BE PROJECT REPORT

On

"Exploring Non-normality Detectors for Automated Powerline RFI Identification"

Submitted by,

Avantika Iyengar (B190453045)

UNDER THE GUIDANCE OF

Prof. Snehal J. Koparde

(Internal Guide)

Mr Kaushal D. Buch

(External Guide)

Dr Divya Oberoi

(External Guide)

IN THE PARTIAL FULFILLMENT OF

BE (ELECTRONICS AND TELECOMMUNICATION)

SAVITRIBAI PHULE PUNE UNIVERSITY

2022 - 23

MARATHWADA MITRA MANDAL'S COLLEGE OF ENGINEERING, PUNE - 52

DEPARTMENT OF ELECTRONICS AND TELECOMMUNICATION ENGINEERING

MARATHWADA MITRA MANDAL'S

COLLEGE OF ENGINEERING, PUNE

Permanently Affiliated to SPPU

Accredited with an "A" Grade by NAAC

CERTIFICATE

This is to certify that the project report entitled

"**Statistical Analysis of Powerline Radio Frequency Interference"**

Submitted by

Ms Avantika Iyengar (Roll No. BE134, Exam Seat no.- B190453046)

is bonafide work carried out by them under the supervision of Prof. Snehal Koparde. It is approved for the partial fulfilment of the requirement of Savitribai Phule Pune University for the award of the degree of Bachelor of Engineering (Electronics and Telecommunication). This report has not been submitted earlier to any other Institute or University for the recognition of any degree or diploma.

Prof. S. J. Koparde	Dr G. S. Gawande	Dr V. N. Gohokar
Project Guide	Head	Principal
Department of E&TC	Department of E&TC	MMCOE, Pune

Place: Pune

External Examiner Name & Signature:

Date:

Support Letter

National Centre for Radio Astrophysics
TATA INSTITUTE OF FUNDAMENTAL RESEARCH

NCRA, PUNE UNIVERSITY CAMPUS, POST BAG 3, GANESHKHIND, PUNE 411 007, INDIA
Telephone Direct: + 91 020 2571 9245;Gen: 2571 9000; Fax +91 020 2569 7257 div@ncra.tifr.res.in

Dr. Divya Oberoi
Associate Professor

Dear Sir/Madam,

This is to certify that Avantika Iyengar, a final year student of the MMCOE, has been accepted at the National Centre for Radio Astrophysics to carry out her final year BE project in the academic year 2022-2023. The expected period of her project is from 01/10/2022 to 31/03/2023. She will work on the joint guidance of Kaushal Buch and Prof. Divya Oberoi. Her project will be focused on understanding and mitigating the impact of broadband radio frequency interference observed at the Giant Metrewave Radio Telescope and will be designed to meet the requirement of her curriculum.

Divya Oberoi
26 September, 2022

Acknowledgements

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. Their expertise and insights were instrumental in shaping the direction of the project and ensuring its successful completion.

Though I was unfamiliar with some of the concepts required for this project, Mr Kaushal Buch took me on board and made sure to bridge the gap with all the necessary resources. The thought-provoking technical discussions held with him were enough to screw my head and keep me engaged in the work. I would also like to thank Dr Divya Oberoi for his valuable comments and suggestions on the project. His critique remarks have always helped me revisit and ponder over my problem-solving approach.

I take a moment to extend my gratitude to Ms Snehal Koparde for her support and feedback throughout the project. Her contributions have been invaluable in helping me to refine my ideas and develop a better understanding of the subject matter.

Furthermore, I am deeply grateful to the staff and management of NCRA Pune, GMRT Khodad, and MMCOE Pune that provided me with the resources and facilities necessary for the completion of this project. Their support has been essential in facilitating the research and analysis required for this project.

Avantika Iyengar

Table of Contents

List of Figures

List of Tables

List of Flowcharts

List of Equations

Abstract

This project particularly aims to study the statistical properties of power-line Radio Frequency Interference (RFI) occurring due to various sources around the Giant Metrewave Radio Telescope (GMRT) array. The agro-industrial manufacturing units and farming establishments surrounding GMRT require high voltage operation which is provided by the high-tension power lines that are a potential source of broadband RFI, causing interference to the GMRT observations. It affects the receiver system by degrading the signal-to-noise ratio and therefore is a major concern a sensitive instrument like the upgraded GMRT (uGMRT).

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 input to the correlator block of the GMRT wideband backend system.

This project forms a part of a bigger project which further aims to use the latest technologies of Machine Learning and Artificial Intelligence, to automatically detect and mitigate broadband RFI. The RFI bunch which occurs quite frequently in the time series is automatically identified using kurtosis. To make this automatic bunch detection algorithm more robust, an additional method of modeling the signal envelope is also planned for research. This project provides a foundational study for the use of recent trends in machine learning for RFI mitigation at the uGMRT.

CHAPTER 1: INTRODUCTION TO uGMRT AND BROADBAND RFI

The Giant Metrewave Radio Telescope (GMRT) is a state-of-the-art and indigenously designed telescope, located around 80 km 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 25 km. This telescope operates in the meter wavelength part of the radio spectrum, to minimize 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 modern telecommunication methods rely extensively on radio frequency bands and they are a valuable resource to any organization. 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 60 dB more sensitive to radio signals than a typical radio communication receiver [2].

Figure 1: One of the antennas of GMRT *Courtesy: NCRA Archives*

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 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.

1.1 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. Figure 2 below represents narrowband RFI at GMRT. The narrow peaks observed in the image represent Narrow band RFI in the 325MHz observing band of GMRT.

1.2. 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. Figure 4 represents the broadband RFI observed at GMRT which is mainly caused due to high voltage powerlines, and hence also known as powerline RFI.

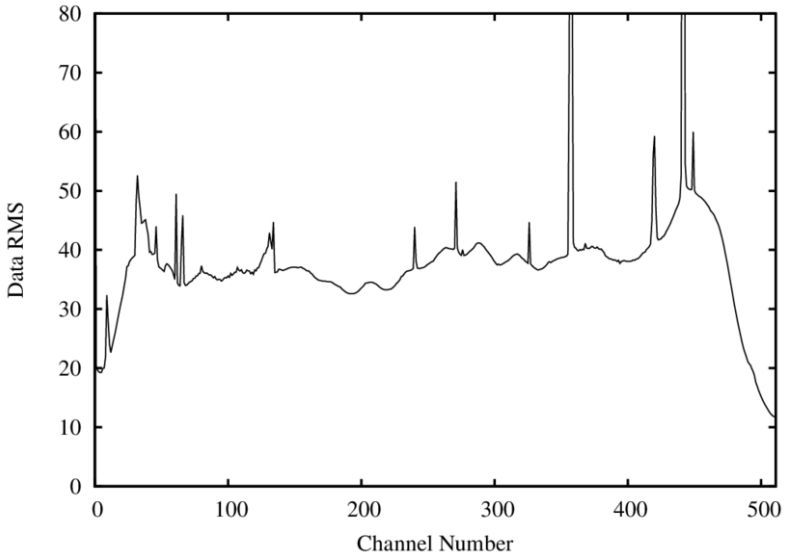


Figure 2: Narrowband RFI at GMRT *Courtesy: Mayuresh Surnis*

1.3 LITERATURE SURVEY

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

Table 1: Frequency bands at GMRT

GMRT is a versatile instrument carrying out observations in the varied subfields of Astronomy. For the past two decades, GMRT has been used in the discoveries of numerous galaxies, the study of atomic hydrogen, the study of Fast Radio Bursts, galaxy mergers etc. One of the scientific goals as envisioned by Swarup et al. [4] of GMRT, is to observe and search for short-period pulsars. Pulsars are highly magnetized 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 neighboring industries and the subsystems of GMRT generate a significant amount of broadband RFI.

1.4 PROPERTIES OF POWERLINE RFI

As explained briefly in the above section, power-line RFI occurs due to gap discharge or corona discharge. The gap discharge results in 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. This arc produces a low impedance path and creates a short circuit across the gap. 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^2R power losses. In the presence of water vapor, the air surrounding the conductors is ionized quickly and gives rise to corona discharge. This type of discharge generates gasses like ozone O3 and various nitrogen oxides (NOX) [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, 50 Hz in India). Figure 3 visualizes the observed RFI in the actual data captured in Band 2 at GMRT.

Figure 3: Corona discharge near GMRT

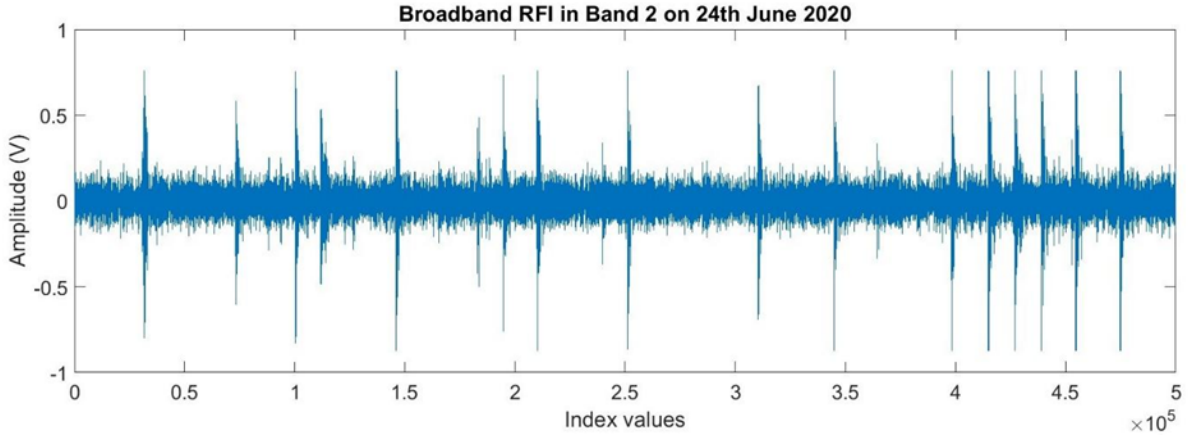


Figure 4: Power-line RFI observed in B2 at GMRT at a sampling rate of 5 ns

CHAPTER 2: STATISTICAL CHARACTERISTICS OF POWERLINE RFI

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… X_n,

$$MAD = median(|Xi - median(X)|).$$

The threshold values based on MAD is 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 thresholds is given by

$$Upper\ threshold\ Tup = \text{median(data)} + 3 * 1.4826 * MAD(data)$$
$$Lower\ threshold\ Tlo = \text{median(data)} - 3 * 1.4826 * MAD(data).$$

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.

2.1 BLOCK DIAGRAM OF THE SYSTEM

Figure 5 below describes the overall block diagram of this project. In this project, different statistical techniques are used to classify RFI and noise samples. This block diagram explains the overall flow of the project.

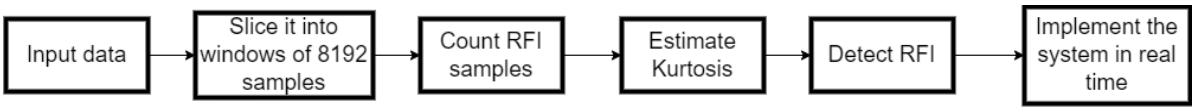


Figure 5: Block diagram of the project

2.2 STATISTICS

Statistics is an important branch of mathematics used to define, interpret, investigate and infer any practical problem. 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 into the nature and distribution of the data. It helps a user to interpret and analyze the data better and reach more precise conclusions.

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 a 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 behavior 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

2.3 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 the ADC channel, before sending them as an input to the correlator system of GMRT. The output of the ADC channel is sampled at 400 x 10^6 Hz sampling frequency and is converted into instantaneous voltages. These signals are observed and recorded from the LeCroy oscilloscope available at GMRT in the form of text files. The text files are further processed in MATLAB to visualize the recorded data. Figure 6 shows the block diagram of the system and the location from where the data is tapped for further analysis. Figure 7 is the actual image of LeCroy Oscilloscope from the receiver room of GMRT.

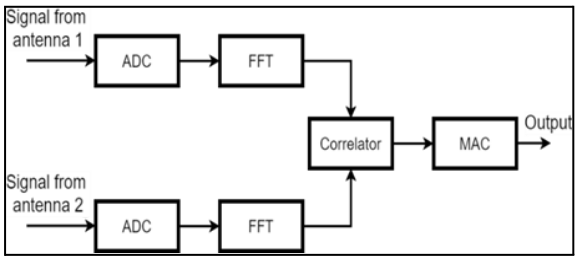


Figure 6: Block diagram of the system

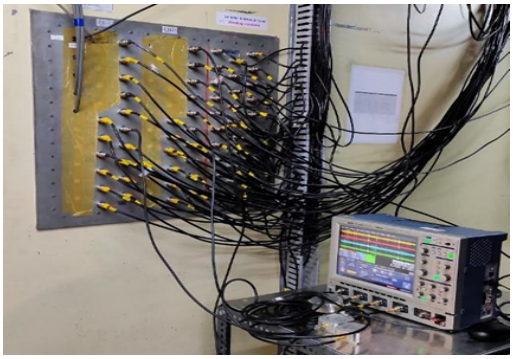


Figure 7: Setup for recording high time-resolution powerline RFI using oscilloscope at GMRT

The data visualized from GMRT is represented by the following Figure 8. Datasets recorded in different seasons (has varying powerline RFI) are visualized in MATLAB in order to study the pattern and occurrence of power-line RFI

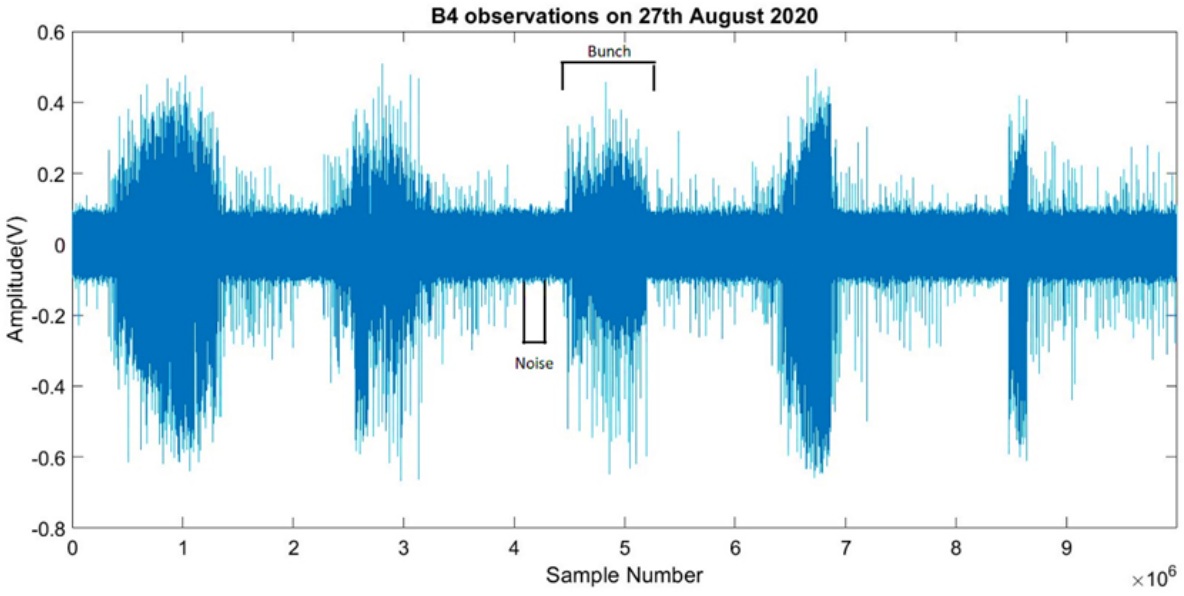


Figure 8: Signal in Band 4

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 an 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.

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 Figure 9 and 10.

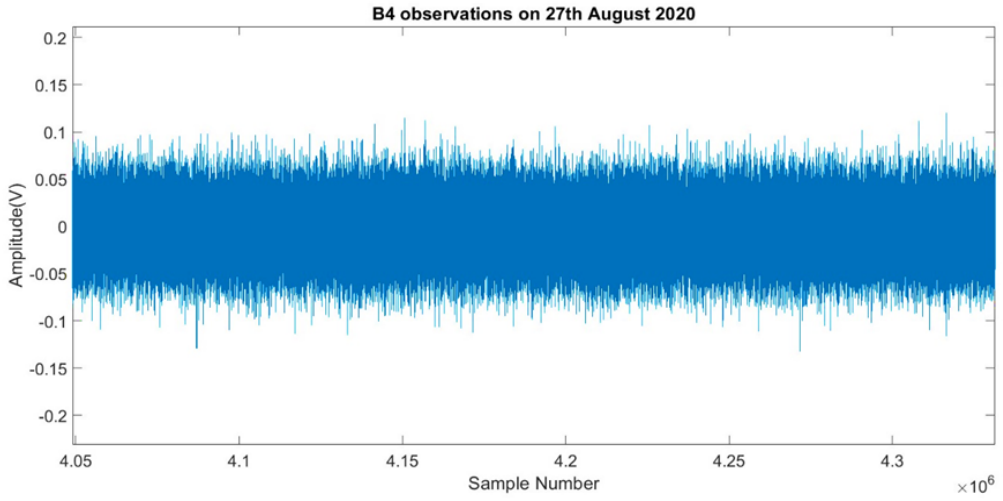


Figure 9: Noise Sample

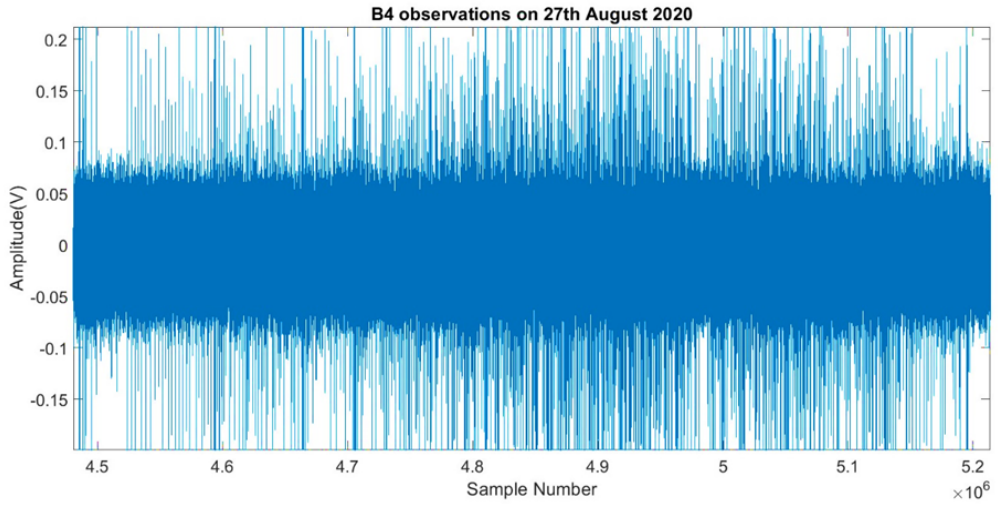
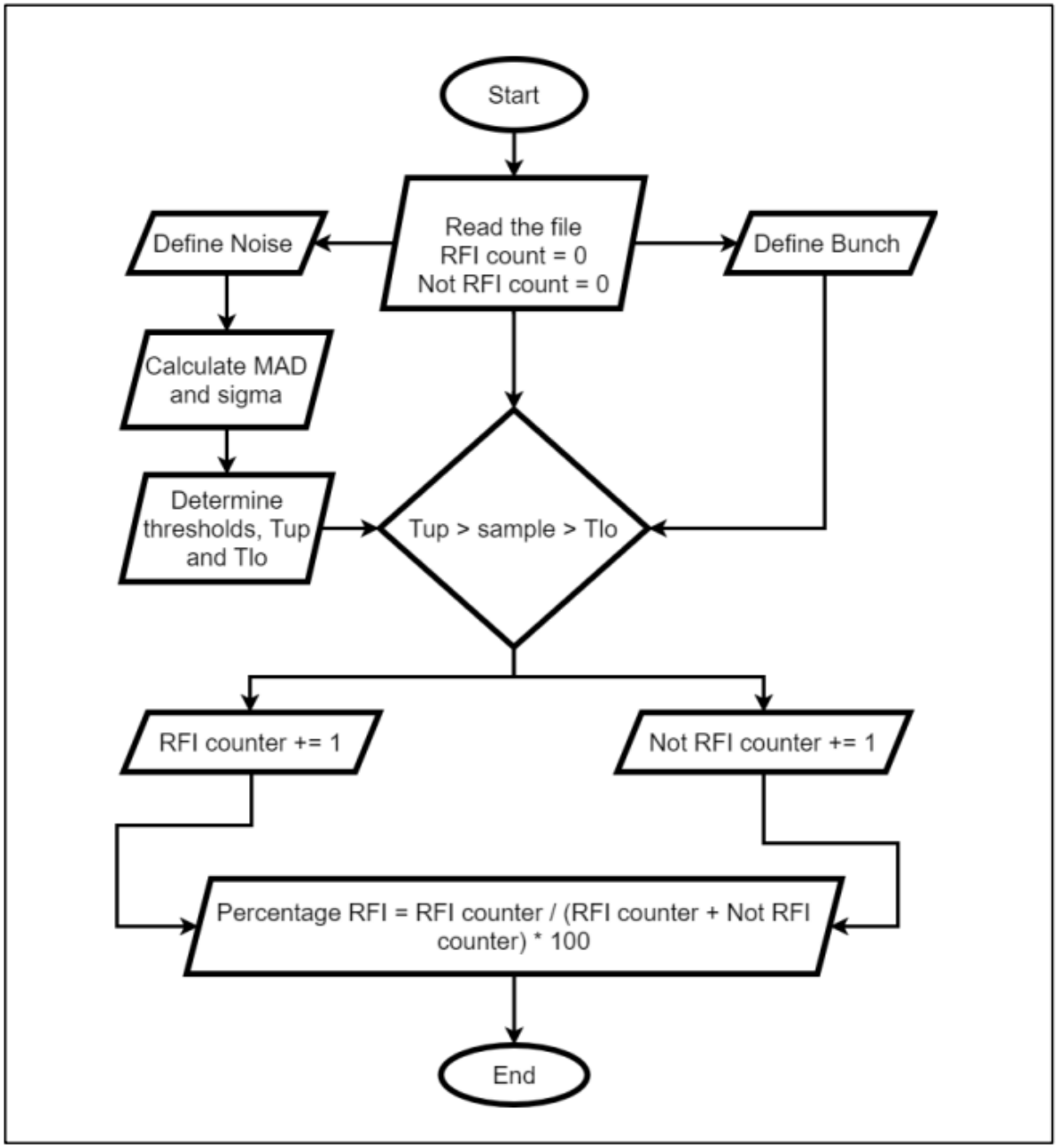


