A Comparative Analysis to Deal with Missing Spectral Information Caused by RFI in Cosmological H I 21 cm Observations

Arnab Chakraborty 1,2,3, Abhirup Datta 1, and Aishrila Mazumder 1

1 Discipline of Astronomy, Astrophysics and Space Engineering, Indian Institute of Technology Indore, Indore 453552, India; phd1601121009@iiti.ac.in, arnab.chakraborty2@mail.mcgill.ca
2 Department of Physics, McGill University, 3600 rue University, Montreal, QC H3A 2T8, Canada
3 McGill Space Institute, McGill University, 3550 rue University, Montreal, QC H3A 2A7, Canada

Received 2020 October 5; revised 2022 January 29; accepted 2022 March 7; published 2022 April 18

Abstract

We investigate the effect of radio-frequency interference (RFI) excision in estimating the cosmological H I 21 cm power spectrum. Flagging of RFI-contaminated channels results in a nonuniform sampling of the instrumental bandpass response. Hence, the Fourier transformation of visibilities from frequency to delay domain contaminates the higher foreground-free delay modes, and separating the spectrally fluctuating H I signal from spectrally smooth foregrounds becomes challenging. We have done a comparative analysis between two algorithms, one-dimensional CLEAN and least-squares spectral analysis (LSSA), which have been used widely to solve this issue in the literature. We test these algorithms using the simulated SKA-1 Low observations in the presence of different RFI flagging scenarios. We find that, in the presence of random flagging of data, both algorithms perform well and can mitigate the foreground leakage issue. But CLEAN fails to restrict the foreground leakage in the presence of periodic and periodic plus broadband RFI flagging and gives an extra bias to the estimated power spectrum. However, LSSA can restrict the foreground leakage for these RFI flagging scenarios and gives an unbiased estimate of the H I 21 cm power spectrum. We have also applied these algorithms to observations with the upgraded GMRT and found that both CLEAN and LSSA give consistent results in the presence of realistic random flagging scenarios for this observed data set. This comparative analysis demonstrates the effectiveness and robustness of these two algorithms in estimating the H I 21 cm power spectrum from data sets affected by different RFI scenarios.

Unified Astronomy Thesaurus concepts: Reionization (1383)

1. Introduction

The redshifted 21 cm “spin-flip” transition (Field 1958) of the neutral hydrogen (H I) is considered to be a promising tool to understand the global phase transition of the universe from neutral phase to ionized phase (the Epoch of Reionization, EoR) over the redshift range 6 < z < 15 (Field 1958; Madau et al. 1997; Furlanetto et al. 2006; Pritchard & Loeb 2012; Loeb & Furlanetto 2013; Dayal & Ferrara 2018). One promising way to study the EoR is by measuring the H I 21 cm power spectrum along with tomographic imaging of the intergalactic medium using low-frequency large interferometric arrays. (Bharadwaj & Sethi 2001; Morales & Hewitt 2004; Zaldarriaga et al. 2004; Bharadwaj & Ali 2005). Several upcoming and ongoing telescopes have been designed to study the 21 cm power spectrum, such as the Donald C. Backer Precision Array to Probe the Epoch of Reionization (PAPER, Parsons et al. 2010), the Low Frequency Array (LOFAR, van Haarlem et al. 2013), the Murchison Wide-field Array (MWA, Li et al. 2018), the Hydrogen Epoch of Reionization Array (HERA, DeBoer et al. 2017), and the Square Kilometre Array (SKA-1 Low, Koopmans et al. 2015). Also, a statistical detection of post-reionization (post-EoR) H I 21 cm signal holds the key to understanding the formation of large-scale structure and source clustering, and will constrain the cosmological parameters (Bharadwaj & Pandey 2003; Bharadwaj & Ali 2005; Wyithe & Loeb 2008; Bull et al. 2015). However, the redshifted H I signal from the EoR and post-EoR epoch is hidden behind radio emissions from different Galactic and extragalactic foregrounds that are several orders of magnitude brighter (Di Matteo et al. 2004; Zaldarriaga et al. 2004; Bharadwaj & Ali 2005; Jelić et al. 2008). Several methods are available in the literature to estimate the H I power spectrum amidst this bright sea of foregrounds. One can model the foregrounds accurately and then subtract them from the data set (Yatawatta et al. 2013; Procopio et al. 2017). Previous studies have shown that foreground is spectrally smooth, whereas the redshifted H I 21 cm signal has spectral structure. Due to this spectral smoothness and chromatic primary beam response, foregrounds are localized within a “wedge-shaped” region in a two-dimensional (2D) cylindrically averaged power spectrum. A traditional way is to avoid these foregrounds dominated by Fourier modes inside the wedge and search for cosmological H I signal in the so-called “EoR window.” This technique is known as foreground avoidance (Datta et al. 2010; Morales et al. 2012, 2019; Parsons et al. 2012; Vedantham et al. 2012). Along with that, there are techniques to separate spectrally smooth foregrounds from rapidly fluctuating H I signal without any prior modeling of foregrounds (Harker et al. 2009; Chapman et al. 2013; Mertens et al. 2018). However, all these methods rely on the fact that the spectral and spatial properties of foregrounds are very well known. Suppose there is some oscillating feature across the frequency band, introduced by instrumental effects or by any postprocessing methods. In that
case, it will pose a hindrance in isolating the redshifted $H\text{I}$ signal from the foregrounds.

One possible cause that can introduce rapidly fluctuating components is the excision of radio-frequency interference (RFI). In general, RFI is localized in the time or frequency domain. It is usual practice to measure visibilities with high time and frequency resolution, which helps to identify RFI across time and frequency. Then those RFI-contaminated samples get flagged in postprocessing of the data analysis. This results in a nonuniform sampling of the instrumental bandpass response, which otherwise is smooth in general. Now, each visibility is the Fourier transform of the product of the sky signal and the instrumental bandpass response. Hence, RFI flagging introduces irregularities in visibilities. To estimate the power spectrum of the cosmological signal, one needs to Fourier transform the measured visibilities from the frequency domain. This results in a nonuniform sampling of the instrumental time and frequency resolution, which helps to identify RFI in the delay domain. It is usual practice to measure visibilities with high time resolutions, and hence this RFI excision causes contamination by leakage of foregrounds.

