STATISTICAL ANALYSIS OF POWER-LINE RADIO FREQUENCY INTERFERENCE AT GMRT

Student Project

Submitted by

Avantika Iyengar

B.E. Electronics and Telecommunication Engineering
Marathwada Mitramandal College of Engineering, Pune

GIANT METREWAVE RADIO TELESCOPE, NCRA - TIFR, KHODAD

May 2022 - August 2022

CERTIFICATE

This is to certify that this report, titled **'Statistical Analysis of Power-line Radio Frequency Interference at GMRT'** authored by **Avantika Iyengar**, is a bonafide work carried out by the author under our supervision, at the Giant Metrewave Radio Telescope.

Mr. Kaushal D. Buch
Engineer E
Giant Metrewave Radio Telescope
(NCRA-TIFR)
Khodad, Maharashtra, India

Dr. Divya Oberoi
Associate Professor
National Centre for Radio Astrophysics
(NCRA-TIFR)
Pune, Maharashtra, India

Abstract

This project particularly aims to study the statistical properties of power-line Radio Frequency Interference (RFI) occurring due to various sources at GMRT. The agro-industrial manufacturing units surrounding GMRT require high power voltage which is transmitted by the high-tension power lines and cause broadband RFI which is picked up by the sensitive antennas of GMRT. It corrupts the original signal by degrading the signal-to-noise ratio coming from an astronomical source and is a major concern for astronomers.

To detect, classify and mitigate this power-line RFI, statistical data analysis of its time series is done using GMRT data. This data is collected in the form of digitized samples, at the output of the analogue to digital converter of the GMRT wideband backend system. It is processed in the time domain using MATLAB, before giving it as an input to the correlator block of the GMRT wideband backend system. The LeCroy oscilloscope at GMRT is used to visualise and select different types of bands and antenna polarizations. The statistical properties such as kurtosis, skewness, and probability distribution functions are studied and understood. Based on this understanding, the RFI density, its typical duration of on and off, the inter-arrival time of sparks inside a bunch, and the inter-arrival time of the RFI bunch in the entire time series are also estimated.

This project forms a part of a bigger project which aims to use the latest technologies of Machine Learning and Artificial Intelligence, to automatically detect and mitigate broadband RFI. Keeping this aim in the mind, various RFI sparks are identified and modelled using different curve fitting techniques. The RFI bunch which occurs quite frequently in the series is automatically identified using kurtosis. To make this automatic bunch detection algorithm more robust, an additional method of modelling the signal envelope is also proposed. This project provides a foundational study about recent trends in machine learning in radio astronomy, which is summarised in the end for better implementation of these algorithms in future.

Acknowledgement

I would like to express my sincere gratitude to my project guides Mr. Kaushal Buch and Dr. Divya Oberoi, for their admirable guidance and constant encouragement throughout this project. Though I was unfamiliar with some of the concepts required for this project, Mr. Kaushal Buch took me onboard and made sure to bridge the gap with all the necessary resources. The thought-provoking technical discussions held with him during tea time were enough to screw my head and keep me engaged in the work. His small gestures of appreciation throughout the project duration motivated me to accelerate the pace of my learning curve. Apart from work, his jovial nature made my days at GMRT full of cheer.
I would also like to thank Dr. Divya Oberoi for his valuable comments and suggestions on the project. His critique remarks always helped me revisit and ponder over my problem-solving approach. He helped me identify some of the loopholes in my solution, and suggested making it more generalised in various aspects.

I also extend my gratitude to Mrs. Nutan Deshmukh, for providing me with the opportunity to work at GMRT. Her cooperation and support helped me to pursue this project which was in my interest area. My overall experience at GMRT was highly memorable because of my mates, Srishti, Shubham and Sohan. They were always on their toes to help me solve any technical problem and made the working environment alive and youthful. Every other technical and non-technical staff member at GMRT had been very supportive throughout the project. I had some of my best learnings from every lab member at GMRT and I am very grateful for it.

Avantika Iyengar

Table of Contents

Chapter 1: Introduction

1.1 Overview

The Giant Metrewave Radio Telescope (GMRT) is a state-of-the-art and indigenously designed telescope, located around 80km North of Pune, Maharashtra India. The GMRT comprises an array of 30 fully steerable parabolic dishes, each with a diameter of 45m and spread across distances of up to 25km. This telescope operates in the meter wavelength part of the radio spectrum, to minimise the effects of man-made radio frequency interference occurring from the towns surrounding the telescope site. [1]

The radio telescopes operate in the radio band of the electromagnetic spectrum, in the frequency range of 3 kHz to 3000 GHz. The wavelength of radio waves ranges from 1mm to 100 km, and thus they can be transmitted over long distances. All the modern telecommunication methods rely extensively on radio frequency bands and they are a valuable resource to any organisation. The 4G and 5G mobile standards in communication operate in the frequency bands from 800 MHz – 2000 MHz which have recently been auctioned at gigantic costs.

The sensitivity of radio telescopes is very high compared to any radio communication receiver. This is due to the fact that radio telescopes are designed to observe some of the faintest celestial objects, emitting low-frequency radio waves. The signal strength received at a radio telescope is specified in the units of Jansky (Jy). It is the unit of power per unit area per unit bandwidth. From the sensitivity analysis described by Swarup, it is observed that a radio telescope is approximately 60db more sensitive to radio signals than a typical radio communication receiver [2].

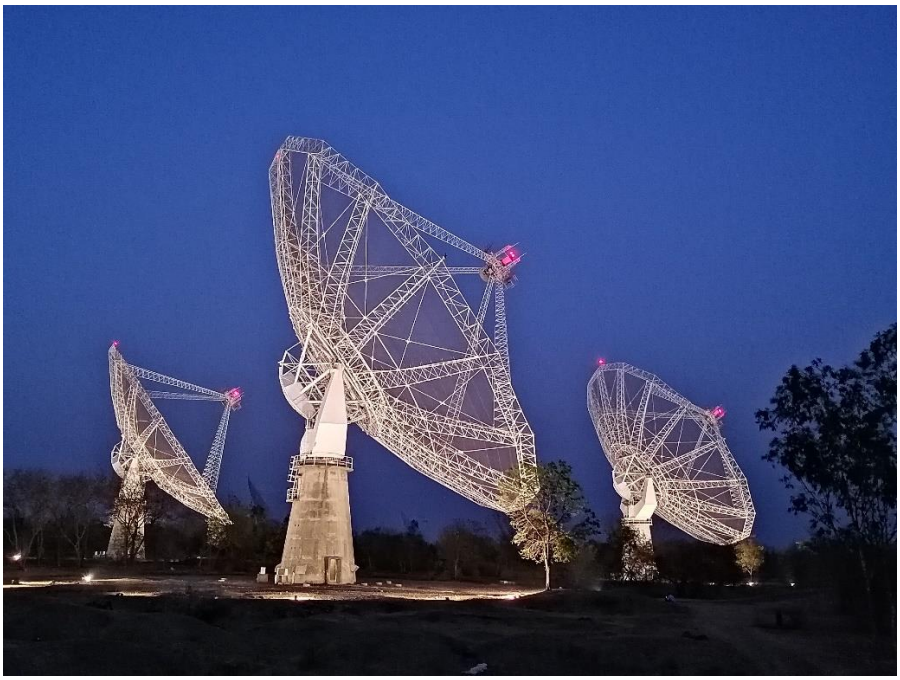


Figure 1: One of the antennas of GMRT

1.2 Radio Frequency Interference (RFI)

Since a radio telescope is very sensitive to radio signals, it needs to be located away from densely populated areas, and one of the major challenges in setting up a radio observatory is the identification of the Radio Quiet Zone. A Radio Quiet Zone (RQZ) should be far from heavily populated areas like cities and towns and no mobile or radio transmitter should be allowed in that area. GMRT is located in such a quiet radio environment, near the remote village of Khodad.

Although the radio telescopes like GMRT are located in RQZ, they experience unwanted interference from man-made radio emissions. In addition to that due to high tension power lines, this interference is mainly due to telecommunication operators operating in the same frequency range, as well as many man-made devices like television, automobiles, and irrigation pumps. This is known as Radio Frequency Interference (RFI). RFI is a subset of electromagnetic interference which occurs from other electronic components.

Any radio signal observed by the telescope is broadly made up of three components, explained with the following equation:

Signal = Source power + System power + RFI

The source power is received from the astronomical source and has very low intensity. The system power is generated by the internal electronic circuits installed at the telescope site. The intensity of the signal is quite high and it is Gaussian distributed in nature. The RFI is generated from many man-made devices. This is the unwanted part of the signal and it overwhelms the power coming from a weak astronomical source.
Radio Frequency Interference is broadly classified into two types: Narrowband RFI and Broadband RFI.

1.2.1 Narrowband RFI

These types of RFI emissions occur in a narrow region of the spectrum. This type of signal is often consistent over a large time and usually does not cause any permanent damage to the system. Narrowband RFI often occurs due to the crossing of frequencies from mobile towers and other communication devices.

1.2.2 Broadband RFI

This type of RFI is impulsive and contains high energy signals, enough to wash out the underlying astronomical signal. This occurs in an impulsive form, in a short period of time. Broadband RFI signals may cause permanent damage to the electronic receiver systems due to their high intensity.

1.3 RFI at GMRT

Right from the commission of the GMRT in 2001, this telescope provided a frequency coverage from 38MHz – 1420 MHz. In order to keep up with the other radio observatories in the world and to incorporate better technology, the GMRT project was upgraded in 2020[3]. With the upgradation of GMRT, it provides an excellent facility for frontline radio astronomical research in the frequency range of 120 – 1500 MHz, operating in the meter wavelengths.

The frequency range covered by uGMRT is divided into four sub-bands;

Number	Frequency range
Band 2	120 – 250 MHz
Band 3	250 – 500 MHz
Band 4	550 – 850 MHz
Band 5	1050 – 1450 MHz

GMRT is a versatile instrument carrying out observations in the varied subfields of Astronomy. From the past two decades, GMRT is being used in the discoveries of numerous galaxies, study of atomic hydrogen, study of Fast Radio Bursts, galaxy mergers etc. One of the science goals as envisioned by Swarup et al. [4] of GMRT, is to observe and search for the short period pulsars. Pulsars are highly magnetised neutron stars which emit a beam of strong electromagnetic radiation at specific time intervals. Some of the short-period pulsars have periods smaller than ~10ms and emit radiation in the low-frequency range. The low-frequency bands provided by uGMRT are ideal for studying pulsars, as observed recently by Bhattacharyya et al. [5]. However, the low-frequency regions of the radio spectrum at GMRT are severely affected by broadband RFI. Observations recorded by engineers in 2008 indicate the gravity of RFI observed at the GMRT. With the effect of modernization, many industries are coming up in the vicinity of GMRT. These industries require high voltage supply and modern switching electronic machines which generate RFI. The high voltage power lines which supply power to the neighbouring industries and the subsystems of GMRT generate a significant amount of broadband RFI.

1.4 Power-line RFI

Power-line RFI is a type of broadband RFI which occurs in the form of high intensity and short duration pulses in the received data signal. This type of RFI is generally produced due to the gap between the connectors of power cables and the capacitive action of the conductors [6]. Some of the unique properties of power-line RFI are discussed in the section 1.4.1 of this report. Figure [2] shows the corona discharge at an HV electric pole near GMRT.

Band 2 (B2) of the GMRT experiences a lot of power-line RFI, with an intensity of 10 – 20 dB more than the system's overall operating temperature. It is also known that the power flux density of power-line RFI is inversely proportional to the frequency. It decreases with increasing frequency as f^{-2}. The recordings of power-line RFI observed at GMRT by Swarup et al. [2] depict some of the adverse effects of HV transmission lines on the observing bands of GMRT. It also proposes some methods to mitigate this RFI and co-exist in the surrounding radio environment. Fig. [3] visualises the observed RFI in the actual data captured in Band 2 at GMRT.

Figure 2: Corona discharge near GMRT

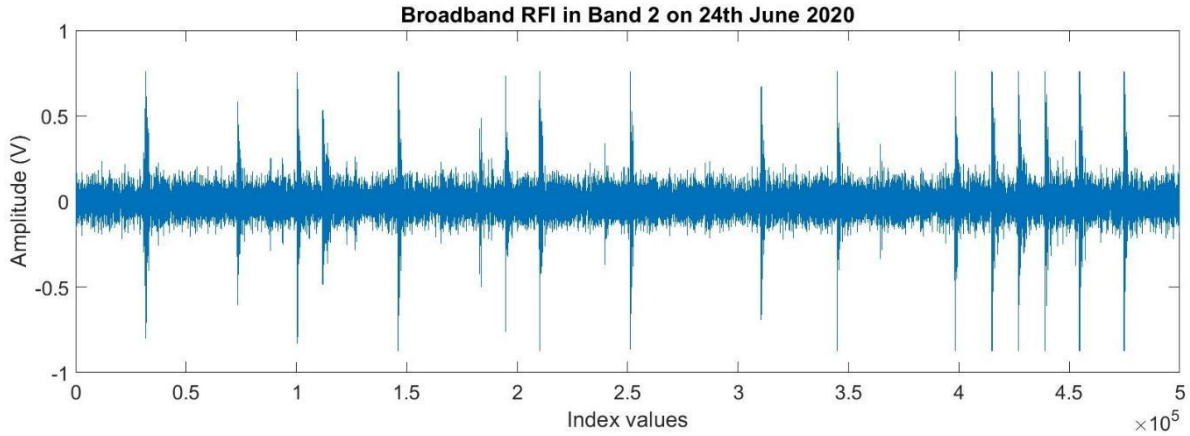


Figure 3: Power-line RFI observed in B2 at GMRT at a sampling rate of 5 ns

1.4.1 Properties of power-line RFI

As explained briefly in the above section, power-line RFI is of two types; gap discharge and corona discharge. The gap discharge is a result of electromagnetic radiation from the gap present between the two conductors of the HV line. The small gap produces an arc of electrons between the two conductors and radiates high power noise in the surroundings. Gap discharge can also occur due to the capacitive action between two conducting elements of the power-line.

Corona discharge is a form of RFI generated due to the ionization of surrounding air molecules around the conductor. The capacitive current generated at these conductors gives rise to some I^2 R power losses. In presence of water vapour, the air surrounding the conductors is ionised quickly and gives rise to corona discharge. This type of discharge generates gases like ozone O_3 and various nitrogen oxides (NO_X) [6]. These gaseous emissions are responsible for the corrosion of various man-made structures and are harmful to human beings. From the above

references, it is also understood that power-line RFI is more intense during the monsoon season and has less intensity during the dry seasons.

It is inferred from the research and experiments carried out by Loftness [7] that power-line RFI does not have any special characteristic sound associated with it. It also further debunks the myth that power-line RFI results from the harmonics of 60 Hz frequency (as per U.S standards).

1.5 Real-time filtering of RFI at GMRT

As the GMRT is now upgraded to incorporate a large frequency range, it is more susceptible to broadband and narrowband RFI. The existing algorithm deployed at GMRT mitigates RFI in real time using some robust statistical methods. The algorithm designed by Buch et al. [8] uses a Median Absolute Deviation (MAD) threshold-based algorithm to identify RFI instances in real-time. This system is implemented on the FPGA and CPU-GPU platforms and filters broadband and narrowband RFI by replacing RFI instances with Gaussian noise.
The Median Absolute Deviation is a robust measure of the variability of a univariate sample of quantitative data. It can be computed using the relation given by the following equation.

