Observational Techniques 2: Assignment 2

Exploration of MeerKAT Data

Thambiran, Abigail
THMABI003@myuct.ac.za

Campher, Francois
CMPFRA003@myuct.ac.za

October 18, 2024

Abstract

We present a short exploration of a radio data set from the MeerKAT telescope, observing PKS1934-63. This report focuses on exploring the data set and investigating the various sources of RFI found within the data set. Additionally we explore processes such as detrending, RFI filtering and identifying satellites that could have impacted the results. This report is meant to highlight the difficulty in dealing with RFI, highlighting the sophistication of modern RFI detection and filtering pipelines.

1 Introduction

We are presented with a .ms file which contains a multitude of data about the source observed. We make use of the python package `dask-ms[xarray]` to extract the data from the .ms file. The code for this assignment was created using Google Colab. Note that the notebook will contain various repeated import statements due to the nature of the notebook environment and not wanting to rerun all code every time Colab disconnects. This is sloppy programming, but due to the need to use Google Colab to make `dask-ms` function it was the only option. The Colab can be found with this link: https://colab.research.google.com/drive/1TsY5hZTpmChHmrSLV-23tP-n1uJMJ3oO#scrollTo=weHAKygQlSUS and the GitHub that contains the .ipynb file can be found: https://github.com/Darkabyss1702/OT2_Assignment2

2 Results

Exploratory data analysis (EDA)

1. The start and end date of the observation:
   this was obtained by extracting the observation table from the .ms data folder, and then extracting start and end time from the observation table. The time obtained just by doing this was given in seconds (Modified Julian Date) and so we had to convert the seconds to days, and then converted that to datetime.

   (a) **start: 2019-01-31 12:56:12.122526**

   (b) **end: 2019-01-31 12:59:48.031192**

2. The target that was observed: **PKS1934-63**
   This was obtained by extracting the field table form the.ms data folder first, and then extracting the actual target name (”NAME” in the field table).

3. The bandwidth of the observation: **856.0 MHz**

   This was obtained by extracting the spectral window table similar to the above steps, and then extracting the bandwidth (”BANDWIDTH” in the spectral window table) and then converting to MHz.

4. The channel width in kHz: **835.9375 kHz**

   Similar to the above step for bandwidth, but extracting the channel width (”CHAN_WIDTH” in the spectral window table) and converting to kHz.

5. The total number of antennas used: 50

   This was obtained by extracting the antenna 1 and 2 list from the measurement folder, and finding the length of the combined array.

Understanding the UVW

1. **UV coverage plot and description:**

   Refer to Figure 1. The blue points represent the measured UV points corresponding to the baselines between antenna pairs, while the red points represent their conjugate (mirrored) counterparts, due to the complex nature of visibility data. The distribution is denser near the center, indicating that the observation captured more low spatial frequencies (larger-scale structures), with sparser coverage towards the edges, representing higher spatial frequencies (smaller-scale structures). The relatively symmetric and spread-out coverage suggests a good range of baseline lengths and orientations, which will help produce a higher-quality image with fewer artifacts. The plot is also symmetrical around the origin so the resultant image will likely have good quality in all directions. There are more shorter baselines than longer ones since the distribution is denser near the centre. In summary, this UV coverage plot shows a dense, elliptical pattern, indicating that the antennas observed the target for an extended period, allowing Earth’s rotation to fill in many baselines. The short baselines near the center suggest the array captured large-scale structures, while longer baselines farther out will help resolve finer details. However, there are a few gaps along the longer baselines, meaning the final image may have some directional artifacts [3].

2. Which antenna pairs make the shortest and longest baseline

   Shortest baseline is between antennas: **m000 and m002**. Shortest baseline length: 29.26 meters

   Longest baseline is between antennas: **m048 and m060**. Longest baseline length: 7697.58 meters

   We first find the indices of the shortest non-zero baseline, and maximum baseline using `np.argmin` and `np.argmax` respectively. Then we find the first and last element in each of those lists to find out the pairs of telescopes that make up the shortest and longest baselines. Finally, we determined the names of the two pairs by referencing them in the antenna table of the .ms folder.

3. What is the expected resolution on the short and long baseline?

   Resolution on shortest baseline: **1480.55 arcseconds**.