Several ways exist in the literature to deal with this issue. PAPER analyses have used a wideband iterative deconvolution algorithm to remove the foreground components in delay space and then estimate the delay space power spectrum (Parsons et al. 2014; Jacobs et al. 2015; Kerrigan et al. 2018). Kolopanis et al. (2019) used a Blackman–Harris tapering window function during fast Fourier transformation to the delay domain to reduce foreground leakage. Trott et al. (2016) used least-squares spectral analysis (LSSA) to perform the line-of-sight transformation from frequency to delay space instead of the Fourier transform to mitigate the issues related to nonuniform bandpass shape for MWA observation. Patil et al. (2017), Gehlot et al. (2019), and Mertens et al. (2020) have also applied LSSA to the LOFAR data set to put an upper limit on the EoR $H\text{I}$ 21 cm power spectrum. Offringa et al. (2019) used Gaussian-weighted interpolation of missing visibility samples due to RFI flagging before any averaging of data. However, they used the same LSSA during the transformation from frequency to delay space.

Deshpande et al. (1996) have used the 1D CLEAN algorithm (Högå 1974) to estimate the power spectrum of pulsar timing residuals from nonuniformly sampled timing measurements. They showed that the CLEAN algorithm could help reconstruct the power spectra from nonuniformly sampled time sequences and it dramatically improved the dynamic range of the estimated power spectra. Deshpande et al. (2000) have used the same CLEAN algorithm to estimate the power spectrum of the optical depth of the cold $H\text{I}$ gas in the Galaxy toward Cass A and Cygnus-A from the incomplete sampling of the optical depth over the extent of the source. In this work, we implement the same 1D CLEAN algorithm in the delay space to deconvolve the effects of side lobes of the delay space PSF. Then we reconstruct the delay spectra of visibilities and estimate the 2D and 3D power spectra of the cosmological $H\text{I}$ signal.

LSSA has been used in digital signal processing as well as in astronomy to deal with nonuniformly sampled data sets (Barning 1963; Lomb 1976; Stoica et al. 2009). The spectrum of nonuniformly sampled data is difficult to interpret because the true peak in the spectrum produces several other peaks (aliases) of various heights distributed throughout the delay spectrum. In other words, the alias structure of the major peak creates confusion in estimating the true underlying spectrum (Lomb 1976). This is similar to foreground contamination of large $k_\parallel$ modes beyond the horizon limit as described before. LSSA is a commonly used method, where sinusoids of different frequencies are fitted to the nonuniform data via the traditional least-squares method, resulting in a reasonably good approximation to the spectrum of nonuniform data. The least-squares spectrum provides the best measure of the power contributed by different frequencies to the overall variance of the data and can be regarded as an extension of the Fourier methods to nonuniform data. In the limit of an equally spaced data set, this method reduces to the Fourier power spectrum.

In this work, we show the application of these widely used CLEAN and LSSA algorithms to simulated data (SKA-1 Low) for different flagging scenarios and compare our results with those from a uniformly sampled data set (i.e., no flagging). We also apply these to the actual observation of the upgraded Giant Metrewave Radio Telescope (uGMRT) data set in the presence of realistic flagging. This comparative analysis shows which algorithm gives a more robust and improved estimation of the cosmological $H\text{I}$ power spectrum in the presence of RFI flagging, without adding extra bias due to the issue of missing samples.

This paper is structured as follows. The simulation of data for the SKA-1 Low array configuration with observational setup and sky model is discussed in Section 2. In Section 3, we discuss the basic formalism to estimate the power spectrum from the simulated observation, including missing samples due to flagging. The application of the CLEAN and LSSA for different flagging scenarios is shown in Section 4. The application to the uGMRT observation is presented in Section 5. Finally, we conclude in Section 6.

2. Simulations

To investigate the effect of missing frequency channels on the estimation of the power spectrum and to demonstrate the performance of the CLEAN and LSSA, we use the data from simulated observations using the SKA-1 Low. We follow the procedure given in Li et al. (2019) for the simulation of the EoR and foreground signal and briefly describe it here. Li et al. (2019) extrapolate the low-frequency observation to make foreground sky maps that exhibit realistic structure of the sky. Foreground sources include the contribution from diffuse Galactic synchrotron emission, free–free emission, extragalactic point sources, and radio halos (Wang et al. 2010; Li et al. 2019). The Galactic synchrotron emission is simulated by extrapolating the reprocessed 408 MHz Haslam map (Remazeilles et al. 2015) using a power-law spectrum, the index of which is taken from Giardino et al. (2002). The Galactic free–free radiation map has been constructed from an Hα survey map (Finkbeiner 2003), where the close relation between Hα and free–free emission has been employed. Galactic diffuse
emission is simulated at a central position of \((R.A., \text{decl.}) = (0^\circ, -27^\circ)\), which is at high Galactic latitude and a suitable candidate field for EoR observation (Beardsley et al. 2016). The extragalactic point sources are simulated by using the SKADS catalog (Wilman et al. 2008). The radio halos are simulated by generating a sample of galaxy clusters with the Press–Schechter formalism (Press & Schechter 1974) and then using the scaling relation (cluster mass–X-ray temperature and X-ray temperature–radio power relations) to estimate their radio emission. For more details regarding foreground simulation, see Wang et al. (2010) and Li et al. (2019).

For the simulation of the EoR signal, we have used the Evolution of 21 cm Structure\(^5\) data set (Mesinger et al. 2016) and extracted the image slices at corresponding frequencies from the light-cone cube of the recommended “faint galaxy” models. These EoR sky maps are tiled and resampled to a field of view of \(5^\circ \times 5^\circ\) with a pixel resolution of 20\(''\). Note that this rescaling is done to match the same sky coverage and resolution of the foreground sky maps.

