Analyzing the Clustering of Point Sources in the 74 MHz VLSS Survey

John Capodilupo

July 15, 2010

1 Part One: Personal Section

Looking up at night, it is easy to get lost in the grandeur of the view. Space seems infinite and the myriad stars inspire a multitude of feelings. Ever since I can remember, I wanted to understand the great mysteries of the universe first getting excited via buzz words like "black holes" and "curved space time". Such fascinating ideas easily captivated my childhood curiosity. I was fortunate to have a mother and father who supported me and helped develop my interests in science and mathematics.

My interest started to become more tangible in elementary school when the school's librarian recommended me to read Stepehen Hawking's "'A Brief History of Time."' With the help of my father I finished the book and became noticeably more interested in astrophysics. I began dreaming of becoming a research scientist much like Einstein and discovering how the universe "'works."' This interest stayed with me all the way to high school becoming more and more real as my knowledge increased.

The high school I attended, Grand Rapids Catholic Central, offered a unique class that quickly grabbed my attention. The newly formed three-year "'Research Seminar"' was designed to allow students to conduct a research project with a mentor from a local university. After being accepted into the class, I began an in depth study of astrophysics and cosmology. The instructor, Mr. Andrew Moore, proved to be the perfect man for this situation constantly encouraging and pushing the students. Personally, he helped guide my development in mathematics leading to my self-study of AP Calculus BC which proved absolutely critical to being able to understand astrophysical concepts. From the study of calculus, I realized that I had a very strong interest in mathematics itself and continued to pursue it taking classes at the local community college.

In the summer between my junior and senior year, I was extremely fortunate to have been selected to be one of the 70 students to attend the Research Science Institute (RSI) held annually at MIT. There, I was teamed up with my mentor, Dr. Angelica de Oliveira-Costa, who posed several different project ideas to me. The actual project itself was conducted almost entirely during the RSI. Because of the rather short duration of RSI, coming in with programming skills and a level of mathematical maturity helped expedite the learning process and made it possible to finish the project. I had to learn a whole new type of mathematics

(the two-point correlation function) as well as understanding calculus as it applied to galaxy formation and distribution. The project I initially started work on expanded to become the one I submitted to the Intel, Siemens, and Junior Science and Humanities Symposium (JSHS) competitions.

My advice to future students is to always keep asking questions and to have perseverance. There were many times during the project in which the code I wrote did not work as intended after three or four debuggings. Even harder was the ability to grasp the ideas being presented at a level to do research in the field. Only through continued effort was this research project possible. The ability to bounce ideas and have new concepts explained by a mentor helps tremendously. I would also advise aspiring researchers to get involved in the science community whether it is through a science camp like RSI or private conversations with a professor at a local university. Nothing beats discussion of ideas to help iron out tough concepts and it can also lead to new and exciting things to explore. Although research is a lot of work, the reward is well worth it. Gaining new insight and knowledge that no one else had before is a truly unique feeling that rivals the greatest pleasures on earth. As Richard Feynman said, the "'kick in the discovery"' is what kept him researching.

2 Part Two: Research Summary

3 Overview

My research project looked at a mathematical function called the two-point correlation function and applied it to measure the clustering of galaxies in a radio survey of the sky. This is important because it was the first time such an analysis was done in a frequency relevant to a new area of astrophysics called 21 cm tomography which hopes to give us precise measurements of cosmological parameters and insight into the very early universe that would not be available otherwise. To observe these features, we must remove the contamination from foreground sources and our paper is the first step in the point source removal effort. What follows is an abbreviated version of my research paper.

4 Introduction

The main goal of cosmology is to study the large-scale structure and dynamics of our Universe. The leading cosmological model, known as the Hot Big Bang theory, states that the Universe has expanded from a primordial hot and dense initial state at some finite time in the past and continues to expand to this day. This model is supported by four observations: the expansion of space (*i.e.*, the observation that distant objects are being redshifted), the observed homogeneity and isotropy on large scales, the abundance of light elements, and the detection of the Cosmic Microwave Background (CMB) radiation (for more details, see Ryden [1]). Through a continued analysis of these four observations, constraints on cosmological parameters

have been improved; however, these observations alone represent an incomplete account of the Universe's history.