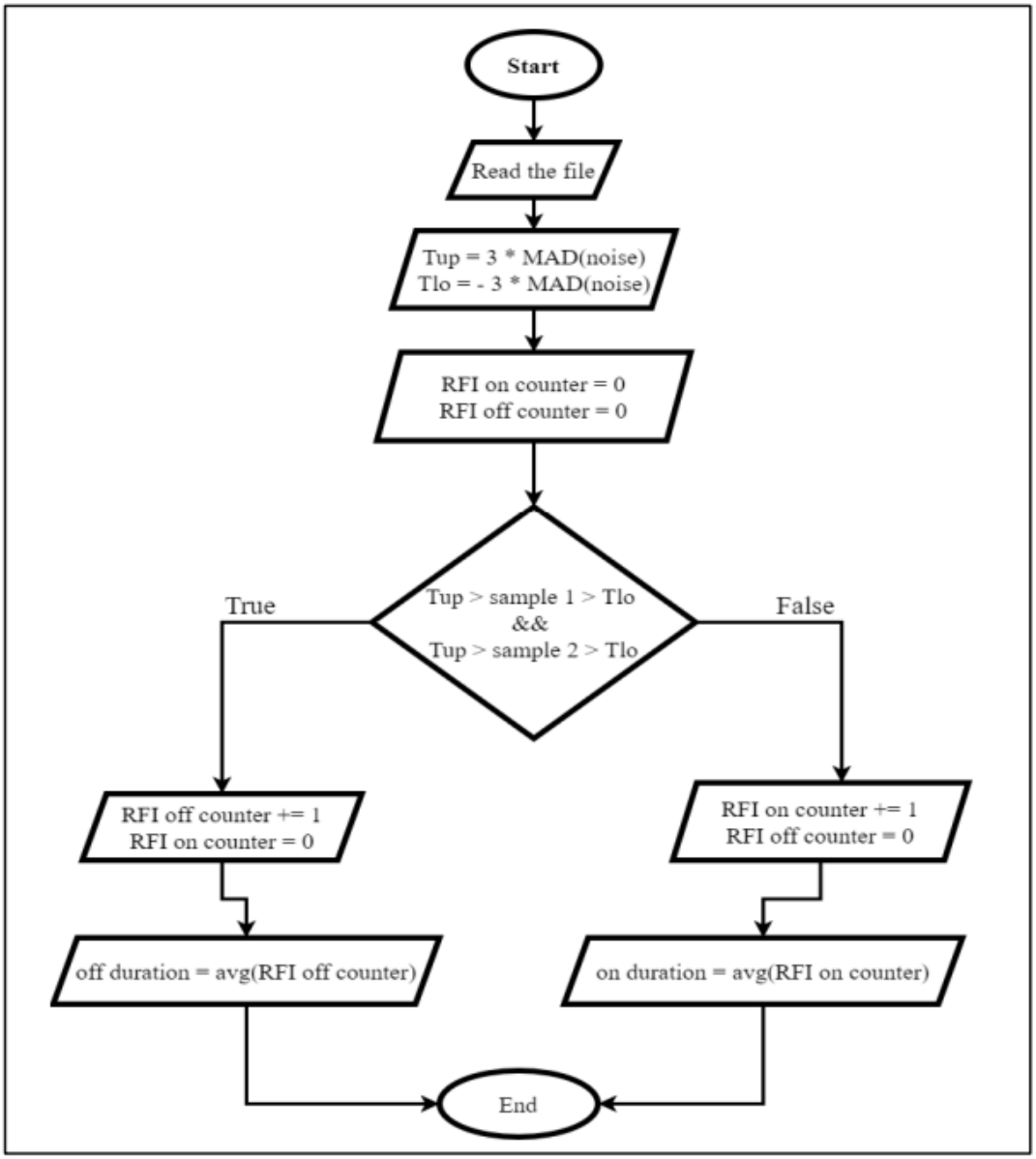
Figure 10: Bunch sample

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 use of median absolute deviation to determine a threshold. These thresholds (generally 3σ) were used to count the number of outliers present in the dataset and thereby compute the density of RFI in the data. It is visualized 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.

It is observed by the implementation of this algorithm on numerous datasets that although the input dataset seems (visually) to be heavily corrupted, the actual RFI density in the signal is anywhere between 0.5 – 2% (MAD based flagging). It is rational to infer that, in general, power-line RFI occurring at GMRT typically corrupts up to 2% of the data. The RFI density inside the heavily corrupted bunch was also found to be around 1% (MAD based flagging) of the total number of samples present in the bunch. This gives an important insight into 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. These results can be viewed in the section C of the Appendix of this report. 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 with a sparse temporal occupancy.

CHAPTER 3: STATISTICAL TOOLS FOR RFI IDENTIFICATION

It is generally observed that radio frequency interference follows a non-Gaussian distribution, and the properties of RFI are statistically different from the properties of a normal distribution. On the other hand, the astronomical signal (noise) follows a Gaussian distribution. These two distinct properties of signal and RFI open up a new window to explore non-normality detection in the time series. With this idea, a set of samples or blocks 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 11 and Figure 12 represent the boxplots of noise and RFI bunch from GMRT data. The central line of the box plot indicates the mean value of distribution while the red markers indicate the outliers. The upper boundary of the box indicates the 75th percentile of distribution and the lower boundary indicates the 25th percentile of the 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 Probability Density Function (PDF) curve. Thus, these boxplots help to visualize the PDFs of noise and RFI.

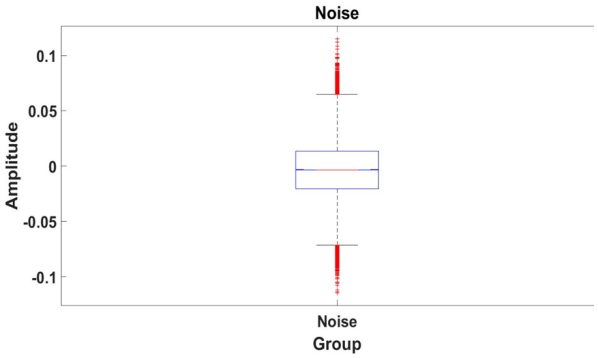


Figure 11: Boxplot of noise

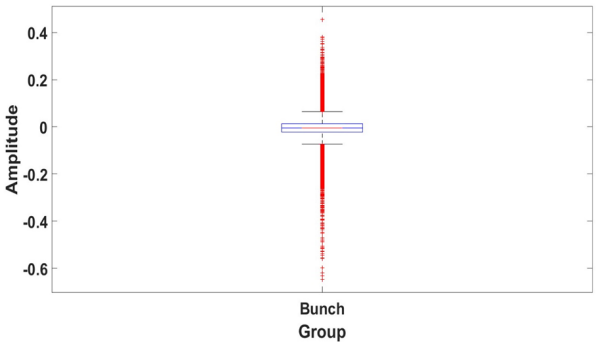


Figure 12: Box plot of a RFI bunch

In order to investigate and identify RFI further, various non-normality detection tests were studied. Statistical identification RFI can be done using two different methods; sample-based outlier detection, and block-based outlier detection.

3.1 SAMPLE-BASED RFI DETECTION

This method involves the Median Absolute Deviation (MAD) based filtering technique for RFI mitigation. As described in Chapter 2, the GMRT backend signal processing system uses MAD-based RFI filter to remove outliers. This technique operates on individual samples in the dataset and flags them either as an outlier or noise samples. It is observed that the concentration of actual outliers for sample-based detection is very less in the entire time series. This method might be useful for narrow impulses, but in order to detect and flag the entire RFI bunch, there is a need for a different detection technique. Results attached in part B of the Appendix illustrate sample based and block based RFI detection.

3.2 BLOCK-BASED RFI DETECTION

The input time series data is divided into blocks of 8192 samples (hereafter known as a window) in order to detect the outliers. It can be inferred from the box plots 10 and 11 that even a small concentration of outliers can change the shape of the overall distribution of the sample. This change can be measured using different non-normality tests. It is evident from the experiments carried out in this project that non-normality tests like kurtosis, and sixth-moment are more useful for RFI bunch detection. Each of these individual tests is explained in the further sections of this chapter.

3.3 KURTOSIS FOR NON-NORMALITY DETECTION

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. Kurtosis is calculated on every window. Figures 13 to 16 illustrate the kurtosis of different datasets taken at GMRT.

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

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

$$Where\ N = Number\ of\ samples$$

Equation 1 Estimation Error for kurtosis

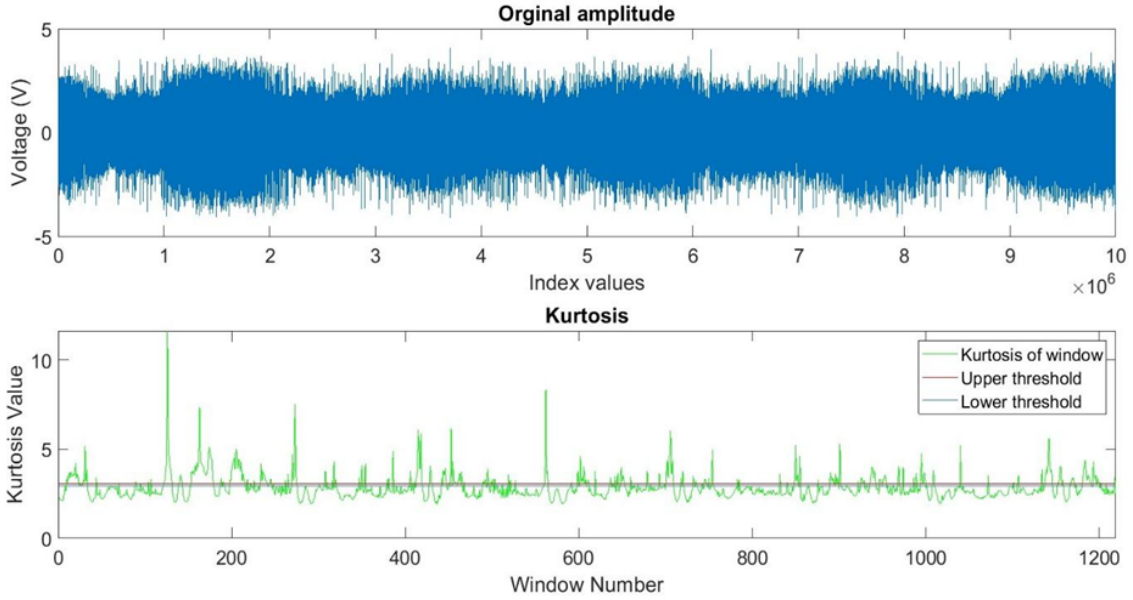


Figure 13: Band 2 - Heavily corrupted by RFI

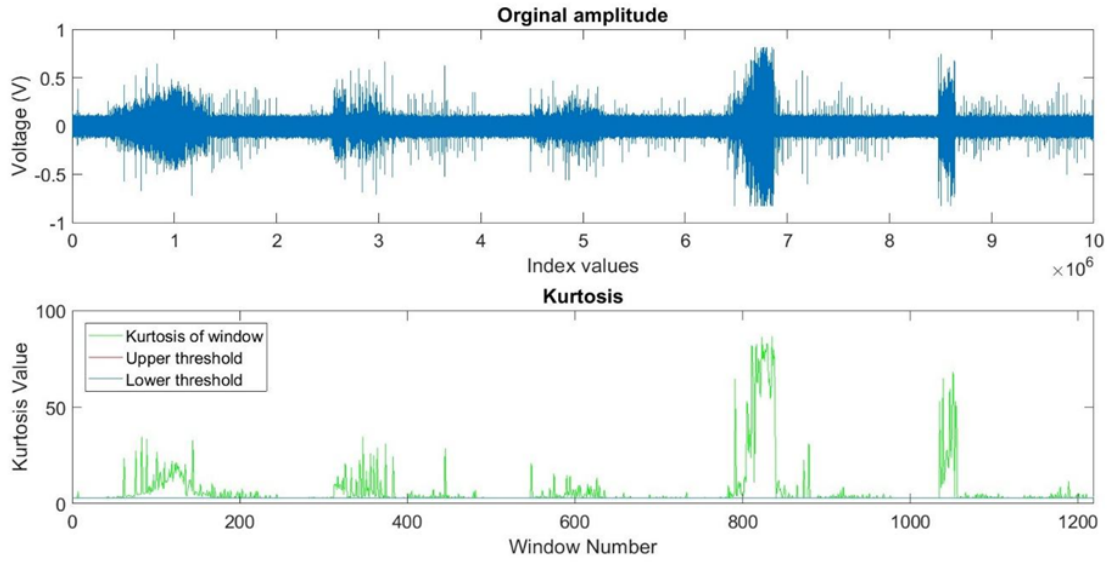


Figure 14: Band 3 - Moderately corrupted by RFI

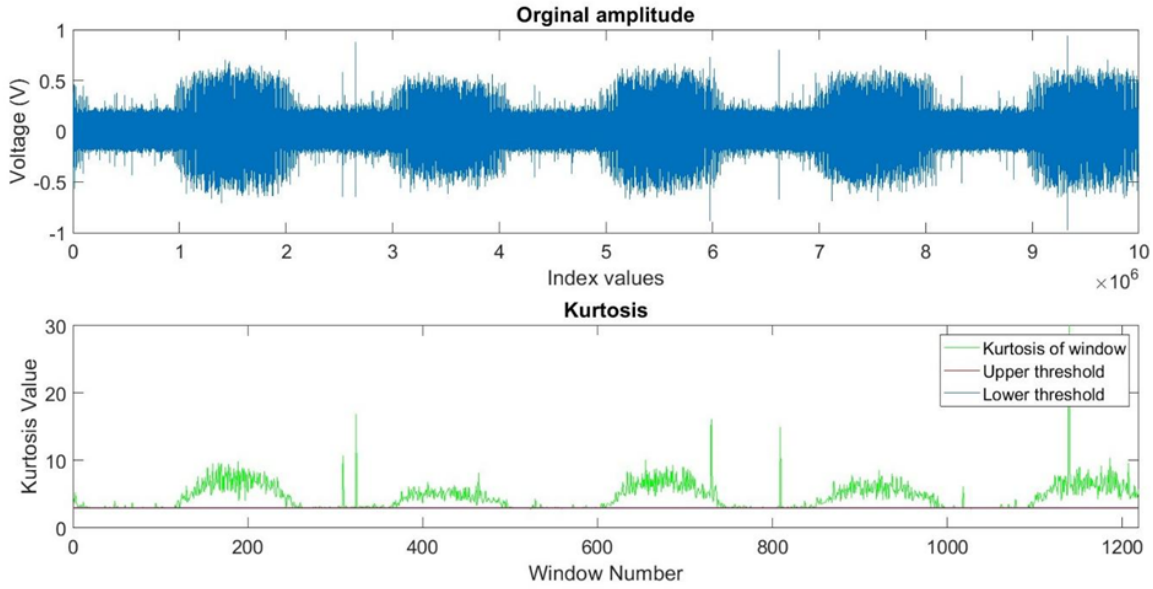


Figure 15: Band 4 - Moderately corrupted by RFI

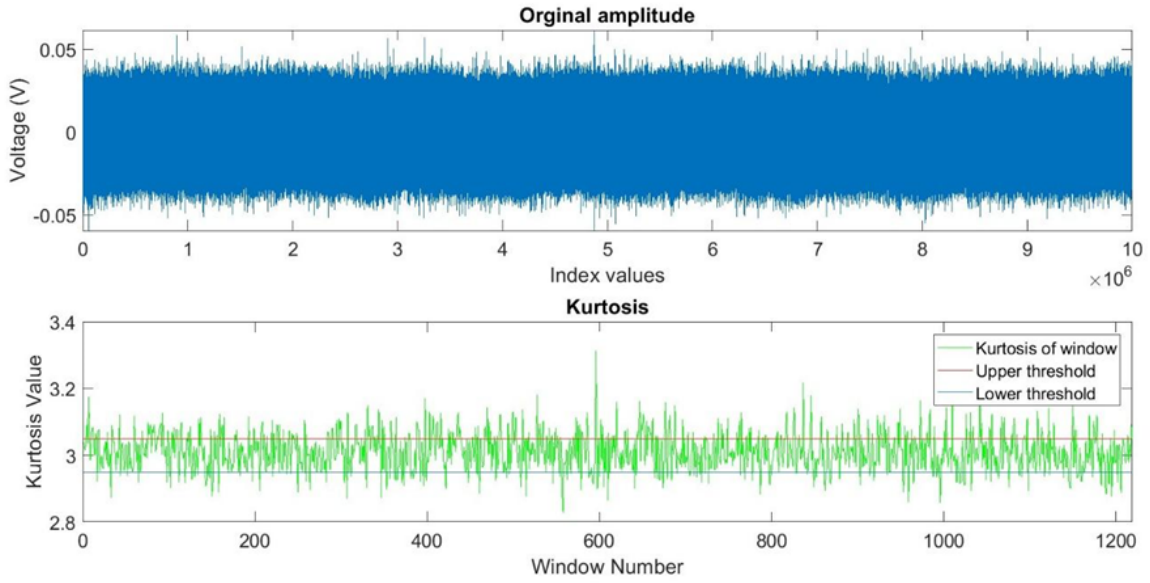


Figure 16: Band 5 - Lightly corrupted by RFI

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 the kurtosis calculated on the window sample is dependent on the nature of the window. If the window

contains Gaussian 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 behavior and variation of kurtosis for a 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 a per cent increase in outliers was studied. Figures 16 and 17 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 visualized by zooming Figure 16 and is shown in Figure 17 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 kurtosis gives false positive values of the data being Gaussian. This limitation of the 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 [14] 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 analyzed so far does not have such a density of outliers. The average density of outliers due to power-line RFI, detected using kurtosis at GMRT is 77.34% (window based). Hence it does not intersect with the kurtosis blind region. This method is explained in the further section. The kurtosis increases if the window contains more outliers and it is equal to 3 if there are a greater number of noise (Gaussian distributed) samples. The following Figure 17 and 18 illustrate this exact behavior.

