Hunting Elusive Excess Variance in Big LOFAR Data

Gan, Hyoyin

DOI:
10.33612/diss.240437107

Publication date:
2022

1

INTRODUCTION

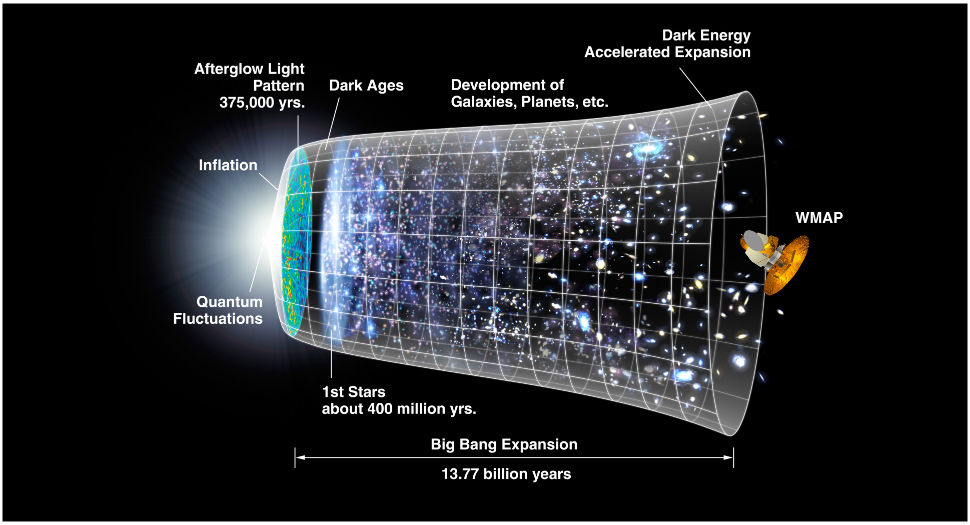

Figure 1.1 | An illustration of the timeline of the universe. Credit: NASA/WMAP Science Team.

1.1. Beginning of Time

About 13.8 billion years ago, the entire universe was squeezed into a singularity, a point of infinite density and temperature (Starobinsky, 1980; Guth, 1981; Sato, 1981; Planck Collaboration et al., 2016). Suddenly, an explosive expansion began, expanding the universe at a speed faster than the speed of light. This event is known as the Big Bang (see Kolb & Turner, 1990; Kane et al., 2015; Allahverdi et al., 2021, for a review). The universe went through a number of crucial phases after the Big Bang:

Up to 10^{-6} **seconds:** the universe was an extremely hot and dense soup of light relativistic sub-atomic particles such as quarks, electrons, photons and neutrinos. The four fundamental forces (i.e., electromagnetic, strong, weak and gravitational forces) were even partially entangled at this time.

After 1 second: as the universe expanded and cooled, the four fundamental forces emerged. Protons and neutrons formed.

After 3 minutes: protons and neutrons combined to form light elements such as hydrogen, helium and lithium. The universe was still too hot and dense for photons to travel. Atoms constantly collided into a hot and dense plasma, scattering photons. The universe remained opaque.

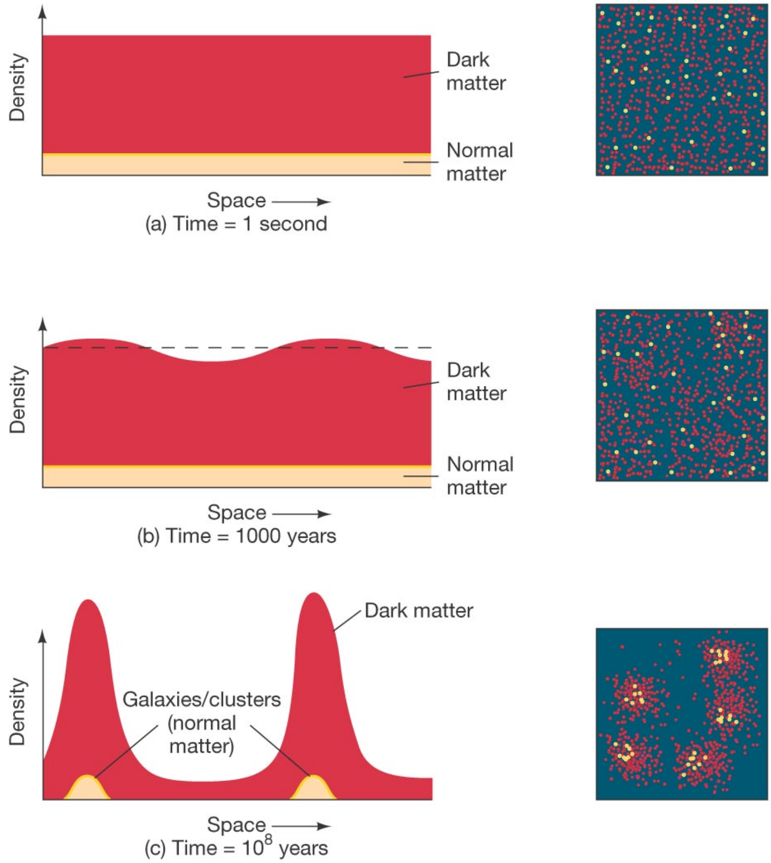

Figure 1.2 | Illustration structure formation. It depicts how tiny irregularities in dark matter and hydrogen gas in the CMB grow over time and collapse to more complex structures. Credit: Pearson Education.

After 380,000 years: matter cooled sufficiently to form neutral atoms, this period is known as ‘recombination’. Photons could freely travel in space without colliding into nuclei, and the universe became transparent. This ‘afterglow’ is the Cosmic Microwave Background (CMB; Komatsu et al., 2009; Hinshaw et al., 2013; Bennett et al., 2013). The CMB was nearly uniform with fluctuations of fractions of Kelvins (i.e., 1 part in $10^4 - 10^5$). There were no stars or galaxies. The universe was still dark apart from the CMB radiation itself.