In the past few years, many authors have argued that 21 centimeter tomography, *i.e.* the three-dimensional mapping of highly redshifted 21 centimeter emission, will be the ultimate cosmological probe (see Loeb [2]). These claims stem from the 3D nature of this technique allowing for the potential to measure as many as 10^{16} independent Fourier modes. By comparison, the 2D WMAP data has about 10^6 modes.

From the time the Universe was just four hundred thousand years old to when it was one billion years old, most of the hydrogen was neutral and thereby an emitter of 21 centimeter radiation. Elementary particles have a property called spin which is analogous to angular momentum. When the electron and proton in a hydrogen atom change from having parallel spins to anti-parallel spins, radiation with a wavelength of 21 centimeters is emitted. The energy for this transition comes from interaction of hydrogen with CMB photons. The spectrum of this radiation is the only known way to probe the "dark ages" from recombination to reionization (see Barkana [3]).

It is important to note that although the 21 centimeter line occurs at a frequency of 1.42 GHz, due to cosmological redshift, this line will be observed at frequencies from about 80 MHz to about 300 MHz on Earth[1]. Therefore, observing this radiation at different frequencies (or redshifts) gives snapshots of the Universe at different periods of its history. The utility of the 21 centimeter signal for measuring cosmological parameters has just started to be fully appreciated [4]. Besides being a tool for learning about the epoch of reionization, it might also be possible to use it to obtain the sharpest constraints ever on inflation, dark energy, dark matter, and neutrinos.

However, measurement of the 21 centimeter radiation presents a multitude of problems. Foremost, the 21 centimeter signal is orders of magnitude weaker than sources of contamination such as Galactic synchrotron emission, free-free emission, extragalactic point sources, and detector noise. Figure 1 depicts an illustration of the various foreground components. Strategic placement of detectors is also crucial because, in addition to foregrounds from space, terrestrial contamination also affects the signal. Foreground reduction techniques have been recently developed in an effort to help realize the potential of 21 centimeter tomography (see Liu *et al.* [5] and references within). Although there was initially concern that foreground contamination would completely prevent any hope of 21 centimeter tomography, foreground reduction techniques have been shown to be viable and capable of high efficiency [5].

To account for galactic foreground emission, cosmologists are conducting multiple frequency analysis. Galactic emission is spectrally smooth as a function of frequency whereas the 21 centimeter signal oscillates

[1] A quantitative description for the lengthening of wavelengths due to the expansion of space is known as the cosmological redshift, denoted z, which is computed as

$$1 + z = \frac{\lambda_{obs}}{\lambda_{emit}}, \tag{1}$$

where λ_{obs} is the observed wavelength and λ_{emit} is the emitted wavelength. Observing the 21 centimeter radiation on Earth requires detectors sensitive to radiation with wavelength λ in the range of $\lambda = 21(1+z)$ centimeters where z is the cosmological redshift.

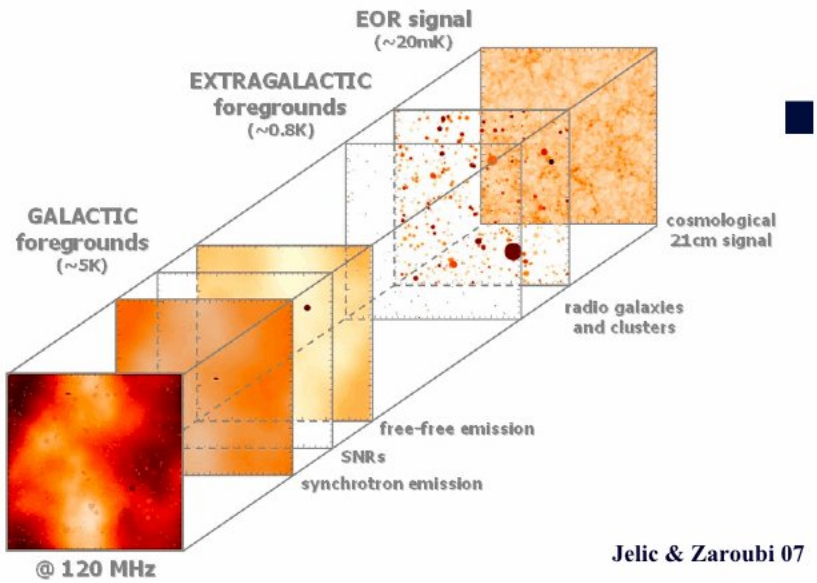

Figure 1: A visual representation of the different components of the various foregrounds contaminating the 21 centimeter signal [9]. Name of the foreground is on the bottom whereas foreground types and average temperatures are labeled on top.

