Universitat Autònoma de Barcelona

Departament de Física

Institut de Ciències de l'Espai (CSIC-IEEC)

PhD Thesis:

Populating cosmological simulations with galaxies using the HOD model

by

Jorge Carretero Palacios

Advisors:
Francisco Javier Castander
Enrique Gaztañaga

Tutor:
Enrique Fernández

Barcelona, November, 2012

A thesis submitted by Jorge Carretero Palacios
for the degree of Doctor of Philosophy in the
Universitat Autònoma de Barcelona

Jorge Carretero Palacios

Advisors:

Dr. Francisco Javier Castander

Dr. Enrique Gaztañaga

Tutor:

Dr. Enrique Fernández

Barcelona, November, 2012

Agradecimientos

En primer lugar quiero agradecer a Enrique Gaztañaga, Francisco Castander y Enrique Fernández por permitirme realizar una tesis doctoral. A Enrique Gaztañaga por haberme dado la oportunidad de trabajar junto a él y el resto de compañeros del grupo de investigación del Institut de Ciències de l'Espai (CSIC-IEEC). Por haberme mostrado su confianza en un primer momento para comenzar la tesis y a continuación para emprender otra etapa profesional en el Port d'Infomatió Científica. A Francisco por aguantarme, por escucharme, por ayudarme y por enseñarme tantas cosas. Por hacérmelo pasar bien mientras investigaba, por darme su confianza, por siempre estar dispuesto a perder el tiempo y por enseñarme lo útil que puede llegar a ser la función arcotangente.

Gracias a todo mi grupo de investigación por hacerme la vida más agradable.

Gracias a la gente de mi despacho actual con la que he pasado un año fenomenal y que me han aguantado durante el proceso de escribir la tesis. Y también a los que estaban en el despacho del ETSE y aunque hace mucho tiempo de ello, a los del lujoso cyber del ICE.

Quiero recordar mi estancia en Philadelphia en donde conocí a gente extraordinaria. Gracias a Ravi por dejarme colaborar con él y a todo el grupo de investigación en el que me sentí como uno más.

Gracias a mi familia. A mis padres y a mi hermana por quererme, por cuidarme, darme ánimos y buenos consejos. Gracias a mis tí@s y a mis prim@s.

Quiero darle las gracias a mi novia (y a mi pacs) por ser tan buena, por cuidarme, quererme, confiar en mí y por aguantarme en los malos momentos. En cuanto acabe nos pegamos unas buenas vacaciones.

Gracias a mis amigos y a mis compañeros de diferentes pisos y equipos de fútbol de Barcelona. Entre todos han conseguido que me guste vivir en ella. Incluso estando el Barça de por medio.

Gracias a mis amigos de toda la vida. A mis colegas de Jaca, a los de Torremayor y la gente que conocí en Grenoble.

Abstract

This thesis presents a method to build mock galaxy catalogues by populating N-body simulations using prescriptions based upon the halo occupation distribution model (HOD). The catalogues are constructed to follow some global local properties of the galaxy population already observed, such as the luminosity function, the colour-magnitude diagram and the clustering as a function of luminosity and colour. The observed data constraints come from the Sloan Digital Sky Survey (SDSS).

The theoretical framework in which the production of the catalogues is based on, the halo model and the HOD, are described. Our mock catalogues are built from halo catalogues extracted from the Marenostrum Institut de Ciències de l'Espai N-body dark matter simulations (MICE). We characterize our input halo catalogues by computing their halo mass function, two-point correlation function and linear large scale halo bias.

The HOD provides prescriptions of how galaxies populate haloes. The HOD can be parameterized in several ways. We start by following the HOD recipes given by Skibba & Sheth 2009 to generate galaxy catalogues. Since the luminosity function of the catalogue does not fit observations, we investigate an analytical derivation of two HOD parameters, M_{min} and M_1 (α is assumed to be 1), by only using two observed constraints: the galaxy number density and bias. Then, a grid of 600 mock galaxy catalogues that covers a wide range of values of the three HOD parameters, M_{min}, M_1 and α, is generated to obtain the best-fit HOD parameters that match the observed clustering of galaxy luminosity threshold samples. As we cannot match observations we introduce additional ingredients: the SubHalo Abundance Matching (SHAM) and a modified NFW density profile.

A *unique* mock galaxy catalogue that follows at the same time the clustering at all luminosities and colours is produced using the halo catalogue extracted from the snapshot at z=0 of the MICE Grand Challenge run. The catalogue is built by following a new algorithm in which several modifications are introduced: scatter in the halo mass - central luminosity relation, the HOD parameter M_1 is set as a function of M_h, and three Gaussian components (instead of only two) are included to describe the colour-magnitude distribution. A derivation of the luminosity function and the linear galaxy bias of the mock catalogue is shown. How galaxy velocity affects the galaxy clustering and an estimation of the angular correlation function at the BAO scale are presented too. Finally, different versions of the catalogue currently used in PAU and DES projects, which include specific characteristics such as shear information or 42 different magnitudes in narrow band filters and also morphological properties for each galaxy, are briefly described.

Contents

Chapter 1

Introduction

1.1 Motivation

Observational cosmology unlike theoretical cosmology is a very "young science". Its origin, despite chinese astronomers documenting a "new star" in 185 AD that appeared in the sky and was visible for eight months, could be dated when Vesto Slipher detected the first redshift of spectral lines from distant galaxy spectra [1]. The difference between cosmology and almost every other science is that cosmologists observe an already completed experiment, which is our universe. There is no way, for instance, of controlling with some determined conditions a star, or recreating the formation of cosmic structure in a laboratory. This last example can be taken as the main motivation of this work: in almost every study related to large scale structure of the universe a mock galaxy catalogue is essential. Moreover, relevant physical processes, where perturbation theory fails, are within the nonlinear regime. In order to reproduce the non-linear regime of matter clustering and achieve the accuracy of observations it is mandatory to use numerical simulations with enough precision and realistic modeling.

The first evidence of the expansion of the universe was Hubble's law. General relativity theory predicts this phenomenon but it was Edwin Hubble, in 1929[2], who discovered this law by studying spectra of galaxies that contained cepheid stars. These stars allow us to derive how far their host galaxies are from us. Hubble observed a systematic redshift in almost all the spectra of the selected galaxies and realized that they were all moving away from us. When he compared the velocities and the distances of the galaxies he found a linear relation between them. At small redshifts a linear proportionality to velocity v can be used, $v = cz$ (where c is the speed of light in a vacuum). Figure 1.1 shows the velocity-distance relation originally published for extra galactic Nebulae. This relation is known as Hubble's law. Recent studies using different ways of measuring distances show the same trend found by Hubble (see figure 1.2).

[1] The redshift (or blueshift) of a galaxy is the shifting of its spectral features to longer (or shorter) wavelengths primarily due to the combination of Doppler motions and the general expansion of the universe.

[2] Although a recent article written by Livio 2011 demonstrates that Edward Lemaître was actually the first one who derived the law from the General Relativity equations in a 1927 article.

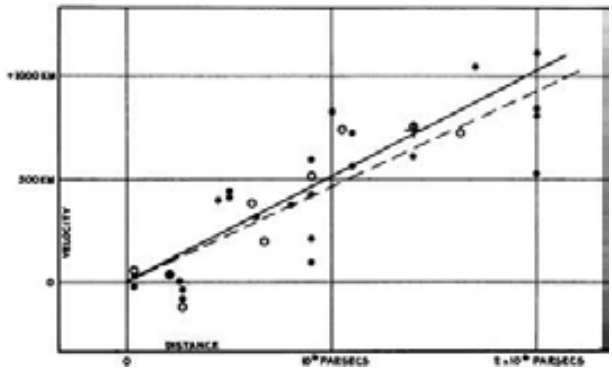

Figure 1.1: Velocity-Distance relation among Extra-Galactic Nebulae. Figure from Hubble 1929.

Saul Perlmutter, Brian P.Schmidt and Adam Riees won the Nobel Prize for Physics in 2011 "for the discovery of the accelerating expansion of the universe through observations of distant supernovae". Two independent international teams, the High-z Team and the Supernova Cosmology Project (Riess et al. 1998; Perlmutter et al. 1999), reached the conclusion of accelerating expansion. This is one of the most significant discoveries in cosmology, and it was an unexpected fact. Cosmologists had been trying to probe the decceleration of the universe due to the gravitational force for decades. Type Ia Supernovae appeared fainter than would be expected for a decelerating universe and the reason for this phenomenon is still an open debate. Data suggest that since around $z = 0.5$ the expansion of the universe has been accelerating. Significant efforts have been made in observations and theory in the last decade in order to understand this phenomenon. The accelerating expansion has been verified by a variety of independent types of observations: anisotropies of the Cosmic Microwave Background (CMB) (Jaffe et al. 2001, Pryke et al. 2002), regarding the large scale structure (LSS), the Baryon Acoustic Oscillations peak (BAO) (Eisenstein et al. 2005), the large-angle anisotropy in the CMB through the integrated Sachs-Wolf (ISW) effect (Boughn & Crittenden 2004, Fosalba & Gaztañaga 2004, Afshordi et al. 2004, Scranton et al. 2003), correlated distortions of galaxy shapes due to gravitational lensig (Schneider 2006, Munshi et al. 2008, Bacon et al. 2000, Kaiser et al. 2000, Van Waerbeke et al. 2000, Wittman et al. 2000), recent supernova results (Knop et al. 2003, Riess et al. 2001, Riess et al. 2004, Riess et al. 2007, Astier et al. 2006, Miknaitis et al. 2007), ratio of X-ray emitting gas to total mass in galaxy clusters (Allen et al. 2004, Allen et al. 2008), the age of the universe (Krauss & Chaboyer 2003, Tegmark et al. 2006). Citing Weinberg et al. 2012, "The question is no longer *whether* the universe is accelerating, but *why*".

The physical explanation for the accelerating expansion is still a mystery. Following the theory of general relativity, if the universe contains only ordinary matter and radiation, one would expect a slowing of the expansion due to gravity. At this point there exist two possibilities for explaining the accelerating expansion, either General Relativity theory fails at cosmological scales and it has to be replaced by another more complete theory or there exists an entity, which is called dark energy (DE), with peculiar physical properties that exhibits a gravitational force opposite

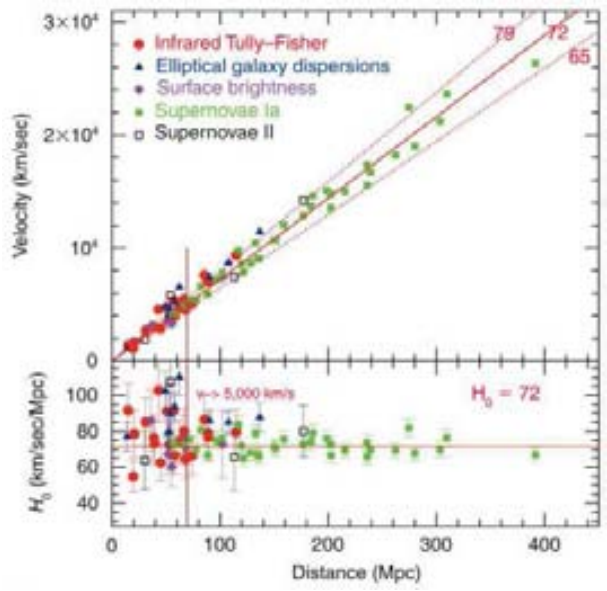

Figure 1.2: Top panel: Hubble diagram of distance vs. velocity for secondary distance indicators calibrated by Cepheids. Bottom panel: Value of H_0. Figure from Freedman et al. 2001.

to the gravity of ordinary matter.

The most straight forward and simple candidate for this dark energy component is vacuum energy. Zel'dovich in 1968 realized that the cosmological constant, introduced by Einstein and many times rejected and reincorporated in Einstein's equations (which is equivalent to the stress-energy of the vacuum) cannot be removed from the field equations of General Relativity. If one computes from quantum field theory this energy density, it is larger than what one observes by a factor of 10^{120}. Although there exists this disagreement between theory and observations the idea of an accelerating universe based on observations, is completely accepted as it provides the missing element needed to complete the current cosmological model. The standard cosmological model is based on the cosmological principle: our location is not special and on large enough scales, the universe looks homogenous in all directions and from every location. In this model the universe is flat and is in accelerating expansion. This model also assumes that the universe is formed of 23% dark matter, only 4.6% of baryonic matter and the remaining 72% is this entity called dark energy (based on data of the Wilkinson Microwave Anisotropy Probe (WMAP)). At the beginning it was hot and dense with an early expansion phase where light elements were formed by nucleosynthesis. It is also assumed that a previous accelerated expansion, called inflation, produced density perturbations from quantum fluctuations, leaving a trace of temperature fluctuations on the CMB which later will be the origin of the formation of large scale structure governed by gravitational instabilities. The standard model attempts to describe the observable universe. Present advances permit to measure cosmological parameters with high accuracy. The basic set of cosmological parameters according to Lahav & Liddle 2010 is shown in table 1.1.

Parameter	Symbol	Value
Hubble parameter	h	0.72 ± 0.03
Total matter density	Ω_m	$\Omega_m h^2 = 0.133 \pm 0.006$
Barion density	Ω_b	$\Omega_b h^2 = 0.0227 \pm 0.0006$
Cosmological constant	Ω_Λ	$\Omega_\Lambda = 0.74 \pm 0.03$
Radiation density	Ω_r	$\Omega_r h^2 = 2.47 \times 10^{-5}$
Neutrino density	Ω_ν	lower limit of $\Omega_\nu \approx 0.001$
Density perturbation amplitude	$\Delta^2_R(k = 0.002$ Mpc)	$(2.41) \pm 0.11) \times 10^{-9}$
Density perturbation spectral index	n	$n = 0.963^{0.014}_{-0.015}$
Tensor to scalar ratio	r	$r < 0.43(95\%$ conf.)
Ionization optical depth	τ	$\tau = 0.087 \pm 0.017$
Bias parameter	b	depends on galaxy type and scale

Table 1.1: Basic set of cosmological parameters from Lahav & Liddle 2010.

Observational cosmology has a fundamental role to understand the nature of dark energy. Dark energy has an impact on the geometry of the universe and on the growth of structure, both of which depend on the expansion history of the universe. The inhomogeneous large-scale distribution of matter is not directly observable because, as mentioned, it is basically formed of dark matter. If one assumes that galaxies are tracers of the underlying distribution of dark matter the properties of the large scale structure of matter could be studied by observing the galaxy distribution in the universe. This assumption is based on several plausible reasons. The most powerful method to study both expansion history and growth of structures is through galaxy redshift surveys, and mock galaxy catalogues are essential (calibrating errors or exploring systematic effects) to fully exploit data from future galaxy redshift surveys.

1.2 Outline of the thesis

The first chapter summarizes the motivation of this work, introduces the different probes to measure some cosmological parameters and describes briefly galaxy surveys, in particular future photometric galaxy surveys such as DES, PAU and Euclid in which our group is involved. The second chapter gives an idea of the state of the art of mock galaxy catalogues and concisely presents some methods to build them. The third chapter introduces the basics of the halo model and focuses on the halo occupation distribution model on which the mock galaxy catalogues of this work are based. The fourth chapter gives a brief summary of the basics of N-body simulations and presents the Marenostrum Institut de Ciències de l'Espai (MICE) cosmological simulations that are used to generate the mock galaxy catalogues. The fifth chapter explains the algorithm used to build our mock galaxy catalogues, the modifications and improvements to the method and the properties followed by the mocks such as the luminosity function and the color-magnitude diagram[3]. The sixth

[3]The galaxy luminosity function gives the mean number density of galaxies per luminosity interval and is the most fundamental statistic describing the galaxy population. The galaxy colour-

chapter presents some checks, validations and possible scientific applications of the generated mock galaxy catalogues such as the galaxy spatial two-point correlation function, the large scale galaxy bias, the effects of galaxy velocities in the galaxy clustering and the galaxy clustering at the BAO scales. The seventh and last chapter deals with the conclusions and future work.

1.3 Cosmological probes

This section describes some of the techniques used to constrain various parameters that govern the universe: CMB anisotropies, Ia-type Supernovae (SNe), weak lensing shear, cluster counts, the matter two-point correlation function, redshift space distortions and baryon acoustic oscillations.

1.3.1 Cosmic microwave background (CMB)

One of the most powerful probes of Cosmology is the comic microwave background radiation (CMB). According to the standard cosmological model it is the radiation that originated at an early stage in the development of the universe. It consists of photons that last interacted with matter when recombination occurred (at $z \approx 1000$). Along time photons travel and get colder and nowadays they are at a very low temperature, $T = 2.728K$. This electromagnetic radiation is remarkably uniform over the sky. The origin of the structures we observe in the universe is assumed to come from inhomogeneities at a very early stage of the universe. These inhomogeneities induce (small) anisotropies in the temperature of the CMB. In 1992 the COBE satellite first detected these temperature fluctuations (with an amplitude of about $\Delta T/T \sim 2 \times 10^{-5}$). To describe the statistical properties of the CMB anisotropies the two-point correlation function of the temperature on the whole sky is commonly employed, in the same way as it is used to study density fluctuations of galaxies. Since temperature fluctuations are defined in a sphere, spherical harmonics work analogously to the Fourier modes in a flat space. The power spectrum of temperature fluctuations describes the amplitude of the fluctuations on an angular scale $\theta \sim \pi/l$, where l is the fluctuation mode. Figure 1.3 shows the angular power spectrum of the CMB measured after 7 years of WMAP. These data are consistent with a cosmological model where the universe is flat and formed of a dark energy component that provides 73% of the total density of the universe, another 23% of the density is dark matter and only 4% of the density is ordinary matter. In addition, and not less curious, only 0.5% is composed of stars.

By itself, the angular power spectrum of the temperature fluctuations can constrain many of the cosmological parameters. In fact, CMB anisotropies depend on nearly all cosmological parameters, so in principle, by comparing the angular distribution of the CMB with theoretical expectations all these parameters can be determined.

magnitude diagram shows the relationship between luminosity or absolute magnitude and galaxy colour.

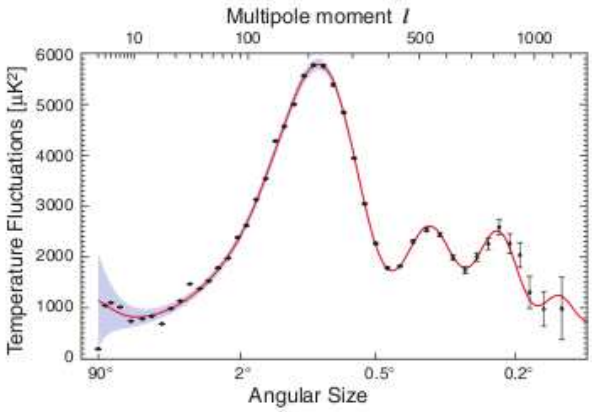

Figure 1.3: Angular power spectrum of the temperature fluctuations with 7 years of WMAP full-sky map. Source: http://map.gsfc.nasa.gov/. Credit: NASA Web Page.

1.3.2 Supernovae SNe Ia. Relation between distance and velocity

Type Ia Supernova (SNe Ia) are supposed to be the result of an explosion of a white dwarf that has crossed the critical mass threshold (Chandrasekhar limit) by accreting mass from a companion star (see review from Hillebrandt & Niemeyer 2000). The assumption is that this threshold should be the same for all type Ia Supernovae, and therefore all these objects have approximately the same intrinsic luminosity. If this is the case SNe Ia would be an ideal candidate for standard candles[4]. After correcting their light curves using an empirical correlation found between the peak luminosity and the width of the light curve SNe Ia become standardized candles.

Since the observed flux of a source depends on its luminosity, L, and its luminosity distance, d_L, which also depends on the cosmological model (and of course redshift), one can determine cosmological parameters by measuring the luminosity distance as a function of redshift. Different cosmological models with different values of Ω_m and Ω_Λ predict different forms for the relation. Models with $\Omega_\Lambda \neq 0$ tend to have luminosity distances that are larger at any given redshift and therefore objects appear systematically fainter than one would expect without dark energy. In order to detect deviations from the Hubble flow many SNe Ia in a wide range of redshift are necessary. Current data favour a universe dominated by dark energy. SNe Ia observations combined with CMB data provide the tightest constraints on the dark energy equation of state and by themselves are currently the best established probe of dark energy. Figure 1.4 shows the Hubble diagram for 557 SNe.

1.3.3 Weak gravitational lensing

When light passes through or near a mass distribution, according to general relativity, it suffers a deflection. If the deflection is strong this effect can produce giant arcs or multiple images. In the case of weak lensing the effect only produces distortions in the observed shapes of galaxies (Bartelmann & Schneider 2001, Schneider 2006,

[4] A standard candle is an astronomical object with known intrinsic luminosity.

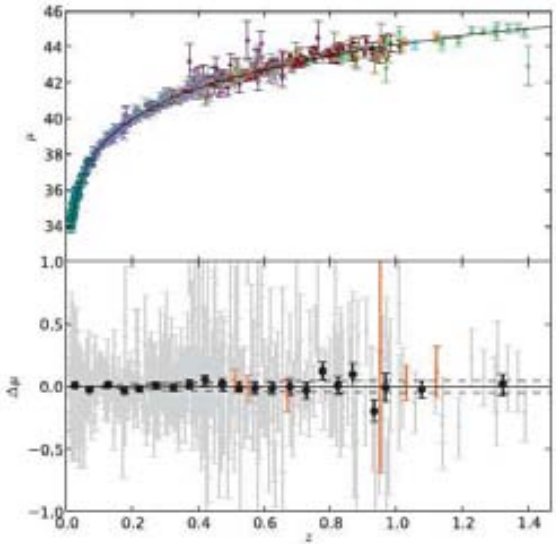

Figure 1.4: Upper panel: Hubble diagram for 557 SNe. The solid line represents the best-fitted cosmology for a flat universe including the CMB and BAO constraints. Colors represent different data sets. Lower panel: Hubble diagram residuals where the best-fitted cosmology has been subtracted from the light-curve shape and color corrected peak magnitudes. The gray points show the residuals for individual SNe, while the black points show the binned values in redshifts bins of 0.05 for $z < 1.0$ and 0.2 for $z > 1.0$. The orange points show the previously unpublished SNe introduced in the work of Amanullah et al. 2010. The dashed lines show the expected Hubble diagram residuals for cosmological models with $w \pm 0.1$ from the best-fitted value. Figure from Amanullah et al. 2010.

Munshi et al. 2008 for reviews). Light from distant galaxies gets deflected on the way to us by the mass distribution in between. Therefore, images of high-redshift sources suffer *shear* (stretching of the galaxy image) and *magnification* (ratio of the image area to the source area) because of gravitational lensing due to fluctuations of the gravitational potential. These effects, shear and magnification, can be related to the underlying density field of the large scale structure. Weak gravitational lensing provides a way of measuring the mass of astronomical objects without making any assumption about their composition or dynamical state.

To describe the distortion between a lensed and an unlensed system it is common to use a Jacobian matrix of this form:

$$A_{ij} = \frac{\partial \theta_i^S}{\partial \theta_j} = \begin{pmatrix} 1 - \kappa - \gamma_1 & -\gamma_2 \\ -\gamma_2 & 1 - \kappa + \gamma_1 \end{pmatrix} \tag{1.1}$$

where θ_i^S are the angular coordinates of the source and θ_j are the observed angles, κ is the convergence that is the responsible for the image size and $\gamma = \gamma_1 + i\gamma_2$ is the complex shear and controls the distortion of the shape. If one can measure the distortion or the ellipticity of observed galaxies, one can estimate the corresponding shear field via the following formula (Bartelmann & Schneider 2001):

$$\epsilon = \frac{\epsilon_S + g}{1 + g^* \epsilon_S} \tag{1.2}$$

where ϵ is a complex quantity that represents the observed ellipticity of a galaxy, ϵ_S is the source ellipticity and $g = \gamma/(1+\kappa)$ is the reduced shear. Since the intrinsic ellipticity is unknown, one has to measure as many as galaxy ellipticities as possible to average and infer the shear. If one assumes that galaxies are randomly oriented, then $\langle\epsilon\rangle = g$. Convergence and shear are related to the deflecting or gravitational potential via:

$$A_{ij} - \delta_{ij} = 2\int_0^{\chi} d\chi' \frac{r(\chi-\chi')}{r(\chi)r(\chi')} \frac{\partial^2\phi(\chi', r(\chi')\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j} \tag{1.3}$$

where χ is the comoving distance to the source and $r(\chi)$ is the corresponding coordinate distance. Cosmologists estimate from observations the angular shear power spectrum as a function of the multipoles, $P_\gamma(l)$. One can derive $P_\gamma(l)$ from equation 1.3 applying the Poisson equation and assuming the Limber approximation following Bartelmann & Schneider 2001:

$$P_\gamma(l) = \frac{9H_0^4\Omega_m^2}{4}\int_0^{\infty} dz \frac{W(z)(1+z)^2}{H(z)} P_\delta\left(\frac{l}{r(z)}, z\right) \tag{1.4}$$

where $W(z)$ is a weighting function, which depends on the source-distance distribution, and P_δ is the power spectrum of the density fluctuations.

1.3.4 Cluster counts

Another method used to constrain cosmological parameters comes from the study of the number of clusters of galaxies. Clusters of galaxies are the biggest gravitationally bound structures in the universe. The mass function of galaxy clusters[5] provides an effective method for constraining cosmological parameters (Press & Schechter 1974, Peebles 1993). The number of objects above a mass per unit comoving volume, can be predicted theoretically and depends on dark energy in two aspects, first geometrically through the comoving volume element, and second through the growth of the amplitude of the density perturbations, on which the cluster abundance depends exponentially.

The main issue is that one cannot directly observe the mass of a cluster, instead one has to estimate it using different methods. The simplest method is using stellar light and assuming a mass-to-light ratio. The first mass estimates were based on the velocity dispersion along the line-of-sight of the galaxies of the cluster and assuming that clusters are relaxed and in virial equilibrium. X-ray emission from hot ionized gas inside the cluster permits one to measure the temperature of the gas and one can estimate the mass assuming (again) that the cluster is in virial equilibrium. Weak gravitational lensing also permits one to derive the mass of the cluster. When light from farther away galaxies passes near a cluster, gravity generates a slight apparent distortion in the shape of the galaxies. One can derive the mass of the cluster by measuring the average shape of background galaxies in different annuli around the cluster and measure the surface mass profile of the cluster. In addition, the richness

[5]The galaxy cluster mass function defines the mean number density of galaxy clusters as a function of its mass.

of a cluster, given the number of galaxies inside the cluster, can be used as a measure of the cluster mass once a richness-mass relation is calibrated, e.g. by weak lensing measurements (Rozo et al. 2009).

Since galaxy clusters are assumed to be the high peaks of the initial density perturbations on comoving scales of 10 Mpc/h, their abundance is sensitive to the amplitude of the fluctuation power spectrum on those scales. Press & Schechter 1974 showed that the number density of galaxy clusters of a given mass provides a constraint on $\sigma_8\Omega_m^\gamma$, where σ_8 is the r.m.s. fluctuation amplitude within a sphere of 8 Mpc/h radius, Ω_m is the matter density parameter, and the exact value of the exponent γ depends on the mass range of the haloes that are considered.

1.3.5 The two-point correlation function and the power spectrum of matter

The two-point correlation function and the power spectrum are two complementary statistical methods that describe the spatial distribution of the matter density of the universe. The two-point correlation function is used to describe the clustering of galaxies. It is defined as the excess probability of finding a galaxy at some distance away from another galaxy. Another way of thinking about it is how different is the spatial distribution of galaxies from that if galaxies were random distributed. If the average number density of galaxies is $\bar{n}$, the probability of finding a galaxy in the volume element dV around $\bar{\mathbf{x}}$ is then (following Schneider, P. 2006):

$$P_1 = \bar{n}dV, \tag{1.5}$$

independent of $\bar{\mathbf{x}}$ if we assume that the universe is statistically homogeneous.

The probability of finding a galaxy in the volume element dV at a location $\bar{\mathbf{x}}$ and at the same time finding another galaxy in the volume element dV at location $\bar{\mathbf{y}}$ is then:

$$P_2 = (\bar{n}dV)^2 \left[1 + \xi_g(\bar{x}, \bar{y})\right] \tag{1.6}$$

Equation 1.6 defines the two-point correlation function or autocorrelation of galaxies $\xi_g(\bar{x}, \bar{y})$.

The two-point correlation function when dealing with variations in the density field, $\delta(x) = (\rho(x)/\bar{\rho}) - 1$ (where $\rho(x)$ is the density at position x, $\bar{\rho}$ is the average density field), is defined as:

$$\xi(r) = \langle \delta(x)\delta(x+r) \rangle \tag{1.7}$$

There exist several methods to statistically estimate $\xi(r)$. The idea is to count the number of pairs of galaxies at a given distance in a box of a given size. One of the problems with these techniques is the underestimation of pairs for those galaxies placed near the boundary of the box. The simplest method, proposed by Peebles & Hauser 1974 to correct for that problem, is given by the expression:

$$\hat{\xi}_g(r) = \frac{DD(r)}{RR(r)} - 1 \tag{1.8}$$

where $DD(r)$ is the number of pairs of galaxies and $RR(r)$ is the number of pairs of random points (placed in the same volume as the galaxies), each in the range $[r, r + dr]$.

Davis & Peebles 1983 presented another improvement for the estimator using

$$\hat{\xi}(r) = \frac{DD(r)}{DR(r)} - 1 \tag{1.9}$$

where $DR(r)$ is the number of pairs between galaxies and random points.

Other estimators were introduced by Hamilton 1993 and Landy & Szalay 1993:

$$\hat{\xi}_{Ham}(r) = \frac{DD(r)RR(r)}{(DR(r))^2} - 1 \tag{1.10}$$

$$\hat{\xi}_{LS}(r) = \frac{DD(r) - 2DR(r) + RR(r)}{RR(r)} - 1 \tag{1.11}$$

Kerscher et al. 2000 made a fair comparison of nine of the most important estimators known for the two-point correlation function. They recommend the Landy & Szalay estimator in comparison with the other indicators. At small scales all the estimators are comparable.

An alternative (an equivalent) method to describe the statistical properties of the structure of the universe is the power spectrum, $P(k)$. It describes the power of fluctuations on different scales in k-space. To relate $\xi(r)$ to $P(k)$, first of all we define the Fourier transform pair for $\delta(r)$ (following Longair 1998):

$$\delta(x) = \frac{V}{(2\pi)^3} \int \delta_k e^{-ikx} d^3k \tag{1.12}$$

$$\delta_k = \frac{1}{V} \int \delta(x) e^{ikx} d^3x \tag{1.13}$$

Using Parseval's theorem to relate the integrals of squares of $\delta(x)$ and its Fourier transform δ_k:

$$\frac{1}{V} \int \delta^2(x) d^3x = \frac{V}{(2\pi)^3} \int |\delta_k|^2 d^3k \tag{1.14}$$

$|\delta_k|^2$ is the power spectrum of the fluctuations, $P(k) = |\delta_k|^2$, therefore:

$$\left\langle \delta_k^2 \right\rangle = \frac{V}{(2\pi)^3} \int |\delta_k|^2 d^3k = \frac{V}{(2\pi)^3} \int P(k) d^3k \tag{1.15}$$

The two-point correlation function is spherically symmetric so the volume element of k-space can be written as $d^3k = 4\pi k^2 dk$,

$$\left\langle \delta_k^2 \right\rangle = \frac{V}{(2\pi^2)} \int |\delta_k|^2 k^2 dk = \frac{V}{(2\pi^2)} \int P(k) k^2 dk \tag{1.16}$$

The density field can be expressed as a sum via:

$$\delta(x) = \sum_k \delta_k e^{ikx} \tag{1.17}$$

Since $\delta(x)$ is a real function, we can write $|\delta(x)|^2 = |\delta(x)\delta^*(x)|$ where $\delta^*(x)$ is the complex conjugate of $\delta(x)$. From eq. 1.7 and taking into account that cross terms all vanish in the summation except for those for which $k = k'$,

$$\xi(r) = \left\langle \sum_k \sum_{k'} \delta_k \delta^*_{k'} e^{-i(k-k')x} e^{ik'x} \right\rangle = \sum |\delta_k|^2 e^{ikr} \tag{1.18}$$

We now convert the summation into a Fourier integral:

$$\xi(r) = \frac{V}{(2\pi)^3} \int |\delta_k|^2 e^{ikr} d^3k \tag{1.19}$$

Finally, since $\xi(r)$ is a real function, the two-point correlation function is spherically symmetric, and defining the power spectrum as $P(k) = |\delta_k|^2$ we obtain the final relation between the two-point correlation function and the power spectrum of the matter density fluctuations:

$$\xi(r) = \frac{V}{(2\pi)^3} \int P(k) \frac{sin(kr)}{kr} 4\pi k^2 dk \tag{1.20}$$

Therefore, one is able to determine the power spectrum of a sample of galaxies by deriving the two-point correlation function of the galaxies. In the framework of Cold Dark Matter models (CDM), the power spectrum of density fluctuations can be predicted under several assumptions (Newtonian framework and linear perturbation theory in the matter dominated era of the universe) except for its normalization, which has to be measured empirically. In order to estimate the matter power spectrum of the universe from the galaxy power spectrum it is necessary to assume a relation between (dark) matter and galaxies. If galaxies trace the distribution of dark matter, the power spectrum can be determined from the galaxy distribution. Since the formation and evolution of galaxies is not yet a well understood process, the simplest relation between galaxies and dark matter is a linear bias factor:

$$\delta_g \equiv \frac{\Delta n}{\bar{n}} = b \frac{\Delta \rho}{\bar{\rho}} = b\delta \tag{1.21}$$

The bias factor is the ratio of the relative overdensities of galaxies to dark matter. Such a linear relation is not justified by theory, but it is a reasonable assumption on scales where the density field is linear. The bias factor should depend on redshift, galaxy type and the scales that are considered.

If one assumes that galaxy bias does not depend on the scale, the shape of the dark matter power spectrum can be determined from the galaxy power spectrum. The shape depends on the shape parameter $\Gamma = h\Omega_m$ (in the CDM models framework). One can compare the shape of the power spectrum of galaxies with that of the CDM models and obtain a value for Γ.

1.3.6 Baryon acoustic oscillations

The same oscillatory structure found in the angular power spectrum of the CMB is also present in the power spectrum of the matter distribution (with a much smaller

amplitude). Before the universe became neutral, baryons, photons and electrons were coupled. The oscillations of this plasma due to gravity and radiation pressure travelled as sound waves. At the time of recombination the oscillations of the plasma stopped. The sound waves provide a feature of known size in the structure of matter and galaxies. The fundamental mode of the sound waves of the oscillations was imprinted on the matter distribution and is now visible as a bump in the correlation function or as tiny wiggles (called baryon acoustic oscillations (BAO)) in the galaxy power spectrum. This fundamental mode is given by the sound horizon at recombination ($z \approx 1100$):

$$s = \int_0^{t_{rec}} \frac{c_s}{H(z)} dz \sim 150\text{Mpc} \tag{1.22}$$

where c_s is the sound speed of the baryon-photon fluid, t_{rec} is the time of recombination and $H(z)$ is the Hubble expansion rate.

This fundamental mode has been detected in the large-scale correlation function of galaxies (Eisenstein et al. 2005, Cole et al. 2005). Figure 1.5 shows the two-point correlation function measured from a spectroscopic sample of luminous red galaxies from the SDSS. There is a peak at 100 Mpc/h which agrees perfectly with the predicted shape and location of the imprint of the recombination-epoch acoustic oscillations on the low-redshift clustering of matter. Gaztanaga & Cabre 2008 studied the bidimensional galaxy correlation function $\xi(\pi, \sigma)$, separated in perpendicular σ and line-of-sight π distances of 75000 LRG galaxies. They find a significant detection of the BAO peak at r $\sim$110 Mpc/h. Figure 1.6 shows the bidimensional galaxy correlation function, $\xi(\pi, \sigma)$, where the BAO peak appears as a circular ring.

In order to explain how standard rulers[6] can be used to constrain cosmological parameters, first it is useful to define some common cosmological distance definitions.

The line-of-sight comoving distance, χ, is defined by the expansion history of the universe. It is the distance between two objects that remains constant with the expansion of the universe,

$$\chi = \int_t^{t_0} \frac{cdt'}{a(t')} = \int_a^1 \frac{cda'}{a'^2 H(a')} = \int_0^z \frac{cdz'}{H(z')} \tag{1.23}$$

where $a(t)$ is the scale factor of the universe as a function of time, c is the speed of light in a vacuum and H(z) is the Hubble expansion rate.

The angular diameter distance, d_A, is defined as the ratio of the transverse physical size of an object to its angular size in radians:

$$d_A = \frac{l}{\theta} = \frac{D_M}{1+z} \tag{1.24}$$

where D_M is the transverse comoving distance, which is also defined as the distance separating two objects at the same redshift but separated on the sky:

$$D_M = \begin{cases} \frac{c}{H_0\sqrt{\Omega_k}} \sinh\left[\sqrt{\Omega_k}\frac{\chi H_0}{c}\right] & \text{for } \Omega_k > 0 \\ \chi & \text{for } \Omega_k = 0 \\ \frac{c}{H_0\sqrt{\Omega_k}} \sin\left[\sqrt{|\Omega_k|}\frac{\chi H_0}{c}\right] & \text{for } \Omega_k < 0 \end{cases} \tag{1.25}$$

[6]Standard rulers are astronomical objects or features whose approximate size is known.

where Ω_k is the spatial curvature density. The luminosity distance, d_L, is defined as:

$$d_L \equiv \sqrt{\frac{L}{4\pi S}} \tag{1.26}$$

where L is the bolometric luminosity[7] and S is the bolometric flux. d_L can also be expressed as a function of χ or d_A:

$$d_L = (1+z)\chi = (1+z)^2 d_A \tag{1.27}$$

Standard rulers, (χ), can be used to derive the angular diameter distance, d_A, by using the subtended angle by the ruler, θ:

$$\theta = \frac{\chi}{d_A(z)} \qquad d_A(z) = \frac{d_L(z)}{(1+z)^2} = \frac{1}{1+z}\int_0^z \frac{dz'}{H(z')} \tag{1.28}$$

By measuring the redshift interval (Δz) associated with this distance one can map out the Hubble parameter, $H(z)$:

$$c\Delta z = H(z)\chi \tag{1.29}$$

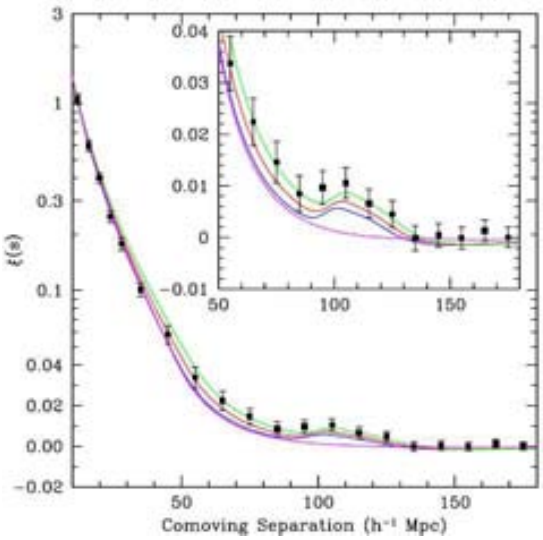

Figure 1.5: Two-point correlation function in redshift-space of the SDSS LRG sample. Vertical axis mixes logarithmic and linear scalings. The inset panel shows a zoom view with linear vertical axis in order to better note the interest region where BAO peak appears. Coloured lines represent different cosmological models: $\Omega_m h^2 = 0.12$ (top green), 0.13 (red), 0.14 (bottom blue, with peak). Magenta line represents a pure CDM model ($\Omega_m h^2 = 0.105$) with no acoustic peak. The bump at 100 Mpc/h scale is statistically significant. Figure from Eisenstein et al. 2005.

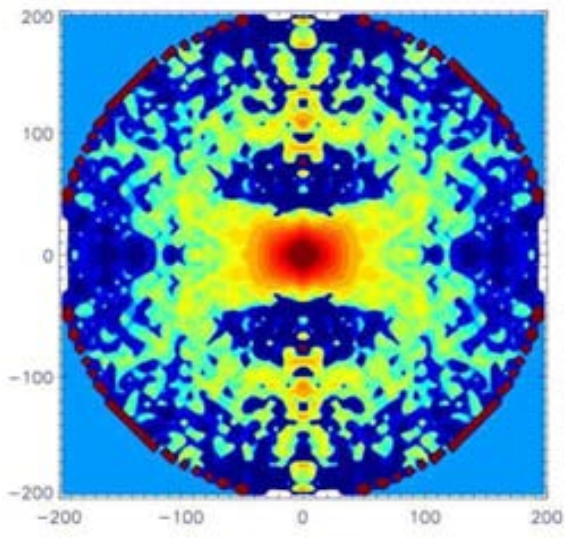

Figure 1.6: Two-point correlation function from LRG SDSS measurements for a slice at $z = 0.15 - 0.30$. The vertical axes shows radial direction, π, while the horizontal axes shows the transverse direction σ_p. Data shows a prominent BAO ring at radius of about 110 Mpc/h Figure from Gaztanaga & Cabre 2008.

1.3.7 Redshift space distortions

Redshift space distortions (RSD) are caused by the peculiar velocities of galaxies not associated to the Hubble flow. When measuring galaxy redshifts, the velocity of a galaxy is the superposition of the Hubble flow velocity and its peculiar velocity v along the line-of-sight. Deviations from the Hubble flow are caused by local density fields, and such fields are in turn generated by local density fluctuations. The observed redshift of a galaxy is therefore given by the expression:

$$cz = H_0 d + v \tag{1.30}$$

One expects that the correlation function of galaxies is not isotropic in redshift space, however the angular position of the galaxies is not affected by the peculiar velocity.

There are two important contributions to the radial displacements caused by peculiar velocities. The first one, on large scales, is due to coherent bulk motion. One can see bigger voids and denser walls that produce a squashing effect along the line-of-sight in the two-point correlation function. This effect is known as the Kaiser effect (Kaiser 1987). The second, at small scales, is produced by random velocities inside clusters and groups of galaxies which create a radial stretching. This effect is known as the Fingers-of-God. When one is computing the correlation function, $\xi(r)$, in redshift space it is common to decompose distance between galaxy pairs, r,

[7] The bolometric luminosity or magnitude refers to the luminosity of an object taking into account all electromagnetic wavelengths.

into two components, a line-of-sight component, π, and a transverse component, σ_p. The transverse component, σ_p, is a true measure of distance while π is distorted.

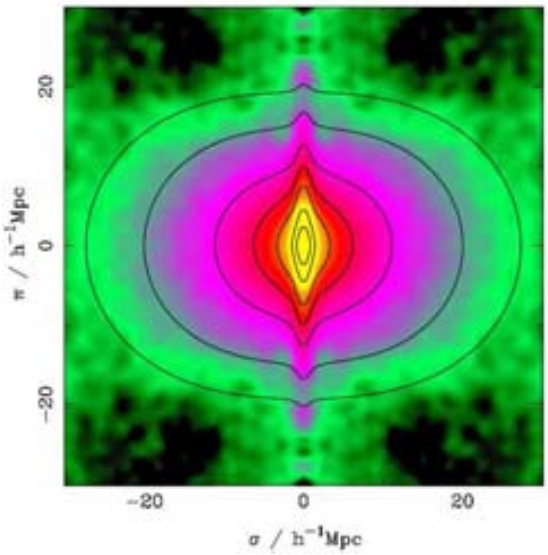

Figure 1.7: Two-point correlation function from 2dFGRS. The vertical axes shows radial direction, π, while the horizontal axes shows the transverse direction σ. Figure from Peacock et al. 2001.

Developing the fluctuations of the density field using linear theory one can get to an expression that relates the velocity field and the density field (with some approximations for the dependence on Ω_m and Ω_Λ):

$$\mathbf{u}(\mathbf{x},t) = \frac{\Omega_m^{0.6}}{4\pi} aH(a) \int d^3y \delta(\mathbf{y},t) \frac{\mathbf{y}-\mathbf{x}}{|\mathbf{y}-\mathbf{x}|^3} \tag{1.31}$$

Equation 1.31 shows that the velocity field can be derived from the density field. At $z \sim 0$ and defining the factor $\beta \equiv \frac{\Omega_m^{0.6}}{b}$, where b is the linear bias factor that relates the galaxy density field to the matter density field, one can get to the equation:

$$\mathbf{u}(\mathbf{x}) = \beta \frac{H_0}{4\pi} \int d^3y \delta_g(\mathbf{y}) \frac{\mathbf{y}-\mathbf{x}}{|\mathbf{y}-\mathbf{x}|^3} \tag{1.32}$$

The anisotropy caused by peculiar velocities in the two-point correlation function is now caused by this correlation between the density field $\delta_g(\mathbf{x})$ and the velocity field $\mathbf{u}(\mathbf{x})$. If one measures the correlation function in both components, radial and transverse, the parameter β can be determined at large separations where the Kaiser effect emerges and depends directly on β. However, one has to take into account that galaxies do not perfectly follow the velocity field, instead they have a velocity dispersion, σ, at small scales due to gravitational interactions around the velocity field (as predicted by linear theory). Therefore the derived value of β is related to σ. One can model the two-point correlation function using different values of β and σ and determine both parameters.

1.4 Galaxy surveys

1.4.1 General Description

The main objective of galaxy surveys is to measure the positions of distant galaxies. Thinking naively, one would like to derive the positions of every galaxy in the universe with infinite accuracy (it could be named the Last Galaxy Survey or LGS). Being more realistic, one can attempt to get the redshifts of as many galaxies as possible in a determined region of the sky. There are, in principle, two main possibilities of deriving the redshift of a galaxy, using spectroscopy or using photometry. Spectroscopic surveys can measure galaxy redshifts with a very high precision compared to photometric surveys. On the other hand, obtaining spectra takes much more time than obtaining photometry. Moreover, in order to get the spectroscopic spectra one needs the position (in angle) of the galaxies where the spectroscopic instrument will be pointed. As in many other branches of Astronomy, redshift surveys evolve with the development of technology. In the early 1980s, with the introduction of the charge-couple devices (CCDs), optical detectors experienced a very important improvement in sensitivity and accuracy. In the beginning only one or very few galaxies were recorded simultaneously, as it was only possible to use a single slit in the spectrograph. It was also a great advance when the multi-spectrograph was introduced and the spectra of many objects (up to thousand) in the field of view of the telescope could be observed at the same time. The first redshift galaxy survey was completed in 1982 at the Center for Astrophysics (CfA), the CfA Survey. The CfA Survey measured radial velocities of about 14000 galaxies up to redshift cz∼15000 km/s. Figure 1.8 shows the first slice of CfA Survey, where galaxies actually appear to be distributed in space in a manner that is anything but random.

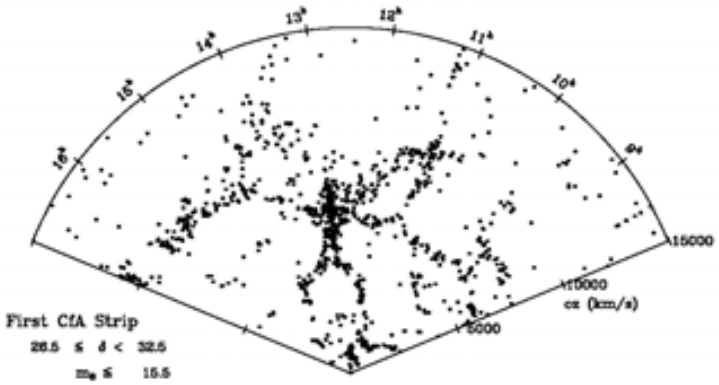

Figure 1.8: First CfA strip. Source: https://www.cfa.harvard.edu/∼dfabricant/huchra/zcat/. Credit: John Huchra.

In 1991, Las Campanas Redshift Survey (LCRS) measured the redshift of 26418 galaxies from an area of 700 square degrees in six strips of 1.5 deg × 80 deg, three each in the North and South galactic caps up to $z = 0.2$. The goal of the survey was to provide a large galaxy sample with properties such as the clustering and luminosity in the local universe. In figure 1.9, the distribution of galaxies is displayed and it shows the typical bubble structure. Because of the volume covered one can see that no structures exist with scales comparable to the extent of the survey. Using targets from the IRAS satellite, which compiled the first infrared map of the sky at 12, 25, 60 and 100 μm, various different redshift surveys such as QDOT and PSCz were performed. The difference with optical surveys is that these galaxies are nearly unaffected by Galactic absorption. The PSCz, as an "all-sky" survey, shows the galaxy distribution in a sphere around us. The Canada-France Redshift Survey (CFRS), performed at the Canada-France-Hawaii Telescope, using a multi-object spectrograph (up to 100 objects simultaneously) compiled the redshifts of over 1000 faint galaxies ($17.5 \leqslant I(AB) \leqslant 22.5$) with a median redshift of about 0.5. After the first galaxy redshift surveys, cosmologists thought about the production of bigger surveys with two main goals: extending the volume will improve the statistics of the galaxy distribution and also higher redshifts will provide more information to study galaxy evolution.

The Two-degree-Field Galaxy Redshift Survey (2dFGRS), is a spectroscopic survey that measured the spectra of a total number of 232155 galaxies between 1997 and 2002 from the 3.9m Anglo-Australian Telescope at the Anglo-Australian Observatory (AAO) using a spectrograph that can simultaneously measure 400 spectra. The area covered was about 1500 square degrees from the north and the south galactic poles taking targets sources from the previous Automated Plate Measuring catalogue (APM). Figure 1.10 shows the large scale structure, with filaments and voids, of the 2dF galaxy redshift survey. This survey measured the density parameter of non-relativistic matter. It detected the baryonic acoustic oscillations and it measured the contribution of massive neutrinos to dark matter.

The Sloan Digital Sky Survey (SDSS) is an imaging and spectroscopic survey. It was performed at the Apache Point Observatory in New Mexico using a dedicated 2.5 m optical telescope. The field of view of the camera covers 1.5 degrees of sky. It uses a new photometric system with 5 filters (u, g, r , i and z) built such that their transmission curves have a very little overlap. Using this photometric information, targets were selected to measure their spectra with a multi-object spectrograph of 640 optical fibers. During eight years of operation it obtained more than 930000 galaxies and more than 120000 quasars from over a quarter of the sky. The volume of the collected data is huge and its management and reduction required an extraordinary effort. It is one of the most influential surveys in history and it will continue operating until 2014. Among many discoveries some science highlights, which have to do with this work, made by the SDSS are: a detailed characterization of small and intermediate scale clustering of galaxies, precision measurements of large scale clustering and cosmological constraints, the detection of acoustic oscillation signatures in the clustering of galaxies, mapping extended mass distributions around galaxies with weak gravitational lensing, the demonstration of ubiquitous substructure in the outer Milky Way, etc.

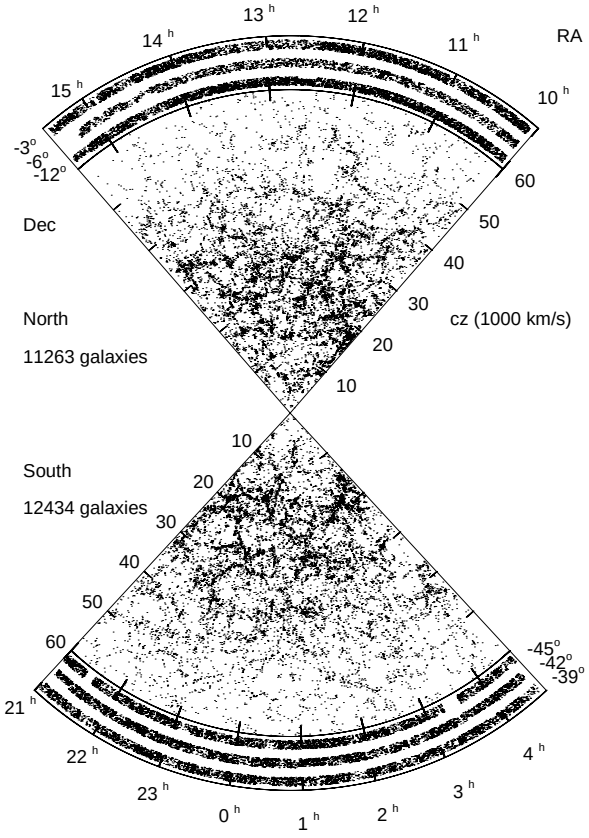

Figure 1.9: The six fields in Las Campanas Redshift Survey. Three in each Galactic pole. Each field is a strip 1.5^o wide and 80^o long. Figure from Lin et al. 1996.

Nowadays, there is a large effort to carry out large cosmological surveys to study dark energy, in particular the projects aim to measure the history of expansion and growth of structure with percent-level precision or higher. Albrecht et al. 2006 in the report of the Dark Energy Task Force (DETF) described the different experimental approaches to study dark energy and provide a quantitative framework to compare their capabilities. More recently Weinberg et al. 2012 update the report of DETF and review in detail the four most well stablished cosmological probes: Type Ia supernovae, BAO, weak gravitational lensing and the abundance of galaxy clusters. In their review one can realize the large number of projects attempting to find out what is causing the accelerating expansion.

1.4.2 DES, PAU and Euclid

This section very briefly describes three future galaxy redshift surveys in which our scientific group is involved: the Dark Energy Survey (DES), the Physics of the

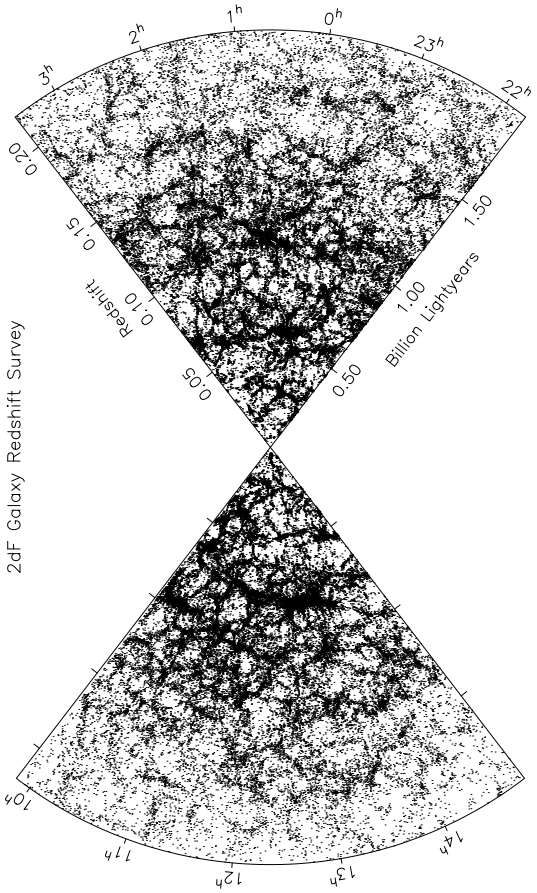

Figure 1.10: 2dF Galaxy Redshift Survey. Source: The 2dF Galaxy Redshift Survey Web Page.

Accelerated Universe project (PAU) and the Euclid project. These surveys will study the accelerating expansion of the universe besides other science that can be done with their data. Our group has already provided mock galaxy catalogues for PAU and DES projects. The idea is also to produce galaxy catalogues for Euclid.

The Dark Energy Survey's main goal is to probe the origin of the accelerating universe and understanding the nature of dark energy. It will use five filters, grizY, which cover a spectral range from the visible to the infrared. The project aims to detect around 300 million galaxies, 15000 galaxy clusters up to z$\sim$1.2 and over 1000 distant type-Ia SNe. The camera, which contains 62 CCDs, will be placed at Cerro Tololo on the Blanco telescope. The expected area covered will be 5000 $\deg^2$ composed of images of 3 $\deg^2$ on the sky. The survey will be performed over 525 nights in 5 years and the commissioning started in September 2012 and it is not finished yet. DES will combine four probes of Dark Energy: Type Ia Supernovae (SN), Baryon Acoustic Oscillations (BAO), Galaxy Clusters (GC) and Weak Gravitational Lensing (WL). Our group at the "Institut de Ciències de l'Espai" (IEEC-CSIC) is involved mainly in the photometric redshift, large scale structure and simulations science groups.

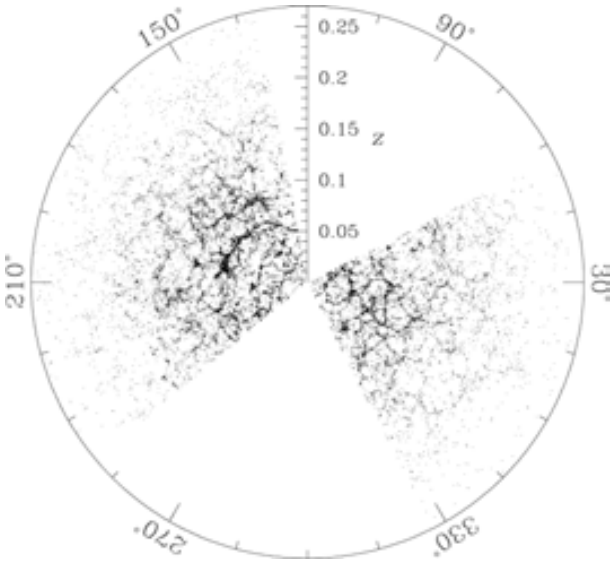

Figure 1.11: Distribution of galaxies from the SDSS redshift survey out to redshift 0.25. Source: Sloan Digital Sky Survey Web Page.

Figure 1.12: Lateral view of the inside of the PAUcam. The top piece is the Copper cold-piece that maintains the CCDs at constant and uniform temperature.

The Physics of the Accelerated Universe (PAU) is a photometric redshift survey performed to study the nature of dark energy. Its principal characteristic is the use of 42 narrow band filters that will provide very high accuracy photometric redshifts ($\sigma_z \leqslant 0.0035(1+z)$, which corresponds to $d \leqslant 10$ Mpc for the whole redshift range of the survey). The camera will cover the electromagnetic spectra from $\sim$ 4500Å up to $\sim$ 8500Å and will also have 5 broad band filters, ugriz which will cover from $\sim$ 3000Å up to $\sim$ 8575Å. The expected area will be $120\,\mathrm{deg}^2$. The camera, which contains 18 CCDs, will have a field of view of 40 arcmin without vignetting and 60 arcmin with vignetting. It will be placed at the $4.2m$ William Herschell Telescope (WHT) in La Palma. Our scientific group is involved in the whole project.

Euclid is a space mission that has been accepted as an M-class mission by the

European Space Agency (ESA). Euclid's main target is to understand the nature of dark energy and dark matter too. It will measure with very high accuracy the accelerated expansion of the universe through different independent methods such as weak gravitational lensing (WL) and baryon acoustic oscillations (BAO) by mapping the large scale distribution of the dark matter. The 1.2m space telescope is planned to be launched in 2019 and will cover $15000\,\mathrm{deg}^2$ (and a deep survey of $40\,\mathrm{deg}^2$). It will have two instruments, VIS and NISP. The former will observe galaxies in the visible range down to magnitude 24.5 for measuring weak gravitational lensing and the latter will take images in the near infrared with three bands (Y, J, H) and spectra slitless for each field of view. Our scientific group is involved in the photometric redshift, simulations and data management science groups, and it is the responsible of building the filter wheel for the near-infrared imaging.

Chapter 2

Mock Catalogues

Solving the problem of galaxy formation is one of the most exciting challenges in Astrophysics. According to the current paradigm of structure formation it is well established that galaxies are formed and reside in overdensities of the underlying dark matter field. These overdensities are called dark-matter haloes. N-body simulations and analytical models give a detailed picture of the abundance and clustering of dark-matter haloes. The link between the properties of the halo population and the galaxy population is a key ingredient in order to understand galaxy formation. Galaxies form by the cooling and condensation of gas inside these dark-matter haloes. Besides cooling, complicated physical properties such as star formation, merging, tidal interactions and several feedback processes, determine galaxy formation and its evolution. To *simulate* these processes and reproduce galaxy formation there exist several methods, which by themselves, and together with other ingredients, are used to build mock galaxy catalogues.

Mock catalogues are useful for a variety of cosmological applications such as calibrating errors or exploring systematic effects for galaxy surveys, calibrating cluster finders and photometric redshift estimators, and, of course, learning about processes driving galaxy evolution. Moreover they are useful for testing and improving the models and algorithms with which they are built. There are many different ways of building mock galaxy catalogues. This chapter is an attempt to show the state of the art of building mock galaxy catalogues and to summarize the most important techniques used to generate them. One could say that methods to build mock galaxy catalogues can be separated into the ones involving N-body simulations and the ones that do not use them. The former are more common than the latter. Methods that do not use N-body simulations avoid the problem of managing and solving the equations of motion of a huge number of particles, which is computationally expensive. One of the reasons why the majority of galaxy catalogues are built using N-body simulations is because one obtains much more information about the underlying dark matter distribution. In fact the methods that do not use N-body simulations usually use results from N-body simulations (such as the dark-matter halo density profile to place galaxies). For both methods, in most cases, the essential ingredients are a correct number density and spatial distribution of dark-matter haloes.

