectrum with visible light being only a fraction of the spectrum (Miller 1998).*

The way that astronomers observe EM waves depends on the portion of the spectrum they wish to study. Before 1930 doing astronomy meant studying astronomical objects that emit radiation in the visible part of the spectrum (Liddle 2015). It was only from the late 1930s that technological advances allowed astronomers to exploit other parts of the EM spectrum. We are now in a regime where we have different kinds of instrument sensitive to different parts of the EM spectrum.

The Earth's atmosphere is opaque to some of the electromagnetic wavelengths. Therefore, space-based telescopes (e.g. Chandra X-ray Observatory [1]) are used to study such wavelengths, while ground-based telescopes (e.g. MeerKAT Radio Telescope) are used to study those that penetrate through the atmosphere. There are two wavelength regimes that can go through the atmosphere, those are the optical/near-infrared, and radio as shown in Fig. 1.3. Over the long path that EM waves travel from the astronomical sources to our instruments lie atoms and molecules and tiny particles of dust referred to as the interstellar medium (ISM).

[1] https://chandra.harvard.edu/about/

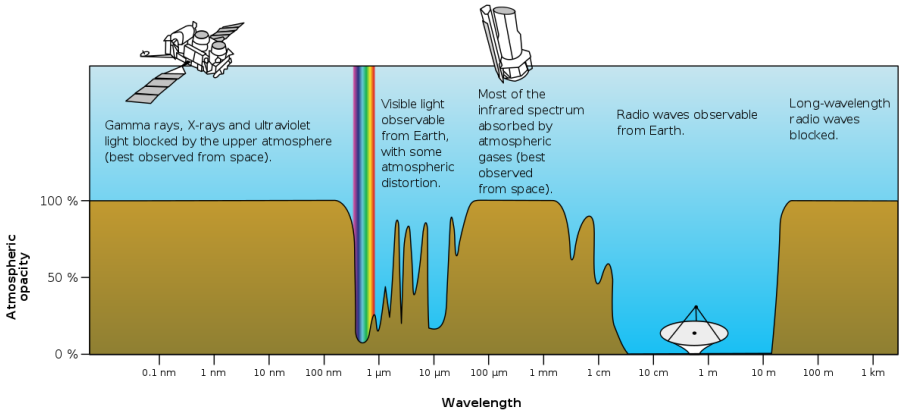


Figure 1.3: *The Earth's atmospheric transmittance (or opacity) towards electromagnetic waves. The millimetre to a decametre window allows the use of ground-based radio telescopes.(Source:https://phys.libretexts.org/Bookshelves/University_Physics/Book3A_Physics_(Boundless)/233A_Electromagnetic_Waves/23.13A_The_Electromagnetic_Spectrum)*

The ISM causes radiation from the cosmic source to be scattered or absorbed. For those wavelengths that can penetrate through the atmosphere, the long radio wavelengths are least affected by the ISM.

1.2 The birth of Radio Astronomy

In the early 1930s, a young physicist by the name of Karl Guthe Jansky was tasked by Bell Laboratories to investigate the sources of radio waves which were interfering with their telephone communications system. To do the research, Jansky built a bridge-like structure that held antenna wires set on a spinning base, Fig. 1.4. The antenna listened to the static radio interference throughout the day and night. Jansky noticed that the pattern of radiation he observed arrived 4 minutes earlier each day, which is the timescale of the sidereal day. The sidereal day is the time it takes for the Earth to complete one rotation about its axis relative to background stars.

The observations led Jansky to conclude that the radiation is not of terrestrial origin; further research allowed him to pinpoint that it is coming from the centre of our Milky Way galaxy. Hence, Jansky's telescope serendipitously detected extraterrestrial radio radiation for the first time, making it the first radio astronomy telescope (Jansky 1958).

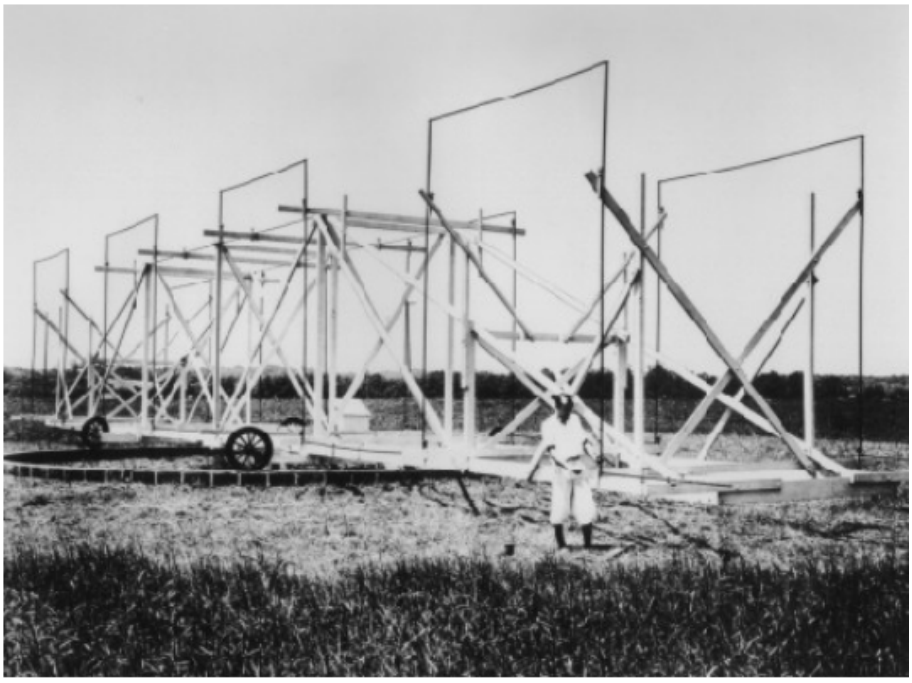


Figure 1.4: *The First Radio Telescope that lead Jansky to his serendipitous discovery of radio emission from the Milky Way (Sullivan III 1984).*

This discovery opened a new whole window on the study of celestial objects, and it marks the birth of radio astronomy. Radio astronomy is a sub-field of astronomy which studies astronomical phenomena often invisible in other portions of the electromagnetic spectrum, that give off radio waves. Radio astronomers, as a way of honouring Karl Jansky for his discovery, have introduced Jansky (Jy) units to measure the flux density of astronomical objects defined as follows:

$$1Jy = 10^{-26}Wm^{-2}Hz^{-2} \tag{1.1}$$

Radio astronomy measurements are recorded using large radio antennas referred to as radio telescopes. The radio telescopes can used as a single dish like FAST (Nan et al. 2011), or with multiple linked antennas like MeerKAT, as discussed by Booth & Jonas (2012). The ability of a telescope to discern fine details of astronomical objects is determined by its angular resolution, ($\theta_{angular}$). This angular resolution depends upon the dish aperture size/diameter (D), as well as the observing wavelength (λ) (Thompson et al. 1986), given by:

$$\theta_{angular} = 1.22\frac{\lambda}{D} \tag{1.2}$$

The bigger the value of $\theta_{angular}$, the harder it becomes to resolve fine details, whereas small values indicate fine resolution. Since radio waves have long wavelengths, they pose a considerable challenge to astronomers to discern fine details. One way to get a fine resolution is to build telescopes with a larger diameter. However, due to mechanical constraints and cost, it is impractical to build steerable single-dish radio telescopes much bigger than 100 meters (Thompson et al. 1986). Examples of well known and fully steerable large radio dishes are the 100 meter Green Bank Telescope, the Effelsberg 100-m Radio Telescope near Bonn in Germany and the 76-meter Lovell Telescope.

The modern way to achieve higher resolution is to use multiple linked radio telescopes, that utilise the techniques of radio interferometry and aperture synthesis (Wiaux et al. 2009).

1.3 Radio Interferometer Technique

The basic idea of radio interferometry involves combining the output signals of two radio telescopes separated by some distance $\|\vec{b}\|$, where $\vec{b}$ is referred to as the baseline. Following Equation 1.2, the angular resolution of an interferometer is given by:

$$\theta_{angular} = \frac{\lambda}{\|\vec{b}\|} \tag{1.3}$$

In an interferometer, we have a set of baselines, and as a result, different resolutions can be achieved. At a fixed wavelength, the shortest baseline gives the coarse resolution while the longest baseline gives us the finest resolution. This angular resolution is closely related to the fringe spacing produced by the interferometer. The relation between fringe spacing and angular resolution will be discussed in detailed in Section 3.2.

Let us consider a simple interferometry system consisting of two antennae separated by a vector $\vec{b}$ as shown in Fig. 1.5. The two radio antennas receive electromagnetic radiation from an astronomical source in the sky, which is in the direction of the unit vector $\hat{s}$. Now suppose that the source is sufficiently far away, such that the incident wavefront can be considered to be a plane wave. Both antennas will receive the same plane wavefront from the source of interest but it will not reach

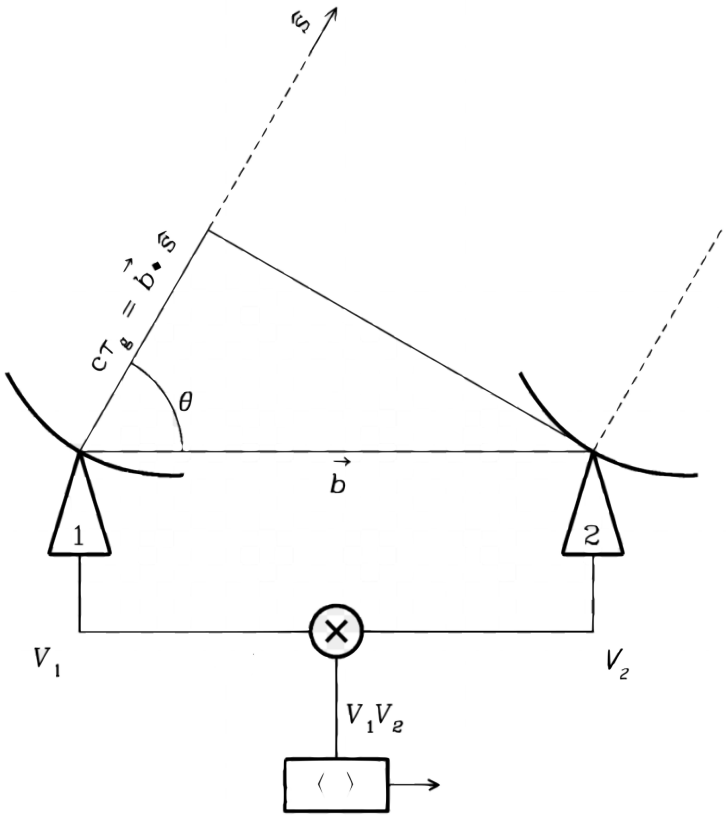


Figure 1.5: *A basic setup of interferometry with two antennas labelled 1 and 2. The vector $\vec{b}$ is the separation distance between the antennas which is the baseline, $c\tau_g$ is the extra distance that the signal must travel to reach antenna 1 where c is the speed of light, and τ_g is the delay. Vector $\hat{s}$ is the unit vector pointing to the source in the sky. V_1 and V_2 are the source output voltages as measured by each antenna, and X is the correlator device.*

both of them simultaneously due to the spacing between them. There will be a time delay for the signal to reach antenna 1 as annotated in Fig. 1.5. The term geometric delay is used to describe this phenomenon and is denoted by, τ_g. The angle between the unit vector $\hat{s}$ and the antenna separation vector $\vec{b}$, is θ.

In order to get the value of the delay, τ_g, we need to know the extra distance that the plane wave had to travel to reach antenna 1. Inferring from Fig. 1.5 and using the cosine rule and the product rule we can write the following equations respectively:

$$cos(\theta) = \frac{c\tau_g}{||\vec{b}||} \tag{1.4}$$

$$cos(\theta) = \frac{\hat{s} \cdot \vec{b}}{||\vec{b}||\,||\hat{s}||} \tag{1.5}$$

Equating Equation 1.4 to 1.5, and also using the fact that $\hat{s}$ is a unit vector we get:

$$\frac{c\tau_g}{\|\vec{b}\|} = \frac{\vec{s}\cdot\vec{b}}{\|\vec{b}\|} \tag{1.6}$$

$$\therefore \tau_g = \frac{\vec{b}\cdot\hat{s}}{c} \tag{1.7}$$

The correlator (denoted by letter X in Fig. 1.5) is a piece of hardware or software that combines the output voltage time series from antenna 1 and antenna 2 to get an output response, $C_{1,2}$. The correlator first multiplies voltages followed by averaging over time. If the output voltages from antenna 1 and antenna 2 are $V_1(t)$ and $V_2(t)$ respectively, then the output response of the correlator, C, can be written as

$$C_{1,2} = \langle V_1(t)V_2(t)\rangle, \tag{1.8}$$

where $\langle\cdot\rangle$ represents time averaging. Assuming that most of the energy of the wavefront is confined within a single frequency ν, we can represent the signal in the following form $V_1(t) = v_1 cos 2\pi\nu(t-\tau_g)$ and $V_2(t) = v_2 cos 2\pi\nu(t)$. The output of the correlator can then be written as follows:

$$C_{1,2}(\tau_g) = v_1 v_2 cos 2\pi\nu\tau_g. \tag{1.9}$$

We now represent the response of the interferometer in terms of the source brightness integrated over the sky. Suppose $S_i(\hat{s})$ is the intrinsic source brightness in the direction of unit vector $\hat{s}$ measured at frequency ν. Assuming that the bandwidth $\Delta\nu$ is sufficiently narrow such that the variation of $S_i(\hat{s})$ with ν is negligible, then the power of the signal in bandwidth $\Delta\nu$ from the source element $d\Omega$ can be calculated as follows (Thompson et al. 2017),

$$P = S_i(\hat{s})\Delta\nu d\Omega. \tag{1.10}$$

But, the actual power received by the correlator is attenuated by the primary beam (primary beam is the sensitivity of the telescope as a function of direction) of the dish in the direction of the unit vector $\hat{s}$, $A(\hat{s})$. We define the attenuated source brightness as $S_a(\hat{s}) = A(\hat{s}) \times S_i(\hat{s})$. Assuming that the variation of $A(\hat{s})$ with ν is negligible, we can represent the output from the correlator for the signal from a solid angle $d\Omega$ as

$$\begin{aligned} dC &= S_a(\hat{s})\Delta\nu d\Omega cos2\pi\nu\tau_g \\ &= A(\hat{s})S_i(\hat{s})\Delta\nu d\Omega cos2\pi\nu\tau_g. \end{aligned} \tag{1.11}$$

Taking into consideration that we have defined τ_g in terms of the baseline and the source position vector in Equation 1.7, we can now write Equation 1.11 as follows:

$$C = \Delta\nu \int_S A(\hat{s})S_i(\hat{s})cos\frac{2\pi\nu\vec{b}\cdot\hat{s}}{c}d\Omega \tag{1.12}$$

When making radio images, it is always common to specify the position of which the field of view is centred referred to as the phase centre. We define the vector $\hat{s}_0$ as the vector pointing to the phase centre such that $\hat{s} = \hat{s}_0 + \sigma$. By substituting $\hat{s}$ into Equation 1.12 and invoking the trigonometric identity [2] we can write:

$$\begin{aligned} C &= \Delta\nu \int_S A(\sigma)S_i(\sigma)cos[\frac{2\pi\vec{b}\cdot(\hat{s}_0+\sigma)}{c}] \\ &= \Delta\nu cos(\frac{2\pi\nu\vec{b}.\hat{s}_0}{c}) \int_S A(\sigma)S_i(\sigma)cos\frac{2\pi\nu\vec{b}\cdot\sigma}{c}d\Omega \\ &\quad -\Delta\nu sin(\frac{2\pi\nu\vec{b}.\hat{s}_0}{c}) \int_S A(\sigma)S_i(\sigma)sin\frac{2\pi\nu\vec{b}\cdot\sigma}{c}d\Omega \end{aligned} \tag{1.13}$$

Thus, we can now define the complex visibility (using Euler's formula) that the correlator will measure from the source in the sky as

$$V = \int_S A(\sigma)S_i(\sigma)e^{\frac{-2\pi i\nu\vec{b}\cdot\sigma}{c}}d\Omega. \tag{1.14}$$

[2] $cos(A+B) = cosAcosB - sinAsinB$

1.4 The Van Citter-Zernike Relation

In order to make use of Equation 1.14 in a practical way we need to define a new coordinate system. Let us define (u, v, w) as the coordinate system representing baseline length. Further, we can write these coordinates in units of wavelength as $\frac{\vec{b}}{\lambda} = (u, v, w)$ where u is the East-West component of the baseline and v is the North-South component of the baseline and w is pointing towards the phase centre. Similarly, we define, (l, m, n) as the coordinate system for the source direction vector, where l is the east-west direction on the sky, m is the North-South direction and the n component comes from the fact that $\hat{s}$ is a direction vector and has a unit length, hence, $l^2 + m^2 + n^2 = 1$. It follows that $n = \sqrt{1 - l^2 - m^2}$; Then we can write:

$$\begin{aligned} \frac{\vec{b}}{\lambda} \cdot \hat{s} &= ul + vm + w\sqrt{1 - l^2 - m^2} \\ d\Omega &= \frac{dldm}{\sqrt{1 - l^2 - m^2}} \end{aligned} \tag{1.15}$$

Following that $\hat{s}_0$ is a unit vector, we define the phase centre coordinates as follows, $\hat{s}_0$ = (0,0,1). Then, in order for us to shift the coordinates of the phase centre to be at the origin (0,0,0), we have to subtract one from the n component on the sky coordinates. Then, the response of an interferometer as a function of (u, v, w) integrating over the angular size of the source can be written as follows:

$$V(u, v, w) = \int\int A(l, m) S(l, m) e^{-2\pi i[ul + vm + w(\sqrt{1 - l^2 - m^2} - 1)]} \frac{dldm}{\sqrt{1 - l^2 - m^2}} \tag{1.16}$$

where $A(l, m)$ is the product of the responses of the primary beams of the antennas and $S(l, m)$ is the intrinsic brightness of the source. The term $A(l, m)$ and $S(l, m)$ are always grouped and being perceived as the intensity of the source such that Equation 1.16 can be written as follows.

$$V(u, v, w) = \int\int I(l, m) e^{-2\pi i[ul + vm + w(\sqrt{1 - l^2 - m^2} - 1)]} \frac{dldm}{\sqrt{1 - l^2 - m^2}} \tag{1.17}$$

If we are only interested in a small area on the sky then, the extent of l and m is sufficiently small such that $l^2 + m^2 \ll 1$. This implies that $\sqrt{1 - l^2 - m^2} \approx 1$, since $\sqrt{1 - l^2 - m^2} = (1 - (l^2 + m^2))^{\frac{1}{2}} = 1 + \frac{1}{2}(l^2 + m^2) + ...$, Equation 1.17 becomes:

$$V(u, v) = \int\int I(l, m) e^{-2\pi i(ul + vm)} dldm \tag{1.18}$$

Therefore, the measured visibility function $V(u,v)$ is related to the perceived brightness on the sky $I(l,m)$ via a 2D Fourier Transform as shown in Equation 1.18 which is called the **Van Cittert-Zernike** relation (Taylor et al. 1999).

The complex visibility of a single source is composed of two parts which are the visibility amplitude and the visibility phase. The former encodes apparent flux density, meanwhile, the latter encodes source position. In essence, by measuring the amplitude and phase of the visibility as a function of u and v we can make radio images. The points in the uv-plane are a result of various projections of the baseline.

1.5 Radio Emission Mechanisms

Celestial objects emit electromagnetic radiation as discussed in Section 1.1 and radio telescopes (Section. 1.2 & 1.3) can be used to observe the radio emission. Various physical processes are responsible for the different kinds of radio emission, as summarised in Fig. 1.6. In order to differentiate between radio emission from celestial objects and spurious emission that can affect an observation, an understanding of the radio mechanism is essential. In this section, we will discuss some of the radiation mechanisms.

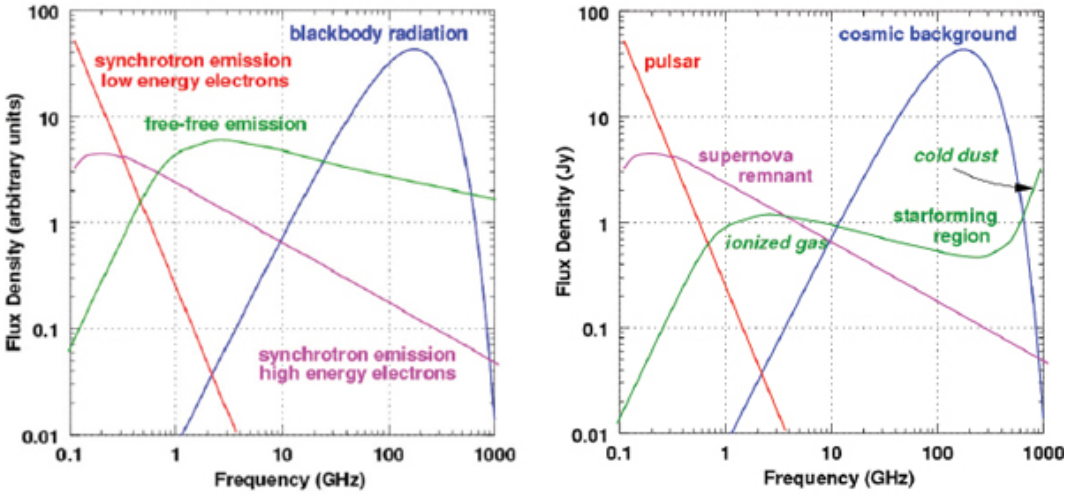


Figure 1.6: *Left: Comparison of continuum spectrum produced by different radiation mechanisms. Right: Examples of astronomical sources that give off the radiation via the mechanisms on the left figure (Council et al. 2007).*

1.5.1 Radio Continuum

Continuum radiation is distributed over a wide range of frequencies. This type of radiation arises from three major mechanisms, which are the thermal (blackbody), free-free and non-thermal emission (Council et al. 2007).

1.5.1.1 Thermal Blackbody Radiation

Thermal blackbody radiation is a form of continuum emission that all matter with a temperature greater than absolute zero emits. An ideal object that absorbs all radiation that it receives without transmitting or reflecting is referred to as a blackbody (Pratap & McIntosh 2005). The temperature of a blackbody can be estimated based on the frequency of the radiation it emits. This relationship is given by Wien's law,

$$\lambda_{max} = \frac{const}{T} \tag{1.19}$$

Where λ_{max} is the wavelength of the radiation that gives the maximum intensity, T is the temperature of the object, and $const$ is the Wien's displacement constant. The Stefan-Boltzmann Law gives us the power that a blackbody radiates summed over frequencies in the EM spectrum:

$$P = AkT^4 \tag{1.20}$$

Where A is the surface area of the object, T is the temperature of the object, and k is the Boltzmann constant. From Equation 1.19 and 1.20 one can notice that as a blackbody becomes hotter, the wavelength of the radiation that gives the peak intensity decreases and the brightness increases. Thus, blackbody objects have to be sufficiently cold to radiate in the radio; hence, there are very few astronomical objects that radiate thermal radiation in radio and can be detected with modern instruments.

One example that gives off thermal radiation is the cosmic microwave background (CMB). This is leftover radiation from the Big Bang. It is observable in all directions, with almost uniform intensity and is the most perfect blackbody ever observed. The CMB is the earliest radiation that humans can detect, it carries with it a snapshot of the physical conditions in the universe at the epoch at which matter and radiation decoupled.

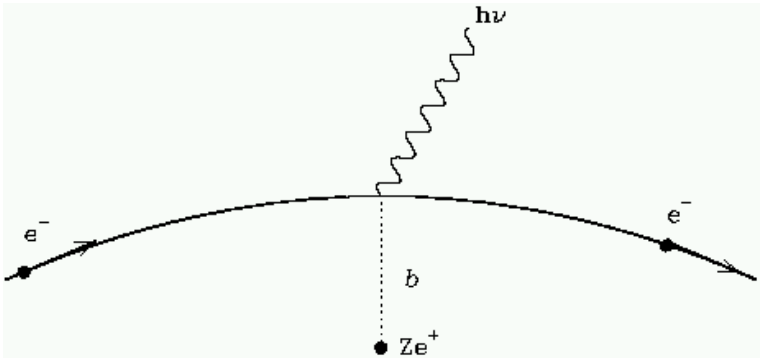


Figure 1.7: *Free-Free emission. An electron e^- tend to recobine with an ion, Ze^+. As they accelerate towards each other, the electron gets deflected by the electric field of an ion, hence produces EM radiation, hν. (Source: http://www.astro.utu.fi/~cflynn/astroII/l3.html).*

1.5.1.2 Free-Free Emission

Ionised gas emits thermal radiation via a different mechanism to the thermal blackbody. Atoms in a gas get ionised when they collide with another atom that has sufficient energy to knock off an electron from the atom, creating a negatively charged electron and positively charged ion. After the separation, the charged particles tend to recombine, and as they accelerate towards each other, the electron gets deflected by the electric field of an ion, hence produces EM radiation (Miller 1998). The deflection of an electron by an ion tends to keep the atoms separated again, making the process continue indefinitely. The rate of recombination depends on the density and the acceleration of the electrons, Fig 1.7 depict the phenomenon. Ionised interstellar gas (e.g. an HII regions) is an example of an astrophysical environment that the free-free radiation mechanism (Miller 1998) governs the emission.

1.5.1.3 Non-Thermal Radiation

In contrast to thermal radiation, the non-thermal radiation mechanism are not dependant on the temperature of the object. Theoretically, many radiation mechanisms can account for non-thermal emission from radio sources, but in practice, synchrotron emission dominates the non-thermal radiation. The synchrotron emission depends on the interaction of charged particles with strong magnetic fields. When a particle travelling with relativistic speed encounters a magnetic field, the particle accelerates along the spiral path following the magnetic field lines and radiates energy (Pratap & McIntosh 2005). Here, the intensity of the radiation is related to the strength of the magnetic field and the energy of the charged particle

within the field. Fig. 1.8 is a schematic diagram showing the formation of synchrotron radiation, as explained above. Some of the astronomical objects that give off synchrotron radiation are pulsars, supernova remnants and active galactic nuclei (AGNs).

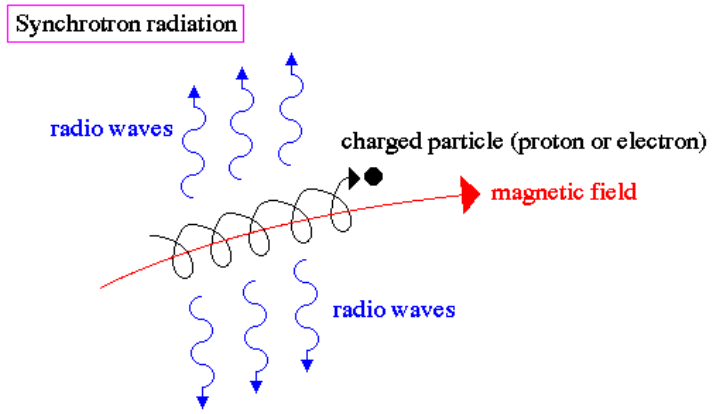


Figure 1.8: *Schematic diagram illustrating the formation of synchrotron radiation. (Source:* `http://abyss.uoregon.edu/~js/glossary/synchrotron_radiation.html` *)*

The radio sky at short wavelength is dominated by sources that emit thermal radiation, while non-thermal processes dominate at long radio wavelengths as shown in Fig. 1.6. One of the fundamental differences between thermal and synchrotron radiation is that with thermal radiation the energy or intensity of the radiation increases with the frequency, whereas with synchrotron radiation an opposite effect is observed as shown in Fig. 1.6.