wildly [6, 7]. Di Matteo, *et al.* [7] have found that efficient foreground modeling will be able to remove contamination to a sufficient extent to allow detection of angular fluctuations in the signal. Recently, de Oliveira-Costa, *et al.* [8] compiled multiple catalogs of Galactic emission and created a model to simulate Galactic emission at any frequency in the range of 10 MHz – 100 GHz. The model closely fits experimental results and, therefore, can be used to reduce unpolarized Galactic contamination. Galaxies and clusters of galaxies (known as extragalactic point sources) also obstruct the 21 centimeter signal. Liu *et al* [5] have made improvements on old methods of foreground removal by using a weighted fit to more accurately subtract contamination and leave an improved signal. Thus, by using a multiple frequency analysis and low-order polynomial fits to the foreground signal, there has been encouraging results in foreground removal.

Bright point sources can be resolved and removed from measurements of 21 centimeter radiation without damaging the signal by masking all sources above a certain flux level. However, some point sources, which are near the detector's noise level cannot be distinguished directly from the desired signal. By analyzing the distribution of extragalactic objects (including a quantitative measure of clustering), accurate simulations of the point source sky can be made. The point sources then can be correctly modeled and removed while keeping the signal intact.

We analyze distributions of extragalactic point sources in the radio range, specifically from the 74 MHz VLA Low-Frequency Sky Survey (VLSS) catalog. Detailed statistics to detect clustering of the point sources are investigated. In Section 2, a review of the VLSS catalog is given. In Section 3, the angular two-point correlation function is introduced and computed. We measure, for the first time, the clustering of point

sources at frequencies relevant to detectors of the 21 centimeter signal. We find evidence of clustering on scales less than 0.9° and fit a power law to the two-point correlation function.

5 The Point Sources

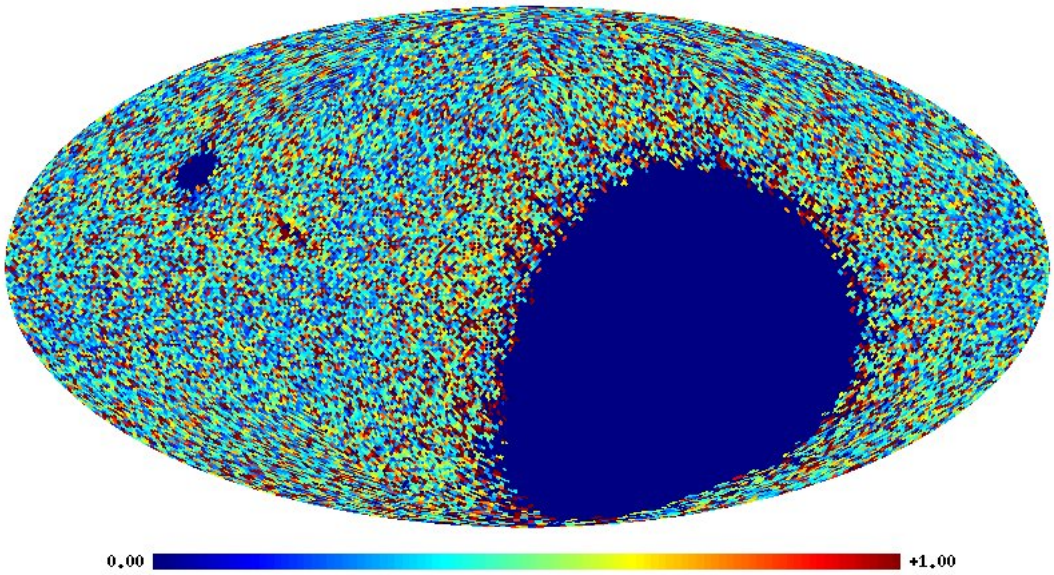

Figure 2: The 74 MHz 4MASS point source catalog. The large spot void of sources is due to the location of the VLSS detector. It is located in the Northern Hemisphere of Earth and due to the curvature of Earth, observations of the southern region of the sky cannot be made.