There are, in general, three different approaches to relate haloes with galaxies: hydrodynamical simulations, semi-analytical models, and extensions of the halo

model such as the conditional luminosity function (CLF) or the halo occupation distribution (HOD). Hydrodynamical simulations, which are another type of numerical simulations in a cosmological context, study galaxy formation and evolution. These simulations incorporate gas dynamics. The involved physics in these simulations is more complicated than in the pure gravitational case. Unlike gravity, which is a long range force, hydrodynamic mechanisms are important on small scales ($\leqslant 1$ Mpc/h), e.g. in the formation of galaxies and in linking the matter distribution of the universe to the observed light distribution. Hydrodynamical simulations are not going to be described in this work. Semi-analytical models attempt to simulate galaxy formation and evolution using simple analytic prescriptions to approximate star formation and feedback processes. We will briefly describe them in this chapter. Extensions of the halo model like the CLF as well as other methods such as the halo abundance matching (HAM), which are useful to build mock galaxy catalogues, will be discussed later in this chapter.

The chapter is structured as follows: it starts by discussing methods that, in principle, do not use N-body simulations to build galaxy catalogues followed by methods that do use N-body simulations as input. The last section shows some examples of catalogues used in the literature and in past, present and future galaxy surveys.

2.1 Mock galaxy catalogues *without* N-body simulations

Scoccimarro & Sheth 2002 cite in their conclusions some methods developed along the history of making mock galaxy catalogues without using as input N-body simulations: Soneira & Peebles 1978 show an algorithm to reproduce the observed two-point statistics at small-scales. Their method does not include velocities nor large-scale correlations. By adjusting the parameters of the model they produce a galaxy map that visually appears as a first approximation of the Lick survey. Coles & Jones 1991 study a model for the distribution of matter using a lognormal random field. Sheth & Saslaw 1994 use a technique to distribute cluster centers using a Poisson distribution and place galaxies around them following a specific recipe for the shape, the radius, the number of particles and the radial density profile of the cluster. Bond & Myers 1996 describe a rapid Monte-Carlo method to identify collapsed structures based on the peak-patch picture. This method attempts to distinguish "fast (nonlinear) internal dynamics from slower (hopefully linear) external dynamics of collapsing structures".

Another approach to model galaxy formation, and to build mock galaxy catalogues, are semi-analytic models. They need a halo merger history tree and, in principle, they do not necessary use N-body simulations. In its origin the halo merger history tree was obtained using the extended Press-Schechter prescription[1], but more recently halo merger history trees are obtained by using N-body simulations.

In this section semi-analytic models and PTHALOS method are described.

[1] It will be explain in chapter 3.

2.1.1 Semi-analytic models of galaxy formation

Semi-analytic models (SAMs) try to describe the physical processes involved in the formation and evolution of galaxies assuming that structure grows hierarchically. These models approximate the assumed physics that occur during the formation and evolution of galaxies, and they offer the possibility of predicting a wide range of properties for the galaxy population at any redshift. One of the advantages of semi-analytic models is that they include baryon physics in a fast, transparent and easy way in comparison with hydrodynamical simulations. However, in addition to the cosmological parameters, SAMs, in the most general case, introduces a limited set of free parameters because some of the processes have to be dealt empirically. These parameters can be tuned to reasonably match the selected observational properties such as the galaxy luminosity function. In the most complex models the number of parameters is so large that models are non-unique and hardly are able to make predictions. As mentioned, there exist gas dynamics simulations, which one can say that have many aspects in common with SAMs. Gas dynamics numerical simulations are helpful to refine recipes used by SAMs. Both, SAMs and gas dynamics simulations, are complementary ways of approaching the galaxy formation problem. At scales where gas simulations fail because of lack of resolution or because the physical mechanisms are not (completely) understood, SAMs are used to extend the modeling.

A summary of the galaxy formation and evolution processes within the framework of hierarchical galaxy formation may be described as follows: initial temperature perturbations observed in the CMB can be related to density perturbations which are amplified by the gravitational force and collapse and form virialized structures called dark-matter haloes. Smaller haloes form first (in all cold dark matter scenarios[2]), and bigger haloes form from the collapse of smaller haloes. Gas is collisionally heated when it collapses in the potential wells of haloes (White & Rees 1978). Inside these haloes the gas cools through atomic line transitions (depending on the properties of the gas) and collapses and forms rotationally supported discs (Fall & Efstathiou 1980, Dalcanton et al. 1997, Mo et al. 1998). Cold gas forms stars with a determined efficiency depending on the density of the gas (Kennicutt 1989). The most massive stars and SNe reheat, and maybe expel, the gas, and therefore quench subsequent star formation (Dekel & Silk 1986). Galaxy mergers occur and trigger star bursts, galaxy tidal disruption, or galaxy morphological modifications. From the star formation rate history and initial mass function[3], models of spectrophotometric evolution of the stellar population provide luminosities and spectra (Bruzual 1981). If one put all these ingredients together with the mass accretion history of dark-matter haloes, which is represented by "merger trees" (see figure 2.1), which can be described analytically (e.g. extended Press-Schechter prescription), one can obtain spectral energy distributions of stellar populations as well as many other properties of the galaxy population.

[2]In this scenario most part of matter in the universe is formed by components that do not interact with electromagnetic radiation and therefore they "move" slow (*cold*).

[3]The initial mass function gives the distribution of stellar masses formed in a burst of stellar formation (SF) in terms of the initial mass where they were formed.

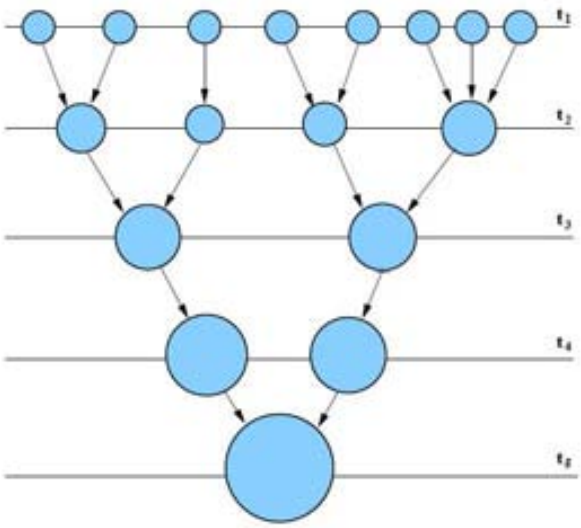

Figure 2.1: Figure from Baugh 2006. Figure shows a schematic merger tree for a dark-matter halo. The horizontal lines represent snapshots in the evolution of the history of the halo, corresponding to timesteps in an N-body simulation or Monte-Carlo realization of the merger tree ($t_1 < t_2$).

Earlier semi-analytical models failed mainly in two related aspects: they formed too many galaxies because of an overcooling/cooling problem in galaxies and clusters, and they did not reproduce the observed color-magnitude diagram. Even if new models introduce new elements to solve these problems, the approach itself still has the inconvenient for building realistic mock catalogues due to the lack of information regarding halo spatial and velocity distributions. New 'hybrid' models use merging history trees from large very high-resolution N-body simulations to include spatial and velocity of dark-matter haloes and subhaloes, and keep the usual semi-analytic approach to model physics of galaxy formation assuming that baryons do not strongly affect the dynamics of the dark matter. The first attempt in using a "hybrid" model was made by White et al. 1987 and Roukema et al. 1997 using less than 10 snaphots of a N-body simulation. Up to date many more models have been proposed including new ingredients in order to match more and different observational constraints. There are also many different codes in the literature (e.g. GALICS, GALFORM, Galacticus, etc.). The review of Baugh 2006 summarizes some of the successes and failures of the models and the areas in which future developments are most likely to be made.

2.1.2 PTHALOS

Scoccimarro & Sheth 2002 present a new method, PTHALOS, to generate mock galaxy catalogues in a faster way (orders of magnitudes) than N-body simulations. Galaxy evolution models divide the problem basically in two parts that in summary are: first, dark-matter particles evolve and form virialized dark-matter haloes and second, inside haloes, gas cools and forms stars. The first step is usually carried

out using numerical simulations and the second step uses prescriptions to mimic non-linear physics of gas cooling in gravitational potential wells to include star and galaxy formation into the simulations. Scoccimarro & Sheth 2002 argue that since the behavior of baryons inside dark-matter haloes is not completely known it is possible to obtain similar galaxy distributions without solving the equation of motion for the dark matter due to gravitational clustering. The idea is to obtain a correct number and space distribution of virialized dark-matter haloes without using the computationally expensive method used in N-body simulations. PTHALOS method combines perturbation theory with halo models of the non-linear density and velocity fields. They claimed that the point distributions the method produces reproduce the observed galaxy clustering statistics (the galaxy counts-in-cells, the power spectrum and the bispectrum). They generate a realization of the large scale dark matter density field by using the second-order perturbation theory, which reproduces correctly the two- and three-point statistics at large scales but does not reproduce correctly the correlations at small scales. They proposed an algorithm that place and partition the large scale dark matter density field created using the 2LPT into haloes and place galaxies around them by using NFW density profiles. They claimed that the two- and three-point statistics of the resulting fields are exact at large scales and rather accurate in the non-linear regime, specially for two-point statistics.

2.2 Mock galaxy catalogues with N-body simulations

Most of the present galaxy catalogues are built using N-body simulations. And most of these kind of catalogues use a halo catalogue as input. It is not important how one obtains the halo catalogue (or no matter how one obtains a halo merger history tree for semi-analytic models). One of the advantages of using N-body simulations is the knowledge of space and velocity dark matter distributions. If one is interested in producing mock galaxy catalogues with the correct clustering the essential point is to have a halo population with the correct number density and the correct clustering. Then, one can populate it with galaxies. The most extended approach to describe how galaxy clustering depends on galaxy type (i.e. what we call galaxy bias), is the halo model. If the goal is to determine the luminosity dependence of clustering there are usually three different methods: the Halo Occupation Distribution (HOD), the Conditional Luminosity Function (CLF) and the Subhalo Abundance Matching (SHAM). There are also many catalogues that do not populate haloes but they *populate* directly dark-matter particles, these catalogues are the ones that we call dark-matter mock galaxy catalogues. This section describes some methods and examples of mock galaxy catalogues that are built using N-body simulations. CLF and SHAM are discussed in the next sections. The HOD method will be described in detail in the next chapter (as well as the halo model) since it is the main method we use for generating our mock galaxy catalogues.

2.2.1 Dark matter mock galaxy catalogues

Populating N-body simulations without previously using any method to locate haloes or overdensities is another way of generating mock galaxy catalogues. The idea is to

put galaxies in the same place of dark-matter particles following some property of the galaxy population. There are actually a very large number of mock galaxy catalogues built only using dark-matter particles. In general they are built in a simple way to follow a single property. As an example, it is easy to build a galaxy catalogue that follows a given redshift distribution, $\frac{dN}{dz}$. By computing the probability of a particle to *be* a galaxy depending on its redshift, one is able to build a realization of the simulation that follows $\frac{dN}{dz}$.

Other methods, instead of locating haloes, also use directly the dark matter field. For instance the ADDGALS (Adding Density Determined GAlaxies to Lightcone Simulations) algorithm. This method relates local dark matter density to galaxy magnitudes. It was developed by Michael T. Buscha and Risa Wechsler in order to produce mock photometric galaxy catalogues. The algorithm is able to reproduce the luminosity function and the luminosity dependence of the two-point correlation function. After luminosities are assigned they add colours to galaxies in a way that satisfies the colour-environment relation observed. This method has been developed within the context of DES.

2.2.2 Subhalo / halo abundance matchting technique

The subhalo / halo abundance matching technique (SHAM) automatically fits the empirical observational constraint and provides a relation that can be checked against observations. Assuming that galaxies form from the gas falling into dark-matter haloes or subhaloes, we know that the mass of a halo is strongly correlated with some properties of galaxies, such as their luminosity, the star formation rate or the stellar mass. The relation between the galaxy property and the mass of its parent dark halo or subhalo is assumed monotonic. Using as an example the luminosity function, one usually computes the cumulative function of the luminosity function and the cumulative function of the halo mass function. Then one relates the most massive halo with the most luminous galaxy, the second most massive halo with the second most luminous galaxy and so on. Trujillo-Gomez et al. 2011a use the halo abundance matching technique (HAM) to fit, at least on average, all basic statistics of galaxies with circular velocities $V_{circ} > 80$ km s^{-1} calculated at a radius of ~ 10 kpc. Moreover they introduce a small scatter in the halo mass-luminosity relation ($\sim 0.15 - 0.2$ dex in luminosity at a given halo mass) that improves the results of the two-point correlation function of bright galaxies[4]. The most probably cause of the scatter is given by differences in mass accretion histories. Up to date this technique provides excellent agreement with a wide range of galaxy statistics, such as the two-point and three-point correlation functions (e.g. Trujillo-Gomez et al. 2011a, Nuza et al. 2012). Another ingredient in the relation is the time. Is this relation constant in time? One of the ideas is that as haloes grow hierarchically, some of the galaxy properties grow in parallel. The HOD parameters also seem to be almost constant for a fixed number density of objects up to $z \sim 1$.

[4]In our approach to generate mock galaxy catalogues we also have to introduce a scatter of 0.15 dex in luminosity in the relation between halo mass and galaxy luminosity for bright galaxies in order to better fit observed galaxy clustering.

2.2.3 Conditional Luminosity Function

The Conditional Luminosity Function (CLF) is another technique to populate dark-matter haloes with galaxies. It is an statistical approach and one could say that it is an extension of the so called halo occupation distribution (HOD). The HOD model quantifies the relation between dark-matter haloes and galaxies. The HOD model gives the probability of having N_{gal} (with some specific property) in a dark-matter halo with mass M_h, $P(N_{gal}|M_h)$. $P(N_{gal}|M_h)$ can be constraint by using clustering observational data. Yang et al. 2003 and van den Bosch et al. 2003 extended the approach and presented the CLF method. The CLF technique gives the number of galaxies that resides in a halo of mass M_h as a function of luminosity. The conditional luminosity function, $\Phi(L|M_h)$, is the average number of galaxies with luminosities in the range $L \pm dL/2$ as a function of the halo mass. Kravtsov et al. 2004 studied the HOD and two-point correlation function of galaxy-size dark-matter haloes using high-resolution dissipationless simulations and they found that the probability for a halo of mass , M_h, to host a number of subhaloes, N_{sh}, is similar to that found in semi-analytic and N-body + gas dynamics studies. They showed that the HOD can be thought as a combination of the probability for a halo of mass, M_h, to host a central galaxy, N_{cen}, and the probability to host a given number of satellite galaxies, N_{sat}. This idea, in the context of CLF, is expressed as follows:

$$\Phi(L|M_h) = \Phi_{cen}(L|M_h) + \Phi_{sat}(L|M_h) \tag{2.1}$$

where $\Phi_{cen}(L|M_h)$ an $\Phi_{sat}(L|M_h)$ are the number of central galaxies and satellite galaxies as a function of luminosity respectively. Using the CLF one can relate the luminosity function of galaxies, $\Phi(L)$, and the halo mass function, $\frac{dn}{dM_h}$:

$$\Phi(L) = \int_0^\infty \Phi(L|M_h)\frac{dn}{dM_h}dM_h \tag{2.2}$$

One can put constraints in $\Phi(L|M_h)$ with equation 2.2 but there is an infinite number of functions that follow 2.2. Additional constraints are necessary in order to remove degeneracies. Data from luminosity dependence of galaxy clustering are commonly used for this purpose.

2.3 Examples of mock galaxy catalogues

This section is a compilation of some catalogues. It is an attempt to show the huge number of different mock galaxy catalogues that exist in the literature but it is by no means complete. Some of them are publicly available.

Cole et al. 1998 present mock galaxy catalogues for the 2dF and SDSS galaxy redshift surveys using N-body simulations that cover a wide range of cosmologies. They reproduce the radial selection functions and the geometry of the surveys and also approximately reproduce clustering at small scales ($1 - 10$ Mpc/h).

Coil et al. 2001 build mock galaxy catalogues to study the clustering within the DEEP2 survey volume using the GIF Virgo Consortium simulations. They use semi-analytic models (Kauffmann et al. 1999) and also produce mocks by introducing some galaxy bias using directly dark-matter particles following Yoshida et al. 2001.

Yang et al. 2004 build mock galaxy redshift surveys for 2dFGRS. They follow the CLF model established by Yang et al. 2003 and van den Bosch et al. 2003 and use a set of simulations with $N = 512^3$ particles each.

In 2005 the Millenium Simulation project produces the largest ever simulation of structure formation and generates mock galaxy catalogues using 2 different semi-analytical models (De Lucia & Blaizot 2007 and Bower et al. 2006). In 2008 the projects also generates a second simulation with 125 higher mass resolution and, in 2010, makes public new mock galaxy catalogues using an extended semi-analytical model (Guo et al. 2011).

Croton et al. 2006 using lightcones of the Millenium simulation (Springel 2005) and a semi-analytic galaxy model produce SDSS and DEEP2/AEGIS mock galaxy and halo catalogues.

Li et al. 2007 present the algorithm to build mock galaxy catalogues using the semi-analytic models and compare them to SDSS data. They also build mock galaxy catalogues for the LAMOST project.

Huff et al. 2007 study baryon oscillations using a simple HOD prescription populating N-body simulations.

Cai et al. 2009 describe a method to build mock galaxy catalogues using semi-analytic galaxy formation model of Bower et al. 2006 implemented in the Millenium N-body simulation.

Jouvel et al. 2009 build realistic galaxy spectro-photometric catalogs for future dark energy space missions. The catalogues are built in a way that they do not use N-body simulations. They produce two different types of catalogues. Ones that fit observed luminosity functions and others that are based on the galaxy distribution of the COSMOS survey.

Cabré & Gaztañaga 2011 produce mock galaxy catalogues to study BAO using the Marenostrum Institut de Ciències de l'Espai (MICE) cosmological simulations assigning directly galaxies to dark-matter particles in order to cover similar volume, density and bias as the real data.

Las Damas project is a set of large dark-matter N-body simulations to follow the evolution of large scale structure. They are designed to model the clustering of the SDSS galaxies in a wide luminosity range. They use a HOD prescription to populate them. They have several mocks publicly available.

Font-Ribera et al. 2012 simulate Lyα absorption spectra, that can be used to measure large scale structure. They present a new efficient method to generate lines of sights to the survey sources only.

Manera et al. 2012 produce 600 mock galaxy catalogues using a HOD prescription populating a halo population generated using PTHALOS algorithm. The catalogues are generated to compute covariance matrices of large-scale clustering and test the methods of analysis for the Baryon Oscillation Spectroscopic Survey (BOSS).

Our scientific group also produces mock galaxy catalogues using MICE N-body simulations. Recently, we have produced the DES-MICE galaxy catalogue. It is constructed using HOD prescriptions by populating the halo catalogue extracted from the MICE N4096L3072 dark-matter simulation lightcone output. It contains more than 100 million galaxies in one octant of the sky. We include in the information of each galaxy, in addition to the positions, magnitudes and colours, the lensing

information also extracted from the MICE simulations. The catalogue is available to the DES collaboration.

2.4 Summary

The importance of generating mock galaxy catalogues and their profits in cosmology have been presented. Besides being a key tool to understand the galaxy formation and evolution processes, the mock galaxy catalogues are essential for galaxy surveys: calibrating errors and exploring systematic effects, calibrating photometric redshift estimators, etc. The most important methods to produce mock galaxy catalogues, either using N-body simulations or not, have been described. The idea of having as input a correct and fair halo population has been stressed. Using or not N-body simulations mainly depends on the information one wants the mock galaxy catalogue to contain. If one uses N-body simulations to produce the halo catalogue, the output mock galaxy catalogue will contain the clustering information. If one does not use N-body simulations, the mock galaxy catalogue will be limited in the galaxy clustering information. Finding out another way of including the galaxy clustering information is not an easy task. We have also described two of the three common different methods in order to determine the luminosity dependence of galaxy clustering, which are the conditional luminosity function (CLF) and the subhalo abundance matching (SHAM). The third method, the halo occupation distribution model will be analyzed in the next chapter. Finally, a compilation of some mock galaxy catalogues, which are built using different methods, is listed.

Chapter 3

Halo Occupation Distribution model (HOD)

This chapter describes the HOD approach, which is the theoretical framework on which the mock galaxy catalogues presented in this work are based. The chapter is structured in 2 main sections. The first one is a description of the halo model, which is a previous and necessary concept to introduce the HOD model. The motivation of the halo model was born in order to explain the observed power spectrum and correlation function of galaxies. A probe that was not described well by the dark matter models. They predict a power spectrum which is too steep at small scales and have features at intermediate scales that enter in contradiction with the observed galaxy power spectrum that follows almost perfectly a power-law. Three common different approaches are used to connect galaxies with dark-matter haloes. (i) The semianalytic models (SAMs) try to describe galaxy formation and evolution using simple prescriptions to describe star formation and feedback processes (e.g. White & Rees 1978, White & Frenk 1991, Kauffmann et al. 1993, Cole et al. 2000, Hatton et al. 2003, Springel 2005, Somerville et al. 2008, Benson 2012). They use dark-matter simulations to follow the evolution of dark-matter haloes and they populate them with galaxies using recipes depending on their formation history. Many of the physical processes are still poorly understood which causes large uncertainties in various recipes used to follow the formation and evolution of galaxies. It is not computationally expensive but there are many degeneracies in the SAMs parameters. (ii) N-body simulations with hydro-dynamics also try to understand qualitatively galaxy properties from first principles but it is computationally too expensive to develop cosmological simulations with enough resolution and volume (Katz et al. 1992, Evrard et al. 1994, Frenk et al. 1996, Weinberg et al. 1997, Springel & Hernquist 2003, Springel 2005, and for a review, e.g. Springel 2010). (iii) The basic halo model of large-scale distribution of galaxies, which is discussed later in more detail, describes galaxy formation and evolution and is a statistical approach that links galaxies with their host dark-matter haloes.

The second section treats the HOD framework. It is not very easy to put the limit between whether one is talking about the halo model or the halo occupation distribution model. One could say that the HOD approach is a logical extension, or

a component, of the halo model when one wants to relate dark matter and baryonic matter. In order to build mock galaxy catalogues from N-body dark-matter simulations one needs "something" to relate dark matter and galaxies, and the HOD model is one of the methods that deals with this relation. Following Martínez & Saar 2002, there are several components necessary for describing dark matter clustering using the halo model: (1) a population of virialized dark-matter haloes which samples the large-scale linear density field (described by its power spectrum, $P_{lin}(k)$), (2) a mass density profile of the haloes, which is usually taken from simulations (e.g. NFW and M99 density profiles which are briefly discussed later), (3) a concentration parameter which defines how concentrated is the mass inside the halo and depends on its mass, (4) the halo mass function, which defines the average number density of haloes of a given mass and (5), in order to know what is the relation between the underlying dark matter distribution and dark-matter haloes, the halo bias function, b_h.

But what we actually can observe is the large scale galaxy distribution, so if one wants to compare predictions to observations one would need some other ingredients, which could be defined as the Halo Occupation Distribution model (HOD). This relation between the galaxy and matter distributions (which can be called as a whole as "galaxy bias") is fully defined by: (6) the probability distribution, $P(N_g|M_h)$, that a halo of mass M_h contains N_g galaxies, (7) a relation between the galaxy and dark matter spatial distribution and (8) a relation between the galaxy and dark matter velocity distribution. All these ingredients will be treated in this chapter of the thesis. They are necessary elements to build the mock galaxy catalogues if one follows the HOD model.

3.1 Halo model

This section is mainly based on the review "Halo models of large scale structure" wrote by Cooray & Sheth 2002. It describes the halo model approach, which is a formalism to describe the non-linear gravitational clustering. The first work on the idea of describing the statistics of large scale structure of dark matter like the halo model was developed by Scherrer & Bertschinger 1991. Description of different elements of the halo model and its relation with baryon matter can be found in several papers (Peacock & Smith 2000, Ma & Fry 2000, Seljak 2000, Scoccimarro et al. 2001, Sheth et al. 2001a, Sheth et al. 2001b, Cooray & Sheth 2002, Berlind & Weinberg 2002, Yang et al. 2003, Tasitsiomi et al. 2004, Mandelbaum et al. 2005, Mandelbaum et al. 2006, Yoo et al. 2006, Skibba & Sheth 2009, Cacciato et al. 2009, Leauthaud et al. 2011). The main idea of the approach is that all the matter in the Universe is contained in virialized dark-matter haloes. The distribution of matter and the distribution of velocities inside the haloes with the number and spatial distribution of these dark-matter haloes are used to estimate the statistics of the large scale structure of the density and velocity fields as a consequence of the non-linear gravitational clustering. The halo model approach has the means for modeling and interpreting many different observational probes such as the large and small scale distribution of galaxies, which constrain the dark matter distribution, the weak gravitational lensing, which provides a direct detection of the dark matter

density field, the distribution of the pressure on large scales given by the Sunyaev-Zel'dovich effect, and the signature on small angular fluctuations on the CMB of the density, velocity, momentum and pressure fields of the dark or/and baryonic matter.

As mention above, what one can observe is the spatial distribution of galaxies rather than dark matter. The first work on trying to describe the spatial distribution of galaxies was made by Neyman & Scott 1952 (similar works by McClelland & Silk 1977 and Peebles 1974 to describe galaxy clustering as superposition of randomly distributed clusters with specified profiles and a range of masses). They built a theory based on four main assumptions: (i) galaxies occur only in clusters; (ii) the number of galaxies varies from cluster to cluster, subject to a probabilistic law; (iii) the distribution of galaxies within a cluster is also subject to a probabilistic law; and (iv) the distribution of cluster centers in space is subject to a probabilistic law described as quasi-uniform. So in order to study the statistical properties they needed the distribution of cluster sizes, the spatial distribution of clusters and the spatial distribution of galaxies inside these clusters of galaxies[1]. At that time there were no data to infer this information. Nowadays it is assumed that most of the matter is dark and is homogeneously and rather smoothly distributed at large scales. The idea is that galaxies trace the underlying dark matter distribution. This *complicated* relation between dark matter and galaxies is called "galaxy bias", b_g. The HOD formalism describes the "bias" relation between galaxies and mass at the level of individual dark-matter haloes.

Linear and higher order perturbation theory description of the gravitational evolution of density perturbations (of dark matter) from Gaussian initial conditions (there is evidence of thinking that initial perturbations were not far from a Gaussian distribution, Guth & Pi 1982, Starobinsky 1982, Bardeen et al. 1983) have been developed (see the review by Bernardeau et al. 2002), but they fail when one wants to study clustering in the non-linear regime (density contrast $\delta \gtrsim 1$). Perturbation theory does not provide a rigorous framework for describing the difference between galaxy clustering and dark matter clustering.

Non-linear dark matter clustering has been studied using numerical simulations. They provide information about the distribution of dark-matter particles inside the haloes (halo density profile) using very high resolution simulations (Navarro et al. 1996, Moore et al. 1999) and they also describe the evolution of large scale structures such as filaments, voids, clusters and haloes (e. g. Jenkins et al. 2001, Crocce et al. 2010) using simulations with larger volume such as, for exemple, the Marenostrum Institut de Ciències de l'Espai simulations (MICE) (Fosalba et al. 2008) or the Hubble Volume simulations (Thomas et al. 1998). The approach where every mass in the universe is contained in haloes is becoming very interesting since the idea of galaxies are formed inside dark-matter haloes is the most reliable. White & Rees 1978 proposed a galaxy formation model in which dissipation plays a role as well as purely gravitational processes. The galaxy formation takes place in two stages, the distribution of the dominant mass component on all scales arises from purely gravitational clustering and the observed sizes and luminosity functions of galaxies are defined by gas-dynamical dissipative processes. Structure is formed in a hierarchical

[1]Galaxies were treated as discrete points in order to derive statistical properties

way, starting from little objects that merge and form bigger structures. Gas inside the resulting potential wells cools, collapses and forms galaxies.

In the halo-based description, one of the most important things to remember, is that physical properties of galaxies are completely determined by the halo in which they are formed. In the work of White & Rees 1978 galaxies of different type are formed in different haloes, so the halo model provides a natural way of explaining galaxy clustering as a function of the galaxy type. Assuming that galaxies are bias tracers of the underlying dark matter distribution one can assume that physical properties of galaxies in a volume (e.g. velocity) can tell us information about the content of dark matter in the same volume. The halo model assumes that haloes are in a state of virial equilibrium and therefore one can also inferred physical properties of galaxies for any given halo.

As mentioned before the approach considers that all the mass in the universe is contained into different entities, which are usually called dark-matter haloes, or simply haloes. If these units can be distinguished it is reasonable to assume that they are small in comparison with the common separation between them. One can then infer that the statistics of the mass density field on small scales are basically governed by the distribution of matter inside the halos. The spatial distribution of haloes is not relevant for the statistics at small scales. Similarly halo mass density profiles are not relevant at scales larger than the size of a typical halo. Therefore the most important thing on the large scale structure is the spatial distribution of the haloes. Moreover, the halo model assumes that even physics can be separated into two regimes. Physics from scales larger than the typical size of a halo are described by the linear perturbation theory. And physics inside the haloes can be inferred since the model assumes that the haloes are in a state of virial equilibrium. There is a key point in this approach, the transition from physics inside haloes and physics between haloes. This limit is given by the definition of the halo. There are several ways of defining a dark-matter halo in numerical simulation and it is not straight forward (e.g. Anderhalden & Diemand 2011). The two main definitions come from the spherical overdensity method (SO, e.g. Warren et al. 1992) and the friends-of-friends algorithm (FoF, e.g Davis et al. 1985a).

In oder to build the mock galaxy catalogues it is necessary to know the spatial distribution, the abundance, the bias and the mass density profile of haloes. These properties will be described in the next sections.

3.1.1 Spatial distribution of haloes

To visualize the two regimes, the one inside haloes and the one between haloes, one can use the two-point correlation function of dark-matter haloes. The two-point correlation function of any distribution of particles refers to the two-point autocorrelation function. It measures how the particles are distributed with respect to an hypothetical random distribution of particles, and is given by the expression:

$$\xi(r) \equiv < \delta(\mathbf{x})\delta(\mathbf{x}+r) > \tag{3.1}$$

where

$$\delta(\mathbf{x}) = \frac{\rho(\mathbf{x}) - \overline{\rho}}{\overline{\rho}} \tag{3.2}$$

and $\delta(\mathbf{x})$ is the density contrast at the position $\mathbf{x}$, and $\rho(\mathbf{x})$ and $\overline{\rho}$ are the density at the position $\mathbf{x}$ and the mean density respectively. The two-point correlation function of haloes can be divided into 2 terms:

$$\xi(r) = \xi^{1h}(r) + \xi^{2h}(r) \tag{3.3}$$

The 1-halo term takes into account contributions to the density that come from the same halo and the 2-halo term takes into account contributions to the density that come from different haloes. They can be expressed as:

$$\xi(r)^{1h} = \int dm \frac{m^2 n(m)}{\overline{\rho}^2} \int d^3\mathbf{x}\; u(x|m)\; u(|\mathbf{x}+\mathbf{r}||m) \tag{3.4}$$

$$\begin{aligned} \xi(r)^{2h} = &\int dm_1 \frac{m_1 n(m_1)}{\overline{\rho}} \int dm_2 \frac{m_2 n(m_2)}{\overline{\rho}} \cdot \\ &\cdot \int d^3\mathbf{x}_1\; u(x_1|m_1) \int d^3\mathbf{x}_2\; u(x_2|m_2)\; \xi_{hh}(|\mathbf{x}_1 - \mathbf{x}_2 + \mathbf{r}||m_1, m_2) \end{aligned} \tag{3.5}$$

where $n(m)$ is the mean number density of haloes as a function of m and $u(r|m)$ is defined as the normalized density profile of a halo as a function of distance from its center for a given halo of virial mass,

$$u(r|m) \equiv \frac{\rho(r|m)}{m}, \tag{3.6}$$

which by definition satisfies $\int_0^{r_{vir}(m)} d^3\mathbf{x} u(x|m) = 1$. According to equations 3.4 and 7.2, in order to compute the correlation function one also needs the abundance, and spatial distribution of haloes and their density profile depending on the halo mass. Eq. 3.4 is just the convolution of two similar profiles of shape $u(r|m)$, weighted by the total number of pairs contributed by halos of mass m. To derive the 2-halo term is more complicated and moreover it requires the two-point correlation function of halos of mass m_1 and m_2, $\xi_{hh}(|\mathbf{x}_1 - \mathbf{x}_2 + \mathbf{r}||m_1, m_2)$. However if one assumes that the profiles are very sharp, one can model them as delta functions, so integrals over $\mathbf{x}_1$ and $\mathbf{x}_2$ yield $\xi_{hh}(r|m_1, m_2)$. On large scales one can assume that the halo bias is linear and therefore the two-point correlation function of haloes of mass m_1 and m_2 is related to the two-point correlation function of the underlying dark matter distribution in this manner:

$$\xi_{hh}(r|m_1, m_2) \approx b(m_1)b(m_2)\xi_{DM}(r) \tag{3.7}$$

where $\xi_{DM}(r)$ is the two-point correlation function of the dark matter. Now $\xi_{DM}(r)$ can be taken outside of the integrals over m_1 and m_2, making integrals separable. Thus, the two-point correlation function of haloes at large scales (scales much larger than the typical halo) is well approximated by the two-point correlation function of the dark matter distribution, $\xi^{2h}(r) \approx \xi_{DM}(r)$. Moreover, on large scales,

$\xi_{DM} \approx \xi^{lin}(r)$, and the 2-halo term become very simple, $\xi^{2h}(r) \approx \xi^{lin}_{DM}(r)$. Using the approximation of linear halo bias, $\xi_{hh}(r|m_1, m_2) \approx b(m_1)b(m_2)\xi_{DM}(r)$, will overestimate the value of the two-point correlation function of haloes on intermediate scales. On small scales, due to the so-called "halo exclusion" (haloes are spatially exclusive, haloes are like hard small billiard balls) the halo-halo correlation function must eventually turn over. Assuming that the two-point correlation function of haloes scales like $\xi_{DM}(r)$ at small scales is a large overestimate. Nevertheless using $\xi_{hh}(r|m_1, m_2) \approx b(m_1, m_2)\xi^{lin}_{DM}(r)$, i.e., using the linear rather than the non-linear correlation function, even on the smallest scales, is a crude but convenient way of accounting for this overestimate. This assumption is not very relevant because at small scales the main contribution comes from the 1-halo term anyway.

At this moment it is convenient to say that the easiest way of working on these two-point statistics is the Fourier space. Convolutions in real-space density profiles become simple multiplications in the Fourier transforms of the halo profiles. Therefore, in Fourier-space one can write:

$$P(k) = P^{1h}(k) + P^{2h}(k) \tag{3.8}$$

where

$$P^{1h}(k) = \int dmn(m) \left(\frac{m}{\bar{\rho}}\right)^2 |u(k|m)|^2 \tag{3.9}$$

$$P^{2h}(k) = \int dm_1 n(m_1) \left(\frac{m_1}{\bar{\rho}}\right) u(k|m_1) \int dm_2 n(m_2) \left(\frac{m_2}{\bar{\rho}}\right) u(k|m_2) P_{hh}(k|m_1, m_2) \tag{3.10}$$

where $u(k|m)$ is the Fourier transform of the dark matter distribution within a halo of mass m, which for spherically symmetric profiles truncated to the virial radius is given by:

$$u(k|m) = \int_0^{r_{vir}} dr 4\pi r^2 \frac{\sin kr}{kr} \frac{\rho(r|m)}{m} \tag{3.11}$$

and $P_{hh}(k|m_1, m_2)$ is the power spectrum of haloes of mass m_1 and m_2. One can also approximate it, like discussed before for the two-point correlation function, as [2]

$$P^{hh}(k|m_1, m_2) \approx \prod_{i=1}^{2} b_i(m_i) P^{lin}(k) \tag{3.12}$$

Figure 3.1 shows the two-point correlation function for different halo mass thresholds of the halo catalogue derived from the Grand Challenge run of the MICE simulations[3]. One can realize in the figure that there are no haloes closer than a certain distance. It depends on the mass. The bigger the halo the bigger the distance. This is the already mentioned characteristic of the halo exclusion.

[2] The one-loop perturbation theory estimate of the power spectrum of dark matter may be more accurate than $P^{lin}(k)$.

[3] see chapter 4 of Cosmological Simulations.

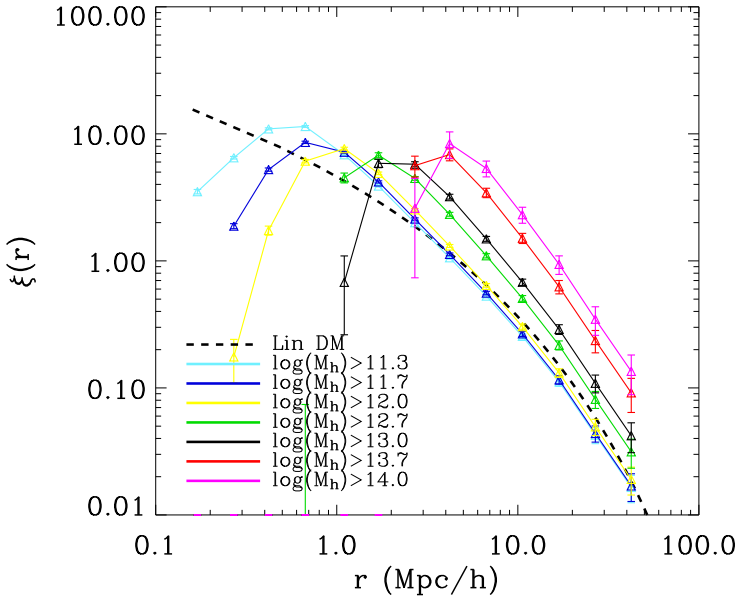

Figure 3.1: Two-point correlation function of dark-matter haloes for different halo mass thresholds of the MICE Grand Challenge run.

The halo model approach can also model higher order clustering such as the bispectrum and the trispectrum but they are not the scope of this work and they are not going to be discussed here.

3.1.2 Halo abundance

The mean number density of haloes as a function of their mass is given by the halo mass function. These haloes are formed in a hierarchical way, little objects start first and then by merging bigger structures are formed. In order to understand the structure evolution it is necessary to know the initial conditions. The amplitude of the CMB temperature fluctuations suggests that the density fluctuations at redshift $z \sim 1000$, where photons and matter interacted for the last time[4], must have had very small amplitudes. The small density perturbations increase over time due to the gravitational force; while underdense regions decrease their density contrast, overdense regions increase their density contrast over the course of time. In both cases the absolute value of the density contrast, $|\delta|$, increases.

The model of gravitational instability describes the evolution of structure in the universe. The abundance, spatial distribution and internal density profiles of haloes depend mainly on the halo mass. When the density contrast $\delta \gtrsim 1$, i.e. in the non-linear regime, perturbation theory fails when trying to describe the structure evolution. To explore the non-linear regime one can start for studying an overdense sphere. One can assume a simple model in which haloes (non-linear objects) formed

[4]It is not completely true because photons suffer several other scatters during their travel towards us.

from a spherical collapse of a initially density perturbation (first studies by Gunn & Gott 1972). The dynamics of a gravitating sphere of mass in a matter-dominated universe ($\Omega_\Lambda = 0$) of mass M and a radius R can be described by the Newtonian approach:

$$\frac{d^2R}{dt^2} = -\frac{GM(R)}{R^2} \tag{3.13}$$

If one assumes that the mass is constant over time one can integrate the equation 3.13 over time and it yields:

$$\frac{1}{2}\left(\frac{dR}{dt}\right)^2 = \frac{GM(R)}{R} + C \tag{3.14}$$

where C is a constant that defines the total energy of the sphere. Equation 3.14 is practically the same as the one of the Friedmann equations, which governs the expansion of space in homogeneous and isotropic models of the universe, because any spherical region in the universe expands at the same rate as the overall expansion factor. If $C > 0$ then the sphere will expand forever, while if $C < 0$, the sphere will expand for a certain period of time until a time (or a radius) where it will begin to contract. The solution for the latter case ($C < 0$) is given by:

$$t = A(\theta - \sin\theta) \qquad \text{and} \qquad R = B(1 - \cos\theta) \tag{3.15}$$

where θ is a parameter and A and B are constants. When $\theta = \pi$ the expansion will stop, this time (maximum expansion) is defined as the turnaround, t_{turn}, which is the point at which expansion ceases and collapse begins via:

$$\frac{t}{t_{turn}} = \frac{1}{\pi}(\theta - \sin\theta) \qquad \text{and} \qquad \frac{a(t)}{a_{turn}} = \frac{1}{2}(1 - \cos\theta) \tag{3.16}$$

where $a(t)$ is the scale factor of the universe. If one expands equation 3.16 to second order gives:

$$\frac{t}{t_{turn}} = \frac{1}{\pi}\left(\frac{\theta^3}{6} - \frac{\theta^5}{120}\right) \qquad \text{and} \qquad \frac{a(t)}{a_{turn}} = \frac{\theta^2}{4} - \frac{\theta^4}{48} \tag{3.17}$$

Taking the leading order of the left expression of equation 3.17 gives:

$$\theta^2 = \left(\frac{6\pi t}{t_{turn}}\right)^{\frac{2}{3}} \tag{3.18}$$

Substituting 3.18 back into the left expression of equation 3.17 gives:

$$\theta^2 \simeq \left(\frac{6\pi t}{t_{turn}}\right)^{\frac{2}{3}} \left[1 + \frac{1}{30}\left(\frac{6\pi t}{t_{turn}}\right)^{\frac{2}{3}}\right] \tag{3.19}$$

And substituting again 3.19 into right expression of equation 3.17 gives the linearised scale factor

$$\frac{a_{lin}(t)}{a_{turn}} \simeq \frac{1}{4}\left(\frac{6\pi t}{t_{turn}}\right)^{\frac{2}{3}} \left[1 - \frac{1}{20}\left(\frac{6\pi t}{t_{turn}}\right)^{\frac{2}{3}}\right] \tag{3.20}$$

The linear density contrast is defined as:

$$\delta_{lin} = \frac{\rho - \bar{\rho}}{\bar{\rho}} = \left(\frac{a_{back}(t)}{a_{lin}}\right)^3 - 1, \tag{3.21}$$

where a_{back} is the background evolution and:

$$\frac{a_{back}(t)}{a_{turn}} \simeq \frac{1}{4}\left(\frac{6\pi t}{t_{turn}}\right)^{\frac{2}{3}} \tag{3.22}$$

Dividing 3.20 by 3.22 one obtains:

$$\frac{a_{back}(t)}{a_{turn}} \simeq \frac{1}{4}\left(\frac{6\pi t}{t_{turn}}\right)^{\frac{2}{3}} \tag{3.23}$$

Substituting 3.23 into 3.24 one obtains the linear density contrast evolution:

$$\delta_{lin} = \frac{3}{20}\left(\frac{6\pi t}{t_{turn}}\right)^{\frac{2}{3}} \tag{3.24}$$

And therefore the density contrast at turnaround is $\delta_{lin}^{turn} = (3/20)(6\pi)^{2/3} = 1.06$. At $t = 2t_{turn}$ the collapse will occur and one finds that the density contrast at that time is $\delta_{lin}^{coll} = \delta_c = (3/20)(12\pi)^{2/3} = 1.686$. This value of the density contrast sets the perturbation threshold at which the density field collapses.

Press & Schechter 1974 developed a formalism which predicts the comoving number density of haloes of a certain mass M at redshift z, $n(M, z)$. This approach starts searching for objects of mass M by smoothing (filtering) the initial overdensity field by a filter of radius $R_f = R_f(M)$. The fraction of mass in collapsed objects more massive than some mass M is related to the fraction of regions where the smoothed initial density fluctuations are above some density threshold, $\delta(\mathbf{x}) > \delta_c$. This assumption follows from the spherical collapse solution described previously. A given point belongs to a collapsed object assuming Gaussianity of the initial overdensity field with a probability:

$$P_G(\delta > \delta_c | R_f) = \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{\delta_c}{\sqrt{2}\sigma(R_f)}\right)\right], \tag{3.25}$$

where R_f is the characteristic length of the filter function and $\sigma(R_f)$ is the linear rms in the filtered version of δ. The total probability from equation 3.25 is only equal to 1/2 and therefore Press & Schechter introduced a "fudge factor[5]" and wrote for the fraction of mass in haloes with mass greater than M:

$$F(> M_h) = 1 - \mathrm{erf}\left(\frac{\nu}{\sqrt{2}}\right) \tag{3.26}$$

where $\nu \equiv \delta_c/\sigma(M_h)$ (in the original Press-Schechter theory), and $\sigma^2(M_h)$ is the variance on the initial density fluctuation field when smoothed with a tophat filter of scale $R = (3M_h/4\pi\bar{\rho})^{1/3}$, extrapolated to the present time using linear theory:

[5]It is not clear the assumption that all the mass in the universe should be concentrated in objects.

$$\sigma^2_{lin}(M_h) \equiv \int \frac{dk}{k} \frac{k^3 P^{lin}(k)}{2\pi^2} |W(kR)|^2, \tag{3.27}$$

where $W(x) = (3/x^3)[\sin(x) - x\cos(x)]$.

The mass function, $n(M_h)$ is defined as the comoving number density of haloes of mass M_h, so:

$$n(M_h) = \frac{dn}{dM_h} = \frac{\bar{\rho}}{M_h} \left| \frac{dF}{dM_h} \right| \tag{3.28}$$

where $\bar{\rho} = 3H_0^2\Omega_m/8\pi G$ is the mean comoving density. Multiplying equation 3.28 by M_h and substituting 3.26 leads to:

$$n(M_h) = \frac{dn}{dM_h} = \frac{\rho}{M_h^2} \left| \frac{dF}{d\log M_h} \right| = \frac{\bar{\rho}}{M_h^2} \frac{d}{d\log M_h} \left[1 - \mathrm{erf}\left(\frac{\nu}{\sqrt{2}} \right) \right] \tag{3.29}$$

which leads to the standard form of the mass function:

$$n(M_h) = \sqrt{\frac{2}{\pi}} \frac{\bar{\rho}}{M_h^2} \nu \left| \frac{dn\sigma(M_h)}{d\ln M_h} \right| \exp\left(-\frac{\nu^2}{2} \right) \tag{3.30}$$

This mass function depends on the cosmological model only via the mean density $\bar{\rho}$ and on the assumed power spectrum through the density variance, $\sigma^2(M_h)$ (equation 3.27). After the analytical model proposed by Press & Schechter 1974, which described not only the abundance of dark-matter haloes as a function of mass but also the evolution, several other analytical works were developed (Bond et al. 1991, Lee & Shandarin 1998, Sheth et al. 2001c). Early comparisons with numerical experiments showed that the Press & Schechter 1974 formalism describes simulations well. The problem of the "fugded factor" was solved by Peacock & Heavens 1990 and Bond et al. 1991. In comparison with more recent numerical simulations the standard Press & Schechter 1974 mass function tends to overpredict largest haloes and underpredict smaller haloes and showed that these predictions were in general not sufficiently accurate for cosmological applications. Sheth & Tormen 1999 and Jenkins et al. 2001 by using numerical results obtained a more accurate calibration of the mass function. Specifically Sheth & Tormen 1999 found a function that fits better the observed mass function:

$$f_{ST}(\nu) = A\sqrt{\frac{\alpha\nu^2}{2\pi}} \left[1 + \frac{1}{(\alpha\nu^2)^p} \right] \exp\left(\frac{-\alpha\nu^2}{2} \right) \tag{3.31}$$

where $\alpha = 0.707$, $p = 0.3$ and A is determined by requiring that

$$\int n_{ST}(M_h) M_h dM_h = \bar{\rho} = 2 \int_0^\infty \bar{\rho} f_{ST}(\nu) d\ln\nu \tag{3.32}$$

which leads to $A = 1/\left(1 + \frac{\Gamma(0.5-p)}{2^p\pi}\right)$.

More recent calculations were done by Warren et al. 2006 and Tinker et al. 2008 and got errors in the halo mass function of a few percent. Crocce et al. calibrate the

most massive regime of the halo mass function using the MICE simulations, which are the simulations used to generate the mock galaxy catalogues presented in this work. In the next chapter of the thesis we will treat the best halo mass function fit found for the MICE simulations.

3.1.3 Halo bias

Dark matter haloes are biased tracers of the underlying dark matter distribution. The two previous sections describe two concepts, the halo abundance and the halo clustering, that are essential in the halo model. Following Sheth & Tormen 1999 the two are not independent, and this is explained in an idea that has come to be called the peak-background split (Bardeen et al. 1986, Cole & Kaiser 1989, Mo & White 1996). This idea can be visualized as follows (from Peacock 1999): one conceptually decomposes the density field into short-wavelength terms, which generate the peaks, plus terms of much longer wavelength, which modulate the peak number density (see figure 3.2). There is an environmental dependence of structure formation. The large-scale environment is often thought of as providing an effective background cosmology or say it in another way, on large scales, perturbed regions of the matter field are treated as universes with slightly different mean density and Hubble constant (see Martino & Sheth 2009 for an accurate derivation). In the Press-Schechter analytical model if a region has a density greater than a certain density threshold, δ_c, it will collapse. The effect of having two density field components can be seen in an example: where the large-scale component is $\delta = \epsilon$, the small-scale component only needs to reach $\delta = \delta_c - \epsilon$ in order to collapse. The number density of haloes is modulated by the large-scale component. In addition, the large-scale perturbation will move haloes closer together where ϵ is large, and therefore, as mentioned above, the halo abundance and the halo clustering are not independent.

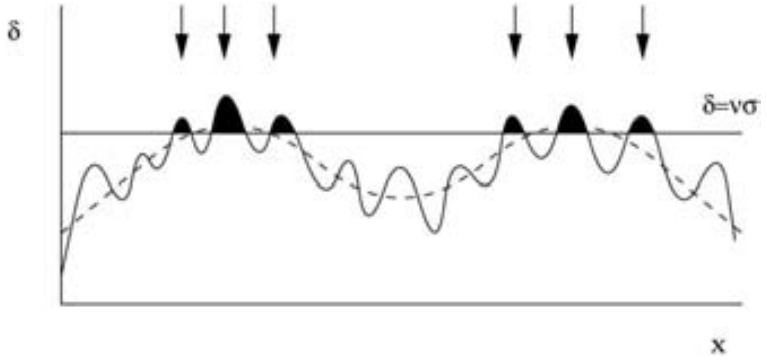

Figure 3.2: Environmental dependence of structure formation. Large- and small-scale components of the density field. Figure from Peacock 1999.

Mo & White 1996 following the spherical collapse model showed that haloes trace the dark matter field with bias, $b(M_h)$, given by:

$$b(\nu(M_h)) = 1 + \frac{\nu^2 - 1}{\delta_c} \tag{3.33}$$

Sheth & Tormen 1999 extended equation 3.34 to a more general form to:

$$b(\nu(M_h)) = 1 + \frac{a\nu^2 - 1}{\delta_c} + \frac{2p}{\delta_c[1 + (a\nu^2)^p]} \tag{3.34}$$

3.1.4 Halo density profiles

Since galaxies are assumed to be formed inside dark-matter haloes, the internal density profile of haloes is very important. Knowledge of the distribution of matter within haloes is crucial for understanding the observed interactions of galaxies with their underlying dark matter density field. Several functions for the density profile have been studied, such as:

$$\rho(r|M) = \frac{\rho_s}{(r/r_s)^\alpha(1 + r/r_s)^\beta} \qquad \text{or} \qquad \rho(r|M) = \frac{\rho_s}{(r/r_s)^\alpha\,[1 + (r/r_s)^\beta]} \tag{3.35}$$

Navarro et al. 1997 found a universal profile to describe the density profile of dark-matter haloes studying high resolution numerical simulations. Galaxies of the mock galaxy catalogues created in this work follow the NFW density profile when populating MICE cosmological simulations. Navarro et al. 1997 used the left expression of equation 3.35 with $(\alpha, \beta) = (1, 2)$. Moore et al. 1999 used $(\alpha, \beta) = (3/2, 3/2)$ in the expression on the right in equation 3.35 and it also provides very good descriptions of the density run around virialized haloes in numerical simulations. It is called M99 profile. NFW profile is described by:

$$\rho_{NFW} = \frac{\rho_s}{(r/r_s)(1 + r/r_s)^2} \tag{3.36}$$

where r_s is the inner scale radius, and ρ_s is the amplitude of the density profile. For $r << r_s$, $\rho_{NFW} \propto r^{-1}$, whereas for $r >> r_s$, $\rho_{NFW} \propto r^{-3}$. If one assumes the definition of a halo as a spherical region within which the average density is ~ 200 times the critical density[6] at the respective redshift, $\rho_{cr}(z)$, one can define the concentration parameter,

$$c \equiv \frac{r_{200}}{r_s} \tag{3.37}$$

The larger the value of c, the more concentrated is the mass towards the inner regions. If one assumes that the origin of dark-matter haloes come from peaks in the initial density field (Kaiser 1984, Hoffman & Shaham 1985), one can also assume that more massive haloes correspond to the highest density peaks. The density around small peaks is steeper than the density around higher peaks and therefore high peaks are less centrally concentrated than smaller peaks. One can generalize that to massive haloes and think that massive virialized haloes are also less concentrated than low mass haloes. Simulations show that given a halo mass the distribution of concentrations is well fitted by a log-normal distribution (Jing 2000, Bullock et al. 2001):

[6] $\rho_{cr}(z)$ is the critical density which is the averaged density required for flatness at redshift z, $\rho_{cr}(z) = 3H^2(z)/(8\pi G)$.

$$p(c|m,z)dc = \frac{d\ln c}{\sqrt{2\pi\sigma_c^2}} \exp\left[-\frac{\ln^2[c/\bar{c}(m,z)]}{2\sigma^2_{\ln c}}\right] \tag{3.38}$$

For the NFW profile,

$$\bar{c}(m,z) = \frac{9}{1+z}\left[\frac{m}{m^*(z)}\right]^{-0.13} \qquad \text{and} \qquad \sigma_{\ln c} \approx 0.25 \tag{3.39}$$

where $m_*(z)$ is the characteristic mass scale at which $\nu(m,z) = 1$. The mass of the halo is related to r_{200}, (where, again, is the radius at which the average density is ~ 200 times the critical density) by

$$M = \frac{4\pi}{3} r_{200}^3 200 \rho_{cr}(z) \tag{3.40}$$

and using the definition of the critical density one can relate the mass of the halo with the expansion history factor, $H(z)$, by:

$$M = \frac{100 r_{200}^3 H^2(z)}{G} \tag{3.41}$$

One can also define V_{200}, as the circular velocity at r_{200},

$$V_{200}^2 = \frac{GM}{r_{200}} \tag{3.42}$$

and one can express the halo mass and r_{200} as a function of V_{200}

$$M = \frac{V_{200}^3}{10GH(z)} \tag{3.43}$$

$$r_{200} = \frac{V_{200}}{10H(z)} \tag{3.44}$$

and therefore since the Hubble function increases with redshift, r_{200} at a fixed V_{200} decreases with redshift. Also, given a halo mass, r_{200} decreases with redshift, which means that haloes of a given mass are more compact at higher redshift than they are nowadays.

No good analytical reasons of why NFW profile fit dark matter density distribution of dark-matter haloes has yet been found. The main idea is that density depends mainly on its mass and also the concentration parameter (which, in turn, is also determined by the mass of the halo).

3.1.5 Halo model assumptions and drawbacks

Before getting into the details of the HOD model it is also important to summarize some of the assumptions and drawbacks of the halo model. First of all when one is following this approach, one is assuming that all the mass in the universe is contained in dark-matter haloes which is not true.

The model also assume that all the haloes have the same parameterized smooth spherical-symmetric profile which depends only on halo mass. This approximation is

not true because it is well known that haloes with the same mass can have different concentration parameters at a fixed mass[7]. Moreover, regarding the shape, haloes are not usually spherically symmetric and they are not usually smooth. Assuming a spherical averaged profile when studying the power spectrum is adequate, however the dependence of the bispectrum configuration on the spherical assumption may cause errors at the $20-30\%$ level between the halo model predictions and configuration dependence of the bispectrum in the mildly non-linear regime measured in simulations.

The contribution of the substructures on the total mass of a dark-matter halo is about 15% and it will have a contribution in the power spectrum, $P(k)$, and higher order statistics on small scales. The contribution of smoothing the profile for $k \leqslant 10k_{nonlin}$ is not important for deriving the power spectrum and the bispectrum according to studies done with simulations.

3.2 Halo Occupation Distribution

This section describes the Halo Occupation Distribution model (HOD). This formalism relates galaxy and matter distributions, or saying the same in other words, the HOD studies galaxy bias in terms of a probability distribution, $P(N_{gal}|M_h)$, that a halo of mass M_h contains N_{gal} galaxies of a given type. It also provides prescriptions to describe galaxy positions and velocities within dark-matter haloes. Mock galaxy catalogues are built in order to follow observational probes. In this work the mocks are going to match the luminosity function, the colour-magnitude diagram and the luminosity and colour dependence of clustering of SDSS data. The description of the luminosity dependence of clustering is usually done in three rather different ways, the halo occupation distribution (HOD, Jing et al. 1998, Benson et al. 2000, Seljak 2000, Scoccimarro et al. 2001, Berlind & Weinberg 2002, Bullock et al. 2002, Zheng et al. 2002, Berlind et al. 2003, Magliocchetti & Porciani 2003, Zehavi et al. 2005, Zehavi et al. 2011, Leauthaud et al. 2011), the conditional luminosity function (CLF, Peacock & Smith 2000, Yang et al. 2003, Cooray 2005, Cooray 2006, Cooray & Milosavljević 2005, van den Bosch et al. 2007, Wang et al. 2010) and the subhalo abundance matching (SHAM, Klypin et al. 1999, Kravtsov et al. 2004, Tasitsiomi et al. 2004, Vale & Ostriker 2004, Vale & Ostriker 2006, Conroy et al. 2006, Guo et al. 2010, Trujillo-Gomez et al. 2011b). As mentioned, the HOD describes the galaxy bias of a type of galaxies using the probability $P(N_{gal}|M_h)$ that a halo of virial mass M_h hosts N_{gal}. The CLF is an extension of the basic halo model. It describes the averaged number of galaxies as a function of luminosity and halo mass, $\phi(L|M_h)$ that resides in a halo of mass M_h. It tries to match the observed luminosity function by specifying how the luminosity distribution in haloes changes as a function of halo mass. One can derive the HOD from the CLF and vice-versa. The SHAM technique is used to populate dark-matter haloes with galaxies. It assumes a monotonic relation between observable properties of galaxies, such as luminosity or

[7] It is not very complicated to include the distribution of the concentration parameter as a function of the halo mass in the formalism. It will change the non-linear regime (small scales) for the two-point correlation function and will have a higher impact on the bispectrum.

stellar mass, and dynamical properties of dark matter substructures. The first step to use SHAM is to identify the substructures (subhaloes) within virialized haloes in simulations and use the subhalo properties to match the subhalo abundances to the observed distribution of luminosities. Then, one can measure CLFs or HODs in the simulation too.

Studying the relation between the galaxy and dark matter distribution offers the possibility of better understanding the physics of galaxy formation. Berlind & Weinberg 2002 studied galaxy bias in the framework of the HOD, and examine the sensitivity of different galaxy clustering statistics to properties of the HOD. They claimed that reproducing the shape of the two-point correlation function of galaxies is completely non-trivial, and that other clustering statistics provide complementary information about galaxy bias. They found that an incorrect cosmological model will not reproduce the observations for any choice of HOD.

Gasdynamics, radiative cooling and star formation can strongly influence the distribution of galaxies within haloes. Notice that the spatial distribution and the abundance of haloes are determined mainly by gravitational dynamics of dark matter. The HOD prescription can help to separate the cosmological model and the theory of galaxy formation in predictions of galaxy clustering: the cosmological model defines the properties of the halo population, whereas the theory of galaxy formation determines how galaxies populate those haloes. The three ingredients necessary to define a galaxy population from a halo population using the HOD model, the number of galaxies in each halo, their positions and velocities, will be explained in the next sections, as well as the possible physical meaning of the HOD parameters related to galaxy formation. The idea is that for a given cosmological model the HOD can tell us plenty of information a theory of galaxy formation has to say regarding statistics of clustering at every scale and in real and redshift space. So, if one can determine the HOD parameters that best match some observed galaxy clustering one can infer information about the physics of galaxy formation of that data set. Moreover, if one assumes that the content of haloes of the same mass is independent of the large scale environment (i.e. they have on average the same galaxy populations), HOD completely describes the bias between galaxies and mass[8].