1.5.2 Line emission

Spectral line radiation is emitted in a very narrow frequency range, as opposed to continuum radiation which covers a wide range of frequencies. Here, an electron within the atom transits to a different energy level and emits radiation that is characterised by the energy which is required for the transition to occur. Molecules or atoms can produce spectral lines, and the transition happens when an atom collides with an atom or electron. Examples of some spectral lines which have been detected at radio frequencies are the 1612 MHz line of hydroxyl (OH), the 230 GHz carbon monoxide spectral line (CO), the 22 GHz water spectral line H_2O, and the

1420 MHz line which is emitted by the neutral hydrogen (HI).

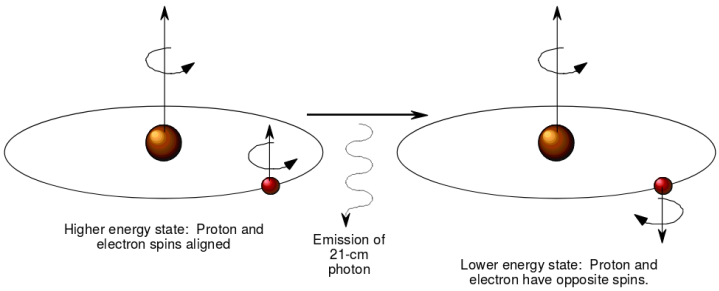


Figure 1.9: *Formation of the 21-cm Line of Neutral Hydrogen (Miller 1998).*

The 1420 MHz line is one of the familiar lines in astronomy due to the abundance of hydrogen in the Universe. When a hydrogen atom gains energy via a collision, the spins of the proton and electron in the hydrogen atom align and point in the same direction (i.e. spins are parallel), leaving the atom in a slightly excited state. When the atom returns to its ground state and the spins of the proton and electron point in opposite directions (i.e. anti-parallel) and a 1420 MHz photon is emitted, Fig. 1.9 depicts this phenomenon. One of the major science surveys planned with the MeerKAT is Looking At the Distant Universe with the MeerKAT Array (LADUMA), which will do ultra-deep HI observations.

Table 1.1 gives an overview of the typical radiation processes in the radio spectrum and the environment which they are found in.

Wavelength	Spectral line	Continuum
metre, cm and mm	• Neutral Hydrogen (HI) 21cm fine structure line-neutral gas • Hydrogen recombination lines -ionized gas • OH, H_2O, SiO Masers - dense, warm molecular gas • Molecular rotation lines - cold molecular gas	• Thermal Bremsstrahlung (free-free emission) – HII regions • Synchrotron Radiation – Jets in radio Galaxies, pulsars, shocks in supernovae, cosmic ray electrons in the magnetic fields of normal galaxies , acceleration of electrons in stellar and planetary systems • Thermal emission from dust – cold, dense gas.
sub-mm and far infrared	• Molecular Rotation Lines – warm, dense gas. • Solid State features (silicates) – dust • Hydrogen recombination lines – ionized HII regions.	• Thermal emission - warm dust

Table 1.1: *Overview of common radiation process in the radio spectrum and their environment highlighted by the red colour in the text.*

1.6 MeerKAT

Radio telescopes such as MeerKAT are used to detect radiation from astronomical objects as discussed in the previous sections. In this section, we will present an overview of the part of the MeerKAT telescope that is relevant to our work.

1.6.1 Overview

The MeerKAT is an interferometry array of 64 radio dishes. The South African Government has built MeerKAT under the administration of the Department of Science and Innovation (DSI) as a demonstrator for the implementation scenario for the Square Kilometer Array (SKA) radio telescope. MeerKAT is located in the Karoo desert in the Northern Cape province of South Africa at approximately $21^{\circ}23^{'}$ East, $30^{\circ}42^{'}$ South.

Figure 1.10: *A picture showing some of the MeerKAT antennas in the Karoo Northern Cape (Source: https://www.ska.ac.za/gallery/meerkat/).*

MeerKAT is the most sensitive radio telescope of its kind (see Table. 1.2) and will observe the sky with unprecedented depth and detail (Camilo et al. 2018). Sensitivity is a measure of the weakest signal that an instrument can distinguish above the random background noise. For an interferometer with N number of antennas, the sensitivity is given by the following equation (Taylor et al. 1999):

$$S_{rms} = \frac{2kT_s}{A\eta_A\eta_Q\sqrt{2N(N-1)\Delta\nu t_{int}}} \tag{1.21}$$

where k is Boltzmann's constant, T_s is the system temperature, A is the geometric antenna area, η_A is the antenna efficiency, η_Q is the correlator efficiency, $\Delta\nu$ is the observation bandwidth and t_{int} is the integration time. Equation 1.21 can be simplified to:

$$S_{rms} = \frac{SEFD}{\sqrt{2N(N-1)\Delta\nu t_{int}}} \tag{1.22}$$

where $SEFD$ is an acronym which stands for System Equivalent Flux Density. It encodes the efficiency, the collecting area and the system temperature of the antenna. An increase in the collecting array (i.e number of antennas (N)), increase of the bandwidth and the decrease of system temperature result in better sensitivity.

In order to determine the sensitivity of the MeerKAT telescope, special drift scan observations of the Crab Nebula were carried out. The values of the $SEFD$ as a function of frequency were measured, Fig.1.11 shows the best estimate of the $SEFD$.

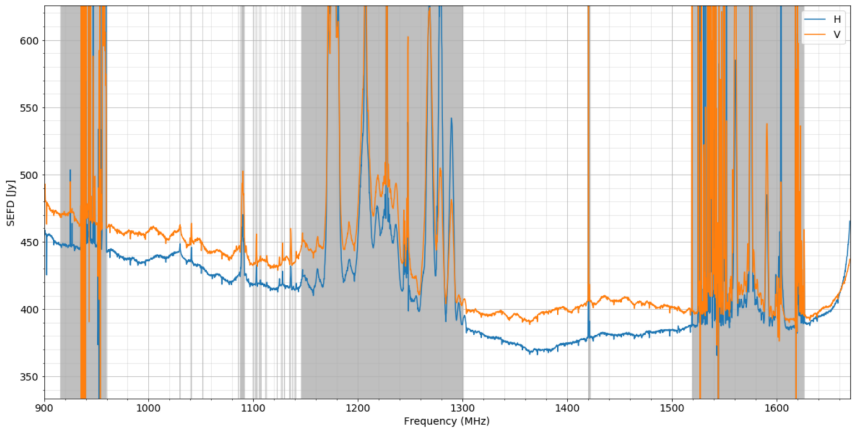


Figure 1.11: *The SFED for MeerKAT]System Equivalent Flux Density (SEFD) for the MeerKAT antenna at L-band. The grey areas are the removed frequencies due to presence of known radio frequency interference. The orange and the blue curves are the vertical and horizontal polarisations respectively (Source: https://drive.google.com/file/d/168KP4W9qtc2H7Too2-6cMyxVSZACnVy5/view).*

The MeerKAT L-band receiver has a bandwidth of 856 MHz and the measured average $SEFD$ of one receptor on the cold sky is approximately 460 Jy (Camilo et al. 2018). Thus, we can now approximate the MeerKAT sensitivity for a 10 minutes observation as follows:

$$S_{rms} = \frac{460Jy}{\sqrt{2 \times 64(64-1) \times 856 * 10^6 s^{-1} \times 600s}} \tag{1.23}$$

$$= 7.15\mu Jy \tag{1.24}$$

Taking into account that about 40% of the MeerKAT L-band is corrupted due to radio frequency interference as evident from Fig. 1.11 , then the sensitivity becomes:

$$S_{rms} = 9.23\mu Jy \tag{1.25}$$

Table 1.2 shows the sensitivity of some of the well known radio telescopes. The key properties such as number of antennas and the SEFD were taken from the official websites of each of the telescope.

Telescope name	**Number of antennas**	S_{rms} **(μJy)**
VLA	27	15.1
uGMRT	30	45.0
ASKAP	36	33.4

Table 1.2: *Sensitivity of some of the well known radio telescopes (Very Large Array (VLA), Upgraded Giant Metrewave Radio Telescope (uGMRT) (Gupta et al. 2017) and the Australian Square Kilometre Array Pathfinder (ASKAP)).*

1.6.2 Specification

The diameter of each MeerKAT antenna is 13.5 m. The array has a dense core with 48 antennas located within a diameter of 1 km, and the remaining 16 are spread out giving a maximum baseline length of 8 km. Fig. 1.12 shows the spatial distribution of the antennas.

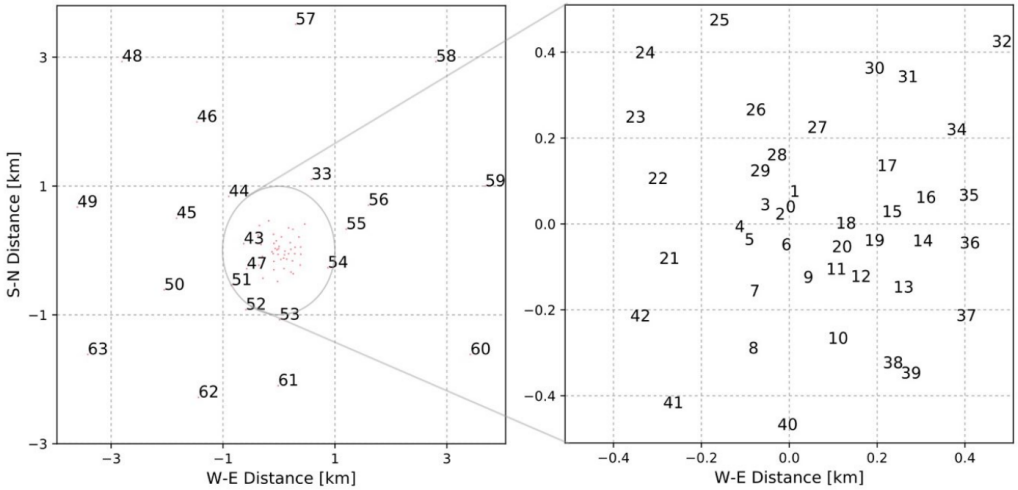


Figure 1.12: *MeerKAT antenna spatial distribution.]Left plot: MeerKAT antenna distribution, with each number representing the antenna. Right: Zoom-in of the central 1 km in diameter dense core with 48 antennas (Asad et al. 2019).*

The MeerKAT dishes use an offset Gregorian configuration. This kind of design was chosen over the competing symmetric centre-fed because it ensures excellent optical performance, the sub-reflector (Annotated sub-reflector in Fig 1.13) is positioned in a way that it does not block radiation from reaching the main reflector.

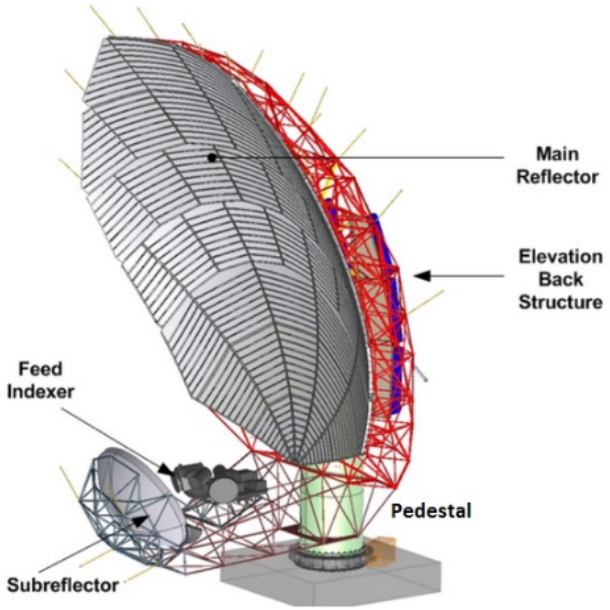


Figure 1.13: *Schematic diagram of the MeerKAT showing the major elements of the dish (Jonas et al. 2018).*

The position of the sub-reflector and the way it is shielded (see wires underneath the sub-reflector in Fig 1.13) ensures the rejection of the man-made signals from the ground. The MeerKAT antennas can accommodate up to four receivers and digitisers situated near the sub-reflector. Currently, the L-band receiver that covers frequency ranging from 900 - 1670 MHz and the UHF-band receiver which covers (580 - 1015 MHz) frequency range are installed, with the S-band (1.75 to 3.5 GHz) planned to be installed soon (Camilo et al. 2018). All three receivers are linearly polarised.

1.6.3 Data Transfer

The receiver captures the radio signal and converts it into a voltage which is then filtered and amplified (Foley et al. 2016). The amplified voltage is then digitised (at the receptor itself) and sent to the correlator/beamformer (CBF) situated at the Karoo Array Processor Building (KAPB) via underground optical fibres (Asad et al. 2019). The KAPB building is located 12 km away from the core. The correlator implements the FX/B signal processing style (Camilo et al. 2018). As explained in (Mauch et al. 2020) the F-engine coarsely aligns the voltages and corrects for both geometrical and instrumental delays and splits the data into frequency channels. The aligned voltages from pairs of antennas can undergo various processes like

the correlation of the signal by the X-engine or beamforming of the signal by the B-engine (Mauch et al. 2020). The raw visibilities are then transferred via a 10 Gb/s optical-fibre backhaul from the KAPB to the MeerKAT archive hosted by Center for High-Performance Computing (CHPC) in Cape Town (Camilo et al. 2018). The MeerKAT Science Data Processing (SDP) team is responsible for doing quality control and quality assurance of the data before it is released to astronomers.

1.6.4 Early Science with MeerKAT

The MeerKAT radio telescope has already been publishing new science results, even though the Radio Frequency Interference (RFI) as measured by the telescope has not yet been fully understood. A detailed overview and analysis of the RFI will be discussed in Chapter 3. In this section we will discuss some of the early great science done using the MeerKAT telescope. During the first few years of doing science with MeerKAT, 70% of the observation time will be allocated to Large Survey Projects (LSPs), which require a minimum of 1000 hours of observing time per project. The remaining 30% of observation time will be allocated to the smaller PI driven proposals. The major science surveys with MeerKAT include ultra-deep HI observations and the testing of Einstein's theory of gravity and gravitational radiation via measurements of pulsars.

MeerKAT was inaugurated on the 13^{th} of July 2018 by the Deputy President of the Republic of South Africa, Mr David Mabuza. The inauguration was coupled with the unveiling of an image produced from MeerKAT data (Fig. 1.14), which shows unprecedented detail of the region surrounding the black hole, four million times massive than our Sun, at the centre of our Milky Way Galaxy. The colour scale ranging from white through yellow to orange and faint red represents the strength of radio emission.

Before the MeerKAT telescope, the region imaged (Fig. 1.14) used to be obscured from the Earth by clouds of gas and dust. The image provides the view of what is happening around the black hole, and it shows a star-forming region, supernova remnant, and also the narrow filaments, which are for the first time being detected.

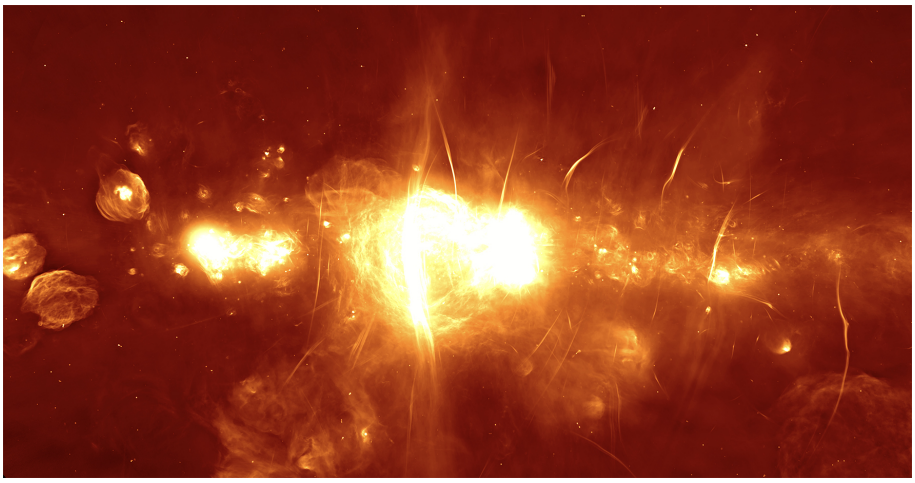


Figure 1.14: *Image showing the clearest view ever of the center of the MilkWay galaxy as measured by the MeerKAT telescope. (Source: SARAO communication team.)*

1.6.5 The MeerKAT site

The Karoo desert was chosen for the MeerKAT site because of its low population density, low economic activity, and the site is shielded by hills. Thus, the site is considered to be a quiet radio zone. The South African government has declared part of the Northern Cape Province as a Radio Astronomy Reserved Zone under the Astronomy Geographic Advantage Act, 2007 (Act No. 21 of 2007); in particular where the MeerKAT telescope is located. Avoidance of man-made instruments that produce radio emission was the main factor used to choose the MeerKAT site, which will also form the core of the SKA in Africa.

To carry out their research, astronomers need access to part of the spectrum which is least affected by the Radio Frequency Interferenc. Unfortunately, due to industrial and technological developments, radio astronomy observations are being threatened by increasing levels of radio emission from man-made instruments. Even though the MeerKAT telescope is built in a quiet radio zone, sources such as satellites, aeroplanes and the instruments of the telescope itself produce radio emission that cannot be entirely avoided (Ford & Buch 2014). Therefore, astronomers require other techniques to filter such unwanted radio signals from their data.

1.7 Radio Frequency Interference in Radio Astronomy

In radio astronomy, Radio Frequency Interference (RFI), is defined as any signal received by a radio telescope which does not come from an astronomical source (Offringa et al. (2010), Ellingson (2005)). Radio astronomers are facing an increasing challenge of RFI due to the increased number of transmitters and broader bandwidth of radio observatories (Hamidi et al. 2011). Radio emission from human activity is orders of magnitude brighter than that of astronomical sources. Since MeerKAT has a large effective collecting area of approximately 1960.88 m^2, it has high sensitivity and as the sensitivity of the instruments increases, so does the sensitivity to unwanted signals (i.e. satellites, mobile cellphones, aeroplanes).

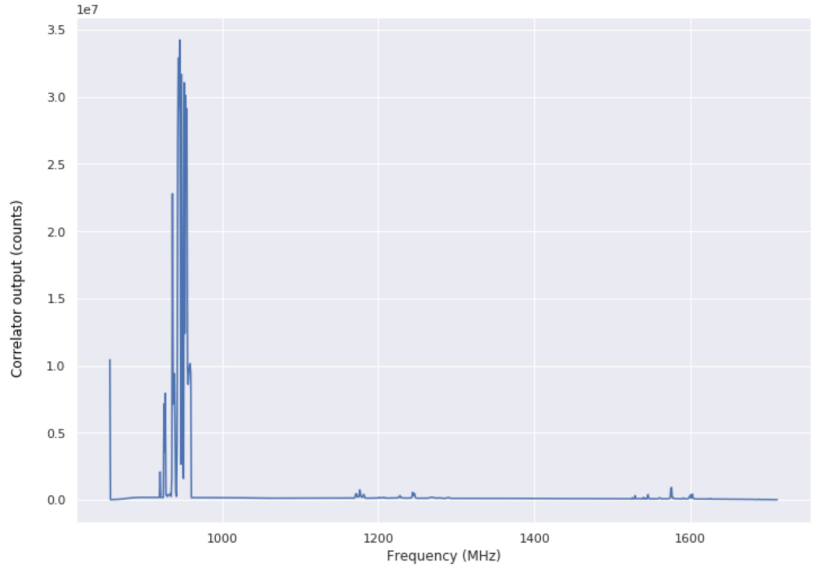
For continuum observations, it is possible to remove some of the data which is contaminated by RFI and use the remaining clean data for science. However, for spectral line observations, it is impossible to remove the culprit if it is emitting precisely at the same frequency as the astronomical source of interest.

1.7.1 Types of Radio Frequency Interference

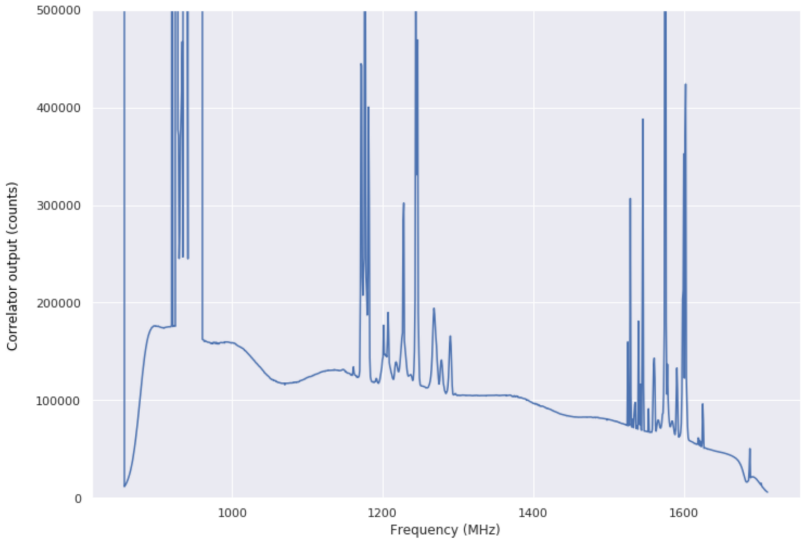
There are different categories of RFI sources that pose a problem to radio astronomy. The two main types of RFI sources are channelised and broadband. Channelised RFI sources are often present for a long duration but in few frequency channels. Broadband RFI sources, in contrast, occur typically for a short duration but across many observed frequencies (Fridman & Baan 2001).

Examples of such interference are respectively, aircraft communication system and lightning. Some of the RFI sources are present 100% of the time in observation, hence rendering the data unusable. Fig. 1.15 shows a typical cross-section of a MeerKAT bandpass with some of the band being contaminated by RFI. Due to strong RFI sources between 935 MHz and 960 MHz, it is difficult to see some of the much weaker sources, therefore Fig. 1.15b shows the same data with reduced y-axis range.

Three of the major known RFI culprits categories of the MeerKAT telescope in the L-band region are shown in Table. 1.3, corresponding to Global System for Mobile Communication (GSM), the aircraft Distance Measuring Equipment (DME) and the Global Positioning System (GPS) satellites.



(a) *MeerKAT bandpass with strong RFI between 935 MHz and 960 MHz.*



(b) *Same plot as Fig. 1.15a with reduced y-scale showing some of the weaker sources at around 1190 MHz - 1300 MHz and 1500 MHz - 1620 MHz that were not clearly visible in Fig. 1.15a due brighter RFI source between 935 MHz - 960 MHz.*

Figure 1.15: *The cross-section of the typical MeerKAT bandpass with some part of the band corrupted by RFI.*

1.7.2 Detection and Mitigation of Radio Frequency Interference

The MeerKAT L-band digitiser samples at a rate of about 1.7×10^6 samples per second, that approximates to 2 gigabytes of data being sampled every second. This large volume and high data rate make the detection and mitigation of RFI one of

RFI Sources	Band-width(MHz)	Central Frequencies (MHz)
GPS satellites: L1 L2 L3 L4 L5	 1567.75 - 1583.10 1217.10 - 1228.10 1376.05 - 1386.05 1374.90 - 1382.90 1170.02 - 1182.70	 1575.42 1222.60 1381.05 1379.90 1176.45
Mobile Cellphones: GSM-900 Uplink GSM-900 Downlink GSM-1800 Uplink	 890.00 - 915.00 935.00 - 960.00 1710.00 - 1785.00	
Aircraft Communications (DME): Ground-to-Air 1 Air-to-Ground Ground-to-Air 2	 962.00 - 1024.00 1025.00 - 1150.00 1151.00 - 1213.00	

Table 1.3: *Summary of major RFI contaminated regions of the L-band, for MeerKAT telescope. The GSM and the DME band does not have a central frequency (Source : SARAO internal documents: T. Mauch)*

the biggest challenges when analysing the observational data collected from the current generation of telescopes such as MeerKAT. The protection of current radio observatories from RFI is of great importance to astronomers. This requires a package of measures including among other things global regulatory protection, strong national and local protection, and efficient RFI mitigation techniques. Several techniques have been exploited to perform the challenging task of the mitigation, detection and excision of RFI from astronomical data. In the following sub-sections, we will discuss in general some of those techniques and regulations.

1.7.3 Pre-observation

The South African government has put laws in place to protect the MeerKAT and SKA radio astronomical observatory against man-made RFI signals. There are spectrum management policies which are aimed at reducing the observatoryś vulnerability. The spectrum management policies prohibit some of the radio emission, put restrictions on allowed power levels of RFI sources and also facilitate access to the alternative radio communication channels to the local people. For example, the government has helped the residents from the nearby towns such as Carnarvon, Vanwyksvlei and Brandvlei to migrate from the transmission of analogue signals to digital.

1.7.4 During observation (High time resolution RFI detection)

High time resolution RFI detection happens before averaging of data. Section 2.1.1 details in depth how this step is implemented in the MeerKAT RFI detection pipeline. It is a crucial step before data averaging, the only step that you can access the high time resolution data. The technique used on this stage need to process huge amount of data over a short period and, as a result, this process is computationally expensive. Thus, the methods used here can only process limited amount of data, for example data from a single scan or a single baseline (Offringa et al. 2010).

1.7.5 Post observation

Regardless of the several techniques that astronomers have developed and put into place to mitigate RFI signals from the observational data, and government laws and regulations to protect radio astronomy observatory from RFI, nearly all the data recorded into a disk will require extra processing to detect the presence of RFI. Astronomers use flagging techniques to mitigate RFI before doing any science. Flagging is the masking of the unwanted RFI signals from data in time, frequency and antenna space. The points which are believed to be contaminated with RFI are marked but not removed from the data set; thus, it is left as an astronomers choice to enable or disable the flags of a given data set depending on their science goals.

Traditionally, the flagging was done by hand. Hence, the process was time-consuming and tedious. With the advancement of astronomical instruments, the amount of data produced is increasing exponentially. The old techniques of flagging RFI signals by hand were no longer sufficient. More sophisticated and automated techniques were required. Astronomers have developed automated methods to flag data that uses thresholding techniques. One of the commonly used thresholding algorithm in radio astronomy is AOFlagger. The MeerKAT team uses a similar framework. Below we will briefly describe the AOFlagger algorithm.

- **AOFlagger**

  AOFlagger is a framework originally developed for the Low-Frequency Array (LOFAR) that implements several methods to deal with RFI, the detailed description of the methods can be found in Offringa et al. (2010). As already

mentioned, some of the RFI signals are orders of magnitude brighter than the astronomical signals, therefore thresholding techniques are used to remove such bright RFI sources. Threshold methods check if the data point in question is outside the statistical distribution of the overall data, and if the check is true then the data is flagged as bad and is not used further in the pipeline of data processing. The threshold value can be set in different ways. One can choose to set it globally or set it as a function of variance per baseline. The setback of using only the thresholding method is that good data that is outside the specified threshold can be flagged as bad data. Thus, as a result, we run a risk of throwing away good astronomical data.

In order to avoid throwing away good data, surface fitting and smoothing method are used. This method works under the assumption that the amplitude of astronomical continuum sources does not change rapidly as a function of frequency and time. However, the amplitude of RFI sources varies rapidly as a function of time and frequency. Hence, if we can fit a smooth function to the output visibility $V(\nu, t)$ we can produce $\hat{V}(\nu, t)$ which encodes the information about the signal from the source of interests. The difference between the fit and data is the system noise $N_{noise}(\nu, t)$ and the RFI signal $N_{RFI}(\nu, t)$, unlike the blind thresholding method discussed above, here we reduce chances of flagging good astronomical data with a higher amplitude as a corrupt data.

Let us define the the surface that encodes the astronomical information as two dimensional, low-order, dimension-independent polynomial that is iteratively fitted to time-frequency tiles in the data using a least-squares fit as,

$$\hat{V}_k(\nu, t) = \sum_{i=1}^{N_\nu} a_{k,i}\nu^i + c\sum_{i=1}^{N_t} b_{k,i}t^i + c_k \tag{1.26}$$

where k is the tile, N_t and N_ν are the polynomial order for time and frequency respectively, and $a_{k,i}$, $b_{k,i}$, c_k are the coefficient of the fit.

The samples which are detected as RFI in tile k are not used in the fit of tile $k+1$. This can be achieved by introducing a weight function, $W_F(\nu, t)$, such

that if $W_F(\nu, t) = 0$ then the value will not be used in next iteration, and if $W_F(\nu, t) = 1$ then the value is accepted. The goal is to find the best fit to the data by minimising the error function E_k for each tile:

$$E_k = \sum_{\nu} \sum_{t} W_F(\nu, t)[\hat{V}_k(\nu, t) - V(\nu, t)]^2 \tag{1.27}$$

The surface created via this method to an extent does represent the astronomical information. However, this method does not generalise well at the boundaries of the tiles as polynomial fits tend to show deviation at the boundaries. To solve the problem, sliding window methods are used. Sliding window methods are considered to more accurate compared to tile-based methods.