   Resolution on longest baseline: **5.63 arcseconds**

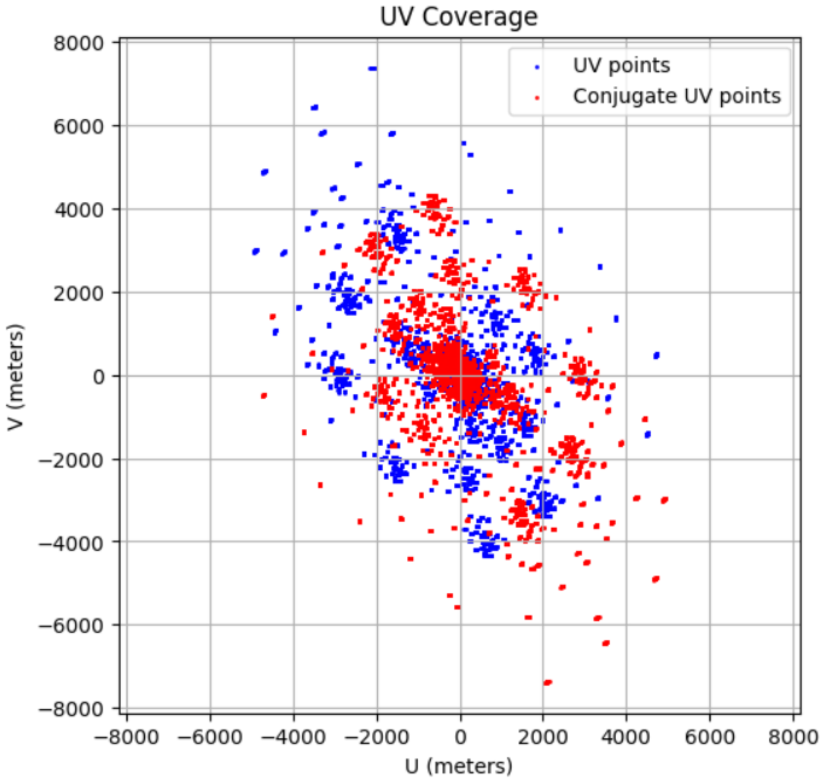

Figure 1: UV Coverage

This is achieved by calculating the baseline lengths for shortest and longest baselines (extracted from antenna table), then calculate both resolutions using the fact that the resolution of baseline = wavelength / baseline. This resolution is given in radians, so we finally convert to arcseconds.

4. Calculate the delay that needs to be applied on the shortest and longest baseline if you are detecting a radio source in the sky.

   Delay on shortest baseline: **9.75e-08 seconds**

   Delay on longest baseline: **2.57e-05 seconds**

   The delays are calculated such that they are the baselines divided by the speed of light.

5. What is the relationship between the UVW and the ENU position?

   The UVW coordinate system is dynamic and is tied to the direction of the source being observed and changes as the Earth rotates. The East-North-Up (ENU) system is a fixed coordinate system used to describe the positions of the antennae on Earth, where E and N point in East and North directions respectively, and up points towards zenith.

   The relationship between the two coordinate systems are such that ‘U’ points along the East-West (E-W) direction towards the East, ‘V’ points along the North-South (N-S) direction towards the North Celestial Pole, and the ‘W’ coordinate points to and follows the source phase tracking center and represents the delay distance, τ, between the two antennas. [3]

Radio Frequency Interference (RFI)

1. Define 3 regions where you notice RFI and explain what are the major sources of RFI in these regions:

   Consider Figures 3 and 4: We note three very distinct regions of high amplitude which likely corresponds to RFI. These sources appear around channels 100, 410 and 813, which corresponds to frequencies of roughly (970 MHz, 1207 MHz and 1500 MHz). We will investigate the RFI around channel 410 mostly for the purposes of this project. As to the source of the RFI, this is a little more tricky to ascertain, considering the number of factors that can play a role. For the RFI observed at 970 MHz (channel 100), a likely candidate is 3G mobile networks as they operate near this range [6]. For the RFI seen at 1207 MHz (roughly channel 410) the most likely candidate is some kind of Global Navigation Satellite System (GNSS) since they are known to operate in the L2 band [2]. Lastly, the RFI seen at roughly 1500 MHz (channel 813) is also likely from satellites.

2. What can you infer from the histogram distribution of the RFI signal and a 'clean' region?

   We select channels 401 for RFI and channel 600 for the clean region. We loop through all 50 antenna baseline pairs to produce a complete histogram of the data set. The result can be seen in figure 2. Both the clean and contaminated regions have relatively low counts, close to 0. This is due to the fact that most radio sources are not very bright emitters. The counts of the contaminated region is comparable to that of the clean, suggesting that the signal is coming from non-astronomical signals or interference contaminating the data. The clean region shows a more compact distribution than the contaminated, meaning it only contains expected noise or weaker astronomical signals.

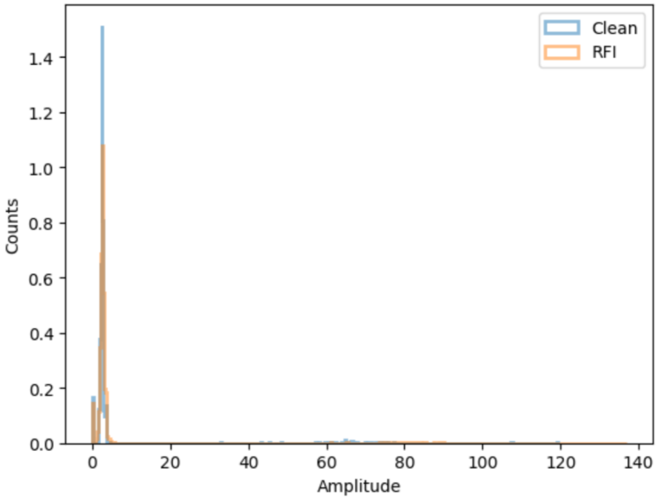

Figure 2: Histogram distribution of the RFI signal and a clean region.

