UNIVERSITY *of the*
WESTERN CAPE

Probing the large-scale structure of the Universe with Correlation Functions

Siyambonga Donald Matshawule

16 May 2014

A thesis submitted in partial fulfilment of the requirements for the degree of Magister Scientiae
in the Department of Physics.
UNIVERSITY OF THE WESTERN CAPE

Supervisors: Prof. R. Maartens and Dr. R. W. Johnston

Abstract

The current standard concordance model of cosmology represents the most concise model to date, combining astronomical observations with theoretical predictions to explain the origins, evolution, structure and dynamics of the Universe. One of the most fundamental and powerful probes of the standard model is the two-point correlation function (2PCF). The 2PCF measures the clustering of galaxies as a function of their spatial separation and determines if the number of sources is larger or less than that expected if the galaxies are distributed randomly on the sky. Measuring the 2PCF from the recent galaxy surveys such as the Sloan Digital Sky Survey (SDSS) and the Baryon Oscillation Spectroscopic Surveys (BOSS) allow us to probe a wide range of cosmological properties as well as galaxy evolution: from galaxy bias, which compares the clustering of baryonic matter with the underlying dark matter distribution, to placing powerful constraints on the Hubble rate, the dimensionless matter density constant, Dark Energy and primordial non-Gaussianity. As we move into an era where we will have access to overwhelmingly large data-sets, applying such tools becomes a computational challenge. In this project we explore a new statistical package called KSTAT that computes the 2PCF and higher order correlation functions, such as the 3PCF, on BOSS, making use of high performance computing facilities to improve optimization. The higher order statistics such as the 3PCF can be used to probe primordial non-Gaussianity. We present the first and most precise measurements of the reduced 3PCF (Q) measured using KSTAT on the DR10 data release at higher redshift, DR10 CMASS ($\bar{z} \sim 0.5$) and LOWZ ($\bar{z} \sim 0.3$). Our reduced 3PCF results at low redshift $z = 0.3$ are consistent with those of McBride et al. (2011a) at $z = 0.104$ in the SDSS LRG sample. In this initial analysis, we have found no evidence of evolution of the 3PCF on small and large scales, however we do observe the characteristic U shape of the reduced 3PCF as one increases scale.

Acknowledgements

First and foremost I would like to express my gratitude to the Lord, who gave me the strength and courage to do this project. I would like to express my sincere gratitude to my supervisors, Russell Johnston and Roy Maartens. Russell, thank you for your unwavering support, guidance, and enthusiasm throughout this project. I gained invaluable knowledge from you, in and out of research. I enjoyed every moment and yes we managed to lay down the rails just as the train arrived at the station. It was a close one!.

To Roy, thank you very much for the opportunity you granted me to further my studies. Thank you for keeping your door open and being willing to assist every time I came to you. Thanks to Cristiano Sabiu for allowing to work with him, thanks for your prompt response to my emails, your inputs were greatly appreciated. I would also like to thank UWC Astrophysics Group, for creating an environment where you are encouraged to become the best you can be. Thanks to Mathew Smith for proofreading this beast and the support to write more. I will master the skill one day.

I would like to also give a special thanks my loving parents, I couldn't have done it without them. Another special thanks goes to my loving sister Fezeka Matshawule, who has supported me throughout the writing process, nangamso Qhawekazi!! I must also thank my friends Elethu Mvusi, Siyasanga Njambatwa and Sisanda Loleka for keeping me sane and encouraging me to follow my dreams. Uyabulela uMnqabe.

Thank you to the South African Square Kilometre Array Project and the South African National Research Foundation for the financial assistance.

Plagiarism Declaration

I, *Siyambonga Donald Matshawule*, know the meaning of plagiarism and declare that all of the work in the document titled '*Probing the large-scale structure of the Universe with Correlation Functions* ', save for that which is properly acknowledged, is my own.

Signed: ____

Date: ____

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The current Standard Model of cosmology

The current standard model cosmology, known as ΛCDM or the 'concordance' model, is a remarkably simple model which nevertheless has been extremely successful to date, able to accommodate a wide range of astronomical observations and theoretical predictions to describe the origins, evolution and structure of the Universe. Here we give only a brief summary of the main features of the model that are relevant for the thesis. For further details, see for example Dodelson (2003), Liddle (1999) and Ryden (2006).

The key features of the concordance model may be summarized as follows.

1.1.1 The Cosmological Principle

The Cosmological Principle asserts that the Universe appears the same at all points and in all directions to any fundamental observer (i.e., an observer who is comoving with the galaxies). More precisely, at any fixed cosmic time, all points in the universe are equivalent (the same density, temperature, etc.) – known as homogeneity, and at any point all directions are equivalent – known as isotropy. This means that the geometry of the spatial universe is as simple as possible. There are three possible geometries that are homogeneous and isotropic– flat, open and closed, and the concordance model has a flat spatial geometry. The spacetime metric is then

$$ds^2 = -dt^2 + a^2(t)\left[dx^2 + dy^2 + dz^2\right], \tag{1.1}$$

where $a(t)$ is the scale factor, which allows for evolution of the spatial universe with time. This metric is known as the Friedmann-Lemaître-Robertson-Walker (FLRW) metric.

The Cosmological Principle effectively says that there are no special points in the universe, and that the only variation occurs with time. Strictly speaking, it needs to be understood statistically and on large enough scales, since clearly the universe is not homogeneous on galactic and smaller scales.

1.1.2 The expanding universe

The cosmic microwave background (CMB) blackbody radiation (see below for more detail) is very strong evidence that the universe was hotter in the past and that it has cooled through expansion. The expansion of the universe is further evidenced by the redshift z in the spectra of all distant galaxies:

$$1 + z = \frac{\lambda_o}{\lambda_e} = \frac{a(t_0)}{a(t_e)}, \tag{1.2}$$

where λ_o is the observed wavelength, λ_e the emitted wavelength, and t_0, t_e are the times of observation and emission. In the expanding universe, $z > 0$ so that a is increasing. The expansion of the universe stretches photon wavelengths such that $\lambda \propto a$. Similarly, the distance between any 2 galaxies, measured at fixed time, increases as $d \propto a$, since the galaxies are at rest in the FLRW metric, i.e. they have $(x, y, z) =$ const (this ignores any peculiar velocities).

We are free to choose $a(t_0)$, the scale factor today, i.e. with t_0 equal to the age of the universe ($t_0 \approx 14\,\text{Gyr}$). By convention, we choose $a(t_0) = 1$. Then the redshift corresponding to a cosmic time t is given by

$$1 + z = \frac{1}{a(t)}. \tag{1.3}$$

The temperature of the CMB blackbody radiation is given by

$$T = \frac{T_0}{a} = T_0(1 + z), \tag{1.4}$$

where $T_0 = 2.7255 \pm 0.0006\,\text{K}$.

The rate of expansion of the universe is described by the Hubble parameter

$$H = \frac{\dot{a}}{a}. \tag{1.5}$$

Its value today is given by

$$H_0 = 100h\,\text{km}\,\text{s}^{-1}\text{Mpc}^{-1}, \quad h = 0.679 \pm 0.015, \tag{1.6}$$

where the observed value with uncertainty is from the Planck CMB experiment (Planck Collaboration et al. 2013).

1.1.3 Einstein's equations

Einstein's equations for the concordance model reduce to two ordinary differential equations:

$$\begin{aligned} H^2 &= \frac{8\pi G}{3}\rho + \frac{\Lambda}{3}, \qquad (1.7)\\ \frac{\ddot{a}}{a} &= -\frac{4\pi G}{3}(\rho + 3p) + \frac{\Lambda}{3}, \qquad (1.8)\end{aligned}$$

where $\rho = \rho_m + \rho_r$ is the total density of matter and radiation, $p = p_m + p_r$ is the total pressure, and Λ is the cosmological constant. The first is the Friedmann equation, governing the rate of expansion of the universe in response to the total density. The second is the acceleration equation. This shows that matter and radiation make a purely negative contribution, $\ddot{a}$ – i.e. the expansion of the universe always decelerates if the universe is dominated by matter and / or radiation. By contrast, the cosmological constant, which is positive, always has an accelerating effect, $\ddot{a} > 0$.

We are interested in this thesis in structure formation in the universe, which takes place when radiation makes a negligible contribution compared to matter. Therefore the Einstein equations become

$$\begin{aligned} H^2 &= \frac{8\pi G}{3}\rho + \frac{\Lambda}{3}, \qquad (1.9)\\ \frac{\ddot{a}}{a} &= -\frac{4\pi G}{3}\rho + \frac{\Lambda}{3}, \qquad (1.10)\end{aligned}$$

where $\rho = \rho_m$. The matter has negligible pressure, $p_m = 0$, i.e. it is cold, non-relativistic matter. The evolution of the matter density is governed by the conservation (or 'continuity') equation,

$$\dot{\rho} + 3H\rho = 0, \qquad (1.11)$$

which follows from the Einstein equations. The solution of this equation is

$$\rho = \frac{\rho_0}{a^3} = \rho_0(1+z)^3, \qquad (1.12)$$

which simply expresses the fact the the volume evolves as $\propto a^3$ and the density evolves as $\propto V^{-1}$. The density today is $\rho_0 \sim 10^{-30}\,\mathrm{g\,cm^{-3}}$.

We can define dimensionless density parameters for matter and dark energy as

$$\Omega_m = \frac{8\pi G\rho_0}{3H_0^2}, \quad \Omega_\Lambda = \frac{\Lambda}{3H_0^2}. \qquad (1.13)$$

By the Friedmann equation

$$\Omega_m + \Omega_\Lambda = 1. \qquad (1.14)$$

Planck data (Planck Collaboration et al. 2013) gives

$$\Omega_m = 0.307 \pm 0.019 \,. \tag{1.15}$$

1.1.4 Cold matter

During the era of structure formation, i.e. for redshifts z less than about 100, the matter in the universe is cold and non-relativistic, so its thermal pressure is negligible. The baryonic matter (protons and neutrons) in the form of hydrogen and other elements, makes up the stars and gas clouds in galaxies. It becomes cold after decoupling from radiation at $z = 1090$ (before decoupling, baryonic matter is coupled to radiation via the scattering of photons off free electrons and the coupling between electrons and protons). There is very strong evidence for the existence of another type of cold matter, known as cold dark matter (CDM)(Zwicky 1933, Rubin et al. 1980, Sofue & Rubin 2001). This is 'dark' because it does not interact with photons and so cannot be directly detected by optical or radio telescopes. It also does not interact with baryonic matter – except through gravity.

So the cold matter is made up of two parts:

$$\rho = \rho_b + \rho_{cdm}, \quad \text{and} \;\; \Omega_m = \Omega_b + \Omega_{cdm}. \tag{1.16}$$

By measuring the orbital velocities of stars in spiral galaxies, it can be shown that these velocities are too high if the galaxies are made up only of baryonic matter. Furthermore, the growth of galaxies after decoupling to the present time cannot match the observations if only baryonic matter is present. These two features require that CDM should make up about 80% of the total cold matter. CMB observations confirm this and give a more precise determination (Planck Collaboration et al. 2013):

$$\Omega_b h^2 = 0.022, \quad \Omega_{cdm} h^2 = 0.12 \,. \tag{1.17}$$

1.1.5 Dark energy

Λ plays the role of 'dark energy' – the quantity that drives the current acceleration of the universe. We can interpret $\rho_\Lambda = \Lambda/(8\pi G)$ as the energy density of the vacuum. This remains constant as the universe expands, $\dot{\rho}_\Lambda = 0$.

At earlier times, the matter dominates over dark energy, $\rho > \rho_\Lambda$, and the universe continues to decelerate, $\ddot{a} < 0$. But ρ is decreasing with expansion while ρ_Λ remains constant. Eventually, the dark energy will dominate and the expansion of the universe will begin to speed up, $\ddot{a} > 0$. This happens at time t_{acc}, where $\ddot{a}(t_{acc}) = 0$, i.e. when $\rho(t_{acc}) = \Lambda/(4\pi G)$ by

Eq. (1.10). Using Eq. (1.12), the redshift when acceleration starts is

$$z_{acc} = \left(\frac{\Lambda}{4\pi G\rho_0}\right)^{1/3} - 1 = \left(\frac{2\Omega_\Lambda}{\Omega_m}\right)^{1/3} - 1. \tag{1.18}$$

1.1.6 Structure formation

After decoupling of matter and radiation, the baryonic matter is freed from radiation pressure and able to collapse under gravity, leading to the formation of the first stars, and then galaxies and clusters of galaxies. This process requires the existence of 'seed' perturbations, i.e. small inhomogeneities in the matter. The slightly overdense regions become more overdense as more matter falls in under gravity. Eventually the overdensity grows large enough for a star to form (and similarly for galaxies and clusters of galaxies).

In the concordance model, the original (or primordial) inhomogeneities are generated by quantum fluctuations in the very early universe during a brief high-energy period of 'inflation'. The nature of the perturbations generated by inflation has been confirmed by CMB observations.

The other critical ingredient in the formation of large-scale structure (galaxies and clusters) is CDM. The CDM is immune from radiation pressure and so is able to undergo gravitational collapse before decoupling. The collapse does not form stars or galaxies, which are baryonic – rather, it forms dark 'halos' of condensed CDM. The CDM halos speed up the process of galaxy formation: galaxies form in pre-existing large halos.

The fractional matter overdensity is

$$\delta = \frac{\delta\rho}{\bar{\rho}} = \frac{\rho - \bar{\rho}}{\bar{\rho}}, \tag{1.19}$$

where $\bar{\rho}$ is the average density – which obeys the equations (1.9)–(1.11). Inside galaxies and clusters we have $\delta \gg 1$ and the process of galaxy formation itself is nonlinear and requires N-body simulations. On larger scales, δ is small and we can use perturbation theory about the FLRW background. Then δ obeys the equation

$$\ddot{\delta} + 2H\dot{\delta} - 4\pi G\bar{\rho}\delta = 0. \tag{1.20}$$

This shows that expansion ($H > 0$) slows down the growth of structure compared to a non-expanding universe, as expected. Dark energy speeds up the expansion, which further suppresses the rate of growth of structure. By measuring the distribution of galaxies at different redshifts, we can obtain estimates of the rate of growth of structure and the strength of clustering, and these will estimates will contain information about the dark energy.

1.2 Observational Probes

Cosmic Microwave Background

The Cosmic Microwave Background (CMB) is isotropic microwave radiation that travelled from the Big Bang, 14 billion years ago. The initial state of the Universe was that of a very hot plasma composed of mostly protons, electrons, photons and neutrons. During the first 380 000 years the photons were not able travel large distances as they were regularly interacting with the free electrons, this caused the Universe to become opaque. The expansion of the Universe caused the temperature to cool to approximately 3000 K. It was then possible for protons to combine with the free electrons to form hydrogen atoms. With the dissipation of the free electrons, the photons were able to travel freely throughout the Universe, making the Universe transparent. Over the 14 billion year period, the photons have been redshifted to longer wavelengths of roughly 1 millimetre and have cooled to a temperature of about 2.7 K. Today the CMB is observed as a background glow that fills the entire Universe (Planck Collaboration et al. 2013). The discovery of the CMB was made accidentally in 1964 by American radio astronomers Arno Penzias and Robert Wilson. This find provided confirmation of the existence of the Big Bang and earned both astronomers the 1978 Nobel Prize for Physics. Other major studies of the CMB making used of space probes, the first was the Cosmic Background Explorer (COBE) launched by NASA in 1989, followed by the Wilkinson Microwave Anisotropy Probe (WMAP) in 2001 and Planck launched by the European Space Agency in 2009. See Figure 1.1 for the evolution of these CMB probes

Observations of the CMB provide scientists with an opportunity to learn about how the structure of the Universe ∼ 400 000 years after the Big Bang; the furthest any instrument can probe using light. The tiny density fluctuations present in the CMB are thought to be seeds of structure formation. Using the light from the CMB and other cosmological probes scientists can describe the evolution of the Universe from the CMB to the large scale structure we observe today. These observations also provide a way to test current models about the origin, structure and evolution of the Universe.

Baryon Acoustic Oscillations

Baryon Acoustic Oscillations (BAO) are sound waves that propagated in the early Universe and were imprinted on the CMB fluctuations. Figure 1.5 illustrates the basic picture of the BAO on the CMB, the figure is not drawn to scale for illustration purposes. The BAO is observed as a feature on the correlation function with characteristic scale length of 110 Mpc (also referred to as the BAO peak), where the large scale structure (stars, galaxies, clusters) we observe today grew from. This characteristic scale length also provides a way to use the BAO as "standard rulers" with which cosmological parameters can be constrained. (see e.g. Eisenstein et al. 2005, Bassett & Hlozek 2010, Blake et al. 2011, Sánchez et al. 2012, Anderson et al. 2012, Tojeiro et al. 2014, Nuza et al. 2013, Anderson et al. 2014).

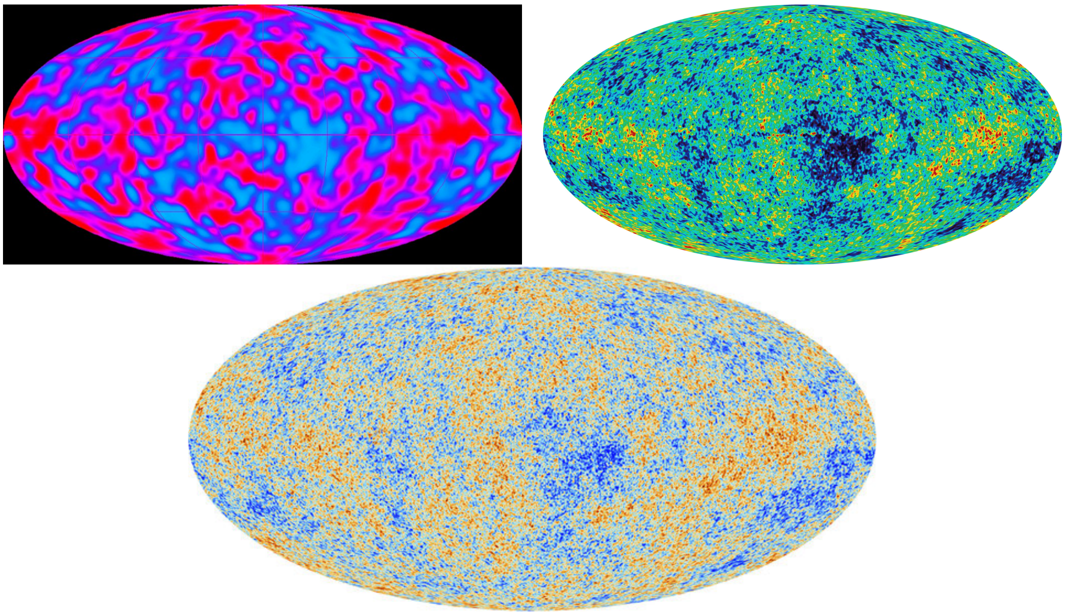


Figure 1.1: Temperature fluctuations on the CMB over a period of $\sim 400,000$ years. On the top left is CMB observations by COBE (Smoot 1995), and on the top right is CMB observations by WAMP (Wright 2003). The latest measurements which were made by the Plank ESA mission (Planck Collaboration et al. 2013).

Type Ia Supernovae

Type Ia supernovae are thought to be explosions of white dwarfs in binary systems. In this scenario, the white dwarf accretes matter from the companion star until it reaches the Chandrasekhar mass limit of $1.44 M_{\odot}$, when nuclear reaction occurs and the star explodes. The energy emitted is 10^{51} ergs, making these amongst the brightest known astrophysical events, and allows them to be used as "standard candles" (i.e. objects that give out a known absolute luminosity). The characteristic luminosity and brightness of type Ia supernova makes them ideal to measure distances to high redshift. By comparing the measured luminosity distance to that inferred from the redshift of the supernova, tight constraints on the cosmological model can be obtained. Observations of type Ia supernova at $z \sim 1$ have shown that the Universe is currently undergoing a period of accelerated expansion, driven by the existence of Dark Energy in the Universe. (Riess et al. 1998, Perlmutter et al. 1999, Conley et al. 2011, Sullivan et al. 2011).

1.3 Mapping the Large Scale Structure of the Universe

Since the 1970s galaxy redshift surveys have always provided important of information to astronomers, not only about the distribution of galaxies within the Universe, but also providing clues to fundamental questions about structure formation and galaxy evolution. Before this era, models of how the galaxies were distributed in the Universe were based on 2D images of galaxies projected onto the plane of the sky. While the distribution of galaxies in the Universe could be inferred from these 2D catalogs (e.g. Charlier 1922), 3D information of the spatial distribution was required in order to fully understand the clustering nature of galaxies. This 3D information came in the form of the redshift. The redshift of a galaxy is obtained from spectroscopy by selecting spectral lines (typically emission lines), noting their wavelength λ. By measuring the relative shifts of these lines from their known position from laboratory measurements λ_0 we obtain a measure of the redshift

$$z = \frac{\lambda - \lambda_0}{\lambda_0}. \tag{1.21}$$

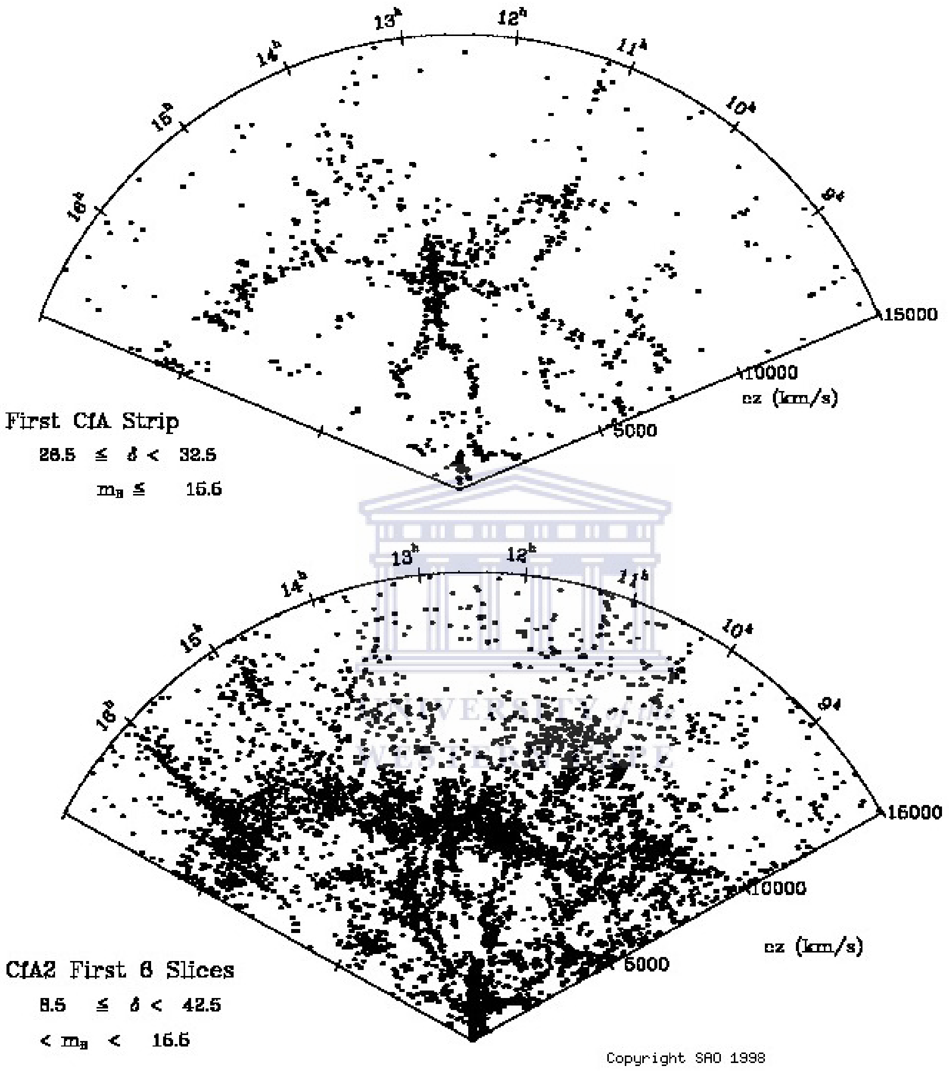
The first major survey to map the nearby large scale structure of the Universe was the CfA galaxy redshift survey (Huchra et al. 1983) which ran from 1977 to 1982. It was the first large area and moderately deep survey obtaining redshifts for 2401 galaxies with an apparent magnitude limit brighter than $m_{\rm lim} = 14.5$. This survey served as confirmation of the 3D properties of the clustering nature of galaxies predicted from 2D images of galaxy distributions. CfA1 (Geller & Huchra 1989) was then closely followed by CfA2 which obtained spectroscopic redshifts for 18,000 objects with a $m_{\rm lim} = 15.5$ out to 15,000 km s^{-1}. Figure 1.2 shows the galaxy distribution wedges from CfA1 (top panel) and CfA2 (bottom panel).