The sliding window method works by calculating an average or median of a window of size $N \times M$. *AOFlagger* algorithm uses weighted average. Let us introduce a weight function $W_d(i, j)$ that is dependent on two components i and j that represents the distances from the centre of the window in time and frequency respectively. Then we can define the surface fit as,

$$\hat{V}(\nu, t) = \frac{\sum_{i=-\frac{1}{2}N}^{\frac{1}{2}N} \sum_{j=-\frac{1}{2}M}^{\frac{1}{2}M} W_d(i, j)(W_F \odot V)(\nu_i, t_j)}{weight} \tag{1.28}$$

where

$$weight = \sum_{i=-\frac{1}{2}N}^{\frac{1}{2}N} \sum_{j=-\frac{1}{2}M}^{\frac{1}{2}M} W_d(i, j) W_F(\nu + i\Delta\nu, t + j\Delta t). \tag{1.29}$$

Equation 1.28 and 1.29 are a convolution operation $W_d \circledast (W_F \odot V)$ and $W_d \circledast W_F$ respectively, which gives

$$\hat{V} = \frac{W_d \circledast (W_F \odot V)}{W_d \circledast W_F}, \tag{1.30}$$

where $\odot$ is an element-wise multiplication operator and $\circledast$ is the convolution operator. A two-dimensional Gaussian function is chosen to be the best choice for the weight function W_d, which is defined as follows:

$$W_d(i,j) = e^{-\frac{i^2}{2\sigma_v^2} - \frac{j^2}{2\sigma_t^2}} \tag{1.31}$$

Substituting Equation 1.31 into Equation 1.30 we get a weighted surface fit with Gaussian smoothing.

As already discussed in Section 1.7.1 RFI can be classified into two main categories which are the broadband and the narrow band. The broadband RFI implies that RFI samples are often connected in frequency space, whereas narrowband indicate that RFI samples are connected in time. We now introduce a better thresholding technique that makes use of this knowledge instead of flagging samples individually (i.e. per frequency channel, per time stamp). Here, a sample combination will be flagged when that combination is beyond a certain threshold. Let us assume that A and B are adjacent samples. In standard thresholding discussed above samples are treated individually, a sample would be flagged if it is outside specified threshold range χ_1.

If sample A or B do not exceed the single sample threshold χ_1, they can still be flagged if the combination of A and B is greater than some lower threshold χ_2. If the combination of A and B does not exceed the threshold χ_2, they can be combined with third adjacent sample C, and be thresholded at χ_3. The more connected the samples the lower the threshold.

Because the threshold, $\{\chi_i\}_{i=1}^{N}$, is strictly decreasing, a value would be flagged as RFI if it belongs to the combination of i values, in which all the values are above the threshold χ_i. The following rule is applied to determine as to whether the sample, $R(\nu,t)$, should be flagged because of the RFI sequence it belongs to, in frequency space.

$$\begin{aligned} flag\nu_m(\nu,t) = \exists i \in \{0...M-1\} : \forall j \in \{0...M-1\} : \\ |R(\nu + (i+j)\Delta\nu, t)| > \chi_M, \end{aligned} \tag{1.32}$$

where M is the number of samples in a combination. This method is known as the VarThreshold method. The MeerKAT RFI detection pipeline uses a

variation of the VarThreshold method which is called SumThreshold method and shall be discussed in detailed in Section 2.1.2.

1.8 Objectives

Radio frequency Interference is one of the important steps in data reduction pipelines for radio telescopes. This because RFI corrupts the scientific data, hence, reducing the quality of the scientific insight that could be possible extracted from the telescope in question.

Currently, there is no way to keep track of RFI changes as measured from the MeerKAT telescope or to quantify the RFI environment statistically around the MeerKAT site using the telescope data. This has prompted the need for a new data product for the MeerKAT telescope that would summarise the statistics of the RFI environment for the site.

To provide the RFI statistics as measured from historical MeerKAT observations, we propose to design and implement a framework that will allow the historical RFI statistics from MeerKAT to be easily accessible by various stakeholders. We will design a framework that produces a statistically quantified data set for MeerKAT RFI environment. We call our framework **MeerKAT Historical Probability of Radio Frequency Interference** which we will refer to as **KATHPRFI** from now on. The following are the key deliverables of this project.

- Software that computes the probability of RFI occupancy as a function of frequency, time of the day, baseline, elevation and azimuth angle.
- A dataset that can be used as a baseline for alert triggering when unusual events occur.
- Study the RFI environment for MeerKAT.
- Study time-evolution of the RFI environment.

Chapter 2

Methodology

This chapter gives an overview of how the KATHPRFI framework (discussed in Section 1.8) was designed, with details pertaining to the design choices, as well as software and hardware constraints. We start by discussing the MeerKAT Science Data Processing (SDP) pipeline in Section 2.1. The MeerKAT visibility data structure is discussed in Section 2.2. The extraction, transforming, and the loading of the data is covered in Section 2.3. We describe the potential use case of the KATHPRFI in Section 2.4. The goals of KATHPRFI framework are discussed in Section 2.5 and we further show our design approach in Section 2.6. The resources used and the limitations of the available resources are discussed in Section 2.8 and 2.9 respectively. Finally, we describe the product in Section 2.10. The full code of the framework can be found at the author's Github repository (https://github.com/SihlanguI/kathprfi).

2.1 MeerKAT SDP Pipeline

A brief description of the MeerKAT data flow and data transfer was given in Section 1.6. Figure 2.1 is a schematic diagram showing the data flow from the antenna to the final visibility product, through the ingest. Data coming from the digitiser (DIG) is sent to the correlator/beamformer (CBF). These are further piped into the ingest at half a second dump period which then processes the data to produce the visibility data product (which is usually averaged to 8s or according to the user's observing parameters- known as $L0$). A calibration pipeline is run on the $L0$ visibilities to produce the calibrated visibility, known as the $L1$ data product. A redis database[1] from telescope state (*telstate*) that contains meta-data data is finally writ-

[1] https://redis.io/

ten into the archive in **.rdb** format.

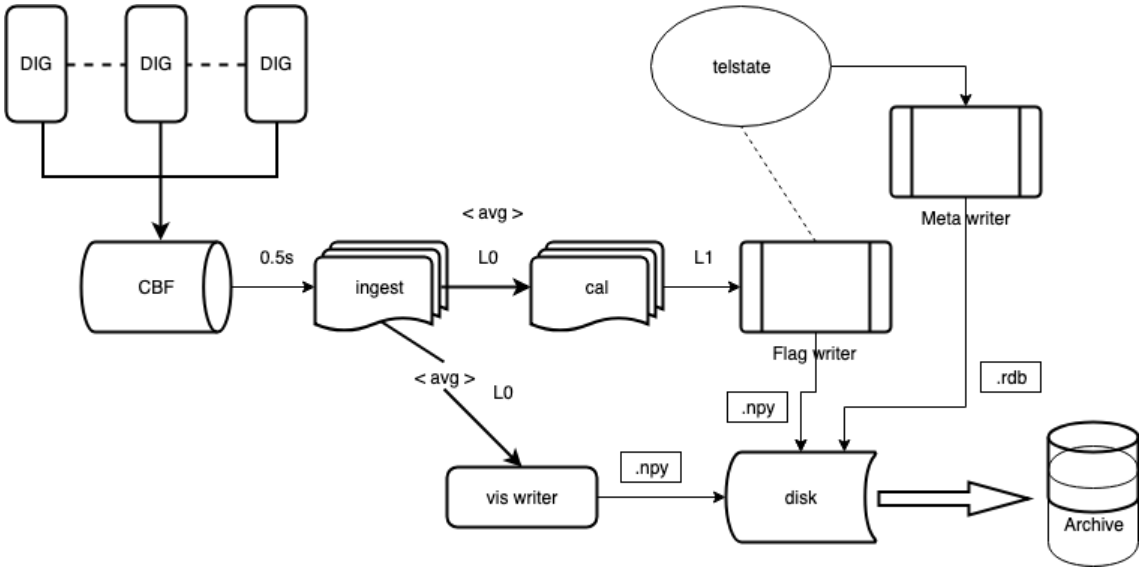


Figure 2.1: *Schematic diagram showing MeerKAT data flow from the telescope digitizer to the final visibility product in the archive (Adapted from the MeerKAT SDP documents.).*

Below is the definitions of the acronyms used in Fig. 2.1 :

- DIG: Is the digitiser that converts the analogue signal that is measured by the antenna to a digital signal.
- CBF: The Correlator Beam Former (CBF), is a piece of hardware that correlate signals from multiple antennas.
- Ingest: It is a processing unit where the high time RFI detection algorithm is run on the correlated signal from the CBF, as explained in subsection 2.1.1.
- *L*0: These are time-averaged CBF uncalibrated visibilities
- Cal: The SDP RFI detection algorithm and the initial calibration pipeline that is run on *L*0 visibilities to produce the *L*1 visibilities and flags.
- *L*1: The calibrated visibilities product.
- <avg>: Is the time averaging processing
- *.npy*: Serialized Numpy arrays that have the actual data that is stored in the archive.

RFI detection in the MeerKAT SDP pipeline happens at two stages, that is the ingest step and the calibration step. Below we shall discuss each of those stages.

2.1.1 MeerKAT High time Resolution RFI detection

The MeerKAT high time resolution RFI detection happens during the ingest step annotated as *ingest* in Fig. 2.1. Here, the strong RFI is detected by checking for outliers in the frequency axis in individual correlator dumps. At this stage averaging of data is carried out, samples which are detected as RFI are excised, only unflagged samples are averaged in time as per the observation requirement and further used in the data processing pipeline. The output of the ingest step is, therefore, an averaged RFI excised data-set with pertinent meta-data stored in *telestate*.

To account for data loss due to the ingest excision each visibility data point has an associated weight ($\mathbf{W_{SDP}}$) [2] which tells us how many samples were averaged to produce that visibility. Let us define **N** as number of correlator samples, $\mathbf{V_{CBF}}$ as the visibility sample from the correlator/beamformer (CBF), **U** as the set of indices of unflagged correlator visibilities. Thus we can calculate the SDP visibility sample $\mathbf{V_{SDP}}$ as follows:

$$V_{SDP} = W_{SDP} \sum_{i \in U}^{N} V_{CBF[i]} \tag{2.1}$$

where i represents correlator/beamformer sample index and $W_{SDP} = \frac{1}{N_U}$, with N_U being defined as the number of unflagged samples by the ingest. The ingest flags become **TRUE** when all N samples are flagged as RFI by the ingest, then at this point there is no excision of data. With partial flagging or excision of data, the ingest flags becomes **FALSE**.

The ingest RFI detection algorithm usually only detect narrow regions around the brightest RFI spikes in the data, and further flagging is required.

2.1.2 The MeerKAT RFI Flagger

The MeerKAT in-house developed Science Data Processing RFI flagger is based on the variation of VarThreshold method used in the *AOFlagger* algorithm explained in Section 1.7.5. This method is called SumThreshold. In this approach, samples

[2] SDP - Science Data Processing: Is the MeerKAT team which is responsible for quality control and quality assurance of the data.

from the VarThreshold method are combined and the sum of one or more sample combinations is used as a threshold, χ_M. M is the number of samples that are summed together. As opposed to the VarThreshold method, here, a sample can be flagged as RFI even if the sum of all the combined samples in a sequence does not exceed the threshold value.

Samples in this method are flagged using an increasing threshold. When a lower threshold has identified samples as RFI, the samples will not be included on the sum and will be replaced by the average threshold level. This is in done to avoid flagging too many samples.

The MeerKAT RFI flagger works on a quasi-real-time model. It runs on a two-dimensional data array of time and frequency. The Data is loaded into the flagger per scan, where a scan is defined as a collection of SDP visibilities over a certain period. For MeerKAT a scan is on average between 5 - 15 minutes. Therefore, for a 5 minutes scan, one will get around forty samples. Furthermore, the algorithm treats each baseline in each scan independently. Hence, this allows the parallelisation of the algorithm along the baseline axis.

The data for a single scan and baseline is further divided into frequency chunks and processed independently. This is because it is challenging to describe the whole data set statistically due to the wide variation of outliers in the bandpass as a function of frequency. Unequal frequency chunks were chosen to cover parts of the band with similar amplitude and also with edges outside the range of the static mask[3] so that the background may be interpolated over masked regions if necessary.

A smooth background fit is applied to the data by convolving it with a 2-D Gaussian whose widths are larger than expected RFI spike widths in both time and frequency and are smaller than any variations in the bandpass or changes in amplitude with time as shown in Equation 1.31 and 1.30. This ensures that the smooth background ignores any potential RFI spikes in the data but follows the true shape of the background. Data already flagged (usually from the ingest flagger or from the static RFI mask) are given zero weight and therefore do not contribute to the

[3]Static mask: Is a mask that is applied to regions of known RFI transmitters (e.g. GPS satellites band, GSM band) on the MeerKAT bandpass.

background estimation. Any data that is wholly masked over the smoothing area will have an undefined background as all contributing weights are zero, these data are interpolated in frequency.

The fitted smooth background is subsequently subtracted from the data, and the standard deviation is measured from the masked residual. This standard deviation is used as the basis for the threshold for spike detection. First, the data are averaged in time over the whole scan and outliers in the resulting 1-D frequency spectrum are located; this is to find faint spikes in the time axis that would otherwise be missed. RFI channels found in the 1-D spectrum are flagged for all times in the scan. Finally, the full data are flagged in the time and frequency dimensions.

2.2 MeerKAT visibility and flags structure

The MeerKAT visibility data files are stored using the MeerKAT Visibility File Version 4 (MVFV4) system. The MVFV4 files live in the MeerKAT archive referred to as the chunk store. A Redis database (**.rdb**) is used to store the files. The **.rdb** object does not contain the actual visibility data themselves, but only contain a database of referenced meta-data of the MeerKAT visibilities. These can be accessed when requested by a user.

The data set is built around the concept of a three-dimensional visibility array with dimensions of time, frequency and correlation products. From the example shown in Fig. 2.2 the shape of the data set is [27, 4096, 544]. Therefore, we can say that this observation had 27 **time dumps**, 4096 **frequency channels** and 544 **correlation products**.

The MeerKAT RFI flag files are also built around the concept of three-dimensional flag array; they have the very same structure and shape as the visibility files. The flags are stored as 8-bits unsigned integers (*uint*8), where each bit is a different kind of flag. In the MeerKAT data pipeline, there are eight different tags that a data point can be flagged as an RFI, these are:

- **Reserved rfi** = Reserved for future use.

- **Static Flags** = Mask of known RFI (e.g known satellites, GSM band) is applied to all the visibility files. Is only applied on short baselines of length less than 1 km.
- **CAM** = Flags produced in control and monitoring (CAM) system, for example when an antenna is down or correlator is broken.
- **Data lost** = This becomes true when there is missing CBF heaps in ingest and also when the correlator switches off in the middle of an SDP dump. The visibility data itself is zeroed in this instance.
- **Ingest RFI** = As explained in Section 2.1.1.
- **Predicted RFI** = It is a feature that is currently not used.
- **Cal RFI** = Flags that are produced by MeerKAT in-house developed Science Processing RFI flagger as discussed in Section 2.1.2.
- **Postproc** = This represents failed or invalid calibration solutions.

2.3 Extract, Transform and Loading (ETL) MeerKAT data

To access the actual data, *katdal* [4] methods are used to call the Numpy arrays from the chunk store. *Katdal* is an in-house-built python library that is used to access and manipulate MeerKAT visibility data files easily. A summary of the content of the *katdal* dataset can be inspected via its string representation, Fig. 2.2.
The first segment of the printout displays the static information of the data set, including the observer's name, dump rate, all the available sub-arrays and spectral windows in the data set. The second segment (between the dashed lines) highlights the active selection criteria. The last portion displays dynamic information that is influenced by the selection, including the overall visibility array shape, antennas, channel frequencies, targets and scan info.

Subsets of the data are selected via the *data.select* method. The subset of the data can be selected based on time, frequency, and correlation products. This applies a set of selection criteria to the data set, which changes the view on the dataset to

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

```
================================================================================
Name: /var/kat/archive2/data/MeerKATAR1/telescope_products/2018/02/18/1518941264.h5 (version 3.0)
================================================================================
Observer: sarah  Experiment ID: 20180218-0003
Description: 'MKAIV-405 Generic AR1 phaseup'
Observed from 2018-02-18 10:07:45.570 SAST to 2018-02-18 10:11:21.479 SAST
Dump rate / period: 0.12505 Hz / 7.997 s
Subarrays: 1
  ID  Antennas                              Inputs  Corrprods
   0  m000,m002,m003,m006,m007,m008,m011,m012,m013,m019,m022,m023,m027,m029,m032,m034  32      544
Spectral Windows: 1
  ID Band Product  CentreFreq(MHz)  Bandwidth(MHz)  Channels  ChannelWidth(kHz)
   0 L    bc856M4k   1284.000         856.000          4096       208.984
--------------------------------------------------------------------------------
Data selected according to the following criteria:
  subarray=0
  ants=['m019', 'm008', 'm003', 'm002', 'm012', 'm013', 'm007', 'm006', 'm029', 'm023', 'm022', 'm032', 'm027', 'm011
', 'm000', 'm034']
  spw=0
--------------------------------------------------------------------------------
Shape: (27 dumps, 4096 channels, 544 correlation products) => Size: 481.296 MB
Antennas: m000,m002,m003,m006,m007,m008,m011,m012,m013,*m019,m022,m023,m027,m029,m032,m034  Inputs: 32  Autocorr: yes
Crosscorr: yes
Channels: 4096 (index 0 - 4095,  856.000 MHz - 1711.791 MHz), each 208.984 kHz wide
Targets: 2 selected out of 2 in catalogue
  ID  Name         Type      RA(J2000)     DEC(J2000)  Tags                       Dumps  ModelFlux(Jy)
   0  PKS 1934-63  radec     19:39:25.03  -63:42:45.7  bfcal single_accumulation     24      14.46
   1  Nothing      special           -             -                                 3
Scans: 6 selected out of 6 total       Compscans: 2 selected out of 2 total
  Date        Timerange(UTC)       ScanState  CompScanLabel  Dumps  Target
  18-Feb-2018/08:07:49 - 08:08:29    0:slew     0:un_corrected      6    0:PKS 1934-63
              08:08:37 - 08:09:25    1:track    0:un_corrected      7    0:PKS 1934-63
              08:09:33 - 08:09:49    2:stop     0:un_corrected      3    1:Nothing
              08:09:57 - 08:10:05    3:stop     1:corrected      2    0:PKS 1934-63
              08:10:13 - 08:10:13    4:slew     1:corrected      1    0:PKS 1934-63
              08:10:21 - 08:11:17    5:track    1:corrected      8    0:PKS 1934-63
```

Figure 2.2: *A typical summary of the katdal dataset.*

match the selection.

The selection criteria are divided into three groups, based on whether they affect the time, frequency or correlation product dimension, whereby:

- Time: *dumps*, *timerange*, *scans*, *compscans*, *targets*
- Frequency: *channels*, *freqrange*
- Correlation product: *corrprods*, *ants*, *pol*

2.4 The Purpose of KATHPRFI

The purpose of the KATHPRFI framework is to provide MeerKAT users with a tool that will aid them to keep track of changes in the RFI statistics over a long period as measured by the telescope. Such information is very useful for various users including astronomers, telescope operators, RFI Engineers and anyone interested in the RFI health of the observatory. Below we describe some of the use cases that such a tool can provide to the users mentioned above.

2.4.1 Observation Planning

RFI is a nuisance for astronomers as it corrupts some of the observing band where scientific observation needs to be carried out. Astronomers must have a better understanding of the RFI environment since such information is vital for them in preparing observation proposals and to carry out scientific analysis of their experiment. The observing bandwidth is one of the key parameters that determine the sensitivity of a radio telescope. Sensitivity is defined as the RMS fluctuations of the noise in a radio image as shown by Equation 1.24.

A tool that can aid astronomers to know how much of the bandwidth will be lost due to the RFI is essential as that would help them to calculate the optimal integration time for a particular observation and to plan the best time of the day and time of the year to observe their targets.

Moreover, statistical data that give a more in-depth insight into the RFI environment of a site is essential to optimise the flagging algorithms. For example currently for MeerKAT, RFI static mask is applied on short baselines when flagging the data in some of the regions in the spectrum.

2.4.2 Telescope operations

Telescope operation involves two teams, the astronomer on duty (AOD) and the operators (OPS). The AOD will take requirements from the astronomers and build a scheduling block (SB) for the operators. An SB is a parameter file for an observation that defines how that particular observation should be run. The SB enables an operator to manage the system and conduct observations. The building of SB requires an understanding of the antenna layouts, position of the targets and other hardware (e.g. correlator) settings.

Operators, on the other hand, will run the SB and monitor the progress of the observation. They also need to log any peculiar events during the observing run, which includes the presence of RFI. Knowing properties such as which baselines on average show high level of RFI could help the AOD/operators for better building of the sub-arrays and for monitoring purposes.

The statistical information such as which hour of the day has high RFI occupancy could be useful for avoiding such times. Furthermore, knowing the azimuth and elevation of RFI sources will be useful for scheduling and for monitoring purposes.

Combing the RFI statistics can allow us to build an intelligent scheduler and monitoring system of the RFI on site. A tool that can help the telescope operation team not to only understand the RFI environment on site but also aid them to detect and understand system problems such as correlator failure and any other telescope electronics issues is vital to them. This can in return lead to less corrupted scientific data.

2.4.3 Site Monitoring

The MeerKAT RFI working group is responsible for keeping the MeerKAT site as free as possible from RFI. A tool that will provide RFI alerts when certain unusual events are detected or when specific known RF emitters go beyond a critical threshold will be very useful. KATHPRFI dataset provides a benchmark from which anomalous RFI events can be detected. This information will support the RFI Working Group to detect and investigate those RFI culprits.

2.5 High level Requirements for KATHPRFI Framework

Discussions with various users with interests described above were held. Following those discussions, the high-level requirements for our framework were elicited.

The first defined requirement of our framework is to develop software that can be used to analyse RFI data as measured from the MeerKAT telescope. The software will produce a data set that statistically summarises the RFI environment for MeerKAT. The design of the framework should be very agile in such a way that if a new need arises, we should be able to add the new attribute.

The second requirement is to create a dataset that would be used as a baseline for alert triggering when unusual events occur. The alert system will depend on the extracted statistics produced from KATHPRFI framework, which will provide a

benchmark for anomaly detection.

The third requirement is to give stakeholders easy access to the data produced from the KATHPRFI and any analysis report about the MeerKAT RFI environment that can be derived from it. Users can download the data set and manipulate it according to their needs.

2.6 Design Approach and Design Decisions

We chose an evolutionary prototyping model in our design instead of a throw-away approach. The evolutionary prototyping is a life cycle model wherein the concept of the system is developed as the project progresses (Carter et al. 2001). The evolutionary prototyping approach allows easy modification of the system in response to the user's inputs. However, the throw-away approach as the name suggests is an approach where prototypes are used to test out ideas, but they do not form part of the final system solution.

Our motivation behind choosing an evolutionary prototyping approach instead of the throw-away approach is due to the complicated nature of RFI signals, as it is difficult to frame the specifications of our system from the word go. The evolutionary prototyping model is very suitable for research projects like ours. In our approach, data set releases and feature additions will be done in phases while requesting end-user feedback. In Carter et al. (2001), they have shown that this kind of design approach allows walkthrough with the stakeholders to elicit and validate requirements, to reduce incompleteness, and inconsistency during requirements capture. The data set quality and resolution shall be improved as per stakeholders inputs.

In our framework, we have designed and implemented three prototypes to provide and test the quality of KATHPRFI data set and accessibility. These prototypes incrementally add dimensions to the dataset and therefore increasing the complexity of the dataset. This is discussed below,

- In the first prototype we have produced a data set with only two attributes, that is, the **time of the observation** (T) in Unix timestamps and the **frequency** (F) with a bandwidth of 856 MHz in steps of 208 kHz. However, keeping all

the timestamps posed a challenge (see Section 2.9) to our framework and hence we used hour of the day [0 - 23 hours], as our time resolution.

- The second prototype we have added the **baseline (*B*)** attribute (i.e. 2016 for MeerKAT). This can help to track down possible localisation of the RFI to a given antenna.
- The final prototype we added the direction attributes which are the **elevation (*El*)** and the **Azimuth (*Az*)** angle, with a resolution of 10° and 15° respectively.

2.7 KATHPRFI Algorithm Design

The KATHPRFI framework should be able to access the MeerKAT chunk store and retrieve the visibility data. After retrieving the data from the archive, the framework should run the MeerKAT RFI detection algorithm on the visibility data set. It should compute the *Master* and the *Counter* array, which would contain the descriptive statistics about the RFI environment on site. The *Master* array contains the number of RFI points per voxel, whereas the *Counter* array contains the total number of observations per voxel. Both arrays are built around the concept of a five-dimensional array with dimensions of **time of the day**, **frequency channels**, **baseline length**, **elevation** and **azimuth**. The shape of the data is [24, 4096, 2016, 24, 10]. Finally, both 5-D arrays must be saved, and they must be easily accessible to the end-users. Figure 2.3 shows the high-level sequential flow of the process followed by the KATHPRFI framework.

2.7.1 MeerKAT Archive search

For us to quantify the RFI statistics of the MeerKAT site, we need to have access to MeerKAT observational data. The data is stored in the chunk store, as mentioned in Section 2.2. We have written an automated archive search script *ArchiveSearch.py*, that takes in the description of the type of observation, the date range and the correlator mode that one is looking for, to return the **.rdb** observational data files. There is an option to download the **.rdb** files and to save the list of all the downloaded files.

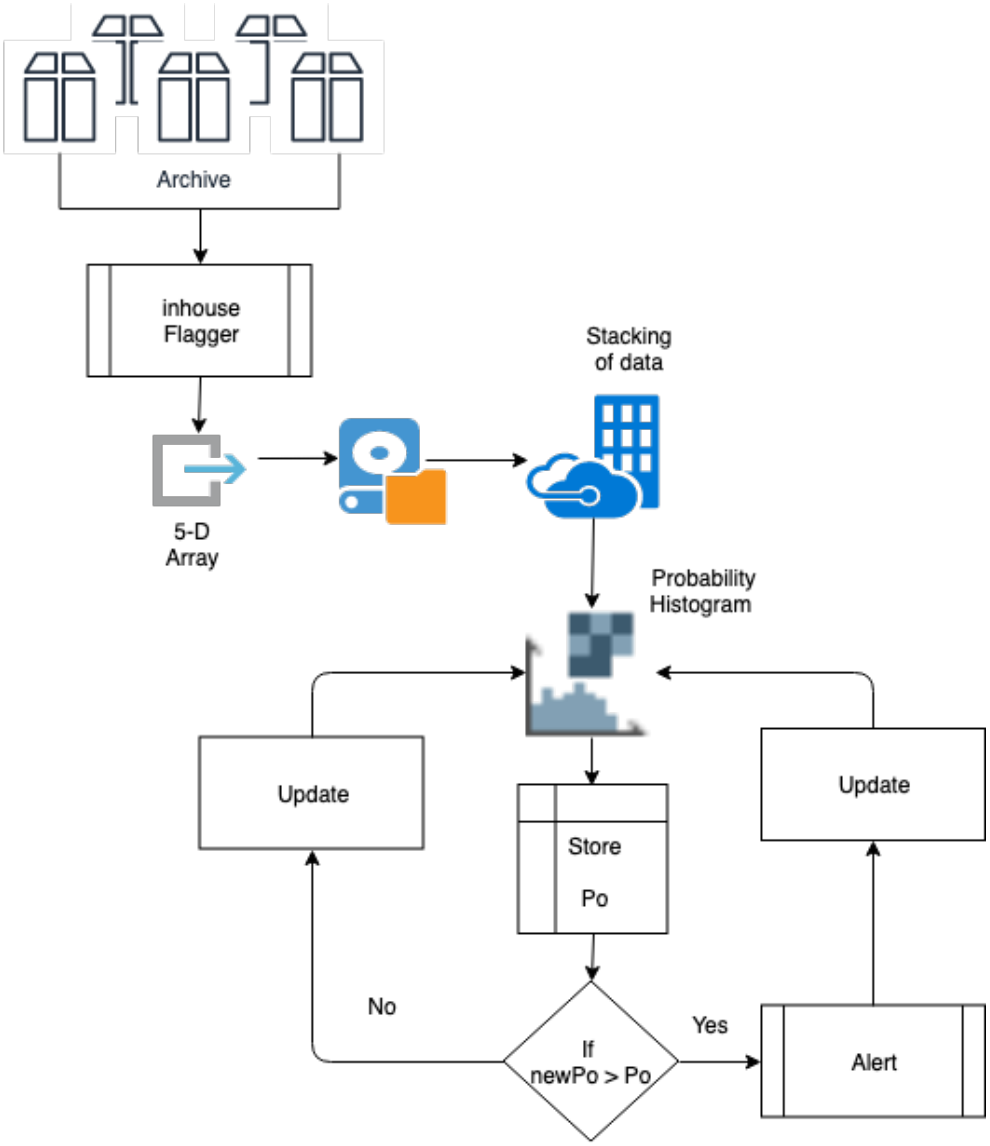


Figure 2.3: *High-level schematic diagram showing the sequential flow of the process followed by the KATHPRFI framework. P_0 is the average RFI occupancy and newP_0 is the RFI occupancy from new observation.*

2.7.2 MeerKAT RFI detection Algorithm

We run the offline MeerKAT RFI detection algorithm that produces the RFI flag files which we use for our analysis. The offline flagger was run because the MeerKAT online flagger produces flags with the static mask already applied to them. The static mask is the masking of some part of the MeerKAT bandpass due to the well known RFI transmitters that have 100% duty cycle and band usage. Thus, the static mask is not determined by the data, hence its name. Therefore the offline flagger was run to remove the static mask so that we can get the real RFI detected flags

without the mask. The details on how the algorithms works were explained in Section 2.1.2.