For a given dataset $X_1, X_2, X_3 \ldots X_n$

$$MAD = median(|X_i - median(X)|)$$

The thresholds based on MAD are calculated using the empirical rule of statistics, which incorporates 99% of the data points within the 3 units of standard deviation from the mean of a Gaussian distribution. The formula for calculating the threshold is given by

$$Upper\ threshold\ Tup = 3 + 1.4826 * MAD(data)$$

$$Lower\ threshold\ Tlo = 3 - 1.4826 * MAD(data)$$

The existing method of real-time filtering of RFI at GMRT has increased the signal to noise ratio and improved the observation results typically for low frequency bands. This method has two default settings of RFI filtering, one with a 4σ threshold for continuum observations and the other with a 3σ threshold for pulsar observations. For the current system of RFI filtering, the user has the choice to replace the detected RFI instances either with a constant value or with noise.

1.6 Motivation of the project

After discussing some of the prerequisites required to understand this project, this particular section describes the motivation and scope behind the statistical analysis RFI at GMRT. The main objective of this project is to study the statistical properties of power-line radio frequency interference occurring at GMRT. The statistical analysis reveals some of the hidden properties of the real signals.

The current method of RFI filtering classifies every data point outside the predefined threshold as a possible RFI candidate. To reduce these false detections, it is necessary to implement a smarter technique of RFI detection and mitigation. This can be achieved by using some of the latest technologies from machine learning, to automatically detect and mitigate RFI in the time domain.

However, to implement machine learning for RFI detection, it is necessary to analyse the signal based on statistical parameters and design a smart algorithm to classify noise and RFI.
The aim of this project is to initiate the automatic RFI detection process by understanding power-line RFI statistics and thereby providing a foundation for future implementation of machine learning techniques.

Chapter 2: Statistical measures and signal understanding

Statistics is an important branch of mathematics used to define, interpret, investigate and infer any practical problem. The statistical analysis helps to decide and take necessary actions regarding the solution to the problem. Some of the important statistical measures are explained in this chapter.
Statistics is broadly divided into two classes; descriptive statistics and inferential statistics. Descriptive statistics relies mainly on the representation of the data in the form of tables, charts and graphs. Various conclusions are derived from the representation of the data.
Inferential Statistics on the other hand, provides a deeper insight to the nature and distribution of the data. It helps a user to interpret and analyse the data better, and reach more precise conclusions.
This project focuses on the use of inferential statistics in order to identify and mitigate RFI using machine learning.

2.1 Statistical moments

In addition to the measure of central tendency (Mean, Median and Mode), statistical moments are used to describe the characteristics of the distribution of the data. For any given data, the statistical moments are its expected values raised to some power.
Statistical moments are analogous to moments described in physics. In physics, moments refer to mass and describe the physical properties of the given object. Once the mass of an object is known, it is easy to compute other physical parameters like the inertia, force, acceleration etc which describe the overall behaviour of the object in the environment.
The statistical moments are very similar to physical moments, and reveal the properties of the data and its distribution. The four major statistical moments are –

i) Expectation or the Mean
ii) Variance
iii) Skewness
iv) Kurtosis

i) The Expectation (M_1):

The expectation of the given dataset is the generalisation average value of all the elements in the dataset. Intuitively, it is equivalent to the mean if the probability of occurrence of all elements in the dataset is equal.
The expectation value of a continuous random variable is given by the equation:

$$E(g(x)) = \int_{-\infty}^{\infty} P(x).g(x).dx$$

Where P(x) is the probability density function of g(x)
ii) Variance (M_2):

Variance is the measure which describes how the elements are spread from the mean value of the dataset. Mathematically, it is the square of standard deviation and gives the average squared distance from the mean of the dataset. It is usually represented by σ^2. A higher value of variance indicates that the samples from the data are widely spread across its mean value. This can be visualised by comparing the probability distribution function (PDF) of a Gaussian curve and

the probability distribution function of highly varied dataset. Figure [4] illustrates the PDFs of Gaussian random distributions with different variances.

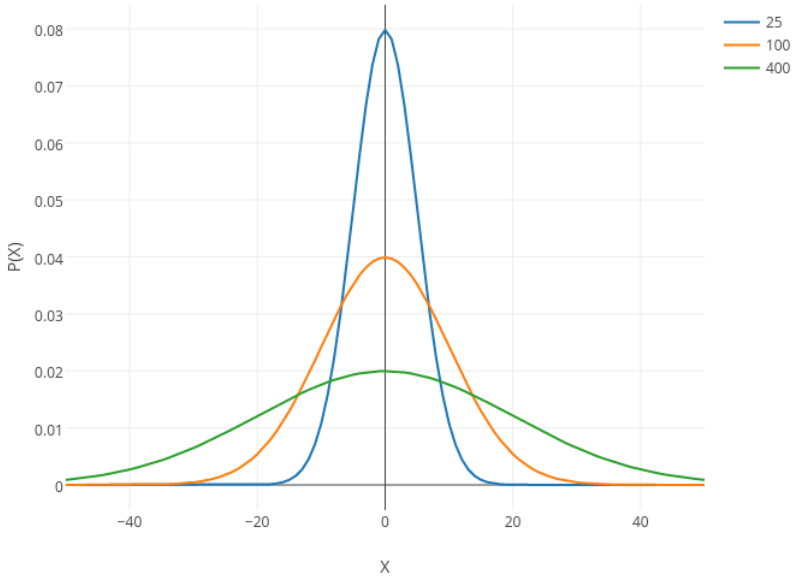


Figure 4: High variance v/s low variance courtesy: scribbr

The variance is mathematically given by the equation:

$$Var(X) = E(X - \mu)^2 = \frac{\sum_{i=1}^{n}(x - \mu)^2}{n}$$

iii) Skewness (M_3)

Skewness is the statistical moment which measures the asymmetricity of the data from the mean. A normal distribution has zero skewness. A distribution is said to have zero skewness if the mean, median and mode have equal values. This type of distribution is known as symmetric distribution. For asymmetric distribution, the values of mean, median and mode are not equal and thus it is further classified as either positively skewed or negatively skewed distribution. The positive skewed distribution has a value greater than 0 and its mean value is the highest among all three measures of central tendency. Vice versa, negative skewed distribution has a value less than 0 and its mode is greatest among all the three central tendencies. Figure [5] visualises positive, negative and zero skewed distributions.

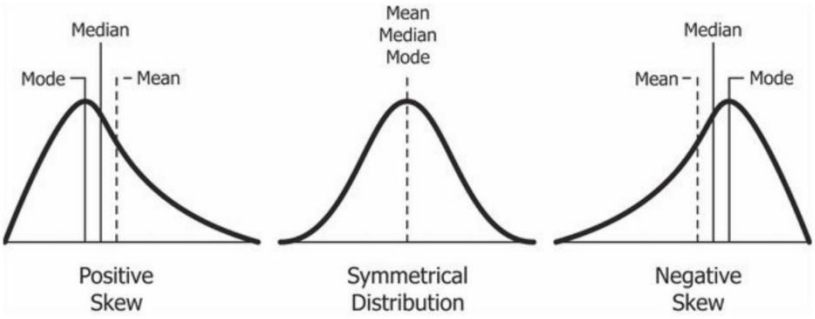


Figure 5: Skewness courtesy: Wikipedia

iv) Kurtosis (M_4):

Kurtosis is the fourth and an important statistical moment, used to characterise a given distribution based on its peakedness. In other words, Kurtosis can also be defined as the measure of tailedness of the distribution. Heavy tailed data has greater or positive value of kurtosis and it indicates the presence of outliers in the dataset. Light tailed data has negative kurtosis and has a blunt peak of its probability density function. The kurtosis of pure normal Gaussian distribution is equal to 3.
The mathematical expression for kurtosis is given by equation:

$$M_4 = K[X] = \sum_{i=1}^{n} \frac{(x - \mu)^4}{\sigma^4}$$

Figure [6] explains the kurtosis variation for different types of curves and fig. [7] illustrates the heavy tailed and light tailed distribution.

	Category		
	Mesokurtic	**Platykurtic**	**Leptokurtic**
Tailedness	Medium-tailed	Thin-tailed	Fat-tailed
Outlier frequency	Medium	Low	High
Kurtosis	Moderate (3)	Low (< 3)	High (> 3)
Excess kurtosis	0	Negative	Positive
Example distribution	Normal	Uniform	Laplace

Figure 6: Kurtosis of different curves courtesy: scribbr.com

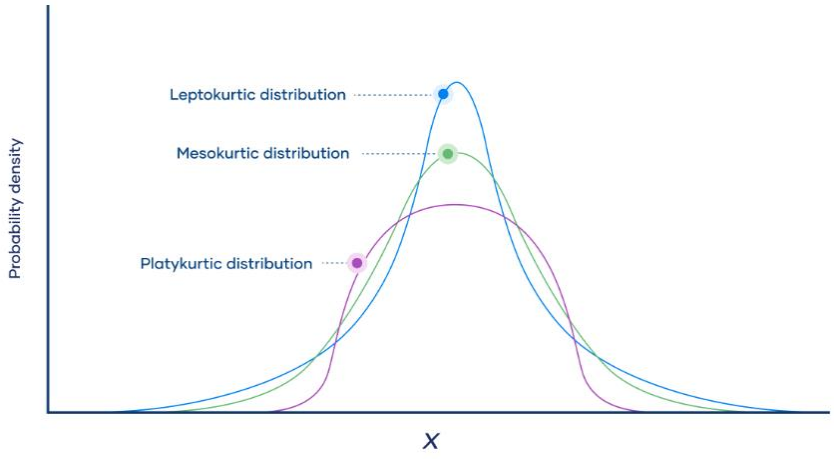


Figure 7: Heavy tailed and light tailed distribution; courtesy: scribbr.com

2.2 Understanding the signal at GMRT

The raw signals received at the antenna output are time domain signals sampled at 5ns sampling rate. The signals used for statistical analysis of broadband RFI are recorded at the output of ADC channel, before sending them as an input to correlator system of GMRT. The output of ADC channel is sampled at 200 x 10^6 Hz sampling frequency and is converted into instantaneous voltages. These signals are observed and recorded from the LeCroy oscilloscope present at GMRT in the form of text files. The text files are further processed in MATLAB to visualise the recorded data. Figure [8] shows the block diagram of the system and the location from where the data is tapped for further analysis.

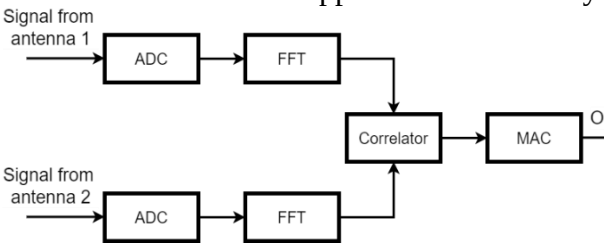


Figure 8: Block diagram of the system

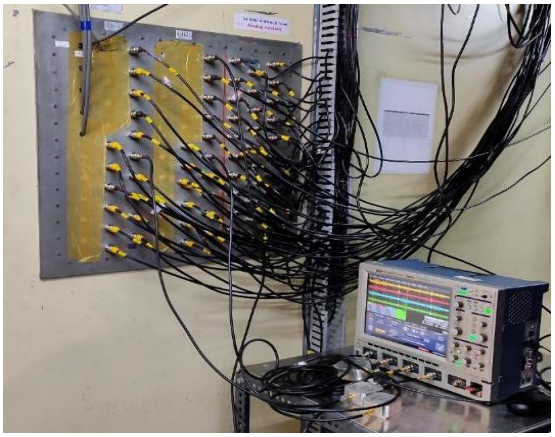


Figure 9: LeCroy Oscilloscope at GMRT

The data visualised from GMRT is represented by the following figures. Datasets recorded in different seasons are visualised in MATLAB in order to study the pattern and occurrence of power-line RFI.

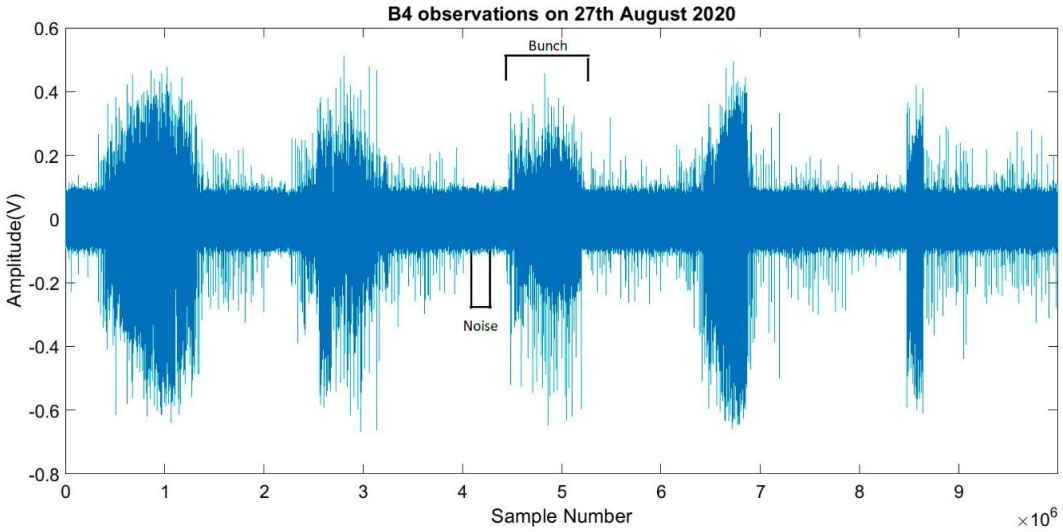


Figure 10: Signal in Band 4

A bunch of broadband power-line RFI in the recorded observations is defined as the high intensity signal lasting for a few milliseconds in the entire duration of the dataset. The noise sample can be identified visually as it has amplitude larger than the average amplitude of a desired signal. Anything which is not impulsive in nature and has Gaussian distribution, is identified as noise in the dataset. Impulsive broadband RFI has a non-normal distribution.

Chapter 3: Time Series Analysis of Power-line RFI

The time series analysis of power-line RFI is carried out with MATLAB. A noise sample is generally defined as the subset of an entire time series which has a normal distribution of values of the signal. It is also termed as a healthy data sample. A bunch majorly consists of outliers and impulsive bursts of voltage. A bunch can also be termed as the corrupted signal, as it dominates the underlying signal from an astronomical source.
A primary analysis is done using some noise and bunch samples, in order to study the RFI pattern, distribution and its frequency of occurrence in the overall recorded duration.
The bunch and noise samples are identified by actually zooming in to the time series and noting the sample number range from the dataset. This is illustrated in fig. [11] and [12].