The 2 Degree Field Galaxy Redshift Survey (2dFGRS)

The second major survey conducted near the end of the 20th century was 2dFGRS. This survey ran from 1998-2003, using the Anglo-Australian Telescope(AAT) to obtain 245,591 redshifts, in which 220,000 of these were galaxies with an extinction corrected magnitude limit of $b_j = 19.45$ mag. The AAT made these measurements using a multi-fibre spectrograph which was able to obtain 400 redshifts simultaneously. Some of the 2dfGRS science survey goals included measuring the matter power spectrum up to 300 h^{-1} Mpc (Percival et al. 2001), measuring the galaxy bias parameter from the galaxy distribution and higher order correlations (Verde et al. 2002) and making measurements of luminosity functions and spectral types, (Folkes et al. 1999). Figure 1.3 shows the large scale distribution of galaxies within the 2dfGRS survey.

Sloan Digital Sky Survey (SDSS)

The SDSS survey York et al. (2000) is the most utilised and influential survey to date. The survey was conducted with the dedicated 2.5 meter aperture Sloan Foundation Telescope at the Apache Point Observatory, New Mexico (Gunn et al. 2006). The survey mapped a quarter of the sky out to $z \sim 0.2$ using multiband photometry. Since the first data release DR1 to DR7, SDSS has obtained over 2 million spectroscopic redshifts of galaxies and imaged over 300 million galaxies over a coverage area of $\sim 10,000$ deg^2. Various studies have been conducted using SDSS data, from large scale clustering, galaxy evolution, star formation and constraining cosmological parameters, to name a few. Figure 1.4 shows the galaxy distribution of the SDSS survey.



Copyright SAO 1998

Figure 1.2: Galaxy distribution wedges from the CfA redshift surveys. Each wedge on the sky is 6 degrees and 130 degrees long with the observer situated at the apex of the wedge. This was before multi-object spectrographs, redshifts were measured for one galaxy at a time. This survey provided confirmation of the clustering nature of galaxies. [Image courtesy of the Smithsonian Astrophysical Observatory (Falco et al. 1999)]

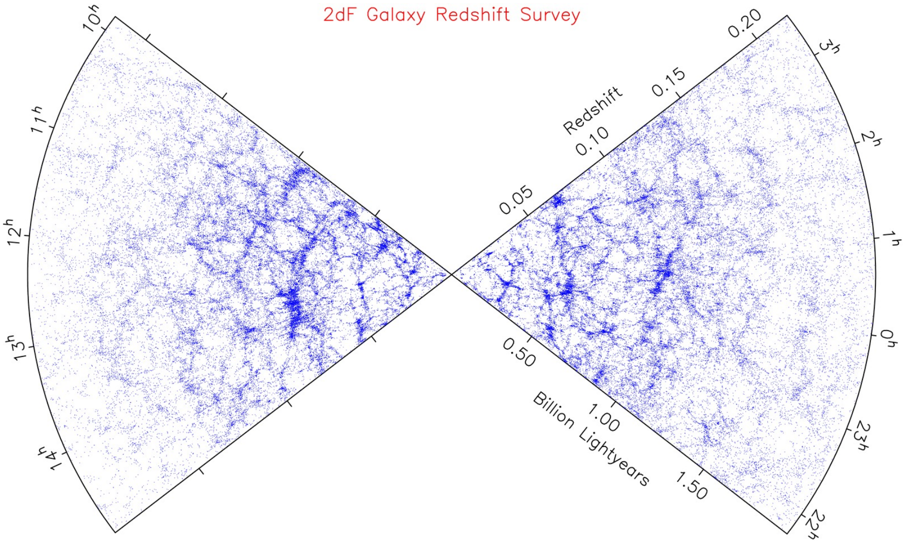


Figure 1.3: The 2dfFGRS final galaxy distribution wedge. 2dFGRS obtained spectra of 220,000 galaxies up to $z \sim 0.3$. [Image courtesy of http://www.aao.gov.au]

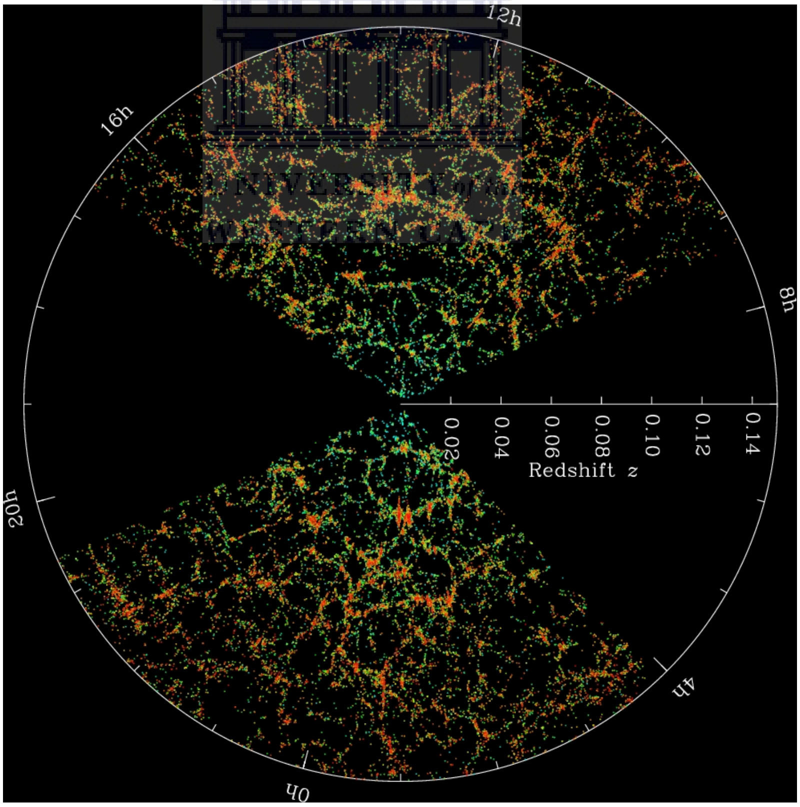


Figure 1.4: The Sloan Digital Sky Survey (SDSS) distribution of galaxies with spectroscopic redshifts. [Image courtesy of http:www.sdss.org]

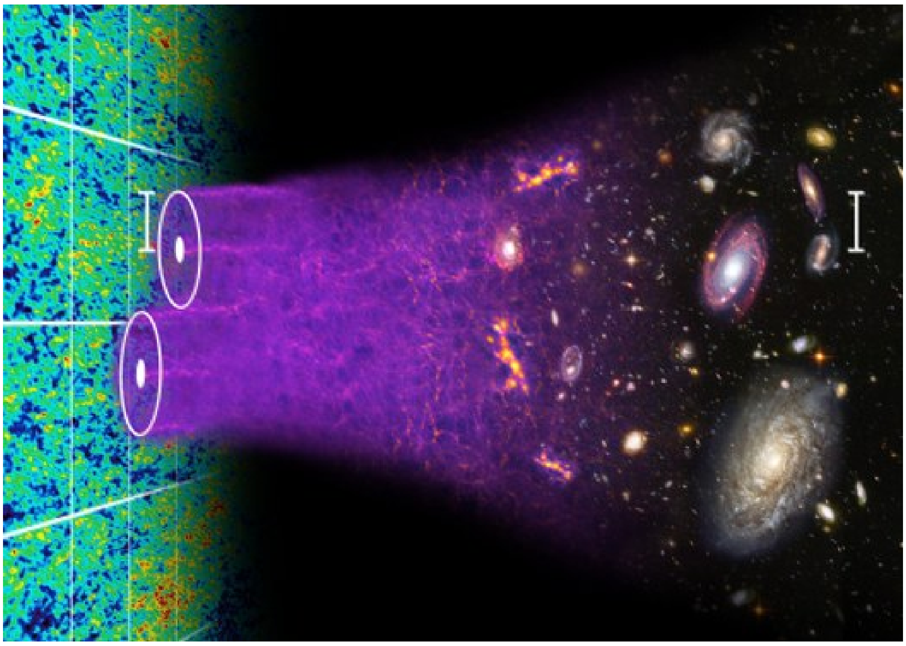


Figure 1.5: Baryon Acoustic Oscillations in the early Universe (white concentric circles) on the Cosmic Microwave Background, which can still be seen today in galaxy surveys such as BOSS [Image courtesy of http://www.sdss3.org/surveys/boss.php]

Baryon Oscillation Spectroscopic Survey (BOSS)

The Baryon Oscillation Spectroscopic Survey (BOSS) (Schlegel et al. 2007), is an ongoing galaxy redshift survey that is part of the SDSS III program (DR8 -DR11) which runs 2009-2014. The telescope makes use of the upgraded SDSS fibre-fed multi-object spectrographs to map the distribution of 1.5 million luminous galaxies over 10,000 square degrees of the sky to spectroscopic redshifts $z < 0.7$ (Dawson et al. 2013). Simultaneously it makes observations of 160,000 high redshift quasars using Lyα forest in the redshift range $2.2 < z < 3$. These spectra will be used to probe the BAO at much higher redshifts. Once completed, BOSS will have mapped out the BAO signature to high precision out to ($z < 0.7$). Measurements provided with BOSS will provide new improved constraints on the cosmological model e.g. Dark Energy, the curvature of space, the angular diameter distance, and the expansion rate H(z) to percentile precision at $z < 0.7$ and z $\approx$ 2.5 (Schlegel et al. 2009).

1.3.1 Future Surveys and Telescopes

eBOSS

eBOSS is the Extended Baryon Spectroscopic Survey will have the largest volume of the Universe surveyed, obtaining spectroscopic redshifts for 1,750,000 LRGs and quasars up to $z = 2.5$ with coverage area of 7500 deg^2. Its primary objectives are to measure the expansion

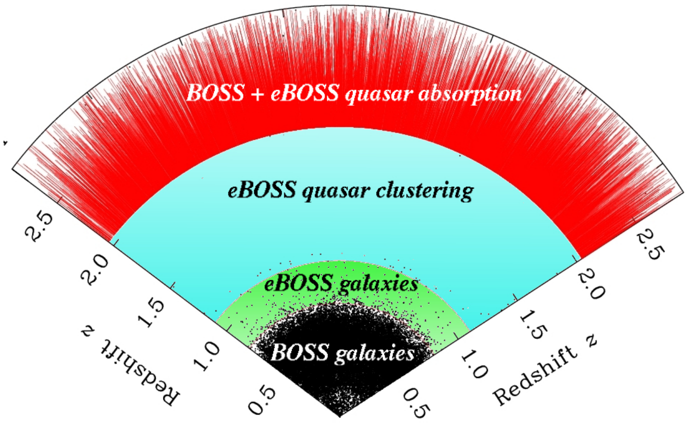


Figure 1.6: The different regions to be probed by the eBOSS survey out to a redshift of $z = 2.5$. From $z < 1$ is the region probed currently by the BOSS survey. [Image courtesy of `www.sdss3.org/future/eboss.php`]

history of the Universe back to when the Universe was less than 3 billion years old, and to improve cosmological constraints, such as the nature of dark energy (Kneib,J.P 2014). Some of the key science goals to be investigated in this survey are,

- At what epoch does the Universe transition from deceleration to acceleration in its expansion and whether the current dark energy theories can predict an equivalent transition at this epoch.
- The growth of structure in this epoch. Are there any deviations from the general theory of gravity, linked to the acceleration at this period?
- To probe the clustering nature of galaxies in the early Universe.

Square Kilometre Array (SKA)

The Square Kilometre Array will be the largest and most sensitive radio telescope in the world with 50 times more sensitivity and 10,000 times faster than currently available radio telescopes. The total collecting area of the SKA will be 1 square kilometre. It will consist of $\sim$ 1000s of dishes. The SKA telescope is shared by two countries 70% of the telescope will be spread across Africa and 30 % will be in Australia. Some of the science goals of the SKA are to,

- Mapping the HI content of the Universe.
- To investigate the formation and evolution of galaxies.
- Testing Einstein's theory of general relativity.
- Probing the nature of dark energy and dark matter.
- Search for Extraterrestrial life.

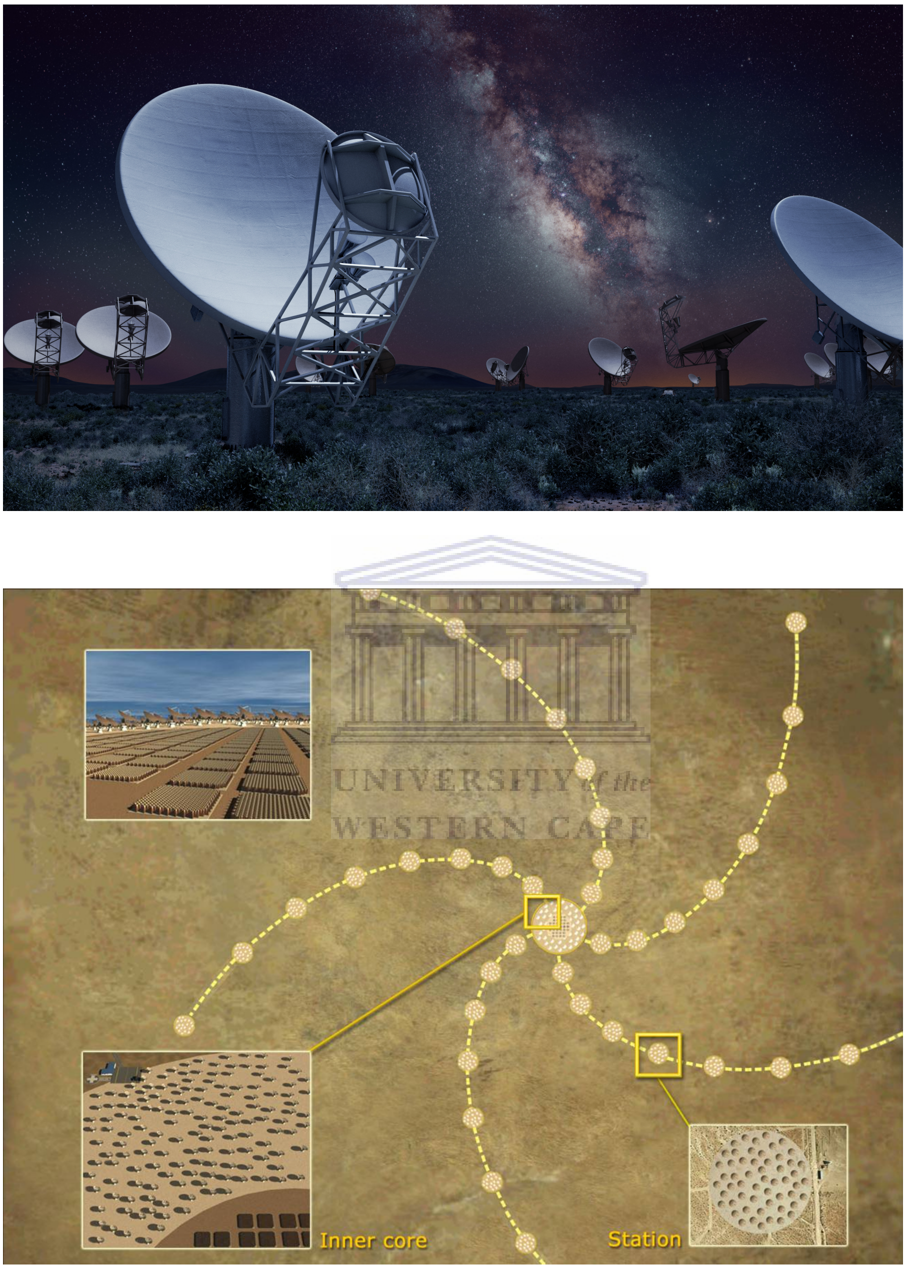


Figure 1.7: The top image show an artist impression of the SKA radio telescope dishes. The bottom image shows an artists impression of the full Square Kilometre Array. [Image courtesy of `http:www.skatelescope.org`]

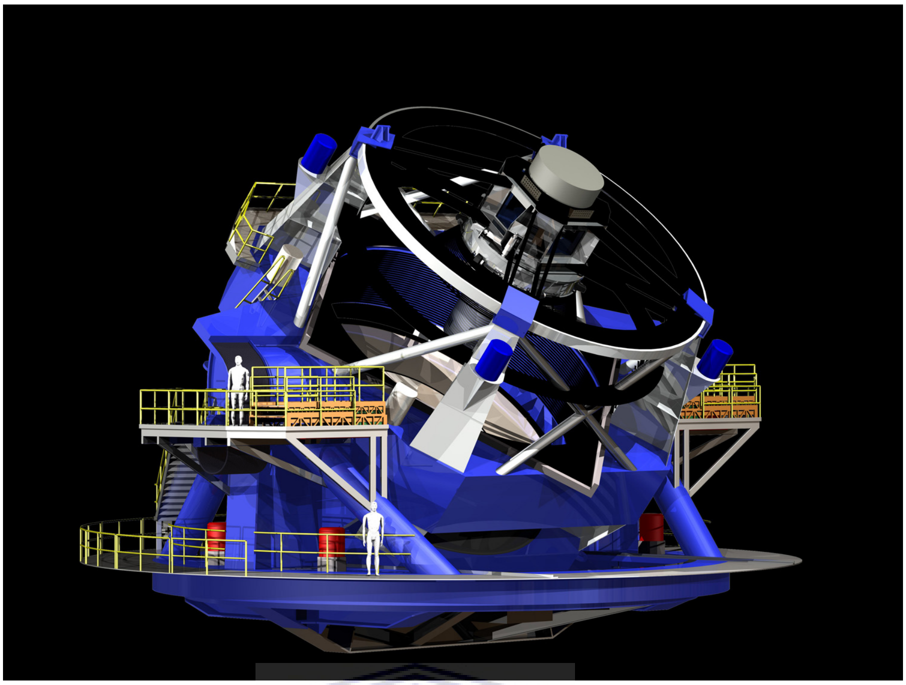


Figure 1.8: An artists impression of the LSST telescope to be operational by 2020. [Image courtesy of `www.lsst.org/lsst/`]

The Large Synoptic Survey Telescope

The Large Synoptic Survey Telescope (LSST) (Ivezic et al. 2008) will be an optical ground based telescope system with an aperture of 8.4 meters. LSST will produce a wide and deep field survey with a total area coverage of 30,000 deg^2 in 6 image bands (*urigzy*) corresponding to a wavelength range of $320 - 1500$ nm. The main science objectives of LSST are to

- probe the nature of dark energy and dark matter,
- exploring the transient optical sky,
- mapping the Milky Way galaxy,
- and making observations of the Solar System.

EUCLID

Euclid is a space based telescope which will be launched in 2020 by the European Space Agency. Euclids primary objective will be to map the geometry of the Universe, probing the deep Universe around $z \sim 2$, this is equivalent to the period when the Universe was roughly 10 billion years. This is the same period in which dark energy played a very important role in the accelerated expansion of the Universe. Euclid will carry out this mission by using two cosmological probes, the BAOs and gravitational weak lensing, to investigate how cosmic

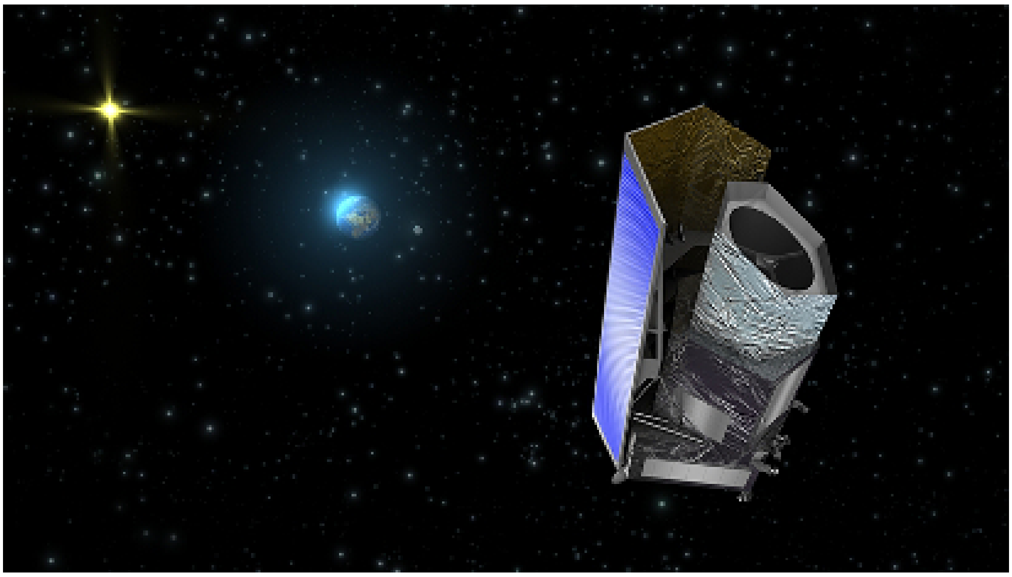


Figure 1.9: Artist impression of the Euclid telescope due to be launched in 2020. Euclid will measure the redshifts, shapes and distribution of galaxies out to $z \sim 2$.[Image courtesy of `http://sci.esa.int/euclid/`]

structures evolve through measuring galaxy redshifts, shapes of galaxies, the distribution of galaxies and measuring the distance-redshift relationship (ESA 2014).