To incorporate the instrumental effects into these simulated sky maps, we use the SKA-1 Low\(^6\) array configuration and the OSKAR\(^6\) package (Mort et al. 2017) to simulate the instrumental observation. The array configuration consists of 512 stations, each of which consists of 256 antennas randomly distributed within a circle of 35 m in diameter. The stations are distributed in two parts: (1) 224 stations are distributed in a core region of diameter 1 km and the remainder distributed in three spiral arms extending up to 35 km. The array layout of the inner core region is shown in Figure 1. OSKAR uses the radio interferometer measurement equation of Hamaker et al. (1996) to simulate the full Stokes visibility data set; however, here we only simulate the RR component of visibilities. We provide the array configuration of SKA-1 Low to the telescope model and simulate the visibilities for each sky map at \((R.A., \text{decl.}) = (0^\circ, -27^\circ)\). We also include the uncorrelated Gaussian noise in our simulation corresponding to 1000 hr of observation with SKA-1 Low. The system temperature \((T_{\text{sys}})\) has contributions from both sky and the instrument receiver, i.e., \(T_{\text{sys}} = T_{\text{sky}} + T_{\text{rcvr}}\). The assumed sky temperature is \(T_{\text{sky}} = 60 \text{ K} \times (300 \text{ MHz}/\nu)^{2.25}\) (Thompson et al. 2017) and the receiver temperature is \(T_{\text{rcvr}} = 40 \text{ K}\) (Bacon et al. 2020). We choose an 8 MHz bandwidth around the central frequency of 142 MHz \((z \sim 9)\) with 64 kHz channel resolution (126 channels). The \(uv\)-plane corresponding to the simulated data is shown in Figure 1. We then use WSCLEAN\(^7\) (Offringa et al. 2014) to make a dirty image cube of size \(300 \times 300 \times 126\) (pixels \(\times\) pixels \(\times\) \(N_{\text{cham}}\)) for the EoR signal and foreground emission. We then convert these cubes from Jy beam to kelvin units (Patil et al. 2017) and then merge them to make a combined dirty image cube, \(I'(\theta_x, \theta_y, \nu)\), where \(\theta_x, \theta_y\) are the sky positions and \(\nu\) is the frequency. We use this image cube to estimate the HI 21 cm power spectrum and briefly discuss it below. The image slices corresponding to the EoR H I 21 cm signal (left panel), foreground (middle panel), and instrumental Gaussian noise (right panel) at \(z = 9\) are shown in Figure 2. The details of our simulation is mentioned in Table 1.

### 3. Estimation of Power Spectra in the Case of Missing Frequency Channels

In this section, we will briefly describe the formalism used in this work to estimate the power spectra from the dirty image cube, \(I'(\theta_x, \theta_y, \nu)\), which also suffers from missing frequency channels, due to RFI flagging. We follow the methodology as described in Morales & Hewitt (2004), Datta et al. (2010), and Patil et al. (2017).

We first perform the 2D spatial Fourier transform of the image cube to make the visibility cube,

\[
V(U, \nu) = \int I'(\theta_x, \theta_y, \nu) e^{-\imath 2 \pi U \theta} d\theta,
\]

where \(U = (u, \nu)\) and \(\theta = (\theta_x, \theta_y)\) are vectors.

We introduce the RFI flagging by multiplying the visibility cube by the frequency-dependent sample weights \((S(\nu))\), given

---

\(^5\) http://homepage.sns.it/mesinger/EOS.html

\(^6\) https://astronomers.skatelescope.org/DOCUMENTS

\(^7\) https://gitlab.com/arooffringa/wsclean
by

\[ S(\nu) = 0, \ \forall \text{flagged frequency channels} \]
\[ = 1, \ \forall \text{channels without flagging}. \]  \hspace{1cm} (2)

Here \( S(\nu) \) essentially corresponds to the flagging mask due to RFI, which has been applied to the visibility data cube. Then we transform the visibility cube to the delay domain (\( \eta \)-domain) by Fourier transforming the data along the frequency axis, as given by Morales & Hewitt (2004):

\[ V(U, \eta) = \int V(U, \nu) B(\nu) S(\nu) e^{-ij2\pi\eta\nu} d\nu, \]  \hspace{1cm} (3)

where \( B(\nu) \) is the instrumental bandpass response, taken to be flat in this case.

According to the Fourier convolution theorem, visibility \( V(U, \eta) \) can be expressed as the convolution of \( \mathcal{F}[V(U, \nu)] \) with the convolving kernel \( \mathcal{F}[B(\nu)S(\nu)] \) as

\[ V(U, \eta) = \mathcal{F}[V(U, \nu)] * \mathcal{F}[B(\nu)S(\nu)], \]  \hspace{1cm} (4)

where \( \mathcal{F} \) denotes the Fourier transform operator and * denotes the convolution.

We refer to the Fourier transform of \( B(\nu)S(\nu) \) to the delay domain as the delay space PSF and denote it as \( D_{\text{psf}} \). Since the excision of RFI-contaminated channels results in a nonuniform sampling of the instrumental bandpass, which is encoded in \( S(\nu) \), the simple Fourier transform of visibilities (Equation (3)) along the frequency axis will result in spectral leakage. This is the leakage of smooth foregrounds into the EoR window. In other words, the side lobes of the delay space PSF (\( D_{\text{psf}} \)) due to RFI flagging will be large. Convolution of \( D_{\text{psf}} \) with \( \mathcal{F}[V(U, \nu)] \) will result in leakage of foregrounds beyond the horizon limit into the EoR window and will show an oscillating pattern above the horizon delay.

To understand the problem, we show the delay spectrum of a single baseline of length 162\( \lambda \) (arbitrarily chosen) from our simulated data cube (see Figure 3). We first Fourier transform the data cube to the delay domain without using any flagging, i.e., \( S(\nu) = 1, \ \forall \text{channels} \). The resulting delay spectrum of the baseline is shown in Figure 3 in gray. We refer to this as the no-RFI case in the plot. The vertical dashed lines are the delay horizon limits.

Figure 3. Delay spectrum of an arbitrarily chosen baseline of length 162\( \lambda \). The top blue curve is the delay spectrum with 10% of channels randomly flagged. The green dashed curve is the delay spectrum after using the 1D CLEAN algorithm across the delay axis. The red dashed curve is the delay spectrum of the baseline after applying the LSSA algorithm to this flagged data set. The gray curve is the delay spectrum of the baseline when there is no flagging (no RFI). The vertical dashed black lines are the horizon limit, which is 1.14 \( \mu s \), for this particular baseline.

