IFT

Instituto de Física Teórica
Universidade Estadual Paulista

TESE DE DOUTORAMENTO IFT–T.010/17

Cosmological analysis of optical galaxy clusters

Jose David Rivera Echeverri

Advisor

Prof. Dr. Maria Cristina Batoni Abdalla Ribeiro

Co-Advisor

Prof. Dr. Filipe Batoni Abdalla

January 15, 2018

Dedicated to my mom, my dad, my sister and my wife...

Acknowledgements

In the road to make this thesis, I found several people which contributed in differents ways. I thank to my advisor prof. Dr. Maria Cristina Batoni Abdalla Ribeiro for all advice regarding my work and my life. Her support was very important for the development of this thesis. I thank to my co-advisor prof. Dr. Filipe Batoni Abdalla for the academic discussions, for the skype meetings, for the ideas and recommendations about the research and for the excellent reception in the two academic visits performed at the University College London (UCL). I thank to prof. Dr. Marcos Lima for allowing me to join his group meetings and for his enriching conversations. I thank to Dr. Bruno Moraes, Dr. Walter Augusto dos Santos Junior, Dr. Samuel Farrens and Dr. Alex Merson for their collaboration in the development of this work. I thank to Hugo Camacho Chavéz and Michel Aguena for their tips and friendship all these years. I thank to Segundo Parra, Fernando Serna, John Sandoval, Prieslei Goulart and Natália Maia for their friendship and support for both in the university and outside it. I thank to Julian Vargas, Marcela Herrera, Diego Rengifo and Carlos Palechor for their friendship from the beginning of our difficult and wonderful adventure that began 12 years ago at the Universidad del Valle (2005). I thank to Mabel Ortiz and her family Andrea, Freddy, Juanjo and Sebas for their great friendship which has allowed me to feel a family atmosphere here in Brazil. I thank to my wife, Angy Davalos, for her support, friendship, collaboration in the most difficult moments and her unconditional love. I thank to my family, my mom Luz Echeverri, my dad Libardo Rivera and my sister Adriana Rivera for always being by my side despite the distance. I thank to my wife's family (Luz Alba Macias, Taylor Davalos and Maicol Davalos) for receiving me as another member of their family and give me their support. I thank to the UCL astrophysics Group for receiving me in two academic visits. I thank to IFT-UNESP for accepting me in the postgraduate program and allowing me to carry out the PhD research work. And finally, I thank to CAPES for financial support.

Resumo

Os aglomerados de galáxias são os maiores objetos ligados que observamos no universo. Dado que as galáxias são consideradas traçadores de matéria escura, os aglomerados de galáxias nos permitem estudar a formação e a evolução de estruturas em grande escala. As contagens do número de aglomerados de galáxias são sensíveis ao modelo cosmológico, portanto são usadas como observáveis para restringir os parâmetros cosmológicos. Nesta tese estudamos os aglomerados de galáxias óticos. Iniciamos o trabalho analisando a degradação da precisão e a exatidão no desvio para o vermelho fotométrico estimado através de métodos de aprendizagem de máquina (machine learning) `ANNz2` e `GPz`. Além do valor singular do desvio para o vermelho fotométrico clássico (isto é, valor médio ou máximo da distribuição), implementamos um estimador baseado em uma amostragem de Monte Carlo usando a função de distribuição cumulativa. Mostramos que este estimador para o algoritmo `ANNz2` apresenta a melhor concorância com a distribuição do desvio para o vermelho espectroscópico, no entanto, uma maior dispersão. Por outro lado, apresentamos o buscador de aglomerados VT-FOFz, o qual combina as técnicas de Voronoi Tessellation e Friends of Friends. Estimamos seu desempenho através de catálogos simulados. Calculamos a completeza e a pureza usando uma região de cilindrica no espaço 2+1 (ou seja, coordenadas angulares e desvio para o vermelho). Para halos maciços e aglomerados com alta riqueza, obtemos valores elevados de completeza e pureza. Comparamos os grupos de galáxias detectados através do buscador de aglomerados VT-FOFz com o catálogo RedMaPPer SDSS DR8. Recuperamos $\sim 90\%$ dos aglomerados de galáxias do catálogo RedMaPPer até o desvio para o vermelho de $z \approx 0.33$ considerando galáxias mais brilhantes com $r < 20.6$. Finalmente, realizamos uma previsão cosmológica usando um método MCMC para um modelo plano de wCDM por meio da abundância de aglomerados de galáxias. O modelo fiducial é um universo ΛCDM plano. Os efeitos devidos à massa observável estimada e aos deslocamentos para o vermelho fotométricos são incluídos através de um modelo de auto-calibração. Empregamos a função de massa de Tinker para estimar o número de contagens em uma faixa de massa e um bin de deslocamento para o vermelho. Assumimos que a riqueza e a massa do aglomerado estejam relacionadas através de uma lei de potência. Recuperamos os valores fiduciais com nível de confiança de até 2σ para os testes considerados.

Palavras Chaves: Desvio para o vermelho fotométrico, estrutura em grande escala, restrição de parâmetros cosmológicos.

Áreas do conhecimento: Cosmologia e astrofísica

Abstract

The galaxy clusters are the largest bound objects observed in the universe. Given that the galaxies are considered as tracers of dark matter, the galaxy clusters allow us to study the formation and evolution of large-scale structures. The cluster number counts are sensitive to the cosmological model, hence they are used as probes to constrain the cosmological parameters. In this work we focus on the study of optical galaxy clusters. We start analyzing the degradation of both precision and accuracy in the estimated photometric redshift via `ANNz2` and `GPz` machine learning methods. In addition to the classical singular value for the photometric redshift (i.e., mean value or maximum of the distribution), we implement an estimator based on a Monte Carlo sampling by using the cumulative distribution function. We show that this estimator for the `ANNz2` algorithm presents the best agreement with the distribution for spectroscopic redshift, nonetheless a higher scattering. On the other hand, we present the VT-FOFz cluster finder, which combines the techniques Voronoi Tessellation and Friends of Friends. Through mock catalogs, we estimate its performance. We compute the completeness and purity by using a cylindrical region in the 2+1 space (i.e., angular coordinates and redshift). For massive haloes and clusters with high richness, we obtain high values of completeness and purity. We compare the detected galaxy clusters via the VT-FOFz cluster finder with the redMaPPer SDSS DR8 cluster catalog. We recover $\sim 90\%$ of the galaxy clusters of the redMaPPer catalog until the redshift $z \approx 0.33$ considering brighter galaxies with $r < 20.6$. Finally, we perform a cosmological forecasting by using a MCMC method, for a flat wCDM model through galaxy cluster abundance. The fiducial model is a flat ΛCDM Universe. The effects due to the estimated observable mass and the photometric redshifts are included via a self-calibriation model. We employ the Tinker's mass function to estimate the number counts in a range of mass and a redshift bin. We assume that the richness and the cluster mass are related through a power law. We recover the fiducial values at 2σ confindence level for the considered tests.

Table of contents

1 Introduction **1**

2 Basic concepts of modern cosmology **5**

2.1 Metric of the Universe . . . 5

2.2 Dynamic equations in the background . . . 9

2.2.1 Fluid equation . . . 9

2.2.2 Friedmann equations . . . 11

2.2.3 Solution of Friedmann equations . . . 13

2.3 Radiation in the background . . . 14

2.3.1 Geodesic equation for the photons . . . 15

2.3.2 Redshift . . . 16

2.3.3 Thermal description of the radiation . . . 16

2.4 Distances . . . 18

2.4.1 Proper distance . . . 18

2.4.2 Angular diameter distance . . . 20

2.4.3 Comoving volume . . . 21

2.4.4 Luminosity distance . . . 22

2.5 The Universe beyond background . . . 23

2.5.1 Metric perturbations . . . 23

2.5.2 Boltzmann equation . . . 26

2.5.3 Perturbed Einstein equations . . . 29

2.6 Large-scale structure . . . 31

2.6.1 Power spectrum of matter . . . 31

2.6.2 Growth factor of matter perturbations . . . 34

2.6.3 Spherical collapse model . . . 35

2.6.4 Halo abundance . . . 39

3 Degradation analysis in the estimation of photometric redshifts 43

3.1 Data 46

3.1.1 Mock galaxy catalog 46

3.1.2 GAMA survey 48

3.1.3 SDSS DR12 sample 49

3.2 Estimating photometric redshifts 49

3.2.1 ANNz2 50

3.2.2 GPz 51

3.2.3 Metrics 52

3.2.4 Further photometric redshift estimators 53

3.3 Results 54

3.3.1 Comparison of real data and mock catalog with a complete training set 55

3.3.2 Performance degradation 61

3.4 Implication for detection of galaxy clusters 70

3.5 Conclusions and remarks 73

4 Galaxy cluster finder combining VT and FOF techniques 76

4.1 Data 78

4.1.1 Mock catalogs 78

4.1.2 GAMA DEEP survey 80

4.1.3 redMaPPer SDSS DR8 cluster catalog 81

4.2 Galaxy cluster finder 82

4.2.1 Candidate galaxies to form clusters 82

4.2.2 Finding galaxy clusters through FOF 84

4.3 Performance of the cluster finder 86

4.3.1 Matching between halos and galaxy clusters 87

4.3.2 Calibrating free parameters of the cluster finder 88

4.3.3 Completeness and purity 93

4.4 GAMA DEEP galaxy cluster catalog 98

4.4.1 Comparison between GAMA DEEP cluster catalog and redMaPPer SDSS DR8 98

4.5 Summary and conclusions 109

5 Cosmological forecast via abundance of galaxy clusters 112

5.1 Galaxy cluster abundance 114

5.1.1 Mass function 114

5.1.2 Observable clusters and theoretical haloes 116

5.2 Statistical tools for analysis 118

5.2.1 Bayes' theorem 118

5.2.2 Poisson statistics 120

5.2.3 Markov chain Monte Carlo methods 121

5.3 Fiducial model 125

5.4 Results and discussion 128

5.4.1 Constraining the parameters 137

5.5 Summary and conclusions 143

6 Conclusions 144

A General relativity for cosmology 148

B Newtonian treatment for matter perturbations 155

References 159

List of figures

2.1 Hubble law. 19

2.2 Angular diameter distance. 20

2.3 Luminosity distance. 22

2.4 Matter power spectrum for different cosmological parameters. 33

2.5 Growth factor for different values of cosmological parameters. 36

3.1 Statistical analysis of $z_{\rm phot}$ computed for GAMA survey by using magnitudes and colors. 59

3.2 Statistical analysis of $z_{\rm phot}$ computed for the mock catalog by using magnitudes and colors. 60

3.3 Comparison of global metrics for GAMA survey and the mock catalog for each photometric redshift estimator by using magnitudes and colors. 61

3.4 Scatter plots of z-spec against z-phot for r-band cuts in the range $[19.4, 20.9]$ by using the mock catalog. 64

3.5 z-spec and redshift estimators distributions for r-band cuts in the range $[19.4, 20.9]$ by using the mock catalog. 65

3.6 Comparison between the spectroscopic distribution and the stacking of photometric redshift PDFs estimated through `ANNz2` and `GPz` algorithms for slices of photometric redshift in all r-band magnitude cuts. 66

3.7 Bias and σ for r-band cuts in the range $19.4 \leq r \leq 20.9$ by using the mock catalog. 67

3.8 σ_{68} and $\mathrm{FR}_{e=0.15}$ for r-band cuts in the range $19.4 \leq r \leq 20.9$ by using the mock catalog. 68

3.9 Global metrics of each photometric redshift estimator as function of r-band cut. 69

3.10 Comparison of number of galaxies per comoving volume element computed by using each photometric redshift estimator with the spectroscopic redshift case. 72

4.1 Scatter plot between halo mass and number of galaxies inhabiting the halo. 79

4.2 Number of galaxies per comoving volume versus the spectroscopic redshift (true redshift) in each r-band cut for the mock catalog. . . . 80

4.3 Number of galaxies per comoving volume for true redshifts from the mock catalog, the estimated photometric redshifts from the mock catalog and the estimated photometric redshifts from the GAMA DEEP survey. 81

4.4 Example of the cleanup method used for a Voronoi tessellation in 2d. 84

4.5 Flow diagram of steps used in the VT-FOFz cluster finder. 87

4.6 Likelihood function against $P_{\rm cut}$ free parameter for each r-band magnitude cut. 91

4.7 Number counts versus redshift for haloes and clusters with $P_{\rm cut}$ = 1.0, and clusters with $P_{\rm cut}$ equal to the value which maximizes the likelihood function. 92

4.8 Completeness versus the mass of dark matter haloes for three redshift bins in the range $z \in [0.0, 0.5]$. 94

4.9 Purity versus the number of cluster members for three redshift bins in the range $z \in [0.0, 0.5]$. 95

4.10 Scatter plot of the mass of haloes as a function of the number of cluster members for the matches. 96

4.11 Scatter plot of the redshift of haloes as a function of the redshift of clusters for the matches. 97

4.12 Completeness and purity between redMaPPer clusters and GAMA DEEP clusters as function of redshift for $P_{\rm cut} = \{1.0,\ 0.5,\ 0.2\}$ and $b_r = 0.8b_r^*$. 101

4.13 Scatter plot of the number of galaxy members of the GAMA DEEP clusters (GD) versus the number of galaxy members of the redMaPPer clusters (rM) for $P_{\rm cut} = \{1.0,\ 0.5,\ 0.2\}$ and $b_r = 0.8b_r^*$. 102

4.14 Scatter plot of the photometric redshift of the GAMA DEEP clusters (GD) versus the photometric redshift of the redMaPPer clusters (rM) for $P_{\rm cut} = \{1.0,\, 0.5,\, 0.2\}$ and $b_r = 0.8b_r^*$. 103

4.15 Scatter plot of the number of galaxy members and the cluster radius in arcmin for the matched GAMA DEEP clusters (red) versus the non-matched GAMA DEEP clusters (blue) by using 4 redshift bins in the range $z_{\rm GD} \in [0.0,\, 0.5]$ ($P_{\rm cut} = \{1.0,\, 0.5,\, 0.2\}$ and $b_r = 0.8b_r^*$). . . 104

4.16 Completeness and purity between redMaPPer clusters and GAMA DEEP clusters as function of redshift for $P_{\rm cut} = 0.7$ and $b_r = \{0.8b_r^*,\, 0.7b_r^*,\, 0.6b_r^*\}$. 105

4.17 Scatter plot of the number of galaxy members of the GAMA DEEP clusters (GD) versus the number of galaxy members of the redMaPPer clusters (rM) for $P_{\rm cut} = 0.7$ and $b_r = \{0.8b_r^*,\, 0.7b_r^*,\, 0.6b_r^*\}$. 106

4.18 Scatter plot of the photometric redshift of the GAMA DEEP clusters (GD) versus the photometric redshift of the redMaPPer clusters (rM) for $P_{\rm cut} = 0.7$ and $b_r = \{0.8b_r^*,\, 0.7b_r^*,\, 0.6b_r^*\}$. 107

4.19 Scatter plot of the number of galaxy members and the cluster radius in arcmin for the matched GAMA DEEP clusters (red) versus the non-matched GAMA DEEP clusters (blue) by using 4 redshift bins in the range $z_{\rm GD} \in [0.0,\, 0.5]$ ($P_{\rm cut} = 0.7$ and $b_r = \{0.8b_r^*,\, 0.7b_r^*,\, 0.6b_r^*\}$).108

4.20 Number counts versus the redshift for the sample of the redMaPPer cluster catalog and the three best cases of the detected GAMA DEEP clusters ($P_{\rm cut} = 1.0,\ b_r = 0.8b_r^*$), ($P_{\rm cut} = 0.7,\ b_r = 0.8b_r^*$), ($P_{\rm cut} = 0.7,\ b_r = 0.7b_r^*$). 109

5.1 Cluster number counts for the fiducial model as well as for the synthetic data for every bin of redshift and richness. 126

5.2 Position of each walker as function of the number step with fixed nuisance parameters (Top) and with fixed cosmological parameters (Bottom). 131

5.3 Position of each walker as function of the number step with Gaussian priors (Top) and without priors (Bottom). 132

5.4 Convergence test in each parameter with fixed nuisance parameters (Top) and with fixed cosmological parameters (Bottom). 133

5.5 Convergence test in each parameter with Gaussian priors (Top) and without priors (Bottom). 134

5.6 Percentage relative error for three thinning tests in each parameter with fixed nuisance parameters (Top) and with fixed cosmological parameters (Bottom). 135

5.7 Percentage relative error for three thinning tests in each parameter with Gaussian priors (Top) and without priors (Bottom). 136

5.8 Constraint of model parameters for both 1σ and 2σ confidence regions with fixed nuisance parameters (Top) and with fixed cosmological parameters (Bottom). 139

5.9 Constraint of cosmological parameters for both 1σ and 2σ confidence regions with Gaussian priors (Top) and without priors (Bottom). . . . 140

5.10 Cluster number counts for every bin of redshift and richness comparing the fiducial model with the best-fitting at 1σ level, for the case with priors (Top) and the case without priors (Bottom). 141

A.1 Representation of a manifold. 149

List of tables

2.1 Values for the total density parameter in each possible geometry. . . . 13

3.1 Values for the characteristic magnitude ($m^{\star}_{X}$), the normalization coefficient ($\sigma^{\star}$), the bright magnitude slope (γ_o) and faint magnitude slope (γ_s) used to compute photometric noise in each photometric band (X) in the SDSS mock data. 47

3.2 The used values for the parameters of the `ANNz2` (**Top**) and `GPz` (**Bottom**) codes. 55

3.3 Number of galaxies and the threshold r-band for every used subsample in the training of the real data and the mock data. We use the same r-band magnitude range for both datasets. 56

3.4 Number of objects for each cut in the r-band. 62

4.1 Values obtained for the free parameters, threshold redshift and galaxy clusters detected after maximizing the global completeness and the likelihood function. 89

5.1 Fiducial values used to obtain the observed synthetic data of the cluster abundance per bin. 127

5.2 Walkers, steps by walker, burn-in and thinning for every performed case. 130

5.3 Best-fitting values at 1σ confidence level, the mean value and the standard deviation for the model parameters. 142

Chapter 1

Introduction

From the Maya civilization passing through the Egyptians and arriving to the ancient Greeks, humanity became interested in understanding and explaining the Universe. Several theories were developed to describe the cosmic events which are observed from the Earth. Initially, the underlying nature of these phenomena were interpreted as divine intervention. Nicolaus Copernicus (1473-1543), in his book "*De revolutionibus orbium coelestium*", broke the classical model of planetary motion. In his work, he proposed that the Sun is placed at the center of the Universe and the planets revolved around it in perfect orbits and not around the Earth. Later, Johannes Kepler (1571-1630), by using the observations of Tycho Brahe (1546-1601), described the laws which govern the planetary orbits in a heliocentric model with great accuracy. Galileo Galilei (1564-1642) discovered the existence of satellites around planets. He also observed the four largest moons of Jupiter (i.e., Io, Europa, Ganymede, and Callisto) thanks to a telescope he constructed himself. In addition, he observed that the Moon has craters and that Venus presents phases such as the Moon, among other important contributions. He is considered the father of observational astronomy.

Celestial mechanics was mathematically described with high accuracy by Sir Isaac Newton (1643-1727) thanks to the laws of motion and universal gravitation presented in his book "*PhilosophiæNaturalis Principia Mathematica*". He showed that the dynamics of the objects on Earth and in the Universe in general should be described by the very same principles. Through his theory of gravity, he derived Kepler's laws of planetary motion. His theory is also able to estimate the trajectories of comets, the tides, the equinoxes and other cosmic events. His ideas on mechanics and gravitation remain to date. Nevertheless, it was not until the early 20th century that the Newtonian principle of gravity was replaced by a more general concept. In

1915, Albert Einstein (1879-1955) presented his theory of general relativity, which describes the gravity not as force or action at a distance, but rather as a consequence of the curvature of spacetime, which is due to the energy-momentum of the objects in the Universe. Einstein's equations generalized the Newton's equations of gravitation allowing to predict several phenomena (e.g., gravitational waves, light deflection, gravitational time delay, gravitational lensing, black holes, among others).

The theory of general relativity laid down the foundations of modern cosmology. Despite the fact that Einstein's field equations describe an expanding Universe, according to solutions found by Alexander Friedmann (1888-1925) in 1922 and later independently by Georges Lemaître (1894-1966) in 1927, Albert Einstein favored a theory of a static Universe. Hence, he introduced a constant term in his field equations, which is mathematically allowed to force a static solution. It was not until that Edwin Hubble (1889-1953) set a linear relationship between the recession speed of distant galaxies and their distances through observational evidence in Hubble (1929). The model of the expanding Universe enables the hypothesis of a early stage in the lifetime of the Universe, here the Universe was in a hot, dense state, the so-called the Big Bang. After Hubble's observations, the static model was abandoned, and Albert Einstein stated that the cosmological constant was his "biggest blunder" (perhaps he was referring to his failure to notice the solution for a static Universe was instable). The Einstein-de Sitter model (i.e., a flat expanding Universe full of matter) was generally accepted to describe the evolution of the Universe until the discovery performed at the end of the 90's, in which it is shown that currently we inhabit in an accelerated expanding Universe, see Riess et al. (1998); Perlmutter et al. (1999).

The study of the evolution of the Universe depends on its geometry as well as its components (i.e., photons, neutrinos, baryons, dark matter and dark energy). Several observations suggest that the Universe is currently formed by $\sim 70\%$ of dark energy, $\sim 25\%$ of dark matter, $\sim 5\%$ of baryonic matter and a negligible fraction of radiation, see Allen et al. (2011); Planck Collaboration (2016). Note that currently the Universe is composed by $\sim 95\%$ of energy density of unknown nature. Therefore, one of the key objectives of modern cosmology is to understand these exotic components and thus able us to understand the past history of the Universe and to predict its evolution.

In order to understand the formation of the large structures observed in the Universe, it is necessary to model the evolution of the density of the matter field. The observed galaxies and galaxy clusters contain information that enables us to analyze the nature of this large scale structure. Zwicky (1933) suggested the existence of dark matter in the Coma Cluster via the observation of the velocity dispersions of galaxies in that same structure. Abell (1958) assembled the first large sample of clusters in the Northern sky. Abell et al. (1989) presented a large galaxy cluster catalog including the Southern sky. The detection of galaxy clusters is a hard task which is in constant development. Among the methods currently used to detect clusters are: X-ray observations, the Sunyaev-Zeldovich effect, gravitational lensing and optical observations, see Allen et al. (2011); Kravtsov & Borgani (2012). Galaxy cluster abundances are extremely sensible to the cosmological model, hence they allow us to constrain cosmological parameters, see Battye & Weller (2003); Mantz et al. (2010); Allen et al. (2011); Mana et al. (2013).

The main aim of this work is to attempt to perform a complete analysis of optical galaxy clusters as cosmological probes. To detect clusters we need to have an accurate and precise measurements of redshifts. To measure this, the ideal would be to use the Doppler shift in the wavelengths of known features in the spectrum of the galaxies, spectroscopically. Nonetheless, the above process is quite costly, due to long integration times. Hence, an alternative solution to this problem is to use the multi-band photometry of the galaxies. We estimate the photometric redshift of the galaxies in the Sloan Digital Sky Survey Data Release 12 (SDSS DR12) by using the following machine learning methods: `ANNz2` (Sadeh et al. (2016)) and `GPz` (Almosallam et al. (2016b)). We analyze the degradation of both precision and accuracy in the estimated photometric redshift for several samples obtained from a mock catalog with a representative and non-representative training data set in magnitude space. Moreover, we perform an analysis about the impact in the detection of galaxy cluster in these cases. The following step in this investigation is to detect the galaxy clusters from a photometric redshift survey of galaxies. To achieve it, we use the VT-FOFz cluster finder which combines two popular techniques, Voronoi Tessellation (VT) (Voronoi (1907)) and Friends of Friends (FOF) (Huchra & Geller (1982)) on photometric redshift survey. We verify the performance of the cluster finder by computing the completeness and purity on a mock catalog. We compare the obtained clusters when applying the VT-FOFz on SDSS DR12 galaxy survey with the clusters in the redMaPPer SDSS DR8 catalog from Rykoff et al. (2014). To finish this work we perform a cosmological forecasting using the galaxy cluster abundance. We assume a flat ΛCDM fiducial model and we constrain the cosmological parameters for a flat wCDM model. We employ the mass-richness relation given in Rozo et al. (2010); Mana et al. (2013); Simet et al. (2017). In addition, we utilize the self-calibration method proposed by Lima & Hu (2005); Lima & Hu (2007) to consider the effects of both the mass-observable relation and the photometric redshifts.

This thesis is split in 6 chapters, as following: In this chapter, Chapter 1, we perform a historical review of astronomy and physical cosmology, in addition we perform a general introduction to our work. In Chapter 2 we present the basic concepts of modern cosmology, we describe background evolution and linear perturbations solution in the case of standard cosmology. In Chapter 3 we analyze the accuracy and precision of photometric redshifts estimated by machine learning methods with non-representative training sets. We estimate the photometric redshift of the galaxies in the SDSS DR12 survey. In Chapter 4 we present the VT-FOFz cluster finder, we assess its performance through a mock catalog. We present a galaxy cluster catalog from the SDSS DR12 survey. In Chapter 5 we perform a cosmological forecasting using the abundance of galaxy clusters. In Chapter 6 we draw conclusions to our work.

Chapter 2

Basic concepts of modern cosmology

The cosmological principle of modern cosmology proposes that the observable Universe looks homogeneous and isotropic at large scales. Observations of the large scale structure (LSS) of the Universe and the cosmic microwave background (CMB) show that the Universe looks isotropic. Assuming that we are not privileged observers of the Universe, we can infer that the Universe is homogeneous and isotropic in all its extension. Both properties together imply that the Universe can be seem as a manifold which is maximally symmetric. The observations indicate that the Universe is homogeneous and isotropic in space, but not in all of spacetime.

2.1 Metric of the Universe

In order to describe the observations, we see that the Universe is maximally symmetric, but it evolves in time. In other words, we consider that the Universe can be described on spacelike slices each being three-dimensional, which are maximally symmetric. Therefore, we can say that the spacetime is a manifold ($\mathbb{R} \times \Sigma$), where $\mathbb{R}$ represents time and Σ is a maximally symmetric 3-manifold. The spacetime metric is given by

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + \mathbf{R}^2(t)\mathrm{d}\sigma^2, \tag{2.1}$$

where t is the timelike coordinate. This is commonly known as cosmic time. $\mathbf{R}(t)$ is called scale factor and $\mathrm{d}\sigma^2$ is the metric in Σ, which can be expressed as

$$\mathrm{d}\sigma^2 = \gamma_{ij}(u)\mathrm{d}u^i\mathrm{d}u^j. \tag{2.2}$$

Here (u^1, u^2, u^3) are comoving coordinates on Σ and γ_{ij} is a maximally symmetric 3-dimensional metric. The information about the size of the spacelike slice in a

t time is contained whitin the scale factor. An observer who is motionless in u^i coordinates is said to be comoving. Only comoving observers see the Universe as being isotropic.

By using Riemannian geometry, we know that the Riemann tensor for a 3-dimensional manifold maximally symmetric can be written as

$$^{(3)}R_{\rho\sigma\mu\nu} = \kappa\left(\gamma_{ik}\gamma_{jl} - \gamma_{il}\gamma_{jk}\right), \tag{2.3}$$

where k is a normalized measure of scalar curvature

$$\kappa = \frac{^{(3)}R}{n(n-1)}, \quad n = 3, \tag{2.4}$$

here R is constant over the manifold. The Ricci tensor is given by

$$^{(3)}R_{jl} = {^{(3)}R^i_{jil}} = 2\kappa\gamma_{jl}. \tag{2.5}$$

A maximally symmetric space is spherically symmetric, it can be shown that its metric has the following form:

$$\mathrm{d}\sigma^2 = \gamma_{ij}\mathrm{d}u^i\mathrm{d}u^j = \exp\left(2\beta(\overline{r})\right)\mathrm{d}\overline{r}^2 + \overline{r}^2\mathrm{d}\Omega^2, \tag{2.6}$$

where $\overline{r}$ is the radial coordinate and $\mathrm{d}\Omega^2 = \mathrm{d}\theta^2 + \sin^2(\theta)\mathrm{d}\phi^2$ is the metric on a 2-sphere. For more details about spherically symmetric, see the Schwarzschild solution in Carroll (2004). The components of the Ricci tensor, using equation (A.9), are given by

$$^{(3)}R_{11} = \frac{2}{\overline{r}}\frac{\partial\beta}{\partial\overline{r}}, \quad {^{(3)}R_{22}} = \exp(-2\beta)\left(\overline{r}\frac{\partial\beta}{\partial\overline{r}} - 1\right) + 1, \quad {^{(3)}R_{33}} = R_{22}\sin^2\theta. \tag{2.7}$$

Using equation (2.5) we have

$$^{(3)}R_{11} = 2\kappa\exp(2\beta), \; {^{(3)}R_{22}} = 2\kappa\overline{r}^2, \; {^{(3)}R_{33}} = 2\kappa\overline{r}^2\sin^2\theta. \tag{2.8}$$

Combining equation (2.7) and equation (2.8) we obtain the following equations

$$2\kappa\exp(2\beta) = \frac{2}{\overline{r}}\frac{\partial\beta}{\partial\overline{r}}, \tag{2.9}$$

$$2\kappa\overline{r}^2 = \exp(-2\beta)\left(\overline{r}\frac{\partial\beta}{\partial\overline{r}} - 1\right) + 1. \tag{2.10}$$

Then the expression for the β function is given by:

$$\beta(\overline{r}) = -\frac{1}{2}\ln\left(1-\kappa\overline{r}^2\right), \tag{2.11}$$

therefore the metric on Σ can be written as:

$$\mathrm{d}\sigma^2 = \frac{\mathrm{d}\overline{r}^2}{1-\kappa\overline{r}^2} + \overline{r}^2\mathrm{d}\Omega^2. \tag{2.12}$$

The parameter κ establishes the curvature and the size of spatial surfaces. In cosmology it is common to normalize this parameter, so

$$\kappa \in \{+1, 0, -1\}, \tag{2.13}$$

and we have made the choice that the scale factor $\mathbf{R}(t)$ absorbs the physical size of the manifold. The geometry of the 3-manifold Σ is determined by κ. For $\kappa = +1$ the curvature is positive and Σ has a close geometry. For $\kappa = -1$ the curvature is negative and Σ has an open geometry. For $\kappa = 0$ the curvature is zero and Σ has a flat geometry. Therefore, the spacetime metric is given by

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + \mathbf{R}^2(t)\left[\frac{\mathrm{d}\overline{r}^2}{1-\kappa\overline{r}^2} + \overline{r}^2\mathrm{d}\Omega^2\right]. \tag{2.14}$$

The above expression is known as Friedmann-Lemaître-Robertson-Walker (FLRW) metric. In modern cosmology it is common to work with a dimensionless scalar factor a, thus we perform the following substitutions

$$a = \frac{\mathbf{R}(t)}{\mathbf{R}_0}, \quad r = \mathbf{R}_0\overline{r}, \quad \kappa = \frac{k}{\mathbf{R}_0^2}, \tag{2.15}$$

where $\mathbf{R}_0$ is a parameter with distance units and k can take any value. Hence the metric (2.14) can be rewritten as

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + a^2(t)\left[\frac{\mathrm{d}r^2}{1-kr^2} + r^2\mathrm{d}\Omega^2\right]. \tag{2.16}$$

Henceforth the dimensionless scale factor will be called scale factor only. The Christoffel symbols (A.6) for the above metric are given by

$$\begin{aligned} \Gamma^0_{\ ij} &= \frac{\dot{a}}{a} g_{ij}, & \Gamma^k_{\ 0i} &= \frac{\dot{a}}{a}\delta^k_i, & \Gamma^1_{\ 11} &= \frac{kr}{1-kr^2}, \\ \Gamma^1_{\ 22} &= -r\left(1-kr^2\right), & \Gamma^2_{\ 12} &= \frac{1}{r}, & \Gamma^1_{\ 33} &= -r\left(1-kr^2\right)\sin^2\theta, \\ \Gamma^2_{\ 33} &= -\sin\theta\cos\theta, & \Gamma^3_{\ 13} &= \frac{1}{r}, & \Gamma^3_{\ 23} &= \cot\theta. \end{aligned} \tag{2.17}$$

The Ricci tensor (A.9) and the scalar curvature (A.10) are given by

$$R_{00} = -3\frac{\ddot{a}}{a}, \quad R_{ij} = \frac{g_{ij}}{a^2}\left(a\ddot{a} + 2\dot{a}^2 + 2k\right), \quad R = 6\left[\frac{\ddot{a}}{a} + \left(\frac{\dot{a}}{a}\right)^2 + \frac{k}{a^2}\right]. \tag{2.18}$$

Another way to write the FLRW metric is by defining a new radial coordinate. Let χ be defined by equation

$$\mathrm{d}\chi = \frac{\mathrm{d}r}{\sqrt{1-kr^2}}. \tag{2.19}$$

By integrating the above equation, we have

$$r = S_k(\chi) \equiv \begin{cases} \frac{\sin(\sqrt{|k|}\chi)}{\sqrt{|k|}} & \text{for} \quad k>0, \\ \chi & \text{for} \quad k=0, \\ \frac{\sinh(\sqrt{|k|}\chi)}{\sqrt{|k|}} & \text{for} \quad k<0. \end{cases} \tag{2.20}$$

The FLRW metric can be written as

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + a^2(t)\left(\mathrm{d}\chi^2 + S_k^2(\chi)\mathrm{d}\Omega^2\right). \tag{2.21}$$

Another way to write the FLRW metric is

$$\mathrm{d}s^2 = a^2(t)\left[-\mathrm{d}\eta^2 + \left(\mathrm{d}\chi^2 + S_k^2(\chi)\mathrm{d}\Omega^2\right)\right], \tag{2.22}$$

where

$$\mathrm{d}\eta \equiv \frac{\mathrm{d}t}{a(t)} \tag{2.23}$$

is called conformal time. In perturbation theory it is more convenient to use this definition rather than the cosmic time.

2.2 Dynamic equations in the background

The scale factor contains the information about the dynamics of the Universe. In this section we will solve the field equations which allow us to describe the behavior of the scale factor as a function of time.

2.2.1 Fluid equation

The cosmological principle implies that the matter distribution in the Universe can be described by a perfect fluid. For an isotropic fluid, in a reference frame, the fluid is at rest in comoving coordinates. Here, the fluid 4-velocity of the fluid is given by:

$$U^{\mu} = (1, 0, 0, 0), \tag{2.24}$$

and the energy-momentum tensor is given by

$$T_{\mu\nu} = (\rho + P)\, U_{\mu} U_{\nu} + P g_{\mu\nu}, \tag{2.25}$$

where P is the pressure and ρ is the fluid density. By using the Christoffel symbols (2.17) and the energy conservation equation (A.22), we obtain

$$\dot{\rho} + 3\frac{\dot{a}}{a}\,(\rho + P) = 0. \tag{2.26}$$

This equation is known as fluid equation. We note that the expansion of the Universe can lead to local changes in the energy density. To solve the fluid equation it is necessary to establish how the pressure and energy density are related to each other. For a perfect fluid we assume that

$$P(\rho) = w\rho, \tag{2.27}$$

where the parameter w can depend on a scalar factor in isotropic models of dark energy and inflation, see Dodelson (2003); Mukhanov (2005). However, in the simplest models it is constant. The above expression is known as the equation of state. The general solution for equation (2.26) is

$$\rho_i(a) = \rho_i \exp\left(-3\int_1^a \frac{\mathrm{d}\tilde{a}}{\tilde{a}}\,(1 + w_i(\tilde{a}))\right), \tag{2.28}$$

where the index i indicates the type of component (non-relativistic matter, radiation, dark energy or some exotic component). If w_i is constant we have that

$$\rho_i(a) = \rho_i a^{-3(1+w_i)}. \tag{2.29}$$

Note that to indicate whether a parameter evolves with the factor scale we write $p(a)$, otherwise we write only p. By convention we set the scalar factor today a_0 to one.

Non-relativistic matter is characterized by having $P \approx 0$, therefore $w_m = 0$. In this component we includ the visible matter (i.e. baryons), which also has negligible pressure in comparison with its energy density, and cold dark matter (i.e. non-visible matter which does not interact with baryons). When the energy density is mainly due to this component, we say that the Universe is matter-dominated. The energy density of non-relativistic matter is given by

$$\rho_m(a) = \rho_m a^{-3}. \tag{2.30}$$

The relativistic component includes electromagnetic radiation (i.e. photons), massless neutrinos and massive particles with relativistic velocities (e.g. relativistic massive neutrinos, hot dark matter and warm one). The energy-momentum tensor for electromagnetic field is given by

$$^{(\mathrm{EM})}T^{\mu\nu} = F^{\mu\lambda}F^{\nu}_{\lambda} - \frac{1}{4}g^{\mu\nu}F^{\lambda\sigma}F_{\lambda\sigma}, \tag{2.31}$$

where $F^{\mu\nu}$ is the electromagnetic tensor. It is straightforward to show that the trace of tensor (2.31) is zero. On the other hand, this component is a perfect fluid in the Universe, then

$$T^{\mu}_{\mu} = -\rho + 3P = {}^{(\mathrm{EM})}T^{\mu}_{\mu} = 0, \tag{2.32}$$

therefore we have $w_r = 1/3$ for electromagnetic radiation (we show later on that the relativistic particles satisfy the same equation of state). If the energy density is mainly given by this component, we say that the Universe is radiation-dominated. Here the energy density is

$$\rho_r(a) = \rho_r a^{-4}. \tag{2.33}$$

In addition to these cosmological fluids, the observations of type Ia supernova performed by Riess et al. (1998) and Perlmutter et al. (1999) suggest that the Universe is expanding at an accelerated rate. To explain this phenomenon one includes an exotic fluid called dark energy. We will discuss this in more detail later on.

2.2.2 Friedmann equations

The Friedmann equations are derived from the Einstein field equations given in (A.25), the fluid tensor given in (2.25), the metric given in (2.16) and the results obtained for the Ricci tensor given in (2.18):

$$\begin{aligned} \left(\frac{\dot{a}}{a}\right)^2 &= \frac{8\pi G}{3}\sum_i \rho_i(a) - \frac{k}{a^2}, & (2.34)\\ \frac{\ddot{a}}{a} &= -\frac{4\pi G}{3}\sum_i \left(\rho_i(a) + 3P_i(a)\right). & (2.35) \end{aligned}$$

The first equation relates the expansion rate to the total energy density and the geometry of the Universe. The second equation, also known as the acceleration equation, allows us to compute the acceleration rate of the Universe regardless its geometry. The Friedmann equations are a differential set of equations that, together with the equation of state for each component, describe the evolution of the scale factor with the time.

As we mentioned above, the Universe has currently a positive acceleration, so according to equation (2.35) it is necessary that there exists at least one component with negative pressure and its equation of state must satisfy

$$\frac{P_{\rm DE}}{\rho_{\rm DE}} = w_{\rm DE} < -\frac{1}{3}. \qquad (2.36)$$

A candidate for this component is the vacuum energy which has an equation of state $w_\Lambda = -1$ with a constant energy density. Vacuum energy appears as a modification of Einstein field equations (see equation (A.27)) and the Friedmann equations take the following form:

$$\begin{aligned} \left(\frac{\dot{a}}{a}\right)^2 &= \frac{8\pi G}{3}\sum_i \rho_i(a) - \frac{k}{a^2} + \frac{\Lambda}{3}, & (2.37)\\ \frac{\ddot{a}}{a} &= -\frac{4\pi G}{3}\sum_i \left(\rho_i(a) + 3P_i(a)\right) + \frac{\Lambda}{3}, & (2.38) \end{aligned}$$

where Λ is known as cosmological constant. There are however alternative theories to explain dark energy. We can consider a minimally coupled scalar field, see

Copeland et al. (2006) which satisfies condition (2.36), or we can modify the laws of gravity by introducing a general function $f(R)$ in the Einstein-Hilbert action (A.14), see Amendola et al. (2007). In order to understand the nature of the dark energy Chevallier-Polarski-Linder (CPL) proposed in Chevallier & Polarski (2001); Linder (2003); an empirical parameterization for the dark energy equation of state

$$w_{\mathrm{DE}}(a) = w_0 + (1-a)w_a, \tag{2.39}$$

where w_0 and w_a are constants. If $w_0 = -1$ and $w_a = 0$ we obtain to vacuum energy model. We define here some useful parameters:

- Hubble parameter:
$$H \equiv \frac{\dot{a}}{a}. \tag{2.40}$$
This parameter characterizes the expansion rate of the Universe. The current value is $H_0 = 100\,h\,\mathrm{km/s\ Mpc}^{-1}$, where $h \sim 0.67$, see Planck Collaboration (2016).

- Critical density:
$$\rho_{cr}(a) = \frac{3H^2}{8\pi G}. \tag{2.41}$$

- Density parameter:
$$\Omega_i(a) = \frac{\rho_i(a)}{\rho_{cr}(a)}, \tag{2.42}$$
where the index i may be non-relativistic matter, radiation, dark energy or another exotic component.

- Deceleration parameter:
$$q(a) = -\frac{a\ddot{a}}{\dot{a}^2}. \tag{2.43}$$

- Curvature density:
$$\rho_k = -\frac{3k}{8\pi G}. \tag{2.44}$$

- Vacuum energy density:
$$\rho_\Lambda = \frac{\Lambda}{8\pi G}. \tag{2.45}$$

By using these definitions and equation (2.30) and equation (2.33) we can write equation (2.37) as

$$H^2 = H_0^2\left(\frac{\Omega_r}{a^4} + \frac{\Omega_m}{a^3} + \frac{\Omega_k}{a^2} + \Omega_\Lambda\right). \tag{2.46}$$

Table 2.1: Values for the total density parameter in each possible geometry.

$\Omega_k > 0$	$k < 0$	$\Omega < 1$	$\rho < \rho_{cr}$	Hyperbolic
$\Omega_k = 0$	$k = 0$	$\Omega = 1$	$\rho = \rho_{cr}$	Flat
$\Omega_k < 0$	$k > 0$	$\Omega > 1$	$\rho > \rho_{cr}$	Spherical

Therefore, for $a = a_0$ the above equation is given by

$$\Omega_r + \Omega_m + \Omega_\Lambda + \Omega_k = 1. \tag{2.47}$$

Let $\Omega = \Omega_m + \Omega_r + \Omega_\Lambda$ be the total density parameter, we have:

$$\Omega_k = 1 - \Omega. \tag{2.48}$$

This equation relates the geometry with the total energy. Table 2.1 shows the values of Ω for each scenario. Here we observe that the critical density determines the geometry of the Universe.