3. Extract the measure of spread in the contaminated and 'clean' region and explain your findings.

   We calculate the standard deviation to quantify the measure of spread in the contaminated and clean regions. We calculate the following finding the mean of the two regions we chose (`np.std`) and then the mean (to measure the spread from the mean) using *np.std* from the `numpy` package.

   Standard Deviation (RFI region): **14.87**

Standard Deviation (Clean region): **12.58**

For the contaminated region, a standard deviation of 14.87 relative to the mean (5.39 - See in code) indicates that the data points are widely spread around the mean (high variability). This wide spread suggests the presence of strong interference. RFI signals are often spiky, meaning they introduce extreme values or bursts into the data, causing a high standard deviation.

For the clean region, a standard deviation of 12.58 relative to the mean (4.85 - See in code) indicate that the data points also have high variability. This probably suggests that there is still RFI in our defined clean region or natural variations in the flux measurements.

4. Extract the kurtosis and skewness of the 2 regions. What can you conclude from your findings?

   Clean Region Kurtosis: 27.05

   RFI Region Kurtosis: 25.93

   Clean Region Skewness: 5.21

   RFI Region Skewness: 5.14

   We calculate the kurtosis and skewness using the built-in `kurtosis` and `skew` functions from `scipy.stats`.

   Kurtosis is a measure of whether data are heavy-tailed or light-tailed relative to a normal distribution [5], i.e. a measure of outliers in the data. Both regions exhibit extremely high kurtosis (> 3), meaning that both datasets contain many extreme values. This is intuitive for the RFI, but not for the clean region, suggesting that there is still RFI in our defined clean region or natural variations in the flux measurements. Either way this is consistent with result of question 3 in RFI section.

   Skewness is a measure of asymmetry of the data. Both distributions (clean and contaminated) are highly positively skewed, meaning that there are several high-amplitude events (likely RFI contamination) causing a long right-hand tail. The similarity in skewness suggests that both datasets exhibit asymmetry towards high values, with extreme events in both the clean and RFI regions.

5. Select a short baseline and a long baseline visibilities and explain what you notice in the contaminated region.

   See Figures 3 and 4. Contaminated regions are clearly seen around channel number 810, 100 in both plots. However, more RFI contamination is seen in the shorter baseline than the longer one. The RFI in the prior mentioned channels are persistent, so these channels are consistently contaminated during the observation. Shorter baselines are more sensitive to large-scale (diffuse) emission and ground-based interference, so the contamination (horizontal bands) [1] could indicate persistent RFI sources affecting specific frequencies. In the longest baseline plot, the contamination appears more localised in time and frequency (thin bright lines). This indicates sporadic RFI bursts that affect only a few channels.

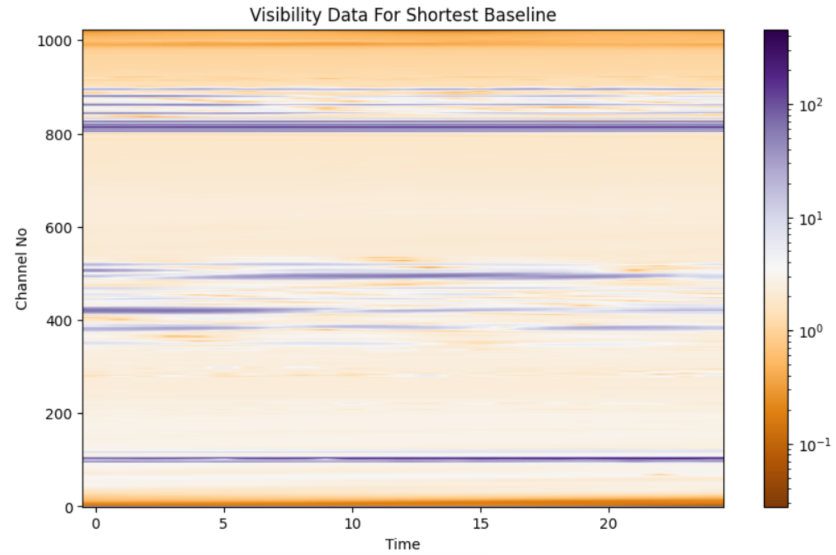

Figure 3: Visibility Data For Shortest Baseline

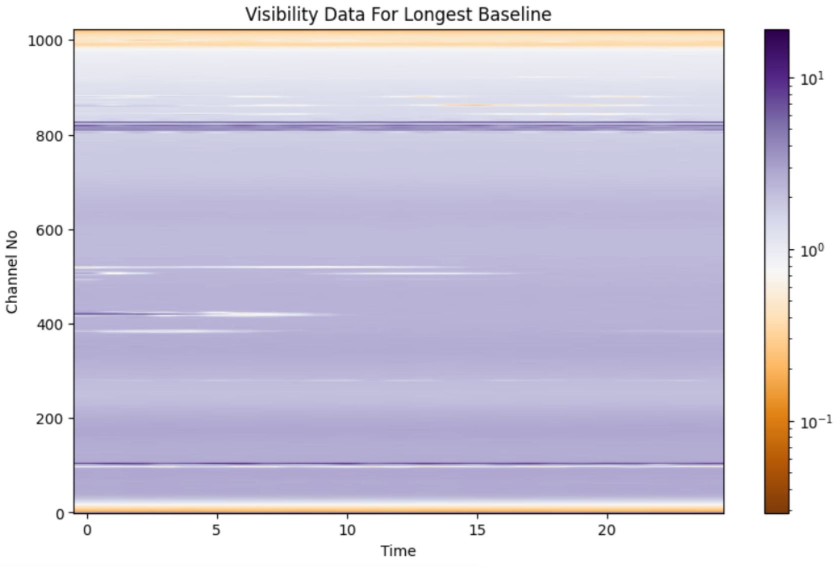

Figure 4: Visibility Data For Longest Baseline

Owing to the more pronounced RFI effects in the short baseline visibility, we opt to stick to this antenna pair for the investigation of the RFI we observe going forward in this report.