Table 1
Details of Simulations

| Simulation Parameters | SKA-1 Low |
|-----------------------|-----------|
| Telescope             | SKA-1 Low |
| Min. baseline         | 42 m      |
| Max. baseline         | 2.8 km    |
| Frequency             | 138–146 MHz |
| Number of channels    | 126       |
| Central redshift (z)  | 9.0       |
| Cosmological signal:  | simulated HI signal cube |
| Sky model             | +         |
| Foreground:           | +         |
| extragalactic point   | +         |
| sources and diffuse   | +         |
| foregrounds           | +         |
| Gaussian noise        | +         |
horizon limits for this particular baseline. Then we randomly flagged 10% of channels and performed the delay transformation.

One can use a window function \( W(\nu) \) during the Fourier transformation to the delay domain to reduce the spectral leakage (Thyagarajan et al. 2013; Kolopanis et al. 2019). Then Equations (3) and (4) will become

\[
V(\mathbf{U}, \eta) = \int V(\mathbf{U}, \nu) B(\nu) S(\nu) W(\nu) e^{-i2\pi\nu\eta} d\nu,
\]

and

\[
V(\mathbf{U}, \eta) = \mathcal{F}[V(\mathbf{U}, \nu)] \ast \mathcal{F}[B(\nu)S(\nu)W(\nu)].
\]

Here, \( W(\nu) \) will act as a spectral taper during delay transformation and will reduce the spectral leakage to some extent. We show in Figure 3 the resulting delay spectrum of the baseline with a Blackman–Harris window as a spectral weighting function (in blue). As seen from the figure, there is still significant spectral leakage beyond the horizon limit even after using the window function. This is mainly due to the convolution of visibility with \( \mathcal{F}[B(\nu)S(\nu)W(\nu)] \), where the missing channels due to flagging are encoded inside \( S(\nu) \).

We employ two methods that have previously been widely used on data, CLEAN (Parsons et al. 2014) and LSSA (Trott et al. 2016; Patil et al. 2017), to mitigate this spectral leakage issue due to missing samples (RFI flagging). Kolopanis et al. (2019) used a Blackman–Harris tapering function to reduce the spectral leakage in their reanalysis of PAPER-64 data. However, as we have seen above, a simple spectral tapering is unable to mitigate foreground spillover due to missing samples in the presence of only 10% RFI flagging. We briefly discuss the other two methods, CLEAN and LSSA, below and compare our results here.

CLEAN. To restrict the spectral leakage beyond the horizon limit, we deconvolve \( D_{\text{psf}} \) using a one-dimensional Hogbom CLEAN algorithm across the delay axis (Högborn 1974; Parsons & Backer 2009; Parsons et al. 2014; Kerrigan et al. 2018). The CLEAN algorithm iteratively searches for the maximum in the delay spectrum of visibility and then subtracts a fraction (0.1; gain term) of the delay space PSF from that until a relative threshold in the residual is reached. This threshold is determined by the ratio of the initial and final amplitudes in the spectrum and is provided as the tolerance parameter in the CLEAN algorithm (Kerrigan et al. 2018; Ewall-Wice et al. 2021). In other words, the tolerance parameter sets the degree of foreground subtraction during CLEANing and can be set as low as we want in order to CLEAN deeper (Ewall-Wice et al. 2021). The choice of this tolerance parameter for various flagging scenarios is discussed in Section 4. We first zero-pad the visibilities by the same number of channels as that of the simulation and apply a Tukey window (with \( \alpha = 0.2 \)) as a tapering function before performing the CLEAN (Kern et al. 2019; Ewall-Wice et al. 2021). After reaching the threshold in the residual, we add the residual to the CLEAN model components. This delay space CLEANing effectively deconvolves the delay space PSF and reduces the side lobes of the delay spectrum due to nonuniform sampling of the bandpass and gives the final delay spectrum. In Figure 3, we show the delay spectrum after using 1D CLEAN across the delay axis as a green dashed line. It is clear from the figure that, with this simple deconvolution technique, the leakage of foreground beyond the horizon limit due to missing channels can be kept to a minimum.

Least-squares spectral analysis (LSSA). Another widely used approach to perform the line-of-sight Fourier transform of the data from frequency space to delay space in the case of missing samples is LSSA (Trott et al. 2016; Patil et al. 2017). LSSA estimates the Fourier spectrum by fitting sinusoids to the data samples using the traditional least-squares method. A discrete data set is approximated by a weighted sum of sinusoids of progressively determined frequencies using a standard linear regression or least-squares fit. If one knows the covariance of the data, then that can be used to weight the data vector to get the approximated functional form of the discrete data samples (Lomb 1976; Stoica et al. 2009). Trott et al. (2016) first used this method in their pipeline for H I 21 cm power spectrum estimation and applied it to the MWA data set. Later, Patil et al. (2017), Gehlot et al. (2019), and Mertens et al. (2020) also used this methodology in their analysis of the LOFAR data set to put an upper limit on the H I 21 cm power spectrum. We have implemented this method here as well to mitigate the spectral leakage in the delay spectrum of visibilities due to missing frequency samples. A smooth model, which is a Fourier series up to a specified order, is fitted to the complex visibility data cube with flags, using a linear least-squares solver (scipy.optimize.least_squares). We use 300 Fourier modes to fit the visibility spectrum in this analysis. The delay spectrum after application of LSSA for a single baseline is shown by the red dashed line in Figure 3. We find that, similar to CLEAN, the LSSA approach is also able to reduce the spectral leakage in the presence of random RFI, and we can go down to the signal level (without RFI) in the delay spectrum beyond the horizon limit.

Power spectrum estimation. After getting the delay visibility cube \( (V(\mathbf{U}, \eta)) \), we proceed to estimate the cylindrically averaged 2D power spectrum using the formula (Morales & Hewitt 2004)

\[
P(k_\perp, k_\parallel) = \frac{\lambda^2}{2k_B} \left( \chi^2 Y \Omega B \right) |V(\mathbf{U}, \eta)|^2,
\]

where \( \lambda \) is the wavelength corresponding to the band center, \( k_B \) is the Boltzmann constant, \( \Omega \) is the sky-integral of the squared antenna primary beam response, \( B \) is the bandwidth, and \( X \) and \( Y \) are the conversion factors from angle and frequency to transverse comoving distance \( (D(z)) \) and the comoving depth along the line of sight, respectively (Morales & Hewitt 2004). The power spectrum \( P(k_\perp, k_\parallel) \) is in units of \( K^2 \text{Mpc}^3 \). The wavenumbers \( k_\perp \) and \( k_\parallel \) are related to the baseline vector \( \mathbf{U} \) and delay \( \eta \) by