5.1 Catalog Overview

We chose the 74 MHz VLA Low-Frequency Sky Survey (VLSS) [10] as the galaxy survey for our study. The VLSS measures a large portion of the sky (north of declination $\delta = -30°$ with a resolution of 80 arcseconds) and the detector has an average noise of $\sigma_{\text{rms}} \approx 110$ mJy/beam, where mJy are milli-Janskys a unit of flux-density for radio sources. The source catalog contains 68,311 point sources. The entire point source catalog is visualized in Figure 2.

All full sky observations in the solar system are obstructed by the Milky Way Galaxy. The Galaxy, when viewed far enough away, appears as a flat disc. This disc forms the center of the Galactic coordinate system and is known as the Galactic plane. Although precautions to mask Galactic emission are taken when producing the catalogs, we perform a check on the data set to ensure that there were no spikes of intensity due to Galactic sources. Slices of the sky were taken at regular intervals and the intensity of the point sources was graphed as a function of galactic latitude, b. A sample graph is shown in Figure 3. Because of

the suspicious behavior of sources close to the plane, we take a cautionary measure and make cuts in the sky at $|b| = 5°$ and $|b| = 10°$.

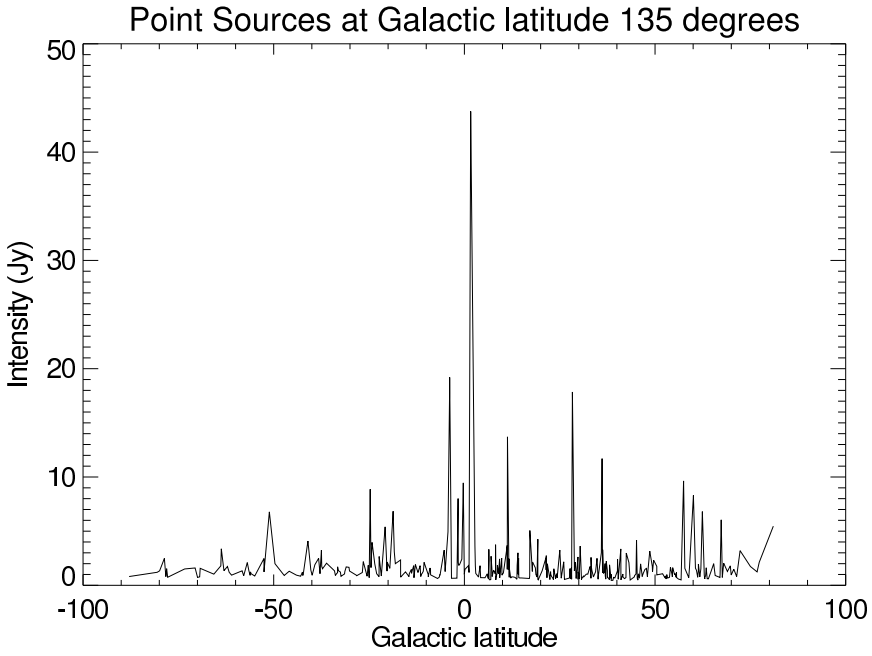

Figure 3: A sample graph showing point sources along a 1° strip in Galactic longitude of 135° The intensity from the Galactic plane seems abnormally high. As a precautionary measure, we mask a 5° and 10° region of the Galactic plane.

All maps were created using Fortran programs supplemented with HEALPix [11] routines.

6 Clustering Analysis

6.1 Two-Point Correlation Function

The main goal of the work was to measure the clustering of galaxies as a function of angular distance. Clustering of galaxies is standardly measured through a mathematical function called the spatial two-point correlation function. The analogous angular two-point correlation function is especially useful when redshift data is not available, as with the VLSS survey. The angular two-point correlation function gives the excess probability of finding two galaxies separated by an angular distance θ with respect to what is expected from a random distribution. An overview of the derivation of the function from Peebles (1980) [12] is presented.

Consider an element of solid angle, $\delta\Omega_1$, in the sky. The probability of a galaxy being within the element is given by

$$\delta P_1 = n\delta\Omega_1, \tag{2}$$

where n is the mean surface density of galaxies. The probability of finding a galaxy in another element of solid angle, $\delta\Omega_2$, is:

$$\delta P_2 = n\delta\Omega_2. \tag{3}$$

In a Poisson distribution, the two events are independent. Therefore, the joint probability is

$$\delta P = n^2\delta\Omega_1\delta\Omega_2. \tag{4}$$

However, looking at galaxy surveys, it is apparent that galaxies are not uniformly distributed but instead form clusters. There is a greater chance of finding another galaxy if it is closer to the first galaxy. The measure of this excess probability is the angular two-point correlation function, $\omega(\theta)$, defined as:

