OT2 - Assignment 2

K. Elliott - ELLKIR005
L. Landsberg - LNDLLO002

Aim
The aim of this report was to conduct a comprehensive analysis of a MeerKAT observation of PKS1934-63, focusing on detecting and analysing radio frequency interference. A jupyter notebook is available on the repository along with a ".yml" requirements file.

Hypothesis
Radio Frequency Interference (RFI) will be found throughout the signal.

---

1 Exploratory Data Analysis

Table 1: EDA Questions: 1-5

1. Observation Dates	31/01/2019 (12:56 - 12:59)
2. Target	PKS1934-63
3. Bandwidth (MHz)	L-band [856, 1711]
4. Channel Width (kHz)	835.938
5. No. of Antennas	50

An observation of PKS1934-63, a Seyfert 2 galaxy, was done between 12:56 and 12:59 on the 31st of January 2019, using a collection of 50 antennas (13.5 m diameter) at MeerKAT. The observation was conducted in the L-band (856 to 1712 MHz) using 1024 channels, resulting in a channel width of 835.94 kHz. The source, PKS1934-63 is found at a right ascension ($\alpha = 5.1462°$) and a declination ($\delta = -1.1120°$) in J2000.

2 Understanding the UVW

UVW and ENU represent two coordinate systems. The ENU system describes positions relative to a point on the Earth's surface using three axes: East, North, and Upwards (towards the zenith) from a reference point on the Earth's surface. This system is used to describe the relative positions of the antennas on the Earth's surface.

The UVW system depends on a baseline and describes positions relative to the source in the plane of the sky. Again, this system uses 3 axes: U, which points along increasing right ascension, V, which points along increasing declination, and W, which points along the line of sight towards the source. This system is used to 'project' the baselines onto the plane of the sky.

These two coordinate systems are connected through a 2D Fourier Transform. The visibility, $V(u, v)$ data collected is in the uv plane of the UVW coordinate system acts as samples for a Fourier transform, which converts them into sky brightness, $I(l, m)$ data in the lm plane of the ENU coordinate system as follows:

$$I(l,m) = \int\int V(u,v)e^{i2\pi(ul+vm)}dudv \tag{1}$$

The visibility data can also be calculated from sky brightness samples by performing the inverse Fourier transform.

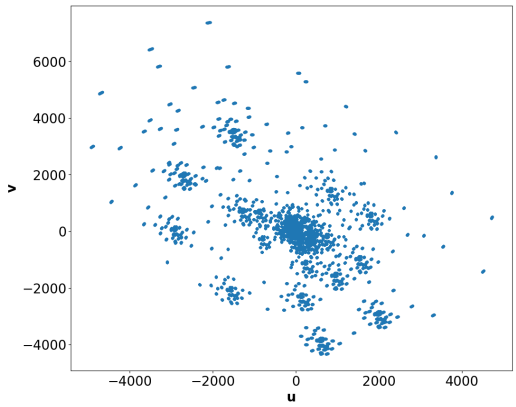


Figure 1: UV Coverage Map of the 50 antennas used in the MeerKAT observation of PKS1934-63

The UV coverage map shows the distribution of the projected baselines in the UV plane. Each point on the map indicates a baseline, therefore ideally this map would be completely filled for optimal image quality. Due to physical constraints this is not possible, the map shown in Fig. 1 is 'relatively' full. The points seem to be distributed in an elliptical shape with a major axis running from NW to SE. Baselines are also found in the outskirts of the map, indicative of larger baselines. This will provide higher-resolution measurements. The elliptical shape observed is due to the rotation of the Earth, which causes new baseline configurations relative to our source.

The expected resolution of the observation is found using Eqn. 2, where θ is the angular resolution (radians), λ is the central wavelength, ν is the associated central frequency, c is the speed of light, and B is the baseline length. A baseline is defined as the Euclidean distance between 2 antennas.

$$\theta = \frac{\lambda}{B} = \frac{c}{\nu B} \tag{2}$$

Due to the two antennas not being at the same physical position, the radiation will reach each antenna in a pair at a slightly different time. This time delay can be calculated using Eqn. 3 below, where $\mathbf{B} = (x_1 - x_2, y_1 - y_2, z_1 - z_2)$ is the position vector describing the baseline distance, $\hat{\mathbf{s}}$ is a unit vector pointing in the direction of the source and c is the speed of light.

$$\tau = \frac{\mathbf{B} \cdot \hat{\mathbf{s}}}{c} \tag{3}$$

The unit vector is found in spherical coordinates, where δ is the declination (deg) and α is the right ascension (deg).

$$\hat{\mathbf{s}} = (\cos\delta\cos\alpha, \cos\delta\sin\alpha, \sin\delta) = (0.99, 0.09, -0.02) \tag{4}$$

Table 2: UVW Questions: 2-4

	Short Baseline	Long Baseline
Distance (m)	29.257	7697.580
B (m)	(-0.702, 25.347, 14.594)	(4401.247, -4959.608, 3909.482)
2. Antenna Pair	m000 and m002	m047 and m060
3. Expected resolution (rad) at 856 MHz	1.199×10^{-2}	4.557×10^{-5}
Expected resolution (rad) at 1711 MHz	5.993×10^{-3}	2.278×10^{-5}
4. Delay (s)	4.315×10^{-9}	1.278×10^{-5}

3 Radio Frequency Interference

A plot is created of a 1D cut along the visibility data with polarization = 0 and a baseline created from the autocorrelation of antennae (daskms default), at a time of 3.45 seconds into the observation. It is shown in Fig. 2. This spectrum contains distinct radio frequency interference (RFI), as seen by the sharp spikes of high amplitude. Three regions of the proposed RFI are given as follows:
920-970 MHz, which is believed to be associated with GSM
1190-1220 MHz believed to be associated with Galileo 1
1595-1610 MHz believed to be associated with Glonass L1