6. Extract the HPBW (Half Power Beam Width) of the RFI at ~~1380 MHz~~ 1500 MHz.

When considering figure 5, we note that there is no obvious source of RFI at 1500 MHz. Instead, to keep consistency we opt to calculate the HPBW at 1207 MHz as this is the RFI we have been focusing on in this project so far. We implement a rather simple method to calculate the HPBW for this peak. Firstly, we find the index of the channel that corresponds to the frequency (or is the closest). Next we can extract the visibility data around this frequency. Then we can calculate the half power value, which is just `peak_value - 0.5`, since the plot is in log scale. From here we can calculate the indices that cross the signal at the half-power value, and then finally we can calculate the half power values in MHz by making use of these indices: `hpbw_mhz = (frequencies[right_idx] - frequencies[left_idx]) / 1e6` We obtain a HPBW at 1207 MHz of 10.03 MHz. The results are also shown in figure 6.

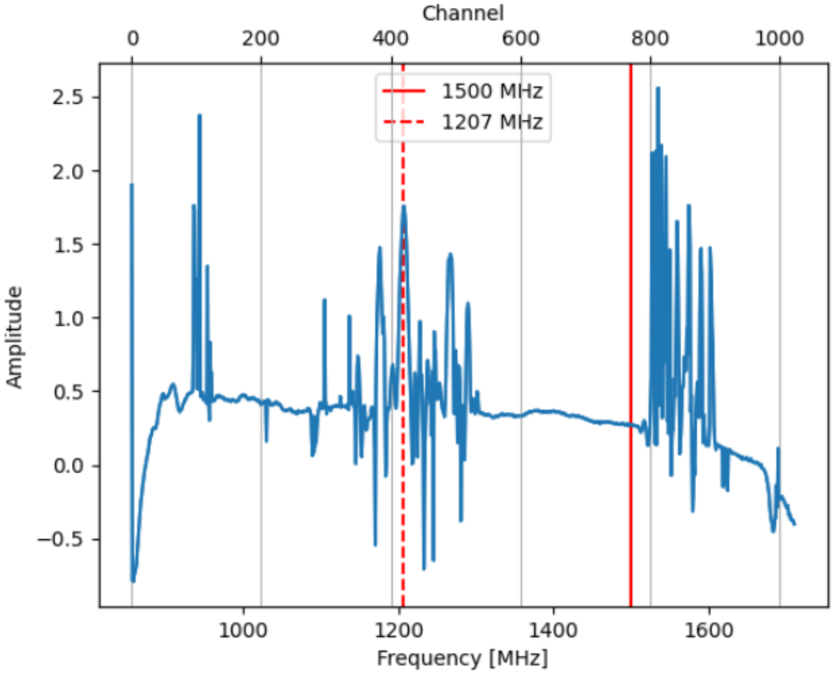

Figure 5: This plot represents a 1D time slice of the visibility data for the shortest baseline at a time index of 5.