2.7.3 KATHPRFI Construction Algorithm

The KATHPRFI is a 5-D array (see Fig.2.3) that contains the RFI statistics from the MeerKAT observational data as per the user requirements discussed in Section 2.5. In this subsection, we will focus on the code design (see Algorithm. 1), the process followed and the reasoning behind.

The *kathprfi.py* script starts by initialising the *master* and the *counter* arrays, as shown in the Algorithm 1. Block 1 of the algorithm reads in the visibility file and the offline flag file, followed by applying the offline flags (see Section 2.2). The final step of Block 1 is the pre-processing stage whereby we remove any bad antennas from the data. An antenna is flagged as a bad one if, during an observation, it fails for some reasons. The *katdal* library is used to get the information about the antenna activities during the observation, and if we find a "stop" state on the activity lists that antenna would also be flagged as a bad one.

In the second block, we chose the parameters of interest to produce a subset of the data. The selected parameters are summarised in Table 2.1. The *katdal* library allows us to do the selection. Thereafter, the flag array is returned with the applied selection criteria.

Due to computational limitations discussed in Section 2.9, data-binning is required. This is carried out in the final step of Block 2. These limitations cause us not to be able to retain the full resolution of certain attributes, such as the time of observation, azimuth and elevation. We have only managed to maintain the full resolution of two axes, which is the frequency axis and the baseline axis. Time is binned per hour into 24 hours of a day, the elevation is binned into 10° interval,s and lastly, azimuth is binned in 15° intervals.

Block 3 to Block 5 are nested loops that go over timestamps, frequency channels and baselines respectively. This is done to update the *master* and the *counter* array based on the indices extracted from a particular observation file in question. From the example shown in Fig.2.2, the observation started at 10:07 until 10:11, therefore the *kathprfi.py* script will put all of the data in the 10^{th} hour bin. It will also check

Algorithm 1: The KATHPRFI Algorithm creating 5-D Arrays

Initialise: $Master \leftarrow zeros(24, 4096, 2016, 8, 24)$ ▷ ; // (T,F,B,El,Az)
Initialise: $Counter \leftarrow zeros(24, 4096, 2016, 8, 24)$ ▷ ; // (T,F,B,El,Az)
/* Loop going through each visibility and flag file to be processed */
1 **for** *each visibility and flag file in Directory* **do**
  Read in visibility and flag file
  if *frequency channels == 4096* ***and*** *dumprate==8* **then**
    Apply the cal flags on the vis file
    if *bad antenna == True* ▷ ; // Check bad antennas and remove them
    then
      Remove the antenna
      return *List of good antennas*
    else
      return *List of all the antennas*
    /* Selecting subset of the data to be used */
2   visibility.select(corrprods = "cross", pol = "HH", ant=ant_list, scan="track")
    return *visibility.flags* ▷ ; // Flags with selection criteria applied
    /* Extracting indices */
    $T \leftarrow visibility.timestamps$ ▷ ; // Binning timestamps to 1 hour interval
    $El \leftarrow visibility.elevation$ ▷ ; // Binning Elevation to 10° interval
    $Az \leftarrow visibility.azimuth$ ▷ ; // Binning azimuth to 15° interval
    $B \leftarrow visibility.corrprods$ ▷ ; // Extracting indices available baselines
    $F \leftarrow visibility.freqs$ ▷ ; // Extracting all frequency channels
    /* Looping through all the timestamps, all the frequency channels and all the baselines */
3   **for** $k \leftarrow 1$ *to* T **do**
4   **for** $j \leftarrow 1$ *to* F **do**
5   **for** $l \leftarrow 1$ *to* B **do**
          $Master[T[k],F[j],B[l],El[k],Az[k]] \leftarrow Flags[k,j,l]$
          $Counter[T[k],F[j],B[l],El[k],Az[k]] \leftarrow 1$
          return *Master, Counter*

Parameter name	Selected parameter and reasoning
pol	The **HH** was chosen for demonstration purposes.
corrprods	We have chosen **Cross**, because the flagger only works on the cross correlation products.
flags	We have chosen the **cal_rfi** and **ingest_rfi** because these two flag type are the only real RFI detection flags. The others indicate system problems.
scans	We have chosen **Track** to ensure only stable telescope pointing contributions to our results.
ants	We do not want to use bad antennas in our analysis.

Table 2.1: *Parameters used to select the data*

which antennas were present during the observation and update the baseline array accordingly.

2.7.4 Data Access and Data visualisation

Providing data access and data visualisation to the stakeholders is one of the requirements of our framework. We need to create an interface that our stakeholders would be able to access and visualise the relevant data quickly. To provide a very flexible functionality of data access for the stakeholders, the KATHPRFI data set can be accessed using the SARAO intranet web interface [5].

2.8 Resources

The python[6] programming language and various python packages have been used to create the KATHPRFI data set. Python is one of the most popular, high-level programming languages in use today and within the SARAO organisation. It is advantageous to use python because its compatibility with many other languages and runs on the major operating systems. Moreover, python has many wrappers for popular languages such as MATLAB, Java, C++ and CUDA, this makes it very

[5] https://sites.google.com/a/ska.ac.za/intranet/teams/data-science/hprfi
[6] https://www.python.org/

useful for interfacing with diverse systems.

However, Python is generally considered to be slower when performing heavy computation as compared to compiled languages such as C, C++ and FORTRAN. The main reasons for its slowness lie in being an interpreted programming language.

Python has plenty of libraries that were developed to overcome the problem of its slowness, and they have different optimisation techniques depending on the task they need to perform. The libraries offer a wide range of methods that can be used for data analysis, manipulation and visualisation. Hence, Python has become the to-go robust environment for scientific computing. Below we will briefly discuss some of the main libraries that we have used in this project.

2.8.1 NumPy

NumPy[7] is considered to be the core modern scientific and computing library in Python. It provides a robust multidimensional array object and tools for working with these arrays. Numpy is very good for numerical processing of multidimensional arrays. Its linear algebra methods combined with its array broadcasting capabilities allow for fast vectorised processing of large multi-dimensional arrays.

2.8.2 Dask

Dask[8] is a python parallel computing library that is composed of two parts. The first part is a general framework for distributing complex computations on many nodes and the second part is a set of convenient high-level APIs to deal with out-of-core computations on large arrays, this gives us the ability to compute on arrays that are larger than memory using all cores. Dask provides data structures resembling NumPy arrays (dask.array) that efficiently scale to huge datasets. The core idea of Dask is to partition large Numpy arrays that do not fit in memory into smaller arrays referred to as chunks. Dask arrays support most of the NumPy array interface and pandas DataFrames.

[7] https://numpy.org/index.html
[8] https://docs.dask.org/en/latest/

The Dask user interface is lazy, meaning that it does not evaluate until you explicitly use the *dask .compute()* method to load data in memory. The *dask.compute()* method converts dask arrays into Numpy array. The method only works if the results would fit into memory though intermediate computations may be performed on larger arrays due to its chucking capability. On the other hand, *dask.persist()* methods are used in a cluster environment, wherein the results from the computation are stored in distributed memory. This method returns a new Dask object that points to the already computed results distributed over the cluster memory.

2.8.3 Xarray and Zarr Arrays

Xarray[9] is an open-source Python package that provides a toolkit and data structures for N-dimensional labelled arrays, strongly inspired by Pandas[10]. Key features of the package include label-based indexing and arithmetic, interoperability with the core scientific Python packages such as pandas, NumPy, Matplotlib, and Dask. Xarray is built on top of Dask, and as a result, it can efficiently perform out-of-core task that does not fit into memory.

Although Numpy Ndarray is widely used data structures in scientific computation, they lack label-based indexing; as a result, they do not have a meaningful representation of the meta-data associated with their data. It becomes a tedious process for users of the data to figure out which axis corresponds to which index position (e.g. Is the frequency axis of the array in the first or second index). In our framework, we have used Xarrays, and all our axes are labelled and have the associated meta-data. The Xarray has a Zarr [11] back-end that can be used to write and read datasets to disk. Here, the Zarr is used to interface with Dask to support parallel reading and writing of data from disk. For that reason, the Zarr is used to store and read the data from the drive.

2.8.4 Numba

Numba[12] is a Just-in-Time (JIT) compiler for CPython. It is written in C and is implemented in such a way that it can be loaded by programs running in the CPython interpreter and does not replace the interpreter itself. As discussed above,

[9] http://xarray.pydata.org/en/stable/
[10] https://pandas.pydata.org/
[11] https://zarr.readthedocs.io/en/stable/api/storage.html
[12] https://numba.pydata.org/

Python is slow, therefore for better performance or speed, developers would have to rewrite their python code in low-level languages. Numba offers a solution that does not require users to rewrite their python code by making use of heavy Ndarrays operations and numeric scalars in loops. Numba can generate specialised loops for arrays in machine code by extracting information from the Ndarrays such as the dimensions, data type and data layout. This allows Numba to avoid unnecessary indirection for accessing data in Ndarray.

In our framework, Numba was used to speed up the updating of the 5-D array when adding a new observation file.

2.9 Computation Limitations

Since our algorithm produces multi-dimensional arrays, we will have limitations on computational resources. The following subsections explain some of the limitations of our current system.

2.9.1 Random Access Memory (RAM)

Computers, in general, have a limited maximum amount of RAM that is controlled by the hardware, software and economic factors. This was the greatest limitation of our framework because we could only store a limited amount of data in a finite amount of RAM. Thus, RAM has influenced the resolution of our dataset as explained in Section 2.7.3.

Sufficient memory was required upon the construction of the *master* and *counter* array. Then the arrays were stored into a disk and data access is lazy. There are two concepts relating to the memory of a computer, and those are the dimension of the array and the size of the data type. As stated in Section 2.7 our array has dimensions of [24, 4096, 2016, 8, 24]. Each element, in our array, is a 16-bit unassigned integer which takes 8 bytes in memory. Below is the calculation of the memory that is required to store the required arrays in RAM:

$$\begin{aligned} \text{Required RAM per array} &= dimension \times data - type - size && (2.2)\\ &= 24 \times 4096 \times 2016 \times 8 \times 24 \times 8 \text{ bytes} && (2.3)\\ &= 7.610 \times 10^{10} \text{ bytes} && (2.4) \end{aligned}$$

The base two binary system [13] (i.e. 1 Megabyte = 2^{20} bytes) is used as the unit of data measurement, instead of the standard SI unit system where 1 Megabyte is equivalent to 1×10^6 bytes. Hence, as a result, the total memory required to store the array is:

$$\text{Required RAM per array} = 7.610 \times 10^{10} \text{ bytes} \times \frac{1 Megabyte}{2^{20} \text{ bytes}} \quad (2.5)$$

$$= 283.5 \text{ Gigabytes} \quad (2.6)$$

Thus, for the two arrays, the total RAM required is 567 Gigabytes. The machine that was made available for our framework has 756 Gigabytes of RAM.

2.9.2 Disk storage

Since RAM is temporary storage, the results from our software need to be stored on permanent disk. The problem of storing the 5-D data array for each observation was that we would eventually run out of disk space. As a result, instead of having a 5-D array for each observation we decided to have only one array that we will update as the new observation comes in by incrementing a counter. This means that we do not have access to data from a single observation and at the same time increases the speed of execution from input/output (I/O) interactions.

2.10 Data Description

Randomly chosen imaging observations were used to create the KATHPRFI data set. The data set used is equivalent to 1500 hours of observing time which was collected from May 2018 to December 2018.

2.11 Conclusion

This chapter gave high-level requirements of the KATHPRFI framework driven by the needs of the MeerKAT users. The design approach and the design decision of the framework was covered in detailed considering software and hardware constraints.

In the following chapter, we now going to look at the results of the KATHPRFI.

[13] https://www.gbmb.org/bytes-to-mb

Chapter 3

MeerKAT Site RFI Status

In this chapter, we present the results of the estimate of the probability of observing RFI as a function of time of the day, frequency, baseline and pointing direction (i.e. elevation and Azimuth) as measured by the MeerKAT telescope during imaging runs. We will use the data produced from the KATHPRFI framework as described in Chapter 2 to compute these probability estimates. We, therefore, begin with an explanation of how we calculated the probabilities.

Suppose that α is the number of RFI samples as obtained from the *Master* array and β is the number of NON-RFI samples (i.e. Total of *Counter* array - Total of *Master* array); where *Counter* array is the total number of observed samples. Then we can compute the probability estimate, P(RFI), in a voxel as follows:

$$P(RFI|t,\nu,b,el,az) = \frac{\alpha_{t,\nu,b,el,az}}{\alpha_{t,\nu,b,el,az} + \beta_{t,\nu,b,el,az}} \tag{3.1}$$

where t, ν, b, el, az are the indices of the time of the day, frequency, baseline length, elevation and azimuth in a given voxel respectively.

In order for us to compute the probability of RFI for a given dimension, we need to marginalise over all other dimensions. For instance, if we want to compute the probability of observing RFI as a function of the frequency, we sum both *master* and *counter* array in all other axes except the frequency axis, and then we divide one by the other, and the resulting array will be the probability of observing RFI as a function of frequency. Mathematically it can be written as,

$$P(RFI|\nu) = \frac{\sum_{t,b,el,az}(\alpha_\nu)}{\sum_{t,b,el,az}(\alpha_\nu + \beta_\nu)}. \tag{3.2}$$

To calculate the average, we used two different methods. The first method was to update the *master* and *counter* array every time we get a new observation. At the end using Equation 3.2 we computed the average RFI probability, effectively combining all observations into a single long observation. This is referred to as the Combibe Average (CA). The other method consists of computing the average of each individual file using Equation 3.2 and finally computing the average of those probabilities. We call this method the Average of Average (AoA). These two averages will coincide if all the files have the same length or if the average of each file is the same, but in general the CA and AoA will differ. In the following sections, we will discuss the RFI occupancy for each of the attributes in our data set.

3.1 RFI Occupancy versus Time of the day and Frequency

Figure 3.1 shows the distribution of RFI probability as a function of time of the day in Coordinated Universal Time (UTC) and frequency in megahertz (MHz). Our results show a clear pattern between the hour of the day and the RFI probability. We see a drop of RFI probability during the night time (i.e. 20:00 - 04:00 UTC) as compared to the day time (i.e. 05:00 - 20:00 UTC). A maximum variation of 4% is observed between hours of the day in the RFI occupancy with an average of 23%.

We noticed that at 05:00 UTC the RFI occupancy goes up, this may be related to when activities begin in the nearby towns and cities, and at times even on-site. At 10:00 UTC we see a drop in the RFI occupancy, similarly, at 14:00 UTC we see another drop. These two times correspond to lunchtime and the end of the working day in South Africa respectively. We cannot conclusively say that the observed increase in RFI occupancy is caused by these human activities, however, a correlation clearly exists.

We also noticed the RFI probability at the following frequencies: 1018 MHz, 1031 MHz, 1041 MHz, 1090 MHz and 1103 MHz increases during the day time and drops at night time. These frequencies are confined within the DME band (see Table 1.3) which is allocated to the aircraft communication system. Therefore, these findings suggest that the observed increase in RFI probability during the day is most probably due to the aircraft passing over a region of the site.

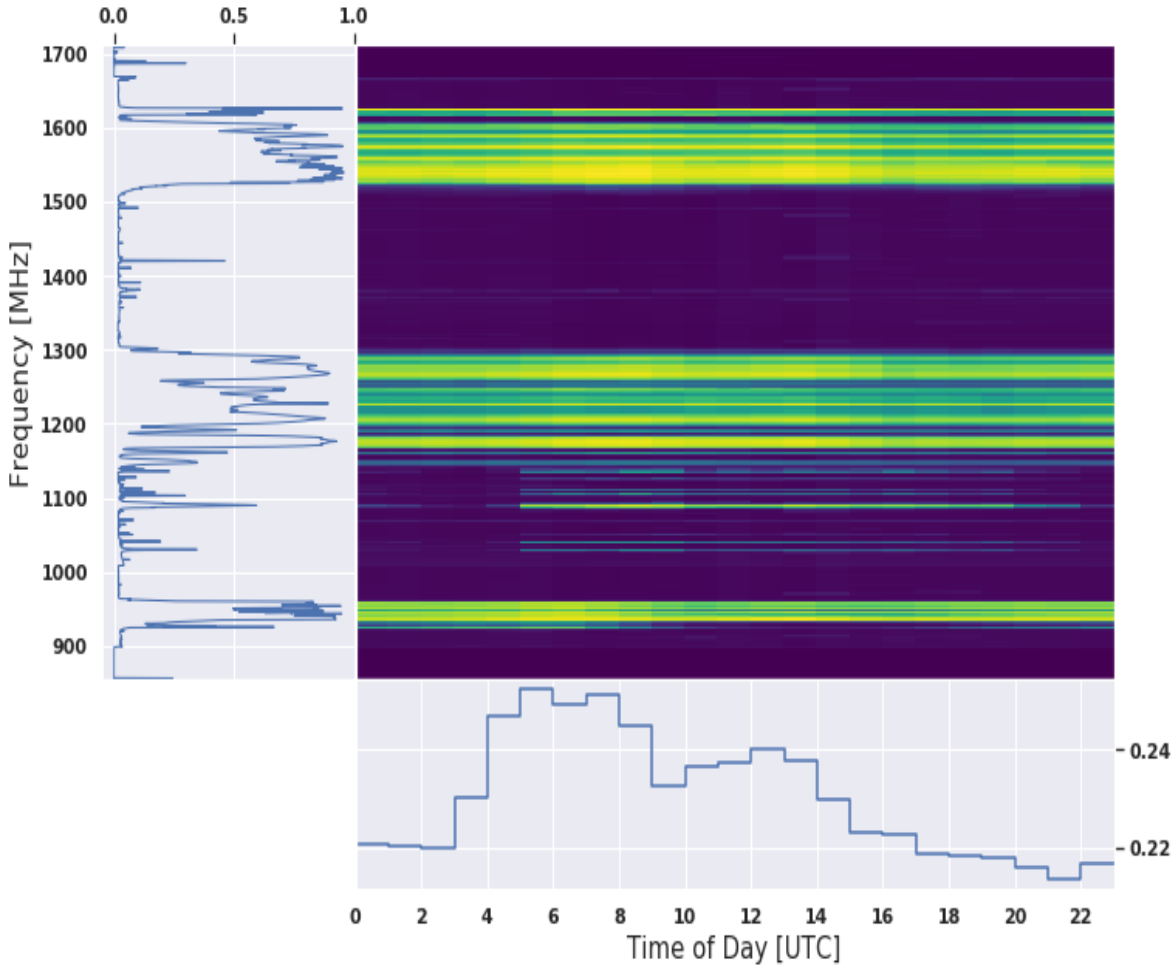


Figure 3.1: *The distribution of RFI probability as a function of frequency and time of the day. The time of the day has an average RFI of about 23%. The colour scale indicates the amount of RFI detected in a specific time-frequency bin with yellow being the highest probability and purple the lowest probability.*

It can also be noticed that there is a great deal of variation of RFI occupancy as a function of frequency. At some frequency bands (e.g. 900 - 960 MHz) we see 100% RFI whereas at others (e.g. 1320-1500MHz) the RFI occupancy is down to less than 10%. We can see the three main frequency bands showing the highest probability of RFI in the MeerKAT site, as discussed in Section 1.7.1. Those are the Global System for Mobile Communication(GSM) (900 - 960 MHZ), aircraft transponders (1000 - 12000 MHz) and Global Positioning System(GPS) satellites (1482 - 1600 MHz & 1169 - 1280 MHz). From our analysis 36.6% of the band at all the time, all the baseline is always flagged as RFI.

In this thesis, we are primarily interested in the RFI, which for most known per-

sistent sources (such as GPS satellites, DMEs and GSM) are fairly constant, predictable and regular. As a result, the variation in the probability of RFI from such sources is expected to be considerably small. For us to understand whether the observed fluctuations are statistically significant or are due to noise fluctuations, we computed the 68 percentile which corresponds to 1-sigma confidence interval for a Gaussian distribution. On the other hand, the 95% confidence interval will include all sorts of outliers. While these are potentially interesting for MeerKAT in general, we will show in Chapter 4 that they are mostly limited to specific months in our data. As a result, the RFI variability is more accurately captured by the 68% confidence limits which are much more tightly constrained around the mean.

In the following subsections, we will investigate the statistical consistency of the RFI probabilities.

3.1.1 Average RFI as a function of time of the day

Figure 3.2 shows the average RFI probability as a function of the time of the day, with the green region representing the 68% confidence interval. We used two methods to calculate the average as explained in the introduction of Chapter 3. The blue line represents the CA, while the orange line represents the AoA. We observe a similar distribution of RFI from the two methods.

As mentioned at the beginning of this section, at 10:00 UTC we observe a drop in the RFI occupancy. For this time of the day, we also found that the data is noisy as shown by the 68% confidence interval. To understand this noisiness, we looked at the histogram of the RFI probabilities at a noisier and quieter hour of the day (Fig 3.3); a kernel density estimate (KDE) (Lee & Park 2012) fit is also shown. A KDE is a non-parametric method that estimates the probability density function (PDF) without assuming any underlying distribution for the variable.

The huge variation of RFI probability that is observed could be explained by the long tail in the distribution of the RFI probabilities as shown in Fig 3.3a. The long tail indicates some form of an anomaly and this could be a result of several issues such as the correlator outputting zero visibilities as explained in Appendix A.2. In contrast Fig 3.3b shows the distribution for a quieter time (07:00 UTC) with respect to the noisier time (10:00 UTC). As a consequence, the RFI probabilities are tightly

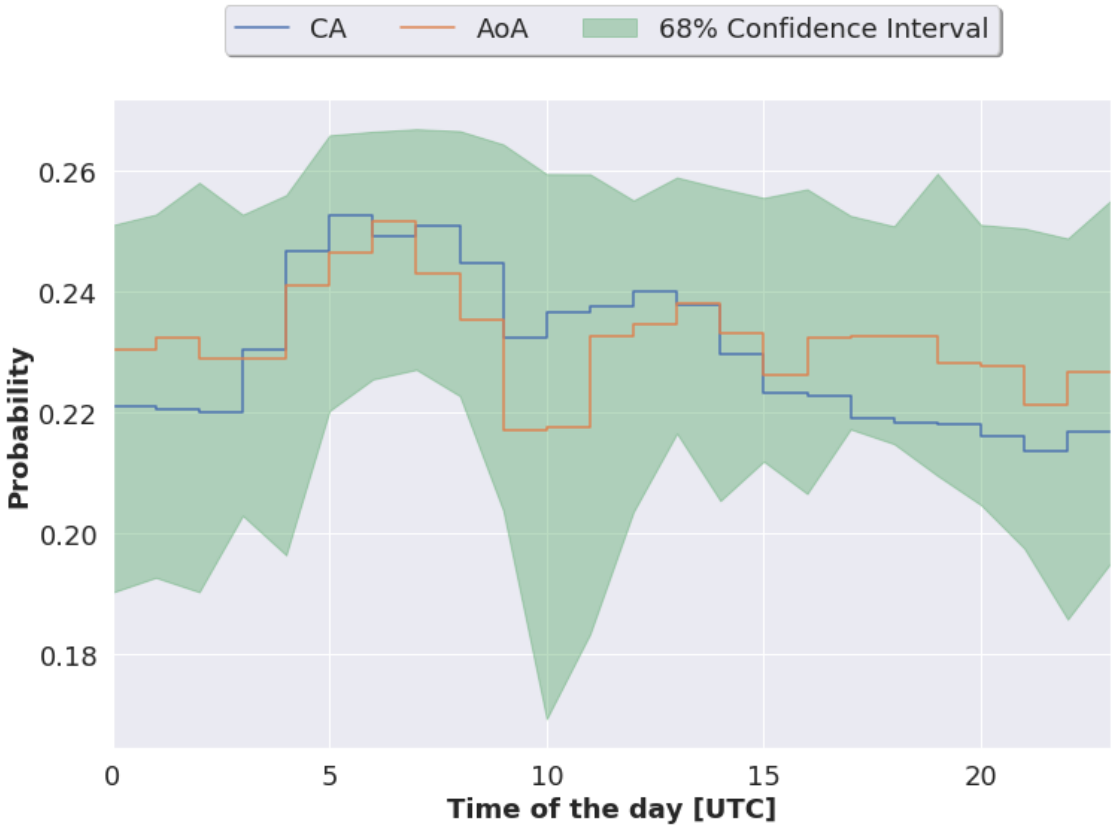
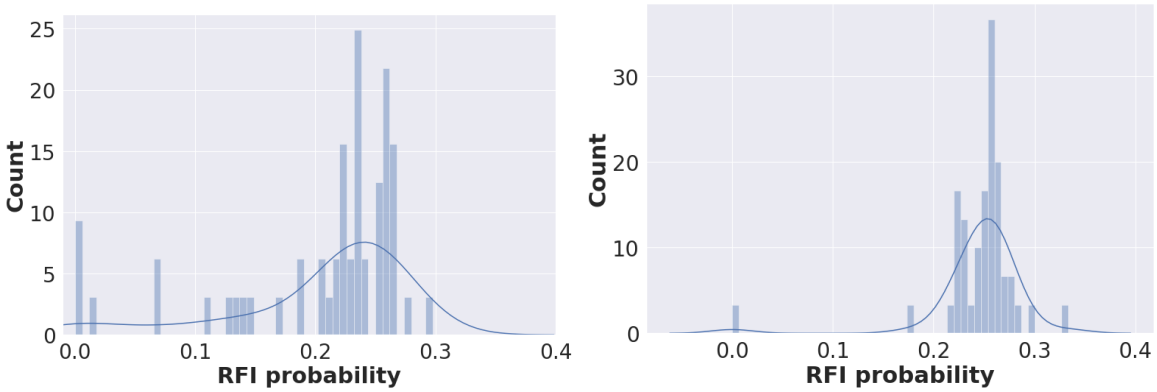


Figure 3.2: *The distribution of RFI occupancy as a function of time of the day. The green region represents 68% confidence interval. The blue line is the Combined Average (CA) and the orange line is the Average of Average (AoA), the difference is explained in the introduction of Chapter 3.*

concentrated around the mean.

Looking at the distribution of the RFI probabilities (Fig. 3.3) at specific hours of the day, we found that some of the observation had zero probabilities. The results imply that no RFI was detected on any baseline and at any frequency by the algorithm; something essentially impossible because of the permanent presence of RFI sources. This is an indication of a potential system problem, such as the correlator outputting zero visibilities. The MeerKAT RFI detection algorithm does not detect any RFI when such events happen. Hence, we see a zero probability of observing RFI. Even though the investigations of the cause for the outliers is beyond the scope of this thesis, we investigated some of the files that had zero visibilities (Appendix A.2).



(a) *The 10^{th} hour distribution.* ***(b)*** *The 07^{th} hour distribution.*

Figure 3.3: *The histogram of RFI probabilities with a kernel density estimate (KDE) fit. The RFI probability distribution at a noisy hour (10^{th} hour, left) and quieter time of the day (07^{th} hour, right). The distribution of the probabilities in the left plot has a long tail that is indicating some form of an anomaly. The RFI behaviour is well defined by the average probability.*

3.1.2 Average RFI occupancy as a function of Frequency

When performing a similar analysis on the frequency axis, we decided to split the spectrum into the corrupted band and the clean band. The corrupted band is defined as the range of frequencies in which the major RFI sources (GSM, DME and GPS satellites) emit. Whereas the clean band is the less corrupted part of the spectrum. We divided the clean band into lower and upper frequencies which are, between 980 MHz - 1070 MHz and 1310 MHz - 1500 MHz respectively, as depicted in Fig 3.1. Figure 3.4 shows the RFI probability variation as a function of frequency with the 68 percentile. We noticed a small variation in the RFI occupancy in the corrupted band as shown by the 68% confidence interval limits which are tightly constrained around the mean. However, as for the lower and the upper clean bands we do find frequencies (e.g. 1030 MHz, 1040 MHz, 1381 MHz, 1390 MHz and 1492 MHz) in which the RFI occupancy is greater than 10%, these are depicted by spikes in those regions. We observe a huge variation in RFI occupancy at these particular frequencies when looking at the 68% confidence interval.