In the works of Tinker et al. 2008, Tinker 2010, they claim that although there is no physics in the HOD method, it is possible to test ideas about galaxy formation. They say that in all environments, a halo of a given mass always has the same galaxies inside (with some scatter). Thus the observed change in the LF with density is a consequence of the change in the mass of dark-matter haloes, and not due to any change in galaxy formation with environment. The model does not contain a physical theory of galaxy formation, it merely posits that this theory - whatever it may be - is independent of large-scale environment at fixed halo mass. They conclude that the luminosity and color of field galaxies are determined predominantly by the mass of the halo in which they reside and have little direct dependence on the environment

[8]The assumption that haloes at fixed mass in different environments have similar properties and formation histories is expected on the basis of fairly general theoretical arguments (Lemson & Kauffmann 1999, Berlind et al. 2003, Sheth & Tormen 2004.). The correlation between galaxy properties and environment is naturally explained with the environmental dependence of the halo mass function (Berlind et al. 2005)

in which the host halo formed.

It is very important to remember that every different type of galaxy, depending for instance on its luminosity or colour or spectral type, has its own HOD. The parameters of the HOD are also known to be self-similar, thus it is possible to scale the form of the HOD between galaxy samples of different luminosities (Zehavi et al. 2005; Zheng et al. (2007)).

More recently Leauthaud et al. 2011 proposed and developed a theoretical model that describes the galaxy-dark matter connection based on the standard halo occupation distribution technique. They made an effort on combining different cosmological probes rather than just use one individually. The theoretical model combines measurements of galaxy-galaxy lensing, galaxy clustering and the galaxy stellar mass function, which will permit to explore galaxy-dark matter relation, constrain cosmological parameters and understand the nature of dark energy.

As mentioned before, the next section describes, according to the HOD model, the number of galaxies in each halo, their positions and velocities, and the galaxy two-point correlation function.

3.2.1 Number of galaxies

The first HOD prescription is the number of galaxies inside each halo. Some authors claimed that they were able to reproduce the clustering of galaxies populating haloes from N-body simulations using semianalytic models according to a predicted probability distributions of galaxies within haloes, $P(N_g|M_h)$ (e.g. Zheng et al. 2005, Benson et al. 2000, Kauffmann et al. 1997, Kauffmann et al. 1999). Hydrodynamic simulations also predicted $P(N_g|M_h)$ (e.g. Zheng et al. 2005, White et al. 2001, Yoshikawa et al. 2001). The probability distribution that a halo of mass M_h contains N_{gal} is the most important part of the HOD regarding the effect on galaxy clustering.

Berlind & Weinberg 2002 studied the physical meaning of the HOD parameters. Different physical mechanisms will influence in the formation of galaxies within virialized dark matter structures. They used two different average mean number galaxies per halo as a function of the halo mass, M_h. A power law (eq. 3.45), and a broken power law (eq. 3.46) which depends on the halo mass range:

$$N_{gal} = \begin{cases} 0 & \text{if } M_h < M_{min} \\ (M_h/M_1)^{\alpha} & \text{otherwise,} \end{cases} \tag{3.45}$$

where α is the power-law index, M_{min} is the cutoff halo mass below which haloes cannot form galaxies, and M_1 defines the amplitude of the relation and corresponds to the mass of haloes that contain, on average, just one galaxy.

$$N_{gal} = \begin{cases} 0 & \text{if } M_h < M_{min} \\ (M_h/M_1)^{\alpha} & \text{if } M_{min} \leqslant M_h \leqslant M_{crit} \\ (M_h/M_1^{'})^{\beta} & \text{otherwise,} \end{cases} \tag{3.46}$$

Parameters M_{min}, α, β and M_{crit} are related to the galaxy formation efficiency. Haloes that do not contain enough gas will not be able to host a galaxy. Therefore it

is necessary to have a minimum mass above which haloes will have the possibility of hosting, at least, one galaxy. α and β are the low- and high-mass power law indices. M_{crit} is the halo mass at which the power-law slope changes. For instance, the typical cooling time for gas increases with halo mass and therefore, $\alpha < 1$ is favored. On the contrary, hierarchical structure formation will produce at the beginning many galaxies which will end up in high-mass haloes. Many physical mechanisms could affect the average number of galaxies depending on the galaxy type (mergers, morphological transformations, etc.) and will change the number of galaxies. Comparison of the HOD set of parameters for different galaxy types can provide information about the formation process which produce such galaxy properties. Once one has the average number density of galaxies per halo, it is necessary to investigate what function will determine the probability distribution, $P(N_{gal}|M_h)$. Berlind & Weinberg 2002 studied 3 different probability density distribution functions: a Poisson distribution, a very narrow distribution where the actual number of galaxies is the integer either above or bellow N_{gal}, and a negative binomial distribution. They do not consider the possibility of changing the probability distribution with the mass. In the low-mass regime the probability density distribution will depend on the cooling and star formation rates, but in the high-mass end galaxy mergers will have a fundamental role in those haloes.

Kravtsov et al. 2004 found that in the HOD framework it is very useful to separate galaxies into contributions of central galaxies and satellite galaxies. They also found that the probability distribution for a halo of mass M_h to host a number of subhaloes N_{sub} is similar to that found in semi-analytic and N-body+gasdynamics studies. The HOD can be understood as a combination of the probability for a halo of mass M_h to host a central galaxy and the probability to host a given number N_{sat} of satellite galaxies. The probability to host a central galaxy can be approximated by a step-function while the probability of hosting satellite galaxies is well described by a simple power law $N_{sat} \propto M^{\beta}$ with $\beta \approx 1$ for a wide range of number densities, redshifts, and different power spectrum normalizations.

More recently Zehavi et al. 2011 interpret the luminosity and colour dependence of galaxy clustering in the seventh realease of the main sample of the Sloan Digital Sky Survey (SDSS) using the HOD approach. They separate galaxies as centrals or satellites (Kravtsov et al. 2004, Zheng et al. 2005). They model the mean number of galaxies that a halo of mass M_h following the funcional form with five free parameters (also Leauthaud et al. 2011 use five parameters to model $\langle N_{cen} \rangle$ and another five parameters to model $\langle N_{sat} \rangle$):

$$\langle N_{gal} \rangle = N_{cen} + N_{sat} = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\log M_h - \log M_{min}}{\sigma_{\log M_h}}\right)\right]\left[1 + \left(\frac{M_h - M_0}{M_1'}\right)^{\alpha}\right] \tag{3.47}$$

where erf is the error function $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$. The number of central galaxies is a step function with a smoothed cutoff in order to account for the scatter between galaxy luminosity and halo mass (see also More et al. 2009). The mean number of satellite galaxies is a power law modified by a similar cutoff profile. This functional form uses five free parameters. This form is motivated by a theoretical study pre-

sented in Zheng et al. 2005, and is identical to the five-parameter model adopted in Zheng et al. 2007 (see also Zheng et al. 2009). According to Zheng et al. 2005 five free HOD parameters model perfectly the mean number of galaxies predicted by hydrodynamic simulations and semi-analytic models. M_{min} characterizes the minimum mass of a halo hosting a central galaxy above the luminosity threshold. Depending on the HOD parameterization M_{min} can vary. In eq. 3.47 is the mass for which half of the haloes which have the minimum mass contains one central galaxy with luminosity greater or equal the luminosity threshold. Another characteristic mass scale M_1, defines the haloes which on average have one additional satellite galaxy above the luminosity threshold, defined as $\langle N_{cen}(M_{min})\rangle = 0.5$. In eq. 3.47 $M_1^{'}$ and M_0 are related to M_1. Zehavi et al. 2011 used the nearest-integer and Poisson distributions for central and satellite galaxies, respectively. Kravtsov et al. 2004, Zheng et al. 2005, Boylan-Kolchin et al. 2010 claimed that, in high mass haloes, the distribution of satellite galaxies become super-Poisson at high $\langle N_{sat}\rangle$. Zehavi et al. 2011 argued that this feature does not impact quantitatively on the clustering derivations.

3.2.2 Galaxy positions

The second point in the HOD is the spatial distribution of galaxies. As mentioned before, Kravtsov et al. 2004 found that it is very useful to separate galaxies into contributions of central galaxies and satellite galaxies. When building mock galaxy catalogues using N-body simulations, there are different possibilities to place the galaxies inside the haloes. The common way is to place the central galaxy at the center of its host halo, it can be either the most bound dark-matter particle or the center of mass of the halo. Satellite galaxies are placed either following a mass density profile, which is usually the NFW profile, or assigning directly the position of the satellite galaxy to a dark-matter particle within the halo. Placing satellite galaxies following a mass density profile solves the problem of resolution given by the linking-length of the simulation. On the other hand if one places satellite galaxies in the same place as dark-matter particles one is fairly following the spatial distribution of dark matter. According to Berlind & Weinberg 2002 the effect of the spatial distribution of galaxies within haloes on the two-point correlation function has a relatively modest impact that is confined to small spatial scales.

3.2.3 Galaxy velocities

In the halo model one assumes that dark-matter haloes are relaxed and virialized structures, and therefore the virial theorem holds. Physical properties of galaxies can be inferred given the halo mass. In order to build mock galaxy catalogues using N-body simulations one can choose between following the virial theorem or taking velocities from particles. If one decides to place satellite galaxies in the same place as dark-matter particles the logical approach is to assign their velocities too. In this method satellite galaxies fairly follow the dark matter velocity field and no approximation has to be done. Sheth & Diaferio 2001 found that to a good approximation the motion of a particle in an N-body simulation can be described as the sum of the virial motion of the particle within its host halo and the bulk motion of the halo as a whole:

$$v = v_{vir} + v_h \tag{3.48}$$

The virial motions within a halo are well approximated by velocities that are independent Gaussians in each of the three cartersian components, with rms values that depend on the halo mass: $\sigma^2_{vir} =< v^2_{vir} >\propto GM_h/r_{vir} \propto M_h^{2/3}$.

The next section describes the galaxy two-point correlation function according to the halo model approach.

3.2.4 Galaxy two-point correlation function

The galaxy two-point correlation function in the halo model can be also decomposed in two terms as well as the halo two-point correlation function. The 1-halo and the 2-halo terms. In the 1-halo term only galaxy pairs that reside in the same halo contribute, and the 2-halo term is dominated by galaxy pairs that reside in separate distinct haloes. The 1-halo term dominates at small scales while the 2-halo term is the main responsible of the total correlation function at scales $\gtrsim 5$ Mpc/h. The main difficulty of modeling the correlation function comes from the transition region between both terms. Zehavi et al. 2004 measuring the projected correlation function $w_p(r_p)$ from the Sloan Digital Sky Survey for a flux-limited sample of 118.000 galaxies and a volume-limited sample subset of 22000 with absolute magnitude $M_r < -21$ found systematics departures from the best-fit power law, in particular a change in slope at $r_p \sim 1 - 2$ Mpc/h. The departures are stronger for the relatively luminous galaxies. They naturally explained this departure using the HOD fitting approach to model the correlation function. They interpret the inflection point as the transition from large to small scale regimes, i.e. the different contributions from galaxy pairs in separate dark-matter haloes and galaxy pairs in the same dark-matter haloes. One can realize this effect of the halo exclusion regarding the two-point correlation function of haloes depending on the halo mass. As already shown, figure 3.1 shows the two-point correlation function of haloes from the Grand Challenge run of the MICE simulations. The higher the mass of the haloes the larger the distance at which the amplitude drops drastically towards smaller scales because of the halo exclusion. No pair counts of haloes at distances smaller than the typical virial radius are found.

Berlind & Weinberg 2002 showed the analytic model of the galaxy two-point correlation function in the halo bias model in real space. They presented the idea already developed in Seljak 2000, Ma & Fry 2000, Scoccimarro et al. 2001 and Sheth et al. 2001b. The analytical model of the two-point correlation function shown later is similar to that presented by Zheng(2004) and used also in Zehavi et al. 2004, 2005 for modeling the SDSS projected two-point correlation function, $w_p(r_p)$, but it also includes the improvement in the treatment of the halo exclusion developed by Tinker et al. 2005. Following the appendix B of Tinker et al. 2005 (where the analytical model of the galaxy clustering is explained from start to finish) the 1-halo term can be expressed by the expression:

$$\xi_{gg}(r) = [1 + \xi^{1h}_{gg}(r)] + \xi^{2h}_{gg}(r) \tag{3.49}$$

The "1+" comes from the pairs counts, proportional to $1+\xi_{gg}^{1h}$ and $1+\xi_{gg}^{2h}$, that sum to give the total pairs counts, proportional to $1+\xi_{gg}$. Following Berlind & Weinberg 2002 the 1-halo term in real space is given by the expression:

$$1+\xi_{gg}^{1h}(r) = \frac{1}{2\pi r^2 \bar{n}_g^2} \int_0^\infty dM_h \; n(M_h) \frac{\langle N(N-1)\rangle_{M_h}}{2} \frac{1}{2R_{vir}(M_h)} F'\left(\frac{r}{2R_{vir}}\right) \quad (3.50)$$

where $\bar{n}_g$ is the mean number density of galaxies, $n(M_h)$ is the halo mass function, and $\langle N(N)-1\rangle_{M_h}/2$ is the average number of pairs in a halo of mass M_h. The function $F(x)$ is the average fraction of galaxy pairs in a halo of mass M_h that have separation less than r. This function is related to the mass halo density profile, $\rho_m(r)$, and $F'(x)$ is its derivative. For the central-satellite galaxy pairs the distribution of pairs is proportional to the volume-weighted density profile, $F'(x) \propto \rho_m r^2$, normalized to 1. The satellite-satellite galaxy pairs come from the convolution of the mass density profile with itself. The average number of pairs in the 1-halo term in the range $(x, x+dx)$ in haloes of mass M_h can be expressed as:

$$\frac{\langle N(N-1)\rangle_{M_h}}{2} F'(x)dx = \langle N_{cen}N_{sat}\rangle_{M_h} F'_{cs}(x)dx + \frac{\langle N_{sat}(N_{sat}-1)\rangle_{M_h}}{2} F'_{ss}(x)dx \quad (3.51)$$

where the subscripts "cs" and "ss" mean central-satellite and satellite-satellite pairs respectively.

Zheng 2004 treated the halo exclusion in the 2-halo term only taking into account haloes with virial radii less than half the value of r for which $\xi_{gg}(r)$ is being calculated, which is the same as taking just haloes with mass smaller than M_{lim},

$$M_{lim} = \frac{4}{3}\pi \left(\frac{r}{2}\right)^3 \rho_{crit}\Omega_m \Delta \quad (3.52)$$

where Δ is the virial overdensity of the halo. This treatment does not take into account pairs between haloes more massive than M_{lim} and smaller haloes. Tinker et al. 2005 showed that this treatment, together with a not accurate halo bias, overpredicts galaxy bias by $\sim 15\%$ and the galaxy correlation function by $\sim 30\%$ at scales $r \gtrsim 10$ Mpc/h. At smaller scales underpredicts the number of 2-halo pairs with errors greater than 50% at 1 Mpc/h. Taking into account this method for the halo exclusion the galaxy 2-halo term of the power spectrum[9] is given by:

$$P_{gg}^{2h}(k,r) = P_m(k) \left[\frac{1}{\bar{n}'_g} \int_0^{M_{lim}} dM_h n(M_h) \langle N\rangle_{M_h} b_h(M_h,r) y_g(k,M_h)\right]^2 \quad (3.53)$$

where $P_m(k)$ is the non-linear matter power spectrum, $y_g(k,M_h)$ is the Fourier transform of the halo density profile, $b_h(M_h,r)$ is the halo bias as a function of the halo mass and separation r and $\bar{n}'_g$ is the average number density of galaxies which reside in haloes of mass $M_h \leqslant M_{lim}$:

[9] It is easier to work in Fourier space rather than real space because convolutions in real space become simple multiplications in Fourier space.

$$\bar{n}'_g = \int_0^{M_{lim}} dM_h n(M_h) \langle N \rangle_{M_h} \tag{3.54}$$

Tinker et al. 2005 improved the halo exclusion treatment taking into account all halo pairs for which the sum of virial radii is smaller than the separation and therefore $R_{vir1} + R_{vir2} \leqslant r$. In the new method they assumed that haloes are spherical. Introducing this improvement in the halo exclusion treatment, the expression of the 2-halo term of the galaxy power spectrum is given by:

$$\begin{aligned} P_{gg}^{2h}(k,r) = P_m(k)\frac{1}{\bar{n}_g^2} \int_0^{M_{lim,1}} dM_{h,1} n(M_{h,1}) \langle N \rangle_{M_{h,1}} y_g(k, M_{h,1}) \cdot \\ \cdot \int_0^{M_{lim,2}} dM_{h,2} n(M_{h,2}) \langle N \rangle_{M_{h,2}} y_g(k, M_{h,2}) \end{aligned} \tag{3.55}$$

where $M_{lim,1}$ is the maximum halo mass such that $R_{vir}(M_{lim,1}) = r - R_{vir}(M_{min})$ and $M_{lim,2}$ is related to $M_{h,1}$ by $R_{vir}(M_{lim,2}) = r - R_{vir}(M_{h,1})$. In the new method the average number density of galaxies, $\bar{n}'_g$, is given by the expression:

$$\bar{n}_g'^2 = \int_0^{M_{lim,1}} dM_{h,1} n(M_{h,1}) \langle N \rangle_{M_{h,1}} \int_0^{M_{lim,2}} dM_{h,2} n(M_{h,2}) \langle N \rangle_{M_{h,2}} \tag{3.56}$$

Equation 3.55 is a double integral but it's more accurate than equation 3.53. The error at the scale of 1 Mpc/h using Zheng's method was of $\sim 50\%$ and correcting for the closer pair counts reduces the error down to $\sim 25\%$. At the scale of 2 Mpc/h the error is completely removed against the previous $\sim 20\%$ error. Tinker et al. 2005 also studied the case in which haloes are not spherical but triaxial objects, and they improved the error at $r = 1$ Mpc/h down to $\sim 10\%$.

Coming back to the HOD parameters and their effects in the galaxy two-point correlation function, $\xi_g(r)$, Berlind & Weinberg 2002 found that the amplitude on large scales is determined by the relative number of galaxies in high mass and low-mass haloes, and is sensitive to α, and to a lesser extent, M_{min}. On small scales the amplitude and the shape of $\xi_g(r)$ are highly sensitive to α, M_{min}, and $P(N|N_{gal})$, although each feature affects $\xi_g(r)$ in a different way. The power-law in the shape of $\xi_g(r)$ in the HOD framework requires a very specific combinations of parameters. The sensitivity of $\xi_g(r)$ to the HOD parameters implies that its observed power-law shape is a strong constraint on the physics of galaxy formation, and that the success of semianalytic models and hydrodynamical simulations in reproducing this form is completely non-trivial. They also studied the influence of the HOD parameters on other clustering statistics like the bispectrum, the galaxy-mass cross correlation or the void probability function. These statistics are not going to be explored in this work.

3.3 Summary

The halo model and the halo occupation distribution model (HOD) have been presented in this chapter. Both are complementary concepts and are the theoretical

framework in which the method we use to generate the mock galaxy catalogues is based. As a consequence, we have also introduced the actual kwnoledge of the scenario of galaxy formation, which is basically that galaxies form inside overdense regions of the dark-matter field (usually called dark-matter haloes). In the gravitational potential wells of these haloes the gas collapses, cools and form stars, and galaxies.

Some basic properties of the halo population, which are included in the halo model, have been studied. In particular, the spatial distribution, the abundance, the bias and the mass density profile of haloes. One of the most relevant properties of the halo clustering is that the two-point correlation function of haloes can be divided in two terms, the 1-halo and the 2-halo terms. The 1-halo term takes into account contributions to the density that come from the same halo and the 2-halo term takes into account contributions to the density that come from different haloes. We have also shown that even physics can be separated in these two regimes. Physics from scales larger than the typical size of a halo are described by the linear perturbation theory. And physics inside the haloes can be inferred since the model assumes that the haloes are in a state of virial equilibrium. We have explained the different approaches in order to derive the mean number density of haloes. We have also introduced the concept of halo bias and shown the internal mass density profile of haloes. To finish with the halo model we have presented some of its limitations and drawbacks.

We have explained in detail the HOD, which is an attempt to relate the dark matter and the galaxy field. The HOD is one of the most powerful methods to describe galaxy bias. It basically describes the number of galaxies with a certain property inside each halo depending on its mass, and it also provides recipes to place galaxies within dark-matter haloes and to include the velocity of galaxies.

Chapter 4

Cosmological Simulations

In this chapter we introduce one of the ingredients we use to generate the mock galaxy catalogues, the Marenostrum Institut de Ciències de l'Espai simulations (MICE). The chapter is structured as follows: in the first section we discuss about the motivation to simulate the universe and provide a brief history of N-body simulations. In the second we summarize the basics of how to produce N-body simulations and in the third and last sections we describe and characterize in particular the MICE simulations.

4.1 Introduction

Since one cannot examine the universe in the laboratory (let alone "several universes"), it is necessary to produce simulations of the universe in order to implement a given experiment. Simulations bridge the gap that often exists between observations and theory. Moreover it is also one way to validate that future projects are feasible (or not) beyond theoretical approaches. The main goal and challenge of cosmological simulations is to reproduce the evolution of the universe from the initial conditions to the present day. The initial conditions are prescribed by a cosmological model, which provides the composition of matter, radiation, and exotic fields, such as a cosmological constant, and primordial fluctuations in the matter, radiation, and spacetime geometry.

The evolution of the universe is characterized by the formation of large structure such as filaments and voids, also dark-matter haloes, clusters of galaxies, galaxies, etc. The most common explanation for the formation of this structure is as the result of the gravitational force in an expanding universe on small primordial density fluctuations together with the action of other physical processes such as adiabatic gas dynamics, radiative cooling, photoionization, recombination and radiative transfer (Peebles 1982, Blumenthal et al. 1984, Davis et al. 1985b).

In our case, we use numerical N-body simulations to trace the evolution of the universe. An N-body simulation is a type of numerical simulation that has a large number of particles interacting under different physical forces forming a dynamical system. Mainly, there are two different kinds of N-body simulations, one composed just of dark matter and one that also takes into account the physics of the gas (hydrodynamical simulations).

The Wilkinson Microwave Anisotropy Probe (WMAP), just by itself, is able to measure the basic parameters of the Big Bang model. It suggests that the content of non-baryonic matter is 23% of the total content of the universe and the content of baryonic matter is just 5%. One can say that the dynamics of the universe is governed by dark matter. One can also say that only gravitational forces and electromagnetic forces can act over large distances. Since cosmic matter is electrically neutral on average, electromagnetic forces do not play any significant role on large scales. Dark matter simulations in comparison with hydrodynamical simulations are simpler because the only physical force involved is the gravitational force. They have become an essential tool in Cosmology and they help us to predict the evolution of the structure in the universe not only in the linear regime but also in the non-linear regime.

In the era before electronic computing one might say that the first N-body simulation was performed by Erik Holmberg in 1941 using an analog computer. He was studying the interaction of galaxies. He used 37 light-bulbs to simulate each galaxy. The intensity of every light-bulb represented gravity. Both intensity of light and gravitational force follow the same inverse-square law. The brighter the light-bulb, the greater the gravitational force. The cumulative intensity measured at a given position (where a light-bulb is placed) was used as an indicator of the gravitational force. He computed the acceleration over a time-step for each light-bulb and changed its position. His experiment was the first to clearly demonstrate effects of tidal interactions that occur during the encounter of two galaxies.

In the early 1960's the first digital N-body simulations in astronomy, studying the N-body problem to the fields of stellar dynamics, were performed by Sebastian Von Hoerner (1960, globular star clusters) and Sverre Aarseth (1963, galaxy clusters). The simulations were limited to systems with no more than 100 particles. The same problems related to computational efficiency and numerical accuracy that we encounter nowadays were discussed by Aarseth and Hoyle in 1964. Through the 1960s and 1970s direct N-body methods using several hundreds of particles were used for several astronomical and cosmological problems. Then the early 1980s saw significant theoretical and technical progress such as the theory of inflation, which provided the initial conditions for simulations, and also the first use of grid-based N-body algorithms, which allowed more than 10^5 particles to be traced. Since the mid 1980s advances in simulations have relied on the increase of computing power and the availability of computers that work in parallel. In the 1990s showed an impressive growth in the size of cosmological simulations and also in the sophistication of the physics. Since the first computer calculations (1960) the number of particles involved has nearly doubled every two years in accordance with Moore's law (fig. 4.1 from Dehnen & Read 2011). The latest collisional N-body calculations have reached over 10^6 particles while collisionless calculations can now reach more than 10^9 particles (Dehnen & Read 2011). Here one realizes the difference in complexity between both kinds of simulations. Projects regarding very large N-body simulations are, for example, those generated by the Virgo Consortium (Frenk et al. 2000), the Millenium I and II (Springel et al. 2005, Boylan-Kolchin et al. 2009), the MICE simulations (Fosalba et al. 2008), the Horizon run (Kim et al. 2009) or the Horizon project (Teyssier et al. 2009).

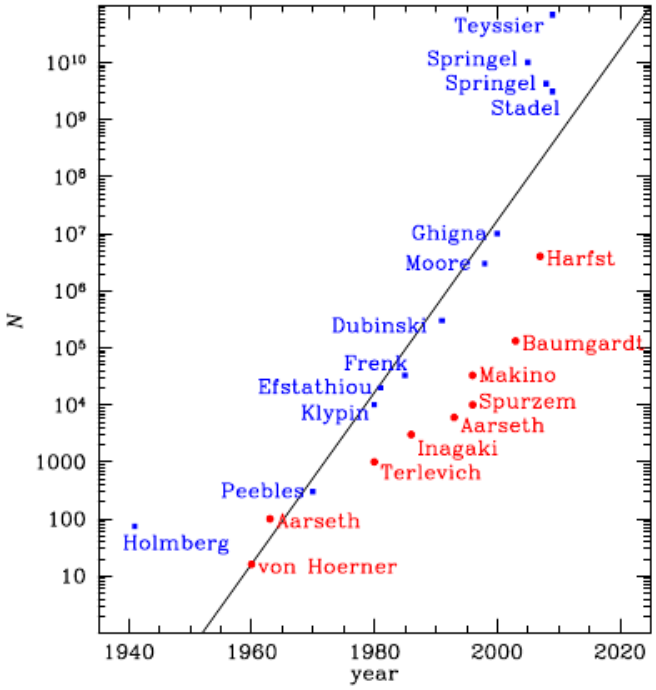

Figure 4.1: Number of particles as a function of time over the past 50 years. Red points represent collisional simulations and blue points collisionless simulations. Figure from Dehnen & Read 2011).

4.2 Basics of Dark Matter N-body Simulations

The basic idea of N-body simulation is to know the position and the velocity of every particle at "any" given time. One starts with a dynamical system of N particles with known masses [1], initial positions and initial velocities in an expanding universe. Then we let the gravitational force act. In the case of just two particles, the motion can be studied analytically. With three or more particles, we get to the N-body problem and there is in general no analytic solution for the problem, so we have to solve it numerically.

The first step is to know the initial conditions of the simulation. One has to specify the background cosmological model and the perturbations imposed on this background. The cosmological model determines the content of dark matter and baryons, a possible cosmological constant, the shape of the universe (in general flat), etc. More recently there are simulations that include different models of dark energy or modified gravity. When baryons decouple from radiation at very high redshift ($z \sim 1100$) small density fluctuations are found in every component of the universe, baryons, photons, massless neutrinos and dark matter. These small amplitude density fluctuations depend on how they were created. There are mainly two classes of early universe models considered, inflation (Guth 1981) with Gaussian

[1] Once one assumes some cosmological model and once one knows the volume and the total number of dark-matter particles in the simulation, one can derive the mass of each particle if they are of equal mass.

fluctuations and topological defects (Vilenkin & Shellard 1994) with non-Gaussian fluctuations.

The matter power spectrum, $P(k)$, completely determines Gaussian fluctuations. Given a position $\vec{x}$ and its density $\delta(\vec{x})$, the probability of having a certain fluctuation at that position is given by:

$$p(\delta)d\delta = \frac{\delta}{\sqrt{2\pi}\sigma} \exp -\frac{\delta^2}{2\sigma^2} \tag{4.1}$$

where $\sigma^2 =< \delta^2 >$. The probability of having certain densities, $\vec{\delta}$, in N points is given by a multidimensional Gaussian function:

$$p(\vec{\delta})d\vec{\delta} = \frac{\vec{\delta}}{\sqrt{2\pi}|C|} \exp -\frac{\vec{\delta}^T C^{-1}\vec{\delta}}{2} \tag{4.2}$$

where $C_{ij} =< \delta_i\delta_j >= \xi_{ij}$, and $\xi(r)$ is the Fourier transform of the power spectrum:

$$\xi(r) = \frac{1}{2\pi^2}\int_0^\infty k^2 P(k)\frac{sin(kr)}{kr}dk \tag{4.3}$$

so, as previously stated, Gaussian fluctuations are fully specified by the power spectrum.

In real space the joint probability distribution at N points is a multidimensional Gaussian function (equation 4.2). In Fourier space the covariance matrix of this Gaussian becomes diagonal and therefore one can sample a Gaussian random field by sampling its Fourier components on a cartesian grid (Peacock & Heavens 1985, Bardeen et al. 1986). Non-Gaussian models are much more complicated and are beyond the scope of this thesis, since our simulations assume Gaussian fluctuations. Once one knows the linear density fluctuation field at some initial time, typically $z \sim 100$, one has to obtain the position and velocity of the particles. The standard approach is to displace equal-mass particles from a uniform Cartesian lattice using the Zel'dovich (1970) approximation (Doroshkevih et al 1980, Dekel 1982, Efstathiou et al 1985):

$$\vec{x} = \vec{q} + D(t)\vec{\psi}(\vec{q}) \tag{4.4}$$

$$\vec{v} = a\frac{dD}{dt}\vec{\psi} = aHfD\vec{\psi} \tag{4.5}$$

where $\vec{q}$ labels the unperturbed lattice position, $D(t)$ is the growth factor of the linear growing mode, and $f = d\ln D/d\ln a \approx \Omega^{0.6}$ is its logarithmic growth rate (Peebles 1980). The irrotational (curl-free) displacement field $\vec{\psi}$ is computed by solving the linearized continuity equation:

$$\vec{\nabla}\cdot\vec{\psi} = -\frac{\delta}{D(t)} \tag{4.6}$$

where $\delta(\vec{x},t) = [\rho(\vec{x},t) - \overline{\rho}]/\overline{\rho}$ is the relative density fluctuation. The displacement field is evaluated from equation 4.6 using Fourier transform methods. Some details of the method to generate the initial conditions are not described here. To find out more about it one can read, for example, the review of Edmund Bertschinger (1998).

Moreover, there are alternative methods to generate initial conditions (Scoccimarro 1997).

Once the particle positions and velocities are known, one has to derive, at each time-step, the gravitational forces experienced by each particle due to the other $(N-1)$ particles. As mentioned before, only gravitational and electromagnetic forces can act over large distances. Since, on average, cosmic matter is neutral, gravity has to be considered as the driving force of the universe. The laws of gravity are described by the theory of General Relativity (GR), however Newton's theory of gravitation has been proven to be eminently successful in certain systems (e.g. motion of planets). Contributions from GR can be ignored since typical dynamical timescales are long compared to the light crossing time for the simulation, and the space-time curvature induced by the particles and the particle velocities are small. The gravitational force $\vec{F_i}$ acting on a particle i of mass m_i is given by Newton's law:

$$\vec{F_i} = -\sum_{j \neq i} G \frac{m_i m_j (\vec{r_i} - \vec{r_j})}{|\vec{r_i} - \vec{r_j}|^3} \tag{4.7}$$

If one divides the force by the mass of the particle one gets its acceleration. Multiplying the acceleration by the time step one finds its change in velocity. Multiplying the average velocity by the time step one finally arrives at the change in position of the particle. This derivation must be done for every particle, so there are $N(N-1)$ calculations per time-step.

One can easily realize that equation 4.7 presents a singularity when the distance between two particles approaches zero and therefore can lead to very large relative velocities. The singularity can be avoided (at the price of altering the dynamics of the system) by introducing the so-called softening length, that is by modifying the gravitational force at small scales and suppressing it below a typical distance, ϵ.

$$\vec{F_i} = -\sum_{j \neq i} G \frac{m_i m_j (\vec{r_i} - \vec{r_j})}{|\vec{r_i} - \vec{r_j}|^2 + \epsilon^2}^{3/2} \tag{4.8}$$

A problem related to the CPU time comes from the huge amount of particles involved in sampling a wide dynamic range of structures. Nowadays, as stated before, state-of-the-art simulations have about $N = 10^{10}$ particles (collisionless simulations). If we compute every particle-particle interaction the number of interactions to compute increases as N^2. So, there are several refinement techniques which aim at obtaining a reliable numerical solution with the minimum amount of CPU time depending on the astrophysical problem.

The two important factors that have to be taken into account to treat this kind of problem are the timescale and the collisionality. This work is concerned with the large scale structure of the universe. A dynamical system of N particles that interact gravitationally reaches a state of dynamic equilibrium on a time scale comparable to a few times the typical cross time[2]. This is the time the system needs to settle down to a state of virial equilibrium. Once the system is in dynamic equilibrium, long term evolution is possible, driven by two-body encounters. The relaxation time

[2]The cross time, T_{cr}, is the time one particle needs to cross the system.

for a dynamical system such as galaxies or dark-matter haloes is much longer than the life of the universe, so they are considered collisionless systems.

There are several methods employed for solving Newtonian gravity systems. The easiest and most intuitive method to derive the gravitational accelerations is the so called direct summation approach. It does not introduce any approximation and therefore you get the highest accuracy solution. The big disadvantage is that the computational cost per body is $O(N)$, so one needs $O(N^2)$ operations to compute the forces on all bodies. A second order accurate leapfrog integration is usually used to make the time integration of trajectories and it only requires one force evaluation per time-step (Efstathiou et al. 1985). One can get more accuracy in the trajectories using higher order schemes with longer time-steps. They are rarely used because of CPU time cost. Another algorithm to compute the gravitational force is the hierarchical tree algorithm (Appel 1985, Barnes & Hut 1986). It divides the system recursively into a hierarchy of cells (i.e."tree"), each containing one or more particles. It is a method for collisionless systems when close encounters are not important and where the contribution to the net force from very distant particles does not need to be computed with very high accuracy. Strong interactions at small scales are typically softened, while the potentials due to distant groups of particles are approximated by multipole expansions about the group centers of mass. The resulting computation time scales as $O(N \log(N))$ but the approximations introduce small force errors. The criteria to separate strong and distant interactions comes from the parameter θ. If a cell of size s and distance d (from the point where acceleration has to be derived) satisfies $s/d < \theta$, the particles in this cell are treated as a pseudoparticle located at the center of mass of the cell. As already stated, the set of particles in the cell is replaced by a low-order multipole expansion depending on the cell mass distribution. The particle-mesh (PM) algorithm is another method to compute the direct summation for collisionless systems in less time. It is based on representing the gravitational potential on a Cartesian grid (with a total of N_g grid points) starting from the density field and by solving Poisson's equation on this grid. The Poisson equation is typically solved using a Fast Fourier Transform (FFT) requiring $O(N_g \log(N_g))$ operations (Miller & Prendergast 1968, Hohl & Hockney 1969, Miller et al. 1970). The method has three steps:

The first step is called mass-assignment. The mass density field, $\rho(\vec{x}, t)$, is computed on a grid from discrete particles positions and masses. The simplest method assigns each particle to the nearest grid point (NGP) with no contribution of mass to any other grid point. This method produces rather large truncation errors (Efstathiou et al. 1985, Hockney & Eastwood 1988). The most used assignment method is Cloud-in-Cell (CIC), which uses multilinear interpolation on the 8 grid points defining the cubical mesh containing the particle. Another method uses 27 grid points to reduce force fluctuations due to the sharp edges.

The second step is to solve Poisson's equation for the gravitational potential:

$$\vec{\phi}(\vec{k}, t) = -4\pi G a^2 \frac{\hat{\rho}(\vec{k}, t)}{k^2} \tag{4.9}$$

where $\hat{\rho}$ and $\vec{\phi}$ are the discrete Fourier transforms of the mass density and potential, respectively. The gravity field is then obtained by transforming the potential back

to the spatial domain and approximating the gradient by finite differences, or by multiplying by $i\vec{k}$ in the Fourier domain.

The third step is to interpolate the gravity field from the grid back to the particles. The same interpolation scheme should be used here as in the first step to ensure that particles self-forces vanish. The total number of required operations in the PM method is $O(N) + O(N_g \log(N_g)$ to evaluate the forces on all particles. As the force is a poor approximation of Newton's law up to several grid spacings, there is a net loss of short range accuracy. There exists an adaptive mesh refinement method, which uses an adaptive rather than a static grid to solve the Poisson equation. The grid elements are concentrated where a higher resolution is needed (in the highest density regions for instance).

The main drawback of the PM method is the limitation of the force resolution due to the spatial grid. In order to increase this force resolution one can supplement the forces with a direct sum over pairs separated by less than two or three grid spacings, resulting in the particle-particle/particle-mesh (P^3M) algorithm. Hockney et al. 1974 developed this hybrid method and Efstathiou & Eastwood 1981 first used it in cosmology. Problems arise when clustering becomes strong because the cost of direct summation dominates. One solution is to replace the direct summation by a tree code, leading to a hybrid PM-Tree scheme. These methods are generally extremely well suited for cosmological simulations, for example see GADGET-II (Springel 2005). Another method is to create a subgrid mesh. Pair summation is still done but only for pairs whose separation is less than two or three spacings of the subgrid mesh, resulting in a substantial reduction. As the pair summation no longer dominates as it does in P^3M, the force computation of this new adaptive P^3M scales as $O(N \log(N))$, similar to a tree code.

4.3 MICE Simulations

The Marenostrum Institut Ciències de l'Espai simulations (MICE) are a set of N-body simulations carried out at the Marenostrum supercomputer at the Barcelona Supercomputing Center [3]. The GADGET-II code (Springel 2005) is used to generate all simulations, in particular the Lean-GADGET-II, which is a highly memory efficient version of GADGET-II. It computes gravitational forces with a hierarchical tree algorithm (optionally in combination with a particle-mesh scheme for long range gravitational forces). The MICE simulations are a suite of different runs that assume a flat concordance ΛCDM model with parameters $\Omega_m = 0.25$, $\Omega_\Lambda = 0.75$, $\Omega_b = 0.044$, $h = 0.7$ and the spectral index $n_s = 0.95$. The linear power spectrum is normalized to yield $\sigma_8 = 0.8$ at $z = 0$. Initial conditions were given by either using the Zel'dovich approximation or $2nd$ order Lagrangian Perturbation Theory (2LPT) (Scoccimarro 1998, Crocce et al. 2006). The main goal of the simulations is to study the formation and evolution of structure at very large scales and also to determine how well future astronomical surveys, such as DES or PAU, can answer the current cosmological questions. Figure 4.2 shows the suite of simulations in the mass resolution-volume plane. They sample a wide range of cosmological vol-

[3] Barcelona Supercomputing Center, www.bsc.es.

umes, from the SDSS main sample ($0.1h^{-3}\mathrm{Gpc}^3$) to DES ($9h^{-3}\mathrm{Gpc}^3$) or even Euclid ($\sim 100h^{-3}\mathrm{Gpc}^3$), and mass resolutions, from $3\times10^{12}h^{-1}\mathrm{M}_\odot$ down to $3\times10^{8}h^{-1}\mathrm{M}_\odot$. Table 4.1 shows the suite of MICE simulations with their main characteristics.

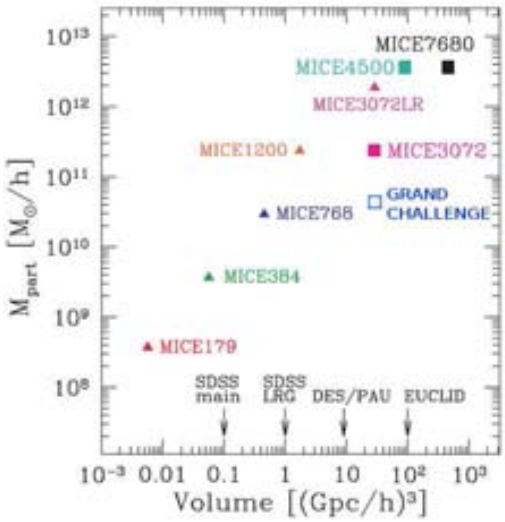

Figure 4.2: Set of MICE simulations in the mass particle-volume plane. Big squares map the halo mass function at the high mass end whereas small triangles (test simulations) extend the dynamical range down to haloes of $10^{10}M_\odot$.

Run	$L_{box}/h^{-1}Mpc$	N_{part}	$m_p/h^{-1}M_\odot$	$l_{soft}/h^{-1}Kpc$	z_i
GRAND CHALLENGE	3072	4096^3	2.93×10^{10}	50	100
MICE7680	7680	2048^3	3.66×10^{12}	50	150
MICE3072	3072	2048^3	2.34×10^{11}	50	50
MICE4500	4500	1200^3	3.66×10^{12}	100	50
MICE3072LR*	3072	1024^3	1.87×10^{12}	50	50
MICE768*	768	1024^3	2.93×10^{10}	50	50
MICE384*	384	1024^3	3.66×10^{9}	50	50
MICE179*	179	1024^3	3.70×10^{8}	50	50
MICE1200* (x20)	1200	800^3	2.34×10^{11}	50	50

Table 4.1: MICE N-body simulation description. The first column is the name of the simulation. L_{box} is the size of the box, N_{part} is the number of particles, m_p gives the particle mass, l_{soft} is the softening lenght and z_{in} denotes the initial redshift of the simulations. Simulations with the mark * were done for completeness or testing.

The MICE simulations have been used to produce both comoving and lightcone outputs. In particular a lightcone output of the simulation MICE3072 has been built, which has $N_p = 2048^3$ and $L_{box} = 3072$ Mpc/h. About 200 comoving outputs separated by constant spacing in cosmic time (≈ 70 Myr) has been used to build the

lightcone. Figure 4.3 shows an octant of the MICE3072 run lightcone, from $r = 0$ to $r = 150$ Mpc/h. Blue points are dark-matter particles. Different structures, such as filaments and voids, can be seen in the image.

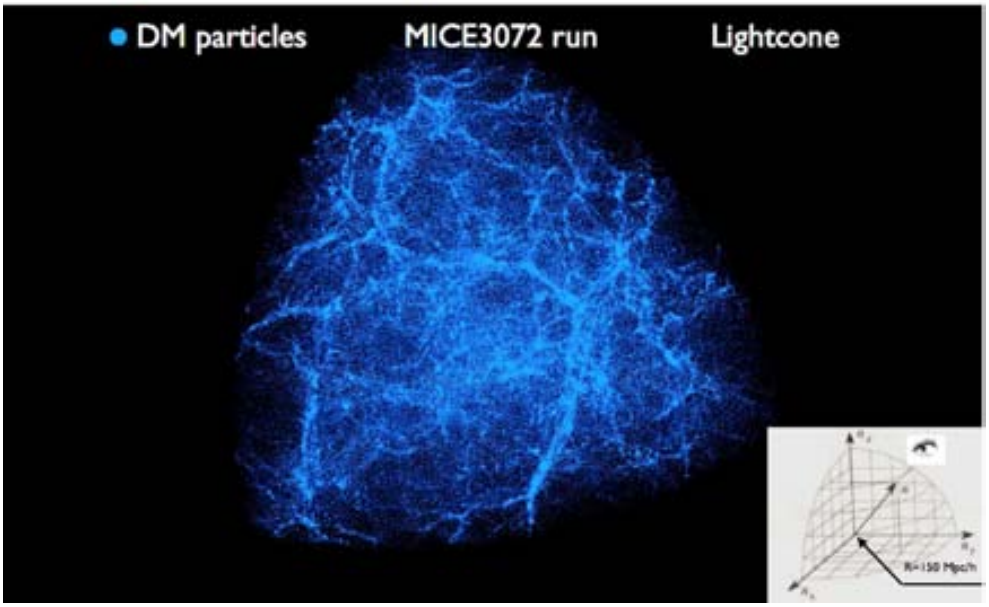

Figure 4.3: Lightcone image from MICE3072 run. Blue points represent dark-matter particles.

Projected density and weak lensing maps have also been produced from the main MICE runs (Fosalba et al. 2008). Some of the MICE simulations have already been used to study the clustering of LRG galaxies with multiple-band photometric surveys such as PAU (Benítez et al. 2009), the large-scale clustering in the spectroscopic LRG SDSS sample through redshift space distortions (Cabré & Gaztañaga 2009a, Cabré & Gaztañaga 2009b), baryonic acoustic oscillation in the three-point correlation function (Gaztañaga et al. 2009a) and in the radial direction (Gaztañaga et al. 2009b), and also to study the high-mass end of the halo mass function (Crocce et al. 2010).

MICE simulations are ideal for studying very large scale statistics because of their volume. One can study very long distance effects or the baryonic acoustic scale. Having such a big difference in volume between simulations helps to study volume effects. Another competitive aspect of the simulations is the fact that replication is not required out to $z = 1.4$ for the simulations with the largest box-size.

During the process of generating mock galaxy catalogues we have used both the MICE3072 and Grand Challenge (GC) runs. At the beginning we used the MICE3072 run and then we used the MICE GC run when it was performed. The MICE3072 run has, as previously stated, 2048^3 particles and a box-size $L_{box} = 3072$ Mpc/h. This is the same number of particles as the Millenium simulation (Springel 2005) but 216 times more volume. To put some numbers, the snapshot of MICE3072 run at $z = 0$ occupies 250GB, and contains ~ 25 million haloes more massive than

$3.9 \times 10^{12} h^{-1} M_{\odot}$ (using a FoF code and setting b=0.2)[4]. The MICE GC run traces the evolution of 4096^3 particles, which is, along with the work of (Teyssier et al. 2009), the largest N-body simulation ever produced until recently with box-size $L_{box} = 3072$ Mpc/h. Every snapshot occupies $\sim$ 2TB. In this case the number of haloes in the $z = 0$ output contains more than 350 million haloes.

The theoretical linear power spectrum of fluctuations at $z = 0$, P(k), used in the MICE simulations is shown in figure 4.4 (it is already normalized). Figure 4.5 shows the two point correlation function, $\xi(r)$, derived using eq. 4.10.

$$\xi(r) = \frac{1}{2\pi^2} \int_0^{\infty} k^2 P(k) \frac{sin(kr)}{kr} dk \tag{4.10}$$

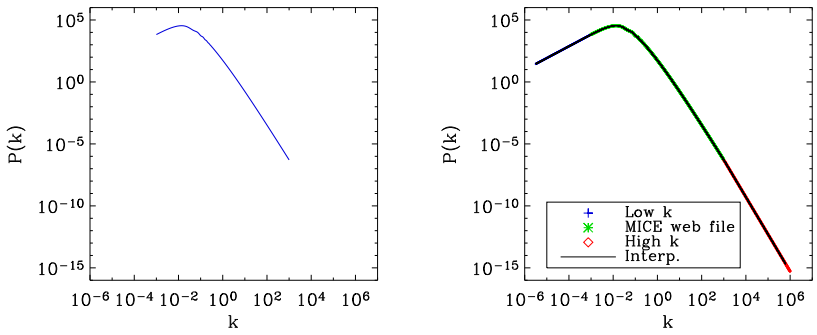

Figure 4.4: The left panel shows the linear power spectrum at z= 0 (normalized), $P_0(k)$, used in MICE simulations. The right panel also shows $P_0(k)$ adding points in order to cover the whole range in k's and derive its Fourier Transform.

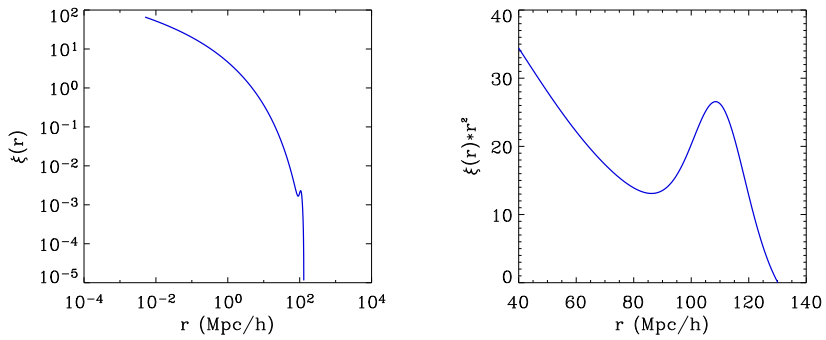

Figure 4.5: The left panel shows the two point correlation function of dark matter from MICE simulations, $\xi(r)$, derived using equation 4.10. The right panel is a zoom of the left panel and multiplied by the square of the separation in order to emphasize the BAO peak.

[4]The parameter b determines the maximum distance at which dark-matter particles in the simulation are defined as *friends*. In particular the distance is usually expressed as the mean inter-particle distance of the simulation times the parameter b.

The P(k) data are available on the MICE webpage, http://www.ice.cat/mice. This file contains two columns, the first one is k and second one is P(k). The lowest k is 0.001 and the largest one is 1000.0. Some additional points either at low k ($P(k) \propto k^{0.95}$) and high k ($P(k) \propto k^{n_s-4}$) are added, and also an interpolation of the power spectrum with a different constant step in $\ln(k)$, $\Delta(\ln(k)) = 0.001$, in order to derive the correlation function (figure 4.4).

The P(k) file is not normalized. Equation (4.11) is used to derive the normalization constant assuming $\sigma_8 = 0.8$. The normalized $\sigma(M_h, z = 0)$ is shown in figure 4.6 and is defined as the variance of the linear density field smoothed with a top hat filter of radius R and enclosing an average mass $M = \rho_b 4\pi R^3/3$.

$$\sigma^2(M_h, z) = \frac{D^2(z)}{2\pi^2} \int_{k_{min}}^{k_{max}} k^2 P(k) W^2(kR) dk \tag{4.11}$$

where D(z) is the linear growth factor between $z = 0$ and the redshift of interest and $W(kR)$ is the Fourier transform of the top hat filter of radius R.

$$D(z) = \left(\frac{x}{x_0}\right) \left(x^3 + 1\right)^{1/2} \int_0^1 dy \left(1 + x^3 y^{6/5}\right)^{-3/2} \tag{4.12}$$

$$x = \frac{\left(\Omega_0^{-1} - 1\right)^{\frac{1}{3}}}{1 + z} \tag{4.13}$$

where Ω_0 is the matter density parameter at $z = 0$, and x_0 is the value of x at $z = 0$.

$$W(x) = \frac{3}{x^3} \left(\sin(x) - x\cos(x)\right) \tag{4.14}$$

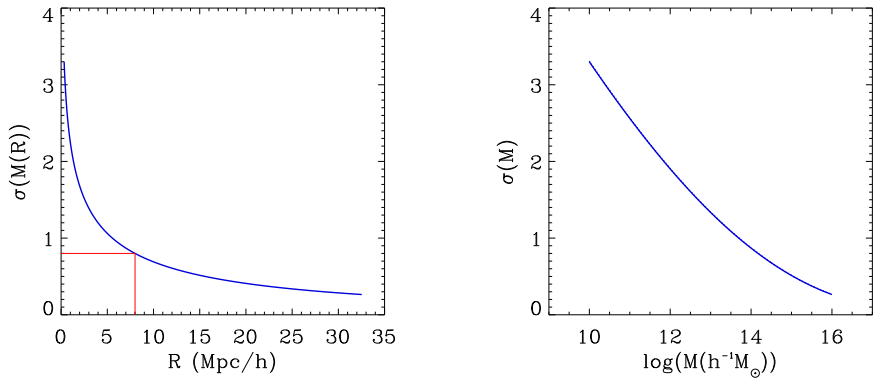

Figure 4.6: Variance of the linear density field as a function of radius (left panel) and mass (right panel).

4.4 MICE Halo Catalogues

To generate the mock galaxy catalogues using the HOD model (previously discussed in chapter 3) it is essential to derive a dark-matter halo population. As mentioned

before, defining a halo is not a simple question (e.g. haloes in simulations may be in a state of merging or accreting). There are two main methods to define a halo, the spherical overdensity (SO, Lacey & Cole 1994) and the Friends-of-Friends (FoF, Davis et al. 1985a) algorithms.

Spherical overdensity defined haloes are spherical regions around matter density peaks, which have an inside mean matter density larger than a given threshold. This threshold is usually taken as a fixed multiple of the critical or background density. The FoF algorithm looks for neighbours of a given particle. Neighbours (or "friends") are defined as particles that are closer than a certain distance to the given particle. The algorithm follows this procedure recursively with the neighbours until no more "friends" are found. The result is a group of dark-matter particles that are closer than a given distance, the linking length, which is some fraction, b, of the mean interparticle distance. The standard value $b = 0.2$ produces haloes that match the mean density of the SO mass definition quite well (for comparisons of the halo mass obtained with SO and FoF see White 2001, White 2002, Warren et al. 2006, Tinker et al. 2008, Lukić et al. 2009, Robertson et al. 2009). The abundance of FoF haloes seems to be universal at the $< 10\%$ level (Jenkins et al. 2001, Reed et al. 2003, Reed et al. 2007, Heitmann et al. 2006, Lukić et al. 2007, Tinker et al. 2008, Crocce et al. 2010) with some dependence on the linking length (Jenkins et al. 2001, Tinker et al. 2008), so there is no need to simulate every cosmological model.

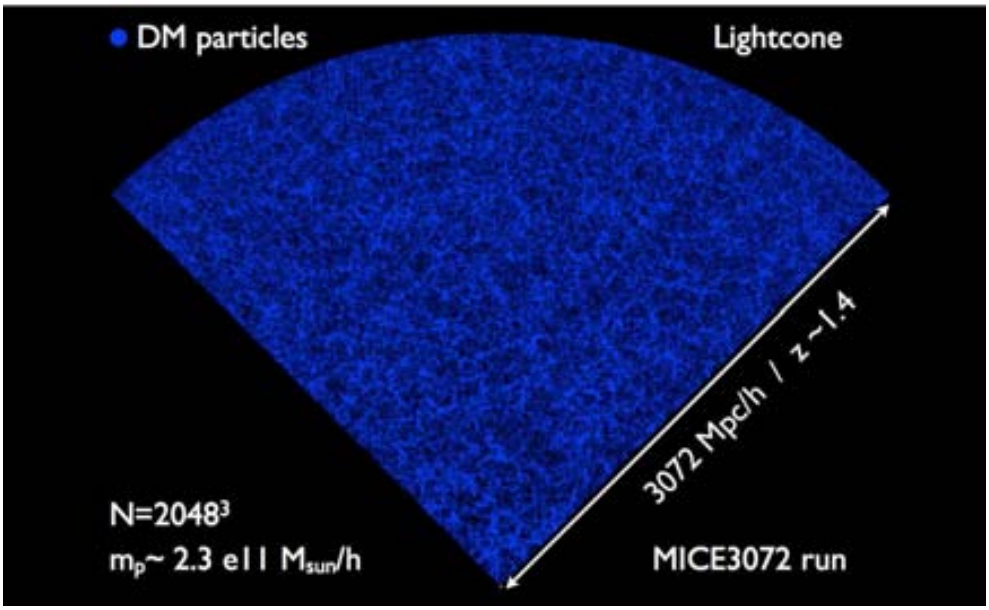

Figure 4.7: Slice through the MICE3072 lightcone map of the distribution of dark-matter particles, from $z = 0$ to $z \sim 1.4$. The distance from the vertex represents the real space of the dark-matter particle.

MICE halo catalogues are derived using the public halo finder code available at the N-body Shop (http://www-hpcc.astro.washington.edu/) with some modifications to gain CPU efficiency because of the large amount of particles involved. Every halo contains all the information needed to generate the mock galaxy catalogues, not

only the mass, position and velocity of the center of mass, and the virial velocity, but also the position and velocity of every particle that belongs to the halo. This information is essential when one wants to place galaxies inside the haloes. One easy and visual check, beyond data analysis or theoretical matches, is plotting the halo catalogue and the dark-matter particles at the same time. In this case, to visualize the simulations we use the public available partiview[5] software developed at the American Museum of Natural History. It is also very useful not only as another test for the halo catalogues but also to observe different structures at large scales such as voids, filaments or clusters.

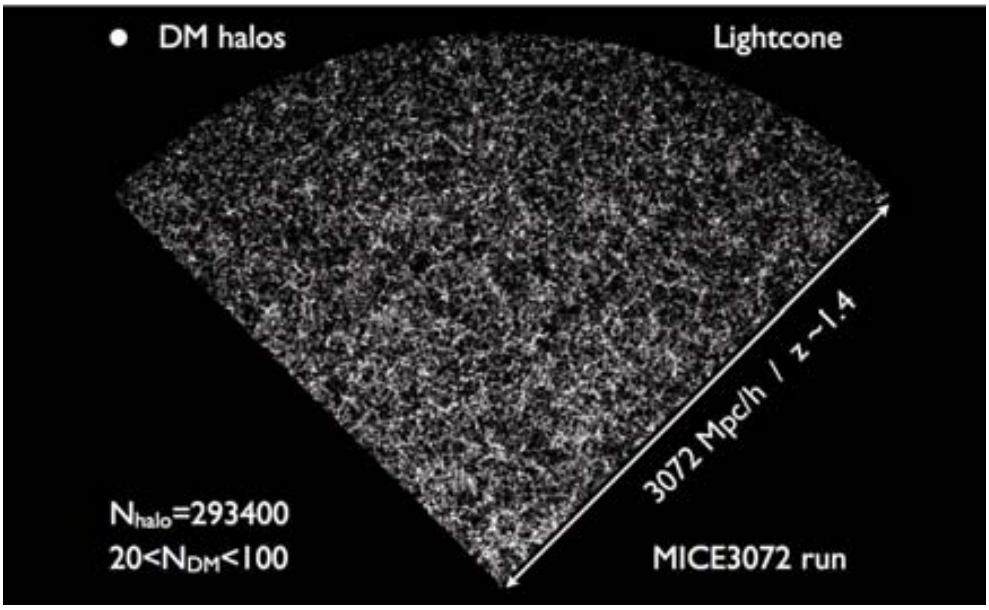

Figure 4.8: Slice through the MICE3072 lightcone of the distribution of dark-matter haloes. The positions of 148928 haloes with no more than 100 dark-matter particles are plotted in this diagram.

A slice through the MICE3072 lightcone of the distribution of dark-matter particles is shown in figure 4.7. The blue points represent dark-matter particles. It actually shows a diluted dark matter field, i.e. just 1 over 500 of the total number of dark-matter particles. The lower vertex corresponds to $z = 0$, the maximum radial distance is 3072 Mpc/h which is $z \sim 1.4$ and the thickness of the slice is 50 Mpc/h. Large density regions and voids can clearly be distinguished. The clustering time evolution cannot be clearly appreciated from figure 4.7, i.e. to observe more overdense regions close to the vertex ($z = 0$) rather than far away ($z = 1.4$). In figure 4.8 white points represent dark-matter haloes in the same slice. The plotted haloes do not have more than 100 dark-matter particles. There are a total of 148928 haloes in the map. It is easy to see the inhomogeneities and in this case one can observe that there are more haloes at low redshift than at high redshift (clustering evolution). This feature is even clearer in figure 4.9 where red points represent dark-matter haloes consisting of more than 100 dark-matter particles in the same slice.

[5] http://www.haydenplanetarium.org/universe/partiview.

Virtually no very massive haloes can be found at high redshift (there has been no time to build such big structures). In figure 4.10 one can see a zoomed-in region of the previous figures. Blue points represent dark-matter particles. One can observe clearly that there is a concentration of dark-matter particles at the center of the image, and with more difficulty, and to a lesser degree, another concentration of dark-matter particles at the bottom right (notice that the density field is diluted so not every particle is plotted). Figure 4.11 shows the same region as the previous one but in this case dark-matter haloes are also plotted. The central concentration of blue points form a massive halo (red point), and the bottom right blue points are part of a less massive halo (white point).