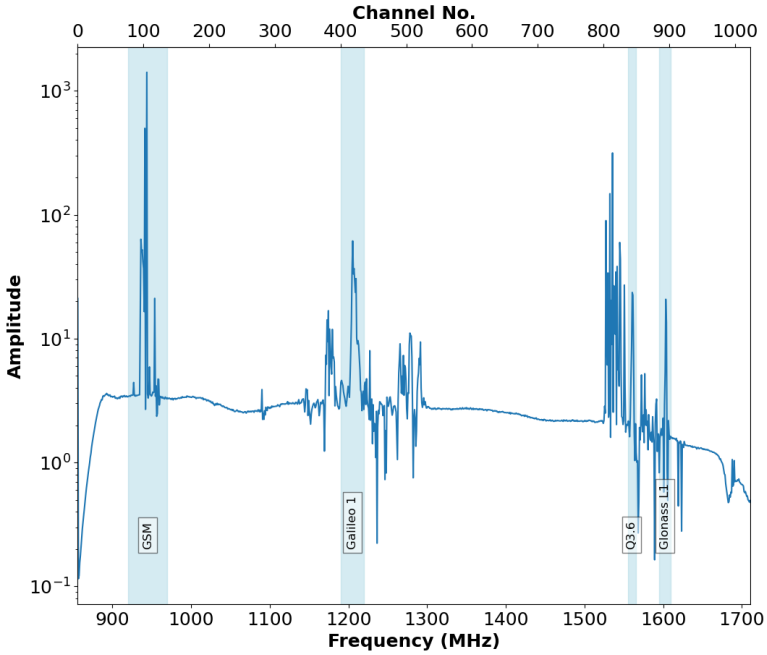


Figure 2: 1D cut across the visibility showing the amplitude of the signal received across a range of frequencies (MHz) in the L-band. Three proposed regions of RFI have been highlighted, as well as the RFI signal localised at 1575 MHz

3.1 Contaminated vs. Clean Signal

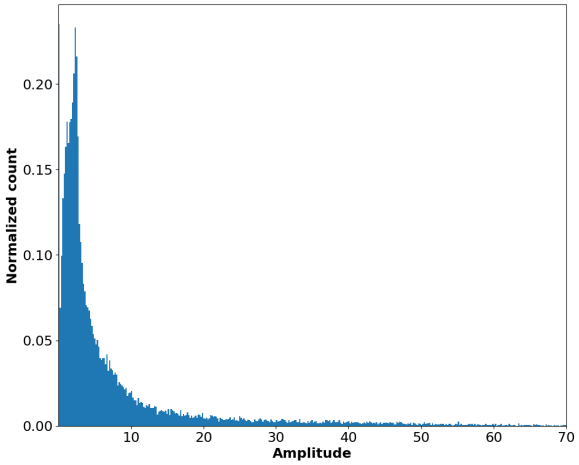
The visibility distribution from a 'contaminated' signal is expected to show marked differences from the visibility distribution of a 'clean' signal. A contaminated signal assumed to be associated with RFI is expected to be much stronger than astronomical background noise and any astronomical sources.

As described below, two frequency ranges highlight these differences between a clean signal and a contaminated signal.

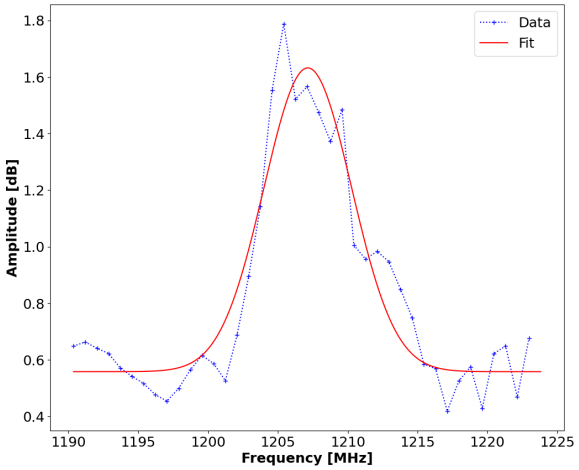
Table 3: Contaminated and clean signal bandwidths

	Contaminated	Clean
Bandwidth (MHz)	(1190, 1223)	(1466, 1499)
Central Channel No.	420	750

A histogram showing the normalised number of points within certain amplitude ranges is created for both the 'contaminated' and the 'clean' regions. The histogram of a 'clean' signal is expected to concentrate at lower amplitudes, while an RFI signal is expected to dominate the higher amplitude ranges.

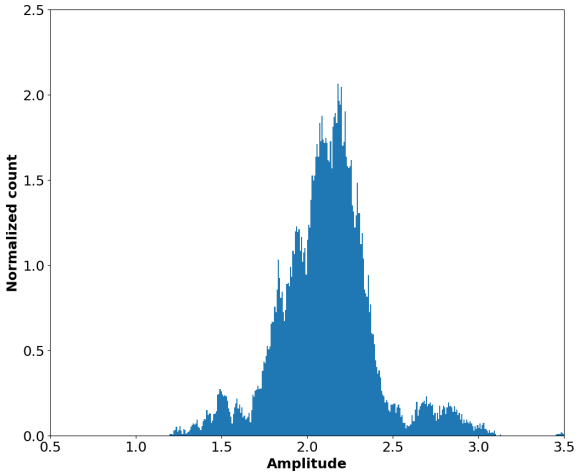


(a) Histogram of the amplitude distribution of an RFI signal at channel 420 across all antennas

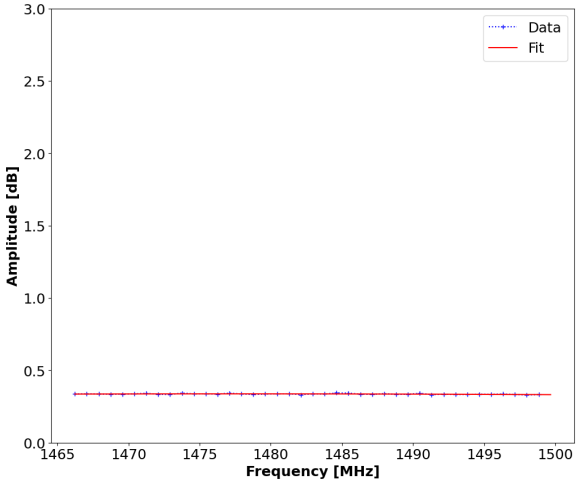


(b) 1D cut across the Visibility of the RFI signal centred at 1207 MHz with a Gaussian fit overlaid.