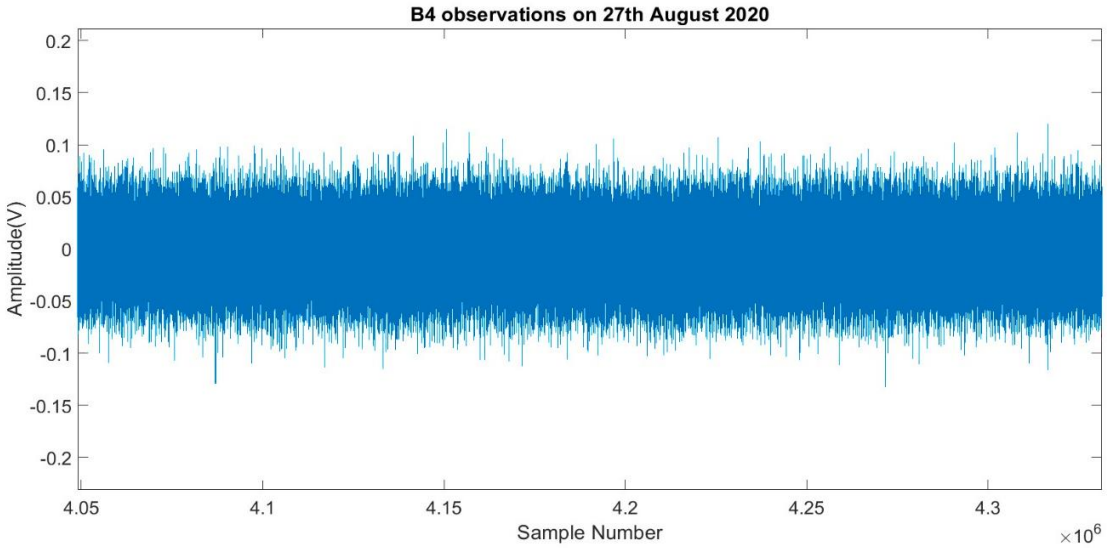


Figure 11: Noise Sample

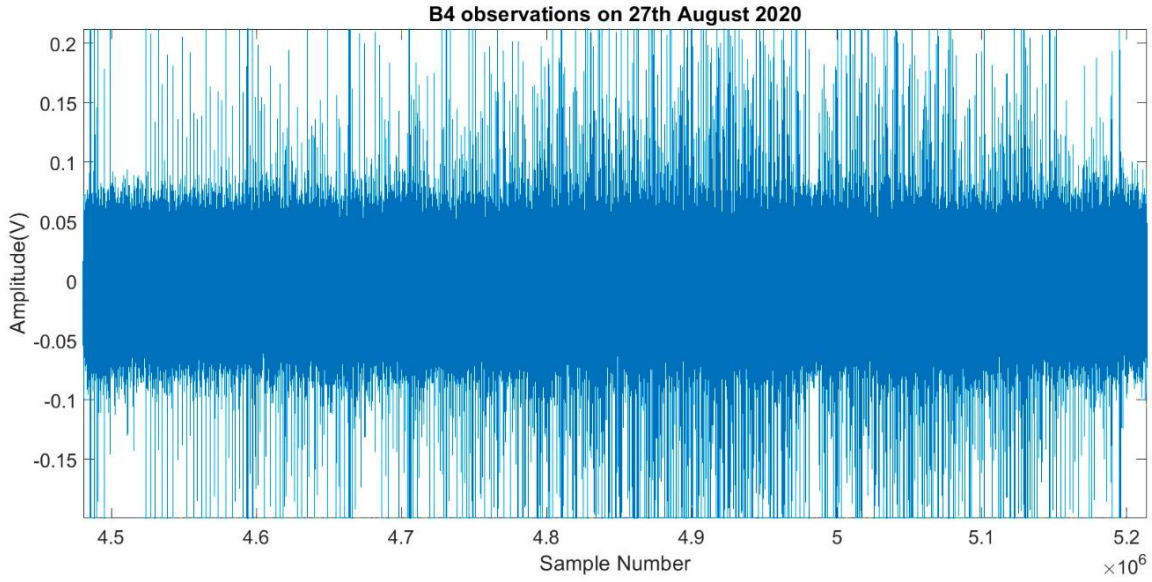
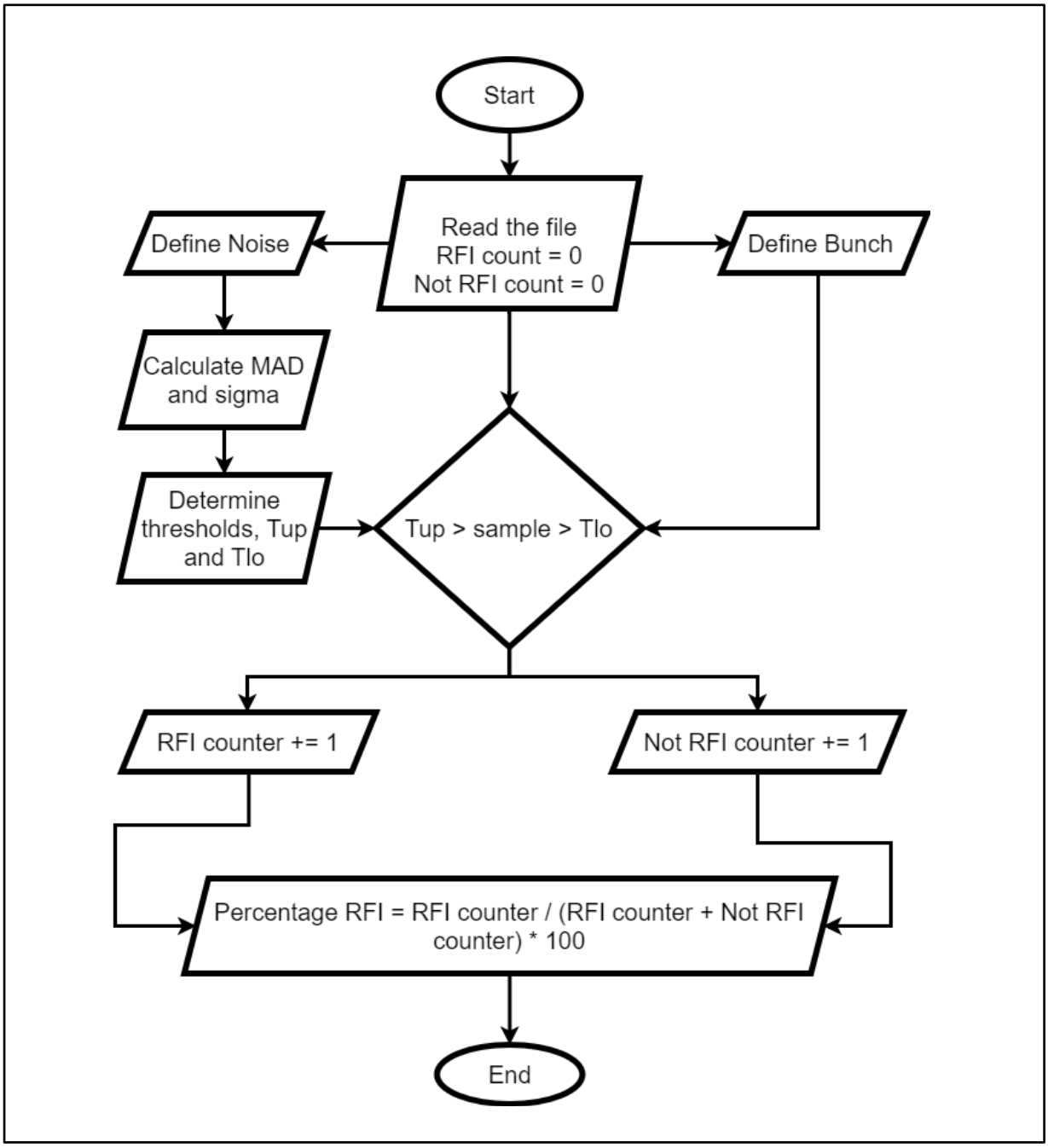
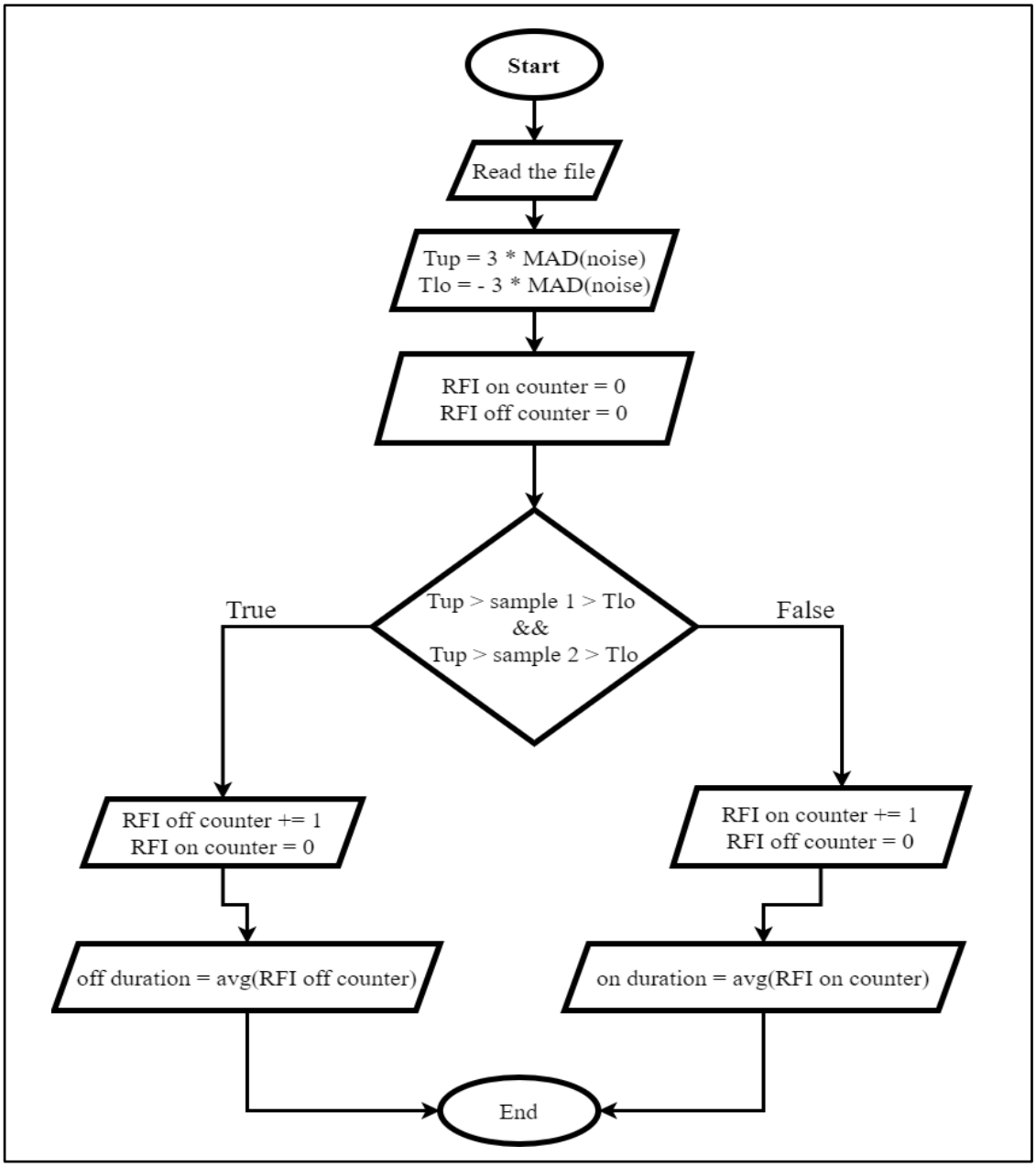


Figure 12: Bunch sample

The method explained above is a long and tedious process and impractical for the analysis of significant quantities of data. However, in the initial phase, we identified the noise and bunch samples visually for some datasets.
The algorithm implemented on MATLAB makes the use of median absolute deviation to determine a 3σ threshold. These thresholds were used to count the number of outliers present in the dataset and thereby compute the density of RFI in the data. It is visualised in parts by the flowcharts [1] and [2] given below.

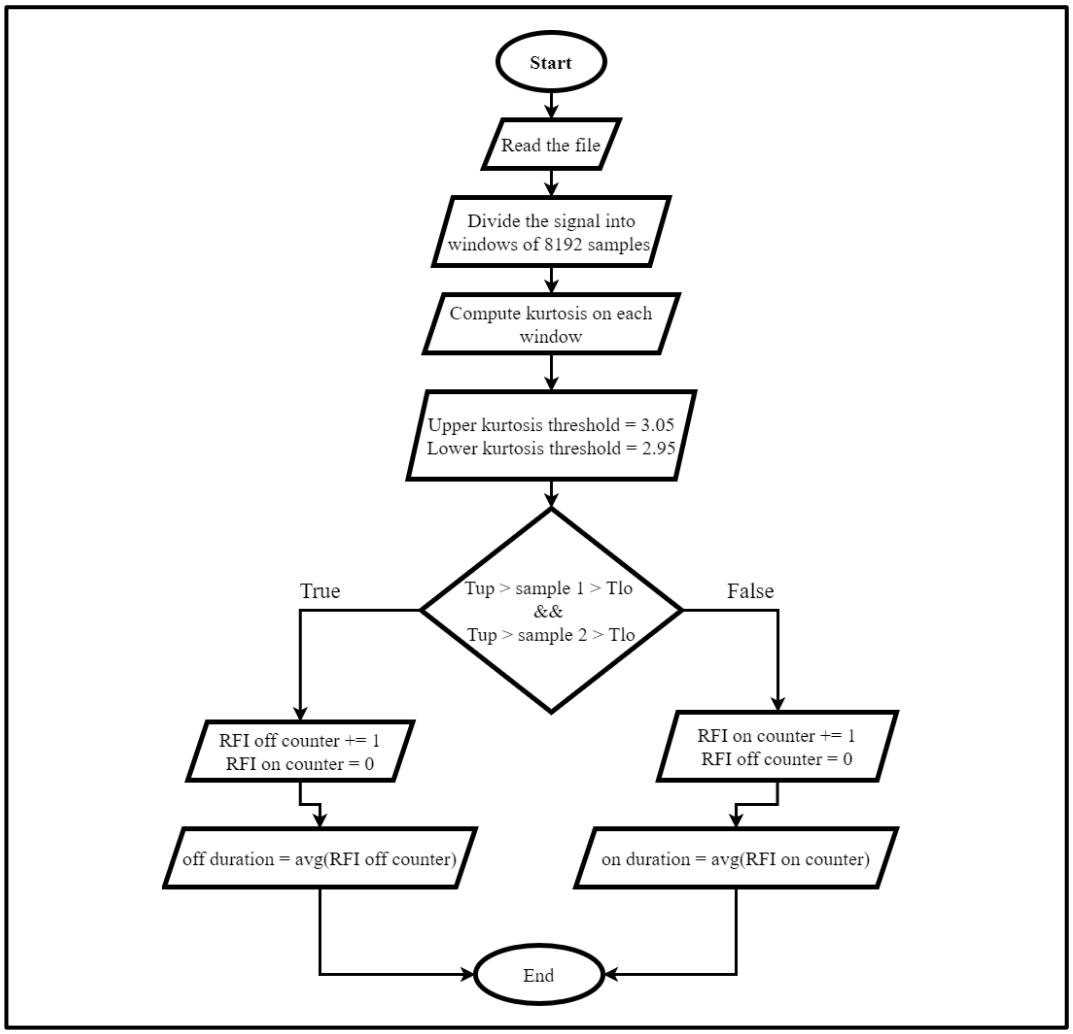


Flowchart 1: RFI Density



Flowchart 2: RFI duration

Above flowcharts give a crisp description of the fundamental algorithms used in this project. The RFI density is an important factor, which unmasks the information of the percentage of RFI dominated data present in the dataset. The duration of RFI is also observed in different bands, in datasets recorded by different antennas, and in different polarizations.
The inter-arrival time of bunches in a signal is computed to get an insight into the frequency of occurrence of power-line RFI. The inter-arrival time is expressed in seconds, and basically represents the duration for which RFI is absent. This is estimated using kurtosis, on a window size of 8192 samples. The flowchart [3] explains the algorithm behind the computation of inter-arrival time.



Flowchart 3: Inter arrival time of bunch

It is observed by the implementation of this algorithm on numerous datasets that although the input dataset seems to be heavily corrupted, the actual RFI density in the signal is anywhere in between 0.5 – 2%**.** It is rational to infer that power-line RFI occurring at GMRT corrupts up to 2% of the data. The RFI density inside the heavily corrupted bunch was also found to be around 1% of the total number of samples present in the bunch. This gives an important insight about the quantitative measure of RFI in the data.