\[
k_\perp = \frac{2\pi|\mathbf{U}|}{D(z)},
\]

\[
k_\parallel = \frac{2\pi\nu_{21}H_0 E(z)}{c(1+z)^2},
\]

where \( \nu_{21} \) is the rest-frame frequency of the 21 cm spin-flip transition of H I, \( z \) is the redshift to the observed frequency, \( H_0 \) is the Hubble parameter, and \( E(z) \equiv [\Omega_M(1+z)^3 + \Omega_\Lambda]^{1/2} \). \( \Omega_M \) and \( \Omega_\Lambda \) are matter and dark energy densities, respectively (Hogg 1999). In this work, we use the best-fitted cosmological parameters of the Planck 2018 analysis (Planck Collaboration et al. 2020).
The 3D power spectrum can be calculated by spherically averaging \( P(k_{1}, k_{2}) \) in independent \( k \)-bins and the dimensionless 3D power spectrum can be written as (Datta et al. 2010)

\[
\Delta^{2}(k) = \frac{K^{3}}{2\pi^{2}}\langle P(k) \rangle_{k},
\]

where \( k = \sqrt{k_{1}^{2} + k_{2}^{2}} \). Throughout this work we will follow this formalism to estimate the 2D and 3D power spectra. We estimate the uncertainty in the estimated power spectrum by calculating the variance of the measured power spectrum and dividing by the number of independent modes averaged together in each \( k \)-bin (Tegmark 1997). We quote this 1\( \sigma \) uncertainty as the error in the 3D power spectrum at each \( k \)-mode.

4. Application to Simulated Data

To understand the effect of RFI flagging in estimating the cosmological HI power spectrum, we first flagged frequency channels before performing the FT along the frequency axis using Equation (5) and then estimated the cylindrically and spherically averaged power spectrum. Below we discuss the different cases we consider in our analysis and the corresponding results.

4.1. Case Studies

To understand the effectiveness of the CLEAN and LSSA algorithms in reducing the spectral leakage for realistic flagging scenarios, we study the following cases.

Case I. There is no flagging, i.e., the no-RFI scenario. The power spectrum is estimated from this unflagged sample.

Case II. We randomly selected 20\% of channels and flagged them. This is the most simplistic scenario and can be observed in data taken with any radio telescope. Here, we conservatively choose 20\% of random channels to flag. However, note that some low-frequency radio telescopes were built in radio-quiet zones, where the percentage of random flagging may be less than this. After applying the random flags to the visibility data cube, we use CLEAN and LSSA algorithms to go to the delay domain and estimate the power spectrum.

Case III-A. The FPGA-based digital receiver employed at MWA digitizes the beam-formed signal and performs a first-stage coarse frequency channelization of the data, i.e., 1.28 MHz coarse bands across 32 MHz bandwidth of MWA data (Prabu et al. 2015). This step applies a filter shape and aliases channels on the edge of each coarse band (25 coarse bands across 32 MHz bandwidth) (Prabu et al. 2015; Beardsley et al. 2016). To avoid this known aliasing effect, Dillon et al. (2015) flagged the 160 kHz channel (after smoothing the data by two channels) in every 1.28 MHz coarse band in their analysis of the MWA data. Beardsley et al. (2016) also flagged two 40 kHz channels on either side of the coarse-band edges (a total of four channels per coarse-band edge) to avoid the aliasing effect. In addition, they also flagged the central 40 kHz channel of each coarse band of MWA data, which corresponds to coarse-band DC mode. These channels have been observed to contain anomalous power, likely due to very low-level rounding errors in the polyphase filter bank of the digital receivers. Barry et al. (2019) also used the same flagging strategy in their updated analysis of MWA data to get a more robust upper limit on the H1 power spectrum. The flagging of channels at either side of each coarse band and at its center creates a periodic flagging scenario in addition to the random flagging due to persistent RFI. To incorporate this in our analysis, we flag 20\% of randomly selected channels, and also 200 kHz flags are introduced every 1.28 MHz across the 8 MHz bandwidth of our simulated data. We call this flagging scenario random plus periodic flagging.

Case III-B. This is similar to the previous Case III-A, with the change in periodicity and increase in flagging channel width. In this case, in addition to random flagging, we introduced 300 kHz flags every 0.64 MHz across the 8 MHz bandwidth.

Case IV. In this case, we introduce the low-level (faint) broadband RFI flagging corresponding to a digital television (DTV) signal as reported in MWA observation (Barry et al. 2019; Wilensky et al. 2019). Wilensky et al. (2019) developed a pipeline, called SSINS, which can identify such ultraintense RFI and remove the contaminated data. Barry et al. (2019) found a significant improvement in the estimated power spectrum after removing the low-level broadband RFI-contaminated data using SSINS in their reanalysis of the MWA data set. To incorporate this broadband RFI in our analysis, we flag a 2 MHz wide frequency chunk at the center of the band, in addition to random and periodic flagging.

The different cases are also summarized in Table 2. We compare the results from these different cases in the following section.

4.2. Results

Case I. We show in Figure 4 (left panel), the cylindrically averaged 2D power spectrum in \( k_{1} \) and \( k_{2} \) space (as given by Equation (7)) when there is no RFI, i.e., no flagging involved. We find that the spectrally smooth bright foregrounds coupled with instrumental response are confined within a “wedge-shaped” region in this 2D space. The black dashed line is the horizon limit with an additional buffer due to the Blackman–Harris (BH) window used during FT in the frequency direction. The region above this line contains less power than the
foreground wedge and is popularly known as the “EoR window.” One can avoid the foreground-dominated modes inside the “wedge” and use only the modes inside the “EoR window” to search for HI 21 cm signal. This technique is called “foreground avoidance.” The spherically averaged 3D power spectrum using the modes outside the foreground wedge is shown in the right panel of Figure 4. The black dashed line shows the injected HI 21 cm power spectrum from 21cmFAST \(^8\) (Mesinger et al. 2016) modulated by the instrument beam. Here, we assume a simple Gaussian beam of the instrument. We find that the 3D power spectrum estimated by avoiding the foreground for the no-RFI scenario is consistent with the input HI 21 cm signal. This case study also provides the null test, where we can recover the injected HI 21 cm power spectrum using our power spectrum estimation methodology as described earlier.