Figure 3: 'Contaminated' signal at channel 420 in the bandwidth range (1900, 1223) MHz



(a) Histogram of the amplitude distribution of the clean signal at channel 750 across all antennas



(b) Visibility of the clean signal centred at 1477 MHz with a Gaussian fit overlaid.

Figure 4: 'Clean' signal at channel 750 in the frequency range (1466, 1499) MHz

The histogram in Fig. 3a shows that the 'contaminated' signal peaks at smaller amplitudes (< 10). This is expected as noise at smaller amplitudes as still affecting this signal. An interesting feature of this histogram is the continuation of the histogram into larger amplitudes (> 20). This is indicative of an RFI signal which is generated to have strong amplitudes. The histogram in Fig. 4a shows a distribution that peaks at small amplitudes (≈ 2.1). This matches the peak in Fig. 3a, giving evidence to the suggestion that the peak in the 'contaminated' signal was in part due to noise. The 'clean' signal is restricted to amplitudes within 0.5 to 3.5, which is expected of faint background sources and noise. The two distributions are plotted together in Fig. 5 to highlight the differences in spread across the amplitude range between these two distributions.

The 1D cut through the visibility shown in Fig. 2, has been sliced at the two channel numbers in Table 3 to better show their finer details. A Gaussian is then fitted to this data slice and overlaid onto its visibility spectrum as seen in figures 3b and 4b.

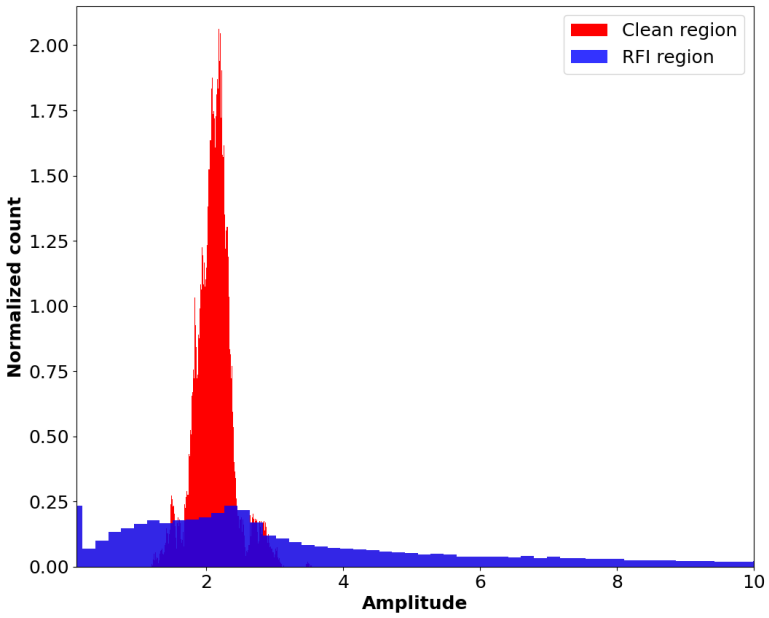


Figure 5: Histogram showing the amplitude distribution of the 'contaminated' and 'clean' signal across all antennas

The statistical moments of the 'contaminated' and 'clean' Gaussian's are then found and detailed in Table 4.

Table 4: Statistical moments for contaminated and clean signal

	Contaminated	Clean
Mean (MHz)	1207.1426	1477.4683
Standard Deviation, σ (MHz)	3.1337	623.8501
Kurtosis	-0.0841	0.2262
Skewness	1.2043	-1.1757

The mean describes the central frequency observed for the signal. The 'contaminated' distribution is centred at 1207 MHz, while the 'clean' region is centred at 1477 MHz.
Standard deviation, σ, describes the spread of the distribution about its mean. Higher values of σ indicate a larger spread in the distribution. The 'contaminated' distribution has a much smaller spread (3 MHz) than the clean distribution, which is much more widespread (624 MHz). This large standard deviation is expected as the clean signal in this smaller frequency range does not show a strong 'Gaussian' shape. Therefore, we have a smaller peak and a lot of spread by fitting a Gaussian. Conversely, The 'contaminated' signal shows a clearer Gaussian shape. RFI signals are generated to only radiate within a certain bandwidth, therefore the small spread is expected.

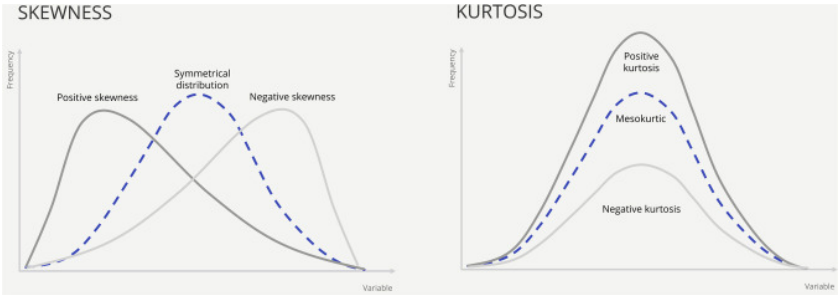
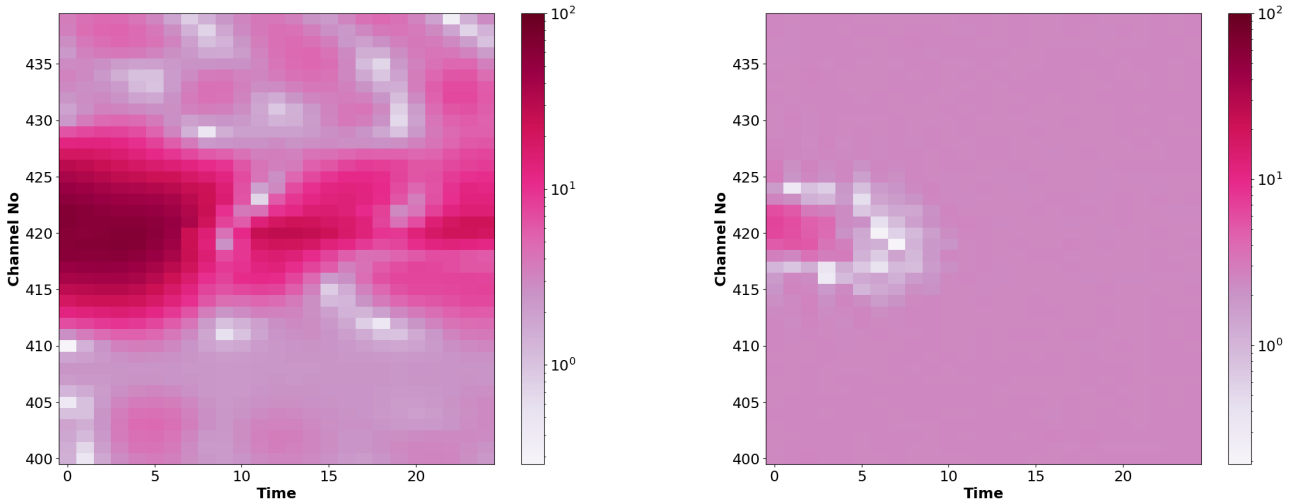