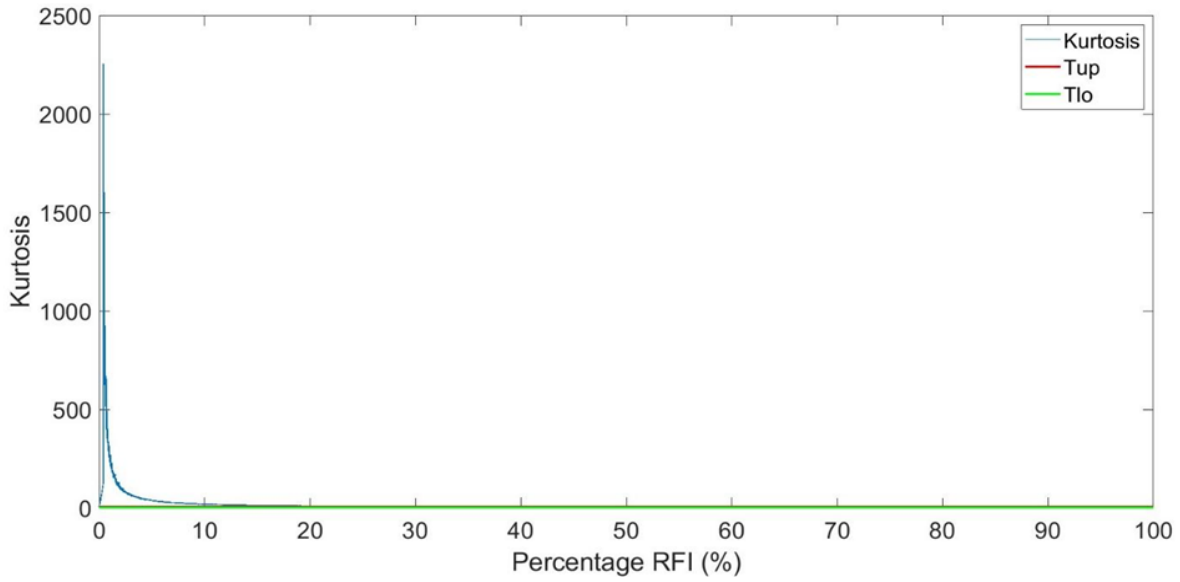


Figure 17: Variation of kurtosis over per cent increase in outliers

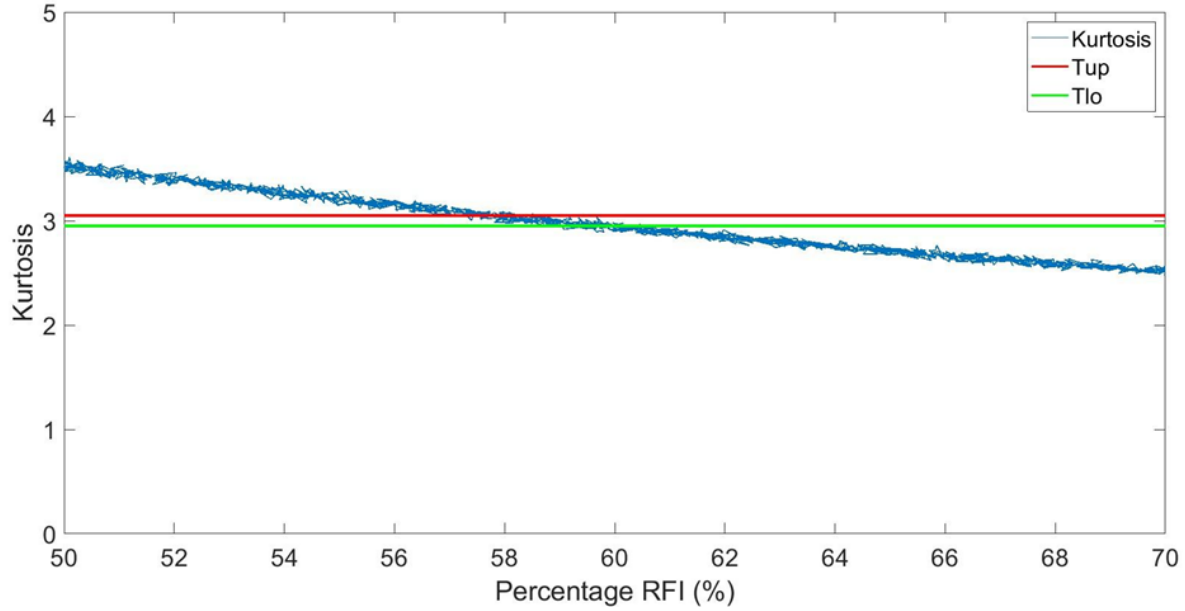


Figure 18: An observed blind spot in kurtosis

3.4 SIXTH MOMENT FOR NON-NORMALITY DETECTION

The sixth moment is the higher-order statistic which is used to measure the overall shape of the distribution. Higher-order moments are usually difficult to estimate and require a large

samples in order to obtain estimates of similar accuracy. The general equation to calculate the statistical moments is given by equation 2.

$$M_k = \frac{E[(X-\mu)]^k}{E[(X-\mu)^2]^{(k/2)}}$$

Where: k = *order of the moment*
X = *sample*
μ = *mean of the sample set*

Equation 2: Statistical moment calculation

$$M_6 = \frac{E[(X-\mu)]^6}{E[(X-\mu)^2]^3}$$

Equation 3: Sixth moment

From mathematical derivations and empirical evidence, it is clear that odd-power higher-order moments give information about the skewness of the distribution and even-power higher-order moments give information about the peaked-ness of the distribution.

The sixth moment recapitulates the information captured by the fourth moment, about the tailedness of the distribution. Higher-order statistics provide valuable insights about the distribution and are extremely useful while detecting non-gaussian datasets. A brief tutorial by Jerry Mendel [17] also states the use of higher-order moments in signal processing.

In order to validate the method proposed by Roger D. Roo and Sidharth Misra [14], the higher statistical moment of sixth order is used over the raw voltage time series to detect powerline RFI. The sixth moment is calculated using the equation[3]. It is considered to be 0 for normal distribution. The sixth moment is computed over every window, each window having 8192 samples. The noise margin for this method is defined with the help of thresholds and is very narrow as compared to kurtosis. The thresholds are defined with the help of equation 4, and for the window size of 8192 samples, the noise margin is from -0.763 to 0.763.

$$Tmup = +\sqrt{\frac{4770}{N}} \qquad Tmlo = -\sqrt{\frac{4770}{N}}$$

Where N = Number of samples inside the window

Equation 4: Estimation error in the sixth moment

As with kurtosis, the sixth moment also has some blind regions, where its value falls within the threshold limit even for a significant density of outliers. Although, as stated by Roger D. Roo and Sidharth Misra [14] and in the previous section, the blind spots of the sixth moment are not only limited to 13% and 61% of outlier density. This fact is verified by the experiment which analyzes different statistical distributions for the sixth moment. The results of this experiment are represented by Figures 19 to 21.

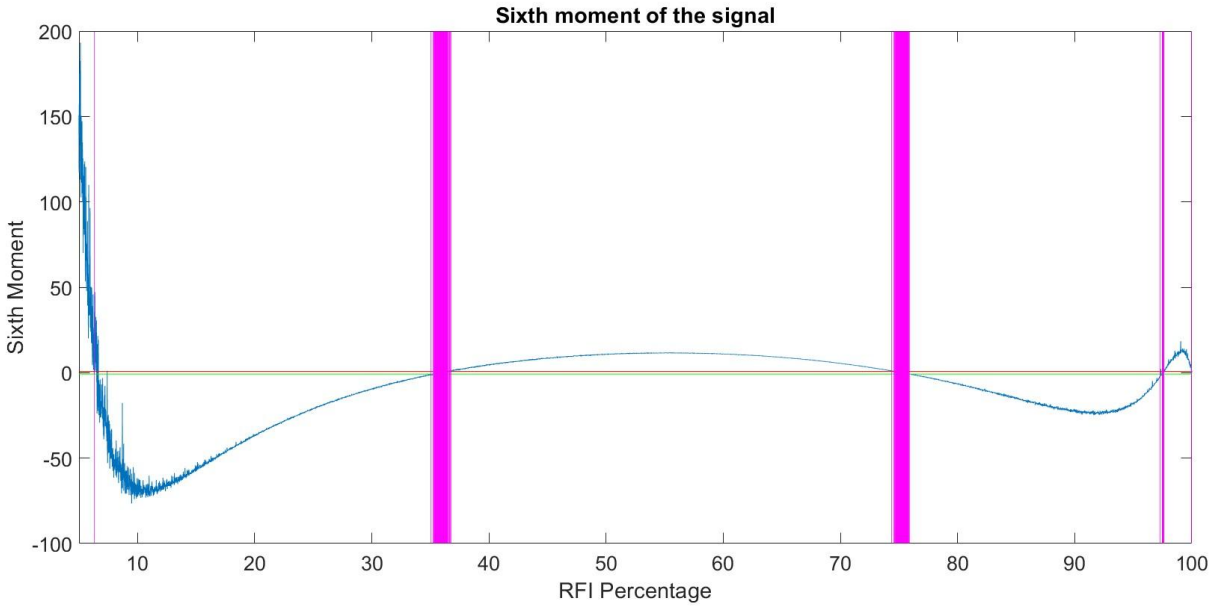


Figure 19: Sixth moment - Chi-square distribution of outliers

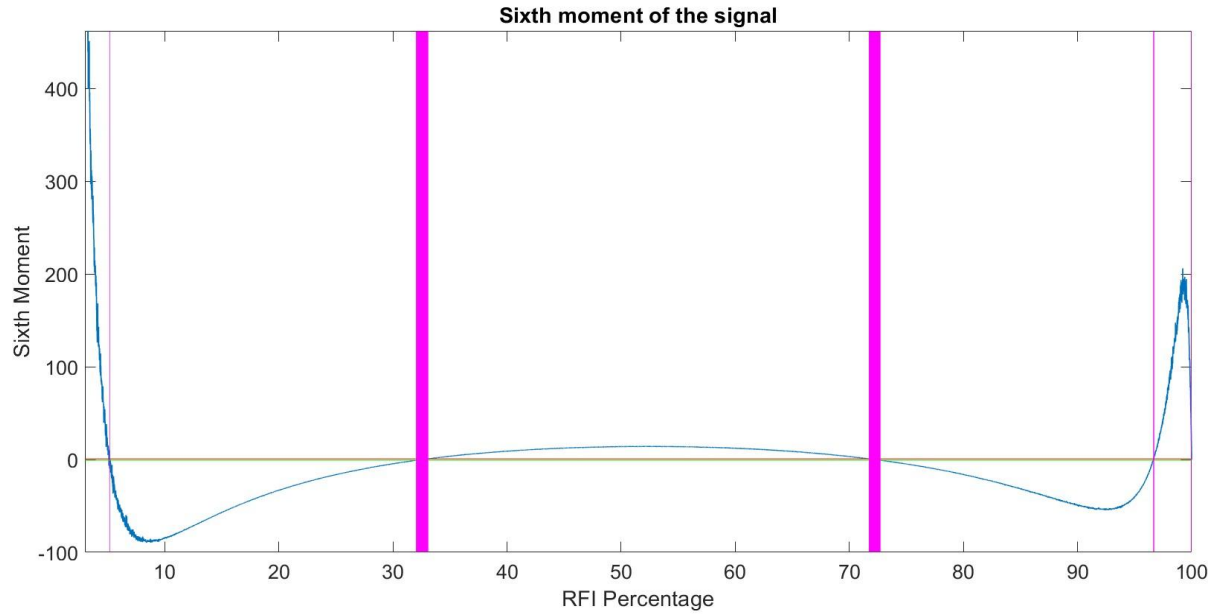


Figure 20: Sixth moment - Poisson distribution of outliers

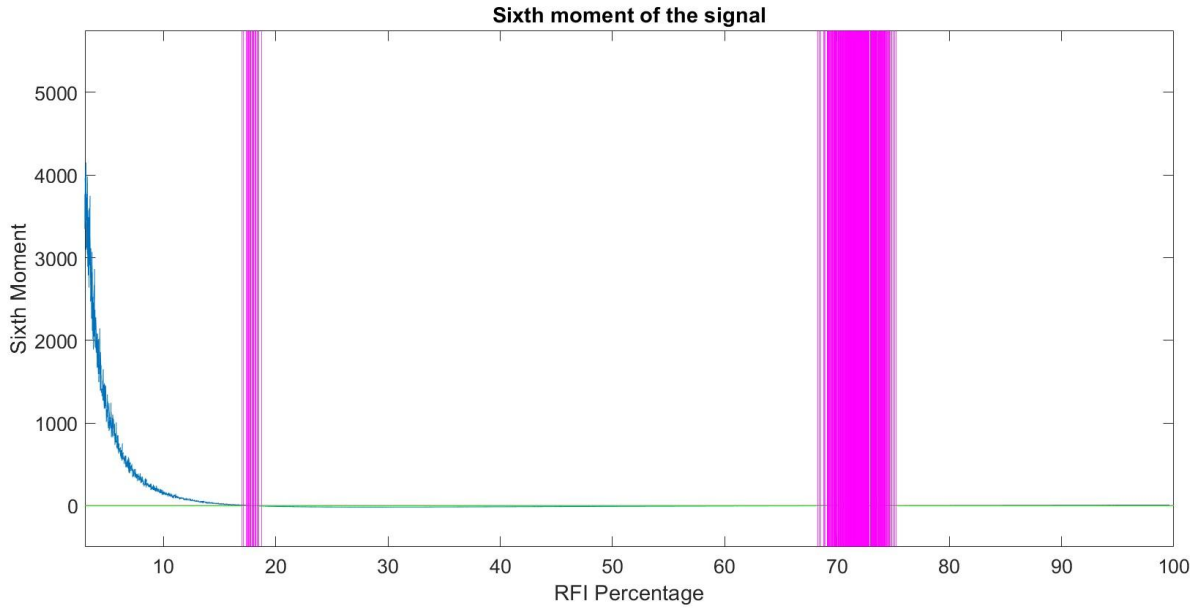


Figure 21: Sixth moment - Uniform distribution of outliers

The above figures illustrate the overall pattern of the sixth moment for increasing the density of outliers. The pink stripes indicate the blind region in the figures. The blind region of the sixth moment is dependent on the nature of the distribution of outliers. As is seen in the figure, there are 3 blind regions for chi-square distribution, 4 for Poisson and 2 for uniform distribution of outliers. This means that if the outliers come from any of the respective distributions, the sixth-moment estimation of the entire window will be blind for non-normality detection at the respective concentration of outliers. However, in any case, the blind regions of the sixth moment do not coincide with the blind spot observed for kurtosis.

Figure 22 illustrates the sixth moment computed for RFI data in Band 2 of GMRT. It can be seen that it has a sharp increase and decrease in the instances where there is RFI and noise respectively. This sudden change and narrow threshold margin is characteristic of the sixth moment.

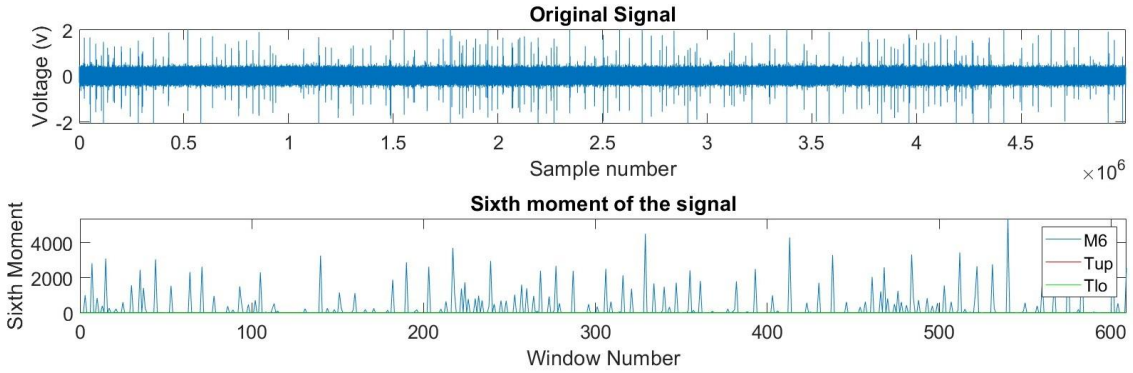


Figure 22: Sixth moment of Band 2

From the extensive study of the sixth moment and its variation with respect to outlier concentration, it can be concluded that it is a better estimate of non-normality, and hence can be used for RFI detection. One of the main reason of using sixth moment over kurtosis is that it can pick up low-level sparks at the edge of the bunch, making it possible to demarcate the bunch boundaries accurately. Figures 23 and 24 can be compared to see the RFI bunch detection using kurtosis and sixth-moment respectively.

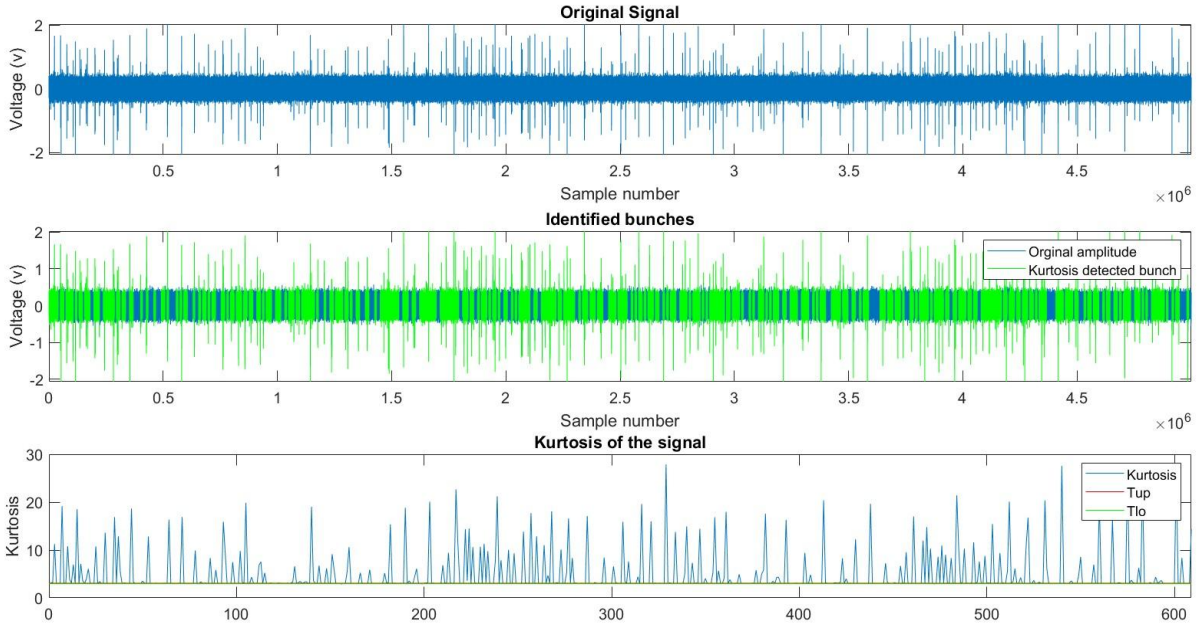


Figure 23: Bunch detection using Kurtosis

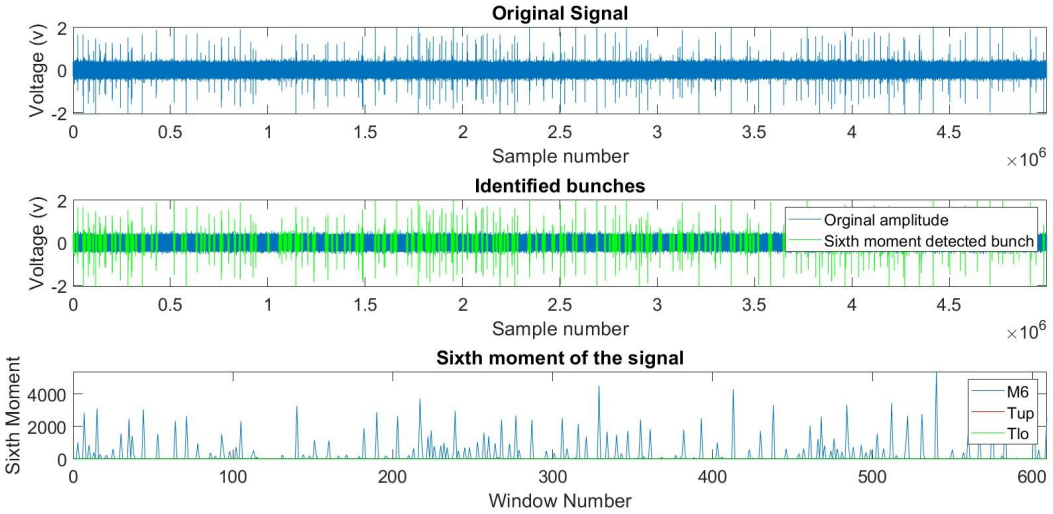


Figure 24: Bunch detection using sixth-moment

3.5 SHAPIRO-WILK TEST FOR NON-NORMALITY

The Shapiro-Wilk test is a test for normality detection, which assumes a null hypothesis is true. The null hypothesis is that the population is normally distributed. The p-value, which gives the probability that the sample is part of the population, considering the null hypothesis is true. If the p-value of that particular sample is less than the alpha level, the null hypothesis is rejected, and there is sufficient evidence that the population is not normal. On the other hand, if the p-value is greater than the chosen alpha level, the null hypothesis is accepted and it can be concluded that the population has normal distribution. The alpha level is the user-defined threshold level for p-values. It is generally taken as 0.05, which indicates that all the samples having more than 5% probability are accepted as samples originating from the normal distribution.