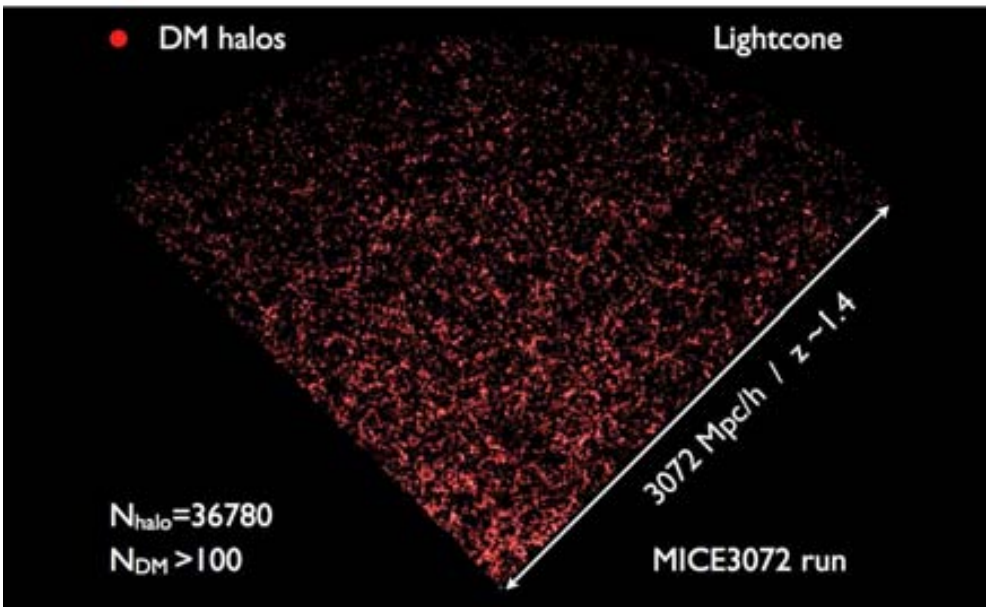

Figure 4.9: Slice through the MICE3072 lightcone of the distribution of very massive dark-matter haloes. Red points represent dark-matter haloes with more than 100 dark-matter particles.

4.4.1 Halo mass function

The halo mass function (MF) describes the halo mass distribution in the universe. It usually refers to the cumulative halo MF, i.e. the number density of haloes with mass greater than some mass M_h, $n(> M_h) = \frac{dn}{dM_h}(> M_h)$. Structure formation models predict that high density peaks of the density field evolve and collapse into dark-matter haloes once the density contrast reaches a certain threshold. According to the spherical collapse model (Gunn & Gott 1972) a region is going to collapse into a point when its mean density reaches ~ 200 times the background density at formation:

$$\langle \rho \rangle > 200 \rho_{cr} = 200 \rho_{cr}^{0} (1+z)^3 \tag{4.15}$$

The Press-Schechter model (Press & Schechter 1974) uses the model of spherical

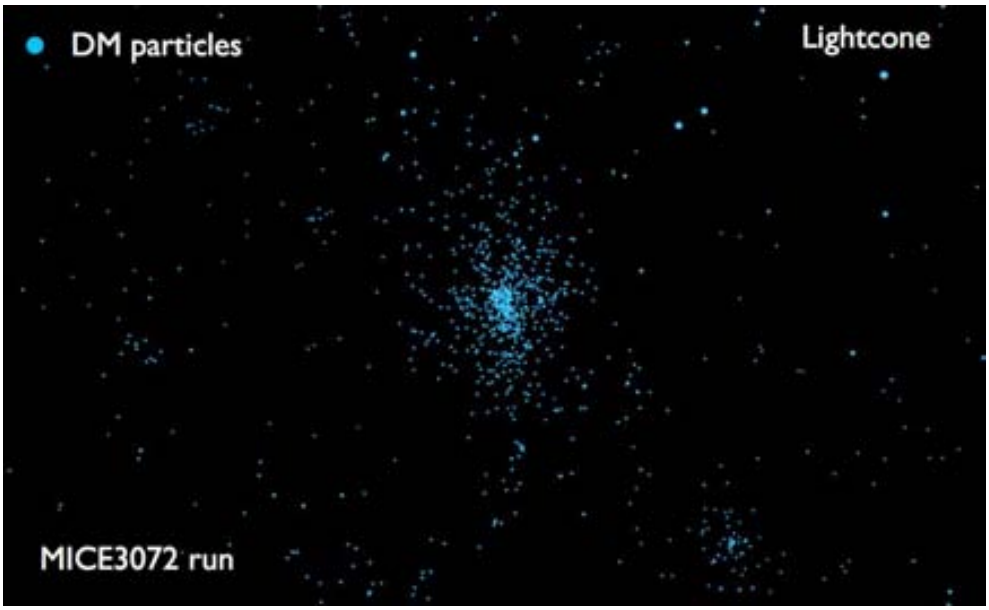

Figure 4.10: Zoom from previous images. Blue points represent dark-matter particles. There exist an overdensity at the center of the image and to a lesser degree another overdensity at the bottom right.

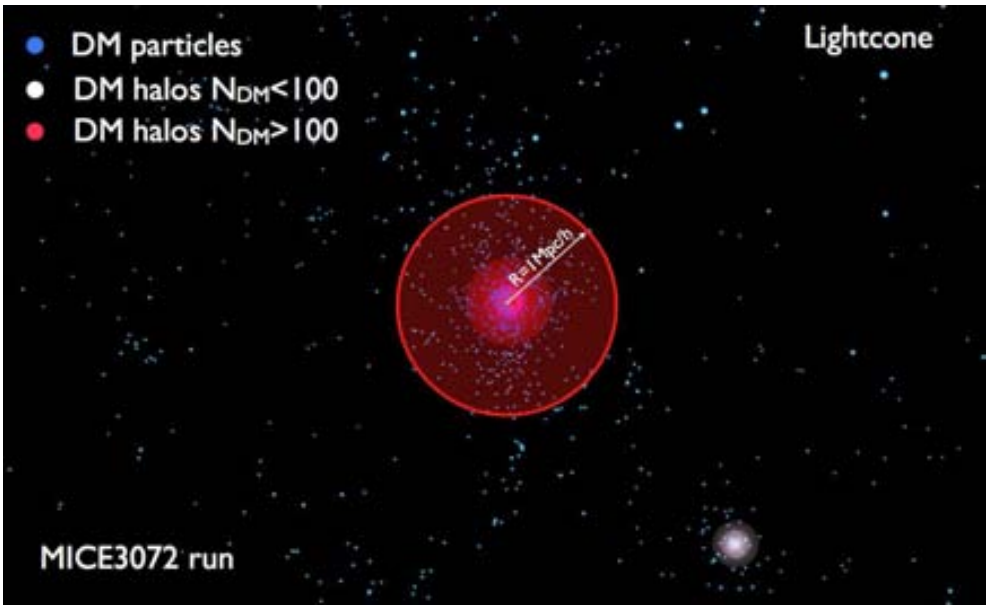

Figure 4.11: Idem as figure 4.10 but including dark-matter haloes. The central concentration of blue points form a massive halo (red point), and the bottom right blue points are part of a less massive halo (white point).

collapse to approximately compute the number density of dark-matter haloes as a function of their mass and redshift. Let $\delta_0(\mathbf{x})$ be considered as a density fluctuation field, with fluctuations on all scales following the power spectrum, $P_0(k)$. This field

is smoothed with a smoothing length R. Let $\delta_R(\mathbf{x})$ be the smoothed density field linearly extrapolated to the present day. Fluctuations on scales $\lesssim R$ are removed because they have been smoothed out. Each maximum in $\delta_R(\mathbf{x})$ corresponds to a peak with characteristic scale $\gtrsim R$ and according to the spherical collapse model also corresponds to a mass M$\sim (4\pi R^3/3)\rho_0$. If δ_R is sufficiently large a spherical region of radius R will decouple from the linear growth of density fluctuations and will begin to grow non-linearly. If $\delta_0(\mathbf{x})$ is a Gaussian distribution (which is expected for several reasons) the statistical properties of the field are completely described by the power spectrum, as explained in previous chapter, and one can compute the number density of maxima with $\delta_R \geqslant \delta_{min}$ (where δ_{min} determines the minimum value for a region to collapse and form a halo and can be computed for any cosmological model) and therefore the number density of relaxed dark-matter haloes in the universe as a function of mass and redshift, $n(M, z)$. In the mid-1990s numerical simulations of structure formation reached a precision that found significant discrepancies between the number density of haloes in simulations and the Press-Schechter predictions. Analytical predictions have also been improved (Bond et al. 1991, Sheth et al. 2001c). Instead of using the spherical collapse model a more realistic ellipsoidal collapse have been introduced by which the number density of haloes is modified in comparison to the Press-Schechter model and it is in very good agreement with the numerical results. After Press-Schechter, semianalytic fits by Sheth & Tormen 1999, Jenkins et al. 2001, Warren et al. 2006 and many more have been done using simulations.

Crocce et al. 2010 studied the abundance of massive haloes in the MICE simulations[6]. They provide a re-calibration of the MF over 5 orders of magnitude in mass (10^{10} <M/$(h^{-1}M_\odot < 10^{15})$), that accurately describes its redshift evolution up to $z = 1$. They used the very large volumes of MICE to make a fair sample of the very high mass end of the halo MF. As a first check they compared the halo abundance in MICE3072 to that in the Hubble Volume Simulation (HVS) (Jenkins et al. 2001, Evrard et al. 2002) because they simulate almost the same volume. They found that, after correcting for different cosmologies (HVS assumes $\sigma_8 = 0.9$ vs. $\sigma_8 = 0.8$ of the MICE simulations) and also for artificially lower abundance in the HVS simulations due to its late start $z_i = 34$, the difference between the measured ratio and the prediction are within the claimed accuracy for the Jenkins fit (10-15%). They also found that available fits to the local abundance of haloes (Warren et al. 2006) match the abundance estimated in the large volume of MICE up to $M \sim 10^{14}h^{-1}M_\odot$ but significantly deviate for larger masses, underestimating the MF by 10% (30%) at $M = 3.16 \times 10^{14}h^{-1}M_\odot$ ($10^{15}h^{-1}M_\odot$). The same underestimation occurs when one extrapolates the Sheth & Tormen 1999 fit to high redshift (assuming universality). The cluster abundance decreases by 30%, 20% and 15% at z=0, 0.5, 1 for fixed $\nu = \delta_c/\sigma \approx 3$ (corresponding to $M \sim [7 - 2.5 - 0.8] \times 10^{14}h^{-1}M_\odot$ respectively).

The differential MF is defined as:

$$f(\sigma, z) = \frac{M_h}{\rho_b} \frac{dn\,(M_h, z)}{d\ln\sigma^{-1}\,(M_h, z)} \tag{4.16}$$

where $n\,(M_h, z)$ is the comoving number density of haloes with mass M_h, $\sigma(M_h, z)$

[6]The MICE GC run was not available at that time.

is the variance of the linear density field (defined previously in equation 4.11) and ρ_b is the background density. R is given in comoving distance, ρ_b is also a comoving density (it is constant in time) so the mass does not change with time for each radius.

The number density of objects in a halo mass bin is given by the expression:

$$n_{bin} = \int_{M_{h_1}}^{M_{h_2}} \left(\frac{dn(M_h, z)}{dM_h} \right) dM_h \tag{4.17}$$

Using equation 4.16,

$$\frac{dn(M_h, z)}{dM_h} = -\frac{\rho_b}{\sigma(M_h, z)} \frac{1}{M_h} \frac{d\sigma(M_h, z)}{dM_h} f(\sigma, z) \tag{4.18}$$

Crocce et al. 2010 fit a Sheth & Tormen-like Mass Function (ST-like MF) to their simulations:

$$f(\sigma, z) = f_{Warren}(\sigma(M_h), z) = A(z) \left[\sigma^{-a(z)} + b(z) \right] \exp \left[-\frac{c(z)}{\sigma^2} \right] \tag{4.19}$$

In table 2 of Crocce et al. 2010 the MF best-fitting parameters for $f(\sigma, z)$ for the MICE simulations are shown. They derive how the parameters evolve with redshift: $A(z) = 0.58(1+z)^{-0.13}$, $a(z) = 1.37(1+z)^{-0.15}$, $b(z) = 0.3(1+z)^{-0.084}$, $c(z) = 1.036(1+z)^{-0.024}$.

$$\frac{dn(M_h, z)}{dM_h} = -\frac{\rho_b}{\sigma(M_h, z)} \frac{1}{M_h} \frac{d\sigma(M_h, z)}{dM_h} A(z) \left[\sigma^{-a(z)} + b(z) \right] \exp \left[-\frac{c(z)}{\sigma^2} \right] \tag{4.20}$$

$$n_{bin} = \int_{M_{h_1}}^{M_{h_2}} \left(\frac{dn}{dM_h} \right) dM_h = \int_{M_{h_1}}^{M_{h_2}} \frac{-\rho_b}{M_h} \frac{1}{\sigma} \frac{d\sigma}{dM_h} f(M_h, z) dM_h \tag{4.21}$$

In order to compare the MF with predictions we measure the MF of the MICE GC run at $z = 0$. In practice we measure the number of haloes ΔN in a given mass bin $[M_1, M_2]$ of width ΔM_h and define (as in Crocce et al. 2010):

$$\frac{dn}{d \ln M_h} = \frac{\Delta N}{V_{box}} \frac{M_h}{\Delta M_h} \tag{4.22}$$

The halo mass is computed taking into account the Warren correction, given by equation 4.23:

$$M_h^W = m_p(N_p(1 - N_p^{-0.6})) \tag{4.23}$$

that corrects a systematic problem that FOF halo masses suffer when haloes are sample by relatively small number of particles. The mass bins are equally space in log-mass, with $\Delta \log_{10} M_h/(h^{-1}M_\odot) = 0.1$. The corresponding value of the mass of the bin is obtained as:

$$M_{bin} = \frac{\int_{M_1}^{M_2} \left(\frac{dn}{dM_h} \right) M_h dM_h}{\int_{M_1}^{M_2} \left(\frac{dn}{dM_h} \right) dM_h} \tag{4.24}$$

from theory and as a mass weighted average:

$$\overline{M}_{bin} = \frac{\sum_{bin} M_h^i}{\Delta N} \tag{4.25}$$

from the simulation. To estimate the errors of the MF we use the Jack-knife re-sampling method. We divide the simulation volume into $N_{JK} = 10^2$ non-overlapping regions, and compute the MF in the full volume excluding one of the non-overlapping regions at a time. We choose as a non-overlapping region a prism of square base with size 307.2 Mpc/h and height 3072 Mpc/h. Crocce et al. 2010 perform a detailed study of different methods to estimate the error or variance in mass function measurements. They found that at smaller masses the sampling variance becomes increasingly important rapidly dominating the total error, and that the Jack-knife re-sampling method does capture this trend but only partially, in particular for the smaller box-sizes ($\lesssim$ 500 Mpc/h, which is not our case), where sampling variance from the absence of long-wavelength modes is more significant. For the Jack-knife method Crocce et al. 2010 used $N_{JK} = 5^3$ and they checked that the estimates have already converged for those values with varying N_{JK}. The variance (defined as the relative error squared) in the i-bin of the number of haloes is obtained as:

$$\sigma_{JK}^{(i)^2}(\Delta N) = \frac{1}{\overline{n}^{(i)^2}} \frac{N_{JK}-1}{N_{JK}} \sum_{j=1}^{N_{JK}} \left(n_j^{(i)} - \overline{n}^{(i)} \right)^2 \tag{4.26}$$

where $\overline{n}^{(i)}$ is the mean number density of haloes for that bin. We checked our errors with the measurements done by Crocce et al. 2010. The top panels of figure 4.12 show the errors measured using different methods (Jack-knife, sub-volumes and Poisson methods) by Crocce et al. 2010 for MICE3072 and MICE7680 runs. The bottom panel of 4.12 shows the errors we estimate for the MICE GC run, respectively. They agree in the same halo mass range and the same simulation size ($L_{box} = 3072$ Mpc/h).

Figure 4.13 shows the MF for the MICE GC run at $z = 0$. Red triangles represent $dn/d\log(M_h)$ in the simulation. Errors are not visible because they are very small. The blue solid line is the best fit using a Schechter-like function (equation 4.27), where $a_0 = 6.1065 \times 10^{-5}$, $a_1 = -0.8438$, $a_2 = 0.6946$ and $M_h^* = 2.3880 \times 10^{14}$.

$$\frac{dn}{d\log M_h} = a_0 \left(\frac{M_h}{M_h^*} \right)^{a_1} \exp\left[- \left(\frac{M_h}{M_h^*} \right)^{a_2} \right] \tag{4.27}$$

4.4.2 Linear large scale halo bias

There are several different ways of measuring halo bias (see e.g. Manera et al. 2010). It measures the relation between the matter density field and the halo density field. The linear large scale halo bias, b_h^{Lin}, can be derived using the two point correlation function. In this section we derive the large scale MICE halo bias using this method and compare our results with fits from Manera et al. 2010.

Expression 4.28 relates the two point correlation function of haloes and dark matter assuming a linear bias relation between the haloes and dark matter distributions, $\delta_h(\mathbf{r}) = b_h^{Lin}\delta_m(\mathbf{r})$:

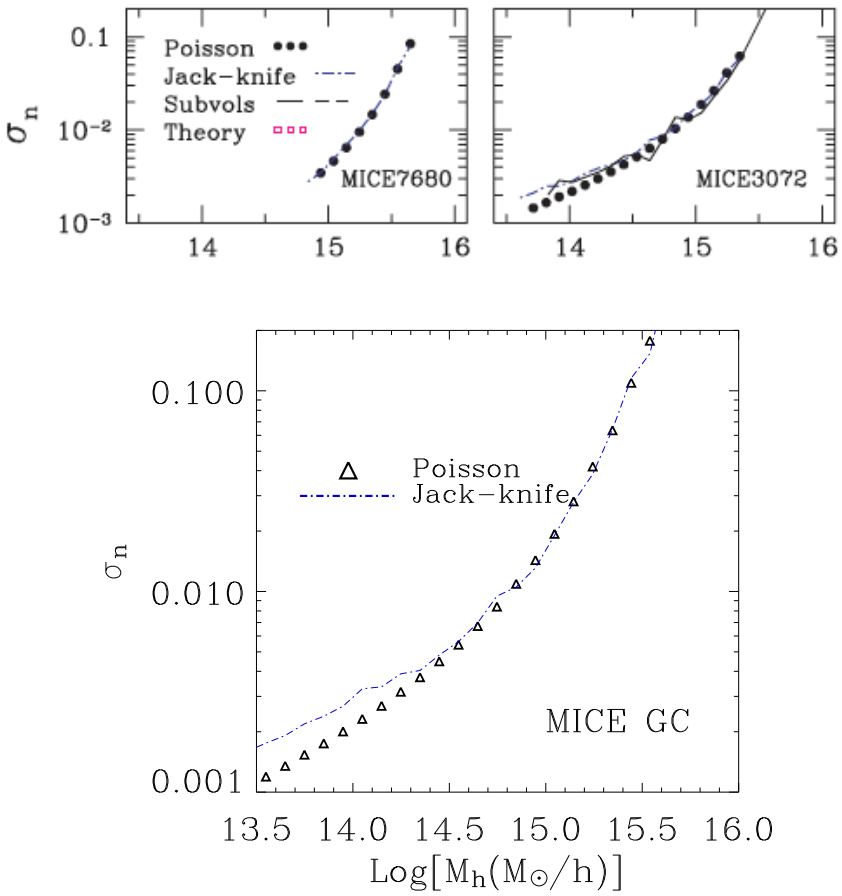

Figure 4.12: Top panels: Different estimates for the variance in the halo mass function within MICE3072 and MICE7680 runs using different methods for two different box-sizes (3072 − 7096 Mpc/h) (two panels of a figure from Crocce et al. 2010). Bottom panel: Poisson and Jack-Knife errors from MICE GC run.

$$\xi_h(r) = \left(b_h^{Lin}\right)^2 \xi_{DM}(r) \tag{4.28}$$

Manera et al. 2010 presented a maximum-likelihood method for fitting parametric functional forms to halo abundances. Manera et al. 2010 showed a linear bias function for haloes applying the peak background split to a Warren-form halo mass function:

$$b^W\left(\nu\left(\sigma\left(M_h\right)\right)\right) = 1 + \frac{c^{'}\nu - 1}{\delta_c} + \frac{2a^{'}b^{'} + b^{'} + \left(c^{'}\nu\right)^{a^{'}}}{\delta_c\left(b^{'} + \left(c^{'}\nu\right)^{a^{'}}\right)} \tag{4.29}$$

where ν is the natural scaling variable according to the spherical evolution model and is defined as:

$$\nu \equiv \frac{\delta_{sc}^2\left(\Omega_z, \Lambda_z, z\right)}{\sigma^2(M_h)} \tag{4.30}$$

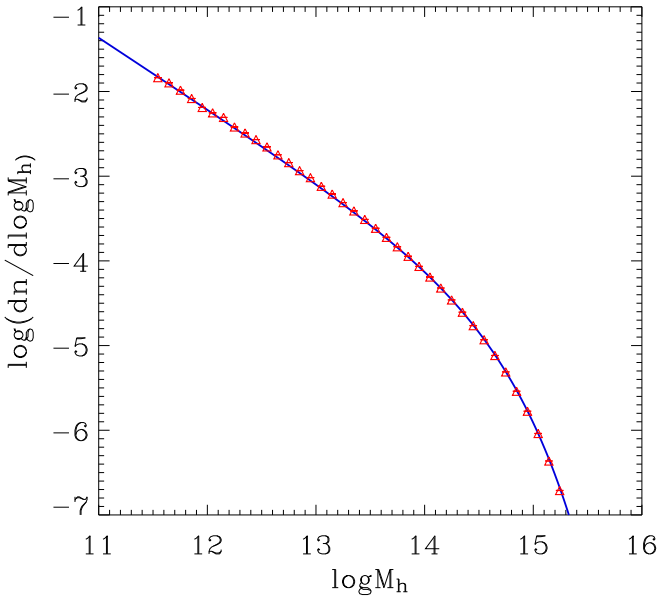

Figure 4.13: Differential halo mass function $\frac{dn}{dlogm}$. The red triangles are the average HMF derived from 100 JK volumes. The black solid line is a fit Schechter-like function (equation 4.27)

where δ_{sc} is the critical density required for spherical collapse given a cosmological model (Ω_z, Λ_z):

$$\delta_{sc}\left(\Omega_z, \Lambda_z, z\right) = \frac{\delta_{sc}(0)}{D(z)} \tag{4.31}$$

where $\delta_{sc}(z=0) = 1.686$ in an Einstein-de Sitter cosmology, and $\delta_{sc}(z=0) = 1.673$ in the MICE cosmology.

Manera et al. 2010 also suggested an expression for the MF depending on the variable ν:

$$\nu f_W(\nu) = \frac{M_h}{\rho_b}\frac{dn(M_h)}{d\ln M_h}\frac{d\ln M_h}{d\ln \nu} = \frac{M_h}{\rho_b}\frac{dn(M_h)}{d\ln \nu} = A^{'}\left[1 + b^{'}(c^{'}\nu)^{-a^{'}}\right]\exp\left(-c^{'}\nu/2\right) \tag{4.32}$$

In order to derive the MICE parameters of halo bias we use the MF expressions from Crocce et al. 2010 and Manera et al. 2010. We use equation 4.16, equation 4.32 and the relation between ν and σ (eq. (4.30)) to get equation 4.33:

$$\nu f_W(\nu) = \frac{1}{2} f(\sigma, z) \tag{4.33}$$

Parameters are related in this way:

$$c = c^{'}\frac{\delta_{sc}^2}{2D^2} \qquad c^{'} = c\frac{2D^2}{\delta_{sc}^2} \tag{4.34}$$

$$a = -2a^{'} \qquad a^{'} = -\frac{a}{2} \tag{4.35}$$

$$A = 2A^{'}b^{'}\left(c^{'}\frac{\delta_{sc}^2}{D^2}\right)^{-a^{'}} \qquad A^{'} = \frac{1}{2}Ab \tag{4.36}$$

$$b = \frac{1}{b^{'}}\left(\frac{c^{'}\delta_{sc}^2}{D^2}\right)^{a^{'}} \qquad b^{'} = \frac{(2c)^{-\frac{a}{2}}}{b} \tag{4.37}$$

We divide the MICE GC run into 10^3 cubes of box-size $L_{box} = 307.2$ Mpc/h and we estimate the two-point correlation function of haloes, $\xi_h(r)$, from very small scales (~ 0.1 Mpc/h) up to intermediate scales (10-30 Mpc/h) using the following estimator for each cube:

$$\xi_h(r) = \frac{DD(r)}{RR(r)} - 1 \tag{4.38}$$

where $DD(r)$ refers to the number of pairs in the simulation and $RR(r)$ is the number of pairs as if haloes were spatially randomly distributed. We run a FoF code in the MICE GC run going down to $N_p^{min} = 10$ to form a halo. The corresponding minimum halo mass is $\log_{10}(M_{min}) = 11.34$. The maximum halo mass is $\log_{10}(M_{max}) = 15.70$, that corresponds to a number N_p of about 170K dark-matter particles. We estimate the two point correlation function for seven different halo mass thresholds and three halo mass bins. Table 4.2 shows the chosen halo mass ranges. We developed a C code to compute the two point correlation function (see Appendix A for more details.)

$\geqslant N_p$	$\geqslant \log_{10} M_h$	$N_{p1} \geqslant N_p > N_{p2}$	$\log_{10} M_{h1} \geqslant \log_{10} M_h > \log_{10} M_{h2}$
10	11.3	10 to 39	11.3 to 12.0
21	11.7	39 to 359	12.0 to 13.0
39	12.0	359 to 3443	13.0 to 14.0
179	12.7		
359	13.0		
1728	13.7		
3443	14.0		

Table 4.2: Halo mass thresholds and halo mass bins.

Figure 4.14 shows the average of $\xi_h(r)$ for the different halo mass thresholds taking into account the 10^3 volumes. Errors bars show the error on the mean value between volumes. The different colours correspond to different halo mass thresholds. The first three halo mass thresholds (up to $\log_{10}(M_h) \geqslant 12.0$) have the same amplitude. More massive haloes show a faster increase in the amplitude, and they also show larger errors due to the small number of massive objects in comparison with less massive haloes. One can easily notice the halo exclusion effect (see section 3.1.1). There are no pairs at very small distances. The scale, at which the correlation function turns, approximately shows the size of the haloes. It depends on their

masses. The more massive the halo is, the larger its radius becomes. For comparison, the linear two-point correlation function of the MICE simulations (calculated as the FT of the linear MICE power spectrum) is shown as a black dashed line.

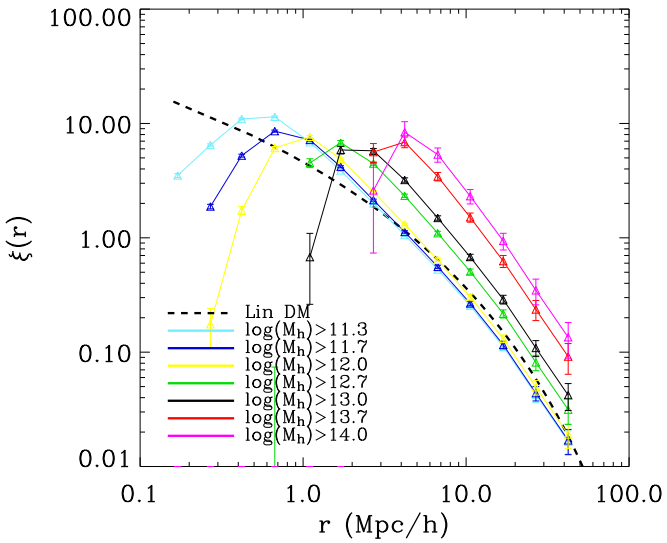

Figure 4.14: Mean value of the real space two-point correlation function for seven different halo mass thresholds for the 10^3 volumes with box-size $L_{box} = 307.2$ Mpc/h. The error bars are the errors on the mean value between volumes. The dashed black line is the two-point correlation function of the linear MICE correlation function calculated as the FT of the linear MICE power spectrum.

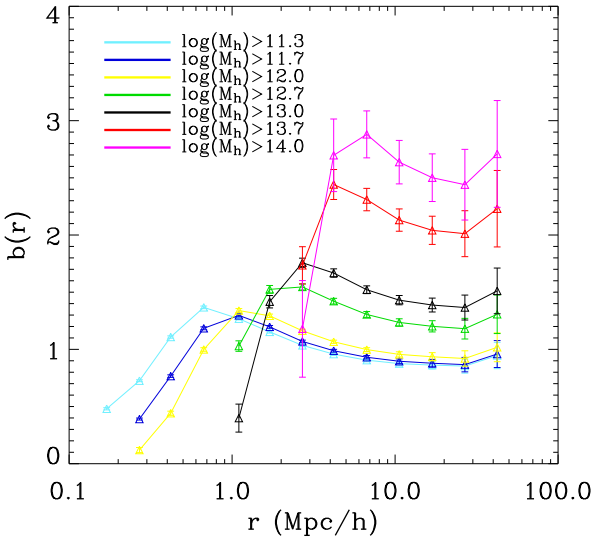

Figure 4.15: Scale dependent linear halo bias for seven different halo mass thresholds.

Figure 4.16 shows the average of $\xi_h(r)$ for the different halo mass bins. The errors

bars show again the error on the mean value between volumes. The different colours correspond to different halo mass bins. There are almost no difference between the first two low halo mass bins. However the more massive bin shows a considerable larger amplitude. The black dashed solid line is the linear two-point correlation function used in MICE simulations.

We estimate the linear halo bias as a function of the scale, $b_h^{Lin}(r)$, following equation 4.28 and using the linear MICE correlation function calculated as the FT of the linear MICE power spectrum. Figure 4.15 and figure 4.17 show the scale dependent halo bias for halo mass thresholds and halo mass bins respectively. The error is larger at large scales because the less number of pairs at those distances. The value of the bias at large scales does not converge into a constant because of the so-called integral constraint effect. Because of the way we divide the simulation into little boxes not all the scales are well represented in the correlation function.

In order to compute the halo bias as a function of the mass we have estimated the value of the bias at large scales. We have chosen the values at $r = 16.9$ and $r = 26.8$. Figures 4.18 and 4.19 show the large scale halo bias as a function the mass for different halo mass thresholds and bins using the value of the correlation function at $r = 16.9$ and $r = 26.8$, respectively. The black solid line is the expression of the halo bias given by Manera et al. 2010 and using the parameters found by Crocce et al. 2010 for the MICE simulations. The blue solid line is the cumulative bias derived by integrating equation 4.29:

$$b_h(\geqslant M_h) = \frac{\int_{M_h}^{\infty} \frac{dn}{dM'_h} b_h dM'_h}{\int_{M_h}^{\infty} \frac{dn}{dM'_h} dM'_h} \tag{4.39}$$

The derivation of the large scale halo bias from MICE simulations and the analytical expression from Manera et al. 2010 agree quite reasonably.

4.5 Summary

Dark-matter N-body simulations, which are used in this work to generate the mock galaxy catalogues, have been introduced and characterized in this chapter. We have briefly described the history and the basics of N-body simulations. The importance of computational advances has been stressed, as well as some improvements in the algorithms to compute the gravitational force between particles.

We have described and analyzed in particular the MICE simulations, which are essential for the process of building our mock galaxy catalogues. The MICE simulations assume a flat concordance ΛCDM model with parameters $\Omega_m = 0.25$, $\Omega_\Lambda = 0.75$, $\Omega_b = 0.044$, $h = 0.7$ and the spectral index $n_s = 0.95$. The linear power spectrum is normalized to yield $\sigma_8 = 0.8$ at $z = 0$, and the initial conditions are given by either using the Zel'dovich approximation or $2nd$ order Lagrangian Perturbation Theory (2LPT). The MICE simulations have been already used to study different aspects of galaxy clustering, such as LRG clustering, weak gravitational lensing, BAO, redshift space distortions, etc. The whole set of simulations has been presented. We have introduced and derived some of the ingredients for generating the mock galaxy

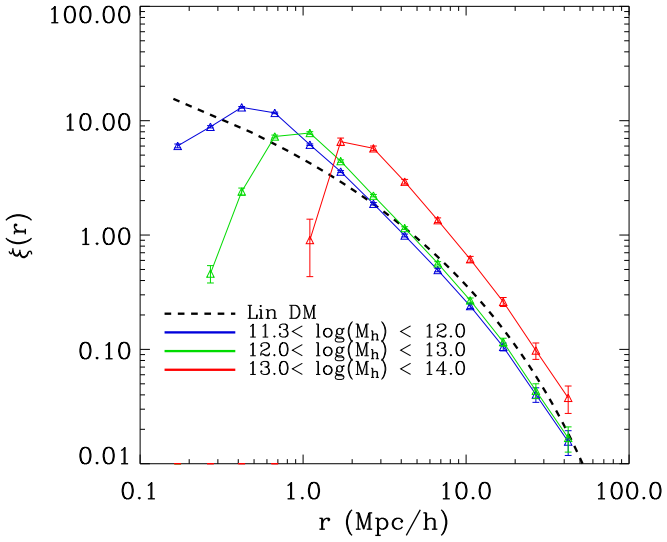

Figure 4.16: Mean value of the real space two-point correlation function for three different halo mass bins for the 10^3 volumes with box-size = 307.2 Mpc/h. Error bars are the errors on the mean value between volumes. The dashed black line is the two-point correlation function of the linear MICE correlation function calculated as the FT of the linear MICE power spectrum.

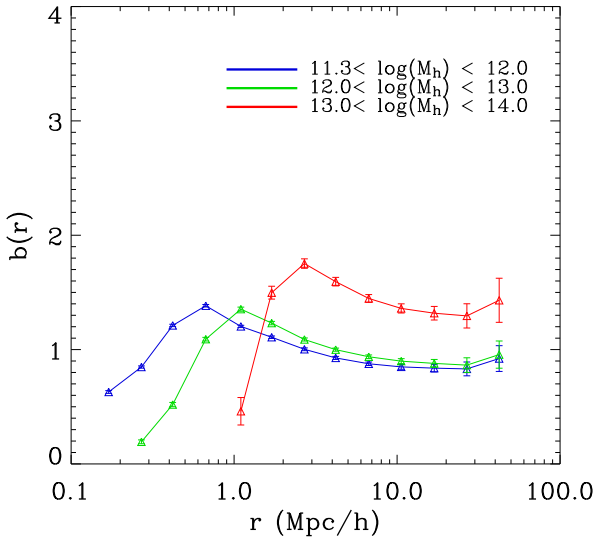

Figure 4.17: Scale dependent linear halo bias for seven different halo mass bins.

catalogues, such as the halo mass function, the two-point correlation function and the halo bias.

The linear MICE two-point correlation function is calculated as the FT of the linear MICE power spectrum. The halo catalogue extracted from the MICE simula-

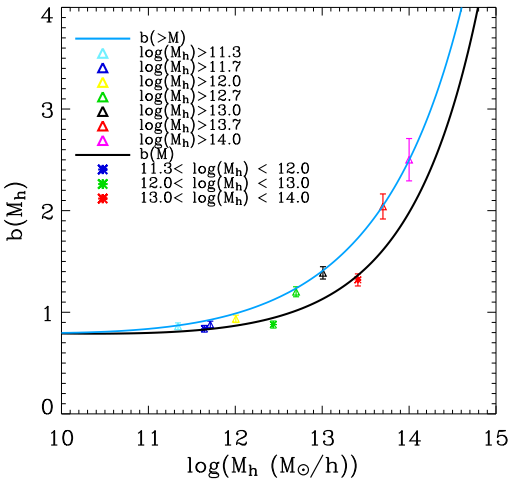

Figure 4.18: Linear bias at large scales for the halo mass thresholds and halo mass bins using the value of the correlation function at $r = 16.9$ Mpc/h. The black solid line is the expression of the halo bias given by Manera et al. 2010 and using the parameters found by Crocce et al. 2010 for the MICE simulations. The blue solid line is the cumulative bias derived by integrating equation 4.29.

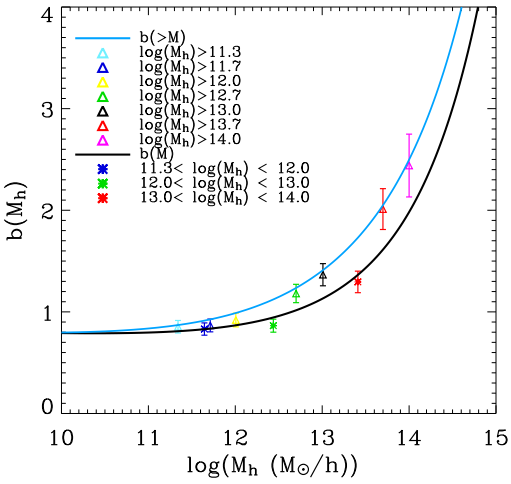

Figure 4.19: Idem as figure 4.18 but using the value of the correlation function at $r = 26.8$ Mpc/h.

tions has been described. Haloes are derived using a FoF code with a value of b=0.2. We have used expressions given by Crocce et al. 2010, which characterized the mass function of MICE simulations. We used also the analytical expression for the halo bias found by Manera et al. 2010 and we compare it to the MICE simulations. We have derived the errors of the halo mass function of the MICE GC run using the Jack-knife re-sampling method. A comparison against Poisson errors is done. In order to derive the halo bias we derived the correlation function of haloes dividing

the whole simulation volume into 10^3 non-overlapping cubes of box-size $L_{box} = 307.2$ Mpc/h. We showed the correlation function for different halo mass threshold and halo mass bin samples and we estimated the halo bias using the linear MICE correlation function of the dark matter density field (derived as the FT of the linear MICE power spectrum). The results show a very good agreement between the halo bias derived from simulations and the analytical expressions from Manera et al. 2010.

During the process of generating the catalogues we have used the MICE3072 and the Grand Challenge runs. Both simulations have a box-size $L_{box} = 3072 Mpc/h$. The first one contains 2048^3 particles and the MICE GC run has 8 times more particles and therefore a better mass resolution. The minimum luminosity of the mock galaxy catalogue depends strongly on the mass resolution of the simulation. We will see in section 5.3.3 of chapter 5 that the minimum luminosity of the mock galaxy catalogue decreases ~ 2 orders of magnitude (from $M_r \sim -21.0$ to $M_r \sim -19.0$) when going from the mass resolution of the MICE3072 run to the MICE GC one.

Chapter 5

Building mock galaxy catalogues

This chapter describes the process of producing mock galaxy catalogues populating dark-matter N-body simulations. We want the catalogue to follow some global properties of the observed galaxy population, such as the luminosity function, the colour-magnitude diagram and the clustering as a function of luminosity and colour. All observational data come from the Sloan Digital Sky Survey (SDSS). In particular we use the luminosity function in the $^{0.1}$r band derived by Blanton et al. 2003 using a catalogue of 147986 galaxy redshifts and fluxes, the galaxy clustering as a function of the luminosity and colour derived by Zehavi et al. 2011, and the colour-magnitude diagram from the NYU DR7 catalogue.

Our idea, since the beginning, was to be as simple as possible. Before getting to the last version of the mock galaxy catalogue, previous and different attempts, including a mock galaxy catalogue built by using a different approach than the HOD model, have been developed. The first section of the chapter describes the generation and reproduction of the mock galaxy catalogue presented by the Institut de Cienciès de l'Espai (ICE) as a collaborating group in the "Large Scale Structure Simulation Challenge" of the Dark Energy Survey (DES) within the Galaxy Clustering Working Group. The second section describes the algorithm presented by Skibba & Sheth 2009 that extends the halo occupation distribution model and include colours into the galaxies, which is a previous step to include spectral energy distribution to the galaxies. The algorithm also reproduces the colour-magnitude diagram and the clustering as a function of colour as well as the luminosity function and the clustering as a function of luminosity. We show the results and issues we found using the method of Skibba & Sheth 2009 by populating halo catalogues extracted from the MICE simulations. The second section also describes the attempt to analytically compute two HOD parameters using the MICE simulations. The third section shows the derivation of three HOD parameters by generating a grid of 600 catalogues populating with galaxies a halo catalogue with a box-size $L_{box} = 307.2$ Mpc/h extracted from the MICE GC run. In section 4 we describe the method we use in order to produce a *unique* mock galaxy catalogue that follows all the observed galaxy properties previously mentioned at the same time. It also shows the results we obtain when following the method populating again a halo catalogue with a box-size $L_{box} = 307.2$ Mpc/h of the MICE GC run.

5.1 DES LSS Challenge I

The first attempt to generate a mock galaxy catalogue was done to participate in the "LSS Simulation Challenge" of the Dark Energy Survey within the Galaxy Clustering Working Group. The working group issued a challenge with the following rules: the parent N-body dark-matter simulation must follow the concordance model, the simulation corresponds to a snapshot at $z = 0$, and the box-size must be greater or equal 1 Gpc/h ($L_{box} \geqslant 1$ Gpc/h). The mock galaxy catalogue has to follow a determined redshift distribution, and contains a total number of galaxies, N_{gal}, between 10^7 and 10^8.

The idea of the challenge was to generate a mock galaxy catalogue that simulates a photo-z catalogue of an octant of the sky ($0° < \mathrm{RA} < 90°$ and $0° < \mathrm{DEC} < 90°$) with the correct number density in a redshift range $0.2 < z < 1.4$. The final purpose was to roughly simulate the redshift measured using photometry. The plan was to include two different photometric redshift precisions, $\sigma_z = 0.01(1 + z)$ and $\sigma_z = 0.03(1 + z)$, and therefore we generated two catalogues.

As mentioned, the mean number density of particles in the octant in comoving coordinates (assuming the mean density of particles is constant) had to follow a particular distribution for the challenge. It was given by the following expression:

$$dN_{dm} = \bar{\rho}_r dV(r) = \bar{\rho}_r \frac{1}{8} 4\pi r^2 dr = \bar{\rho}_r \frac{\pi}{2} r^2 dr \tag{5.1}$$

Therefore, using the relation between the redshift and the comoving-distance,

$$dr = \frac{cdz}{H(z)} \tag{5.2}$$

yields,

$$\frac{dN_{dm}}{dz} = \frac{dN_{dm}}{dr}\frac{dr}{dz} = \frac{dN_{dm}}{dr}\frac{c}{H(z)} \tag{5.3}$$

In order to generate the mock galaxy catalogue we use one of the MICE runs. In particular the snapshot at $z = 0$ with $N_{dm} = 2048^3$ and $L_{box} = 3072$ Mpc/h. In our case no replication is necessary to go up to $z \sim 1.4$ because L_{box} corresponds to $z \sim 1.4$. Figure 5.1 shows a slice of the simulation in comoving coordinates where it is possible to visualize the large scale structures formed, such as filaments, voids and overdensities. It is a z-axis section of thickness 96 Mpc/h.

In particular for the challenge the distribution of galaxies as a function of redshift has to follow the expression:

$$\frac{dN_{gal}}{dz} = 1.5 N_{total} \frac{z^2}{0.5} \exp\left[-\left(\frac{z}{0.5}\right)^{1.5}\right] \tag{5.4}$$

As already said, the mock galaxy catalogue must have a total number of "particles" (or galaxies) between $10^7 - 10^8$. We normalize the distribution in order to have the needed number of galaxies. We derive the probability for a dark-matter particle "being" a galaxy depending on its comoving distance, r, which is given by the expression:

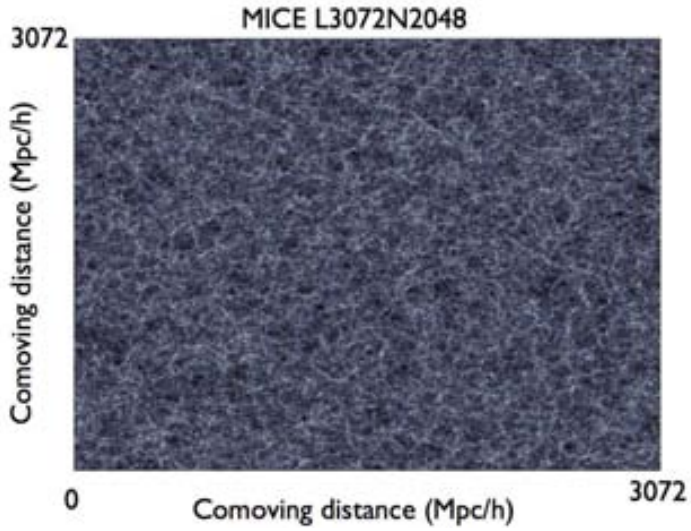

Figure 5.1: Complete z-axis section of the N2048L3072 MICE run with a thickness of 96 Mpc/h. Horizontal and vertical axis refer to the x and y axis respectively of the simulating box in comoving distance in units of Mpc/h. The pixel resolution is 5 Mpc/h. Each pixel shows the mean number density of dark-matter particles. The whiter the pixel appears, the largest the mean number density is.

$$P_{dm}(N_{gal}|r) = \frac{N_{gal}(r)}{N_{dm}(r)} \tag{5.5}$$

We divide the volume in constant comoving-distance bins of width Δr and therefore the probability of a dark-matter particle of "being" a galaxy in the redshift range (z_1, z_2) or the equivalent comoving distance range (r_1, r_2) is given by the expression:

$$P_{dm}(z_1, z_2) = \frac{\int_{z_1}^{z_2} \frac{dN_{gal}}{dz} dz}{\int_{z_1}^{z_2} \frac{dN_{dm}}{dr} \frac{c}{H(z)} dz} = P_{dm}(r_1, r_2) = \frac{\int_{r_1}^{r_2} \frac{dN_{dm}}{dz} \frac{H(z)}{c} dr}{\int_{r_1}^{r_2} \bar{\rho}_r \frac{\pi}{2} r^2 dr} \tag{5.6}$$

In practice we draw a random number u_0 uniformly distributed between 0 and 1 for each dark-matter particle and if $u_0 < P_{dm}(r)$ we assign one galaxy to that dark-matter particle.

Figure 5.2 shows the expected and the actual distribution of galaxies in comoving coordinates. Red solid line is the predicted distribution and the blue noisy line is the actual distribution of galaxies.

Figure 5.3 shows the spatial distribution of galaxies in a slice through the simulation of the distribution of galaxies. The top panel shows the distribution of galaxies with no error in the galaxy positions. The bottom panels show the distortions in the distribution when photo-z are introduced. Clusters and voids in the top panel become blurred in the left bottom panel, and completely disappear in the right one. Only slightly elongated patterns in the line-of-sight hold.

The effects of the photo-z errors are (also) found in the two-point correlation function. Figure 5.4 shows the two point correlation function as a function of the radial or line-of-sight distance, π (in vertical axis), and the perpendicular distance,

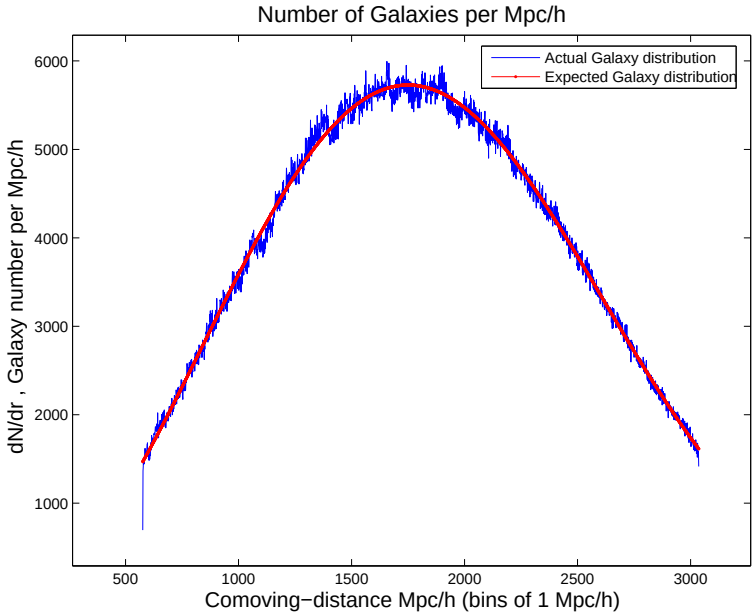

Figure 5.2: Actual and predicted galaxy distributions. Red solid line refers to the predicted galaxy distribution in comoving coordinates and the blue solid (noisy) line is the actual distribution of the mock galaxy catalogue.

σ (in horizontal axis). It is usually called the sigma-pi correlation function, $\xi(\sigma, \pi)$. Figure 5.4 comes from the MICE webpage. They was produced by Anna Cabré. The left panel shows the sigma-pi correlation function for the true galaxy radial distances. The red colour corresponds to large values of the correlation ($\xi > 1$) and the blue colour to negative values ($\xi < 0$). The BAO scale can be seen as dark yellow at the scale ~ 100 Mpc/h. The right panel shows the same plot but in this case a photo-z error of $\sigma(z) = 0.01(1 + z)$ is included. The BAO amplitude is now very difficult to see in the σ axis (as a yellow contour) and completely removed in the π direction. Horizontal stripes are due to redshift discretization ($\Delta z = 0.0002$).

5.2 HOD mock galaxy catalogues

This section describes the next attempt to create a mock galaxy catalogue by using a more sophisticated method than the previous one. The goal is to populate the MICE cosmological simulations using a HOD model. The simplest and faster way to proceed is to follow already developed algorithms. In particular we follow the algorithm presented in Skibba & Sheth 2009 (hereafter SS09). SS09's work is a very interesting approach since they present a method to introduce colours in the HOD, which is a previous step to introduce galaxy spectral energy distributions. Section 5.2.1 describes the SS09's algorithm which we have followed to produce mock galaxy catalogues. Section 5.2.2 describes the issues found when trying to match the correct luminosity function following the SS09's algorithm. Section 5.2.3 shows the attempt of analytically constraining two HOD parameters using only two conditions, the

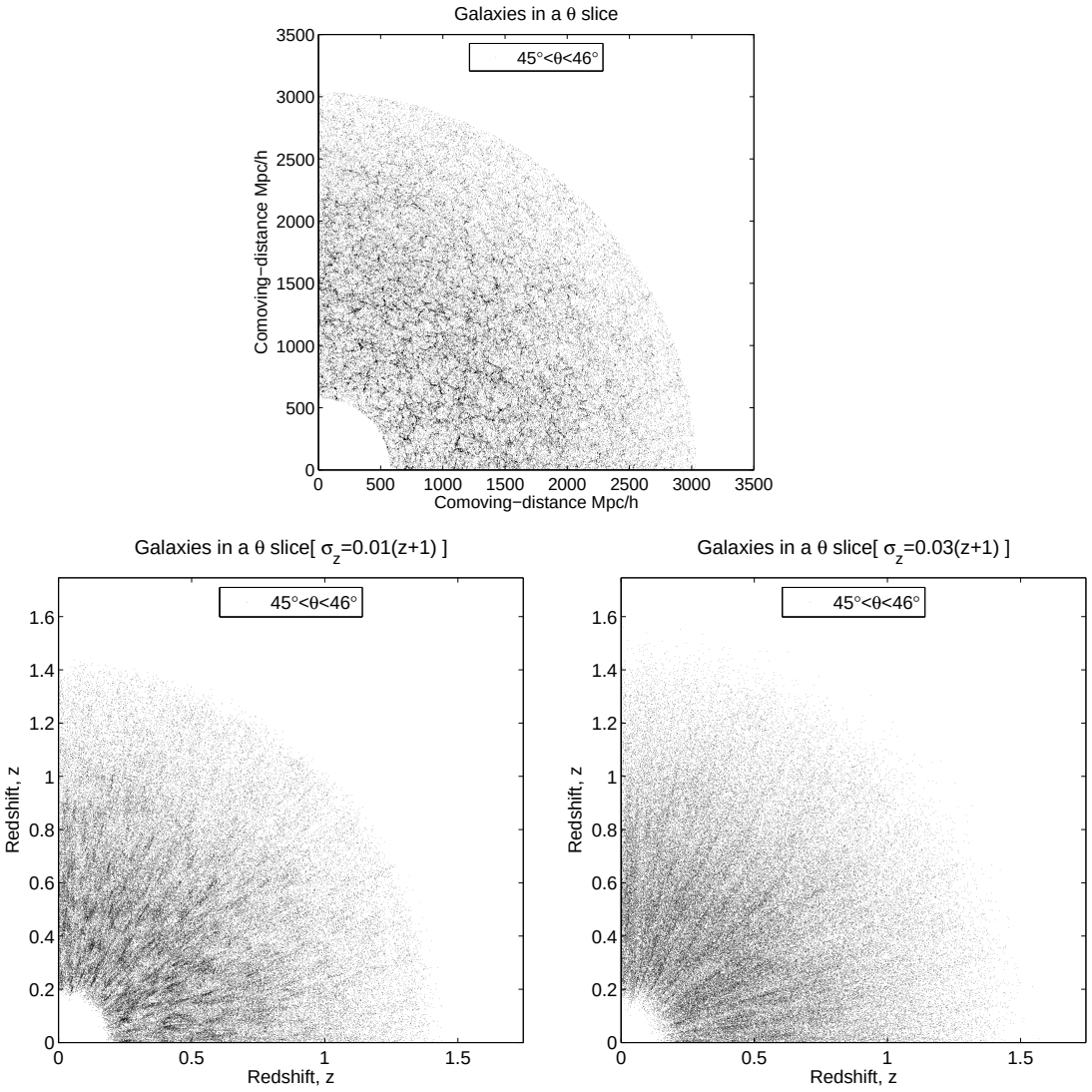

Figure 5.3: Slices through the simulation of the distribution of galaxies. The top panel, in comoving distances, without error in the galaxy positions. The bottom panels, in redshift space, include Gaussian errors in galaxy redshifts.

observed mean number density of galaxies and the average linear large scale galaxy bias at a certain scale.

5.2.1 The Skibba and Seth's algorithm. Colours in the halo model approach

Skibba & Sheth 2009 (SS09) presented an algorithm to create mock galaxy catalogues. These catalogues are constructed to have the correct luminosity function as well as the correct luminosity dependence of the two-point correlation function. In addition, SS09 introduced galaxy colours in the framework of the HOD. Their model also reproduces the galaxy clustering dependence on colour. In their method they

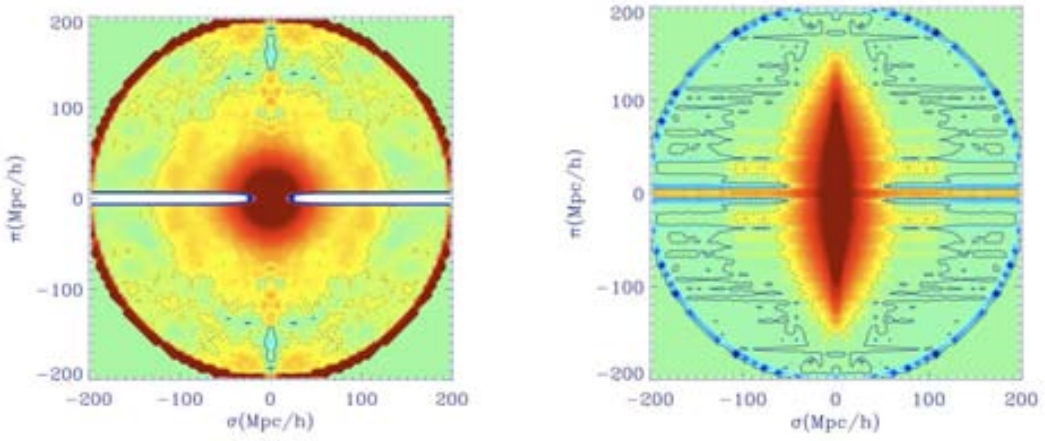

Figure 5.4: Two point correlation function of radial or line-of-sight distance, π (in vertical axis), and perpendicular distance, σ (in horizontal axis). Left panel shows the sigma-pi correlation function for the true galaxy radial distances and the right panel shows exactly the same with a photo-z error of $\sigma(z) = 0.01(1+z)$. Both panels come from the MICE webpage (http://maia.ice.cat/mice/) produced by Anna Cabré.

assume that: (i) the bimodality of the colour distribution at fixed luminosity is independent of the halo mass and, (ii) the fraction of satellite galaxies which populate the red sequence increases with luminosity.

Colour and absolute magnitude are two of the most useful variables to characterized galaxies. Different galaxy types, elliptical, lenticular and spiral, show different relationships between the colour and the absolute magnitude. When one considers all galaxy types together, the colour function can be approximated by two Gaussian functions. Baldry et al. 2004 reported this trend for the distribution of rest frame $u-r$ colour at a fixed r-magnitude, and e.g., Blanton et al. 2005, found the same for the distribution of $g-r$ colour. SS09 provide simple equations to describe this dependence[1]:

$$\begin{aligned} \langle g-r|M_r\rangle_{red} &= 0.932 - 0.032(M_r+20.0) \\ rms\,(g-r|M_r)_{red} &= 0.07 + 0.01(M_r+20.0) \end{aligned} \tag{5.7}$$

$$\begin{aligned} \langle g-r|M_r\rangle_{blue} &= 0.62 - 0.11(M_r+20.0) \\ rms\,(g-r|M_r)_{blue} &= 0.12 + 0.02(M_r+20.0) \end{aligned} \tag{5.8}$$

Figure 5.5 shows the bimodal distribution of $g-r$ colour using the data from the NYU DR7 catalogue. The red and blue lines refer to the mean values of the red and blue sequences given by the expressions 5.7 and 5.8 respectively.

We follow, in principle, the SS09's algorithm to populate the MICE simulations. The SS09's algorithm is made up of three steps. First, the galaxy luminosities are generated. Second, they are used to generate colours. And, third, galaxies are spatially distributed. The method used to generate luminosities is not relevant. One can

[1]Restframe magnitudes in SS09's work are associated with SDSS filters shifted to $z = 0.1$.

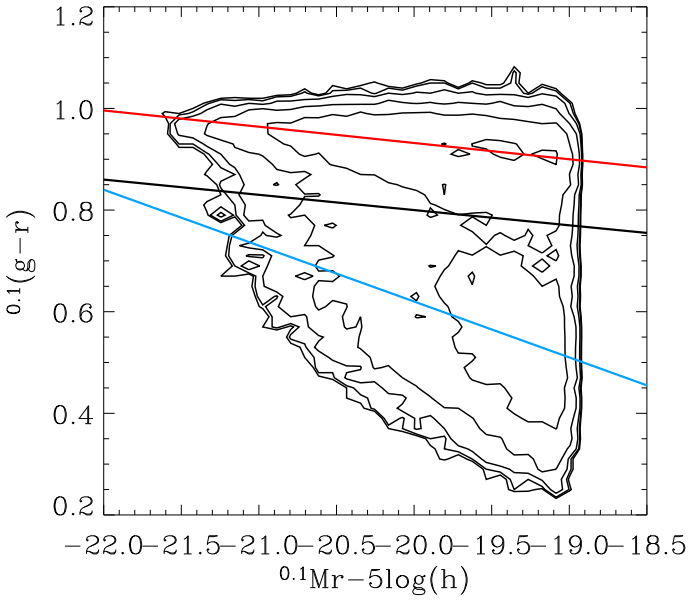

Figure 5.5: Colour-magnitude diagram in the NYU DR7 catalogue for the $M_r < -19.5$ volume limited sample. The red and blue lines show the mean values of the red and blue sequences given by SS09 following equations 5.7 and 5.8 respectively. Black line is a crude cut which some authors use to separate galaxy population into the red and blue sequences.

use either the Conditional Luminosity Function (CLF), the Halo Occupation Distribution (HOD) or the SubHalo Abundance Matching (SHAM) methods (described in previous chapters). SS09 proposed a complete algorithm to build mock galaxy catalogues and these are the steps to follow:

(i) Specify a minimum luminosity, L_{min}, smaller than the minimum luminosity one wishes to study.

(ii) Select the subset of haloes in the simulation that have a halo mass $M_h \geqslant M_{min}$, where M_{min} is the mass of the halo that corresponds to L_{min}. The way to relate galaxy luminosities and halo masses is given by the following equation:

$$\left(\frac{M_h}{M_*}\right) \approx \exp\left(\frac{L_r}{L_*}\right) - 1 \tag{5.9}$$

where $M_* = 10^{12} h^{-1} M_\odot$ and $L_* = 1.12 \cdot 10^{10} h^{-2} L_\odot$. Expression 5.9 comes from the relation between the minimum halo mass, M_{min}, and the corresponding luminosity, L_{min}, found by Zehavi et al. 2005 (or Skibba et al. 2006) when they studied the SDSS galaxy clustering using the HOD approach[2]. Zehavi et al. 2005 (hereafter ZH05) computed the parameters that best match the luminosity function and the luminosity dependence of the galaxy clustering and SS09 derived an approximate relation between M_{min} and L_{min}. Table 5.1 shows the best-fit HOD parameters for luminosity threshold samples estimated by ZH05. Figure 5.6 shows the HOD

[2] M_{min}, M_1 and α are three free HOD parameters explained in chapter 3.

parameters as a function of the luminosity threshold galaxy samples from ZH05. The left panel shows M_{min} (blue squares), M_1 (red triangles) and the relation between M_{min} and L_{min} given by expression 5.9 (orange solid line) as a function of the luminosity threshold of the galaxy sample. The right panel shows the parameter α as a function of the luminosity threshold of the galaxy sample[3].

(iii) Every selected halo contains a central galaxy with a luminosity given by inverting the relation between halo mass and luminosity using the expression 5.9.