Figure 6: The effects of kurtosis and skewness on normal distributions. Credit: Green et al. 2023

As shown above in Fig. 6, the kurtosis of a normal distribution is expected to equal 0. Positive kurtosis values imply a sharper peak with higher weighted tails, while negative values imply a flatter distribution with lower weighted tails. The 'contaminated' distribution was found to have a kurtosis of -0.08, indicative of a slightly flatter distribution with slightly smaller tails compared to the normal distribution. Again, this is expected as these signals are generated to radiate within a specific bandwidth. The 'clean' distribution was found to have a kurtosis of 0.22, implying a slightly sharper peak ad heavier tails compared to the normal distribution. This is, however, an inaccurate description of this flat, non-Gaussian 'clean' distribution.
The skewness of a normal distribution is expected to equal 0. A positive skewness value implies more weight in the distribution's right tail (Zwillinger 2002). The 'contaminated' distribution was found to have a skewness of 1.2, which indicates that the distribution peaked on the left and had a large right tail. This shape can be seen in the underlying Data plot in Fig. 3b. The 'clean' distribution was found to have a skewness of -1.2. This would ordinarily indicate a distribution peaked on the right with a larger left tail. This is not what is observed in Fig. 4b, again due to this distribution's flat, non-Gaussian shape causing an inaccurate analysis of these moments.

3.2 Long vs Short Baseline

Contaminated Signal Bandwidth (1190, 1220) MHz:



(a) Shortest baseline
Antenna pair : m000 & m002

(b) Longest baseline
Antenna pair: m038 & m060

Figure 7: Visibility over time between channels 400 and 440 (1190-1220 MHz) for a short and long baseline.

In figures 7a and 7b, dark regions (high amplitude measurement) indicate RFI as these measured signals are much larger than the expected fluxes for astronomical objects. We note a much more apparent RFI in Fig. 7a (short baseline) than we do for Fig. 7b (long baseline). This results from the RFI being correlated at distances $\lesssim 1$ km and uncorrelated for longer baselines. This significantly reduces observed RFI from our shortest baseline (29 m) to our longest baseline (7.7 km).

3.3 Interesting RFI Signal 1: ≈ 1575 MHz

We will now investigate two interesting RFI signals: a signal localised near 1575 MHz and a signal with a known rest frequency of 153 MHz.
There is a distinct RFI signal around 1575 MHz, indicated in Fig. 2 as 'Q3.6'. This frequency would correspond to a BeiDou E1 band signal, part of the BeiDou Navigation Satellite System (BDS) in China, with a known central frequency of 1561 MHz.

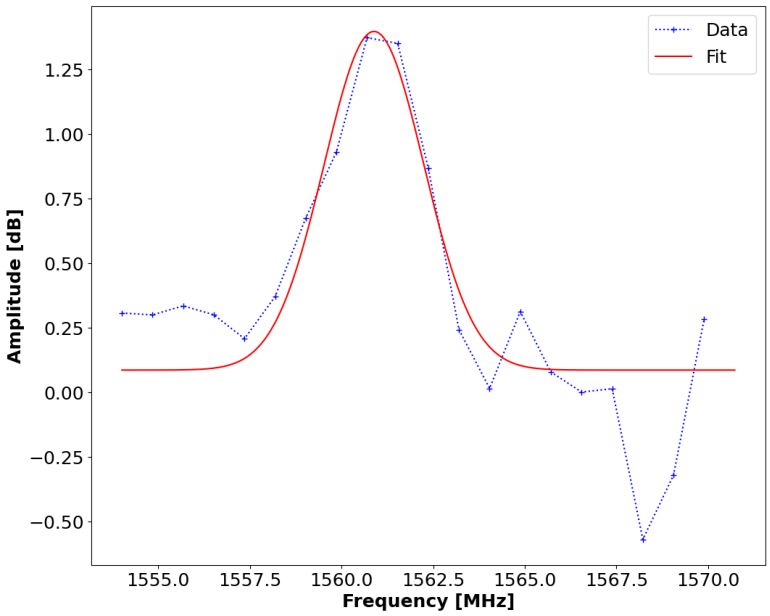


Figure 8: 1D cut across the visibility of the interesting RFI signal centred at 1561 MHz with a Gaussian fit overlaid.

A Gaussian distribution was fitted to the RFI signals seen in Fig. 8, producing a central frequency of 1560.885 MHz with an FWHM of 3.354. This reached a maximum of 1.396 dB, which affected four adjacent channels.