Table 1.1 gives a summary of galaxy redshift surveys from the past, present and next generation surveys.

Table 1.1: Summary of major Galaxy Redshift Surveys

Survey	**Number of Galaxies**	**Coverage**	**Redshift (z)**	**Period**
	Galaxies	**(deg^2)**		
Past				
CfA-1	1,100	17,000	0.05	1977-1982
CfA-2	18,000	17,000	0.05	1985-1995
HDF-North	3,000	2.38	~6.0	1995
IRAS PSCz	15,400	1,500	0.1	1992-1999
Stromlo-APM	1,797	4,300	0.13	1992-2000
2MASS	>1,000,000	37,000	0.02	1997-2001
HDF-South	2,016	2.38	~5.6	1998
2dFGRS	220,000	1,500	~0.3	1998-2003
HUDF	10,000	0.00305	8.0	2003-2004
SDSS I &II -DR1 to 7	> 2,000,000	11,663	$\sim 5.5(z_{photo})$	2003-2008
6dFGRS	125,000	17,000	~ 0.15	2004-2009
WiggleZ	400,000	1,000	0.5 to 1.0	2006-2012
GAMA	~300,000	240	0.5	2008-2013
Present				
DES	$\sim 300,000,000$	5,000	1.3	2012-2014
BOSS (SDSS III, DR8-DR11)	1,500,000	10,000	0.8	2009-2014
Future				
eBOSS	1,750,000	7,500	3.5	2014-2020
JPAS	> 100,000,000	8,000	1.3	2014-2018
SKA	~ 10x10^6	survey dependent	>1.2	2024
EUCLID	~ 50x10^6	survey dependent	~ 2	2020
LSST	~ 2x10^{11}	30,000		2015

1.4 Thesis Summary

The aim of this thesis is to show how correlation functions can be used to understand the clustering nature of galaxies. We present a package that computes correlation functions and for the first time apply this package to the BOSS DR9 and DR10 data releases. In chapter 2 we introduce correlation functions which are commonly used tools to probe distribution of galaxies in the Universe, and review the two and three-point correlation functions and error estimation techniques. We end the chapter by reviewing current of studies which have applied correlation functions to real data samples to probe e.g. the BAO, cosmological parameters and the 3PCF. In chapter 3 we show how we apply the different types of correlation functions practically using the KSTAT package which is designed to compute correlation functions. We also give a brief overview on other available packages that compute correlation functions and lastly perform some benchmark tests to see how the package performs in a high performance computing environment. In chapter 4, we test KSTAT with the BOSS DR9 and DR10 data releases and discuss the results we obtain from the computations. Finally, in chapter 5 we provide a summary and discuss future work.

Chapter 2

Probing the Large Scale Structure of the Universe

One of the most powerful and fundamental tools for probing the large scale structure of the Universe is the 2-Point Correlation Function (2PCF). In essence, this statistic counts the number of pairs of galaxies separated by a distance r in some given galaxy distribution when compared to a random galaxy distribution. Through such a process we can study a variety of cosmological processes such as how the baryonic content of the Universe clusters with dark matter (bias), probing primordial non-Gaussianity, studying Dark Energy through Baryonic Acoustic Oscillations (BAO), redshift space distortions, constraining cosmological parameters, and galaxy evolution.

We begin this chapter by describing the development of correlation functions estimators, followed by a review some of the recent key work to probe the BAO and 3PCF.

2.1 Two-Point Correlation Function (2PCF)

The formation of structure has been studied extensively over the last century. It was in the early 1960s that theories of structure formation emerged. Eggen et al. (1962) established a theory known as the monolithic collapse, which suggested that galaxies were created from large regions of primordial baryonic gas. When the central regions of these baryonic masses gravitationally collapsed, the first stars and most heaviest galaxies were formed. A decade later, James Peebles in his seminal work (see e.g. Peebles 1970; 1973; 1974; 1980b, Gunn & Gott 1972) established an alternative theory known as the gravitational instability theory. This theory proposed that the structure we observe today was caused by tiny density fluctuations that were enlarged by gravitational instabilities in the primordial Universe. Consequently, smaller objects were formed first and merged to form larger objects, this new theory is also known as the Hierarchical clustering model (see e.g. Searle & Zinn 1978).

From Peebles we can define the density contrast as

$$\delta(\vec{x}) = \frac{\rho(\vec{x})}{\bar{\rho}} - 1, \tag{2.1}$$

where ρ is the observed density and $\bar{\rho}$ is the average density of the Universe. If the cosmological principle is assumed (i.e. homogeneity and isotropy), the 2PCF can be expressed as

$$\xi(r_{12}) = \langle \delta(\vec{x_1}) \delta(\vec{x_2}) \rangle, \tag{2.2}$$

where the $\langle ... \rangle$ describes an ensemble average, $r_{12} = |\vec{x_1} - \vec{x_2}|$ is the separation of the two positions. $\xi(r)$ can also be described as the excess probability with respect to a Poisson distribution, of finding two galaxies in volumes δV_1 and δV_2 separated by a distance r,

$$\delta P = N[1 + \xi(r)] \delta V_1 \delta V_2, \tag{2.3}$$

where N is the mean surface density of galaxies. When $\xi(r) = 0$ then the distribution of galaxies is homogeneous, while $\xi(r) < 1$ implies that the distribution of galaxies is less clustered compared to the random distribution, and $\xi(r) > 1$ implies that the distribution of galaxies is more clustered compared to the random distribution. In practice, there are a number of statistical approaches that have emerged over the years for estimating the 2PCF from galaxy surveys.

One of the first was developed by Davis & Peebles (1983), which computed the correlation function using

$$\xi_{DP}(r) = \frac{n_R}{n} \frac{DD(r)}{DR(r)} - 1, \tag{2.4}$$

where $\frac{n_R}{n}$ is the ratio of the mean density in the random catalog to that of the real dataset, DD represents the data-data pair counts and DR is the data-random pair counts.

Hamilton (1993) showed that the estimator given above needed to be improved, since at large scales the uncertainty in the mean galaxy density was very difficult to measure, due to the density fluctuations being small on those scales. A new estimator was proposed which did not have this limitation and was more accurate than the Davis & Peebles (1983). The Hamilton (1993) is defined as

$$\xi_{HAM}(r) = \frac{DD(r) RR(r)}{[DR(r)]^2} - 1, \tag{2.5}$$

where RR has the additional the random-random pair counts.

However, the most popular method currently applied for estimating the 2PCF was developed by Landy & Szalay (1993)(hereafter referred to as LS93)

$$\xi_{LS}(r) = \frac{DD(r) - 2DR(r) + RR(r)}{[RR(r)]^2}, \tag{2.6}$$

which has been found useful to minimize errors due to e.g boundary effects and holes in a galaxy survey. This estimator has been used extensively in cosmology, (see Cress et al. 1996, Blake & Wall 2002, Georgakakis et al. 2000, Tegmark et al. 2004, Eisenstein et al. 2005, Contreras et al. 2013, Anderson et al. 2013, to name a few). It is this estimator that we apply in forthcoming chapters.

A more simpler application of the 2PCF is the angular correlation function $\omega(\theta)$, which is the projected form of the spatial 2PCF. It has been applied widely through the astronomy literature, particularly in radio continuum surveys. In these types of surveys no redshift information can be obtained since there are no absorption or emission lines observed to estimate redshifts from.

In a similar way as the spatial 2PCF, the ACF can be defined as the joint probability δP of finding two galaxies separated by an angular separation θ with respect to that of a random distribution

$$\delta P = N[1 + \omega(\theta)]\delta\Omega_1\delta\Omega_2, \tag{2.7}$$

where θ is the angular separation, $\delta\Omega_1$ and $\delta\Omega_2$ are the elements of a solid angle, N is the mean surface density of galaxies. Thus the Landy & Szalay (1993) estimator can be recast in the angular form as

$$\omega(\theta) = \frac{DD(\theta) - 2DR(\theta) + RR(\theta)}{[RR(\theta)]^2}, \tag{2.8}$$

where θ is the angular separation.

2.2 Three-Point Correlation Function (3PCF)

The discovery of the CMB, has provided important information about the early Universe. It has shown that the Universe was not entirely uniform, that it had small density fluctuations; non-uniformities were vital in the formation of the large scale structure we observe presently. If the distribution of galaxies were gaussian then the 2PCF would be sufficient to accurately describe the clustering of galaxies. However, we can explore the next higher order statistic which can provide insights into deviations from gaussianity.

The 3PCF is defined as the joint probability of there being a galaxy in each of the volume elements δV_1, δV_2 and δV_3 given that these elements are arranged in a configuration

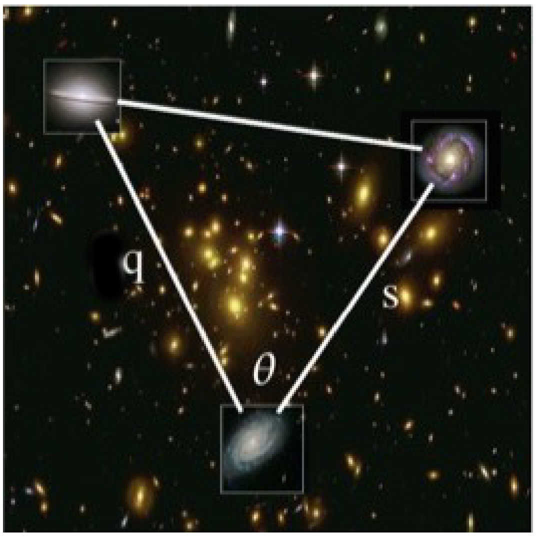


Figure 2.1: The shape parameters (s, q, θ) used to distinctively describe a triangular configuration in order to calculate the reduced 3PCF. [Image courtesy: Sabiu & Nichol (2009)]

defined by three sides of a triangle, r_1, r_2 and r_3. It measures the probability of finding three galaxies separated by a distance r within a given triangular volume when compared to a randomly distributed sample. This joint probability can be expressed as

$$\delta P_{1,2,3} = n^3[1 + \xi(r_1) + \xi(r_2) + \xi(r_3) + \zeta(r_1, r_2, r_3)]\delta V_1 \delta V_2 \delta V_3. \tag{2.9}$$

The above equation is given as a sum of 2PCFs, the order of the correlation function n and the full 3PCF ($\zeta(r_1, r_2, r_3)$). Expressing the joint probability in this manner assumes the Copernican Principle and allows for $\zeta(r_1, r_2, r_3)$ to be a symmetric function of the three lengths (Peebles 1980a) such that

$$\zeta(r_1, r_2, r_3) = Q[\xi(r_1)\xi(r_2) + \xi(r_2)\xi(r_3) + \xi(r_1)\xi(r_3)], \tag{2.10}$$

where Q is a free parameter constrained from observations. To estimate the 3PCF we make use the estimator developed by Szapudi & Szalay (1998)

$$\zeta = \frac{DDD - 3DDR + 3DRR - RRR}{RRR}, \tag{2.11}$$

where each term represents the normalized triplet counts for the data (D), and the random (R) fields that satisfy a particular triangular configuration.

An alternative way to express the 3PCF is through its shape parameters, these can be the sides of the triangle (r_1, r_2, r_3) or expressing the sides as

$$s = r_1 \tag{2.12}$$

$$q = \frac{r_2}{r_1}, \tag{2.13}$$

$$\theta = cos^{-1}(\hat{r_1} \cdot \hat{r_2}), \tag{2.14}$$

where s and q vary from 0° and 180° and $\hat{r_1} \cdot \hat{r_2}$ is the unit vector of the two sides of the triangle as illustrated in Figure 2.1.

2.3 Construction of Random Catalog

When estimating both the 2PCF and the 3PCF, one requires to construct a random catalog to compare against the observed data. To ensure that correlation functions are estimated correctly there are a number of aspects to be taken into consideration. Firstly, one must ensure that when constructing the random catalog, the survey area of the random catalog must match that of the observed data. Secondly, it is crucial to match as closely as possible the survey selection function when constructing the random catalog. In practice, the number of objects in the random catalog is chosen to be several times greater than the number of objects in the data, to reduce poisson errors (shot noise). Other effects that must be accurately matched in the random catalog are the overlapping of fibre plates or slit masks, edge effects and gaps in the data. Figure 2.2 illustrates the SDSS DR3 galaxy field and its corresponding random sample below it.

2.4 Error Estimation

2.4.1 Jackknife Resampling

Jackknife resampling (Scranton et al. 2002) is an error estimation method that involves partitioning of a survey area into N equal areas or volumes as illustrated in panel (a) of Figure 2.1. One creates N jackknife subsamples by successively removing and replacing each partition from the whole sample. The first 3 jacknife subsamples are illustrated by panels (b)-(d).

One then applies the correlation function estimator to each jackknife sample. The variance and the mean are computed by calculating the correlation function N times, giving a set of N values for the correlation function $\{\xi_i, i = 1..., N\}$ as error estimates. The variance is computed using this equation by Lupton (1993),

$$\sigma^2_\xi(r_i) = \frac{N_{jk} - 1}{N_{jk}} \sum_{j=1}^{N_{jk}} [\xi_j(r_i) - \bar{\xi}(r_i)]^2, \tag{2.15}$$

where N_{jk} is the number of Jacknife samples, ξ is the 2PCF or 3PCF and r_i represents the jacknife bin, the error is being calculated for.

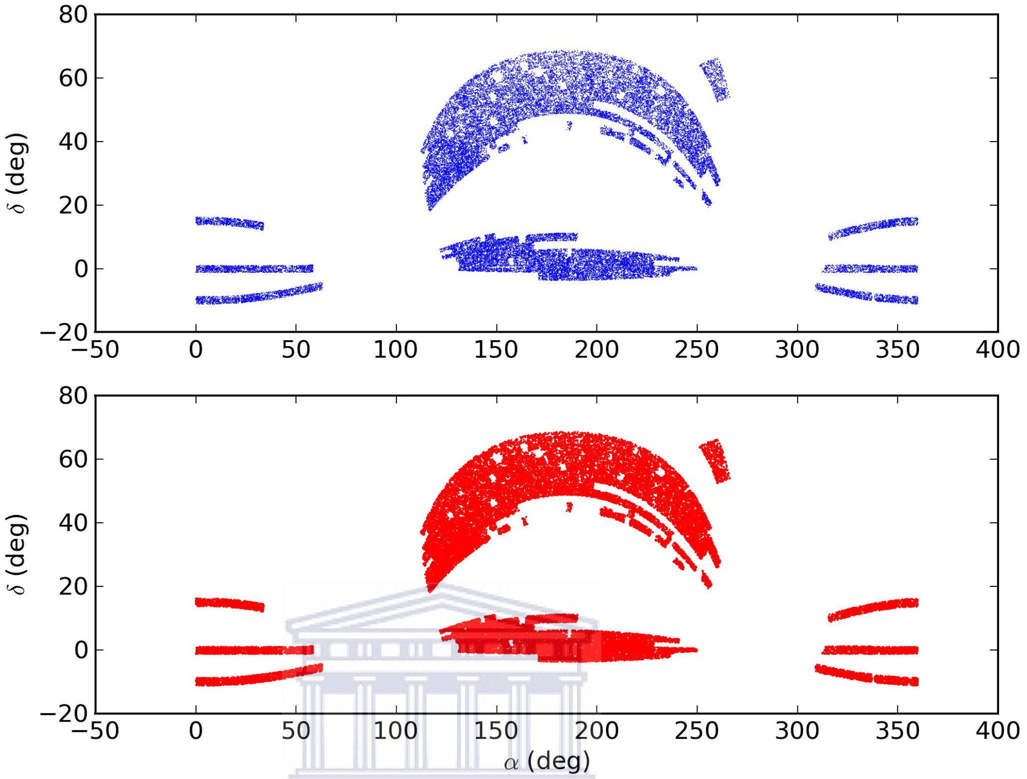


Figure 2.2: Coverage map of the SDSS DR3 galaxy field in blue (top). Below it, is the random field (red) field which has been constructed to match the survey area dimensions, selection effects and gaps in the data have been considered in the construction of the random field. The random field contains more data points than the galaxy field.

2.4.2 Bootstrapping

An equally popular approach to error estimation is the bootstrapping technique developed by Efron (1979), and first applied in correlation analyses by Ling et al. (1986) and Fisher et al. (1994). In essence, the bootstrapping error estimation is based on randomly resampling data points from the parent dataset to create new bootstrapped samples with the same number of data points as the original sample. Since we are randomly drawing from the sample, a galaxy located at (i, j) can be selected more than once. It can also happen that some are not selected to be in the bootstrapped samples. In a similar way to Jackknifing, this process is repeated to produce a set of N bootstrapped samples from which the correlation function can be calculated. Using this approach, for each spatial (r) or angular (θ) bin, a uniform set of correlation function estimates is produced. As with jackknifing we can make use of equation 2.13 to estimate the uncertainty in the measurement of the correlation function.

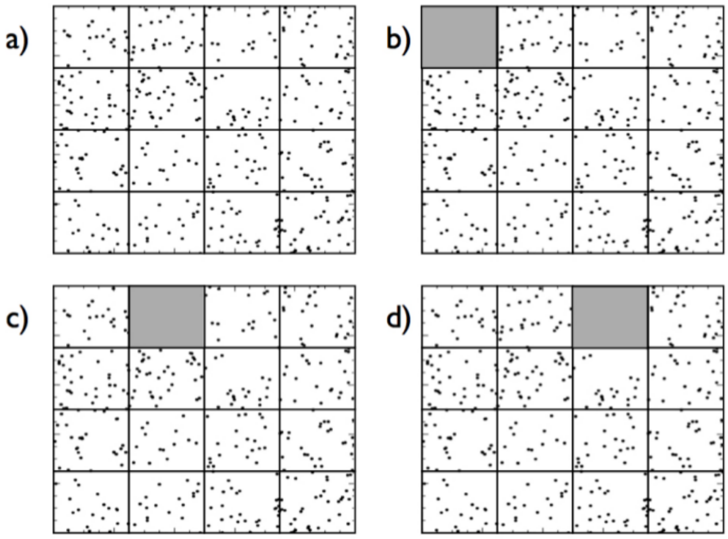


Figure 2.3: The jacknifing error estimation process. First 3 jacknife subsamples shown in panels (b)-(d) [Image courtesy of Sabiu & Nichol (2009)]

2.5 Probing the BAO with the spatial 2PCF

Whilst landmark redshift surveys such the 2dFGRS (Colless 1999) provided accurate constraints on the matter power spectrum via correlation function (see Figure 2.4) measurements, it would take the SDSS survey to probe the BAO for the first time. In Eisenstein et al. (2005) using the SDSS DR3 Luminous Red Galaxy (LRG) sample, they made a landmark first detection of the BAO at 3.4σ significance. The number of objects in the sample was 46 748, over an effective survey area of 3816 deg^2. This sample was able to probe the BAO over a redshift range from $0.16 < z < 0.4$. In this work, they applied the LS93 estimator using a random sample with 16 times number of objects in the LRG sample. The radial and angular selection functions were modelled using methods described in Zehavi et al. (2004).

The BAO peak was detected at a 100 h^{-1} Mpc separation and was found to compare well with standard CDM models as shown in Figure 2.5. The figure shows the redshift space 2PCF measurements (black square points). Also shown in the figure are various CDM models were with varying matter densities. The top green line is a model in which the matter density $\Omega_m h^2 = 0.12$, the red line $\Omega_m h^2 = 0.13$, the blue $\Omega_m h^2 = 0.14$. In the above models the baryon density and spectral tilt was set to $\Omega_b = 0.024$, $n_s = 0.98$ respectively. The magenta line is a pure CDM model with $\Omega_m h^2 = 0.105$, in this case, the BAO peak is not observed. Even with this relatively small sample of galaxies, we can clearly see the now familiar BAO ‘bump’.

The detection of the BAO peak allowed cosmologists to use it as a standard ruler for con-

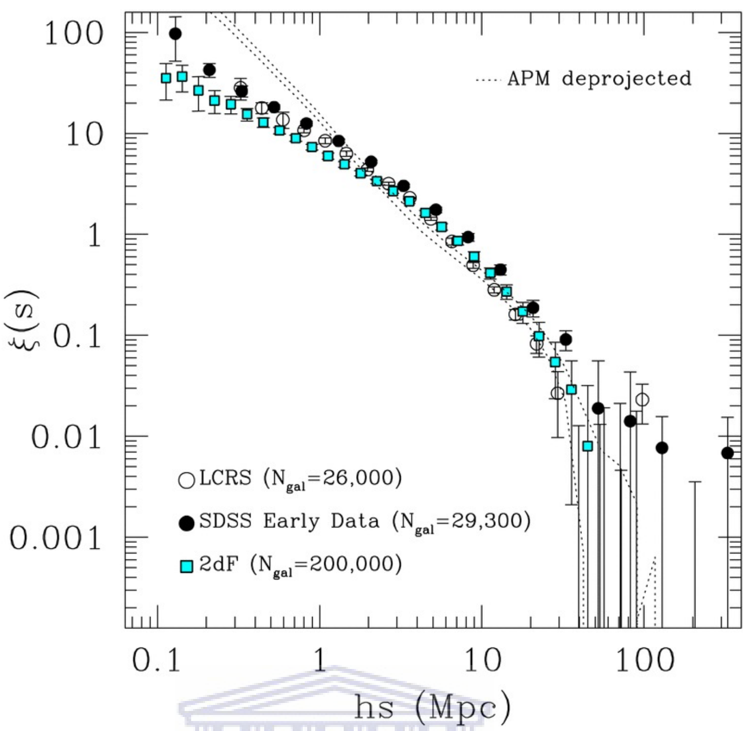


Figure 2.4: The 2PCF measurement of 220,000 galaxies in the 2dfGRS survey.[Image courtesy of `http:www2.aao.gov.au/2dfgrs/`]

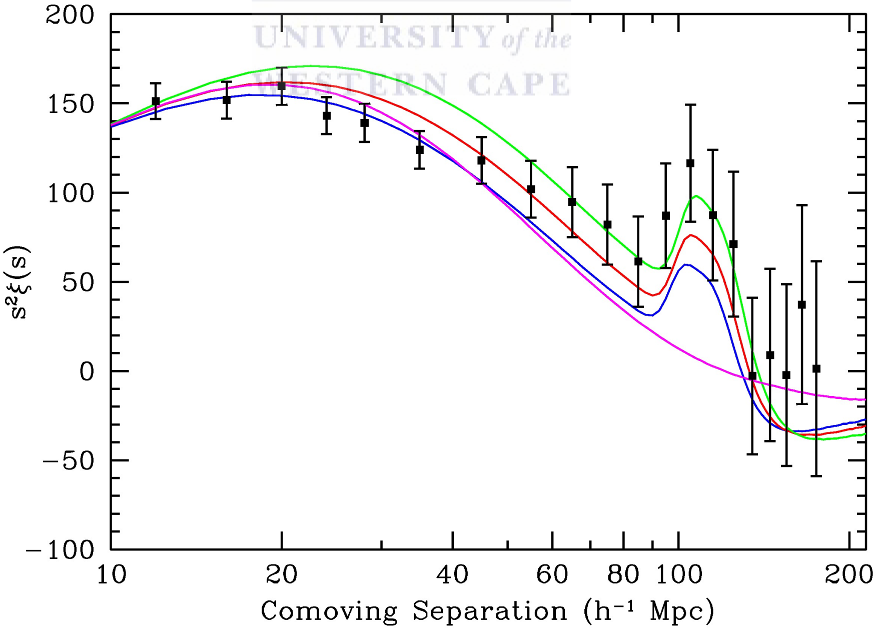


Figure 2.5: The first detection of the BAO peak from LRG sample in the SDSS DR3 data release. The coloured lines are standard CDM models with the matter density allowed to vary. The BAO peak observed at 100 h^{-1} Mpc separation (Eisenstein et al. 2005).

straining cosmological parameters. In this study the absolute distance d_A at $z = 0.35$ was measured to an accuracy of 5%. Using the shape of the measured correlation function, they were able to measure the matter density to be $\Omega_m h^2 = 0.135 \pm 0.008$ and the spatial curvature of the Universe $\Omega_k = -0.01 \pm 0.009$, provided that dark energy is a cosmological constant. With this discovery, it was not long before a more targeted survey was devised, specifically designed to provide the most precise constraints to date on the BAO.