(iv) Specify a number of satellite galaxies per halo choosing an integer from a Poisson distribution with mean value $N_{sat}(> L_{min}|M_h)$, which is the mean number of galaxies with luminosity brighter than L_{min} that reside in a halo of mass M_h. In our case we use the expressions of ZH05 to derive the number of galaxies inside each dark-matter halo, a step function for the central galaxies and a simple power law for the mean value of the number of satellite galaxies[4]:

$$N_{gal}(> L_r|M_h) = N_{cen}(> L_r|M_h) + N_{sat}(> L_r|M_h) \tag{5.10}$$

$$N_{gal}(> L_r|M_h) = 1 + N_{sat}(> L_r|M_h) = 1 + \left[\frac{M_h}{M_1(L_r)}\right]^{\alpha} \quad \text{if} \quad M_h \geq M_{min}(L_r)$$

$$N_{gal}(M_h) = 0 \qquad \text{otherwise,} \tag{5.11}$$

(v) The luminosity of satellite galaxies is assigned by generating a random number u_0 uniformly distributed between 0 and 1, and finding that L_r for which

$$\frac{N_{sat}(> L_r|M_h)}{N_{sat}(> L_{min}|M_h)} = u_0, \tag{5.12}$$

this ensures that satellite galaxies follow the correct luminosity distribution.

(vi) Assign colours to central and satellite galaxies in order to follow the correct colour-magnitude diagram. SS09 showed that if one uses equation 5.7 to assign satellite colours, the results do not provide a very good agreement with observations. Instead SS09 proposed the next relation in which satellite galaxies have colours bluer than the red sequence at faint luminosities:

$$\langle g - r|M_r\rangle_{sat} = 0.83 - 0.08(M_r + 20.0) \tag{5.13}$$

which provides better agreement with the observations. The selection of the function is merely empirical. The implementation of SS09 to assign colours to satellites is generating a uniformly random number between 0 and 1, u_1, per satellite. If $u_1 \leqslant p(\text{red sat}|L_r)$, the satellite is drawn from the red sequence (equation 5.7). The colour of satellite galaxies is assigned to the blue population, i.e. equation 5.8, otherwise (it is important to remark that colours only depend on luminosity; the halo mass does not appear explicitly; the halo mass is implicitly included because galaxy luminosity depends on the halo mass). $p(\text{red sat}|L_r)$ is given by the following equation:

[3] ZH05 do not show the error bars of the HOD parameters in their work.

[4] Galaxy luminosities are given in the r-band.

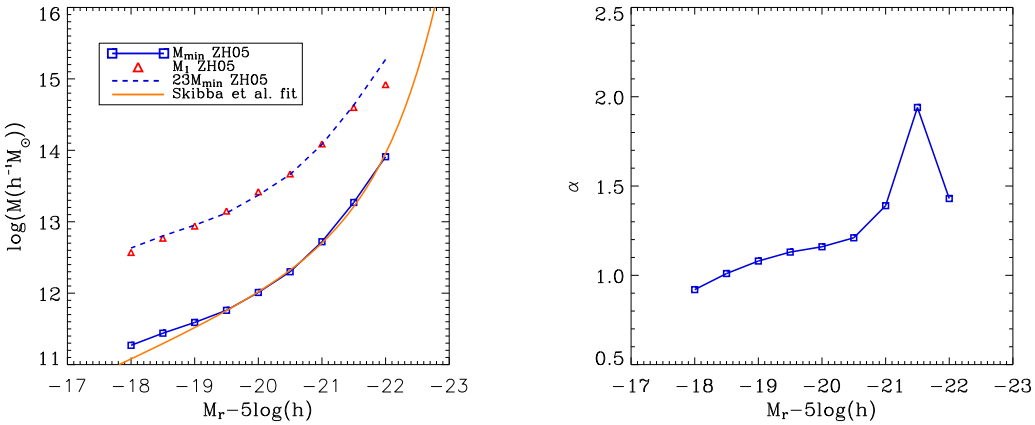

Figure 5.6: HOD parameters as a function of the luminosity threshold of galaxy samples. The left panel shows M_{min} (blue squares) and M_1 (red triangles). The dashed blue line is the $M_{min} - M_r$ curve scaled by a factor of 23, and the orange solid line is the SS09's fit. The right panel shows the HOD parameter α.

$$p(\text{red sat}|L_r) = \frac{\langle c|L_r\rangle_{sat} - \langle c|L_r\rangle_{blue}}{\langle c|L_r\rangle_{red} - \langle c|L_r\rangle_{blue}} \tag{5.14}$$

where all terms can be derived from data except $\langle c|L_r\rangle_{sat}$ which, as mentioned, is set by SS09 to follow equation 5.13. To assign colours to central galaxies, SS09 use another uniformly random number between 0 and 1, u_2. If $u_2 > f_{\rm blue}(L_r)/f_{\rm cen}(L_r)$, then the central galaxy *belongs* to the red sequence, and therefore one uses equation 5.7 to generate a red colour for this galaxy. Otherwise, it *belongs* to the blue sequence and one uses equation 5.8 instead. $f_{\rm blue}(L_r)$ is the fraction of objects in the blue component, and can be derived from observations. SS09 provided the following expression:

$$f_{\rm blue}(L_r) \approx 0.46 + 0.07(M_r + 20) \tag{5.15}$$

The fraction of central objects as a function of the luminosity in the catalogue, $f_{\rm cen}(L_r)$, can be derived from the HOD model. SS09, in the appendix A, showed a derivation of $f_{\rm cen}(L_r)$. It can be expressed as a function of $n_{sat}(L_r)/n_{cen}(L_r)$ for which SS09 proposed the following reasonable fit using their actual halo model values:

$$\frac{n_{sat}(L_r)}{n_{cen}(L_r)} \approx 0.35\,[2 - \text{erfc}\,[0.6\,(M_r + 20.5)]] \tag{5.16}$$

(vii) Place a central galaxy at the center of its halo and distribute satellite galaxies in the halo following a NFW density profile.

M_r^{max}	$\log_{10} M_{min}$	$\log_{10} M_1$	α
-22.0	13.91	14.92	1.43
-21.5	13.27	14.60	1.93
-21.0	12.72	14.09	1.39
-20.5	12.30	13.67	1.21
-20.0	12.01	13.42	1.16
-19.5	11.76	13.15	1.13
-19.0	11.59	12.94	1.08
-18.5	11.44	12.77	1.01
-18.0	11.27	12.57	0.92

Table 5.1: Best-fit HOD parameters for luminosity threshold samples from ZH05.

SS09 claimed that following their algorithm the resulting mock galaxy catalog is constructed to have the correct luminosity function as well as the correct luminosity dependence of the galaxy two-point correlation function. SS09 also pointed out that there are some assumptions in their algorithm to build mock galaxy catalogues. They assume haloes are spherical which are not. The number of galaxies in a halo, the spatial distribution of galaxies within a halo and the galaxy luminosity only depend on halo mass. And the same is true for colours. Therefore the environmental dependence only comes from the halo population dependence on environment. SS09 also point that they are not taking into account the halo formation history, i.e. the number of galaxies and their stellar populations are not correlated with the formation history.

We have described the starting algorithm we have used to generate the galaxy mocks. The next sections will show the issues we have found during the process, their possible causes and how we have tried to solve them.

5.2.2 Matching the luminosity function

If one follows the SS09's algorithm, the generated mock galaxy catalogue has to match the correct luminosity function. In particular galaxy luminosities are assigned to fit the luminosity function at $z = 0.1$ measured by Blanton et al. 2003 using a catalogue of 147986 galaxy redshifts and fluxes from the SDSS. The luminosity function gives the space density of galaxies within a luminosity bin. We use the best-fit Schechter function in the ${}^{0.1}r$-band. It is given by the expression:

$$\phi(M_r) = \frac{dn}{dM_r} = \frac{2}{5}\phi^* \ln(10) \left(10^{\frac{2}{5}(M_r^* - M_r)}\right)^{\alpha+1} \exp\left(-10^{\frac{2}{5}(M_r^* - M_r)}\right) \tag{5.17}$$

where $\phi^* = (1.49 \pm 0.04) \times 10^{-2} h^3 Mpc^{-3}$, $M_r^* - 5\log_{10}(h) = -20.44 \pm 0.01$, and $\alpha = -1.05 \pm 0.01$. Figure 5.7 shows the best Schechter function fit to the SDSS data in the ${}^{0.1}r$-band assuming $\Omega_0 = 0.3$ and $\Omega_\Lambda = 0.7$.

We follow the SS09's algorithm and populate with galaxies the halo catalogue extracted from the snapshot at $z = 0$ of the N2048L3072 MICE simulation. The mass of each dark-matter particle is $m_p = 23.418 \times 10^{10} M_\odot/h$. We choose the minimum

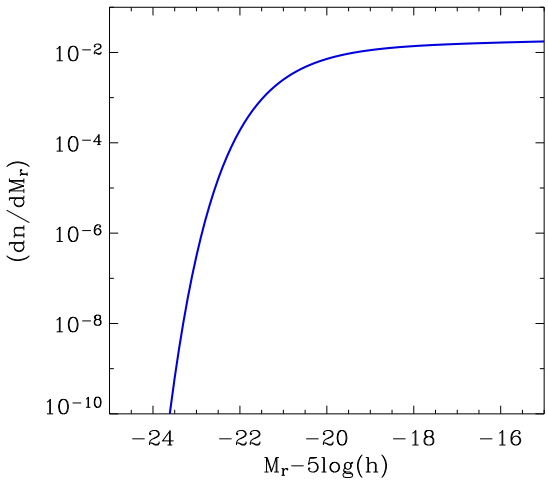

Figure 5.7: Best Schechter function fit to the SDDS data in the $^{0.1}r$ band.

number of dark-matter particles, $N_p^{min} = 20$, to form a halo, which correspond to a minimum halo mass $M_{min} = 390.7468 \times 10^{10} M_\odot/h$. HOD fitting to observed luminosity limited galaxy samples shows that at small halo masses the number of galaxies is well modeled by a step function. We decide that all haloes in our simulation will have, at least, one central galaxy at its center (instead of choosing a minimum luminosity we choose a minimum halo mass, which is actually equivalent given the recipe to relate halo mass and galaxy luminosity). The luminosity of central galaxies is given by inverting the relation between halo mass and luminosity given by equation 5.9. Therefore the luminosity of the minimum halo mass is $L_{min} = 1.78 \times 10^{10} L_{r_\odot}/h^2$, which corresponds to an absolute magnitude in the r-band $M_r - 5\log(h) = -20.87$.

The mass of each halo is computed counting the number of dark-matter particles it contains, N_p. We estimate their mass taking into account the Warren mass correction:

$$M_h^W = m_p(N_p(1 - N_p^{-0.6})) \tag{5.18}$$

Equation 5.18 corrects a systematic problem that FoF halo masses suffer when haloes are sampled by relatively small number of particles. If a given halo has too few particles, its FoF halo mass turns out to be systematically too high. The fact that the number of particles that form a halo is discrete causes that galaxy luminosities (specially close to L_{min}) are also discrete. We introduce a scatter of half a dark-matter particle in every halo mass equation to solve this issue. We generate a uniformly distributed random number, $-(1/2) > u_1 > (1/2)$, which we multiply by a linear weighted dark-matter particle mass depending on the Warren correction[5]:

$$M_h = M_h^W + \Delta M_h \tag{5.19}$$

[5]Note: It is easier to introduce the scatter in the number of particles and then apply the Warren mass correction.

$$\Delta M_h = \begin{cases} \left[(N_p+1)\left(1-(N_p+1)^{-0.6}\right) - N_p\left(1-N_p^{-0.6}\right)\right] m_p & \text{if } u_1 > 0 \\ \left[N_p\left(1-N_p^{-0.6}\right) - (N_p-1)\left(1-(N_p-1)^{-0.6}\right)\right] m_p & \text{if } u_1 < 0 \end{cases} \tag{5.20}$$

Before generating the galaxy catalogue one can analytically compute the contribution of the central galaxies to the total luminosity function. The luminosity function is formed by two components, central and satellite galaxies. If we assume that every halo has one galaxy at its center and one knows the halo mass function and the relation between halo mass and central luminosity (equation 5.9), one can derive the central galaxy luminosity function.

$$\left(\frac{dn}{dM_r}\right)_{cen} = \left(\frac{dn}{dM_h}\right)\left(\frac{dM_h}{dM_r}\right) \tag{5.21}$$

more explicitly,

$$\left(\frac{dn}{dM_r}\right)_{cen} = \left(\frac{dn}{d\log M_h}\right)\left(\frac{d\log M_h}{dM_h}\right)\left(\frac{dM_h}{dL_r}\right)\left(\frac{dL_r}{dM_r}\right) \tag{5.22}$$

$$\left(\frac{dn}{dM_r}\right)_{cen} = \left(\frac{dn}{d\log M_h}\right)\left(\frac{1}{\ln(10)M_h}\right)\left(\frac{dM_h}{dL_r}\right)\left(\frac{dL_r}{dM_r}\right) \tag{5.23}$$

where

$$\left(\frac{dM_r}{dL_r}\right) = -\frac{2.5}{\ln(10)}\frac{1}{L_r} \tag{5.24}$$

and

$$\left(\frac{dM_h}{dL_r}\right) = \left(\frac{M_h^*}{L_r^*}\right)\exp\left(\frac{L_r}{L_r^*}\right) \tag{5.25}$$

We use the halo mass function fit derived by Crocce et al. 2010 (showed in section 4.4.1 of chapter 4 for the MICE simulations). Figure 5.8 shows the contribution of central galaxies to the total luminosity function (red line) in comparison to the luminosity function derived by Blanton et al. 2003 using a Schechter function to fit SDSS data (blue line). The halo mass-central luminosity relation does not provide a good agreement with observations. One can realize in the figure that neither the amplitude nor the shape agree. Looking at the very brightest part of the luminosity function in figure 5.8 (galaxies brighter than $M_r \sim -22.0$) one realizes that in order to match the observed luminosity function it is necessary to add satellite galaxies. Following the SS09's algorithm using the MICE simulations makes our catalogue to have satellite galaxies as the brightest galaxies. This is not what we want since we assume the brightest galaxies are, in general, central galaxies.

We would also like to highlight other different aspects about figure 5.8. First, we have used a snapshot of the simulation at $z = 0$, and the relation is derived to match the luminosity function at $z = 0.1$. Figure 5.9 shows the difference between the HMF at $z = 0$ and $z = 0.1$. There exists a little difference in the halo mass function for haloes more massive than $10^{13.5} M_\odot$. The reason is because there are more massive haloes at $z = 0$ than at $z = 0.1$ due to time evolution. We shift the MICE HMF from $z = 0$ to $z = 0.1$ but it does not produce a significant change in the fit of the luminosity function since the difference is not much in logarithmic scale

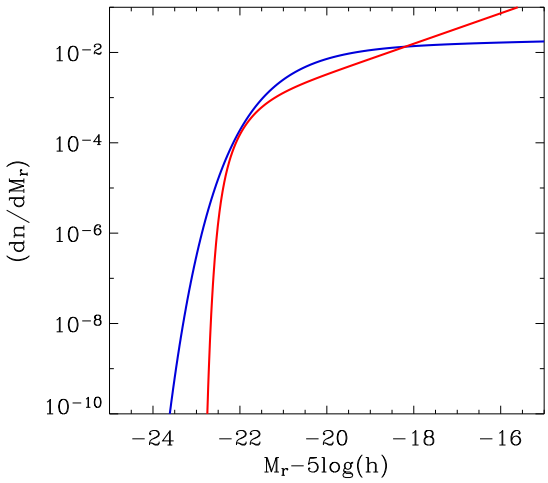

Figure 5.8: Contribution of central galaxies to the luminosity function using the halo mass - luminosity relation from equation 5.9 (red solid line) compared to the best Schechter function fit to the SDDS data (blue solid line).

and figure 5.8 is also plotted in logarithmic scale. Second, the assumed cosmological model when generating the MICE simulations is not the same as the model assumed when computing the luminosity function of SDSS (Blanton et al. 2003). Blanton et al. 2003 assume $\Omega_m = 0.3$ to measure distances, and MICE simulations assume $\Omega_m = 0.25$. This change in Ω_m has only a 1% effect on the distance-redshift relation even up to $z = 0.25$. Third, the luminosity - halo mass relation we used is proposed by SS09 (derived from results of ZH05) and the assumed cosmological parameters in their simulation are not the same as the one assumed in the MICE simulations. ZH05's cosmological parameters are $\Omega_m = 0.3$, $\sigma_8 = 0.9$ and $n_s = 1$. The parameter that can affect the most is σ_8. We correct for this issue by modifying the MICE halo masses. We derived a relation between halo masses for the two different cosmologies:

$$\log(M_h^{ZH05}) = \log(M_h^{MICE}) + 0.145 + 0.069 \arctan\left(1.25(\log(M_h^{MICE} - 14.0))\right) \tag{5.26}$$

Using their different cumulative halo mass functions (abundance matching technique). We compute both HMFs using first the CAMB code to obtain their power spectrum and then the genmf.f code[6] to generate their halo mass function. We apply the relation found to shift the MICE halo masses *into* the masses in the ZH05 cosmology. Figure 5.10 shows the difference in the HMF between the different cosmologies. Black and red lines refer to the ZH05 and MICE cosmologies, respectively. The mean value of the MICE GC run is shown as a reference (red triangles) and also its corresponding value once the mass is shifted to the ZH05 cosmology (black triangles).

We derived the corresponding central galaxy luminosity using equation 5.9. The

[6] http://icc.dur.ac.uk/ reed/genmf.html.

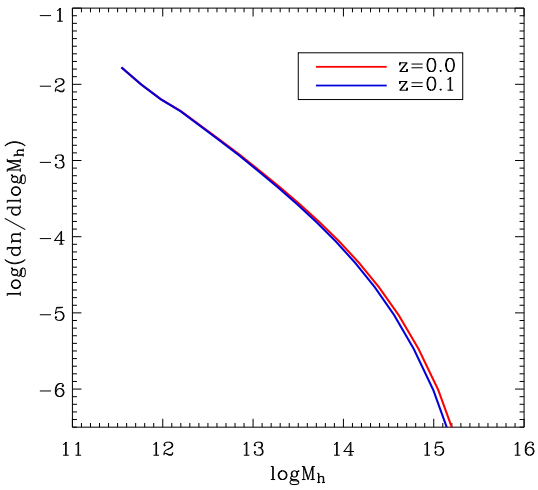

Figure 5.9: The MICE halo mass function at $z = 0$ (red line) and $z = 0.1$ (blue line) derived from a box with $L_{box} = 302.7$ Mpc/h of the MICE GC run.

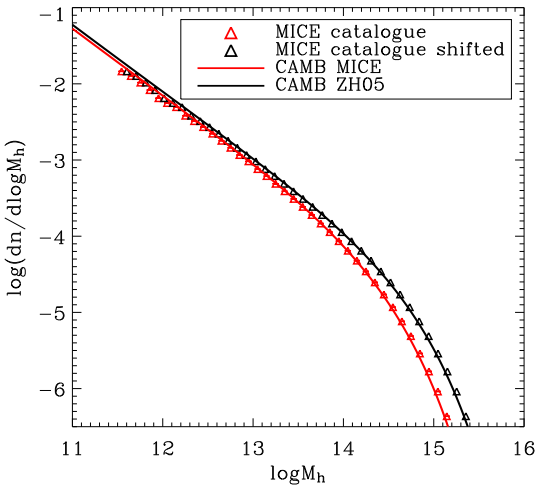

Figure 5.10: HMF of different cosmologies using CAMB.

result is shown in figure 5.11 as well as the original fit before correcting for the masses. We do not find a good agreement either.

We also explored the possibility of modifying the M_h^* and L_r^* parameters of equation 5.9 to match the SDSS luminosity function but we did not find the agreement we wanted. We think that it is necessary more free parameters in the relation between galaxy luminosity and halo mass to match the observed luminosity function using the MICE simulations.

Actually, all previous checks were done after generating a mock galaxy catalogue following SS09's algorithm and remarking that the resulting galaxy catalogue luminosity function does not match with observations. When generating the galaxy

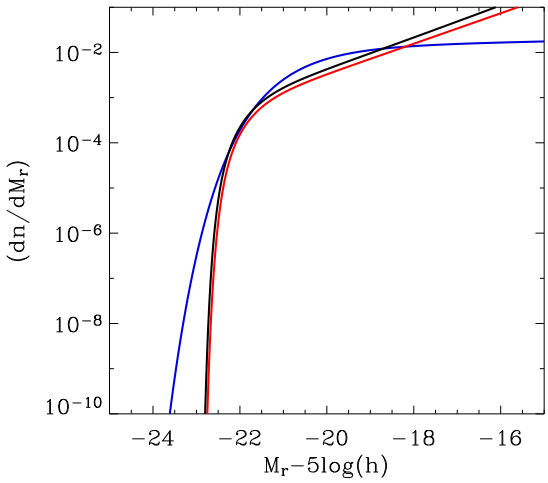

Figure 5.11: Contribution of central galaxies to the luminosity function using the halo mass - luminosity relation from equation 5.9 compared to the best Schechter function fit to the SDDS data (blue line) for both cosmologies, and therefore for two different halo mass functions: the one assumed in ZH05's work and the MICE cosmology.

catalogue we used ZH05's HOD parameters, M_{min}, M_1 and α. At this point we explore the possibility of deriving *our own* HOD parameters using the MICE simulations. As mentioned before we assume $\alpha = 1$ for simplicity (and also because previous works show that α is almost constant with respect to the luminosity and close to 1, $\alpha \sim 1 \neq \alpha(L)$). Therefore we only have to derive two parameters. The next section shows the process to analytically derive M_{min} and M_1.

5.2.3 Constraining M_{min} and M_1

In order to constrain M_{min} and M_1 (we assume $\alpha = 1$) we use two conditions, the observed mean number density of galaxies, and the average linear large scale galaxy bias at a certain scale. Both observational constraints come from the work of Zehavi et al. 2011 (hereafter ZH11) and depend on the absolute magnitude threshold of different galaxy samples. Tables 1 and 2 of ZH11's work show the mean number density of galaxies for galaxy samples corresponding to luminosity bins and luminosity thresholds respectively. Table 3 of ZH11 shows the HOD and other derived parameters (such as the galaxy bias) for luminosity threshold samples. They used volume-limited samples, extracted from a parent sample of $\sim$ 700000 galaxies over 8000 deg^2 up to redshift $z = 0.25$.

We assume that the mean number density of galaxies that reside in a halo of mass M_h is given by the expression 5.11 and therefore one can compute the number of galaxies per halo as a function of mass:

$$n_{cen}(> M_{min})(M_{min}) = \int_{M_h^{min}}^{M_h^{max}} \left(\frac{dn}{dM_h}\right) dM_h \tag{5.27}$$

$$n_{sat}(> M_{min})(M_{min}, M_1) = \int_{M_h^{min}}^{M_h^{max}} \left(\frac{dn}{dM_h}\right) \left(\frac{M_h}{M_1}\right)^{\alpha} dM_h \tag{5.28}$$

where we assume $\alpha = 1$. The first constraint is the mean number density of galaxies and therefore:

$$n_{obs}(> L_r) = n_{cen}(> M_{min})(M_{min}) + n_{sat}(> M_{min})(M_{min}, M_1) \tag{5.29}$$

We use the following equation for the differential halo mass function (see chapter 4):

$$\frac{dn(M_h, z)}{dM_h} = -\frac{\rho_b}{\sigma(M_h, z)} \frac{1}{M_h} \frac{d\sigma(M_h, z)}{dM_h} f(\sigma, z) \tag{5.30}$$

where $f(\sigma, z)$ is given by the expression:

$$f(\sigma, z) = f_{Warren}(\sigma(M_z), z) = A(z) \left[\sigma^{-a(z)} + b(z)\right] \exp\left[-\frac{c(z)}{\sigma^2}\right] \tag{5.31}$$

using the best-fit parameters derived by Crocce et al. 2010 for the MICE simulations.

The second constraint is given by the linear galaxy bias at large scales. We know the projected correlation function of galaxies as a function of luminosity, $w_p^{gal}(r_p| > L_r)$, from SDSS. We know that real space correlation function, ξ_r, is related to the projected correlation function, $w_p(r_p)$, as follows:

$$w_p(r_p) = 2 \int_0^{\infty} \xi(r) dr_\pi = 2 \int_0^{\infty} \xi(r = (r_p^2 + r_\pi^2)^{1/2}) dr_\pi \tag{5.32}$$

We assume a linear bias relation between the galaxy and dark matter distributions:

$$\xi^{gal}(r| > L_r) = (b_{gal}^{Lin}(> L_r))^2 \xi^{DM}(r) \tag{5.33}$$

and therefore:

$$w_p^{gal}(r_p| > L_r) = 2 \int_0^{\infty} \xi^{gal}(r| > L_r) dr_\pi = 2(b_{gal}^{Lin}(> L_r))^2 \int_0^{\infty} \xi^{DM}(r) dr_\pi \tag{5.34}$$

We know $w_p^{gal}(r_p| > L_r)$ from data. We can derive $\xi^{DM}(r)$ as the FT of the assumed linear MICE power spectrum. And therefore, we can derive the linear galaxy bias as a function of the luminosity:

$$b_{gal}^{Lin}(> L_r) = \left(\frac{w_p^{gal}(r_p| > L_r)}{2 \int_0^{\infty} \xi^{DM}(r) dr_\pi}\right)^{1/2} \tag{5.35}$$

On the other hand we can also derive the linear galaxy bias as a function of the luminosity assuming a linear relation between the halo field and the dark matter density field, $\delta_{halo}(M_h) = b(M_h)\delta_{DM}$, and following the next expression:

$$b_{gal}^{Lin}(> L_r) = \frac{\int_{M_{min}}^{\infty} \left(\frac{dn}{dM_h}\right) N_{gal}(M_{min}, M_1) b(M_h) dM_h}{\int_{M_{min}}^{\infty} \left(\frac{dn}{dM_h}\right) N_{gal}(M_{min}, M_1) dM_h} \tag{5.36}$$

We use equations 5.35 and 5.36 as a second constraint.

We sample the parameter space by generating a grid of possible values of M_{min} and M_1. The minimum halo mass to host a central galaxy, M_{min}, covers the range $\log_{10}(M_{min}) \in (9-15)(M_\odot/h)$ with $\Delta \log M = 0.03$. The second parameter M_1 is determined by the factor f_{M_1} following $M_1 = f_{M_1} M_{min}$. The factor f_{M_1} varies from 5 to 25 with $\Delta f_{M_1} = 0.1$.

The left panel of figure 5.12 shows the halo bias from Manera et al. 2010 using the parameters for the MICE simulations (see chapter 4 for more details). The right panel shows the mean number density of central galaxies as a function of M_{min} (solid line). The figure also shows the cumulative of the mean number density of central galaxies as a function of M_{min} but including the halo bias in the integral (dashed line), given by the expression 5.37. It is the mean number density of central galaxies (or haloes since we assume that each halo contains one central galaxy) weighted by the halo bias.

$$\int_{M_{min}}^{\infty} \left(\frac{dn}{dM_h}\right) N_{cen}(M_{min}) b(M_h) dM_h \tag{5.37}$$

The left panel of figure 5.13 shows the cumulative mean number density of satellite galaxies as a function of M_{min} and M_1. When M_1 increases the mean number density of satellites decreases. The right panel shows the cumulative mean number density of satellite galaxies as a function of M_{min} and M_1 but including the halo bias in the integral, i.e:

$$\int_{M_{min}}^{\infty} \left(\frac{dn}{dM_h}\right) N_{sat}(M_{min}, M_1) b(M_h) dM_h \tag{5.38}$$

The left panel of figure 5.14 shows the cumulative mean number density of total galaxies as a function of M_{min} and M_1 (black solid line). The dashed line shows the same but, as before, including the bias in the integral. It is the sum of expressions 5.37 and 5.38:

$$\int_{M_{min}}^{\infty} \left(\frac{dn}{dM_h}\right) N_{gal}(M_{min}, M_1) b(M_h) dM_h \tag{5.39}$$

In the left panel of figure 5.14 the cumulative number density of galaxies as a function of M_{min} and M_1 appears as a unique line (solid *lines*, right part of eq. 5.29). To explain why it appears as a unique solid line we can look at the values of $n_{cen}(> M_{min})$ and $n_{sat}(> M_{min}, M_1)$ at a fix M_{min} and at the largest and smallest values of M_1. Assuming $M_{min} = 10^{12}(M_\odot/h)$, $M_1 = 5 \times M_{min}$ and $M_1 = 24.5 \times M_{min}$, we have $n_{cen}(> M_{min} = 10^{12}(M_\odot/h)) = 2.97 \times 10^{-3}$, $n_{sat}(> M_{min} = 10^{12}(M_\odot/h), f_{M_1} = 5.0(M_\odot/h)) = 5.94 \times 10^{-4}$ and $n_{sat}(> M_{min} = 10^{12}(M_\odot/h), f_{M_1} = 24.9(M_\odot/h)) = 1.19 \times 10^{-4}$. Therefore at $M_{min} = 10^{12}$, the minimum and maximum value of n_{gal} are 3.10×10^{-3} and 3.60×10^{-3} respectively. These values in a logarithm scale are

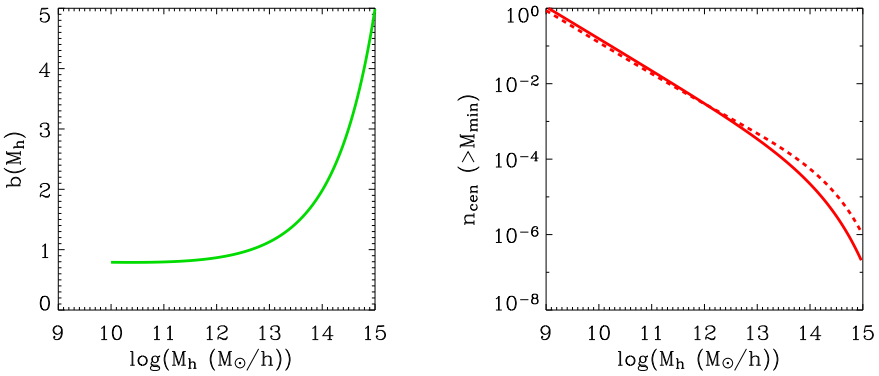

Figure 5.12: Left panel: Expression of halo bias from Manera et al. 2010 for MICE simulations. Right panel: mean number density of central galaxies as a function of M_{min} (solid line, eq. 5.27) and the same but including the halo bias in the integral (dashed line, eq. 5.37).

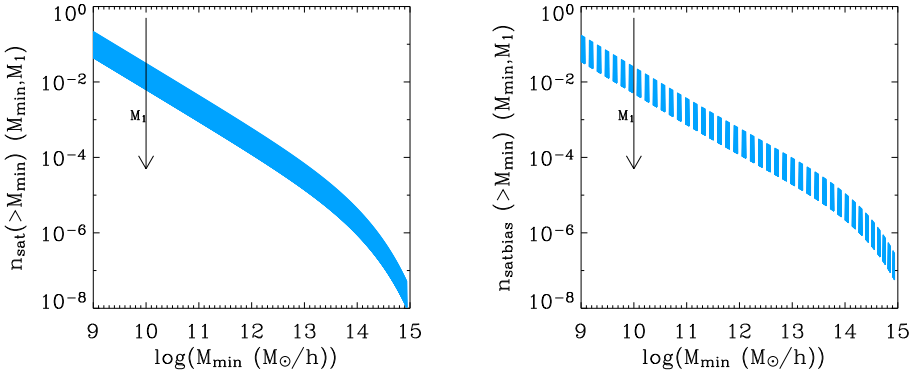

Figure 5.13: Left panel: Mean number density of satellite galaxies as a function of M_{min} and M_1 (eq. 5.28). Right panel: Idem but including the halo bias (eq. 5.38). Note that many curves are plotted, corresponding to different values of the parameter M_1, and that is why it appears as a thick solid line.

almost similar and this is why $n_{gal}(> M_{min}, M_1)$ seems to be a unique line in the left panel of figure 5.14. The same happens for expression 5.39 (dashed *lines*). The contribution of satellites is smaller and almost negligible compared to the central term.

As mentioned before, table 3 from Zehavi et al. 2011 shows the galaxy bias derived for luminosity threshold samples. The right panel of figure 5.14 shows the analytical galaxy bias derived using equation 5.36. Again, the whole grid appears like a solid line and we find big degeneracies to constrain the HOD parameters M_{min} and M_1 by only using two conditions, the mean number density of galaxies as a function of the luminosity and the average linear large scale galaxy bias at a certain scale.

Up to now we have shown that using the MICE simulations and following SS09's recipes do not provide a good fit to the observed luminosity function and also that

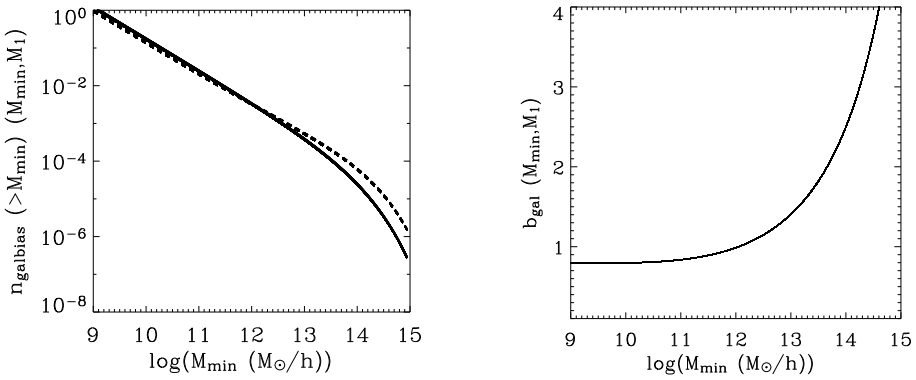

Figure 5.14: Left panel: Mean number density of total galaxies as a function of M_{min} and M_1 (solid line, right part of eq. 5.29) and the same but including halo bias in the integral (dashed line, eq. 5.39). Right panel: Galaxy bias as a function of M_{min} and M_1 derived using eq. 5.36.

we are not able to analytically constrain two HOD parameters by only using two conditions. The next section describes a method that overcomes these limitations. It describes a way to assign galaxy luminosities in order to improve the luminosity function fit. It also describes how we constrain the three HOD parameters (M_{min}, M_1 and α) by generating a grid of mock galaxy catalogues, each of them with a different set of HOD parameters. In this case we use as constraints the observed mean number density of galaxies and the whole observed luminosity dependence of clustering and not just the linear galaxy bias at a given scale.

5.3 MICE HOD parameters

The next approach consists on using more constraints in order to derive the HOD parameters that match the best-fit Schechter luminosity function from Blanton et al. 2003 and the luminosity dependence of clustering using MICE simulations. We generate a grid of mock catalogues sampling the HOD parameter values (M_{min}, M_1 and α). We vary the factor f_{M_1}, that multiplies M_{min} to determine M_1, from 10 to 29.5 in steps of 0.5, and α from 0.50 to 1.50 in steps of 0.05, and we build a catalogue for each set of parameters (a total number of 600 galaxy catalogues), using in this case a little box of the Grand Challenge simulation with a box-size $L_{box} = 307.2$ Mpc/h and taking into account haloes with a minimum number of dark-matter particles $N_p = 10$. We follow the next steps to build the catalogues: (1) Select haloes from the halo catalogue with mass greater than M_{min} and assume that every halo will contain a central galaxy (each halo will have one central galaxy). (2) Compute the number of satellite galaxies per halo given its mass assuming that the mean number density of satellite galaxies follows a Poisson distribution with mean value given by equation 5.11. (3) Assign galaxy luminosities following a method which matches the luminosity function (we use the SHAM technique that will be explain in detail later). (4) Place central galaxies at the center of their host haloes and place satellite

galaxies assuming that they follow a modified NFW density profile (it will also be described in detail later).

Once the catalogues are derived we compute the projected correlation function of galaxies depending on luminosity, $w_p(< M_r)$, and compute the best HOD set of parameters that matches at the same time the luminosity function and $w_p(< M_r)$. The next section describes, in particular, step (3), the way we assign galaxy luminosities.

5.3.1 Assigning galaxy luminosities

This section describes a new method to assign galaxy luminosities using the abundance matching technique in order to fit correctly the observed luminosity function. We choose a set of HOD parameters and using the MICE HMF we derive the cumulative number density of objects (central and satellite galaxies) that inhabit haloes of mass greater than a certain value, $n_{gal}(> M_{min}) = n_{cen}(> M_{min}) + n_{sat}(> M_{min})$:

$$n_{cen}(> M_{min}) = \int_{M_{min}}^{\infty} \frac{dn}{dM_h} dM_h \tag{5.40}$$

$$n_{sat}(> M_{min}) = \int_{M_{min}}^{\infty} \frac{dn}{dM_h} \left(\frac{M_h}{M_1}\right)^{\alpha} dM_h \tag{5.41}$$

On the other hand we compute the mean number density of galaxies with luminosity brighter than a certain value, $n_{gal}(< M_r)$, using the best-fit Schechter function derived by Blanton et al. 2003:

$$n_{gal}(> L_r) = \int_{L_r}^{\infty} \frac{dn}{dL_r} dL_r \tag{5.42}$$

Having both cumulative functions we are able to derive a relation between mass and galaxy luminosity, $M_{gal} - L_{gal}$. This is called the SubHalo Abundance Matching technique (SHAM):

$$n_{gal}(> M_{min}) = n_{gal}(> L_r) \tag{5.43}$$

We decide to model the relation $M_{gal} - L_{gal}$ by using a function of five free parameters (a more complex function than equation 5.9 proposed by SS09):

$$\log(L_r) = c_0 + c_1 \sinh\left[c_2 (\log(M_h - c_3))^2 + c_4 (\log(M_h) - c_3)\right] \tag{5.44}$$

Once we have this relation between mass and luminosity, one can proceed in two different ways. One can assume that only central galaxies follow the $M_{gal} - L_{gal}$ relation or one can assume that all galaxies follow the same relation. If we assume that all galaxies follow the relation we can derive the contribution of satellite galaxies to the total galaxy luminosity function introducing the relation $M_{gal} - L_{gal}$ into equation 5.41:

$$n_{sat}(> M_{min}(L_{gal})) = \int_{M_{min}(>L_{gal})}^{\infty} \frac{dn}{dM_h} \left(\frac{M_h}{M_1}\right)^{\alpha} dM_h \tag{5.45}$$

Once we have $n_{sat}(> L_r)$ we can subtract it from the observed total cumulative galaxy luminosity function and obtain the central contribution:

$$n_{cen}(> L_r) = n_{gal}(> L_r) - n_{sat}(> L_r) \tag{5.46}$$

If we assume that only central galaxies follow the relation we can compute in this case the contribution of central galaxies to the total galaxy luminosity function introducing the relation $M_{gal} - L_{gal}$ into equation 5.40:

$$n_{cen}(> M_{min}(L_{gal})) = \int_{M_{min}(>L_{gal})}^{\infty} \frac{dn}{dM_h} dM_h \tag{5.47}$$

and then derive the contribution of satellite galaxies to the total galaxy luminosity function by subtracting $n_{cen}(> M_{min}(L_{gal}))$ from the observed total cumulative galaxy luminosity function:

$$n_{sat}(> L_r) = n_{gal}(> L_r) - n_{cen}(> L_r) \tag{5.48}$$

We have to remark that we have tested both approaches in the construction of the mock galaxy catalogues and we did not find big differences between them, but we have obtained slightly better fits using the first one, all galaxies follow the relation $M_{gal} - L_{gal}$. In the following we show the results when we assume that all galaxies follow the $M_{gal} - L_{gal}$ relation.

Using the HOD parameters we know that every halo more massive than M_{min} host a central galaxy, and therefore having the halo catalogue we can derive the total number of central galaxies in the catalogue, N_{cen}. Since we assume that central galaxies follow the relation $M_{gal} - L_{gal}$, given the halo mass we can derive the luminosity of its central galaxy. We can also compute the number of satellite galaxies inside each halo given its mass, and therefore we know the total number of satellite galaxies in the catalogue, N_{sat}. Finally, for each mock galaxy catalogue, we know the number of total galaxies that it contains, $N_{gal} = N_{cen} + N_{sat}$. To assign satellite luminosities, we generate randomly N_{sat} "available" luminosities sampling the cumulative satellite luminosity function, $n_{sat}(> L_r)$. We assume that, in general, the most luminous galaxy in a halo is the central galaxy (although there are several works that claim that a non negligible number of haloes does not contain the brightest galaxy at its center (e.g. Skibba et al. 2011). We enforce satellite galaxies not to be brighter than a certain value times the luminosity of the central galaxy. In particular we assign one of the "available" and sampled luminosities for each satellite galaxy if $L_{sat} \leqslant 1.05 \times L_{cen}$. Following this method one gets to a point in which no "available" satellite luminosity fulfill the constraint. It is very important to remark that the number of objects left is very small. We have explored two possibilities, either assigning the minimum luminosity of the catalogue or assigning one of the "available" luminosities (even without following the assumed condition). We do not find significant differences between both methods (again, the number of satellite galaxies with no assigned luminosity at this moment is very small) regarding the luminosity function and the clustering as a function of luminosity. In the following the results we show are derived using the second possibility.

Following the method described in this section guarantees that each mock galaxy catalogue follows the observed luminosity function. The next section explains step (4) of the algorithm, how we place galaxies inside the halo.

5.3.2 Positions and velocities of galaxies

This section describes the method we follow to place central and satellite galaxies inside their host haloes. It also describes how we assign a velocity to each galaxy.

In principle we place galaxies in order to follow a NFW density profile. High-resolution N-body simulations (Navarro et al. 1996) show that the spherically average density around the center of a virialized dark-matter halo containing mass M with virial radius r_{vir} is well fitted by

$$\frac{\rho(r)}{\bar{\rho}} = \frac{\Delta_{vir}(z)}{3\Omega(z)} \frac{c^3 f(c)}{x(1+x)^2} \tag{5.49}$$

where $x \equiv c(M_h)\frac{r}{r_{vir}}$, $\bar{\rho}$ is the average density of the background universe at redshift z, $\bar{\rho}(z)\ \Delta_{vir}(z)/\Omega_z$ is the average density within the virial radius and $f(c) = [\ln(1+c) - c/(1+c)]^{-1}$. The parameter c is often called the central concentration of the halo. More massive haloes are less centrally concentrated. We used the parametrization of this trend provided by Bullock et al. 2001:

$$c(M_h) \approx \frac{9}{(1+z)} \left(\frac{M_h}{M_{*0}}\right)^{-0.13}, \tag{5.50}$$

where M_{*0} is the standard non-linear mass scale: $\sigma(M_*) \equiv \delta_{sc}(z)$. We assume $M_{*0} = 2 \times 10^{13} M_\odot$ for $z = 0.1$,

$$\Delta_{vir}(z) = 18\pi^2 + 82d - 39d^2; \quad d = \Omega(z) - 1. \tag{5.51}$$

$$r_{vir} = \left(\frac{3M}{4\pi\Delta_{vir}(z)\rho_{crit.}}\right)^{(1/3)} \tag{5.52}$$

We place central galaxies at the center of their host haloes and satellite galaxies following a NFW density profile. When implementing the mass density profile, after trying with different values, we set the cut-off of the integral to be 5 times the virial radius. For some haloes, where their centers are very close to the simulation limits, satellite galaxies may be outside of the simulation. In this case we assume periodicity conditions. If we follow the method described before we do not correctly fit observations. The galaxy clustering of the mocks at very small scales (from ~ 0.1 to ~ 0.5 Mpc/h) has a smaller amplitude than observations. Δ_{vir} and the concentration parameter modify strongly the 1-halo term of the $w_p(r_p)$. We have tested several values of M_{*0} in equation 5.50 and also different values for Δ_{vir}. If the concentration parameter increases the amplitude of the correlation function at very low scales becomes more cuspy. We still do not find the expected results after these tests.

The position of satellite galaxies inside the halo affects both regimes of the galaxy clustering, the non-linear and the linear scales. But it is specially important for the

clustering at the scales equal or smaller than the size of the haloes (the biggest haloes have virial radius of $\sim$ 1Mpc). At larger scales what is important for the satellite clustering is the number of satellite galaxies rather than their position inside their host haloes. We therefore have to modify the NFW density profile to better fit observations. Central galaxies are still placed at the center of their host haloes but satellite galaxies are placed following a modified NFW density profile. As mentioned we find that the galaxy clustering of the mocks at very small scales (from ~ 0.1 to ~ 0.5 Mpc/h) has a smaller amplitude than observations. We modify how we place satellite galaxies inside the halo. We derive the distance to the center of the halo. Then, we apply a factor to the distance, f, depending on the luminosity of the satellite galaxy, M_r^{sat}. This factor makes the satellite galaxies to be more concentrated than a NFW density profile. The faintest galaxies are even more concentrated than the brightest ones. The factor f follows this equation:

$$\begin{array}{ll} f = 0.30 & \text{if } M_r^{sat} > -19.5 \\ f = 0.30 + \frac{0.65-0.30}{10.30-9.60}(\log L_r^{sat} - 9.60) & \text{if } -21 \leqslant M_r^{sat} \leqslant -19.5 \\ f = 0.65 & \text{if } M_r^{sat} < -21.0 \end{array} \tag{5.53}$$

Watson et al. 2012 study the radial distribution of galaxies within their host dark matter halos by modeling their small-scale clustering, as measured in the Sloan Digital Sky Survey. They use the HOD framework to model the number and the spatial distribution of galaxies inside the haloes. They assume that the central galaxy is at the center of the halo and satellite galaxies follow a NFW density profile, except that they allow to vary the concentration and the inner slope. They find the same trend as we find. Satellite galaxies are more concentrated than a NFW density profile, although the dependence on luminosity is different. We concentrate more the fainter galaxies and they concentrate more the brighter galaxies.

We also assign velocities to galaxies. It is a good approximation to assume that the motion of a particle in a N-body simulation can be described as the sum of the virial motion of the particle within its host halo and the bulk motion of the halo as a whole (Sheth & Diaferio 2001),

$$v = v_{vir} + v_h \tag{5.54}$$

The virial motions within a halo are well approximated by velocities that are independent Gaussians in each of the three cartesian components, with rms values that depend on halo mass: $\sigma_{vir}^2 =< v_{vir}^2 > \propto GM_h/r_{vir} \propto M_h^{2/3}$. To obtain the velocity dispersion of the halo we use in particular (Bryan & Norman 1998, Sheth & Diaferio 2001):

$$\sigma_{vir} = 476 f_{vir} [\Delta_{vir} E^2(z)]^{1/6} \left(\frac{M_h}{10^{15} M_\odot h^{-1}} \right)^{1/3} kms^{-1}, \tag{5.55}$$

where f_{vir}=0.9 and $E^2(z) = \Omega_0(1+z)^3 + \Omega_\Lambda$ for a flat model with a cosmological constant.

When implementing the NFW density profile we are also assuming that haloes are spherical. We know this is not true and therefore we have implemented triaxial

profiles for the haloes (non-spherical haloes) with a constant ratio between the axis. We do not find big differences in the clustering though.

In the following section we show the results of the galaxy clustering as a function of the luminosity when placing the satellite galaxies following the modified NFW mass density profile we have just described.

5.3.3 Matching the clustering as a function of luminosity

As it is explained in section 5.3.1 the method we use to assign luminosities matches the observed luminosity function for every set of HOD parameters. Therefore we are interested in the set of parameters that, in addition, matches the luminosity dependence of clustering. What we actually match is the projected correlation function, $w_p(r_p)$, from SDSS data (see appendix A for a detailed explanation of the estimation of $w_p(r_p)$ in our mock galaxy catalogues). We use as constraint the projected correlation function, w_p, of volume-limited samples. We use the data from table B7 of Zehavi et al. 2011 where they show the diagonal terms of the error covariance matrix. We use these errors to compute χ^2 for every set of HOD parameters:

$$\chi^2\left(M_{min}, M_1, \alpha\right) = \sum_i \left(\frac{w_p^{SDSS}(r_i) - w_p^{HOD}(r_i)}{\sigma_i} \right)^2 \tag{5.56}$$

where $i = 13$ is the number of scale bins in the projected correlation function, and w_p is defined as:

$$\omega_p(r_p) = 2\int_0^\infty \xi(r) dr_\pi = 2\int_0^\infty \xi(r_p, r_\pi) dr_\pi \tag{5.57}$$

where $\xi(r)$ is estimated on a two-dimensional grid of pair separations parallel, r_π, and perpendicular, r_p, to the line-of-sight.

The generated grid of mock galaxy catalogues does not yields good χ^2 values and one of the reasons is that the catalogues have higher amplitude at faint luminosities compared to observations. Since we assume a univocal relation between central luminosities and their host halo masses we decrease the clustering of the catalogue by decreasing the clustering of satellite galaxies. In order to do it we penalize the faintest satellite galaxies to inhabit the most massive haloes. Therefore, we introduce another condition, in addition to the maximum satellite luminosity with respect to the central luminosity (explained in section 5.3.1), when we assign the "available satellite luminosities" to the satellite galaxies. We implement this condition as follows. We set a straight line in the luminosity-halo mass plane and we compute the distance, d, for its "available" satellite luminosity and the halo mass:

$$\log L_{sat} = a_0 \log M_h + b - d\sqrt{a_0^2 + 1} \tag{5.58}$$

where $a_0 = 0.36$, $b = 4.86$ and d is the value that we use to impose the conditions. For each satellite galaxy we draw a uniformly distributed random number between 0 and 1, u_3. Then if $d < 0$ or $u_3 > 1 - \exp\left(-7d\right)$ then we assign the available satellite luminosity to the satellite galaxy. Basically, following this method makes the faintest

satellite luminosities not to be in the most massive haloes. We obtain better χ^2 values when we fit the projected correlation function as a function of luminosity.

Table 5.2 shows the best HOD parameters found to fit $w_p(r_p)(< M_r)$ including the described previous modifications. Due to the halo mass resolution in our simulations (and because of the $M_{gal} - L_{gal}$ relation used) our catalogue has a minimum galaxy luminosity of $M_r = -18.78$ (using the MICE GC run), and therefore we cannot compare the mock to data from the two faintest samples ($M_r < -18.00$ and $M_r < -18.50$). The left panel of figure 5.15 shows the parameters M_{min} and M_1 as a function of the luminosity threshold of the sample compared to the best HOD parameters found by ZH05. The values we found of the parameter M_{min} are smaller at all luminosities compared to values from ZH05. However, the difference is very small for the two brightest samples. This means that galaxies with the same luminosity inhabit more massive haloes in ZH05's simulations than in MICE simulations, except for the two brightest samples where the haloes will have roughly the same mass. Looking at the parameter M_1 the trend is similar, i.e. the values we found are smaller at all luminosities compared to values from ZH05 except for the brightest sample. First, at the fainter luminosities, it decreases slowly, then at $M_r < -20.0$ it decreases faster down to $M_r = -20.50$, it stays almost constant and finally for the two brightest samples it increases fast making the value of M_1 to be larger than the value in ZH05's work at $M_r < -22.00$. The factor f_{M_1} is not constant as it is in data from ZH05 ($M_1 \sim 23 M_{min}$). This can also explain the fact that the relation between halo mass and luminosity proposed by SS09 using data from ZH05 does not work correctly when using the MICE simulations. The right panel shows the parameter α. The values we found are smaller at all luminosities. Except for the threshold sample at $M_r < -21.5$, our results are very close to 1. This result means that the number of satellites are approximately proportional to the mass of the halo. At $M_r < -21.5$, where ZH05 found the highest value, we find the smallest value. The parameter α seems to have a large scatter at the brightest luminosities.

$< Mr$	$\log M_{min}$	$\log M_1$	f_{M_1}	α	χ^2	N
-19.00	11.35	12.79	22.50	1.05	18.77	13
-19.50	11.58	12.89	18.50	1.00	18.13	13
-20.00	11.79	13.16	18.50	1.10	7.15	13
-20.50	12.17	13.25	10.50	0.95	26.15	13
-21.00	12.48	13.77	11.00	0.95	103.66	13
-21.50	13.17	14.40	15.50	0.70	63.40	13
-22.00	13.82	15.18	23.50	1.00	26.21	13

Table 5.2: Best HOD parameters found when matching the projected correlation function of SDSS luminosity threshold samples. The first column is the luminosity threshold, the second column shows the value of $\log(M_{min})$ using $M_{gal} - L_{gal}$ relation. The third column is $\log(M_1)$ and the fourth column is the factor f_{M_1} that multiplies M_{min} to determine M_1. The fifth column shows the values of the α parameter. The sixth column is the value of χ^2 and the last column refers to the number of points used to derive the χ^2.

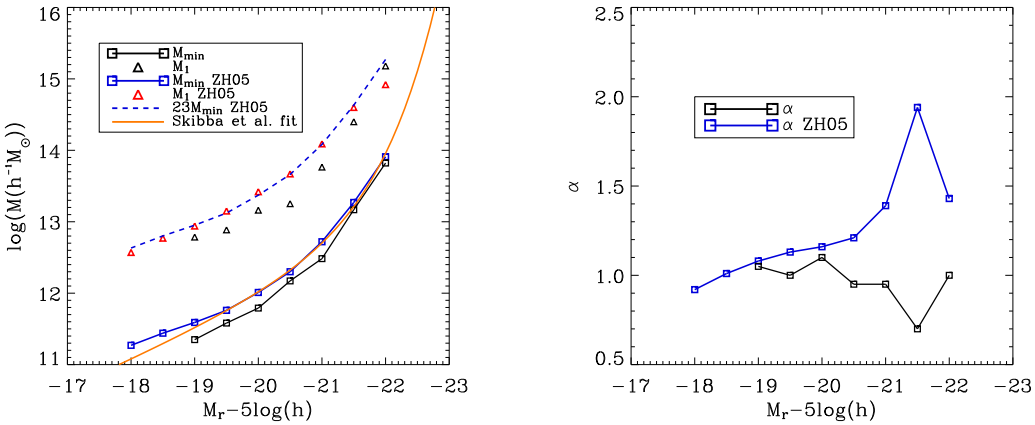

Figure 5.15: HOD parameters as a function of the luminosity threshold of galaxy samples. The left panel shows M_{min} and M_1 for our best fit (black squares and triangles respectively) and for ZH05 best fit (blue squares and triangles). The dashed blue line is the $M_{min} - M_r$ curve scaled by a factor of 23, and the orange solid line is the SS09's fit. The right panel shows the HOD parameter α.

Figure 5.16 shows the projected correlation function depending on luminosity threshold samples (from $M_r < -19.0$ to $M_r < -22.0$) for the best HOD parameters that we found. The fits are quite reasonable for luminosities fainter than $M_r^* = -20.44$. For all luminosities, at large scales, the fits have higher amplitude compared to the observed data but this feature is even more significant for the brightest samples ($M_r < M_r^*$) (in addition, the number of brightest galaxies is small compared to the faintest one and it produces the correlation function to have much more noise). This result is not very surprising since we are only using three free parameters to describe the number of galaxies inside each halo. The step function used in equation 5.11 for central galaxies may be not accurate enough to reproduce the correct clustering. Recent works (e.g. Zehavi et al. 2011 or Leauthaud et al. 2011) point to a number of at least five free parameters to a better fit, which seems to be specially necessary for the brightest samples since the mean number of galaxies in each halo cannot be perfectly described by a step function and a power law. The method we use to build the mock galaxy catalogues places the most brilliant galaxies at the most massive haloes, and mostly they will be central galaxies, which are most of the galaxies. A possible method to decrease the amplitude of the clustering of the most brilliant galaxies (instead of using five free parameters) is to introduce scatter in the halo mass - central luminosity relation which will cause bright galaxies to reside in less massive haloes and therefore to have smaller bias. We will implement this scatter in the next section.

It is also very important to remark that the HOD parameters are useful for a specific type of galaxies which follow a certain property, such as the luminosity in our case. This means that if one uses a HOD set of parameters to build a catalogue that best matches luminosity dependence of clustering for galaxies brighter than a certain value, $M_r < M_{r1}$, and one computes the clustering for galaxies brighter than

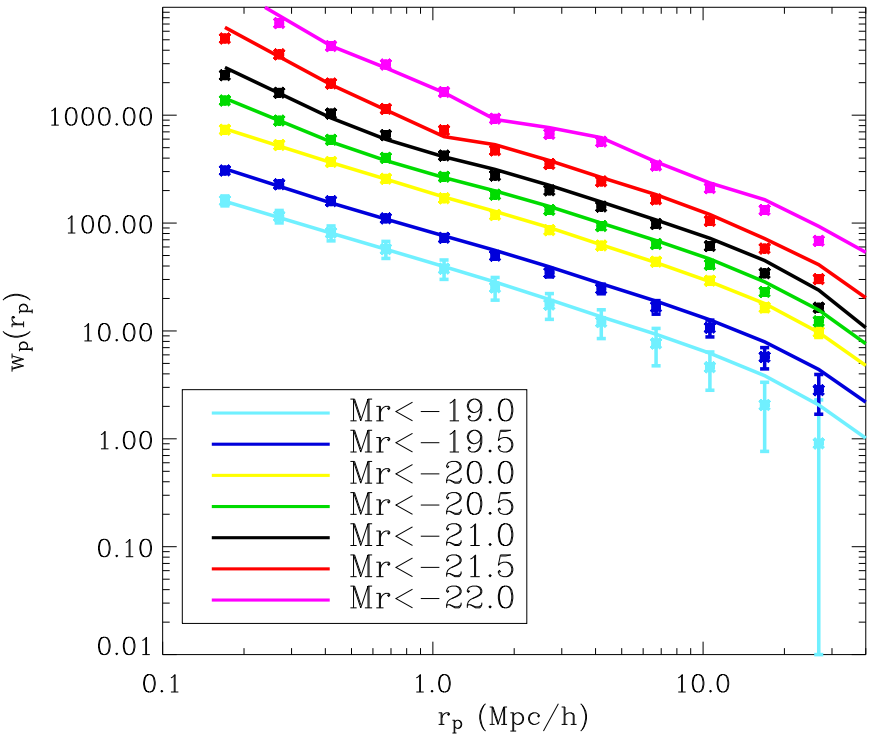

Figure 5.16: Projected correlation function for luminosity threshold samples using the best-fit HOD parameters. The different colours represent the different magnitude limits, from $M_r < -19.0$ to $M_r < -22.0$. The different samples are staggered for clarity.

another brighter value, $M_r < M_{r2}$, with $M_{r2} < M_{r1}$, the clustering may not be necessarily correct. Our purpose is to build a *unique* catalogue that has the correct clustering for *all* luminosities (and also colours). We will deal with this problem in the next section.

5.4 Building a *unique* galaxy catalogue

This section describes the method we have followed to build a *unique* catalogue that reproduces, not only the clustering for one luminosity threshold sample or one luminosity bin sample, but a galaxy catalogue that follows at the same time the galaxy clustering for all luminosity threshold and bin samples (as well as the luminosity function, the colour-magnitude diagram and the colour dependence of clustering). The process is similar to the previous section, in the way that we continue using the parameterization of the number of galaxies followed by SS09, i.e. haloes with mass greater than M_{min} hosts a central galaxy, and the number of satellite galaxies is given by the expression:

$$\left[\frac{M_h}{M_1(L_r)}\right]^{\alpha} \quad \text{if} \quad M_h \geq M_{min}(L_r) \tag{5.59}$$

The main difference is that instead of using discrete values of the HOD parameters we assume the parameter M_1 to be a function which depends on M_h. In addition, we have also introduced some other changes in the method to be able to reproduce the observed data. We have introduced scatter in the relation between halo mass and central galaxy luminosity in order to decrease the clustering of the brightest galaxies to better reproduce their clustering at large scales. We have included another Gaussian component to better fit the distribution of the $g-r$ colour vs. the absolute magnitude in the r-band. As a consequence, we have also changed the recipes to determine the galaxy colours in order to fit correctly the colour-magnitude diagram and at the same time the clustering as a function of colour.

5.4.1 HOD function

As it was mentioned the main difference in the new method is that instead of using discrete values of the HOD parameters we assume the parameter M_1 to be a function which depends on M_h. We still assume that all haloes have one central galaxy and the parameter $\alpha = 1$. Our best HOD parameters shown in the previous section did not show a constant relation between the parameters M_1 and M_{min}, as it was found in ZH05 (see figure 5.15). The factor f_{M_1}, which multiplies M_{min} to obtain M_1, varies with luminosity. After making different tests we set the factor f_{M_1} to follow the expression:

$$f_{M_1} = 0.5((a_1-a_2)\tanh(s_1(b_1-\log M_h))+(a_3-a_2)\tanh(s_2(\log M_h-b_2))+(a_1+a_3)) \tag{5.60}$$

where $a_1 = 25.0$, $a_2 = 11.0$, $a_3 = 14.0$, $b_1 = 11.5$, $b_2 = 12.5$, $s_1 = 2.0$ and $s_2 = 2.50$. We found that using expression 5.60 provides a reasonable fit to observations.