The redshift that a hypothetical RFI at a frequency of 1380 MHz interferes with the HI signal, the most widely observed emission in the radio. RFI interfering with continuum emission is problematic but not detrimental as this can be masked and science can still be done, but when the HI line contains RFI, we can not recover the signal and can not simply look at a different frequency as HI is a hyperfine transition line. To find the redshift, we use:

$$z + 1 = \frac{\nu_{rest}}{\nu_{obs}} \tag{5}$$

Where z is the redshift, ν_{rest} is the rest frequency, and ν_{obs} is the observed frequency. Then, the redshift of HI affected by RFI at 1380 MHz would be

$$z + 1 = \frac{1420.405752}{1380}$$
$$z \simeq 0.029$$

3.4 Interesting RFI Signal 2: 153 MHz

We could observe RFI at 153 MHz as a harmonic of the fundamental frequency:

$$f_n = n f_0 \tag{6}$$

where f_n is the frequency of the n^{th} harmonic and f_0 is the fundamental frequency. We could, therefore, see the following frequencies in the L-band:

Table 5: Summary of potential higher order harmonics that fall inside MeerKAT L-Band

Harmonic	Frequency (MHz)
6	918
7	1071
8	1224
9	1377
10	1530
11	1683

However, RFI is not necessarily visible across all antennae pairs and might only be seen by select antennae. Because of this, it's possible to see one or more of the harmonics in our dataset hidden in a specific antennae combination. Due to this project's limited time, it is impossible to individually investigate every 1225 existing pairs for the 50 antennae used. We present the default antennae pair overlaid with the expected location of the higher order harmonics summarised in Table 5:

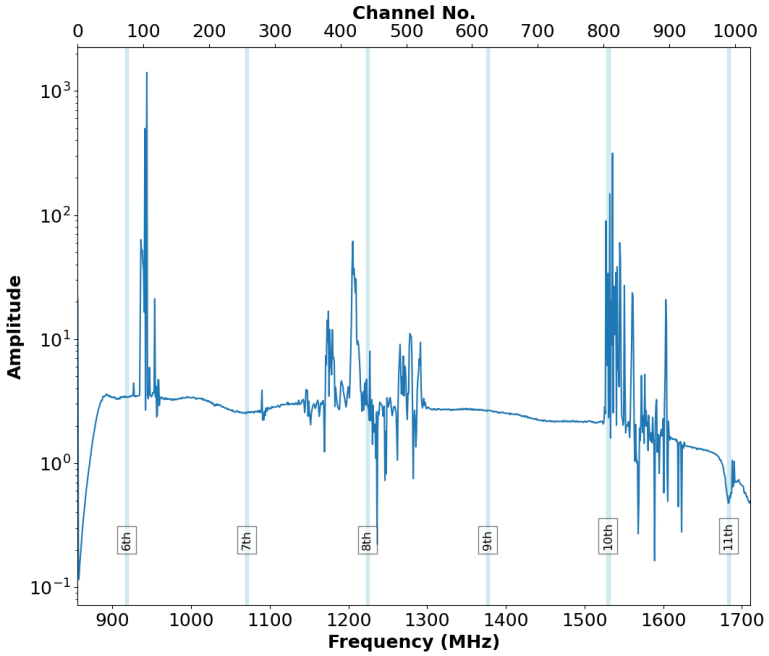


Figure 9: 1D cut across the visibility showing the amplitude of the signal received across a range of frequencies (MHz) in the L-band with the harmonic frequencies highlighted

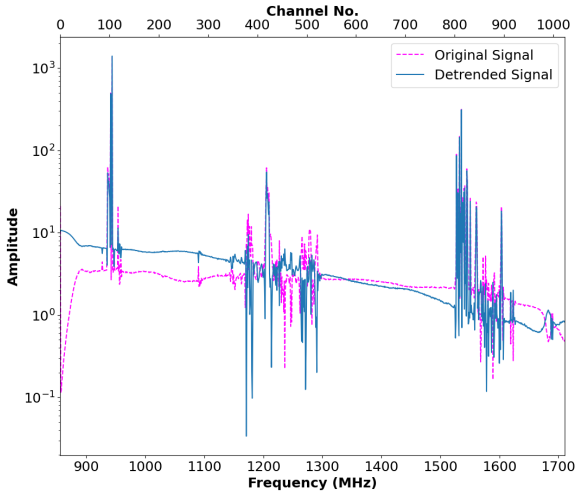
We note that both the 8th and 10th harmonic could potentially be present, but a conclusive detection is not possible due to the abundance of RFI in the corresponding regions. In addition to the default pair, the shortest baseline pair was investigated and available in Appendix B in Fig. 14. The shortest baseline was chosen as a likely candidate as RFI of this nature tends to only be coherent at short baselines. Again, while it is possible that one of the harmonics is present, no conclusive detection can be claimed.

3.5 Detrending Analysis

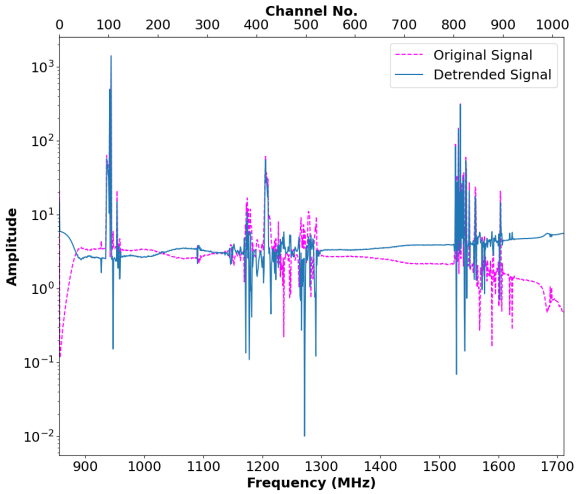
Detrending is the process of removing background noise as well as RFI signals to allow the comparatively weaker astronomical sources to be detected. There are many ways to handle the detrending process. In this report, we attempted to detrend the data using three methods:

The first method involved using the default linear `detrend` function in the `scipy.signal` package,

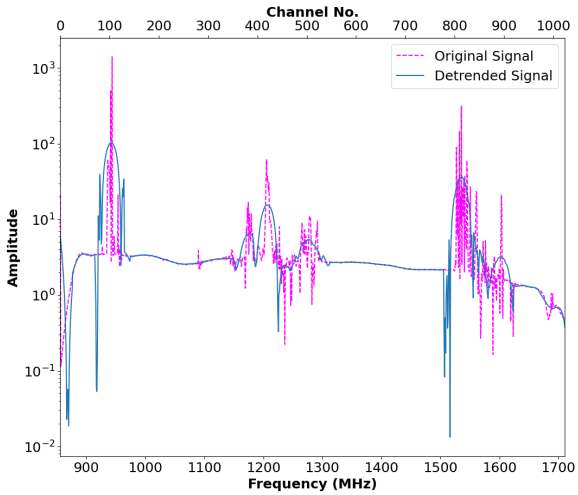
which removes a linear trend from the data. The second method involved once again using the `detrend` function but this time setting `type='constant'`, which removes a constant value from the data. The third method involved using the `savgol_filter` function in the `scipy.signal` package, which applies a Savitzky-Golay filter that smooths the data by fitting successive sub-sets of adjacent data points with a low-degree polynomial using a method of linear least squares regression (Luo, Ying, and Bai 2005). The `savgol_filter` function requires two arguments, namely, `window_length = 50` and `polyorder = 3`, which are the length of the filter window and the order of the polynomial used to fit the samples. The detrended data from each method is plotted here in Fig. 11.