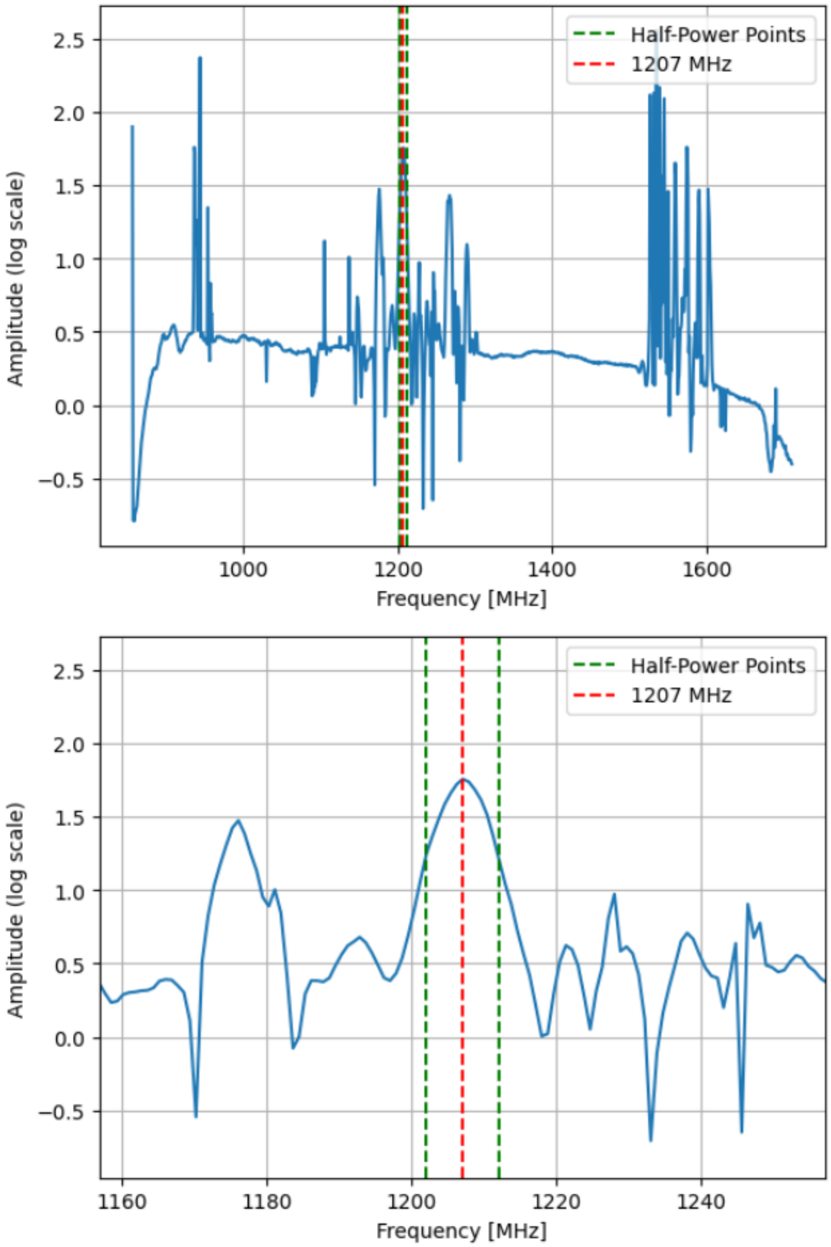

Figure 6: Plots showing the results of the HPBW extraction.

7. How many channels are actually impacted by this source of RFI?

   Number of channels impacted by RFI: 13

   This is determined by first finding the visibility data around 1207 MHz, the peak amplitude of the RFI at 1207 MHz, the half-power value, and indices where the signal crosses the half-power value. Then we find the number of impacted channels by subtracted the max and min indices calculated.

8. What redshift will this frequency correspond to?

   The redshift corresponding to 1207 MHz is: 0.18

This is calculated using equation 1, where f_{em} is the emitted frequency and f_{obs} is the observed frequency. We assume here that the f_{obs} correlates to HI, since we know the source is a radio galaxy which are known for generally having high levels of HI.

$$z = (f_{em}/f_{obs}) - 1 \tag{1}$$

9. Assuming a source of RFI has a fundamental frequency of 153 MHz, will you see any effect of this RFI on the current data that you have? Explain your reasoning and also any plot if you notice the impact of this RFI.

   RFI appears at its fundamental frequency, but it can also be seen at harmonics, or integer multiples of the fundamental frequency. Harmonics can cause interference at these higher frequencies and thus must be accounted for too. Equation 2 describes how harmonic frequencies (f_n) are calculated from fundamental frequencies (f_0).

   $$f_n = nf_0 \tag{2}$$

   We find that the 8th harmonic (n=8) is close to 1207 MHz, specifically (1224 MHz) which happens to be the reference frequency that we have discussing above in detail. This implies that we do indeed see an effect of RFI of a fundamental frequency with 153 MHz in our data. This RFI at 1207 MHz is plotted in Figure 5.

10. What is detrending in signal processing?
    Detrending refers to the process of removal of a systematic bias (trend) in a signal. This is done to isolate real fluctuations in the data set so as to better understand the underlying behaviour of the data [4].

    - Using a time slice of your data for any baseline, detrend your signal and justify any method you have used:
      For this task, we simply made use of the `scipy.signal` packages' `detrend` function. This function uses a simple linear detrending function. Owing to time restrictions, we did not investigate more sophisticated detrending methods, however we instead focus on several methods to test the quality of the detrend applied. The results of the detrend are show in figure 7. We note that the detrending appears to be quite unsuccessful. To emphasise just how poor the performance was, we make use of a number of statistical tests

    - Explain how confident you are with your detrended signal?:
      We opted to show a number of statistical tests to emphasize how poor the detrending is:
      Firstly, We make use of `statsmodels.graphics.tsaplots`'s `plot_acf` function to plot the Autocorrelation Function (ACF). This shows the autocorrelation of residuals, so if the deterending was effective we would expect to see that the autocorrelation trends towards zero, which is clearly not the case in figure 8.

      We also make use of `sklearn.linear_model`'s `LinearRegression` function to fie a linear regression line to the data, and therefore extract the R^2 value. We get a value of $R^2 = 0.001$, which is clearly very poor.

Additionally, we also check for stationarity by making use of an Augmented Dickey-Fuller Test implemented through `statsmodels.tsa.stattools`'s `adfuller` function. Ideally, your detrended data should become stationary, i.e. it's statistical properties (such as mean and variance) should become constants over time. The ADF test is a well known way to determine this. Essentially if your p-value is less than 0.05, then the residuals can be said to be stationary, otherwise they are not. In this case we obtain a p-value of 0.95 which is clearly not satisfactory.