After 400 million years: under the interaction of gravity, smaller irregularities in dark matter and hydrogen gas amplified over time, collapsing to form the first stars and galaxies (Becker et al., 2001; Fan et al., 2006a). Fig. 1.2 shows an illustration of structure

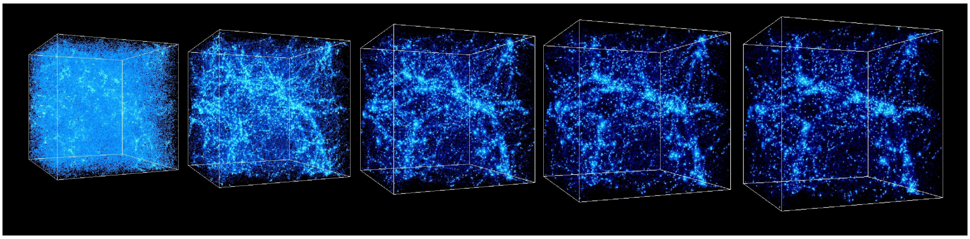

Figure 1.3 | The evolution of large-scale structures in a 43 Mpc box from redshift of 10 to the present epoch (from left to right). Credit: Kavli Institute for Cosmological Physics.

formation[1]. This period is known as the Cosmic Dawn, which ended in the Epoch of Reionisation (EoR), which is also the main interest of this thesis. The EoR is a watershed period in the history of the universe where neutral intergalactic medium (IGM) was ionised and the first luminous sources emerged Furlanetto et al. (2006); Barkana & Loeb (2007). Many of the details of the EoR scenario are not yet known. The observation of the EoR is promised to revolutionise our understanding of fundamental astrophysical processes and properties related to the formation of the first generation of stars, galaxies, and quasars.

After 1 billion years: gravity continued to aggregate galaxies and stars to form groups and clusters.

After 9 billion years: our solar system formed from a cloud of dust and gas. Our Sun is one of 100 billion stars in our Milky Way galaxy.

Today: the current universe has a complex structure, containing various objects such as planets, stars, galaxies, dust and gas. On smaller scale, gas and dust particles attract each other to make stars and stars attract each other to make galaxies; on larger scales, galaxies and matter attract each other into patterns of filaments and voids that are much larger than individual or groups of galaxies, referred to as the Large Scale Structure (LSS) of the universe. Fig. 1.3 shows the formation of the large-scale structure in the universe[2]. Our universe is still expanding (Lemaître, 1927; Hubble, 1929; Lemaître, 1931) and observations show that the expansion has accelerated over the last five billion years (Riess et al., 1998; Perlmutter et al., 1999).

1.2. 21-cm emission of Epoch of Reionisation

The main interest of this thesis is the study of the EoR. The EoR can be observed through the 'redshifted' 21-cm emission from neutral hydrogen atoms (Madau et al., 1997). Be-

[1] https://pages.uoregon.edu/imamura/123/lecture-8/wmap.html
[2] http://cosmicweb.uchicago.edu/filaments.html

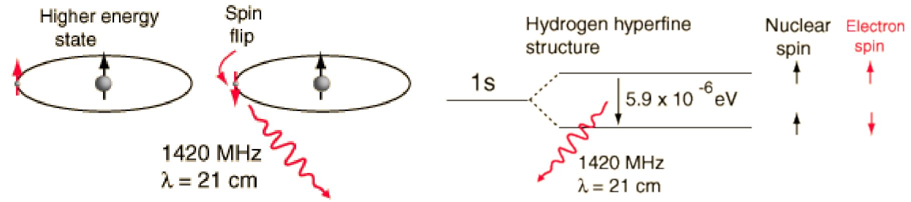

Figure 1.4 | Illustration of the 21-cm hyper-fine transition of neutral hydrogen. Credit: Hyper Physics.

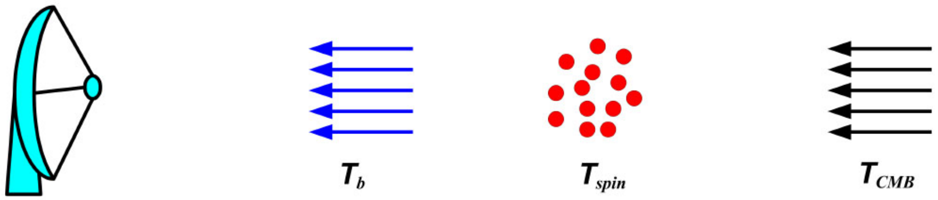

Figure 1.5 | Configuration of the 21-cm observation scheme with various components. The signal observed by our instrument is the temperature T_b after the CMB radiation goes through hydrogen gas. Credit: Zaroubi (2012).

cause the universe is expanding, the wavelength of the 21-cm signal we observe on earth is stretched, meaning that the electromagnetic emission is shifted to longer wavelengths.

The 21-cm emission is a transition between the two hyper-fine states of neutral hydrogen atoms. There are two states existing between the proton and electron of neutral hydrogen atoms: parallel spin and anti-parallel spin. When there is a spin-flip to a lower energy state, photon will be emitted with an energy equal to the difference between the two states, $\Delta\mathrm{E} \sim 5.88 \times 10^{-6}$ eV. An illustration of the 21-cm emission process is shown in Fig. 1.4. The radiation intensity is proportional to the density of the neutral hydrogen atoms and the ratio between the populations of the two spin states, given by

$$\frac{n_1}{n_0} = \frac{g_1}{g_0} e^{-\Delta\mathrm{E}/kT_\mathrm{s}}, \tag{1.1}$$

where n_0 and n_1 are the number densities of neutral hydrogen atoms in the low and high energy spin states (i.e., singlet and triplet), g_0 and g_1 are the statistical weights, and T_s is the spin temperature. The 21-cm emission is observed with respect to the background of the CMB. Depending on the CMB and spin temperatures, the 21-cm line can be observed either in emission or absorption: if the spin temperature T_s is higher than the CMB temperature T_CMB, i.e., $T_\mathrm{s} > T_\mathrm{CMB}$, the 21-cm line will be observed as an excess above the CMB temperature and vice versa. Fig. 1.5 shows a configuration of the 21-cm observation setup with various radiation components (Zaroubi, 2012).