2.2.3 Solution of Friedmann equations

We note that the Friedmann equation (2.46) cannot be solved analytically for all cosmological models. In fact there are few models that enable us to obtain an analytical solution. Here we show solutions for three scenarios which occurred during the history of the Universe.

- **Radiation-dominated Universe:** a Universe composed by radiation and relativistic particles. This scenario is similar to the early Universe where the photons interact strongly with the baryons. Here all particles have speeds close to the light speed. The Friedmann equation is given by

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{H_0^2}{a^4}, \ \text{so} \ a \propto t^{1/2}. \tag{2.49}$$

- **Matter-dominated Universe:** a Universe composed by non-relativistic matter. This is another scenario in the history of the Universe. It is in this period that the first neutral atoms are formed as well as the structures that we observe today. This solution is known as Einstein-de Sitter Universe. The Friedmann equation is given by

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{H_0^2}{a^3}, \quad \text{so} \quad a \propto t^{2/3}. \tag{2.50}$$

- **Vacuum-dominated Universe:** According to current observations this is the currently preferred scenario. Here the Universe has an accelerated expansion. This solution is known as a de Sitter Universe where the density parameter is constant. The Friedmann equation is given by

$$\left(\frac{\dot{a}}{a}\right)^2 = H_0, \quad \text{so} \quad a \propto \exp(H_0 t). \tag{2.51}$$

The ΛCDM model is widely accepted to describe the Universe in the modern cosmology. This model tell us that we live in a flat Universe dominated by cold dark matter (CDM) and vacuum energy. $\{H_0, \Omega_r, \Omega_m, \Omega_\Lambda\}$ are free parameters which may be constrained via a cosmological observable (e.g. distance of type Ia supernovae, CMB, galaxy clusters, among others). The Friedmann equation (2.46) for the ΛCDM model is given by

$$\left(\frac{\dot{a}}{a}\right)^2 = H_0^2\left(\frac{\Omega_r}{a^4} + \frac{\Omega_m}{a^3} + \Omega_\Lambda\right). \tag{2.52}$$

Currently the value of the free parameters according to Planck collaboration (2016) are $H_0 = 67.74 \pm 0.46$, $\Omega_m = 0.3089 \pm 0.0062$, $\Omega_\Lambda = 0.6911 \pm 0.0062$ at the 1σ confidence level. The density parameter of radiation is also negligible. From this point forwards, we shall consider the Universe as flat, unless otherwise stated.

2.3 Radiation in the background

Photons are particles which can provide us information to understand the Universe, therefore it is very important to understand their nature and their behavior in the cosmological background. Here we are going to show the properties of the photons and relativistic particles in a homogeneous and isotropic Universe.

2.3.1 Geodesic equation for the photons

The geodesic equation (A.11) for a particle in a homogeneous and isotropic Universe is given by

$$E\frac{\mathrm{d}\mathbf{P}^{\nu}}{\mathrm{d}t}+\Gamma^{\nu}{}_{\alpha\beta}\mathbf{P}^{\alpha}\mathbf{P}^{\beta}=0. \tag{2.53}$$

Here $\mathbf{P}^{\nu}$ is the 4-momentum and it is defined as

$$\mathbf{P}^{\nu}\equiv\frac{\mathrm{d}x^{\nu}}{\mathrm{d}\lambda}=(E,p_x,p_y,p_z), \tag{2.54}$$

where E is the energy. We know that the photon is a massless particle, thus we have the following condition for the 4-momentum

$$g_{\mu\nu}\mathbf{P}^{\mu}\mathbf{P}^{\nu}=-E_{\gamma}^{2}+g_{ij}\mathbf{P}^{i}\mathbf{P}^{j}=-m_{\gamma}^{2}=0. \tag{2.55}$$

So that the zero component of equation (2.53) for photons is given by

$$E_{\gamma}\frac{\mathrm{d}E_{\gamma}}{\mathrm{d}t}+E_{\gamma}^{2}\frac{\dot{a}}{a}=0. \tag{2.56}$$

The above equation allows us to observe that the photon energy is inversely proportional to the scale factor

$$E_{\gamma}\propto\frac{1}{a}. \tag{2.57}$$

The energy density of photons may be seen as the product between their energy and the number density n_{γ}, therefore

$$\rho_{\gamma}=E_{\gamma}n_{\gamma}. \tag{2.58}$$

Moreover, the number density is inversely proportional to the volume. Thus $n_{\gamma}\propto a^{-3}$, then

$$\rho_{\gamma}\propto\frac{1}{a^{4}}. \tag{2.59}$$

The above result has already been derived, see equation (2.33).

2.3.2 Redshift

Equation (2.57) indicates that photons lose energy as they propagate from any emitter to any observer, so we have

$$\frac{E_\gamma^{(\mathrm{emit})}}{E_\gamma^{(\mathrm{obs})}} = \frac{a_{(\mathrm{obs})}}{a_{\mathrm{emit}}}. \tag{2.60}$$

By using the relation between energy and frequency we can rewrite equation (2.60) as

$$\frac{a_{(\mathrm{obs})}}{a_{\mathrm{emit}}} = \frac{\nu_{\mathrm{emit}}}{\nu_{\mathrm{obs}}} = 1 + z, \tag{2.61}$$

where z is the cosmological redshift of the emitter. If we place the observer $a_{\mathrm{obs}} = 1$ and the emitter is any observed object in the Universe $a_{\mathrm{emit}} = a$, we have

$$1 + z = \frac{1}{a}. \tag{2.62}$$

The redshift is an useful property in modern cosmology, as it allows us to measure distances of distant objects. In addition, we can obtain measure by observing the spectrum of an object. This quantity is often used instead of time. In chapter 3 we focus on this property in detail.

2.3.3 Thermal description of the radiation

From statistical mechanics we know that the particle number between the position $\vec{x}$ and $\vec{x} + \mathrm{d}\vec{x}$, and between the momentum $\vec{p}$ and $\vec{p} + \mathrm{d}\vec{p}$ are given by

$$\mathrm{d}N = f_i(\vec{x}, \vec{p}, t)\frac{\mathrm{d}^3x\mathrm{d}^3p}{(2\pi)^2}, \tag{2.63}$$

where $f_i(\vec{x}, \vec{p}, t)$ is the distribution function and i indicates the nature of the particles, bosons (Bose-Einstein statistics) or fermions (Fermi-Dirac statistics). Therefore, the number density can be written as

$$n_i(\vec{x}, t) = g_i \int \frac{\mathrm{d}^3p}{(2\pi)^3} f_i(\vec{x}, \vec{p}, t), \tag{2.64}$$

where g_i is the degeneracy of the species. The energy density can be written as

$$\rho_i(\vec{x}, t) = g_i \int \frac{\mathrm{d}^3p}{(2\pi)^3} f_i(\vec{x}, \vec{p}, t)E(p), \tag{2.65}$$

and the pressure can be written as

$$P(\vec{x},t) = g_i \int \frac{\mathrm{d}^3 p}{(2\pi)^3} f_i(\vec{x},\vec{p},t) \frac{p^2}{3E(p)}. \tag{2.66}$$

At equilibrium, the distribution function only depends on $\|\vec{p}\|$, so that for bosons we have

$$f_{\mathrm{BE}} = \frac{1}{\exp\left[(E-\mu)/T\right]-1}, \tag{2.67}$$

and for fermions

$$f_{\mathrm{FD}} = \frac{1}{\exp\left[(E-\mu)/T\right]+1}, \tag{2.68}$$

where μ is the chemical potential and it is related to particle interactions.

The energy density for photons at equilibrium is given by the Bose-Einstein distribution with zero chemical potential. We can therefore write that

$$\rho_\gamma = g_\gamma \frac{1}{2\pi^2} \int \frac{p^3 \mathrm{d}p}{\exp\left(p/T_\gamma\right)-1}, \tag{2.69}$$

where $g_\gamma = 2$ is the degeneracy for a given state, and we use $E = p$ for photons. Solving the above integral we obtain

$$\rho_\gamma \propto T_\gamma^4, \tag{2.70}$$

which combined with equation (2.59) gives us

$$T_\gamma \propto \frac{1}{a}. \tag{2.71}$$

For the case of relativistic particles at equilibrium with negligible chemical potential (in most astronomical cases, the chemical potential is much smaller than the temperature), the pressure is given by

$$P = g_i \int \frac{\mathrm{d}^3 p}{(2\pi)^3} f_i(E/T) \frac{p^2}{3E}. \tag{2.72}$$

As $p \gg m$ the energy satisfies $E = \sqrt{p^2+m^2} \approx p$, thus equation (2.72) can be rewritten as

$$P = g_i \int \frac{\mathrm{d}^3 p}{(2\pi)^3} f_i(E/T) \frac{p}{3} = \frac{1}{3} g_i \int \frac{\mathrm{d}^3 p}{(2\pi)^3} f_i(E/T) E = \frac{1}{3}\rho. \tag{2.73}$$

Therefore, the equation of state $w = P/\rho$ for the relativistic particles is the same that for the photons independently of their nature.

2.4 Distances

In order to identify the precise position of an object in the Universe it is necessary to measure its distance from us. Here we outline the different definitions of distance between two observers in cosmology.

2.4.1 Proper distance

The physical distance or proper distance is the simplest distance definition in cosmology. Here we measure the distance between an object and an observer in the 3-manifold Σ for an specific time, see above discussion of equation (2.1). Taking equation (2.21) we define the proper distance as

$$d_p(t) \equiv \int_c \mathrm{d}\sigma\, a(t), \quad \text{with } \mathrm{d}\sigma^2 = \mathrm{d}\chi^2 + S_k^2(\chi)\mathrm{d}\Omega^2, \tag{2.74}$$

where c is the spacelike geodesic. Given the isotropy of space we have that θ and ϕ constant along the geodesic, therefore equation (2.74) can be rewritten as

$$d_p(t) = \int_0^\chi \mathrm{d}\tilde{\chi}\, a(t) = a(t)\chi, \tag{2.75}$$

where $\tilde{\chi} = 0$ is the position of the observer and $\tilde{\chi} = \chi$ is the radial position of the object. We note that the proper distance is the scale factor times the radial coordinate, so that χ is the distance measured by a comoving observer. This is called the *comoving distance*, D_c. In order to compute χ, we assume that the object emits a photon which travels along a null geodesic towards the observer, therefore

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + a^2(t)\mathrm{d}\chi^2 = 0. \tag{2.76}$$

Integrating the above equation, we have

$$D_c = \chi = \int_a^1 \frac{\mathrm{d}\tilde{a}}{\tilde{a}^2 H(\tilde{a})} = \int_0^z \frac{\mathrm{d}\tilde{z}}{H(\tilde{z})}. \tag{2.77}$$

The velocity for an object is given by

$$v = \frac{\mathrm{d}d_p}{\mathrm{d}t} = a\dot{\chi} + \dot{a}\chi = a\dot{\chi} + Hd_p. \tag{2.78}$$

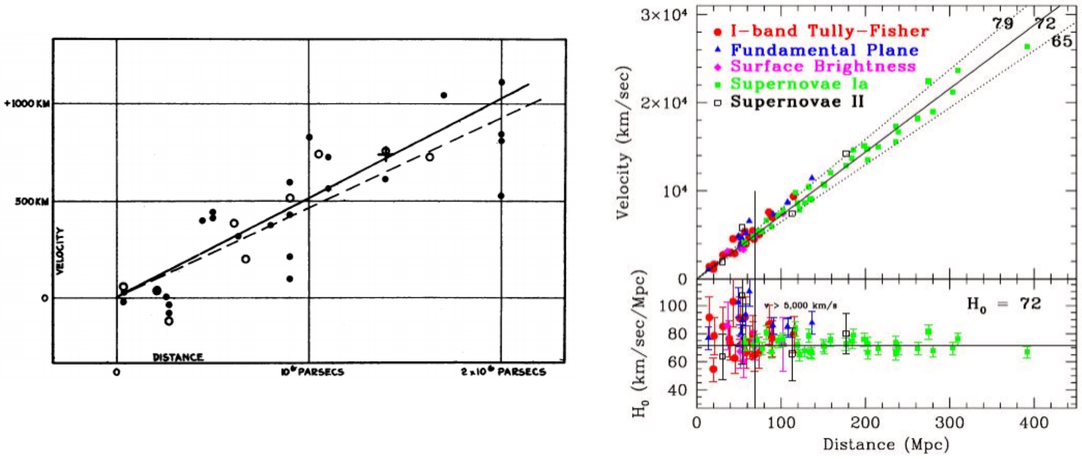

Figure 2.1: Left: The original Hubble diagram presented in Hubble (1929). The two lines use a different correction for the movement of the sun. Right: Velocity against distance for galaxies until 400 Mpc calibrated by Cepheid distance scale, see Freedman et al. (2001). An adjustment to the slope yields a value of $H_0 = 72\,km/s\,\mathrm{Mpc}^{-1}$.

The term $a\dot{\chi}$ is known as peculiar velocity and it is due to the individual speed of objects relative to the background expansion. The term Hd_p is known as Hubble flow and it is due to the expansion of the Universe. In Hubble (1929) measured the recession velocity of nearby galaxies in low redshift and he showed that the velocities of galaxies are proportional to their distances, see figure 2.1. The relation between velocity and distance is known as Hubble's law and it is given by

$$v = H_0 d_p, \tag{2.79}$$

where H_0 is the Hubble's constant. The velocity satisfies the above equation if we neglect the peculiar velocity in equation (2.78) and we consider object in low redshift (i.e., this condition assumes that the variation between the 3-manifolds Σ_1 and Σ_2 is neglected for a $\mathrm{d}t = t_2 - t_1$ very small).

The comoving horizon or just horizon is defined as the maximum comoving distance traveled by photons from the beginning of the Universe, thus we have

$$D_H = \int_0^t \frac{\mathrm{d}\tilde{t}}{a(\tilde{t})} = \int_z^\infty \frac{\mathrm{d}\tilde{z}}{H(\tilde{z})}. \tag{2.80}$$

Therefore, regions separated by distances greater than D_H are not causally connected. The comoving horizon is also identified with the conformal time (2.23). The physical horizon is defined as the scale factor times the comoving horizon, then

$$d_H = a(t)D_H. \tag{2.81}$$

2.4.2 Angular diameter distance

The angular diameter distance d_A is defined as the ratio between the transverse physical length δl of an object and its angular separation $\delta\alpha$. Figure 2.2 is a diagram which shows these amounts for an object. Therefore, we have

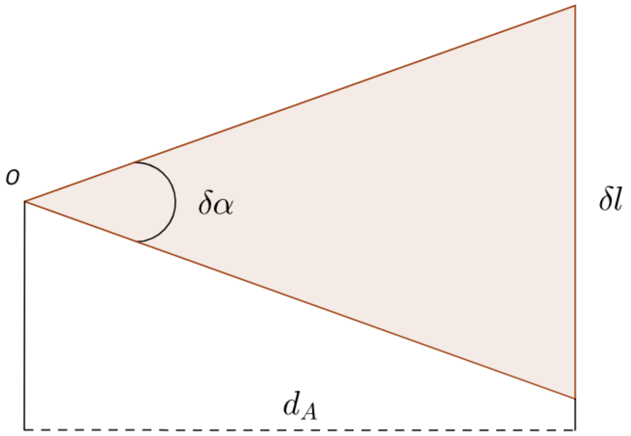

Figure 2.2: Diagram that represents the geometric relation between the angular separation ($\delta\alpha$), the transverse physical length (δl) and the angular diameter distance (d_A). Here o is the observer position.

$$d_A = \frac{\delta l}{\delta\alpha}. \tag{2.82}$$

We can compute the transverse physical length for an object assuming that its extremities are localized in (χ, θ, ϕ) and $(\chi, \theta + \delta\alpha)$ according to the coordinates defined in the metric (2.21), therefore we have

$$\delta l = a(t)\, S_k \int_{\theta}^{\theta+\delta\alpha} \mathrm{d}\tilde{\theta} = a(t)\, S_k \delta\alpha. \tag{2.83}$$

The angular diameter distance is related with the comoving distance via

$$d_A = a(t)\, S_k(\chi) = \frac{S_k(\chi)}{1+z}. \tag{2.84}$$

Note that for a flat Universe the angular diameter distance is the proper distance. We define the angular comoving distance or transverse comoving distance as

$$D_A = \frac{d_A}{a(t)} = S_k(\chi). \tag{2.85}$$

In the flat Universe case we have that $D_A = \chi$. By using the definition of angular comoving distance, the FLRW metric (2.21) can be rewritten as

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + a^2(t)\left[\mathrm{d}\chi^2 + D_A^2 \mathrm{d}\Omega^2\right]. \tag{2.86}$$

2.4.3 Comoving volume

We can define the volume in the 3-manifold Σ which describes the space. Riemannian geometry tells us that the volume element in a manifold is given by

$$\mathrm{d}V = \sqrt{|g|}\mathrm{d}x^1 ... \mathrm{d}x^n. \tag{2.87}$$

The metric for the comoving space is given by the expression between the brackets of equation (2.86), so the comoving volume element can be written as

$$\mathrm{d}V_c = D_A^2 \mathrm{d}\chi \mathrm{d}\Omega = \frac{D_A^2}{H(z)} \mathrm{d}z \mathrm{d}\Omega. \tag{2.88}$$

The physical volume element is defined as

$$\mathrm{d}V_p = a^3(t) \mathrm{d}V_c = \frac{a^3(t) D_A^2}{H(z)} \mathrm{d}z \mathrm{d}\Omega. \tag{2.89}$$

2.4.4 Luminosity distance

The luminosity distance is defined by using the flux from an object with known luminosity. The observed flux F from the source with luminosity L at distance d is given by:

$$F = \frac{L}{4\pi d_L^2}, \tag{2.90}$$

since the total luminosity through of a spherical shell with area $4\pi d^2$ is constant. Here d_L is called luminosity distance. Equation (2.90) is consistent with other distance definitions if we are in a static Universe, but our Universe is expanding. Therefore, this distance does not equal to the proper and angular distances. The expression between the brackets of equation (2.21) is the metric for a comoving sphere with radius $S_k(\chi)$. In order to generalize the flux concept, we assume a source at center of comoving coordinate system, it emits photons with the same energy. They travel through a spherical shell with radius $S_k(\chi)$. The observer is on the shell. Figure 2.3 shows a diagram with the above "Gedanken-experiment". Since the photons emitted are redshifted in an expanding Universe we have

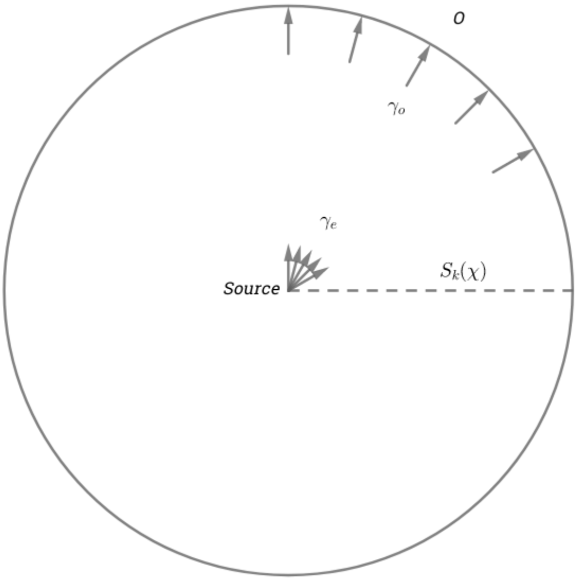

Figure 2.3: Diagram that represents the path of the photons from emission source to spherical shell in which the observed is. Here γ_e are the emitted photons, γ_o are the observed photons, $S_k(\chi)$ is the radius of the comoving spherical shell and o is the observer position.

$$\frac{\mathrm{d}t_e}{a(t_e)} = \frac{\mathrm{d}t_o}{a(t_o)}, \tag{2.91}$$

where the subindex (e) is referred to the emission time and the subindex (o) is referred to the observation time. On the other hand, the luminosity is defined as the product between the photon energy and the number of photons which crosses the spherical shell by unity of time. Therefore the luminosity perceived by the observer is given by

$$L_o = E_o \frac{\mathrm{d}N_o}{\mathrm{d}t_o}. \tag{2.92}$$

Equation (2.57) allows us to assert that $E_o a_o = E_e a_e$, hence by using equation (2.91) we obtain

$$L_o = E_e \frac{a_e}{a_o} \frac{\mathrm{d}N_e}{\mathrm{d}t_e} \frac{a_e}{a_o} = E_e \frac{\mathrm{d}N_e}{\mathrm{d}t_e} a^2 = L_e a^2. \tag{2.93}$$

Here we assume that $a_o = 1$ and $a_e = a$. Therefore, the flux observed is given by

$$F = \frac{L_o}{4\pi S_k(\chi)^2} = \frac{L_e a^2}{4\pi S_k(\chi)^2} = \frac{L_e}{4\pi d_L^2}, \tag{2.94}$$

where d_L is given by

$$d_L = (1+z) S_k(\chi). \tag{2.95}$$

Note that according to equation (2.85) the luminosity distance is related with the angular diameter distance via

$$d_L = (1+z)^2 d_A. \tag{2.96}$$

2.5 The Universe beyond background

In order to explain the formation and evolution of the large scale structure, we need a perturbative description of the Universe. The primordial inhomogeneities were generated by perturbations which were relatively small in the scales of interest for this chapter. We focus here on linear perturbations.

2.5.1 Metric perturbations

In the above sections we considered that the metric in the background is characterized by a single function $a(t)$, which depends only on time and not on space. Here we will consider perturbations around a smooth Universe, therefore the metric needs to be characterized by functions which depend on both space and time.

The decomposition theorem tells us that perturbations in the metric can be divided up into three types: scalar, vector and tensor perturbations. Each of these types of perturbations evolves independently, see Dodelson (2003). In order to study fluctuations to linear order, we need to consider all perturbations in the metric. For early Universe physics (i.e., inflation), it is necessary to consider also tensor perturbations, in addition there are cosmological theories where vector perturbations are important. For more details, see Dodelson (2003); Mukhanov (2005). Here we consider scalar perturbation to the background, given that these are the most important elements in the coupling of matter and radiation perturbations. The metric, with perturbations, is given by

$$\mathrm{d}s^2 = -\left(1+2A\right)\mathrm{d}t^2 - a(t)\partial_i B \mathrm{d}x^i \mathrm{d}t + a(t)^2\left(\delta_{ij}\left[1+2\psi\right] - 2\partial_{ij}E\right)\mathrm{d}x^i\mathrm{d}x^j, \quad (2.97)$$

where $A(x^\mu)$, $B(x^\mu)$, $\psi(x^\mu)$ and $E(x^\mu)$ are scalar perturbations to the metric. We can write the metric (2.97) in a more convenient way by choosing other coordinate system, or gauge. Let x and $\tilde{x}$ be two coordinate systems, then the metric satisfies the following transformation law:

$$\tilde{g}_{\alpha\beta}(\tilde{x})\frac{\partial \tilde{x}^\alpha}{\partial x^\mu}\frac{\partial \tilde{x}^\beta}{\partial x^\nu} = g_{\mu\nu}(x). \quad (2.98)$$

The most general coordinate transformation is

$$\begin{aligned} t &\rightarrow \tilde{t} = t + \xi^0(x^\mu), && (2.99) \\ x^i &\rightarrow \tilde{x}^i = x^i + \delta^{ij}\partial_j \xi(x^\mu), && (2.100) \end{aligned}$$

where ξ^0 and ξ are small linear perturbations. By using the above expressions and the transformation law (2.98) it is straightforward to show that the scalar perturbations of the metric (2.97) satisfy

$$\begin{aligned} A &\rightarrow \tilde{A} = A - \dot{\xi}^0, && (2.101) \\ \psi &\rightarrow \tilde{\psi} = \psi - H\xi^0, && (2.102) \\ B &\rightarrow \tilde{B} = B - \frac{\xi^0}{a} + a\dot{\xi}, && (2.103) \\ E &\rightarrow \tilde{E} = E + \xi. && (2.104) \end{aligned}$$

Note that the scalar perturbations depend on two functions which characterize the coordinate transformations, so we really need only two functions to characterize the metric (2.97). Bardeen (1980) proposed two gauge invariant variables which remain unchanged under a general coordinate transformation. These variables are given by

$$\begin{aligned} \Phi_A &\equiv A+\frac{\partial}{\partial t}\left[a\left(a\dot{E}-B\right)\right], && (2.105)\\ \Phi_H &\equiv -\psi+aH\left(B-a\dot{E}\right). && (2.106) \end{aligned}$$

The Bardeen variables are useful to compute scalar perturbations in different gauges.

Here we are going to use the conformal Newtonian gauge in which $E=B=0$, so the metric (2.97) can be rewritten as

$$\mathrm{d}s^2=-\left(1+2\Psi(x^\mu)\right)\mathrm{d}t^2+a(t)^2\left(1+2\Phi(x^\mu)\right)\delta_{ij}\mathrm{d}x^i\mathrm{d}x^j, \tag{2.107}$$

where $A=\Psi$ and $\psi=\Phi$. The Christoffel symbols for the metric (2.107) are given by

$$\begin{gathered} \Gamma^0{}_{00}=\dot{\Psi},\ \Gamma^i{}_{0j}=\Gamma^i{}_{j0}=\delta_{ij}(H+\dot{\Phi}),\ \Gamma^0{}_{ij}=\delta_{ij}a^2\left[H+2H(\Phi-\Psi)+\dot{\Phi}\right], \quad (2.108)\\ \Gamma^i{}_{00}=\frac{1}{a^2}\partial_i\Psi,\ \Gamma^0{}_{0i}=\Gamma^0{}_{i0}=\partial_i\Psi,\ \Gamma^i{}_{jk}=\delta_{ki}\partial_j\Phi-\delta_{jk}\partial_i\Phi+\delta_{ij}\partial_k\Phi. \end{gathered}$$

The components of Ricci tensor are given by

$$\begin{aligned} R_{00} &= -3\frac{\ddot{a}}{a}+\frac{1}{a^2}\nabla^2\Psi+3H(\dot{\Psi}-2\dot{\Phi})-3\ddot{\Phi}, && (2.109)\\ R_{0i} &= -2\partial_i(\dot{\Phi}-H\Psi),\\ R_{ij} &= \delta_{ij}\left[\left(a\ddot{a}+2a^2H^2\right)\left(1+2(\Phi-\Psi)\right)+a^2H\left(6\dot{\Phi}-\dot{\Psi}\right)+a^2\ddot{\Phi}\right.\\ &\quad\left.-\nabla^2\Phi\right]-\partial_i\partial_j(\Phi+\Psi), \end{aligned}$$

and the scalar curvature is given by

$$\begin{aligned} R &= 6\left(\frac{\ddot{a}}{a}+H^2\right)-\frac{2}{a^2}\nabla^2\left(\Psi+2\Phi\right)-6H\left(\dot{\Psi}-4\dot{\Phi}\right)+6\ddot{\Phi} && (2.110)\\ &\quad -12\Psi\left(\frac{\ddot{a}}{a}+H^2\right). \end{aligned}$$

2.5.2 Boltzmann equation

In order to find the equations for the perturbations of each species in the Universe (i.e., photons, baryons, dark matter and neutrinos) we use the Boltzmann formalism, see chapter 4 of Dodelson (2003). The Boltzmann equation describes the evolution of distribution function of a particular species, and it is given by

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \frac{\partial f}{\partial t} + \frac{\mathrm{d}x^i}{\mathrm{d}t}\frac{\partial f}{\partial x^i} + \frac{\mathrm{d}p}{\mathrm{d}t}\frac{\partial f}{\partial p} = C[f], \tag{2.111}$$

where $f = f(t, \vec{x}, \vec{p})$ is the distribution function and $C[f]$ are the collision terms which describe the interaction process between species in the Universe (e.g., Compton scattering which is due to the interaction between photons and free electrons). Here we have neglected the angular term $\hat{p}$ in equation (2.111), since that this produce a second order term and we consider only first order terms.

Radiation

The photon distribution can be written as

$$f(t, \vec{x}, p, \hat{p}) = \left[\exp\left(\frac{p}{T(t)\,[1 + \Theta(t, \vec{x}, \hat{p})]}\right) - 1\right]^{-1}, \quad \Theta = \frac{\delta T}{T}. \tag{2.112}$$

The above expression can be expand to first order, thus we have

$$f \approx f^{(0)} - p\frac{\partial f^{(0)}}{\partial p}\Theta, \tag{2.113}$$

where $f^{(0)}$ is the photon distribution in the background which is given by the Bose-Einstein distribution (2.67) with zero chemical potential. The photon distribution is mainly affected by the interaction with free electrons through the Compton scattering

$$e^-(\vec{q}) + \gamma(\vec{p}) \leftrightarrow e^-(\vec{q}\,') + \gamma(\vec{p}\,'), \tag{2.114}$$

thus the collision term is given by this interaction. The equation for the temperature perturbation of photons Θ by using the Boltzmann equation in Fourier space is given by

$$\Theta' + ik\mu\Theta = -\Phi' - ik\mu\Psi - \tau'\left[\Theta_0 - \Theta + \mu\, v_b - \frac{1}{2}\mathrm{P}_2(\mu)\Theta_2\right], \tag{2.115}$$

where $'$ indicates the derivative over conformal time. μ is defined as the cosine of the angle between the wavenumber $\vec{k}$ and the photon direction $\hat{p}$:

$$\mu \equiv \frac{\vec{k} \cdot \hat{p}}{k}. \tag{2.116}$$

v_b is the baryon velocity which is defined as irrotational, so $\vec{v_b} = v_b \hat{k}$. τ is the optical depth and it is defined as

$$\tau(\eta) \equiv \int_{\eta}^{\eta_0} \mathrm{d}\eta' n_e \sigma_T a, \tag{2.117}$$

where n_e is the number density of free electrons and σ_T is the Thomson cross-section. The optical depth characterizes the Compton scattering experienced by electrons due to their interactions with photons. At late times, the electron density is small, so $\tau \ll 1$, while at early times, it is very large. $\Theta_l(k, \eta)$ for $l = \{0, 2\}$ are called monopole and quadrupole respectively, in general the photon perturbations are characterized by the lth multipole moments of the temperature field. These are defined as

$$\Theta_l(k, \eta) \equiv \frac{1}{(-i)^l} \int_{-1}^{1} \frac{\mathrm{d}\mu}{2} \mathrm{P}_l(\mu) \Theta(k, \mu, \eta), \tag{2.118}$$

where P_l is the Legendre polynomial of order l. The higher moments describe the small scale structure of the temperature field. Here we drop the polarization terms in equation (2.115).

For neutrinos, we have a similar equation to that of photons, but there is no collision term because neutrinos only interact weakly. The distribution function for neutrinos is given by

$$f_\nu = f_\nu^{(0)} + \frac{\partial f_\nu^{(0)}}{\partial T_\nu} T_\nu N, \quad N = \frac{\delta T_\nu}{T_\nu}, \tag{2.119}$$

where $f_\nu^{(0)}$ is the Fermi-Dirac distribution (2.68) with zero chemical potential. The equation for N, using the Boltzmann equation in Fourier space is given by

$$N' + ik\mu \frac{p}{E} N + \Phi' + ik\mu \frac{E}{p} \Psi = 0. \tag{2.120}$$

Note that for neutrinos with negligible mass, which are relativistic, we have that $E = p$, therefore equation (2.120) can be rewritten as

$$N' + ik\mu N + \Phi' + ik\mu \Psi = 0. \tag{2.121}$$

The above equation is similar as equation (2.115) for Θ, but without the interaction term.

Dark matter

For dark matter it is not necessary to consider the collision term, because dark matter has a tiny and negligible cross-section. In order to obtain the equations which describe the perturbation in the dark matter density field it is necessary to consider the first two moments of the Boltzmann equation, thus we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\int \frac{\mathrm{d}^3p}{(2\pi)^3} f_{\mathrm{DM}}\right) = 0, \tag{2.122}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\int \frac{\mathrm{d}^3p}{(2\pi)^3} f_{\mathrm{DM}} \frac{p\hat{p}^i}{E}\right) = 0, \tag{2.123}$$

since the treatment is linear, we neglect second-order terms in p/E. Therefore, the term $p/E \sim v$ is a linear perturbation. Recall that the dark matter density is defined as

$$n_{\mathrm{DM}} \equiv \int \frac{\mathrm{d}^3p}{(2\pi)^3} f_{\mathrm{DM}} \tag{2.124}$$

and its velocity is given by

$$v^i \equiv \frac{1}{n_{\mathrm{DM}}} \int \frac{\mathrm{d}^3p}{(2\pi)^3} f_{\mathrm{DM}} \frac{p\hat{p}^i}{E}. \tag{2.125}$$

The dark matter density can be written as the sum of the dark matter density in the background and a small perturbation, thus

$$n_{\mathrm{DM}} = n_{\mathrm{DM}}^{(0)} + \delta n_{\mathrm{DM}} = n_{\mathrm{DM}}^{(0)}\left[1 + \delta(t,\vec{x})\right], \quad \delta(t,\vec{x}) = \frac{\delta n_{\mathrm{DM}}}{n_{\mathrm{DM}}^{(0)}}, \tag{2.126}$$

since $\rho_{\mathrm{DM}} = m\, n_{\mathrm{DM}}$ then $\delta = \delta\rho_{\mathrm{DM}}/\rho_{\mathrm{DM}}$ is also identified with the perturbations in the energy density of dark matter. The equations for the dark matter overdensity δ and its velocity $\vec{v}$ are given by

$$\delta' + ikv + 3\Phi' = 0, \tag{2.127}$$

$$v' + \frac{a'}{a}v + ik\Psi = 0, \tag{2.128}$$

by using the Boltzmann equation in Fourier space. Here the velocity is assumed irrotational so $\vec{v} = v\hat{k}$.

Baryons

For baryons (by convention in cosmology, we call to all objects made of normal atomic matter as baryons, although electrons are leptons), the equations for perturbation variables are similar to the equations for dark matter, but here it is necessary to take into account a collision term. Electron and protons are coupled by Coulomb scattering

$$e^- + p \rightarrow e^- + p. \tag{2.129}$$

Since this coupling is strong we can consider that the perturbation variables for electrons and protons have a common value, then

$$\frac{\delta\rho_e}{\rho_e^{(0)}} = \frac{\delta\rho_p}{\rho_p^{(0)}} \equiv \delta_b, \quad \vec{v}_e = \vec{v}_p \equiv \vec{v}_b. \tag{2.130}$$

In order to find the equations for δ_b and $\vec{v}_b$ it is important to note that there are two collision terms involved here in addition to Coulomb scattering: The electrons interacting with photons via Compton scattering and the interaction of protons with photons. The collision term for photons scattering of protons is neglected because it is much smaller than for the Compton scattering (note that the cross section is inversely proportional to the mass squared for the above mentioned processes). Therefore, the interaction between the photons and the set protons and electrons is dominated by Compton scattering of electrons. The equations for perturbation variables by using the Boltzmann equation in Fourier space are given by

$$\begin{aligned} \delta_b' + ikv_b + 3\Phi' &= 0, && (2.131)\\ v_b' + \frac{a'}{a}v_b + ik\Psi &= \frac{\tau'}{R}\left[3i\Theta_1 + v_b\right], \quad \frac{1}{R} \equiv \frac{4\rho_\gamma^{(0)}}{3\rho_b^{(0)}}, && (2.132) \end{aligned}$$

where R is the ratio of photon to baryon density.

2.5.3 Perturbed Einstein equations

Previously we showed the equations for the perturbation variables for each species, here we are going to present the evolution equations for Φ and Ψ. The Einstein equations for perturbations can be written as

$$^{(0)}G^{\mu}_{\ \nu} + \delta G^{\mu}_{\ \nu} = 8\pi G \left(^{(0)}T^{\mu}_{\ \nu} + \delta T^{\mu}_{\ \nu} \right), \tag{2.133}$$

where the background equation allows us to obtain the Friedmann equations, so now we are interested in the perturbed part

$$\delta G^{\mu}_{\ \nu} = 8\pi G \delta T^{\mu}_{\ \nu}. \tag{2.134}$$

Considering the time-time component from equation (2.134) in Fourier space, we obtain the following equation:

$$k^2\Phi + 3\frac{a'}{a}\left(\Phi' - \Psi\frac{a'}{a}\right) = 4\pi G a^2 \left[\rho_{\rm DM}\delta + \rho_b\delta_b + 4\rho_\gamma\Theta_0 + 4\rho_\nu N_0\right]. \tag{2.135}$$

Note that on the right hand side of the above equation, the factor 4 which appears for photons and relativistic neutrinos is due to the relation $\rho \propto T^4$, which implies that $\delta\rho/\rho = 4\delta T/T$. Considering now the longitudinal traceless part from space components of equation (2.134), which can be extracted by contracting this equation with the projection operator $\hat{k}_i\hat{k}^j - (1/3)\delta_i^j$, we obtain the following equation

$$k^2(\Phi + \Psi) = -32\pi G a^2 \left[\rho_\gamma\Theta_2 + \rho_\nu N_2\right]. \tag{2.136}$$

The right hand side in the above equation is known as the anisotropic stress. Note that non-relativistic particles do not contribute to this term. In practice it can neglect the anisotropic stress, we have $\Phi = -\Psi$. This argument is valid for photons due to strong coupling with the electrons via Compton scattering, nonetheless neutrinos develop anisotropic stress after their decoupling. Therefore in the radiation dominated era this contribution is relevant, but in the matter dominated era the neutrino contribution becomes unimportant and the anisotropic stress can be neglected. With equation (2.135) and equation (2.136) we close the differential equation system for the perturbation variables. However, from time-space components of equation (2.134) and equation (2.135) we obtain the following equation

$$k^2\Phi = 4\pi G a^2 \left[\rho_m\delta_m + 4\rho_r\Theta_{r,0} + \frac{3aH}{k}\left(i\rho_m v_m + 4\rho_r\Theta_{r,1}\right)\right], \tag{2.137}$$

where we define

$$\rho_m\delta_m \equiv \rho_{\rm DM}\delta + \rho_b\delta_b, \qquad \rho_r\Theta_{r,0} \equiv \rho_\gamma\Theta_0 + \rho_\nu N_0, \tag{2.138}$$

$$\rho_m v_m \equiv \rho_{\rm DM}v + \rho_b v_b, \qquad \rho_r\Theta_{r,1} \equiv \rho_\gamma\Theta_1 + \rho_\nu N_1. \tag{2.139}$$

Note that for $k\eta \gg 1$, in the sub-horizon limit, equation (2.137) is reduced to the Poisson's equation. In summary, the equations for the perturbation variables are

$$\begin{aligned}
\Theta' + ik\mu\Theta + \Phi' + ik\mu\Psi &= -\tau'\left[\Theta_0 - \Theta + \mu\, v_b - \frac{1}{2}\mathrm{P}_2(\mu)\Theta_2\right], \qquad (2.140)\\
N' + ik\mu N + \Phi' + ik\mu\Psi &= 0,\\
\delta' + ikv + 3\Phi' &= 0,\\
v' + \frac{a'}{a}v + ik\Psi &= 0,\\
\delta_b' + ikv_b + 3\Phi' &= 0,\\
v_b' + \frac{a'}{a}v_b + ik\Psi &= \frac{\tau'}{R}\left[3i\Theta_1 + v_b\right],\\
k^2\Phi + 3\frac{a'}{a}\left(\Phi' - \Psi\frac{a'}{a}\right) &= 4\pi G a^2\left[\rho_m\delta_m + 4\rho_r\Theta_{r,0}\right],\\
k^2(\Phi + \Psi) &= -32\pi G a^2\rho_r\Theta_{r,2}.
\end{aligned}$$

2.6 Large-scale structure

In order to understand the evolution of matter perturbations we need to solve the differential equations set out in the above section. It is not possible to find a general analytical solution for those equations, but there are public codes which allow us to obtain numerical solution. Here we use CAMB by Lewis et al. (2000). There are two interesting limits to study the perturbation variables: Super-horizon fluctuations $k\eta \ll 1$ where the modes are far outside the horizon, and sub-horizon fluctuations $k\eta \gg 1$ where the modes are well within the horizon.