(a) Method 1: detrend, type = linear



(b) Method 2: detrend, type = constant



(c) Method 3: savgol_filter

Figure 11: Plots showing detrending methods applied to the 1D cut across the visibility across the L-band frequencies

Table 6: Variances of the original and detrended datasets

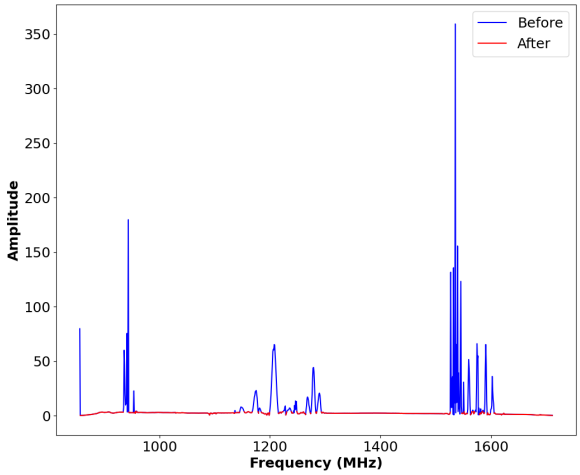
Method	Variance
1. detrend (type=linear)	2314.320
2. detrend (type=constant)	2321.859
3. savgol_filter	224.436
Original	2321.859

To evaluate the effectiveness of this detrending, the variance of the original dataset was found to be 2321.859 and compared to the variance of the detrended datasets. The results are given in the table in Table 6. The `savgol_filter` removed the most variance with the `detrend` function hardly removing any variance, for the constant case, there is no change in the variance.

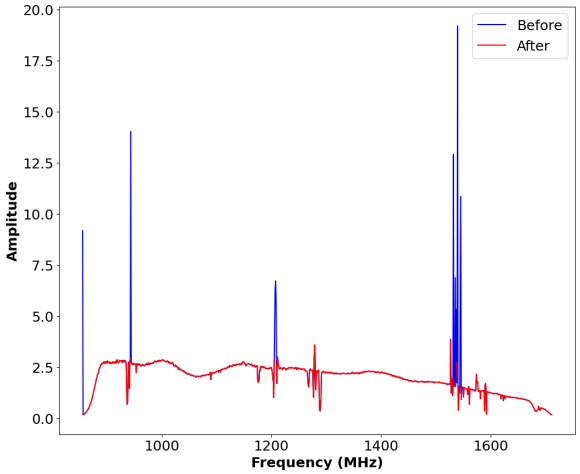
3.6 Filtering Analysis

Because of the variable and strong baseline dependence of RFI, creating a single function to completely eliminate RFI is a formidable task. As a first filtering attempt before a more fine-tuned removal process like hand flagging regions, simple masking of all amplitudes greater than a specified threshold offers a computationally inexpensive starting point. Despite its crude nature, it performed better on a range

of data from short to long baselines without tuning parameters for each baseline. By simply capping the amplitude to 4 dB, the following effect was observed:



(a) Amplitude capped to 4 dB vs no filter for shortest baseline antenna pair: m000 & m002



(b) Amplitude capped to 4 dB vs no filter for longest baseline antenna pair: m038 & m060

Figure 12: Amplitude vs frequency for two different antenna pairs showing an amplitude maximum filter applied to both a short and long baseline

From Fig. 12 we observe a reasonably effective trimming of strong RFI sources for both the short- and long baseline pairs. While stricter amplitude limits can be applied, caution must be taken not to cut the already limited clean data. Other statistical filtering methods like Z-score, which calculates a deviation score at each point:

$$Z = \left| \frac{x - \bar{x}}{\sigma} \right| \tag{7}$$

and then filters for amplitude under a specified threshold does work similarly to cutoff amplitude applied in Fig. 12, but effectiveness is sensitive to the variance in the data, and therefore, the RFI itself leads to effectiveness for short baselines where RFI is much stronger, therefore larger Z scores, than for longer baselines where RFI is weaker and therefore often does not get masked. Additionally, both the computationally and conceptually similar Inter Quartal Range (IQR) clipping and Sigma clipping algorithms were tested with even less success in Z-score statistics. Because of the above-mentioned reasons, the conceptually and computationally simplest method of applying an amplitude cut-off was chosen as a first attempt at filtering strong RFI signals from the dataset. Due to the nature of RFI, it is often impossible to recover data from RFI regions without exact knowledge of the RFI source, and therefore, the data points were cut from the dataset. If amplitude values at these cut frequencies need to be recovered, an interpolation or mean background method could serve as a stopgap.

3.7 Satellite Analysis

Satellite TLE data was requested for a 100 'Weather' and 'Navigation' satellites (with known NORAD identification numbers) from the 30th of January 2019 till the 1st of February 2019 from the Celestrak Special Data Request feature made available on the Celestrak website.

The data was returned in `.txt` files for each satellite requested containing the TLE data. 26 of the requested satellites did not return data due to it not being in commission at the time of the observation.