The actual observable is the differential brightness temperature, $\delta T_\mathrm{b} \equiv T_\mathrm{b} - T_\mathrm{CMB}$, measuring the deviation of the hydrogen gas temperature from T_CMB. At later stages of the

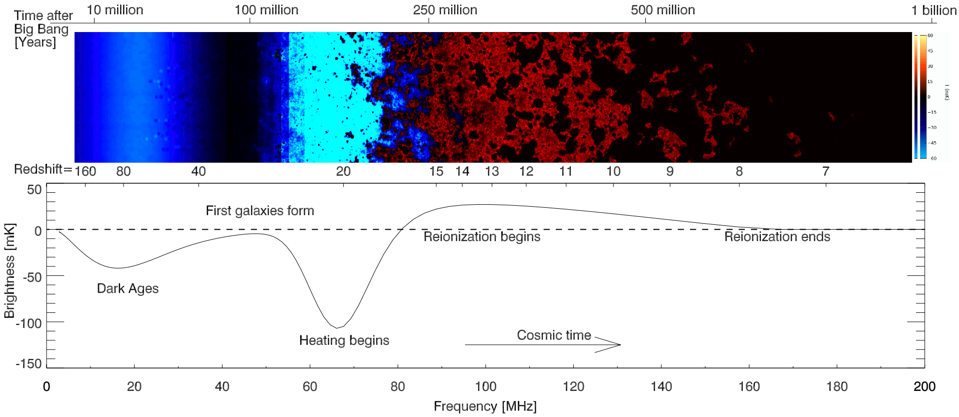

Figure 1.6 | Top: fluctuations in the 21-cm brightness as a function of redshift. Bottom: sky-averaged 21-cm brightness as a function of redshift. Credit: Pritchard & Loeb (2012).

Cosmic Dawn and EoR, we can assume a sufficiently high spin temperature ($T_\mathrm{s} \gg T_\mathrm{CMB}$) and a small optical depth τ of neutral hydrogen (Santos et al., 2008), δT_b becomes

$$\delta T_\mathrm{b} = 28\,\mathrm{mK}(1+\delta)x_\mathrm{HI}\left(1-\frac{T_\mathrm{CMB}}{T_\mathrm{s}}\right)\left(\frac{\Omega_\mathrm{b}h^2}{0.0223}\right)\sqrt{\left(\frac{1+z}{10}\right)\left(\frac{0.24}{\Omega_\mathrm{m}}\right)}\left[\frac{H(z)/(1+z)}{dv_{||}/dr_{||}}\right], \tag{1.2}$$

where (1+δ) is the fractional over-density of baryons, x_HI is the neutral fraction of hydrogen, h is the Hubble constant in units of 100 km s $^{-1}$Mpc^{-1}, z is the redshift, Ω_b and Ω_m are the baryon and mass densities of the critical density, $dv_{||}/dr_{||}$ is the gradient of the proper velocity along the line-of-sight, taken into account the uniform Hubble expansion at high redshifts and the peculiar velocity, corrected by $H(z)/(1+z)$ (for a detailed review, see Furlanetto et al., 2006; Pritchard & Loeb, 2012; Zaroubi, 2012). The equation shows the complex composition of the 21-cm signal. At different stages of the evolution, δT_b is dominated by different contributions: at high redshifts, the neutral hydrogen fraction is close to 1 ($x_\mathrm{HI} \sim 1$), δT_b is dominated by the over-density; at low redshifts, where hydrogen gas is mostly ionised ($x_\mathrm{HI} \ll 1$), δT_b is dominated by the contrast between the neutral and ionised region. Fig. 1.6 shows the evolution of the 21-cm brightness temperature at different redshifts (Pritchard & Loeb, 2012).

1.2.1. 21-cm Power spectrum

One of the most powerful statistical measures of the EoR is the power spectrum of the 21-cm differential brightness temperature, or simply, the 21-cm power spectrum, $P_{21}(k)$. Using homogeneity and isotropy of the universe, the power spectrum of brightness temperature fluctuations should be symmetric in Fourier space and depends only on the

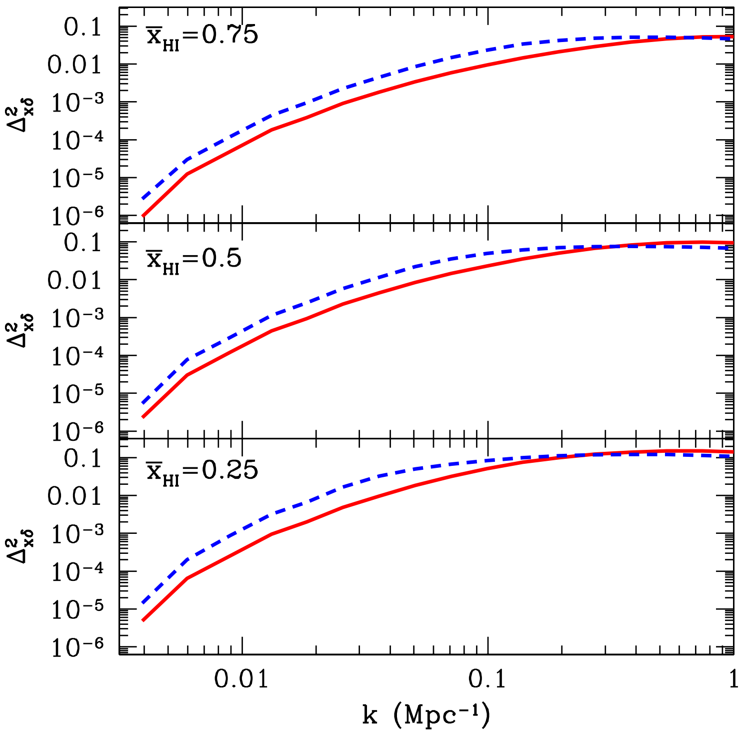

Figure 1.7 | The 21-cm power spectrum during the EoR at different neutral hydrogen fractions, simulated by two different models. Solid red lines are produced by FAINT GALAXIES model and dashed blue lines are produced by BRIGHT GALAXIES. Credit: Mesinger et al. (2016).