2.6.1 Power spectrum of matter

The evolution of the matter perturbations is characterized by the same equations given in the above section, however we must define the initial condition of such perturbations. We assume that the early Universe had a period of inflation in which it experienced an accelerated expansion, it was cooled and then reheated and this process produced the primordial anisotropies, see Mukhanov (2005). Here we suppose that the above process is adiabatic and that the initial fluctuations given by the primordial anisotropies are Gaussian. We can characterize the matter perturbation field by the two-point function, defined as

$$\xi(r) = \langle \delta(\vec{x})\delta(\vec{x}+\vec{r})\rangle, \quad \text{with} \quad \delta(\vec{x}) \equiv \frac{n(\vec{x})}{\langle n(\vec{x})\rangle} - 1, \tag{2.141}$$

where $\langle \delta(\vec{x})\rangle = 0$. For non-Gaussianities it is necessary to consider higher order correlation functions to characterize the matter fluctuation completely. Nevertheless this is not the case for Gaussian fields. In Fourier space we can write

$$\xi(r) = \int \frac{\mathrm{d}^3 k}{(2\pi)^3} P(k) e^{i\vec{k}\cdot\vec{r}}, \tag{2.142}$$

where $P(k)$ is the power spectrum and it is defined as

$$\langle \delta(\vec{k},\eta)\delta^*(\vec{k}',\eta)\rangle = (2\pi)^3 P(k)\delta_D^3(\vec{k}-\vec{k}'). \tag{2.143}$$

Here δ_D^3 is the Dirac delta function. In this formalism, the power spectrum has dimensions of $[L^3]$, we can also define the dimensionless power spectrum as

$$\Delta^2(k) \equiv \frac{k^3 P(k)}{2\pi^2}. \tag{2.144}$$

It is common to rewrite the power spectrum as

$$P(k,a) = P_0(k)T^2(k)\left(\frac{D(a)}{D(a=1)}\right)^2. \tag{2.145}$$

The first factor on the right hand side of the above equation, $P_0(k)$, is the primordial matter power spectrum given by the initial scalar fluctuations. Commonly, for many inflationary scenarios, this factor obeys the following phenomenological parameterization

$$P_0 = A_s \left(\frac{k}{k_0}\right)^{n_s + \frac{1}{2}\frac{\mathrm{d}n_s}{\mathrm{d}\ln k}\ln\left(\frac{k}{k_0}\right) + \frac{1}{6}\frac{\mathrm{d}^2 n_s}{\mathrm{d}^2\ln k}\ln\left(\frac{k}{k_0}\right)^2 + \ldots}, \tag{2.146}$$

where A_s is the amplitude of the initial fluctuations which are related to the amplitude of the fluctuations seen in the large-scale structure, n_s is the scalar spectral index and the other terms in the argument of the exponential of equation (2.146) are the running and the running of the running of the scalar spectra index, see Planck Collaboration (2016). The second factor on the right hand side of equation (2.145), $T(k)$, is known as the transfer function, describes the effects in the growth of the perturbations from their creation until after recombination (i.e., Meszaros effect, acoustic oscillations, Silk damping, free-streaming damping, radiation drag, see Dodelson (2003); Mukhanov (2005). For a Universe dominated by cold dark matter

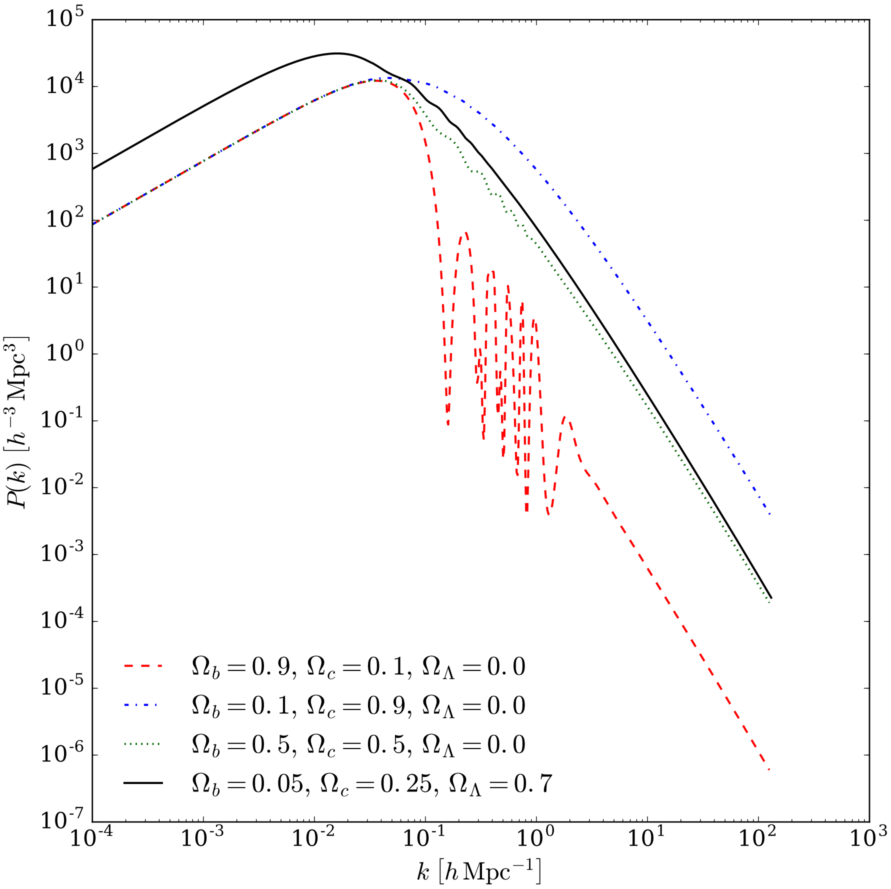

Figure 2.4: Matter power spectrum for a flat Universe with different cosmological parameters at $a = 1$. Note that the baryons include oscillations in the power spectrum for small scales. The dark energy mainly affects the power spectrum in large scale. The spectra are computed by using CAMB (Lewis et al. (2000)).

Bardeen, Bond, Kaiser & Szalay (1986) proposed the following fit for the transfer function

$$T(k) = \frac{\ln(1+2.34q)}{2.34q}\left(1+3.89q+(16.2q)^2+(5.47q)^3+(6.71q)^4\right)^{-1/4}, \qquad (2.147)$$

where

$$q \equiv \frac{k}{\Omega_{\rm DM} h^2 {\rm Mpc}^{-1}}. \tag{2.148}$$

At late times the growth of matter perturbations is independent of scale. The growth factor, $D(a)$, describes the growth of perturbations at this period. Figure 2.4 shows the matter power spectrum for a flat Universe with different cosmological parameters at $a = 1$. Here we consider 4 examples where the first three are cases for a Universe dominated by non-relativistic matter (i.e., baryon dominated, CDM dominated and 50% for baryons and 50% for CDM). We note that the power spectrum is highly depedent on the cosmological parameters, so that it can be used to constrain the model that best describes the Universe. The power spectra are computed by using CAMB, Lewis et al. (2000).

2.6.2 Growth factor of matter perturbations

In order to compute the evolution of matter perturbations at late time, we use equations (2.140) and the equation for the gravitational potential (2.137) in the sub-horizon limit where $k\eta \gg 1$ and we neglect radiation perturbations, then we have

$$\begin{aligned} \delta_m' + ikv_m &= -3\Phi', && (2.149)\\ v_m' + aHv_m &= ik\Phi, && (2.150)\\ k^2\Phi &= 4\pi G a^2 \rho_m \delta_m. && (2.151) \end{aligned}$$

In the equation for the gravitational potential we neglect the velocity term because the perturbations are well within the horizon, therefore $aH/k \ll 1$. With equation (2.149) and equation (2.150) we obtain

$$\delta_m'' - ikv_m aH = k^2\Phi - 3\Phi''. \tag{2.152}$$

The velocity term on the left hand side of the above equation is replaced by using equation (2.149), hence we rewrite equation (2.152) as

$$\delta_m'' + aH\delta_m' = k^2\Phi - 3\Phi'' - 3\Phi' aH. \tag{2.153}$$

Note that both terms $3\Phi''$ and $3\Phi' aH$ are of the order of $\sim 1/\eta^2$, therefore $k^2\Phi - 3\Phi'' - 3\Phi' aH \approx k^2\Phi$. On the other hand using equation (2.151) the evolution of matter perturbations is described by

$$\delta_m'' + aH\delta_m' = 4\pi G a^2 \rho_m \delta_m. \tag{2.154}$$

The above equation is scale independent, therefore in the sub-horizon limit all scales grow at the same rate. We can write equation (2.154) as function of the scale factor by using the relation $\mathrm{d}\eta = \mathrm{d}a/a^2H$, then we have

$$\frac{\mathrm{d}^2D}{\mathrm{d}a^2} + \left(\frac{\mathrm{d}\ln H}{\mathrm{d}a} + \frac{3}{a}\right)\frac{\mathrm{d}D}{\mathrm{d}a} = \frac{3\Omega_m}{2a^5}\left(\frac{H_0}{H}\right)^2 D. \tag{2.155}$$

Here we write the matter perturbations as $\delta_m(\eta, k) = \delta_m(k)D(\eta)/D(\eta = \eta_0)$, where $D(\eta)$ is known as growth factor, and we also use the relation $\rho_m = \Omega_m\rho_{cr}/a^3$. The above expression allows us to find analytic solutions for D in some cases, see Dodelson (2003). It is straightforward to show that the solution in a Universe dominated by matter is $D \propto a$. However, for the general case we have to solve this equation numerically. In order to do this we use the substitution $y = \ln a$, so that equation (2.155) is rewritten as

$$\frac{\mathrm{d}^2D}{\mathrm{d}y^2} + (1-q)\frac{\mathrm{d}D}{\mathrm{d}y} = \frac{3\Omega_m}{2a^3}\left(\frac{H_0}{H}\right)^2 D, \tag{2.156}$$

where q is the deceleration parameter. Figure 2.5 shows the growth factor for different cosmologies. In appendix B we show the Newtonian treatment for matter perturbations in a sub-horizon limit.

2.6.3 Spherical collapse model

So far we have described the linear theory of perturbations where $\delta \ll 1$. However, this theoretical framework is no longer valid when the perturbations are larger and $\delta > 1$. In this case, we cannot describe the evolution of density field through the growth factor $D(a)$ defined above. Therefore, in order to obtain information about the nonlinear regime, we have to consider the model of spherical collapse for dark matter.

We assume a spherical perturbation into a homogeneous Universe dominated by a collisionless fluid (dark matter). We consider the stage of the Universe after the recombination. Previously, we show that a Universe dominated by non-relativistic

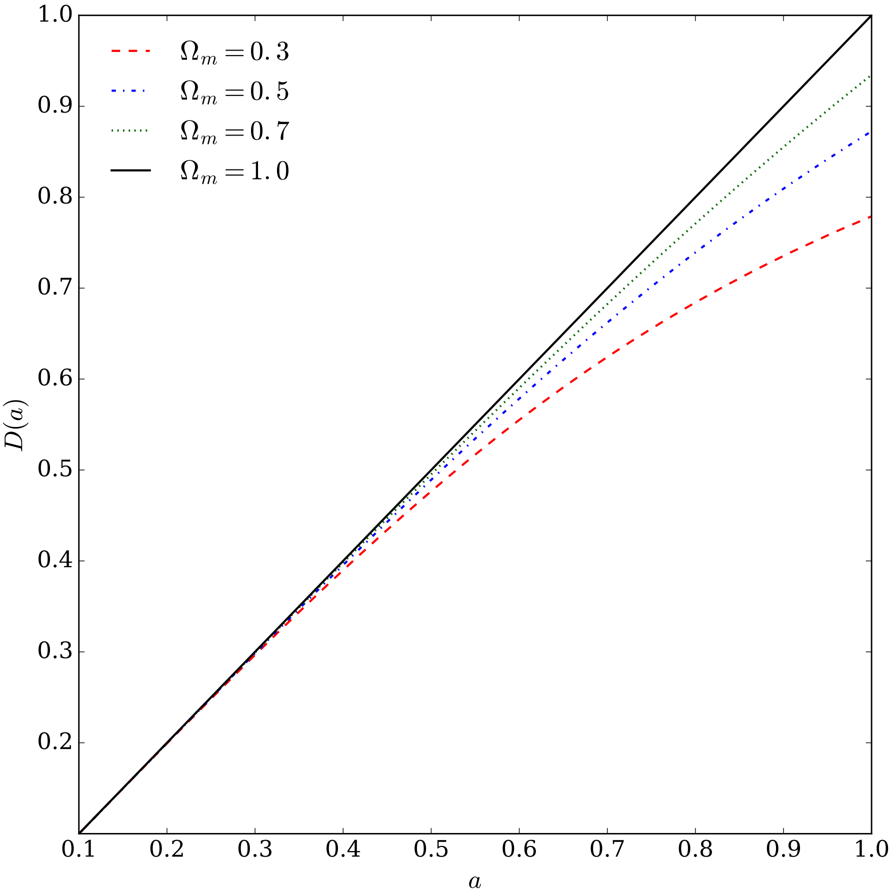

Figure 2.5: Growth factor for a ΛCDM model in a flat Universe without radiation. We consider different values for Ω_m and Ω_Λ. Note that the dark energy slows down the growth of the perturbations.

matter satisfies the following relations

$$a \propto t^{2/3},\ \rho_m(t) = \frac{1}{6\pi G t^2},\ H(t) = \frac{2}{3t},\ D(a) = a \propto t^{2/3}. \tag{2.157}$$

According to Birkhoff's theorem, we can assume a spherical perturbation as a closed Universe. Then, for this "sub-Universe" the first Friedmann equation is given by

$$\left(\frac{\dot{a}_s}{a_s}\right)^2 = H_0^2 \left(\frac{\Omega}{a_s^3} + \frac{(1-\Omega)}{a_s^2}\right), \tag{2.158}$$

where a_s and $\Omega > 1$ are parameters in the spherical perturbation. The parametric solution for the above equation is

$$a_s = A\left(1-\cos\theta\right), \qquad A = \frac{\Omega}{2\left(\Omega-1\right)}, \tag{2.159}$$

$$t = B\left(\theta-\sin\theta\right), \qquad B = \frac{\Omega}{2H_0\left(\Omega-1\right)^{3/2}}, \tag{2.160}$$

for $\theta \in [0, 2\pi]$. We note that the sphere reaches a maximum $a_{s,max} = 2A$ and $t_{max} = \pi B$ for $\theta = \pi$, then it collapses back to $a_{s,min} = 0$ when $\theta = 2\pi$ and $t_c = 2\pi B = 2t_{max}$. The time of maximum size is called turn-around time and this is denoted as $t_{ta} = t_{max}$. By using the critical density definition and mass conservation $M = 4\pi\rho/3$, we show that $A^3 = GMB^2$. The density of perturbation is given by

$$\rho = \frac{3M}{4\pi a_s^3} = \frac{3M}{4\pi A^3}\left(1-\cos\theta\right)^{-3}, \tag{2.161}$$

and the background density can be written as

$$\bar{\rho} = \frac{1}{6\pi G t^2} = \frac{1}{6\pi G B^2}\left(\theta-\sin\theta\right)^{-2}, \tag{2.162}$$

therefore, the overdensity of the spherical perturbation is given by

$$\Delta = 1+\delta = \frac{9\left(\theta-\sin\theta\right)^2}{2\left(1-\cos\theta\right)^3}. \tag{2.163}$$

Here we have used the relation between A and B. Linear theory predicts that the overdensity for a Universe dominated by matter is given by

$$\delta^{lin} \propto D(a) \propto t^{2/3}. \tag{2.164}$$

In order to find the proper initial conditions, we use the parametric solution in the limit $\theta \ll 1$. So that, according to the Taylor series expansion, we have

$$\left(\theta-\sin\theta\right)^2 \approx \frac{\theta^6}{36}\left(1-\frac{\theta^2}{10}\right), \tag{2.165}$$

$$\left(1-\cos\theta\right)^3 \approx \frac{\theta^6}{8}\left(1-\frac{\theta^2}{4}\right). \tag{2.166}$$

Then, the initial overdensity is given by

$$1 + \delta_i = 1 + \frac{3\theta_i^2}{20}, \text{ for } \theta \ll 1, \tag{2.167}$$

therefore, we can assert that $\delta_i = 3\theta_i^2/20$. On the other hand, in this limit we have that $t_i = B\theta^3/6$, moreover we know that $t_{ta} = \pi B$, thus the initial condition is given by

$$\delta_i = \frac{3}{20}(6\pi)^{2/3}\left(\frac{t_i}{t_{ta}}\right)^{2/3}. \tag{2.168}$$

In the linear case, the perturbation can be rewritten as

$$\delta^{lin} = \delta_i\left(\frac{t}{t_i}\right)^{2/3} = \frac{3}{20}(6\pi)^{2/3}\left(\frac{t}{t_{ta}}\right)^{2/3}. \tag{2.169}$$

At the turn-around time the density field for the spherical model is $\delta_{ta} \approx 4.55$ while for the linear case we have $\delta_{ta}^{lin} \approx 1.062$. In the collapse time we note that the spherical model predicts that the density field is infinite, nevertheless in the linear theory we obtain a finite value $\delta_c = 3(12\pi)^{2/3}/20 \approx 1.686$, which is called the critical overdensity for collapse. This parameter sets a limit for identifying the regions in the linear density field which should have collapsed at time t.

In this toy model, the density tends to infinite at the time t_c for the spherical collapse case, yet this does not happen in the Universe. The system relaxes toward virial equilibrium forming objects known as dark matter haloes. The virial theorem sets the following relation between kinetic energy (K) and potential energy (U):

$$K = -\frac{U}{2}. \tag{2.170}$$

Therefore, the total energy of system in virial equilibrium is

$$E_v = K_v + U_v = \frac{U_v}{2} = -\frac{3GM^2}{10R_v}, \tag{2.171}$$

where R_v is the physical size at the virial equilibrium. Here we define $R = a_s X$, where X is the comoving coordinate. According to energy conservation and by using the fact that there is no kinetic energy at turn-around time, we have

$$E_{ta} = U_{ta} = -\frac{3GM^2}{5R_{ta}} = E_v = -\frac{3GM^2}{10R_v}, \text{ so } R_v = \frac{1}{2}R_{ta}. \tag{2.172}$$

Hence, the system is expected to virialize at half its turn-around radius. We note that the average density in the sphere is eight times denser than at turn-around time

$$\rho_v = \frac{3M}{4\pi R_v^3} = \frac{3M}{4\pi (R_{ta}/2)^3} = 8\rho_{ta}. \tag{2.173}$$

If we consider that the virialization occurs at the collapse time (i.e, $t_v = 2t_{max}$), then the overdensity of a virialized dark matter halo is

$$\Delta_v = 1 + \delta_v = \frac{\rho_v}{\bar{\rho}_v} = \frac{8\rho_{ta}}{\bar{\rho}_{ta}/4} = 32(1 + \delta_{ta}) \approx 178. \tag{2.174}$$

Here we use the fact that $\bar{\rho} \propto t^{-2}$ for a Universe dominated by matter. The above result allows us to assert that the collapsed haloes have an average overdensity of around 200 with respect to the mean density of the Universe at that epoch. The overdensity depends on the cosmology. Bryan & Norman (1998) showed that its value is given by

$$\Delta_v \approx (18\pi^2 + 60x - 32x^2)/\Omega_m(t_v), \quad (\Omega_\Lambda = 0), \tag{2.175}$$

$$\Delta_v \approx (18\pi^2 + 82x - 39x^2)/\Omega_m(t_v), \quad (\Omega_\Lambda \neq 0), \tag{2.176}$$

where $x = \Omega_m(t_v) - 1$. Often the value $\Delta_v = 200$ is used in the N-body simulations to identify the virialized dark matter haloes, see Tinker et al. (2008).

2.6.4 Halo abundance

The perturbations in the Universe grow until the linear theory ceases to be valid. According to the spherical collapse model described above, the regions where the linear density field is greater than δ_c should have collapsed to produce dark matter haloes. Here we are going to compute the halo abundance as function of the mass and the redshift. For a volume V with the variance of density field it is defined as

$$\sigma^2 = \langle \delta^2 \rangle = \frac{1}{V} \int \mathrm{d}^3 x \delta^2(\vec{x}). \tag{2.177}$$

By using equation (2.142) we note that

$$\sigma^2 = \xi(0) = \frac{1}{2\pi^2} \int \mathrm{d}k\, P(k) k^2 = \int \mathrm{d}\ln k\, \Delta^2(k). \tag{2.178}$$

Therefore the dimensionless power spectrum is identified with the variance of the field as a function of k. The smoothed density field by a normalized window function $W(\vec{x}; R)$ is defined as a convolution integral

$$\delta(\vec{x}; R) \equiv \int \mathrm{d}^3x\, \delta(\vec{x}')W(\vec{x}-\vec{x}'; R)\,; \qquad \int \mathrm{d}^3x\, W(\vec{x}; R) = 1. \tag{2.179}$$

In Fourier space, we have

$$\delta(\vec{k}; R) = \delta(\vec{k})\tilde{W}(kR); \qquad \tilde{W}(kR) = \int \mathrm{d}^3x\, W(\vec{x}; R)e^{-i\vec{k}\cdot\vec{x}}. \tag{2.180}$$

Here, we use the top-hat filter defined as

$$W(\vec{x}; R) = \begin{cases} \frac{3}{4\pi R^3} & r \leq R, \\ 0 & r > R, \end{cases} \qquad \tilde{W}(kR) = \frac{3}{(kR)^3}\left(\sin(kR) - kR\cos(kR)\right), \tag{2.181}$$

for this filter, the mass is given by

$$M = \frac{4\pi}{3}\bar{\rho}R^3. \tag{2.182}$$

Note that the filter is defined by its radius R or by its mass M. The variance of the smoothed density field is defined as

$$\sigma^2(R) = \langle\delta^2(\vec{x}; R)\rangle = \frac{1}{2\pi^2}\int \mathrm{d}k\, P(k)\tilde{W}^2(kR)k^2. \tag{2.183}$$

Usually, in cosmology, the variance of the linear density field at $z = 0$ filtered by a sphere with radius $R = 8h^{-1}$Mpc is used to characterize the normalization of the power spectrum (i.e., the density field smoothed by a top-hat filter with $R = 8$). This parameter is defined as σ_8 and its current value is ≈ 0.81 according to the observations of Planck Collaboration (2016). Press & Schechter (1974) postulated that the probability of finding perturbations for a smoothed density field with mass M above the threshold δ_c is equal to fraction of haloes with mass greater than M. In addition, they assumed that the density field is a Gaussian random field with variance $\sigma^2(M)$. Hence, we have

$$P(\delta_M > \delta_c) = \frac{1}{\sqrt{2\pi}\sigma_M}\int_{\delta_c}^{\infty} \mathrm{d}\delta_M\, \exp\left(-\frac{\delta_M^2}{2\sigma_M^2}\right). \tag{2.184}$$

According to these assumption only half of all matter in the Universe would be located in haloes, therefore to avoid this problem, the probability is multiplied by an ad hoc factor of 2, thus

$$F(>M,z) = 2P(\delta_M > \delta_c) = \mathrm{erfc}\left(\frac{\nu}{\sqrt{2}}\right), \tag{2.185}$$

where $\mathrm{erfc}(x)$ is the complementary error function and $\nu = \delta_c/\sigma(M,z)$. In order to determinate the number of halos per comoving volume with mass between M and $M + \mathrm{d}M$ we differentiate equation (2.185), we obtain:

$$\mathrm{d}n(M,z) = -\frac{\rho_m}{M}\frac{\mathrm{d}F(>M,z)}{\mathrm{d}M}\mathrm{d}M, \tag{2.186}$$

where the minus sign is due to the fact that $F(>M,z)$ is a decreasing function of the mass M. From a straightforward algebraic manipulation of the above expression, we obtain

$$\frac{\mathrm{d}n}{\mathrm{d}\ln M} = f_{\mathrm{PS}}(\nu)\frac{\rho_m}{M}\frac{\mathrm{d}\ln\sigma^{-1}}{\mathrm{d}\ln M}, \tag{2.187}$$

where

$$f_{\mathrm{PS}}(\nu) = \sqrt{\frac{2}{\pi}}\nu\exp\left(-\frac{\nu^2}{2}\right), \qquad \nu = \frac{\delta_c}{\sigma}. \tag{2.188}$$

Equation (2.187) is called mass function and $f_{\mathrm{PS}}(\nu)$ is a functional factor, here the subindex PS indicates the function obtained through the Press-Schechter formalism. We consider that all mass is contained in halos, then we have

$$\int M\mathrm{d}n = \bar{\rho}_m \quad \text{so} \quad \int \mathrm{d}\nu\frac{f(\nu)}{\nu} = 1. \tag{2.189}$$

The above equation allows us to have a normalization condition for the mass function. Numerical simulations (e.g., Efstathiou et al. (1988)) showed that the PS formalism is was broad in agreement with the mass function computed in such cases, nevertheless currently N-body simulations (e.g., Millennium Simulation, Springel et al. (2005)) have showed that the PS formalism is only an approximation. By using the ellipsoidal collapse approach Sheth et al. (2001) obtained the following functional factor for the mass function

$$f_{\mathrm{SMT}}(\nu) = 2A\left(1 + \frac{1}{\nu'^{2q}}\right)\left(\frac{\nu'^2}{2\pi}\right)^{1/2}\exp\left(-\frac{\nu'^2}{2}\right), \tag{2.190}$$

where $\nu' = \sqrt{a}\nu$, $a = 0.707$, $q = 0.3$, and $A \approx 0.322$ is determined using the condition (2.189). The factor a is determined by numerical simulations, and the parameter q is determined by the shape of the mass function at the low-mass end. Jenkins et al. (2001) find the mass function through a full fitting to N-body simulations, by using Spherical Overdensity (SO) and Friends-of-Friends (FOF) methods in order to detect the haloes.

The number of haloes per comoving volume for a mass bin $[M_\alpha, M_\alpha]$ is obtained integrating equation (2.187), thus we have

$$\bar{n}_\alpha = \int_{M_\alpha}^{M_{\alpha+1}} \mathrm{d}\ln M \frac{\mathrm{d}n}{\mathrm{d}\ln M}. \tag{2.191}$$

The number counts in redshift bin $[z_i, z_{i+1}]$ are given by

$$\bar{N}_{\alpha,i} = \Delta\Omega \int_{z_i}^{z_{i+1}} \mathrm{d}z \frac{\mathrm{d}V_c}{\mathrm{d}z\mathrm{d}\Omega} \bar{n}_\alpha(z), \tag{2.192}$$

where $\Delta\Omega$ is the survey sky coverage and $\mathrm{d}V_c$ is the comoving volume, see equation (2.88).

The halo abundance in the Universe can be used as cosmological probe as we will see in chapter 5. However, given that dark matter only interact with baryons and photons through gravity, we cannot make direct observations of these dark matter haloes. In order to address the above problem, we use the galaxies which are tracers of the dark matter (since the baryons are gravitationally attracted to the structures of dark matter). In chapter 4 we show a optical method for finding the galaxy clusters by using a photometric redshift survey.

Chapter 3

Degradation analysis in the estimation of photometric redshifts

The cosmological redshift of an extra-galactic object is arguably one of the most important directly observable properties. The redshift of an object provides a measure of the recessional velocity of that object relative to an observer, which arises due to the expansion of the Universe. In General Relativity, knowledge of the redshift of an object allows one to connect the spatial and time-dependent components of the space-time metric. In other words, at any given epoch, knowledge of the redshift allows one to relate the spatial extent of the Universe to the current expansion rate of the Universe. A cosmological model provides us with a prediction of how to accurately translate between the redshift of an object and the physical distance to that object, known as the distance-redshift relation, Hubble (1929), see chapter 2. A precise measurement of this relation would allows us to place tight constraints on cosmological parameters and therefore on our fundamental understanding of cosmology, see Riess et al. (1998); Perlmutter et al. (1999). This is a major goal of future cosmological missions that aim to make high precision measurements of *cosmological probes*; including measurements of Baryon Acoustic Oscillations (BAO, Hu & Dodelson (2002); Eisenstein et al. (2005); Percival et al. (2007); Blake et al. (2011); Anderson et al. (2014)), the weak lensing of galaxies (Massey, Kitching & Richard (2010); Bartelmann (2010); Kilbinger et al. (2013)) and the number counts of galaxy clusters (Battye & Weller (2003); Mantz et al. (2010); Rozo et al. (2010); Allen et al. (2011); Mana et al. (2013)). In addition, many of the astrophysical processes governing the growth of the large-scale structure and the formation of galaxies show a strong time-dependence and so measurement of the redshifts of different types of galaxies allows us to test our theories of structure formation.

The redshift of a galaxy can be measured in two ways: either *spectroscopically* or *photometrically*. Spectroscopic determination of redshift involves measuring the Doppler shift of known features in the spectrum of a galaxy, typically absorption or emission lines. Photometric determination of redshift is based upon the assumption that the colors of a population of galaxies of the same type and redshift (i.e. with very similar SEDs) will be clustered in a particular region of the color space. One can therefore estimate the photometric redshift of a galaxy by using multi-band photometry to compare the broad-band colors of that galaxy with the colors of set of galaxies for which redshifts are already known, see Benitez (2000); Collister & Lahav (2004); Ilbert et al. (2006); Almosallam et al. (2016b); Sadeh et al. (2016). Since the measurement of the spectrum of a galaxy is much more costly, due to the requirement of long integration times, photometric redshifts provide a cheaper and much more rapid alternative. Therefore, photometric redshifts are a viable and efficient option to be used in cosmological surveys that plan to observe several billion galaxies, including as the Dark Energy Survey (DES)[*], the Large Synoptic Survey Telescope (LSST)[†], the Euclid[‡] and the Wide Field Infrared Survey Telescope (WFIRST)[§]; photometric redshifts are the most viable option. Note that the Euclid and WFIRST missions will additionally measure spectroscopic redshifts for a subset of galaxies. The major challenge that these surveys face is the problem that photometric redshifts are much less precise than spectroscopic redshifts and will need considerable calibration.

We can split the photometric techniques into two approaches: machine learning and template fitting. Machine learning involves using machine learning methods (MLMs) to establish the relationship between the photometric observables (e.g. colors or magnitudes) and the redshift of a galaxy. This is usually done by training these methods on dataset of galaxies with known redshifts. Among these methods we have the artificial neural networks (ANNs) (Firth et al. (2003); Vanzella et al. (2004); `ANNz` Collister & Lahav (2004); `ANNz2` Sadeh et al. (2016)); nearest-neighbour (Ball et al. (2008)), random forest (`TPZ` Carrasco Kind & Brunner (2013)), and Gaussian process (GPs) (Way et al. (2009); Bonfield et al. (2010); Almosallam et al. (2016a); `GPz` Almosallam et al. (2016b)). The effectiveness of these methods depends on whether the training set is a representative sample of the photometric

[*] ⟨http://www.darkenergysurvey.org⟩

[†] ⟨http://www.lsst.org⟩

[‡] ⟨http://sci.esa.int/euclid/⟩

[§] ⟨https://wfirst.gsfc.nasa.gov⟩

dataset. Moreover, the MLMs are reliable for the redshift range of the used training data set. Therefore, in principle, those methods cannot be employed to estimate high redshifts in which there are no spectroscopic data available. Template methods are based on fitting empirical or synthetic galaxy spectra with the photometric information available (i.e., colors or magnitudes). Specifically, they use the broadband photometry to estimate an approximate galaxy spectral energy distribution (SED), which they then fit against a library of SEDs with known redshifts. Those methods require astrophysical effects, e.g., the dust extinction in the observed galaxy or in our galaxy to be corrected for. A non exhaustive list of codes known for template fitting methods are `HYPERZ` Bolzonella et al. (2000); `ZEBRA` Feldmann et al. (2006); `EAZY` Brammer et al. (2008) and `LE PHARE` Ilbert et al. (2006). Both techniques to estimate photometric redshifts have advantages and limitations depending on the spectroscopic data available and the photometric data set to being evaluated. Abdalla et al. (2008); Hildebrandt et al. (2010); Abdalla et al. (2011); Sánchez et al. (2014) have compared different photometric redshift techniques and their efficiency in ground and space data.

The objective of this work is to analyze the degradation of both precision and accuracy in the estimated photometric redshift for several samples obtained from a mock catalog (constructed using a semi-analytical model of galaxy formation, see section 3.1.1) with a non-representative training data set in magnitude space, we then use these results to guide our choices and perform the same analogies in real data. Here we use the `ANNz2` and `GPz` algorithms, which belong to group of machine learning techniques. We also perform an analysis on the impact of bias in the detection of galaxy cluster with photometric redshifts estimated by using these non-representatives training sets.

We organize this chapter as follows: In section 3.1 we present the mock catalog, and the Sloan Digital Sky Survey (SDSS) and Galaxy and Mass Assembly (GAMA) surveys which are used in this analysis. Section 3.2 describes the `ANNz2` and `GPz` algorithms used in this work and introduce the metrics used to assess the quality of the derived photometric redshifts. Both of these algorithms output for each galaxy a single redshift estimate as well as a redshift probability distribution function (PDF). As such we also introduce two estimators to additionally compute the photometric redshifts using the full PDF information. Section 3.3 compares, for both the mock catalog and observed datasets, the quality of the derived photometric redshifts obtained using the `ANNz2` and `GPz` algorithms and examine the impact of building our training set using either magnitude-space or color-space selection criteria. We then apply sequentially deeper r-band magnitude cuts to the mock catalog in order to analyze the degradation in the quality and completeness of the derived photometric redshifts when the testing set extends to r-band magnitudes significantly deeper than the training set. In section 3.4 we discuss the impact on the detection of galaxy clusters. Section 3.5 summarizes our conclusions. This work is presented in the paper Rivera et al. (2017a) (I-prep.).

3.1 Data

In order to verify the robustness of our results, we use both simulated and real data in this work. Simulated galaxies are from a lightcone mock catalog constructed from a galaxy formation model. By using this mock data set, we can measure the precision and accuracy of the estimated photometric redshifts, as well as performing additional cluster detection completeness tests. To check our results against observations, we apply our methods to a galaxy sample obtained from the SDSS DR12 data release. We tailor a photometric GAMA-like sample going to deeper magnitudes, and train our photometric redshifts with the GAMA survey. This allows us to perform a comparison of photometric redshift results obtained with the mock catalog for a depth comparable to the one in real data. Here we describe the data in more detail.

3.1.1 Mock galaxy catalog

The mock catalog[¶] used in this work was constructed using the lightcone construction method presented in Merson et al. (2013). Briefly, this method involves populating the dark matter halo merger trees extracted from a cosmological N-body simulation with galaxies generated from a *semi-analytical* galaxy formation model. In this case, the merger trees were taken from the Millennium Simulation (Springel et al. (2005)) and populated using the Lagos et al. (2012) version of the `GALFORM` model, which was originally developed by Cole et al. (2000). A lightcone catalog is then constructed by interpolating the galaxy positions between the simulation redshift snapshots to determine when each galaxy crosses the past lightcone of the observer. For further details we refer the reader to Merson et al. (2013). The cosmology used in the Millennium Simulation is a ΛCDM model (Ω_m, Ω_Λ, Ω_b, $h =$

[¶] We use the SDSS_500_photoz catalog available from ⟨http://astro.dur.ac.uk/~d40qra/lightcones/SDSS/⟩.

Table 3.1: Values for the characteristic magnitude ($m^{\star}_{X}$), the normalization coefficient ($\sigma^{\star}$), the bright magnitude slope (γ_o) and faint magnitude slope (γ_s) used to compute photometric noise in each photometric band (X) in the SDSS mock data. See text in section 3.1.1 for details. The magnitude limit for the u-band is from Zou et al. (2015) and the magnitude limits for the g-band, r-band, i-band and z-band are from Raichoor et al. (2016).

X	$m^{\star}_{X}$	$\sigma^{\star}$	γ_o	γ_s
u	22.03	0.2	-0.1	0.25
g	23.10	0.2	-0.1	0.25
r	22.70	0.2	-0.1	0.25
i	22.20	0.2	-0.1	0.25
z	20.70	0.2	-0.1	0.25

0.25, 0.75, 0.045, 0.73), with parameters consistent with the first year results from the Wilkinson Microwave Anisotropy Probe (Spergel et al. (2003)).

The lightcone catalog spans the redshift range $z = 0.0$ to $z = 3.0$ and has a sky footprint of approximately 500 deg^2, centered on position (RA, DEC) = (303.29 deg, -14.48 deg). An SDSS r-band selection ($r \leq 24$) was applied, yielding a total of 15 823 757 galaxies. The (u, g, r, i, z) magnitudes of galaxies reported in the lightcone are AB apparent magnitudes. For each photometric band, X, the magnitudes are perturbed to introduce photometric noise by randomly sampling from a Gaussian with a mean, m_X, equal the AB apparent magnitude of the galaxy in that band, and with a standard deviation, $\sigma_X(m_X)$, which is defined following the approach described in Jouvel et al. (2009) as,

$$\sigma_X = \begin{cases} 10^{0.4(\gamma_o+1)\left(m_X - m^{\star}_X\right)}, & \text{if } m_X < m^{\star}_X, \\ \frac{\sigma^{\star}}{2.72} \exp\left(10^{\gamma_s\left(m_X - m^{\star}_X\right)}\right), & \text{otherwise,} \end{cases} \tag{3.1}$$

where $m^{\star}_X$ is a characteristic magnitude, $\sigma^{\star}$ is a normalization coefficient and γ_o and γ_s are power-law slopes. The values adopted for these parameters are shown in table 3.1. The power-law used in the case $m_X < m^{\star}_X$ corresponds to brighter fluxes, dominated by object noise, whilst the exponential law in the case $m_X \geq m^{\star}_X$ corresponds to fainter fluxes dominated by sky background noise. For further details see Jouvel et al. (2009). In order to obtain a sample similar to our GAMA/SDSS data set, we apply a further i-band magnitude cut $i < 21$, which leaves a total of 1 876 505 galaxies, with a mean redshift of $z_{\text{mean}} \sim 0.35$.

3.1.2 GAMA survey

The Galaxy and Mass Assembly (GAMA) survey[‖] is a sample of optical spectroscopy data for a low-redshift galaxy population. The survey was designed to investigate the galaxy formation and evolution process, aiming at studying the galaxy distribution on scales of 1 kpc to 1 Mpc, see Driver et al. (2009); Baldry et al. (2010); Driver et al. (2011) and Liske et al. (2015). Observations were performed with the AAOmega spectrograph on the Anglo-Australian Telescope (AAT), covering a sky area of $\sim 286\,\mathrm{deg}^2$ split into five survey regions on the sky, with a total of 238 000 objects. The regions observed were split into three equatorial regions (G09, G12, G15) and two southern sky regions (G02 and G23).

The survey consisted of two phases, with slightly different target selection criteria for each of them. GAMA I refers to data collected during the first three years, while GAMA II refers to the full survey, including all of GAMA I. The first phase extended over the three equatorial regions down to (extinction-corrected) Petrosian magnitude of $r_{\rm petro} < 19.4$ in G09 and G15, and $r_{\rm petro} < 19.8$ in G12. Magnitude cuts and target selection were based on photometry from SDSS and additional infrared bands from the UKIRT (United Kingdom InfraRed Telescope) Infrared Deep Sky Survey (UKIDSS), which were introduced to help improve star-galaxy separation. In the second phase, the three existing equatorial survey regions were enlarged and the two southern regions, G02 and G23, were added. The r-band Petrosian magnitude limit was pushed to $r_{\rm petro} < 19.8$ in all survey regions.

Here we use the public Data Release 2 (DR2). This includes the galaxies from GAMA I of survey region G15 ($r_{\rm petro} < 19.4$) and a subset of G09 and G12 survey regions ($r_{\rm petro} < 19.0$) with a total area of $\sim 144\,\mathrm{deg}^2$ for a total of 70 726 targets with secure redshifts download from the GAMA database. For more details, see Baldry et. al. (2010); Liske et al. (2015). To match to the selection criteria of our photometric sample, we then use the SDSS DR12 CasJobs server[**] to match the GAMA data to a clean sub-sample of SDSS DR12 galaxies with additional "GAMA-like" cuts. Our final spectroscopic sample contains 63 226 objects with $r_{\rm petro} < 19.4$.

[‖] ⟨http://www.gama-survey.org/⟩

[**] ⟨https://skyserver.sdss.org/CasJobs/⟩

3.1.3 SDSS DR12 sample

The photometric data set is obtained from a parent sample downloaded from the Sloan Digital Sky Survey Data Release 12 (SDSS DR12) database. Since we consider the GAMA survey as the spectroscopic training sample, the choice of photometric data is performed by using the GAMA target selection cuts in the SDSS DR12 according to Christodoulou et al. (2012). Here we consider two cases for our analysis. In the first case we use the magnitude and color cuts, such that the training set is a fully representative in the magnitude space. This sample is called GAMA MAIN. In the second case we use a deeper magnitude limit keeping a fully representative training set in color space. Hence, the training is performed with 4 colors unlike the above case where we use 5 magnitudes (we use `dered_modelMag` as magnitudes and `modelMagErr` as magnitude errors). This new training criterium is chosen as to ignore the non-representativeness of the training in magnitude space (i.e., the lack of coverage of the training set in the r-band). This sample is called GAMA DEEP. See Moraes et al., 2017 (in prep.) for more details about these samples.

In order to estimate the photometric redshift in the above photometric samples we use the `ANNz2` and `GPz` codes which are described in section 3.2.1 and section 3.2.2 respectively. We use the same training set, validation and test for GAMA MAIN sample (with magnitudes), as well as for GAMA DEEP sample (with colors). Section 3.3 shows the parameters used in each code to estimate the photometric redshift. Section 3.3.1 presents the results of metrics for the testing data set by using both the magnitude space and the color one. Furthermore we perform a comparison between the results obtained for mock galaxy catalog and the testing data.

3.2 Estimating photometric redshifts

In order to estimate the photometric redshifts for galaxies in the GAMA and SDSS surveys and as well as the mock catalogs, we use the `ANNz2` (Sadeh et al. (2016)) and `GPz` (Almosallam et al. (2016b)) public photometric redshift algorithms. These codes apply a set of machine learning methods, using a set of training redshifts to estimate the value of redshift for galaxies without spectroscopic information from their photometry. We briefly describe the `ANNz2` and `GPz` codes.

3.2.1 ANNz2

`ANNz2`[††] (Sadeh et al. (2016)) is a updated version of the original `ANNz` package developed by Collister & Lahav (2004), which used artificial neural networks (ANNs) to estimate the photometric redshifts of galaxies. Given a training set of galaxies, `ANNz2` combines different machine learning techniques (i.e., artificial neural networks, boosted decision/regression trees, among others) to compute a photometric redshift probability distribution function (PDF) for each galaxy in the testing set. The machine learning methods (MLMs) employed are implemented in the TMVA package (Hoecker et al. (2007)).

Like all MLMs, the `ANNz2` code requires training and validation samples from a spectroscopic redshift survey. During each step of the training, the validation sample is used to estimate the convergence of the solution. Once the mapping is established, an independent testing set (i.e., an independent subsample from the spectroscopic redshift survey with photometric information) is used to evaluate the performance of the trained MLM. The methods implemented in this code allow us to optimise the photometric redshift reconstruction, and to estimate their associated uncertainties, which helps mitigate possible problems of non-representativeness. To correct for inaccuracies due to unrepresentative training sets, the `ANNz2` algorithm can use training weights. This method aims to match the distribution of the inputs from the training sample with the testing data following the approach presented in Lima et al. (2008). If the training data set is incomplete (i.e., there are some regions of the input phase-space where the evaluated sample has no corresponding objects for training), this code provides a quality flag, which indicates when unrepresented data are being evaluated.

In order to estimate the photometric redshift PDFs for galaxies, the `ANNz2` algorithm follows an approach called randomised regression, which ranks the different solutions according to their performance based upon the values of various metrics (i.e., bias, scatter, level of outliers). The entire set of solutions is used to construct the photometric redshift PDF. Initially, each solution is folded with a distribution of uncertainty values computed via the *K-nearest neighbours* (KNN) method, see Oyaizu et al. (2008). Later, the solutions are combined with a weighting scheme.

[††] ⟨https://github.com/IftachSadeh/ANNZ⟩

3.2.2 GPz

`GPz`[‡‡] (Almosallam et al. (2016b)) is a machine learning approach which uses sparse Gaussian processes (GPs) to estimate a photometric redshift and its variance. The GPs are probabilistic models for regression. The computational cost can be very high, which can make it impractical - the cost depends of inverting an $n \times n$ covariance matrix for a training sample with n components. Different authors have proposed several techniques in order to reduce this problem. Zhang et al. (2005) showed that in some cases the covariance matrix could have a Toeplitz structure which would relieve the cost in the inversion. Tsiligkaridis & Hero (2013) decomposed the covariance matrix as a sum of Kronecker products to simplify the computation of the inverse. However, these techniques cannot always be applied.