*Case II.* In Figure 5, we show our findings for the random flagging scenario. We plot the ratio of the 2D cylindrically averaged power spectrum for different algorithms applied to the randomly flagged data. At the top left of Figure 5, we show the ratio of the 2D power spectrum of flagged to no-RFI scenario on a log scale. We find a significant spillover of foregrounds into the “EoR window” when there are samples missing due to RFI flagging. Note that, although we use the BH window during FT along the frequency axis in this case, it completely fails to restrict the bleed-in of foregrounds into the “EoR window.” The use of a tapering window function during the FT helps to suppress the effects of the sharp edge of the finite bandpass (Thyagarajan et al. 2013), but it cannot deal with the samples missing due to RFI flagging. In the top right and bottom left panels of Figure 5, we show the ratio after applying the CLEAN and LSSA algorithms to the RFI-flagged data set with respect to the no-RFI scenario, respectively. Here, we take the tolerance parameter in CLEAN to be \(10^{-10}\). We find that there are only a few modes outside the foreground wedge, which have a little foreground contamination in comparison with the no-RFI case when we use the CLEAN algorithm. The contamination is even less when applying LSSA to the flagged data set during FT. Nevertheless, these two plots show that one can restrict the foreground spillover to a minimum after using CLEAN and LSSA algorithms to the flagged data set in estimating the power spectrum. After applying these algorithms, one can choose a foreground-free “window” from this 2D cylindrical power spectrum and spherically average the modes to estimate the 3D power spectrum. At the bottom right of Figure 5, we show the 3D power spectrum after spherically averaging all the modes above the horizon line, i.e., the modes inside the “EoR window.” We compare the result of the 3D power spectrum for CLEAN and LSSA algorithms to the no-RFI scenario (Case I) and the input 21cmFAST HI power spectrum. We find that we can still recover the \(k\)-modes of the HI 21 cm power spectrum in the presence of random flagging after using these two algorithms.

*Case III-A.* The results for random and periodic flagging are shown in Figure 6. Similar to the findings of Case II, here we also find that the presence of a periodic flag along with the random flagging severely contaminates the “EoR window” when only the BH tapering function is being applied (top left panel of Figure 6). In the top right panel, we show the ratio of the 2D power spectrum after the application of CLEAN to the periodic plus randomly flagged data set. We find that the Fourier modes inside the EoR window are still contaminated by foreground spillover after applying the CLEAN algorithm. The extra bias in CLEAN arises due to the unsubtracted foregrounds in the residual, which still contains side lobes from RFI flagging (Ewall-Wice et al. 2021). To CLEAN the spectrum much deeper than Case II, we choose the tolerance parameter to be \(10^{-11}\). But, even after lowering the tolerance parameter, we find significant bias due to foreground leakage into the EoR window. However, when we apply the LSSA to the flagged data set (bottom left panel), we can restrict the foreground contamination of Fourier modes beyond the “wedge” in the presence of periodic flagging. The 3D power spectrum in the bottom right panel also shows that there is an extra bias to the power spectrum for the CLEAN algorithm with respect to the no-RFI scenario. But LSSA can mitigate this foreground leakage in the presence of periodic flagging and it gives an unbiased estimate of the 3D power spectrum.

---

\(^8\) https://homepage.sns.it/mesinger/EOS.html
**Case III-B.** The results for the increased flagging channel width (300 kHz) and a decreased periodicity (every 0.64 MHz) compared to Case III-A are shown in Figure 7. The results are fairly similar to Case III-A when we apply BH and LSSA algorithms to the flagged data set. The application of the BH tapering function shows a significant foreground leakage above the horizon. However, applying the LSSA algorithm shows that this leakage is negligible. We also find that, unlike Case III-A, the application of CLEAN gives higher foreground leakage into the EoR window. Note that, here also, we take the tolerance parameter to be $10^{-11}$ in the CLEAN algorithm to reduce the foreground residual. The spherically averaged 3D power spectrum also shows the presence of a substantial bias in the estimated power spectrum in the case of CLEAN. In contrast, we can still recover the Fourier modes inside the EoR window in the case of LSSA. The LSSA can give an unbiased and robust estimate of the underlying power spectrum in the presence of broadband RFI flagging in addition to random and periodic flags.

The different flagging scenarios used in this comparative analysis prove that LSSA gives a more sensitive and robust way to estimate the cosmological H I power spectrum from the observed data. However, a sufficiently deep CLEANing in delay space can mitigate the foreground leakage issue in the presence of random RFI flagging only.
5. Application to the uGMRT Data

Here, we extend our comparative analyses to a real observation of the ELAIS N1 field observed with the uGMRT (in GTAC cycle 32) during 2017 May. The details of the observation, a brief discussion on data analysis, and the estimation of the power spectrum from the observed visibility data set are discussed in the following subsections.

5.1. Observations and Data Analysis

The data were taken over four nights, where the total integration time, including calibrators, is 25 hr. The total bandwidth (200 MHz) between 300 and 500 MHz is divided into 8192 channels, which results in a 24 kHz frequency resolution. The raw data have a time resolution of 2 s. The observational details are tabulated in Table 3. The details of data analysis are presented in Chakraborty et al. (2019). The raw data were flagged for RFI with AOFLAGGER (Offringa et al. 2012) and then averaged to 16 s time resolution to reduce the data volume. About 40% of the data were flagged after RFI flagging, calibration, and self-calibration steps. We have not averaged the data across the frequency coordinate to get the maximum \( k_\parallel \) modes and work with 24 kHz frequency resolution. This choice will allow us to access a larger region of the EoR window above the foreground wedge.

5.2. Power Spectrum Estimation from the Data

In this work, we show the result of a small 8 MHz chunk around 304 MHz (\( z \approx 3.6 \)) of one night’s data (~6 hr including calibrators). This demonstrates the effect of CLEAN and LSSA algorithms in estimating the power spectrum with realistic flagging applied to the observed visibilities. In our previous work (Chakraborty et al. 2021), we have already analyzed the whole data set and put an upper limit on the post-EoR H I 21 cm power spectrum at redshifts \( z = 1.96, 2.19, 2.62, \) and 3.58. In Chakraborty et al. (2021), we have used the CLEAN algorithm to mitigate the issue of missing samples due to RFI flagging. However, here we also apply the LSSA along with CLEAN on a single night’s data set to compare the effectiveness of these two algorithms in a realistic scenario.