$$\delta P = n^2[1 + \omega(\theta)]\delta\Omega_1\delta\Omega_2, \tag{5}$$

where δP is the joint probability, θ is the angular separation between the two elements, and n is the average surface density of galaxies across the sky. Estimates for the angular two-point correlation function are given by Hamilton (1993) [13]

$$\omega(\theta) = \frac{DD \cdot RR}{DR^2} - 1, \tag{6}$$

and Landy and Szalay (1993) [14]

$$\omega(\theta) = \frac{DD - 2DR + RR}{RR}, \tag{7}$$

where DD is the number of pairs of galaxies from the source catalog (data-data pairs) with separation θ, DR is the number of pairs of galaxies from the source catalog and random generated sources (data-random pairs), and RR is the number of pairs of galaxies from the randomly generated sources (random-random pairs) [2]. DD can be thought of as choosing a random galaxy and counting the number of galaxies within a certain angular distance θ multiplied by a weighting function. DR and RR are similar except for DR the count is the number of galaxies from the randomly generated source catalog that are separated by θ, and RR is the number of pairs of galaxies from the randomly generated sources that are separated by θ from a randomly chosen source from the generated catalog. These measures are conducted across all points in the catalog.

Comparisons of multiple estimators, including the Hamilton and the Landy and Szalay estimators, have been previously conducted in the literature. Kerscher *et al.* [15] have found that both the Hamilton and the Landy and Szalay estimators gave better values than other estimators especially at larger angular scales.

[2]It is important to note that DD, RR, and DR are functions of θ. For readability, the function notation is dropped but assumed.

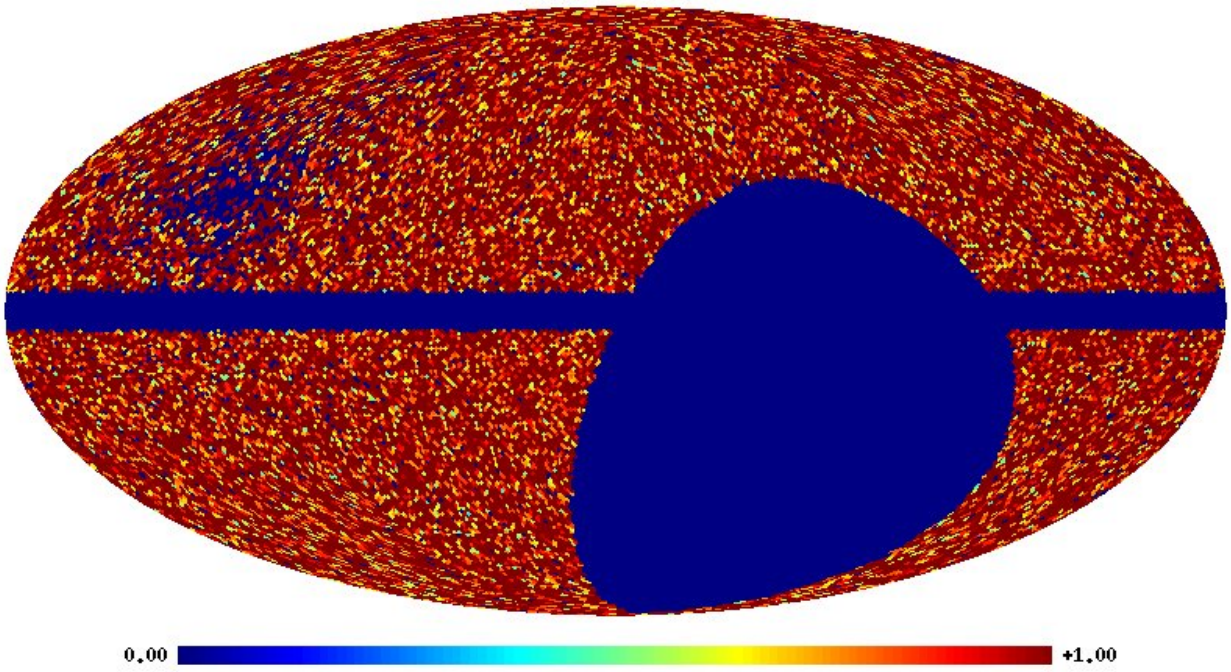

Figure 4: A sample randomly generated map of the VLSS point source sky with a 5° Galactic cut.