magnitude of k. The 21-cm power spectrum is defined as

$$\langle \widetilde{\delta T_b^*}(k)\widetilde{\delta T_b}(k')\rangle \equiv (2\pi)^3 P_{21}(k)\delta_D^3(k-k'), \tag{1.3}$$

where k is a comoving wave vector, $\widetilde{\delta T_b}$ is the Fourier transform of δT_b and δ_D^3 is the Dirac delta function in 3D. The power spectrum is often reduced to a form that denotes the variance in brightness temperate contributed by a range of scales centered on wavenumber k,

$$\Delta_{21}^2(k) = \frac{k^3}{2\pi^2} P_{21}(k), \tag{1.4}$$

where Δ_{21}^2 has the units of mK^2. Fig. 1.7 shows examples of the 21-cm power spectra (Mesinger et al., 2016).

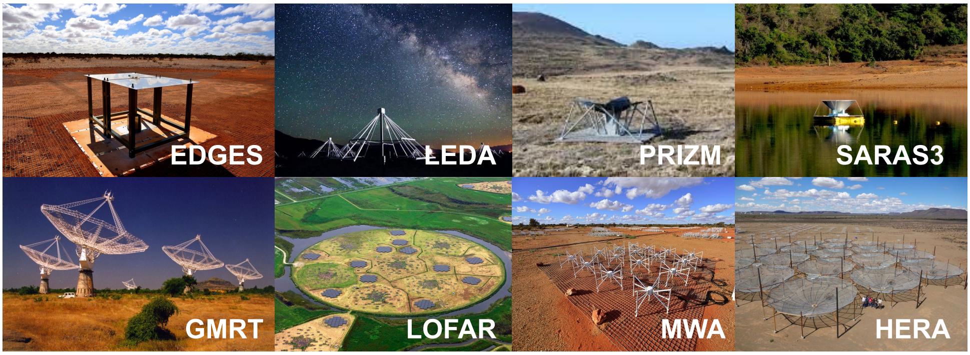

Figure 1.8 | Photographs of the 21-cm signal observation instruments. Top: 21-cm observation instruments, from left to right, EDGES (Credit: CSIRO Australia), LEDA (Credit:the Centre for Astrophysics/Harvard & Smithsonian), PRIZM (Credit: Philip et al., 2018), and SARAS3 (Credit: Ravi Subrahmanyan). Bottom: 21-cm interferometric observation instruments, from left to right, GMRT (Credit: NCRA/TIFR), LOFAR (Credit: van Haarlem et al., 2013), MWA (Credit: Dragonfly Media) and HERA (Credit: MIT Kavli Institute for Astrophysics and Space Research).

1.3. Observation of 21-cm emission

Many instruments have been designed and built to detect the redshifted 21-cm signal from the EoR. There are mainly two approaches to the 21-cm observations: (1) global 21-cm observations aiming at measuring the sky-averaged spectrum of the 21-cm signal with a single receiver, such as EDGES[3] (Bowman et al., 2018), LEDA[4] (Greenhill & Bernardi, 2012), PRIZM[5] (Philip et al., 2018) and SARAS[6] (Singh et al., 2017; Thekkeppattu et al., 2021); (2) interferometric observations aiming at measuring the spatial brightness-temperature fluctuations of the 21-cm signal with arrays of radio antennas, such as GMRT[7] (Paciga et al., 2011, 2013), LOFAR[8] (van Haarlem et al., 2013; Patil et al., 2017; Mertens et al., 2020), MWA[9] (Bowman et al., 2013; Tingay et al., 2013; Wayth et al., 2018; Barry et al., 2019; Li et al., 2019) and PAPER[10] (Parsons et al., 2012; Cheng et al., 2018; Kolopanis et al., 2019), as well as the second generation instruments, HERA[11] (DeBoer et al., 2017; HERA Collaboration et al., 2021) and the upcoming SKA[12] (Mellema et al., 2013; Koopmans et al., 2015). In this thesis, we focus on the second approach, the interferometric observations. Fig. 1.8 shows photographs of the 21-cm signal observation instruments.

[3] Experiment to Detect the Global EoR Signature

[4] the Large aperture Experiment to detect the Dark Ages, `http://www.tauceti.caltech.edu/leda/`

[5] the Probing Radio Intensity at high Z from Marion

[6] Shaped Antenna measurement of the background RAdio Spectrum.

[7] Giant Metrewave Radio Telescope, http://gmrt.ncra.tifr.res.in

[8] Low-Frequency Array, `http://www.lofar.org`

[9] Murchison Widefield Array, `http://www.mwatelescope.org`

[10] the Donald C. Backer Precision Array for Probing the Epoch of Reionisation, `http://eor.berkeley.edu`

[11] Hydrogen Epoch of Reionisation Array, `http://reionization.org/`

[12] the Square Kilometer Array, `http://www.skatelescope.org`

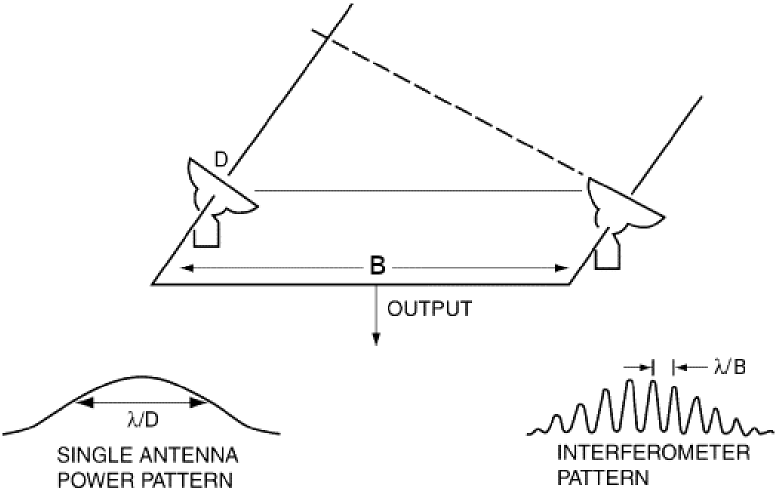

Figure 1.9 | Configuration of a simple interferometer. Credit: Gulkis & de Pater (2003).