The Baryon Oscillation Spectroscopic Survey (BOSS) (Schlegel et al. 2009) is an extension of the SDSS III project primarily designed to probe the BAO by targeting luminous galaxies at the low redshifts and Lyman-α forest spectra from quasars at higher redshifts ($z > 2.5$). BOSS currently has 3 data releases publicly available (DR8, DR9, DR10). The luminous galaxy sample are split into two subsamples, LOWZ and CMASS (a higher redshift sample). The 'LOWZ' subsample consists of galaxies with low redshifts $z < 0.40$ and which have colours similar to LRGs but with lower luminosities. The 'constant mass' (CMASS) subsample, consists of luminous galaxies whose stellar mass is approximately constant over the redshift range $0.4 < z < 0.7$. These subsamples have been utilized widely in the clustering community to probe the BAO and to constrain cosmological parameters in the respective data releases. Here we review work carried out on the latest two data releases, DR9 and DR10.

Work by Sánchez et al. (2012) utilized the spherically averaged redshift space 2PCF on the DR9 CMASS sample (see Figure 2.6). To place strong constraints on cosmological parameters, these measurements were combined with CMB measurements from the WMAP satellite (Hinshaw et al. 2009) and South Pole Telescope (Keisler et al. 2011), together with Type Ia Supernovae (Conley et al. 2011) measurements. The sample used has 262,104 galaxies in the redshift range of $0.43 < z < 0.7$. The coverage area of the sample is placed the left panel of Figure 2.7. Sectors in the survey area with 75 % completeness were selected. Their results showed no deviation from the standard ΛCDM model, and finding constraints on; $\Omega_m = 0.285 \pm 0.009$, $H_0 = 69.4 \pm 0.8$ km s^{-1} Mpc^{-1}, $100\Omega_b = 4.59 \pm 0.09$, $n_s = 0.961 \pm 0.009$ and $\sigma_8 = 0.80 \pm 0.02$. Measurements of the CMB from WMAP and South Pole Telescope, combined with CMASS, were also able to constrain the curvature of the Universe to, $\Lambda_k = -0.0043 \pm 0.0049$, and the dark energy equation of state to, $w_{DE} = -1.033 \pm 0.073$.

In Anderson et al. (2012), clustering measurements were made using the DR9 release, and a more recent analysis in Anderson et al. (2013) utilizes the most recent catalogs from DR10 and 11. These measurements were carried out using the angle-averaged 2PCF, and its Fourier transform counterpart, the power spectrum. The power spectrum carries the same information as the 2PCF, but for our purposes we present only the 2PCF. They also followed a linear density reconstruction technique by Eisenstein et al. (2007) to reconstruct the DR9 CMASS density field in order to see if they could recover the BAO feature. The algorithm for this technique can be found in Padmanabhan et al. (2012) and Anderson et al.

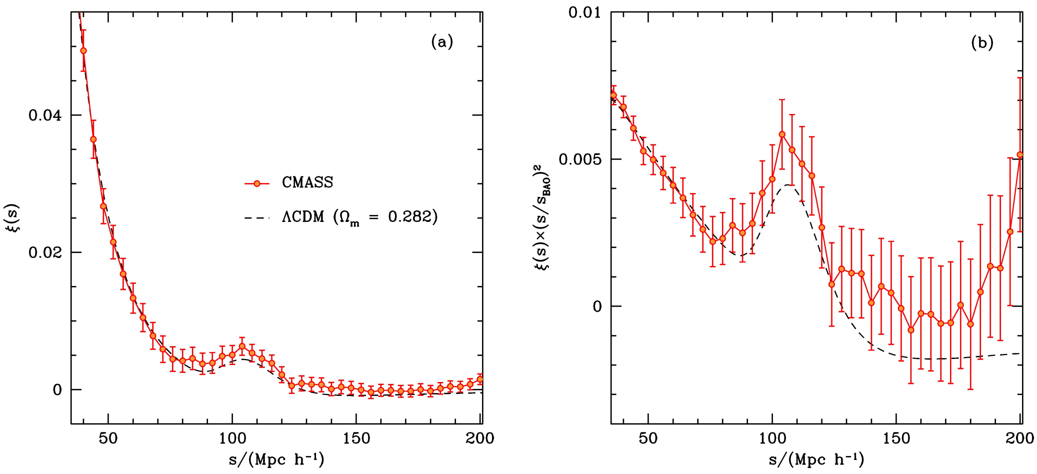


Figure 2.6: The spherically averaged spatial 2PCF of the DR9 CMASS sample by Sánchez et al. (2012). The error bars were obtained from 600 Monte Carlo mock catalogs. From the shape of the correlation function combined with CMB measurements, a best fit ΛCDM model was obtained, represented on the plot by the dashed line. The BAO peak was observed at 107 h^{-1} Mpc. On the right panel is another representation of the redshift space 2PCF, with the y-axis rescaled by $(s = s/s_{BAO})$ where $s_{BAO} = 153.2$ Mpc, to show the BAO peak.

(2012). Figure 2.7 shows the successive increase in objects and coverage of DR9, DR10, DR11 respectively. The effective sky coverage for these samples are 3,275 deg^2, 6,161 deg^2, 8,377 deg^2 respectively. The different colours represent the completeness of a sector, shown in the bottom right panel. Figure 2.7, is the measured BAO from in the DR9 data which was detected at 5σ. On the right panel, is the 2PCF measured from the reconstructed the density field. The BAO is clearly detected using this method and it well within the errors as observed in the χ^2 best fit values.

Recent work by Blake et al. (2011) used the WiggleZ Dark Energy survey to probe the redshift-distance relation by measuring BAO. The WiggleZ Dark Energy survey is a galaxy redshift survey that targeted bright emission line galaxies in the redshift range of $0.2 < z < 1.0$. The survey used 158,741 galaxies over an area of 800 deg^2 around the equatorial region of the sky. The BAO peak was measured at different redshifts $z = 0.44$, $z = 0.6$ and $z = 0.73$. These measurements were then combined with those from the 6-degree Field Galaxy Survey (6DFGS) and Sloan Digital Sky Survey (SDSS) at lower redshifts to produce a stacked survey correlation function that has a statistical significance of 4.9σ, when compared to a zero-baryon mode model with no BAO peak. Figure 2.9 shows the resulting 2PCFs from this work. The lower right hand panel is the combined 2PCF. Cosmological models were fitted to the combined BAO dataset, composed of 6 distance-redshift data points. The results of the fitting were then compared to CMB and Type Ia supernovae (SNe) data (Amanullah et al. 2010, Komatsu et al. 2009). Assuming a flat Universe, they found the equation of state for dark energy to be $w_{DE} = -1.03 \pm 0.08$. Moreover, by

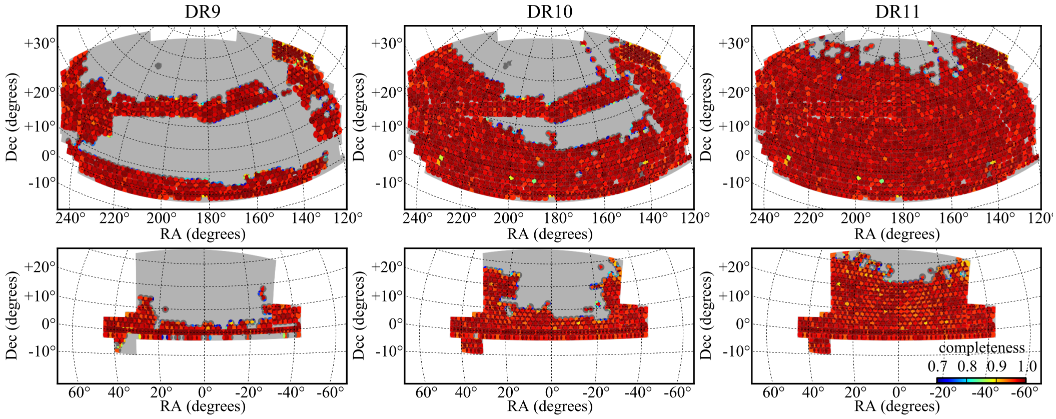


Figure 2.7: Evolution of the BOSS data releases, DR9, DR10 and DR11 (not publicly available yet). The top 3 panels are observations from the North Galactic Cap (NGC) and the bottom 3, are from the South Galactic Cap (SGC) (Sánchez et al. 2012).

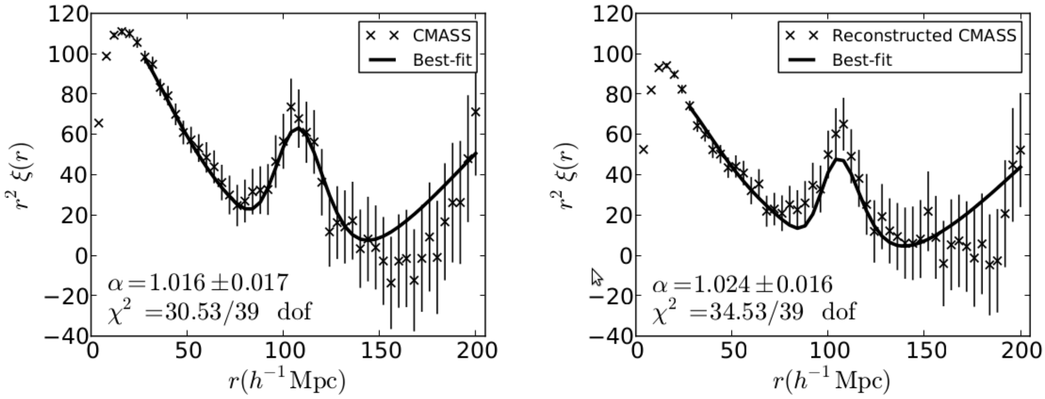


Figure 2.8: The 2PCF measurements in the DR9 CMASS sample (left panel). Along side is 2PCF measured from the reconstructed CMASS density field (Anderson et al. 2013).

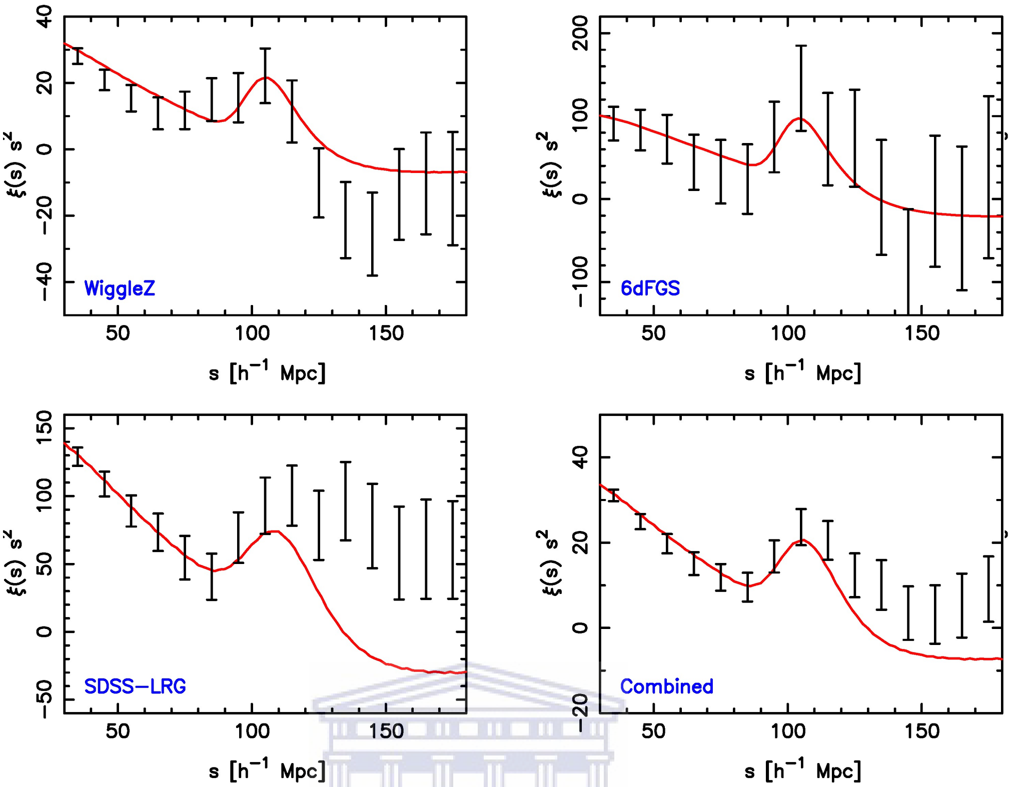


Figure 2.9: Spatial 2PCF measurements from the WiggleZ, 6dFGS and SDSS DR7-Full LRG sample. The lower right hand panel shows the stacked 2PCF measurement (Blake et al. 2011).

assuming that dark energy is a cosmological constant, they found a spatial curvature of the Universe measured is $\Omega_k = -0.004 \pm 0.006$, which was fully consistent with the standard ΛCDM model.

2.6 Probing the higher order correlations with the 3PCF

One of the earliest measurements on the 3PCF were carried out by Groth & Peebles (1977) based on angular catalogs. In this study, the angular form of the 3PCF, analogous to the spatial form in Equation 2.10 was used.

$$\mathcal{L} = \frac{Z(\theta_1, \theta_2, \theta_3)}{[\omega(\theta_1)\omega(\theta_2) + \omega(\theta_2)\omega(\theta_3) + \omega(\theta_3)\omega(\theta_1)]}, \tag{2.16}$$

where $Z(\theta_1, \theta_2, \theta_3)$ is the normalized angular 3PCF and ω denotes the angular form for the correlation measurement. In the denominator we can observe that the angular 3PCF is expressed as a sum of angular 2PCFs. Since $\mathcal{L}$ is a symmetric function, a change of variables can be introduced, where $\theta = \theta_1$, $u = \frac{\theta_2}{\theta_1}$, $1 \leqslant u$, $v = \frac{(\theta_3 - \theta_2)}{\theta_1}$, $0 \leqslant v \leqslant 1$. Thus, θ is a parameter determining size, u is elongation of the triangle, and v is the opening angle of the triangle configuration, which ranges from 0 for an isosceles triangle (with two angles $\geqslant 60°$)

to 1 for a straight line (Peebles & Groth 1975). The angular 3PCF was measured to an average value of $\mathcal{L} = 1.56 \pm 0.22$. They went further by estimating the reduced 3PCF (Q), and found $Q = 1.3 \pm 0.21$ for $r \lesssim 3h^{-1}$Mpc. Figure 2.10, shows their normalized 3PCF as a function of θ. The estimation of the 3PCF was limited by the size of the samples, hence the large error bars observed.

With the completion of larger galaxy surveys (e.g.2df, SDSS), it became possible to make more precise measurements of the 3PCF and in later years, has led debates as to whether there is a luminosity dependence on the value of Q, (e.g. Jing & Boerner 1998, Jing & Börner 2004, Kayo et al. 2004).

Kayo et al. (2004) conducted the first detailed study on the morphology, colour and luminosity dependence of the 3PCF in the SDSS "sample12" data release (Blanton et al. 2001), a data release between DR1 and DR2, which was 1.8 times larger than DR1. To control systematics due to the selection function within this sample, an r-band volume selected sample was constructed within the SDSS survey magnitude limits (14.5<r<17.5). The sample was then partitioned into morphology type (early and late). Their findings showed that the 3PCF aligns well with the hierarchical clustering relation in equation 2.10 for equilateral triangle configurations in the range $1h^{-1}$Mpc< $s < 10h^{-1}$Mpc, however, the reduced 3PCF has a shape and dependence, particularly on large scales as depicted in Figure 2.11. They measured the reduced 3PCF and obtained an almost scale independent value of $Q = 0.5 \sim 1.0$ and from their analysis, they find no statistically significant dependence on morphology, colour and luminosity of the 3PCF.

Interpretation of the spatial 3PCF is made difficult by the presence of redshift space distortions caused by peculiar motions of galaxies arising from the gravitational potentials which they reside in. One of the ways to evade this problem is to measure the angular 3pcf in which the redshift distortions are phased out by integrating over the redshift dimension (Zheng 2004), in similar approach as in the angular 2pcf.

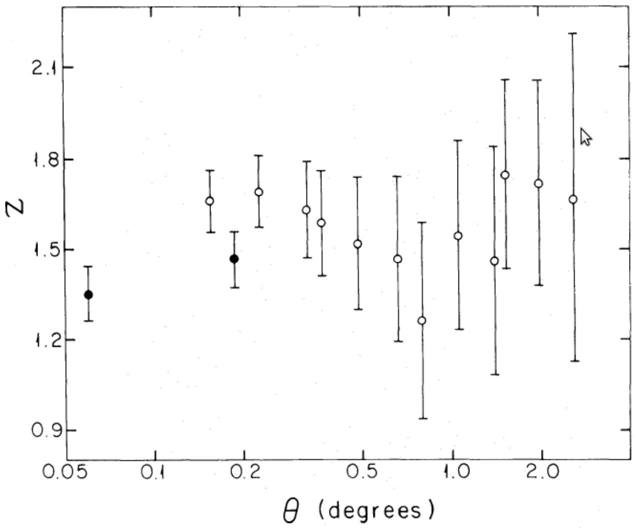


Figure 2.10: The first 3PCF measurements by Groth & Peebles (1977) on angular catalogs. These measurements were limited by small galaxy samples.

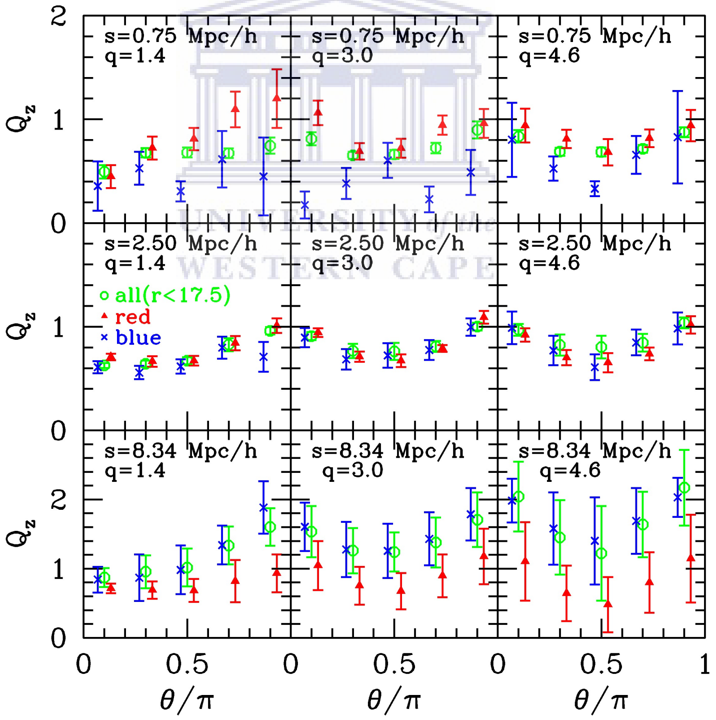


Figure 2.11: The reduced 3PCF (Q) measurements by Kayo et al. (2004) from volume limited SDSS sample 12 data release. These galaxies are divided according to galaxy population(red and blue). A very weak shape dependence on scales $s < 1h^{-1}$Mpc, shape dependence becomes stronger on large scales as observed in the third row of Q plots.

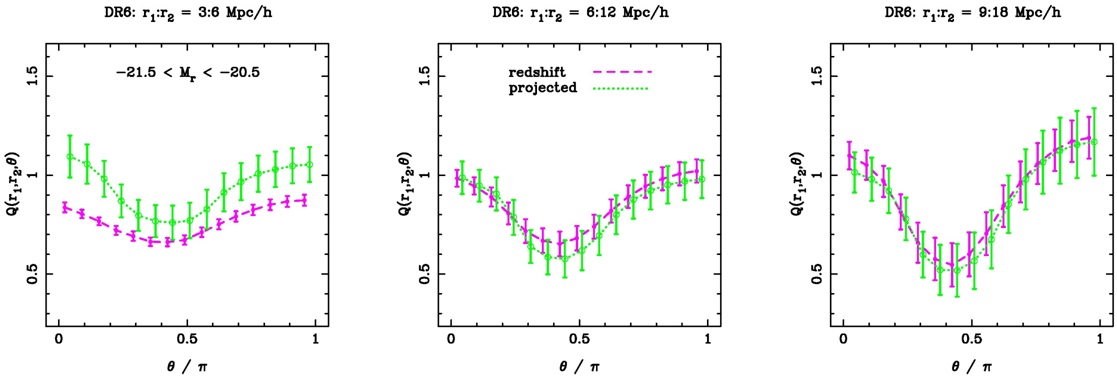


Figure 2.12: An example of the reduced 3PCF measurements of the SDSS DR6 LRG sample.The points in purple represent the spatial (redshift) 3PCF, the green points represent the projected 2PCF.McBride et al. (2011b)

McBride et al. (2011a) were the first people to use the largest sample of galaxies to investigate the shape dependence of the 3PCF in redshift and projected space. They conducted clustering measurements on 220,000 galaxies within 3 volume limited samples in the SDSS main galaxy sample (York et al. 2000). Their findings showed that there is a significant shape dependence of the reduced 3PCF on scales between $3 - 27h^{-1}$Mpc, which is consistent with standard ΛCDM model predictions. This finding disagrees with the hierarchical clustering relation. They suggest that the reason for the reduced 3PCF to appear consistent with the hierarchical ansatz is due to redshift distortions and not galaxy bias. Redshift distortions on scales less than 6 h^{-1}Mpc cause the reduced 3PCF to have a weak shape dependence and smaller amplitude when compared to the projected 3PCF. Figure 2.12 shows the reduced 3PCF in redshift and projected space. One can observe the shape dependence of the reduced 3PCF as the triangle configurations change. Lastly, their analysis demonstrates that measuring the 3PCF is sensitive systematic effects, such as the binning scheme selected, the sky completeness, difficulty in calculating errors and prior assumptions that do not account for the effects consistently.