6.2 Simulations

To estimate the angular two-point correlation function, a randomly generated source of points uniformly distributed on the surface of a sphere is required. To avoid the incorrect over-density of points at the poles that would occur in a random generation of points on a sphere, we use the "Sphere Point Picking Algorithm" [16] [3]. Then, the coordinates on the sky, (θ, ϕ), are defined by:

$$\theta = 2\pi u, \tag{8}$$

$$\phi = \cos^{-1}(2v - 1) \tag{9}$$

where u and v are random variables in [0,1]. Each random vector is assigned to a flux of a point source from the actual catalog. In this way, we generated a simulated sky with a uniform distribution of point sources. Parts of the map with declination less than $\delta = -30°$ were then masked as the boundary conditions of the 74 MHz survey were applied. A sample generated map with a 5° cut is shown in Figure 4.

6.3 Calculations

The values for DD, DR, and RR were computed by using a correlation matrix defined as the symmetric $N \times N$ matrix, where $\frac{N}{2}$ is the number of points in the source catalog and a_{ij} is the correlation between the

[3]Note that we follow the convention that θ is the azimuthal angle ($\theta \in [0, 2\pi)$) and ϕ is the polar angle ($\phi \in [0, \pi]$).

Cut Size	A	γ	$\frac{\chi^2}{DoF}$	S_{min}
0°	0.14	-1.55	1.87	770 mJy
5°	0.44	-0.76	2.38	770 mJy
10°	0.10	-1.2	0.67	770 mJy

Table 1: A summary of the parameters for the angular correlation function at various Galactic cuts. The last column gives the flux density threshold considered.

ith point and the jth point, defined as the flux of point i multiplied by the flux of point j. The computations were made simpler by some basic results from linear algebra relating to the structure of the matrices.

To test the accuracy of our simulations and correlation algorithm, we first compute $\omega(\theta)$ for 2 of the randomly generated maps. It is expected that $\omega(\theta)$ for these random distributions to be approximately equal to zero as we expect no clustering in uniform distributions. Our computations match what is expected with $\omega(\theta) \approx 0$ for $\theta > 0.2°$.

6.4 Results

The angular two point correlation function is known to fit a power law of the form $\omega(\theta) = A\theta^{\gamma}$ particularly at small angular scales where the primary contribution to the function is the clustering of galaxies [12]. By least-squares regression on a plot of the angular two-point correlation function (see Figure 5), we obtain the best power law fit for the data. At $\theta > 0.9°$ the function tends towards zero indicating a lack of clustering at these scales. The four graphs shown in Figure 5 represent the function measured with different Galactic cuts. Since cuts greater than 10° do not significantly affect the power law fit, we choose it to be the reported two-point correlation function for this data in attempt to maximize the point sources considered. In agreement with expectations, the Hamilton estimator gives an indistinguishable fit from the Landy and Szalay estimator. Additionally, the yellow region of Figure 5 is a simple check of the non-clustering feature of our random maps. The region shows the two-point correlation function computed between two of the mock maps and, as expected, the function is zero indicating a lack of clustering. Table 1 displays the parameters of the angular correlation function for some Galactic cuts.

The angular correlation function obeys the power law fit well. A binsize of $\theta = 0.09°$ was used in calculating the two-point correlation function. This choice is safely above the resolution of the detectors in the survey. We see in Table 1 that the best fit for the angular two-point correlation is given with a 10° Galactic cut applied to the data. de Oliveira-Costa *et al.* [17] found that the two-point correlation function is independent of the flux density threshold considered. Thus, all results are given using all point sources above the completeness level of the survey, 770 mJy. On scales smaller than 0.2°, the resolution of the detector prevents accurate data from being obtained since the detector cannot resolve the individual point sources. For $\theta > 0.9°$, we find $\omega(\theta)$ to be consistent with zero showing the uniformity of the survey on these larger