An interferometer consists of arrays of antennas that measure the coherence between the electromagnetic field measured at the two positions of an antenna pair. The angular resolution θ of a single dish radio telescope is roughly

$$\theta \sim \frac{\lambda}{D}, \tag{1.5}$$

where D is its diameter of the telescope and λ is the wavelength of the signal. By combining multiple antennas (or telescopes), separated by a distance B (i.e., baseline), a radio interferometer mimics a single telescope of a very large aperture with a discrete sample. Fig. 1.9 shows an illustration of a simple interferometer.

The measured coherences, are known as complex visibilities $V(u,v)$, to first order, and assuming all receiver lie on a plane, represent a 2D Fourier transform of the sky brightness $I(l,m)$. Theoretically, the sky brightness and visibility are thus connected as

$$\begin{aligned} V(u,v) &= \int\int I(l,m)e^{-i2\pi(ul+vm)}dldm, \\ I(l,m) &= \int\int V(u,v)e^{i2\pi(ul+vm)}dudv, \end{aligned} \tag{1.6}$$

where (l,m) is the (cosine) angular location of the signal in the sky, with $l^2+m^2<1$, and (u,v) is the position of the signal measured on the ground in wavelengths.

Due to the finite number of receivers, the sampling function $\mathscr{S}(u,v)$ covers only part of the uv-plane further helped by the Earth rotation synthesis, and the sky brightness distribution sampled at a discrete (u,v)-plane becomes

$$I_D(l,m) = \int\int \mathscr{S}(u,v)V(u,v)e^{i2\pi(ul+vm)}dudv, \tag{1.7}$$

where I_D is often referred to as the dirty image. Using the convolution theorem, I_D can be written as

$$I_D(l,m) = \mathscr{F}(\mathscr{S}V) = \mathscr{F}(\mathscr{S}) * \mathscr{F}(V) = \mathrm{PSF}(l,m) * I(l,m), \tag{1.8}$$

where $\mathscr{F}$ indicates the Fourier transform, $*$ is the convolution operation and PSF is the Point Spread Function. To reconstruct the sky brightness, one has to deconvolve the dirty image from the PSF. The PSF is the response of the interferometer to a point source and is the Fourier transform of the sampling function $\mathscr{S}(u,v)$, the uv-coverage. The sampling function is limited by the smallest and largest separations (baselines) between antennas in the array. The maximum baseline B_{max} sets the ultimate limit to the angular resolution θ_B as

$$\theta_B \sim \frac{\lambda}{B_{\mathrm{max}}}. \tag{1.9}$$

The discrete sampling function introduces side-lobes into the PSF with nulls and secondary lobes.

The 21-cm power spectrum given by Eq. 1.3 can be estimated from the observed dirty image $I_D(l,m)$ cubes and the PSF (for details see, Harker et al., 2010; Patil et al., 2017; Mertens et al., 2020). Interferometers, by nature, however, are designed to measure the spatial correlations. Alternatively, therefore, one can obtain the power spectrum via a direct Fourier transform of visibilities along the frequency axis, known as the delay transform (Parsons et al., 2012),

$$\widetilde{V}(u,v,\tau) = \int V(u,v,f)e^{-i2\pi f\tau}df, \tag{1.10}$$

where the delay τ is the Fourier conjugate of frequency f. The delay transform is proportional to the 3D power spectrum,

$$P(k) \propto \widetilde{V}(u,v,\tau) \sim \widetilde{V}(B,\tau), \tag{1.11}$$

where B is the baseline at (u,v). The measurement units, (B,τ), can map directly into k modes perpendicular and parallel to the line-of-sight, $(k_\perp, k_{||})$ (Morales & Hewitt, 2004).

The cylindrical power spectrum, or the 2D power spectrum $P(k_\perp, k_{||})$, is another useful form of the power spectrum, enables to separate orientation of the wave number k.

Fig. 1.10 shows an example of the 2D power spectrum measured by an interferometer (Morales & Wyithe, 2010) on left and a 21-cm power spectrum obtained by a LOFAR observation on right. The measured 21-cm power spectrum intensity is averaged spherically in $|k|$ annuli (regions within curved lines, $k^2 = k_\perp^2 + k_{||}^2$). The maximum $k_\perp$ is limited by the maximum baseline and the perpendicular width of each cell is limited by 1/FoV. The line-of-sight is limited by the inverse bandwidth of the instrument. Most importantly, the smooth-foregrounds are confined into the bottom of the power spectrum, known as 'the foreground wedge' (Datta et al., 2010; Trott et al., 2012; Vedantham et al., 2012), leaving the remaining region uncontaminated, known as 'the EoR window'. In the EoR window, one can largely 'avoid' foregrounds without subtracting them (the foreground avoidance, see Morales & Wyithe, 2010; Pober et al., 2013; Kerrigan et al., 2018).