As mentioned in Section 2.2, the 3PCF is a higher order statistic of the 2PCF. Can it be used to detect the BAO peak in the 3PCF? The first detection of the BAO peak in the 3PCF has been claimed by Gaztañaga et al. (2009) using volume limited samples in SDSS DR6 & DR7 data releases. These LRGs were selected by magnitude cuts in the range $-22.5 < M_r < -21.5$ within a redshift range $0.15 < z < 0.8$, giving 40 000 galaxies with a number density of $\tilde{n} \simeq 4\text{x}10^{-5}$. Figure 2.13 shows the results from this work using the reduced 3PCF from $55 - 125$ h^{-1} Mpc, the BAO peak was detected at 105 h^{-1} Mpc with signal to noise ratio, S/N> 6. The triangular configuration used for this measurement was r1 = 88 and r2 = 33 Mpc. In order to verify the significance of this detection, they made use of the MICE simulation by Fosalba et al. (2013), which is a simulation of structure forma-

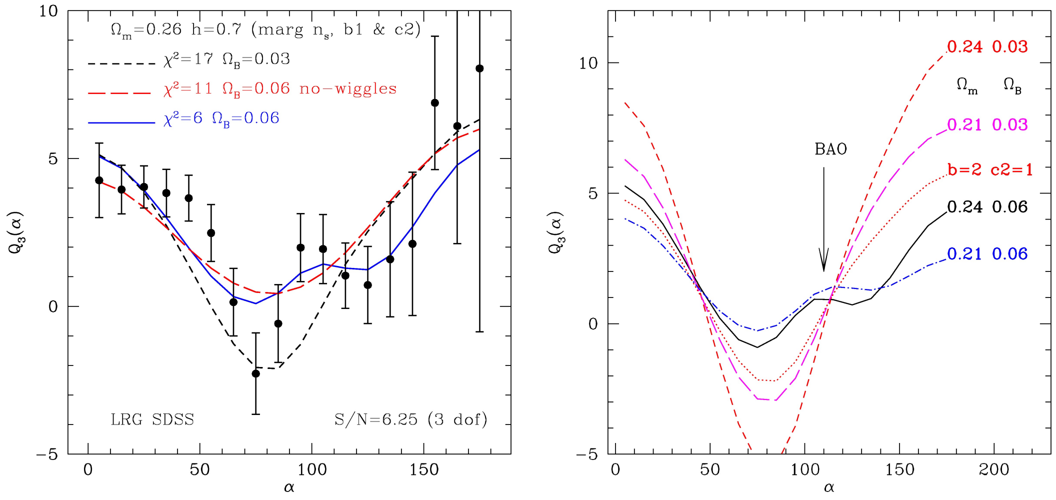


Figure 2.13: On the left panel are measurements of the reduced 3PCF Q (black points), 3 different models with varying Ω_b and Ω_m. The line in blue represents the best fit model to the data as it as the lowest χ^2 value. On the right panel are predictions from perturbation theory with different values of Ω_b and Ω_m. All models expect for the red dotted line show no bias. The BAO peak is detected at $\alpha = 115°$ for high values of Ω_b

tion in the Universe in the linear to the non-linear regime of large scale structure clustering. The simulation makes use of 2048^3 dark matter particles with a cube of side 7680 h^{-1}Mpc. The cosmological parameters used to construct this simulation are $\Omega_m = 0.25$, $\Omega_b = 0.044$, $\sigma_8 = 0.8$, $n_s = 0.95$ and $h = 0.7$. The simulation also computes the Power spectrum by Eisenstein & Hu (1999). Figure 2.13, the left panel shows the measured Q, as a black points. The blue line represents the best fit model with $\Omega_m = 0.26$ and $\Omega_b = 0.06$, the BAO peak was detected at $\alpha = 115°$. The right panel shows predictions from perturbation theory models, for high values of Ω_b, the BAO peak is well detected.

The power of the correlation function which connects the structure we observe today to the physical processes in the early universe can clearly be observed. In the next chapter we take a closer look a new package we are co-developing to compute clustering statistics in a more optimized way for future galaxy surveys.

Chapter 3

Applying the 2 and 3-point correlation function

In this chapter we explore various methods for computing correlation functions. We begin with an $\mathcal{N}^2$ approach to illustrate how the 2PCF is computed. From there, we explore more advanced methods that use kd-tree algorithms, and demonstrate how they can be implemented within parallel processes (MPI). In particular we examine KSTAT [1] and conduct performance tests of KSTAT within a high performance computing environment.

3.1 The $\mathcal{N}^2$ approach

In this section we demonstrate one of the basic ways to calculate the angular 2PCF (ACF). The ACF is a fundamental statistic that is used for describing the galaxy distribution. Its strength lies in the ability to describe the clustering of galaxies by only making use of the positions of galaxies projected onto the plane of the sky (i.e right ascension, declination). As we discussed in the previous chapter, the application of the ACF has been of particular importance for radio continuum galaxy surveys where no redshift information is implicitly available. The earliest attempts recorded of clustering measurements on radio surveys were conducted in the 1970s, since then there has been a number of studies carried out on clustering measurements using e.g. the Faint Images of the Radio Sky at Twenty Centimetres (FIRST) Survey, (Becker et al. 1995, Cress et al. 1996), the Westerbork Northern Sky Survey (WENSS), Rengelink et al. (1997), and the NRAO VLA Sky Survey (NVSS) (Condon et al. 1998, Blake & Wall 2002).

To illustrate how the ACF is computed we select a sample of galaxies from the FIRST survey, similar to the sample described in Passmoor et al. (2013). This subsample was selected in such a way as to avoid gaps and edges of the survey, which provides us with a

[1] https://bitbucket.org/csabiu/kstat

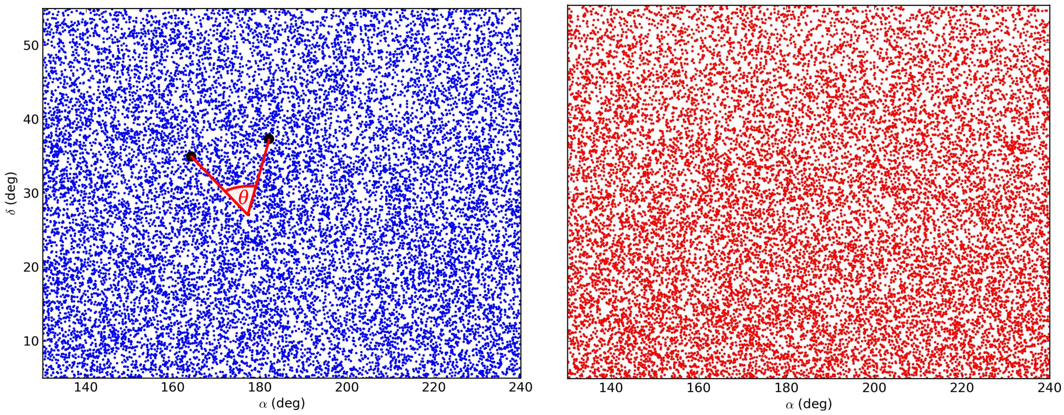


Figure 3.1: On the left is a subsample of galaxies in the FIRST sample (red points). On the right is a randomly drawn distribution of random galaxies (black points).

relatively clean sample to use for this demonstration. The angular coverage of the sample is 130° $< ra <$ 240° and 5.0° $< dec <$ 55.0° and a flux cut was made at 1 mJ giving a total of 17360 galaxies in our test sample. We now need to construct a random catalog to compare against our data. The random catalog is constructed in such a way that it covers the same area covered by the FIRST galaxy sample. Typically the random sample is selected to contain several times more objects than in the real data, in order to reduce effects of shot noise at smaller angular scales, and must have also have the same selection effects and defects (holes and strips) as the observed data. For simplicity the random catalog is homogeneous as shown in the right panel of Figure 3.1. The figure shows the projected positions of our FIRST galaxy sample on the plane of the sky (left panel), and our constructed random sample (right panel). In this example, the number of galaxies in both fields is equivalent.

The left panel of Figure 3.1 also illustrates how we compute the ACF. For each pair of galaxies located at (α_1,δ_1) and (α_2,δ_2) we determine the angular separation θ using the following relation

$$\theta = \frac{180}{\pi}\tan^{-1}\left(\frac{\sqrt{\cos^2(\delta_2)\sin^2(\alpha_2-\alpha_1)+[\cos(\delta_1)\sin(\delta_2)-\sin(\delta_1)\cos(\delta_2)\cos(\alpha_2-\alpha_1)]^2}}{\sin(\delta_1)\sin(\delta_2)+\cos(\delta_1)\cos(\delta_2)\cos(\alpha_2-\alpha_1)}\right). \tag{3.1}$$

Once the angular separation of each and every pair of galaxies has been computed, we use one of the dimensions of the field (the right ascension) and the chosen number of bins to compute the binwidth. The ACF is then computed for each angular bin using the Landy & Szalay (1993) estimator as defined in Equation 2.8 in Chapter 2.

Table 3.1: The ACF computation times on different platforms.

Code	**Number of Processors**	**Runtime (mins)**
$\mathcal{N}^2$ method	1	00:35:00
Tree Code	1	00:02:20
Tree Code	2	00:01:05

On the schematic below is a pseudocode to illustrate how the ACF is computed within a basic $\mathcal{N}^2$ computational framework.

```
***O(NxN) calculations for DR pair counts
***For each object, i, we need to loop over
THE WHOLE catalogue: j index end loop j
for i in range(Ngal):
  for j in range(Ngal):
       **Calculating Angular Seperation
       AngSep=ThetaFunc(FirstRA[i],RandomRA[j],FirstDec[i],RandomDec[j])
       if(DRAngSep >mintheta and DRAngSep<=maxtheta): ***Allocate bin
          binnum = int((log10(DRAngSep/mintheta))/binwidth)
          DRArray[binnum]+=1 ***Add to DR pair count bin
end loop i
*** NEED TO REPEAT ABOVE FOR DD and RR COUNTS
THEN...
for i in angular bin:
       ***Landy and Szalay Estimator
          W[i]=(DDArray[i]-2*DRArray[i]+RRArray[i])/RRArray[i]
```

We applied this approach to our sample data under very simple conditions. The number of objects in both samples was 17,360. We chose 8 angular bins and probed angular scales of $1 - 10^{\circ}$. Table 3.1 details the time taken to compute the angular 2PCF using this basic $\mathcal{N}^2$ approach as written by the author.

From this simple demonstration one can observe that the basic approach used to estimate the ACF would be very computationally expensive for very large samples as the computation times scale with the number of galaxies $O(\mathcal{N}^2)$, where $\mathcal{N}$ is the number of galaxies. Whilst such an approach may be adequate for small samples, as we move into an era where future surveys e.g. SKA, LSST, EUCLID, which will obtain positions of billions of galaxies, we require more sophisticated algorithms to perform these statistics efficiently.

3.2 Optimized Algorithms

3.2.1 The kd-Tree Algorithm: An $O(\mathcal{N}\log\mathcal{N})$ approach

A popular way in which pair counting in the 2PCF has been optimized is through the application of the kd-tree algorithm. kd-Trees were introduced by Friedman et al. (1977), and has since been applied across diverse industries such as, DNA sequencing, pattern recognition, and game development to name a few. kd-Trees are a way of organizing a set of data in k-dimensional space such that, once built, any query requesting a list of points in a neighborhood can be answered quickly without the need for searching through every single point (Moore et al. 2001). The tree is constructed in a "top down" scenario is illustrated in Figure 3.2, where a bounding box is created containing all the data points. This is referred to as the root node. The next step is dividing the data equally along the widest dimension of the box, which creates two child nodes that have distinct data points in each. This process is then repeated by recursively splitting each child node to even smaller scales.

The splitting is halted when the two following criteria are met. A bounding box (node) is considered as a leaf node if the width of its widest dimension in the bounding box is less than a minimum box (MinBoxWidth) width allowed. Secondly this node is left unsplit if the number of data points within are less than some minimum threshold, r_{min}. This type of node is called a leaf node, containing a list of galaxy positions. Moore et al. (2001) suggests that in practice r_{min} should be around 10 and the MinBoxWidth should be set to 1% of the range of data point components. Having these limits is particularly useful in very dense regions of a distribution of data points, as tiny leaf nodes a capable of collecting many data points. The cost of building a tree is small, and also once the tree has been constructed, it can be utilised for various operations. Figure 3.3 shows the basic process of building a 2 dimensional kd-tree, the process starts from the top left with the root node containing of the data points. The bounding box is then split by its widest dimension to create two child nodes, the process is repeated for each child node to even smaller scales until a balanced tree is built, if possible.

Once the tree has been built, how is it accessed? There are different types of ways to search trees, the search method used for this particular work is *Range Searching* as illustrated in Figure 3.3. A query is requesting nodes that satisfy a particular condition begins with the root node. The question asked is whether the node lies within a particular search radius, if the node is inside the radius then the node is divided and the same question is probed to its child nodes, and if the nodes are within the radius they are added to the correlation pair counts. If the nodes do not lie within the search radius, the node is not added to the correlation and the following node is considered. What makes range searching favourable is that very far away points are not searched, as in the green boxes in Figure 3.3, but if the radius is too large then all the points inside that radius (points in the pink boxes) must be visited. The advantage kd-trees have over the $O(\mathcal{N}^2)$ approach is that instead of calculating

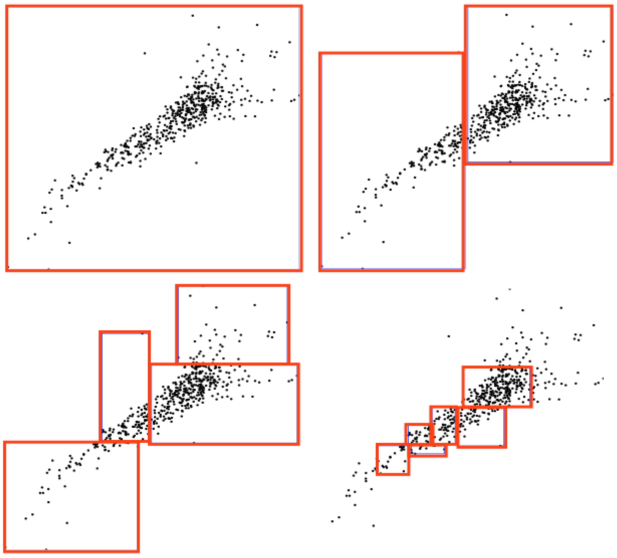


Figure 3.2: Construction of a 2-dimension kd tree. The first 4 tree building stages are illustrated. [Image courtesy: Moore et al. (2001)]

the distance to every pair of points, the distances for entire groups are calculated by looking at the bounding boxes of different groups. The kd-tree approach reduces the computation runtime of correlation functions to $O(\mathcal{N} \log \mathcal{N})$ scales.

As we shall see, one can further optimize kd-trees through Message Passing Interface (MPI) and Open Multiprocessing (Open MP). Open MP is a software package that allows one to distribute a computational problem over several processors. Each processor solves a subset of the problem, once all processors are completed they combine to give the final solution of the computation. Open MPI is an open source version of MPI which facilitates message passing within the different processors involved in the computation.

3.2.2 Current implementation of optimized codes

The application of the kd-tree algorithm in computing correlation functions has been developed in various ways. In this section we briefly summarize publicly available computational packages used for computing correlation functions. It should be noted, however, that a detailed comparison of the performance of these codes is beyond the scope of this work.

Fast Two Point Correlation Code

The Fast two-point Correlation Code (TPACF)[2] by Dolence & Brunner (2008) computes the spatial and angular 2-point correlation function using a modified kd-tree data structure

[2] `http://lcdm.astro.illinois.edu/code.html`

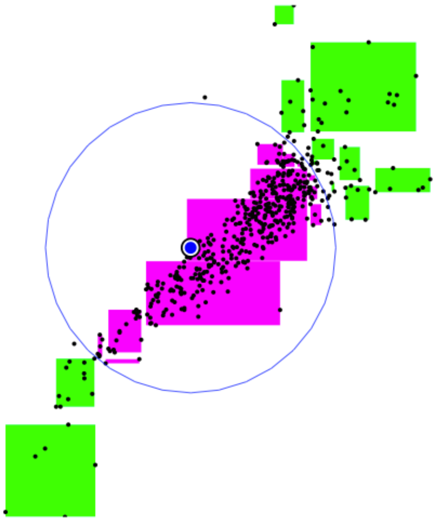


Figure 3.3: Range searching in a 2 dimension kd tree. All the points in the pink boxes are visited, while the points in the green are not considered, this makes searching in a kd-tree faster than the traditional approach. [Image courtesy: Moore et al. (2001)]

called the *dual kd-tree algorithm*. Dual kd-trees provide a way to divide the computation into two separate parts by converting the DD and RR computation into an additional kd tree, this is done by having two references to the same tree in memory. TPACF was first implemented on a sample of galaxies selected from the SDSS DR7 data release (Wang et al. 2013). The code can be parallelized using MPI or Open MP to improve efficiency.

Correlation Utilities and the Two-Point Estimation (CUTE)

CUTE was developed by David Alonso (Alonso 2012) and is a publicly available code for estimating the 2PCF. CUTE is implemented in two forms, the first form makes use of Open MP, allowing it to be parallelized for shared memory machines. The second estimates these correlation functions by utilizing Graphical Processing Units (GPU's).[3] CUTE estimates 4 different types of correlation functions: the monopole, 3-D, the spatial and angular correlation function.

HealPix

Although not using the kd-tree environment, the Hierarchical Equal Area ISO-Lattitude Pixelization (HEALPix) (Górski et al. 2005), is a scientific data processing and analysis package applied to a variety of scientific problems that involve data distributed on a sphere. HEALPix is a support structure which allows for the discretization of data with very high resolution. Initially designed for data analysis and processing, HEALPix has been applied to a many large projects including the latest Planck Cosmic Microwave Background (CMB)

[3] A GPU is hardware that is mainly used in image building and image processing, due to its parallelized structure for shared memory, they are ideal for intensive numerical computations

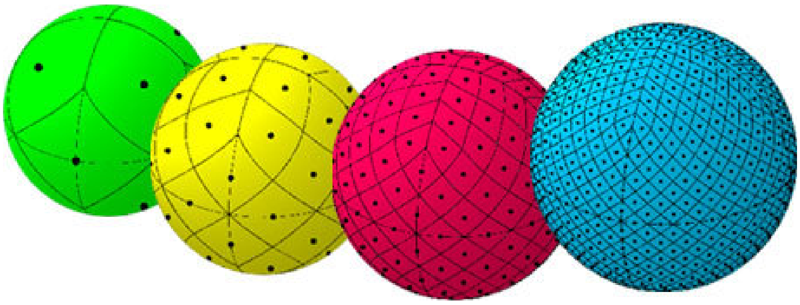


Figure 3.4: An illustration of the partitioning into equal areas of HEALPIX, The green sphere is the has the lowest resolution of 12 pixels of equal size. The blue sphere has the highest resolution of 768 pixels which is similar to 7.3 degree resolution. [Image courtesy: `http://healpix.jpl.nasa.gov`]

anisotropy measurements. For clustering statistics the positions of a galaxy are placed on a sphere such as the one below, with each pixel on the sphere representing an area on the sky that is pixelized. Figure 3.4 shows the division of a sphere into equal pixels, from left to right is the lowest to highest resolution partitioning.

3.3 Kosmo Stats Package (KSTAT)

Finally, the Kosmo Stats Package (KSTAT) is an ongoing development initiated by Cristiano Sabiu at the Korea Institute of Advanced Studies (KIAS), South Korea, specifically designed to compute correlation functions. In collaboration with Sabiu we have been developing this package for public release.

The package is designed for the user to easily compute the angular, spatial 2- and 3-point correlation function, as well as the anisotropic 2PCF. KSTAT has been optimized to make use of high performance facilities and can implement bootstrapping and jackknifing error estimation, by parallelizing the computation of the correlation functions. Since KSTAT is still in the development stage, we have been able, through rigorous testing, to debug the algorithm and begin improvements to the 3PCF calculation by e.g. incorporating standard error estimation such as jackknifing and bootstrapping. In what follows, we have performed various benchmarking tests as well using KSTAT to explore the BAO and the 3pt function with the latest available BOSS data releases.

For our analysis we made use of the Sepnet Computing Infrastructure for Astrophysical Modelling and Analysis (SCIAMA), which is a high performance computing cluster hosted by the Institute of Cosmology and Gravitation (ICG) at the University of Portsmouth. It contains 1008 2,66 GHz Intel Xeon processors, with 26 GBytes of memory per core, giving it a total memory per core of 2 TBytes. The cluster also has fast parallel storage with 85 TBytes and 10 TBytes of NFS storage. Sciama has an open source stack and can be used on 3 types of networks, infiniband, 100bT and GigaBit (Burton,G .2014)

KSTAT computes correlation functions using two script files. The first script file contains commands which specify which correlation function to be computed, the data files to be used (i.e. the galaxy sample and its randomly generated sample), the angular or spatial scales being probed, the number of angular or spatial bins, and the type of error estimation method to be used (jack knifing or bootstrapping).

The second script is used to execute KSTAT specifying how many processors are to be used for a particular computation and how the computation is distributed across the chosen number of processors (which internally executed with Open MPI). The second script file also provides a way to track the progress of the computation by continuously updating in a log file.