Another approach to solve the computational cost is to reduce the size of the covariance matrix by using sparse approximations, such that a set of $m \ll n$ samples to obtain the covariance matrix, see Candela & Rasmussen (2005). The `GPz` method uses *basis function* (BFM) models, which is classified as a sparse GP method. The BFM is a non-linear function set $\Phi(\mathbf{x}_i) = \{\phi_1(\mathbf{x}_i), ..., \phi_m(\mathbf{x}_i)\}$, where $\mathbf{x}_i$ are the vector of inputs (i.e., magnitudes or colors) for the training sample with n objects. The set of target outputs y_i are generated by a linear combination of $\Phi(\mathbf{x}_i)$, then

$$y_i = \Phi(\mathbf{x}_i)\mathbf{w} + \epsilon_i, \tag{3.2}$$

where ϵ_i is an additive noise, $\mathbf{w}$ are the weights, or the parameters of the model, and $m \ll n$. This method chooses the radial basis function (RBF) kernel as basis functions, which are defined as:

$$\phi_j(\mathbf{x}_i) = \exp\left(-\frac{1}{2}\left(\mathbf{x}_i - \mathbf{p}_j\right)^T \Gamma_j^T \Gamma_j \left(\mathbf{x}_i - \mathbf{p}_j\right)\right), \tag{3.3}$$

where $\mathbf{p}_i$ are the set of basis vectors associated with the basis functions and $\Gamma_j^T\Gamma_j$, are the covariance matrices (or bespoke precision matrices) associated with each basis function. The code allows us different modes for different cases of the covariance matrix,

- `GPVC`: Covariance matrix for each basis function (GP with variable covariance).

- `GPGC`: The same covariance matrix in all basis functions (GP with global co-

[‡‡] ⟨https://github.com/OxfordML/GPz/blob/master/python/demo_photoz.py⟩

variance).

- GPVD: Diagonal covariance matrix for each basis function (GP with variable diagonal covariance).
- GPGD: The same diagonal covariance matrix for all basis functions (GP with global diagonal covariance).
- GPVL: The covariance matrix is given by $\Gamma_j = \mathbf{I}\gamma_j$, where γ_j is a scalar for each basis function (GP with variable length-scales).
- GPGL: The same covariance matrix which is given by $\Gamma_j = \mathbf{I}\gamma_j$, where γ_j is a scalar, for all basis functions (GP with global length-scale).

By using the y_i values from training sample, the basis functions and Bayesian statistic, the method infers the values of parameters $\bar{\mathbf{w}}$. With these ingredients, the method sets the predictive distribution for a test sample x_*, which is given by

$$p(y_*|y) = \mathcal{N}(\mu_*, \sigma_*^2), \tag{3.4}$$

$$\mu_* = \Phi(x_*)\bar{\mathbf{w}}, \tag{3.5}$$

$$\sigma_*^2 = \nu_* + \beta^{-1}, \tag{3.6}$$

where $\mathcal{N}(\mu, \sigma^2)$ is a normal distribution, ν is the model variance and β^{-1} is the noise uncertainty due to the data. Note that in this method the variance is an input-dependent function and it is composed of two terms for different sources of uncertainty, the intrinsic uncertainty about the mean function due to data density and the uncertainty due to the intrinsic noise or the lack of precision in the training set. This information is very useful to identify regions of input space where more data is required, versus areas where additional precision or information is required, such that it is possible to develop strategies to increase the photometric accuracy, either by getting more photometric data in different regions or improving the quality of input data. In this work we use the GPVC mode, which allows us to employ all capacity of GPz code and to obtain robust results.

3.2.3 Metrics

In order to assess the quality of photometric redshifts estimated in this work, we define the following set of metrics commonly employed for this purpose: bias, σ, σ_{68}, FR_e.

The bias measures the deviation of the estimated photometric redshift from the true value (i.e., the spectroscopic redshift).

$$\text{Bias} = \left\langle \frac{z_{\text{phot}} - z_{\text{spec}}}{1 + z_{\text{spec}}} \right\rangle . \tag{3.7}$$

The scatter between the true redshift and the photometric redshift is denoted as σ and given by

$$\sigma = \left\langle \left(\frac{z_{\text{phot}} - z_{\text{spec}}}{1 + z_{\text{spec}}} \right)^2 \right\rangle^{1/2} . \tag{3.8}$$

We define σ_{68}, as

$$\sigma_{68} = \max_{i \in U} \left\{ \left| \frac{z^i_{\text{phot}} - z^i_{\text{spec}}}{1 + z^i_{\text{spec}}} \right| \right\} , \tag{3.9}$$

where U is the set of the 68 percent of galaxies which have the smallest value of $|z_{\text{phot}} - z_{\text{spec}}|/(1 + z_{\text{spec}})$. The catastrophic outlier rate, which we call FR_e, is given by

$$\text{FR}_e = \frac{100}{n} \left\{ i : \left| \frac{z^i_{\text{phot}} - z^i_{\text{spec}}}{1 + z^i_{\text{spec}}} \right| < e \right\} , \tag{3.10}$$

where n is the number of galaxies and e is the outlier threshold. This quantity is the percentage of galaxies with good photometric redshift estimated in the sample for a chosen outlier threshold value. We choose $e = 0.15$.

In order to compare the estimated photometric redshift distribution with the spectroscopic redshift distribution, we also define the chi-square measure D_{χ^2} as

$$D_{\chi^2}(P, Q) = \frac{1}{2} \sum_{i=1}^{n} \frac{[p(i) - q(i)]^2}{p(i) + q(i)} , \tag{3.11}$$

where $P(p_i)$ and $Q(q_i)$ are distribution functions. Note that if the two distributions are different, we obtain a high value for the chi-square measure. Therefore, this metric allows us to determine how close is the distribution obtained from the estimated photometric redshifts to the spectroscopic redshift distribution.

3.2.4 Further photometric redshift estimators

The `ANNz2` and the `GPz` codes provide for each galaxy both an individual redshift estimate as well as a full PDF. We describe here two estimators to extract a single redshift estimate based upon the full PDF information.

By integrating over the full PDF information we can estimate the mean photometric redshift, $z_{\rm phot}$, defined as,

$$z_{\rm phot} = \int z \, {\rm PDF(z)} \, {\rm dz}. \tag{3.12}$$

The corresponding uncertainty is assumed to be Gaussian and can be computed in a similar manner as the square root of the variance. When we apply this estimator to the PDFs from `ANNz2` we shall denote these mean redshifts estimates by AvgPDF-ANNz2. Note that the individual redshifts estimated directly from the `GPz` code already assume a Gaussian uncertainty, and so are already equivalent to equation (3.12). As such we do not need to apply this estimate to the PDFs from `GPz`.

Secondly, we derive an estimate for the photometric redshift for each galaxy by summing the PDF to construct the cumulative distribution function (CDF), which we can randomly sample in a Monte-Carlo process. This process consists in estimating the $z_{\rm phot}$ by using the image of a random number between $[0, 1)$ for the inverse of the cumulative distribution function in each galaxy. With this method we ensure that the redshift estimates are representative of the full underlying PDF information. We expect that the distribution function of the single number redshift obtained through this method is equivalent to the stacked PDF of all galaxies in the data set. Moreover we await to reduce the systematic errors compared with any other photometric redshift estimator according to Wittman (2009).

In summary, we have defined the following two pairs of photometric redshift estimators for this work: AvgPDF-ANNz2 and GPz (both assuming a Gaussian uncertainty); and CDF-ANNz2 and CDF-GPz (both estimated using the Monte-Carlo method).

3.3 Results

Initially in section 3.3.1 we compare the quality of the photometric redshifts obtained from our mock data and those from our real SDSS data. We then compare the quality of the photometric redshifts obtained when our real data is trained using a magnitude-selected training set and when the data is trained using a color-selected training set. Magnitude-selected training and color-selected training are additionally applied to the mock catalog. Subsequently, in section 3.3.2, we select

Table 3.2: The used values for the parameters of the `ANNz2` (**Top**) and `GPz` (**Bottom**) codes.

Code	Parameter	Definition	Value
`ANNz2`	`nMLMs`	Number of MLMs	100
	`minValZ`	Min. value for redshift	0.0
	`maxValZ`	Max. value for redshift	1.0
	`nErrKNN`	Near-neighbours for error	90
	`rndOptTypes`	MLM types	ANN_BDT
	`nPDFbins`	Number of PDF bins	200
`GPz`	`method`	GP method	VC
	`m`	Number of BFM	25
	`heteroscedastic`	Heteroscedastic noise	True
	`csl_method`	Cost-sensitive	Normal
	`maxIter`	Max. of iterations	500
	`maxAttempts`	Max. iterations to attempt	50

sequentially deeper r-band selected samples from the mock catalog to analyze the degradation of the photometric redshifts recovered from each estimator when using a non-representative magnitude-selected training set. Table 3.2 shows the used values for the parameters of the `ANNz2` and `GPz` codes.

3.3.1 Comparison of real data and mock catalog with a complete training set

In this section, we apply both `ANNz2` and `GPz` codes to real data and mock catalogs with two different training choices, using the 5 SDSS magnitude bands in one case and 4 colors in the other. Our aim is twofold: firstly, we want to confirm that our analysis with mock data is qualitatively consistent with the results we obtain in real data. Additionally, we wish to compare color and magnitude types of training and assess their relative performance.

When considering our real data, the GAMA MAIN data will be the testing set when training with a magnitude-selected training set and the GAMA DEEP data will be the testing set when training with a color-selected training set. Details of the construction of the GAMA MAIN and GAMA DEEP samples, which we shall refer to collectively as the GAMA test data, are given in section 3.1.3. For the photometric analysis in the GAMA test data, we take `dered_modelMag` (i.e. SDSS model magnitudes corrected for extinction) as the galaxy magnitudes and `modelMagErr` (error in `modelMag`) as the magnitude errors. Since the magnitude

Table 3.3: Number of galaxies and the threshold r-band for every used subsample in the training of the real data and the mock data. We use the same r-band magnitude range for both datasets.

Sample	Training	Validation	Testing	r-band range
GAMA	20 864	20 865	21 497	$r < 19.4$
Mock	20 220	20 222	200 288	$r < 19.4$

limit in the spectroscopic GAMA dataset is $r_{\mathrm{petro}} < 19.4$ we apply a similar r-band cut to the mock catalog, obtaining a mock training sample of 240 730 galaxies. The `GPz` code provides us a function that allows us to split the spectroscopic GAMA sample and the mock catalog in three subsamples: a training data set, a validation data set and a test data set, the last subsample is used to test the training in each case. Table 3.3 shows the number of galaxies in each subsample for both the real data and the mock data.

The photometric analysis of the GAMA sample and mock catalog sample is performed using magnitudes (u, g, r, i, z) and colors (u-g, g-r, r-i, i-z). For each color $C(m_1, m_2) = m_2 - m_1$ the error on the color are obtained via standard error propagation:

$$\delta C(\delta m_1, \delta m_2) = \sqrt{\delta m_1^2 + \delta m_2^2}, \tag{3.13}$$

where δm_1 and δm_2 are the errors on the magnitudes m_1 and m_2.

In the upper grids of panels in Fig. 3.1 and Fig. 3.2 we compare for each photometric redshift estimator the recovered photometric redshifts to the spectroscopic redshifts of the galaxies. We show the results for both a magnitude-selected training and a color-selected training of the mock catalog and GAMA test data. We compare the corresponding redshift distribution functions. In the lower grid of panels we compare the metrics for each estimator. We note an almost insignificant difference between training with a color-selected training set and a magnitude-selected training set. For both the real data and the mock catalog, we note that the CDF cases show slightly more scatter compared to the AvgPDF-ANNz2 and GPz cases. However, when we examine the metrics we see that for both the mock catalog and the GAMA data sets, over the redshift range $0.1 < z < 0.4$ each photometric redshift estimator yields a bias and a fraction of catastrophic outliers that is in excellent agreement with the other estimators, further there is a good agreement between the mock and the data.

For the GAMA data sets, we see that compared to the two Gaussian estimators the CDF cases are typically able to estimate photometric redshifts out to higher redshifts beyond $z_{\rm phot} \sim 0.4$, though these photometric redshifts have a larger scatter, a larger bias and a greater number of outliers (as indicated by a decreasing value for $\mathrm{FR}_{e=0.15}$). In the mock catalog we observe the similar effect in the CDF-ANNz2 estimator compared with the AvgPDF-ANNz2 estimator. Moreover, the estimators based on Gaussian GPz are also able to recover photometric redshifts out beyond $z_{\rm phot} \sim 0.4$. The quantity $\mathrm{FR}_{e=0.15}$ shows different trends in the data of the mocks, it deviates from 1 at $z \sim 0.5$ in the mock and at $z_{\rm phot} \sim 0.4$ on the data. We also note that for the mocks, the $\mathrm{FR}_{e=0.15}$ values for CDF-GPz and GPz estimators remain close to unity out to $z_{\rm phot} \sim 0.65$.

We also note some sample variance features in both the mock and data (e.g. feature at $z \sim 0.3$ in the data stack). These features disappear with some estimations. The plots of the distribution functions show that those distributions based on the single value which are obtained through the Monte-Carlo method fits better with the stacking of galaxy PDFs than the AvgPDF-ANNz2 and GPz estimators. In fact, the previous assertion is more noticeable in the photometric redshifts estimated by the `ANNz2` code than in the photometric redshifts estimated by the `GPz` code. According to the chi-square measure presented in figure 3.3, for the GAMA case, the CDF-ANNz2 distribution fits better the spectroscopic redshift distribution than the distributions obtained through the other estimators. In the case of the mock, the CDF-ANNz2, GPz and CDF-GPz estimators have similar chi-square measures and their distributions fit better the spectroscopic redshift distribution than the AvgPDF distribution.

Figure 3.3 shows the global metric values of each photometric redshift estimator by using magnitudes and colors for GAMA test data and mock catalog. In order to identify the cases and photometric redshift estimator used here, we employ the following notation AvgPDF-ANNz2 (A1), CDF-ANNz2 (A2), GPz (G1), CDF-GPz (G2). The final letter indicates whether we compute the photometric redshift via magnitude-selected training (m) or color-selected training (c). Furthermore, the bottom panel shows the chi-square measure given by equation (3.11) in each case. We observe that the results obtained by using magnitudes and colors for the mock catalog and the GAMA test data are similar. The scatter and the fraction $\mathrm{FR}_{e=0.15}$ for the photometric redshifts in the mock catalog are overall better than the equivalent metrics for the GAMA test data. It is clear that the mock catalog is unable to properly model a ~ 0.5 per cent catastrophic failure rate and have errors that are slightly too optimistic, though the mock catalog has managed to simulate the overall qualities of the real data. We would expect larger photometric errors in the real data due to additional sources of error not included in the mock catalogs, such as the sky background on a given night or the effects of proximity to bright objects in the sky. This similarity gives us confidence that these mock catalogs are suitable for examining the degradation in the next section. The results obtained from the mock catalog show the same qualitative trends as the results for the GAMA test data, and we therefore claim that using the mock catalog for the performance degradation analysis of the next section is suitable to show any degradation trends that would also be observed in real data.

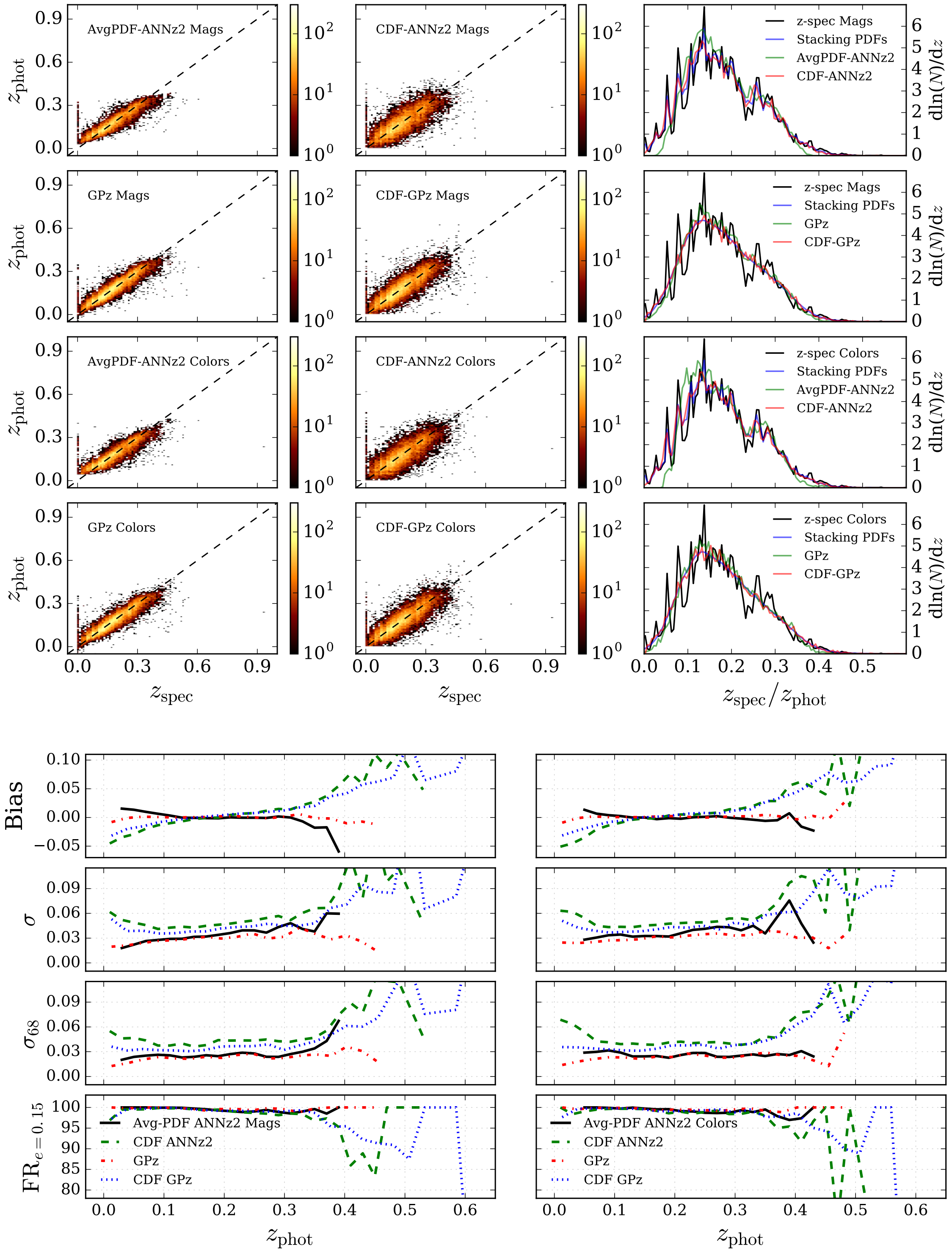

Figure 3.1: Statistical analysis of $z_{\rm phot}$ computed for GAMA survey by using magnitudes and colors. **Top**: The two first columns are the scatter plots $z_{\rm spec}$ against $z_{\rm phot}$ for each photometric redshift estimator. The last column are the $z_{\rm spec}/z_{\rm phot}$ distributions. **Bottom**: Metrics as function of photometric redshift for each estimator (*left*: magnitudes, *right:* colors).

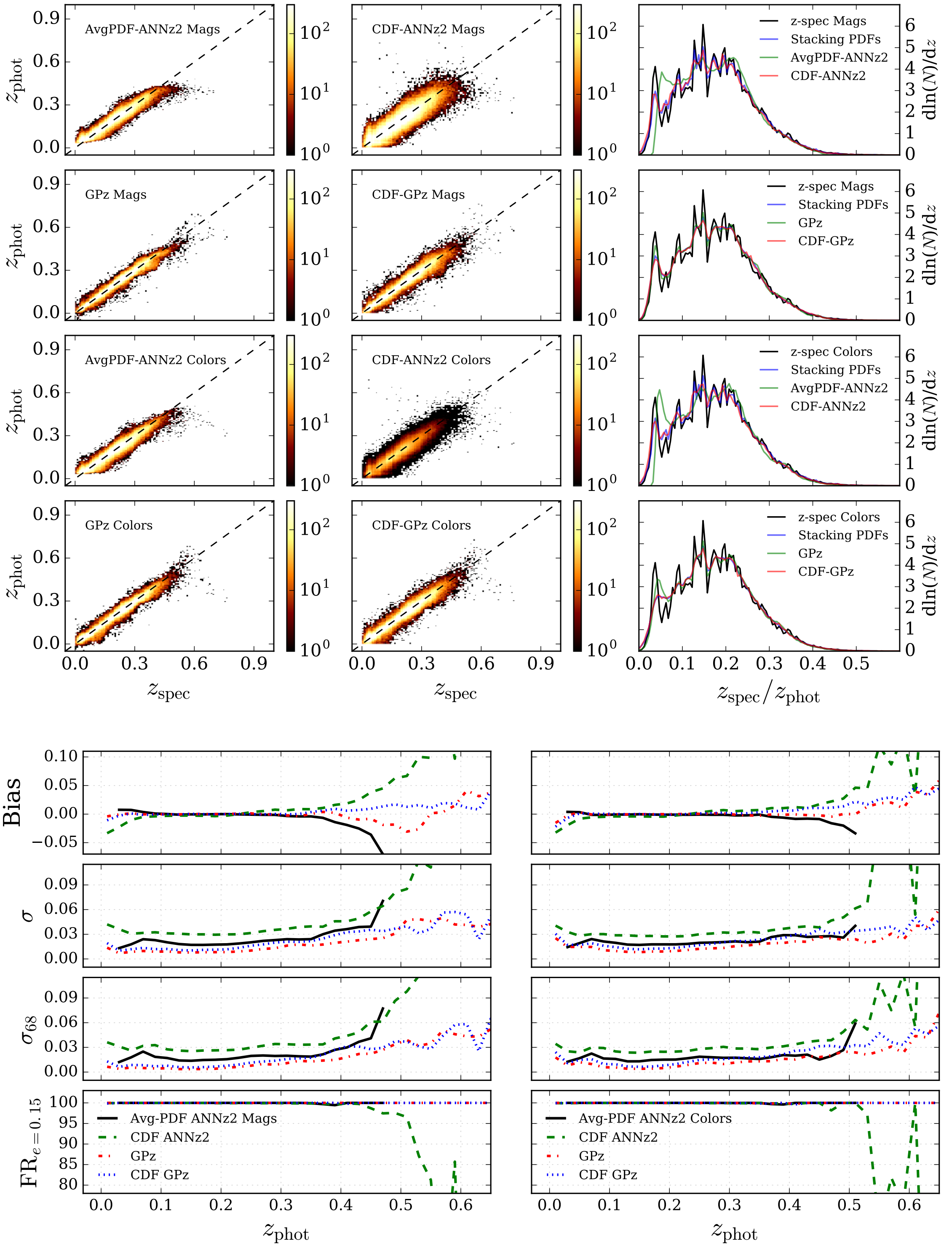

Figure 3.2: Statistical analysis of $z_{\rm phot}$ computed for the mock catalog by using magnitudes and colors. **Top**: The two first columns are the scatter plots $z_{\rm spec}$ against $z_{\rm phot}$ for each photometric redshift estimator. The last column are the $z_{\rm spec}/z_{\rm phot}$ distributions. **Bottom**: Metrics as function of photometric redshift for each estimator (*left*: magnitudes, *right:* colors).

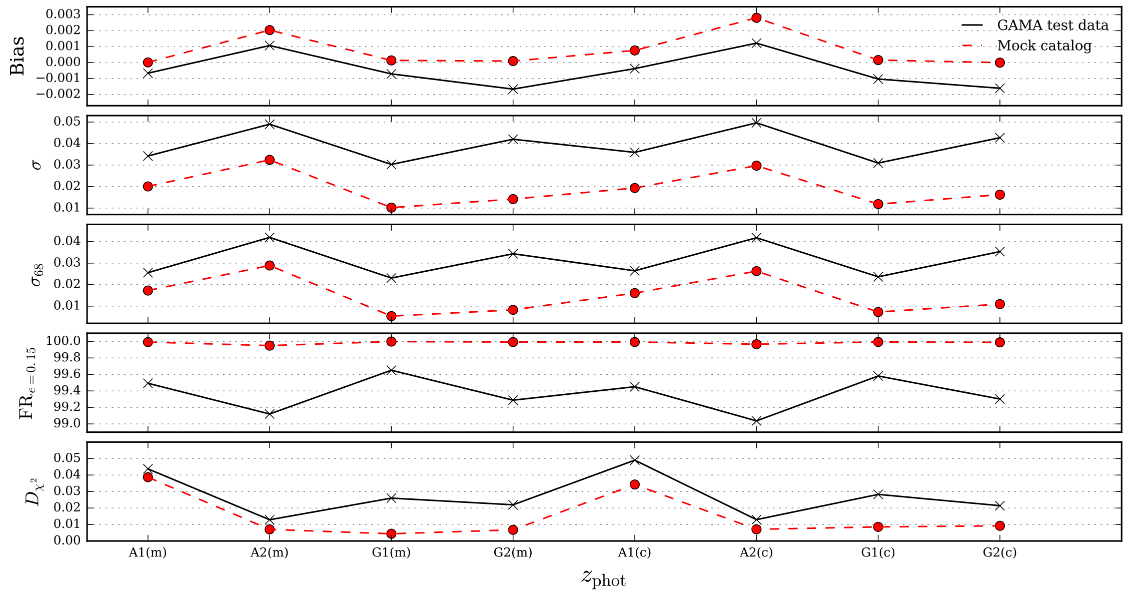

Figure 3.3: Comparison of global metrics for GAMA test data and the mock catalog for each photometric redshift estimator by using magnitudes and colors. Here we use the following notation A1: AvgPDF-ANNz2, A2: CDF-ANNz2, G1: GPz, G2: CDF-GPz. The last letter indicates whether we compute the z_{phot} via the magnitudes (m) or colors (c). The bottom plot is the chi-square measure (D_{χ^2}) which compares the distribution function for every estimator with the spectroscopic distribution function.

3.3.2 Performance degradation

Having established the qualitative equivalence between the observed data and the mock catalogs, we will use the latter to evaluate the performance of the AvgPDF-ANNz2, CDF-ANNz2, GPz and CDF-GPz estimators for a training that is not representative in magnitude space. The idea being that we can safely extrapolate a certain amount in the r-band magnitude given that we have a representative set in color space. More specifically, we define several samples from the mock catalog, by varying the r-band limiting magnitude in the range $[19.4, 20.9]$ in steps of 0.1 magnitudes, i.e. with $\mathrm{dm}_r = 0.1$. Table 3.4 shows the number of objects for each sample used in this analysis. For each testing set we use same the training and validation sets that were used to estimate the photometric redshifts for the previous mock catalog analysis. These training and validation sets are selected from the mock catalog with a magnitude cut of $r < 19.4$. Since the training set is not representative in the magnitude space of the deeper testing sets, we work in color space only to estimate the photometric redshifts. Our goal is to demonstrate that we can obtain

Table 3.4: Number of objects for each cut in the r-band.

Cut of r-band	Number of objects	Cut of r-band	Number of objects
$r < 19.4$	200 288	$r < 20.2$	521 375
$r < 19.5$	228 005	$r < 20.3$	581 154
$r < 19.6$	258 472	$r < 20.4$	647 349
$r < 19.7$	292 526	$r < 20.5$	719 435
$r < 19.8$	330 181	$r < 20.6$	798 152
$r < 19.9$	371 363	$r < 20.7$	884 636
$r < 20.0$	416 572	$r < 20.8$	978 533
$r < 20.1$	466 394	$r < 20.9$	1 079 851

reliable redshift distributions for fainter objects if we ensure representativeness in color space. This can help to mitigate the impact of the non-representativeness problem in the training set of current large-scale structure surveys, where the available spectroscopic data sets are usually shallower than the overlapping photometric surveys.

We estimate the photometric redshifts by applying the same four estimators, as were used in the previous analysis, to the different r-band selected samples. In figure 3.4 we plot, for each sample, the recovered photometric redshifts against the corresponding spectroscopic redshifts. We note that the scatter in the photometric redshift recovery increases with increasing magnitude depth for all methods. Moreover, for fainter flux limits the scatter appears to increase with spectroscopic redshift. This is expected as fainter galaxies will have larger photometric errors and hence higher scatter in the photometric redshift space. On the other hand, we can see that the AvgPDF-ANNz2 estimator is unable to recover photometric redshifts above (i.e. $z_{\rm phot} > 0.5$), an effect that worsens for fainter magnitude cuts. This is also expected due to the nature of the PDF fitting in `ANNz2` and the lack of training galaxies in the sample. This estimator has a higher precision but low accuracy as we tend to fainter magnitudes. Note that the GPz and CDF-GPz estimators also struggle to recover many redshifts beyond $z_{\rm phot} \sim 0.6$. Indeed for every estimator the one-to-one correspondence between spectroscopic and photometric redshift breaks down for redshifts above $z_{\rm phot} \sim 0.6$. For $z > 0.6$ there is significant scatter and bias in the recovered redshifts, particularly in the samples with fainter magnitude selection. Figure 3.5 shows the distribution functions, as a function of limiting magnitude, for each of the photometric redshift estimators that we consider. We note that the CDF-ANNz2 estimator provides a better fit to the spectroscopic redshift distribution than the other photometric redshift estimators in all samples.The distributions from the two GPz estimators show good fit with the spectroscopic redshift distribution for all r-band cuts brighter than $r < 20.0$. For fainter magnitude cuts, the distributions of the estimated photometric redshifts have a peak in $z_{\rm phot} \approx 0.25$ which is not present in the spectroscopic redshift distribution. This peak comes hand in hand with a mismatch at higher redshift. The effect is more prominent for the two `GPz`-based photometric redshift estimators. This peak excess is caused by galaxies that are identified with deeper magnitude selection and have a large spectroscopic redshift, but are estimated to have a smaller photometric redshift. These galaxies can be seen in the lower panels of figure 3.4 as a long tail extending to high spectroscopic redshift. We conclude that the photometric redshift distributions are very similar for magnitude limits brighter than $r < 20.0$.

In cosmological measurements with photometric large-scale structure surveys, much of the information is obtained by splitting the galaxy sample in several photometric redshift bins in order to measure auto- and cross-correlations between the subsamples in the different bins. We are therefore interested in assessing the accuracy of the recovery of the redshift distribution in differential redshift bins. Figure 3.6 compares the stacking of the photometric redshift PDFs estimated through `ANNz2` and `GPz` codes with the spectroscopic distribution for slices of photometric redshift in all r-band magnitude cuts. We consider six photometric redshift bins of width $\mathrm{dz}_{\rm phot} = 0.1$ between $0.0 \leq z_{\rm phot} \leq 0.6$. The selection of galaxies in each redshift slice is performed with the AvgPDF-ANNz2 estimator for the `ANNz2` case and with the GPz estimator for the `GPz` case. Since the specific choice of galaxies in the slices is different for each photometric redshift estimator, we compute the z-spec distribution associated to each algorithm. We observe that the stacking of photometric redshift computed with the `ANNz2` algorithm fits better the spectroscopic distribution than the `GPz` case. In the redshift bins within the range $0.1 \leq z \leq 0.4$, there is good agreement between the stacking for both algorithms and the spectroscopic redshift distribution. However, for deeper cuts the results worsen and the stacking presents differences with the spectroscopic redshift distribution for both cases `ANNz2` and `GPz`. The stacking (GPz) presents the greatest differences with the z-spec distribution in the redshift slices $0.4 \leq z \leq 0.6$.

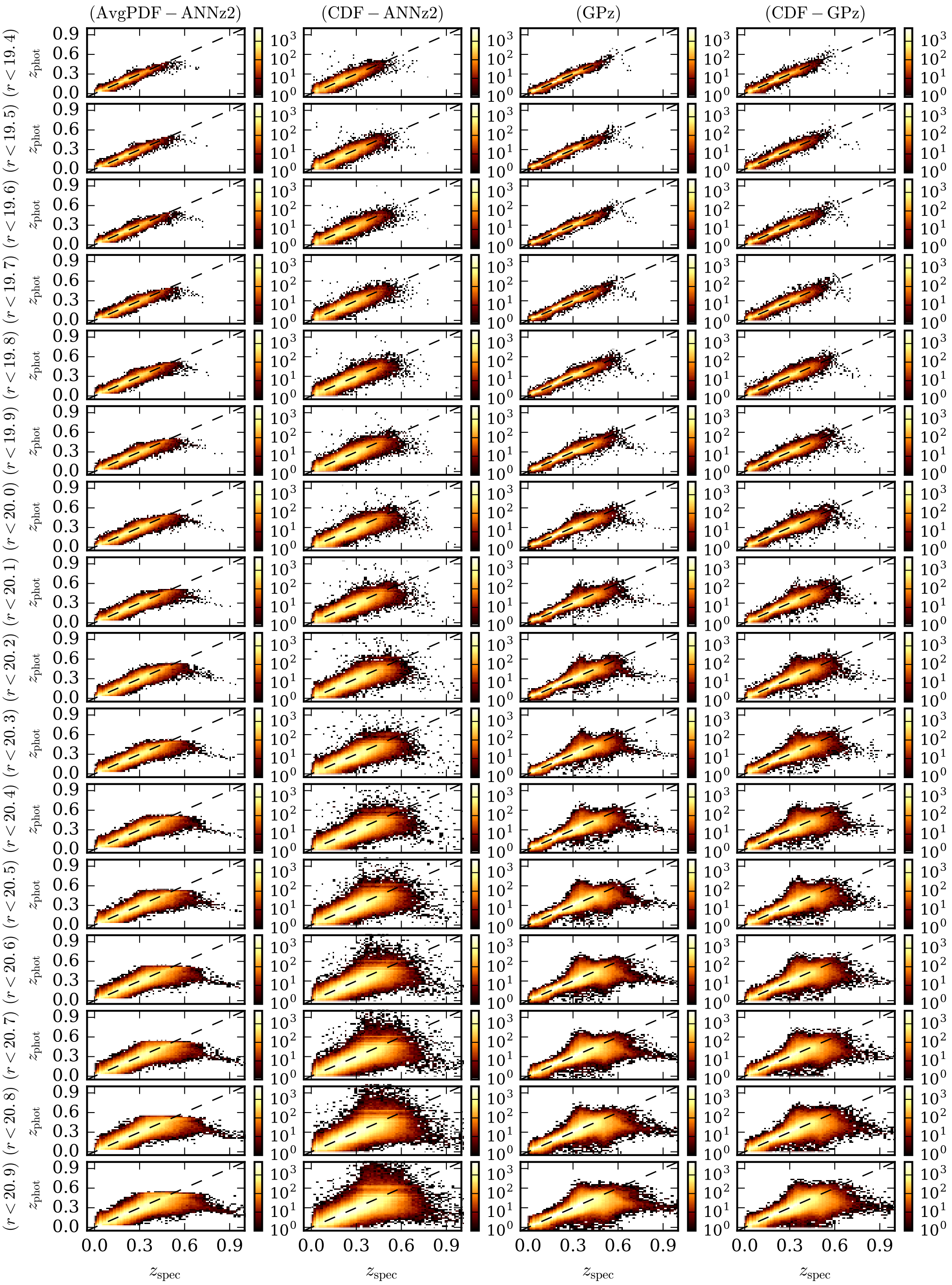

Figure 3.4: Scatter plots of z-spec against z-phot for r-band cuts in the range $[19.4, 20.9]$ by using the mock catalog. Here the colors are used as input for the photometric methods. The training set and validation are obtained for $r < 19.4$. In the horizontal axis, we indicate the photometric redshift estimator used and in the vertical axis we indicate the r-band cut performed on mock catalog. Note that the scatter in the photometric redshift recovery increases with increasing magnitude depth for all methods. Moreover, for fainter flux limits the scatter increases with spectroscopic redshift.

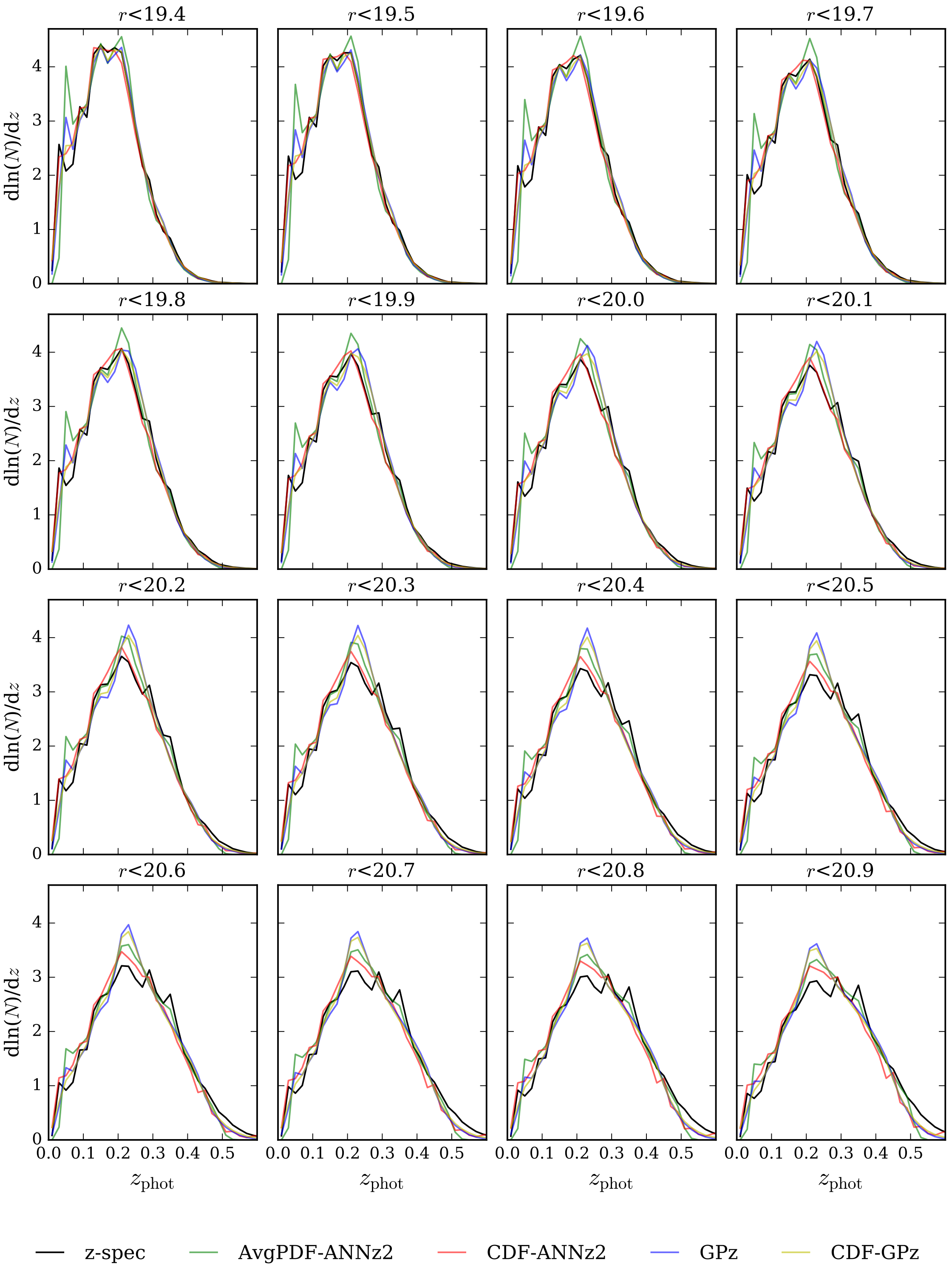

Figure 3.5: z-spec and redshift estimators distributions for r-band cuts in the range $[19.4, 20.9]$ by using the mock catalog. Note that for fainter magnitude cuts (i.e., the cuts in the region $[20.2, 20.9]$), the distributions of the estimated photometric redshifts have a peak in $z_{\rm phot} \approx 0.25$ which is not present in spectroscopic redshift distribution and a tail mismatch at higher z. The effect is greater for the photometric redshift estimators obtained via the GPz algorithm. In generally, the CDF-ANNz2 distribution fits better the z-spec distribution than the distributions obtained through the other estimators.

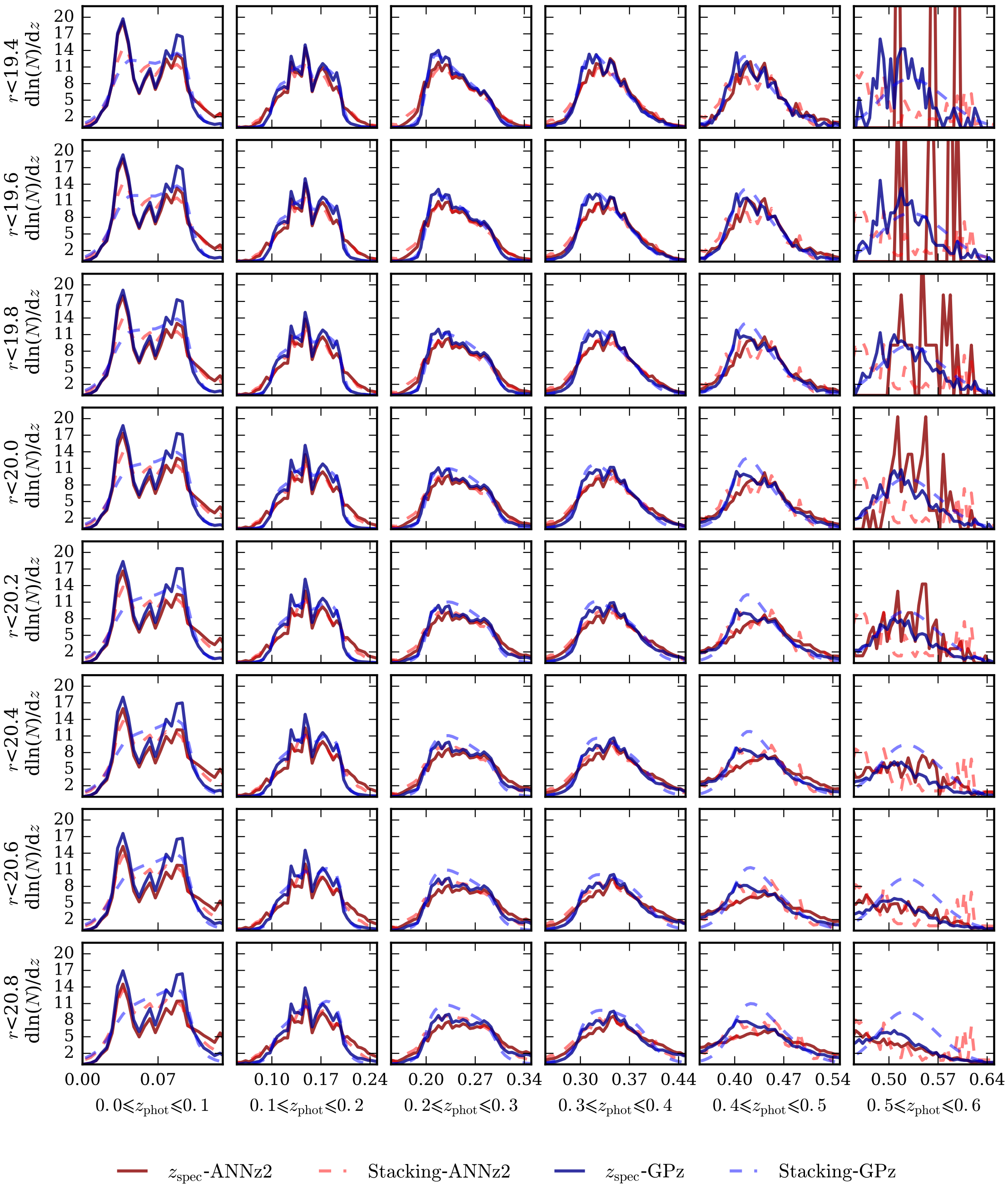

Figure 3.6: Comparison between the spectroscopic distribution and the stacking of photometric redshift PDFs estimated through `ANNz2` and `GPz` algorithms for slices of photometric redshift in all r-band magnitude cuts. We estimate different z-spec distribution for each used algorithm, since that the population of galaxies in the slices is different for each photometric redshift estimator. Note that the stacking of photometric redshift computed with `ANNz2` algorithm fits better the spectroscopic distribution than the `GPz` case. Here we use black solid line for z-spec (ANNz2), red dashed line for stacking (ANNz2), blue solid line for z-spec (GPz) and green dashed line for stacking (GPz).