We use the Shapiro-Wilk test as an estimator for RFI and for bunch detection, the results of which are discussed in Chapter 5 of this report.

CHAPTER 4: HARDWARE IMPLEMENTATION OF KURTOSIS-BASED RFI DETECTION SYSTEM

The algorithm based on the statistical characteristics of powerline RFI needs some advanced computational resources to deploy it into the working system. Field Programming Gate Array (FPGA) is one of the best hardware resources supporting various signal processing techniques. FPGA is a programmable integrated circuit which consists of different configurable logic blocks, surrounded by a system of programmable interconnects, and clock sources. FPGAs are reprogrammable, cost-efficient and capable of handling parallel processing.

The FPGA board used for deploying this kurtosis-based algorithm is developed by the University of California Berkeley's CASPER team. The Collaboration of Astronomical Signal Processing and Electronics Research (CASPER) is a group of highly skilled scientists and engineers who produce instruments and resources to aid astronomy research [18]. The CASPER community has developed a set of software libraries and specially designed FPGA boards to synthesize different signal processing techniques. A detailed description of the hardware boards and software libraries developed by CASPER is given in the following sections.

4.1 ROACH BOARD SPECIFICATIONS

The Reconfigurable Open Architecture Computing Hardware (ROACH) is a standalone FPGA Processing Board developed by the CASPER community and manufactured by Digicom Electronics Pvt. Ltd. The Figure 25 below is the snapshot of the actual ROACH board at GMRT, which is used for this project.

ROACH systems acquire data from some of the most powerful radio telescopes in the world and decipher some of the interesting information about our universe using sophisticated signal processing techniques.

The present hardware processing platform at GMRT, ROACH - 1 system has a Xilinx Virtex 5 XC5VLX110T FPGA with 6 serial and differential pair connectors, 16 GPIOs, and a

separate ethernet port and operates on a standard 12V DC power supply [18]. The entire ROACH system consists of an analogue-to-digital converter and a switched-mode power supply to regulate the input voltages. The detailed specifications of the ROACH - 1 board are listed below, as cited on the official documentation and datasheets of CASPER [18].

Product Name	ROACH 1
FPGA	Xilinx Virtex - 5 XC5VLX110T-1FF1136
FPGA Memory	2x 2M x 18-bit QDRII+ SRAMs
	1x DDR2 DRAM DIMM
FPGA Interfaces	2x Z-DOK+ 40 differential pair connectors
	4x CX4 10Gbps high-speed serial connectors
	1x QSH 40 differential pair connector
	16x GPIO
	4x SMA IO (2x clock-capable)
PowerPC Memory	16KB of on-chip SRAM.
PowerPC External Communication	Two Ethernet 10/100/1000Mbps half- or fullduplex interfaces.
	External peripheral bus (32-bit data) for up to six devices with external mastering
CPU Interfaces	1x RS232 DB9 serial port
	1x 10/100/1000Mbit RJ45 Ethernet
	1x USB2.0
	1x MMC/SD card socket
Power	Voltages of 12V, 5V, 3.3V, 2.5V, 1.8V, 1.5V, 1V and 1.2V aux rails.

Table 2: FPGA Specifications

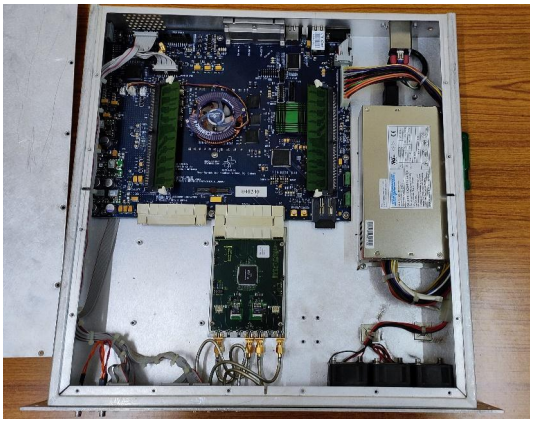


Figure 25: ROACH-1 board at GMRT

4.2 DESIGN OF THE ALGORITHM

Kurtosis is a statistical estimator which determines the peak-ness of the distribution. The distribution confirms a Gaussian distribution if the kurtosis value is within 3 +/- sqrt(24/N), where N is the number of samples used for estimating kurtosis. Powerline RFI results in a heavy-tailed Gaussian distribution, thereby not confirming the kurtosis value for a Gaussian distribution. In order to implement the RFI detection algorithm using kurtosis as an estimator on available FPGA hardware systems, it is necessary to understand some of the fundamentals of binary arithmetic. The algorithm developed by K. Buch [20] simplifies the kurtosis estimation by avoiding the binary division process. Binary division and multiplication of 8-bit signed and unsigned numbers cause an exponential bit growth at every step, consuming the hardware memory. The input analogue samples recorded by GMRT antennas are digitized at a 5ns sampling frequency, by the ADC present on the ROACH system. These digitized samples are in the form of 8-bit signed binary numbers and are stored in Read Only Memory (ROM) of the ROACH – 1 board. These samples are further processed by the FPGA board to estimate the fourth moment, and its detailed process is explained in the following sections.

Computers and digital electronics like FPGA, use binary number systems by converting the input data from decimal to a binary system. Binary multiplication has the worst-case time complexity of $O(n^2)$ where n is the number of bits. A detailed study conducted by R. Brent and H. Kung [21] asserts the fact that multiplication is harder and more time-consuming than addition in VLSI and chip design technology. The binary division is much more complicated since it involves periodic left shifting of bits, depending on the divisor. Each binary arithmetic operation results in bit growth due to extra carry and borrow bits.

4.2.1 Simplified kurtosis

In order to RFI in the observed data from GMRT using CASPER's ROACH - 1 board, the simplified kurtosis approach as suggested by K. Buch [20] is used. Kurtosis is computed on individual windows of 8192 samples. If the value of kurtosis is within the thresholds as defined by equation [1], the window is classified as noise. If the window is corrupted by powerline RFI, the kurtosis is outside the threshold limit and is flagged as RFI.

Conventionally, the kurtosis K is calculated using the equation [4] and the bunch detection algorithm checks for the condition given by the equation [5].

$$(K + Tup) * \left[E(X - \mu)^2\right]^2 > E(X - \mu)^4 > \left[E(X - \mu)^2\right]^2 * (K - Tlo)$$

Equation 5: Inequality condition for the fourth moment

The expectation operation involves summation of all samples, followed by the division with the total number of samples in the window. This repeated addition and multiplication of numbers causes an exponential bit growth. Towards the end of the computation, the dividend $E(X-\mu)^4$ and the divisor $[E(X-\mu)^2]^2$ is greater than 16 bits and thus consumes a lot of time and area. The simplified version of kurtosis avoids this division of binary numbers at the later stage, which results in significant resource savings on FPGA. To make the computation simpler, the mean of window μ is assumed to be 0 and is replaced by integer constant 0.

The block diagram of the model designed using this algorithm and simulated using Simulink libraries are represented in Figure 26.

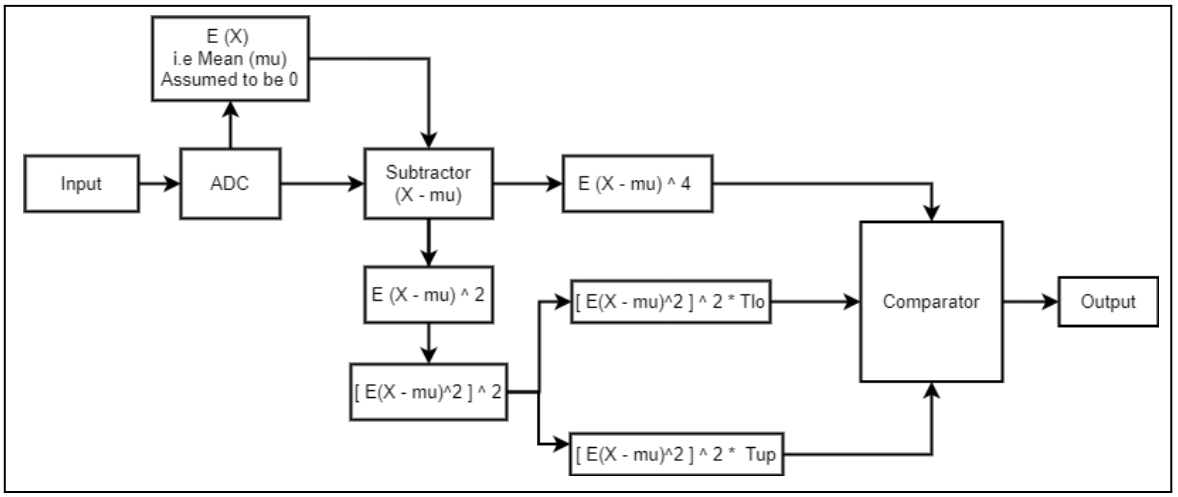


Figure 26: Block diagram of the model

4.2.2 Overview of System Generator library

FPGAs consist of several configurable logic blocks connected via programmable interfaces. For programming different AND, OR, and XOR logic gates inside the CLB of FPGA, different hardware description languages are used. Many hardware description languages like Verilog or VHDL allow the designer to configure specific ports of logic gates and build the digital circuit design, however, they lack the support of a graphical user interface for model-based design. For implementing a model-based design approach, MATLAB's Simulink, and Xilinx have provided an HDL library called System Generator.

The CASPER ROACH boards can be reprogrammed using MATLAB and Simulink's graphical interface. The System Generator library is designed using some primitive digital logic elements and is a higher-level tool which can be easily used to program and synthesize FPGA circuits. This GUI support of Simulink abstracts low-level hardware and simplifies the hardware designing process for non-specialists. The System Generator library contains adder, subtracter, multiplier, and accumulator blocks usually in gray or yellow. These blocks automatically generate a low-level Hardware Description Language (HDL) code when the design is compiled in order to synthesize the FPGA. Using these blocks, users can easily simulate different designs by integrating them with Simulink blocks.

The simplified kurtosis algorithm, implemented on the CASPER's ROACH – 1 system is modeled using the System Generator library in the Simulink environment.

4.3 SIMULINK MODEL

The algorithm explained in section 1.2 is designed in Simulink and is represented in Figure 27. The counter and the ROM blocks are used to store the input data in the read-only memory of the ROACH board. The flow of the logic of this model is as per the block diagram given in Figure 26.

The window control subsystem divides the incoming bit stream of samples into blocks of 8192 samples, known as windows. This subsystem generates a trigger pulse after every 8192 samples, which is then used to reset the accumulator in the further part of the design. This subsystem is visualized in Figure 28.

The expectation operation subsystem represented by Figure 26 consists of an accumulator block. The accumulator performs continuous addition of incoming bits until its reset port is triggered. Accumulator is followed by a shift block which performs a binary left shift operation. Binary division operation can be realized by serially shifting the bits of the dividend by M positions, where the divisor = 2^{M}. Hence, for the window size of 8192 samples, the accumulator output is left shifted by 13 bits. The 2:1 multiplexer is used to hold the sample value till the end of the window cycle.

The comparator subsystem uses relation blocks to compare the fourth moment with the threshold range and gives 1 as the output flag if the input window is corrupted and contains RFI. The output flag is 0 if the window contains noise and lies within the threshold range. This output is stored in the software register of the ROACH – 1 board and read out on a computer. The comparator subsystem is represented in Figure 30.

In addition to the functional blocks, delay blocks are added after every stage in the model. These are required to configure the model according to the system generator's clock frequency. The delay blocks provide the required latency before every functional block, in order to match the 100MHz system clock frequency. This is the standard clock frequency used for ROACH – 1 board, but it can be altered using the system generator token 'XSG core config'. The system generator block and its token are used to set the target FPGA hardware and synthesize the model into low-level HDL code.

Every Simulink System Generator model needs these tokens in order to compile successfully. The final delay-corrected design is given in Figure 31.