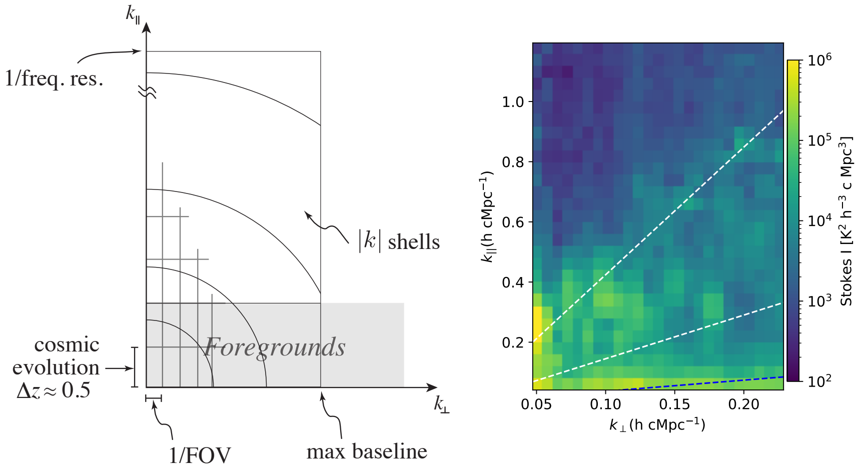

Figure 1.10 | Left: illustration of the 21-cm power spectrum measured by an interferometer. Credit: Morales & Wyithe (2010) (Fig. 10). Right: 21-cm power spectrum after DD-calibration on the NCP field obtained by one-night observation of LOFAR. The three dotted lines indicate the 5° (the primary beam), 20°, 90° (instrumental horizon) delay lines from the phase centre.

1.3.1. OBSERVATIONAL CHALLENGES

The observation of the 21-cm signal in the frequency range of 100-200 MHz is challenging for a number of reasons. First of all, the brightness temperature of the 21-cm signal is expected to be extremely weak, around 10-100 mK (Shaver et al., 1999; Datta et al., 2010).

The signal measured by the instrument is a mixture of this faint 21-cm signal and the bright galactic and extra-galactic foregrounds which are 4-5 orders of magnitude stronger than the 21-cm signal (Shaver et al., 1999). Fig. 1.11 illustrates simulated foreground contaminants in the EoR observation (Jelić, 2010b).

Ionospheric effects further distort the measured signal. The ionosphere is the upper part of the atmosphere, ~50-500 km above the Earth surface, whose electron content changes with time and position. The influence of the ionosphere on electromagnetic wave propagation increases at low frequencies and, if not corrected for, could adversely affect the 21-cm signal power spectrum (Vedantham & Koopmans, 2016; Mevius et al., 2016).

The interferometer has a complex instrumental response, which is dependent on time, frequency and direction and baseline. The instrument observes a wide range of frequency and the beam pattern changes over frequency, known as chromatic beam effects. LOFAR also has an instrumentally-polarised response. These instrumental effects are largely calibrated during processing. If not calibrated correctly or sufficiently, however, they can introduce extra noise and corrupt the signal and need to be mitigated before the data

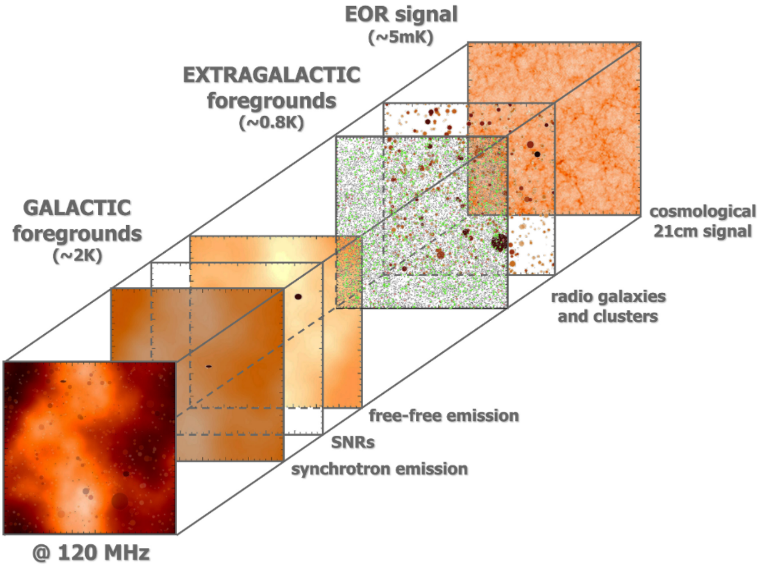

Figure 1.11 | Illustration of various simulated foreground contaminants in the EoR observation. Credit: Jelić (2010b).

analysis starts (Labropoulos et al., 2009; Jelić et al., 2015; Asad et al., 2015; Hothi et al., 2021).

Last but not least, the observed low frequency range is also susceptible to Radio Frequency Interference (RFI), introduced by human activities and wireless communication technologies (Offringa et al., 2013; Wilensky et al., 2019). These effects can severely contaminate the signal.

To be able to detect the 21-cm signal, it is crucial to understand these effects and mitigate them during the calibration process.

1.4. LOFAR-EoR KSP AND EXCESS VARIANCE

Low Frequency Array, LOFAR, is an international radio interferometer designed and constructed by ASTRON[13] in the north of the Netherlands and now spanning several countries across Europe, including France, Germany, Ireland, Latvia, Poland, Sweden, the UK and Italy. Fig. 1.12 shows the distribution of LOFAR stations across Europe. LOFAR operates at a frequency range between 10 and 240 MHz, with two types of antennas, the Low Band Antenna (LBA) and the High Band Antenna (HBA), optimised respectively for the frequency ranges of 30-80 MHz and 120-240 MHz.

The revolutionary multi-beaming capabilities of LOFAR enable astronomers to carry

