Identifying RFI in voltage data from the GMRT

Ruta Kale, Kaushal D. Buch, Bela S. Dixit
[1] National Centre for Radio Astrophysics,
Tata Institute of Fundamental Research, S. P. Pune University Campus,
Ganeshkhind, Pune 411007.
Email: ruta@ncra.tifr.res.in, kdbuch@gmrt.ncra.tifr.res.in, dixitbela@gmail.com

May 19, 2023

Keywords: Observational astronomy, astronomical methods, radio astronomy, antennas
Level and stream: UG/PG (Science)
Level of difficulty: 2

Prerequisites

We recommend knowledge of the following aspects of radio astronomy:

- A qualitative understanding of how a radio antenna works and how observing in the radio is different from observing in the optical spectrum.
- Basic understanding of radio receiver elements such as low noise amplifiers, mixers, filters, detectors and integrators and their effect on electrical signals.
- Basic probability theory and statistics, including knowledge of the Gaussian distribution and the central limit theorem; basics of Fourier transforms.
- Experience with file handling, data manipulation, and plotting using any programming language appropriate for data analysis.

Alternatively, we suggest that you explore and read up on these topics as you attempt the activity. Although one can find plenty of online resources and books covering the prerequisites listed above, here are a few standard references on these topics.

1. Low Frequency Radio Astronomy a.k.a. GMRT Blue Book (http://www.ncra.tifr.res.in/ncra/gmrt/gmrt-users/low-frequency-radio-astronomy).
2. J. J. Condon, SM Ransom, Essential Radio Astronomy (2016). [Online version, https://science.nrao.edu/opportunities/courses/era].
3. J. D. Kraus, Radio astronomy (1966).
4. T. L. Wilson, K. Rohlfs, S. Hüttemeister, Tools of Radio Astronomy (2013).
5. R. N. Bracewell, Fourier transform and its applications (1999).
6. K. F. Riley, M. P. Hobson, S. J. Bence, Mathematical Methods for Physics and Engineering: A Comprehensive Guide, 2 edition, (2002).

1 Aims of this experiment

The following are the aims of this experiment:

1. To examine and compare the statistical properties of the voltage time series data recorded from the GMRT at bands 3 and 5.

2. To construct a power spectrum and dynamic spectrum from the data.

3. To construct auto-correlation and cross-correlation spectra.

4. Identify RFI in the data at bands 3 and 5.

2 Background

Radio telescopes are receiver systems that collect and process electromagnetic waves in radio frequency range (typically tens of MHz to a few hundred GHz) from celestial sources. Since the signals from celestial sources are weak, large dishes or arrays of dishes are often used to increase the collecting area. Giant Metrewave Radio Telescope (GMRT) is an array of antennas that consists of 30 parabolic dish antennas distributed roughly in a "Y"-shape (Fig. 1). There is a central group of 12 antennas distributed in an area of about 1 square km that is referrred to as the *central square*. The rest of the antennas are distributed along three arms in east, west and south. Each dish has a diameter of 45 m and the separations between the dishes range from 100 m to 25 km.

At the prime focus of each dish, there is a dipole that converts the electromagnetic wave into an electrical signal. In the case of the GMRT, these electrical voltages are further processed in a correlator system where the cross-correlation of these from unique pairs of antennas is calculated and stored for each sampling time and frequency channel. This is called an interferometric observation and the stored cross-correlations are termed as visibilities. The visibilities are processed to form an image of the observed patch of the sky.

In this experiment, we are going to use the voltage time series data recorded from the GMRT antennas. The Fig. 2 shows the block diagram of the path of the data from the dipole at the focus of the antenna to the data aquisition system. We will use data from the antennas C11 and C12 from the central square of the GMRT.