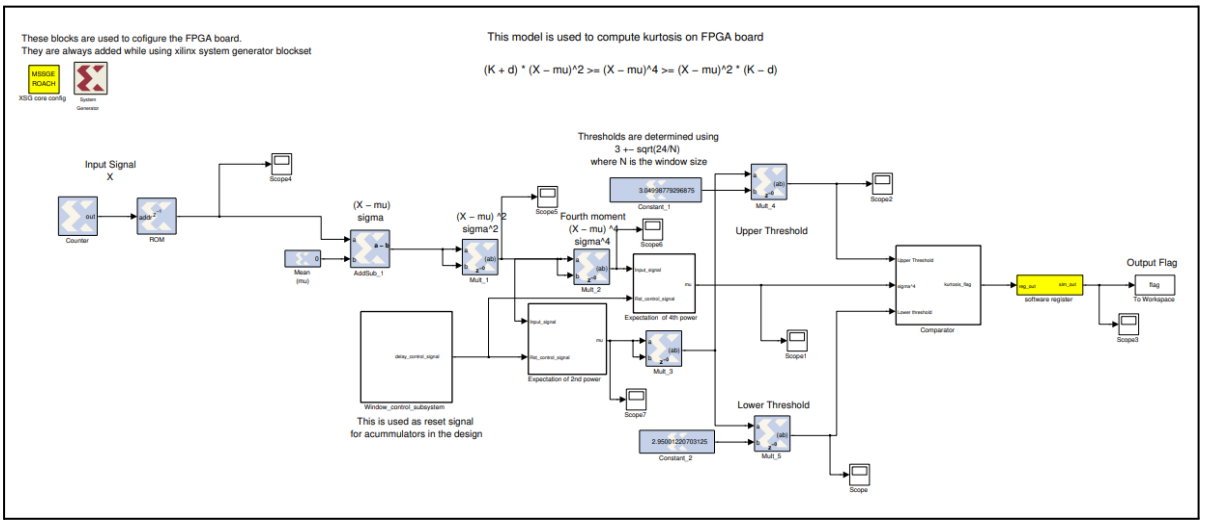


Figure 27: Simulink model of simplified kurtosis algorithm

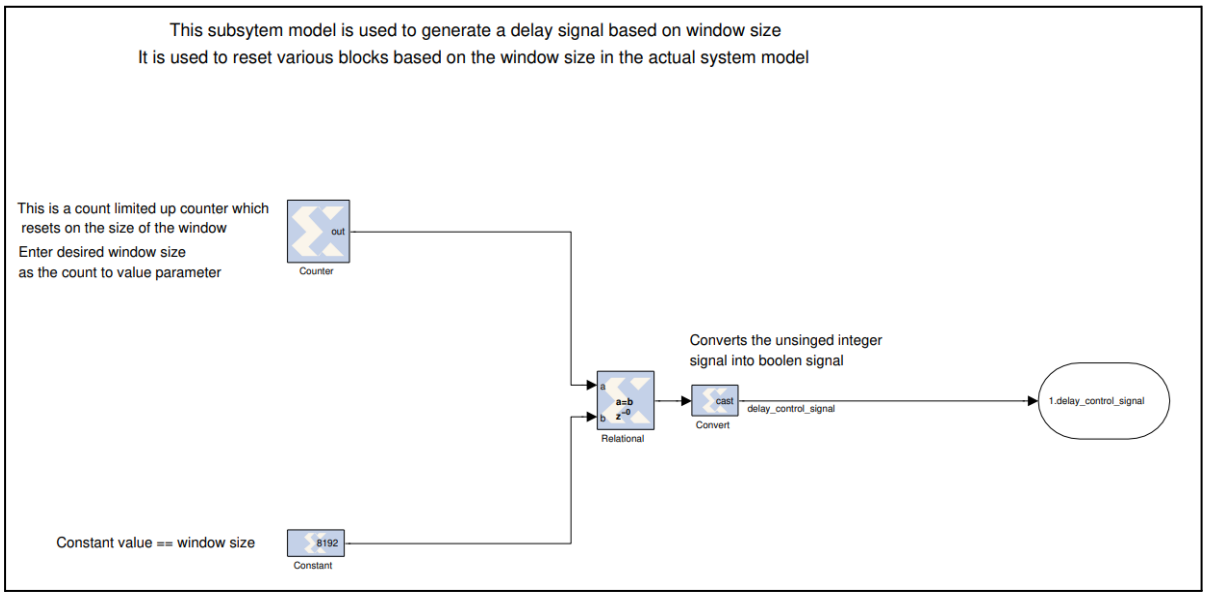


Figure 28: Window control subsystem

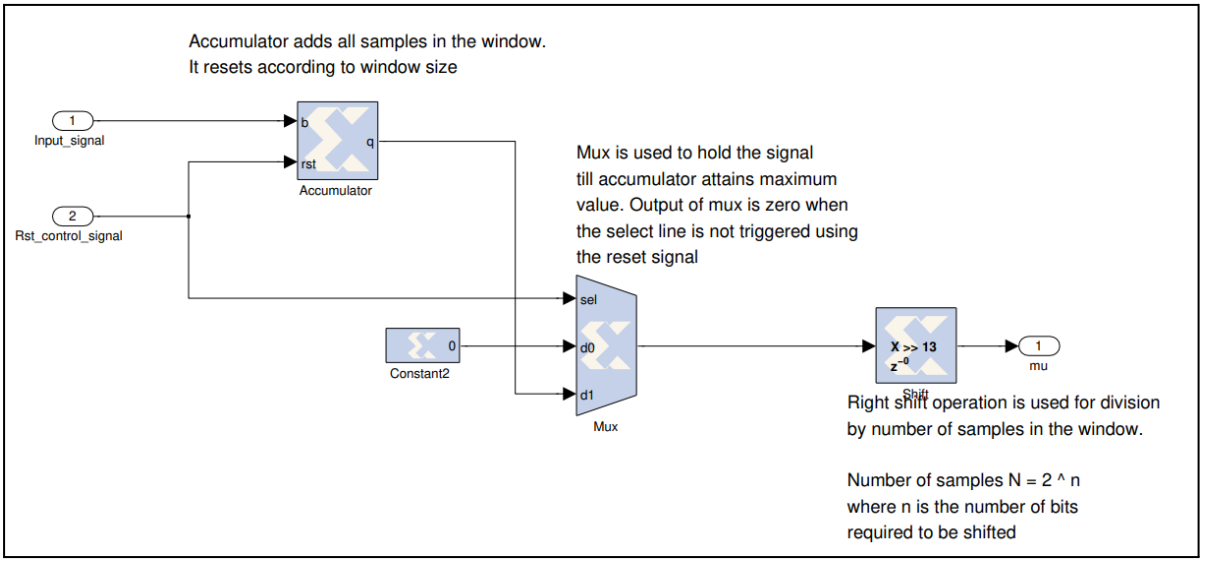


Figure 29: Expectation operation subsystem

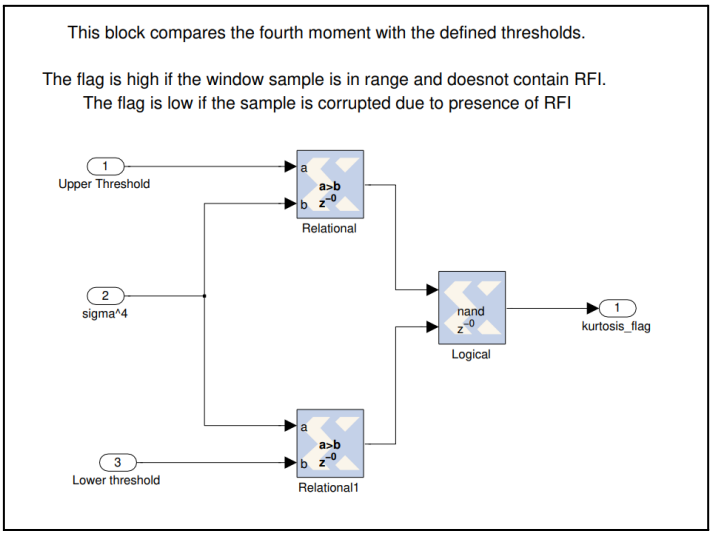


Figure 30: Comparator subsystem

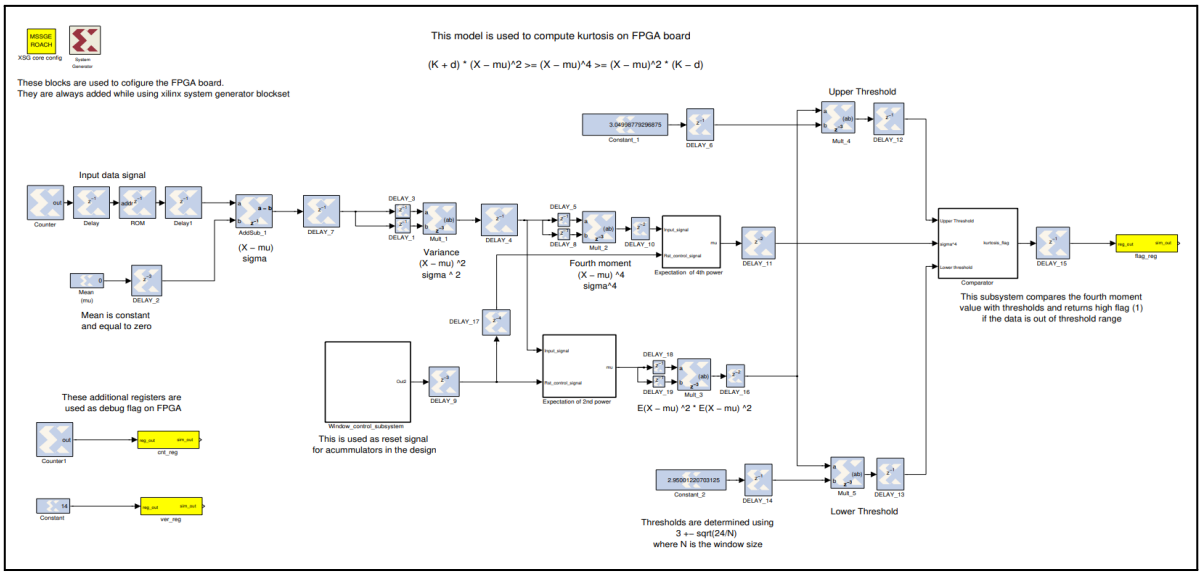


Figure 31: Delay corrected model

4.4 SIMULATION RESULTS

The kurtosis algorithm is used to detect RFI using CASPER's ROACH board. Simulations are carried out on different datasets using the Simulink model described in section 1.3. Some of the results from different observation bands are given in Figures 32 to 35. These simulations are performed using MATLAB code. This code simultaneously analyzes the input data using MATLAB commands as well as the Simulink model. Finally, the output flags from both computational methods are compared and visualized. It can be seen that both computation methods coherently detect RFI windows. Flag 1 indicates that the window is corrupted by RFI and flag 0 indicates that the window contains a noise signal.

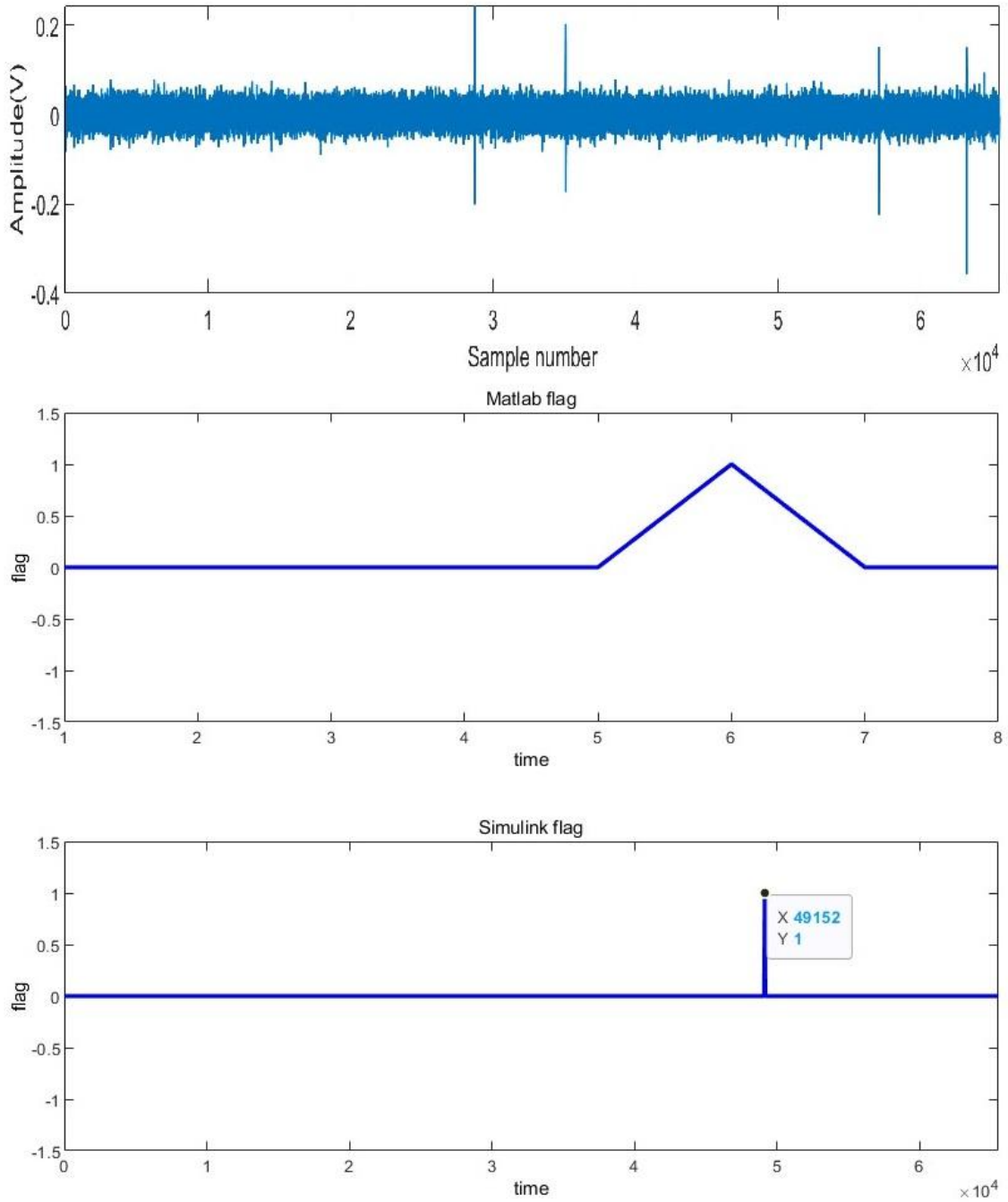


Figure 32: RFI in Band 2

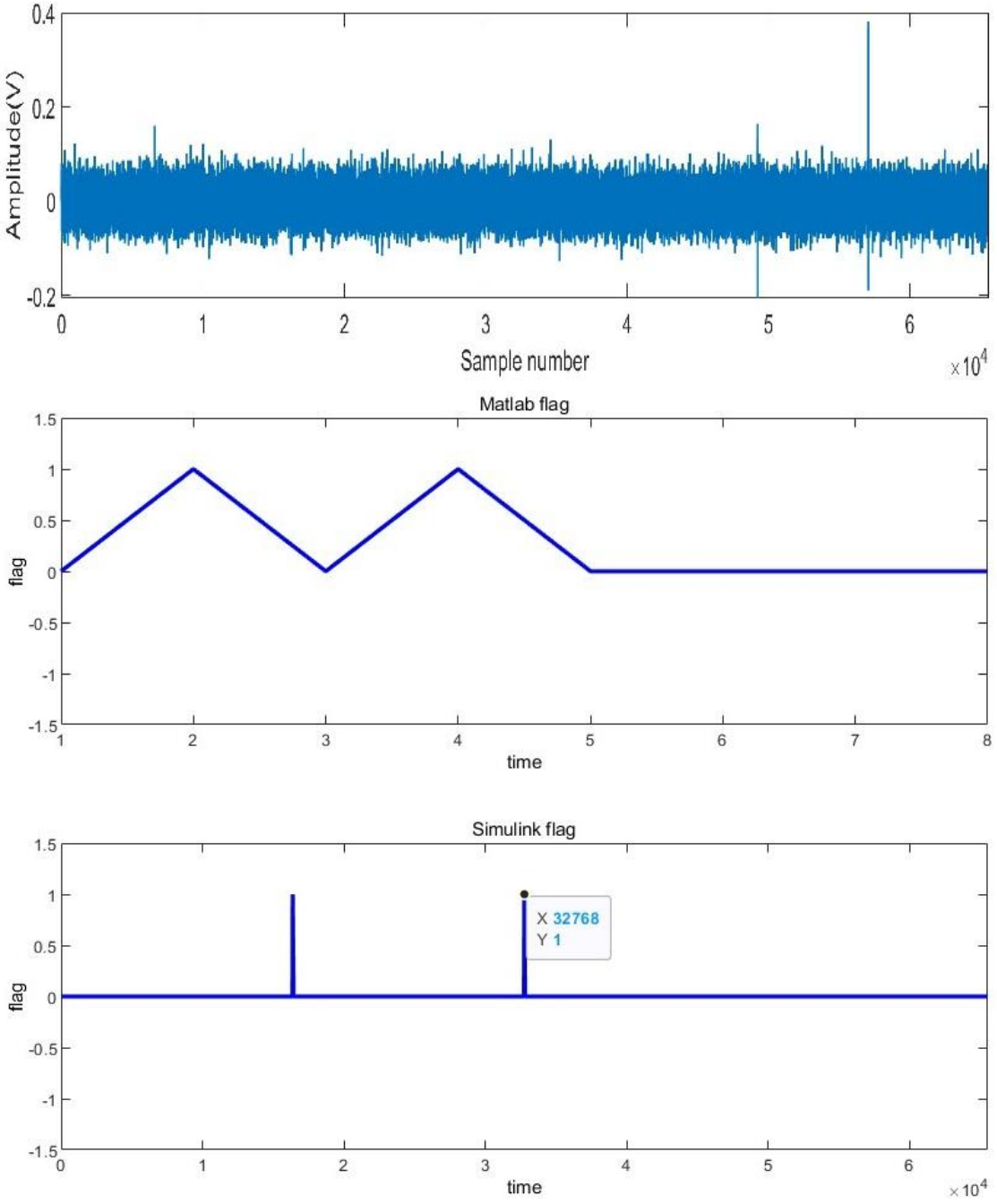


Figure 33: RFI in Band 3

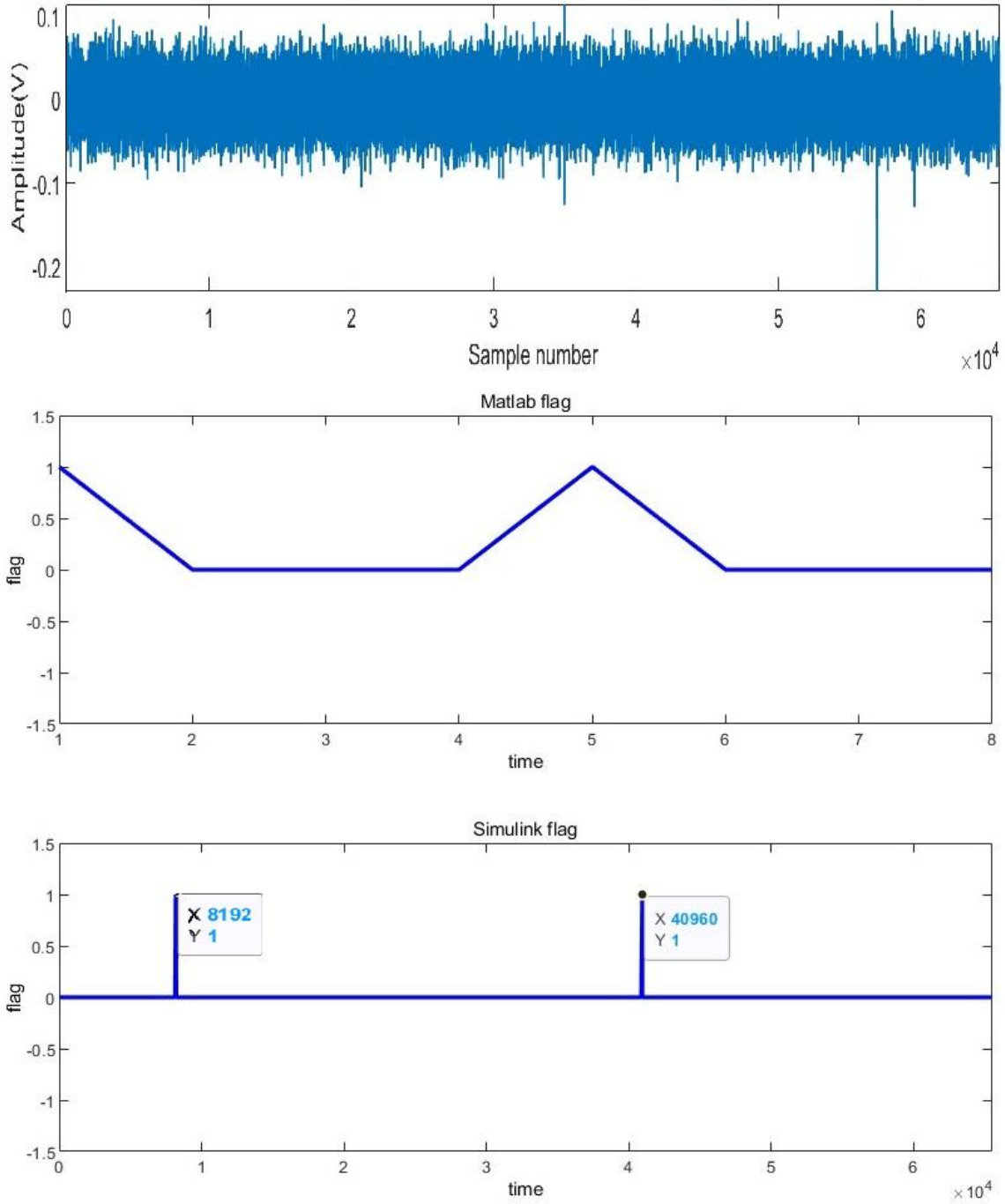


Figure 34: RFI in Band 4

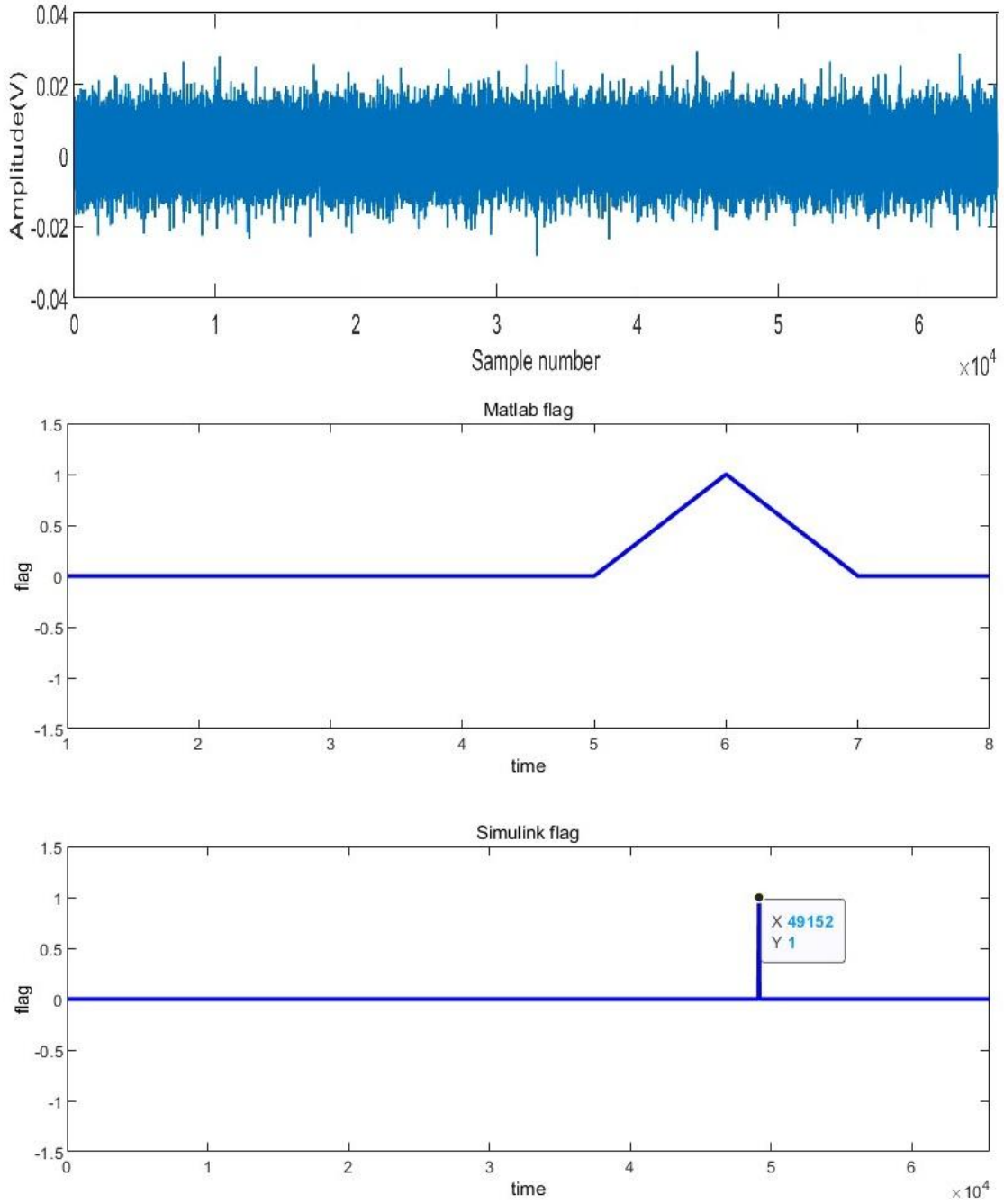


Figure 35: RFI in Band 5

4.5 ROACH BOARD IMPLEMENTATION RESULTS

The Simulink Model described in section 4.3 is implemented on the CASPER's ROACH - 1 board at GMRT. The delay-corrected Simulink model, Figure 31, is first compiled into a low-language design. The compilation is done in two steps, the first step involves Xilinx's System Generator Library, which compiles the Xilinx blocks in the design into the corresponding Hardware Description Language. The second step involves the synthesis of the design through Vivado, which translates the design for physical implementation on the FPGA board. After successful compilation, a final *.bof* file is generated which includes all the information for configuring the FPGA as well as the meta-data containing information about the registers and blocks used in the design [22].