5.4.2 Scatter in the halo mass luminosity relation

We find that generating a *unique* mock galaxy catalogue without introducing scatter makes the clustering of the brighter galaxies much larger than observed data. Previously, in section 5.3.3, we showed that we use the subhalo abundance matching technique to obtain $M_{gal} - L_{gal}$. Basically we compared the cumulative luminosity function of galaxies and the cumulative mass function of haloes/subhaloes. In this case we generate an "unscattered" luminosity function that after introducing the scatter in the central luminosities the cumulative luminosity function of the catalogue will match the observed cumulative luminosity function. We compute $M_{gal} - L_{gal}$ using the same SHAM technique but in this case with the "unscattered" luminosity function and the cumulative number density of objects (central and satellite galaxies), $n_{gal}(> M_{min})$. We assign central luminosities following this relation and we only introduce scatter if the luminosity is brighter than the characteristic luminosity, L_r^* ($M_r^* = -20.44$). In particular, if $\log(L_r^*) < \log(L_r^{cen}) \leqslant \log(L_r^*) + 1$ then we introduce the following scatter:

$$\log(L_r^{cen}) = \log(L_r^{cen}) + u_1 \Delta_{\log(L_r)} (\log(L_r^{cen}) - \log(L_r^*)) \tag{5.61}$$

and if $\log(L_r^{cen}) > \log(L_r^*) + 1$:

$$\log(L_r^{cen}) = \log(L_r^{cen}) + u_1 \Delta_{\log(L_r)} \tag{5.62}$$

where u_1 is a uniformly distributed random number between 0 and 1 and $\Delta_{\log(L_r)}$ is the maximum amplitude of the scatter. We tried different values for $\Delta_{\log(L_r)}$ and we find that using a value of $\Delta_{\log(L_r)} = 0.15$ yields quite good results. We find that using a larger value, $\Delta_{\log(L_r)} \sim 0.2$, makes the "unscattered" luminosity function to almost drop vertically since it is an exponential function. We do not need to have a χ^2 value equal to 0 since, in addition, we are calibrating our method using a box of the MICE GC run with side-box $L_{box} = 307.2$ Mpc/h and it may change from box to box. Figure 5.17 shows the "unscattered" cumulative luminosity function (black line) used to derived the halo mass - luminosity relation compared to the cumulative luminosity function of the best fit Schechter function of SDSS data (blue line). Figure 5.18 shows the relation between halo mass and luminosity obtained using the SHAM technique. The blue line is the relation $M_{gal} - L_{gal}$ we found for the best HOD parameters that best match the luminosity clustering of galaxies brighter than $M_r = -20.0$ using the "common" cumulative luminosity function. The black line is the relation $M_{gal} - L_{gal}$ obtained using the "unscattered" cumulative luminosity function. The green line is the relation suggested by SS09, and the red and orange lines represent M_{min} of the halo population for the MICE N2048L3072 (intermediate) and GC runs, respectively. It is very interesting to realize that, finally, our halo mass - central luminosity relation is quite similar to that proposed by SS09, which was the relation used at the beginning. At a fix halo mass, our central galaxies are always brighter than SS09's galaxies. This is significant at the faintest luminosities and up to $M_r \sim -21.5$. Then, at luminosities brighter than $M_r \sim -21.5$, the relation is almost similar. The main difference comes from the fact that we are now introducing scatter in the relation. This makes possible to have central galaxies as the brightest galaxies in our catalogue.

Figure 5.19 shows the relation between halo masses and luminosities of their central galaxies using a little box ($L_{box} = 307.2$ Mpc/h). The blue points represent central galaxies and the solid red line is the relation $M_{gal} - L_{gal}$ without scatter.

Figure 5.25 shows the luminosity function of the generated mock galaxy catalogue compared to the best fit Schechter function found by Blanton et al. 2003. As expected, by following the method to assign luminosities, the agreement between the mock and the data is almost perfect. And we are also confident of having correctly implemented the scatter between halo mass and central galaxy luminosity.

5.4.3 Galaxy colours: a third Gaussian component

The last change in the algorithm is the introduction of a third Gaussian component to better describe the colour - magnitude distribution ($g - r$ vs. M_r). Skibba & Sheth 2009 presented a method to incorporate colours in the HOD model in order to follow the color magnitude diagram and the clustering as a function of colour as well as the luminosity function and the clustering as a function of luminosity. We explained their method in section 5.2.1. We use the NYU DR7 catalogue to generate figure 5.20 which shows the bimodality in the colour-magnitude diagram (M_r vs.

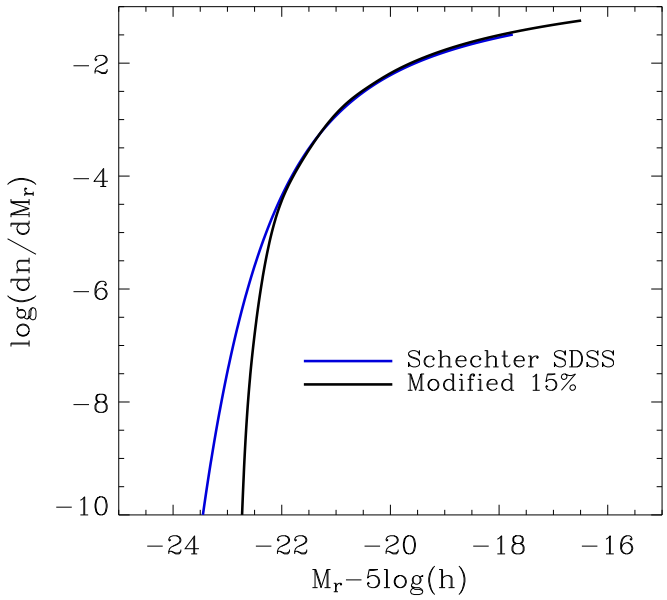

Figure 5.17: Cumulative luminosity functions. The black line represents the "unscattered" cumulative luminosity function and the blue line is the cumulative luminosity function of the best fit Schechter function of SDSS data.

$(g-r)$) by using different histograms of the number of galaxies as a function of colour for different luminosity bins. Following SS09 we fit two Gaussian functions for both, the blue and the red sequences. The fits are not perfect. There is a small excess of galaxies in the so called *green valley* when fitting two Gaussian functions. We find small differences between our best-fit parameters and SS09's ones. Figure 5.22 shows the comparison between the parameters we fit and the SS09's parameters when using two Gaussian functions. These are the fits to the values of the mean and the rms we found of the two Gaussian functions:

$$\begin{aligned} \langle g-r|M_r\rangle_{red} &= 0.923 - 0.020(M_r+20.0) \\ rms\,(g-r|M_r)_{red} &= 0.044 + 0.001(M_r+20.0) \end{aligned} \tag{5.63}$$

$$\begin{aligned} \langle g-r|M_r\rangle_{blue} &= 0.655 - 0.090(M_r+20.0) \\ rms\,(g-r|M_r)_{blue} &= 0.177 + 0.003(M_r+20.0) \end{aligned} \tag{5.64}$$

In order to have a better fit we decide to introduce another *sequence* using a third Gaussian function. Equations 5.65, 5.66 and 5.67 are the fits to the values of the mean and the rms of the three Gaussian functions we fit. Figure 5.21 shows SDSS colour distribution using our fit with three Gaussians functions.

$$\begin{aligned} \langle g-r|M_r\rangle_{red} &= 0.923 - 0.021(M_r+20.0) \\ rms\,(g-r|M_r)_{red} &= 0.041 - 0.003(M_r+20.0) \end{aligned} \tag{5.65}$$

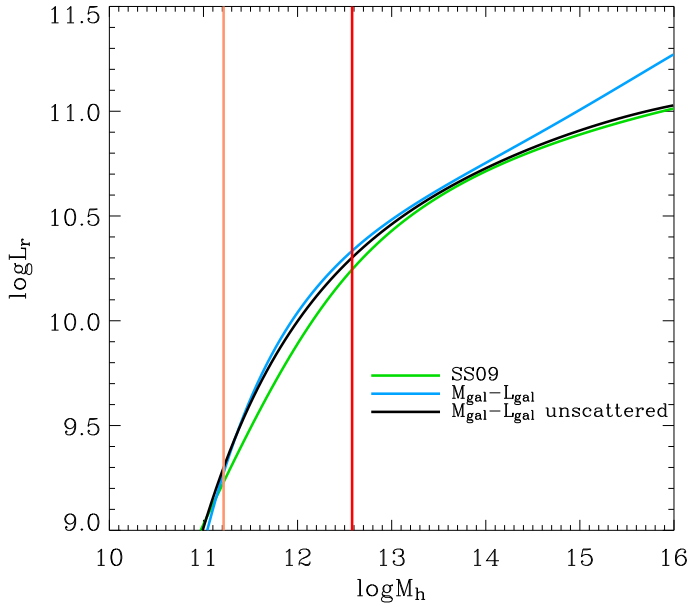

Figure 5.18: Halo mass - luminosity relation obtained using the SHAM technique. The blue line is the relation $M_{gal} - L_{gal}$ we found for the best HOD parameters that best match the luminosity clustering of galaxies brighter than $M_r = -20.0$ using the "common" cumulative luminosity function. The black line is the relation $M_{gal} - L_{gal}$ obtained using the "unscattered" cumulative luminosity function. The green line is the relation suggested by SS09, and the red and orange lines represent M_{min} of the halo population for the MICE N2048L3072 (intermediate) and GC runs, respectively.

$$\begin{aligned}
\langle g-r|M_r\rangle_{green} &= 0.880 - 0.035(M_r+20.0) - 0.062\tanh\left(\tfrac{M_r+22.60}{0.12}\right) \\
rms\,(g-r|M_r)_{green} &= \begin{cases} 0.055 + 0.023(M_r+20.0) & \text{if } M_r \geqslant -22.5 \\ -0.002 - 0.280(M_r+22.5) & \text{if } -22.5 > M_r \geqslant -22.75 \\ 0.067 + 0.020(M_r+22.75) & \text{otherwise} \end{cases}
\end{aligned} \tag{5.66}$$

$$\begin{aligned}
\langle g-r|M_r\rangle_{blue} &= \begin{cases} 0.610 - 0.100(M_r+20.0) & \text{if } M_r \geqslant -22.0 \\ 0.810 - 0.020(M_r+22.0) & \text{otherwise} \end{cases} \\
rms\,(g-r|M_r)_{blue} &= 0.170 - 0.005(M_r+20.0) - 0.037\tanh\left(\tfrac{M_r+22.38}{0.27}\right)
\end{aligned} \tag{5.67}$$

The way we assign colours to galaxies is similar as in SS09's model in the sense that it depends on the galaxy type (whether it is a central or a satellite galaxy) and also on the colour sequence (blue, green or red) it belongs to. First we assign colours to satellite galaxies and later we force central colours to follow the total colour-magnitude diagram. SS09 proposed a function that defines the mean colour of satellite galaxies given the luminosity. They used this function (in addition to the assumption that colours do not depend on the halo mass) to define the fraction

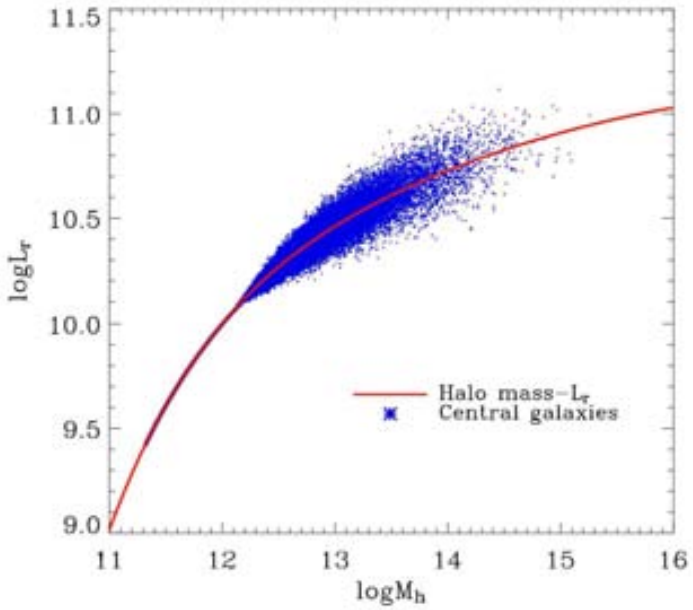

Figure 5.19: Relation between central galaxy luminosities and their host halo mass showing the scatter applied in the halo mass-central luminosity relation. The total number of central galaxies plotted is 343351, which corresponds to a simulation of box-size $L_{box} = 307.2\,\mathrm{Mpc/h}$. The scatter is applied for galaxies brighter than $M_r^* = -20.44$.

of satellite galaxies that belongs to the red sequence. In our case we directly set the fraction of satellite galaxies that belongs to the red sequence to be given by the expression[7] 5.68 since it yields reasonable results. The fraction is not very different from the one proposed by SS09. The physical reason why most satellite galaxies are red and how this function depends on luminosity should be related to galaxy formation and evolution. When satellite galaxies fall into a big halo can experienced an episode of star formation and ram pressure stripping of their gas. Then, they are expected to passively evolve, which bring them to the red branch.

$$f_{sat}^{red}(M_r) = 1.00 - 0.30\tanh\left(\frac{M_r + 22.20}{1.20}\right) \tag{5.68}$$

We also set the fraction of satellite galaxies that belongs to the green sequence which is given by the expression:

$$f_{sat}^{green}(M_r) = 0.05 - 0.05\,(M_r + 20.0) \tag{5.69}$$

and therefore the blue sequence is determined by:

$$f_{sat}^{blue}(M_r) = 1 - \left(f_{sat}^{red}(M_r) + f_{sat}^{green}(M_r)\right) \tag{5.70}$$

[7]When the function is greater than 1 all satellite galaxies are red, and therefore the fraction of green and blue satellite galaxies become 0.

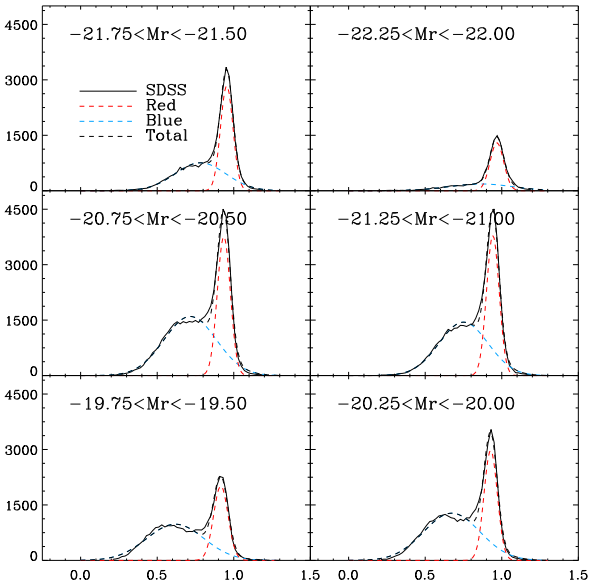

Figure 5.20: Bimodal distribution of $(g-r)$ colour in the SDSS. The black solid line is the SDSS distribution and black dashed line is the sum of two Gaussian functions to the red and blue sequences (dashed lines).

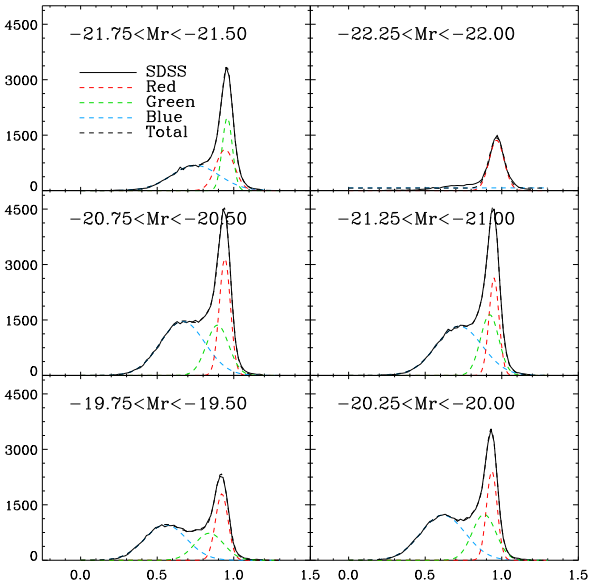

Figure 5.21: Idem as figure 5.20 but using three Gaussian functions instead of two to fit the observed distribution.

We find that setting the fraction of satellite galaxies to follow expressions 5.68, 5.69 and 5.70 provides very good results regarding the colour-magnitude diagram and reasonable good results regarding the clustering as a function of colour.

Figure 5.24 shows the fraction of objects as a function of the absolute magnitude M_r. The black solid line and the black point-dashed line show the fraction of central

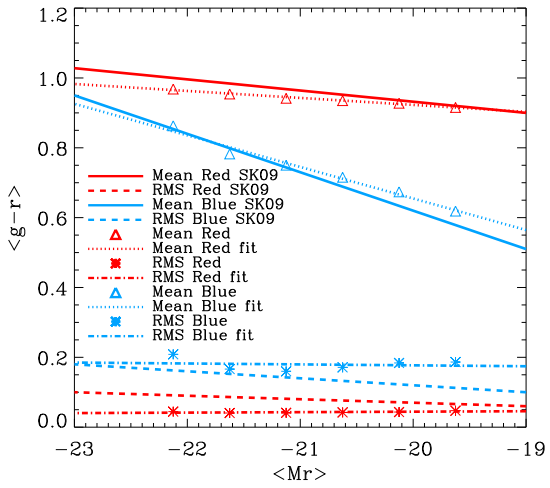

Figure 5.22: Parameters found when fitting two Gaussian functions compared to SS09's fits. The red and blue solid lines are the SS09's fits to the mean values. The dashed red and blue lines are the SS09's fit to the rms values. The red and blue triangles are our best mean values of the two Gaussian functions and the red and blue point-lines are the straight line fits to the points. The red and blue stars are our best rms values of the two Gaussian functions, and the point-dashed lines are straight line fits to the points.

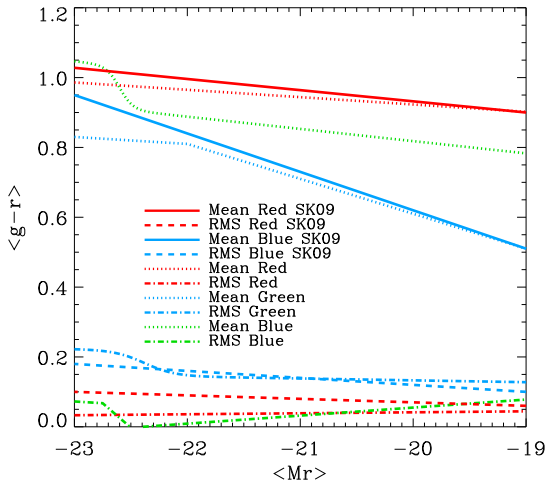

Figure 5.23: Idem as figure 5.22 but using three Gaussian functions instead of two.

and satellite galaxies respectively. It also shows the fraction of central, satellite and total galaxies that belong to the red, green and blue sequences. As mentioned the fraction of red satellite galaxies is very large at all luminosities. At the faintest luminosities this fraction tends to 0.7 and it increases slowly towards higher luminosities, becoming 1 at $M_r \sim -22.0$. The fraction of green satellite galaxies is very small at all luminosities although it was necessary to better match the observations as it was shown in figure 5.23.

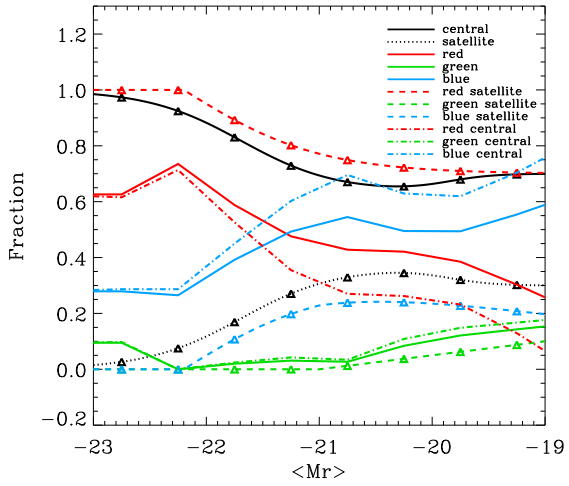

Figure 5.24: Fraction of objects as a function of the absolute magnitude M_r. The black solid line refers to the fraction of central objects. The black dotted line is the satellite fraction of objects. The red, green and blue solid lines correspond to the total (central + satellite) red, green and blue fraction of galaxies. The dashed and dashed dotted lines show the same colour fractions but for central and satellite galaxies respectively.

In detail, we assign colours to the galaxies in the following way: for each satellite galaxy we draw two random numbers, u_0 and u_1. u_0 is a uniformly distributed random number between 0 and 1, and determine to which sequence the satellite galaxy belongs to depending on the expressions 5.68, 5.69 and 5.70. Then we compute the mean value of the colour depending on the luminosity following the expressions 5.65, 5.66 or 5.67. u_1 follows a Gaussian distribution with mean value 0 and rms = 1 and it is used to determine the standard deviation from the mean (by multiplying u_1 times the rms). For central galaxies, as mentioned before, we force their colours to follow the total colour-magnitude diagram: by using our three Gaussians to fit the colour-magnitude diagram (expressions 5.68, 5.69 and 5.70) we can compute what is the fraction of red, green and blue objects (central + satellite galaxies) as a function of the absolute magnitude, $f_{tot}^{red}(M_r)$, $f_{tot}^{green}(M_r)$ and $f_{tot}^{red}(M_r)$, respectively. We can also describe the fraction of red objects as:

$$f_{tot}^{red}(M_r) = f_{cen}^{red}(M_r) f_{cen}(M_r) + f_{sat}^{red}(M_r) f_{sat}(M_r) \tag{5.71}$$

and therefore we can derive the fraction of central galaxies that belongs to the red sequence since we know every term in the following equation:

$$f_{cen}^{red}(M_r) = \frac{f_{tot}^{red}(M_r) + f_{sat}^{red}(M_r) f_{sat}(M_r)}{f_{cen}(M_r)} \tag{5.72}$$

We can derive the green fraction of central galaxies in the same way:

$$f_{cen}^{green}(M_r) = \frac{f_{tot}^{green}(M_r) + f_{sat}^{green}(M_r) f_{sat}(M_r)}{f_{cen}(M_r)} \tag{5.73}$$

and we estimate the blue fraction of central objects as $f_{cen}^{blue}(M_r) = 1 - (f_{cen}^{red}(M_r) + f_{cen}^{green}(M_r))$.

Once we know the probability for central galaxies to belong to one of the three sequences we follow the same method we have followed for satellite galaxies, one random number to determine the sequence and another one to derive a realization of the colour.

5.4.4 Results

This section shows the resulting mock galaxy catalogue generated using the last method explained in the previous section populating a halo catalogue (with a minimum number of dark-matter particles to form a halo equal to 10) extracted from a box-size $L_{box} = 307.2$ Mpc/h of the snapshot at $z = 0$ of the MICE GC run.

Figure 5.25 shows the luminosity function of the catalogue compared to the best Schechter function fit to the SDSS data. The error bars are derived using Poisson errors in each bin. As it was expected the agreement is almost perfect because the method used force the galaxy catalogues to follow the luminosity function. In the next chapter the luminosity function of the whole snapshot, $L_{box} = 3072$ Mpc/h, is derived estimating the errors using the Jack-kinfe resampling method.

Figures 5.26, 5.28 and 5.29 show the color-magnitude diagram for all (central + satellite), central and satellite galaxies respectively. The red and blue solid lines are SS09's fits to the SDSS volume limited samples using two Gaussian functions and dashed lines are our fits when using three Gaussian functions in order to model the colour distribution as a function of the magnitude. The green solid line refers to the satellite sequence set by SS09, equation 5.13. In figure 5.26, in addition, the colour-magnitude diagram in the $M_r < -19.0$ SDSS volume limited sample (black dashed-dot contours) is shown. The agreement between DR7 and the mock is very good. Figure 5.27 shows the bimodal distribution of $g - r$ colour of the mock galaxy catalogue for different absolute magnitude-bin samples.

Figure 5.30 shows the luminosity dependence of galaxy clustering for different luminosity threshold samples compared to the SDSS data. The samples are staggered for clarity since galaxies fainter than M_r^* have almost a similar bias. The fits are reasonably good, although at large scales the amplitude of the clustering is in general slightly larger that SDSS data, specially for galaxies brighter than M_r^*. However, as explained previously, if we do not include scatter in the relation $M_{gal} - L_{gal}$ the amplitude of the clustering at large scales would be much higher. Figure 5.31 shows the projected correlation function of luminosity-bin samples of the catalogue compared to SDSS data. The fits show the same trend at large scales for the brighter galaxies, the amplitude is larger than observations.

Figure 5.32 shows the colour dependence clustering of the catalogue. We show separately red (left panel) and blue galaxies (right panel). When separating galaxies in colour one observes that fixing the luminosity red galaxies are more clustered than blue ones. It is also important to remark another characteristic of red galaxies. They exhibit a higher clustering amplitude on small scales increasing towards low luminosities. This strong clustering of the fainter luminosities can be an indication that most of these galaxies are satellite galaxies that reside in massive haloes that

have stronger bias. Our catalogue does not fairly follow this trend. In addition, the brightest blue galaxies are too clustered compared to observations too. We plan to solve these issues in the near future. We think that introducing a condition depending on the galaxy colour during the satellite luminosity assignation could be a possible solution.

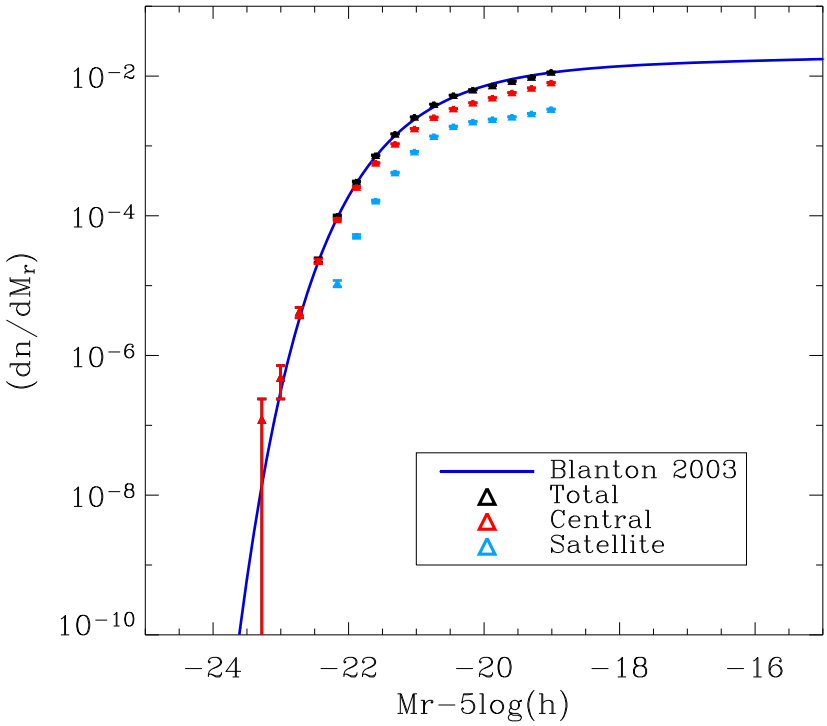

Figure 5.25: Luminosity function of the mock galaxy catalogue. The blue solid line is the best Schechter function fit to the SDSS data shifted to z=0.1. The black, blue and red triangles are the luminosity function of total (central+satellite), central and satellite galaxies of the catalogue respectively. The error bars are derived as Poisson errors.

5.5 Summary

The whole process of generating mock galaxy catalogues has been presented in this chapter. We have developed a method to construct a *unique* mock galaxy catalogue that follows some observed global local properties of galaxies, in which we use both, the HOD and the SHAM models to assign galaxy luminosities. In particular the luminosity function in the r-band, the $g-r$ vs. M_r diagram, and the galaxy clustering as a function of luminosity and colour. The observed properties come from observed SDSS data at $z = 0.1$.

First, we generated a mock galaxy catalogue to participate in the LSS Simulation Challenge of the Dark Energy Survey within the Galaxy Clustering Working Group. In this case, instead of using HOD prescriptions, we selected dark matter particles

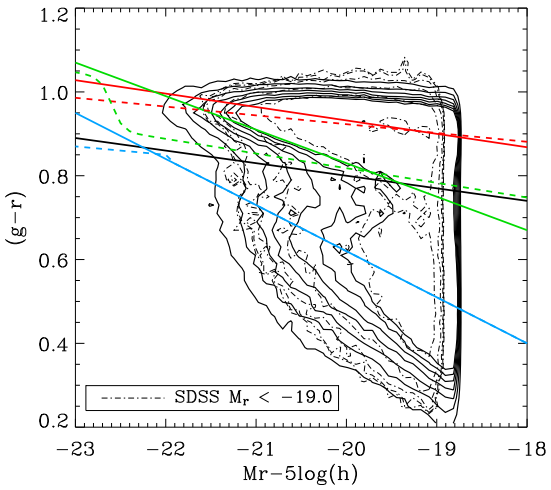

Figure 5.26: Colour-magnitude diagram of all galaxies (central + satellite) of the mock galaxy catalogue. The red and blue solid lines show the mean values of the red and blue sequences derived by SS09, given by equations 5.7 and 5.8, respectively. The black solid line is a crude cut which some authors use to separate the galaxy population into the red and blue sequences. The green solid line refers to the satellite sequence set by SS09, equation 5.13. The black dashed-dot contours refer to the colour-magnitude diagram in the $M_r < -19.0$ SDSS volume limited sample. The red, green and blue dashed lines are the mean values of the red, green and blue sequences we fit to improve the fitting.

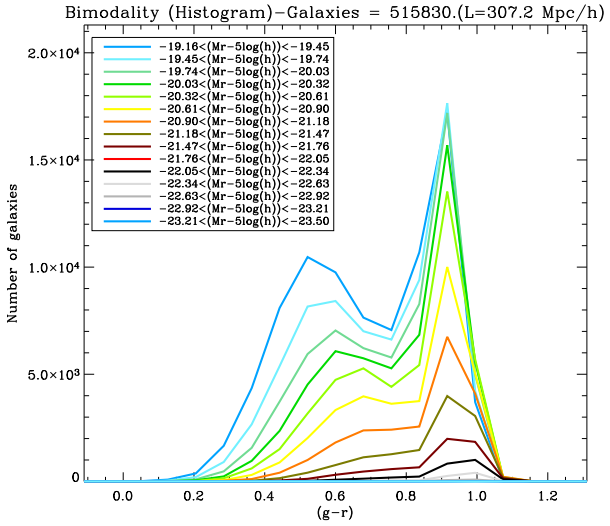

Figure 5.27: Bimodal distribution of (g-r) colour in the mock galaxy catalogue. Different colours refer to different absolute magnitude-bin samples. The total number of galaxies in the catalogue is 515830.

from the MICE2048L3072 run to be galaxies in order to follow a determined number density as a function of redshift (from $z = 0.2$ up to $z = 1.4$). The catalogue was built in an octant of the sky and we introduced different errors in the galaxy redshifts in order to simulate photometric redshifts.

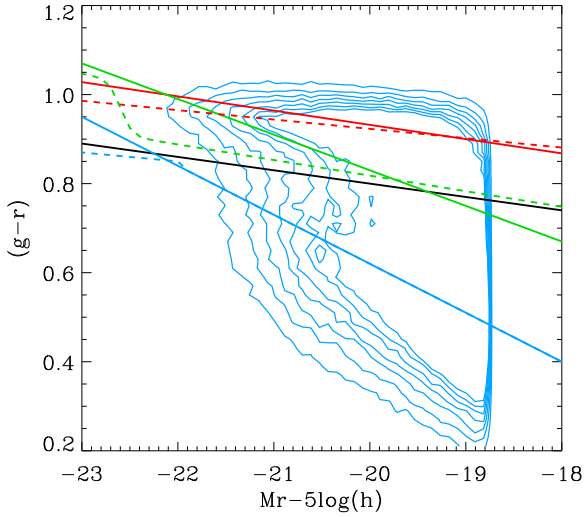

Figure 5.28: Colour-magnitude diagram of central galaxies of the mock galaxy catalogue. Solid and dashed lines are the same as in figure 5.26.

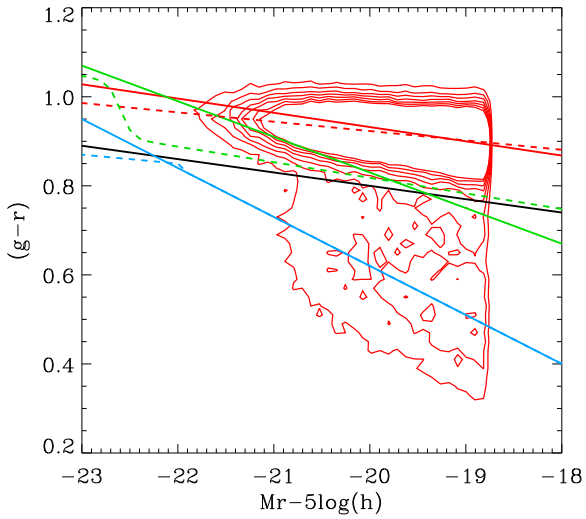

Figure 5.29: Colour-magnitude diagram of satellite galaxies of the mock galaxy catalogue. Solid and dashed lines are the same as in figure 5.26.

We have presented Skibba & Sheth 2009's method to include colours in the halo occupation model, and an algorithm to build mock galaxy catalogues that preserve the mean number density of objects, the global colour-magnitude diagram and the correct clustering as a function of luminosity and colour. We have shown the colour-magnitude diagram and the bimodal distribution of (g-r) colour in the SDSS.

We have generated a mock galaxy catalogue following SS09's algorithm. We populate with galaxies the halo catalogue extracted from the simulation N2048L3072 using the best HOD parameters found by ZH05 and then we assign luminosities to

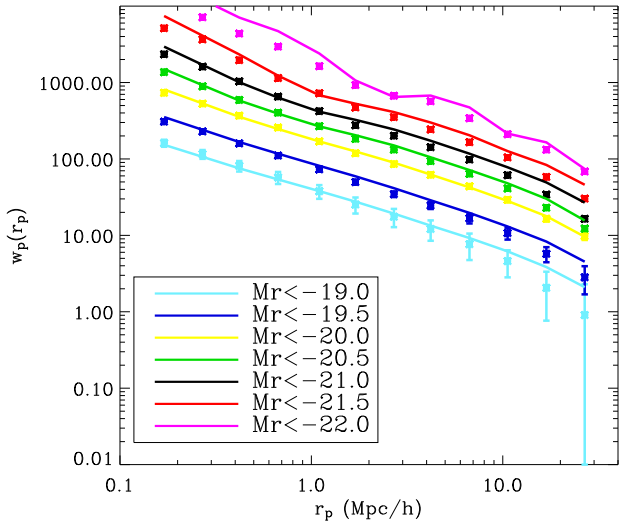

Figure 5.30: Luminosity dependence of galaxy clustering. Figure shows the projected correlation functions for different luminosity threshold samples of SDSS and the mock galaxy catalogues. The samples are staggered for clarity.

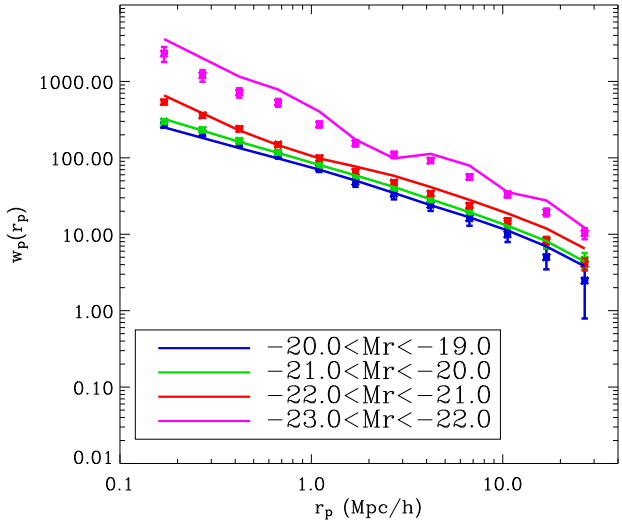

Figure 5.31: Projected correlation function of luminosity-bin samples of the catalogue.

the galaxies following the relation between the parameters M_{min} and L_{min} proposed by SS09 using data from ZH05 in order to reproduce the luminosity function. We found that the prescription does not work correctly in our simulations since we are not able to follow the luminosity function even after correcting for the differences in the cosmological model.

We tried to analytically derive two HOD parameters, M_{min} and M_1, for the MICE simulations using as constraints the mean number density of objects and the average linear galaxy bias at a certain scale (larger than 10 Mpc/h). We found big

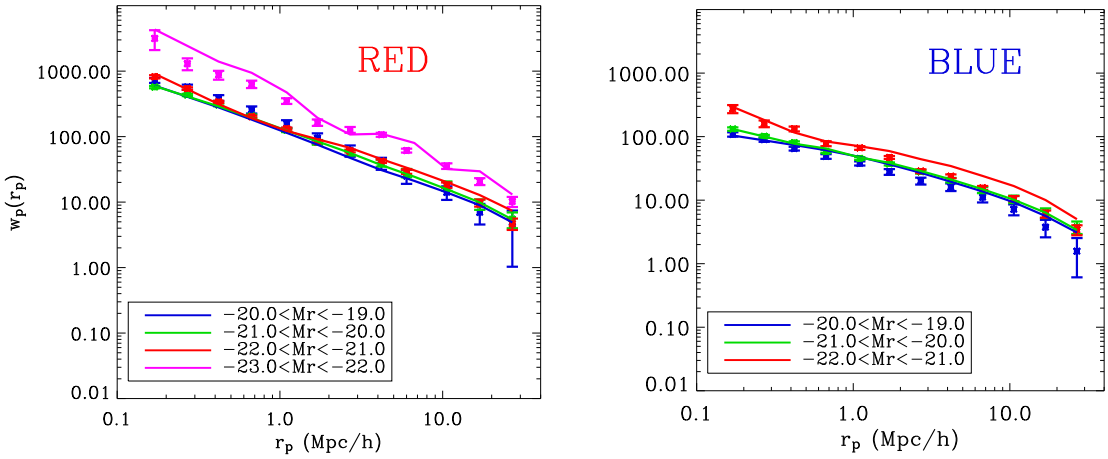

Figure 5.32: Projected correlation function for different luminosity-bin samples of the catalogue. The left panel shows red galaxies and the right panel shows blue galaxies.

degeneracies for both parameters. The mean number density of objects constrain quite well the parameter M_{min} but degeneracies do not disappear when using the average linear galaxy bias at low halo mass.

We have derived the HOD parameters, M_{min}, M_1 and α, for MICE simulations. We use as constraints the mean number density of galaxies as a function of luminosity and the correlation function as a function of luminosity. We use the halo abundance matching technique as well as HOD prescriptions to compute the relation between halo mass and galaxy luminosity. We have also modified the galaxy density profile from the standard NFW that determines the halo dark matter density. We need to place satellite galaxies closer to the center of the halo, i.e. satellite galaxies are more concentrated that a "common" NFW density profile. We also penalized the faintest satellite galaxies to inhabit the most massive haloes since we did not obtain the expected ξ^2 values for the clustering.

We are able to build mock galaxy catalogues that follow the luminosity function, the colour-magnitude diagram and the clustering as a function of luminosity and colour for a determined galaxy population. It is important to remark that these catalogues do not fulfill the properties for other galaxy populations, for instance one catalogue could follow the clustering at a certain luminosity but the same catalogue may not have the correct clustering for a brighter luminosity.

We developed a method to build a unique mock galaxy catalogue that reproduces at the same time the observed galaxy properties already mentioned. We assume $\alpha = 1.0$ but the relation between M_1 and M_{min} is not constant and depends on the halo mass following a certain function. We also introduced scatter in the relation between halo mass and the luminosity of central galaxies. This scatter is needed in order to decrease the amplitude of the clustering for the brighter galaxies. We found that a value of $\Delta_{\log(L_r)} = 0.15$ yields reasonable results. We also incorporate three Gaussian functions, instead of two, to describe the observed SDSS colour-magnitude diagram in order to have a better fit to the data. As a consequence we also modified

and found new recipes to assign colours to central and satellite galaxies to get the correct clustering as a function of colour. We showed the results of implementing this method in a halo catalogue extracted from a box with $L_{box} = 307.2$ Mpc/h of the MICE Grand Challenge run resulted in a catalogue with ~ 0.5 million galaxies. The luminosity function, the colour distribution as a function of the magnitude and the clustering as a function of luminosity and colour are also shown.

We found that it is not straight forward to produce a *unique* catalogue to follow different observed galaxy properties at the same time and it requires some changes in the method, including a lot of tuning in the parameters. The results showed very good agreement in the luminosity function and in the color-magnitude diagram. Regarding the clustering the results show in general quite good agreement with observations. However we have also pointed out different aspects of the galaxy clustering that could be improved. The amplitude of the clustering of bright galaxies at scales larger than $1 - 2$ Mpc/h is, in general, not perfect. The catalogue does not follow properly the trend of the faint red galaxies to be more clustered than the bright ones. The blue brightest galaxies, as well as the whole catalogue, have a slightly larger amplitude of the clustering at the largest scales. We will try to solve these issues in the following versions of the catalogue in the near future.

Chapter 6

Catalogue's validations and applications

This chapter shows the implementation of the method explained in the previous chapter using the halo catalogue extracted from the whole snapshot, covering one octant of the sky with a box-size $L_{box} = 3072$ Mpc/h, at z=0 of the MICE Grand Challenge run (rather than a little cube with a box-size $L_{box} = 307.2$ Mpc/h) and building a *unique* mock galaxy catalogue. The chapter also analyzes different aspects and possible scientific applications of the generated mock galaxy catalogue.

The first section shows the derivation of the luminosity function computing the error bars by using the Jack-knife re-sampling method. The second section shows the correlation function and the projected correlation function computed as an average of 10^3 non-overlapping volumes, each of them with a box-size $L_{box} = 307.2$ Mpc/h, for luminosity threshold, luminosity bin, and colour-luminosity bin samples. It also shows an estimation of the linear galaxy bias by comparing the galaxy correlation function and the linear dark matter correlation function. The third section shows the correlation function of the catalogue including the effects of galaxy velocities (redshift space distortions) at small (~ 0.1 Mpc/h) and intermediate (20-30 Mpc/h) scales. The fourth section shows the correlation function of the catalogue at the BAO scale, and the fifth section presents three applications already carried out of the mock galaxy catalogue.

6.1 Luminosity function of the snapshot at $z = 0$ of the MICE GC run

This section shows the estimation of the luminosity function of a galaxy catalogue generated using the MICE GC snapshot at $z = 0$[1]. To generate the galaxy catalogue we divide the whole halo catalogue into 10^3 non-overlapping cubes with a box-size of $L_{box} = 307.2$, and we populate with galaxies each cube. Then, in order to estimate the errors of the luminosity function (LF) of the whole catalogue we use the Jack-knife re-sampling method in the same way as in chapter 4 when we computed the

[1] Notice that halo masses are *shifted* to $z = 0.1$

errors of the halo mass function. Now we divide the whole catalogue into $N_{JK} = 10^2$ non-overlapping regions, and compute the LF in the full volume excluding one of the non-overlapping regions at a time. Again, we choose as a non-overlapping region a prism of square base with size 307.2 Mpc/h and height 3072 Mpc/h (which is formed by 10 of the cubes with a box-size of $L_{box} = 307.2$). As mentioned in chapter 4, Crocce et al. 2010 used $N_{JK} = 5^3$ to measure the errors in the halo mass function, and they checked that the estimates have already converged with varying N_{JK}. We expect the same behaviour for the luminosity function. The variance (defined as the relative error squared) in the i-luminosity bin of the number of galaxies is obtained as:

$$\sigma_{JK}^{(i)^2}(\Delta N) = \frac{1}{\overline{n}^{(i)^2}} \frac{N_{JK}-1}{N_{JK}} \sum_{j=1}^{N_{JK}} \left(n_j^{(i)} - \overline{n}^{(i)} \right)^2 \tag{6.1}$$

where $\overline{n}^{(i)}$ is the mean number density of galaxies for that luminosity bin. Figure 6.1 shows the LF of the MICE GC run at $z = 0$. The black triangles represent $\log\left(\frac{dn}{dMr}\right)$ in the mock galaxy catalogue. It is estimated as the mean value of the Jack-knife volumes. The red solid lines represent the luminosity function of all prism volumes and the blue solid line is the best fit Schechter function to the SDSS data. The luminosity function of the catalogue and the Schechter function fit are in very good agreement across the whole luminosity range.

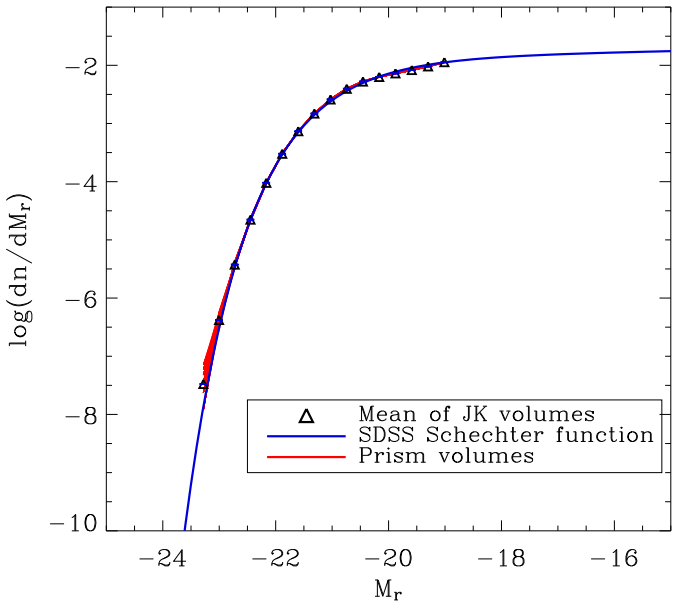

Figure 6.1: Luminosity function of the MICE GC run at $z = 0$. The black triangles represent the LF in the mock galaxy catalogue. The red solid lines shows the luminosity function of all prism volumes and the blue solid line is the best fit Schechter function to the SDSS data at $z = 0.1$.

6.2 Galaxy bias

Halo bias measures the relation between the matter density field and the halo density field. Galaxy bias measures the relation between the matter density field and the galaxy density field. In addition to the non-linear evolution of dark matter, baryonic physics takes an important role in the structure evolution of galaxies. The galaxy distribution depends on the efficiency of haloes to form galaxies. This efficiency depends on halo mass. Very small haloes cannot form galaxies. The inside gas is reheated by feedback processes, whereas in more massive haloes galaxy formation is inhibited by the long cooling time of the gas. Due to these phenomena, among others, there exists a scale dependent bias in the distribution of galaxies relative to the distribution of mass.

We derive the linear large scale galaxy bias in the same way as it is described in section 4.4.2 of chapter 4 where the MICE simulations are introduced. We derive the two point correlation function of galaxies and we estimate their bias assuming a linear bias relation with the dark matter field, $\delta_g(\mathbf{r}) = b_g^{Lin}\delta_m(\mathbf{r})$, and therefore:

$$\xi_g(r) = \left(b_g^{Lin}\right)^2 \xi_{DM}(r) \tag{6.2}$$

We estimate the two-point correlation function of galaxies, $\xi_g(r)$, from small scales ($\sim$ 0.17 Mpc/h) up to intermediate scales (10-30 Mpc/h). We divide the MICE GC run into 10^3 cubes of side 307.2 Mpc/h and we use the following estimator to compute $\xi_g(r)$ (it is called the natural estimator):

$$\xi_g(r) = \frac{DD(r)}{RR(r)} - 1 \tag{6.3}$$

where $DD(r)$ refers to the number of galaxy pairs in the simulation at a certain distance r and $RR(r)$ is the number of galaxy pairs as if they were spatially randomly distributed.

We estimate the two point correlation function, $\xi_g(r)$, and the projected correlation function, $w_p(r_p)$, for seven different luminosity thresholds, four luminosity bin (see table 6.1), and also for different galaxy types (central and satellite galaxies). We also compute $\xi_g(r)$ and $w_p(r_p)$ for red and blue galaxies for four luminosity bins samples.

$< M_r$	$M_{r1} < M_r < M_{r2}$
-19.0	-19.0 to -20.0
-19.5	-20.0 to -21.0
-20.0	-21.0 to -22.0
-20.5	-22.0 to -23.0
-21.0	
-21.5	
-22.0	

Table 6.1: Absolute magnitude thresholds and bins in the r-band.

Figure 6.2 shows the correlation function, $\xi_g(r)$, depending on the galaxy type for luminosity threshold samples. Different colours correspond to different galaxy luminosities. The left panel shows the correlation function for the whole catalogue, including central and satellite galaxies. The central and right panels refer to centrals and satellites respectively. We have not staggered the different luminosity threshold samples this time. The errors are computed as the standard deviation using 10^3 cubes of side 307.2 Mpc/h. Some error-bars are only plotted in one direction to avoid crowding the plot with little extra information. The agreement of the correlation function with observed SDSS data is similar to the previous chapter. However, we have not included SDSS data in the following plots for clarity. The black dashed line is the linear two-point correlation function of the dark matter MICE simulations, $\xi(r)^{MICE}_{Lin}$, calculated as the FT of the linear MICE power spectrum. There are many different interesting aspects in the three panels to look at. At the same luminosity, satellite galaxies are more clustered than central galaxies, except for the brightest sample. This is expected since in the HOD model satellite galaxies inhabit, in general, in haloes that contain enough mass to form satellite galaxies, and therefore these haloes are usually more massive and more biased. The brightest galaxies, centrals or satellites, inhabit in the most massive haloes and therefore they have the same bias. The amplitude of the clustering is very similar between galaxies with luminosity fainter than the characteristic absolute magnitude, $M_r^* = -20.44$. When $M_r < M_r^*$ the luminosity function drops very fast, and therefore there are less number of bright galaxies but their bias increases rapidly, which means that, in general, they reside in the more massive haloes. The change in the slope of the correlation function between scales larger and smaller than 1-2 Mpc/h is significant for the two brightest samples of the whole catalogue and satellite galaxies. For central galaxies, at $r < 1-2$ Mpc/h, due to the halo exclusion effect, there are very few pairs. Central galaxies (or haloes) cannot be placed closer than a certain distance. The bigger the halo, the larger this distance. The amplitude of the correlation function for the whole catalogue compared to $\xi(r)^{MICE}_{Lin}$ is always larger except for the faintest sample at the largest scales where it has almost the same amplitude, which means that the faintest sample is approximately an unbiased sample of galaxies. All luminosity threshold samples of satellite galaxies have a larger amplitude than $\xi(r)^{MICE}_{Lin}$. Only the three central brightest samples have a larger amplitude than $\xi(r)^{MICE}_{Lin}$, and therefore the fainter samples are anti-biased. Figure 6.3 shows the correlation function, $\xi_g(r)$, depending on the galaxy type for luminosity bins samples. The same trends as in figure 6.2 can be seen.

The left panel of figure 6.4 shows the correlation function, $\xi_g(r)$, for red galaxies separated into four luminosity bin samples whereas in the right panel blue galaxies are represented and split into three different luminosity bin samples. One can observe that red galaxies are more clustered than blue ones at the same luminosity, and the difference is larger at small scales. The difference in the slope at small scales is quite significant, while the slope of red galaxies is almost constant at all scales, the slope of blue galaxies changes and it is less steeper at small scales than at large scales. As mentioned in the previous chapter we do not reproduce precisely the small-scale clustering of red galaxies increasing towards low luminosities. We also mentioned that we will try to solve this problem in the following versions of the catalogue in

the near future.

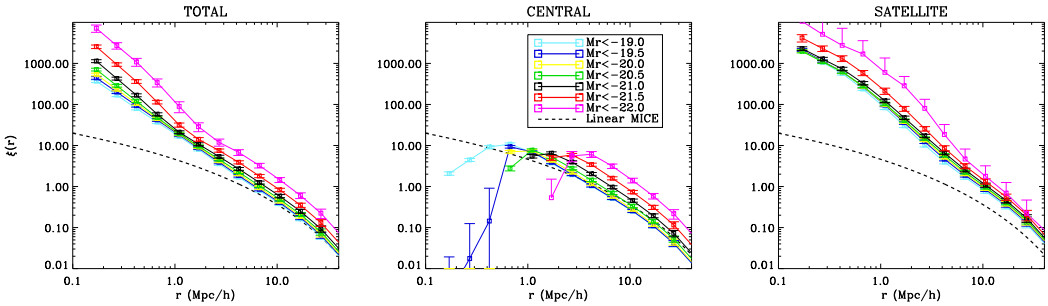

Figure 6.2: Two-point correlation function of total (left panel), central (central panel) and satellite galaxies (right panel) for the whole snapshot of the MICE GC run at z=0. Different colours represent different luminosity thresholds. The black dashed line is the linear two-point correlation function of the MICE simulations derived as the FT of the linear MICE power spectrum.

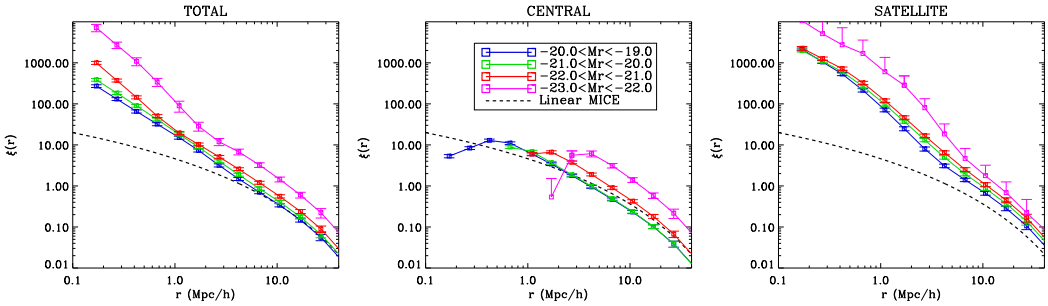

Figure 6.3: Idem as figure 6.2 but for luminosity bin samples.

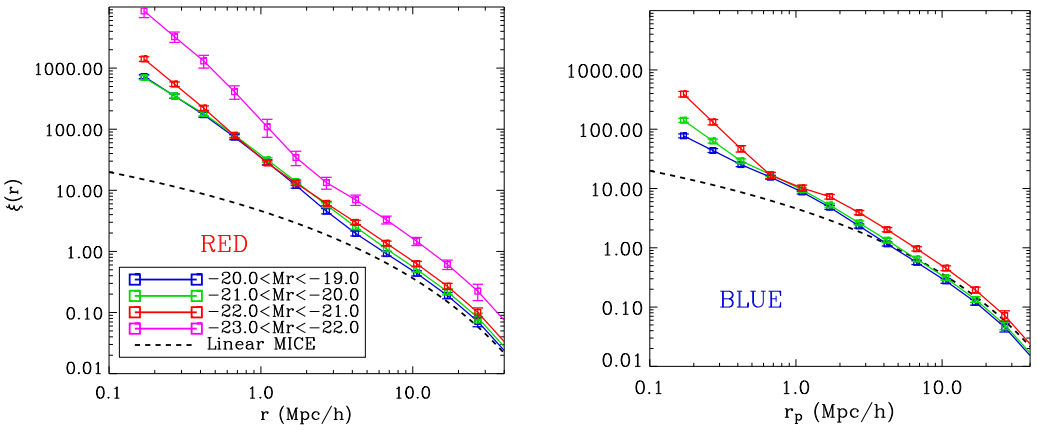

Figure 6.4: Two-point correlation function of red (left panel) and blue galaxies (right panel) for the whole snapshot of the MICE GC run at z=0. Different colours represent different luminosity bin samples. The black dashed line is the linear two-point correlation function of the MICE simulations derived as the FT of the linear MICE power spectrum.

Figures 6.5, 6.6 and 6.7 show the same as the three previous figures but for the projected correlation function, $w(r_p)$, rather than the two-point correlation function.

When looking at figure 6.5 one can see the same trends as in figure 6.2. At the same luminosity, satellite galaxies are more clustered than central galaxies, specially for the fainter samples. The projected correlation function, compared to the two-point correlation function, is flatter. The change in the slope at scales larger and smaller than 1-2 Mpc/h is also present, specially for the two brightest samples of the whole catalogue and satellite galaxies. Again, the amplitude of the clustering is very similar between galaxies with luminosity fainter than the characteristic absolute magnitude, and the bias of the galaxies increases rapidly for galaxies more luminous than $M_r = -20.44$. In this case, the amplitude of $w_p(r_p)$ at scales of the 1-halo term for central galaxies does not drop towards zero since we are integrating in the line-of-sight. Now it is possible to find central galaxies closer than any distance and all galaxies have an almost constant amplitude. The amplitude of the three brightest samples at scales of the 1-halo term, from larger to smaller scales, first slightly decreases down to $r \sim 0.3$ Mpc/h and then slightly increases up to $r \sim 0.17$ and reaches the value at scales $r \sim 1-2$ Mpc/h. Figure 6.6 shows $w_p(r_p)$ depending on the galaxy type for luminosity bin samples. We can see the same trends as in figure 6.5. The left panel of figure 6.7 shows $w_p(r_p)$ for red galaxies separated into four luminosity bins samples and in the right panel, blue galaxies are represented and split into three different luminosity bin samples. One can also observe the same trends as in figure 6.4.

Figure 6.8 and figure 6.9 show the linear galaxy bias of luminosity threshold and luminosity bin samples respectively as a function of scale extracted from the two-point correlation function following equation 6.2. The error-bars are derived propagating the errors derived in the two-point correlation function. Looking at the central panels of both figures, they show the same shape as the one derived in chapter 4 for different halo masses since we place central galaxies in every halo. As it is already mentioned, brighter galaxies are more clustered than fainter ones, and therefore the bias is larger. At scales $r < 1-2$ Mpc/h where there are a very few number of central pairs the bias drops towards zero very fast, exactly in the same way as the two-point correlation function. The whole catalogue (left panel) and satellites (right panel) have, in general, a very similar shape. Again, brighter galaxies are more clustered than fainter galaxies. At large scales, at the same luminosity, satellite galaxies have a larger bias than central galaxies, and central galaxies have a smaller bias than the whole catalogue. The bias for central and satellite galaxies, and for the whole catalogue, tends to a constant value, that depends on the galaxy luminosity and on the galaxy type, at large scales. This is expected since we assume that galaxies at large scales tends to follow the distribution of dark matter more closely.

Figures 6.10 shows the bias as a function of scale for red (left panel) and blue galaxies (right panel). Red galaxies are more clustered than blue galaxies for all luminosities and at all scales. Notice that the clustering at scales of the 1-halo term of red galaxies is not properly followed (explained in previous chapter). The blue line, that corresponds to the faintest sample, should have a larger bias, larger even than the red one in order to follow the observations. The red faintest galaxies are, in general, satellite galaxies that inhabit in the most massive haloes. This is the reason why they have such a large bias.

Finally we show the linear galaxy bias as a function of M_r. We derive the bias

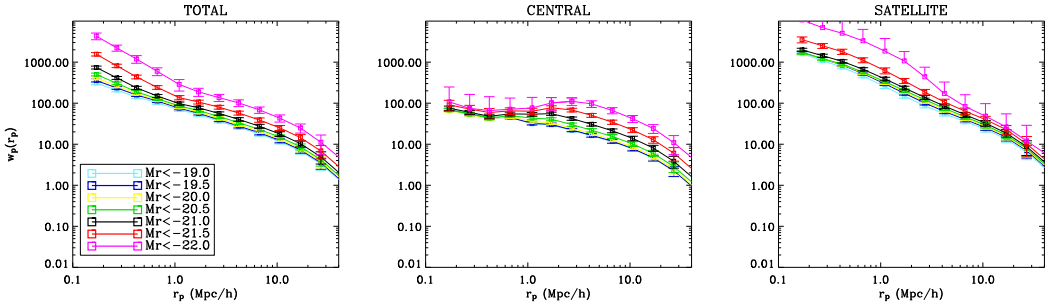

Figure 6.5: Idem as figure 6.2 but showing the projected correlation function.

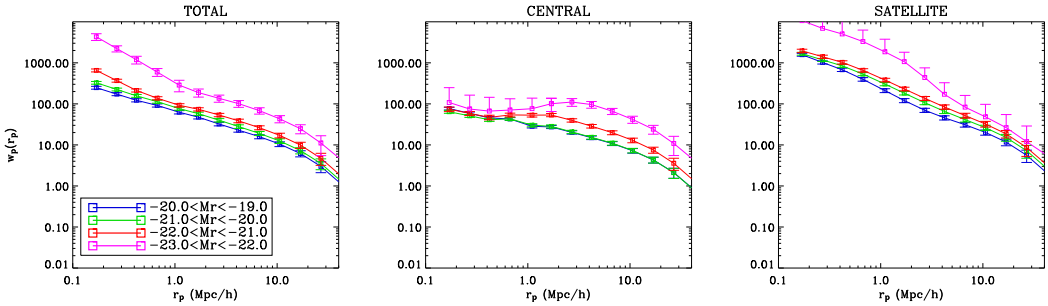

Figure 6.6: Idem as figure 6.3 but showing the projected correlation function.