The TLE files were processed using `ephem`, a Python package designed for performing high-precision astronomy calculations (Rhodes 2023). This package allows the user to specify an 'Observer' object with the following parameters:

Table 7: ephem.Observer() object parameters

Latitude (deg)	37.775
Longitude (deg)	-122.419
Elevation (m)	1087
Date	2019/1/31
Time	12:57:00

and a 'Satellite' Object created from the TLE data in the Observer frame. The satellite object contains many parameters, one of which being the altitude, which was used to determine whether the satellite would have been above the horizon at the time and position of the observation (altitude $> 0°$). 20 of the 74 satellites were found to be above the horizon. The target right ascension, α and declination, δ, were converted from degrees $(5.1462°, -1.1120°)$ to radians (0.0898 rad, -0.0194 rad). Each satellite's right ascension and declination were extracted from each 'Satellite' object. These coordinates were compared, and the angular separation between the target and each satellite was found using the Haversine formula given below:

$$d = 2\arcsin\sqrt{\sin^2\left(\frac{\Delta\delta}{2}\right) + \cos(\delta_1)cos(\delta_2)\sin^2\left(\frac{\Delta\alpha}{2}\right)} \tag{8}$$

The angular separation between the satellite and the source changes slightly over the observation period, but due to the short observation time, a mean of these angular separation values was found for each satellite. These mean angular separations from the 20 potentially visible satellites are given in Table 8 given in Appendix B. The closest separation was measured to be 50.057° by COSMOS 2436 in the GLONASS-M group.

This separation is compared to the resolution, and the satellites that fall within this radius are considered to have possibly been visible during the observation. The largest resolution is found in Table 2 is 1.199×10^{-2} rad which is equivalent to 0.687°. Therefore, no satellites would have crossed our observation, but some emissions may have been picked up in the 'side lobes' and seen as slightly weaker signals in our spectrum. If the RFI signal is particularly strong, it may still be detected, as seen in the GLONASS L1 and L2 bands shown below in Fig. 13.

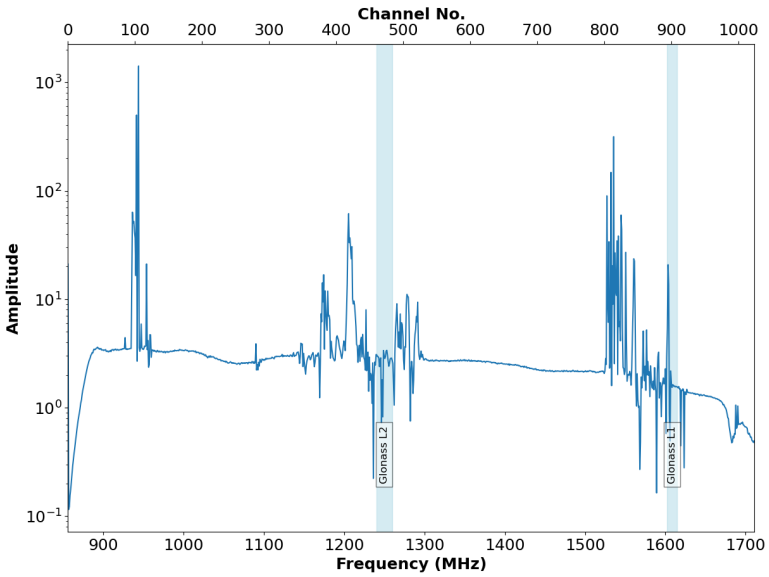


Figure 13: Satellite Bandwidths highlighted on a 1D cut across the visibility showing the amplitude of the signal received across a range of frequencies (MHz) in the L-band.

References

Zwillinger (Nov. 2002). "Probability and Statistics". In: *CRC Standard Mathematical Tables and Formulae* Not available, pp. 631–742. ISSN: Not available. DOI: 10.1201/9781420035346-10. URL: https://dx.doi.org/10.1201/9781420035346-10.

Luo, Jianwen, Kui Ying, and Lijing Bai (July 2005). "Savitzky–Golay smoothing and differentiation filter for even number data". In: *Signal Processing* 85, pp. 1429–1434. DOI: 10.1016/j.sigpro.2005.02.002.

Green, Jennifer L. et al. (2023). "Descriptive statistics". In: *International Encyclopedia of Education(Fourth Edition)*. Elsevier, pp. 723–733. ISBN: 9780128186299. DOI: 10.1016/b978-0-12-818630-5.10083-1. URL: http://dx.doi.org/10.1016/B978-0-12-818630-5.10083-1.

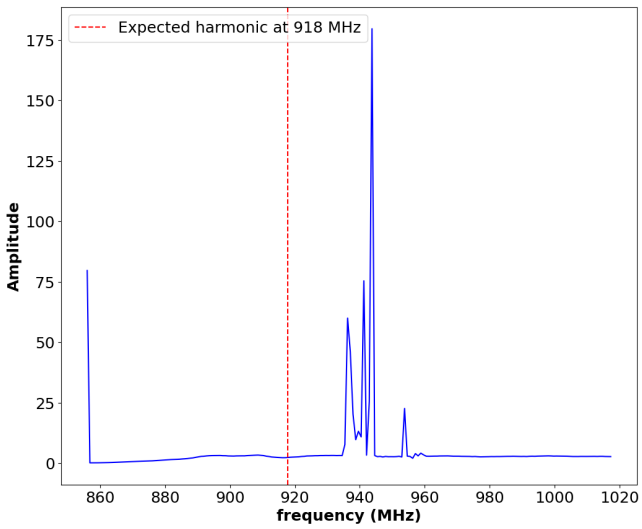
Rhodes, Brandon (2023). *PyEphem: Astronomical computations for Python*. Version 4.1. URL: https://rhodesmill.org/pyephem/.

A Satellite separation

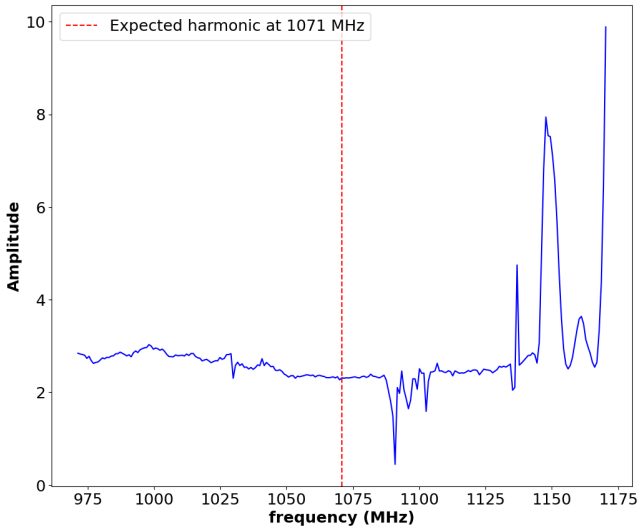
Table 8: Angular separation (deg) between each of the 20 satellites and the source ($5.1462^\circ, -1.1120^\circ$*)*