RFI on and off durations are observed in the range of 10^{-2} – 9 x 10^{-6} seconds. This characteristic is used to study the impulsive nature of RFI occurrence. From the observations carried out on different datasets, it is clear that the average duration for which RFI lasts is much smaller than its off duration. This observation supports the fact that power-line RFI is broadband RFI, and has an impulsive nature.

The inter-arrival time of bunches was computed with the motivation to study the frequency of RFI bunch. RFI bunches are more frequent in band 3 and band 4, whereas band 2 is heavily corrupted with high intensity sparks of RFI. These sparks are randomly spread in band 2 and form some sophisticated bunching in band 3 and band 4 of GMRT. Band 5 is the highest frequency band observed by GMRT. This band does not experience much of the power-line RFI and thus the inter-arrival time of bunches in this particular band is more.

In addition to these parameters, individual sparks are identified from the RFI bunch using a standard deviation-based threshold. The threshold used for classifying a sample point as a spark is 5σ. The spark, along with some noise samples neighbouring the spark is modelled to understand the decay pattern of the spark. The spark decays by following a Fourier curve of order 4 with reasonable goodness of fit statistics. The inter arrival time of sparks is also measured to understand the frequency of arrival of spark. Figure [13] shows the identified sparks in the dataset and fig. [14] shows the roll off curve fitted on a single spark.

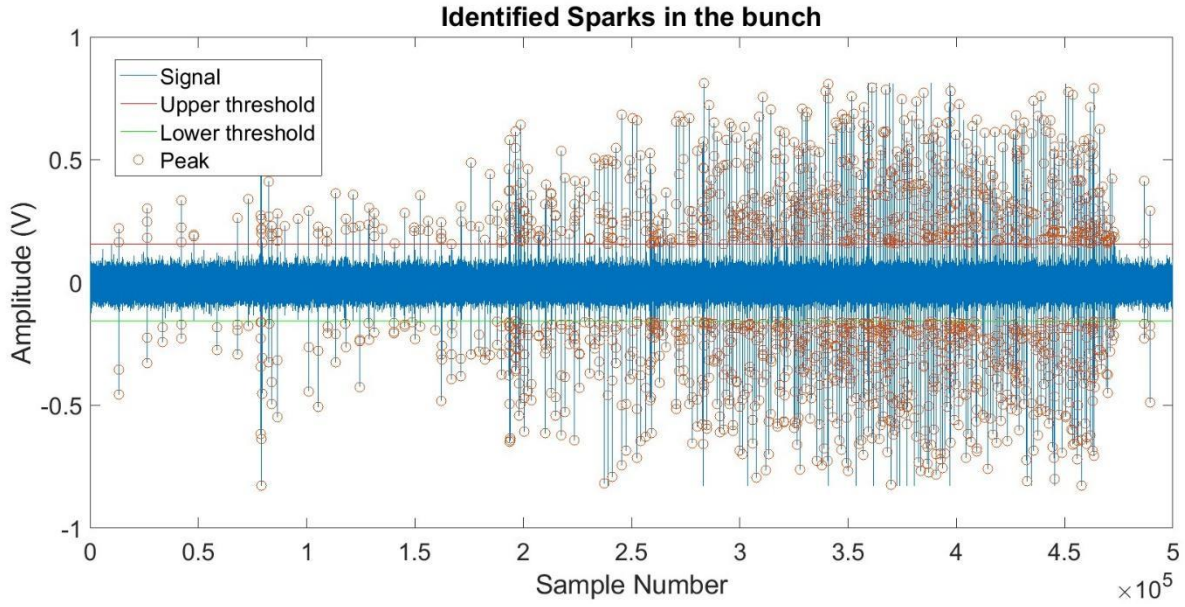


Figure 13: Identified sparks

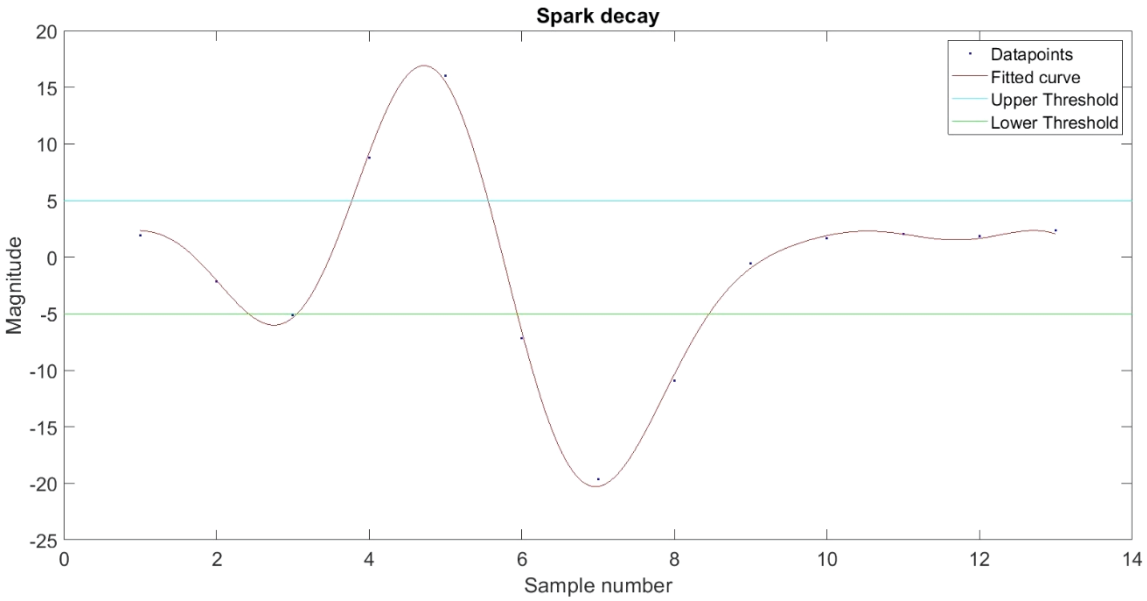


Figure 14: Spark curve

The coefficient of cross correlation of two signals coming from different antennas is also a distinguishing factor between RFI and noise. The cross-correlation coefficient increases if both signals are corrupted by the same RFI source. This correlation is in the regard of the terrestrial radio frequency source only, and not for the radio signal coming from the celestial object. Figure [15] illustrates the correlation of two time series.

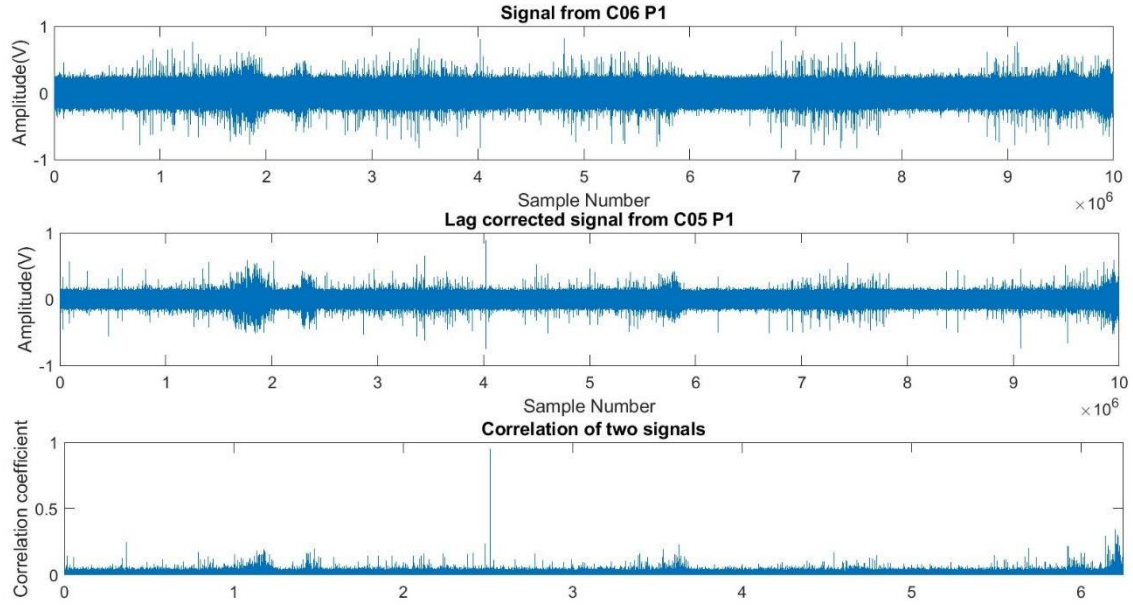


Figure 15: Correlation of time series

Chapter 4: Non-Normality Detection

The statistical analysis done in the previous chapter uses noise and bunch samples which were identified by the user manually from the given dataset. The above method is practical in case of analysis of one or two datasets. However, for a large number of datasets, this solution is impractical, as the user will have to manually update the noise and bunch samples for every dataset. Thus, it is a tedious and time-consuming process, involving a lot of human labour and resources.

It is a well-known fact that radio frequency interference follows a non-Gaussian curve, and the properties of RFI are different from the properties of a normal distribution. On the other hand, the noise follows a Gaussian distribution. These two distinct properties of noise and RFI open up a new window to explore non normality detection in the time series. With this idea, a set of samples which do not follow a normal distribution curve are detected and flagged as RFI. This detected RFI can be further mitigated by using some of the existing RFI mitigation algorithms at GMRT.

Figure [16] and fig. [17] represent the boxplots of noise and RFI from GMRT data. The central line of the boxplot indicates the mean value of distribution while the red markers indicate the outliers. The upper boundary of the box indicates the 75^{th} percentile of distribution and lower boundary indicates the 25^{th} percentile of given distribution. From the images below, it is verified that RFI has more outliers and does not follow a Gaussian distribution. As the number of outliers increases, the data is heavily distributed in the tailed regions of the PDF curve. Thus, these boxplots help to visualise the PDFs of noise and RFI.

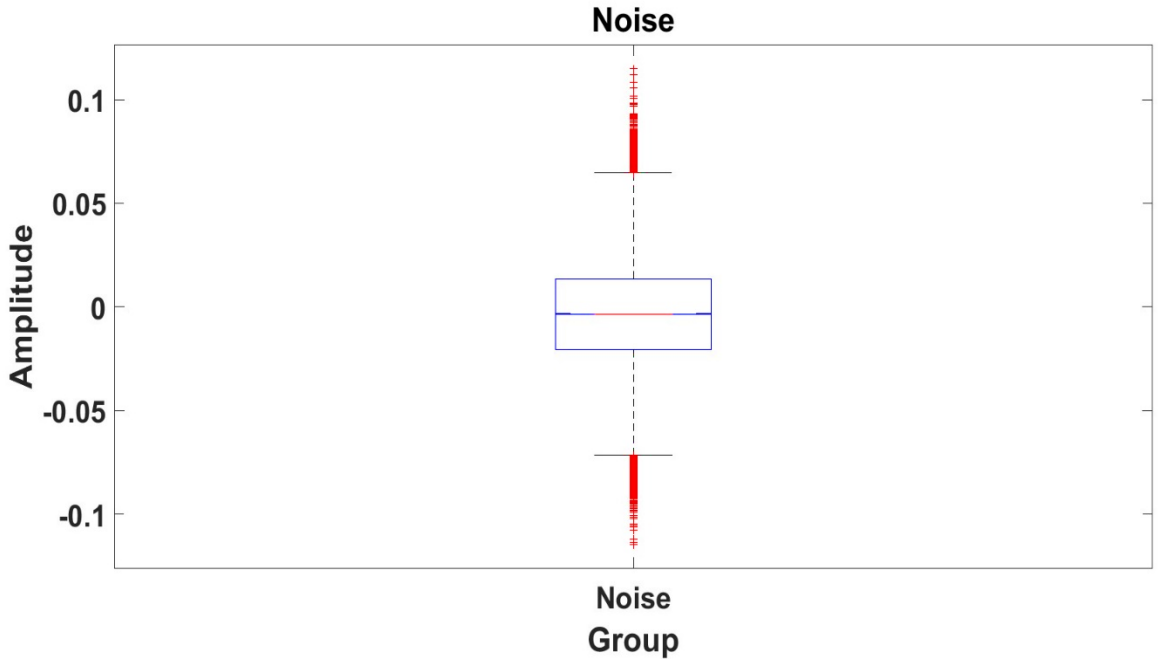


Figure 16: Boxplot of noise

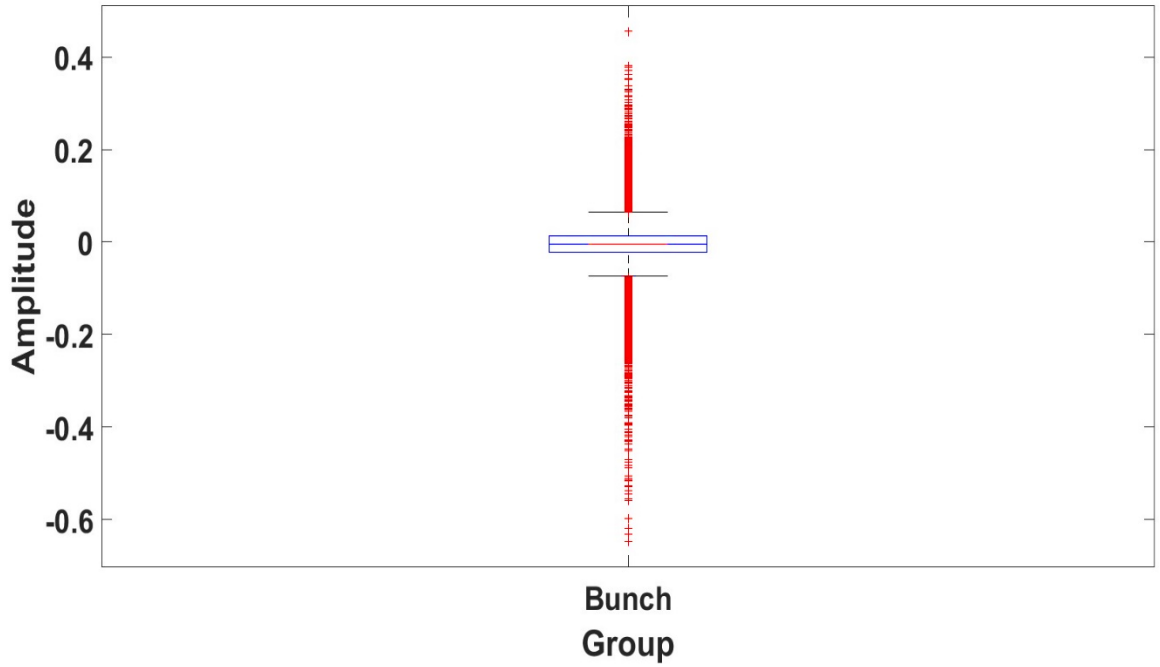


Figure 17: Box plot of bunch

In order to investigate and identify RFI further, various non-normality detection tests were studied. A brief information about these tests is given below:

4.1 Chi-square (χ^2) test:

The χ^2 test is a statistical test used to check if the given sample comes from a Gaussian distribution or not. It is a method to analyse data based on the observations, considering that the null hypothesis of a given data sample is non-Gaussian is true. This test helps to determine the difference between observed values and expected values in one or more classes. It also gives the probabilities of independent variables. Mathematically chi-square test is given by the equation:

$$\chi 2 = \sum \frac{(observed\ value - expected\ value)^2}{expected\ value}$$

The limitation of this test is that it is reliable for the small size of the data.