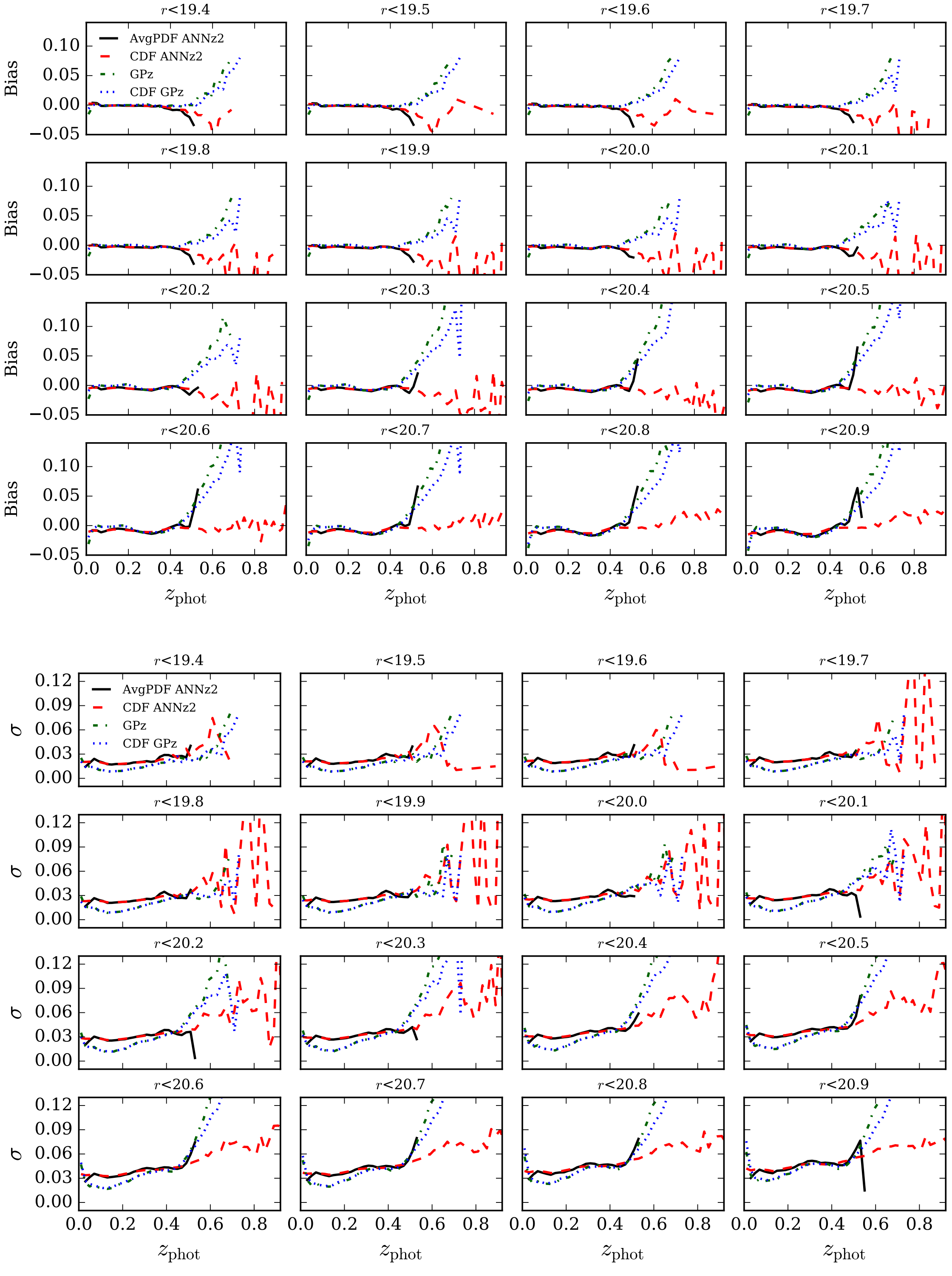

Figure 3.7: Bias (**Top**) and σ (**Bottom**) for r-band cuts in the range $19.4 \leq r \leq 20.9$ by using the mock catalog.

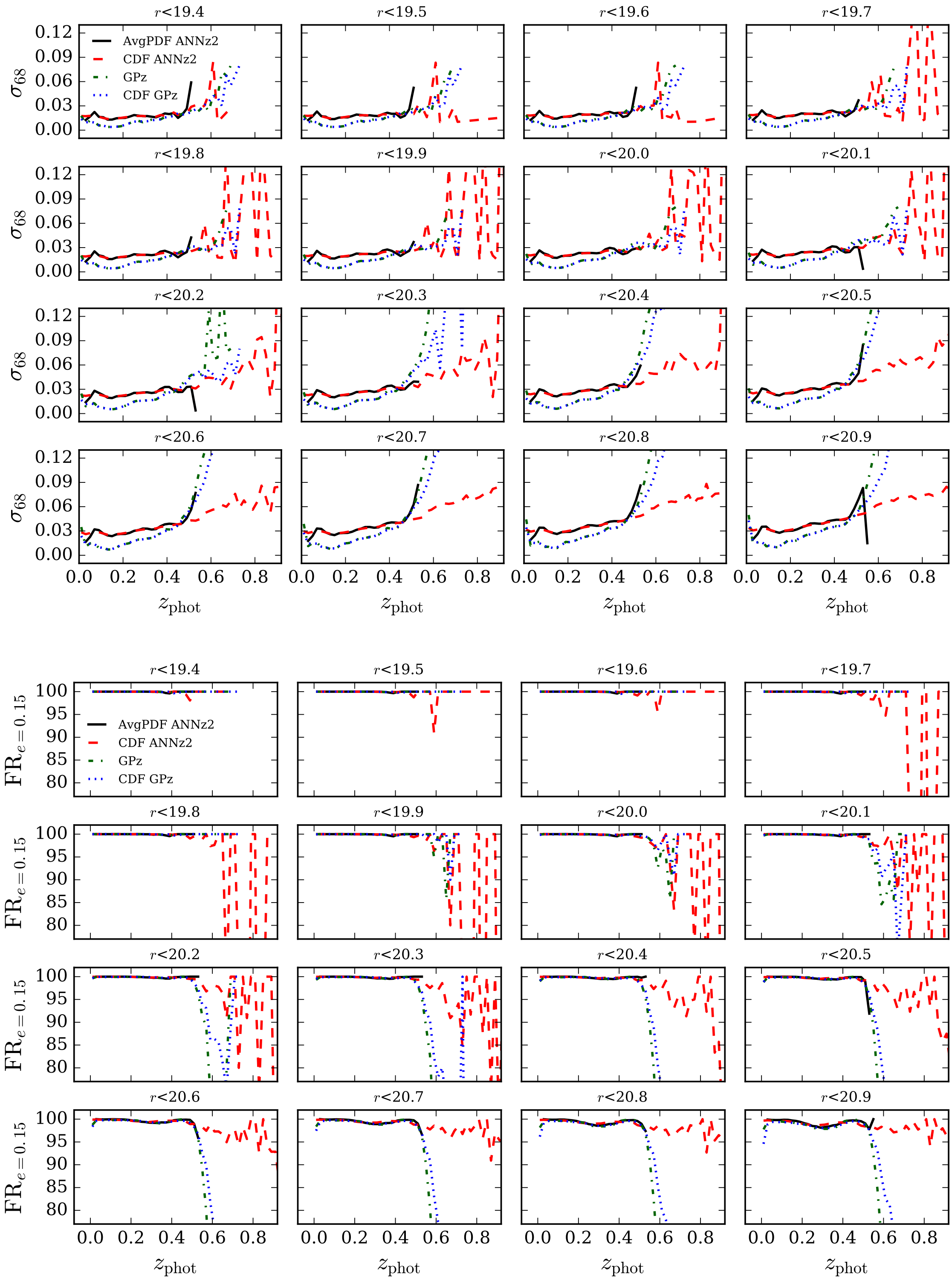

Figure 3.8: σ_{68} (**Top**) and $\mathrm{FR}_{e=0.15}$ (**Bottom**) for r-band cuts in the range $19.4 \leq r \leq 20.9$ by using the mock catalog.

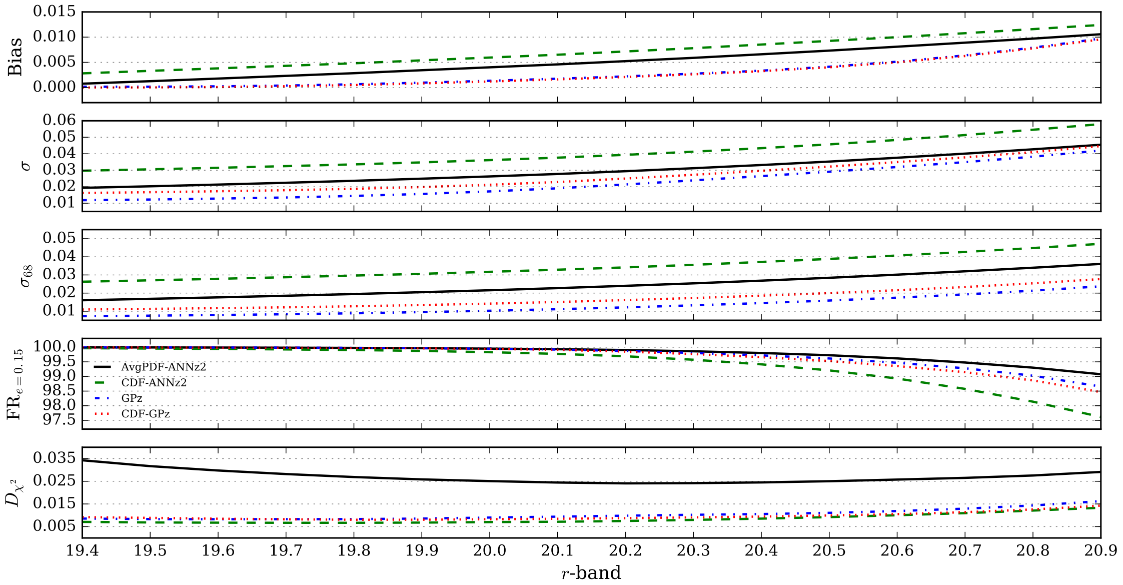

Figure 3.9: Global metrics of each photometric redshift estimator as function of r-band cut. The last figure is the chi-square measure (D_{χ^2}) for the z_{spec} distribution and z_{phot} one. We observe that the CDF-ANNz2 estimator has the best chi-square measure in all r-band cuts. Moreover, we note that the GPz estimators present the lowest global scatter and bias.

Figure 3.7 and figure 3.8 show the metrics in the range $0 < z_{\mathrm{phot}} < 1$. We see that the photometric redshift estimators have good metric values in the range $0 < z_{\mathrm{phot}} < 0.4$ for all r-band cuts. In this redshift range the metrics slightly worse for deeper cuts. We note that the bias and scatter computed for the estimators based on Gaussian GPz grow faster than the CDF-ANNz2 estimator in hight redshift and this is more evident for $r > 20.0$. Figure 3.9 shows the global metric values and the chi-square measure as function of r-band cut. The metric values worsen towards deeper magnitude limits, as we might expect. However, for each of the photometric redshift estimators the fraction $\mathrm{FR}_{e=0.15}$ remains above 99.5 per cent until $r \approx 20.2$. The AvgPDF-ANNz2 and CDF-ANNz2 estimators have the highest scatter and bias, though the CDF-ANNz2 estimator has the best chi-square measure for all r-band cuts. The GPz and CDF-GPz estimators present the lowest global scatter and bias, as well as high values for the global fraction $\mathrm{FR}_{e=0.15}$. These estimators also have a low chi-square measure.

Here the focus has been on comparing the different redshift runs. But for the science applications of these results, the important point is that for the best of the estimators (CDF-ANNz2), we can push the magnitude limit to a deeper range, and the degradation of redshift performance is only gradual. The performed test in slices of redshift shows us that the `ANNz2` code achieves good results in high redshifts for fainter magnitude cuts unlike to `GPz` code. Note that the Monte-Carlo method allows us to improve the accuracy of the photometric redshift values if we know the full photometric redshift PDF for every galaxy in the survey as is the `ANNz2` case.

3.4 Implication for detection of galaxy clusters

The reduced cost of measuring photometric redshifts, compared to spectroscopic redshifts, means that we are able to obtain photometric redshifts for many more objects more rapidly. As such, we are able probe larger cosmological volumes with photometric galaxy surveys, which is statistically beneficial for many cosmological analyses. Galaxy clusters are statistically very rare objects, at the extreme high mass end of the halo mass function, and so to maximize counts we need to probe large volumes. On the other hand, for the detection of galaxy clusters we need to ascertain with as great an accuracy as possible which galaxies are members of the cluster and which are not For this we need as accurate and precise redshift measurements as possible. Furthermore, to measure the halo mass function we need to estimate the halo mass of clusters. One way is to estimate the mass dynamically, for which we need to accurately know the positions of the cluster members to high precision, see Borgani & Guzzo (2001); Borgani et al. (2001); Voit (2005); Allen et al. (2011); Kravtsov & Borgani (2012). Therefore it is very important to estimate the photometric redshifts with accuracy and precision for minimize the impact on the systematic errors in the estimated number cluster count and subsequent cosmological analysis. The main aim here is to examine the implications to use a non-representative training data set for estimating photometric redshift at the time of detecting galaxy clusters with methods that are sensitive to the density of galaxies in a field such as Voronoi Tessellation (VT) or kernel density estimation, see chapter 4. In addition, we want to know the impact in each photometric redshift estimator used in this work.

In redshift regions with higher density of galaxies we expect to find more galaxy clusters. Therefore in order to estimate the number of galaxy clusters that we can detect with a given redshift survey, we first compute the number density of galaxies as a function of redshift, $n(z)$. This is equal to the number of galaxies, N, per comoving volume, V, and given by,

$$n(z) = \frac{\mathrm{dN}}{\mathrm{d}V} = \frac{\mathrm{dN}}{\mathrm{dz}}\frac{\mathrm{dz}}{\mathrm{dV}} = \frac{\mathrm{dN}}{\mathrm{dz}} f_c(z), \tag{3.14}$$

where

$$f_c(z) \equiv \frac{H(z)}{D_c^2(z)\Delta\Omega}. \tag{3.15}$$

Here dN/dz corresponds to the galaxy redshift distribution, $\Delta\Omega$ is the angular area that the galaxy catalogue covers, $H(z)$ is the Hubble parameter and D_c is the comoving distance.

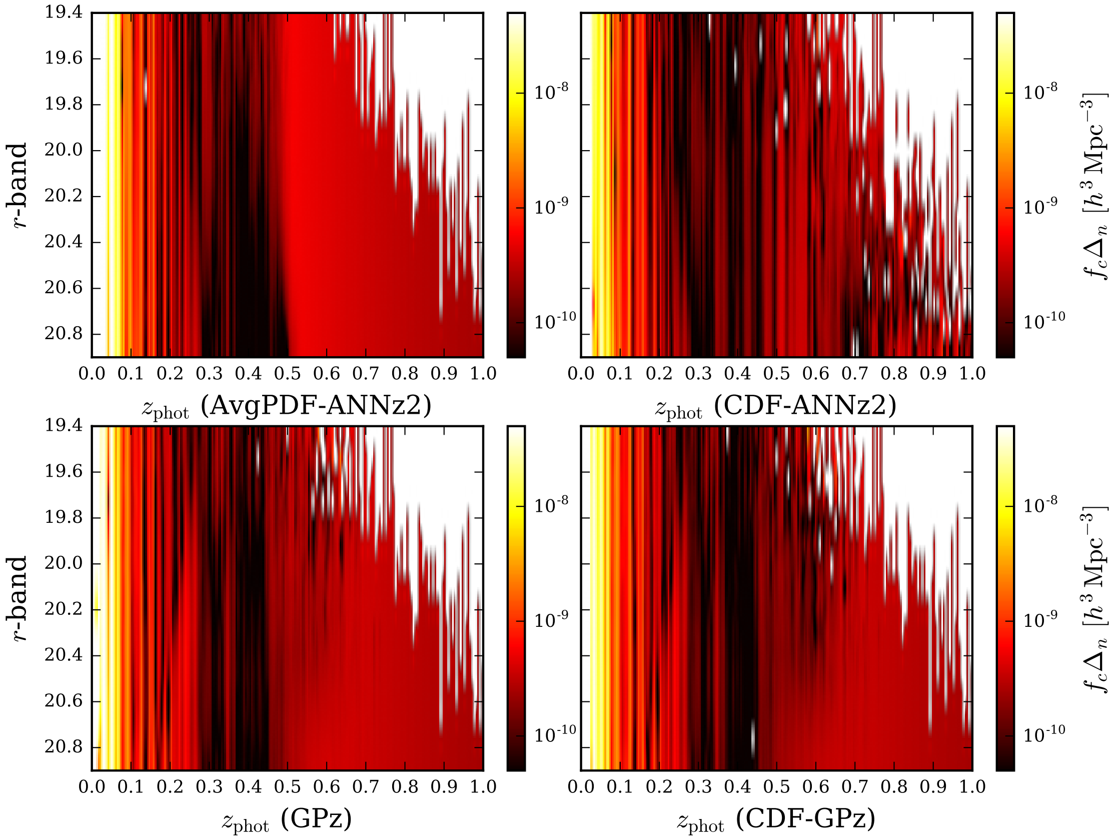

Figure 3.10: Comparison of number of galaxies per comoving volume element computed by using each photometric redshift estimator with the spectroscopic redshift case. We compute the relative error Δ_n times the function $f_c(z)$ (this function contains the cosmological information, see equation (3.15)) between density of galaxies for z-spec and the density of galaxies for each redshift estimator. This process is performed for all r-band cuts. Note that the CDF-ANNz2 estimator allows us to detect galaxy clusters agreement with the z-spec data for deeper cuts in the r-band magnitude and highest redshifts, hence we expect that the galaxy cluster catalog obtained by employing this photometric redshift estimator is purer until high redshifts than in other cases.

We compare the density of galaxies estimated using the photometric redshift distribution for each photometric redshift estimator with the density of galaxies estimated using the spectroscopic redshift. We make this comparison for each of our r-band magnitude cuts. To quantify this comparison, we use the function $f_c(z)$ times the relative error between the two number densities, thus we have

$$f_c\Delta_n \equiv \frac{H}{D_A^2\Delta\Omega}\left|1-\frac{n}{\bar{n}}\right|, \tag{3.16}$$

where n is the number density of galaxies from the photometric redshift estimators and $\bar{n}$ is the number density of galaxies from the spectroscopic redshift. This quantity is relevant as we would like to have a cluster detection method based on density estimation which is not affected by detection in the $n(z)$ function inferring incorrectly a different density of galaxies at that redshift. We note that this calculation is not applicable to color based methods to finding galaxy clusters.

Figure 3.10 shows the amount described in equation (3.16) using AvgPDF-ANNz2, CDF-ANNz2, GPz and CDF-GPz estimators. In each panel darker colors correspond to smaller values for $f_c\Delta_n$, which indicates regions in the magnitude versus redshift space where the number densities derived from photometric redshifts are equal to the number densities from spectroscopic redshifts. Therefore in such regions we could robustly detect a galaxy cluster using both spectroscopic and photometric redshifts. Note that for the CDF-ANNz2 estimator we see more darker regions at higher redshifts suggesting that with this estimator we can more robustly detect galaxy clusters at higher redshift with deeper r-band selected samples. Hence we would expect that a galaxy cluster catalog obtained with this photometric redshift estimator would be purer, out to higher redshift, compared to catalogs build with the other estimators. In other words, this result suggests that of all of the photometric redshift estimators considered, the CDF-ANNz2 estimator would provide the most accurate detection of galaxy clusters. The AvgPDF-ANNz2 estimator has the best results in the region $z \sim [0.25, 0.50]$ and deeper r-band magnitude cuts. The GPz estimators have good results for the initial magnitude cuts. However, for deeper cuts, we observe that in the redshift range $z \sim [0.2, 0.3]$, the `GPz`-based estimators have larger values for $f_c\Delta_n$ than the `ANNz2`-based estimators (i.e. the `GPz`-based estimators have fewer darker regions than the `ANNz2`-based estimators). This is understandable as this within this redshift range where, in the lower panels of figure 3.5, we saw a spurious peak in the photometric redshifts from the `GPz`-based estimators. We conclude that the results presented here can be used to guide parameter optimization of cluster finding algorithms.

3.5 Conclusions and remarks

Photometric redshifts allow us to probe much larger volumes of the Universe than it is possible with spectroscopic redshifts, but they have large measurement uncertainties. Machine learning methods are often used to estimate photometric redshifts, but these estimators must be trained using existing spectroscopically detected datasets, which probe a limited volume. There is much uncertainty regarding the reliability of measured photometric redshifts when the spectroscopic training set is not representative of the photometric dataset. In this chapter we have investigated the degradation in the accuracy and precision of the recovered of photometric redshifts when two machine learning methods, applied to deep photometric datasets, are trained using much shallower and brighter spectroscopic samples. We have used the `ANNz2` and `GPz` machine learning codes for estimating the photometric redshifts with four colors instead of all five magnitudes as input, ensuring representativeness only in this subspace, and evaluated the consequences, both in SDSS DR12 data trained on GAMA spectra, and on mock catalogs. For this analysis, we also utilize the Monte-Carlo random sampling for defining a photometric redshift estimator based on the cumulative distribution function (CDF) of the redshift probability distribution function (PDF). Altogether we use four photometric redshift estimators in this work; AvgPDF-ANNz2, CDF-ANNz2, GPz and CDF-GPz, which we define and introduce in section 3.2.

We start by showing that, for a representative training data set in the magnitude space, the photometric redshifts obtained using the `ANNz2` and `GPz` algorithms display the similar quality, either using magnitude-selected or color-selected training sets as input. We estimate the photometric redshift for the samples GAMA DEEP and GAMA MAIN (subsamples from the SDSS DR12 data with GAMA selections, see section 3.1.3), which are trained by the spectroscopic GAMA survey. In general, we find that the results in the metrics obtained for the mock catalogue display similar trends to the results metrics obtained for the GAMA test data. We observe that the photometric redshift distribution obtained with the CDF-ANNz2 estimator is the most consistent with the spectroscopic redshift distribution for the GAMA test data. We note that the distribution of the photometric redshifts obtained with those estimators that sample the CDF are a better fit to the photometric redshift PDF stacking of all galaxies in the data set. Nonetheless, these estimators yield a greater scatter than the other estimators.

We proceed to analyze samples of the mock catalog selected using progressively deeper cuts in the r-band magnitude in order to study the degradation of the photometric redshifts obtained from the AvgPDF-ANNz2, CDF-ANNz2, GPz and CDF-GPz estimators when the training data set is non-representative of a deeper photometric testing set. In each instance we use the same training data set selected with $r < 19.4$. The AvgPDF-ANNz2 estimator fails at high redshift and worsens for deeper cuts. We consider that this result is due to the low density of galaxies in the training set for this region. Comparatively, the CDF-ANNz2 estimator shows better performance at higher redshifts, albeit with larger scatter. We observe that the CDF-ANNz2 estimator has the best chi-square measure for all r-band selections. The GPz and CDF-GPz estimators, appear to provide more reliable results at low redshifts. Nevertheless for deeper cuts, we observe that these estimators tend to under-estimate the redshifts of high-redshift spectroscopic galaxies leading to an excess of photometric redshifts at the peak of the redshift distribution and a mismatch in the tail of the distribution. For the scatter plots between spectroscopic redshifts and photometric redshifts as well as the $n(z)$ plots up to $r < 20.0$ we observe very good results in all photometric redshift estimators, see figure 3.4 and figure 3.5.

In order to quantify the impact of the photometric redshifts in the detection of galaxy clusters, we compute the number of galaxies per comoving volume for each redshift estimator (i.e., the number density of galaxies). The depth of the cut is directly related with the density of galaxies and hence the number of galaxy clusters detected. For deeper cuts we can detect more galaxy clusters and improve the redshift depth. However, we show that the estimated photometric redshifts become poorer quality for deeper cuts. The density of galaxies given by CDF-ANNz2 estimator has the least error according to the number density of galaxies given by spectroscopic redshift data in deeper cuts and high redshifts. For lower redshifts and cuts, the other estimators have better results, nonetheless the CDF-ANNz2 estimator also has good results. We conclude that the results here can improve detectability of clusters with density based detection methods.

Chapter 4

Galaxy cluster finder combining VT and FOF techniques

Galaxy clusters are the largest gravitationally bound objects observed in the Universe. Given that galaxies are tracers of dark matter, galaxy clusters allow us to study the formation and evolution of large-scale structure. Galaxy cluster number counts have been used to constrain cosmological parameters (Battye & Weller (2003); Mantz et al. (2010); Allen et al. (2011); Mana et al. (2013)). George O. Abell pioneered the observation of galaxy clusters. Abell (1958) assembled the first large sample of clusters in the Northern sky. Abell et al. (1989) added to this sample with a large galaxy cluster catalog including the Southern sky. The above works marked the beginning of a whole science on detection of groups and clusters of galaxies in the Universe. The following methods are commonly used to detect galaxy clusters: X-ray observations of the Intra-Cluster medium (ICM), which is a hot baryonic gas with the high kinetic energy; the Sunyaev-Zel'dovich effect, which is a result of the inverse Compton scattering of CMB photons when they pass through the ICM; gravitational lensing, which is associated with the high mass concentration contained in the galaxy cluster that bends light rays passing near to it according to general relativity; and optical and near-infrared observations, which are based on looking for overdensities in the photometric or spectroscopic surveys of galaxies.

To find optical galaxy clusters in a photometric survey it is necessary to employ cluster-finding algorithms that employ techniques that take into account the angular clustering of the galaxies and their photometric redshifts (Ramella et al. (2001); Botzler et al. (2004); Lopes et al. (2004); Berlind et al. (2006); Soares-Santos et al. (2011); Hung-Yu et al. (2014)) or by utilizing the cluster red-sequence method (Gladders et al. (2007); Gal et al. (2009); Hao et al. (2010); Rykoff et al. (2014);

Rykoff et al. (2016)). The latter technique takes advantage of the fact that generally a giant elliptical galaxy, which is called brightest cluster galaxy (BCG), is nearby the center of cluster. Several projects that are currently being developed, such as the Dark Energy Survey (DES)[*], the Large Synoptic Survey Telescope (LSST)[†], the Javalambre Physics of the Accelerating Universe Astrophysical Survey (J-PAS)[‡], among others, will provide large galaxy catalogs from which we can probe the distribution of clusters in the Universe.

Commonly used geometric algorithms for detecting clusters that use the photometric redshifts of the galaxies include Voronoi Tessellation (VT) (Voronoi (1907)) and Friends of Friends (FOF) (Huchra & Geller (1982)). In this work, our main aim is to detect galaxy clusters by using the photometric redshifts from a sample of the SDSS DR12 survey. Our cluster-finding algorithm, which we call henceforth VT-FOFz, combines both Voronoi Tessellation and Friends of Friends techniques. The galaxies that form part of a cluster are those that lie in overdense regions, hence we select as potential cluster member candidates those galaxies that have small Voronoi cell volume (i.e., galaxies with larger Voronoi density) at a given redshift. Afterwards, in order to detect the galaxy clusters from the sample of candidate galaxies, we employ the FOF algorithm provided by Farrens et al. (2011) which is based on that described in Botzler et al. (2004). To evaluate the efficiency of the method, we run the VT-FOFz cluster finder on a mock galaxy catalog. For this, we compute the completeness and purity. We perform a statistical analysis using the number counts of detected clusters and mock halo number counts, while requiring that the detected clusters match with the largest number of haloes. In addition, we run the cluster finder on a sample of SDSS DR12 survey and we compare the results obtained with the galaxy clusters detected by the redMaPPer cluster finder, see Rykoff et al. (2014). The analysis is performed for r-band magnitude cuts in the range $19.4 < r < 20.9$, in order to evaluate the impact of the photometric redshift via a non-representative training data set in magnitude space, see chapter 3. We expect to recover a large volume of galaxy clusters detected by the redMaPPer code.

This chapter is organized as follows: Section 4.1 describes the mock catalogs used in this work, moreover we describe the selection criteria to choose the galaxies from the SDSS DR12 survey and we present the main features of the galaxy cluster

[*] ⟨http://www.darkenergysurvey.org⟩

[†] ⟨http://www.lsst.org⟩

[‡] ⟨http://www.darkenergysurvey.org⟩

catalog given by Rykoff et al. (2014). Section 4.2 describes the VT-FOFz cluster finder. In section 4.3 we perform a statistical analysis using the mock catalog. Here, we explain the method employed for determining a match between the detected galaxy clusters and the simulated haloes, furthermore we define the completeness and purity. Section 4.4, presents the galaxy clusters detected from the sample of SDSS DR12 data, furthermore we compare the VT-FOFz and redMaPPer cluster catalogs. Finally, in section 4.5 we present the conclusions of this chapter. This work is presented in the paper Rivera et al. (2017b) (II-prep.).

4.1 Data

In this section we describe the mock catalogs used to perform the statistical analysis of the VT-FOFz cluster finder. We employ these catalogs to evaluate the performance of the cluster finder by calculating the completeness and purity which quantify the reliability of the estimated galaxy cluster catalog. In addition, we describe the photometric redshift survey based on the SDSS DR12 survey which is employed to detect the galaxy clusters via the VT-FOFz algorithm.

4.1.1 Mock catalogs

For the performance analysis we use the mock galaxy catalog described in section 3.1.1. We apply a similar i-band magnitude cut $i < 21$ to obtain a subsample with "SDSS-like" cuts. In total we have 1 876 505 galaxies, out to a redshift of $z < 1.62$ and a mean redshift of $z_{\mathrm{mean}} \approx 0.35$. In order to avoid catastrophic results, we only consider galaxies with redshift error smaller than 0.15. We use the photometric redshifts estimated in chapter 3. Here, we choose the photometric redshift estimated by the `ANNz2` code (Sadeh et al. (2016)) the single value for the photometric redshift is determined from the inverse of the cumulative distribution function, see chapter 3. The training data set has a r-band magnitude cut $r < 19.4$ and is used in color space to avoid non-representativeness problems in the magnitude space. We choose this redshift estimator because, despite having large scatter, its associated photometric redshift distribution fits very well with the true redshift distribution, which allows us to improve the detection of galaxy clusters.

We also use the mock dark matter halo catalog associated to the galaxy mock catalog. The dark matter haloes which populate the mock catalog are provided by the *Millennium Simulation*, which is a 2160^3 particle N-body simulation for ΛCDM

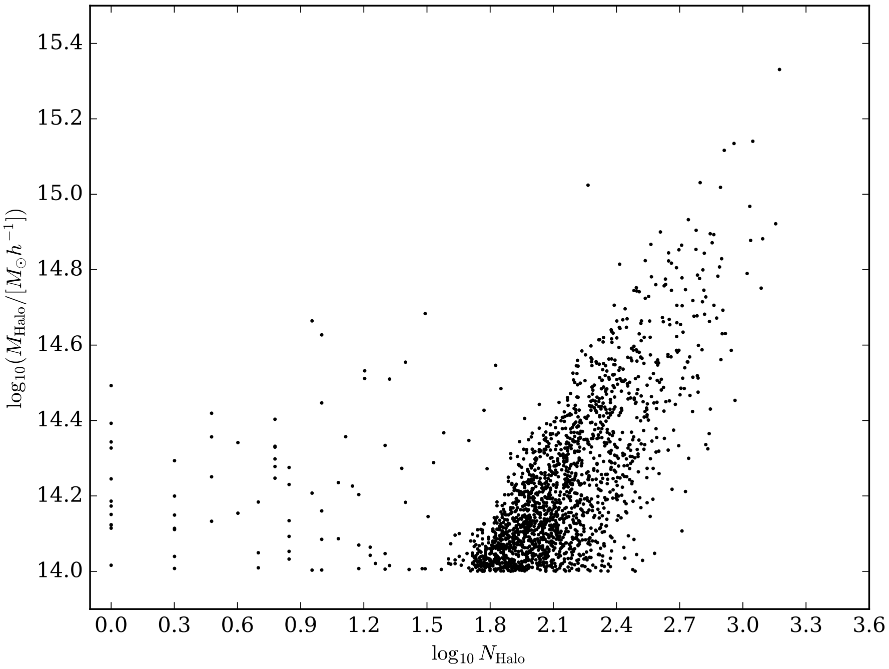

Figure 4.1: Scatter plot between halo mass and number of galaxies inhabiting the halo. Note that the mass selection sets the minimum number of galaxies which must be considered for the cluster finder in the mock catalog. In order to include the scatter in the observable-mass relation and the possible bias, we set that $N_{g,\rm obs,min} = 10$.

cosmology (Springel et al. (2005)). The mock encompasses a cubic volume of side 500 h^{-1} Mpc, in which the dark matter field is evolved from redshift $z = 127$ until the present. The simulation considers haloes with a minimum of 20 particles for a halo resolution of $M_{\rm halo,lim} = 1.72 \times 10^{10} h^{-1} M_\odot$. The adopted ΛCDM model has the following parameters: $\Omega_b = 0.045$, $\Omega_m = 0.25$, $\Omega_\Lambda = 0.75$, $h = 0.73h$, $n_s = 1$ and $\sigma_8 = 0.9$ according to the cosmological parameters estimated from the first year results from the WMAP (Spergel et al. (2003)). For our analysis we consider dark matter haloes with mass $M \geq 10^{14} M_\odot h^{-1}$ and redshift $z \leq 0.5$. This threshold redshift is chosen because the estimated photometric redshifts in the galaxy mock catalog are reliable up to this limit, see chapter 3. Note that the mass selection determines the minimum number of galaxy members required to detect galaxy clusters according to figure 4.1. In order to include the scatter in the observable-mass relation and the possible bias, we set $N_{g,\rm obs,min} = 10$ for this work. After mass and redshift selection we get a halo catalog with 1 903 objects.

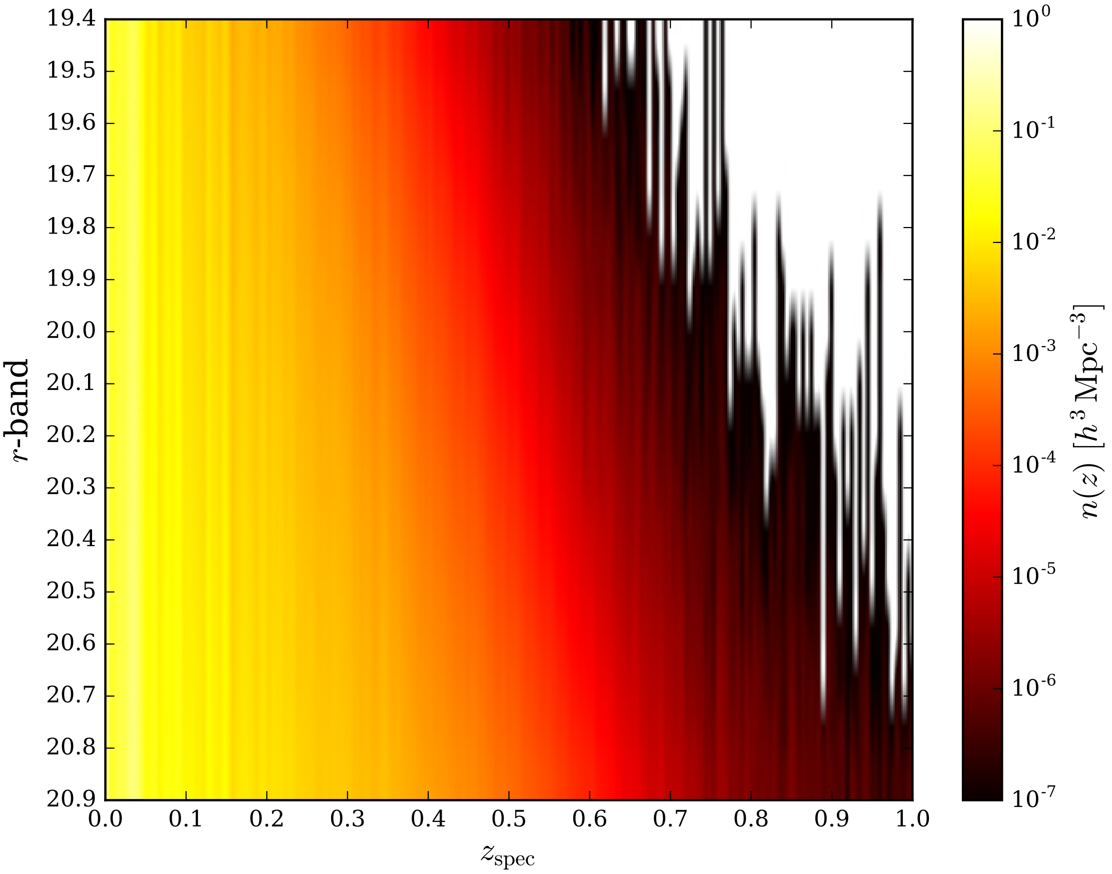

Figure 4.2: Number of galaxies per comoving volume versus the spectroscopic redshift (true redshift) in each r-band cut for the mock catalog. We note that for deeper cuts, the density of galaxies is highest in each spectroscopic redshift. Thus, if we perform a cut in low magnitude, we can lose galaxy clusters and we can increase the error in cosmological studies. For high redshifts, the density of galaxies is low. Therefore, in this region we expect to detect less galaxy clusters.

4.1.2 GAMA DEEP survey

The GAMA DEEP survey is a photometric redshift catalog in which the photometric data are obtained from the SDSS DR12 survey with "GAMA-like" cuts and without magnitude limit, as it is described in section 3.1.3. The photometric redshifts are estimated using the `ANNz2` code. This sample was trained using the color space in order to avoid problems of non-representativeness in the magnitude space. Section 3.3.1 shows that the quality of estimated redshifts is kept when using the magnitudes or the colors as input. The single value for the photometric redshift is estimated through the Monte Carlo sampling method (i.e., the method based on the inverse of the cumulative function). Figure 4.3 shows the number of galaxies per comoving volume for true redshifts from the mock catalog, the estimated photometric redshifts from the mock catalog and the estimated photometric redshifts from the GAMA DEEP survey. We note that the estimated number density of galaxies from the true redshifts and photometric redshifts for the mock catalog agree very well. On the other hand, the number density of galaxies in the GAMA DEEP survey is greater than the mock catalog case in $z \in [0, 0.33]$, especially at $z \approx 0.3$ we observe a little peak in the GAMA DEEP. Hence in this redshift, we expect to detect more galaxy clusters in the GAMA DEEP survey than in the mock galaxy catalog. For higher redshifts, the number density of galaxies drops faster in the GAMA DEEP survey than in the mock catalog. Note that the above result is independent of r-band magnitude cuts. Indeed, the number density of galaxies experiences few changes for cuts with $r \geq 20.6$, see figure 4.2.

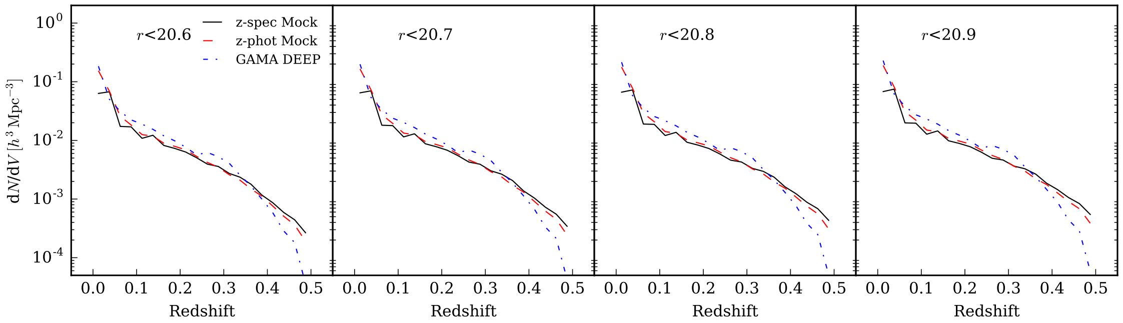

Figure 4.3: Number of galaxies per comoving volume for true redshifts from the mock catalog, the estimated photometric redshifts from the mock catalog and the estimated photometric redshifts from the GAMA DEEP survey. Note that the estimated number density of galaxies from the true redshifts and photometric redshifts for the mock catalog agree very well. On the other hand, the number density of galaxies in the GAMA DEEP survey is greater than the mock catalog case $z \in [0, 0.33]$, especially at $z \approx 0.3$ we observe a little peak in the GAMA DEEP case. For higher redshifts, the number density of galaxies drops faster in the GAMA DEEP survey than in the mock catalog.

4.1.3 redMaPPer SDSS DR8 cluster catalog

The readMaPPer SDSS DR8[§] cluster catalog was obtained by running the redMaPPer cluster-finding algorithm on a sample from SDSS DR8 photometric catalog. The

[§] ⟨http://risa.stanford.edu/redmapper/⟩

red-sequence Matched-filter Probabilistic Percolation (redMaPPer) code is a cluster finder which detects optical clusters through the red sequence technique, see Rykoff et al. (2014); Rykoff et al. (2016) for more details. The used photometric sample covers a total angular area of $\approx 10\,400\,\mathrm{deg}^2$ after applying the selection criteria described in section 2 of Rykoff et al. (2014). The catalog contains galaxy clusters with a richness cut $\lambda \geq 20\,S(z_\lambda)$ in the redshift range $z_\lambda \in [0.08, 0.55]$ which minimizes the edge effects from the training sample. Here the function $S(z_\lambda)$ is called the scale factor and it defines the correction factor on the richness caused by the survey depth. The chosen richness cut implies that the detected clusters should have at least 20 galaxies above the limiting magnitude of the survey. This result agrees to the threshold mass of $M > 10^{14}\,M_\odot$ according to Rykoff et al. (2012). In total, the catalog has 26 111 detected galaxy clusters.