3.3.1 Implementing KSTAT

Angular 2PCF and Spatial 2PCF

The schematic below shows an example script file used to invoke the 2PCF computation.

```
#!/bin/bash
time ../bin/2pcf -gal galaxy_sample.dat -ran random_sample.dat -rmin 1.0
-rmax 20.0 -nbins 5 -out output.dat -wgt .true.
```

The arguments with a '-' sign as prefix are flags. The input files galaxy_sample.dat and random_sample.dat contain positions of galaxies given in rectangular coordinates (x,y,z) measured in Mpc with the last column being the weight, are read in using -gal and -ran. If the input files are in angular coordinates (RA,DEC), they are converted to rectangular using these relations

$$x = D_C \cos(\phi) \sin(\theta), \tag{3.2}$$

$$y = D_C \sin(\phi) \sin(\theta), \tag{3.3}$$

$$z = D_C \sin(\phi), \tag{3.4}$$

where θ is the right ascension (RA), ϕ is the declination (DEC). D_C is the comoving distance given by

$$D_C = \frac{c}{H_0} \int_0^z \frac{dz'}{E(z')}, \tag{3.5}$$

where c is the speed of light, H$_0$ is the Hubble constant and $E(z')$ is given by

$$E(z') = \sqrt{\Omega_m (1+z)^3 + \Omega_K (1+z)^2 + \Omega_\Lambda}, \tag{3.6}$$

where z is the redshift, Ω_M, Ω_Λ are the respective dimensionless matter and dark energy density parameters and Ω_K is the curvature of space parameter (Hogg 1999). Since we have no redshift information for the ACF, D_C is set to 1.

The argument -wgt .true. is the weight given to pairs of galaxies closer than $62''$ due to SDSS spectroscopy limitations, and is the fourth column in the galaxy sample file. In this example, the spatial separations to be probed are from -rmin 1.0 to -rmax 20.0 in Mpc, and the flag -nbins is used to input the number of bins one wishes to probe. With these settings KSTAT will compute the spatial 2PCF as default. To invoke the angular 2PCF, the flag -proj .true. is added to the script. In this new scenario, one must remember that the input data is formatted in spherical coordinates (RA, DEC), and the angular scales to be probed (rmin and rmax) must be in degrees(°). The result of the computation is then written to file output.dat.

Below is a schematic illustrating a typical way to submit a job to the SCIAMA cluster. We will briefly explain the important parts of the script.

```
#!/bin/tcsh
#PBS -l walltime=999:00:00
#PBS -l nodes=6:ppn=6
mpirun -np 36 ./test.script  > test.log
echo "Program finished at: ‘date‘"
```

The wall time is the maximum amount of time the job is on the cluster either in queue or in execution, if this time has passed the job is cancelled. The line PBS -l nodes= 6:ppn= 6 informs the cluster how to distributed the processors over the nodes. The are over 86 nodes available, each node consists of 12 processors. The MPI command "mpirun -np 36 ./test.script > test.log " executes the script file (test.script) explained above, -np is a flag for the number of processors to be used for the computation in this case 36. All the screen output and progress is written into a log file test.log.

The Reduced Spatial 3PCF

To compute the 3pcf we follow a methodology described in Peebles (1980a) and applied in (e.g. Kayo et al. 2004, Sabiu & Nichol 2009, McBride et al. 2011a). In Chapter 2, it was shown that the reduced 3PCF can be expressed as the ratio of the full 3PCF and the combination of 2PCF's for each side of a triangle

$$Q = \frac{\zeta(r_1, r_2, r_3)}{\xi(r_1)\xi(r_2) + \xi(r_2)\xi(r_3) + \xi(r_1)\xi(r_3)}, \tag{3.7}$$

where ζ is the full 3PCF given in Equation 2.9. and ξ is the 2PCF in Equation 2.4. Since $\zeta = \zeta(r_1, r_2, r_3)$ is a function of 3 variables that define a triangular configuration distinctively, these shape parameters can also be expressed as (s, q, θ),

$$s = r_1, \tag{3.8}$$

$$q = \frac{r_2}{r_1}, \tag{3.9}$$

$$\theta = cos^{-1}(\hat{r_1} \cdot \hat{r_2}), \tag{3.10}$$

where in Equation 3.10, $\hat{r_1} \cdot \hat{r_2}$ is the dot product of the unit vectors of two sides of a triangle, s and q are constants while θ varies from 0° to 180°. These shape parameters are illustrated in Figure 2.1.

Similar to the 2PCF, the computation of the reduced 3PCF (Q) on SCIAMA is executed using a script such as the one below, the difference is in how the spatial scales are probed. The input for the spatial scales are lengths of 2 sides of a triangle, with r1min being the minimum length of r_1 and r1max being the maximum length of side r_1, this is similar to the lengths of r2min, and max. The midpoint (rmax+rmin2)/2 gives the spatial range being probed on each of the two sides of the triangular configuration as illustrated in Figure 3.5.

```
time ../bin/3pcf -gal galaxy_sample.dat -ran random_sample.dat -r1min 7
-r1max 8 -r2min 5 -r2max 6  -nbins 8  -wgt .true.  -out output.dat
```

Once the computation has been completed, in the output file are columns describing the lengths of the third side per bin, the normalized triplet counts and the reduced 3PCF, Q. To plot Q, we need to take a step further and convert the length of the third side of the triangular configuration into θ and this is achieved using the following equation based on the geometry of the triangle described in McBride et al. (2011a)

$$\cos\theta = \frac{r_1^2 + r_2^2 - r_3^2}{2r_1r_2}, \tag{3.11}$$

where r_1 and r_2 are the selected lengths provided as input, r_3 is the length of the third side of the triangle given as output by KSTAT, and θ is the opening angle of the triangle deduced from the cosine rule.

3.3.2 KSTAT Benchmarking

To explore the performance of KSTAT on the cluster, we conducted a several test runs, under different conditions. The first runtime test was on the spatial 2PCF using 18 spatial bins and probing small scales between 0 and 20 Mpc, and also moderate spatial scales be-

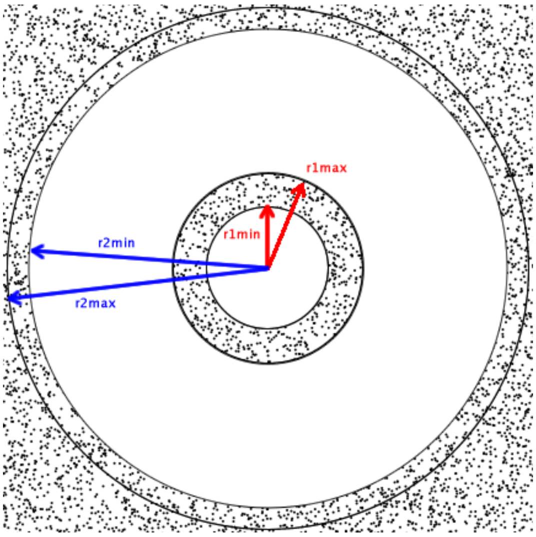


Figure 3.5: The figure illustrates how KSTAT computes Q. Triplet counts are computed for galaxies which reside in the inner and outer circular rings.

tween 0 and 90 Mpc. This test was performed on both the DR9 and DR10 samples, which had 207,246, 203,613 galaxies respectively, 3.5 million randoms for the DR9 and 9.7 million for the DR10 sample. In Figure 3.6 we show how the computation runtime of the 2PCF scales with the number of processors. The solid lines show the 2PCF running with no error estimation and the dashed line with the bootstrapping option invoked using 18 bootstraps in each run. The red lines have been run on scales 0 to 20 Mpc and the blue increased to 90 Mpc. As we can see, when probing out to relative small scales ($<$ 20 Mpc), when using $>$ 10 processors we gain little in computational time. This is most likely due to each processor taking about the same amount of time to compute the correlation function as to build the tree. As expected, with the increase of processors we observe a decrease in the computation runtime, as detailed in Table 3.2. We also observe that there is a point where increasing the number of processors does not give significant gain in the reduction of the computation runtime. This could be a limitation of the algorithm used, or the processor waiting time, where processors have to wait for other processors to finish executing their parts of the computation before returning the final result. It is also interesting to note that invoking the bootstrapping does not seem to add significant computational time over these moderate scales.

In Figure 3.7 we explore computation runtimes for the spatial 3PCF on the same dataset for a standard equilateral triangle configuration of with r1min = 14 Mpc, r1max = 16 Mpc, r2min = 14 Mpc and max = 16 Mpc and with no error estimation invoked (it should be noted that the current version of KSTAT does not have error estimation option for the 3PCF and therefore has to be computed manually). Comparing to Figure 3.6, we achieve similar runtimes.

Table 3.2: 2PCF computation runtimes on the DR9 dataset with 207,246 galaxies and 3.5 million random galaxies. 18 spatial bins selected and the scales probed were between 0 and 90 Mpc.

Number of Processors	**Runtime (mins)**
1	00:09:21
8	00:04:01

Lastly, in Figure 3.8 we investigated how the spatial scales being probed affect the computation runtime. It should be noted that this test was performed with no error estimation. This demonstrates that there is a significant trade off in computational time if one wants to extend to large scales. This also perhaps demonstrates the limitations of the tree code, where probing the largest scales requires to search most of the tree space.

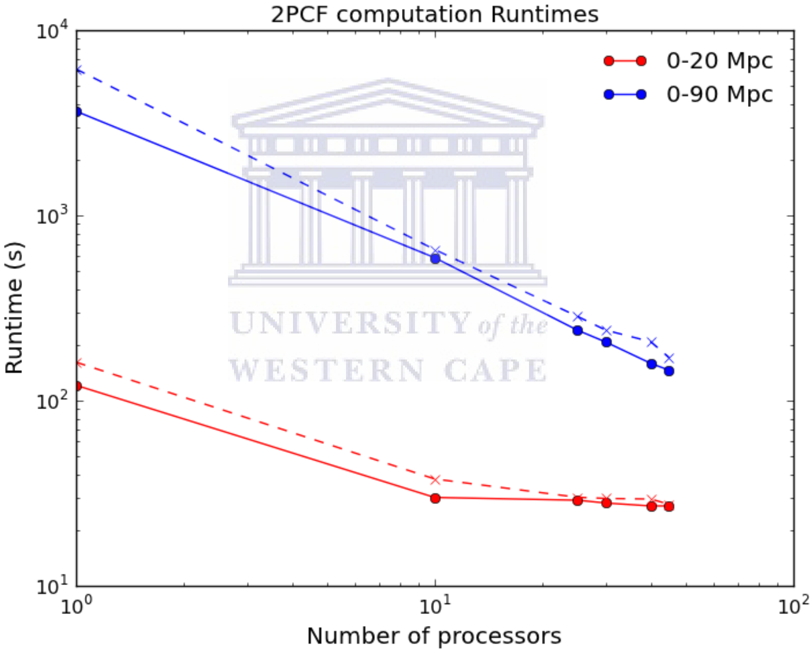


Figure 3.6: The figure shows how the run time compares with the number of processors on KSTAT, the dashed lines represent the computation of the 2PCF with bootstrapped error estimation, solid lines is the computation times without any error estimation. The test was performed on the DR9 dataset with 207,246 galaxies and 3.5 million random galaxies

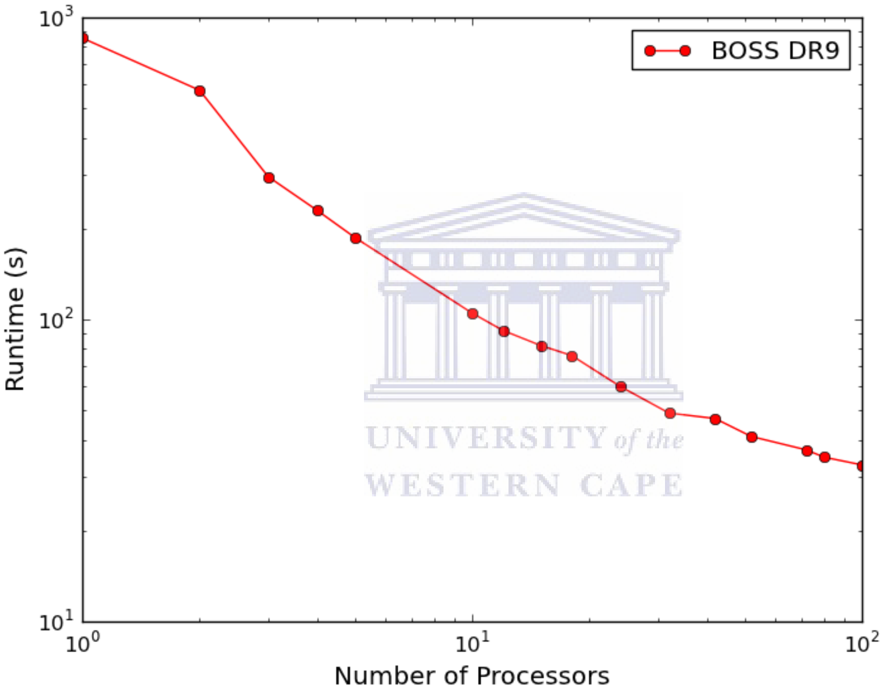


Figure 3.7: 3PCF computation runtimes on KSTAT for the BOSS DR9 dataset. Here we chose equilateral triangular configuration. It should be noted that no error estimation was performed.

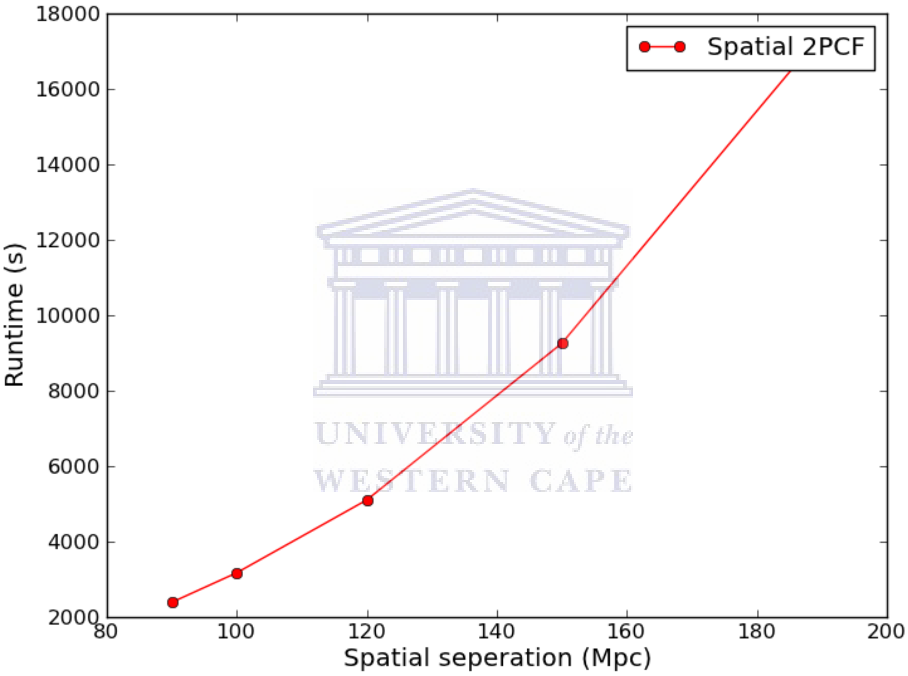


Figure 3.8: Illustration of how the computation runtime of the 2PCF scales with the spatial separation.

Chapter 4

Applying KSTAT to BOSS

In this chapter we compare KSTAT outputs of the 2pcf to that of recent results using BOSS DR9 and D10. We also present the first exploration into the 3pcf as a function of redshift using BOSS DR10. We begin by giving an overview of data selection for these samples.

4.1 Data Selection

For our analysis we apply KSTAT to BOSS DR9, DR10 samples which is summarized in Table 4.1. The DR9 & DR10 CMASS are samples of massive LRGs whose stellar whose stellar mass remains constant over the entire redshift range being probed, hence the acronym CMASS, 'Constant Mass'. The LOWZ sample targets galaxies in the low redshift range $0.15 < z < 0.4$, and the CMASS sample spans a higher redshift range of $0.43 < z < 0.8$. A target selection algorithm described in Eisenstein et al. (2001) was used to obtain these samples. In order to conduct large scale structure analysis, the data samples must be checked for completeness, Anderson et al. (2013) describes the method used for these samples. More details on the galaxy selection and color cuts can be found in (Dawson et al. 2013, Anderson et al. 2013, Tojeiro et al. 2014).

For our clustering measurements we look at galaxies in the Northern Galactic Cap (NGC) in both the DR10 low and high redshift samples and only the high redshift in the DR9 data sample. Table 4.1 also shows summarized the Eisenstein et al 2005 data for comparison.

4.1.1 Weighting the DR9 galaxies

When making observations, there are various limitations on how well one obtains positions of galaxies. To collect galaxy positions and redshift information, BOSS makes use of a fibre-fed multi-object spectrograph. The diameter of a spectroscopic fibre is $62''$ on the focal plane. If two galaxies are separated by less than this angular diameter they cannot be observed with a single fibre plate, this causes a systematic undersampling for galaxies that reside in groups or clusters. This limitation is known as fibre collisions. By default, each galaxy is assigned a weight of $w_{\rm cp} = 1$. When a galaxy is not allocated a fibre, $w_{\rm cp} = 2$.

It can also happen that the spectrum of a galaxy is measured but no redshift is calculated, these are known as redshift failures and are denoted as $w_{\rm rf}$. Angular systematic errors must also be accounted for, and are denoted as $w_{\rm sys}$. These limitations are corrected for by upweighing a galaxy by one unit. Cosmic variance and shot noise errors greatly affect clustering measurements, and weights are also applied to optimise clustering measurements using a relation by Feldman et al. (1994),

$$w_{\rm FKP} = 1 + n(z_i)P_0, \tag{4.1}$$

where $n(z_i)$ of a galaxy distribution at z_i and $P_0 = 20000\ h^{-3}{\rm Mpc}^3$, which is the scale independent value of the power spectrum optimized for the BAO peak. P_0 corresponds to $P(k = 0.1h\ {\rm Mpc}\ ^{-1})$. The total weighting for the galaxy distribution is then given by

$$w_{\rm tot} = w_{\rm FKP} w_{\rm sys}(w_{\rm rf} + w_{\rm cp} - 1), \tag{4.2}$$

where $w_{\rm rf}$ and $w_{\rm cp}$ are set to 1 by default, $w_{\rm FKP}$ and $w_{\rm sys}$ are multiplicative terms which depend on spatial location (Anderson et al. 2012).

To determine the correlation function from a sample of galaxies, one has to obtain the average galaxy density of the sample. A way in which this estimate is found is by creating a random catalog with unclustered objects. This random catalog must take into account the survey geometry, the redshift and angular selection function from the galaxy sample. To limit shot noise errors, the number of objects in the random catalog are several times greater than the number of objects in the galaxy sample as shown in Table 4.1. Objects in the random catalog are weighted by $w_{\rm FKP}$.

4.1.2 Weighting the DR10 galaxies

The weighting for the DR10 galaxies follows a similar method as that of the DR9 detailed in Anderson et al. (2012). Galaxies in close pairs are given a weight of $w_{\rm cp} = 1$, and galaxies in which redshifts were not obtained are also weighted by $w_{\rm rf} = 1$. An additional weight in the DR10 CMASS is given in order to account for the stellar density ($w_{\rm star}$), seeing ($w_{\rm see}$) and number density of galaxies that are observed. The total weight is calculated as

$$w_{\rm tot} = (w_{\rm cp} + w_{\rm rf} - 1)w_{\rm star}w_{\rm see}, \tag{4.3}$$

where $w_{\rm star}$ and $w_{\rm see}$ are set to 1 for all LOWZ galaxies. The random catalog was created the same way as explained in the previous section. The weighting is $w_{\rm tot} = w_{\rm FKP}$.

The weights and galaxy positions are supplied as input to KSTAT when computing the

Table 4.1: Summary of data samples used for this analysis

Dataset	**Number of objects**	**Number randoms**	**Coverage in deg^2**	**redshift (z) range**	**median z**
SDSS DR3	47,063	2,322,580	3,816	$0.15 < z < 0.4$	0.35
BOSS DR9 CMASS North	207,246	3,499,986	3,275	$0.4 < z < 0.7$	0.55
BOSS DR10 LOWZ North	203,613	9,694,563	5,156	$0.15 < z < 0.45$	0.32
BOSS DR10 LOWZ South	76,803	3,688,587	5,156	$0.15 < z < 0.45$	0.32
BOSS DR10 CMASS North	392,372	19,120,990	5,161	$0.43 < z < 0.8$	0.57
BOSS DR10 CMASS South	109,472	5,345,587	5,161	$0.43 < z < 0.8$	0.57
CMASS total (North+South)	501,844	23,466,577	10,322	$0.43 < z < 0.8$	0.57
LOWZ total (North+South)	280,416	13,383,150	10,312	$0.15 < z < 0.45$	0.32

different types of correlation functions. For each pair count we add the weight to it, such that it becomes the pair count + $(\omega_i + \omega_j)$, instead of +1, as in the standard case,where i and j are the indicies of the ith and jth galaxy.

In Figure 4.1, we show the redshift distribution of the datasets used in this analysis. In purple is the DR3 catalog which was used by Eisenstein et al. (2005) providing the first significant detection of the BAO peak was made. In green, red and blue are the BOSS DR9, DR10 CMASS and LOWZ catalogs respectively. In Figure 4.2 and 4.3 we show the coverage maps of these galaxy catalogs. From DR3 to DR10, we can observe the increase in the number of objects cataloged. In the Results section we also refer to the BOSS DR11 sample which has not been publicly released yet. We show 2PCF measurements from the first results obtained from the number of galaxies they have collected thus far. Mock catalogs for these datasets will also have the same coverage footprint as the real galaxy catalogs.

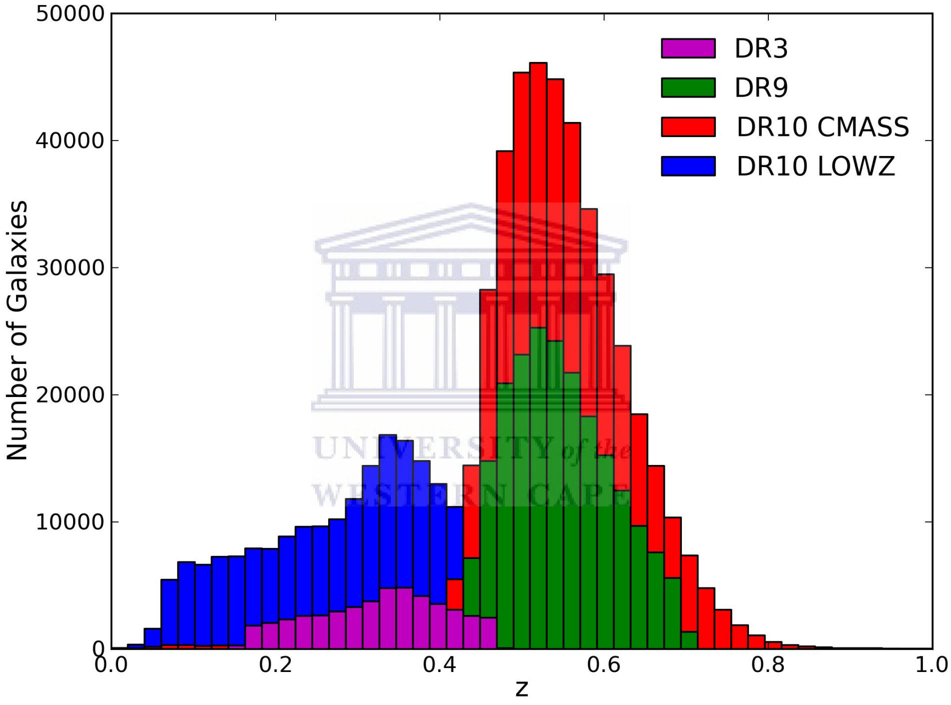


Figure 4.1: The redshift distribution of LRGs in the SDSS DR3, DR9 and DR10 samples.