4.2 Kolmogorov-Smirnov test:

The Kolmogorov-Smirnov test or KS test compares two independent variables and tests if they come from the identical continuous distribution. This is generally more efficient than the chi-square test for goodness of fit for small samples, and it can be used where the chi-square test cannot be applied.

4.3 Shapiro-Wilk test:

The Shapiro-Wilk test considers the null hypothesis is true, unless and until it is rejected when the p-value is less than 0.05. From the research carried out by Guner, Frankford and Johnson

[16], it is observed that the S-W test can be used for RFI detection, but is currently beyond the scope of this project.

4.4 Spectral Kurtosis:

Kurtosis, is the fourth statistical moment, used to measure the tailedness of the distribution. Kurtosis is estimated using an individual dataset and does not bother about the null hypothesis being true or false. The kurtosis of Gaussian distribution is 3. Any distribution with kurtosis less than or greater than 3 is considered to be non-Gaussian.

Block-wise evaluation of kurtosis on the time series is used to detect non-Gaussian noise in this project.

The input time series data is divided into blocks of 8192 samples (hereafter known as a window) in order to calculate the kurtosis. Figures [18] to [21] illustrate the kurtosis of different datasets taken at GMRT.

The threshold values for kurtosis estimation techniques are determined by the equations –

$$Tkup = 3 + \sqrt{\frac{24}{N}} \qquad Tklo = 3 - \sqrt{\frac{24}{N}}$$

Where N = Number of samples

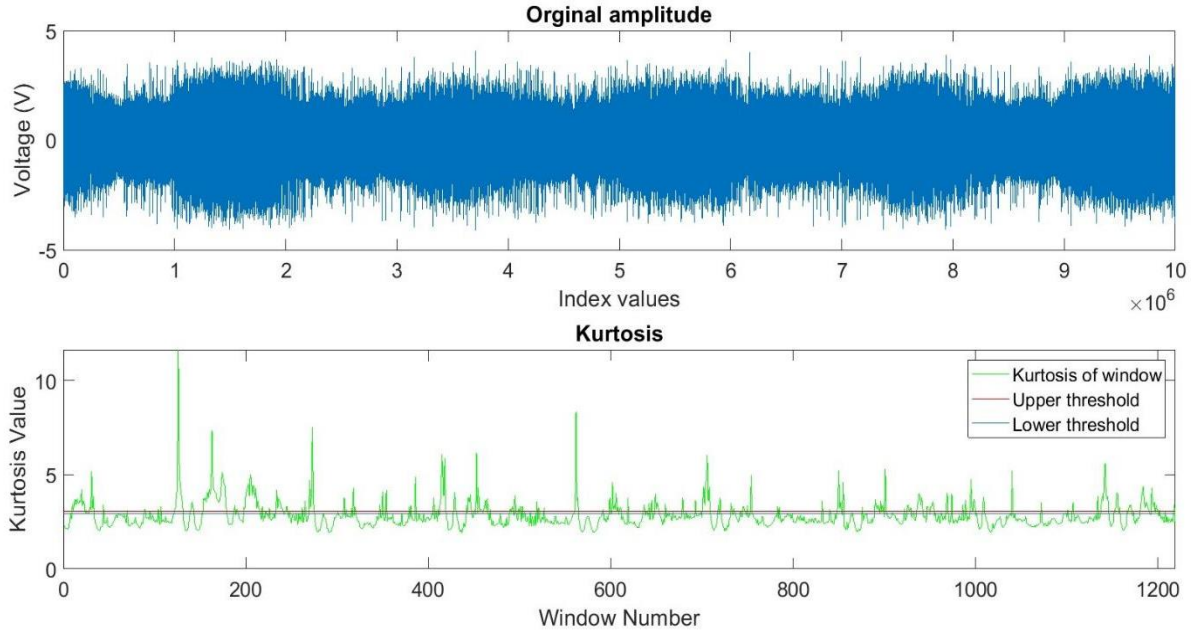


Figure18: Band 2 - Heavily corrupted by RFI

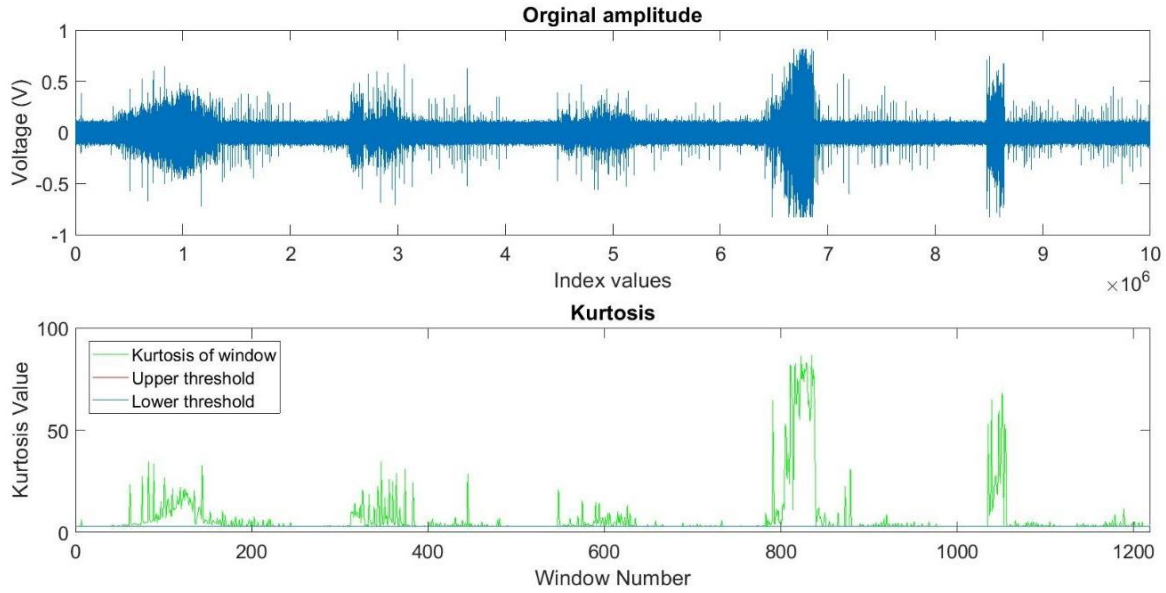


Figure 19: Band 3 - Moderately corrupted by RFI

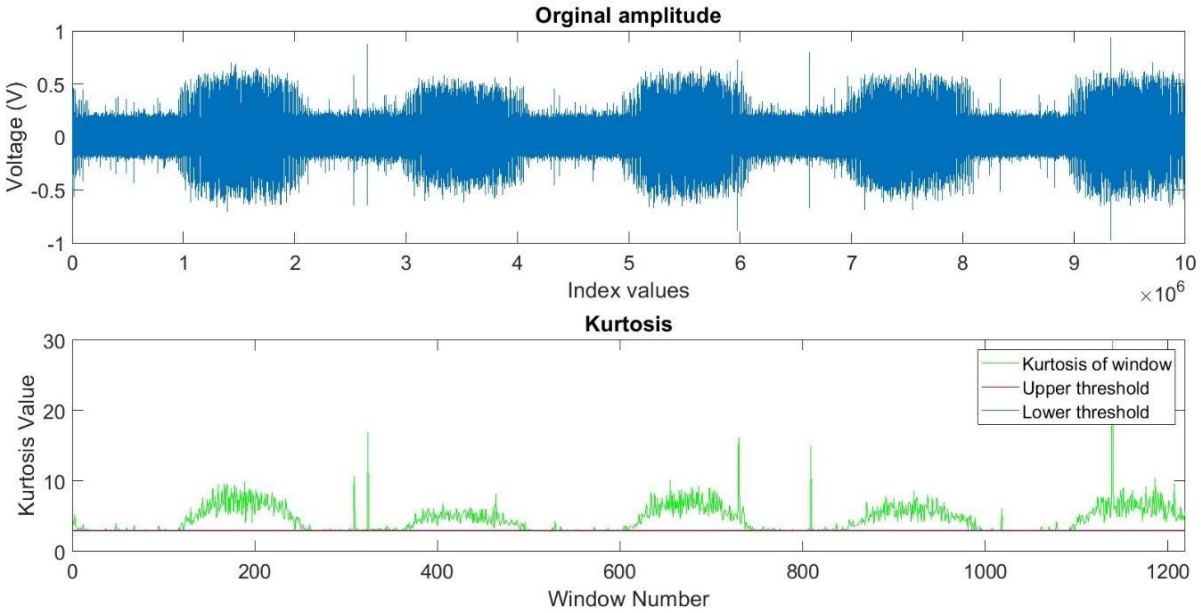


Figure 20: Band 4 - Moderately corrupted by RFI

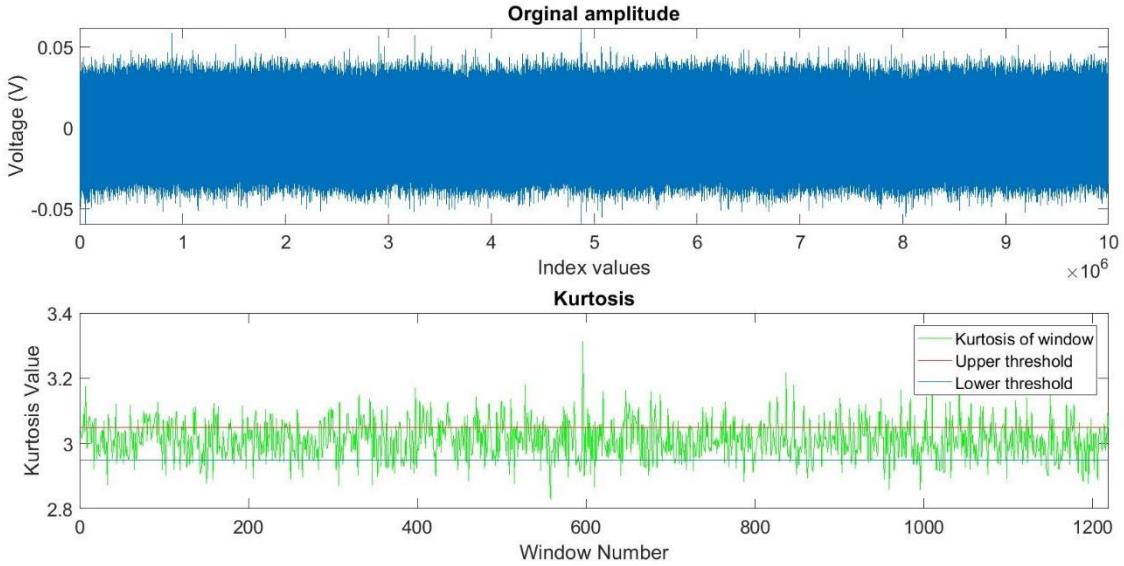


Figure 21: Band 5 - Lightly corrupted by RFI

The above images show the occurrence of RFI in different observing bands of GMRT. Bands 2, 3 and 4 are majorly corrupted by impulsive power-line RFI. It is also seen that kurtosis calculated on the window sample is dependent on the nature of the window. If the window contains normally distributed voltages, its kurtosis is within the thresholds. If the window contains RFI samples, the distribution of the samples in the window is not normal and the kurtosis is either greater or less than 3.

To study the behaviour and variation of kurtosis for fractional increase in outliers, a test experiment was conducted. For the experiment, kurtosis was first estimated on test data of 8192 samples. The input data was then corrupted with some outliers and the variation of kurtosis for per cent increase in outliers was studied. Figures [21] and [22] illustrate the pattern.

It is observed that there is a sharp increase in the kurtosis till the overall outlier count in the sample is less than 2%. After increasing the outlier count above 2%, the kurtosis of the sample decreases gradually and exponentially. A ‘blind spot’ of kurtosis was observed at 58% of the corrupted sample. This is visualised by zooming Fig. [21] and is shown in fig. [22] for better readability.

It can be inferred that the spectral kurtosis estimation technique for Gaussian distribution detection has limitations at 58% of outlier density. At this outlier density, the spectral kurtosis gives false positive values of the data being Gaussian.

This limitation of the spectral kurtosis method can be eliminated by using statistical moments of higher order, as proposed by Roger D. Roo and Sidharth Misra [14]. The sixth moment also has some limitations. The study done by Roger D. Roo and Sidharth Misra indicates that the sixth moment is blind at 13% and 61% of outlier density, and it needs other higher-order moments to eliminate this limitation.

However, the GMRT data analysed so far does not have such a high density of outliers. The average density of outliers due to power-line RFI at GMRT is less than 2%. Hence the kurtosis observed in the time series of GMRT data has very small false positive rate.

The kurtosis increases if the window contains more than 0.05% outliers and it is equal to 3 if there are a greater number of noise (Gaussian distributed) samples. The Following fig. [22] and [23] illustrate this exact behaviour.