2.1 Radio Frequency Interference (RFI)

We would like to record signals from astronomical sources with radio telescopes. RFI is a man-made electromagnetic radiation that inevitably gets recorded along with the astronomical signals. The sources of RFI could be external to the telescope or within the electronics of the telescope itself. RFI can cause distortion of the radio astronomical data and thus needs to be identified and removed from

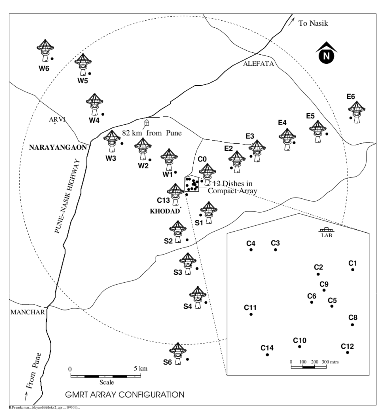

Figure 1: Left: GMRT antennas. Right: The GMRT array of 30 antennas is shown schematically. The inset shows the layout of antennas in the central square. Image credits: NCRA

the data during processing. Identification of RFI in the data can be done using the differences in the statistical properties of the RFI and the astronomical signal.

Based on its spectral occupancy, RFI is broadly classified as broadband and narrowband. In the case of broadband RFI, the interference is spread across the band (spectrum) whereas for narrowband RFI, the interference is confined to a certain part of the spectrum.

Typical sources of broadband RFI are sparking on high-power transmission lines and distribution equipment (e. g. transformers) , automobile sparking and switching of inductive load. A few examples of sources of narrowband RFI are broadcasting transmitters, satellites, air traffic control systems and radar.

For GMRT observations, most of the observing bands are affected by both the kinds of RFI. Broadband RFI from sparking on high power transmission lines and transformers is dominant in Bands-2 (120-240 MHz), 3(250-500 MHz) and 4 (550-850 MHz) of the GMRT. Narrowband RFI from various transmitters is observed in all the GMRT bands including Band-5 (1000-1450 MHz).

3 Introduction to signal processing

3.1 Sampling, mean and distribution

The astronomical signals coming to us in the form of electromagnetic waves are recorded as voltages by a radio telescope. The signal is often termed as *noise* in radio astronomy. The signal is not just a monochromatic wave but a combination of a number of monochromatic waves with different rates and amplitudes that gets recorded over a finite bandwidth. In order to reconstruct the signal from the recorded samples, the sampling of the signal has to be at a rate that is twice that of the bandwidth (the Nyquist-Shannon sampling theorem). In the case where the bandwidth is $\Delta\nu$, the sampling time needs to be,

$$\delta t = \frac{1}{2 \times \Delta\nu}. \tag{1}$$

We use the mean and root mean square noise (rms) to find the properties of the signal. The distribution of the voltages is expected to be a Gaussian with a zero mean. Any offset from zero mean can be measured and removed from the data.

We will also examine the distribution of voltage square which will correspond the power ($P \propto |V|^2$) and that follows a distribution of the form $e^{-P}/\sqrt{P}$ (See B.6 in reference 2). For this distribution, the mean power ($\langle P \rangle$) equals the variance of the voltages and the standard deviation (Root-Mean-Squared deviation in power) is equal to to $\sqrt{2}$ times the mean power. On averaging the Nyquist-sampled power over a certain time interval, τ , the noise power fluctuations (σ_n) reduce by a factor $1/\sqrt{N}$, where $N = 2\Delta\nu\tau$ is the number of samples averaged, and thus $\sigma_n = \sqrt{2}\langle P \rangle/\sqrt{2\Delta\nu\tau} = \langle P \rangle/\sqrt{\Delta\nu\tau}$.

3.2 Time and frequency domains

The time series of voltages can also be analysed in the frequency domain. The operation to convert the time series to frequency domain is to Fourier transform (FT) the time series. We have signals that are sampled usually at regular intervals and are of finite duration. For such a data, Discrete Fourier Transform (DFT) is used. The DFT of the points x_j, where $j = 0, 1, 2, ...N-1$, sampled at uniform intervals is given by,

$$X_k = \sum_{j=0}^{N-1} x_j e^{-2\pi ijk/N}, \tag{2}$$

where $i = \sqrt{-1}$. The DFT of an N-point input time series is an N-point frequency spectrum, with Fourier frequencies k ranging from $-(N/2-1)$, through the 0-frequency or so-called DC component, and up to the highest Fourier frequency $N/2$. Each bin number represents the integer number of sinusoidal periods present in the time series.