4.2 Galaxy cluster finder

Here we describe the VT-FOFz cluster finder. This combines two techniques Voronoi Tessellation (VT) (Voronoi (1907)) and Friends of Friends (FOF) (Huchra & Geller (1982)) on a photometric redshift survey. Both techniques are frequently used to find galaxy clusters. Ramella et al. (2001); Kim et al. (2002); Lopes et al. (2004); Soares-Santos et al. (2011) used VT to find galaxy clusters, whereas Berlind et al. (2006); Botzler et al. (2004); Farrens et al. (2011); Hung-Yu et al. (2014) used FOF algorithm to look groups and clusters on galaxy surveys. The VT-FOFz software package is split into two parts

1. Compute Voronoi regions for all galaxies and to establish the candidate galaxies to form clusters.
2. From the candidate galaxies catalog, we use the FOF code to find the galaxy clusters.

In the remainder of the section we describe in detail the above steps.

4.2.1 Candidate galaxies to form clusters

In order to determine the candidate galaxies, we convert the astronomical coordinates (α, δ, z) to Cartesian coordinates (X,Y,Z), where α is the Right Ascension angle, δ is the Declination angle and z is the cosmological redshift. We use the following transformation equations:

$$X = D_c \cos(\delta)\cos(\alpha), \tag{4.1}$$

$$Y = D_c \cos(\delta)\sin(\alpha), \tag{4.2}$$

$$Z = D_c \sin(\delta), \tag{4.3}$$

where D_c is the comoving distance, see equation (2.77). With the galaxies in Cartesian coordinates, we split the catalog into boxes to parallelize the processes. Our aim is to have approximately the same number of galaxies in each box, hence we use the k-d tree algorithm, which organizes points in a k-dimensional space. To avoid boundary problems at the time of merging the boxes, we introduce an overlap region between nearest neighbors. To build that region, we compute the mean inter-galaxy separation for each dimension

$$d_i = D_i N^{-1/3};\ i = 1, 2, 3, \tag{4.4}$$

where N is the number of objects in the box and D_i are the dimensions of the box. Then, we add $N_t(d_i/2)$ in each boundary of respective dimension, where N_t is a parameter introduced to control the overlap region size. This way of making the overlap allows us to take into account the spatial galaxy distribution for every box.

After the splitting the catalog, we establish Voronoi cells for each galaxy by using the open source software library `Voro++`[¶]. This library computes Voronoi tessellations in 3d (Rycroft (2009)). The Voronoi volume of a galaxy is defined as the sum of the volumes of all the cells that are closer to that galaxy than to any other. Thus, in regions with high density of galaxies the Voronoi volumes should be smaller than in regions with low density. The galaxy clusters with large number of members are found in places with high density, thus the candidate galaxies to form clusters have small Voronoi volumes. We define the Voronoi density as

$$\rho_{\mathrm{voro}} \equiv \frac{1}{V_{\mathrm{voro}}}, \tag{4.5}$$

where V_{voro} is the Voronoi volume of galaxy. To recover the initial catalog in the coordinates (α, δ, z) and to add ρ_{voro} for each galaxy, we clean the galaxies in the overlap region for all boxes. This cleanup method is valid when the overlap is large

[¶] ⟨http://math.lbl.gov/voro++/⟩

such that the Voronoi volume for the galaxies at the boundary is the true value. Figure 4.4 shows an example of this method. Here, we set $N_t = 4$ as a good value in order to build the overlap regions.

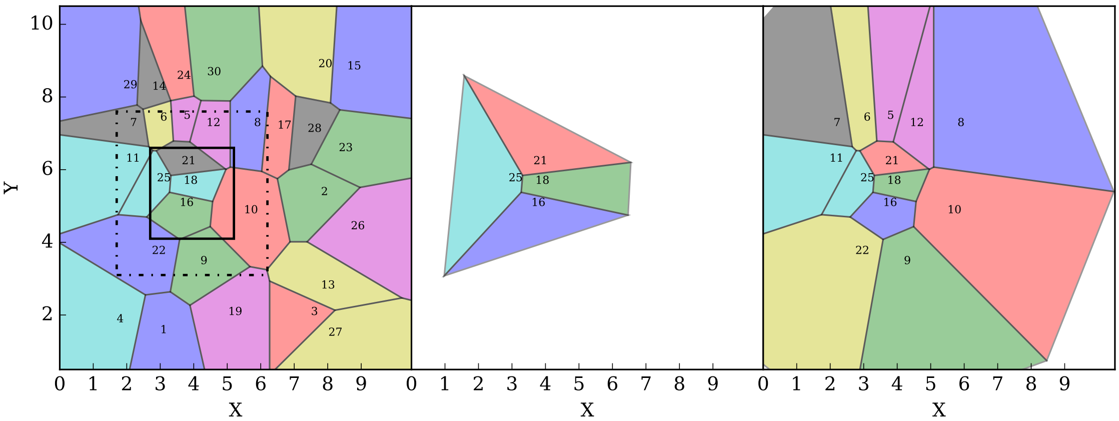

Figure 4.4: Example of the cleanup method used for a Voronoi tessellation in 2d. **Left**: The solid rectangle represents a box without a region of overlap. Note that this split contains the points $\{16, 18, 21, 25\}$. To find the true Voronoi regions of these points it is necessary to add an overlap region (region between the solid rectangle and dashed rectangle), which must contain the points $\{5, 6, 7, 8, 9, 10, 11, 12, 22\}$. **Center and right**: Voronoi regions for the points contained in the box without and with the overlap region, respectively.

According to the above assertion, the candidate galaxies to form clusters are those with large $\rho_{\rm voro}$. However, we know that the redshift distribution of galaxies is not constant (see previous chapter). To avoid mistakes in the selection, we split the catalog into redshift bins, here we set 100 redshift bins. We define the percentage of cut $P_{\rm cut}$ as the fraction of the densest galaxies which are selected to be candidate galaxies. For each redshift bin, the $P_{\rm cut}$ parameter establishes a density threshold $\rho_{\rm thr}$ which allows us to select galaxies with density above this threshold. The percentage of cut is a free parameter.

4.2.2 Finding galaxy clusters through FOF

The FOF algorithm is a percolation method that allows us to find groups and clusters in galaxy surveys. This method looks for galaxy pairs, that are nearer to one another than a given linking length. The FOF method used for this work is the

SFOF[‖] code provided by Farrens et al. (2011) and based on work by Botzler et al. (2004). After selection of the densest galaxies, we run the FOF code on catalog of candidate galaxies. In order to optimize the runtime in this step, we perform an additional splitting of the objects into overlapping patches in $\alpha - \delta$ space. The code can then be run on each patch independently.

SFOF percolates in angular space in bins of photometric redshift. The angular distance Θ between two galaxies in the same bin is given by the following equation

$$\cos(\Theta) = \sin(\delta_1)\sin(\delta_2) + \cos(\delta_1)\cos(\delta_2)\cos(\alpha_1 - \alpha_2). \tag{4.6}$$

These galaxies are considered friends (i.e. linked) if they satisfy the condition $\Theta \leq \Theta_{\text{friend}}$, where Θ_{friend} is the angular linking length. The transverse separation of two objects in a redshift bin can be estimated through the comoving distance, $\delta\ell = D_c(z)\delta\theta$, where $\delta\theta$ is the angular separation. Recall that in this work we consider a flat Universe, thus the angular comoving distance is equal to the comoving distance. The angular linking length is defined as

$$\Theta_{\text{friend}} = \frac{\delta\ell_{\text{friend}}(z)}{D_c(z)}, \tag{4.7}$$

where $\delta\ell_{\text{friend}}$ is the transverse linking length. Note that the mean inter-galaxy separation ℓ_m for a redshift bin increases if the number of galaxies decreases. In order to take into account this effect in the selection of galaxy cluster members, the transverse linking length is defined as a function of the redshift

$$\delta\ell_{\text{friend}}(z) \equiv b_r \frac{\ell_m(z)}{\ell_m(z_{\text{ref}})}. \tag{4.8}$$

Here, b_r is a free parameter for the transverse linking length which must be provided and z_{ref} is the reference redshift where the transverse linking length is equal to b_r. Henceforth, when we talk about the transverse linking length we will be referring to the b_r parameter. The mean inter-galaxy separation is computed by using the mean superficial density of galaxies in the redshift bin, thus

$$\ell_m(z) = \left(\frac{\mathrm{d}N}{\mathrm{d}A}\right)^{-1/2}. \tag{4.9}$$

[‖] ⟨https://github.com/sfarrens/sfof⟩

Since the bins are shallow, the following approximation is used

$$\frac{\mathrm{d}N}{\mathrm{d}A} \propto \frac{\mathrm{d}N}{\mathrm{d}V} = \frac{\mathrm{d}N}{\mathrm{d}z}\frac{\mathrm{d}z}{\mathrm{d}V}. \tag{4.10}$$

By using equation (4.7), equation (4.8) and equation (4.9) the angular linking length can be written as

$$\Theta_{\text{friend}} = \left(\frac{\mathrm{d}N}{\mathrm{d}z}\frac{\mathrm{d}z}{\mathrm{d}V}\right)^{-1/2} \frac{R_{\text{ref}}}{D_a(z)}, \quad R_{\text{ref}} = b_r \left(\frac{\mathrm{d}N}{\mathrm{d}z}\frac{\mathrm{d}z}{\mathrm{d}V}\right)^{1/2}_{z_{\text{ref}}}, \tag{4.11}$$

where $\mathrm{d}N/\mathrm{d}z$ is the photometric redshift distribution and $\mathrm{d}z/\mathrm{d}V$ is given by the comoving volume definition. In order to implement the photometric redshift errors in the analysis, the method allocates a galaxy in a redshift bin if the following condition is satisfied:

$$|z_{\text{gal}} - z_{\text{bin}}| \leq b_z \mathrm{d}z_{\text{gal}}, \tag{4.12}$$

where z_{bin} is the central redshift of the bin, $\mathrm{d}z_{\text{gal}}$ is the photometric redshift error of the galaxy and b_z is the free parameter for the line-of-sight linking length. Here b_z is a factor that determines the number of galaxies which are allocated along the line-of-sight. The process of finding galaxy clusters in each redshift bin is performed independently. As galaxies can exist in multiple bins, it is possible to form proto-clusters in different bins with common members. When finishing the search for clusters in all redshift bins, the proto-clusters with common members are merged to form the final detections and build the galaxy cluster catalog.

The cluster center in Right Ascension, Declination and redshift, is computed as the median of the galaxy members. The errors are computed via the standard error on the median. The richness is defined as the number of galaxy members in cluster. The cluster radius is calculated as the distance from the cluster center to the position of the farthest member.

Figure 4.5 shows all the steps performed to detect galaxy clusters using the VT-FOFz pipeline. In summary, the VT-FOFz cluster finder has three free parameters P_{cut}, b_r and b_z, which must be calibrated.

4.3 Performance of the cluster finder

In order to assess the performance of the VT-FOFz cluster finder we perform a statistical analysis using the mock catalogs described in section 4.1.1. We compute

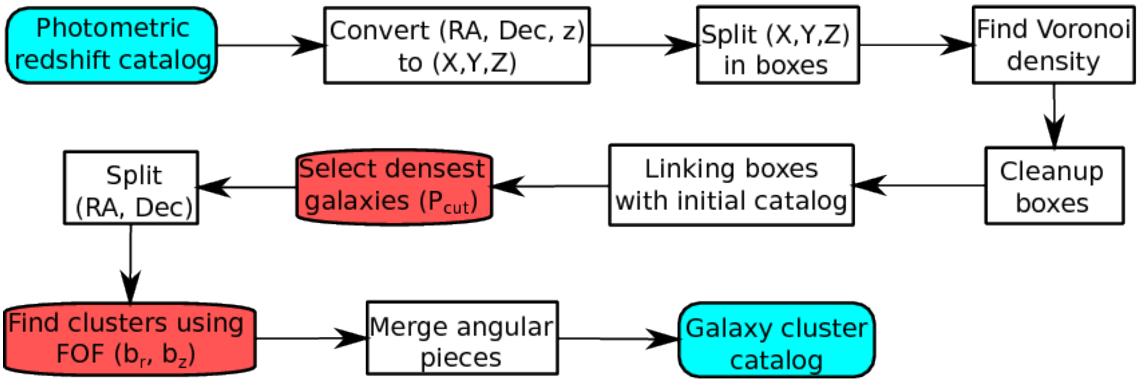
Photometric redshift catalog
Convert (RA, Dec, z) to (X,Y,Z)
Split (X,Y,Z) in boxes
Find Voronoi density
Cleanup boxes
Linking boxes with initial catalog
Select densest galaxies (P_{cut})
Split (RA, Dec)
Find clusters using FOF (b_r, b_z)
Merge angular pieces
Galaxy cluster catalog

Figure 4.5: Flow diagram of steps used in the VT-FOFz cluster finder. The red boxes indicate the key steps and the associated free parameter. Here $P_{\rm cut}$ is the fraction of the densest galaxies which allows us to select the candidate galaxies to form the clusters, b_r is the transverse linking length parameter and b_z is a factor which determines the how galaxies are distributed along the line-of-sight.

the purity (**P**) and the completeness (**C**) properties of the cluster catalogs. The values of **P** and **C** depend on the definition used for determining matches between the galaxy clusters observed and the dark matter haloes, see Farrens et al. (2011); Soares-Santos et al. (2011); Hung-Yu et al. (2014). These quantities characterize the overall fidelity of the resulting galaxy cluster catalogs.

4.3.1 Matching between halos and galaxy clusters

We implement a cylindrical method for identifying matches between the dark matter haloes and the detected galaxy clusters. The mock halo catalog is rank ordered by the number of galaxy members contained within the halo N_g, whereas the cluster catalog is rank ordered by the number of galaxies in each cluster $N_{g,\rm obs}$ (i.e. the richness of the cluster). For each mock halo, we compute the radius R_{200} which is defined by a sphere of overdensity equal to 200 with respect to the mean density of the Universe. We define the matching threshold region for a halo as a cylinder with radius R_{200} around the halo center and a height in the line-of-sight given by $2{\rm d}z_{\rm max}(1+z)$, where dz is related to the associated redshift error in the catalogs. We require that any potential match be within the boundaries defined by the members of the halo (i.e. the maximum extent for RA-Dec set by the galaxy member farthest from the halo center). After that, we take the highest ranked halo and we search the highest ranked clusters detected within the matching threshold region. We repeat the same process for all haloes by keeping the rank order. Only unique matches are allowed. Therefore if there are multiple clusters with equal rank, the object closest to the halo center is chosen as the match.

The completeness is defined as the fraction of haloes which are matched with the detected galaxy clusters. On the other hand, the purity is defined as the fraction of detections that corresponds to the haloes. Thus, we have

$$\mathbf{C}(z_h, M) = \frac{N_{\text{matches}}}{N_{\text{halos}}}, \quad \mathbf{P}(z_c, N_{g,\text{obs}}) = \frac{N_{\text{matches}}}{N_{\text{clusters}}}. \tag{4.13}$$

We observe that the completeness depends on the mass (M) and the halo redshift (z_h), whereas the purity depends on the number of cluster members ($N_{g,\text{obs}}$) and on the cluster photometric redshift (z_c). The global completeness is defined as

$$\mathbf{C}_{\text{global}} \equiv \left(\sum_i^N \frac{\mathbf{C}^2(z_i, > M^*)}{N} \right)^{1/2}, \tag{4.14}$$

and the global purity is defined as

$$\mathbf{P}_{\text{global}} \equiv \left(\sum_i^N \frac{\mathbf{P}^2(z_i, > N^*_{g,\text{obs}})}{N} \right)^{1/2}, \tag{4.15}$$

for a $M > M^*$ and $N_{g,\text{obs}} > N^*_{g,\text{obs}}$ respectively. Note that we are computing the root mean square in a set of redshift bins for the completeness and purity.

4.3.2 Calibrating free parameters of the cluster finder

The VT-FOFz cluster finder has three free parameters (P_{cut}, b_r and b_z) which must be adjusted in order to detect galaxy clusters in a photometric galaxy survey. By using the mock catalogs described in section 4.1 we can calibrate the free parameters. For this analysis, we define several samples from the mock galaxy catalog, by varying the cut in the r-band magnitude in the range $19.4 < r < 20.9$ with $\text{dm}_r = 0.1$. Note that the parameter values depend on the spatial distribution of galaxies, therefore the obtained results are only valid for the mock galaxy sample used. Nevertheless, the values estimated from the mock sample allow us to have an a priori knowledge about the values which must be used at moment to run the cluster finder in the GAMA DEEP catalog.

The method for calibrating the parameter set has two phases. Initially we estimate the parameters (b^*_r, b^*_z) such that the number of matches is maximum for

Table 4.1: Values obtained for the free parameters, threshold redshift and galaxy clusters detected after maximizing the global completeness and the likelihood function.

Cut of r-band	$P^*_{\rm cut}$	b^*_r	b^*_z	z_{th}	$N_{\rm clusters}$	$N_{\rm matches}$
$r < 19.4$	0.347	1.758	1.420	0.2325	229	124
$r < 19.5$	0.400	1.585	1.409	0.2525	424	221
$r < 19.6$	0.360	1.414	1.456	0.2575	447	224
$r < 19.7$	0.401	1.157	1.578	0.2725	562	262
$r < 19.8$	0.322	1.074	1.547	0.2925	507	260
$r < 19.9$	0.434	0.792	1.992	0.3025	719	332
$r < 20.0$	0.371	0.828	1.570	0.3075	714	330
$r < 20.1$	0.345	0.587	1.938	0.3225	529	295
$r < 20.2$	0.416	0.539	1.931	0.3475	788	378
$r < 20.3$	0.387	0.515	1.957	0.3525	864	420
$r < 20.4$	0.400	0.471	1.728	0.3625	873	434
$r < 20.5$	0.350	0.454	1.705	0.3725	873	444
$r < 20.6$	0.344	0.425	1.747	0.3825	975	463
$r < 20.7$	0.341	0.409	1.550	0.4025	1 006	476
$r < 20.8$	0.416	0.318	1.914	0.4125	1 161	540
$r < 20.9$	0.362	0.314	1.910	0.4225	1 177	559

$P_{\rm cut} = 1.0$. In other words, we maximize the global completeness by varying b_r, b_z, without Voronoi cuts, then

$$b^*_r,\, b^*_z = \mathop{\arg\max}_{b_r,\, b_z} \mathbf{C}_{\rm global}(b_r, b_z), \quad P_{\rm cut} = 1.0. \tag{4.16}$$

The redshift range used here is $z \in [0, 0.5]$ and the number of bins is 10. We use the Scipy Differential Evolution method** to maximize the function given in equation (4.16). We repeat this process for all considered r-band magnitude cuts. To determine the matches between haloes and the detected galaxy clusters we set a conservative value for d$z = 0.025$, which is consistent to the value of the estimated photometric redshift scatter for the galaxy mock catalog in $z \in [0.0, 0.5]$ (i.e., $\sigma_{\rm phot} \approx 2{\rm d}z$), see chapter 3. In order to calibrate the $P_{\rm cut}$ parameter, we perform a maximum likelihood analysis by comparing the redshift distribution of the detected galaxy clusters and the halos by using Poisson statistics. We constrain the maximum redshift in the analysis for each magnitude r-band cut. Note that the Voronoi density allows us to set the densest regions. Hence, in this step, we are removing

** ⟨https://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.optimize.differential_evolution.html⟩

impurities and avoiding false detections, thus obtaining more reliable results. The limit is defined by the redshift in which the number density of galaxies according to the mock catalog begins to be very low, see figure 4.2. Here we set that the threshold redshift z_{th} for each r-band cut is defined as

$$z_{th} \equiv \max_{i}, \left\{ z_i \; : \; \frac{\mathrm{d}N}{\mathrm{d}V}(z_i) > 10^{-3}\, h^3 \,\mathrm{Mpc}^{-3} \right\}. \tag{4.17}$$

The likelihood function used in this analysis is given by

$$L(C|H) = \prod_k \frac{H_k^{C_k} \exp\left(-H_k\right)}{C_k!}, \tag{4.18}$$

where C_k are the galaxy cluster number counts and H_k are the halo number counts for the k-th redshift bin.

Figure 4.6 shows $-\ln(L)$ as a function of P_{cut} for all r-band cuts. Note that the values which maximize the likelihood for all magnitude cuts are in the range $P_{\mathrm{cut}} \in [0.32, 0.44]$. This means that approximately 40% of the densest galaxies are candidates to form galaxy clusters, regardless of the cut in r-band magnitude. Table 4.1 shows the threshold redshift, the values obtained for P^*_{cut}, b^*_r and b^*_z, and the detected number of clusters by using the method described in this section for all magnitude cuts. We observe that the b_r parameter decreases for deeper magnitude cuts. This effect is due to the fact that for deeper cuts we have more available galaxies, hence the transversal linking length tends to be smaller. The above effect is opposite in the line-of-sight linking length because this parameter depends on the photometric redshift quality, therefore it is important to remark that, for deeper cuts, b_z tends to be greater as redshift quality gets worse.

Figure 4.7 shows number counts as a function of redshift for the dark matter haloes, galaxy clusters without cut in the Voronoi density and galaxy clusters with cut in the Voronoi density determined by the value which maximizes the likelihood function, equation (4.18). Here, we use the values b^*_r and b^*_z which maximize the global completeness. We observe that the P_{cut} parameter allows us to select the galaxy clusters with the densest galaxies which are better adjusted to the halo number in each redshift bin. This result is important for performing a cosmological analysis. For deeper r-band cuts, we note that the galaxy cluster number counts adjust better to the halo number counts. Nevertheless, we note that in those cases, the estimated photometric redshifts are of low quality, see chapter 3. Therefore, we could have problems to detect the true galaxy clusters in deeper cuts (i.e, galaxy cluster catalog with high completeness and low purity).

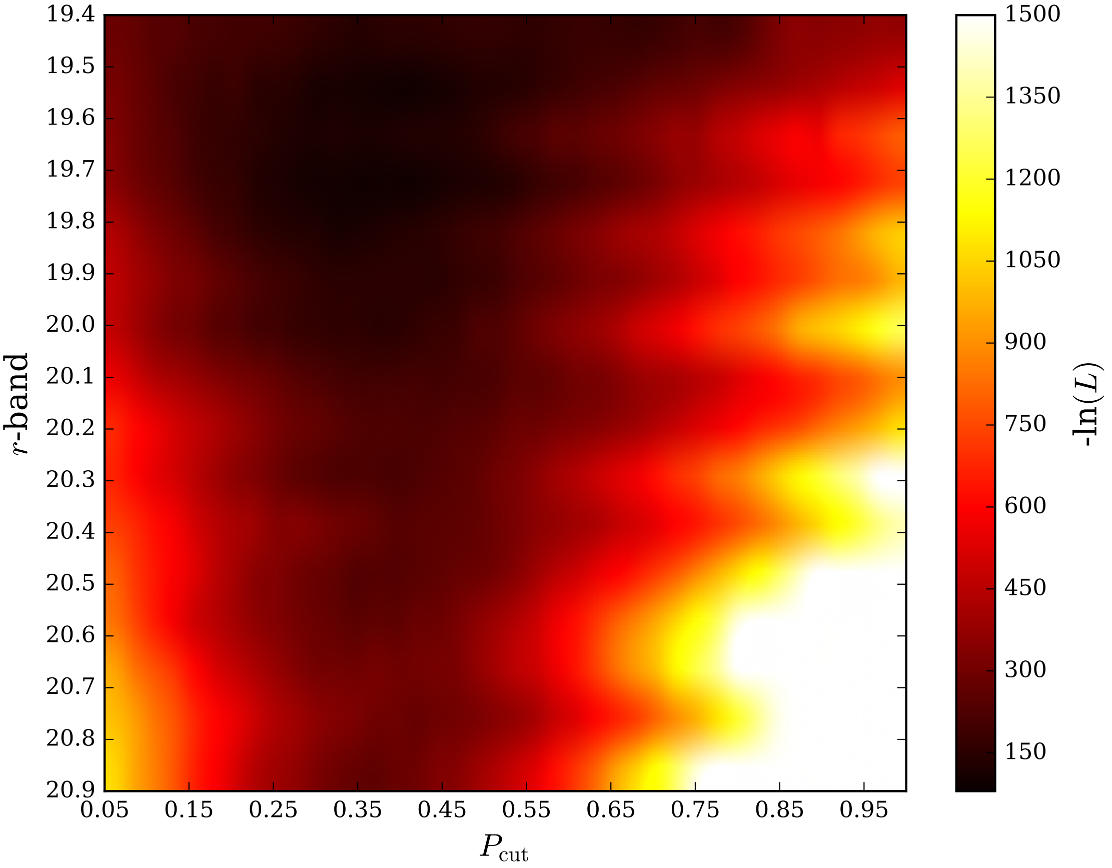

Figure 4.6: Likelihood function against P_{cut} free parameter for each r-band magnitude cut. We observe that the maximum value of likelihood independent of r-band cut is in the range $P_{\mathrm{cut}} \in [0.32, 0.44]$. This means that approximately the 40% of the densest galaxies are candidates to form galaxy clusters, independently of the magnitude cut.

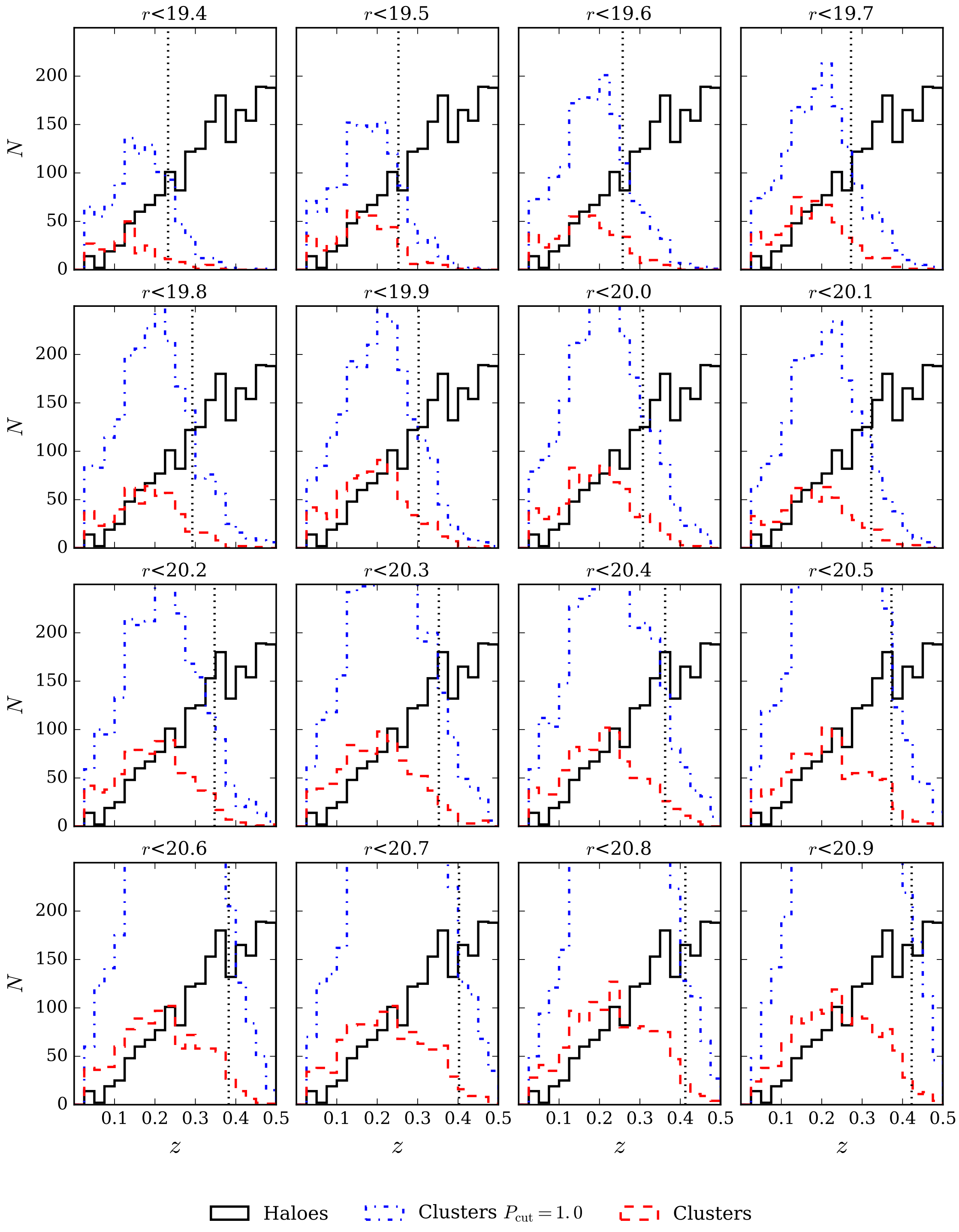

Figure 4.7: Number counts versus redshift for haloes and clusters with $P_{\rm cut} = 1.0$, and clusters with $P_{\rm cut}$ equal to the value which maximizes the likelihood function, equation (4.18). Note that in the case without the cut in the Voronoi density, the galaxy cluster number counts overestimate the halo number counts. Therefore, the value of $P_{\rm cut}$ allows us to select the galaxy clusters that are better adjusted to the halo number in each redshift bin. The dotted black line represents the threshold redshift for each r-band cut.

4.3.3 Completeness and purity

We perform a statistical comparison between the detected galaxy clusters and the dark matter haloes. For assessing the performance of the VT-FOFz cluster finder, we compute the completeness and purity. We employ the values given in table 4.1 for the free parameters $P_{\rm cut}$, b_r and b_z.

Figure 4.8 and figure 4.9 show the completeness and purity respectively according to the definition presented in section 4.3.1. The completeness depends on the mass of haloes and the purity depends on the number of the cluster members. Since both of these quantities depend on the redshift, we split the halo sample and the cluster catalogs in redshift bins with $\mathrm{d}z \approx 0.17$. For both figures we have used the following convention: the black solid line is the bin $z \in [0.0, 0.17]$, the red dashed line is the bin $z \in [0.17, 0.33]$ and the blue dotted line is the bin $z \in [0.33, 0.5]$.

For high redshift, we detect less clusters according to the density of galaxies (see figure 4.2) and as such the completeness worsens. Nevertheless, a large amount of the few detected galaxy clusters are matched with the haloes in that redshift region, thus implying high purity. In deeper cuts we see that the completeness and purity improve for all redshift bins. Since we are adding fainter galaxies to the mock sample, we can detect new clusters in regions where apparently there were no galaxy clusters according to conservatives r-band cuts. On the other hand, despite r-band cuts, we obtain good values of completeness and purity for massive haloes and clusters with a large number of members. Figure 4.10 shows the scatter plot between the mass of haloes and the number of cluster members for the matches. The massive halos tends to match clusters with a large number of members as we expected, nevertheless we note that there is a large scatter in the observable-mass relation.

Figure 4.11 shows the scatter plot between redshift of haloes and redshift of clusters for the matches. For deeper r-band cuts, we achieve matches in higher redshifts, although the dispersion increases. Moreover, in these cases we obtain more pure and complete catalogs. Therefore, we observe that the low quality of the photometric redshift for deeper r-band cuts do not produce a high impact in measuring completeness and purity. The most important factor is to have a photometric redshift distribution in agreement with the true redshift distribution. In the next section we will perform a more robust statistical analysis with the detected galaxy clusters from the GAMA DEEP catalog.

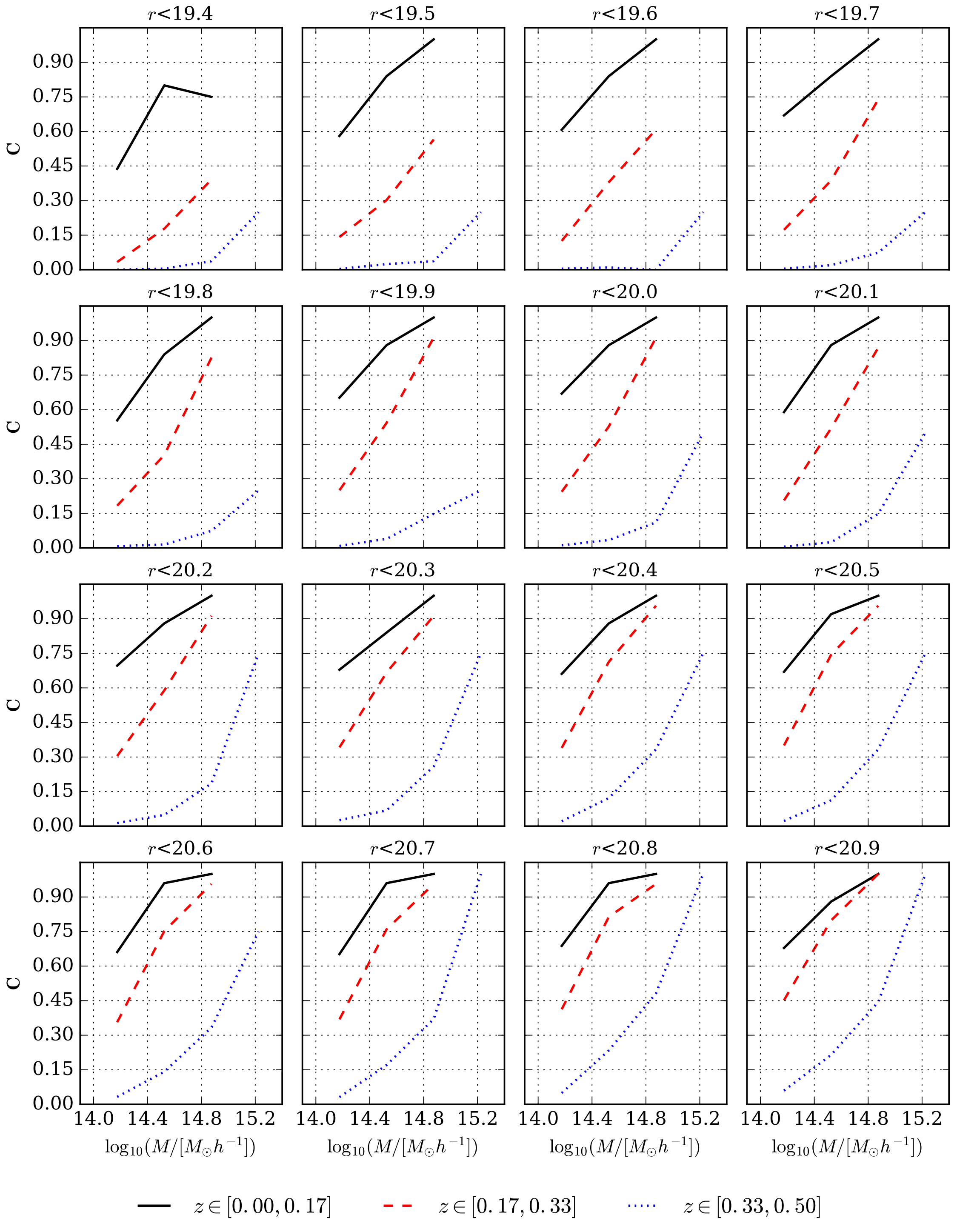

Figure 4.8: Completeness versus the mass of dark matter haloes for three redshift bins in the range $z \in [0.0, 0.5]$. Note that the completeness worsens as the redshift increases for all magnitude cuts. Moreover, for high mass, the completeness is greater than for low mass as expected. For deeper r-band magnitude cuts and low redshifts, we attain high completeness (> 0.8) for $\log_{10} M > 10.4$.

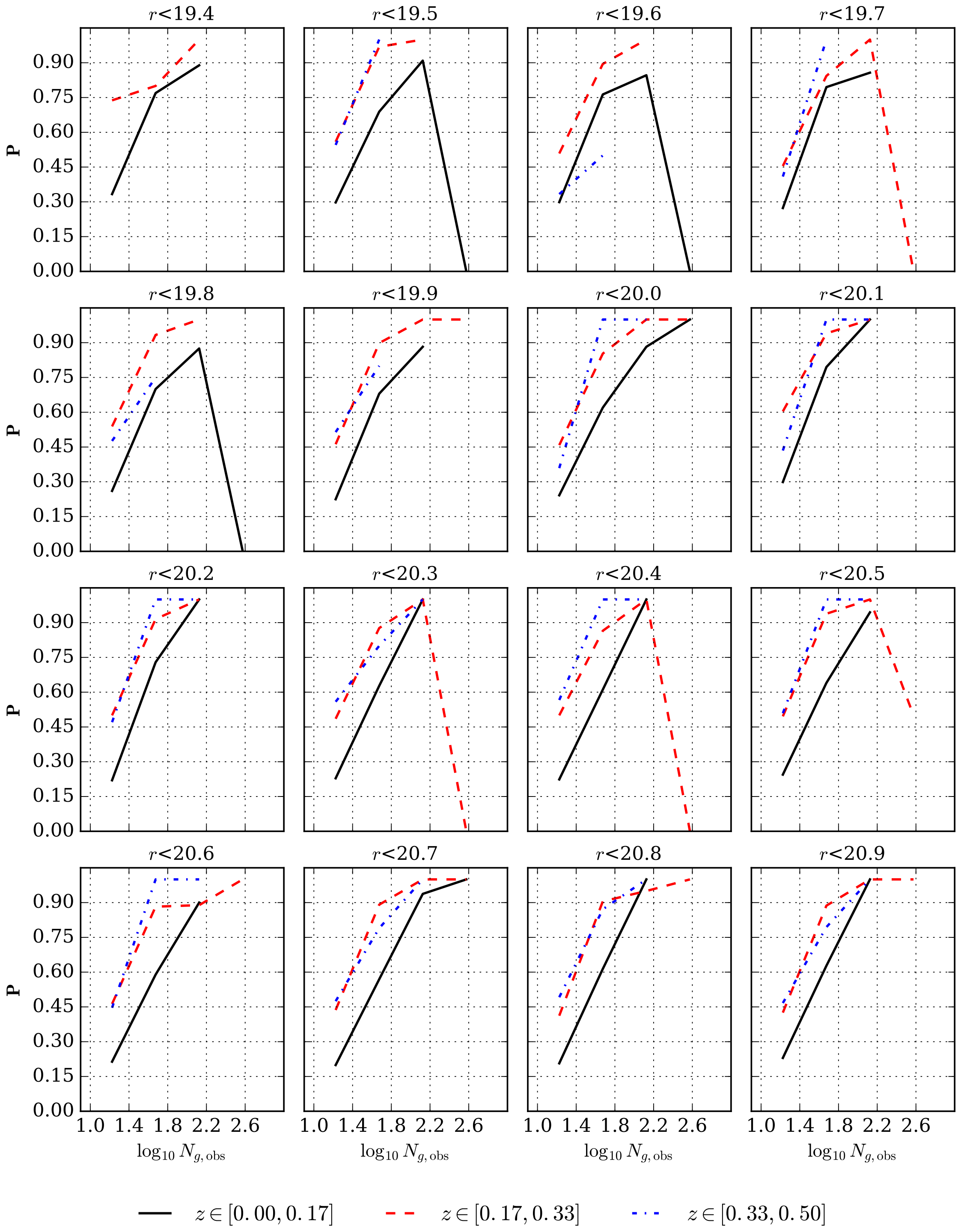

Figure 4.9: Purity versus the number of cluster members for three redshift bins in the range $z \in [0.0, 0.5]$. The purity improves considerably for deeper r-band cuts. Moreover, unlike the completeness, for low redshift we observe that the purity decreases. For the last two redshift bins and for deeper r-band magnitude cuts, we attain high purity (> 0.8) for $N_{g,\mathrm{obs}} > 40$.

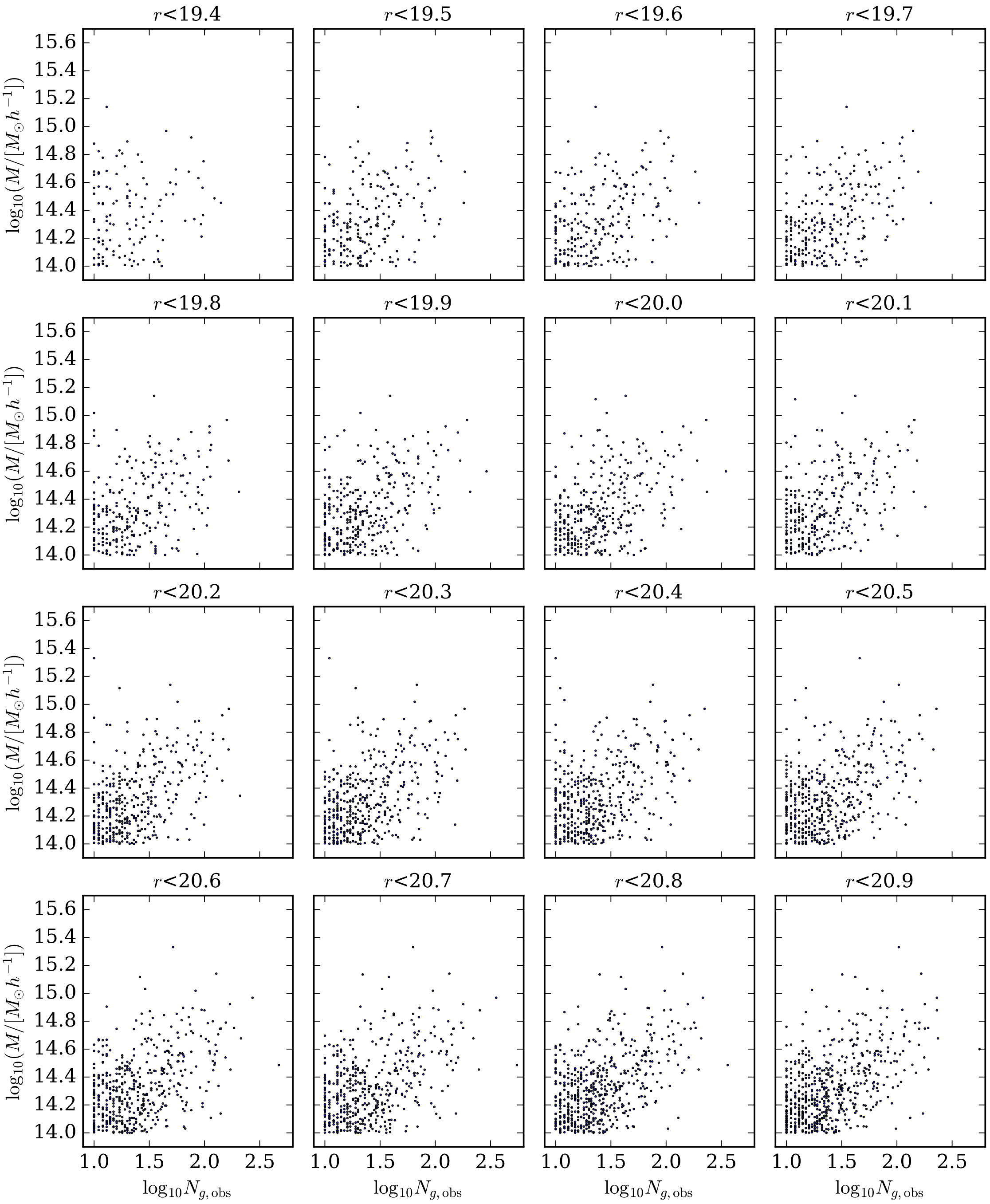

Figure 4.10: Scatter plot of the mass of haloes as a function of the number of cluster members for the matches. The massive halos tend to match clusters with more members as expected, nevertheless we note that there is a large scatter in the observable-mass relation.