[13] Netherlands Institute for Radio Astronomy; https://www.astron.nl/

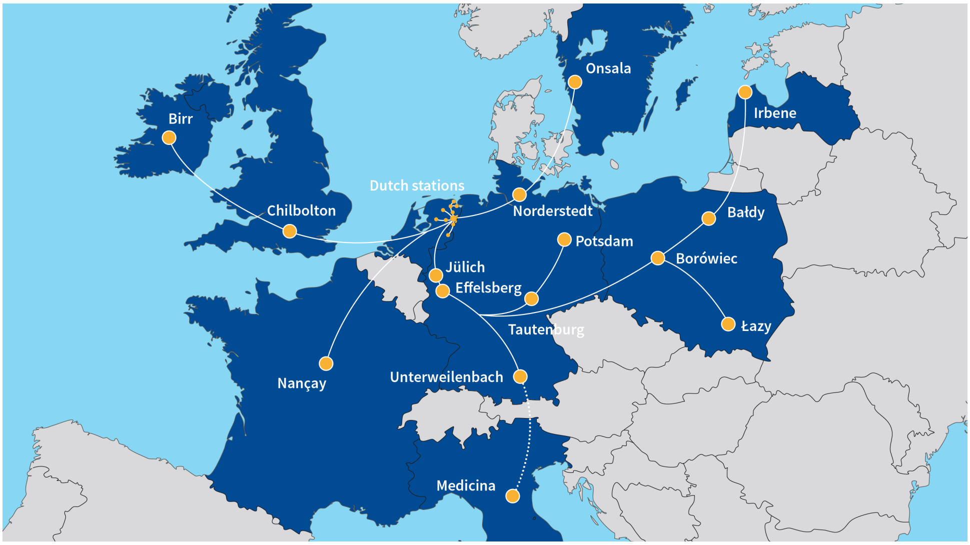

Figure 1.12 | The LOFAR station spanning nine countries, the Netherlands (38 stations), Germany (6 stations), Poland (3 stations), France, Irelend, Latvia, Sweden, the United Kingdom (one station each) and Italy (one station planned to be constructed). Credit: ASTRON.

out a wide range of scientific research including six Key Science Projects (KSPs) (van Haarlem et al., 2013): (1) Epoch of Reionisation (EoR), (2) Survey Key Project[14], (3) Transient sources[15], (4) Ultra high energy cosmic rays, (5) Solar science & space weather[16], and (6) Cosmic magnetism[17]. This thesis focuses on the LOFAR-EoR KSP.

The main science goals of the LOFAR-EoR KSP are (Labropoulos et al., 2009): (1) statistical detection of the global 21-cm signal averaged along line-of-sight (Shaver et al., 1999; Jelić et al., 2008), (2) observation of the spatial-frequency power spectrum of the 21-cm brightness temperature fluctuations in the redshift range $z = 6-11$, and (3) characterisation of ionised bubbles around bright sources and the 21-cm absorption-line forest.

Despite the extreme challenges that we face in the observations, discussed in subsection 1.3.1, impressive progress has been made over the years in analysing LOFAR-HBA 21-cm signal data to increasingly deeper levels, overcoming many of the above challenges. Mertens et al. (2020) reported results on the 21-cm signal power spectrum with better calibration scheme (Mevius et al., 2022): a best 2-σ upper limit of $\Delta^2_{21}{<}(73)^2$ mK^2 at $k = 0.075$ hc MPc^{-1}. This is an improvement by a factor ≈ 8 compared to the previous upper limits set by Patil et al. (2017). Fig. 1.13 shows the current best upper limits of several 21-cm signal observations and comparison with the simulated 21-cm signal.

[14] SKP; https://lofar-surveys.org/
[15] http://www.transientskp.org/
[16] http://www.aip.de/groups/osra/sksp/
[17] MKSP; https://lofar-mksp.org/

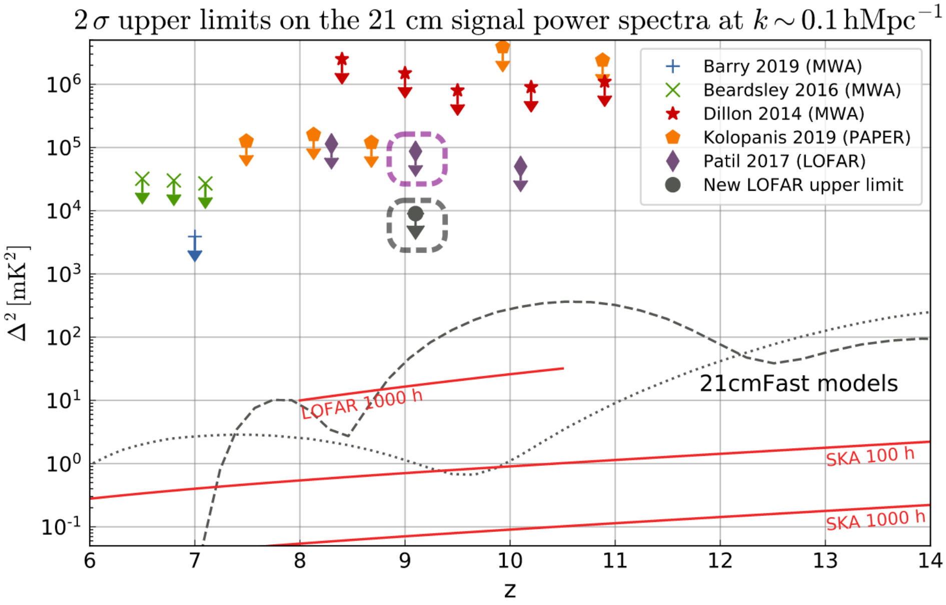

Figure 1.13 | The current best 2-σ upper limits of several 21-cm Epoch of Reionisation observations and comparison with the simulated 21-cm signal at $k = 0.1\mathrm{hMpc}^{-1}$. Credit: Florent Mertens.

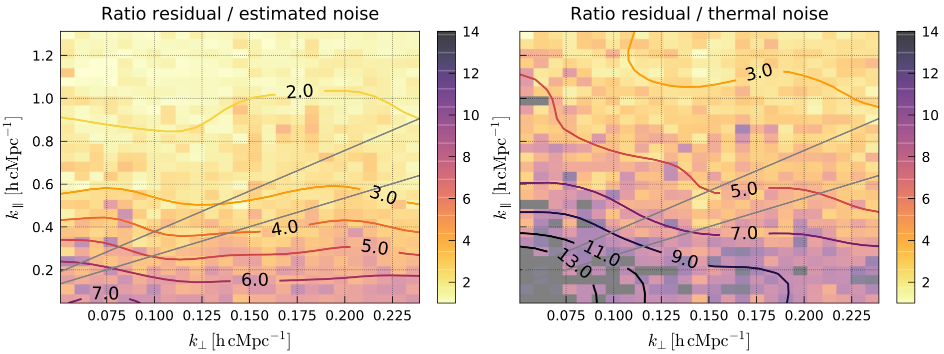

Figure 1.14 | Ratio of 2D Stokes I power spectra after residual foreground removal and the noise estimated by GPR (left). Ratio of 2D Stokes I power spectra after residual foreground removal and the thermal noise estimated from time differenced visibilities (right). Credit: Fig. 12 of Mertens et al. (2020).