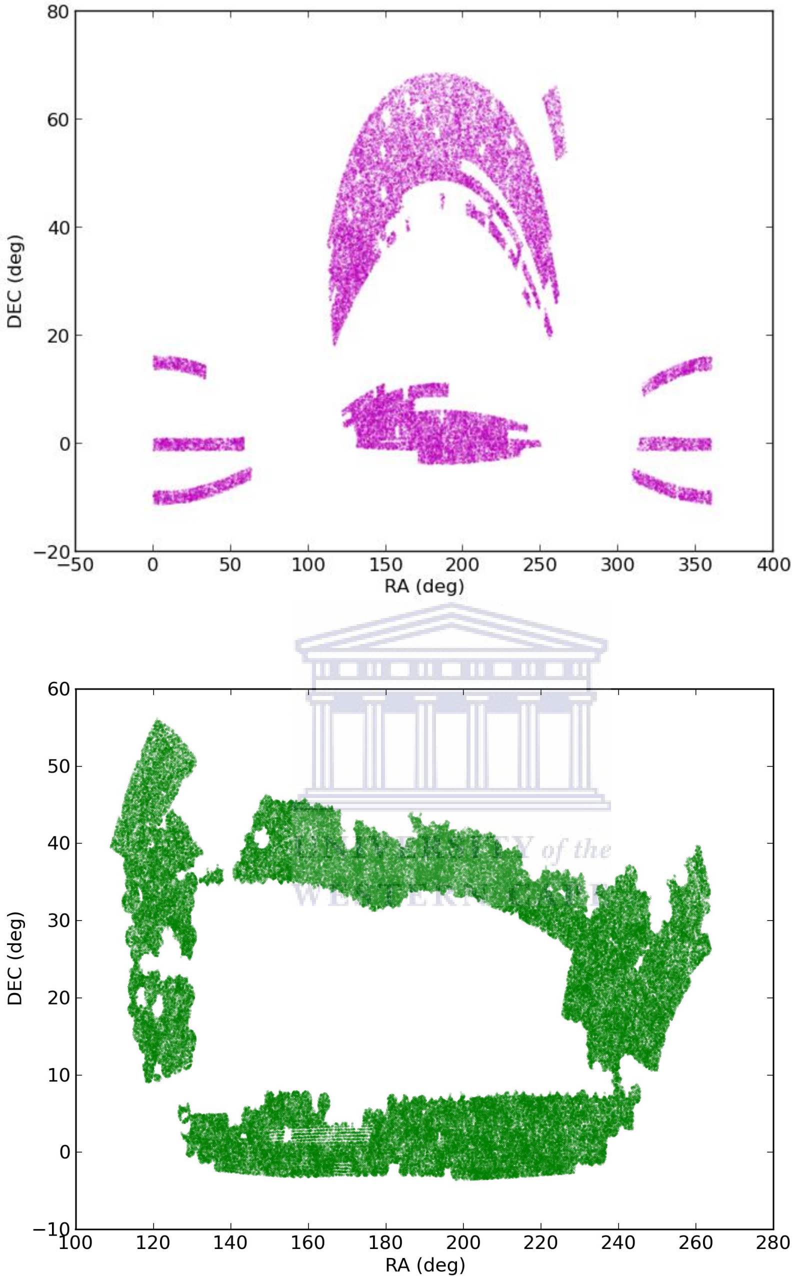


Figure 4.2: On the top panel is the projected sky coverage map of the SDSS DR3 sample used in Eisenstein et al. (2005). On the bottom panel is the DR9 CMASS samples used for this analysis.

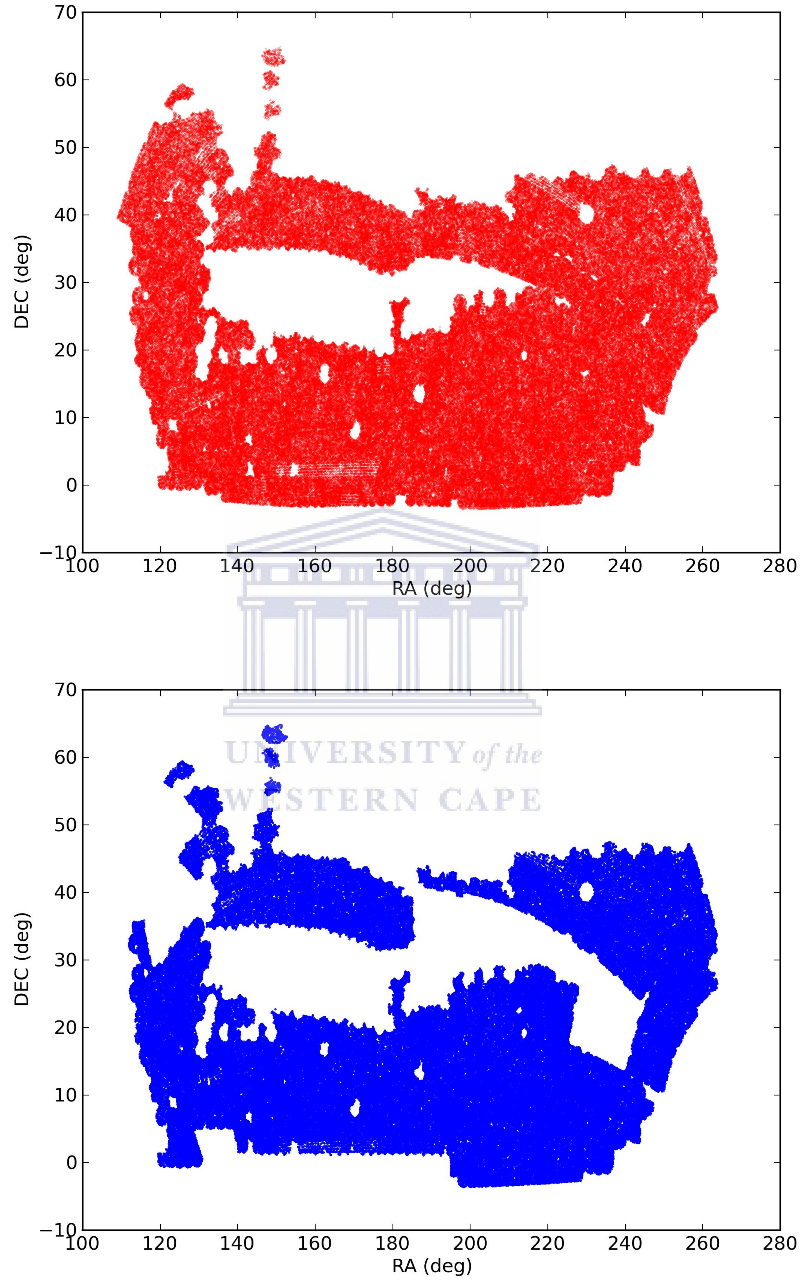


Figure 4.3: The top panel shows the DR10 CMASS galaxy footprint and the bottom shows DR10 LOWZ footprint.

4.2 Results

4.2.1 Probing the BAO with BOSS DR9 and DR10

We present the KSTAT outputs of the 2pcf to probe the BAO in the BOSS data. We apply a spatial range out to d<190 Mpc using 25 equally spaced bins. The top panel of Figure 4.4 shows the results applied to the DR9 sample and the bottom left and right panels show the DR10 LOWZ and CMASS samples respectively. In both panels the blue shaded region is the jackknife error estimation whilst the red region is the bootstrapped error region, with the respective solid lines showing the mean. On the bottom panel, the left plot is the 2PCF measurement of the DR10 LOWZ sample and on the right is the 2PCF measurement for the DR10 CMASS sample. In Figure 4.5 we show how the jackknife regions applied to

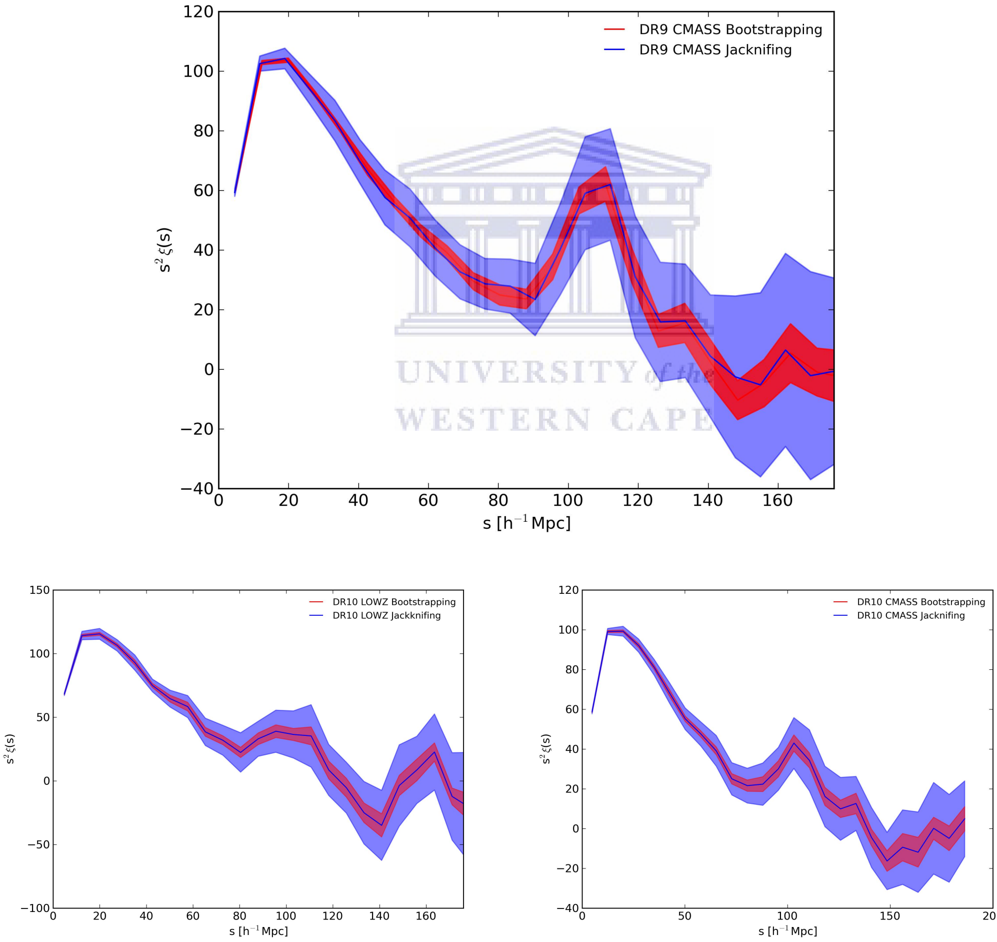


Figure 4.4: The spatial 2PCF measurement of BOSS DR9 (top panel) and DR10 sample (bottom panel). On bottom panel, on the left we show the 2PCF measured from the DR10 LOWZ sample and on the right is the 2PCF measured from the DR10 CMASS sample.

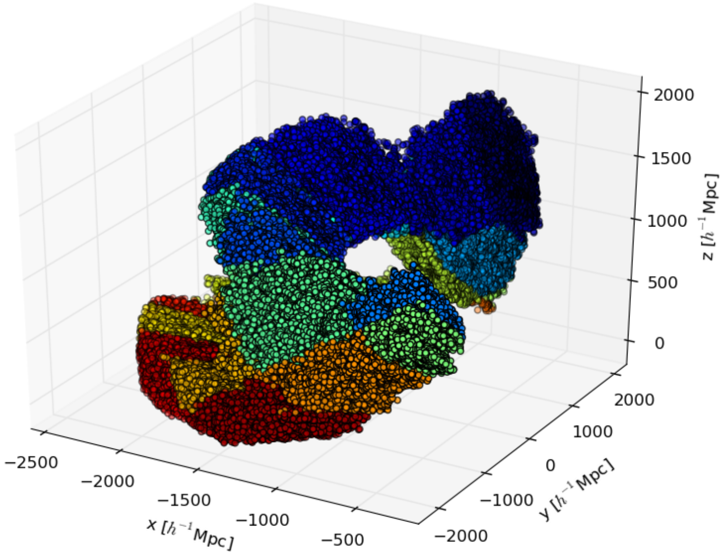


Figure 4.5: The jackknifing error estimation technique on the DR9 dataset in 3 dimensions. The different colours correspond to the different jackknifed samples created.

the DR9 sample sample in (x, y, z) co-ordinates. The various colors represent the different jackknifed regions as determined using HEALPIX.

KSTAT applies the standard approach to compute the 2PCF and we can see that it recovers the BAO peak well at s ∼ 110 Mpc. The bootstrapping technique is a useful probe of the statistical uncertainty when the theoretical distribution is complicated or unknown. However, the Jackknifing technique is used to estimate the bias of a statistic. In the remainder of this work we consider only the statistical uncertainty through the bootstrapping technique. Exploration of potential biases will be explored in future work.

In Figure 4.6 we show our KSTAT 2PCF measurements from the DR9 CMASS sample. On the same plot we show the Anderson et al. (2012) (green) and Sánchez et al. (2012) 2PCF measurements (blue). Also shown on the figure as the black curve is the theoretical prediction derived from *CAMB*. *CAMB* is the Code for Anisotropies in the Microwave Background developed by Lewis et al. (2000). In essence, CAMB models temperature and matter fluctuations observed in the CMB and outputs the matter power spectrum. Using the linear power spectrum measured by the WMAP satellite at $z = 30$ (Komatsu et al. 2009), CAMB models how the matter fluctuations would evolve from $z = 30$ to a particular redshift. We convert the matter power spectrum to a correlation function using the procedure used in Komatsu et al. (2009)[4]. Using the model from CAMB, the code then converts the power

[4] http://www.mpa-garching.mpg.de/~komatsu/crl/list-of-routines.html

spectrum to a spherically-averaged 2PCF as a function of R [h^{-1} Mpc] given by

$$\xi(R) = \int \frac{k^2 dk}{2\pi^2} P(k) \frac{\sin(kR)}{kR}, \tag{4.4}$$

where *P(k)* is the spherically-averaged power spectrum given by

$$P(k) = \int_0^1 d\mu P(k\mu). \tag{4.5}$$

In Figure 4.6 we can see that the Anderson and Sanchez results are consistent with each other although using different approaches to minimize bias in the measurements. Our results using a 'raw' form of the 2PCF is also consistent with their results on scales > 40 Mpc. Below this scale we do find a slight tilt in the 2PCF relative to theirs (at the time of writing this thesis, the corresponding error points for both Anderson and Sanchez were unavailable). However, it is interesting to note that our result on these scales seems to agree well with the theoretical prediction, but all measurements show a higher amplitude in the BAO compared to the CAMB prediction.

We then explored clustering measurements on the BOSS DR10 data as shown in Figure 4.7. 2PCF KSTAT output is shown in magenta, on the same plot we show measurements of the 2PCF on the LOWZ sample made by Sánchez et al. (2013) (blue) and Chuang et al. (2013) (red). Finally, the first measurement of the BAO by Eisenstein et al. (2005) is shown in green. Once again, we add the CAMB prediction in black. What we can observe is that our 2PCF measurements are broadly consistent on scales especially when considering that the error bars by Sánchez et al. (2012) and Anderson et al. (2012) will be highly correlated which have not been considered when plotting their error estimates. The shape of the correlation function is in agreement with Sánchez et al. (2013) and Chuang et al. (2013) on scales $<100\ h^{-1}$ Mpc. What is also interesting to note is the high amplitude in the Eisenstein et al. (2005) result. We also note that as with Sánchez et al. (2012) and Chuang et al. (2013) the BAO recovered peak appears to be quite broad compared to the CMASS sample in Figure 4.8.

In Figure 4.8 are the 2PCF measurements obtained from the BOSS DR10 and DR11 CMASS samples. In the cyan color are our measurements computed by KSTAT, in green is a result from Anderson et al. (2013), the red represents the result from Sánchez et al. (2013) and in blue is the 2PCF measurement made by Sánchez et al. (2013) from the DR11 data release (not yet publicly available). From the plots, it can be clearly observed that our measurements match very well with the other results particularly at large scales, however we do observe a slight departure on scales $<50\ h^{-1}$Mpc. We can also see the results are consistent with the CAMB prediction.

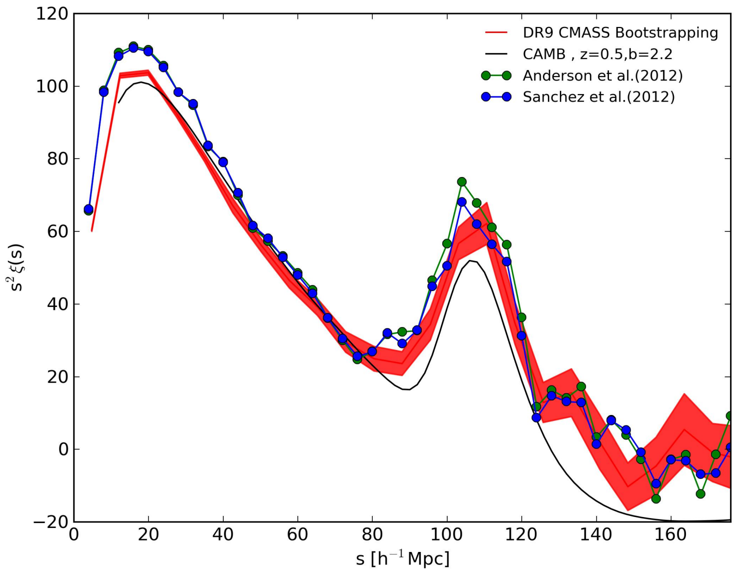


Figure 4.6: The plots shown are the same as Figure 4.4 (top panel) with only bootstrapped error estimation. We add the DR9 CMASS 2PCF results from Anderson et al. (2012), Sánchez et al. (2012) and a theoretical prediction from CAMB in black.

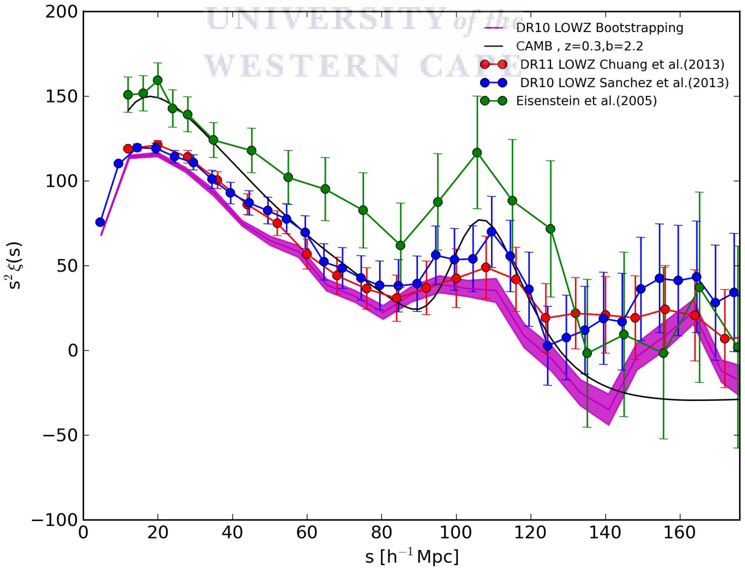


Figure 4.7: On this plot we show our 2PCF measurement (magenta) of the DR10 LOWZ sample, we also plot 2PCF measurements from Sánchez et al. (2013), DR11 2PCF measurement by Chuang et al. (2013) and the first 2PCF measurements to detect the BAO peak by Eisenstein et al. (2005). The 2PCF computed from CAMB is plotted in black

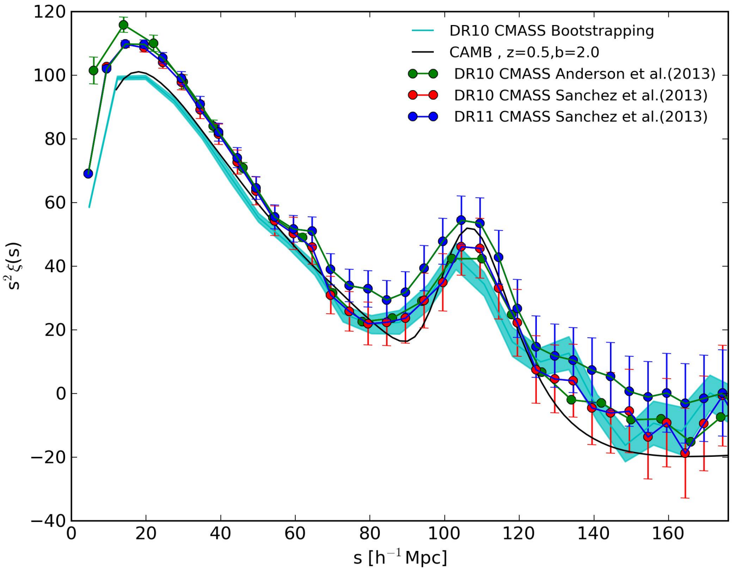


Figure 4.8: In this plot is our 2PCF measurement on the DR10 CMASS sample, to compare we plot other 2PCF measurements from published works by Anderson et al. (2012) and Sánchez et al. (2013). We also plot theoretical 2PCF from CAMB in black.

4.3 Results of the 3PCF

In this section we present the first and most precise measurements of the reduced 3PCF (Q) to date using the recent BOSS DR10 sample at higher redshift ranges than previously probed. The full DR10 CMASS ($\bar{z} \sim 0.5$) and LOWZ ($\bar{z} \sim 0.3$) samples used in this study contain 502,844 and 218,905 galaxies respectively. By using these samples we can, for the first time explore any evolution in the 3PCF at different scales in order to investigate aspects of the large scale structure.

For this study we have applied the reduced spatial 3PCF Q to the DR10 sample as shown in Figure 4.9. The top panel set shows results for the combined North and South regions which maximizes our coverage (5156 deg^2 LOWZ and 6161 deg^2 CMASS). To explore any spatial biases arising from the sample selection we have also tested the 3PCF on the separate northern and southern regions as shown in the bottom two panel sets in the figure. Each panel set shows three plots representing different triangular configurations that probe increasing large scales: r1=3.8, r2=7.5 Mpc (left), r1=7.6, r2 =15.0 Mpc (middle) and r1=15.0, r2=30.0 Mpc (right). The blue and red lines on each plot represent the high (CMASS) and low (LOWZ) redshift samples respectively with the 1σ bootstrapped errors.