Satellite	Angular Separation (deg)
BEIDOU-2 G4	112.548
BEIDOU-2 M3	159.945
BEIDOU-2 M6	129.664
BEIDOU-3 M12	90.727
BEIDOU-3 M14	148.438
BEIDOU-3 M15	147.855
BEIDOU-3 M3	83.853
BEIDOU-3 M8	98.431
BEIDOU-3S M2S	143.854
COSMOS 2432 [GLONASS-M]	87.231
COSMOS 2433 [GLONASS-M]	131.898
COSMOS 2436 [GLONASS-M]	50.057
COSMOS 2476 [GLONASS-M]	86.389
COSMOS 2492 [GLONASS-M]	139.394
COSMOS 2527 [GLONASS-M]	151.101
GSAT0101 [GALILEO-PFM]	125.005
GSAT0204 [GALILEO 8]	106.352
GSAT0212 [GALILEO 16]	128.320
GSAT0213 [GALILEO 17]	72.423
GSAT0214 [GALILEO 18]	173.740

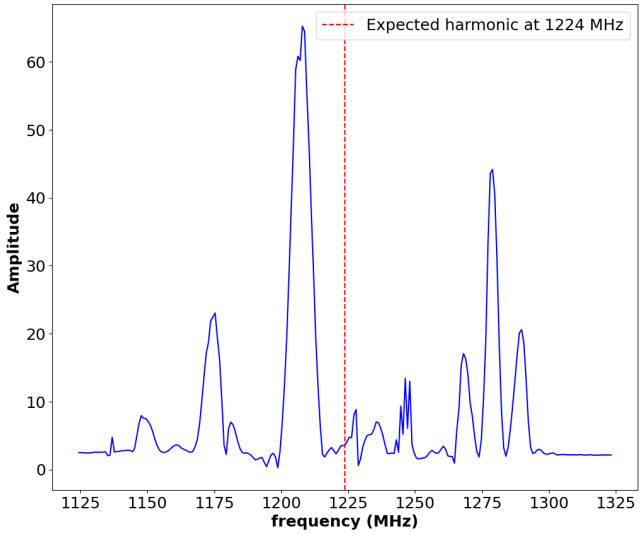
B Harmonic plots



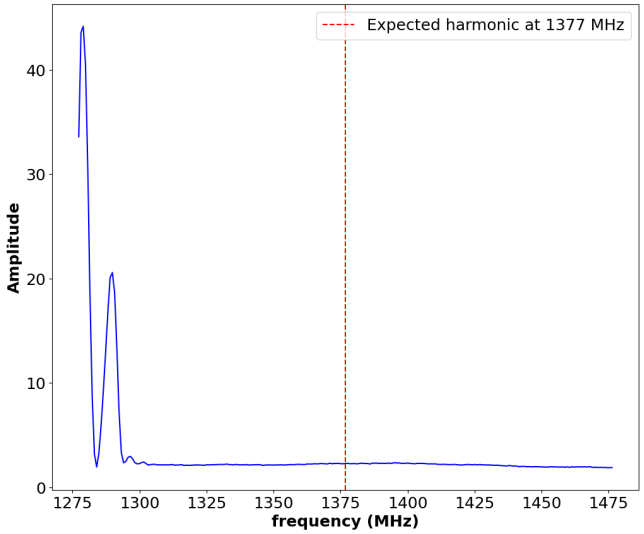
(a) Expected frequency of 6th harmonic



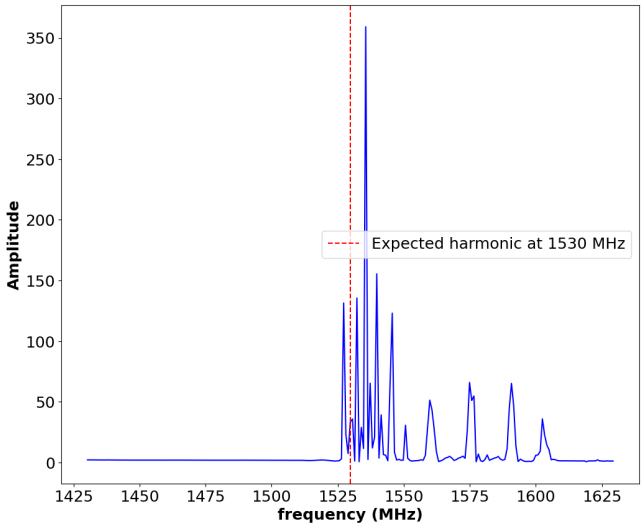
(b) Expected frequency of 7th harmonic



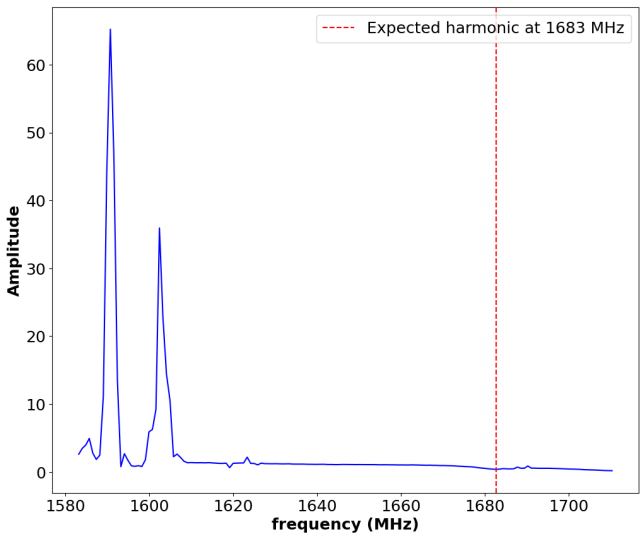
(c) Expected frequency of 8th harmonic



(d) Expected frequency of 9th harmonic



(e) Expected frequency of 10th harmonic



(f) Expected frequency of 11th harmonic

Figure 14: Frequency vs Amplitude centred at the expected frequency of the nth harmonic. Data were taken for the shortest baseline pair.