Selecting N points from the time series one can obtain the DFT. In this experiment our signal is real and the DFT will be an array of complex numbers. The DFT of such data is Hermitian - implying that the real part of the spectrum is even and the imaginary part is odd. The negative frequencies carry no new information and thus we will work with only the positive frequencies.

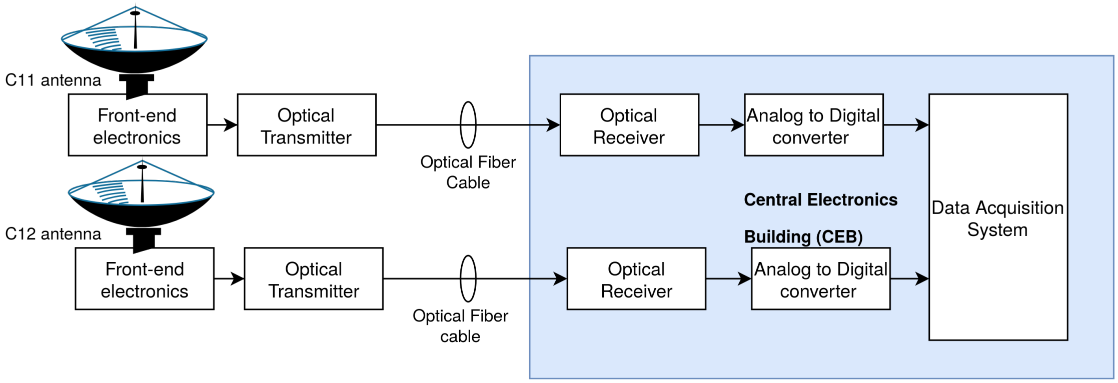

Figure 2: Block diagram showing the receiver chain from which the raw voltage data are acquired. Astronomical signal is collected through parabolic dish reflectors and amplified in the Front-end electronics system located at the prime-focus of the dish. Signal from each antenna is sent through an optical fibre cable to a central location. In the Central Electronics Building (CEB), the optical signals received from the antennas are converted to electrical signal using optical receiver system. The signals are then digitized using Analog-to-Digital converters and recorded on a computer through the data acquisition system.

The number of points chosen for the DFT will decide our frequency resolution. For example, DFT with 4096 points will give a spectrum of 2048 points (frequency bins). If $F(\nu)$ is the DFT of $f(t)$, then the FF^* (product of F and its complex conjugate F^*) is called the power spectrum.

3.3 Cross-correlation and auto-correlation

The cross correlation h(τ) of two signals, f(t) and g(t), is given by

$$h(\tau) = \int_{-\infty}^{\infty} f(t-\tau)g(t)dt. \tag{3}$$

The cross-correlation theorem states that the Fourier transform of the cross-correlation of two functions is equal to the product of the individual Fourier transforms, where one of them has been complex conjugated. If $H(k)$ is the Fourier transform of the cross-correlation $h(\tau)$ and $F(k)$ and $G(k)$ are the respective Fourier transforms of $f(t)$ and $g(t)$, then,

$$H(k) = F(k)G^*(k). \tag{4}$$

This is also referred to as the cross-correlation spectrum. When the two functions are identical ($f(t) = g(t)$), the we get the special case of auto-correlation. The Fourier transform of an auto-correlation function is the power spectrum (same as mentioned in Sec. 3.2), or equivalently, the auto-correlation is the inverse Fourier transform of the power spectrum.

In this experiment we will be calculating examining the Fourier transforms of the auto and cross-correlations (spectra). The plot of the frequency-time plane with the amplitude plotted on a colour scale is called a dynamic spectrum. We will use this kind of a plot to visualise the data.