The details of the estimation of the power spectrum from the observed uGMRT data set are mentioned in Chakraborty et al. (2021) and we briefly mention the methodology here. The simple squaring of visibilities as described in Equation (7) to estimate the power spectrum will result in positive noise bias since the observed data also contain noise (Bharadwaj & Pandey 2003; Ali et al. 2008). To avoid this, we grouped the visibilities within a uv-cell whose dimension is governed by the inverse of the half-power beamwidth of the primary beam (\( \theta_{HPBW} \)). We then cross-correlate the visibilities associated with a particular uv-cell. This results in a correlation matrix for each uv-cell. The diagonal terms of that matrix are the
self-correlation of visibilities and contain the positive noise bias. The off-diagonal terms are the cross-correlation of visibilities that lie inside the \(uv\)-cell. The noise is expected to be uncorrelated between two visibilities, and so the off-diagonal terms do not contain the positive noise bias (Ali et al. 2008). We take the average of the off-diagonal terms of a correlation matrix of a particular \(uv\)-cell and quote it as an estimated power for that baseline. The underlying assumption is that the cosmological H I signal is correlated among visibilities that fall within a \(uv\)-cell whose dimension is the inverse of the primary beam. This formalism was proposed by Bharadwaj & Pandey (2003) and investigated in detail using observations by Ali et al. (2008), Ghosh et al. (2012), Choudhuri et al. (2016), Bharadwaj et al. (2019), Chakraborty et al. (2021), and Pal et al. (2021).

We have applied both CLEAN and LSSA to transform the observed visibility data from frequency to delay domain in the presence of realistic flagging and then to estimate the power spectrum following the method described above. In Figure 9, we show the cylindrically averaged 2D power spectrum after applying CLEAN (left panel) and LSSA (right panel) algorithms to the observed visibilities. We find that both CLEAN and LSSA can restrict the foreground leakage above the foreground wedge and give nearly identical results. Above the horizon line, there is a region where the value of the power spectrum is much less than the power inside the wedge. This region is relatively free from foreground contamination, and one can spherically average the \(k_\perp\) and \(k_\parallel\) modes of that region to estimate the 3D power spectrum. In Figure 10, we show the ratio of the 2D power spectra estimated using CLEAN and LSSA algorithms. The ratio is close to one for most of the \(k\)-modes above the horizon limit. This shows that, in the presence of realistic random flagging of the uGMRT data set, CLEAN and LSSA give nearly identical results and are consistent with each other. Note that, here, we do not estimate the spherically averaged power spectrum of cosmological H I signal with this small data set. We only demonstrate the effectiveness of CLEAN and LSSA algorithms in the case of realistic flagging due to RFI. The 3D spherically averaged power spectra using the modes free from foreground contamination above the horizon line for data sets from different nights as well as for the combined data set are already presented in Chakraborty et al. (2021).

6. Summary and Conclusion

The flagging of RFI-contaminated channels can cause excess power leaking into the EoR window (above the horizon limit) in the 2D cylindrical power spectrum. The nonuniform sampling of the bandpass due to RFI excision causes large side lobes in the delay space PSF. The convolution of this nonideal delay space PSF with the delay spectrum of the data causes foreground leakage into the EoR window. The one-dimensional CLEAN algorithm in delay space and LSSA are two widely used algorithms, that have been applied to the
observed data set to mitigate this foreground leakage issue. In this work, we have done a comparative analysis of these two algorithms using simulations in the presence of different realistic flagging scenarios.

We have simulated the data for the SKA-1 Low array configuration with cosmological H I 21 cm signal and foreground signals as a sky model at redshift $z \approx 9$. We have compared three different RFI flagging scenarios. At first, we randomly flagged 20% of channels of the simulated data and then applied CLEAN and LSSA to restrict the foreground leakage into the EoR window (Case II). We estimate the 2D cylindrically averaged power spectrum and compare the result with the no-RFI scenario, i.e., no flagging. We have found that both of these algorithms can restrict the foreground leakage into the EoR window for random flagging of channels. We spherically averaged the modes inside the “EoR window” to estimate the H I 3D power spectrum. We have compared the 3D power spectrum with the no-RFI scenario as well as with the injected H I 21 cm signal in our simulation, as taken from 21cmFAST. The results are consistent with each other. We conclude that in the presence of random flagging, the application of CLEAN and LSSA can recover the $k$-modes to extract the cosmological H I 21 cm signal power spectrum.

Then we study the case where we have flagged 200 kHz channels every 1.28 MHz across the whole 8 MHz band of our simulated data set in addition to 20% random flagging (Case III-A). This is identical to data observed with the MWA telescope. We have also changed the periodicity and flagging channel width (Case III-B), i.e., we flag 300 kHz channels every 0.64 MHz across 8 MHz bandwidth. We applied CLEAN and LSSA and compared the 2D and 3D power spectra with the no-RFI scenario and with the input H I 21 cm power spectrum. We have found that the CLEAN algorithm fails to restrict the foreground leakage for both cases, and there is an extra bias in the 3D power spectrum. However, the LSSA method can limit the foreground leakage even in the presence of periodic flagging. We can average the modes above the “wedge” and reconstruct the H I 21 cm power spectrum using the LSSA method.

Table 3
Observational Details of the Target Field ELAIS N1 for Four Observing Sessions

| Project code | 32_120 |
|--------------|--------|
| Observation dates | 2017 May 5, 6, 7, 2017 June 27 |
| Bandwidth | 200 MHz |
| Frequency range | 300–500 MHz |
| Channels | 8192 |
| Integration time | 2 s |
| Correlations | RR RL LR LL |
| Total on-source time | ~13 hr (ELAIS N1) |
| Working antennas | 26 |
| Pointing center | $16^h 10^m 01^s + 54^d 30^m 36^s$ (ELAIS N1) |

The Astrophysical Journal, 929:104 (13pp), 2022 April 10 Chakraborty, Datta, & Mazunader
algorithm. We conclude from this analysis that LSSA works better than CLEAN and gives an unbiased estimate of the power spectrum in the presence of periodic flagging.