We, therefore, looked at the distribution of probabilities of some of the clean band frequencies, Fig. 3.5. We expected the distribution of the RFI probabilities in the clean band to be close to zero, as there should not be any contamination. However, the figure shows a long tail distribution towards higher values of RFI probability. This long tail is a result of rare events that are appearing much more frequently

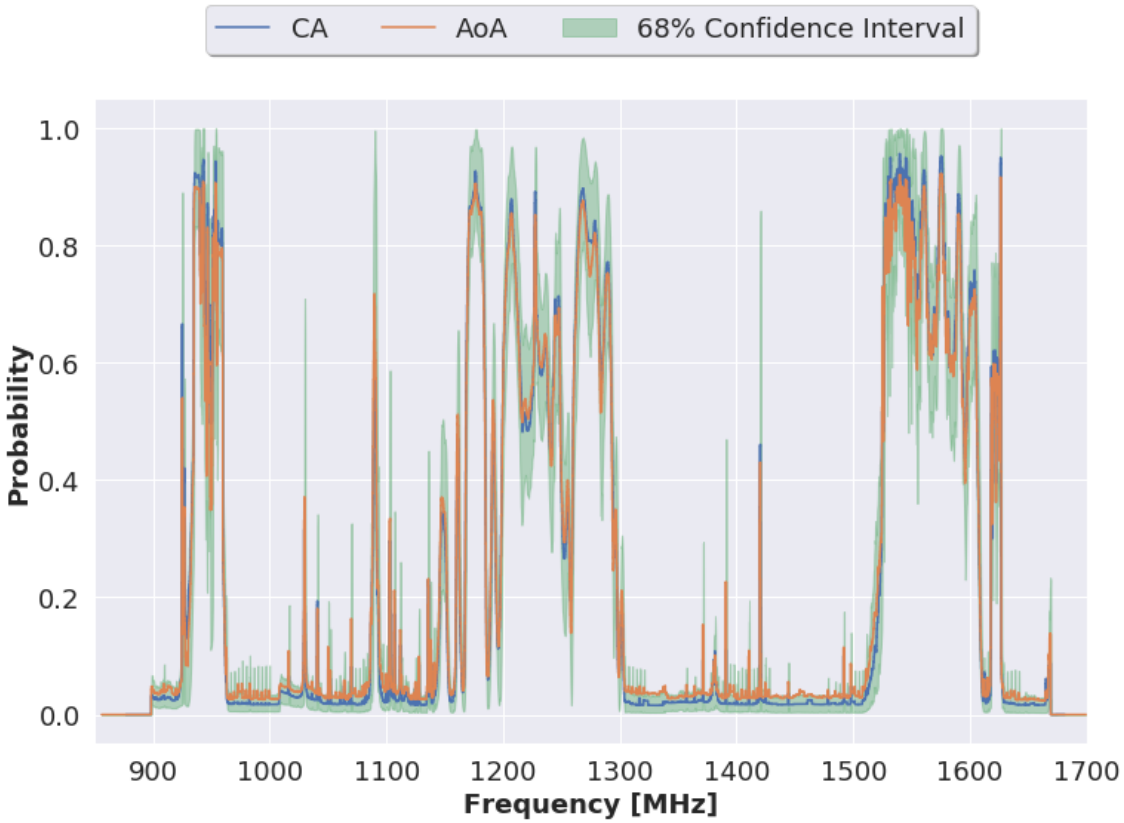
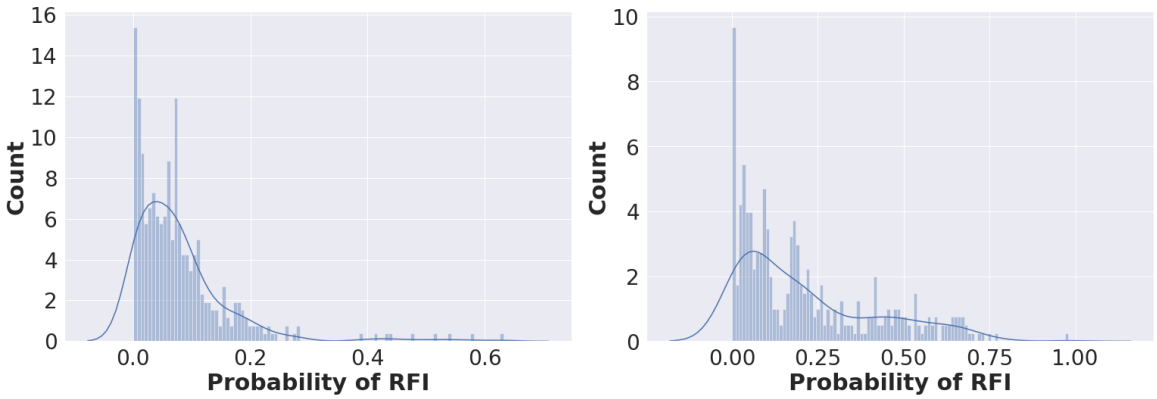


Figure 3.4: *The distribution of RFI probability as a function of frequency. The green region represents the 68% confidence interval. The blue line is the Combined Average (CA) and the orange line is the Average of Average (AoA), the difference is explained in the last paragraph of Section 3.1. The average behaviour of the known RFI source is constant and is well captured by the 68% confidence interval.*

than we expected. For example, the 1380 MHz L3 GPS band which is used for detecting nuclear activity on Earth seems to have been more active. The two frequencies shown are confined within the GPS L3 band.



(a) *Frequency = 1381 MHz* ***(b)*** *Frequency = 1391 MHz*

Figure 3.5: *An example of the RFI probability distribution of some of the contaminated frequencies from the clean band. The distribution of RFI in the clean band is mostly skewed towards zero probability, expected since this band is supposed to be free of RFI, however, we do see some outliers at higher probability values.*

Overall, these primary findings are consistent with our argument that the major RFI sources at the MeerKAT site are the GPS satellites, the GSM and the DME (see Section 1.7.1). Moreover, these findings cast a new light on the activities that are happening in the clean band. A typical MeerKAT bandpass cross-section was shown in Chapter 1, Fig 1.15. The figure shows that for a particular baseline and particular time for a single observation, both the lower and the upper clean band are free from RFI. However, our results (Fig 3.4) shows that on average the bandpass is not as clean as depicted by a single observation. From these results, we can say that the RFI environment has changed over the six months of data we analysed. The clean band is supposed to be as clean as possible from RFI, but collectively, our results provide evidence of some of the activities that are happening in the clean band which are worth investigating. A more detailed analysis of the clean will done in Chapter 4.

3.2 Baseline Length

We investigated the probability of RFI as a function of baseline length in, Fig 3.6. The blue and orange dots are the average RFI probabilities from the two different methods, CA and AoA respectively. Meanwhile the green region represents the usual 68% confidence interval. We noticed that the RFI probability decreases as a

function of baseline length from both methods.

Figure 3.6: *RFI occupancy for the MeerKAT telescope as a function of Baseline length (m). The blue and the orange dots are the mean RFI probability from CA and AoA respectively. While the green region represents the 68% confidence interval. The observed decrease of the RFI probability with an increase of baseline length is due to moving RFI sources with respect to the static sky. Therefore, the phase of the RFI changes rapidly on long baselines compared to short baselines. As a result when a correlation is carried out the RFI amplitude will vanish less on short baselines compared to the long baselines.*

To explain the observed decrease in RFI probability as a function of the baseline length, one needs to have a better understanding of how interferometry works as explained in Chapter 1. The complex visibility (Equation 1.18) of a single source is produced by the multiplying the sky with the fringe pattern produced by the baseline integrated over a solid angle. The angular distance between two consecutive peaks of the fringe pattern is defined as the fringe spacing. The fringe spacing is dependent on the separation between the antennas, with a short baseline giving a large fringe spacing while the long baseline gives smaller fringe spacing.

Some RFI sources are moving with respect to the static sky and as a result, the phase of these RFI sources wraps rapidly on long baselines as compared to the short baselines. Therefore, when a correlation is carried out on long baselines the

RFI amplitude will be considerably reduced. On the contrary, the short-baselines gives a large fringe spacing, hence as a result, when the correlation is carried out the RFI amplitude is reduced less when compared to the longer baselines.(Offringa et al. 2013). Thus we should expect RFI from moving sources to decrease with increasing baseline length.

3.3 Telescope Pointing Directions

When a telescope is pointing at various directions in the sky, the amount of RFI it measures is expected to change depending on the number of radio frequency transmitters that is in the field of view. It is anticipated that RFI due to terrestrial sources should be more dominant at low elevation. Figure 3.8 was used to examine this possibility. Recall as we explained in Chapter 2, the actual azimuth and elevation were binned. That is an elevation bin of 10° corresponds to an actual telescope elevation between 10° and 20°. We noticed that between 20° and 50° elevation the RFI probability is the highest and it gradually drops as we go to higher elevations on both the CA and AoA methods. This results can explain that indeed at low Elevation we do see more RFI as compared to high elevations.

A similar analysis was done on the azimuth angle (Fig. 3.8 b). We notice that the the two methods we used to calculate the average RFI probability gives different results more especially at lower azimuth (30° - 150°). This discrepancy arises since the AoA is actual an unweighted average.

When looking at the CA plot the maximum RFI occupancy is measured between 120°-180°. According to the MeerKAT RFI monitoring system, these directions points towards the closest towns as shown in Fig. 3.7, Carnarvon, Victoria West and Beaufort West.

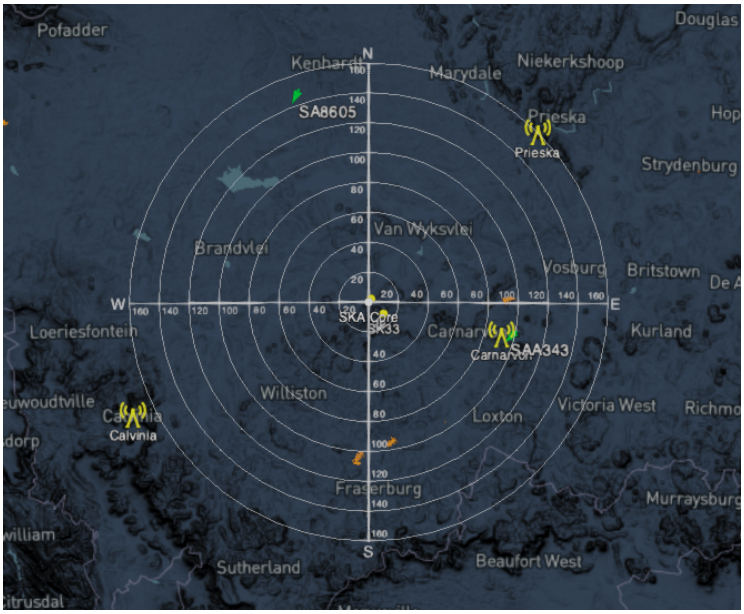
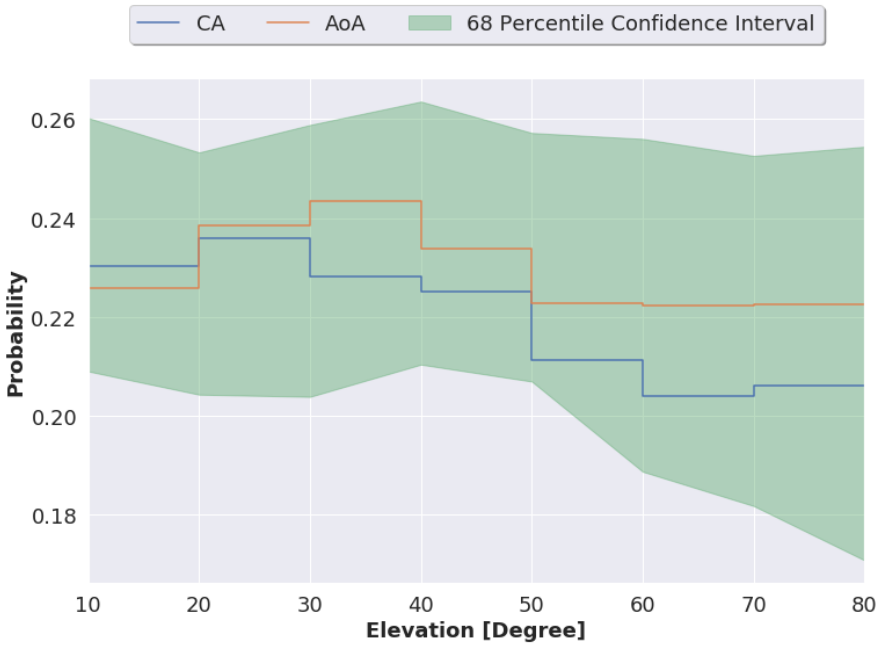
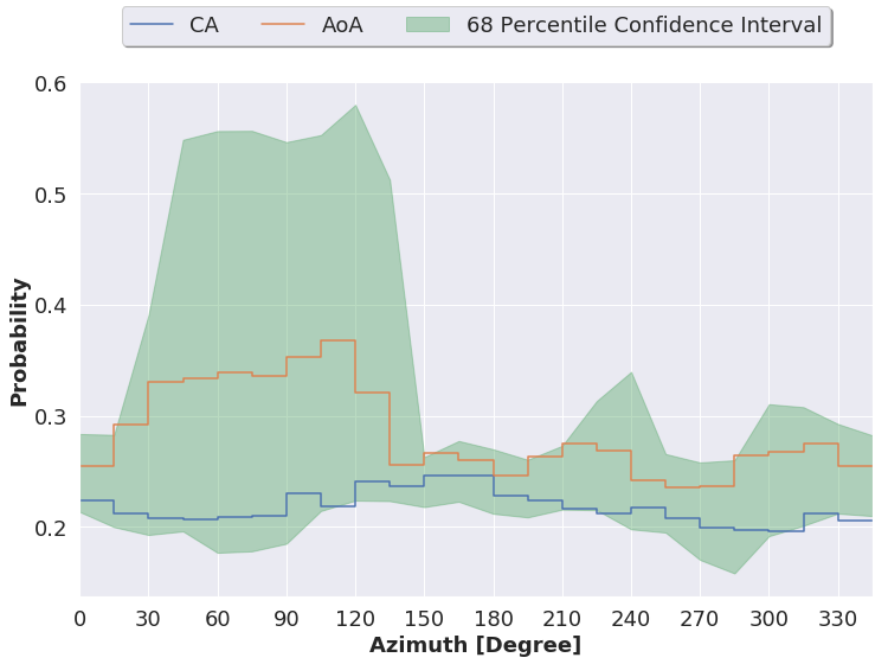


Figure 3.7: *A screengrab of the MeerKAT RFI monitoring system. The MeerKAT array is denoted by the yellow dot at the centre. The yellow tower-like structures are the communication towers and the blue dot at around 350° azimuth is a flying aircraft. The annuli represent the distance from the core in km.*



(a) *Elevation in Degrees*



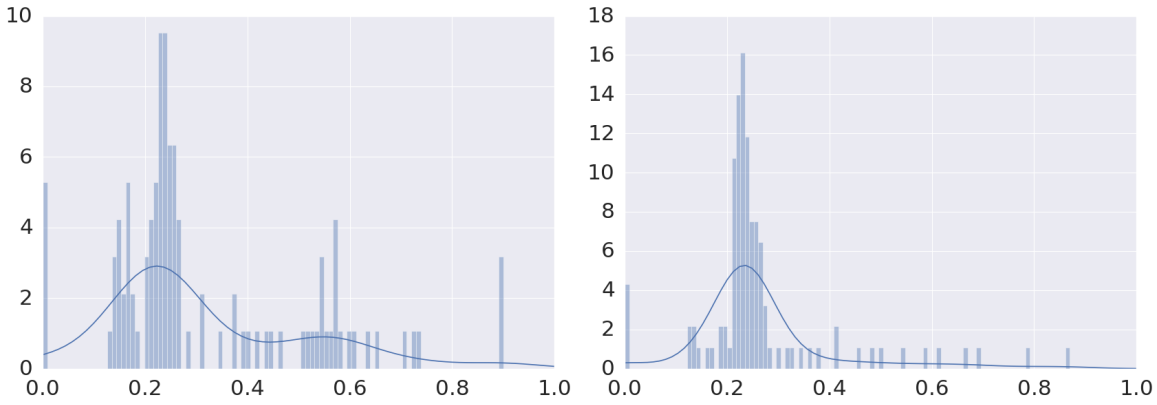
(b) *Azimuth in Degrees*

Figure 3.8: *RFI occupancy for MeerKAT site as a function of telescope pointing. The green region represents the 68% confidence interval. The blue and the orange lines represent the CA and AoA methods. The confidence limits are tightly concentrated around the mean, except for angles between* 30° *and* 140° *on the azimuth plot.*

Likewise, we computed the 68% confidence interval for the elevation axis and the azimuth axis, Fig 3.8. We found that the 68% confidence interval limits on the elevation are tightly constrained around the mean, hence a small variation in RFI probability is observed. As for the azimuth plot we found that some of the directions (30° and 140°) are too noisy. In order for us to understand the observed large variations we took a slice at a specific direction to look at the distribution of the RFI probabilities, Fig 3.9.

Looking at the number of counts for both noisier and quieter azimuth we noticed that the noisier angle has less count. It is worth noting that the count on outliers is comparable for both angles. Thus, this is indicative of the lack of data in those regions.

Figure 3.10 is the polar plot that shows RFI probability as a function of elevation (radial direction) and azimuth (theta direction). The white empty areas (Azimuth: 225° - 240° and 345° - 360°) is indicative of lack of data for these angles in our anal-



(a) *Noisier Azimuth* 30° ***(b)*** *Quieter Azimuth* 150°

Figure 3.9: *The RFI probability distribution of some of the noisier and the quieter azimuth. We see that the number of samples on the noisier azimuth is less compared to the quieter azimuth angle. One can notice that the count of outliers in the noisier and quieter azimuth is comparable.*

ysis. The colour scale ranging from purple through blue to yellow represents the probability of RFI occupancy, with yellow denoting the highest probability while purple is representing the lowest probability of RFI. We see that on average, in all directions the RFI probability is around 30%. We do notice a hot spot at low elevation and azimuths around 135°. This coincidentally points towards the town Beaufort West (Fig. 3.7) and requires further investigation for confirmation.

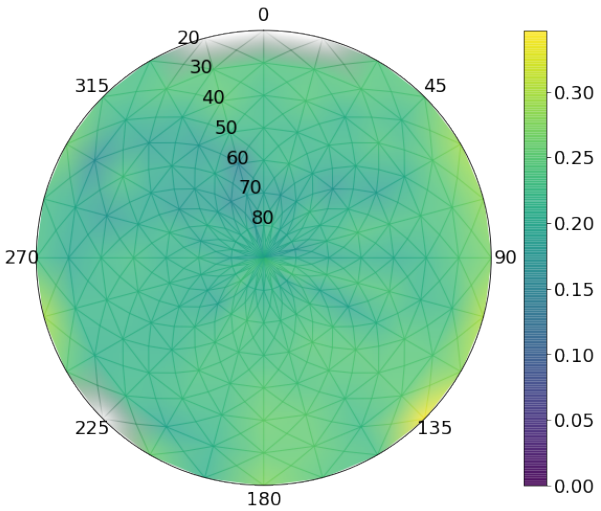


Figure 3.10: *Polar plot for the RFI probability as a function of Elevation (r) and Azimuth (theta). The colour scale represents the RFI occupancy with purple denoting the lowest probability and yellow representing the highest RFI occupancy. The white regions (e.g elevation: 20° and azimuth 0° - 15°) is indicative of lack of data.*

3.4 Conclusion

In this chapter, we presented the results of the RFI probability as a function of key characteristics. Our results are consistent with the claims we made in Chapter 1 that the major RFI sources for MeerKAT site are the GPS satellite, DMEs and the GSM. These primary findings give us an overview of the average RFI in the MeerKAT clean band, as well as the variability of the RFI, that one cannot characterise from a single observation.

Importantly, our results indicate that there is a correlation between the time of the day and RFI occupancy, that is potentially related to human activities around the site. Furthermore, we did find a small correlation between the time of the day and the flights passing over the region of the site. Our results have shown that the highest probability of RFI point towards a region including nearby towns. Overall the detected RFI occupancy for MeerKAT site as function of *time*, *frequency*, *baseline*, *elevation* and *azimuth* as shown by the KATHPRFI data set is **22.9%**.

In the following chapter, we will discuss in details the RFI from the clean band.

Chapter 4

RFI Analysis for the Clean band

The known RFI transmitters that impact the MeerKAT array are allocated to the communication system (GSM), aircraft communication system (DME) and the GPS satellites. The data that are used for scientific analysis need to be from frequency bands which are less corrupted by RFI. In this chapter, we will, therefore, focus on analysing the RFI from the clean band wherein the spectrum is known to be less corrupted by RFI. For the MeerKAT L-band spectrum, we defined two regions where the spectrum is considered and known to be clean those are: 970 - 1080 MHz and 1300 - 1500 MHz bands. However, the GPS L3 band is confined within the upper clean band, hence for this analysis, we removed the L3 band contribution. Some results for the known corrupted band is shown in Appendix B.

An increase in RFI occupancy level in the clean band pose a huge challenge to radio astronomy because the data collected would be corrupted. Therefore, it is important to keep track of changes that might be happening over time. In this section, we will analyse the clean band and see how much RFI we observe as a function of direction and time.

Recall, in Chapter 3 we have shown how to calculate the probability of RFI for a given dimension, for example the RFI probability as a function of telescope pointing direction can be calculated as follows:

$$P(RFI|el,az) = \frac{\sum_{t,\nu,b}(\alpha_{el,az})}{\sum_{t,\nu,b}(\alpha_{el,az} + \beta_{el,az})} \tag{4.1}$$

In order to calculate the probability of RFI in the clean band, we need to condition our probability calculation on the range of frequencies of the clean band. Using

Equation 4.1, the probability of observing RFI in the clean band can be calculated using:

$$P(RFI|el_{\Delta\nu}, az_{\Delta\nu}) = \frac{\sum_{t,\Delta\nu,b}(\alpha_{el,az})}{\sum_{t,\Delta\nu,b}(\alpha_{el,az} + \beta_{el,az})} \tag{4.2}$$

where $\Delta\nu$ represents the range of frequency indices in the clean band and with α and β being the number of RFI and NON-RFI samples as explained in Chapter 3.

4.1 Clean band as a function of pointing direction

We looked at how much RFI is generated in the clean band as a function of telescope pointing direction as represented in Fig. 4.1. We noticed a hot-spot (maximum RFI occupancy) at lower elevations and azimuth angle of 135° that is pointing towards the nearby towns. In addition, the RFI occupancy is quite moderate across the azimuth angles at lower elevations. Looking at higher elevations (elevations > 40°) the average RFI occupancy is about 2%.

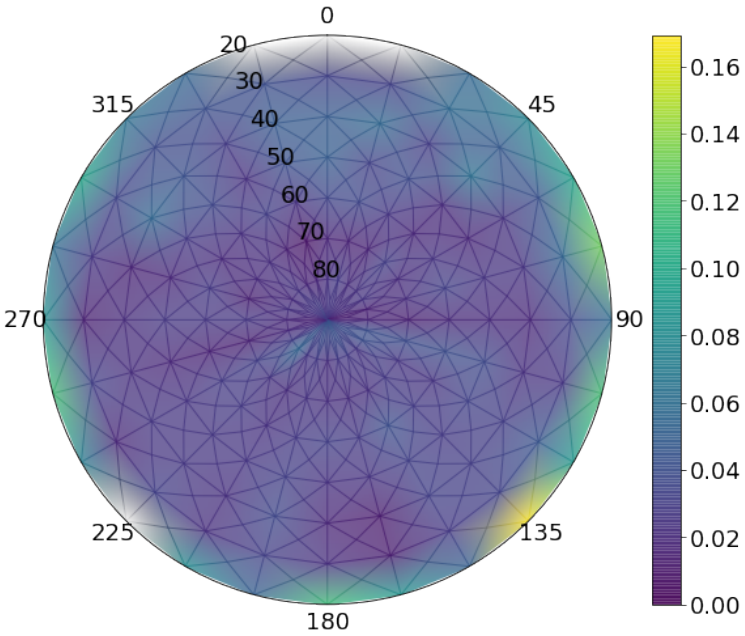


Figure 4.1: *RFI occupancy as a function of the telescope pointing direction for the clean band. We can notice a hot spot at low elevation and azimuth of 135° which is pointing towards nearby towns.*

This result shows that at low elevation angles across all the azimuth we seem to pick more RFI, which implies that the data from these directions would be detrimental due to RFI. Hence, the observer needs to know such information so that corrective measures can be taken into consideration when performing observation planning, as well as data reduction and data analysis.

4.2 Clean band as a function of time

In this section, we discuss the evolution of RFI as a function of time. By looking at the statistical distribution of the RFI probability alerts can be triggered if certain events do not follow the known distribution. In Chapter 3 we showed the histogram of probability distribution (Fig.3.3) of specific hours. It is evident from the distributions that there is some form of anomalies in the data. Hence it is when such anomalous events are observed that they can be used to trigger alerts. Figure 4.2 shows the RFI probability as a function of time of the day for the clean band with the green region representing the 68% confidence interval. The RFI occupancy of this is on average around 2.6%. We cannot see any clear extreme variation between the hours of the day and the RFI occupancy with 68% confidence level.

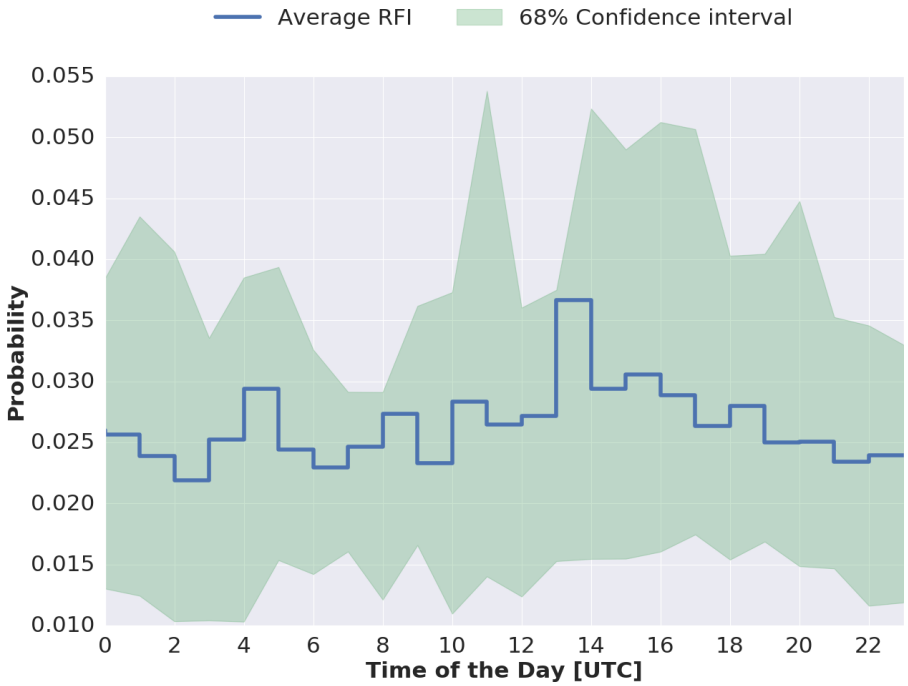


Figure 4.2: *RFI occupancy as a function of time of the day in UTC for the clean band.*

We took a step further and looked at the RFI evolution over months of the year 2018 intending to check whether the RFI status has changed on-site, Fig 4.3 depicts the results. To calculate the RFI average per month, we have taken into account the fact that RFI is dependant on the baseline length as shown in Fig. 3.6. Hence, for all the months we used an equal number of observations in each baseline bin. For example, let us consider a baseline index j, with n being the minimum number of observations, then, the weighted average and the standard deviation can be calculated using Equation 4.3 and 4.4 respectively.

$$\theta_j = \sum_{n=1}^{n} \left(\frac{t_{j,k} \times \theta_{j,k}}{\sum_{k=1}^{n} t_{j,k}}\right) \tag{4.3}$$

$$\sigma_j = \sqrt{\frac{\sum \theta_{j,k} - \theta_j}{n-1}} \tag{4.4}$$

where $t_{j,k}$ is the observation length in seconds and $\theta_{j,k}$ is the probability of RFI of a particular observation k in a particular baseline j. The average was weighted using the length of the observation, such that long observation contribute more as compared to the short observations.