Looking firstly at the top panel in Figure 4.9 we can see the combined North and South results. For the smallest triangular configuration in the left hand plot we can see a fairly constant value of $Q \sim 0.7$ across the θ/π range in the low redshift sample. A similar result is observed for the high redshift sample, showing only a slight decrease in Q at θ/π =0.6. However, as we move to increasingly larger scales in the middle panel, this allows better access to probe the filamentary structure and we begin to see the familiar U shape in the correlation function. This feature is most prominent in the largest configuration shown in the right plot where we see both the low and redshift samples dipping to $Q \sim 3.0$ at θ/π =5.2. This characteristic shape seems consistent with previous results observed by McBride et al. (2011a) (see Figure 2.12) on an SDSS LRG sample of galaxies with a mean smaller redshift of z=0.104. If we now look at the separate north and south regions, overall we observe a consistent picture of the correlation function for the three configurations, indicating no directional dependency in the result.

Looking now at any Q as a function of redshift we can see a fairly consistent picture emerging. In the combined north and south region data we can see that even on the largest scales, the low and high redshift samples show consistency within the error bars, with only a slight departure at θ/π =0.45 in the right hand plot of the top panel set. Although the north region in the middle panel set shows consistency at all scales, the southern region (bottom right panel set) shows a significant difference at θ/π =5.2 where the low redshift sample minimizes at $Q \sim 0.2$ and the high redshift sample shows a value of $Q \sim 0.8$ at the same bin. The most likely reason for this is probably due to the relatively reduced sample size of the southern region which could produce a noisier estimate of the 3PCF. However, this is certainly something that would require deeper exploration in future work.

To fully explore and interpret this result we will require to compare to current theoretical predictions of the 3PCF. In continuation of this work, beyond this project, we can use theoretical predictions from Gil-Marín et al. (2014) to constrain cosmological parameters. The Bispectrum is the 3PCF in Fourier space. In Gil-Marín et al. (2014), they showed that the halo bispectrum as a function of the matter power spectrum as

$$\begin{aligned}B_h(\mathbf{k}_1, \mathbf{k}_2) = b_1^3 P_{\delta\delta}(k_1) P_{\delta\delta}(k_2) 2F_2(\mathbf{k}_1, \mathbf{k}_2) + b_1^2 b_2 P_{\delta\delta}(k_1) P_{\delta\delta}(k_2) + \\ b_1^2 b_{s^2} P_{\delta\delta}(k_1) P_{\delta\delta}(k_2) S_2(\mathbf{k}_1, \mathbf{k}_2) + cyc., \quad (4.6)\end{aligned}$$

where $P_{\delta\delta}$ is the matter power spectrum, b_1 and b_2 are the linear and non linear bias parameters respectively, b_{s^2} is the non-local bias term. F_2 is the second kernel from the standard perturbation theory (SPT) and is expressed as

$$F_2(\mathbf{k}_1, \mathbf{k}_2) = \frac{5}{7} + \frac{1}{2}\cos(\alpha_{ij})\left(\frac{k_i}{k_j} + \frac{k_j}{k_i}\right) cos^2(\alpha_{ij}) \quad (4.7)$$

where $\alpha(i,j)$ is the angle between the Fourier space vectors $\mathbf{k}_i$ and $\mathbf{k}_j$ and S_2 is defined from the tidal tensor and is expressed as

$$S_2(\mathbf{k}_1,\mathbf{k}_2) \equiv \frac{(\mathbf{k}_1 \cdot \mathbf{k}_2)^2}{(k_1 k_2)^2} - \frac{1}{3}. \tag{4.8}$$

From this bispectrum definition above, we can then compute the theoretical 3PCF using a relation from Takada & Jain (2003) ,

$$\zeta(\mathbf{r}_1,\mathbf{r}_2,\mathbf{r}_3) = \int \prod_{i=1}^{3} \frac{d^3\mathbf{k}_i}{(2\pi)^3} B(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3) \exp[i(\mathbf{k}_1 \cdot \mathbf{r}_1 + \mathbf{k}_2 \cdot \mathbf{r}_2 + \mathbf{k}_3 \cdot \mathbf{r}_3)](2\pi)^3 \delta_D(\mathbf{k}_{123}) \tag{4.9}$$

where $\mathbf{k}_{123} = \mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3$. We can then use eq. 4.9 to model the 3PCF to provide parameter constraints and compare with other work in this field and thus make more meaningful analysis of our 3PCF measurements at these redshifts.

In this chapter we have examined in more detail, the performance of the KSTAT package by applying it to the BOSS data to explore to both the BAO and the 3PCF correlation function. By directly comparing to existing data and theoretical predictions we demonstrated the usefulness of KSTAT as a robust statistical probe of the spatial 2PCF having recovered well the BAO signature peak and showing consistency with published works from the BOSS team.

We also examined the spatial reduced 3PCF on the BOSS DR10 data and consequently have provided the first and most precise measurement of the 3PCF to date. For our chosen triangular configuration, we showed that our results are consistent with previous works in this area and also extended work in this field by providing a measurement of the 3PCF as a function of redshift. So far, we have found no evidence for and evolving 3pt function, however future work could examine many different configurations and other selection criteria e.g. magnitude cuts, to provide a more concise picture of the growth of structure on a variety of scales.

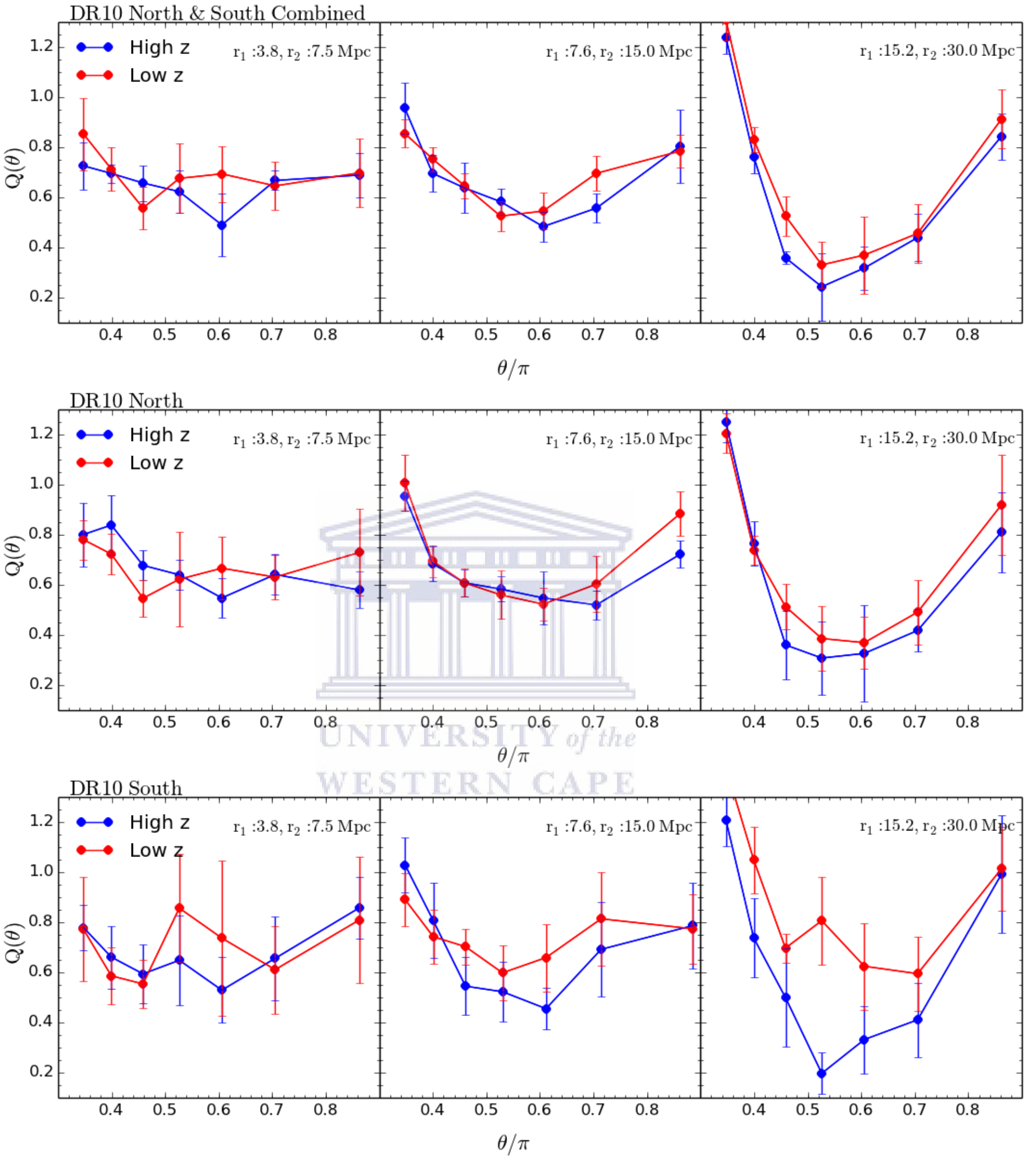


Figure 4.9: The reduced 3PCF measurement from the BOSS DR10 data release. The red color represents the Q from DR10 CMASS sample and the blue is the Q from the LOWZ sample. From the left to the right column are the different triangular configurations, from these values KSTAT computes the opening angle θ which represents the length of the third side of the triangle.

Chapter 5

Conclusion

In this work we have explored the use of correlation functions as a probe of the large scale structure of the universe. In particular we have examined the ongoing development of the software packaged called KSTAT that can perform optimized calculations of a variety of correlation estimators.

We began by describing the framework current standard model of cosmology. We briefly reviewed observational probes that provide measurements that support the standard ΛCDM model. We also took a historical look at the development of galaxy redshift surveys that have revolutionized our understanding of the large scale distribution of galaxies in the Universe. Finally we looked towards the enticing prospects of the next generation surveys from telescopes such as the SKA, LSST, and EUCLID.

In chapter 2 we discussed the statistical tools required to compute clustering statistics of the large scale structure of the Universe, through the 2PCF and 3PCF. We gave a brief overview on the construction of the random catalogs, described the error estimation techniques used (jackknifing and bootstrapping) and reviewed some of the key work surrounding the BAO with the 2PCF and emerging studies probing the higher order 3PCF.

In chapter 3 we illustrated how the 2PCF and 3PCF are applied in practice and looked at methods used to optimise computation of the 2PCF and 3PCF, briefly reviewing some of the most recent publicly available packages for estimating correlation functions. In particular, we introduced the package used in this work called KSTAT, detailing how it can be used to compute the 2PCF and 3PCF. As an ongoing development, we provided a range of benchmarking tests of KSTAT to demonstrate its scalability in a parallel computing environment. We showed KSTAT performs very well with large datasets and scales well in an MPI environment making it promising code for next generation datasets from surveys such as the SKA,LSST, and EUCLID.

Finally, we used KSTAT on some of the most recent BOSS data to probe the BAO and compare with current results. The results we obtain are consistent with that of (Anderson et al. 2012; 2013, Sánchez et al. 2013, Tojeiro et al. 2014, Chuang et al. 2013) on these datasets.

For the first time we performed the most precise measurements of the 3PCF at high redshift on the BOSS DR10 data release. Our reduced 3PCF results at low redshift $z = 0.3$ are consistent with those of McBride et al. (2011a) at $z = 0.104$ in the SDSS LRG sample. In this initial analysis, we have found no evidence of evolution of the 3PCF on small and large scales, however we do observe the characteristic U shape of the reduced 3PCF as one increases scale.

5.1 Future Work

From a practical point of view there a number of ways in which KSTAT could be improved before being made public that were identified during this work. In particular, the calculation of the 3PCF does not implicitly compute error estimation (such as jackknife and/or bootstrapping) as the 2PCF does. This is something that we are currently working on to improve. At its heart, when running KSTAT in an MPI environment requires the kd-tree to be created on each node the code is distributed on. This is a rather memory intensive procedure that could be improved in the future by e.g. building it once and allowing access from each node instead. Scalability to very large datasets, further testing will be carried out inorder to determine whether KSTAT will be able to handle the large datasets from future surveys such as the SKA, LSST, and EUCLID.

From a scientific point of view, we have really only scratched the surface of the potential for which the 3PCF can offer. With the 3PCF, we would like to investigate other possible triangular configurations and obtain theoretical predictions for the 3PCF so that we can make comparisons. As we are about to enter a new age of radio astronomy, there is much scope to apply the 3PCF on simulations. We would also like to apply the projected 3PCF on simulations such as S^3 inorder to conduct forecasts for future surveys such as the SKA. Once we have a framework on 3PCF analysis, we can then use this tool to constrain cosmological parameters. We intend to extend the work of Gaztañaga et al. (2009) to test whether can indeed probe the BAO with the 3PCF using the latest BOSS data releases.

Bibliography

Alonso, D. 2012, ArXiv e-prints, 1210.1833

Amanullah, R. et al. 2010, ApJ, 716, 712, 1004.1711

Anderson, L. et al. 2013, ArXiv e-prints, 1312.4877

——. 2014, MNRAS, 439, 83, 1303.4666

——. 2012, MNRAS, 427, 3435, 1203.6594

Bassett, B., & Hlozek, R. 2010, Baryon acoustic oscillations, ed. P. Ruiz-Lapuente, 246

Becker, R. H., White, R. L., & Helfand, D. J. 1995, ApJ, 450, 559

Blake, C. et al. 2011, MNRAS, 418, 1707, 1108.2635

Blake, C., & Wall, J. 2002, MNRAS, 329, L37, astro-ph/0111328

Blanton, M. R. et al. 2001, AJ, 121, 2358, astro-ph/0012085

Burton,G. .2014, SCIAMA , High Performance Computing Cluster

Charlier. 1922, Astron Fys, 16:1

Chuang, C.-H. et al. 2013, ArXiv e-prints, 1312.4889

Colless, M. 1999, Royal Society of London Philosophical Transactions Series A, 357, 105, astro-ph/9804079

Condon, J. J., Cotton, W. D., Greisen, E. W., Yin, Q. F., Perley, R. A., Taylor, G. B., & Broderick, J. J. 1998, AJ, 115, 1693

Conley, A. et al. 2011, ApJS, 192, 1, 1104.1443

Contreras, C. et al. 2013, MNRAS, 430, 924, 1302.5178

Cress, C. M., Helfand, D. J., Becker, R. H., Gregg, M. D., & White, R. L. 1996, ApJ, 473, 7, astro-ph/9606176

Davis, M., & Peebles, P. J. E. 1983, ApJ, 267, 465

Dawson, K. S. et al. 2013, AJ, 145, 10, 1208.0022

Dodelson, S. 2003, Modern cosmology

Dolence, J., & Brunner, R. 2008, The Fast Two Point Correlation Function

Efron, B. 1979, The Annals of Statistics, 7, 1

Eggen, O. J., Lynden-Bell, D., & Sandage, A. R. 1962, ApJ, 136, 748

Eisenstein, D. J. et al. 2001, AJ, 122, 2267, astro-ph/0108153

Eisenstein, D. J., & Hu, W. 1999, ApJ, 511, 5, astro-ph/9710252

Eisenstein, D. J., Seo, H.-J., Sirko, E., & Spergel, D. N. 2007, ApJ, 664, 675, astro-ph/0604362

Eisenstein, D. J. et al. 2005, ApJ, 633, 560, astro-ph/0501171

ESA. 2014, Euclid Space Telescope, The European Space Ageny Mission

Falco, E. E. et al. 1999, PASP, 111, 438, astro-ph/9904265

Feldman, H. A., Kaiser, N., & Peacock, J. A. 1994, ApJ, 426, 23, astro-ph/9304022

Fisher, K. B., Davis, M., Strauss, M. A., Yahil, A., & Huchra, J. P. 1994, MNRAS, 267, 927, astro-ph/9308013

Folkes, S. et al. 1999, MNRAS, 308, 459, astro-ph/9903456

Fosalba, P., Crocce, M., Gaztanaga, E., & Castander, F. J. 2013, ArXiv e-prints, 1312.1707

Friedman, J. H., Bentley, J. L., & Finkel, R. A. 1977, ACM Transactions on Mathematical Software, 3, 209

Gaztañaga, E., Cabré, A., Castander, F., Crocce, M., & Fosalba, P. 2009, MNRAS, 399, 801, 0807.2448

Geller, M. J., & Huchra, J. P. 1989, Science, 246, 897

Georgakakis, A., Mobasher, B., Cram, L., Hopkins, A., & Rowan-Robinson, M. 2000, A&AS, 141, 89, astro-ph/9910318

Gil-Marín, H., Wagner, C., Noreña, J., Verde, L., & Percival, W. 2014, ArXiv e-prints, 1407.1836

Górski, K. M., Hivon, E., Banday, A. J., Wandelt, B. D., Hansen, F. K., Reinecke, M., & Bartelmann, M. 2005, ApJ, 622, 759, astro-ph/0409513

Groth, E. J., & Peebles, P. J. E. 1977, ApJ, 217, 385

Gunn, J. E., & Gott, III, J. R. 1972, ApJ, 176, 1

Gunn, J. E. et al. 2006, AJ, 131, 2332, astro-ph/0602326

Hamilton, A. J. S. 1993, ApJ, 417, 19

Hinshaw, G. et al. 2009, ApJS, 180, 225, 0803.0732

Hogg, D. W. 1999, ArXiv Astrophysics e-prints, astro-ph/9905116

Huchra, J., Davis, M., Latham, D., & Tonry, J. 1983, ApJS, 52, 89

Ivezic, Z. et al. 2008, ArXiv e-prints, 0805.2366

Jing, Y. P., & Boerner, G. 1998, ApJ, 503, 37, astro-ph/9802011

Jing, Y. P., & Börner, G. 2004, ApJ, 607, 140, astro-ph/0311585

Kayo, I. et al. 2004, PASJ, 56, 415, astro-ph/0403638

Keisler, R. et al. 2011, ApJ, 743, 28, 1105.3182

Kneib,J.P. 2014, eBOSS Survey (SDSS IV)

Komatsu, E. et al. 2009, ApJS, 180, 330, 0803.0547

Landy, S. D., & Szalay, A. S. 1993, ApJ, 412, 64

Lewis, A., Challinor, A., & Lasenby, A. 2000, ApJ, 538, 473, astro-ph/9911177

Liddle, A. R. 1999, An introduction to modern cosmology

Ling, E. N., Barrow, J. D., & Frenk, C. S. 1986, MNRAS, 223, 21P

McBride, C. K., Connolly, A. J., Gardner, J. P., Scranton, R., Newman, J. A., Scoccimarro, R., Zehavi, I., & Schneider, D. P. 2011a, ApJ, 726, 13, 1007.2414

McBride, C. K., Connolly, A. J., Gardner, J. P., Scranton, R., Scoccimarro, R., Berlind, A. A., Marín, F., & Schneider, D. P. 2011b, ApJ, 739, 85, 1012.3462

Moore, A. W. et al. 2001, in Mining the Sky, ed. A. J. Banday, S. Zaroubi, & M. Bartelmann, 71, astro-ph/0012333

Nuza, S. E. et al. 2013, MNRAS, 432, 743, 1202.6057

Padmanabhan, N., Xu, X., Eisenstein, D. J., Scalzo, R., Cuesta, A. J., Mehta, K. T., & Kazin, E. 2012, MNRAS, 427, 2132, 1202.0090

Passmoor, S., Cress, C., Faltenbacher, A., Johnston, R., Smith, M., Ratsimbazafy, A., & Hoyle, B. 2013, MNRAS, 429, 2183, 1211.5589

Peebles, P. J. E. 1970, AJ, 75, 13

——. 1973, ApJ, 185, 413

——. 1974, ApJ, 189, L51

Peebles, P. J. E. 1980a, in Annals of the New York Academy of Sciences, Vol. 336, Ninth Texas Symposium on Relativistic Astrophysics, ed. J. Ehlers, J. J. Perry, & M. Walker, 161–171

——. 1980b, The large-scale structure of the universe

Peebles, P. J. E., & Groth, E. J. 1975, ApJ, 196, 1

Percival, W. J. et al. 2001, MNRAS, 327, 1297, astro-ph/0105252

Planck Collaboration et al. 2013, ArXiv e-prints, 1303.5083

Rengelink, R. B., Tang, Y., de Bruyn, A. G., Miley, G. K., Bremer, M. N., Roettgering, H. J. A., & Bremer, M. A. R. 1997, A&AS, 124, 259

Rubin, V. C., Ford, W. K. J., & . Thonnard, N. 1980, ApJ, 238, 471

Ryden, B. 2006, Introduction to cosmology

Sabiu, C., & Nichol, R. 2009, in Bulletin of the American Astronomical Society, Vol. 41, American Astronomical Society Meeting Abstracts #213, #340.02

Sánchez, A. G. et al. 2013, MNRAS, 433, 1202, 1303.4396

——. 2012, MNRAS, 425, 415, 1203.6616

Schlegel, D., White, M., & Eisenstein, D. 2009, in Astronomy, Vol. 2010, astro2010: The Astronomy and Astrophysics Decadal Survey, 314, 0902.4680

Schlegel, D. J. et al. 2007, in Bulletin of the American Astronomical Society, Vol. 39, American Astronomical Society Meeting Abstracts, #132.29

Scranton, R. et al. 2002, ApJ, 579, 48, astro-ph/0107416

Searle, L., & Zinn, R. 1978, ApJ, 225, 357

Smoot, G. F. 1995, Planet. Space Sci., 43, 1449

Sofue, Y., & Rubin, V. 2001, ARA&A, 39, 137, astro-ph/0010594

Szapudi, S., & Szalay, A. S. 1998, ApJ, 494, L41

Takada, M., & Jain, B. 2003, MNRAS, 340, 580, astro-ph/0209167

Tegmark, M. et al. 2004, Phys. Rev. D, 69, 103501, astro-ph/0310723

Tojeiro, R. et al. 2014, MNRAS, 440, 2222, 1401.1768

Verde, L. et al. 2002, MNRAS, 335, 432, astro-ph/0112161

Wang, Y., Brunner, R. J., & Dolence, J. C. 2013, MNRAS, 432, 1961, 1303.2432

Wright, E. L. 2003, ArXiv Astrophysics e-prints, astro-ph/0306132

York, D. G. et al. 2000, AJ, 120, 1579, astro-ph/0006396

Zehavi, I. et al. 2004, ApJ, 608, 16, astro-ph/0301280

Zheng, Z. 2004, ApJ, 614, 527, astro-ph/0405527

Zwicky, F. 1933, Helvetica Physica Acta, 6, 110