3.4 RFI in the data

To identify RFI in the data we will use the statistical properties of the signal and that of the RFI. For an ideal signal from a celestial source, we expect a Gaussian distribution of the voltage and the power has an exponential distribution. We can use the property of the exponential distribution that the mean is equal to the rms and any deviation from this could be introduced due to a signal that is not of celestial origin (see Deshpande (2005)). The distribution of the magnitude in every channel in the auto correlation spectrum is expected to be exponential. We can take the ratio of the expected rms which will be proportional to the mean divided by the $\sqrt{\Delta\nu\tau}$ and the observed rms and use this distribution to locate the frequencies in the spectrum affected by RFI. In the case when the spectra are not averaged, the $\sqrt{\Delta\nu\tau}$ will be equal to unity.The value of the ratio will deviate from 1 due to contributions of non-random signals. Thus a dip in the value will indicate channels with RFI . A significant deviation like $\geq 10\%$ can be classified as RFI. At the edges of the spectrum the values will typically show large deviations and these are contributed by dominant quantization noise at the digitiser.

4 Data

The data contains voltage time series recorded simultaneously from the two antennas C11 and C12 of the upgraded GMRT (uGMRT) array. The upgraded GMRT allows an instantaneous bandwidth of 200 MHz as compared to the legacy GMRT when the maximum instantaneous bandwidth of only 33 MHz was possible. The upgrade of the GMRT happened over the last few years and it was inaugurated in 2019 (Gupta et al 2017, Current Science, 113, 707).

The observations were carried out for band 3 (300 - 500 MHz) and band 5 (1050 - 1450 MHz). A bandwidth of 200 MHz was used at each of these bands. Since the bandwidth is 200 MHz, the minimum sampling frequency (as per Nyquist criterion) was 400 MHz.

Each data file contains a list of 4194304 numbers. Each number is the digitized voltage recorded and the successive numbers are recorded at a time interval of 2.5×10^{-9}s (1/400MHz). The data files are in ASCII format. The names of the files and links to download them are given in Table 1. The recorded voltages are in arbitrary units.

uGMRT Band	C11 data	C12 data
3	C11_1024_Packets_B3.out	C12_1024_Packets_B3.out
5	C11_1024_Packets_B5.out	C12_1024_Packets_B5.out

Table 1: Summary of the data files. The data files can be found here.

5 Procedure

5.1 Properties of the voltage time series

Plot and visualize the voltage time series from the two antennas. Include the plots in the report with appropriately chosen time ranges to best show the voltage variability. You may use any programming language or software for plotting. Choose the first voltage sample as zero time reference and increasing by the sampling time (1/400MHz). Assign a time coordinate to the X-axis. The Y-axis voltages can remain in arbitrary units. See an example of such a plot in Fig. 3 (left).

Plot the voltage amplitude distribution for the two antennas and calculate the mean and standard deviation of the probability density function (PDF). Fit a Gaussian and obtain the mean and the

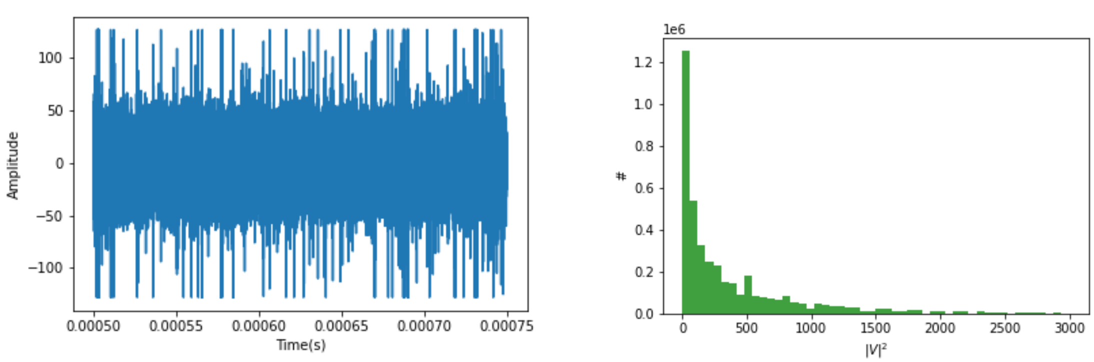

Figure 3: Left: An example of a time series plot from the data. The amplitude is in arbitrary units. The time axis is calculated by assuming the first sample to be at time 0 s and each subsequent one is separated by 2.5×10^{-9}s. Right: Plot of voltage squared ($|V|^2$).