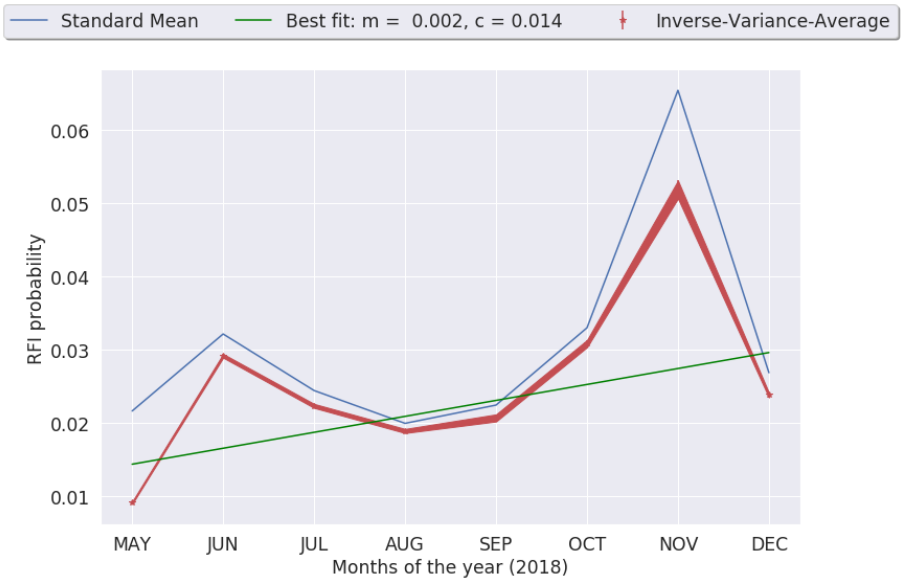


Figure 4.3: *RFI occupancy as a function of month of the year 2018 in the clean band.*

As a result, each month would have an array of RFI probability as function of baseline length with the associated standard deviation. Finally, in order to get the

average probability per months, we calculated the mean of the baseline array in two ways. The first way is to compute the standard average, which defined as follow:

$$\theta_{month} = \frac{\sum_j^N \theta_j}{N} \tag{4.5}$$

where N is the total number of baselines and j is the baseline index in question. Meanwhile, the second way we have taken into account the error associated with each θ_j by calculating the inverse-variance-weighted average and its associated variance using Equation 4.6 and 4.7 respectively. The inverse-variance-weighting method is used when combining two or more values to minimise the variance of the weighted average, each value is weighted in inverse proportion to its variance.

$$\theta_{Inv\_Var\_Mean} = \frac{\sum_j \frac{1}{\sigma_j^2}\theta_j}{\sum_j \frac{1}{\sigma_j^2}} \tag{4.6}$$

$$D^2(\theta_j) = \frac{1}{\sum_j \frac{1}{\sigma_j^2}} \tag{4.7}$$

In Figure 4.3 the blue line is the standard average (Equation 4.5), and the red line is the inverse-weighted-average (Equation 4.6) with the associated error bars. The green line is the linear fit to the data. It can be noticed that there was a significant increase in the fraction of flagged data in November 2018 were the RFI was approximately 300% higher compared to September.

We noticed two peaks for June and November 2018. We, therefore, looked at the distribution of RFI as a function of baseline length for these two months, to check if we can see a difference between the two RFI distributions. Figure 4.4 shows the results, we found that outside the core (baseline > 1000 m), the probability in November is twice as much as in June. Quite interestingly, at the core, we see a much bigger jump. It puzzled us that the probability of RFI in June did not go down as the baseline length increases. This may suggest that the RFI source is internal. Also, between baseline length 1000 m and 5000 m, we do see a jump in probability (blue points with a probability of 0.05). This jump can be explained by the fact that at least one of the imaging observation in June had an antenna where

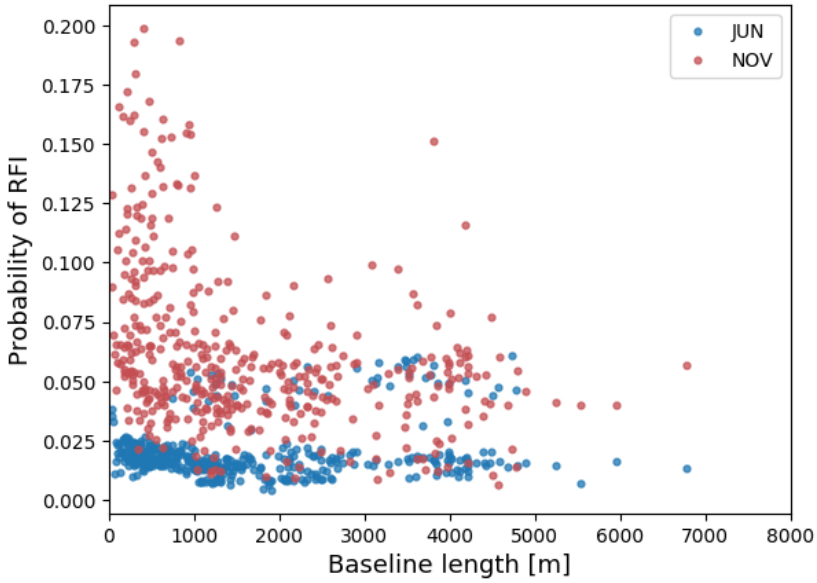


Figure 4.4: *RFI probability as function of baseline length for the MeerKAT clean band for June and November. At the outer core (baseline length > 1000 m) the RFI probability in November is twice as high as in June. At the core we see a bigger jump in November.*

during tracking something went wrong with signal level.

On the other hand, the RFI probability distribution for the full MeerKAT L-band does show a clear reduction of RFI as a function of baseline length as expected (Fig. 4.5). However one can still observe the drift of the probabilities over time for the two months. According to the MeerKAT commissioning report, in November 2018 there were fewer science observations as compared to other months. This is because engineering activities were performed onsite. It is most probable that those activities could have generated RFI while small filler imaging observations were being carried out. However, looking at the MeerKAT telescope activity usage Fig. 4.6, one can notice that for months like May, August and September the telescope was mostly utilised for engineering purposes but the RFI levels are not as high as compared to the month of November. Following the above discussion one can be tempted to rule out the claim that engineering activities may be the cause of the increase in RFI in November, but to make such claim one needs to have a better intuition about the nature and location of the engineering activities. Investigating the nature and locating the RFI activities is beyond the scope of this study.

In Chapter 3 we showed that the average RFI behaviour is well described by 68% confidence interval since the 95% confidence limit includes all sort of outliers. The

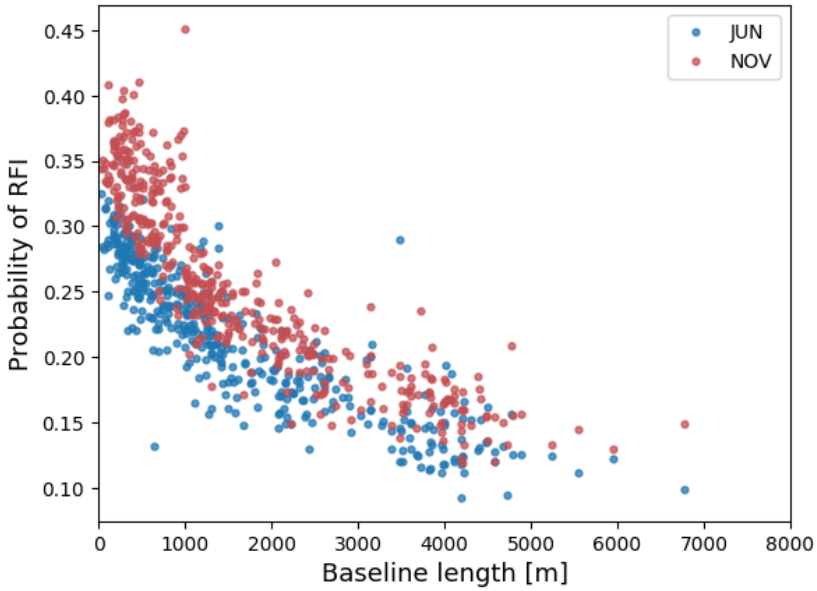


Figure 4.5: *RFI probability as a function of baseline length of the full MeerKAT L-band spectrum for June and November months. We see a clear drop in RFI occupancy as the baseline length increases, however, there is drift is in the RFI occupancy over the two months.*

results from this chapter have shown that there is a great increase in RFI occupancy in November relative to the other months. This increase in RFI occupancy in this month can be used to account for the huge variation in the RFI distribution when looking at the 95% confidence limit. Therefore, we conclude that most of the outliers found in Chapter 3 when looking at the 95% confidence limits are mostly limited to the month of November.

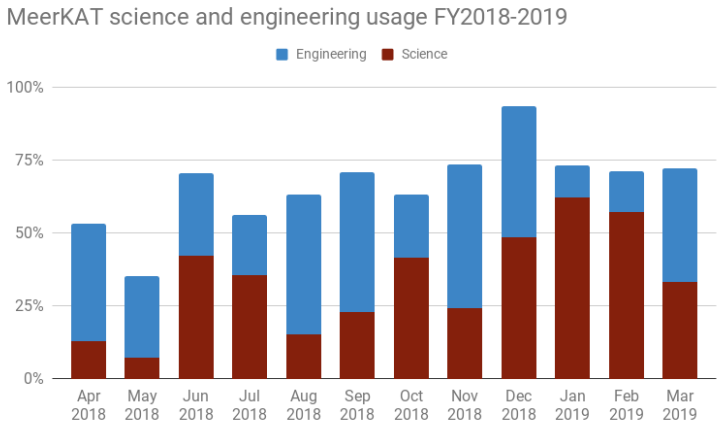


Figure 4.6: *The usage of the MeerKAT telescope for science and engineering (Source : SARAO Internal commissioning documents.)*

4.3 Conclusion

In this chapter, we looked at RFI occupancy in the MeerKAT clean band. In summary, this analysis has shown that at low elevations (elevations < 30°) the RFI occupancy is quite high with an average of about 8%. The highest probability of RFI points towards a region of the nearby towns.

The analysis we carried out allowed us to track the evolution of RFI as a function of time. We found that the RFI occupancy has increased significantly in the month of November 2018 relative to other months. It remains unclear to us what has caused such a huge drift in RFI occupancy in November. However with this historical baseline known, one can also provide alerts about such sudden changes. This could be due to new sources of RFI or stem from any outliers in the data. These outliers could indicate telescope or correlator issues, just to mention a few. Hence the KATHPRFI framework also provides a window into the operational health of the telescope.

Chapter 5

Conclusion and Future Work

In this thesis, we have presented a general framework that summarises statistically the average RFI environment around MeerKAT into a 5-D array database using the most sensitive RFI instrument available, namely the MeerKAT Telescope itself. We refer to this framework as the MeerKAT Historical Probabilities of RFI (KATHPRFI). The dimensions of the array are: *time of the day*, *frequency*, *baseline length*, *elevation* and *azimuth* angle. The framework can be adapted to any radio telescope.

We looked at ways in which RFI is handled at MeerKAT; different methods are discussed ranging from global regulatory protection, national spectrum management policies and RFI detection and excision techniques. In Chapter 2, we described how we created the 5-D array database for RFI. We started by discussing the MeerKAT data pipeline followed by the MeerKAT RFI detection strategy that is based on the *AOFLAGGER* algorithm, originally developed for LOFAR. The MeerKAT RFI detection algorithm was used to generate the data to populate our 5-D array database. We also discussed the MeerKAT data structure and how to extract, transform and load the data. We then described the motivation behind building such a framework focusing on the following use cases; observation planning, telescope operation and site monitoring. Following our motivation the high-level requirements, as well as the design approach and design decision, were discussed. In this thesis we are interested in quantifying statistically the RFI environment of the MeerKAT site, therefore we described in full the KATHPRFI framework. The python programming language together with its big data packages such as Numpy, Dask, Xarrays and Zarrays were used to optimise the creation of the database.

In Chapter 3 we demonstrated the usefulness of the 5-D array database by computing the probability of RFI for different attributes. In general, in all the 1-D probability plots large variation was observed in the distribution of the RFI when looking at the 95 percentile confidence interval. The large variation has been identified as a result of a long tail in the distribution which indicates anomalies that are not related to RFI. In this study we are interested in RFI so the anomalies were not investigated further, however, indications suggest system problems such as correlator spitting zero visibilities, as described in Appendix A.2. Moreover, the RFI environment for MeerKAT is described by average behaviour which is well represented by the 68 percentile confidence interval. One of the outcomes of this work is the discovery of rare RFI in the clean band that is missing from typical observations, illustrating the value of combining a large amount of data across many months, observers and science programs.

This study has shown three regions in the MeerKAT spectrum where the frequencies are correspondingly allocated to the GSM, the DME band and the GPS band. These transmitters are extremely harmful to science observations because of their 100% duty cycle and band usage. In these regions recovering scientific quality observation is almost hopeless.

Finally, this analysis allowed us to validate some of the claims and hypothesis using MeerKAT commissioning imaging observations. These results have confirmed that during the day time, the RFI probability is high as compared to the night time, and it has also confirmed that at low Elevation angles the RFI probability is higher. One finding that came out from this analysis that the highest probability of RFI over the Azimuth point towards regions that include nearby cities. Overall, the detected RFI occupancy for MeerKAT site as function of *time*, *frequency*, *baseline*, *elevation* and *azimuth* as shown by the KATHPRFI data set is **22.9%**.

In Chapter 4 we investigated the RFI occupancy from the clean band. We found that most of the anomalies detected in Chapter 3 were limited to the months of November.

5.0.1 Future work

Radio astronomers always flag the outliers from their data without caring much about their causes. On the other hand, as an observatory, one has to ensure the

best quality of data is obtained from the telescope. However, to do so, characterising the telescope site is quite complex. There are many challenges that one can face in trying to understand the RFI environment of a particular instrument due to the complex nature of signals. This study has presented a broad range of possibilities for future research. The framework can be continuously improved by getting feedback from the telescope users. To reduce the huge observed variation in the probability, a proper selection model of the data must be employed. Data sets that have some system problems should not be included in the framework. Further research is required in order to develop an online visualisation tool that users can efficiently work with and have a quick understanding of the RFI environment.

In this analysis, we randomly choose commissioning imaging observations without proper selection model of good datasets. Recently, the MeerKAT array has started to do science dedicated observations that we can run our framework on. This will help to reduce some of the anomalies that we observed that may not be related to RFI but rather stem from system problems. Moreover, we know that RFI is polarised, but in this analysis, we only looked at the HH polarisation product. Future work can also look at the other polarisations which will help to isolate the RFI.

Finally, additional research is required to integrate the KATHPRFI framework into the MeerKAT data processing pipeline. This framework can be used as a quality assessment of the data.

Appendix A

Modelling of RFI

A.1 Maximum Likelihood Estimate for MeerKAT RFI Occupancy

The KATHPRFI data is generated by a process that only has two possible outcomes:

- RFI $\equiv 1$
- NO RFI $\equiv 0$

In statistics such processes are well described by the **Bernoulli process**, which is a sequence of Bernoulli trials Cauffriez (2017), in which:

- for each trial i, the value of X_i is either 0 or 1
- the trials are independent of each other
- the probability of success is the same for each trial.

The independence of the trials implies that the process is memory-less; that is to say, given that the probability p is known, past outcomes provide no information about future outcomes. We assume that the underlying distribution does not change as a function of time. Now we want to estimate the probability of RFI at one particular voxel, t_i,ν_j, b_k, El_l, Az_m.

Let X be a random variable. We want to estimate the parameter θ, which is the probability of observing RFI. The random variable X has a Bernoulli distribution, and it equals to 1 if the outcome is RFI and 0 if it is not RFI, which can be written

as follows,

$$P(X = 1) = \theta \tag{A.1}$$

$$P(X = 0) = 1 - \theta \tag{A.2}$$

where $0 \leq \theta \leq 1$. Now we can write the probability mass function g of this distribution over all possible x outcomes as follows:

$$g(x; \theta) = \theta^x (1 - \theta)^{1-x} \tag{A.3}$$

where $0 \leq x \leq 1$. The likelihood of observing parameter θ given the data X is.

$$P(\theta|X) = \begin{cases} \theta, & \text{if } X = 1 \\ 1 - \theta, & \text{if } X = 0 \end{cases} \tag{A.4}$$

Since the events are independent, then the total probability of observing all the data is the product of seeing each data point individually (Ross 2014).

$$L(\theta|X) = \prod_{i=1}^{N} \theta^{x_i} (1 - \theta)^{1-x_i} \tag{A.5}$$

where N = total number of ones (i.e RFI) + total number of zeros (i.e Not RFI).

We can now take the natural logarithm of the likelihood expression. This is fine because the natural logarithm is a monotonously increasing function, that is to say, the maximum of the density in question will be the same as the maximum of the natural log transformation.

$$\log(L) = \sum_{i=1}^{N} \log(\theta^{x_i}) + \log((1 - \theta)^{1-x_i}) \tag{A.6}$$

$$= \sum_{i=1}^{N} x_i \log(\theta) + (1 - x_i) \log(1 - \theta) \tag{A.7}$$

Now, we can calculate the value of θ that results in giving the maximum value of the Likelihood function by equating its first derivative to zero.

$$\frac{d}{d\theta} \log(L) = \sum_{i=1}^{N} \left(\frac{x_i}{\theta} - \frac{1 - x_i}{1 - \theta} \right) = 0 \tag{A.8}$$

Suppose that α is the number of RFI (*master* array), and β is the number of NON-RFI data points (*counter* array - *master* array), such that:

$$N = \alpha + \beta \tag{A.9}$$

Then we can split Equation A.8 and substitute the values of X as follows:

$$\frac{d}{d\theta}\log(L) = \sum_{i=1}^{\alpha}(\frac{1}{\theta} - \frac{1-1}{1-\theta}) + \sum_{i=1}^{\beta}(\frac{0}{\theta} + \frac{1-0}{1-\theta}) \tag{A.10}$$

$$= \sum_{i=1}^{\alpha}\frac{1}{\theta} - \sum_{i=1}^{\beta}\frac{1}{1-\theta} \tag{A.11}$$

$$= \frac{\alpha}{\theta} - \frac{\beta}{1-\theta} \tag{A.12}$$

Now we equate the derivative to zero and solve for the parameter θ.

$$\theta = \frac{\alpha}{\alpha + \beta} \tag{A.13}$$

The parameter θ, is a coarse estimate of the probability, especially if $\alpha + \beta$ is small. Intuitively it does not capture the uncertainty about the value of θ. Just like with any other random variables, it often makes sense to hold a distributed belief about the value of θ. To formalise the idea, instead of having a single value (Maximum Likelihood) of the parameter θ we want a distribution. In order to do that we will use the Bayes theorem.

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} \tag{A.14}$$

where,

- $P(\theta|X)$: Is called the **Posterior** term which is the probability of observing RFI after taking into account the evidence
- $P(\theta)$: Is called the **Prior** term which is the probability of seeing an RFI before taking a new evidence into account.
- $P(X|\theta)$: Is called the **Likelihood** term this tells how often we observe the RFI given the data.
- $P(X)$: Is called the **evidence** term which support or opposes the Prior.

but,

$$P(X|\theta) = L(\theta|X) \tag{A.15}$$

Since we have the data, then we can compute the likelihood of observing parameter θ given the data X (i.e $L(\theta|X)$) as shown in Equation A.1. Since the posterior term and the likelihood term are equal, then Equation A.14 becomes:

$$P(\theta|X) = \frac{L(\theta|X)\prod(\theta)}{P(X)} \tag{A.16}$$

Assuming that we know nothing about the prior information, will start with a uniform prior, we know that $\theta \in \{0,1\}$ and we want $\int p(\theta)d\theta = 1$, then this implies that:

$$p(\theta) = 1 \tag{A.17}$$

The evidence term is the joint probability of the data and the parameter which we can write as follows,

$$P(X) = \int p(\theta, X)d\theta \tag{A.18}$$

$$= \int p(X, \theta)d\theta \tag{A.19}$$

Hence, Equation A.16 becomes,

$$P(\theta|X) = \frac{\prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{1-x_i}.1}{\int \prod \theta^{x_i}(1-\theta)^{1-x_i}d\theta} \tag{A.20}$$

Now we will look at each scenario of the two possible outcomes:

Case 1: $x_i = 0$

$$\theta^{x_i}(1-\theta)^{1-x_i} = \theta^0(1-\theta)^1 - 0 \tag{A.21}$$

$$= 1-\theta \tag{A.22}$$

Case 2: $x_i = 1$

$$\theta^{x_i}(1-\theta)^{1-x_i} = \theta^1(1-\theta)^{1-1} \tag{A.23}$$

$$= \theta \tag{A.24}$$

Then we can write Equation A.20 as follows:

$$P(\theta|X) = \frac{\prod_{i=1}^{\beta}(1-\theta)\prod_{j=1}^{\alpha}\theta}{\int \prod_{i=1}^{\beta}(1-\theta)\prod_{j=1}^{\alpha}\theta} \tag{A.25}$$

$$= \frac{(1-\theta)^{\beta}\theta^{\alpha}}{\int (1-\theta)^{\beta}\theta^{\alpha}d\theta} \tag{A.26}$$

Equation A.25 is the definition of the **Beta** distribution if we let,

- a = $\alpha + 1$
- b = $\beta + 1$

$$\textbf{Beta}(a,b) = \theta^{a-1}(1-\theta)^{b-1} \tag{A.27}$$

Thus we can write $P(\theta|X)$ as a beta distribution.

$$P(\theta|X) = \frac{(1-\theta)^{\beta}\theta^{\alpha}}{\int (1-\theta)^{\beta}\theta^{\alpha}d\theta} \tag{A.28}$$

$$= \textbf{Beta}(\alpha+1, \beta+1) \tag{A.29}$$

The **Beta** distribution of $\alpha = 1$ and $\beta = 1$ is a uniform distribution,

$$\text{Beta}(1,1) = \theta^{1-1}(1-\theta)^{1-1} \tag{A.30}$$

$$= 1 \tag{A.31}$$

Then the above allows us to write Equation A.29 as follows,

$$P(\theta|D) = \frac{\theta^{\alpha}(1-\theta)^{\beta}Beta(\alpha+a+1, \beta+b+1)}{\int (1-\theta)^{\beta}\theta^{\alpha}d\theta} \tag{A.32}$$

$$= Beta(\alpha+a+1, \beta+b+1) \tag{A.33}$$

As a result the distribution of our belief about θ before ("prior") and after ("posterior") can both be represented using a Beta distribution. Thus this means that Beta distribution is a conjugate before the Bernoulli likelihood function.

In order for us to calculate the uncertainty associated with the estimate of the parameter θ, we will make use the cumulative distribution function (cdf). The cdf F,

of a real number value X evaluated at θ is the probability that X takes on a value less than or equal to θ.

$$F_X(\theta) = P(X \leq \theta) \tag{A.34}$$

where P is the probability distribution function as define is Equation A.29. Now we will want to calculate the confidence on the values, we will use a 95% confidence interval. This is a range of values that we are 95% certain contains the true mean of the population.

We can calculate the lower and the upper bounds respectively as follows:

$$0.025 = \int_0^{\theta_{low}} P(\theta) d\theta \tag{A.35}$$

$$0.975 = \int_0^{\theta_{upper}} P(\theta) d\theta \tag{A.36}$$

Since we are dealing with a discrete variables we can use numerical approximations methods to estimate the bounds using the following equations,

$$0.025 = \sum_i^{\theta_{lower}} P(\theta) \Delta\theta \tag{A.37}$$

$$0.975 = \sum_i^{\theta_{upper}} P(\theta) \Delta\theta \tag{A.38}$$

For us to calculate the uncertainty associated with the parameter θ, we have divided the master by the counter array without doing the summing over other dimensions to the 5-D probability array. Then we flatten the 5-D array over all other attributes except the one of interest. This would give us the probability values that have contributed to make one particular data point. Below we will discuss the uncertainty of the parameter θ for different attributes.

A.1.1 Uncertainty analysis in Time of day

We have computed the uncertainty of the probability of RFI that is associated with each hour of the day. Figure A.1 shows the probability of RFI as a function of

the day with 95% confidence interval multiplied by 50 to make them visible. This shows that the statistical uncertainties are negligible compared to the systematic variation of the RFI. It can be noticed that the uncertainty around the parameter θ is small, this is due to the nature of the **beta** distribution. So, as a result, we could not use the Beta distribution to model our RFI behaviour.

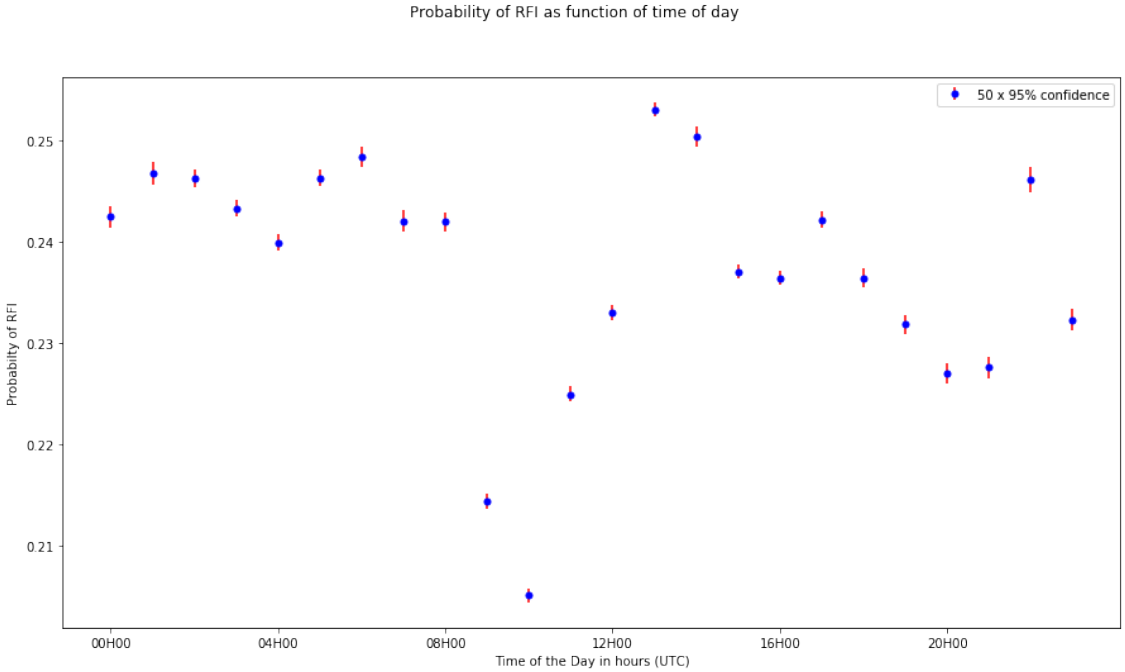
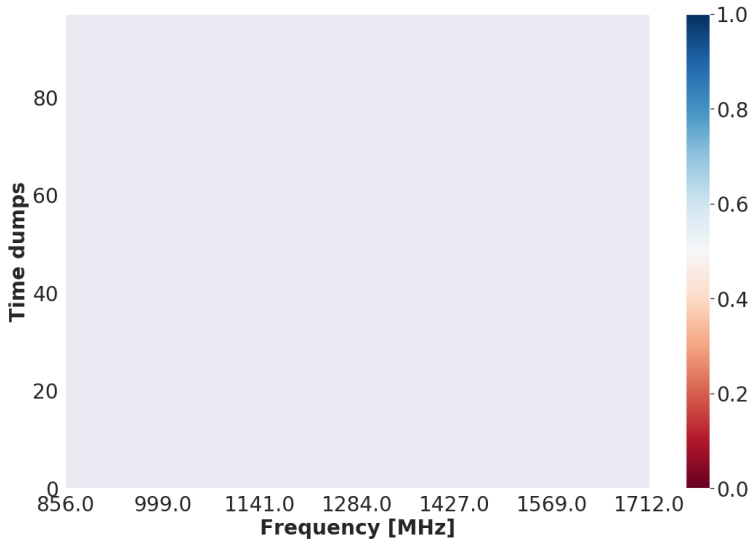


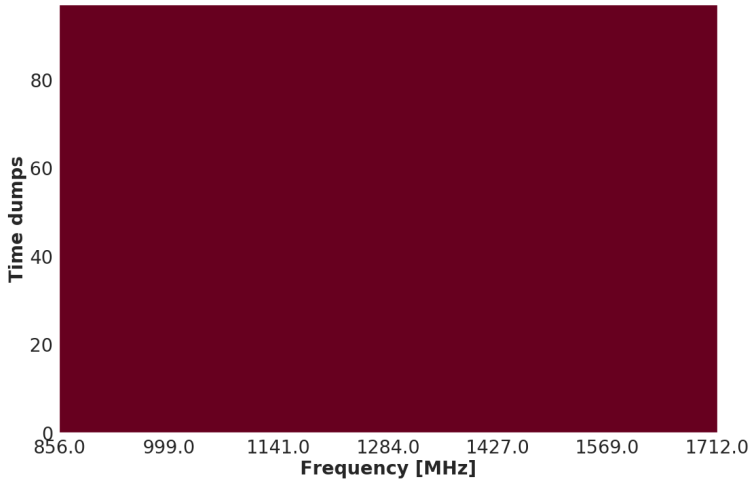
Figure A.1: *Probability of RFI as a function of hour of the day with 95% confidence interval multiplied by 50 to make them visible. This shows that the statistical uncertainties are negligible compared to the systematic variation of the RFI.*

A.2 Analysis of Observation Files that had zero and one RFI Probability.