Lastly, we study the scenario where a 2 MHz wide frequency chunk has been flagged from the center of the band in addition to the random and periodic flagging. This scenario corresponds to the ultrafaint broadband RFI due to the dish TV signal. After applying different algorithms for this case, we found that CLEAN gives a significant bias in the power spectrum and fails to restrict the foreground leakage. But LSSA can constrain the leakage and provide an unbiased estimate of the HI power spectrum.

We have also applied these algorithms to the observed data set from the ELAIS N1 field using uGMRT to verify these findings with realistic flagging. However, we only choose a small 8 MHz chunk around 310 MHz of one night’s data (6 hr) and estimate the 2D power spectrum using CLEAN and LSSA algorithms. Here also, we have found that these two algorithms give identical results in restricting the foreground contamination above the “wedge,” since the uGMRT data contain the random flagging only. We can limit the foreground spillover with these algorithms and extract the cosmological HI signal from the uGMRT observation. The first upper limit on post-EoR HI signal using this data set has already been presented in Chakraborty et al. (2021), where we have used the CLEAN algorithm to mitigate the issue of missing samples. However, this analysis with the data shows that the power spectrum estimated using the LSSA algorithm is consistent with that using the CLEAN algorithm.

In summary, this paper explores the effectiveness of 1D CLEAN and LSSA algorithms to deal with issues of missing frequency channels in Cosmic Dawn (CD)/EoR observations. We show that both of these algorithms are effective in the presence of random flagging. However, CLEAN fails in the presence of periodic and broadband RFI flagging. But LSSA can still restrict the foreground spillover beyond the horizon limit and give a robust estimate of the HI 21 cm power spectrum. This methodology gives a sensitive estimation of the HI power spectrum even in the case where 20% of random channels are flagged, as well as when a periodic flagging of channels has been applied across the band. The comparative analysis presented here establishes the advantages and necessity of using one algorithm over the other to estimate the HI power spectrum from observed data contaminated by RFI. This work is relevant for upcoming telescopes with CD/EoR science goals like the SKA, HERA, and also for ongoing projects like the LOFAR, MWA, and PAPER, etc. However, it should be noted that the issue of missing channels is not limited to CD/EoR frequencies but also affects post-EoR redshifted frequencies, which are important for HI

\[ \text{Figure 9. The cylindrically averaged 2D power spectrum after applying CLEAN (left panel) and LSSA (right panel) to one night’s data set of the ELAIS N1 field observed with the uGMRT. The data span an 8 MHz bandwidth around 310 MHz (} z = 3.58\text{). The black dashed line in both panels is the horizon limit.} \]

\[ \text{Figure 10. The ratio of the 2D power spectra between LSSA and CLEAN algorithms after applying them to the data set of uGMRT. This plot is basically the ratio of the 2D power spectra shown in Figure 9.} \]
21 cm intensity mapping experiments like CHIME,\textsuperscript{14} TIAN-LAI,\textsuperscript{15} HIReX,\textsuperscript{16} SKA-1 Mid\textsuperscript{17} and OWFA.\textsuperscript{18}

We thank the anonymous referee and the scientific editor for helpful comments and suggestions that have helped to improve this work. A.D. would like to acknowledge the support of EMR-II under CSR No. 03(1461)/19. A.M. would like to thank Indian Institute of Technology Indore for supporting this research with Teaching Assistantship.

Software: This work heavily relies on the Python programming language (https://www.python.org/). The packages used here are pyuvdata (https://github.com/RadioAstroSoftwareGroup/pyuvdata), apiy (https://github.com/HERA-Team/apiy/tree/main/apiy; Hazelton et al. 2017), astropy (https://www.astropy.org/; Astropy Collaboration et al. 2013, 2018), numpy (https://numpy.org/; van der Walt et al. 2011; Harris et al. 2020), scipy (https://www.scipy.org/; Virtanen et al. 2020), matplotlib (https://matplotlib.org/; Hunter 2007), OSKAR (https://github.com/OxfordSKASOKAR) and WSCLEAN (https://gitlab.com/aroofringa/wsclean).

Data Availability

The simulated data cube used in this analysis will be shared on reasonable request to the corresponding author. The point source catalog of the ELAIS N1 field at 400 MHz compiled from uGMR observation can be found here: https://github.com/Arnabitiindore/Catalog-of-ELAIS-N1-field and in the link https://doi.org/10.5281/zenodo.5822497.

ORCID iDs

Arnab Chakraborty https://orcid.org/0000-0002-7758-9859
Abhirup Datta https://orcid.org/0000-0002-5333-1095
Aishhila Mazumder https://orcid.org/0000-0003-4641-946X

References

Ali, S. S., Bharadwaj, S., & Chengalur, J. N. 2008, MNRAS, 385, 2166
Astropy Collaboration, Robitaille, T., et al. 2013, A&A, 558, A33
Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, AJ, 156, 123
Bacon, D. J., Batty, R. A., Bull, P., et al. 2020, PASA, 37, e007
Barning, F. J. M. 1963, Bull. Astron. Inst. Netherlands, 17, 22
Barry, N., Wilensky, M., Trott, C. M., et al. 2019, ApJ, 884, 105
Bharadwaj, S., & Sethi, S. K. 2001, JApA, 22, 293
Bharadwaj, S., & Pandey, S. K. 2003, JApA, 24, 23
Bharadwaj, S., & Sethi, S. K. 2001, JApA, 22, 293
Bull, P., Ferreira, P. G., Patel, P., & Santos, M. G. 2015, ApJ, 803, 21
Chakraborty, A., Roy, N., Datta, A., et al. 2019, MNRAS, 490, 243
Chakraborty, A., Datta, A., Roy, N., et al. 2021, ApJL, 907, L7
Chapman, E., Abdalla, F. B., Bobin, J., et al. 2013, MNRAS, 429, 165
Choudhuri, S., Bharadwaj, S., Chatterjee, S., et al. 2016, MNRAS, 463, 4093
Datta, A., Bowman, J. D., & Carilli, C. L. 2010, ApJ, 724, 526
DeBoer, D. R., Parsons, A. R., Aguirre, J. E., et al. 2017, PASP, 129, 045001
Deshpande, A. A., Dwarakanath, K. S., & Goss, W. M. 2000, ApJ, 543, 227
Di Matteo, T., Ciardi, B., & Miniati, F. 2004, MNRAS, 355, 1053
Dillon, J. S., Neben, A. R., Hewitt, J. N., et al. 2015, PRRd, 91, 123011