In conclusion, we are clearly not satisfied with the detrend. A more sophisticated approach is required, but due to time constraints it was not possible to implement one.

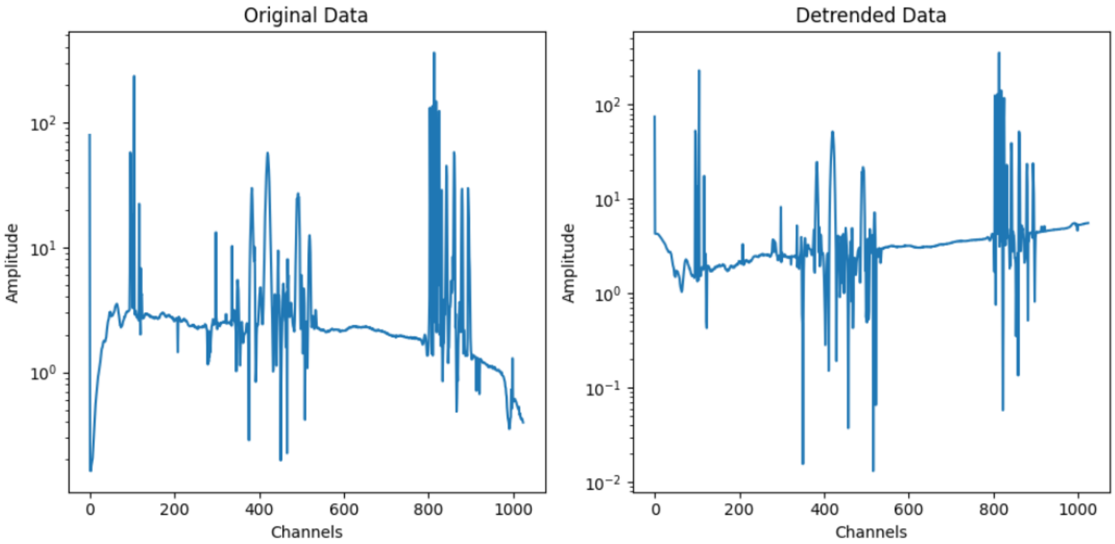

Figure 7: Plots showing before and after detrending. It should be noted that the time slice is take at a time 5 seconds an that the visibility data used is that for the shortest baseline as the RFI is more pronounced at this baseline.

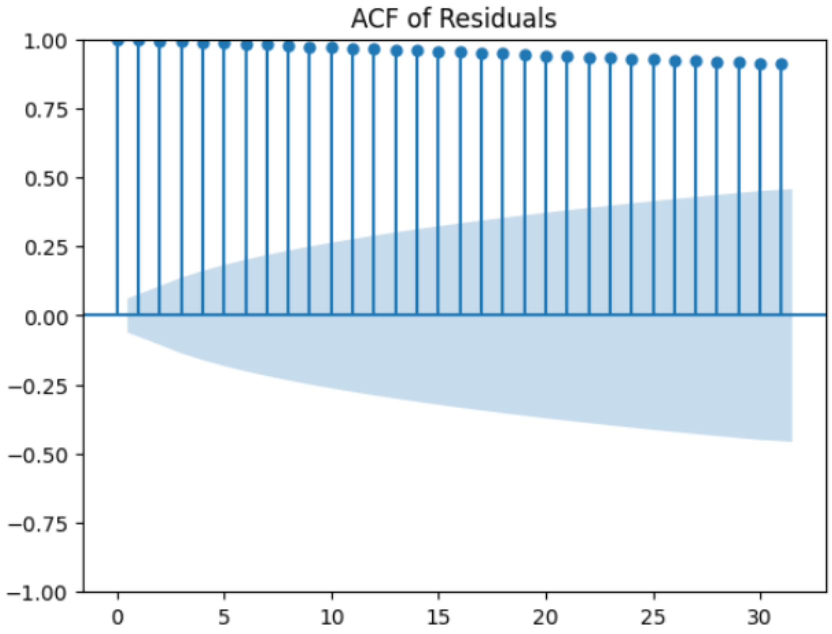

Figure 8: Plot showing the ACF of the detrended data. Note that the lack of axis naming seems to be convention for this metric and so it was left as is.

11. Use one filtering method to detect any RFI, explain the method and how well or not is it in detecting RFI.

    We decided to implement a simple filtration method: Fourier Threshold Clipping. Essentially this method entails performing a fast Fourier transform on the visibility data and then removing the higher power frequencies past a certain threshold, which would most probably correspond to RFI. We made use of the `scipy.fft` for the Fourier transform. We decided to set the threshold frequency as 10% of the maximum power frequency, since this was unlikely to remove the lower amplitude signal from the radio source, while ensuring the removal of very high power sources, which would correspond to strong radio signals which would most likely correspond to RFI. The results of the filtered data is shown in figure 9. We can clearly see that the filtration method did do a good job at removing some RFI, however not all of the RFI is removed. We can clearly see a reduction in the RFI around channel 410, but the RFI at 813 and 100 persist. This is a rather crude way to filter the signal and a more complex method is perhaps required to properly identify and remove all sources of RFI.