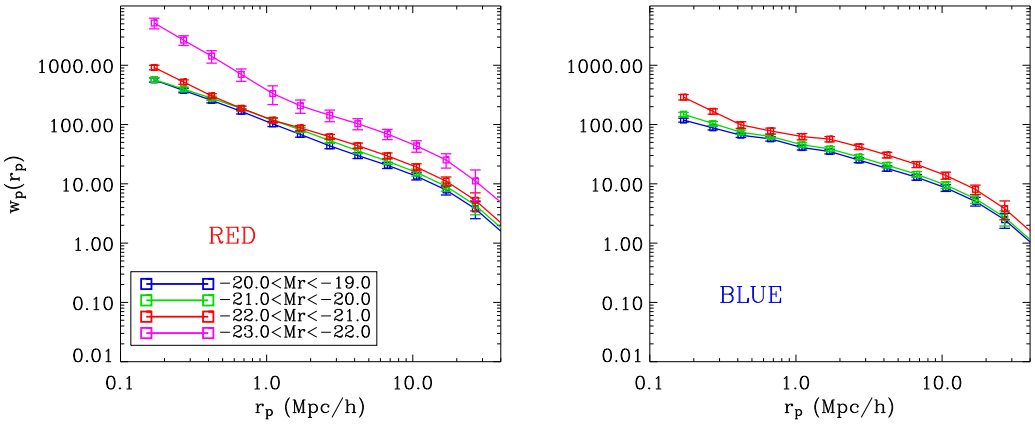

Figure 6.7: Idem as figure 6.4 but showing the projected correlation function.

by averaging the bias at the scales $r_1 = 6.0$ and $r_2 = 10.6$ Mpc/h. The left panel of figure 6.11 shows $b_g(< M_r)$ and $b_g(M_r)$ for the whole catalogue. The triangles refer to luminosity threshold samples and the asterisks represent luminosity bin samples. The blue dashed line is a fit to HOD model bias factors derived by Zehavi et al. 2011 given by the following expression:

$$b_g(> L_r) \times (\sigma_8/0.8) = 1.06 + 0.21(L_r/L_r^*)^{1.12} \tag{6.4}$$

where L_r is the r-band luminosity corrected to $z = 0.1$ and L_r^* corresponds to $M_r^* = -20.44$ (Blanton et al. 2003). The red dashed line is the fit of Zehavi et al. 2011 to

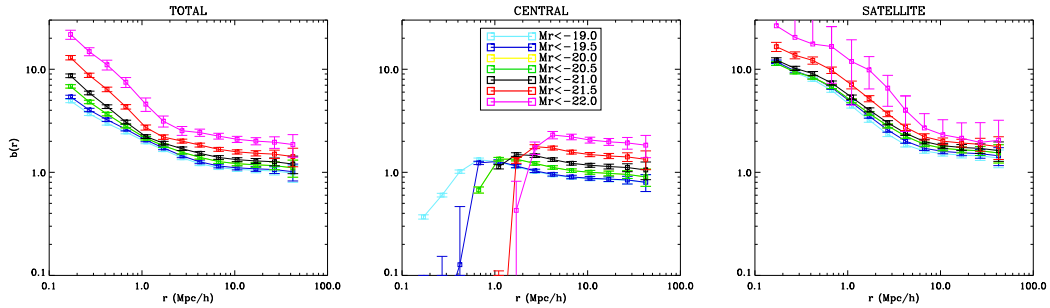

Figure 6.8: Galaxy bias of luminosity threshold samples as a function of scale. The left, central and right panels refer to the whole, central and satellite catalogues.

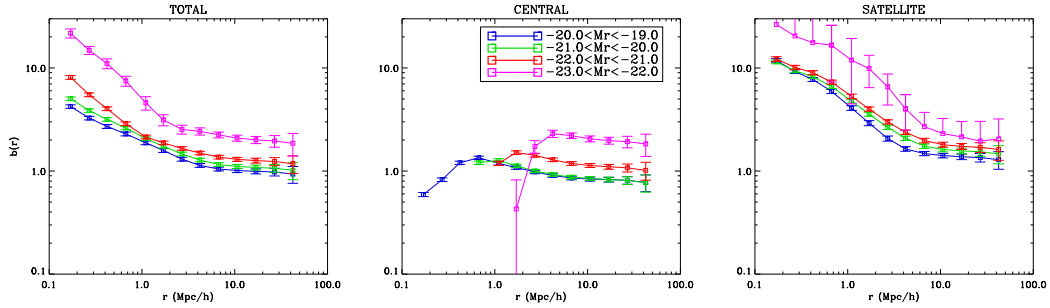

Figure 6.9: Galaxy bias of luminosity bin samples as a function of scale. The left, central and right panels refer to the whole, central and satellite catalogues.

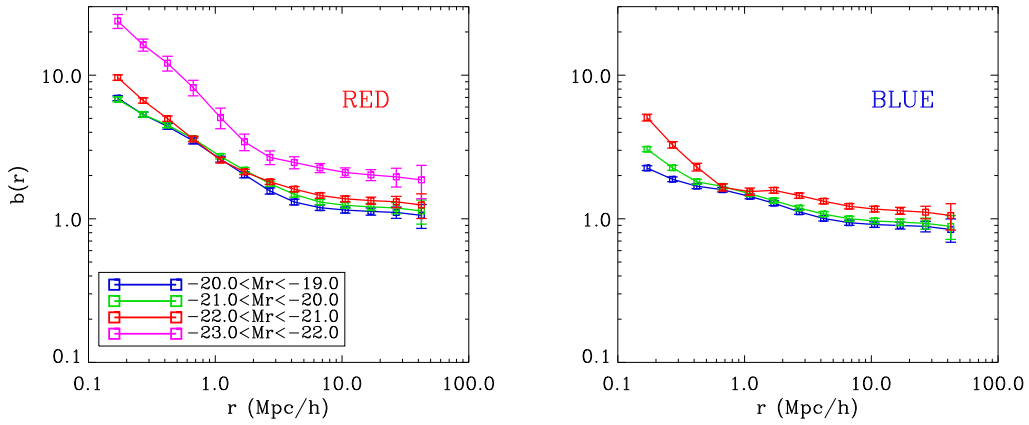

Figure 6.10: Galaxy bias of colour luminosity bin samples as a function of scale. The left panel refers to red galaxies and the right panel to blue galaxies.

the luminosity bin samples and is given by the following expression:

$$b_g(L_r) \times (\sigma_8/0.8) = 0.97 + 0.17(L_r/L_r^*)^{1.04} \tag{6.5}$$

We have also plotted the formula derived by Norberg et al. 2001 to fit the projected correlation functions in the 2dFGRS (black solid line):

$$b_g(> L_r)/b^* = 0.85 + 0.15(L_r/L_r^*) \tag{6.6}$$

where we take $b^* \equiv b_g(L_r^*) = 1.14$ as in Zehavi et al. 2011, and a modified fit to SDSS power spectrum measurements derived by Tegmark et al. 2004 (dotted black line):

$$b_g(> L_r)/b^* = 0.85 + 0.15(L_r/L_r^*) - 0.04(M_r - M_r^*) \tag{6.7}$$

The agreement between the bias of luminosity threshold samples of the catalogue with the fit of ZH11 is quite reasonable. And the agreement between the bias of luminosity bin samples with the fit obtained by ZH11 is also very good except for the brightest samples where our catalogue is more clustered that the fit. At fainter scales the agreement with the fit of Norberg et al. 2001 is almost perfect and very good also with the fit of Tegmark et al. 2004. The right panel of figure 6.11 shows the bias as a function of M_r for red (triangles) and blue galaxies (asterisks). Red galaxies are much more clustered than blue ones at all luminosities.

In figure 6.12 the bias as a function of M_r for central (left panel) and satellite galaxies (right panel) is shown. In both panels, the triangles refer to luminosity threshold samples and the asterisks refer to luminosity bin samples. In the left panel, the blue and black solid lines are derived using the analytical expression for the MICE halo bias (combining the expression of Manera et al. 2010 and the parameters found by Crocce et al. 2010 for the halo mass function) and converting halo masses to luminosities by using the relation, $M_{sh} - L_{gal}$, obtained when building the mock galaxy catalogue. The agreement is very good for both, luminosity threshold and bin samples. The figure should be similar to the results shown in figures 4.18 and 4.19 of chapter 4 since central galaxies are placed at the center of their host haloes. A possible reason of having a smallest bias than the halo bias fit for the brightest sample is the scatter introduced in the relation $M_{sh} - L_{gal}$ when assigning central luminosities. In the right panel we have also plotted the fits derived by Zehavi et al. 2011, Norberg et al. 2001 and Tegmark et al. 2004 for comparison. Satellite galaxies, at the same luminosity, are much more clustered than central galaxies and it is also very easy to observe that galaxies fainter than $M_r^* = -20.44$ have a similar bias and it increases very fast when their luminosity is brighter than M_r^*. Notice that certain central populations can be anti-biased, satellite populations have always $b_{sat}(L_r) > 1$ and the whole population may be almost unbiased.

6.3 Redshift space distortions

This section analyzes the distortions in the galaxy clustering when measuring galaxy redshifts due to their peculiar velocities. When one wants to measure how far a galaxy is instead of measuring directly its distance, one measures its redshift. The observed redshift is the sum of the redshift due to the expansion of the universe and the redshift due to the peculiar velocity of the galaxy. There are two well-known effects, the so-called Fingers-of-God effect, and the Kaiser effect. On small scales galaxy velocities within a cluster are well approximated by velocities that are independent Gaussians in each of the three cartesian components, with rms values that depend on the halo mass. This causes galaxies at the same distance to be at different redshifts and makes the distribution of galaxies in redshift space maps to

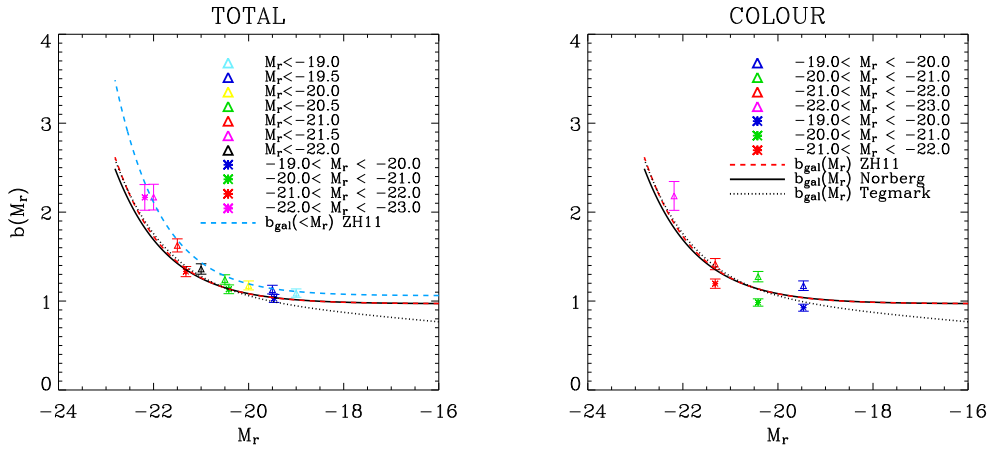

Figure 6.11: Galaxy bias as a function of M_r of luminosity threshold, $b_g(< M_r)$, and bin samples, $b_g(M_r)$, for the whole catalogue (left panel), and $b_g(M_r)$ for red and blue galaxies (right panel). Left panel: the coloured triangles and asterisks represent luminosity threshold and bin samples respectively. The light blue dashed line is the fit to the HOD model bias factors for luminosity threshold samples from Zehavi et al. 2011 (eq. 6.4). The red dashed line is the fit to the HOD model bias factors for luminosity bin samples from Zehavi et al. 2011 (eq. 6.5). The black solid line is a fit to projected correlation functions in the 2dFGRS (eq. 6.6) and the dashed curve is a modified fit to SDSS power spectrum measurements derived by Tegmark et al. 2004 (eq. 6.7). Right panel: the coloured triangles correspond to red luminosity bin samples and the coloured asterisks correspond to blue luminosity bin samples. The lines are the same as in the left panel for the luminosity bin samples.

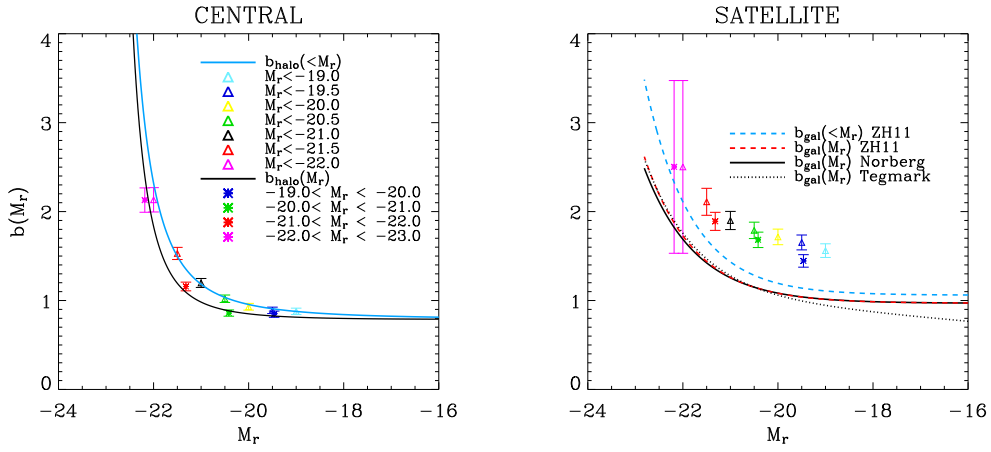

Figure 6.12: Idem as left panel of figure 6.11 but for central (left panel) and satellite galaxies (right panel). Left panel: the black line is the halo bias derived by Crocce et al. 2010 for the MICE simulations relating halo masses to absolute magnitudes by using the relation $M_h - L_{gal}$ presented in chapter 5 derived using the subhalo abundance matching technique. The blue line is derived by integrating the black line. Right panel: the lines are the same as the ones plotted in figure 6.11.

be elongated along the line-of-sight (like fingers). The Kaiser effect is evident at larger scales. Galaxies are falling into overdense regions and this makes the galaxies

between the overdensity and us to appear further away. However, galaxies that are placed further away than the overdensity appear to be closer. This effect produces the overdensity to be squashed in the line-of-sight. We can visualize these two effects when computing the two-point correlation function as a function of radial or line-of-sight, r_π, and perpendicular distance, r_p, in the $\xi(r_p, r_\pi)$ diagram since both effects are going to produce distortions in the positions and therefore distortions in the number of pairs depending on the distance at any scale.

We analyze the distribution of the galaxy velocities in a cube with a box-size $L_{box} = 307.2$ Mpc/h of the snapshot at $z = 0$ of the MICE GC run. Figure 6.13 shows the normalized distribution of the three different velocity components and the distribution of the velocity modulus for central (solid lines) and satellite galaxies (dashed lines). Since we have assumed that the velocity of satellite galaxies is derived as the sum of their host halo velocity plus the velocity due to virial motions inside the halo, the dispersion in the satellite distribution is larger than for centrals. It is interesting to notice that the mean values for all velocity components are negative. This means that even at such very large scales ($\sim L_{box}$) there exists a bulk motion. Galaxies, in average, in this particular cube, are moving towards the observer.

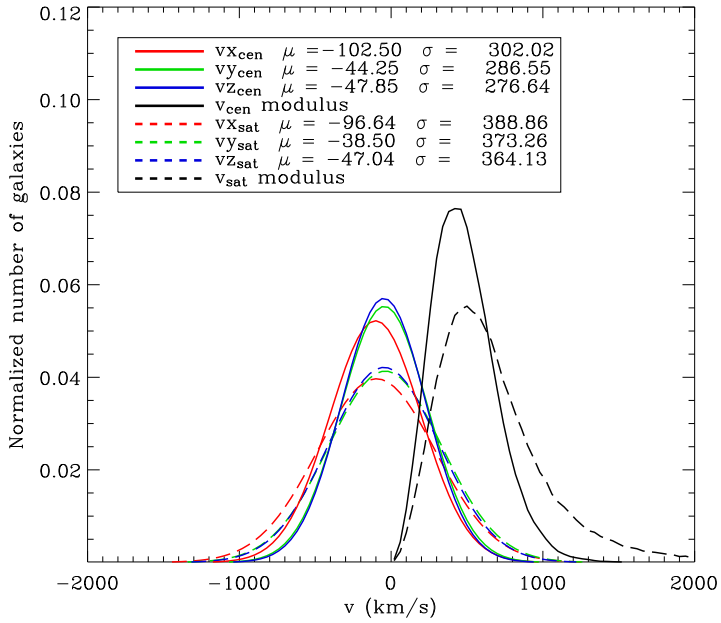

Figure 6.13: Galaxy velocity histogram. Solid and dashed lines show the Gaussian fits to the three different components of the central and satellite galaxy velocity, respectively. Black lines show the modulus of the velocities.

We use the expression 5.55 to derive the velocity dispersion of satellite galaxies. Figure 6.14 shows the normalized number of satellites as a function of the velocity dispersion of haloes. When two or more satellite galaxies inhabit in the same halo they will have the same value of σ.

Once we know the comoving-distance of a given galaxy to obtain the cosmological

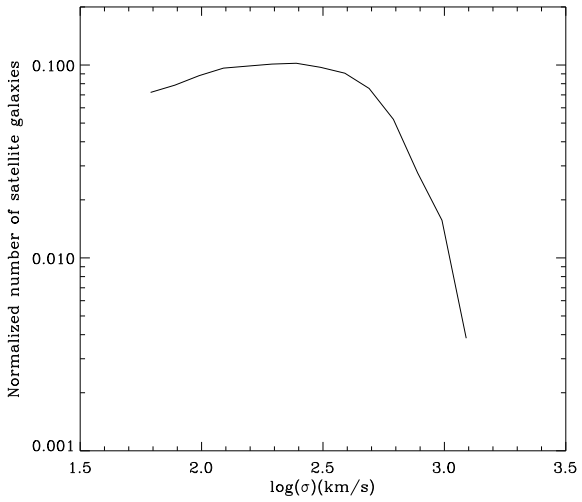

Figure 6.14: Normalized velocity dispersion histogram of satellite galaxies from a cube of box-size $L_{box} = 307.2$ Mpc/h extracted from the snapshot of the MICE GC run at z=0.

redshift we use a table generated by using the relation between the comoving-distance and the redshift:

$$r_{com}(z) = \int_0^z \frac{cdz'}{H(z')} \tag{6.8}$$

We locate the comoving-distance between two values in the table and then we linearly interpolate to get its corresponding redshift. We compute the distortion due to the peculiar velocity in the radial direction by directly using the Hubble parameter:

$$r_{||}^{obs} = r_{||}^{com} + \frac{v_{||}}{aH(z)} \tag{6.9}$$

where $a = \frac{1}{1+z}$ is the scale factor and we assume that all galaxies in the snapshot are at the same redshift, $z = 0.1$.

Figure 6.15 and 6.16 show the correlation function $\xi(r_p, r_\pi)$ of galaxies (central + satellite) brighter than $M_r < -19.0$ without taking and taking into account the velocity of galaxies respectively. We have mirrored the first quadrant along both axes in order to stress deviations from circular symmetry. The colours of the plots refer to different values of the two-point correlation function. The larger the scale the smaller the value of $\chi(r)$ (a few values of $\chi(r)$ are shown to see the trend). Both, the Fingers-of-God and the Kaiser effect, are clearly visible in figure 6.16. We can observe at small perpendicular scales, virial velocities of satellite galaxies in haloes causes a larger amplitude in the clustering (Fingers-of-God). In addition, at large scales the infall motion of galaxies towards overdensities causes the correlation function to be "flatter".

Figures 6.17 and 6.18 show the correlation function $\xi(r_p, r_\pi)$ of the total (left panel), central (center panel) and satellite (right panel) galaxies brighter than $M_r < -19.0$ without taking and taking into account the redshift space distortions respec-

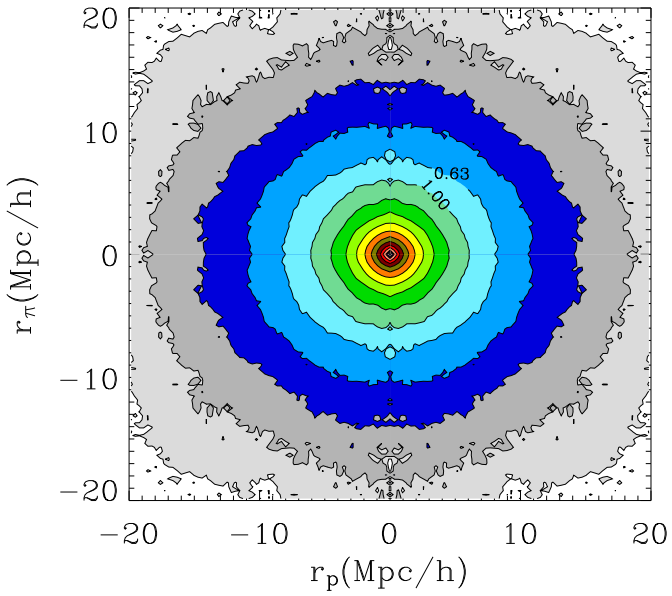

Figure 6.15: Two-point correlation function $\xi(r_p, r_\pi)$ as a function of radial or line-of-sight (r_π in vertical axis) and perpendicular distance (r_p in horizontal axis) for a sample of galaxies (central+satellite) brighter than $M_r < -19.0$ without taking into account galaxy velocities. Orange and red colours correspond to large values of the correlation whereas blue and grey colours to small values.

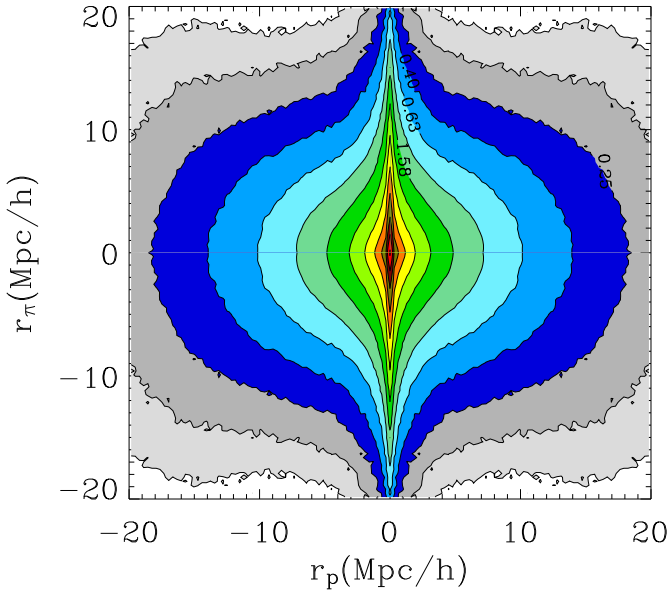

Figure 6.16: Idem as 6.15 but taking into account galaxy velocities. Both phenomena, the Fingers-of-God and the Kaiser effect, are clearly shown at small and large scale respectively.

tively. In this case we have not mirrored the first quadrant along both axes. Again, colours refer to different values of the correlation function as in previous figures, at larger scales the amplitude is smaller than at small scales. In figure 6.17 one can observe the same trends explained when measuring the correlation function in the previous section. Basically at the same luminosity ($M_r < -19.0$) satellite galaxies (right panel) are more clustered than the whole population (centrals + satellites), and the whole population is more clustered than central galaxies. The three panels show symmetry in the correlation function for both distances, along the line-of-sight, r_π and perpendicular to the line-of-sight, (r_p). The effects of redshift space distortions in figure 6.18 are evident. In the left and right panels one can observe both effects, the Fingers-of-God and the Kaiser effect. However in the central panel, where only the correlation function of central galaxies is shown, the Fingers-of-God effect is not visible as it is expected due to the way we assign velocities to central galaxies. We only add dispersion to satellite velocities. Central galaxies "only" feel the large scale Kaiser effect.

It is also interesting to see the effect of galaxy bias in the redshift space distortions. Following Dodelson 2003, working in the context of linear theory and assuming the distant observer approximation, one can reach to an expression that relates the Fourier transform of the redshift space overdensity, $\tilde{\delta}_s(\vec{k})$, and the real space overdensity, $\tilde{\delta}(\vec{k})$:

$$\tilde{\delta}_s(\vec{k}) = [1 + f\mu_{\boldsymbol{k}}]\,\tilde{\delta}(\vec{k}) \tag{6.10}$$

where $\mu_{\boldsymbol{k}}$ is defined to be $\hat{z} \cdot \hat{k}$, the cosine of the angle between the line of sight and the wavevector $\hat{k}$. And therefore one can relate the power spectrum in redshift space, $P_s(\vec{k})$, and in real space, $P(\vec{k})$:

$$P_s(\vec{k}) = P(\vec{k})\left[1 + \beta\mu_k^2\right]^2 \tag{6.11}$$

where it follows that $P_s(\vec{k})$ depends not only on the magnitude of $\vec{k}$ but also on its direction. The parameter β is not only the linear growth rate, $f = \frac{d\ln D}{d\ln a}$. There is an additional factor in β, though, due to the fact that the mass overdensity δ is not necessarily equal to the overdensity in galaxies, δ_g. If one defines galaxy bias as $\delta_g \equiv b_g\delta$, then the correction due to redshift space distortions in equation 6.11 is proportional to the parameter β. The smaller the galaxy bias, the more the galaxy population feels the Kaiser effect. We have just shown that satellite galaxies are always more clustered than central galaxies at a given luminosity and therefore central galaxies should feel more the Kaiser effect. One can observe in the central and right panels of figure 6.18 that the Kaiser effect (the squashing effect) is slightly larger for central galaxies than for satellite galaxies.

6.4 Baryonic acoustic oscillations

This section shows the angular two-point correlation function, $w(\theta)$, of the mock galaxy catalogue at the scale of the Baryonic Acoustic Oscillations (BAO). We do not intend to do an exhaustive study of the large scale clustering of the catalogue, we just want to show another check. We use, again, the catalogue generated by populating the halo catalogue extracted from snapshot at $z = 0$ of the MICE GC

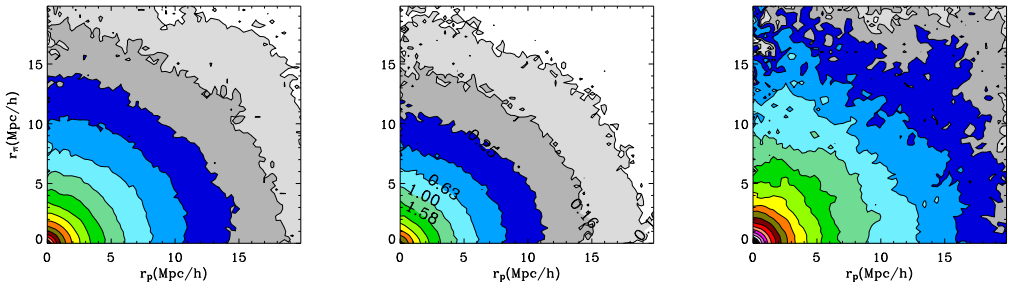

Figure 6.17: Two-point correlation function $\xi(r_p, r_\pi)$ as a function of radial or line-of-sight (r_π in vertical axis) and perpendicular distance (r_p in horizontal axis) for a sample of galaxies brighter than $M_r < -19.0$ without taking into account the velocity of galaxies. Orange and red colours correspond to large correlation and blue and grey colours to small correlation. Left, central and right panels correspond to the total (central+satellite), central and satellite catalogues respectively.

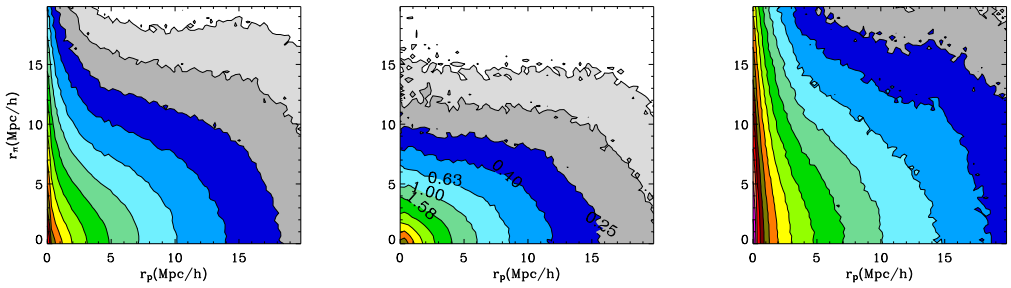

Figure 6.18: Idem as figure 6.17 taking into account the velocity of galaxies.

run. In particular we derive $w(\theta)$ for a comoving-distance spherical shell of width 300 Mpc/h, covering the range $(2772 - 3072)$ Mpc/h. The mean value of the shell is $z \sim 1.35$ in the MICE cosmology. We take into account in the analysis all galaxies ($L_r > L_r^{min}$) in the chosen region. The number of central galaxies is 48837263 ($\sim 49M$), the number of satellite galaxies is 22846895 ($\sim 23M$), so the total number of galaxies in the shell is 71684158 ($\sim 72M$). The solid angle of an octant of the sky is $(4\pi/8)$ rad^2 (or 5156.62 deg^2). The number of objects per $arcmin^2$ is 2.63, 1.23 and 3.86 for central, satellite and total galaxies respectively.

The BAO scale is assumed to be at $r_{com}^{BAO} \sim 100$ Mpc/h (e.g. Eisenstein et al. 2005). If we assume the observer to be at one vertex of the box, and we know that the mean comoving-distance of the shell is $\bar{r}_{com} = 2922$ Mpc/h we can compute at which angle the BAO feature should be in the angular correlation function.

$$\theta_{BAO} \sim \arctan \frac{r_{com}^{BAO}}{\bar{r}_{com}} \tag{6.12}$$

We obtain $\theta_{BAO} \sim 1.96$ deg.

In this case we compute the angular correlation function starting from `Healpix`[2] angular maps (Górski et al. 2005). We build a grid with a number of pixels of

[2]http://healpix.jpl.nasa.gov

$N_{side} = 512$ and a pixel size of ~ 0.01 deg^2, and we use a standard pixel estimator (Barriga & Gaztañaga 2002, Eriksen et al. 2004) to compute $w_{BAO}(\theta)$:

$$\hat{w}_{BAO}(\theta) = \frac{1}{N_{pairs}(\theta)} \sum_i \sum_j \delta_G^i \delta_G^j \tag{6.13}$$

where $\delta_G^i = N_{gal}^i / \hat{N}_{gal} - 1$ is the fluctuation in number of galaxies in the i-*th* pixel with respect to the mean in the angular map, pixels i and j are separated by an angle θ and $N_{pairs}(\theta)$ is the corresponding number of pixel pairs. The error bars are estimated using the Jack-Knife resampling method. We only use about 92% of the total area to compute the errors due to the way the code generates the Jack-knife volumes. Then we scale them to be appropriate for the whole octant.

Figure 6.19 shows the $w(\theta)$ for the total (black), central (blue) and satellite (red) galaxies. The cross-correlation between central and satellite galaxies (green) is also shown. The black dashed line is the theoretical angular correlation function of the MICE simulations. In order to compute it we use the MICE power spectrum, then we compute its FT to get the two point correlation function (we normalize it to get $\sigma_8 = 0.8$). Finally, we integrate to obtain the projected correlation and transforming the distance into angles using the mean value of the redshift bin. The inset shows an expanded view of figure 6.19 in order to better show the region where the acoustic peak is present and where the correlation function becomes negative. Looking at the outer figure one can observe that the angular correlation function decreases very fast at small angles until $\theta \sim 0.5$ and then becomes almost flat around zero. Looking at the inset, $w(\theta)$ is decreasing until $\theta \sim 1.5$ where the amplitude starts to be almost flat at ~ 0.5 deg, in this angle interval is where the θ_{BAO} should be, then decreases and becomes negative ($\theta \sim 2.3$), and then increases slowly towards zero. As described in previous sections, one can observe that, at the same luminosity, satellite galaxies (red) are the most clustered. It is interesting to notice that the cross-correlation between central and satellite galaxies is less clustered than satellite galaxies themselves but it is more clustered than the autocorrelation of the whole catalogue and the central catalogue. Looking for example at the satellite clustering, another aspect to notice, is that the amplitude of the clustering when the correlation is positive is the largest, however it is the smallest when the correlation is negative. This behaviour is also expected since the integral of the correlation function over the whole angle range is zero. The error-bars are larger for satellite galaxies because the number of satellites is smaller than the number of objects in the other galaxy samples. Figure 6.20 also shows the angular correlation function but multiplied by θ^2 which flatten out the curve and at the same time magnified the "bump" at the scale θ_{BAO} in the correlation.

6.5 Applications

We expect the catalogue to be used in many different areas and for many different purposes since it follows some global and important properties of the galaxy population. We have covered in this chapter without entering into detail different scientific aspects that can be studied using our mock galaxy catalogues, such as

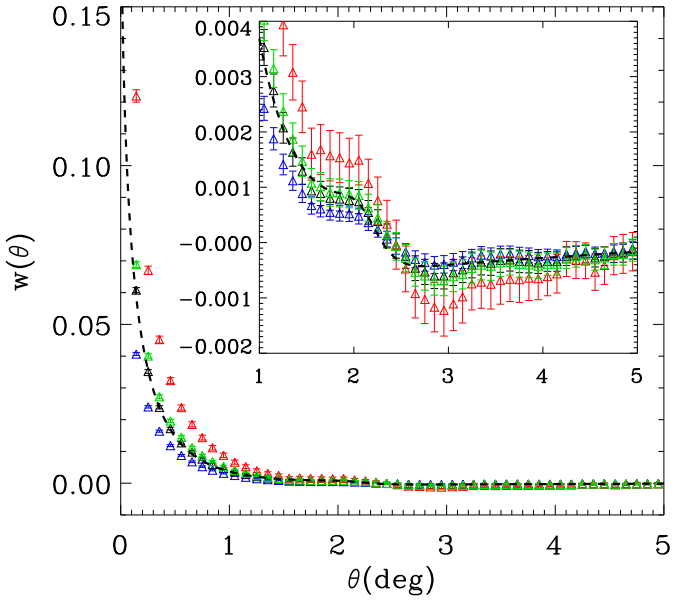

Figure 6.19: Angular correlation function for the whole (black), central (blue) and satellite (red) galaxies. The green triangles refer to the cross-correlation between central and satellite galaxies. The black dashed line is the theoretical angular correlation function of MICE simulations. The inset is an expanded view.

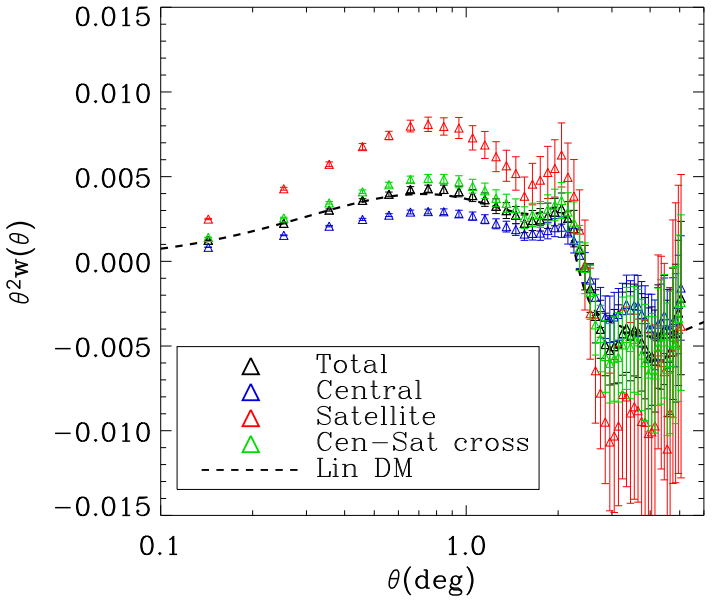

Figure 6.20: Angular correlation function times θ^2. Colours refer to the same correlations as in the upper figure.

galaxy bias, redshift space distortions or baryonic acoustic oscillations. Currently the mock galaxy catalogues are being used in several applications. This section briefly describes them.

The first versions of the catalogue were used to test the cluster finder code developed by Farrens et al. 2011. This first test helped us to realize that the catalogue had problems with the luminosities and colours of the galaxies. It was the time when the catalogue did not fit correctly the luminosity function.

The Physics of the Accelerated Universe project (PAU) expects to obtain photometric spectra from $\sim$ 10M galaxies. The PAU Data Management group is developing different pipelines which will reduce the observed images taken at the William Herschell telescope at La Palma and produce galaxy catalogues. Our mock galaxy catalogues are being used to feed the pipelines, the data base and the storage system with formatted data before real data arrives. This is useful to test and optimize the data reduction algorithms with known data and to identify possible survey-specific handicaps. Morphological properties, that follow empirical distributions extracted from the stuff code[3], and spectral energy distributions (SED), that are derived using a relation between galaxy colour $(g - r)$ and 66 templates of galaxy spectra, are included in the catalogue. Figure 6.21 shows the first simulated image produced by the PAU data management pipeline.

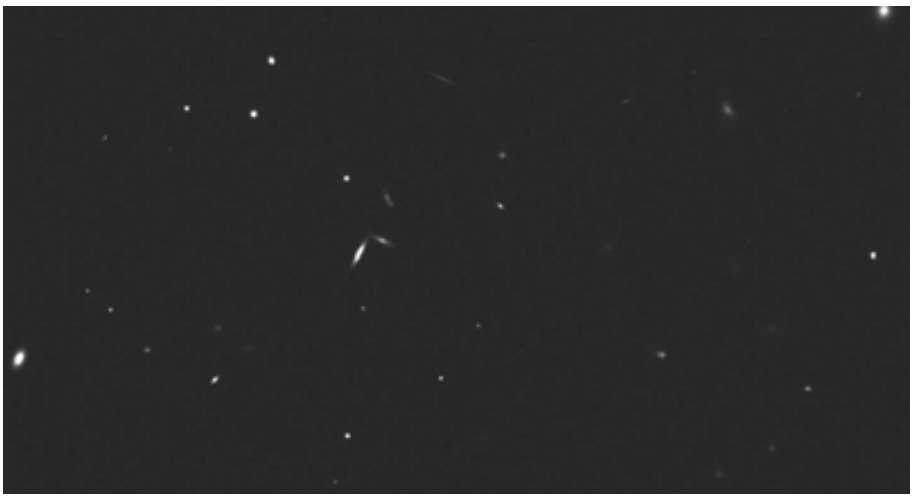

Figure 6.21: First simulated image generated by the PAU data management pipeline using our mock galaxy catalogue as input.

We have also generated a version of the catalogue that has been made available to the DES collaboration for DES-related studies. The catalogue contains 117 million galaxies. It is generated by populating a halo catalogue extracted from the lightcone output of the MICE Grand Challenge run. It approximately reproduces the magnitude limits of the DES survey. The catalogue provides information about positions, magnitudes and colours. In addition, we have included shear information for all galaxies out to $z = 1.4$ extracted from the input MICE simulation. Figure

[3]Stuff code is part of the software developed by Emmanuel Bertin (webpage:http://www.astromatic.net/software)

6.22 shows the shear amplitude of dark matter (left panel) and galaxies (right panel) for sources at redshift z=1.

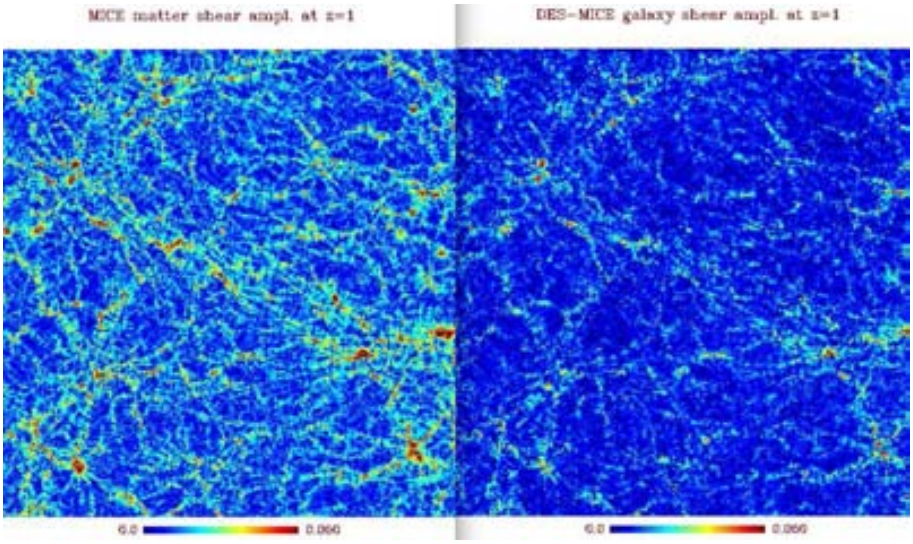

Figure 6.22: Left panel shows the shear amplitude of the dark matter simulation at z=1 from the lightcone of the MICE GC run. Right panel shows the same but for the galaxy catalogue.

6.6 Summary

Different properties of the last version of the generated mock galaxy catalogue have been analyzed: the luminosity function, the linear large scale galaxy bias, the effect of the redshift space distortions in the galaxy clustering and the angular correlation function at the BAO scale.

The resulting mock galaxy catalogue is generated by using the halo catalogue extracted from the snapshot at z=0 of the MICE GC run. The total number of haloes in the halo catalogue is $\sim 350M$, with $\log M_h^{min} = 11.31$. Since one central galaxy is placed at the center of each halo, the catalogue contains the same number of central galaxies. The number of satellite galaxies is $\sim 150M$ and therefore the total number of galaxies in the catalogue is $\sim 500M$. The minimum absolute magnitude of the catalogue is $M_r = -18.78$.

The luminosity function of central, satellite and total (central+satellite) galaxies has been computed. We have derived the errors of the luminosity function using the Jack-knife re-sampling method in the same way as we did in chapter 4 for the halo mass function of the MICE GC run. The fit against the SDSS data derived from Blanton et al. 2003 is almost perfect as expected since the method forces the galaxies to follow the observed luminosity function.

In order to compute the galaxy bias we have derived the correlation function of galaxies dividing the whole simulation volume into 10^3 non-overlapping cubes of box-size $L_{box} = 307.2$ Mpc/h. We showed the correlation function for different luminosity threshold, luminosity bin, and colour-luminosity bin samples and we have estimated

the galaxy bias using the linear MICE correlation function of the dark matter density field derived as the FT of the linear MICE power spectrum. We have computed the linear galaxy bias as a function of the luminosity and colour assuming a linear relation between the dark matter field and the galaxy field.

We have shown what are the effects of including the velocity of galaxies in the determination of galaxy redshift in the two-point correlation function, and we have also shown the clustering of galaxies at the BAO scale for a sample of galaxies at high redshift (~ 1.35).

We have also presented 3 applications already carried out. The first versions of the catalogues were used to test the cluster finder code developed by Farrens et al. 2011. A modified version of the catalogue where morphological properties and spectral energy distributions are assigned to every galaxy is used to feed the Data Management pipelines in the PAU project. Our science group has also provided a mock galaxy catalogue to the Euclid project collaboration including shear information to each galaxy.

Chapter 7

Discussion and conclusions

The main goal of this thesis was to present a method to build mock galaxy catalogues using the halo occupation distribution model. We populate with galaxies halo catalogues extracted from the Marenostrum Institut de Ciències de l'Espai Simulations (MICE). The catalogues follow some observed global properties of the galaxy population already observed, such as the luminosity function, the colour-magnitude diagram and the clustering as a function of luminosity and colour. The observed data constraints come from the Sloan Digital Sky Survey (SDSS). We hope that the catalogues are going to be used for many different purposes in future and current galaxy surveys such as DES, PAU or Euclid. This chapter summarizes the work presented in this thesis and lists the conclusions. The last section discusses the ongoing and future work.

7.1 Summary and conclusions

The following list enumerates and discuss the steps followed during the thesis:

- We have presented some of the most important present cosmological probes to measure cosmological parameters, and the benefits and needs of building mock galaxy catalogues to be able to fully exploit data from future galaxy surveys. We have shown different methods to produce mock galaxy catalogues using (or not) N-body simulations.

- We have described in more detail the halo model and the halo occupation distribution (HOD) model. The halo model describes the non-linear gravitational clustering. This approach assumes as the main idea that all the matter in the universe is comprised in virialized dark-matter haloes. We have explained some ingredients of the halo model, such as the spatial distribution of haloes, their mean number density, the amplitude of the clustering of haloes as a function of their mass and their mass density profile. All these ingredients are necessary when generating the mock galaxy catalogues using the HOD model. The HOD is one of the methods to relate the dark matter and galaxy distributions. It treats galaxy bias in a statistical way. The HOD describes the number of galaxies in a halo in terms of a probability distribution that depends on some

free parameters and the mass of the halo. It also gives prescriptions to place galaxies in the halo and to describe galaxy velocities within dark-matter haloes.

➤ We have briefly described the history and the basics of dark matter N-body simulations. The importance of computational advances has been stressed as well as some improvements in the algorithms to compute the gravitational force between particles.

➤ We have presented the Marenostrum Institut de Ciències de l'Espai simulations (MICE). The MICE simulations assume a flat concordance ΛCDM model with parameters $\Omega_m = 0.25$, $\Omega_\Lambda = 0.75$, $\Omega_b = 0.044$, $h = 0.7$ and the spectral index $n_s = 0.95$. The linear power spectrum is normalized to yield $\sigma_8 = 0.8$ at $z = 0$, and the initial conditions are given by either using the Zeldovich approximation or $2nd$ order Lagrangian Perturbation Theory (2LPT), depending on the run. We have introduced the suite of the MICE simulations, which covers a wide range of halo masses and volumes, from volumes comparable to the SDSS main sample (0.1 $h^{-3}Gpc^3$) to DES (9 $h^{-3}Gpc^3$) or even Euclid ($\sim$ 100 $h^{-3}Gpc^3$), and mass resolutions, from $3 \times 10^{12} h^{-1} M_\odot$ down to $3 \times 10^8 h^{-1} M_\odot$.

➤ We have specially analyzed the Grand Challenge (GC) run, which is one of the largest N-body simulations up to now with $N = 4096^3$ particles and $L_{BOX} =$ 3072 Mpc/h. Using as input the halo catalogue extracted from the snapshot of the GC run at $z = 0$, we have computed its halo mass function by using the Jack-knife resampling method. We have compared our derivation of the errors with the previous work developed by Crocce et al. 2010. They also analyzed the MICE simulations but they did not include the MICE GC run since it had not been performed yet. Our error measurements agree with the errors computed by Crocce et al. 2010 in the same halo mass range and the same simulation size.

➤ We have also fitted a Schechter-like function to the halo mass function of the GC run at $z = 0$:

$$\frac{dn}{d\log M_h} = a_0 \left(\frac{M_h}{M_h^*}\right)^{a_1} \exp\left[-\left(\frac{M_h}{M_h^*}\right)^{a_2}\right] \tag{7.1}$$

where $a_0 = 6.1065 \times 10^{-5}$, $a_1 = -0.8438$, $a_2 = 0.6946$ and $M_h^* = 2.3880 \times 10^{14}$.

➤ We have derived the two-point correlation function at small scales ($\sim$ 0.1 Mpc/h) up to intermediate scales ($\sim$ 20 − 30 Mpc/h) of haloes depending on their mass. We have computed the large scale halo bias extracted from the two-point correlation function assuming a linear relation between the dark matter field and the dark matter halo field, and we have shown the scale dependence of bias for different halo masses. More massive haloes are more clustered than the lighter ones. Haloes less massive than a certain mass ($\log(M_h) \leqslant 12.5$) have a similar bias ($b_h \sim 1$) and haloes more massive than that characteristic mass have a bias that increases very fast with mass. Therefore rare haloes, the most massive ones, are the most clustered. We compared our bias results

to analytical expressions from Manera et al. 2010. These authors presented a method to fit functional forms to halo abundances and linear halo bias. We have used their bias expression and the parameters found by Crocce et al. 2010 when they fitted the MICE halo mass function using a Sheth & Tormen-like mass function, to compare to our linear bias derivations from simulations. The results show a very good agreement between the halo bias derived from simulations and the analytical expressions.

➤ We have populated the MICE simulations with galaxies and we have produced different mock galaxy catalogues. As a first approach we have reproduced the mock galaxy catalogue that was released by our scientific group to the DES collaboration for a DES simulation challenge. In this case we assigned galaxies to dark matter particles in the MICE N2048L3072 run to follow a determined galaxy redshift distribution.

➤ We have also built mock galaxy catalogues by populating with galaxies different halo catalogues extracted from different MICE runs using the HOD model, which is a much more complex approach than the previous one. We started by using the intermediate run, MICE N2048L3072, with $N = 2048^3$ particles and $L_{BOX} = 3072$ Mpc/h, and we finished by using the GC run. The latter contains 8 times more particles than the former in the same volume and therefore we gained 8 times in mass resolution. This is very important since in our method to build mock catalogues, central galaxy luminosities depend strongly on the halo mass.

➤ We have had in mind, since the beginning, to be as simple as possible when building the mock galaxy catalogues. Because of this we started by using the simplest HOD model, which only uses 3 free parameters to describe the number of galaxies inside the haloes (M_{min}, M_1 and α), and by following algorithms from previous works. We have described the algorithm developed by Skibba & Sheth 2009 to produce mock galaxy catalogues. They introduced galaxy colours in the framework of the HOD, which is a previous step to include the full spectral energy distribution. They claimed that the mock galaxy catalogues produced by following their algorithm fit the luminosity function, the luminosity dependence of the two-point correlation function and, in addition, the galaxy clustering dependence on colour.

➤ We have found that it is not straight forward to follow prescriptions from other authors when they assume a different cosmological model and therefore their input cosmological simulations are different too. We found that the halo mass luminosity relation proposed by Skibba & Sheth 2009 from SDSS clustering data obtained by Zehavi et al. 2005 does not work perfectly when applied to the MICE simulations. It does not provide a good fit to the luminosity function either in the shape or in the amplitude. The main difference comes from the halo mass function due to different cosmologies. Even correcting for this issue by *shifting* masses from one cosmology to another it still does not work. Little changes in the relation produce big changes in the luminosity function. We

reached to the conclusion that it is necessary a more complicated relation (rather than an exponential function with 2 free parameters) to correctly fit the SDSS luminosity function using MICE simulations.

➤ We have derived *our own* HOD parameters. First, we have tried to analytically derive only two HOD parameters, M_{min} and M_1, by only using two conditions, the mean number density of objects as a function of the luminosity and the linear mean large scale galaxy bias at a certain scale (we assumed $\alpha \sim 1$). We have found that the first condition can constrain M_{min} but the second one is not enough to constrain M_1. We have found big degeneracies and therefore we need more constraints. Second, we have derived the best HOD parameters, M_{min}, M_1 and α, to match the luminosity dependence of the clustering using the MICE simulations, but in this case we have used as constraints the whole two-point correlation function from small scales $r \sim 0.1$ Mpc/h up to scales $r \sim 30$ Mpc/h from SDSS data shown in Zehavi et al. 2011 (in addition to the mean number density of galaxies). We have built a grid of 600 catalogues each of them using a different set of HOD parameters, M_{min}, M_1 and α, covering a wide range of values, by populating a halo catalogue extracted from the snapshot of the MICE GC run at $z = 0$ with box-size $L_{BOX} = 307.2$ Mpc/h. We have still followed the algorithm proposed by Skibba & Sheth 2009 to produce the catalogues but we have introduced some modifications:

- We have used the SubHalo Abundance Matching technique (SHAM) to match the best fit Schechter function to the SDSS data derived by Blanton et al. 2003. We have derived and modelled a relation between mass and galaxy luminosity, $M_{gal} - L_{gal}$, by comparing two cumulative functions, the observed mean number density of galaxies with luminosity brighter than a certain value, $n_{gal}(< M_r)$, using the best-fit Schechter function derived by Blanton et al. 2003, and the cumulative number density of objects (central and satellite galaxies) that inhabit in haloes of mass greater than a certain value, $n_{gal}(> M_{min}) = n_{cen}(> M_{min}) + n_{sat}(> M_{min})$.
- We have assumed that all galaxies follow the $M_{gal} - L_{gal}$ relation and we have been able to derive the contribution of satellite galaxies to the total galaxy luminosity function by using the relation $M_{gal} - L_{gal}$ in the equation $n_{sat}(> M_{min}) = \int_{M_{min}}^{\infty} \frac{dn}{dM_h} \left(\frac{M_h}{M_1}\right)^{\alpha} dM_h$. Finally we have subtracted $n_{sat}(> M_{min})$ from the observed total cumulative galaxy luminosity function and we have obtained the central contribution.
- We have derived the total number of central galaxies in the catalogue, N_{cen}, having the halo mass function and the HOD parameters, and we have assigned a luminosity for each central galaxy using the relation $M_{gal} - L_{gal}$. We have also computed the total number of satellite galaxies, N_{sat}, and we have generated randomly N_{sat} "available" luminosities sampling the cumulative satellite luminosity function, $n_{sat}(> L_r)$, already computed.
- We have assumed that central galaxies are, in general, the brightest galaxies and therefore we have enforced satellite galaxies not to be brighter than

1.05 times the central galaxy luminosity.

- We have placed satellite galaxies in the haloes following a NFW density profile. We did not obtain the expected correlation function at very small scales ($r \leqslant 0.5$ Mpc/h). Therefore we have had to modify the NFW density profile. We have made the satellite galaxies to be more concentrated than a NFW density profile. The faintest satellite galaxies are even more concentrated than the brightest ones.
- We have had to decrease the clustering of the catalogues since they still did not yield the expected χ^2 values. We decreased the clustering of the catalogue by decreasing the clustering of satellite galaxies since we assumed a univocal relation between central luminosities and their host halo masses. We penalized the faintest satellite galaxies to inhabit the most massive haloes by setting a straight line in the luminosity-halo mass plane, computing a distance, d, for every available satellite luminosity and the halo mass, and imposing a condition to the distance that basically makes the faintest satellite galaxies not to be placed at the most massive haloes.
- We have derived the best set HOD parameters by computing the minimum χ^2 to fit the luminosity dependence of clustering. Every mock galaxy catalogue generated followed the luminosity function no matter how many galaxies it contains and the halo mass - central luminosity relation changed depending on the HOD parameters accordingly. Mocks and observations agree although we still have found small discrepancies. The fits are quite reasonable for luminosities fainter than $M_r^* = -20.44$. For all luminosities, at large scales, the fits have higher amplitude compared to the observed data. This feature is even more significant for the brightest samples ($M_r < M_r^*$). The results were not surprising since we have only used three HOD parameters to describe the number of galaxies inside each halo. Recent works (e.g. ZH11 or Leauthaud et al. 2011) point to a number of at least five free parameters for a better fit, which seem to be specially necessary for the brightest samples since the mean number of galaxies in each halo cannot be perfectly described by a step function and a power law.

- We have realized that the fact that every galaxy population had its own set of HOD parameters has its advantages and drawbacks. One is able to produce a mock galaxy catalogue given a set of HOD parameters that follows perfectly the global luminosity function and the clustering for galaxies brighter than a certain luminosity, L_{r1}, but it will not follow correctly, in general, the clustering for galaxies brighter than another given luminosity, $L_{r2} > L_{r1}$. Specially if $L_{r1} < L_r^* < L_{r2}$ because galaxies fainter than M_r^* have almost the same bias and the bias grows much faster when $M_r < M_r^*$.
- We have developed a method to generate a *unique* catalogue that reproduces, not only the clustering for one luminosity threshold sample or one luminosity bin sample, but a galaxy catalogue that follows at the same time the galaxy

clustering for all luminosity threshold and bin samples, as well as the luminosity function, the colour-magnitude diagram and the colour dependence of clustering. We have introduced some changes with respect to the previous method:

- We have assumed the parameter M_1 to be a function that depends on M_h since we have found that the factor f_{M_1}, that multiplies M_{min} to obtain M_1, is not constant and varies with M_h. We modeled it with the following function since it provides a reasonable fit to observations:

$$f_{M_1} = 0.5((a_1 - a_2)\tanh(s_1(b_1 - \log M_h)) + \\ +(a_3 - a_2)\tanh(s_2(\log M_h - b_2)) + (a_1 + a_3)) \tag{7.2}$$

  where $a_1 = 25.0$, $a_2 = 11.0$, $a_3 = 14.0$, $b_1 = 11.5$, $b_2 = 12.5$, $s_1 = 2.0$ and $s_2 = 2.50$.

- We have had to introduce scatter in the relation between halo mass and central galaxy luminosity since the clustering of the mock galaxy catalogue of the brighter galaxies was much larger than observed data. In order to implement the scatter we have generated an "unscattered" luminosity function that after introducing the scatter in the central luminosities the cumulative luminosity function of the catalogue will match the observed cumulative Schechter luminosity function. We have computed $M_{gal} - L_{gal}$ using the same SHAM technique but in this case with the "unscattered" luminosity function and the cumulative number density of objects (central and satellite galaxies), $n_{gal}(> M_{min})$. We then assigned central luminosities following this relation and we only introduced scatter if the luminosity is brighter than the characteristic luminosity, L_r^* ($M_r^* = -20.44$).

- We have included another Gaussian component to better fit the distribution of $g - r$ colour vs. the absolute magnitude in the r-band. We have changed the recipes to determine the galaxy colours depending on the galaxy type (whether it is a central or a satellite galaxy) in order to fit correctly the colour-magnitude diagram and at the same time the clustering as a function of the colour.

➤ We have shown the results we obtain after implementing our method to build a *unique* mock galaxy catalogue by populating a halo catalogue extracted from the snapshot of the MICE GC run at $z = 0$ with box-size $L_{BOX} = 307.2$ Mpc/h, compared to SDSS data. We have derived the contributions to the luminosity function of central and satellite galaxies. The best fit Schechter function to the SDSS data derived by Blanton et al. 2003 and the luminosity function of the generated mock agree almost perfectly. The agreement between the SDSS colour-magnitude diagram and that of the mocks is also very good. The luminosity dependence of galaxy clustering for different luminosity threshold samples compared to the SDSS data agree quite reasonable although at large scales the amplitude of the clustering is still slightly larger that SDSS data,

specially for galaxies brighter than M_r^*. The same occur with the luminosity bin samples. Regarding the colour dependence of clustering our catalogue does not reproduce precisely the small-scale clustering of red galaxies increasing toward low luminosities and in addition, the brightest blue galaxies are too clustered compared to observations. We have mentioned that introducing a condition depending on the galaxy colour during the satellite luminosity assignation could be a possible solution. We plan to solve these issues in the near future.

➤ We have done some validations of the generated *unique* catalogue:

- We have derived the luminosity function computing the error bars by using the Jack-knife re-sampling method in the same way as we have computed the errors of the halo mass function. The luminosity function of the catalogue and the Schechter function fit are in very good agreement across the whole luminosity range.
- We have estimated the two point correlation function, $\xi_g(r)$, and the projected correlation function, $w_p(r_p)$, for seven different luminosity thresholds, four luminosity bins (see table 6.1), and also for different galaxy types (central and satellite galaxies). We have also computed $\xi_g(r)$ and $w_p(r_p)$ for red and blue galaxies for four luminosity bin samples.
- We have measured the linear large scale galaxy bias extracted from the two-point correlation function assuming a linear bias relation with the dark matter field as a function of the scale and as a function of the absolute magnitude. We have estimated the two-point correlation function of galaxies, $\xi_g(r)$, from small scales ($\sim$ 0.17 Mpc/h) up to intermediate scales (10-30 Mpc/h) by dividing the MICE GC run into 10^3 cubes of side 307.2 Mpc/h and we have used the natural estimator to compute $\xi_g(r)$. We have estimated the errors as the standard deviation using the 10^3 cubes of side 307.2 Mpc/h. The agreement between the bias of luminosity threshold samples of the catalogue with the fit of Zehavi et al. 2011 is quite reasonable. And the agreement between the bias of luminosity bin samples with the fit obtained by Zehavi et al. 2011 is very good except for the brightest samples where our catalogue is more clustered that the fit. At fainter scales the agreement with the fit made by Norberg et al. 2001 is almost perfect and also very good with the fit proposed by Tegmark et al. 2004.
- We have shown the distortions in the galaxy clustering when measuring galaxy redshifts due to their peculiar velocities by deriving the two-point correlation function of radial or line-of-sight distance, r_π, and perpendicular distance, r_p. Both known effects, the Fingers-of-God and the Kaiser effects, are also shown depending on the galaxy type. We have also pointed to the effect of galaxy bias in the redshift space distortions.
- We have derived the angular correlation function up to the BAO scale of a sample of galaxies from the catalogue generated by populating the halo catalogue extracted from the snapshot at $z = 0$ of the MICE GC

run. In particular we have taken into account all galaxies in a comoving-distance spherical shell of width 300 Mpc/h, covering the range (2772 − 3072) Mpc/h. We also computed the analytical prediction of the angular correlation function for the MICE simulations and the agreement with the mock is very good.