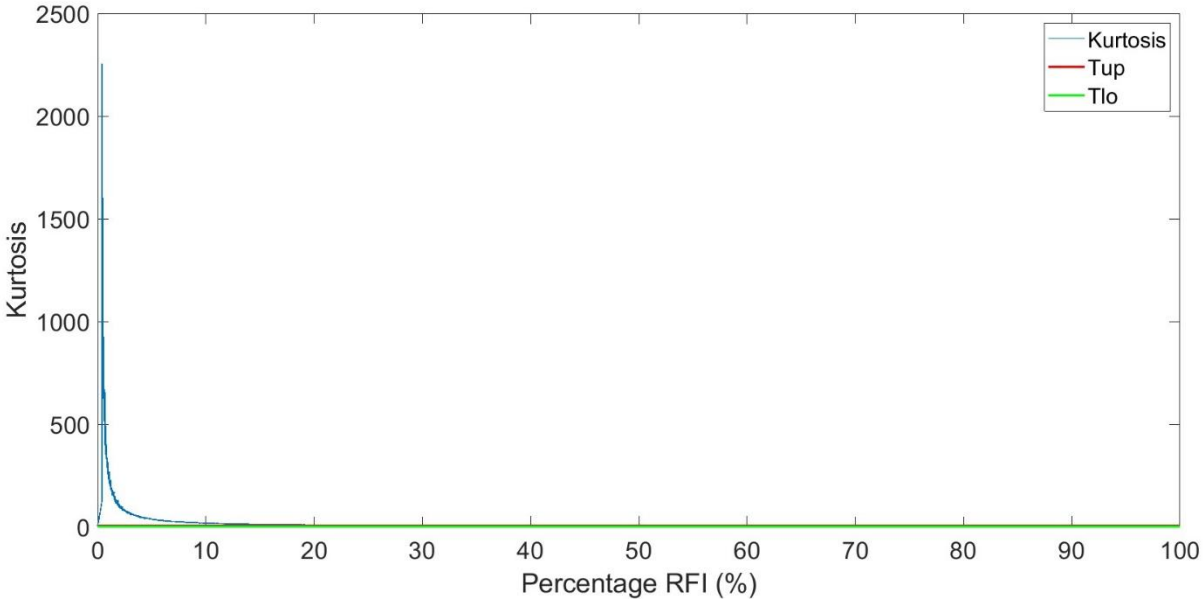


Figure 22: Variation of kurtosis over percent increase in outliers

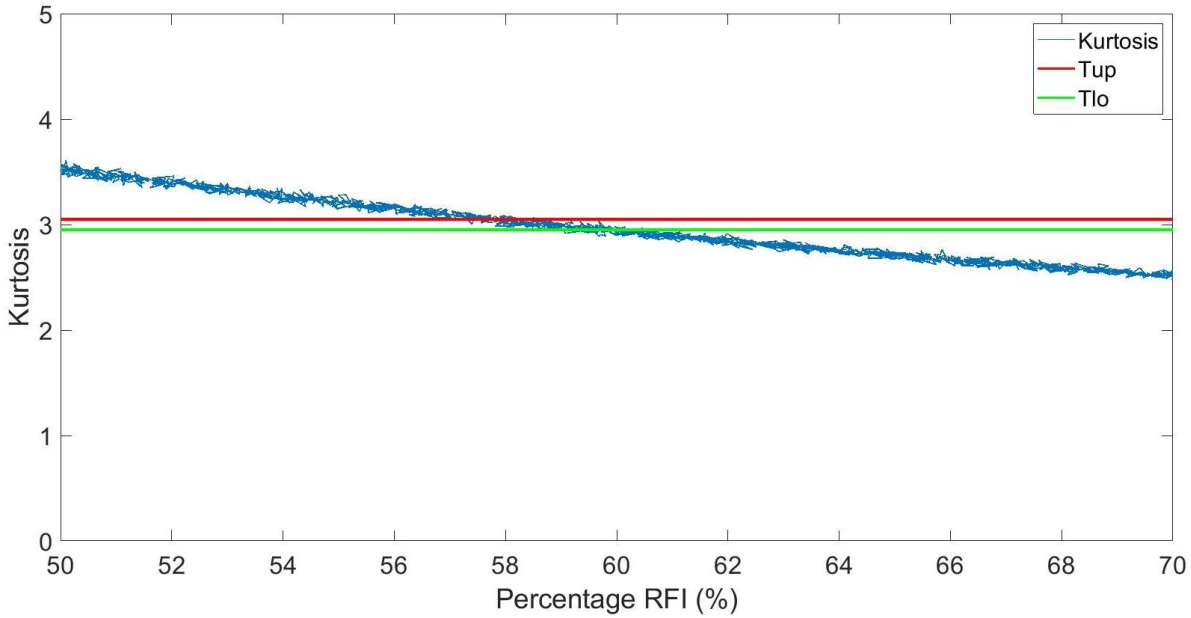
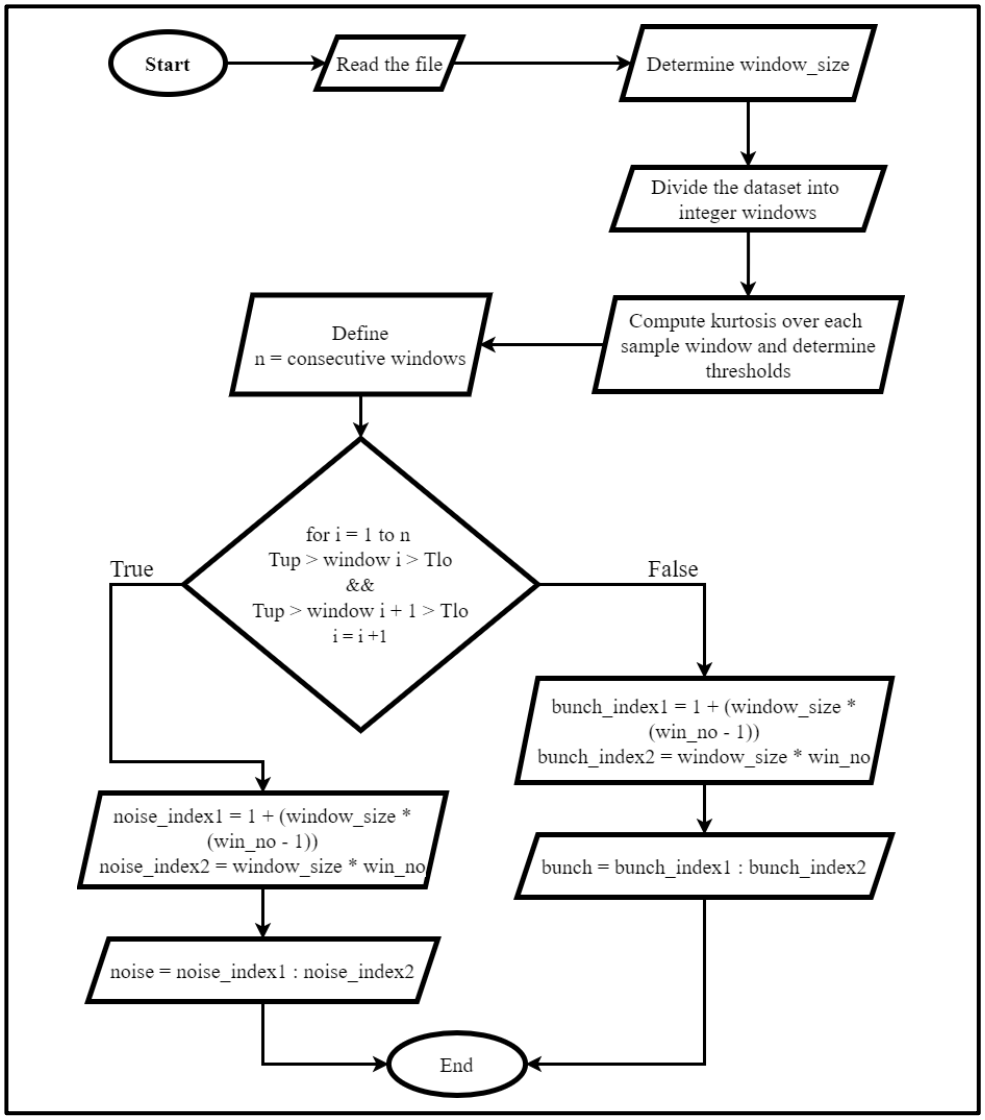


Figure 23: Observed blind spot in kurtosis

Chapter 5: Automatic Bunch Detection

From the experiments and nature of spectral kurtosis studied in the previous chapters, an algorithm was designed which detects a RFI bunch automatically from the given dataset. Kurtosis was computed on a window size of 8192 samples. The thresholds for this sample size are 3.05 and 2.95 respectively. These thresholds change if the window size is changed.
The algorithm of automated bunch detection is explained using flowchart [4].



Flowchart 4: Noise and Bunch detection algorithm

The algorithm represented above is tested on various datasets to automatically detect and flag RFI and noise samples. The samples which are identified as RFI bunch can be statistically evaluated and mitigated further in the signal processing chain.

Some figures are represented below which have RFI bunches automatically detected by the algorithm.

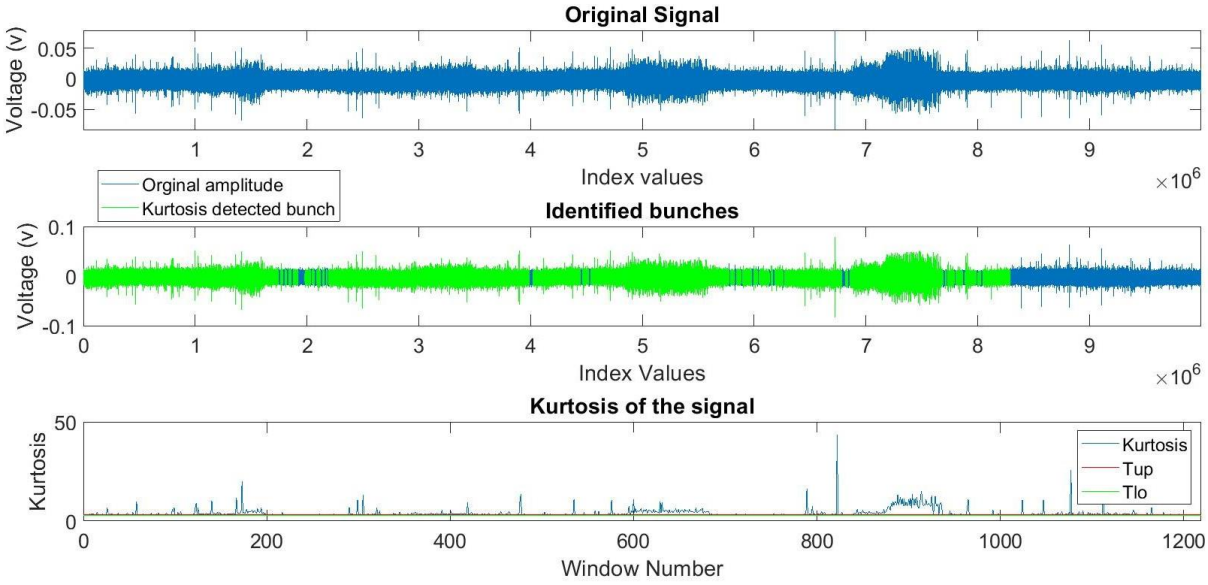


Figure 24: Identified bunches in band 2

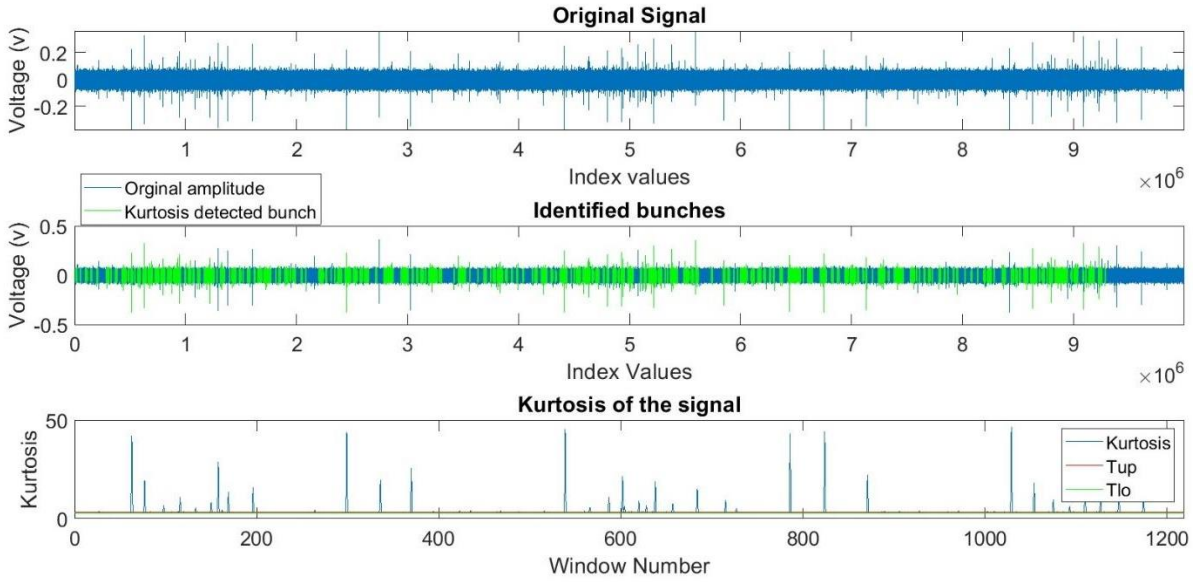


Figure 25: Identified bunches in band 3

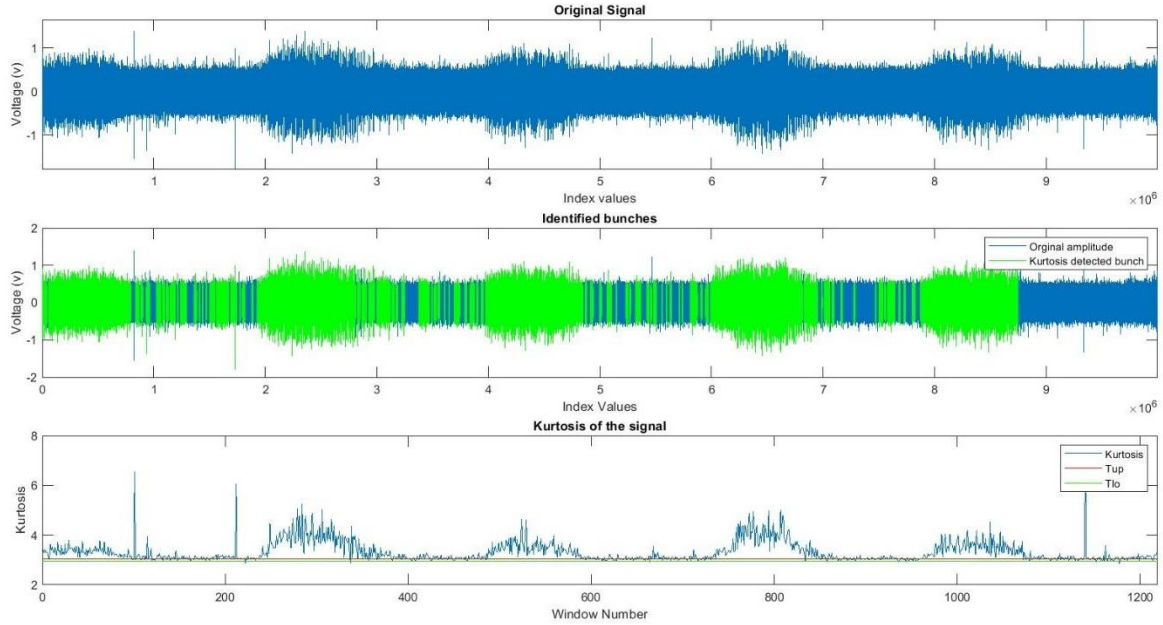


Figure 26: Identified bunches in band 4

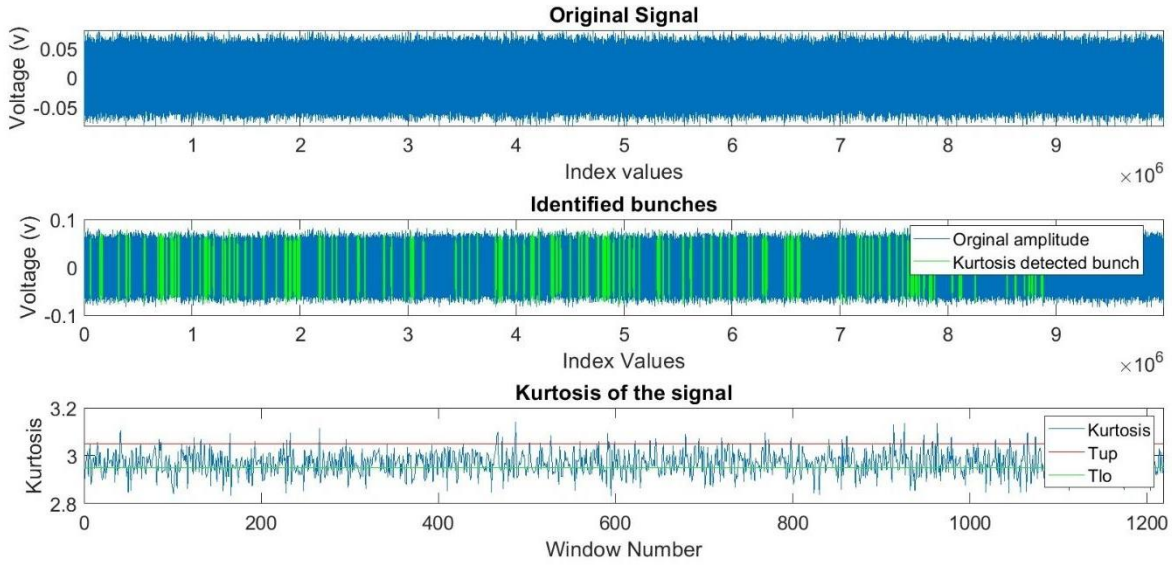


Figure 27: Identified bunches in band 5

The designed algorithm is highly dependent on some of the factors which are explained as follows.