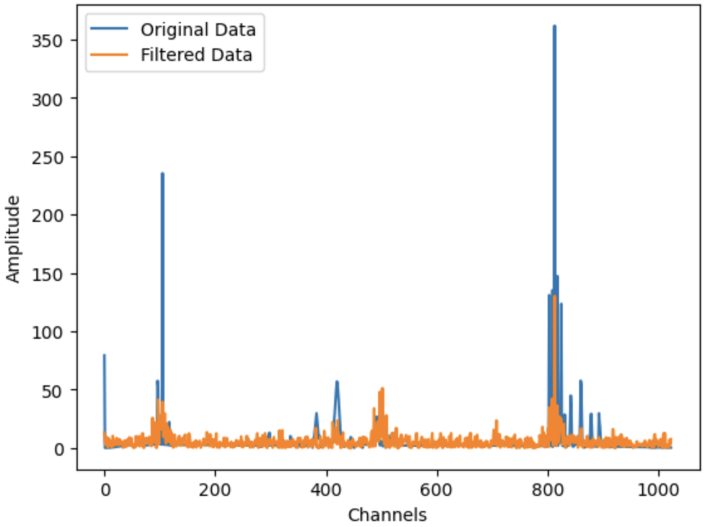

Figure 9: Plot showing the original signal vs the filtered signal.

12. Locate any satellite/s that could have impacted that specific observation and calculate the angular separation of the satellite/s from your phase center.

    To achieve this we make use of `skyfield.api`'s `load` , `EarthSatellite`, `Topos` functions. We made use of the TLE finder website (`https://celestrak.org/NORAD/elements/`) and requested data around the dates of observation. Our code for this section calculates the angular separation between the satellite positions and the radio source PKS1934-63 at a specific time. We choose to measure angular separation from the source since it is assumed to be in the phase centre, and this is also slightly simpler since the source has well defined RA and DEC coordinates. We first load the TLE data (Two-Line Element sets) from a text file, which contains satellite orbit information. For each satellite, we compute the RA and Dec coordinates at the observation time (2019-01-30 12:56:12). Using trigonometry, we calculate the angular separation between the satellite and PKS1934-63. The code ensures that the angular separation is printed only once for each unique satellite name to avoid redundancy.

    The angular separation is calculated using the spherical law of cosines as defined in Equation 3, where RA_1 and Dec_1 refer to the coordinates of the source and RA_2 and Dec_2 refer to the coordinates of each satellite.

$$\cos\theta = \sin(Dec_1)\sin(Dec_2) + \cos(Dec_1)\cos(Dec_2)\cos(RA_1 - RA_2) \tag{3}$$

    We present the results of the angular separation in section 2. We note that this is not a perfect method to identify satellites that could have impacted the observations since we would need to look at the frequency ranges at which these satellites operate and compare it to the data set. Owing to time constraints we were unable to implement this selection.

Angular Separations from PKS1934-63

- Angular separation to BEIDOU-2 IGSO-7: 59.26 degrees
- Angular separation to BEIDOU-3 M2: 64.24 degrees

- Angular separation to GSAT0221 (GALILEO 25): 108.24 degrees
- Angular separation to BEIDOU-3 M9: 97.59 degrees
- Angular separation to GSAT0204 (GALILEO 8): 64.72 degrees
- Angular separation to BEIDOU-2 IGSO-5: 149.82 degrees
- Angular separation to BEIDOU-2 G5: 87.46 degrees
- Angular separation to GSAT0201 (GALILEO 5): 166.46 degrees
- Angular separation to BEIDOU-2 M6: 113.79 degrees
- Angular separation to BEIDOU-3 M4: 58.58 degrees
- Angular separation to COSMOS 2485 [GLONASS-M]: 132.10 degrees
- Angular separation to BEIDOU-3S M2S: 84.21 degrees
- Angular separation to BEIDOU-3 M14: 132.37 degrees
- Angular separation to GSAT0208 (GALILEO 11): 116.01 degrees
- Angular separation to BEIDOU-2 G4: 116.92 degrees
- Angular separation to BEIDOU-2 IGSO-6: 101.48 degrees
- Angular separation to BEIDOU-3S IGSO-2S: 145.21 degrees
- Angular separation to COSMOS 2432 [GLONASS-M]: 131.37 degrees
- Angular separation to GSAT0202 (GALILEO 6): 13.19 degrees
- Angular separation to BEIDOU-3 M13: 93.86 degrees
- Angular separation to BEIDOU-3 M8: 125.67 degrees
- Angular separation to BEIDOU-2 M4: 11.22 degrees
- Angular separation to BEIDOU-3 M16: 53.54 degrees
- Angular separation to BEIDOU-3 M1: 47.64 degrees
- Angular separation to BEIDOU-3S M1S: 32.03 degrees
- Angular separation to GSAT0215 (GALILEO 19): 116.84 degrees
- Angular separation to BEIDOU-2 IGSO-4: 63.86 degrees
- Angular separation to BEIDOU-3 M7: 167.55 degrees
- Angular separation to GSAT0206 (GALILEO 10): 42.44 degrees