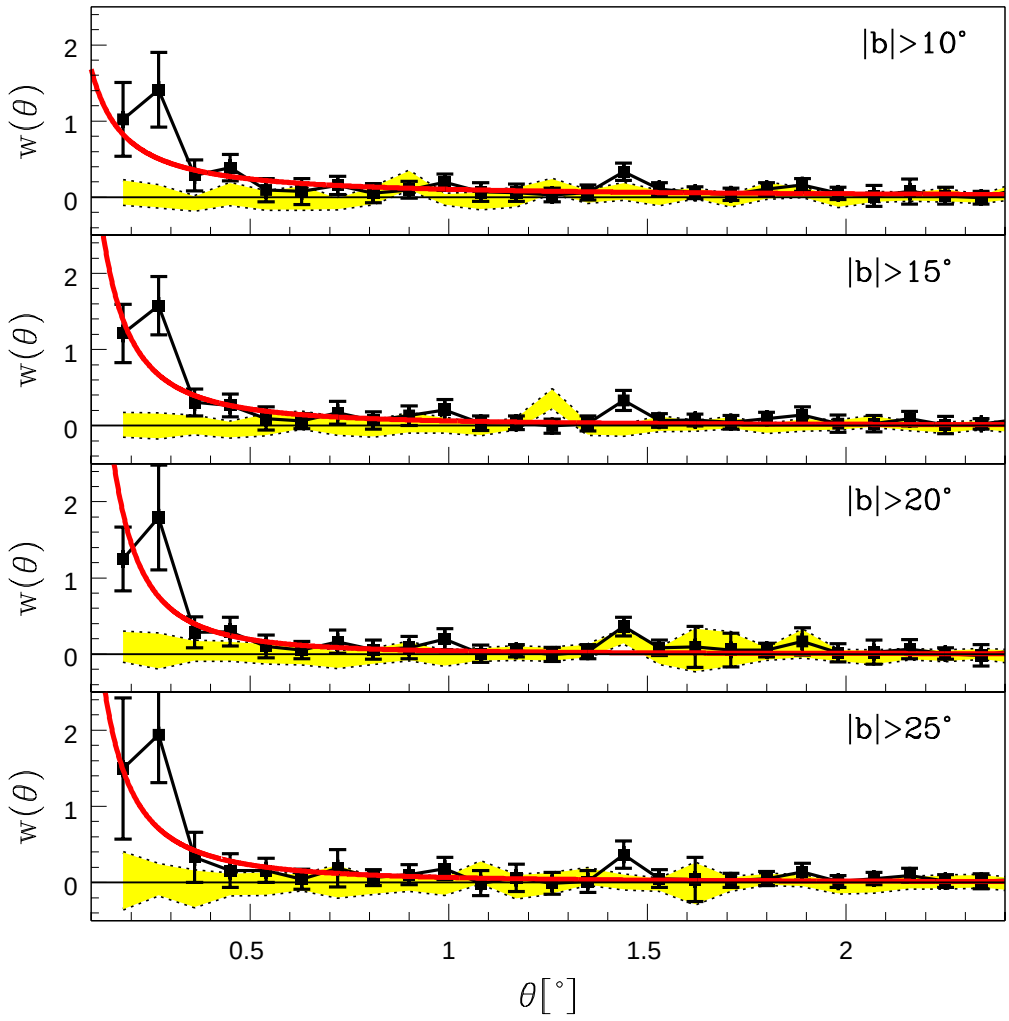

Figure 5: The Landy and Szalay estimator for the two-point correlation function. A single power law is fitted to the data and shown as the red line. From the fitted line, we derive the two parameters of $\omega(\theta)$. The yellow region is the two-point correlation function as measured for two random maps.

angular scales.

7 Conclusion

We analyzed the distribution of point sources in the 74 MHz VLSS survey and we presented, for the first time, a measure of the angular clustering of galaxies in a frequency relevant to 21 centimeter detection. In addition, we found a model for the differential source counts for the survey checking the completeness level and obtaining information about the intensity of the sources. We found evidence of clustering on scales $\theta \leq 0.9°$. On scales greater than $0.9°$, the two-point correlation is approximately zero showing a lack of clustering on larger angular sizes. By applying a $5°$ Galactic cut and using the Landy and Szalay estimator we show that $\omega(\theta) = 0.44\theta^{-0.76}$. However, the best fit is with a $10°$ Galactic cut giving a power law with slope $\gamma = -1.25 \pm 0.35$.