5.1 Window size

The kurtosis is calculated on the entire dataset by dividing it into subsets called windows. These subsets contain samples in the multiples of 1024, commonly referred to as window of 1k samples. The error in kurtosis estimation is inversely proportional to the square root of the number of samples in the estimation. Based on this the uncertainty in the Kurtosis estimation is decided.

Other than this, the window size is also decided based on the typical density of RFI present at a given sampling rate. Kurtosis can be falsely estimated if the sample size is too high. From the relation of number of outliers and kurtosis present in the dataset, it is clearly known that kurtosis have extreme values even at the density of 2% outliers.

If the window size is too large, the kurtosis of the window increases and the Gaussian distributed noise from the dataset is washed out. This is because of the high density of RFI present in the window.

If the window size is too small, the kurtosis of the window is estimated accurately. The windows with kurtosis in between thresholds are correctly classified as noise sample. However, this increases the computation speed of the machine, and individual sparks, having values between 3σ and 4σ are falsely modelled as bunch. To avoid this some approximation is needed in term of window size.

Thus, the user needs to determine window size optimally, according to the statistical properties of RFI. After verifying different window sizes, a window of 8192 samples was found to be optimum for the uGMRT data analysis

5.2 Number of consecutive windows

The 'consecutive_windows' is a user defined parameter of integer data type, required to confidently classify a window as the bunch in the dataset. This is the most important parameter along with window size, in the algorithm, and avoids unnecessary detection and classification of sparks as a bunch in the dataset.

Supporting the fact that kurtosis can have a very high peak value at 2% outlier density, a window of samples can be flagged as RFI if it consists of few elements contributing to 2% of total outliers. There might be cases in the dataset that a window with kurtosis greater than threshold has two adjacent windows with kurtosis less than threshold. In such a case, there is a high possibility that the window with higher kurtosis, has a very small number of outliers and they can be ignored.

To avoid false detection of such windows as a bunch, the parameter 'consecutive_windows' adds a condition to check the kurtosis of a number of adjacent windows before classifying the given window as a bunch.

If the kurtosis of consecutive windows is greater than the threshold, the window is classified as a bunch.

This parameter has made the algorithm more robust and accurate to determine the bunch. If the value of 'consecutive_windows' is large, the bunches having shorter width of windows are ignored, as they fail to satisfy the condition. Also, with this more samples at the end of the dataset are ignored by the algorithm. If the value of 'consecutive_windows' is small, every RFI spark satisfies the condition and is inaccurately modelled as a RFI bunch.

This problem can be effectively stated by the Venn diagram as shown in the fig. [28]. If b is the total number of windows having kurtosis greater than threshold, n is the number of consecutive windows, and w is the number windows that satisfy the condition that at least n windows are corrupted, then the actual number of windows identified as bunch B is given by:

$$B = b \cap n \cap w$$

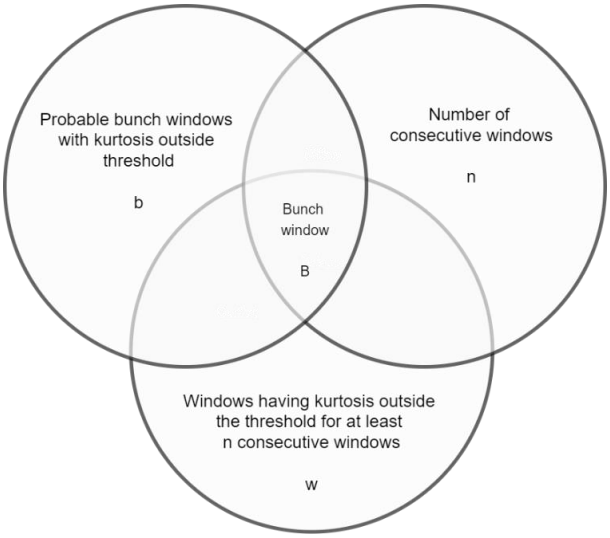


Figure 28: Venn diagram of consecutive windows

Hence in order to identify an RFI bunch using kurtosis accurately, the user needs to tune the above-mentioned parameters.

While doing the data analysis for this project, the setting of 'consecutive_windows' parameter is set to 150 for a window size of 8192 samples.

The user needs to tune this parameter according to the window size and robustness required in the bunch detection.

5.3 Study of signal envelope for improved accuracy

The algorithm of automatic bunch detection has some limitations and drawbacks as discussed in the previous section. To overcome the drawbacks of the kurtosis bunch detection system, the envelope of the entire time series was modelled on different bands of GMRT data.

Figures [29] and [30] illustrate the envelope of the signal observed in various bands.

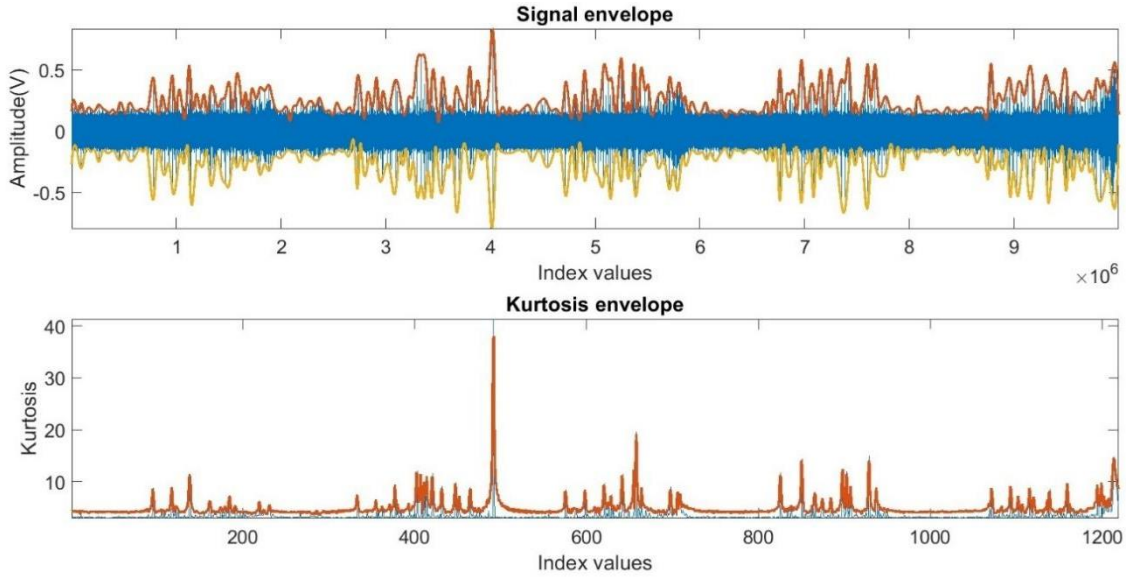


Figure 29: Band 4

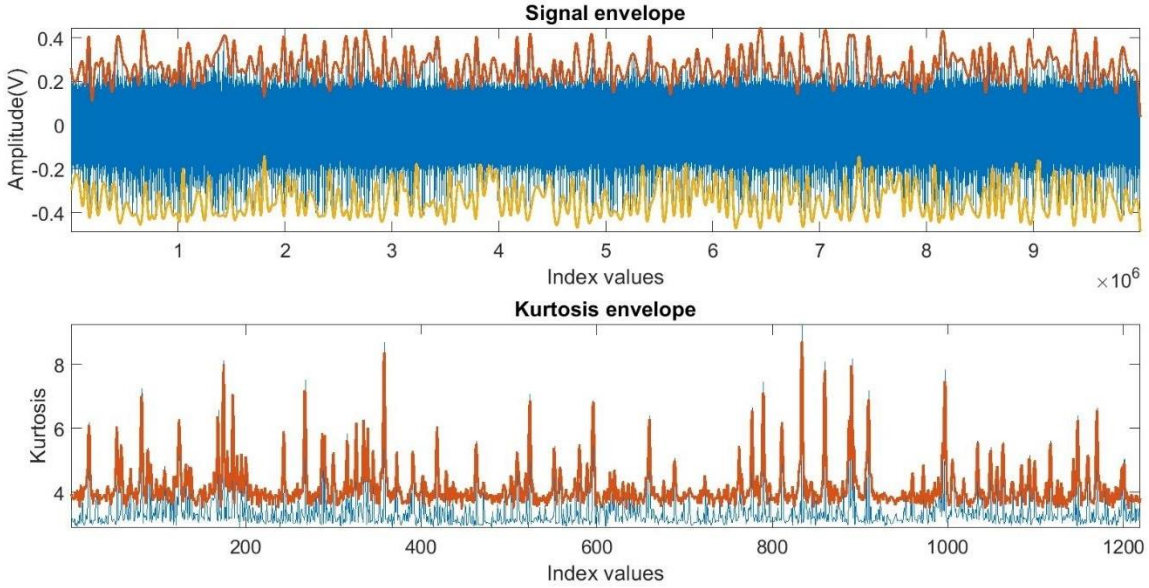


Figure 30: Band 3

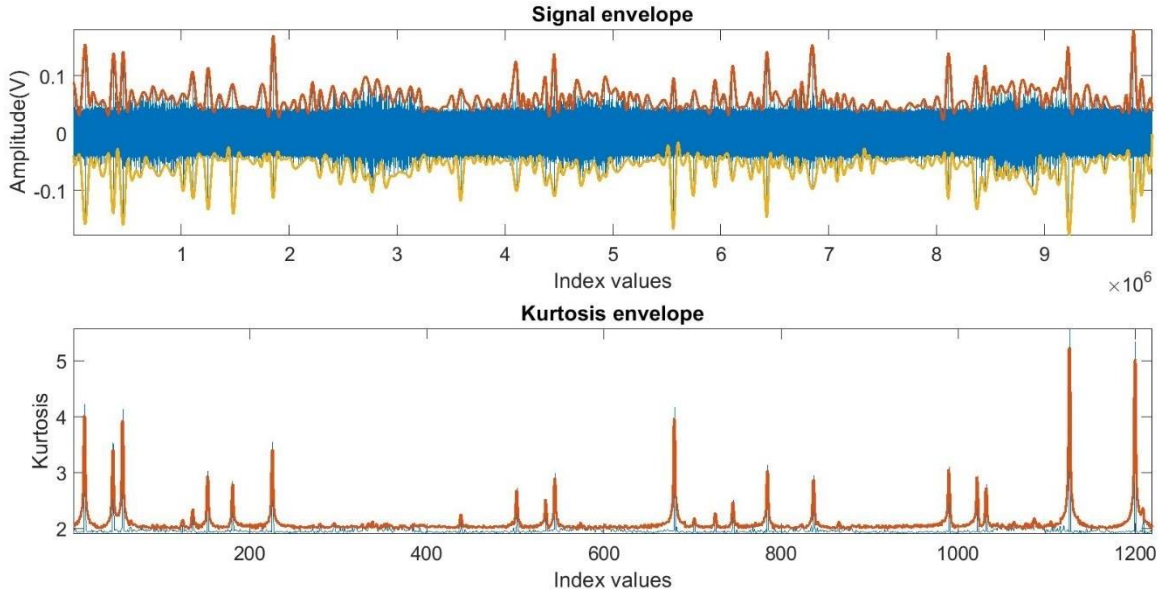


Figure 31: Band 2

The signal envelope follows a cubic spline interpolation, and models every occurrence of RFI. To use it along with kurtosis to estimate a bunch accurately, the thresholds of the signal envelope must be determined. This envelope threshold determination is currently out of the scope of this project, but can be added as a task for the solution of automatic RFI bunch detection problem. The regression curve of cubic spline interpolation of the envelope can be used to predict signal values and mitigate them if they are possible RFI candidates. The system can then be implemented in real time using regression analysis, along with kurtosis estimation to identify and mitigate instances of RFI. Figure [32] shows the cubic spline interpolation curve fitted over the signal envelope.

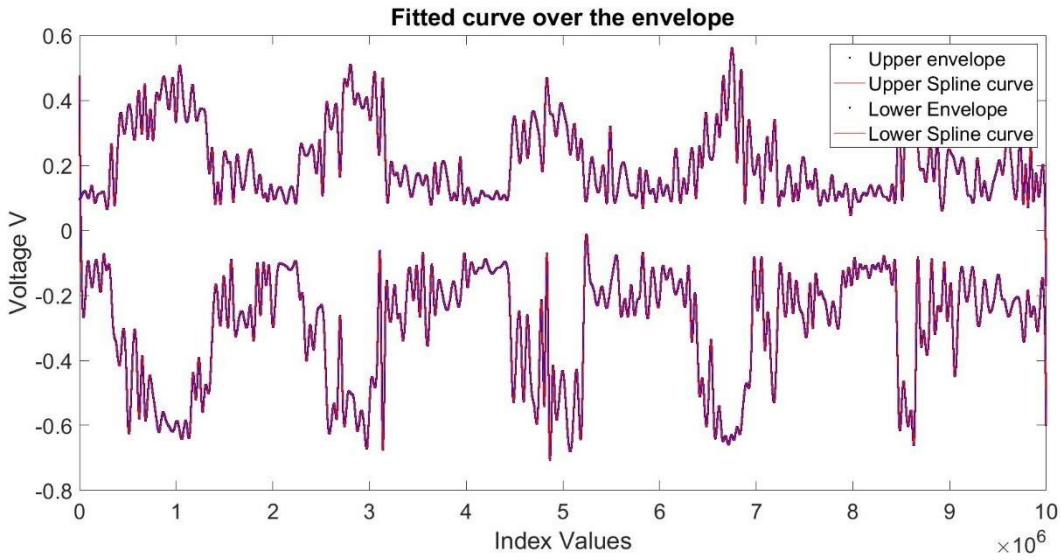


Figure 32: Band 4

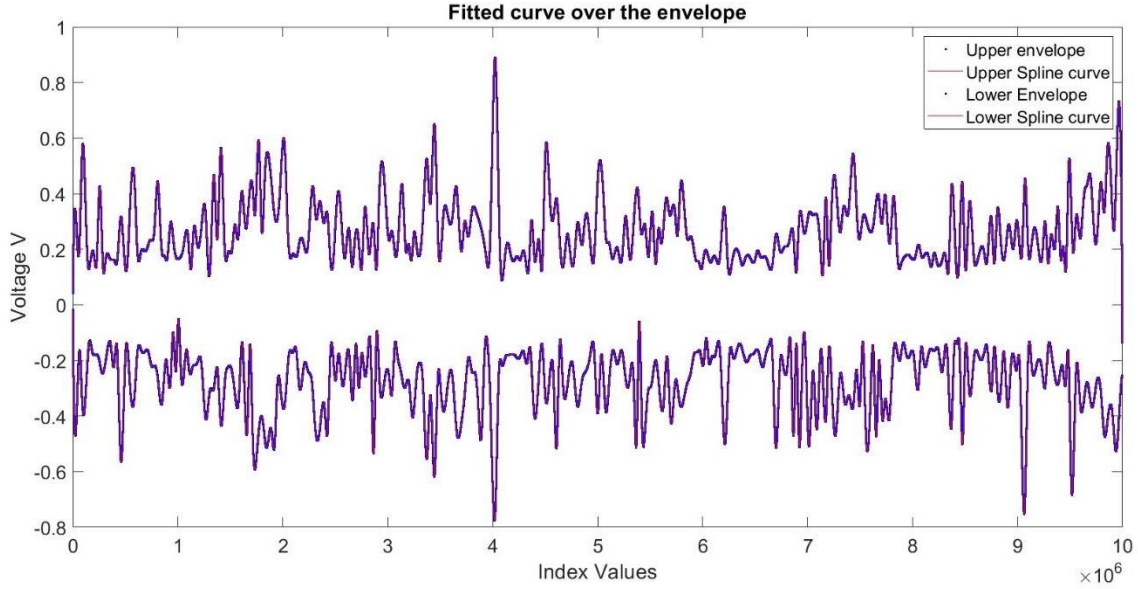


Figure 33: Band 2

Chapter 6: Machine Learning in RFI detection

Apart from the statistical analysis of power-line RFI at GMRT, various machine learning techniques were also studied as a part of this project. Machine learning in the areas of Radio Frequency Interference has opened up a new area for research.
The statistical analysis done so far in this project is the foundation to apply high complexity machine learning algorithms to automatically detect and mitigate power-line RFI.
Broadly, there are two methods of classification and identification in machine learning:

6.1 Supervised learning-

Supervised machine learning techniques use class labels like 'noise', 'pulsar' and 'RFI' to classify the input sample. The machine is trained using a set of labelled data and tested against the set of unlabelled data. In case of RFI mitigation, supervised machine learning models are used to train the machine according to classes. This is a fairly easy method which is based on providing a large number of labelled samples of RFI ridden and clean data. Given the enormous amount of data collected by observatories like GMRT, a supervised machine learning approach can provide moderate accuracy in automatic detection and mitigation of RFI.