As discussed in Chapter 3 we have investigated the observation file, *1531304170-sdp-l0.full.rdb*, that had zero RFI probabilities. We randomly choose a baseline that had zero RFI probability and plot its visibility and its corresponding detected RFI flags by the in-house flagger, Fig A.2 shows the results. We can note that the visibilities from this baseline are zero, as a consequence the MeerKAT RFI flagger algorithm does not detect any RFI for all the frequencies for the whole observation. Hence, as a result in our analysis, such baseline has zero RFI probability. It is important to note that when we observe zero RFI probability it is an indication that the correlator was outputting zero visibilities.



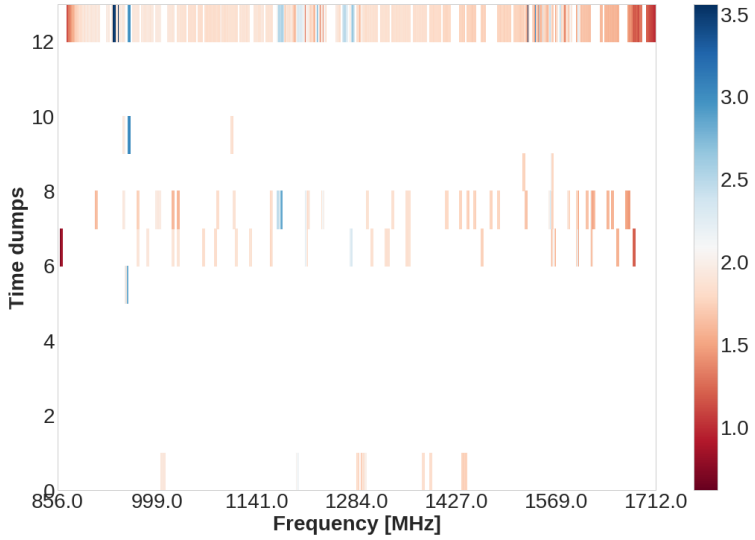
(a) *Visibilitiies from baseline ['m023h'-'m060h'] where everything is zero for whole 10 minutes observation.*



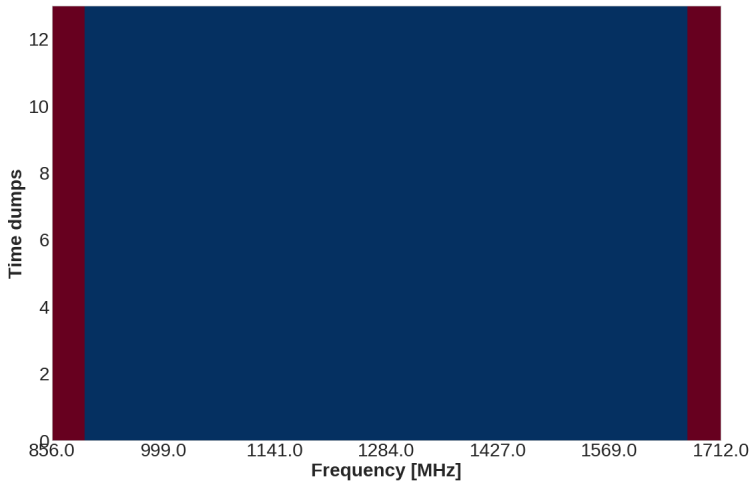
(b) *RFI flags generated by the in-house MeerKAT RFI flagger. The flagger did not detect anything in all the frequencies at all times due to zero visibilities from the correlator.*

Figure A.2: *A typical example when the correlator output zero visibilities. Therefore, the MeerKAT RFI flagger does not detect any RFI, hence, we see such baseline with zero RFI probability in our analysis.*

Similarly, we have looked at observation file, *1539872205-sdp-l0.full.rdb*, that had RFI probability greater than 80% on a particular baseline, Fig. A.3 depicts the results. It can be noticed that if the visibilities are not smooth, then the RFI detection algorithm flags everything in that baseline as RFI.



(a) *Visibilitiies from baseline where the background is not smooth.*



(b) *RFI flags generated by the in-house MeerKAT RFI flagger. The flagger detected 90% of the spectrum as RFI leaving out only the edges.*

Figure A.3: *A typical example when the MeerKAT RFI flagger algorithm flags all the data from a particular baseline*

Appendix B

Analysis of the known RFI sources

In this chapter we will perform similar analysis as described in Chapter 4 for the known RFI transmitters in fixed bands which are allocated to the mobile communication system, aircraft distance measurement equipment and GPS satellites. In Section B.1 we will discuss the impact from Global System for Mobile Communication (GSM) band, the aircraft Distance Measuring Equipment (DME) will be discussed in Section B.2 and finally the Global positioning System (GPS) satellites shall be discussed in Section B.3.

In each section, we investigated how much RFI we observed from those culprits as a function of time and direction.

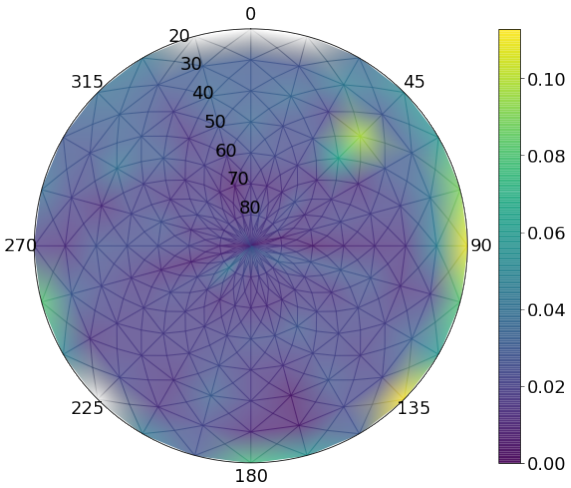
B.1 GSM band

The Global System for Mobile Communication (GSM) is an open, digital cellular technology used for transmitting mobile voice and data services. The GSM band is divided into two major sub-bands, based on their frequencies, the GSM-900 and the GSM-1800. In a GSM network, the term uplink frequency is used for a band of frequencies dedicated for transmitting data from the mobile phones to the Base Transceiver Station (BTS) towers, whereas the downlink frequency is used for frequencies that transmit data from the BTS to the mobile cellphone.

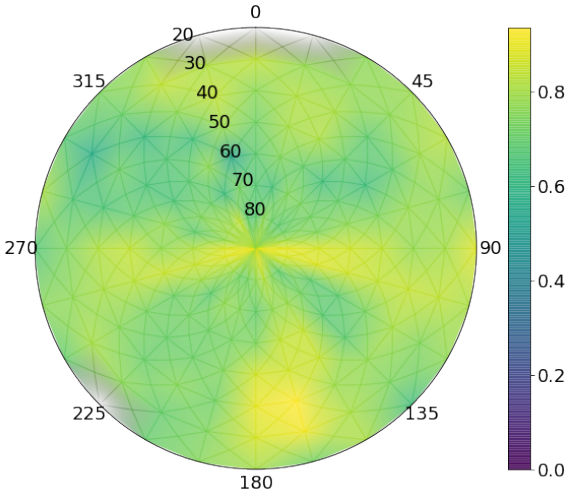
B.1.1 RFI occupancy as a function of telescope pointing direction

The uplink frequency in a GSM-900 network generally lies between a range of 890 and 915 MHz, making a bandwidth of 25 MHz. This band contains multiple

frequencies assigned to different users to facilitate communication. As per the Independent Communications Authority of South Africa (ICASA) document [1], the frequency band 880 MHz - 960 MHz are fully utilised by the mobile phone service providers in South Africa which are Cell-C, MTN, Telkom and Vodacom. Figure B.1 shows the RFI probability as function of direction for the GSM sub-bands.

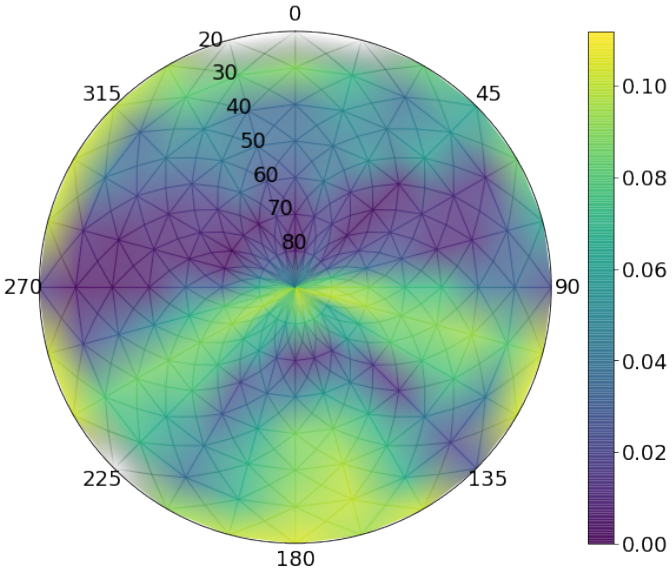


(a) *The GSM-900 Uplink band, the three hot-spots are pointing towards nearby towns.*



(b) *The GSM-900 Downlink band.*

[1] https://www.icasa.org.za/uploads/files/SpectrumUsage_Q1-2013.pdf



(c) GSM-1800 Uplink

Figure B.1: *The probability of RFI for GSM sub-bands bands as a function of azimuth and elevation as measured from MeerKAT.*

From Fig. B.1a we noticed three hot-spots (Az:90° and El: 20° - 30°, Az: 135° and El: 20° - 30°, Az: 45° and El: 40° - 50°). The hot-spot at azimuth 45° is pointing towards a communication tower at the nearby town called Prieska (Fig. 3.7). Meanwhile, the hot-spots at 90° and 135° are pointing towards Vosburg town and Loxton town respectively. We found that at low the RFI generated from this band is concentrated at lower elevations. This reveals that the GSM-900 uplink band is quite directional.

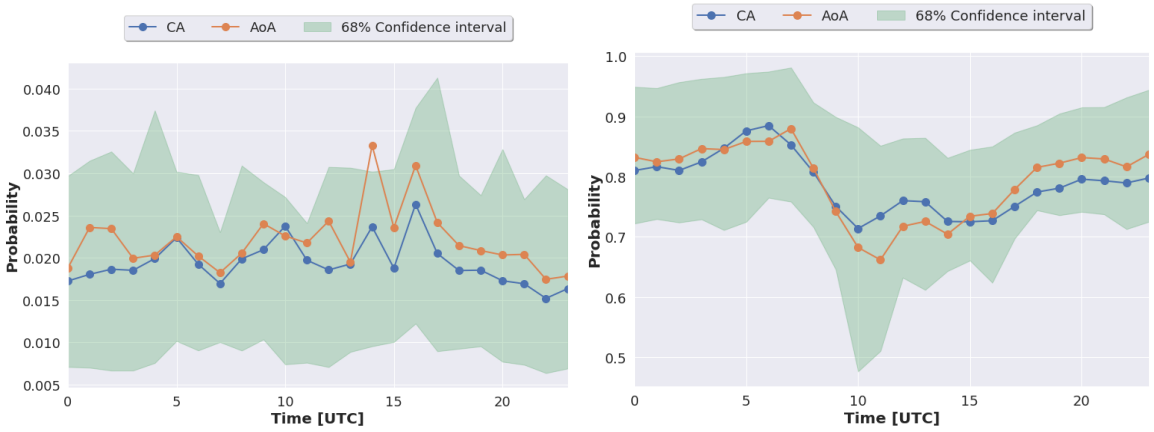
Figure B.1b shows the distribution of RFI occupancy for GSM-900 downlink band that is 935 MHz - 960 MHz. This result shows that the signal from this band is quite distributed oversite, with a hot-spot between 135° and 180° azimuth. This direction corresponds to the one that we observed higher RFI occupancy in the 1-D plot (Fig.**??**) of azimuth. We have shown that according to the MeerKAT RFI monitoring system these directions point towards nearby towns. We also notice a dipole-like structure at azimuth around 110° and 260° were the RFI occupancy is quite high, this point towards two communication towers in Carnavon town and Calvinia town respectively.

The GSM-1800 uplink frequency band ranges from 1710 MHz to 1785 MHz. The MeerKAT L-band receiver can only detect 2 MHz of this frequency range. Quite

strangely and interestingly, the RFI occupancy in this band is also quite distributed over site (Fig. B.1c). We notice a similar dipole-like structure as in the GSM-900 downlink. This suggests that the source of RFI is most probably the same as the one for the GSM-900 downlink band.

B.1.2 RFI occupancy as a function of time of the day

It is anticipated that during the day the GSM band should be more active as compared to the night time due to human activities. We looked at the sub-bands of the GSM to assess the truth of this hypothesis. Figure B.2 shows the RFI occupancy of the GSM sub-bands as a function of time. The blue and the orange line represent the two different methods that we used to calculate the average RFI probability as explained in the introduction of Chapter 3. The distribution of RFI from the two methods follows each other.



(a) *The GSM-900 Uplink band.* ***(b)*** *The GSM-900 Downlink band.*

Figure B.2: *The probability of RFI for GSM sub-bands bands as a function of time of the day.*

The GSM-900 uplink (Fig. B.2a) shows the average RFI occupancy contribution of 2.5%. When looking at the AoA method we notice that at 14H00 UTC the average RFI is outside the 68% confidence interval. This is fine because unlike the median, the mean is sensitive to outliers. Hence, the mean can be outside the cofidence interval limit. Furthermore, these results indicate there is no significant correlation between the hour of the day and the RFI occupancy due to less usage of cellphone

at night when looking at the 68% confidence limit.

Similarly, Fig. B.2b shows the RFI occupancy for the GSM-900 downlink, with an average of 75%. We found that the night time RFI occupancy is somewhat higher than the day time. The maximum occupancy is observed between 20:00 and 8:00 UTC, while the lowest occupancy is between 8:00 and 19:00 UTC. The figures clearly, shows that the GSM-900 downlink band is generating more RFI as compared to its counterpart, the uplink band.

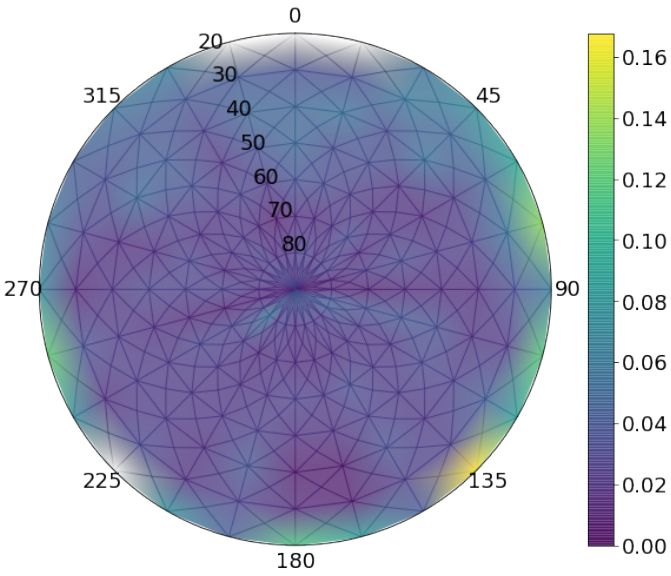
B.2 Distance Measurement Equipment (DME)

The Distance Measurement Equipment (DME) system is used for aircraft navigation. It is a combination of ground and airborne equipment that transmit signals confined between 962 to 1213 MHz. This frequency band is also utilised by the military Tactical Air Navigation (TACAN) system. The ground-to-air transmission is confined within 962 MHz-1024 MHz and 1151 MHz-1213 MHz bands. Whereas, the air-to-ground band is within 1025 MHz - 1150 MHz (Fisher 2004).

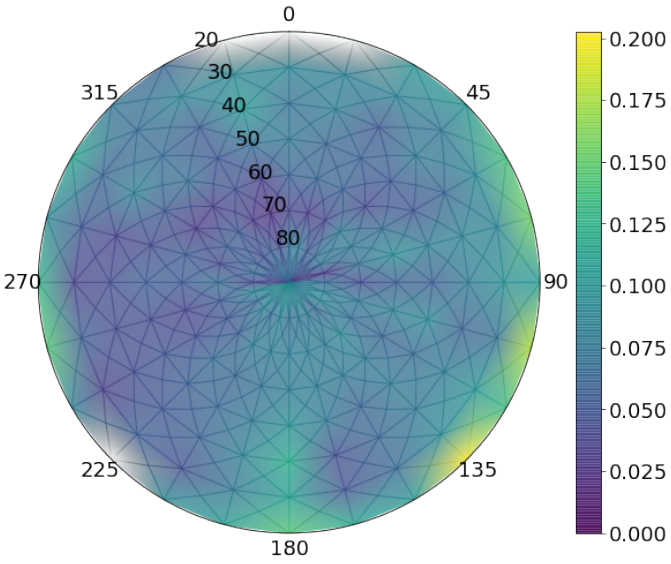
B.2.1 RFI occupancy as a function of pointing direction for the DME band.

We investigated the RFI occupancy contributed by the DME signal as measured by MeerKAT, as they fall within our L-band. Figure B.3 shows the probability of RFI as a function of the pointing direction of the telescope for the DME sub-bands. We noticed that the distribution of RFI occupancy is high at low elevations for the first ground-to-air band, Fig. B.3a. As we go to higher elevations, the RFI occupancy becomes moderate with an average of approximately 3%. On this basis, we can conclude that this sub-band is quite directional.

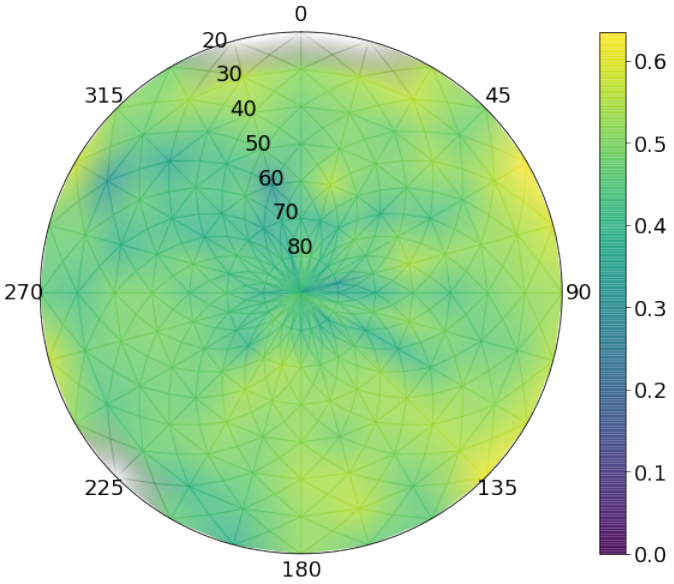
Likewise, we looked at the RFI probability from the air-to-ground DME emission as depicted by Fig. B.3b. The RFI distribution is quite moderate over the site with an average of 1%.



(a) *Ground-to-Air transmission [962-1024 MHz]*



(b) *Air-to-Ground [1025-1150 MHz]*



(c) *Ground-to-Air transmission [1151-1213]*

Figure B.3: *The probability of RFI for DME sub-bands as function of azimuth and elevation as measured from MeerKAT.*

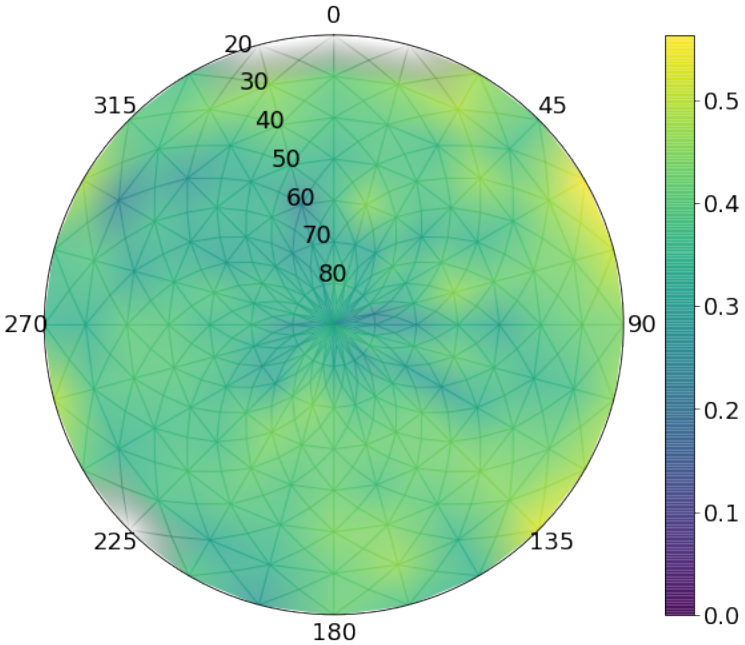


Figure B.4: *The RFI probability of the DME Ground-to-Air 2 band without the L5 frequencies.*

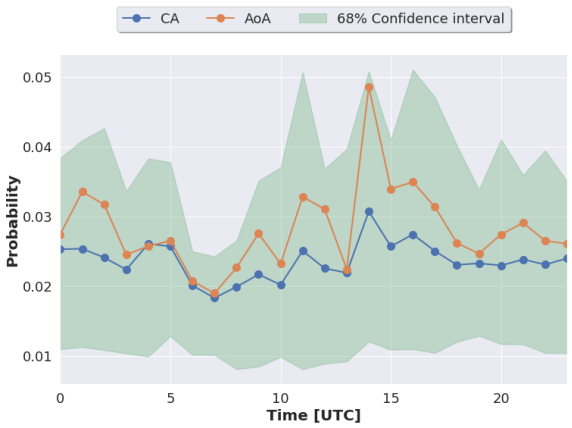
The distribution of the RFI from the second DME ground-to-air sub-band as observed from Fig B.3c is quite distributed over the site, with an average occupancy of more than 50% in all directions. It should be noted that there is an overlap

in frequencies between the GPS L5 band (discussed in Section B.3) and the DME ground-to-air band. To understand the contribution due to the latter band only, we removed the overlapping L5 band frequencies. The result is shown in Fig B.4. We found that the distribution does not change, however, the maximum frequency went down by 10%.

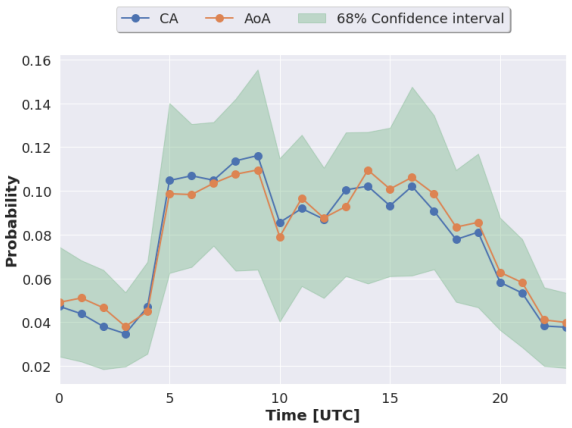
This results may hint that flights crossing site from time to time contribute to corrupting the lower part of the MeerKAT L-band.

B.2.2 RFI occupancy as a function of time of the day for the DME band

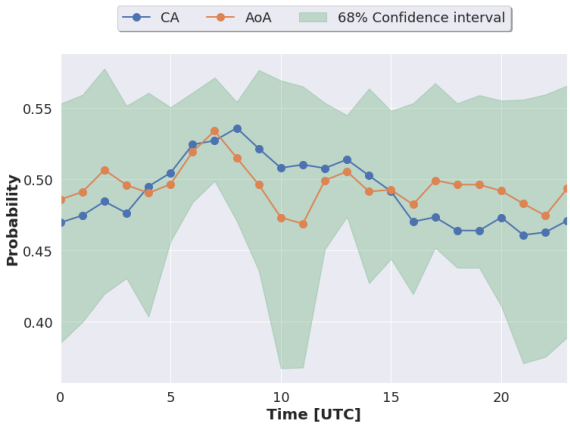
The MeerKAT site is along the Cape Town - Johannesburg flight path, which is the busiest in South Africa. Following the previous analysis, one might expect that the probability of RFI due to the DME transmitters will decrease at night since in general domestic flights operate mostly during the day time. Figure B.5 was used to assess this hypothesis. Figure B.5a and B.5c shows the ground-to-air transmission of the DME band, whereas the Fig. B.5b shows the air-to-ground emission. A clear increase of the RFI probability is observed between the day time and the night time observations for the air-to-ground transmission. This is also evident in the 2-D plot of time and frequency, Fig. 3.1, which show an obvious increase of RFI occupancy during the day, within frequencies that are allocated to the DME band.



(a) *Ground-to-Air transmission 1.*



(b) *Air-to-ground transmission.*



(c) *Ground-to-Air transmission 2*

Figure B.5: *RFI occupancy as a function of time in UTC for the DME sub-bands. The blue and the orange line represents two methods that we used to calculate the average as explained in introduction of Chapter 3*

The RFI distribution of the two DME ground-to-air bands does not show any clear relationship between the hour of the day and the RFI probability.

B.3 Global Positioning System

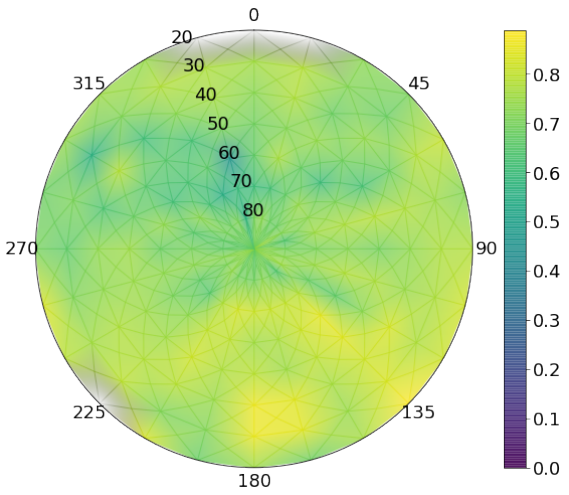
The Global Positioning System (GPS) is currently a system of 31 satellites that uses triangulation techniques to accurately pinpoint the geographical location of the users anywhere on Earth. The GPS satellites transmit radio signals in the L-band region to the GPS receiver (e.g cellphones, car GPS devices). At least between five to eight satellites should be detected by the receiver at any position on earth at any one time to accurately determine the position of the receiver (Misra & Enge 2006).

The radio signals from GPS satellites are subdivided into multiple ranges of frequencies from L1 to L5. All the GPS satellites broadcast at the two frequencies L1 and L2 that has a bandwidth of 15.3 MHz and 11 MHz, with a central frequency of 1575.4 MHz and 1227.6 MHZ respectively. The L1 is the oldest and well developed GPS signal, whereas the L2 is a newer signal compared to the L1, but its infrastructure is not well established relative to the L1. Hence, the L2 and the L1 are used together. The L3 band has a central frequency of 1381.05 MHz and is used by the United States to detect Nuclear Detonation (NUDET) and has a bandwidth of 10 MHz. The L4 band is utilised for ionospheric studies and has a central frequency of 1379.9 MHz and spans 10 MHz of bandwidth. The L5 band has a central

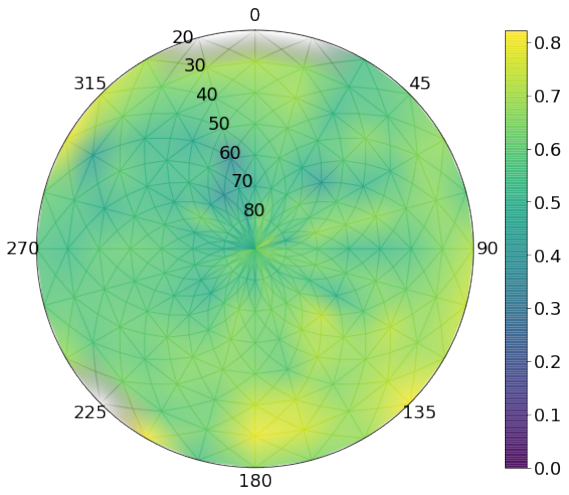
frequency of 1176.5 MHz with a bandwidth of 12.5 MHz. The L5 frequency is the most advanced Global Navigation Satellite System (GNSS) signal but still in development stages and is going to be utilised for the demanding application such as aviation and eventually used for civilisation purposes. In the subsequent sections, we would look at how these sub-bands affect the MeerKAT RFI environment.