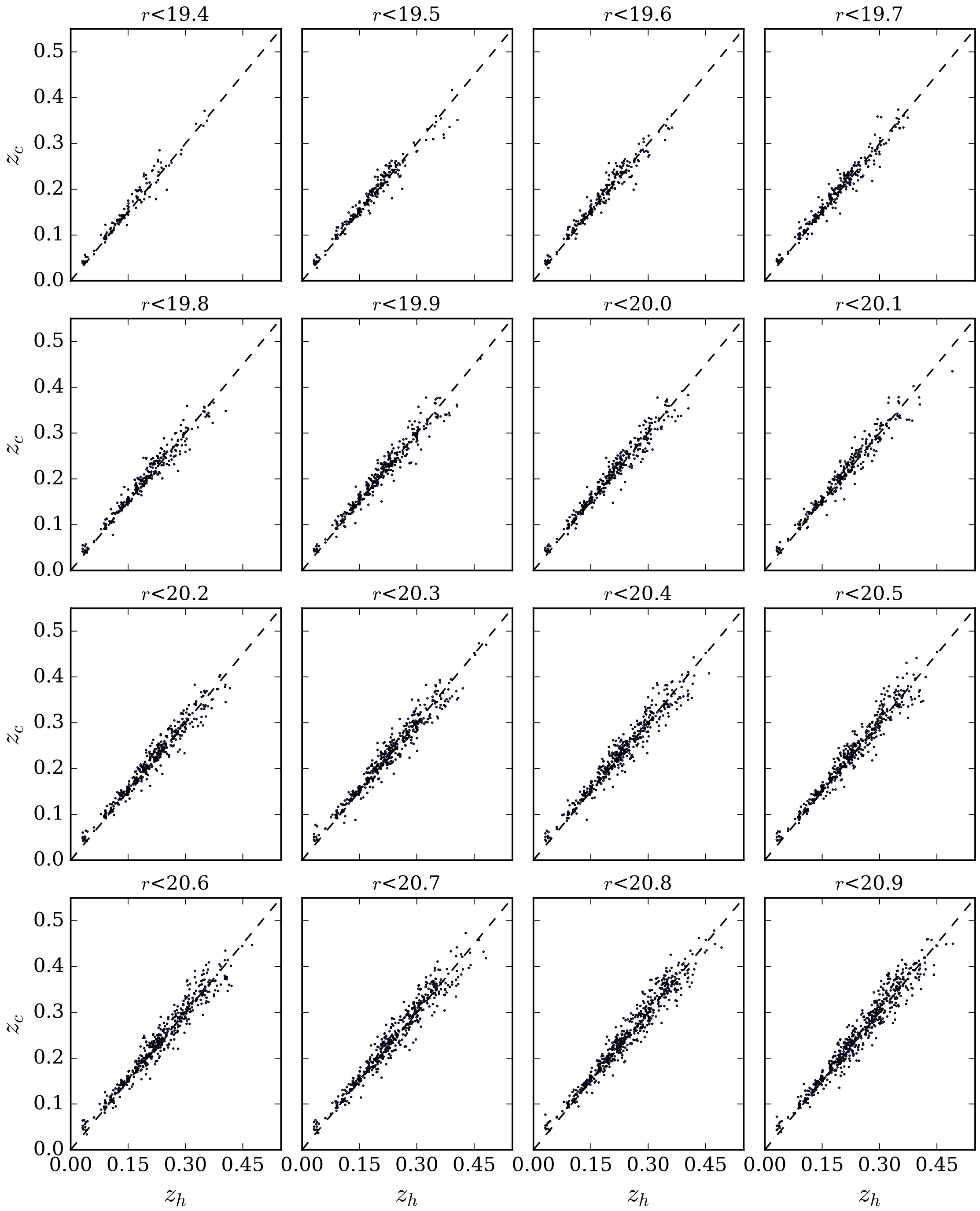

Figure 4.11: Scatter plot of the redshift of haloes as a function of the redshift of clusters for the matches. We note that the scatter is low including deeper r-band cuts. For $r > 20.6$, we achieve matches at higher redshift, although the dispersion increases. The low scatter is related to the conservative selection of dz.

4.4 GAMA DEEP galaxy cluster catalog

We consider a subsample from the GAMA DEEP survey, we select the galaxies that lie in a circle centered at RA, Dec = [180, 30] and radius 30 deg for a total area $\approx 2763\,\mathrm{deg}^2$ in the sky. For this analysis we consider r-band magnitude cuts in the range $r \in [20.6, 20.9]$ with $\mathrm{d}m_r = 0.1$. Recall that for these magnitude cuts we obtain the best results of completeness and purity in the mock catalogs, see section 4.3.3. Moreover, in these cases, the number density of galaxies experiences few changes, see figure 4.2 and figure 4.3.

Given that the spatial distribution in the GAMA DEEP catalog is different to the spatial distribution in the mock catalog, we choose different values for the fraction of the densest galaxies and the transverse linking length parameter. In order to analyze the impact of the Voronoi selection and the linking length parameter, we employ the following set of values: $P_{\mathrm{cut}} = \{1.0,\ 0.7,\ 0.5,\ 0.2\}$ and $b_r = \{0.8b_r^*,\ 0.7b_r^*,\ 0.6b_r^*\}$, where b_r^* is the value given in table 4.1 for each r-band magnitude cut. Here, we use the b_z^* value estimated for the mock catalogs. Unlike the mock catalog, we set the condition $N_g > 5$ for the number of galaxy members in the detected clusters.

4.4.1 Comparison between GAMA DEEP cluster catalog and redMaPPer SDSS DR8

In this section we compare the galaxy clusters detected using the VT-FOFz cluster finder in the GAMA DEEP catalog with those listed in the redMaPPer SDSS DR8 catalog (Rykoff et al. (2014)). We select a sample of the redMaPPer SDSS DR8 which lies in the same angular region chosen for the sample of the GAMA DEEP catalog given previously.

In order to find the matches between the galaxy clusters in the GAMA DEEP catalog and the redMaPPer SDSS DR8 catalog, we develop a similar code to that described in section 4.3.1. Initially the redMaPPer SDSS DR8 catalog is rank ordered by number of galaxy members contained within the cluster N_g and we set a matching threshold region for every cluster in the catalog. The region is defined by $\Delta z = 2\mathrm{d}z(1+z)$, $\Delta\mathrm{RA} = \mathrm{RA}_{\mathrm{max}} - \mathrm{RA}_{\mathrm{min}}$ and $\Delta\mathrm{Dec} = \mathrm{Dec}_{\mathrm{max}} - \mathrm{Dec}_{\mathrm{min}}$. To consider the photometric redshift error in both redMaPPer clusters and GAMA DEEP clusters, we set $\mathrm{d}z = 0.075$. Afterwards, we look for all the GAMA DEEP clusters which lie in the matching threshold region. All matches must satisfy the following criterion

$$Cl_{\mathrm{match}} = \min_{i \in U} \left\{ \left| 1 - \frac{N^i_{g,\mathrm{rM}}}{N^i_{g,\mathrm{GD}}} \right| D^i_{\mathrm{ang}} \right\}, \tag{4.19}$$

where U is the set of galaxy clusters that inhabit the threshold region, $N^i_{g,\mathrm{rM}}$ is the number of galaxy members in the redMaPPer cluster, $N^i_{g,\mathrm{GD}}$ is the number of galaxy members the GAMA DEEP cluster and D^i_{ang} is the angular distance between the cluster candidate to be matched and the center of the threshold region. To avoid multiple matching with the same galaxy cluster, we remove from the GAMA DEEP catalog those clusters which have been selected as matching. We use equation (4.13) to estimate the completeness and purity. We treat redMaPPer clusters in the same way as the simulated dark matter haloes, hence the completeness measures the fraction of redMaPPer clusters that we have been detected using the VT-FOFz cluster finder. We run the VT-FOFz cluster finder on the GAMA DEEP catalog performing two tests. In the first test, we fix the value of the transverse linking length and we vary the P_{cut} parameter, we assume the following values $b_r = 0.8b^*_r$ and $P_{\mathrm{cut}} = 1.0, 0.5, 0.2$. In the second test, we fix the value of the P_{cut} parameter and we vary the transverse linking length, we assume the following values $P_{\mathrm{cut}} = 0.7$ and $b_r = 0.8b^*_r, 0.7b^*_r, 0.6b^*_r$. Then, the GAMA DEEP catalogs obtained in both tests are compared with the redMaPPer catalog. Figure 4.12, figure 4.13, figure 4.14 and figure 4.15 correspond to the first test. Figure 4.16, figure 4.17, figure 4.18 and figure 4.19 correspond to the second test. These figures show the completeness and purity, the scatter plot between $N_{g,\mathrm{GD}}$ and $N_{g,\mathrm{rM}}$, the scatter plot between z_{GD} and z_{rM}, and the scatter scatter plot between $N_{g,\mathrm{GD}}$ and R_{ang} for matched (red) and non-matched (blue) GAMA DEEP clusters. We observe that the Voronoi selection as well as the transverse linking length affect the detection of galaxy clusters in the same way. By selecting a fraction of the galaxies using the P_{cut} (i.e., those regions which contains the densest galaxies), we avoid detecting large fake galaxy clusters and merging effects.

Figure 4.15 and figure 4.19 show that the P_{cut} and b_r parameters affect the size and the richness (i.e., number of galaxy members for our cluster finder) in the same way. For lower values of P_{cut} and b_r, the detection of large and/or rich galaxy clusters decreases, mainly those clusters which do not match with the redMaPPer clusters. Nevertheless, low values of these parameters imply missed detections of true galaxy clusters, therefore we must choose the values with care. Figure 4.12 and figure 4.16 show that the completeness and purity vary with the Voronoi selection and the transverse linking length. We note that the purity increases and the completeness decreases for low values of P_{cut} and b_r. This effect is stronger for the P_{cut} parameter because a strong Voronoi selection would only allows us to choose the galaxies that lie in the densest regions and we could be rejecting true galaxy clusters with low Voronoi density. Note that for the cases ($P_{\text{cut}} = 1.0,\ b_r = 0.8b_r^*$), ($P_{\text{cut}} = 0.7,\ b_r = 0.8b_r^*$) and ($P_{\text{cut}} = 0.7,\ b_r = 0.7b_r^*$) with $N_{g,\text{GD}} \geq 5$, we recover a large number of RedMaPPer clusters ($\mathbf{C} > 0.9$) up to $z \approx 0.33$, however the purity is very low. The two cases where Voronoi selection was used present higher purity, this result is in agreement with the above assertion. It shows us that we can keep high values of completeness and improve the purity by using the Voronoi selection.

Figure 4.13 and figure 4.17 show that $N_{g,\text{GD}}$ presents a positive correlation with $N_{g,\text{rM}}$ as is expected. Both parameters P_{cut} and b_r have a similar effect on the richness. The low values avoid detecting rich clusters, which improves the scatter relation. However, in extreme cases (i.e., very low values, $P_{\text{cut}} = 0.2$ and $b_r = 0.6b_r^*$) we begin to observe a constant bias as the detected clusters can only be formed by the closest galaxies to each other and several members are rejected. Figure 4.14 and figure 4.18 show the scatter relation between the photometric redshifts of the GAMA DEEP clusters and the photometric redshifts of the redMaPPer clusters. We observe that for high redshifts there is a bias, which is due to the low quality of galaxy photometric redshifts in this region. Moreover, we note that the scatter in the relation $z_{\text{GD}} - z_{\text{rM}}$ is less sensitive to the variation of b_r parameter whereas the Voronoi selection relieves the large scatter. This effect is expected because, when choosing the densest galaxies, we are constraining the regions wherein the SFOF code would search for clusters, thus avoiding satellite members which could affect the selection of the redshift.

Figure 4.20 compares the redshift distribution in the three cases described above, in which we recover a large number of redMaPPer galaxy clusters (i.e., for $\mathbf{C} > 0.9$). We consider the following threshold richness $N_{g,\text{GD}} \geq 15$ and $N_{g,\text{GD}} \geq 20$ according to the threshold richness used in the redMaPPer catalog (Rykoff et al. (2014)). For $P_{\text{cut}} = 1.0,\ b_r = 0.8b_r^*$ case, we observe a large excess of clusters detected by the VT-FOFz for both cases $N_{g,\text{GD}} \geq 15$ and $N_{g,\text{GD}} \geq 20$. The reduction of P_{cut} improves the results, however the histogram of GAMA DEEP clusters does not match with the histogram of redMaPPer. If we combine the effect of two parameters (i.e., we reduce the transverse linking length and we select less densest galaxies) the results improve. Note that for the cases ($P_{\text{cut}} = 0.7,\ b_r = 0.7b_r^*,\ N_{g,\text{GD}} \geq 15,\ r < 20.8$)

and ($P_{\rm cut} = 0.7,\ b_r = 0.7b_r^*,\ N_{g,{\rm GD}} \geq 20,\ r < 20.6$) the GAMA DEEP histograms match the redMaPPer histogram up to $z \approx 0.33$. Therefore, we can assert that these combinations of parameters are reliable results. We remark that the GAMA DEEP cluster catalog must not necessarily have the same cluster selection as the redMaPPer catalog, because the used optical methods are different. For GAMA DEEP cluster catalog we employ a spatial method (i.e., we use the angular position and the photometric redshifts), whereas the redMaPPer catalog utilizes the cluster red sequence method.

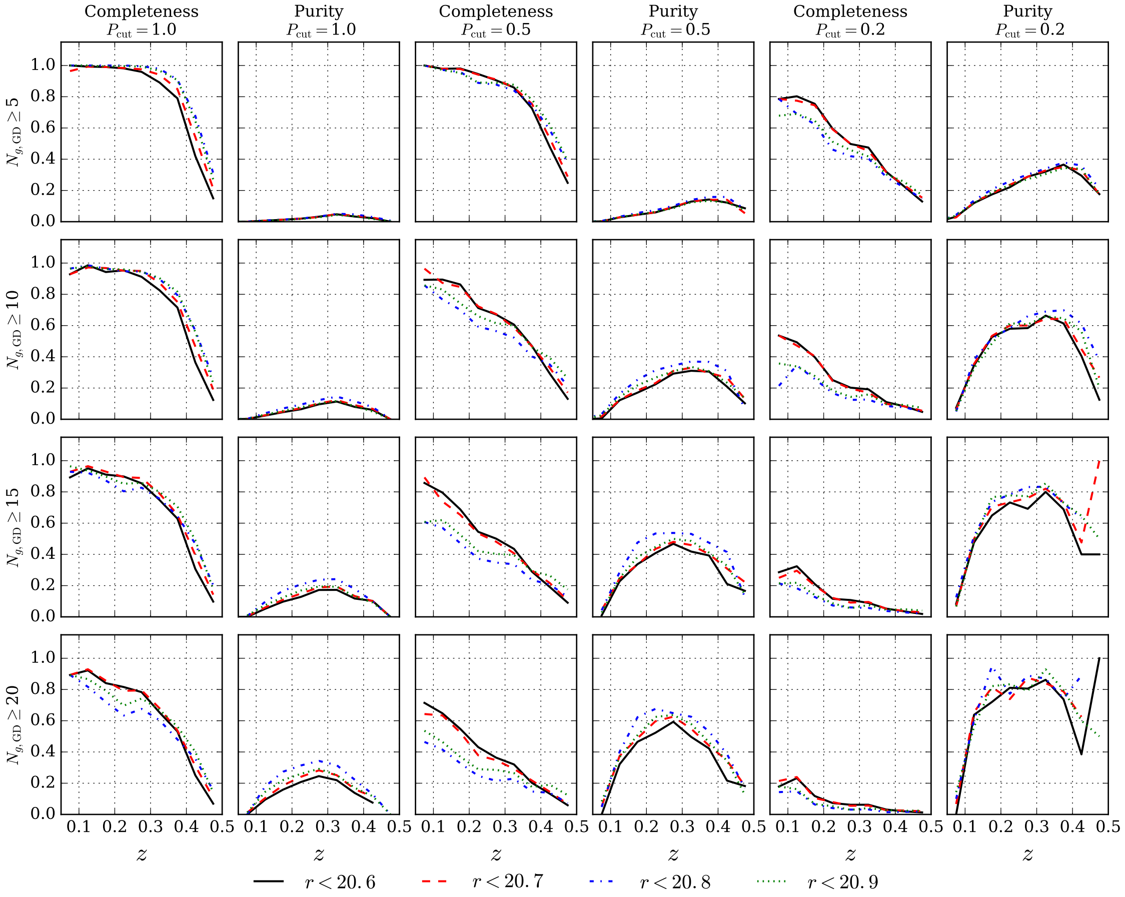

Figure 4.12: Completeness and purity between redMaPPer clusters and GAMA DEEP clusters as function of redshift for $P_{\rm cut} = \{1.0,\ 0.5,\ 0.2\}$ and $b_r = 0.8b_r^*$. For $N_{g,{\rm GD}} \geq 5$, $P_{\rm cut} = 1.0$ and $b_r = 0.8b_r^*$, we recover all redMaPPer clusters up to $z \approx 0.3$, nevertheless the cluster catalog has too many impurities. If we consider a higher threshold of the number of cluster members, we improve the purity but we lose in completeness mainly at high redshifts. For the other cases we observe that the Voronoi selection improves the purity, although the completeness decreases.

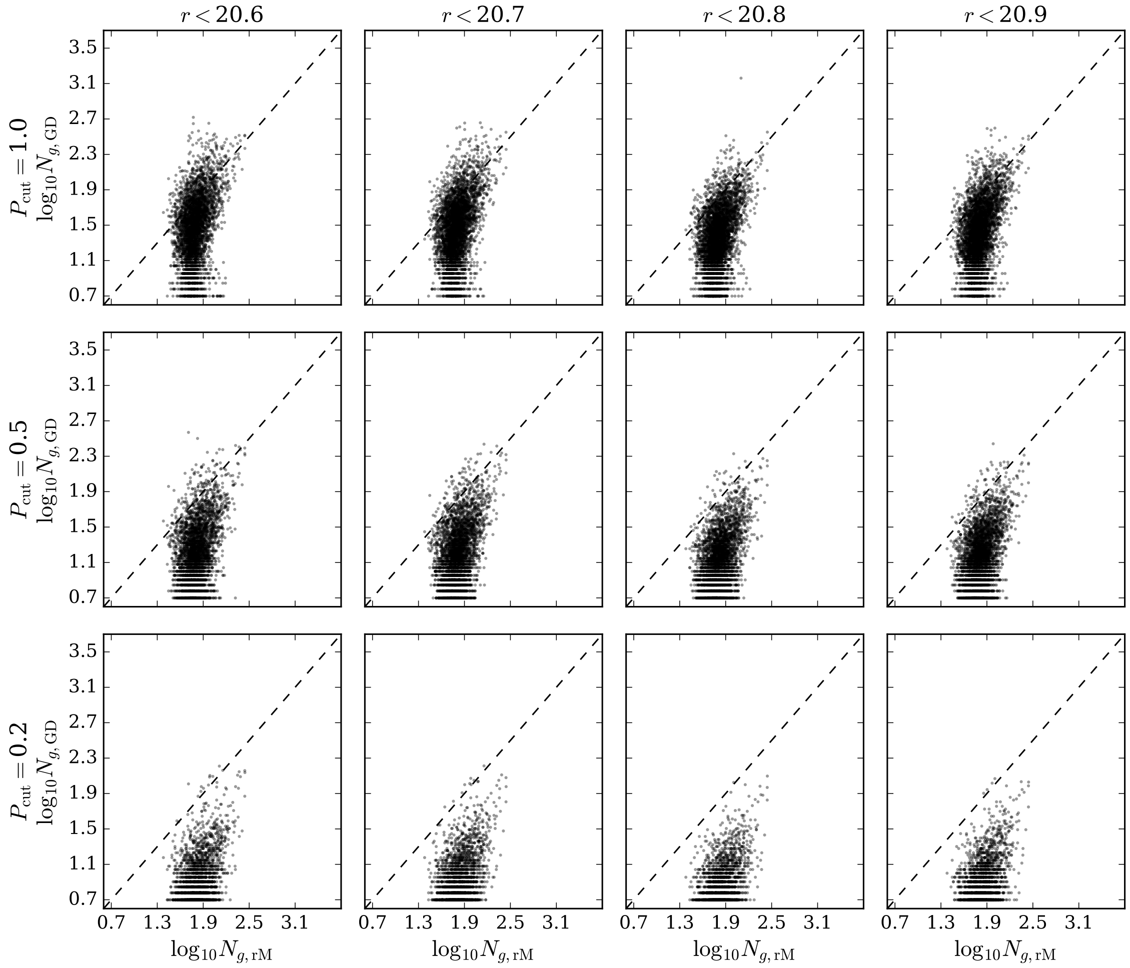

Figure 4.13: Scatter plot of the number of galaxy members of the GAMA DEEP clusters (GD) versus the number of galaxy members of the redMaPPer clusters (rM) for $P_{\rm cut} = \{1.0,\ 0.5,\ 0.2\}$ and $b_r = 0.8b_r^*$. The Voronoi selection improves the scatter relation $N_{g,\rm GD} - N_{g,\rm rM}$. However, for $P_{\rm cut} = 20$, we begin to observe a bias, which is due to the low threshold in the selection of the densest galaxies. This effect is more evident in deeper r-band magnitude cuts.

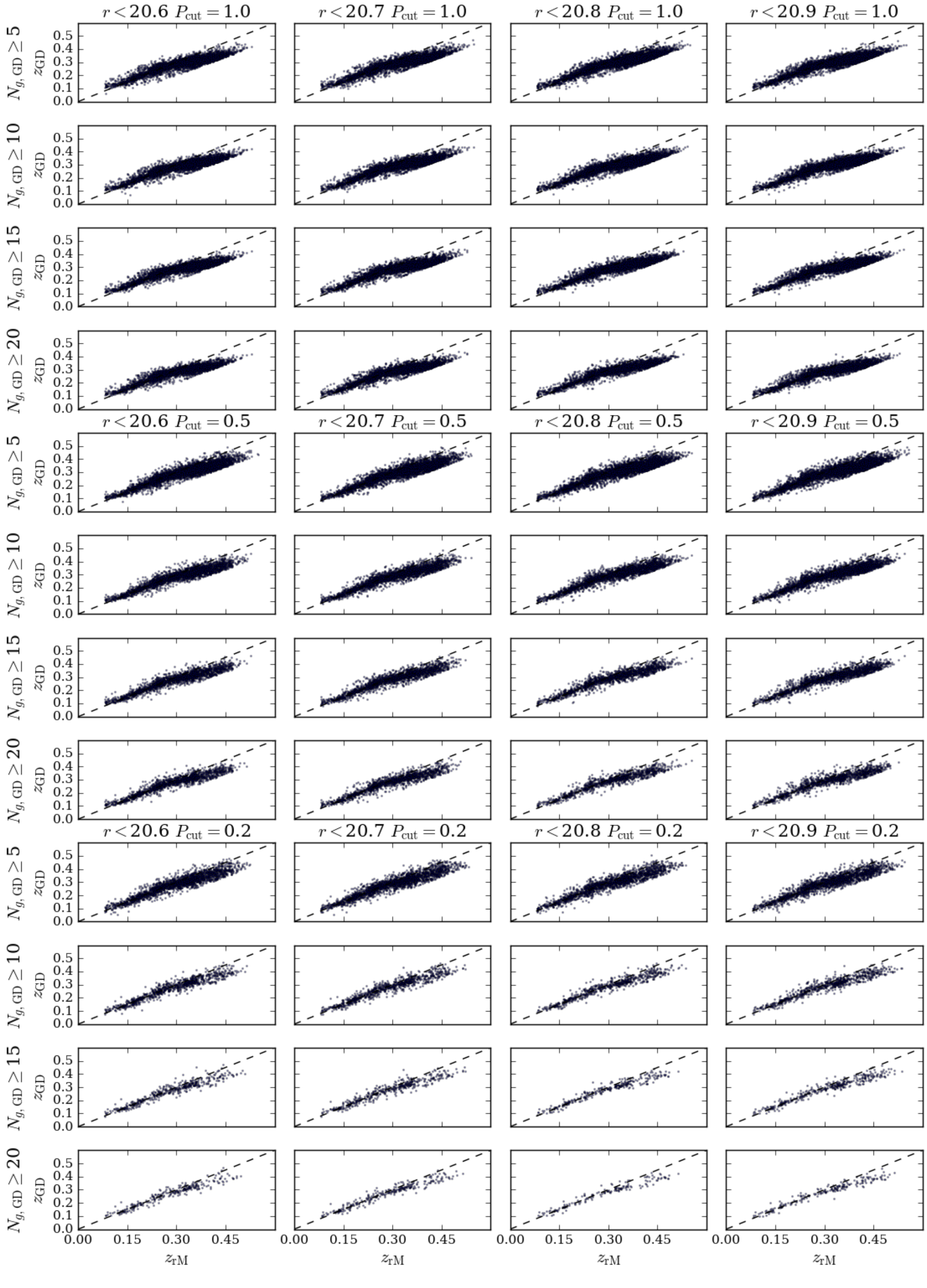

Figure 4.14: Scatter plot of the photometric redshift of the GAMA DEEP clusters (GD) versus the photometric redshift of the redMaPPer clusters (rM) for $P_{\rm cut} = \{1.0,\ 0.5,\ 0.2\}$ and $b_r = 0.8 b_r^*$. Note that the scatter improves for higher threshold of the number of cluster members. The Voronoi selection relieves the large scatter. For high redshifts, the relation presents a bias due to the low quality of galaxy photometric redshifts in this region.

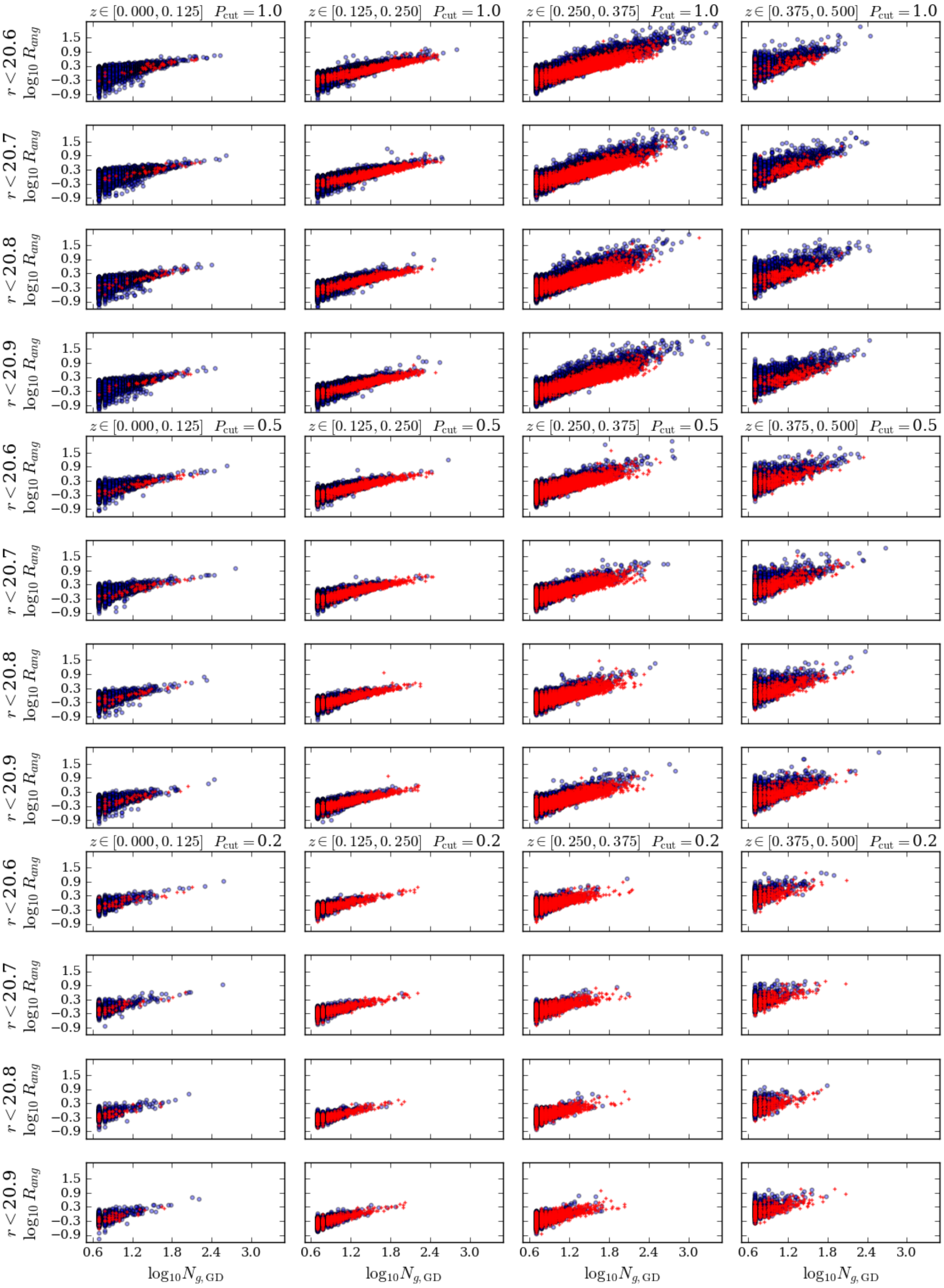

Figure 4.15: Scatter plot of the number of galaxy members and the cluster radius in arcmin for the matched GAMA DEEP clusters (red) versus the non-matched GAMA DEEP clusters (blue) by using 4 redshift bins in the range $z_{\rm GD} \in [0.0,\ 0.5]$ ($P_{\rm cut} = \{1.0,\ 0.5,\ 0.2\}$ and $b_r = 0.8b_r^*$). Note that for high values of the $P_{\rm cut}$ parameter we detect a large number of big and rich galaxy clusters which are not matched with the redMaPPer clusters. The Voronoi selection relieves this problem. On the other hand, for the lowest redshift bin and the highest redshift bin we detect many non-matched clusters and the effect is greater for low redshifts.

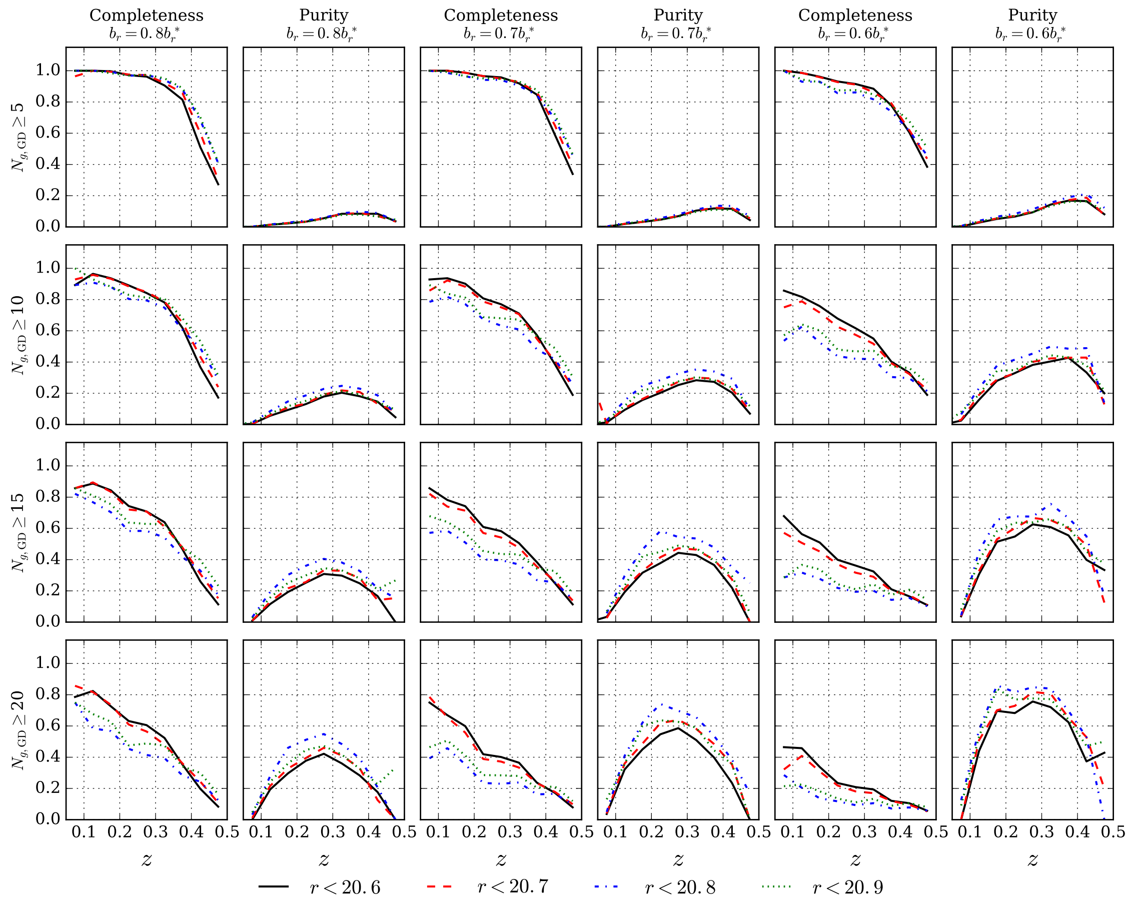

Figure 4.16: Completeness and purity between redMaPPer clusters and GAMA DEEP clusters as function of redshift for $P_{\mathrm{cut}} = 0.7$ and $b_r = \{0.8b_r^*,\ 0.7b_r^*,\ 0.6b_r^*\}$. For $N_{g,\mathrm{GD}} \geq 5$, $P_{\mathrm{cut}} = 0.7$, $b_r = 0.8b_r^*$ and $b_r = 0.7b_r^*$, we recover most redMaPPer clusters until $z \approx 0.3$, nevertheless the cluster catalog has many impurities. If we consider a higher threshold of the number of cluster members, we improve the purity but we lose in completeness mainly at high redshifts. Note that the reduction of the transverse linking length improves the purity, although the completeness decreases.

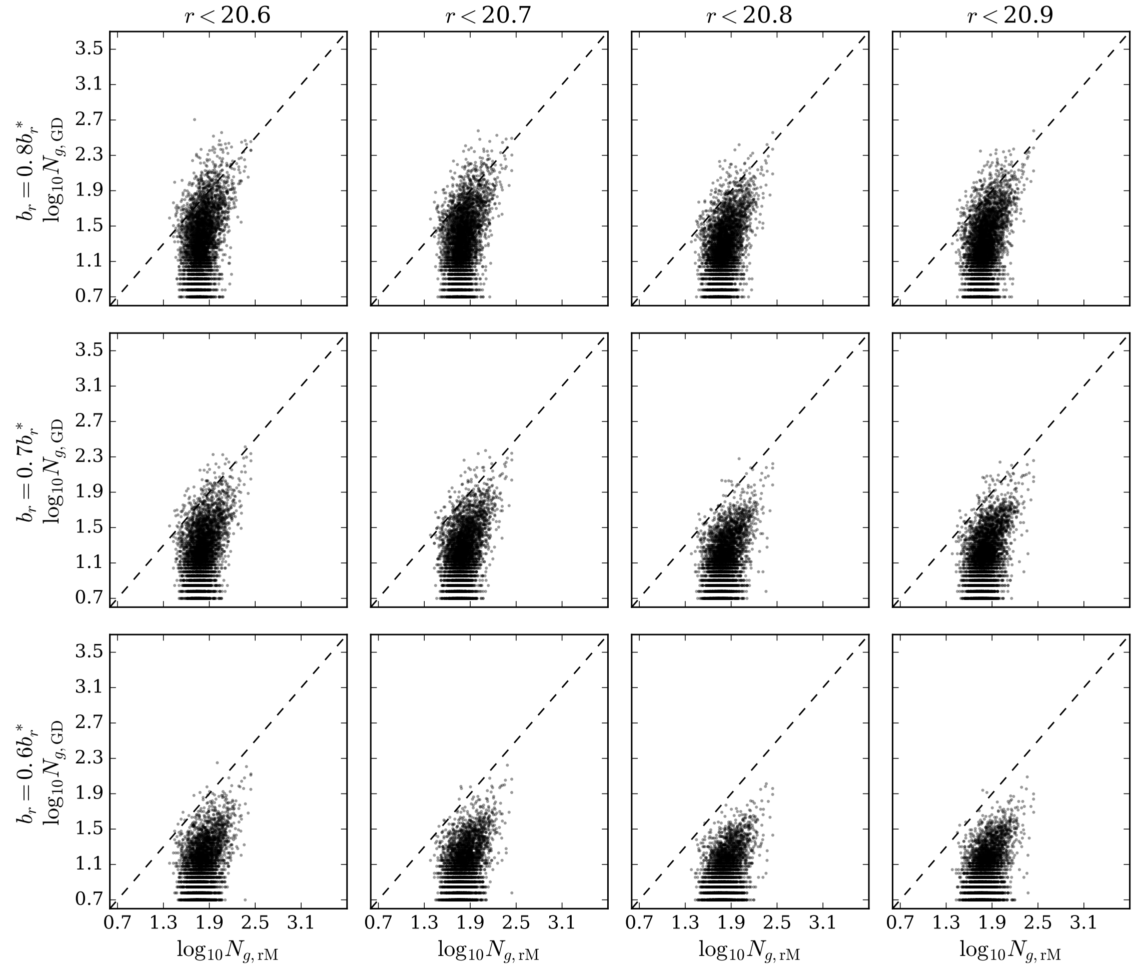

Figure 4.17: Scatter plot of the number of galaxy members of the GAMA DEEP clusters (GD) versus the number of galaxy members of the redMaPPer clusters (rM) for $P_{\mathrm{cut}} = 0.7$ and $b_r = \{0.8b_r^*,\ 0.7b_r^*,\ 0.6b_r^*\}$. The scatter in the relation $N_{g,\mathrm{GD}} - N_{g,\mathrm{rM}}$ improves for lower b_r values. Nevertheless, for $b_r = 0.6b_r^*$, we begin to note a bias due to the reduction of the transverse linking length. For deeper r-band magnitude cuts this effect is more evident.

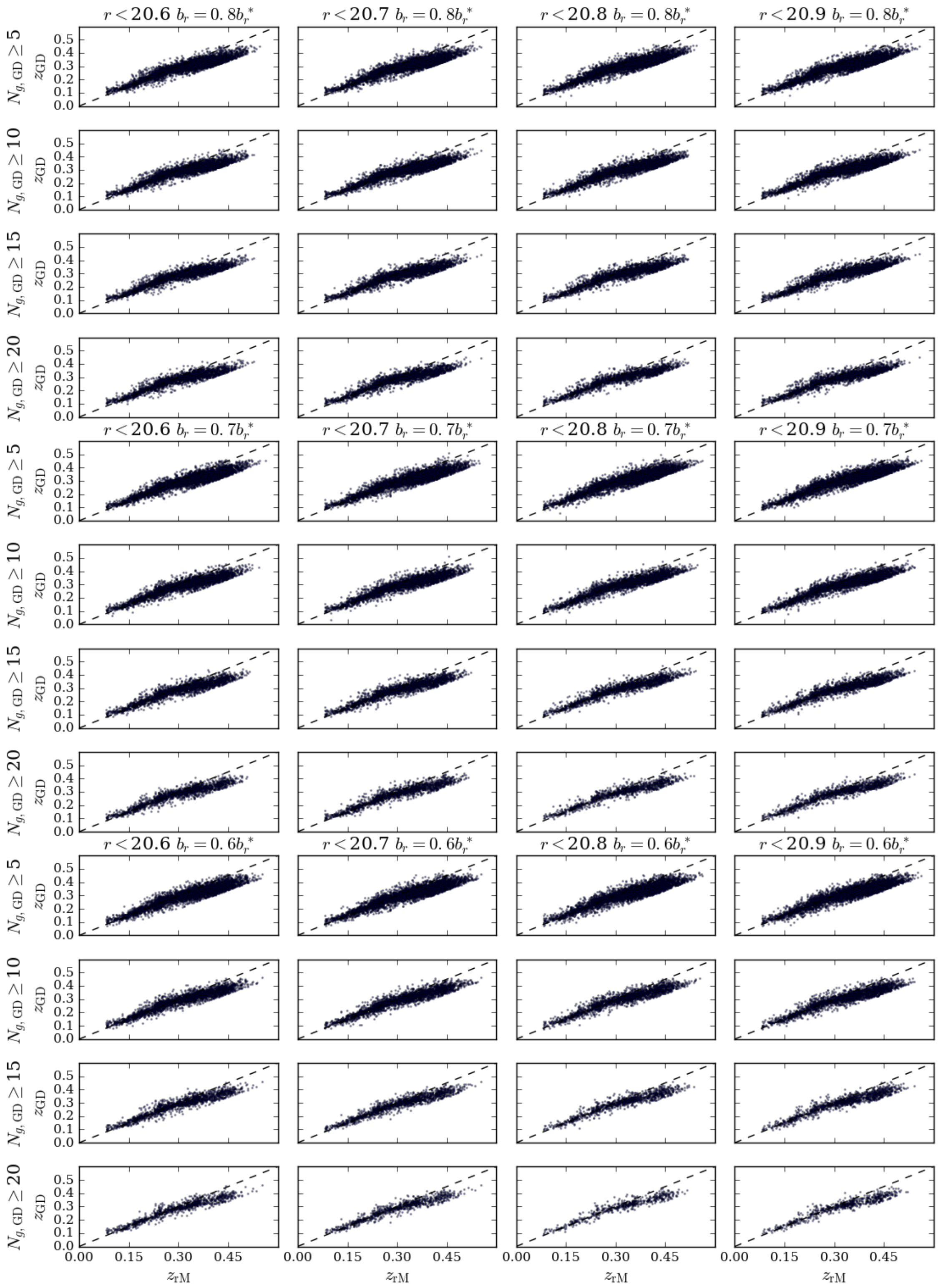

Figure 4.18: Scatter plot of the photometric redshift of the GAMA DEEP clusters (GD) versus the photometric redshift of the redMaPPer clusters (rM) for $P_{\mathrm{cut}} = 0.7$ and $b_r = \{0.8b_r^*,\ 0.7b_r^*,\ 0.6b_r^*\}$. The scatter in the relation $z_{\mathrm{GD}} - z_{\mathrm{rM}}$ is less sensitive to the variation of the b_r parameter. For high redshifts, the relation presents a bias due to the low quality of galaxy photometric redshifts in this region.