Figure 36 shows the actual working system of ROACH boards in the receiver room of GMRT. These ROACH boards form an independent computational node, called the snode at GMRT.

Figure 36: snode server inside the receiver room

The *.bof* file is executed within the BORPH operating system of the ROACH board. The Berkeley Operating System for ReProgrammable Hardware (BORPH) is an OS specially designed for FPGA-based reconfigurable computers. BORPH is an extended version of the Linux Kernel which handles the FPGA as CPU and makes it easy to deploy the hardware design and run it just like a normal user program.

Figure 37 describes the communication of ROACH board with the host computer, via different communication protocols. The Xilinx FPGA on the ROACH board is connected to its Power PC through a hardware link. The ROACH is physically connected to the hoast computer using 1Gb Ethernet link.

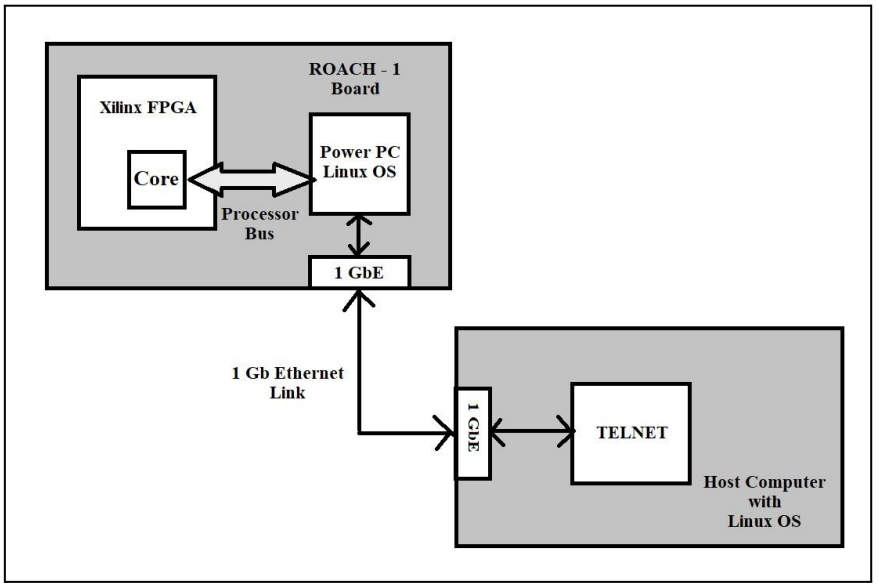


Figure 37: ROACH Communication

To run the design on the FPGA, The ROACH is programmed using a set of Python commands which install and initialize some of the necessary packages. The python library used for controlling ROACH's FPGA is known as *Corr* and needs python 2.5 or later versions. Once the ROACH is successfully programmed, the *.bof* file is copied to the local machine and executed using the *progdev* command.

The first trial run of the model was carried out by logging into the snode machine at GMRT. The snode is a local computer inside the firewall of GMRT, which can be accessed remotely using the TELNET protocol. Figure 39 represents the setup of the ROACH board and its host computer. This setup is located inside an RFI-shielded environment, at the E-lab of GMRT. The results of the hardware experiment are shown in Figures 41 and 42.
The timing and resource utilization reports for this design are provided in Figure 38 and 39.

```
Derived Constraint Report
Derived Constraints for TS_sys_clk_n
+-------------------------------+-------------+-------------+-------------+-------------+-------------+-------------+-------------+
|                               |   Period    |       Actual Period       |      Timing Errors        |      Paths Analyzed       |
|           Constraint          | Requirement |-------------+-------------|-------------+-------------|-------------+-------------|
|                               |             |   Direct    | Derivative  |   Direct    | Derivative  |   Direct    | Derivative  |
+-------------------------------+-------------+-------------+-------------+-------------+-------------+-------------+-------------+
|TS_sys_clk_n                   |     10.000ns|      8.332ns|      8.096ns|            0|            0|            0|      7854466|
| TS_infrastructure_inst_infrast|     10.000ns|      8.096ns|          N/A|            0|            0|      7854466|            0|
| ructure_inst_sys_clk_dcm      |             |             |             |             |             |             |             |
+-------------------------------+-------------+-------------+-------------+-------------+-------------+-------------+-------------+
```

Figure 38: Timing report for Simulink model

```
Device Utilization Summary:

   Number of BUFGs                          3 out of 32      9%
   Number of DCM_ADVs                       2 out of 12     16%
   Number of DSP48Es                       18 out of 640     2%
   Number of ILOGICs                       45 out of 800     5%
   Number of External IOBs                 52 out of 640     8%
      Number of LOCed IOBs                 52 out of 52    100%

   Number of OLOGICs                       17 out of 800     2%
   Number of RAMB36_EXPs                    2 out of 244     1%
   Number of Slice Registers             1902 out of 58880   3%
      Number used as Flip Flops          1902
      Number used as Latches                0
      Number used as LatchThrus             0

   Number of Slice LUTS                  1244 out of 58880   2%
   Number of Slice LUT-Flip Flop pairs   2166 out of 58880   3%


Overall effort level (-ol):   High
Router effort level (-rl):    High

Starting initial Timing Analysis.  REAL time: 8 secs
Finished initial Timing Analysis.  REAL time: 9 secs
```

Figure 39: System utilization report

Sr. No	Register Name	Function
1	ver_reg	Preserve the version number of the design
2	flag_reg	Used to store boolean output flag about the input data. if the input data contains noise, the flag register has value 0x0 and if it contains RFI, the flag register has value 0x1
3	cnt_reg	Stores the value from free running 32 bit counter and is used for parity checking

Table 3: Functions of Registers

Table 3 lists the functions of different software registers used in the Simulink model.

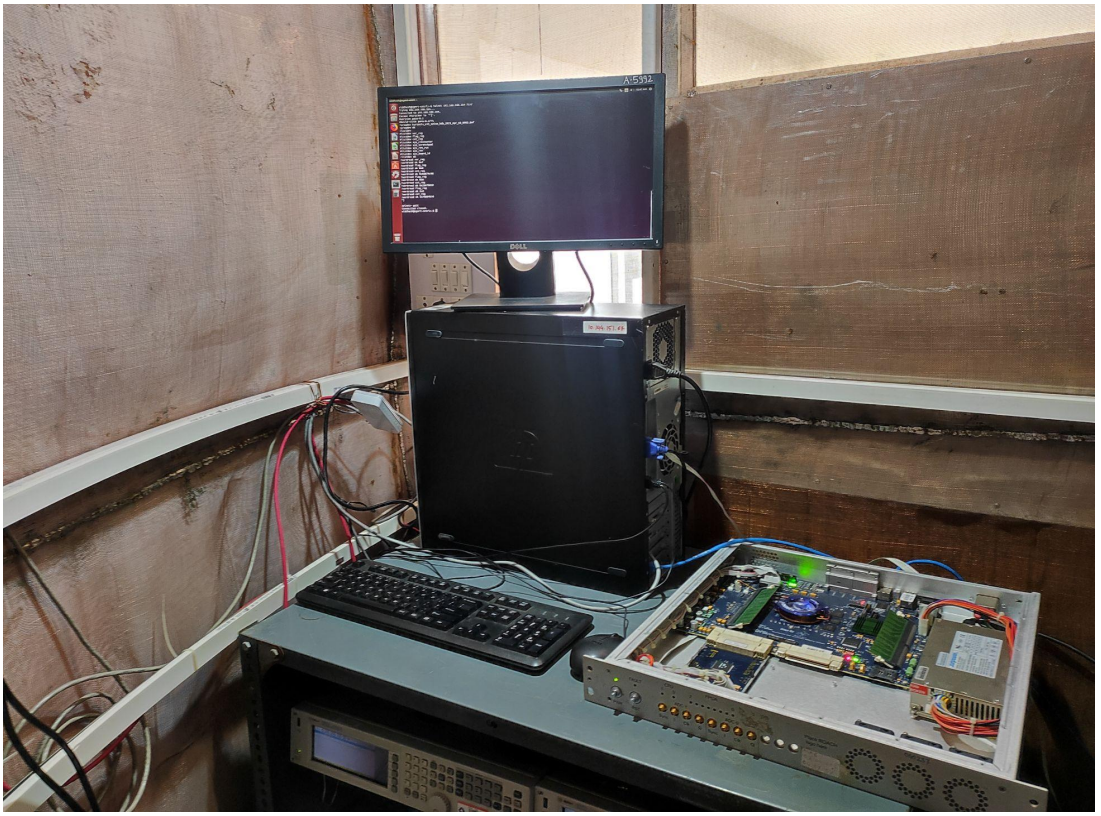


Figure 40: ROACH Setup at GMRT

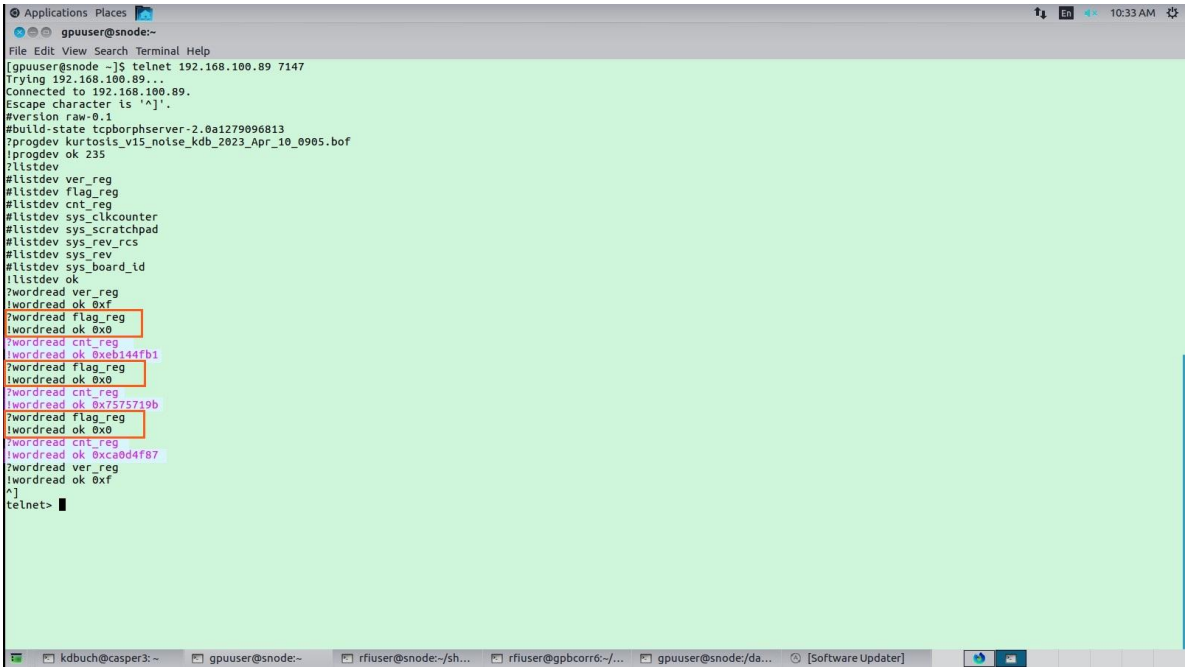


Figure 41 : ROACH output for the noise model

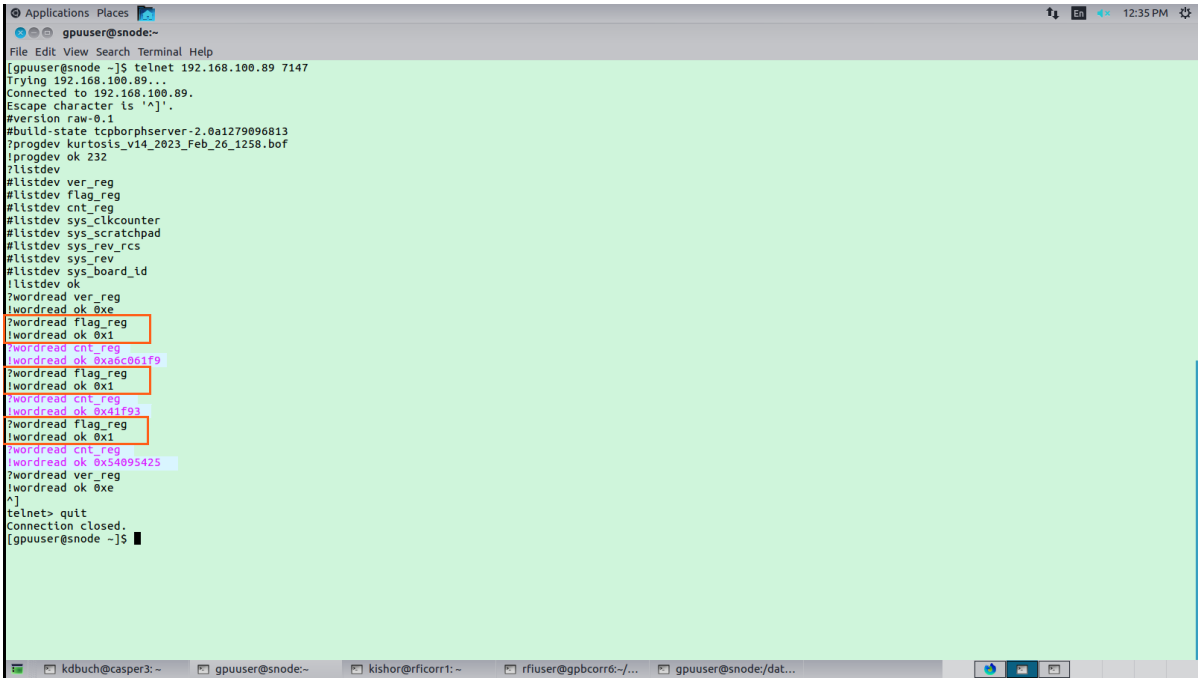


Figure 42: ROACH output for the RFI model

While compiling, the ROM can store only one data window of 8192 samples. Hence, this hardware platform can process only the first window of the input data. The model can be redesigned and optimized for incorporating the continuous time domain data, but currently, it operates only a block of pre-stored data.

The delay-corrected model contains three software registers. Each of the software registers is read out from the command line control of the host computer. The first register, called *ver_reg*, maintains the version number of the running model. This prints out the hexadecimal value of the version which is configured while designing the model.

The second register, *flag_reg* holds the output flag, which determines if the dataset contains RFI or not. The difference in the values of *flag_reg* can be seen in Figures 35 and 36. The first model is synthesized with the dataset having no RFI, whereas the second model is synthesized with the dataset containing RFI. A high flag (0x1) indicates the presence of RFI and a low flag (0x0) indicates that the first few samples of the dataset are not corrupted with RFI.

The third software register *cnt_reg* is connected to a free-running 32-bit counter, which increments its value for every clock cycle. This register is used for parity checking and ensuring that the communication between the host server and the FPGA board is live and accurate. At every instance, the value of *cnt_reg* increments and this value can be seen on the command line by using the wordread command. Table 4 lists all the functions of the commands used for reading these registers.

Sr. No	Command Name	Syntax	Function
1	progdev	?progdev <my_borph_file.bof>	Programs the FPGA with a BORPH bitstream
2	listdev	?listdev [size]	Displays available device registers. Only useful if the FPGA is programmed.
3	wordread	?wordread <reg_name>	Reads data words from a named register.

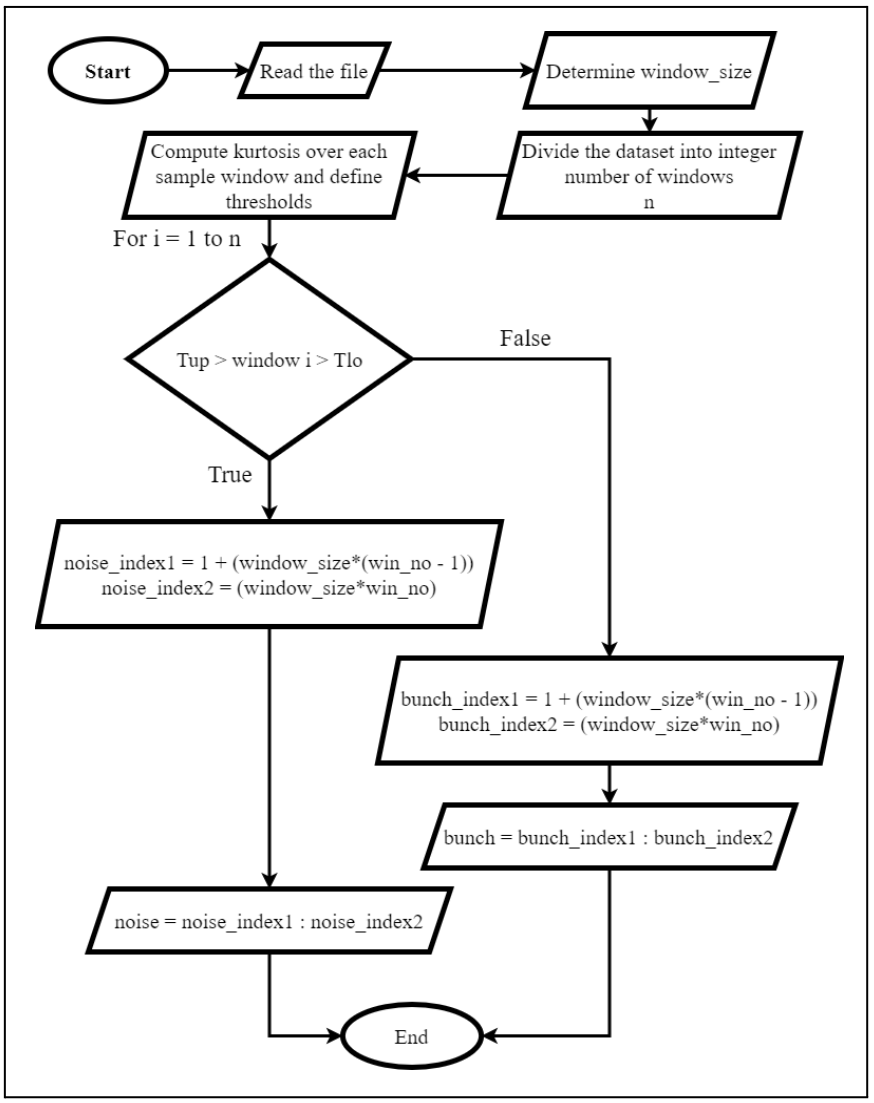
Table 4: List of commands used for interacting with the ROACH board

CHAPTER 5: RFI BUNCH DETECTION

5.1 BUNCH DETECTION USING KURTOSIS

From the experiments and nature of spectral kurtosis studied in the previous chapters, an algorithm was designed that detects an 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 3.



Flowchart 3: RFI and Noise 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.

Figures 43 to 45 are represented below which have RFI bunches automatically detected by the algorithm.