6.2 Unsupervised learning-

Unsupervised learning requires large amount of data to train the model. The predictive model learns from unlabelled dataset and is majorly used for clustering and classification. The machines identify patterns in the dataset on their own and make decisions based on their interpretation of the data. In order to classify RFI accurately, unsupervised machine learning models require every occurrence of RFI with similar properties. This is practically impossible as RFI source is random in nature. The random occurrence of RFI makes it very complicated to apply unsupervised machine learning technique over the system.
In addition to these, regression analysis can be used to effectively predict the next occurrence of RFI. Regression analysis can identify and predict a typical broadband RFI source which occurs frequently in the dataset. It can then be mitigated to retain the signal to noise ratio of the astronomical source.
Machine learning techniques in the field of RFI detection and mitigation are studied and implemented by various observatories around the world. Some of the techniques used particularly for RFI mitigation are explained as follows:

6.3 Random Forest Classifier

Random forest classifier (RFC) is an ensemble learning method belonging to the class of supervised learning technique. RFC uses decision trees to classify samples of a subset of the entire series of data. It combines multiple decision trees of multiple subsets to find the ensemble mean and improve the accuracy of the decision. The more are the number of subsets; more are the decision trees and greater is the accuracy of the algorithm.
The data analysis of Karoo Array Telescope done by Mosaine, Oozer and Aniyan [9] uses Random Forest classifier to confidently detect man-made RFI. They further state the accuracy of RFC irrespective of antenna polarization for an 8-second segment sliding window.

6.4 k - Nearest Neighbour (kNN)

kNN is a nonparametric algorithm and classifies data based on k data points which are nearby to the given data point. It identifies the similarity between datapoints and defines a class according to it.

The statistical classification of RFI done by Wolfaardt, Davidson and Niesler [10] uses kNN classification method on data observed by SKA telescope in four frequency bands. RFI data is collected in the time domain using a real time analyser. The captured data was from all 4 bands of the telescope, i.e., from 50Hz - 2550Hz. The time series was processed by segmenting it into windows of 1024 samples. 10-fold cross-validation was used. The dataset was divided into 10 parts, out of which 6 were used for training, 2 for tuning, and 2 for testing the algorithm. The accuracy obtained so far using this algorithm is 20.80%

KNN classifier predicts the class of new data points by selecting the most prevalent class among the k nearest neighbours in the training dataset. The parameter k needs to be tuned.

The data from HERA and LOFAR telescopes analysed by Mesarcik et al. describes an inverted RFI detection algorithm which relies on weak labels. The uncontaminated dataset is labelled as weak. These weak labels are identified using Nearest Latent Neighbour (NLN). The main objective of this method is to address the problem of unavailability and high cost of labelled datasets for ML algorithms, by proposing an inverted approach using NLN to this problem.

6.5 Gaussian mixture model (GMM)

A Gaussian mixture model is a machine learning algorithm used to classify different dataset based on its probability distribution function. It classifies the data based on distribution. It can used to classify the distribution of pure random Gaussian noise from RFI corrupted signal.

Wolfaardt, Davidson and Niesler [10]. GMM classifier creates a model of the data points in a window by fitting the Gaussian distribution over that window. It is then used to calculate the probability that a new data point belongs to a certain class. The tuning parameters for this classifier are the number of distributions and the constraint on the covariance matrix of each distribution. The accuracy achieved so far using this algorithm is 65.86% [10].

6.6 Deep and convolutional neural network-based learning methods

Convolutional neural networks are a subset of machine learning, and comprise of node layers, containing an input layer, some hidden layers and an output layer. CNNs are regularised versions of multilayer perceptrons, which are usually fully connected networks. Convolutional neural networks are inspired by biological processes of decision making.

The implementation of convolution neural networks in RFI mitigation is done by Joel Akeret [11] et al. The data from 7m single dish telescope at Blein observatory is processed using U-Net, a type of convolutional neural network. It classifies RFI signatures in 2D time ordered data, learns a set of features and gives out the probability for each pixel, whether it is contaminated by RFI or not. It is implemented using TensorFlow.

In another comparative study conducted by Haomin Sun [12] et al., it is observed that CNN model performs better than AOFlagger software, but has some limitations on the computational speed. The dataset used for training the model was artificially simulated from the SKA Radio Astronomy Simulation, Calibration and Imaging Library. The observation time was 10 min and the frequency range from 100 MHz to 200 MHz. While implementing the model on the LOFAR & MeerKAT dataset, it was observed that these datasets are not labelled for CNN.

Hence AOFlagger with its default SumThreshold was used to flag RFI and its results were used as the reference truth for the CNN model. This analysis was done on amplitude v/s phase data for different frequency channels.

Chapter 7: Conclusions and future scope

Radio Frequency Interference pose a serious problem to world class sensitive radio telescopes like GMRT. These observatories constantly need to monitor radio frequency interference and maintain radio quiet zones to be in the forefront of scientific research. The present RFI mitigation and monitoring systems at GMRT are very robust. They flag every sample point as an outlier if it falls outside the category of thresholds. This robust estimation may sometimes lead to loss of valuable data, when the threshold is applied using the empirical rule of statistics. The information gathered with this project has opened up new windows in the exciting new fields of machine learning and artificial intelligence. With the entire world transitioning towards automation and artificial intelligence, the challenging problem of RFI can be mitigated using these techniques.

The experiments and analysis carried out in this project forms the foundation for the advanced machine learning techniques required to be deployed in the future for RFI mitigation. A huge amount of data needs to be processed to train various machines for RFI detection. If seen in the big picture, to mitigate broadband RFI occurring from man-made sources like power-lines, the input data should be passed through a machine, which processes time domain signals. These signals come directly from the analogue to digital converter and are processed in real time to detect RFI bursts. Before the online deployment of such a model, it is important to train the machine, according to certain statistical properties of the signal. This project marks the next step towards providing a detailed recipe to begin the modelling and simulation of offline collected data for GMRT. To highlight, the following points can be considered in future of this project:

- Identify RFI bunches with more accuracy and confidence, by studying the signal envelope patterns and their thresholds.
- Using regression analysis, predict and classify the incoming sample either as RFI or as noise
- Implement different machine learning techniques on the huge amount of data available at GMRT
- Automatically select a threshold based on the nature of incoming signal
- Train the machine to make smart decisions on window size based on the selected frequency band of observation.

References and Citations

1. NCRA website; http://www.ncra.tifr.res.in/ncra/gmrt
2. G. Swarup, “Power-line Radio Frequency Interference at the GMRT”, 2008
3. Y. Gupta, B. Ajithkumar, H. S. Kale, et al. “The upgraded GMRT: opening new windows on the radio Universe”, 2017
4. G. Swarup, S. Ananthakrishnan, V. K. Kapahi et al., “The Giant Metrewave Radio Telescope”, Current Science, 1991.
5. B. Bhattacharyya, Jane Roy et al., “Serendipitous Discovery of Three Millisecond Pulsars with the GMRT in Fermi-directed Survey and Follow-up Radio Timing”, 2022
6. Edward, “Seven bad effects of corona on transmission lines”; https://electrical-engineering-portal.com/7-bad-effects-corona-transmission-lines
7. Marvin o. Loftness, “Power Line RF Interference- Sounds, Patterns, and Myths”, 1997
8. Kaushal D. Buch, Ruta Kale, Mekhala v. Muley et al. “Real-time RFI Filtering for uGMRT Observations: Shared-risk Release and Optimal System Configuration”, 2022
9. Olorato Mosaine, Nadeem Oozeer, Arun Aniyan et al., “Radio Frequency Interference Detection using Machine Learning”, 2012
10. Cornelis Johannes Wolfaardt, David Davidson, Thomas Niesler, “Statistical classification of radio frequency interference (RFI) in a radio astronomy environment”, 2016
11. Joël Akeret, Chihway Chang, Aurelien Lucchi, “Radio frequency interference mitigation using deep convolutional neural networks”, 2017
12. Haomin Sun, Hui Deng, Feng Wang, “A Robust RFI Identification for Radio Interferometry based on a Convolutional Neural Network”, 2022
13. Miller and Fruend, “Probability and Statistics for Engineers and Scientists”, Richard Johnson
14. Roger D. Roo, Sidharth Misra, “Effectiveness of the sixth moment to eliminate a kurtosis blind spot in detection of interference in radiometer.
15. Jafar Ali Habshee, “A Simulation-based Modelling and study of the effects of Broadband RFI on Astronomical Data”, 2019
16. B. Guner, M. Frankford, J. T. Johnson, “On the Shapiro Wilk Test for the detection of pulsed sinusoidal Radio Frequency Interference”, 20

Appendix

A] Analysis on different datasets

Sr No	Date	Dataset	Tup	Tlo	ENTIRE SIGNAL					
					Percent RFI (%)	RFI_count (samples)	Not_RFI_count (samples)	Avg duration of RFI_on (seconds)	Avg duration of RFI_off (seconds)	Avg inter arrival time of bunch (seconds)
1	27 August 2020	C01 / P-1 / B2	0.0597	-0.0597	0.7923	79227	9920762	7.75E-06	0.0051	0.1048
		C05 / P-1 / B3	0.0931	-0.0931	0.4199	41988	9958001	9.16E-06	1.32E-02	1.55E-01
		C06 / P-1 / B4	0.0765	-0.0765	0.729	72895	9927094	1.14E-05	0.0054	0.0953
		C08 / P-1 / B5	0.018	-0.018	0.4348	43481	9956508	5.14E-06	0.044	0.1473
2	12 November 2020	C05 / P-1 / B1	0.122	-0.122	0.385	38495	9961494	6.96E-06	0.0223	0.1327
		C06 / P-1 / B2	0.129	-0.129	0.4166	41655	9958334	7.41E-06	0.0159	0.1436
		C05 / P-2 / B3	0.1991	-0.1991	0.4232	42317	9957672	6.50E-06	0.0307	0.0931
		C06 / P-2 / B4	0.1186	-0.1186	0.3824	38240	9961749	6.99E-06	0.0208	0.1476
3	14 July 2022	C08 / P-1 / B3	0.2704	-0.2704	1.0745	107452	9892550	8.30E-06	0.0037	0.0847
		C09 / P-1 / B3	0.4981	-0.4981	0.3714	37143	9962859	5.63E-06	0.0152	0.0989
4	21 October 2021	C00 / P-1 / B2	0.0035	-0.0035	1.9902	199019	9800975	6.06E-06	0.0017	0.2153
5	19 June 2021	C00 / P-1 / B4	0.2204	-0.2204	0.3057	30570	9969427	5.67E-06	0.0844	0.2076

Table 1: Statistical analysis of different datasets

Sr No	Date	Dataset	Tup	Tlo	INSIDE BUNCH					
					RFI_count (samples)	Not_RFI_count (sampels)	Percent RFI (%)	Avg duration of RFI_on (seconds)	Avg duration of RFI_off (seconds)	Avg IAT Spark (seconds)
1	27 August 2020	C01 / P-1 / B2	0.0597	-0.0597	68.8688	8.12E+03	0.8407	1.50E-04	0.1255	2.21E-06
		C05 / P-1 / B3	0.0931	-0.0931	45.016	8.15E+03	0.5495	2.18E-05	2.42E-02	2.45E-06
		C06 / P-1 / B4	0.0765	-0.0765	67.772	8.12E+03	0.8273	1.32E-04	0.1342	6.42E-07
		C08 / P-1 / B5	0.018	-0.018	41.3384	8.15E+03	0.5046	8.84E-06	0.0259	5.74E-08
2	12 November 2020	C05 / P-1 / B1	0.122	-0.122	35.0817	8.16E+03	0.4282	2.01E-05	0.0392	1.60E-06
		C06 / P-1 / B2	0.129	-0.129	37.7646	8.15E+03	0.461	2.07E-05	0.0352	1.50E-06
		C05 / P-2 / B3	0.1991	-0.1991	35.3809	8.16E+03	0.4319	4.36E-05	0.1032	1.16E-06
		C06 / P-2 / B4	0.1186	-0.1186	34.237	8.16E+03	0.4179	2.51E-05	0.0482	1.13E-06
3	14 July 2022	C08 / P-1 / B3	0.2704	-0.2704	93.9633	8.10E+03	1.147	1.32E-04	0.133	1.50E-06
		C09 / P-1 / B3	0.4981	-0.4981	32.0088	8.16E+03	0.3907	4.95E-05	0.1005	2.79E-06
4	21 October 2021	C00 / P-1 / B2	0.0035	-0.0035	232.073	7.96E+03	2.8329	1.02E-05	0.0096	2.17E-06
5	19 June 2021	C00 / P-1 / B4	0.2204	-0.2204	29.0997	8.16E+03	0.3552	7.25E-06	0.0205	7.30E-07

Table 2: Statistical analysis of different datasets

B] List of generic parameters

Sr. No.	Parameter	Type
1	Input file	.txt
2	Window size	Integer
3	Signal frequency	Integer
4	Dataset duration	Float

Table 3: List of generic parameters

C] Major versions of design

Sr No	Type	File Name	Version	Remarks	Status
1	Statistical analysis	statistical_analysis_automated.m	1	Computes RFI count, Not RFI count, RFI percent, IAT of bunch and spark inside the signal and bunch respectively	Final
		statistical_analysis_manual.m	1	Computes RFI count, Not RFI count, RFI percent, IAT of bunch and spark inside the signal and bunch respectively	Final
2	Spark morphology	Spark_morphology_pre_final.m	1	Fits fourier power 4 curve on identifed sparks inside the dataset. Save the outptu in PDF file	Final
3	Signal Envelope	Signal_Envelope.m	1	Models the signal envelope and kurtosis envelope. Also fits a cubic spline interpolation curve on the signal envelope	Final
4	Blind spot study	Kurtosis_blind_spot.m	1	Studies the variation of kurtosis for fractional increase in outliers	Final
		sixth_moment_v1.m	1	Uses sixth moment to study the variation of fractional increase in outliers	Final
5	Correlation	correlation_time_series_v1	1	Computes the correlation coefficient on window size of 16 samples by lag correcting two signals	Final

Table 4: Major code versions

D] Input file format

Supported file type	Delimiter	Number of columns	Number of rows	Number of rows ignored
.txt With floating literals	,	2	>100	First 5 containing string literals

Table 5: Input file format

E] Limitations of the proposed model

Some of the limitations of this project are listed below-

- The automatic bunch detection algorithm ignores some samples at the end of the dataset.
- The goodness of fit statistics of spark curves are low for some of the identified sparks.