➤ Finally we have very briefly described some already done and in progress applications of the mock galaxy catalogues such as testing a cluster finder code or being used for different purposes in projects such as DES and PAU.

We finish this section by listing what we think they are the main conclusions of our work:

★ The halo model and the halo occupation distribution model are together a useful approach to deal with the production of mock galaxy catalogues. Although the approach is a statistical method and it is not based on physical principles one can infer and draw conclusions about the processes involved in galaxy formation.

★ The fact that every galaxy population has its own set of HOD parameters has its advantages and drawbacks. On the one hand, a specific galaxy population has its own set of HOD parameters and the differences with another set from another galaxy population may reveal characteristics and differences from both populations. On the other hand, if one uses a HOD set of parameters to build a catalogue that best matches, for instance, the luminosity dependence of clustering for galaxies brighter than a certain value, $M_r < M_{r1}$, and one computes the clustering for galaxies brighter than another brighter value, $M_r < M_{r2}$, with $M_{r2} < M_{r1}$, the clustering will probably not be the same because the best set of HOD parameters for the brighter sample will be different than the one used for the fainter catalogue.

★ Using HOD recipes and algorithms from other works, specially if they assume different cosmologies, and therefore different simulations, cannot be followed word by word. It is not straight forward to implement already developed recipes using the MICE cosmological simulations since we have found several difficulties in the process. In particular we cannot match the luminosity function derived by Blanton et al. 2003 using the relation between halo mass and galaxy luminosity derived from the HOD parameters shown in Zehavi et al. 2005.

★ We have had to introduce different modifications to the *common* HOD model in order to obtain reasonable results in the galaxy clustering as a function of luminosity. Basically we modify the way to assign galaxy luminosities since, as mentioned previously, we did not match the total luminosity function.

★ If we assume that satellite galaxies follow a NFW profile we cannot match the observed clustering at small scales. Therefore we modify the way to place satellite galaxies inside the haloes.

- We have had to introduce more modifications in the *common* HOD model to generate a *unique* catalogue that has the correct clustering for *all* luminosities (and also for colours). We somehow introduce continuity in the discreteness of the HOD parameters, in particular by modeling the variation of M_1 with a continous function. We introduce scatter in the relation between galaxy luminosity and halo mass for the brighter galaxies. This is equivalent to introducing more free HOD parameters in the model. We also include another Gaussian component to describe the colour-magnitude diagram. Therefore we need to apply different recipes depending on the galaxy type to better describe at the same time the colour-magnitude relation and the colour dependence of clustering.

- The generated *unique* galaxy catalogue can be useful for a variety of cosmological applications such as calibrating cluster finders and photometric estimators, studying galaxy clustering at small and large scale and therefore galaxy bias or exploring the effects of redshift space distortions. It is also useful to explore systematics effects and calibrate errors for galaxy surveys, and of course to learn and improve the models and algorithms with which they are built.

- The MICE simulations are ideally suited for studying very large scale structures and specially to study the most massive part of the halo mass function (Crocce et al. 2010). When following the HOD model to generate galaxy catalogues the mass resolution of the input N-body simulation constrain strongly the minimum luminosity of the catalogue, which mainly depends on the minimum halo mass. By using the MICE GC run, which divides by 8 the minimum halo mass of the previous run, we have decreased M_r^{min} from $M_r \sim -21.0$ to $M_r \sim -19.0$. However we find that it is necessary to reach to an even better mass resolution in order to produce galaxy catalogues with a fainter M_r^{min} to be complete at small redshifts when simulating deep surveys. At the same time the catalogues we have generated use haloes formed by only 10 dark-matter particles, and we know that structures formed by such a few number of particles may not be virialized. The interesting point is that even by populating these haloes one is able to generate galaxy catalogues which have quite reasonable results.

- The degree of success of a mock galaxy catalogue depends on the complexity one wants to reproduce. It can be straight forward to fit only one observable but we have found that fitting many of the galaxy properties requires the implementation of complicated recipes.

7.2 Ongoing and future work

There are many interesting improvements and benefits left to be explored regarding the mock galaxy catalogues that we have generated. In future research we intend to improve the mock galaxy catalogues in different ways. We think that one of the most important defficiency of the mock galaxy catalogues is the *almost complete absence* of evolution. We aim to introduce evolution in the algorithm. We have said *almost*

complete absence since we have already used and we are currently using the lightcone output of the MICE GC run to produce mock galaxy catalogues. The lightcone, by itself, contains evolution. The halo mass function evolves as a function of time since we assume that structure formation is hierarchical, which means little things collapse first and big things collapse later. There are more massive haloes at $z = 0$ than at high redshift. Several works point out that the values of the HOD parameters do not change significantly from $z = 0$ to $z \sim 1$. If this is correct we will not have to make big modifications in our algorithm to fit intermedium and high redshift data. We are on the way of validating the galaxy catalogues produced populating the lightcone output. We already know that these catalogues have some issues related to the distribution of colours compared to observed data. This is not surprising since the observations we fit are local data. Several efforts on validating galaxy clustering and also studies of cross-correlations functions between galaxy positions and shear information extracted from the MICE simulations are being developed too. We plan to compare data from medium and high redshift surveys such as DEEP2, VVDS, COMBO-17, NDWFS or more recently COSMOS and CFHTLS. One can use Schechter fits to luminosity functions as a function of redshift from COMBO-17 and DEEP2 (Faber et al. 2007), or from NDWSFS (Brown et al. 2008). Also colour-magnitude diagrams at different redshifts (Brown et al. 2008). Evolution also plays a very important role in the spectra of galaxies. We plan to assign spectral energy distributions, that evolve with redshift, to galaxies in the lightcone too.

As it was mentioned in the previous section we have been producing the catalogues that feed the data reduction pipelines of the PAU project. In the very near future we aim to improve the morphological properties as a condition for the next Data Challenge inside the Data Management working group of the PAU project. We plan to use data from the stuff code[1] and COSMOS or CFHTLS surveys for that purpose.

Since our group has wide access to supercomputing facilities such as the Port d'Informació Científica (PIC) in Barcelona, which is a center of excellence for scientific data processing and storage supporting scientific groups that require massive computing resources for the analysis of large sets of distributed data, we would like to incorporate the code to produce mock galaxy catalogues in a *mass production way*, and from different MICE runs, since the MICE simulations are already stored at PIC. This massive production of catalogues could be useful in particular to produce the required catalogues for future galaxy surveys such as the Euclid mission.

We would also like to explore the different scientific applications that we have covered, without entering into detail, in the previous chapter. We think that studies of galaxy redshift space distortions, galaxy clustering at small and large scales and cross-correlations between galaxy positions and galaxy shear information, or galaxy positions and galaxy magnitudes could be carried out using our mock galaxy catalogues. Some of these scientific applications are being studied in particular inside the PAU project.

Another interesting idea is the possibility of producing mock galaxy catalogues using different methods than the HOD model. Most of the methods that produce

[1]Stuff code is part of the software developed by Emmanuel Bertin (webpage:http://www.astromatic.net/software).

mock galaxy catalogues use as input a halo catalogue, and some of them, in addition, a merger history tree. It would be very interesting to derive the merger history tree of the MICE simulations and be able to produce mock catalogues using a different method than the HOD model. We think it would be useful to implement a semi-analytic model once the merger history tree is obtained, for instance.

We hope to be involved in the development and production of the mock galaxy catalogues that are going to be essential in current and future galaxy surveys.

Chapter 8

Bibliography

Afshordi, N., Loh, Y.-S. & Strauss, M. A., 2004. Cross-correlation of the cosmic microwave background with the 2MASS galaxy survey: Signatures of dark energy, hot gas, and point sources. *PRD*, **69**(8), 083524.

Albrecht, A., Bernstein, G., Cahn, R., Freedman, W. L., Hewitt, J., Hu, W., Huth, J., Kamionkowski, M., Kolb, E. W., Knox, L., Mather, J. C., Staggs, S. & Suntzeff, N. B., 2006. Report of the Dark Energy Task Force. *ArXiv Astrophysics e-prints*.

Allen, S. W., Rapetti, D. A., Schmidt, R. W., Ebeling, H., Morris, R. G. & Fabian, A. C., 2008. Improved constraints on dark energy from Chandra X-ray observations of the largest relaxed galaxy clusters. *MNRAS*, **383**, 879–896.

Allen, S. W., Schmidt, R. W., Ebeling, H., Fabian, A. C. & van Speybroeck, L., 2004. Constraints on dark energy from Chandra observations of the largest relaxed galaxy clusters. *MNRAS*, **353**, 457–467.

Amanullah, R., Lidman, C., Rubin, D., Aldering, G., Astier, P. et al., 2010. Spectra and Hubble Space Telescope Light Curves of Six Type Ia Supernovae at $0.511 < z < 1.12$ and the Union2 Compilation. *ApJ*, **716**, 712–738.

Anderhalden, D. & Diemand, J., 2011. The total mass of dark matter haloes. *MNRAS*, **414**, 3166–3172.

Appel, A. W., 1985. An Efficient Program for Many-Body Simulations. *Journal on Scientific and Statistical Computing*, **6**, 85–103.

Astier, P., Guy, J., Regnault, N., Pain, R., Aubourg, E. et al., 2006. The Supernova Legacy Survey: measurement of Ω_M, Ω_Λ and w from the first year data set. *A&A*, **447**, 31–48.

Bacon, D. J., Refregier, A. R. & Ellis, R. S., 2000. Detection of weak gravitational lensing by large-scale structure. *MNRAS*, **318**, 625–640.

Baldry, I. K., Balogh, M. L., Bower, R., Glazebrook, K. & Nichol, R. C., 2004. Color bimodality: Implications for galaxy evolution. In R. E. Allen, D. V. Nanopoulos, & C. N. Pope, ed., *The New Cosmology: Conference on Strings and Cosmology*, vol. 743 of *American Institute of Physics Conference Series*, 106–119.

Bardeen, J. M., Bond, J. R., Kaiser, N. & Szalay, A. S., 1986. The statistics of peaks of Gaussian random fields. *ApJ*, **304**, 15–61.

Bardeen, J. M., Steinhardt, P. J. & Turner, M. S., 1983. Spontaneous creation of almost scale-free density perturbations in an inflationary universe. *PRD*, **28**, 679–693.

Barnes, J. & Hut, P., 1986. A hierarchical O(N log N) force-calculation algorithm. *Nature*, **324**, 446–449.

Barriga, J. & Gaztañaga, E., 2002. The three-point function in large-scale structure - I. The weakly non-linear regime in N-body simulations. *MNRAS*, **333**, 443–453.

Bartelmann, M. & Schneider, P., 2001. Weak gravitational lensing. *PHyRev.*, **340**, 291–472.

Baugh, C. M., 2006. A primer on hierarchical galaxy formation: the semi-analytical approach. *Reports on Progress in Physics*, **69**, 3101–3156.

Benítez, N., Gaztañaga, E., Miquel, R., Castander, F., Moles, M. et al., 2009. Measuring Baryon Acoustic Oscillations Along the Line of Sight with Photometric Redshifts: The PAU Survey. *ApJ*, **691**, 241–260.

Benson, A. J., 2012. GALACTICUS: A semi-analytic model of galaxy formation. *na*, **17**, 175–197.

Benson, A. J., Baugh, C. M., Cole, S., Frenk, C. S. & Lacey, C. G., 2000. The dependence of velocity and clustering statistics on galaxy properties. *MNRAS*, **316**, 107–119.

Berlind, A. A., Blanton, M. R., Hogg, D. W., Weinberg, D. H., Davé, R., Eisenstein, D. J. & Katz, N., 2005. Interpreting the Relationship between Galaxy Luminosity, Color, and Environment. *ApJ*, **629**, 625–632.

Berlind, A. A. & Weinberg, D. H., 2002. The Halo Occupation Distribution: Toward an Empirical Determination of the Relation between Galaxies and Mass. *ApJ*, **575**, 587–616.

Berlind, A. A., Weinberg, D. H., Benson, A. J., Baugh, C. M., Cole, S., Davé, R., Frenk, C. S., Jenkins, A., Katz, N. & Lacey, C. G., 2003. The Halo Occupation Distribution and the Physics of Galaxy Formation. *ApJ*, **593**, 1–25.

Bernardeau, F., Colombi, S., Gaztañaga, E. & Scoccimarro, R., 2002. Large-scale structure of the Universe and cosmological perturbation theory. *PHyRev.*, **367**, 1–248.

Blanton, M. R., Eisenstein, D., Hogg, D. W., Schlegel, D. J. & Brinkmann, J., 2005. Relationship between Environment and the Broadband Optical Properties of Galaxies in the Sloan Digital Sky Survey. *ApJ*, **629**, 143–157.

Blanton, M. R., Hogg, D. W., Bahcall, N. A., Brinkmann, J., Britton, M. et al., 2003. The Galaxy Luminosity Function and Luminosity Density at Redshift z = 0.1. *ApJ*, **592**, 819–838.

Blumenthal, G. R., Faber, S. M., Primack, J. R. & Rees, M. J., 1984. Formation of galaxies and large-scale structure with cold dark matter. *Nature*, **311**, 517–525.

Bond, J. R., Cole, S., Efstathiou, G. & Kaiser, N., 1991. Excursion set mass functions for hierarchical Gaussian fluctuations. *ApJ*, **379**, 440–460.

Bond, J. R. & Myers, S. T., 1996. The Peak-Patch Picture of Cosmic Catalogs. I. Algorithms. *ApJS*, **103**, 1.

Boughn, S. & Crittenden, R., 2004. A correlation between the cosmic microwave background and large-scale structure in the Universe. *Nature*, **427**, 45–47.

Bower, R. G., Benson, A. J., Malbon, R., Helly, J. C., Frenk, C. S., Baugh, C. M., Cole, S. & Lacey, C. G., 2006. Breaking the hierarchy of galaxy formation. *MNRAS*, **370**, 645–655.

Boylan-Kolchin, M., Springel, V., White, S. D. M. & Jenkins, A., 2010. There's no place like home? Statistics of Milky Way-mass dark matter haloes. *MNRAS*, **406**, 896–912.

Boylan-Kolchin, M., Springel, V., White, S. D. M., Jenkins, A. & Lemson, G., 2009. Resolving cosmic structure formation with the Millennium-II Simulation. *MNRAS*, **398**, 1150–1164.

Brown, M. J. I., Zheng, Z., White, M., Dey, A., Jannuzi, B. T., Benson, A. J., Brand, K., Brodwin, M. & Croton, D. J., 2008. Red Galaxy Growth and the Halo Occupation Distribution. *ApJ*, **682**, 937–963.

Bruzual, G., 1981. *Spectral evolution of galaxies*.

Bryan, G. L. & Norman, M. L., 1998. Statistical Properties of X-Ray Clusters: Analytic and Numerical Comparisons. *ApJ*, **495**, 80.

Bullock, J. S., Dekel, A., Kolatt, T. S., Kravtsov, A. V., Klypin, A. A., Porciani, C. & Primack, J. R., 2001. A Universal Angular Momentum Profile for Galactic Halos. *ApJ*, **555**, 240–257.

Bullock, J. S., Wechsler, R. H. & Somerville, R. S., 2002. Galaxy halo occupation at high redshift. *MNRAS*, **329**, 246–256.

Cabré, A. & Gaztañaga, E., 2009a. Clustering of luminous red galaxies - I. Large-scale redshift-space distortions. *MNRAS*, **393**, 1183–1208.

Cabré, A. & Gaztañaga, E., 2009b. Clustering of luminous red galaxies - II. Small-scale redshift-space distortions. *MNRAS*, **396**, 1119–1131.

Cabré, A. & Gaztañaga, E., 2011. Have baryonic acoustic oscillations in the galaxy distribution really been measured? *MNRAS*, **412**, L98–L102.

Cacciato, M., van den Bosch, F. C., More, S., Li, R., Mo, H. J. & Yang, X., 2009. Galaxy clustering and galaxy-galaxy lensing: a promising union to constrain cosmological parameters. *MNRAS*, **394**, 929–946.

Cai, Y.-C., Angulo, R. E., Baugh, C. M., Cole, S., Frenk, C. S. & Jenkins, A., 2009. Mock galaxy redshift catalogues from simulations: implications for Pan-STARRS1. *MNRAS*, **395**, 1185–1203.

Coil, A. L., Davis, M. & Szapudi, I., 2001. Galaxy Correlation Statistics of Mock Catalogs for the DEEP2 Survey. *PASP*, **113**, 1312–1325.

Cole, S., Hatton, S., Weinberg, D. H. & Frenk, C. S., 1998. Mock 2dF and SDSS galaxy redshift surveys. *MNRAS*, **300**, 945–966.

Cole, S. & Kaiser, N., 1989. Biased clustering in the cold dark matter cosmogony. *MNRAS*, **237**, 1127–1146.

Cole, S., Lacey, C. G., Baugh, C. M. & Frenk, C. S., 2000. Hierarchical galaxy formation. *MNRAS*, **319**, 168–204.

Cole, S., Percival, W. J., Peacock, J. A., Norberg, P., Baugh, C. M. et al., 2005. The 2dF Galaxy Redshift Survey: power-spectrum analysis of the final data set and cosmological implications. *MNRAS*, **362**, 505–534.

Coles, P. & Jones, B., 1991. A lognormal model for the cosmological mass distribution. *MNRAS*, **248**, 1–13.

Conroy, C., Wechsler, R. H. & Kravtsov, A. V., 2006. Modeling Luminosity-dependent Galaxy Clustering through Cosmic Time. *ApJ*, **647**, 201–214.

Cooray, A., 2005. A divided Universe: red and blue galaxies and their preferred environments. *MNRAS*, **363**, 337–352.

Cooray, A., 2006. Halo model at its best: constraints on conditional luminosity functions from measured galaxy statistics. *MNRAS*, **365**, 842–866.

Cooray, A. & Milosavljević, M., 2005. What is L_*? Anatomy of the Galaxy Luminosity Function. *ApJ*, **627**, L89–L92.

Cooray, A. & Sheth, R., 2002. Halo models of large scale structure. *PHyRev.*, **372**, 1–129.

Crocce, M., Fosalba, P., Castander, F. J. & Gaztañaga, E., 2010. Simulating the Universe with MICE: the abundance of massive clusters. *MNRAS*, **403**, 1353–1367.

Crocce, M., Pueblas, S. & Scoccimarro, R., 2006. Transients from initial conditions in cosmological simulations. *MNRAS*, **373**, 369–381.

Croton, D. J., Springel, V., White, S. D. M., De Lucia, G., Frenk, C. S., Gao, L., Jenkins, A., Kauffmann, G., Navarro, J. F. & Yoshida, N., 2006. The many lives of active galactic nuclei: cooling flows, black holes and the luminosities and colours of galaxies. *MNRAS*, **365**, 11–28.

Dalcanton, J. J., Spergel, D. N. & Summers, F. J., 1997. The Formation of Disk Galaxies. *ApJ*, **482**, 659.

Davis, M., Efstathiou, G., Frenk, C. S. & White, S. D. M., 1985a. The evolution of large-scale structure in a universe dominated by cold dark matter. *ApJ*, **292**, 371–394.

Davis, M., Efstathiou, G., Frenk, C. S. & White, S. D. M., 1985b. The evolution of large-scale structure in a universe dominated by cold dark matter. *ApJ*, **292**, 371–394.

Davis, M. & Peebles, P. J. E., 1983. A survey of galaxy redshifts. V - The two-point position and velocity correlations. *ApJ*, **267**, 465–482.

De Lucia, G. & Blaizot, J., 2007. The hierarchical formation of the brightest cluster galaxies. *MNRAS*, **375**, 2–14.

Dehnen, W. & Read, J., 2011. N-body simulations of gravitational dynamics. *The European Physical Journal Plus*, **126**, 1–28. ISSN 2190-5444. URL `http://dx.doi.org/10.1140/epjp/i2011-11055-3`. 10.1140/epjp/i2011-11055-3.

Dekel, A. & Silk, J., 1986. The origin of dwarf galaxies, cold dark matter, and biased galaxy formation. *ApJ*, **303**, 39–55.

Dodelson, S., 2003. *Modern cosmology*.

Efstathiou, G., Davis, M., White, S. D. M. & Frenk, C. S., 1985. Numerical techniques for large cosmological N-body simulations. *ApJS*, **57**, 241–260.

Efstathiou, G. & Eastwood, J. W., 1981. On the clustering of particles in an expanding universe. *MNRAS*, **194**, 503–525.

Eisenstein, D. J., Zehavi, I., Hogg, D. W., Scoccimarro, R., Blanton, M. R. et al., 2005. Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies. *ApJ*, **633**, 560–574.

Eriksen, H. K., Lilje, P. B., Banday, A. J. & Górski, K. M., 2004. Estimating N-Point Correlation Functions from Pixelized Sky Maps. *ApJS*, **151**, 1–11.

Evrard, A. E., MacFarland, T. J., Couchman, H. M. P., Colberg, J. M., Yoshida, N., White, S. D. M., Jenkins, A., Frenk, C. S., Pearce, F. R., Peacock, J. A. & Thomas, P. A., 2002. Galaxy Clusters in Hubble Volume Simulations: Cosmological Constraints from Sky Survey Populations. *ApJ*, **573**, 7–36.

Evrard, A. E., Summers, F. J. & Davis, M., 1994. Two-fluid simulations of galaxy formation. *ApJ*, **422**, 11–36.

Faber, S. M., Willmer, C. N. A., Wolf, C., Koo, D. C., Weiner, B. J. et al., 2007. Galaxy Luminosity Functions to $z \sim 1$ from DEEP2 and COMBO-17: Implications for Red Galaxy Formation. *ApJ*, **665**, 265–294.

Fall, S. M. & Efstathiou, G., 1980. Formation and rotation of disc galaxies with haloes. *MNRAS*, **193**, 189–206.

Farrens, S., Abdalla, F. B., Cypriano, E. S., Sabiu, C. & Blake, C., 2011. Friends-of-friends groups and clusters in the 2SLAQ catalogue. *MNRAS*, **417**, 1402–1416.

Font-Ribera, A., McDonald, P. & Miralda-Escudé, J., 2012. Generating mock data sets for large-scale Lyman-α forest correlation measurements. *jcap*, **1**, 1.

Fosalba, P. & Gaztañaga, E., 2004. Measurement of the gravitational potential evolution from the cross-correlation between WMAP and the APM Galaxy Survey. *MNRAS*, **350**, L37–L41.

Fosalba, P., Gaztañaga, E., Castander, F. J. & Manera, M., 2008. The onion universe: all sky lightcone simulations in spherical shells. *MNRAS*, **391**, 435–446.

Freedman, W. L., Madore, B. F., Gibson, B. K., Ferrarese, L., Kelson, D. D., Sakai, S., Mould, J. R., Kennicutt, Jr., R. C., Ford, H. C., Graham, J. A., Huchra, J. P., Hughes, S. M. G., Illingworth, G. D., Macri, L. M. & Stetson, P. B., 2001. Final Results from the Hubble Space Telescope Key Project to Measure the Hubble Constant. *ApJ*, **553**, 47–72.

Frenk, C. S., Colberg, J. M., Couchman, H. M. P., Efstathiou, G., Evrard, A. E., Jenkins, A., MacFarland, T. J., Moore, B., Peacock, J. A., Pearce, F. R., Thomas, P. A., White, S. D. M. & Yoshida, N., 2000. Public Release of N-body simulation and related data by the Virgo consortium. *ArXiv Astrophysics e-prints*.

Frenk, C. S., Evrard, A. E., White, S. D. M. & Summers, F. J., 1996. Galaxy Dynamics in Clusters. *ApJ*, **472**, 460.

Gaztañaga, E., Cabré, A., Castander, F., Crocce, M. & Fosalba, P., 2009a. Clustering of luminous red galaxies - III. Baryon acoustic peak in the three-point correlation. *MNRAS*, **399**, 801–811.

Gaztañaga, E., Cabré, A. & Hui, L., 2009b. Clustering of luminous red galaxies - IV. Baryon acoustic peak in the line-of-sight direction and a direct measurement of H(z). *MNRAS*, **399**, 1663–1680.

Gaztanaga, E. & Cabre, A., 2008. The anisotropic redshift space galaxy correlation function: detection on the BAO Ring. *ArXiv e-prints*.

Górski, K. M., Hivon, E., Banday, A. J., Wandelt, B. D., Hansen, F. K., Reinecke, M. & Bartelmann, M., 2005. HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere. *ApJ*, **622**, 759–771.

Gunn, J. E. & Gott, III, J. R., 1972. On the Infall of Matter Into Clusters of Galaxies and Some Effects on Their Evolution. *ApJ*, **176**, 1–+.

Guo, Q., White, S., Boylan-Kolchin, M., De Lucia, G., Kauffmann, G., Lemson, G., Li, C., Springel, V. & Weinmann, S., 2011. From dwarf spheroidals to cD galaxies: simulating the galaxy population in a ΛCDM cosmology. *MNRAS*, **413**, 101–131.

Guo, Q., White, S., Li, C. & Boylan-Kolchin, M., 2010. How do galaxies populate dark matter haloes? *MNRAS*, **404**, 1111–1120.

Guth, A. H., 1981. Inflationary universe: A possible solution to the horizon and flatness problems. *PRD*, **23**, 347–356.

Guth, A. H. & Pi, S.-Y., 1982. Fluctuations in the new inflationary universe. *Physical Review Letters*, **49**, 1110–1113.

Hamilton, A. J. S., 1993. Toward Better Ways to Measure the Galaxy Correlation Function. *ApJ*, **417**, 19.

Hatton, S., Devriendt, J. E. G., Ninin, S., Bouchet, F. R., Guiderdoni, B. & Vibert, D., 2003. GALICS- I. A hybrid N-body/semi-analytic model of hierarchical galaxy formation. *MNRAS*, **343**, 75–106.

Heitmann, K., Lukić, Z., Habib, S. & Ricker, P. M., 2006. Capturing Halos at High Redshifts. *ApJ*, **642**, L85–L88.

Hillebrandt, W. & Niemeyer, J. C., 2000. Type IA Supernova Explosion Models. *ARA&A*, **38**, 191–230.

Hockney, R., Goel, S. & Eastwood, J., 1974. Quiet high-resolution computer models of a plasma. *Journal of Computational Physics*, **14**, 148–158.

Hockney, R. W. & Eastwood, J. W., 1988. *Computer simulation using particles*.

Hoffman, Y. & Shaham, J., 1985. Local density maxima - Progenitors of structure. *ApJ*, **297**, 16–22.

Hohl, F. & Hockney, R. W., 1969. A Computer Model of Disks of Stars. *Journal of Computational Physics*, **4**, 306–+.

Hubble, E., 1929. A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae. *Proceedings of the National Academy of Science*, **15**, 168–173.

Huff, E., Schulz, A. E., White, M., Schlegel, D. J. & Warren, M. S., 2007. Simulations of baryon oscillations. *Astroparticle Physics*, **26**, 351–366.

Jaffe, A. H., Ade, P. A., Balbi, A., Bock, J. J., Bond, J. R. et al., 2001. Cosmology from MAXIMA-1, BOOMERANG, and COBE DMR Cosmic Microwave Background Observations. *Physical Review Letters*, **86**, 3475–3479.

Jenkins, A., Frenk, C. S., White, S. D. M., Colberg, J. M., Cole, S., Evrard, A. E., Couchman, H. M. P. & Yoshida, N., 2001. The mass function of dark matter haloes. *MNRAS*, **321**, 372–384.

Jing, Y. P., 2000. The Density Profile of Equilibrium and Nonequilibrium Dark Matter Halos. *ApJ*, **535**, 30–36.

Jing, Y. P., Mo, H. J. & Boerner, G., 1998. Spatial Correlation Function and Pairwise Velocity Dispersion of Galaxies: Cold Dark Matter Models versus the Las Campanas Survey. *ApJ*, **494**, 1.

Jouvel, S., Kneib, J.-P., Ilbert, O., Bernstein, G., Arnouts, S., Dahlen, T., Ealet, A., Milliard, B., Aussel, H., Capak, P., Koekemoer, A., Le Brun, V., McCracken, H., Salvato, M. & Scoville, N., 2009. Designing future dark energy space missions. I. Building realistic galaxy spectro-photometric catalogs and their first applications. *A&A*, **504**, 359–371.

Kaiser, N., 1984. On the spatial correlations of Abell clusters. *ApJ*, **284**, L9–L12.

Kaiser, N., 1987. Clustering in real space and in redshift space. *MNRAS*, **227**, 1–21.

Kaiser, N., Wilson, G. & Luppino, G. A., 2000. Large-Scale Cosmic Shear Measurements. *ArXiv Astrophysics e-prints*.

Katz, N., Hernquist, L. & Weinberg, D. H., 1992. Galaxies and gas in a cold dark matter universe. *ApJ*, **399**, L109–L112.

Kauffmann, G., Colberg, J. M., Diaferio, A. & White, S. D. M., 1999. Clustering of galaxies in a hierarchical universe - I. Methods and results at z=0. *MNRAS*, **303**, 188–206.

Kauffmann, G., Nusser, A. & Steinmetz, M., 1997. Galaxy formation and large-scale bias. *MNRAS*, **286**, 795–811.

Kauffmann, G., White, S. D. M. & Guiderdoni, B., 1993. The Formation and Evolution of Galaxies Within Merging Dark Matter Haloes. *MNRAS*, **264**, 201.

Kennicutt, Jr., R. C., 1989. The star formation law in galactic disks. *ApJ*, **344**, 685–703.

Kerscher, M., Szapudi, I. & Szalay, A. S., 2000. A Comparison of Estimators for the Two-Point Correlation Function. *ApJ*, **535**, L13–L16.

Kim, J., Park, C., Gott, III, J. R. & Dubinski, J., 2009. The Horizon Run N-Body Simulation: Baryon Acoustic Oscillations and Topology of Large-scale Structure of the Universe. *ApJ*, **701**, 1547–1559.

Klypin, A., Gottlöber, S., Kravtsov, A. V. & Khokhlov, A. M., 1999. Galaxies in N-Body Simulations: Overcoming the Overmerging Problem. *ApJ*, **516**, 530–551.

Knop, R. A., Aldering, G., Amanullah, R., Astier, P., Blanc, G. et al., 2003. New Constraints on Ω_M, $\Omega_?$, and w from an Independent Set of 11 High-Redshift Supernovae Observed with the Hubble Space Telescope. *ApJ*, **598**, 102–137.

Krauss, L. M. & Chaboyer, B., 2003. Age Estimates of Globular Clusters in the Milky Way: Constraints on Cosmology. *Science*, **299**, 65–70.

Kravtsov, A. V., Berlind, A. A., Wechsler, R. H., Klypin, A. A., Gottlöber, S., Allgood, B. & Primack, J. R., 2004. The Dark Side of the Halo Occupation Distribution. *ApJ*, **609**, 35–49.

Lacey, C. & Cole, S., 1994. Merger Rates in Hierarchical Models of Galaxy Formation - Part Two - Comparison with N-Body Simulations. *MNRAS*, **271**, 676–+.

Lahav, O. & Liddle, A. R., 2010. The Cosmological Parameters. *ArXiv Astrophysics e-prints*.

Landy, S. D. & Szalay, A. S., 1993. Bias and variance of angular correlation functions. *ApJ*, **412**, 64–71.

Leauthaud, A., Tinker, J., Behroozi, P. S., Busha, M. T. & Wechsler, R. H., 2011. A Theoretical Framework for Combining Techniques that Probe the Link Between Galaxies and Dark Matter. *ApJ*, **738**, 45.

Lee, J. & Shandarin, S. F., 1998. Large-Scale Biasing and the Primordial Gravitational Potential. *ApJ*, **505**, L75–L78.

Lemson, G. & Kauffmann, G., 1999. Environmental influences on dark matter haloes and consequences for the galaxies within them. *MNRAS*, **302**, 111–117.

Li, C., Jing, Y. P., Kauffmann, G., Börner, G., Kang, X. & Wang, L., 2007. Luminosity dependence of the spatial and velocity distributions of galaxies: semi-analytic models versus the Sloan Digital Sky Survey. *MNRAS*, **376**, 984–996.

Lin, H., Kirshner, R. P., Shectman, S. A., Landy, S. D., Oemler, A., Tucker, D. L. & Schechter, P. L., 1996. The Power Spectrum of Galaxy Clustering in the Las Campanas Redshift Survey. *ApJ*, **471**, 617.

Livio, M., 2011. Lost in translation: Mystery of the missing text solved. *Nature*, **479**, 171–173.

Longair, M. S., ed., 1998. *Galaxy formation*.

Lukić, Z., Heitmann, K., Habib, S., Bashinsky, S. & Ricker, P. M., 2007. The Halo Mass Function: High-Redshift Evolution and Universality. *ApJ*, **671**, 1160–1181.

Lukić, Z., Reed, D., Habib, S. & Heitmann, K., 2009. The Structure of Halos: Implications for Group and Cluster Cosmology. *ApJ*, **692**, 217–228.

Ma, C.-P. & Fry, J. N., 2000. Deriving the Nonlinear Cosmological Power Spectrum and Bispectrum from Analytic Dark Matter Halo Profiles and Mass Functions. *ApJ*, **543**, 503–513.

Magliocchetti, M. & Porciani, C., 2003. The halo distribution of 2dF galaxies. *MNRAS*, **346**, 186–198.

Mandelbaum, R., Seljak, U., Kauffmann, G., Hirata, C. M. & Brinkmann, J., 2006. Galaxy halo masses and satellite fractions from galaxy-galaxy lensing in the Sloan Digital Sky Survey: stellar mass, luminosity, morphology and environment dependencies. *MNRAS*, **368**, 715–731.

Mandelbaum, R., Tasitsiomi, A., Seljak, U., Kravtsov, A. V. & Wechsler, R. H., 2005. Galaxy-galaxy lensing: dissipationless simulations versus the halo model. *MNRAS*, **362**, 1451–1462.

Manera, M., Scoccimarro, R., Percival, W. J., Samushia, L., McBride, C. K. et al., 2012. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: a large sample of mock galaxy catalogues. *ArXiv e-prints*.

Manera, M., Sheth, R. K. & Scoccimarro, R., 2010. Large-scale bias and the inaccuracy of the peak-background split. *MNRAS*, **402**, 589–602.

Martínez, V. J. & Saar, E., 2002. *Statistics of the Galaxy Distribution*. Chapman &amp.

Martino, M. C. & Sheth, R. K., 2009. On the equivalence between the effective cosmology and excursion set treatments of environment. *MNRAS*, **394**, 2109–2112.

McClelland, J. & Silk, J., 1977. The correlation function for density perturbations in an expanding universe. II - Nonlinear theory. *ApJ*, **217**, 331–352.

Miknaitis, G., Pignata, G., Rest, A., Wood-Vasey, W. M., Blondin, S. et al., 2007. The ESSENCE Supernova Survey: Survey Optimization, Observations, and Supernova Photometry. *ApJ*, **666**, 674–693.

Miller, R. H. & Prendergast, K. H., 1968. Stellar Dynamics in a Discrete Phase Space. *ApJ*, **151**, 699–+.

Miller, R. H., Prendergast, K. H. & Quirk, W. J., 1970. Numerical Experiments on Spiral Structure. *ApJ*, **161**, 903–+.

Mo, H. J., Mao, S. & White, S. D. M., 1998. The formation of galactic discs. *MNRAS*, **295**, 319–336.

Mo, H. J. & White, S. D. M., 1996. An analytic model for the spatial clustering of dark matter haloes. *MNRAS*, **282**, 347–361.

Moore, B., Quinn, T., Governato, F., Stadel, J. & Lake, G., 1999. Cold collapse and the core catastrophe. *MNRAS*, **310**, 1147–1152.

More, S., van den Bosch, F. C., Cacciato, M., Mo, H. J., Yang, X. & Li, R., 2009. Satellite kinematics - II. The halo mass-luminosity relation of central galaxies in SDSS. *MNRAS*, **392**, 801–816.

Munshi, D., Valageas, P., van Waerbeke, L. & Heavens, A., 2008. Cosmology with weak lensing surveys. *PHyRev.*, **462**, 67–121.

Navarro, J. F., Frenk, C. S. & White, S. D. M., 1996. The Structure of Cold Dark Matter Halos. *ApJ*, **462**, 563.

Navarro, J. F., Frenk, C. S. & White, S. D. M., 1997. A Universal Density Profile from Hierarchical Clustering. *ApJ*, **490**, 493.

Neyman, J. & Scott, E. L., 1952. A Theory of the Spatial Distribution of Galaxies. *ApJ*, **116**, 144–+.

Norberg, P., Baugh, C. M., Hawkins, E., Maddox, S., Peacock, J. A. et al., 2001. The 2dF Galaxy Redshift Survey: luminosity dependence of galaxy clustering. *MNRAS*, **328**, 64–70.

Nuza, S. E., Sanchez, A. G., Prada, F., Klypin, A., Schlegel, D. J. et al., 2012. The clustering of galaxies at z˜0.5 in the SDSS-III Data Release 9 BOSS-CMASS sample: a test for the LCDM cosmology. *ArXiv e-prints*.

Peacock, J. A., 1999. *Cosmological Physics*.

Peacock, J. A., Cole, S., Norberg, P., Baugh, C. M., Bland-Hawthorn, J. et al., 2001. A measurement of the cosmological mass density from clustering in the 2dF Galaxy Redshift Survey. *Nature*, **410**, 169–173.

Peacock, J. A. & Heavens, A. F., 1985. The statistics of maxima in primordial density perturbations. *MNRAS*, **217**, 805–820.

Peacock, J. A. & Heavens, A. F., 1990. Alternatives to the Press-Schechter cosmological mass function. *MNRAS*, **243**, 133–143.

Peacock, J. A. & Smith, R. E., 2000. Halo occupation numbers and galaxy bias. *MNRAS*, **318**, 1144–1156.

Peebles, P. J. E., 1974. The Nature of the Distribution of Galaxies. *A&A*, **32**, 197.

Peebles, P. J. E., 1980. *The large-scale structure of the universe*.

Peebles, P. J. E., 1982. Large-scale background temperature and mass fluctuations due to scale-invariant primeval perturbations. *ApJ*, **263**, L1–L5.

Peebles, P. J. E., 1993. Book Review: Principles of physical cosmology / Princeton U Press, 1993. *Physics Today*, **46**, 87.

Peebles, P. J. E. & Hauser, M. G., 1974. Statistical Analysis of Catalogs of Extragalactic Objects. III. The Shane-Wirtanen and Zwicky Catalogs. *ApJS*, **28**, 19.

Perlmutter, S., Aldering, G., Goldhaber, G., Knop, R. A., Nugent, P. et al., 1999. Measurements of Omega and Lambda from 42 High-Redshift Supernovae. *ApJ*, **517**, 565–586.

Press, W. H. & Schechter, P., 1974. Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation. *ApJ*, **187**, 425–438.

Pryke, C., Halverson, N. W., Leitch, E. M., Kovac, J., Carlstrom, J. E., Holzapfel, W. L. & Dragovan, M., 2002. Cosmological Parameter Extraction from the First Season of Observations with the Degree Angular Scale Interferometer. *ApJ*, **568**, 46–51.

Reed, D., Gardner, J., Quinn, T., Stadel, J., Fardal, M., Lake, G. & Governato, F., 2003. Evolution of the mass function of dark matter haloes. *MNRAS*, **346**, 565–572.

Reed, D. S., Bower, R., Frenk, C. S., Jenkins, A. & Theuns, T., 2007. The halo mass function from the dark ages through the present day. *MNRAS*, **374**, 2–15.

Riess, A. G., Filippenko, A. V., Challis, P., Clocchiatti, A., Diercks, A. et al., 1998. Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant. *AJ*, **116**, 1009–1038.

Riess, A. G., Nugent, P. E., Gilliland, R. L., Schmidt, B. P., Tonry, J. et al., 2001. The Farthest Known Supernova: Support for an Accelerating Universe and a Glimpse of the Epoch of Deceleration. *ApJ*, **560**, 49–71.

Riess, A. G., Strolger, L.-G., Casertano, S., Ferguson, H. C., Mobasher, B. et al., 2007. New Hubble Space Telescope Discoveries of Type Ia Supernovae at z $\geq$ 1: Narrowing Constraints on the Early Behavior of Dark Energy. *ApJ*, **659**, 98–121.

Riess, A. G., Strolger, L.-G., Tonry, J., Casertano, S., Ferguson, H. C. et al., 2004. Type Ia Supernova Discoveries at z $\geq$ 1 from the Hubble Space Telescope: Evidence for Past Deceleration and Constraints on Dark Energy Evolution. *ApJ*, **607**, 665–687.

Robertson, B. E., Kravtsov, A. V., Tinker, J. & Zentner, A. R., 2009. Collapse Barriers and Halo Abundance: Testing the Excursion Set Ansatz. *ApJ*, **696**, 636–652.

Roukema, B. F., Quinn, P. J., Peterson, B. A. & Rocca-Volmerange, B., 1997. Merging history trees of dark matter haloes - A tool for exploring galaxy formation models. *MNRAS*, **292**, 835.

Rozo, E., Rykoff, E. S., Koester, B. P., McKay, T., Hao, J., Evrard, A., Wechsler, R. H., Hansen, S., Sheldon, E., Johnston, D., Becker, M., Annis, J., Bleem, L. & Scranton, R., 2009. Improvement of the Richness Estimates of maxBCG Clusters. *ApJ*, **703**, 601–613.

Scherrer, R. J. & Bertschinger, E., 1991. Statistics of primordial density perturbations from discrete seed masses. *ApJ*, **381**, 349–360.

Schneider, P., 2006. Part 1: Introduction to gravitational lensing and cosmology. In G. Meylan, P. Jetzer, P. North, P. Schneider, C. S. Kochanek, & J. Wambsganss, ed., *Saas-Fee Advanced Course 33: Gravitational Lensing: Strong, Weak and Micro*, 1–89.

Schneider, P., ed., 2006. *Extragalactic Astronomy and Cosmology*. Springer Berlin / Heidelberg.

Scoccimarro, R., 1997. Cosmological Perturbations: Entering the Nonlinear Regime. *ApJ*, **487**, 1–+.

Scoccimarro, R., 1998. Transients from initial conditions: a perturbative analysis. *MNRAS*, **299**, 1097–1118.

Scoccimarro, R. & Sheth, R. K., 2002. PTHALOS: a fast method for generating mock galaxy distributions. *MNRAS*, **329**, 629–640.

Scoccimarro, R., Sheth, R. K., Hui, L. & Jain, B., 2001. How Many Galaxies Fit in a Halo? Constraints on Galaxy Formation Efficiency from Spatial Clustering. *ApJ*, **546**, 20–34.

Scranton, R., Connolly, A. J., Nichol, R. C., Stebbins, A., Szapudi, I. et al., 2003. Physical Evidence for Dark Energy. *ArXiv Astrophysics e-prints*.

Seljak, U., 2000. Analytic model for galaxy and dark matter clustering. *MNRAS*, **318**, 203–213.

Sheth, R. K. & Diaferio, A., 2001. Peculiar velocities of galaxies and clusters. *MNRAS*, **322**, 901–917.

Sheth, R. K., Diaferio, A., Hui, L. & Scoccimarro, R., 2001a. On the streaming motions of haloes and galaxies. *MNRAS*, **326**, 463–472.

Sheth, R. K., Hui, L., Diaferio, A. & Scoccimarro, R., 2001b. Linear and non-linear contributions to pairwise peculiar velocities. *MNRAS*, **325**, 1288–1302.

Sheth, R. K., Mo, H. J. & Tormen, G., 2001c. Ellipsoidal collapse and an improved model for the number and spatial distribution of dark matter haloes. *MNRAS*, **323**, 1–12.

Sheth, R. K. & Saslaw, W. C., 1994. Synthesizing the observed distribution of galaxies. *ApJ*, **437**, 35–55.

Sheth, R. K. & Tormen, G., 1999. Large-scale bias and the peak background split. *MNRAS*, **308**, 119–126.

Sheth, R. K. & Tormen, G., 2004. On the environmental dependence of halo formation. *MNRAS*, **350**, 1385–1390.

Skibba, R., Sheth, R. K., Connolly, A. J. & Scranton, R., 2006. The luminosity-weighted or 'marked' correlation function. *MNRAS*, **369**, 68–76.

Skibba, R. A. & Sheth, R. K., 2009. A halo model of galaxy colours and clustering in the Sloan Digital Sky Survey. *MNRAS*, **392**, 1080–1091.

Skibba, R. A., van den Bosch, F. C., Yang, X., More, S., Mo, H. & Fontanot, F., 2011. Are brightest halo galaxies central galaxies? *MNRAS*, **410**, 417–431.

Somerville, R. S., Hopkins, P. F., Cox, T. J., Robertson, B. E. & Hernquist, L., 2008. A semi-analytic model for the co-evolution of galaxies, black holes and active galactic nuclei. *MNRAS*, **391**, 481–506.

Soneira, R. M. & Peebles, P. J. E., 1978. A computer model universe - Simulation of the nature of the galaxy distribution in the Lick catalog. *AJ*, **83**, 845–849.

Springel, V., 2005. The cosmological simulation code GADGET-2. *MNRAS*, **364**, 1105–1134.

Springel, V., 2010. Smoothed Particle Hydrodynamics in Astrophysics. *ARA&A*, **48**, 391–430.

Springel, V. & Hernquist, L., 2003. Cosmological smoothed particle hydrodynamics simulations: a hybrid multiphase model for star formation. *MNRAS*, **339**, 289–311.

Springel, V., White, S. D. M., Jenkins, A., Frenk, C. S., Yoshida, N., Gao, L., Navarro, J., Thacker, R., Croton, D., Helly, J., Peacock, J. A., Cole, S., Thomas, P., Couchman, H., Evrard, A., Colberg, J. & Pearce, F., 2005. Simulations of the formation, evolution and clustering of galaxies and quasars. *Nature*, **435**, 629–636.

Starobinsky, A. A., 1982. Dynamics of phase transition in the new inflationary universe scenario and generation of perturbations. *Physics Letters B*, **117**, 175–178.

Tasitsiomi, A., Kravtsov, A. V., Wechsler, R. H. & Primack, J. R., 2004. Modeling Galaxy-Mass Correlations in Dissipationless Simulations. *ApJ*, **614**, 533–546.

Tegmark, M., Blanton, M. R., Strauss, M. A., Hoyle, F., Schlegel, D. et al., 2004. The Three-Dimensional Power Spectrum of Galaxies from the Sloan Digital Sky Survey. *ApJ*, **606**, 702–740.

Tegmark, M., Eisenstein, D. J., Strauss, M. A., Weinberg, D. H., Blanton, M. R. et al., 2006. Cosmological constraints from the SDSS luminous red galaxies. *PRD*, **74**(12), 123507.

Teyssier, R., Pires, S., Prunet, S., Aubert, D., Pichon, C., Amara, A., Benabed, K., Colombi, S., Refregier, A. & Starck, J.-L., 2009. Full-sky weak-lensing simulation with 70 billion particles. *A&A*, **497**, 335–341.

Thomas, P. A., Colberg, J. M., Couchman, H. M. P., Efstathiou, G. P., Frenk, C. S., Jenkins, A. R., Nelson, A. H., Hutchings, R. M., Peacock, J. A., Pearce, F. R. & White, S. D. M., 1998. The structure of galaxy clusters in various cosmologies. *MNRAS*, **296**, 1061–1071.

Tinker, J., Kravtsov, A. V., Klypin, A., Abazajian, K., Warren, M., Yepes, G., Gottlöber, S. & Holz, D. E., 2008. Toward a Halo Mass Function for Precision Cosmology: The Limits of Universality. *ApJ*, **688**, 709–728.

Tinker, J. L., 2010. There Is No Void Problem. In L. Verdes-Montenegro, A. Del Olmo, & J. Sulentic, ed., *Galaxies in Isolation: Exploring Nature Versus Nurture*, vol. 421 of *Astronomical Society of the Pacific Conference Series*, 95.

Tinker, J. L., Weinberg, D. H., Zheng, Z. & Zehavi, I., 2005. On the Mass-to-Light Ratio of Large-Scale Structure. *ApJ*, **631**, 41–58.

Trujillo-Gomez, S., Klypin, A., Primack, J. & Romanowsky, A. J., 2011a. Galaxies in ΛCDM with Halo Abundance Matching: Luminosity-Velocity Relation, Baryonic Mass-Velocity Relation, Velocity Function, and Clustering. *ApJ*, **742**, 16.

Trujillo-Gomez, S., Klypin, A., Primack, J. & Romanowsky, A. J., 2011b. Galaxies in ΛCDM with Halo Abundance Matching: Luminosity-Velocity Relation, Baryonic Mass-Velocity Relation, Velocity Function, and Clustering. *ApJ*, **742**, 16.

Vale, A. & Ostriker, J. P., 2004. Linking halo mass to galaxy luminosity. *MNRAS*, **353**, 189–200.

Vale, A. & Ostriker, J. P., 2006. The non-parametric model for linking galaxy luminosity with halo/subhalo mass. *MNRAS*, **371**, 1173–1187.

van den Bosch, F. C., Yang, X. & Mo, H. J., 2003. Linking early- and late-type galaxies to their dark matter haloes. *MNRAS*, **340**, 771–792.

van den Bosch, F. C., Yang, X., Mo, H. J., Weinmann, S. M., Macciò, A. V., More, S., Cacciato, M., Skibba, R. & Kang, X., 2007. Towards a concordant model of halo occupation statistics. *MNRAS*, **376**, 841–860.

Van Waerbeke, L., Mellier, Y., Erben, T., Cuillandre, J. C., Bernardeau, F., Maoli, R., Bertin, E., McCracken, H. J., Le Fèvre, O., Fort, B., Dantel-Fort, M., Jain, B. & Schneider, P., 2000. Detection of correlated galaxy ellipticities from CFHT data: first evidence for gravitational lensing by large-scale structures. *A&A*, **358**, 30–44.

Vilenkin, A. & Shellard, E. P. S., 1994. *Cosmic strings and other topological defects*.

Wang, L., Cooray, A. & Oliver, S., 2010. A Halo Model of Local IRAS Galaxies Selected at 60 Micron Using Conditional Luminosity Functions. *ArXiv e-prints*.

Warren, M. S., Abazajian, K., Holz, D. E. & Teodoro, L., 2006. Precision Determination of the Mass Function of Dark Matter Halos. *ApJ*, **646**, 881–885.

Warren, M. S., Quinn, P. J., Salmon, J. K. & Zurek, W. H., 1992. Dark halos formed via dissipationless collapse. I - Shapes and alignment of angular momentum. *ApJ*, **399**, 405–425.

Watson, D. F., Berlind, A. A., McBride, C. K., Hogg, D. W. & Jiang, T., 2012. The Extreme Small Scales: Do Satellite Galaxies Trace Dark Matter? *ApJ*, **749**, 83.

Weinberg, D. H., Hernquist, L. & Katz, N., 1997. Photoionization, Numerical Resolution, and Galaxy Formation. *ApJ*, **477**, 8.

Weinberg, D. H., Mortonson, M. J., Eisenstein, D. J., Hirata, C., Riess, A. G. & Rozo, E., 2012. Observational Probes of Cosmic Acceleration. *ArXiv e-prints*.

White, M., 2001. The mass of a halo. *A&A*, **367**, 27–32.

White, M., 2002. The Mass Function. *ApJS*, **143**, 241–255.

White, M., Hernquist, L. & Springel, V., 2001. The Halo Model and Numerical Simulations. *ApJ*, **550**, L129–L132.

White, S. D. M., Davis, M., Efstathiou, G. & Frenk, C. S., 1987. Galaxy distribution in a cold dark matter universe. *Nature*, **330**, 451–453.

White, S. D. M. & Frenk, C. S., 1991. Galaxy formation through hierarchical clustering. *ApJ*, **379**, 52–79.

White, S. D. M. & Rees, M. J., 1978. Core condensation in heavy halos - A two-stage theory for galaxy formation and clustering. *MNRAS*, **183**, 341–358.

Wittman, D. M., Tyson, J. A., Kirkman, D., Dell'Antonio, I. & Bernstein, G., 2000. Detection of weak gravitational lensing distortions of distant galaxies by cosmic dark matter at large scales. *Nature*, **405**, 143–148.

Yang, X., Mo, H. J., Jing, Y. P., van den Bosch, F. C. & Chu, Y., 2004. Populating dark matter haloes with galaxies: comparing the 2dFGRS with mock galaxy redshift surveys. *MNRAS*, **350**, 1153–1173.

Yang, X., Mo, H. J. & van den Bosch, F. C., 2003. Constraining galaxy formation and cosmology with the conditional luminosity function of galaxies. *MNRAS*, **339**, 1057–1080.

Yoo, J., Tinker, J. L., Weinberg, D. H., Zheng, Z., Katz, N. & Davé, R., 2006. From Galaxy-Galaxy Lensing to Cosmological Parameters. *ApJ*, **652**, 26–42.

Yoshida, N., Colberg, J., White, S. D. M., Evrard, A. E., MacFarland, T. J., Couchman, H. M. P., Jenkins, A., Frenk, C. S., Pearce, F. R., Efstathiou, G., Peacock, J. A. & Thomas, P. A., 2001. Simulations of deep pencil-beam redshift surveys. *MNRAS*, **325**, 803–816.

Yoshikawa, K., Taruya, A., Jing, Y. P. & Suto, Y., 2001. Nonlinear Stochastic Biasing of Galaxies and Dark Halos in Cosmological Hydrodynamic Simulations. *ApJ*, **558**, 520–534.

Zehavi, I., Weinberg, D. H., Zheng, Z., Berlind, A. A., Frieman, J. A. et al., 2004. On Departures from a Power Law in the Galaxy Correlation Function. *ApJ*, **608**, 16–24.

Zehavi, I., Zheng, Z., Weinberg, D. H., Blanton, M. R., Bahcall, N. A., Berlind, A. A., Brinkmann, J., Frieman, J. A., Gunn, J. E., Lupton, R. H., Nichol, R. C., Percival, W. J., Schneider, D. P., Skibba, R. A., Strauss, M. A., Tegmark, M. & York, D. G., 2011. Galaxy Clustering in the Completed SDSS Redshift Survey: The Dependence on Color and Luminosity. *ApJ*, **736**, 59.

Zehavi, I., Zheng, Z., Weinberg, D. H., Frieman, J. A., Berlind, A. A. et al., 2005. The Luminosity and Color Dependence of the Galaxy Correlation Function. *ApJ*, **630**, 1–27.

Zheng, Z., 2004. Interpreting the Observed Clustering of Red Galaxies at z $\sim$ 3. *ApJ*, **610**, 61–68.

Zheng, Z., Berlind, A. A., Weinberg, D. H., Benson, A. J., Baugh, C. M., Cole, S., Davé, R., Frenk, C. S., Katz, N. & Lacey, C. G., 2005. Theoretical Models of the Halo Occupation Distribution: Separating Central and Satellite Galaxies. *ApJ*, **633**, 791–809.

Zheng, Z., Coil, A. L. & Zehavi, I., 2007. Galaxy Evolution from Halo Occupation Distribution Modeling of DEEP2 and SDSS Galaxy Clustering. *ApJ*, **667**, 760–779.

Zheng, Z., Tinker, J. L., Weinberg, D. H. & Berlind, A. A., 2002. Do Distinct Cosmological Models Predict Degenerate Halo Populations? *ApJ*, **575**, 617–633.

Zheng, Z., Zehavi, I., Eisenstein, D. J., Weinberg, D. H. & Jing, Y. P., 2009. Halo Occupation Distribution Modeling of Clustering of Luminous Red Galaxies. *ApJ*, **707**, 554–572.

Appendices

Appendix A

Estimating $w_p(r_p)$

To estimate $w_p(r_p)$ at small scales (~ 0.1 Mpc/h) up to intermediate scales ($20-30$ Mpc/h), first we estimate the two-point correlation function, $\hat{\xi}(r)$, and then we integrate it. We implement a code that measures distances between particles or galaxies (not between pixels). We optimize the code by dividing the box of the simulation in a grid with a pixel-size ~ 60 Mpc/h and compute the number of pairs of galaxies that are placed in the same pixel of the grid, and also in the nearest neighbors. The code does not count pairs twice and assumes periodicity conditions. We use the Peebles & Hauser 1974's estimator to derive the two-point correlation function, $\xi(r)$. On small scales, all the estimators are comparable, (Kerscher et al. 2000). This estimator computes the ratio of the number of pairs of galaxies separated at a distance r counted in the sample, $DD(r)$, to that expected for a random distribution, $RR(r)$, with the same mean density and sampling geometry, suitably normalized:

$$1 + \hat{\xi}(r) = \frac{DD(r)}{RR(r)} \tag{A.1}$$

Since we know the total number of galaxies, N_{gal}, and the total volume of the box, V_{box}, we are able to analytically estimate $RR(r)$ by following the expression:

$$RR(r) = N_{gal} \cdot \overline{n} \cdot \Delta V \cdot \frac{1}{2} = N_{gal} \frac{N_{gal}}{V_{box}} \frac{4}{3} \pi \left(r_2^3 - r_1^3\right) \frac{1}{2} \tag{A.2}$$

where r_1 and r_2 are the limits of a spherical shell around each galaxy ($r_1 < r_2$) and the factor $\frac{1}{2}$ is introduced in order not to count twice each pair.

The projected correlation function, w_p, can be computed following:

$$\omega_p(r_p) = 2 \int_0^{\infty} \xi(r) dr_\pi \tag{A.3}$$

In practice to derive $\omega_p(r_p)$ one has to put a limit in the integral, $r_{\pi_{max}}$. In our case we integrate up to $r_{\pi_{max}} \sim 60$ Mpc/h (as in ZH11's work), which is large enough to include most correlated pairs and suppresses noise from distant uncorrelated pairs:

$$\omega_p(r_p) = 2 \int_0^{r_{\pi_{max}}} \xi(r) dr_\pi \tag{A.4}$$

This is a way of computing the projected correlation function. However, in the thesis, we also derive the correlation function of radial or line-of-sight distance, r_π, and perpendicular distance, r_p. Therefore we can also compute the projected correlation function by using $\widehat{\xi}(r_p, r_\pi)$. Since we are working with a simulation and because we assume small angle approximation, we can place the observer at minus infinite in the z axis and taking it as the line-of-sight. The distance between particle 1 and 2 is given by the expression:

$$\vec{r}_{12} = \vec{r}_2 - \vec{r}_1 \tag{A.5}$$

and we can decompose $\vec{r}_{12}$ in a perpendicular, $\vec{r}_p$, and parallel, $\vec{r}_\pi$, components to the line-of-sight:

$$\vec{r}_{12} = \vec{r}_p + \vec{r}_\pi \tag{A.6}$$

The component in the line-of-sight of the distance is derived as:

$$\vec{r}_{\pi_{12}} = \vec{r}_{\pi_2} - \vec{r}_{\pi_1} = |z_2 - z_1|\vec{u}_{r_\pi} \tag{A.7}$$

And the component in the perpendicular component of the distance as:

$$\vec{r}_{p_{12}} = |(x_2 - x_1)^2 + (y_2 - y_1)^2|\vec{u}_{r_p} \tag{A.8}$$

In order to compute the number of random pairs $RR(r_p, r_\pi)$ we proceed in the "same" way as previously described for $RR(r)$. We analytically compute $RR(r_p, \pi)$ (after checking with the code that the expression used is consistent using random catalogues). The number of galaxies that are placed at a perpendicular distance r_p and a parallel distance r_π from a random galaxy are placed in two rings with radius r_p and height r_π:

$$RR(r_p, r_\pi) = \frac{1}{2} \cdot N_{gal} \cdot \overline{n} \cdot \Delta V \tag{A.9}$$

where ΔV is given by the expression:

$$\Delta V = 2\pi \left[r_{p_2}^2 - r_{p_1}^2\right] \left[r_{\pi_2} - r_{\pi_1}\right] \tag{A.10}$$

and therefore:

$$RR(r_p, r_\pi) = N_{gal} \frac{N_{gal}}{V_{box}} \pi \left[r_{p_2}^2 - r_{p_1}^2\right] \left[r_{\pi_2} - r_{\pi_1}\right] \tag{A.11}$$

Then, one can estimate the projected correlation function:

$$1 + \widehat{\xi}(r_p, r_\pi) = \frac{DD(r_p, r_\pi)}{RR(r_p, r_\pi)} \tag{A.12}$$

and finally, we integrate over r_π:

$$\widehat{\omega}_p(r_p) = 2 \int_0^{r_{\pi max}} \widehat{\xi}(r_p, r_\pi) dr_\pi \tag{A.13}$$