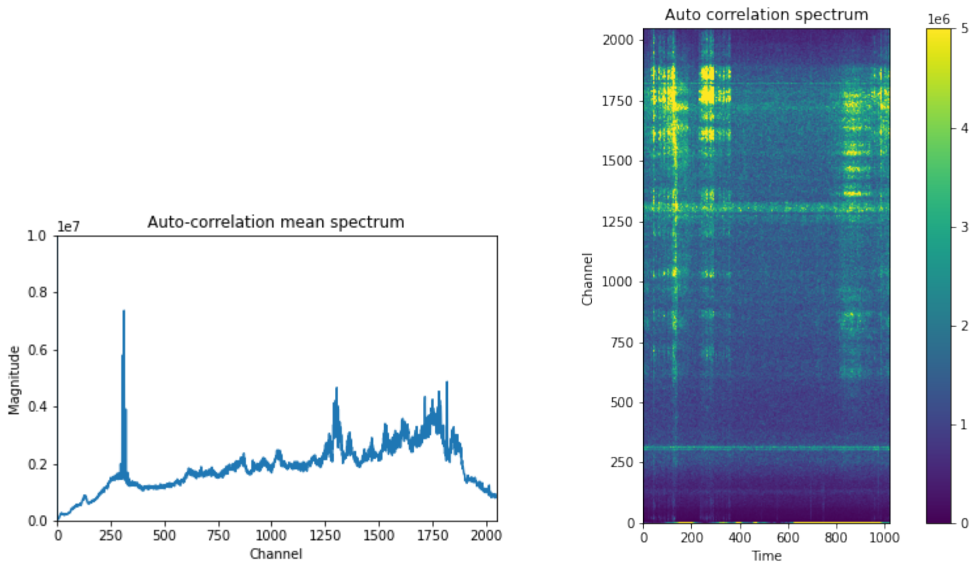

Figure 4: Left: A mean auto-correlation spectrum. Right: The dynamic auto-correlation spectrum is plotted. The frequency bins (channels) are on the y-axis and the time on the x-axis. The color bar shows the magnitude in arbitrary units.

standard deviation. Use the calculated mean and standard deviation to plot a Gaussian probability density function curve over the histogram. The PDF curve may require appropriate scaling to match the histogram. Mention the formulae used to obtain the distribution parameters, provide the estimated values, and mark those in the plots. The mean, if non-zero is to be subtracted from the samples (to remove the DC offset) and the new array of voltages thus obtained will be used for the next steps.

5.2 Properties of the power

Obtain the power from the voltages (will be in arbitrary units). Plot the power time series. Plot the power distribution (histogram) and estimate the parameters characterizing the probability density function (Fig. 3, right). Determine the functional form of the distribution followed by the power. Note that in our case the voltages are real. Use the calculated mean and standard deviation to plot a probability density function curve over the histogram. The PDF curve might require appropriate scaling to match the histogram. Mention the formulae used to obtain the distribution parameters, provide the estimated values, and mark those in the plots.

5.3 Power spectrum

The calculation of DFT is numerically carried out using the Fast Fourier Transform (FFT) algorithm. Taking 4096 points in the file at a time, calculate the FFT. Obtain the power spectrum by multiplying the Fourier transform with its complex conjugate. Plot the power spectrum. You will have 1024 such spectra and those can all be averaged to obtain a mean spectrum (Fig. 4, left).

Plot a dynamic spectrum using the 1024 spectra (frequency Vs time with amplitude in colour scale). See an example in Fig. 4 (right).

Calculate the Fourier transform of the cross-correlation using Fourier transforms of the data from individual antennas. For band 3 and band 5 plot the cross-correlation spectra. Use appropriate colour scale limits to visualise the data.