Although there is intrinsic interest in analyzing the clustering of galaxies in astronomy surveys, this research can also be used to create accurate simulations of the point source sky. Future studies should focus on measuring the two-point correlation function at different frequencies relevant to 21 centimeter detection in order to be able to more completely remove the extragalactic foreground. As more surveys in the 80 MHz - 300 MHz range are analyzed, combining the data into accurate simulations of the point source sky will need to be done. These simulations will serve a critical role in isolating the 21 centimeter signal while reducing contamination from point sources near the noise level of the detectors. With a clear detection of the 21 centimeter signal, details on the formation of structure as well as tighter constraints on cosmological parameters will be possible.

8 Acknowledgments

I would like to thank Dr. Angelica de Oliveira-Costa, for her guidance while providing me with an opportunity that has helped deepen my interest in cosmology. I would also like to acknowledge the support from the Center for Excellence in Education and Massachusetts Institute of Technology for hosting the 2009 Research Science Institute. I would also like to thank Dr. John Rickert, Molly Peeples, and David Levary for their advice on several related issues. I am especially grateful for the generosity received from the Leonetti/O'Connell Family Foundation which helped make this experience possible. My gratitude extends to Mr. Moore and Mrs. Eriks for their instruction and support of the Research Seminar. Finally, I thank my family who has given me invaluable support throughout my life.

References

[1] B. Ryden. Introduction to Cosmology. 1st ed. Addision Wesley, San Francisco, CA (2003).

[2] A. Loeb. The Dark Ages of the Universe. Scientific American (2006), 47-54.

[3] R. Barkana and A. Loeb. The Physics and Early History of the Intergalactic Medium. Reports on Progress in Physics 70 (2007), no. 4, 627-657.

[4] Y. Mao, M. Tegmark, M. McQuinn, M. Zaldarriaga, and O. Zahn. How accurately can 21 cm tomography constrain cosmology? Physics Review D 78 (2008), no. 2.

[5] A. Liu, M. Tegmark, J. Bowman, J. Hewitt, and M. Zaldarriaga. An Improved Method for 21cm Foreground Removal. Monthly Notices of Royal Astronomical Society (in press).

[6] A. Liu, M. Tegmark, and M. Zaldarriaga. Will point sources spoil 21 cm tomography? Monthly Notices of Royal Astronomical Society 394 (2009), no. 3, 1575–1587.

[7] T. Di Matteo, B. Ciardi, and F. Miniati. The 21 centimeter from the reionization epoch: extended and point source foregrounds. Monthly Notices of the Royal Astronomical Society 355 (2004), no. 4, 1053–1065.

[8] A. de Oliveira-Costa, M. Tegmark, B.M. Gaensler, J. Jonas, T.L. Landecker, and P. Reich. A Model of Diffuse Galactic Radio Emission from 10 MHz to 100 GHz. Monthly Notices of Royal Astronomical Society 388 (2008), no. 1, 247–260.

[9] S. Zaroubi. Probing the Epoch of Reionization with LOFAR. Available at http://confs.obspm.fr/Blois2007/PresentationsPDF/Zaroubi.pdf.

[10] A. S. Cohen, W. M. Lane, W. D. Cotton, N. E. Kassim, T. J. W. Lazio, R. A. Perley, J. J. Condon, W. C. Erickson. The VLA Low-Frequency Sky Survey. The Astronomical Journal 134 (2007), no. 3, 1245–1262.

[11] K.M. Gorski, E. Hivon, A.J. Banday, B.D. Wandelt, F.K. Hansen, M. Reinecke, and M. Bartelmann. HEALPix: A Framework for High-resolution Discretization and Fast Analysis of Data Distributed on the Sphere. Astrophysical Journal 622 (2005) 759–771

[12] P.J.E. Peebles. The Large-Scale Structure of the Universe. Princeton University Press (1980).

[13] A.J.S. Hamilton. Toward Better Ways to Measure the Galaxy Correlation Function. Astrophysical Journal 417 (1993), 19–35

[14] S. D. Landy and A. S. Szalay. Bias and variance of angular correlation functions. Astrophysical Journal 412 (1993), no. 1., 64–71

[15] M. Kerscher, I. Szapudi, and A. Szalay. A Comparison of Estimators for the Two-Point Correlation Function. Astrophysical Journal 535 (2000), L13–L16

[16] Weisstein, Eric W. "Sphere Point Picking." From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/SpherePointPicking.html.

[17] A. de Oliveira-Costa and J. Capodilupo. Clustering at 74 Mhz. Submitted to Monthly Notices of the Royal Astronomical Society. arXiv:0908.4248v1 [astro-ph.CO].