B.3.1 RFI probability as function pointing direction for the GPS band

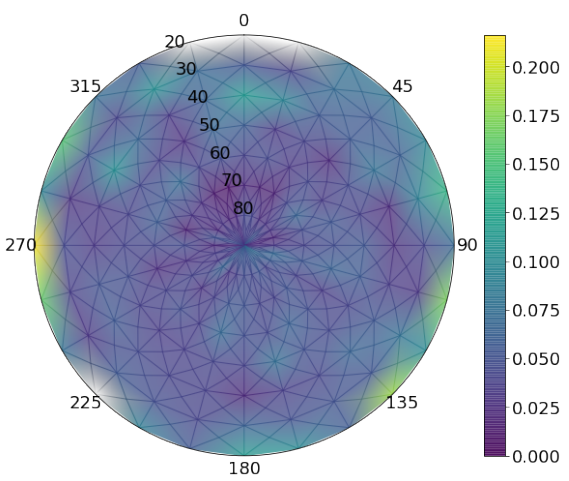
As explained above, everywhere you are on Earth, you should be able to receive signals from at least four GPS satellites. As a result, we expect all the MeerKAT antennas to pick up the GPS signals at all times. Figure B.6 depicts the individual contribution of the GPS sub-bands as a function of Elevation and Azimuth.



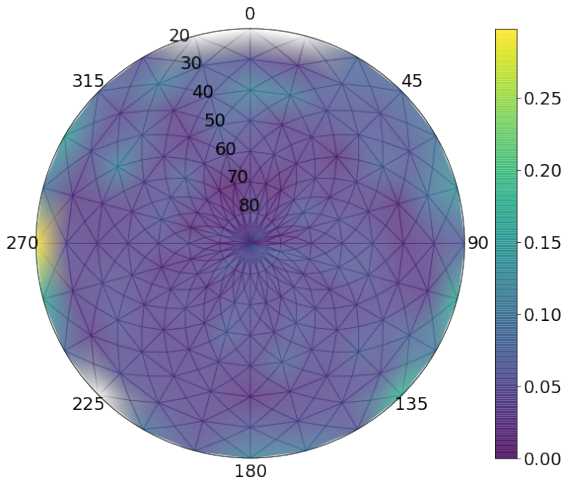
(a) *L1 band*



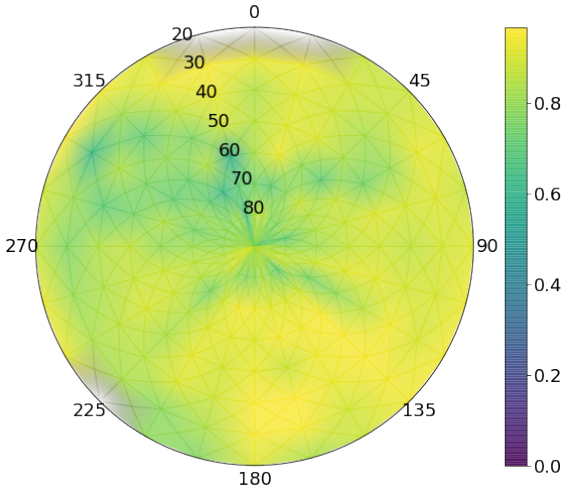
(b) *L2 band.*



(c) *L3 band.*



(d) *L4 band.*



(e) *L5 band*

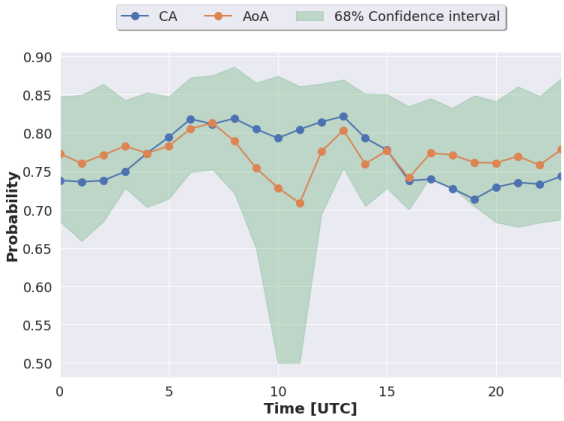
Figure B.6: *The probability of RFI for GPS sub-bands bands as a function of azimuth and elevation.*

It can be noticed that the distribution of the RFI probability for the L1, L2 and the L5 band is fairly smeared overall directions as compared to the L3 and L3 which are quite moderated over the site, we see a hot-spot at lower Elevation (10° - 20°) and Azimuth of 270°. It is interesting to note that the RFI detected from the L5 band is almost constant in all directions with an average of approximately 80°.

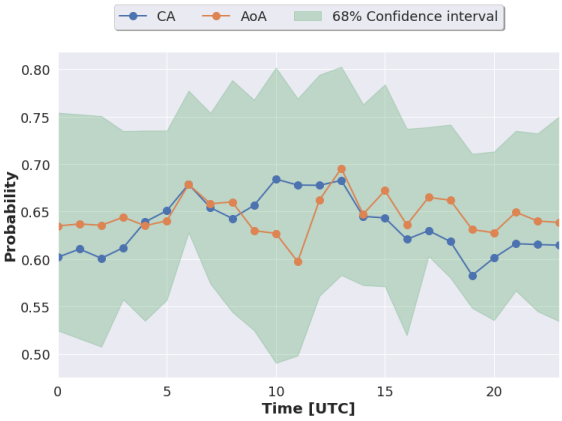
B.3.2 The RFI probability as a function of time of the day for GPS band

One can expect to see a fairly constant RFI occupancy as a function of time of the day. This is because, at all times, all antennas will at least pick GPS satellites in

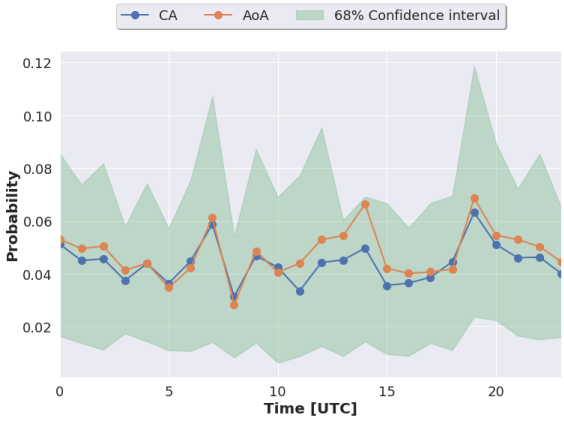
the main or sidelobes. Figure B.7 shows the RFI occupancy as a function for GPS-satellite sub-bands where the blue and orange line represents the two methods used to calculate the average namely, CA and AoA respectively. We found that the L1, L2 and L5 bands have similar RFI occupancy distribution. Furthermore, the L3 and L4 have almost identical RFI occupancy distribution, this may be due to the overlapping of frequencies of the two bands. Collectively, when looking at the 68% confidence level the variation that we are observing of the RFI occupancy as a function of time is negligible. Hence, on this basis, we conclude that indeed the GPS emissions is constant over time.



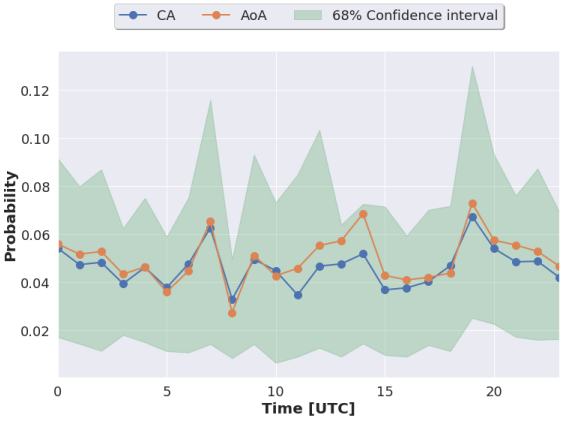
(a) L1 band



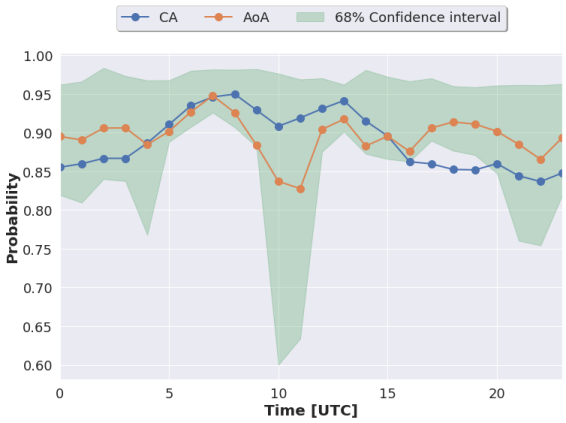
(b) L2 band.



(c) L3 band.



(d) L4 band.



(e) *L5 band*

Figure B.7: *The probability of RFI for GPS sub-bands bands as a function of time of the day. The blue and orange line represents the two methods used to calculate the average namely, CA and AoA. The difference between the two method is explained in the introduction of Chapter 3.*

B.4 Conclusion

In summary, to an extent this analysis allowed us to make the following conclusion, the highest RFI occupancy from GSM band points towards the communication towers in the nearby towns. We did not find a statistically significant correlation between RFI occupancy and the time of the day for the GSM band. As for the DME band, we found that the air-to-ground band is quite directional and is most dominant at low elevations and becomes moderate as we go to higher elevations. Our results have shown that the RFI occupancy from the DME air-to-ground band increases during the day and drops at night time. These times correspond to operational time for the domestic flights as discussed in Chapter 3.

The three GPS sub-bands (L1, L2, L5) are quite smeared all over the site while the L3 and L4 are quite directional and dominant at low elevations. The RFI occupancy for the GPS sub-bands as a function of time is almost constant when looking at the 68% confidence limits.

Appendix C

KATHPRFI Code

```
#!/usr/bin/env python
# coding: utf-8
import katdal
import h5py
import numpy as np
import matplotlib as plt
import xarray as xr
import pandas as pd
import pylab as plt
import time as tme
from dask import array as da
from dask import delayed
from numba import jit, prange
import argparse,os
import six



def readfile(pathfullvis, pathflag):
    '''
    Reading in the full visibility file, and also the flag file.

    Arg : path2full and path to the flag file
    '''

    visfull = katdal.open(pathfullvis)

    flagfile = h5py.File(pathflag)

    #da.from_array(flagfile['flags'], chunks='auto')


    return visfull, flagfile



def remove_bad_ants(fullvis):
    '''
```

```
    This function is going to extrcat the list of all goos antennas.

    Input: Take a fullvis rdb file

    Output: List of good antennas
    '''
    # This pull all the antenna used for observation
    AntList = []
    for ant in fullvis.ants:
        AntList.append(ant.name)

    # This will give the antenna activity list
    AntsActivity = []
    for AntName in AntList:
        AntsActivity.append((AntName, fullvis.sensor[AntName+'_activity']))

    for i in range(len(AntsActivity)):
        if 'stop' in AntsActivity[i][1]:
            AntList.remove(AntsActivity[i][0])
        else:
            pass

    return AntList



def select_and_apply_with_good_ants(fullvis,flagfile, pol_to_use, corrprod,scan,clean_ants):
    from dask import array as da
    '''
    This function is going to select correlation products and apply flag table
    to the full visibility file.

    Arg: full visibility, flagfile, pol to use, corrproducts and scan and good antennas.

    Output : The flag table with ingest rfi flags and cal rfi flags
    '''

    fullvis.select(reset='TFB')
    flags = da.from_array(flagfile['flags'], chunks=(1, 342, fullvis.shape[2]))

    fullvis.source.data.flags = flags

    fullvis.select(corrprods=corrprod, pol=pol_to_use, scans=scan, ants = clean_ants,flags=['cal_rfi','ingest_rfi'])

    flag =fullvis.flags


    return flag
```

```
87
88
89  def get_az_and_el(fullvis):
90      '''
91      Getting the full the elevation and azimuth of the file.
92
93      Arg: fullvis file
94
95      Return: List of avaraged elevation and azimuth of all antennas
        per time stamp
96      '''
97
98      # Getting the azmuth and elevation
99
100     azmean = np.mean(fullvis.az, axis=1)%360
101     elmean = np.mean(fullvis.el, axis=1)
102
103
104     return elmean, azmean
105
106
107 def get_time_idx(fullvis):
108     import datetime
109     '''
110     This function is going to convert unix time to hour of a day
111
112     Input : full visibility data file object
113
114     Output : list with time dumps converted to hour of a day
115     '''
116     unix  = fullvis.timestamps
117
118     local_time = []
119     for i in range(len(unix)):
120         local_time.append(datetime.datetime.fromtimestamp((unix[i]))
        .strftime('%H:%M:%S'))
121
122     # Converting time to hour of a day
123     hour = []
124     for i in range(len(local_time)):
125         h = int(round(int(local_time[i][:2]) + int(local_time[i
        ][3:5])/60 + float(local_time[i][-2:])/3600))
126         if h == 24:
127             hour.append(0)
128         else:
129             hour.append(h)
130
131     return np.array(hour, dtype=np.int32)
132
133
134 def get_az_idx(azimuth,bins):
135     '''
136     This function is going get the index of the azimuth
137
```

```
    Input : List of Azimuthal angle and azimuthal bins

    Output : Azimuthal index
    '''
    az_idx = []
    for az in azimuth:
        for j in range(len(bins)-1):
            if bins[j] <= az < bins[j+1]:
                az_idx.append(j)

    return np.array(az_idx)


def get_el_idx(elevation,bins):
    '''
    This function is going get the index of the elevation

    Input : List of elevation angle and bins

    Output : Elevation index

    '''
    el_idx = []

    for el in elevation:
        for j in range(len(bins)):
            if bins[j] <= el < bins[j]+10:
                el_idx.append(j)

    return np.array(el_idx, dtype=np.int32)



def get_corrprods(fullvis):
    '''
    This function is getting the corr products

    Input : Visibility file

    Output : Correlation products
    '''
    bl = fullvis.corr_products
    bl_idx = []
    for i in range(len(bl)):
        bl_idx.append((bl[i][0][0:-1]+bl[i][1][0:-1]))

    return np.array(bl_idx)


def get_bl_idx(corr_prods,nant):
    '''
    This function is getting the index of the correlation products

    Input  : Correlation products, number of antennas
```

```
192
193     Output : Baseline index
194     '''
195     nant = nant
196
197     A1, A2 = np.triu_indices(nant, 1)
198
199     # Baseline antenna cobinations
200     corr_products = np.array(['m{:03d}m{:03d}'.format(A1[i], A2[i])
        for i in range(len(A1))])
201
202
203
204
205     df = pd.DataFrame(data=np.arange(len(A1)), index=corr_products).
        T
206
207     bl_idx = df[corr_prods].values[0].astype(np.int32)
208
209     return bl_idx
210
211 def get_files(path2flags, path2full):
212     '''
213     This file is going to get the list of datafiles[flag files and
        full visibility]
214
215     Input: path to: flagfiles and fullvis
216
217     Ouput: List of flagfiles and fullvis names.
218     '''
219     path = [path2flags, path2full]
220     import os, fnmatch
221     listOfflags = os.listdir(path[0])
222
223     listOffull = os.listdir(path[1])
224
225     patternflags = "*.h5"
226     patternfull = "*.rdb"
227     dataflags = []
228     datafull =[]
229     for entry in listOffull:
230         datafull.append(entry[0:10])
231     for entry in listOfflags:
232          if fnmatch.fnmatch(entry,patternflags):
233             dataflags.append(entry[0:10])
234
235     data = list(set(datafull).intersection(set(dataflags)))
236
237     fullvis = []
238     flags = []
239     for i in range(len(data)):
240         fullvis.append(data[i]+'_sdp_l0.full.rdb')
241         flags.append(data[i]+'_sdp_l0_flags.h5')
242
```

```
243     return flags,fullvis
244 
245 from numba import cuda
246 
247 @jit(nopython=True, parallel=True, debug=True)
248 def update_arrays(Time_idx, Bl_idx, El_idx, Az_idx, Good_flags,
    Master, Counter):
249     '''
250 from numba import cuda
251 @jit(nopython=True, parallel=True, debug=True)
252     This function is gonna update the master and counter array
253 
254     Input: time_idx, bl_idx, el_idx, az_idx, flags_array, master and
     counter arrays
255 
256     Output: update master and counter array
257     '''
258     cstep = 128
259     cblocks = (4096 + cstep - 1) // cstep
260     for cblock in prange(cblocks):
261         c_start = cblock * cstep
262         c_end = min(4096, c_start + cstep)
263         for k in range(c_start, c_end):
264             for i in range(len(Bl_idx)):
265                 for j in range(len(Time_idx)):
266                     Master[Time_idx[j],k,Bl_idx[i],El_idx[j],Az_idx[
    j]] += Good_flags[j,k,i]
267                     Counter[Time_idx[j],k,Bl_idx[i],El_idx[j],Az_idx
    [j]] += 1
268 
269 
270     return Master, Counter
271 
272 
273 
274 if __name__=="__main__":
275     parser = argparse.ArgumentParser(description='This package
    produces two 5-D arrays, which are the counter array and the
    master array. The arrays provides statistics about measured RFI
    from MeerKAT telescope.',)
276 
277     parser.add_argument('-v',
278                         '--vis', action='store',  type=str,
279                         help='Path to the full rdb visibility files')
280     parser.add_argument('-f',
281                         '--flags',
282                         action='store',  type=str,
283                         help='Path to TOM flag files')
284     parser.add_argument('-b',
285                         '--bad', action='store',  type=str,
286                         help='Path to save list of bad files')
287     parser.add_argument('-g',
288                         '--good', action='store', type=str,default =
    '\tmp',
```

```
                    help='Path to save bad files')
    parser.add_argument('-z',
                        '--zarr', action='store', type=str,default =
    '\tmp',
                    help='path to save output zarr file')
    parser.add_argument('-n','--no_of_files', action = 'store', type
    =int,
                        help='Multiple of number of files to save a
    number between 1 and 10',default=1 )

    args = parser.parse_args()


    #Getting the file names
    flag,f = get_files(args.flags,args.vis)

    #f = f[4:]
    #Initializing the master array and the weghting
    master = np.zeros((24,4096,2016,8,24), dtype=np.uint16)
    counter =np.zeros((24,4096,2016,8,24), dtype=np.uint16)

    # Running the Hp code
    badfiles = []
    goodfiles = []

    for i in range(len(f)):
        print('Adding file {} : {}'.format(i, f[i]))
        try:
            pathfullvis=str(args.vis)+'/'+f[i]
            pathflag = str(args.flags)+flag[i]
            fullvis,flagfile = readfile(pathfullvis,pathflag)
            print('File ',i,'has been read')

            if len(fullvis.freqs) == 4096 and fullvis.dump_period>7
    and fullvis.dump_period<8:
                clean_ants = remove_bad_ants(fullvis)
                print('good ants')
                good_flags = select_and_apply_with_good_ants(fullvis
    , flagfile, pol_to_use='HH', corrprod='cross', scan='track',
                                                             clean_ants=
    clean_ants)
                print('Good flags')
                if good_flags.shape[0]* good_flags.shape[1]*
    good_flags.shape[2]!= 0:

                    el,az = get_az_and_el(fullvis)
                    time_idx = get_time_idx(fullvis)
                    az_idx = get_az_idx(az,np.arange(0,370,15))
                    el_idx = get_el_idx(el,np.arange(10,90,10))
                    print('el and az extracted')
                    corr_prods = get_corrprods(fullvis)
                    bl_idx = get_bl_idx(corr_prods, nant=64)
                    # Updating the array
                    s = tme.time()
```

```
                    ntime = good_flags.shape[0]
                    time_step = 5
                    for tm in six.moves.range(0, ntime, time_step):
                        time_slice=slice(tm, tm + time_step)
                        flag_chunk = good_flags[time_slice].astype(
int)
                        tm_chunk = time_idx[time_slice]
                        el_chunk = el_idx[time_slice]
                        az_chunk = az_idx[time_slice]
                        master, counter = update_arrays(tm_chunk,
bl_idx, el_chunk, az_chunk, flag_chunk, master, counter)

                    print(tme.time() - s)
                    goodfiles.append(f[i])

                    if i%args.no_of_files==0:

                        ds = xr.Dataset({'master': (('time','
frequency','baseline','elevation','azimuth') , master),
                            'counter': (('time','frequency','
baseline','elevation','azimuth'), counter)},
                           {'time': np.arange(24),'frequency':
fullvis.freqs,'baseline':np.arange(2016),
                            'elevation':np.linspace(10,80,8),'
azimuth':np.arange(0,360,15)})

                        ds.to_zarr(args.zarr,'w')
                        np.save(args.good,goodfiles)
                        np.save(args.bad,badfiles)
                        print('File has been saved')


                else:
                    print(f[i],'selection has a problem')
                    badfiles.append(f[i])
                    pass


            else:
                print(f[i],'channel has a problem')
                badfiles.append(f[i])
                pass



        except Exception as e:
            print(e)
            continue
```

Listing C.1: *The KATHPRFI Framework Code*

Bibliography

Asad, K., Girard, J., de Villers, M., Lehmensiek, R., Ansah-Narh, T., Iheanetu, K., Smirnov, O., Santos, M., Jonas, J., de Villiers, D. et al. (2019), 'Primary beam effects of radio astronomy antennas-ii. modelling the meerkat l-band beam', *arXiv preprint arXiv:1904.07155* .

Booth, R. & Jonas, J. (2012), 'An overview of the meerkat project', *African Skies* **16**, 101.

Camilo, F., Scholz, P., Serylak, M., Buchner, S., Merryfield, M., Kaspi, V., Archibald, R., Bailes, M., Jameson, A., Van Straten, W. et al. (2018), 'Revival of the magnetar psr j1622–4950: Observations with meerkat, parkes, xmm-newton, swift, chandra, and nustar', *The Astrophysical Journal* **856**(2), 180.

Carter, R. A., Antón, A. I., Dagnino, A. & Williams, L. (2001), Evolving beyond requirements creep: A risk-based evolutionary prototyping model, *in* 'Proceedings Fifth IEEE International Symposium on Requirements Engineering', IEEE, pp. 94–101.

Cauffriez, L. (2017), Modelling of safety instrumented systems by using bernoulli trials: towards the notion of odds on for sis failures analysis, *in* 'Journal of Physics: Conference Series', Vol. 783, IOP Publishing, p. 012057.

Council, N. R. et al. (2007), *Handbook of Frequency Allocations and Spectrum Protection for Scientific Uses*, National Academies Press.

Ellingson, S. W. (2005), 'Introduction to special section on mitigation of radio frequency interference in radio astronomy', *Radio Science* **40**(5).

Fisher, J. (2004), 'Signal analysis and blanking experiments on dme interference', *NRAO Electronics Division, Tech. Rep* **313**.

Foley, A., Alberts, T., Armstrong, R., Barta, A., Bauermeister, E., Bester, H., Blose, S., Booth, R., Botha, D., Buchner, S. et al. (2016), 'Engineering and science highlights of the kat-7 radio telescope', *Monthly Notices of the Royal Astronomical Society* **460**(2), 1664–1679.

Ford, J. M. & Buch, K. D. (2014), Rfi mitigation techniques in radio astronomy, *in* '2014 IEEE Geoscience and Remote Sensing Symposium', IEEE, pp. 231–234.

Fridman, P. & Baan, W. (2001), 'Rfi mitigation methods in radio astronomy', *Astronomy & Astrophysics* **378**(1), 327–344.

Gupta, Y., Ajithkumar, B., Kale, H., Nayak, S., Sabhapathy, S., Sureshkumar, S., Swami, R., Chengalur, J., Ghosh, S., Ishwara-Chandra, C. et al. (2017), 'The upgraded gmrt: opening new windows on the radio universe', *Current Science* **113**(4), 707–714.

Hamidi, Z. S., Abidin, Z., Ibrahim, Z., Shariff, N. N. M., Ibrahim, U. F. S. U. & Umar, R. (2011), Preliminary analysis of investigation radio frequency interference (rfi) profile analysis at universiti teknologi mara, *in* 'Proceeding of the 2011 IEEE International Conference on Space Science and Communication (IconSpace)', IEEE, pp. 311–313.

Jansky, C. (1958), 'The discovery and identification by karl guthe jansky of electromagnetic radiation of extraterrestrial origin in the radio spectrum', *Proceedings of the IRE* **46**(1), 13–15.

Jonas, J. et al. (2018), The meerkat radio telescope, *in* 'MeerKAT Science: On the Pathway to the SKA', Vol. 277, SISSA Medialab, p. 001.

Krupp, E. C. (2003), *Echoes of the ancient skies: The astronomy of lost civilizations*, Courier Corporation.

Lee, J. & Park, M. (2012), 'An adaptive background subtraction method based on kernel density estimation', *Sensors* **12**(9), 12279–12300.

Liddle, A. (2015), *An introduction to modern cosmology*, John Wiley & Sons.

Mauch, T., Cotton, W. D., Condon, J. J., Matthews, A. M., Abbott, T. D., Adam, R. M., Aldera, M. A., Asad, K. M. B., Bauermeister, E. F., Bennett, T. G. H. & et al. (2020), 'The 1.28 ghz meerkat deep2 image', *The Astrophysical Journal* **888**(2), 61. **URL:** *http://dx.doi.org/10.3847/1538-4357/ab5d2d*

Miller, D. F. (1998), 'Basics of radio astronomy for the goldstone-apple valley radio telescope'.

Misra, P. & Enge, P. (2006), 'Global positioning system: signals, measurements and performance second edition', *Global Positioning System: Signals, Measurements And Performance Second Editions*, .

Nan, R., Li, D., Jin, C., Wang, Q., Zhu, L., Zhu, W., Zhang, H., Yue, Y. & Qian, L. (2011), 'The five-hundred-meter aperture spherical radio telescope (fast) project', *International Journal of Modern Physics D* **20**(06), 989–1024.

Nielbock, M. (2017), 'Navigation in the ancient mediterranean and beyond', *arXiv preprint arXiv:1708.07700* .

Offringa, A., De Bruyn, A., Biehl, M., Zaroubi, S., Bernardi, G. & Pandey, V. (2010), 'Post-correlation radio frequency interference classification methods', *Monthly Notices of the Royal Astronomical Society* **405**(1), 155–167.

Offringa, A., De Bruyn, A., Zaroubi, S., van Diepen, G., Martinez-Ruby, O., Labropoulos, P., Brentjens, M. A., Ciardi, B., Daiboo, S., Harker, G. et al. (2013), 'The lofar radio environment', *Astronomy & astrophysics* **549**, A11.

Pratap, P. & McIntosh, G. (2005), 'Measurement of the radiation from thermal and nonthermal radio sources', *American journal of physics* **73**(5), 399–404.

Ross, S. M. (2014), *Introduction to probability models*, Academic press.

Sullivan III, W. T. (1984), 'Karl jansky and the discovery of extraterrestrial radio waves', *The Early Years of Radio Astronomy: Reflections fifty years after Jansky's discovery* pp. 4–12.

Taylor, G. B., Carilli, C. L. & Perley, R. A. (1999), Synthesis imaging in radio astronomy ii, *in* 'Synthesis Imaging in Radio Astronomy II', Vol. 180.

Thompson, A. R., Moran, J. M. & Swenson, G. W. (2017), *Analysis of the Interferometer Response*, Springer International Publishing, Cham.
URL: *https://doi.org/10.1007/978-3-319-44431-4*$_3$

Thompson, A. R., Moran, J. M., Swenson, G. W. et al. (1986), *Interferometry and synthesis in radio astronomy*, Wiley New York et al.

Walker, C. (1996), Astronomy before the telescope., *in* 'Astronomy Before the Telescope'.

Wiaux, Y., Jacques, L., Puy, G., Scaife, A. M. & Vandergheynst, P. (2009), 'Compressed sensing imaging techniques for radio interferometry', *Monthly Notices of the Royal Astronomical Society* **395**(3), 1733–1742.