5.4 Identification of RFI

The RFI is present in the signal coming from each antenna and thus first we will identify RFI in each antenna for a given observing band. We will use the power spectrum for this analysis. The steps are

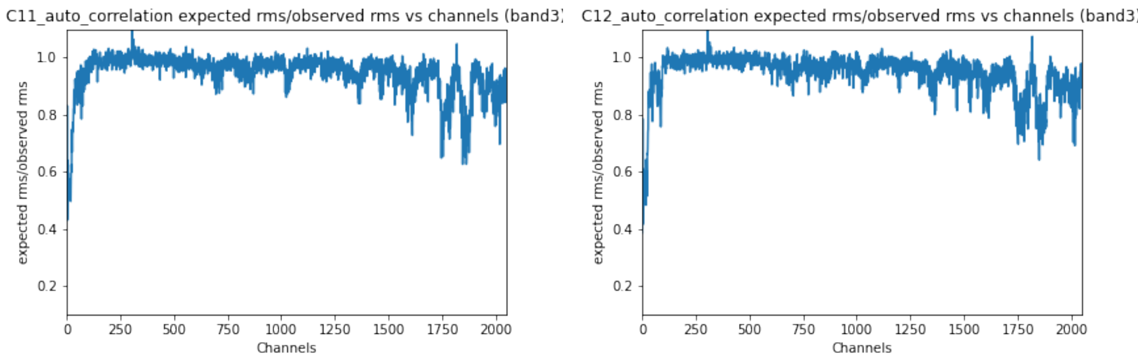

Figure 5: Left: The ratio of expected rms and observed rms is plotted as a function of channels using the auto-correlation spectrum for antenna C11. Right: The same plotted for the antenna C12.

given below.

1. Take the data for C11 for band-3. Obtain a FT using 4096 points at a time from the file. Obtain the power spectrum (auto-correlation spectrum) as given in Sec. 5.3. Consider only the positive frequencies for further analysis.

2. For each frequency bin (channel), obtain the mean and the rms.

3. The expected rms (Exp_rms) is mean divided by the square-root of the number of samples (twice product of the bandwidth and time) (Fig. 5). Here the product of bandwidth and time is equal to unity as the channel bandwidth ($\Delta\nu$) is inverse of the time interval (τ). Thus the mean times square root 2 is the estimate of the expected rms. Plot the ratio of expected rms to the observed rms.

4. Use a threshold of $> 10\%$ to identify the channels containing RFI. Note down the channels. Comment on the threshold to be used and how it would matter.

5. Repeat the procedure for the other antenna and plot it in a different colour in the plot for the earlier antenna for comparison.

6. Compare the power in the channels with RFI in the two antennas. Use an average power spectrum for each antenna (average over the 1024 spectra).

7. Repeat the steps for the two antennas for band-5. Using the same threshold as above, do you find RFI in the data ? Identify and compare RFI in the two antennas.

8. Also compare the two bands for the same antenna and record your findings.

The same procedure of obtaining the ratio can be carried out for the cross-correlation spectrum. However there it will only be possible to infer about the correlations in the RFI itself as there is no correlation expected in the signal. Find out the channels where there is correlated RFI for band-3 and also comment on the same for band-5.

6 Future work

This experiment is designed to expose the users to the properties of the voltages recorded with a radio telescope and the subsequent procedure to obtain the spectra by taking Fourier transform of the data. The task of identifying RFI in the data using this method allows to understand the statistical properties of the signal. The users can find out about other methods of identifying and also excising RFI from the data. For reading about the applications of the concepts the article by Radhakrishnan (1999) is recommended.

Acknowledgements

We thank the staff of the GMRT who have made these observations possible. GMRT is run by the National Centre for Radio Astrophysics of the Tata Institute of Fundamental Research. We thank Rahul Aragade for help in acquiring the raw voltage data for Bands-3 and 5. We are thankful to Avinash Deshpande for discussions on this experiment.

References

A. A. Deshpande, "Correlations of spectral intensity fluctuations: Application to radio frequency interference mitigation," in Radio Science, vol. 40, no. 05, pp. 1-9, Oct. 2005, doi: 10.1029/2004RS003156.

Radhakrishnan, V. 1999. Noise and Interferometry. Synthesis Imaging in Radio Astronomy II 180, 671.