Table 1.1 | Characteristics in modern astronomical sky surveys. Credit: Garofalo et al. (2017). The column volume refers to raw data produced at the end of the experiment. Values regarding Pan-STARRS, LSST, and SKA surveys refer to expected volume and volocity values.

Sky Survey	Volume	Velocity	Variety
SDSS Sloan Digital Sky Survey	50 TB	200 TB per day	images, catalogues, and redshifts
GAIA	100 TB	40 GB per day	more then 100 parameters
Pan-STARRS Panoramic Survey Telescope and Rapid Response System	5 PB	5 TB per day	images and catalogues
LSST Large Synoptic Survey Telescope	60 PB	10 TB per day	images and catalogues
SKA Square Kilometer Array	3 ZB	150 TB per day	images, catalogues, and redshifts

The results show that even after subtraction of the sky model and noise bias, the residual Stokes I shows a non-negligible excess power well above the estimated thermal noise, known as 'the excess variance' (Patil et al., 2016, 2017; Mertens et al., 2020). While the excess power above the expected thermal noise level is reduced in the new results with the improvements in the calibration scheme, the remaining excess is still significant (up to ten times higher than the thermal noise in the foreground wedge), imposing a limit on the detection, and its source remains unclear at the moment. Fig. 1.14 shows the excess level estimated by the ratio between the 2D residual Stokes I and thermal noise (Mertens et al., 2020).

1.5. Efficient handling of massive data sets

In general, astronomical data are rapidly growing in volume, dimensionality, resolution and complexity. LOFAR produces massive data sets each day and a typical volume of one-night observation is roughly 50 TByte (Patil et al., 2017). With forthcoming advances in instrumentation, such as the SKA, the volume of astronomical data will be increased by several orders of magnitude (Mellema et al., 2013; Koopmans et al., 2015; Garofalo et al., 2017, , see Table. 1.1). Traditional analysis techniques no longer can handle these massive data sets and astronomers need new paradigms for the data representation, analysis and visualisation (Ball & Brunner, 2010; Pesenson et al., 2010; Longo et al., 2019; Baron, 2019).

Such problems are addressed and studied by other disciplines of science such as computer science and applied mathematics for many years. Especially, machine learning and image processing techniques are proven to be very powerful for analysing and representing massive data sets. This thesis, in part, is dedicated to finding a solution for the more efficient analysis of big data sets by combining machine learning and image processing techniques, and applying these techniques to LOFAR data sets.

1.6. Thesis outline

This thesis is dedicated to understanding the complex correlations between the excess variance in the LOFAR-EoR 21-cm power spectrum and its potential causes as discussed in section 1.3.1. Although some assumptions about possible sources of the excess variance have been proposed (Patil et al., 2017; Mertens et al., 2020), no quantitative analysis has been conducted on this topic.

In Chapter 2, I study a number of potential causes of the excess variance based on thirteen nights of LOFAR-HBA data from observations of the North Celestial Pole. Among multiple potential causes, I focus on the impact of gain errors, the sky model, and ionospheric effects on the excess variance. I select a number of metrics that can used as measures for these potential causes, such as the gain variance over time or frequency, local sidereal time (LST), diffractive scale, and phase structure-function slope with the level of excess variance, in particular focusing on the impact of Cas A and Cyg A which are two of the brightest sources in the sky at these frequencies.

In Chapter 3, I focus on the direction-dependent gain calibration of the data and whether the 21-cm signal power spectrum results depend on the chosen algorithm. I compare the performance of two different calibration algorithms, SAGECAL and a newly developed algorithm known as DDECAL (Direction-Dependent Calibration). I analyse one of the North Celestial Pole (NCP) flanking fields with two DD-calibration algorithms. Additionally, I test two different strategies for the subtraction of Cas A and Cyg A, inspired by Chapter 2.

In Chapters 2 and 3, I find that strong residuals from Cas A and Cyg A remain in the wide field sky images after calibration and sky-model subtraction. However, the analysis was limited to one or two observations, due to the massive volume of our data. To be able to analyse numerous images automatically, in Chapter 4, I propose a new general-purpose data-science tool, the Self-organising attribute maps and pattern spectra. The self-organising maps enable one to explore clusters in vector attributes of a component tree, constructed from an image or image cubes, the max-tree, with an unsupervised machine learning technique, self-organising maps (SOMs). The method does not require an optimised set of vector attributes for feature extraction or manual thresholding of vector attributes. It allows the system to learn an optimal 2D representation from a large collection of attributes. For testing purposes, the method is applied to a set of medical positron emission tomography (PET) scans used for the lung tumour detection. Our results show that certain neurons in the self-organising attribute maps are sensitive to various morphological features in the PET scans, including organs and lung tumours. This shows great potential as a general-purpose data exploratory tool, which can also be applied the astronomical images.

In Chapter 5, I apply the Self-organising attribute maps and pattern spectra, developed in Chapter 4, to analyse anomalies in residual sky images after DD-calibration from LOFAR. I train a SOM using a subset of vector attributes of connected components from a set of very-wide field sky images. I also study the excess response of SOM neurons over local sidereal time, using self-organising pattern spectra by summing over fluxes of nodes that are assigned to the same winning neuron. The new approach also reveals temporal structures in the images which were not shown in Chapter2.

In Chapter 6, I summarise the findings of this thesis and consider how our understanding of the excess variance in the 21-cm power spectrum and the new general-purpose data-science tool, Self-Organising Attribute Maps, may benefit our future research.

Visible light only accounts for part of electromagnetic spectrum.
Observation of radio waves from dust, stars and galaxies,
reveals a hidden universe that we cannot see with our eyes.