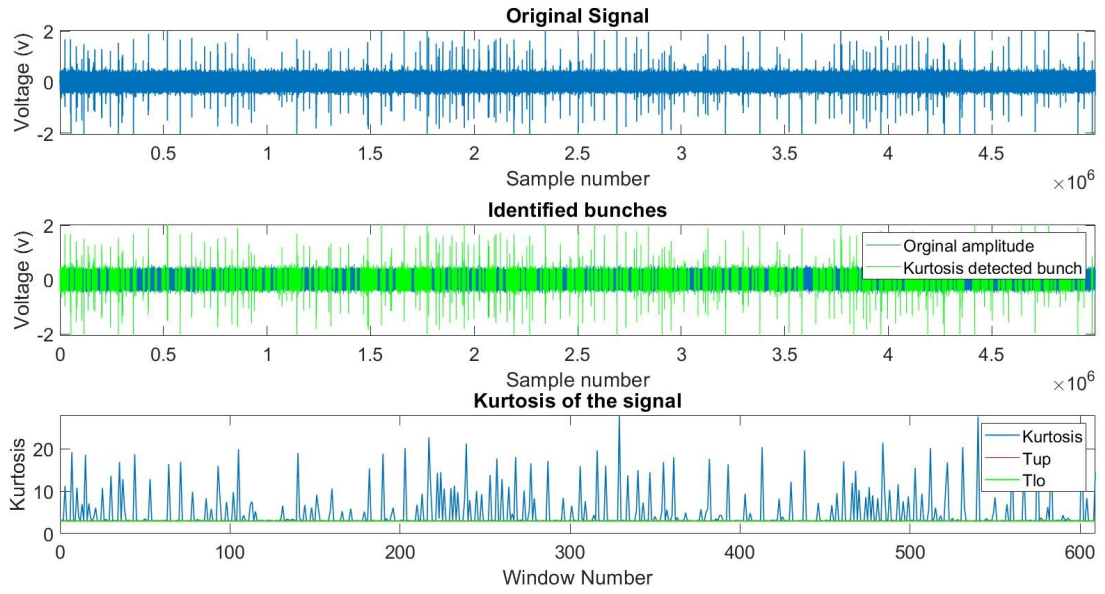


Figure 43: RFI in Band 2

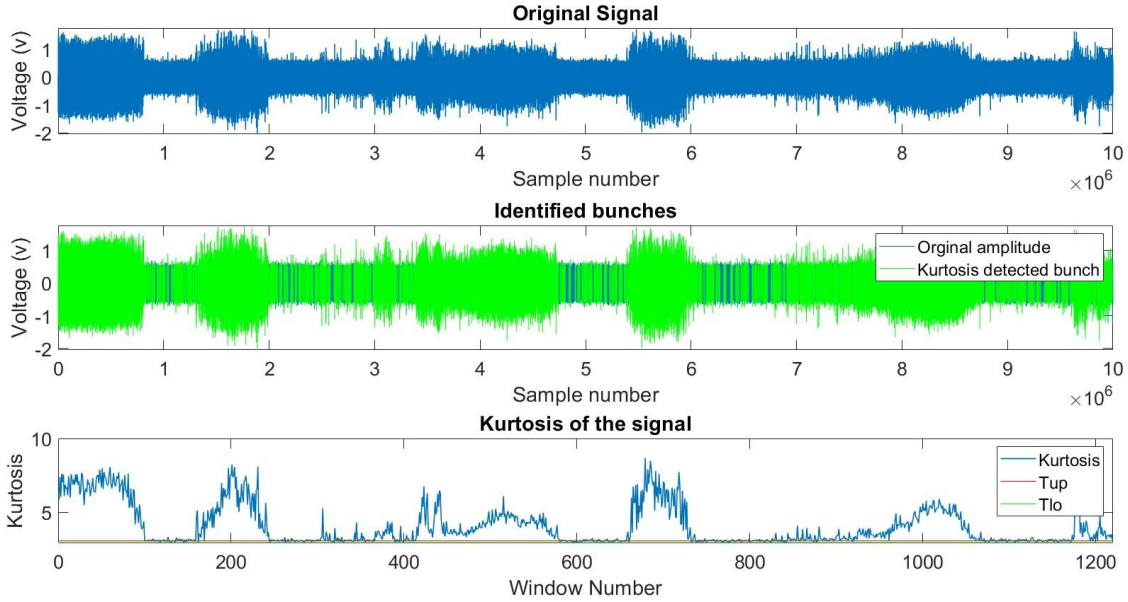


Figure 44: RFI in Band 3

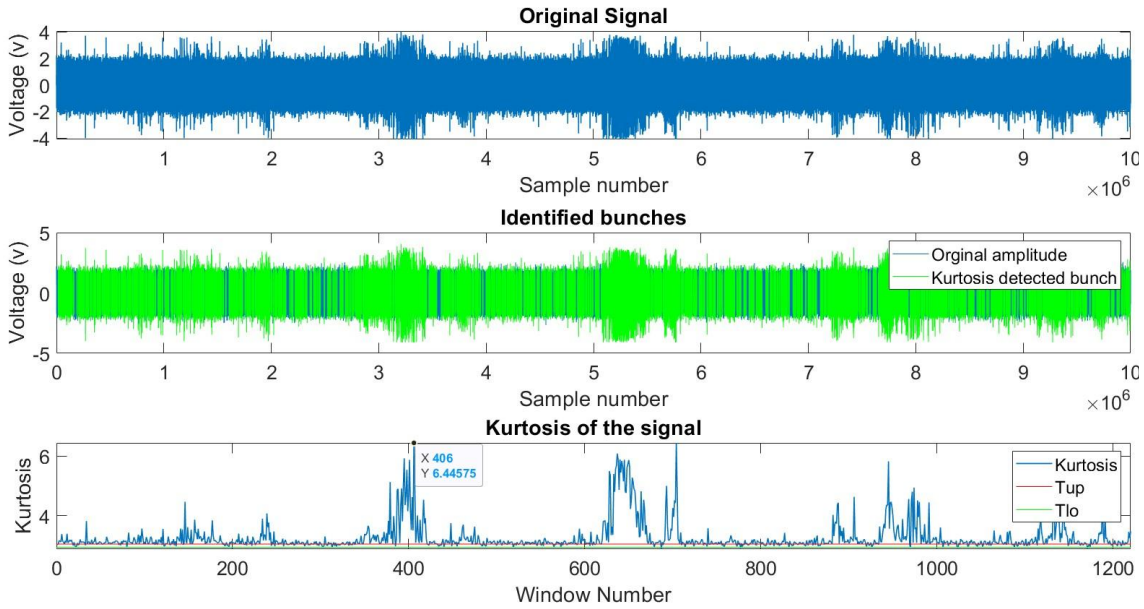


Figure 45: RFI in Band 4

The designed algorithm is highly dependent on the window size, which is explained as follows. 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 a 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, and the thresholds are defined. Other than this, the window size is also decided based on the typical density of RFI present at a given sampling rate. Kurtosis is less efficient as an RFI estimator if the sample size is too large. From the relation of the number of outliers and kurtosis present in the dataset, it is clearly known that kurtosis has 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 samples.

However, this increases the computation speed requirement of the machine, and individual sparks, having values between 3σ and 4σ are falsely modeled as a bunch. To avoid this some

approximation is needed in terms 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 BUNCH DETECTION USING THE SIXTH MOMENT

The sixth statistical moment was computed on the time series of raw voltages observed by GMRT. The window size was considered to be 8192 samples, and the thresholds were computed using the equation 4 described in section 3.4. The logic of the sixth-moment algorithm is similar to the logic of kurtosis-based bunch detection, illustrated by the flowchart 3. The actual thresholds for the 8k sample window are found to be +0.763 and -0.763. It is observed that the sixth moment is more effective in detecting RFI bunch from the dataset, most likely because of the narrow threshold margin for noise. Figures 46 to 48 represent the results of RFI bunch detection using the sixth-moment algorithm.

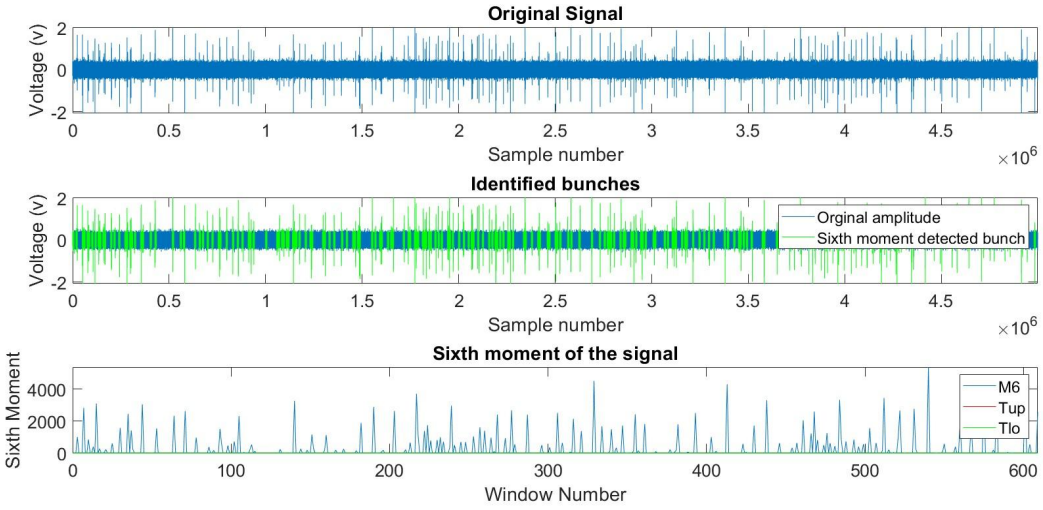


Figure 46: RFI bunch detected by the sixth moment in Band 2

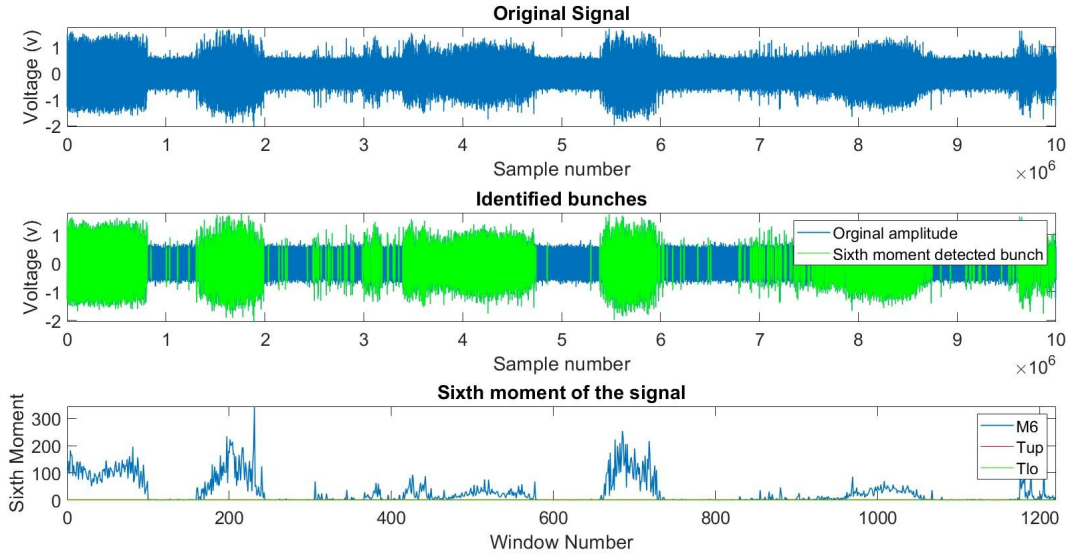


Figure 47: RFI bunch detected by the sixth moment in Band 3

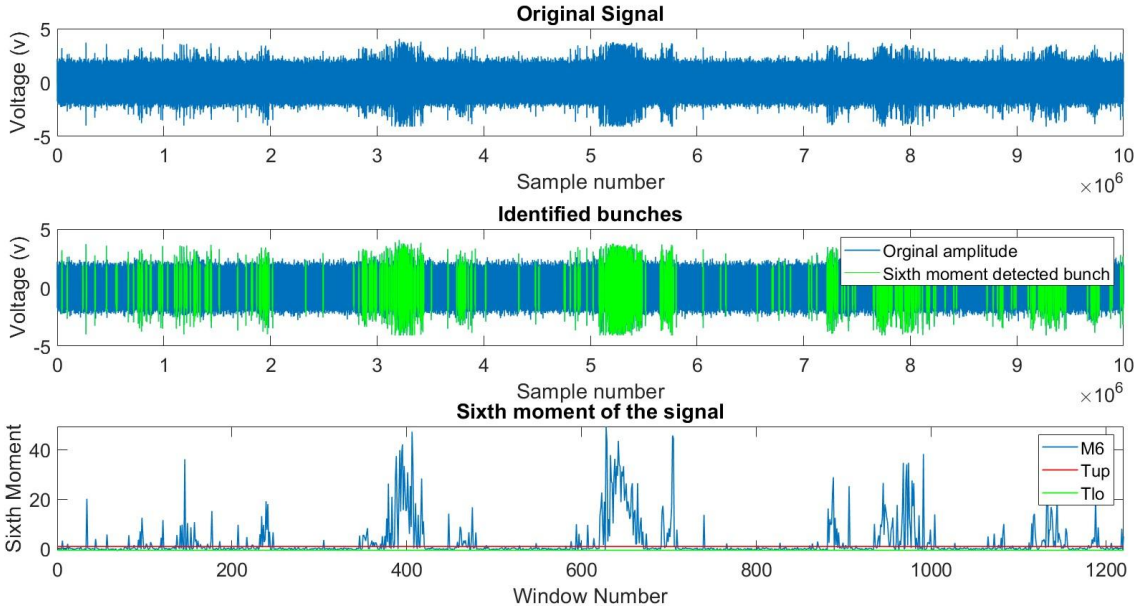


Figure 48: RFI bunch detected by the sixth moment in Band 4

CHAPTER 6: SIGNAL ENVELOPE AND AUXILIARY METHODS FOR BUNCH DETECTION

Apart from the RFI detection techniques described and used for this project, other methods for non-normality detection were also explored and tested for their efficacy. The analysis and results are discussed in the following sections of this chapter.

6.1 SIGNAL ENVELOPE AND SPARK MODELING

An envelope of the signal was studied in order to study the modulation pattern and nature of the sparks occurring in the dataset. The signal envelope follows a cubic spline interpolation, and models every occurrence of RFI. To use it along with kurtosis or sixth moment 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 the 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 50 shows the cubic spline interpolation curve fitted over the signal envelope.

The individual sparks inside the bunch were also modeled to study its decay pattern. It is observed that the sparks follow a Fourier series curve of fourth order. Figure 51 represents the Fourier curve modeled over an identified spark from the bunch.

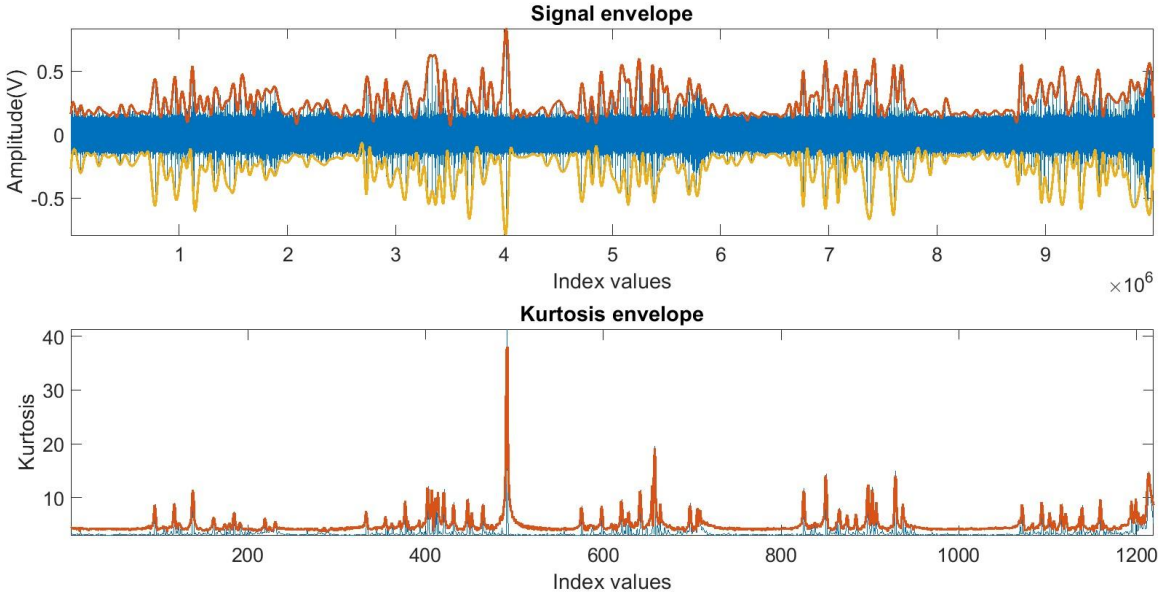


Figure 49: Signal envelope of Band 4

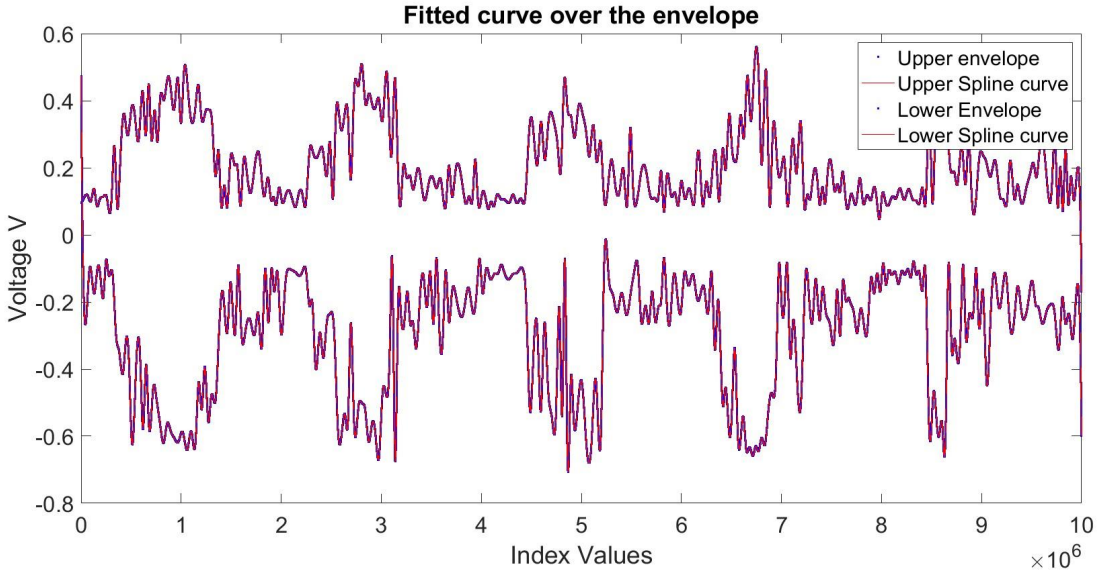


Figure 50: Envelope curve of signal in Band 4

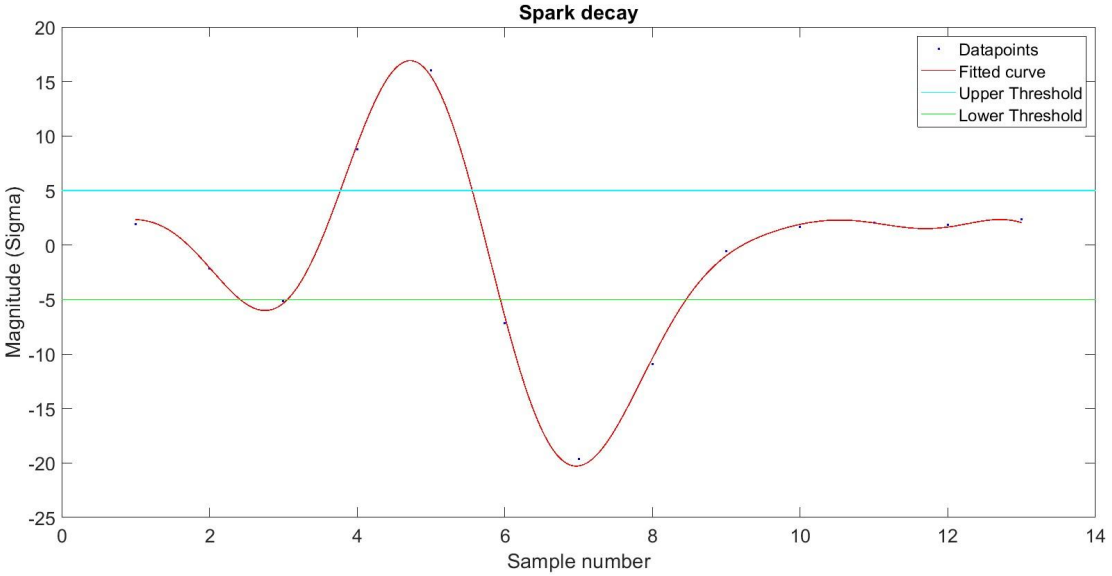


Figure 51: Spark curve from the bunch

6.2 NON-NORMALITY DETECTION BY COMPUTING THE GRADIENT

A gradient is a differentiable function which can be applied to two or three-dimensional vectors to generate a vector.

In order to avoid the use of higher-order statistics and the complexities involved in computing those metrics, a simple approach involving the gradient of the function was explored, for detecting non-Gaussianity.