- Angular separation to COSMOS 2476 [GLONASS-M]: 132.41 degrees
- Angular separation to COSMOS 2477 [GLONASS-M]: 48.08 degrees
- Angular separation to GSAT0203 (GALILEO 7): 159.31 degrees
- Angular separation to COSMOS 2456 [GLONASS-M]: 100.46 degrees
- Angular separation to GSAT0207 (GALILEO 15): 93.65 degrees
- Angular separation to GSAT0212 (GALILEO 16): 123.22 degrees
- Angular separation to GSAT0213 (GALILEO 17): 64.29 degrees
- Angular separation to GSAT0218 (GALILEO 22): 138.27 degrees
- Angular separation to COSMOS 2457 [GLONASS-M]: 47.39 degrees
- Angular separation to BEIDOU-3 M18: 88.68 degrees
- Angular separation to BEIDOU-3 M6: 58.49 degrees
- Angular separation to COSMOS 2529 [GLONASS-M]: 62.82 degrees
- Angular separation to COSMOS 2436 [GLONASS-M]: 81.83 degrees
- Angular separation to BEIDOU-3 G1: 115.35 degrees
- Angular separation to COSMOS 2433 [GLONASS-M]: 179.34 degrees
- Angular separation to COSMOS 2522 [GLONASS-M]: 30.84 degrees
- Angular separation to BEIDOU-3S IGSO-1S: 151.20 degrees
- Angular separation to COSMOS 2492 [GLONASS-M]: 137.23 degrees
- Angular separation to GSAT0210 (GALILEO 13): 148.22 degrees
- Angular separation to BEIDOU-3 M11: 40.79 degrees
- Angular separation to COSMOS 2527 [GLONASS-M]: 74.95 degrees
- Angular separation to BEIDOU-3 M15: 37.99 degrees
- Angular separation to BEIDOU-2 IGSO-2: 146.51 degrees
- Angular separation to COSMOS 2434 [GLONASS-M]: 39.19 degrees
- Angular separation to COSMOS 2500 [GLONASS-M]: 84.44 degrees
- Angular separation to BEIDOU-3 M3: 86.18 degrees
- Angular separation to GSAT0101 (GALILEO-PFM): 28.17 degrees

- Angular separation to COSMOS 2514 [GLONASS-M]: 94.28 degrees
- Angular separation to COSMOS 2501 [GLONASS-K]: 142.93 degrees
- Angular separation to GSAT0216 (GALILEO 20): 63.24 degrees
- Angular separation to BEIDOU-3 M17: 139.81 degrees
- Angular separation to GSAT0205 (GALILEO 9): 101.05 degrees
- Angular separation to GSAT0214 (GALILEO 18): 109.97 degrees
- Angular separation to BEIDOU-2 M3: 59.52 degrees
- Angular separation to BEIDOU-2 IGSO-3: 121.14 degrees
- Angular separation to BEIDOU-2 IGSO-1: 60.24 degrees
- Angular separation to BEIDOU-3 M10: 142.34 degrees
- Angular separation to BEIDOU-2 G6: 98.82 degrees
- Angular separation to GSAT0211 (GALILEO 14): 31.73 degrees
- Angular separation to BEIDOU-3 M12: 90.56 degrees
- Angular separation to COSMOS 2460 [GLONASS-M]: 12.82 degrees
- Angular separation to GSAT0103 (GALILEO-FM3): 56.88 degrees
- Angular separation to COSMOS 2475 [GLONASS-M]: 72.98 degrees
- Angular separation to BEIDOU-3 M5: 56.32 degrees
- Angular separation to BEIDOU-2 G7: 110.05 degrees

3 Conclusions

We were able to extract some basic information about the observation, however when it comes to the more complex task of dealing with RFI in this data set we encountered a number of challenges. This very clearly highlights how complex dealing with RFI can be. This short report has emphasized the need for sophisticated RFI filtering methods when dealing with radio data.

References

[1] Ntsikelelo Charles et al. “On the use of temporal filtering for mitigating galactic synchrotron calibration bias in 21 cm reionization observations”. In: *Monthly Notices of the Royal Astronomical Society* 522.1 (2023), pp. 1009–1021.

[2] RF Integration Inc. *Global Navigation Satellite Systems (GNSS)*. Last accessed 17 October 2024. URL: http://www.rfintegration.com/gps-gnss.html.

[3] Justin Jonas. *Coordinate systems and UV mapping*. Last accessed 17 October 2024. 2021. URL: https://www.sarao.ac.za/lessons/coordinate-systems-and-uv-mapping/.

[4] Amberle McKee. *A Data Scientist's Guide to Signal Processing*. Last accessed 17 October 2024. 2023. URL: https://www.datacamp.com/tutorial/a-data-scientists-guide-to-signal-processing.

[5] National Institude of Standard and Technology. *1.3.5.11. Measures of Skewness and Kurtosis*. Last accessed 17 October 2024. URL: https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm#:~:text=Kurtosis%20is%20a%20measure%20of,would%20be%20the%20extreme%20case..

[6] Yeastar. *GSM/3G/4G LTE Trunk Overview*. Last accessed 17 October 2024. URL: https://help.yeastar.com/en/p-series-appliance-edition/administrator-guide/gsm-3g-4g-lte-trunk-overview.html.