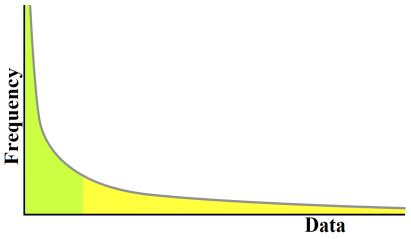
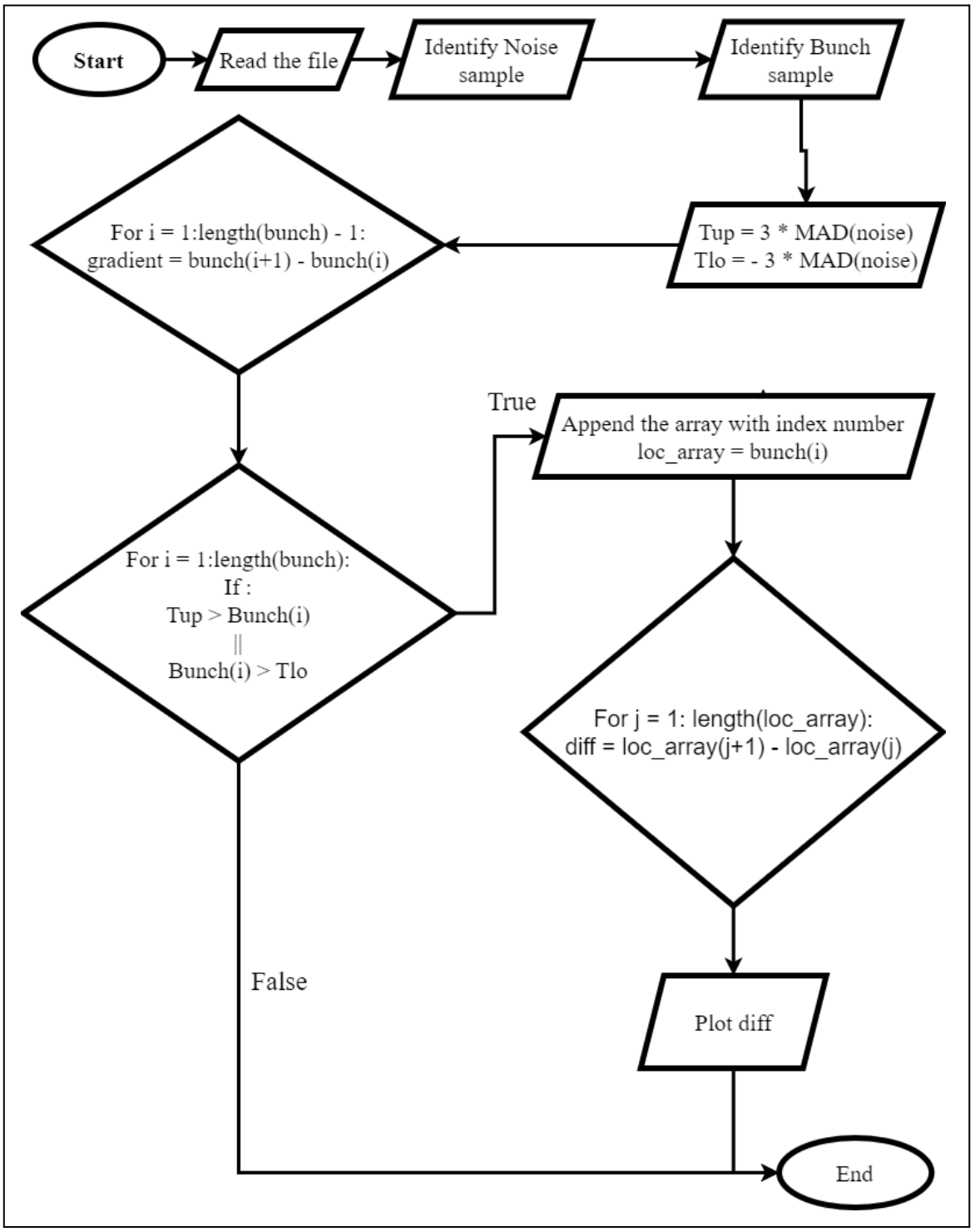


Figure 52: Power law distribution

The aim of this experiment was to study the distribution of gradients, computed over the entire time series and infer the nature of the underlying data. Since it is a well-known fact that RFI does not follow the Gaussian distribution, the gradient curve of samples containing RFI should follow a power-law curve, which will be a differentiating parameter for outlier bunch and uncorrupted noise. The power law distribution is basically an exponential function which indicates that over 95% of the occurrences have small and exponential intensities and very few of the occurrences have large and intensive magnitudes. Many natural phenomena follow the power law distributions. Figure 52 shows the typical curve of power law distribution, having a long tail and a sharp peak.

Some of the samples were analyzed from the individual RFI bunch in the dataset. The following Flowchart 4 explains the algorithm behind this technique.



Flowchart 4: Gradient distribution algorithm

The outputs of this technique are represented by Figures 53 and 54. The frequency v/s delta subplot in the below figures is a histogram of the difference (delta) between the adjacent gradients. This is the distribution observed in a selected bunch from the time series and is plotted using the log-log scale. The results indicate that the distribution of the delta clearly does not follow a power-law curve. Thus, the gradients of outliers might be following some other distribution and need some more research. This can be considered as a parallel approach to solve the problem of classification of non-Gaussian distribution. However, considering the duration constraint and primary goals of this particular project, this task is currently beyond its scope.

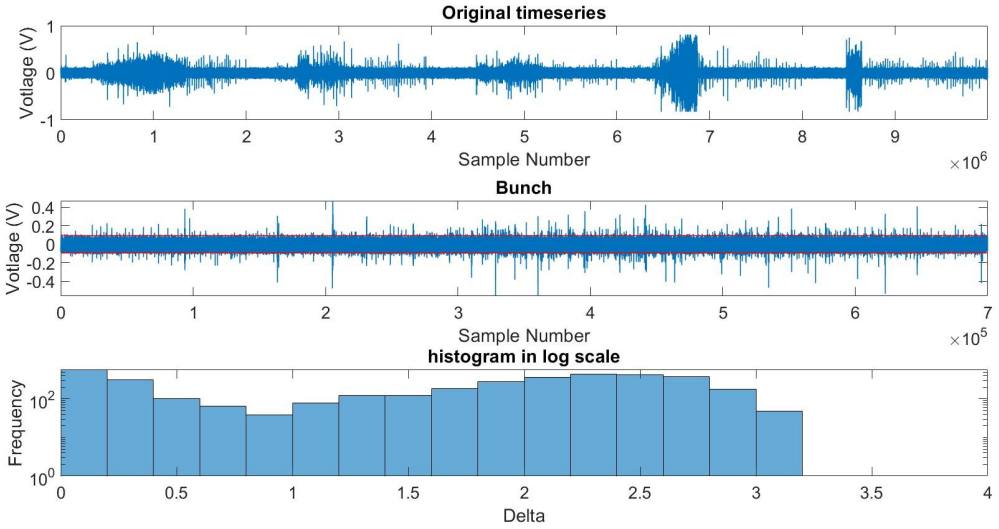


Figure 53: Distribution of Delta in RFI bunch of Band 3

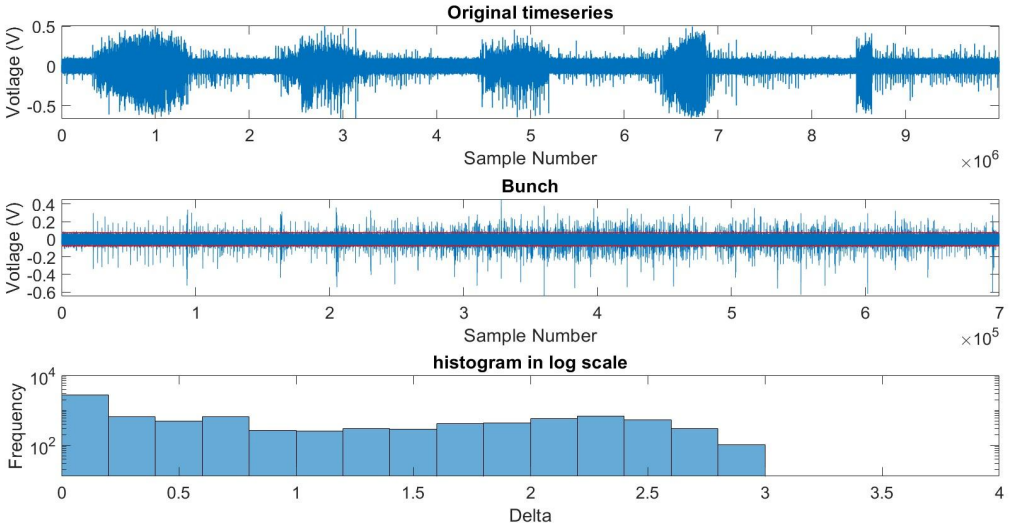


Figure 54: Distribution of Delta in RFI bunch of Band 4

CHAPTER 7: CONCLUSIONS AND FUTURE SCOPE

Radio Frequency Interference poses a serious problem to world-class sensitive radio telescopes like GMRT. Observatories operating in the radio spectrum constantly need to monitor radio frequency interference and maintain radio quiet zones to be at the forefront of scientific research. The present RFI mitigation system at GMRT is robust for flagging most of the strong powerline RFI. This robust estimation may sometimes lead to loss of valuable data, when a constant threshold is applied.

The information gathered with this project will help in opening 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 might get identified in a better way and mitigated using these techniques. 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 is a step towards providing a detailed recipe to begin the modeling and simulation of offline collected data for GMRT.

7.1 CONCLUSION

This project has explored many aspects of the statistical simulation and modeling of radio frequency interference. Statistical signal processing is one of the less attempted problem-solving approaches in the field of RFI mitigation. Some major conclusions from the experiments carried out as part of this project are regarding the higher order statistics and its hardware implementation.

- Kurtosis is a good estimator for the non-normality of the distribution and is efficient in detecting RFI bunch from the time series.

- The kurtosis algorithm is tested and verified over an FPGA hardware platform, which further validates the implementation of this method in real-time. With the power of Simulink and CASPER-enabled ROACH boards, it is possible to quickly simulate and run kurtosis based detection designs and make changes in the algorithm.
- It is observed that the results of the RFI detection technique using sixth-moment statistics are better than the fourth-moment statistics. A fairly reasonable explanation for this could be that the magnitude of the sixth moment has a sharp increase in the case of outliers and it lies within the thresholds in the case of noise. Hence, the sixth moment provides definite outlier boundaries which help in efficient RFI bunch detection.

7.2 FUTURE SCOPE

The experiments and analysis carried out in this project forms a foundation for the advanced machine learning techniques required to be deployed in the future for RFI mitigation. To highlight, the following points can be considered in future of this project:

- Identify RFI bunches with more accuracy and statistical confidence, by studying the signal envelope patterns and their thresholds.
- Design the compatible FPGA systems for deploying the RFI detection technique using the sixth-moment statistics in real-time. The architectural design for the sixth-moment model in Simulink could be very much similar to the fourth-moment, but the model might need some change to achieve desired frequency of operation
- Using regression analysis, predict and classify the incoming sample either as RFI or as noise
- Implement different machine learning techniques on a large amount of high time-resolution data being recorded periodically since last few years at GMRT
- Automate the selection of threshold and window size of the filtering algorithm based on the nature of the incoming signal/RFI

APPENDIX

A] Summary of literature review

A summary of the literature review conducted for this project is given in the table below:

No.	Author	Title	Journal	Year	Summary
1	Y. Gupta, B. Ajithkumar, H. S. Kale, S. Nayak, et al.	The upgraded GMRT: opening new windows on the radio Universe	Current Science	2017	GMRT system upgradation specifications, motivation for the upgrade, technological details
2	G. Swarup, S. Ananthakrishnan, V. K. Kapahi et al.	The Giant Metrewave Radio Telescope	Current Science	1991	GMRT telescope specifications, observational objectives, mechanical details
3	B. Bhattacharyya, Jane Roy et al.	Serendipitous Discovery of Three Millisecond Pulsars with the GMRT in Fermi-directed Survey and Follow-up Radio Timing	IEEE	2022	Pulsar observations using GMRT at 322 - 607 MHz, Discovery of the pulsar, the study of its orbital period
4	Marvin o. Loftness	Power Line RF Interference-Sounds, Patterns, and Myths	IEEE	1997	Causes and effects of powerline RFI are discussed, various sources like television, engine spark plug etc are

No.	Author	Title	Journal	Year	Summary
					identified, common patterns of RFI are studied and myths related to it are busted.
5	Kaushal D. Buch, Ruta Kale, Mekhala v. Muley et al.	Real-time RFI Filtering for uGMRT Observations: Shared-risk Release and Optimal System Configuration	IEEE	2022	RFI filtering system in real-time for GMRT is designed, the accuracy of the system is discussed with known advantages and disadvantages

B] Analysis of Non-normality tests on different datasets

Detection technique	Theoritical blindspots (%)	Observed blindspots (%)	Window size	Threshold		Input data		% RFI (window based)	% RFI (sample based)	Plot
				Lower threshold	Upper threshold					
Kurtosis (without consecutive window condition)	50	58	8192	2.95	3.05	11 Jan 2023	1	62.39	0.3502	Kurtosis bunch detection
							2	52.54	0.3231	
							3	91.78	0.4225	
							4	49.91	0.3092	
						eGMRT 200MHz	1	100	100	
							2	100	100	
						14 July 2022	1	93.93	0.8841	
							2	75.16	0.5352	
							3	94.83	1.3089	
							4	84.01	0.5851	
						20 October 2022	1	99.09	0.248	
							3	77.78	0.2917	
							5	93.03	0.093	
							7	53.68	0.4984	
Sixth Moment	13.56 and 61.44	between 17.5 - 18.2% and between 70 - 75%	8192	-0.763	0.763	11 Jan 2023	1	35.96	0.3502	Sixth moment bunch detection
							2	30.2	0.3231	
							3	28.4	0.4225	
							4	14.61	0.3092	
						14 July 2022	1	74.5	0.8841	
							2	43.85	0.5352	
							3	85.81	1.3089	
							4	60.49	0.5851	
						eGMRT 200MHz	1	100	100	
							2	100	100	
						20 October 2022	1	98.93	0.248	
							3	65.9	0.2917	
							5	30.9	0.093	
							7	3.68	0.4984	
Shapiro - Wilk	NA	NA	8192	0.05	NA	11 Jan 2023	1	43.18	0.3502	Shapiro Wilk bunch detection
							2	68.96	0.3231	
							3	61.9	0.4225	
							4	27.91	0.3092	
						14 July 2022	1	100	0.8841	
							2	100	0.5352	
							3	100	1.3089	
							4	100	0.5851	

C] Statistical Analysis on different datasets

Sr No	Date	Dataset	Tup	Tlo	ENTIRE SIGNAL						INSIDE BUNCH				
					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)	RFI_count (samples)	Not_RFI_count (samples)	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	0.7923	79227	9920762	7.75E-06	0.0051	0.1048	68.8688	8.12E+03	0.8407	1.50E-04	0.1255	2.21E-06
		C05 / P-1 / B3	0.0931	-0.0931	0.4199	41988	9958001	9.16E-06	1.32E-02	1.55E-01	45.016	8.15E+03	0.5495	2.18E-05	2.42E-02	2.45E-06
		C06 / P-1 / B4	0.0765	-0.0765	0.729	72895	9927094	1.14E-05	0.0054	0.0953	67.772	8.12E+03	0.8273	1.32E-04	0.1342	6.42E-07
		C08 / P-1 / B5	0.018	-0.018	0.4348	43481	9956508	5.14E-06	0.044	0.1473	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	0.385	38495	9961494	6.96E-06	0.0223	0.1327	35.0817	8.16E+03	0.4282	2.01E-05	0.0392	1.60E-06
		C06 / P-1 / B2	0.129	-0.129	0.4166	41655	9958334	7.41E-06	0.0159	0.1436	37.7646	8.15E+03	0.461	2.07E-05	0.0352	1.50E-06
		C05 / P-2 / B3	0.1991	-0.1991	0.4232	42317	9957672	6.50E-06	0.0307	0.0931	35.3809	8.16E+03	0.4319	4.36E-05	0.1032	1.16E-06
		C06 / P-2 / B4	0.1186	-0.1186	0.3824	38240	9961749	6.99E-06	0.0208	0.1476	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	1.0745	107452	9892550	8.30E-06	0.0037	0.0847	93.9633	8.10E+03	1.147	1.32E-04	0.133	1.50E-06
		C09 / P-1 / B3	0.4981	-0.4981	0.3714	37143	9962859	5.63E-06	0.0152	0.0989	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	1.9902	199019	9800975	6.06E-06	0.0017	0.2153	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	0.3057	30570	9969427	5.67E-06	0.0844	0.2076	29.0997	8.16E+03	0.3552	7.25E-06	0.0205	7.30E-07

D] List of generic parameters

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

E] Major versions of design

	Sr No	Type	File Name	Version	Remarks	Status
	1	Statistical Analysis	all_in_one.m	2	Computes RFI count, Not RFI count, RFI percent, IAT of bunch and spark inside the signal and bunch respectively	Final
	2	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
PRIMARY CODES:	3	Non normality tests	all_non_normality combined.m	1	Identifies non normal distribution samples and RFI using different non normality tests. Combined code for shapiro wilk, sixth moment and kurtosis	Final
	4	Test data analysis	emulator_analysis.m	1	Kurtosis algorithm for testing the emulator raw voltage data	Final
			simulink_fourth_moment_test.m	3	This code is used to compare kurtosis estimation method and RFI based flagging of data from MATLAB based empirical method and simulink model	Final
	5	Bunch detection	kurtosis_bunch_detect_final.m	1	Detects RFI bunch using kurtosis	Final
			shapiro_wilk_test.m	1	Detects RFI bunch using Shapiro Wilk	Final
			only_sixth_moment_bunch_detect.m	2	Detect bunch using sixth moment	Final
	6	Gradient	gradient_v1.m	1	Computes the gradient and plots its histogram to study the distribution of outliers in the bunch	Final

	Sr No	Type	File Name	Version	Remarks	Status
	1	RFI analysis	bunch_identify_function.m	1	Identifies rfi bunches using kurtosis	Final
	2		iat_bunch.m	1	Calculates the inter arrival time between the bunches	Final
AUXILLARY FUNCTIONS:	3		iat_spark_function.m	1	Calculates the inter arrival time between the sparks inside the bunch	Final
	4		noise_identify_function.m	1	Identifies noise within the dataset using kurtosis	Final
	5		RFI_analysis.m	1	Computes RFI density within the given timeseries	Final
	6		RFI_duration.m	1	Calculates the avareage duration of RFI	Final

F] 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

REFERENCES AND CITATIONS

1. NCRA website [Online], Available: http://www.ncra.tifr.res.in/ncra/main
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://electricalengineering-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. B. W. Yap, C. H. Sim, " Comparisons of various types of normality tests", Journal of Statistical Computation and Simulation, 2011
11. Sidharth Misra, Jonathon Kocz, Robert Jarnot et al. " Development of an On-Board Wide-Band Processor for Radio Frequency Interference Detection and Filtering", IEEE, 2019
12. Irappa M Halagali, "Programming the FPGA using BORPH", 2011.
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 the detection of interference in radiometer", IEEE, 2008.
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", 2009
17. Jerry M. Mendel, 1991, "Tutorial on Higher-order Statistics in Signal Processing and System Theory: Theoretical Results and some Applications", [Online], Available: https://sipi.usc.edu/~mendel/publications/HOSATutorial.pdf
18. CASPER Documentation [Online]: https://casper.astro.berkeley.edu/wiki/Main_Page
19. Datasheet of Xillinx Virtex 5 FPGA [Online]: https://docs.xilinx.com/v/u/en-US/ds100
20. Kaushal D. Buch, 2015, "Simplified Kurtosis Detects Signal Interference", [Online]: https://www.edn.com/simplified-kurtosis-computation-detects-signal-interference/
21. R. P. Brent, H. T. Kung, "The Chip Complexity of Binary Arithmetic", Harvard University, 1980.
22. ROACH Tutorial [Online]: https://casper.astro.berkeley.edu/wiki/Tutorial