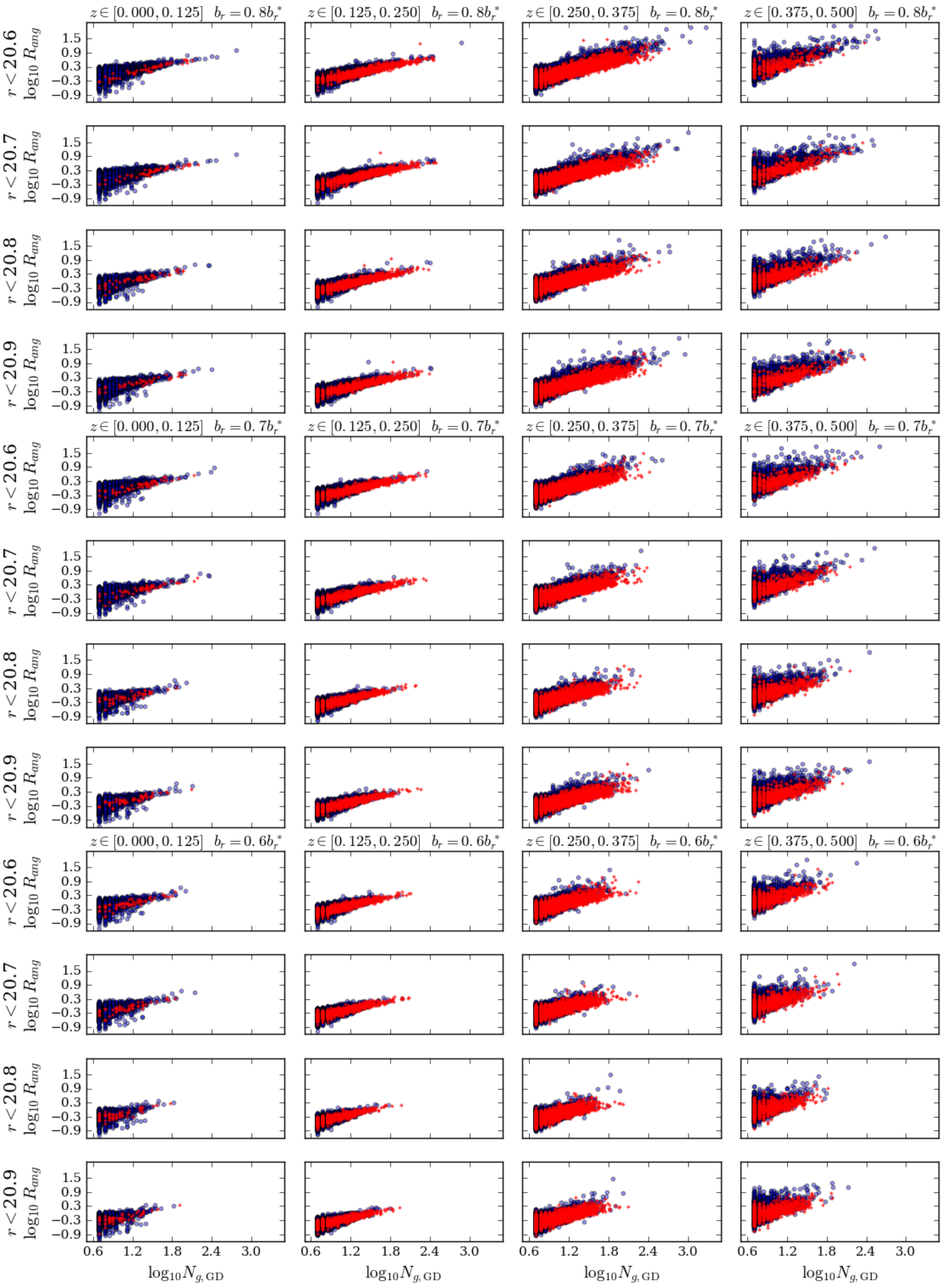

Figure 4.19: Scatter plot of the number of galaxy members and the cluster radius in arcmin for the matched GAMA DEEP clusters (red) versus the non-matched GAMA DEEP clusters (blue) by using 4 redshift bins in the range $z_{\mathrm{GD}} \in [0.0, 0.5]$ ($P_{\mathrm{cut}} = 0.7$ and $b_r = \{0.8b_r^*, 0.7b_r^*, 0.6b_r^*\}$). Note that for high values of the b_r parameter we detect a large number of big and rich galaxy clusters which are not matched with the redMaPPer clusters. The reduction of the transverse linking length relieves this problem. On the other hand, for the lowest redshift bin and highest redshift bin we detect many non-matched clusters and the effect is greater for low redshifts.

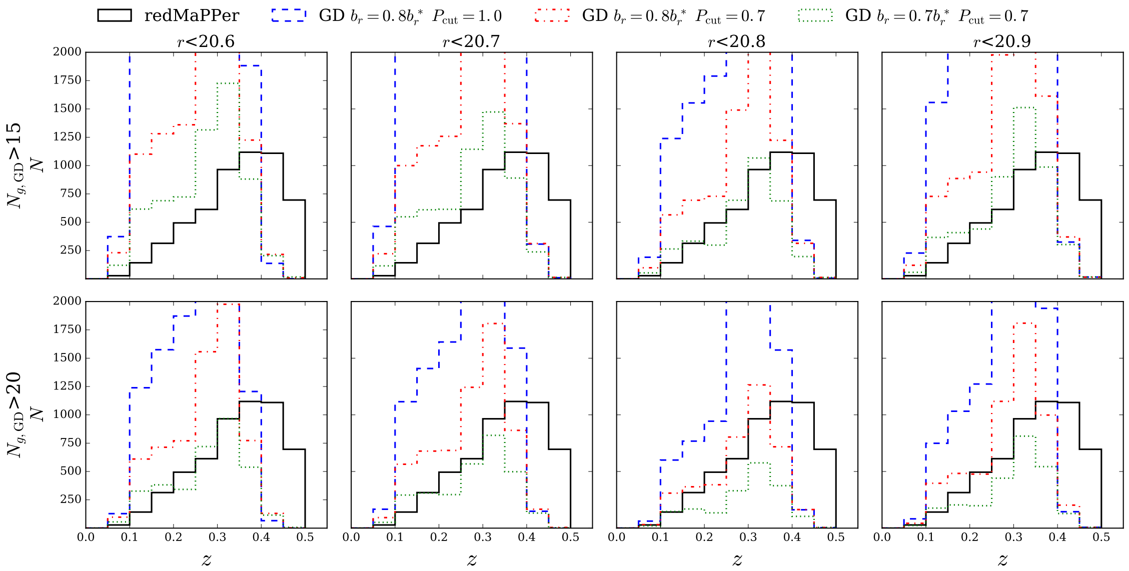

Figure 4.20: Number counts versus the redshift for the sample of the redMaPPer cluster catalog and the three best cases of the detected GAMA DEEP clusters ($P_{\rm cut} = 1.0,\ b_r = 0.8b_r^*$), ($P_{\rm cut} = 0.7,\ b_r = 0.8b_r^*$), ($P_{\rm cut} = 0.7,\ b_r = 0.7b_r^*$). In other words, we select the cases in which we recover a large number of redMaPPer clusters. For $P_{\rm cut} = 1.0,\ b_r = 0.8b_r^*$ case, we observe a large excess of clusters detected by the VT-FOFz for both cases $N_{g,\rm GD} \geq 15$ and $N_{g,\rm GD} \geq 20$. The reduction of the $P_{\rm cut}$ improves the results, however the histogram of GAMA DEEP clusters does not match with the histogram of redMaPPer. If we combine the effect of two parameters (i.e., we reduce the transverse linking length and we select less densest galaxies) the results improve. Note that for the cases ($P_{\rm cut} = 0.7,\ b_r = 0.7b_r^*,\ N_{g,\rm GD} \geq 15,\ r < 20.8$) and ($P_{\rm cut} = 0.7,\ b_r = 0.7b_r^*,\ N_{g,\rm GD} \geq 20,\ r < 20.6$) the GAMA DEEP histograms match the redMaPPer histogram up to $z \approx 0.33$. Therefore, we can assert that these combinations of parameters are reliable results.

4.5 Summary and conclusions

In this chapter, we have introduced the VT-FOFz cluster finder. This technique combines Voronoi Tessellation and Friends of Friends algorithms. We have been able to detect galaxy clusters in every cut in the r-band for the magnitude range $19.4 < r < 20.9$ in the mock galaxy catalog. We observe that the cut in r-band magnitude as well as the quality of estimated photometric redshifts play an important role in the detection of galaxy clusters and in the redshift depth of the cluster catalog. We set that a cluster detected from the mock galaxy catalog matches with a dark matter halo if the galaxy cluster lies in the matching threshold region of the halo.

The matching region is a cylinder defined in the 2+1 space (i.e., angular coordinates and redshift). We require that the detected matches be unique.

For the mock catalogs, the b_r and b_z free parameters from the FOF technique are obtained by finding the value set which maximizes the number of matches for an input galaxy catalog with no Voronoi selection. The $P_{\rm cut}$ free parameter from the Voronoi step is obtained by determining the value which adjusts the galaxy cluster number counts with the halo number counts for a threshold redshift set by the density of galaxies in each r-band magnitude cut. We observe that approximately the 40% of densest galaxies are candidates to form part of galaxy clusters, for all r-band cuts in the mock catalog. On the other hand, we note that the b_r parameter decreases for deeper magnitude cuts. This effect is due to the fact that for deeper cuts we have more available galaxies, hence the transversal linking length tends to be smaller. The above effect is opposite to the line-of-sight linking length because this parameter depends on the photometric redshift quality. Therefore, it is important to remark that for deeper cuts b_z tends to be greater in agreement with the loss in the quality of the estimated photometric redshifts. We compute the completeness and purity by using the best values for the free parameters of the VT-FOFz cluster finder in the mock catalog. At high redshift we detect less clusters, hence the completeness worsens. However, a large number of the few detected galaxy clusters are matched with the haloes in that redshift region, thus implying high purity. The massive haloes tend to match clusters with high richness as we expect (i.e., positive correlation), nevertheless we note that there is a large scatter in the observable-mass relation. We measure a high completeness and purity for massive haloes and clusters with high richness as expected, these measurements improve for deeper r-band cuts. We observe that the low quality of the photometric redshift for deeper r-band cuts does not have a large impact on the measurement of completeness and purity in the mock catalog.

We run the VT-FOFz cluster finder on a sample of the GAMA DEEP survey. We compare the obtained cluster catalog with the redMaPPer SDSS DR8 cluster catalog. We analyze the impact of the Voronoi selection and the choice of the transverse linking length by varying the $P_{\rm cut}$ and b_r parameters in the GAMA DEEP sample. For the comparison between the GAMA DEEP cluster catalog and redMaPPer catalog, we treat the redMaPPer clusters like dark matter haloes and we compute the completeness and purity in the same way as with the mock catalogs. We observe that the Voronoi selection as well as the transverse linking length affect the detection of galaxy clusters in the same way. The selection of low values of P_{cut} and b_r parameters allows us to avoid the detection of large and/or rich galaxy clusters, which do not match with redMaPPer clusters. Nonetheless, low values of these parameters imply missed detections of true galaxy clusters, therefore we must choose the parameters with care. For $N_{g,\mathrm{GD}} \geq 5$ we recover a large number of redMaPPer cluster ($\mathbf{C} > 0.9$) up to $z \approx 0.33$, but with low purity. The results obtained for $z > 0.33$ are less reliable. The Voronoi selection allows us to reduce the scatter in the redshift relation between the GAMA DEEP clusters and redMaPPer clusters. Comparing the redshift distribution for $N_{g,\mathrm{GD}} \geq 15$ and $N_{g,\mathrm{GD}} \geq 20$ in the cases in which we recover a large number of redMaPPer clusters, we show that including Voronoi selection produces the best results. Note that the GAMA DEEP cluster catalog must not necessarily have the same cluster selection as the redMaPPer catalog, because the used optical methods are different.

Chapter 5

Cosmological forecast via abundance of galaxy clusters

In the study of the evolution of large scale structures in the Universe, galaxy clusters appear as candidates for understanding the underlying nature of this density field. In section 2.6.4 we showed that the abundance of dark matter haloes and mainly the halo mass function is intrinsically related with the cosmological model describing the Universe. Therefore the number counts of these large objects is ideal to be used as a cosmological probe, however dark matter structures cannot be directly observed. The observables we can detect are the baryonic structures which lie within the cluster haloes, among which are galaxies as well as baryonic gas (galaxies here are considered as tracers of the dark matter), see Voit (2005); Allen et al. (2011).

Galaxy clusters are the largest detected virialized objects. They are embedded within the filaments in a cosmic web of matter, see Bharadwaj et al. (2004). In 1933 the Swiss astrophysicist Fritz Zwicky suggested the existence of dark matter in the Coma Cluster via the observation of the velocity dispersions of the galaxies in that same structure. Currently this exotic form of matter is also used for explaining the rotation curves of spiral galaxies (Freeman (1970); Rubin, Ford & Thonnard (1980)) as well as the evolution of structures in the Universe (Blumenthal et al. (1984)). Clusters are formed by as many as several hundred galaxies which are gravitationally bound ($\sim$ 5% of the total cluster mass), the hot Intra-Cluster Medium ($\sim$ 15% of the total cluster mass) and the dark matter that composes the halo ($\sim$ 80% of the total cluster mass). Other negligible components in the total energy of the clusters are the relativistic particles and magnetic fields. These systems are usually classified on their mass: Structures between $\sim 10^{12} - 10^{13.7}\,M_{\odot}$ are usually known as *Galaxy Groups*, structures between $\sim 10^{13.7} - 10^{15}\,M_{\odot}$ are usually known as

Galaxy Clusters and structures above $\sim 10^{15}\, M_\odot$ are usually known as *Galaxy Super Clusters*. Clusters are also used for studying energetic hydrodynamic process, the enrichment of metals in the Universe (in astrophysics, metals are referred to any element except for hydrogen and helium) and the physics of *Active Galactic Nuclei* (AGN).

The abundance of galaxy clusters as a cosmological probe has been used in several works (Gladders et al. (2007); Mantz et al. (2008); Rozo et al. (2010); Allen et al. (2011); Mana et al. (2013)). However the determination of the mass in galaxy clusters through observable quantities (e.g., temperature, richness) as well as the other observational effects (i.e., photometric redshift, completeness and purity of cluster catalog, among others) are open problems nowadays, see Allen et al. (2011); Kravtsov & Borgani (2012); Simet et al. (2017). In order to use galaxy clusters in cosmological analysis, it is necessary to model the above effects as functions of mass and redshift. There are several works that have used the self-calibration method to constrain both cosmological parameters of interest and nuisance parameters involved in the observational effects via cluster number counts and the clustering of clusters (i.e., cluster covariance), see Lima & Hu (2004); Lima & Hu (2005); Lima & Hu (2007); Erickson et al. (2011); Aguena & Lima (2016). Our aim in this chapter is to perform a self-calibration cosmological inference test through the galaxy cluster abundance by using a MCMC (Markov-Chain Monte Carlo) statistical method. We apply four tests in this analysis. First, we fix the nuisance parameters and we constrain the cosmological parameters of interest (i.e., we consider a large flat prior probability distribution). Second, we fix the cosmological parameters and we constrain the nuisance parameters. Third, we constrain all parameters involved in the analysis by using a Gaussian prior probability (i.e., normal distribution) for the parameters except the parameter of the equation of state of dark energy and the density parameter of the cold dark matter. Fourth, we constrain all parameters involved in the analysis with a flat prior probability. The used cluster abundance data are generated via random Poisson sampling from a fiducial model.

This chapter is organized as follows: Section 5.1 we describe the relation between the observed galaxy clusters and the halo dark matter. Section 5.2 we perform a brief introduction of the statistical tools which are used to the development of this work. Section 5.3 we present the fiducial model utilized for the cosmological forecasting. Section 5.4 we constrain the model parameters (i.e, the cosmological parameters plus the nuisance parameters), and we perform the discussion of the results. Section 5.5

we conclude this chapter.

5.1 Galaxy cluster abundance

In section 2.6 we presented the concepts of a power matter spectrum of fluctuations and the growth factor. We explained the spherical collapse model to illustrate the formation of dark matter haloes and we derived an estimated the halo abundance through the Press-Schechter formalism. In this section we outline the relation between the abundance of the observed galaxy clusters and the abundance of the dark matter haloes (i.e., the observable-mass relation and the effect due to the photometric redshifts).

5.1.1 Mass function

The mass function describes the number density of collapsed dark matter haloes, in other words, it models the dark matter halo abundance. Section 2.6.4 presents the basic notions to estimate the halo number counts in mass slices and redshift bins. We showed that the mass function is described by the following function form

$$\frac{\mathrm{d}n}{\mathrm{d}\ln M} = f(\sigma)\frac{\rho_m}{M}\frac{\mathrm{d}\ln \sigma^{-1}}{\mathrm{d}\ln M}, \tag{5.1}$$

where ρ_m is the background matter density in the Universe, equation (2.30); $\sigma(M, z)$ is the rms (root mean square) variance of the smoothed density field, equation (2.183); M is the halo mass, dn is the number of haloes per comoving volume for mass value between M and $M + \mathrm{d}M$; and $f(\sigma)$ is a factor that depends on the model used to describe the dark matter collapse, in the PS formalism it is given by equation (2.188). Jenkins et al. (2001); Evrard et al. (2002); Linder & Jenkins (2003); Kuhlen et al. (2005); Crocce et al. (2010); among others showed that the factor $f(\nu)$ can be parametrized through a universal function of the peak height $\nu = \delta_c/\sigma$, hence it is expected to be cosmologically dependent only on $\sigma(z, M)$. Courtin et al. (2011) demonstrated that the critical overdensity for collapse δ_c also depends on redshift and cosmology to describe a more accurate mass function.

In order to improve the accuracy of the functional form, the mass function is measured via a large set of cosmological simulations. Currently, the most used model is given by Tinker et al. (2008). They used a large set of simulations based on a flat, ΛCDM cosmology. The simulations were perfomed by using three independent codes, `GADGET2` (Springel (2005)), `HOT` (Warren & Salmon (1993)), and `ART` (Kravtsov et al. (1997)). The dark matter haloes are identified by a spherical overdensity (SO) algorithm. The functional form calibrated by them is given by

$$f(\sigma) = A\left[\left(\frac{\sigma}{b}\right)^{-a} + 1\right]\exp\left(-c/\sigma^2\right), \tag{5.2}$$

which it is motivated from Sheth & Tormen (1999). The parameter A is the amplitude of the overall mass function, and the parameter c sets the cutoff scale at which the halo abundance exponentially decreases. Moreover, a and b are the slope and amplitude in the limit of the low-mass, respectively. The above parameters were estimated for various values of the overdensity with respect to the background matter density Δ and of redshifts ($\Delta \in [200, 3200]$ and $z \leq 2.5$). They showed that the parameters A, a and b depend on the overdensity and redshift, and the parameter c only depends on the overdensity. The functional form for the parameters is given by

$$\begin{aligned}
A(z) &= A_0(1+z)^{-0.14}, &(5.3)\\
a(z) &= a_0(1+z)^{0.06}, &(5.4)\\
b(z) &= b_0(1+z)^{-\alpha}, &(5.5)\\
\log_{10}\alpha(\Delta) &= -\left[\frac{0.75}{\log_{10}(\Delta/75)}\right]^{1.2}. &(5.6)
\end{aligned}$$

The zero subscript indicates the value at $z = 0$. The fitting functions for the overdensity are

$$\begin{aligned}
A_0 &= \begin{cases} 0.1(\log_{10}\Delta) - 0.05 & \text{if} \quad \Delta < 1600, \\ 0.26 & \text{if} \quad \Delta \geq 1600, \end{cases} &(5.7)\\
a_0 &= 1.43 + (\log_{10}\Delta - 2.3)^{1.5}, &(5.8)\\
b_0 &= 1.0 + (\log_{10}\Delta - 1.6)^{-1.5}, &(5.9)\\
c &= 1.2 + (\log_{10}\Delta - 2.35)^{1.6}. &(5.10)
\end{aligned}$$

The Tinker mass function is calibrated over the range of halo masses $10^{10.5}h^{-1}M_\odot \leq M \leq 10^{15.5}h^{-1}M_\odot$ at $z = 0$. Figure 13 of Tinker et al. (2008) shows the evolution of the mass range with the redshift. The models mentioned previously do not consider effects of baryon physics which can affect the halo mass function at the level of a few percent, see Stanek et al. (2009); Cui et al. (2012). We will use the Tinker mass function for the analysis developed in this chapter.

5.1.2 Observable clusters and theoretical haloes

So far we have talked about the distribution of dark matter haloes (i.e., the mass function for haloes) as well as the use of halo number counts to constrain cosmological parameter. We must relate the halo masses employed in theoretical counts with the observable physical properties of the galaxy clusters. The observational signals of clusters are found in different wavelengths. These include millimeter wavelengths, i.e. the Sunyaev-Zel'dovich effect, optical (richness or velocity dispersion in galaxy members) and X-ray due to the thermal bremsstrahlung (luminosity, temperature, gas mass and/or gas thermal energy), see Allen et al. (2011).

In order to estimate the mass of galaxy clusters, the scaling relations usually depend on the observable. For our cosmological analysis, we assume the scaling relation to be a power law in mass according to Rozo et al. (2010); Mana et al. (2013); Simet et al. (2017). We parametrize

$$\ln M^{\rm obs} = \ln M_0 + \alpha \, \ln \left(\frac{\lambda}{\lambda_0} \right) , \tag{5.11}$$

where $M^{\rm obs}$ is the observed mass inferred from the observable, λ is the cluster richness and M_0 is the normalization factor for the pivot point λ_0. To relate the observed mass with the true mass (i.e., theoretical mass), we assume the following lognormal distribution given by Lima & Hu (2005)

$$P(M^{\rm obs}|M) = \frac{1}{\sqrt{2\pi\sigma^2_{\ln M}}} \exp\left(-x^2\left(M^{\rm obs}\right)\right) , \tag{5.12}$$

where

$$x\left(M^{\rm obs}\right) \equiv \frac{\ln M^{\rm obs} - \ln M - \ln M^{\rm bias}}{\sqrt{2\sigma^2_{\ln M}}} . \tag{5.13}$$

Here the mass bias $\ln M^{\rm bias}$ and the variance $\sigma_{\ln M}$ depend on the redshift and the true mass. The number of galaxy clusters per comoving volume within the observable mass range $\left[M_k^{\mathrm{obs}}, M_{k+1}^{\mathrm{obs}}\right]$ is given by

$$\begin{aligned}
\bar{n}_k &= \int_{M_k^{\mathrm{obs}}}^{M_{k+1}^{\mathrm{obs}}} \mathrm{d}\ln M^{\mathrm{obs}} \int \mathrm{d}\ln M \frac{\mathrm{d}\bar{n}}{\mathrm{d}\ln M} P(M^{\mathrm{obs}}|M) \qquad (5.14)\\
&= \int \mathrm{d}\ln M \frac{\mathrm{d}\bar{n}}{\mathrm{d}\ln M} \frac{1}{2}\left[\mathrm{erfc}(x_k) - \mathrm{erfc}(x_{k+1})\right].
\end{aligned}$$

We define $x_k = x\left(M_k^{\mathrm{obs}}\right)$. The function $\mathrm{erfc}(x_k)$ is the complementary error function. Note that for $\sigma_{\ln M} \to 0$ and $\ln M^{\mathrm{bias}} = 0$ we recover equation (2.191). In order to account the effects of the cluster finder used to detect the galaxy clusters and to improve the accuracy of the results we can consider the completeness and purity in the above equation according to Aguena & Lima (2016). However, for our approach, we do not consider these properties in the cluster number counts, in this analysis we assume the purity and completeness equal to one. In addition to the effects due to the observable-mass relation presented previously, in our approach we consider the uncertainties obtained in the estimation of photometric redshifts. These redshift uncertainties can affect the position of the galaxy clusters in the 2+1 space of angular coordinates and redshift. We also assume a normal distribution in the relation $z - z_{\mathrm{phot}}$ according to Lima & Hu (2007), hence we have

$$P(z_{\mathrm{phot}}|z) = \frac{1}{\sqrt{2\pi\sigma_z^2}} \exp\left(-y^2\left(z_{\mathrm{phot}}\right)\right), \qquad (5.15)$$

where

$$y(z_{\mathrm{phot}}) \equiv \frac{z_{\mathrm{phot}} - z - z_{\mathrm{bias}}}{\sqrt{2\sigma_z^2}}. \qquad (5.16)$$

Here z_{bias} and σ_z are functions of the redshift. To simplify calculations, we neglect the effects in the angular selection, therefore the mean number of the observed galaxy clusters in the redshift bin $[z_{\mathrm{phot},i}, z_{\mathrm{phot},i+1}]$ and mass range $[M_k^{\mathrm{obs}}, M_{k+1}^{\mathrm{obs}}]$ is given by

$$\begin{aligned}
\bar{m}_{k,i} &= \Delta\Omega \int_{z_{\mathrm{phot},i}}^{z_{\mathrm{phot},i+1}} \mathrm{d}z_{\mathrm{phot}} \int \mathrm{d}z \frac{\mathrm{d}V_c}{\mathrm{d}z\mathrm{d}\Omega} P(z_{\mathrm{phot}}|z)\bar{n}_k, \qquad (5.17)\\
&= \Delta\Omega \int \mathrm{d}z \frac{D^2}{H(z)} \frac{1}{2}\left[\mathrm{erfc}(y_i) - \mathrm{erfc}(y_{i+1})\right]\bar{n}_k,
\end{aligned}$$

where $y_i = y(z_{\mathrm{phot,i}})$ and $\Delta\Omega$ is the survey sky coverage. We observe that in the perfect redshift case (i.e., in the absence of photo-z uncertainties and biases), the above expression is given by equation (2.192). Lima & Hu (2007) performed a quantitative and qualitative discussion about the effects on cluster abundance due to the bias and scatter in photometric redshift, see figure 1 from Lima & Hu (2007). We should point out that equation (5.17) allows us to theoretically estimate the observed cluster number counts in a mass range and a redshift bin.

5.2 Statistical tools for analysis

The constraint of physical properties and parameters in a cosmological model via observed data, requires robust statistical analysis for inferring the best fit of the parameters and their associated uncertainties. The following section presents a brief introduction of the statistical tools used in the development of this work.

5.2.1 Bayes' theorem

In Bayesian statistics, the probability represents a degree-of-belief of an assumption instead that of the Frequentist interpretation of the probability in which it is related with the frequency with which an event occurred. Therefore, as our data (i.e., the Universe) cannot be repeated in a cosmological analysis, it is more useful to use the Bayesian inference approach rather a Frequentist inference approach.

Let X and Y be two propositions, then according to probability theory, these variables obey the following rules

$$\mathrm{prob}(X|I) + \mathrm{prob}(\bar{X}|I) = 1, \qquad \text{sum rule,} \tag{5.18}$$

$$\mathrm{prob}(X, Y|I) = \mathrm{prob}(X|Y, I) \times \mathrm{prob}(Y|I), \qquad \text{product rule,} \tag{5.19}$$

where $\bar{X}$ denotes the proposition that X is false. The function $\mathrm{prob}(X|I)$ defines the probability that the proposition X occurs for a given relevant background information I. By using the product rule (5.18), we find the following expression:

$$\mathrm{prob}(X|Y, I) = \frac{\mathrm{prob}(Y|X, I) \times \mathrm{prob}(X|I)}{\mathrm{prob}(Y|I)}. \tag{5.20}$$

This relation is known as Bayes' theorem. By using both sum and product rules, we can obtain the marginalization relation, which is given by

$$\mathrm{prob}(X|I) = \sum_{k=1}^{M} \mathrm{prob}(X, Y_k|I), \tag{5.21}$$

for a whole set of alternative possibilities $\{Y_k\} = Y_1, Y_2, Y_3, ..., Y_M$. Therefore, through the product rule (5.18), we can show that the above expression satisfies the following normalization condition

$$\sum_{k=1}^{M} \mathrm{prob}(Y_k|X, I) = 1. \tag{5.22}$$

In the continuum limit, equation (5.21) is given by

$$\mathrm{prob}(X|I) = \int_{-\infty}^{\infty} \mathrm{pdf}\,(X, Y|I)\mathrm{d}Y, \tag{5.23}$$

where $\mathrm{pdf}\,(X, Y|I)$ is the probability distribution function, and the normalization condition is

$$\int_{-\infty}^{\infty} \mathrm{pdf}\,(Y|X, I)\mathrm{d}Y = 1. \tag{5.24}$$

The probability distribution function is interpreted here as a probability density. The probability that the variable Y lies in the range $[y_1, y_2]$ and with a true proposition X is given by

$$\mathrm{prob}(X, y_1 \leq Y < y_2|I) = \int_{y_1}^{y_2} \mathrm{pdf}(X, Y|I)\mathrm{d}Y. \tag{5.25}$$

Sivia & Skilling (2006) prefer to use "prob" to refer to the probability distribution function and thus keeping the notation between the discrete and continuous cases. In the rest of this work, we denote "pdf" by "prob". Note that the marginalization is a powerful tool for problems which involve nuisance parameters as it is the case in the analysis that we perform in this chapter.

If we replace X by the data $\{d_i\}$ and Y by the hypothesis $\Theta(p_k)$, here $\{p_k\}$ are the parameter set of the hypothesis. Thus, Bayes' theorem can be rewritten as

$$\mathrm{prob}(\Theta(p_k)|\{d_i\}, I) = \frac{\mathrm{prob}(\{d_i\}|\Theta(p_k), I) \times \mathrm{prob}(\Theta(p_k)|I)}{\mathrm{prob}(\{d_i\}|I)}. \tag{5.26}$$

The term $\mathrm{prob}(\Theta(p_k)|I)$ is called the *prior* probability, it represents the initial knowledge we assume about the hypothesis before the statical analysis; the term $\mathrm{prob}(\{d_i\}|\Theta(p_k), I)$ is called *likelihood* function, it modifies the prior probability according to the data; the term $\mathrm{prob}(\Theta(p_k)|\{d_i\}, I)$ is called *posterior* probability, it shows the state of knowledge about the truth of the hypothesis according to the experimental measurements; the term $\mathrm{prob}(\{d_i\}|I)$ is the *evidence*. The evidence plays an important role in the model selection, see chapter 4 of Sivia & Skilling (2006). The above representation of the Bayes' theorem enables us to know the probability that the hypothesis is true given the experimental measurements. We can estimate the degree of believe in the proposed model by using the data.

5.2.2 Poisson statistics

In nature there are problems that involve the counting of discrete events in a finite interval, either of time, distance or other physical properties. The probability of observing N events in a fixed interval, given only the expected value $\langle N\rangle = \mu$ can be expressed by a Poisson distribution

$$\mathrm{prob}(N|\mu) = \frac{\mu^N e^{-\mu}}{N!}. \tag{5.27}$$

We observe that the mean value of N is given by

$$\langle N\rangle = \sum_{N=0}^{\infty} N\mathrm{prob}(N|\mu) = \sum_{N=0}^{\infty} N\frac{\mu^N e^{-\mu}}{N!} = \mu e^{-\mu}\sum_{N=1}^{\infty}\frac{\mu^{N-1}}{(N-1)!} = \mu, \tag{5.28}$$

The variance is given by

$$\begin{aligned} \sigma_N^2 &= \langle (N-\langle N\rangle)^2\rangle = \langle N^2\rangle - \langle N\rangle^2 = \sum_{N=0}^{\infty} N^2\mathrm{prob}(N|\mu) - \mu^2 \qquad (5.29)\\ &= \sum_{N=0}^{\infty} N^2\frac{\mu^N e^{-\mu}}{N!} - \mu^2 = e^{-\mu}\mu\sum_{k=0}^{\infty}(k+1)\frac{\mu^k}{k!} - \mu^2 = \mu. \end{aligned}$$

The Poisson distribution tends to a normal distribution for large μ. In the limit of large mean value, we can say that $N = x = \mu(1+\delta)$, where $\mu \gg 1$ and $\delta \ll 1$. Note that the discrete distribution becomes a continuous probability distribution function for the variable x. Given that N is large we can use the Stirling's approximation, then

$$N! \to \sqrt{2\pi x}e^{-x}x^x, \quad x \to \infty. \tag{5.30}$$

Therefore, equation (5.27) is rewritten as

$$\begin{aligned}\mathrm{prob}(x|\mu) &= \frac{\mu^x e^{-\mu}}{\sqrt{2\pi x}e^{-x}x^x} = \frac{\mu^{\mu(1+\delta)}e^{-\mu}}{\sqrt{2\pi\mu(1+\delta)}e^{-\mu(1+\delta)}\mu(1+\delta)^{\mu}(1+\delta)} \\ &= \frac{e^{\mu\delta}(1+\delta)^{-\mu(1+\delta)-1/2}}{\sqrt{2\pi\mu}} = \frac{1}{\sqrt{2\pi\mu}}\exp\left(-\frac{(x-\mu)^2}{2\mu}\right).\end{aligned} \tag{5.31}$$

Here we use the fact that

$$(1+\delta)^{\mu(1+\delta)-1/2} \approx \exp\left(-\mu\left(\delta+\frac{\delta^2}{2}\right)\right), \tag{5.32}$$

for $\mu \gg 1$ and $\delta \ll 1$. The above approximation is found by taking the natural logarithm and then expanding in δ to second order. Note that in this case the mean value as well as variance are equal to μ as previously shown.

5.2.3 Markov chain Monte Carlo methods

The Markov chain Monte Carlo (MCMC) is a family of techniques which are used to sample probability distributions. The concept of a *Monte Carlo* refers to the algorithms based on a random realizations, whereas that the Markov chain concept is a series of random variables, such that the next value in the chain only depends on the current position and not on previous values. The MCMC methods have been used in several astrophysical and cosmological analysis, for instance Lewis & Bridle (2002); Battye & Weller (2003); Vikhlinin et al. (2009); Planck Collaboration (2016); Simet et al. (2017). In addition to understanding the posterior probability distribution function or likelihood function in detail, the MCMC methods enable us to solve the marginalization problem of the nuisance parameters in the statistical analysis, because these algorithms naturally provide a sampling of values for a specific parameter from the marginalized probability distribution. In general, MCMC methods allow us to determine the maximum of the posterior distribution as well as to estimate the uncertainties of the involved parameters in the analysis. On the other hand, we must point out that an algorithm based on MCMC provides us an efficient method to characterize posterior probability function or likelihood function which depends on a large set of parameters unlike the other methods such as the gridding (i.e., the time cost is lower compared with other methods). For more details, see MacKay (2003).

The Metropolis-Hastings method

The Metropolis-Hastings method (M-H) is the most commonly MCMC algorithm used to estimate the probability distribution in the data analysis. This method is given by the following algorithm:

1. Set an arbitrary initial point in the parameter space Θ_1 to which it is computed the probability $p(\Theta_1|\{d_i\})$.

2. Determine a second point through a proposal probability distribution $q(\Theta_2|\Theta_1)$, which depends on the current point Θ_1. Usually, the proposal distribution is given by a multivariate Gaussian distribution centered on Θ_1 with a general covariance tensor that must be tuned for performance.

3. Estimate the acceptance probability, which is defined as

$$\alpha(\Theta_2|\Theta_1) = \min\left\{1, \frac{p(\Theta_2|\{d_i\})}{p(\Theta_1|\{d_i\})}\frac{q(\Theta_1|\Theta_2)}{q(\Theta_2|\Theta_1)}\right\}. \tag{5.33}$$

   If the proposal probability distribution is symmetric $q(\Theta_2|\Theta_1) = q(\Theta_1|\Theta_2)$, thus the last quotient in the acceptance probability is equal to one.

4. Generate a random number r from a uniform distribution of $(0, 1)$.

5. Compare r with the acceptance probability

   - If $\alpha > r$, Θ_2 is accepted in the chain and it is taken as the initial point for the following step.
   - If $\alpha < r$, Θ_2 is rejected, Θ_1 is again the initial point for the following step.

In order to build a chain, the above process is repeated to obtain the next point. Note that the proposal point Θ_2 only depends on the current point and not on the previous points of the chain. The M-H algorithm has the advantage that it is simple to implement. Nonetheless there are others algorithms with faster convergence rate.

emcee software package

The `emcee` software package* was developed by Dan Foreman-Mackey, see Foreman-Mackey et al. (2013). It is a python implementation of the affine-invariant ensemble

*⟨http://dan.iel.fm/emcee/current/⟩

sampling algorithm proposed by Goodman & Weare (2010), which is commonly called the stretch move. The advantages of this method are that its performance does not depend on the covariances among parameters, its autocorrelation time is shorter than in the M-H method (i.e., the stretch move algorithm requires fewer computations of the probability distribution compared to a M-H sampler to generate the same number of independent samples) and it can be run parallel according to Foreman-Mackey et al. (2013).

The stretch move algorithm consists of simultaneously evolving an ensemble of K walkers $S = \{\Theta_k\}$, such that the proposal probability distribution for each walker is based on the current position of the $k-1$ walkers which belongs to the complementary ensemble $S_{[k]} = \{\Theta_j, \forall\, j \neq k\}$. If $\Theta_j \in S_{[k]}$, then the following step of a walker at position $\Theta_k(t)$ is obtained via

$$\Theta_k(t+1) = \Theta_j + Z\left(\Theta_k(t) - \Theta_j\right), \tag{5.34}$$

where Z is a random variable drawn from a distribution $g(Z = z)$ that does not depend on the covariances between the parameters. Here the acceptance probability is given by

$$\alpha(\Theta_k(t+1)|\Theta_k(t)) = \min\left(1, Z^{n-1}\frac{p\left(\Theta_k(t+1)\right)}{p\left(\Theta_k(t)\right)}\right), \tag{5.35}$$

where n is the number of parameters. The above algorithm is repeated in series for each walker in the ensemble.

In order to parallelize the above algorithm, the full ensemble is split into two subsets ($S^{(0)} = \{\Theta_k, \forall\, k = 1, \ldots, K/2\}$ and $S^{(1)} = \{\Theta_k, \forall\, k = K/2+1, \ldots, K\}$) and simultaneously all the walkers from the $S^{(0)}$ set are updated by the stretch move method by employing only the positions of the walkers in $S^{(1)}$. In the same way, by using the new positions in the $S^{(0)}$ set, it is updated $S^{(1)}$. Thus achieving a valid step for all walkers.

Preparing the chains for the statistical analysis

The chains obtained by a MCMC method allow us to determine the best value and their uncertainty of the model parameters involved in the study (i.e., characterize the probability distribution for each model parameter). However, we should be careful with the MCMC output. We must set whether the results have achieved a converged set of samples or they do not converge yet, as well as to determine the position in which the chains have converged. Here, we present useful concepts which are used in this chapter to prepare the chains before performing the statistical analysis.

- **Burn-in:** The number of initial steps required for that the MCMC output begins to describe a stationary sample around to the maximum of the posterior probability distribution is known as burn-in. This set of points must be rejected to avoid their influence on the estimation of both the best fit and the confidence regions of the model parameters. The selected burn-in depends on the performed experiment. We choose the percent of burn-in by performing a visual inspection in the chains against the number of the iteration. This choice must be satisfied for the involved parameters in the experiment.

- **Convergence Criterion:** In order to determine if the obtained MCMC output has converged, we set the following steps as convergence criterion:

  1. Split the total sample in two subsets with the same number of chains.
  2. Set the burn-in in both subsamples.
  3. Determine the root square of the variance $\sigma_i^{(k)}$ and the mean value $\mu_i^{(k)}$ for each parameter i in both subsamples ($k = 1, 2$).
  4. Compute the ratio

  $$r_i^{(k)} = \frac{|\mu_i^{(1)} - \mu_i^{(2)}|}{\sigma_i^{(k)}}, \quad k = 1, 2, \tag{5.36}$$

  where i represents all parameters.

  If $r_i^{(1)} \approx r_i^{(2)} \to 0.0 \;\forall\, i$ we can say that the dataset enables sampling the probability distribution with a good reliability level. Therefore, as long as the ratios are lower, the reliability level will improve. Note that here we consider finite variance, i.e., $\sigma^2 < \infty$. In addition to the above criterion, we perform a visual comparison between the confidence contours obtained for the two subsamples.

- **Thinning the chains:** This process consists of reducing the number of points in the chains obtained from a MCMC algorithm to decrease their autocorrelation and to provide a more precise estimate of the posterior probability distribution. We thin the sample keeping every n-th point from the chains, e.g., if we have a sample of 50 000 points and we apply a thinning of 5, then the final sample would have 10 000 points. According to Owen (2017), this can improve the statistical efficiency of the chain.

5.3 Fiducial model

In order to carry out the cosmological forecasting we assume a flat wCDM model without radiation as our fiducial cosmology. The mass bias of the observable-mass relation is expected to be a smooth function of redshift. On the other hand, the mass scatter tends to increase for high redshifts and low mass. Therefore, according to Aguena & Lima (2016), the bias and scatter of equation (5.13) can be parametrized by

$$\ln M^{\text{bias}} = A_b + n_b \ln(1+z), \tag{5.37}$$

and

$$\frac{\sigma^2_{\ln M}}{0.2^2} = 1 + B_0 + B_z(1+z) + B_M\left(\frac{\ln M_s}{\ln M}\right). \tag{5.38}$$

Here A_b, n_b, B_0, B_z and B_M are nuisance parameters, and $\ln M_s$ is the pivot mass. It is fixed with value $M_s = 10^{14.2} h^{-1} M_\odot$. For the bias and the scatter due to the estimated photometric redshift (see equation (5.15) and equation (5.16)) we assume a constant value for each considered redshift bin in the analysis. Hence, if we use N_z redshift bins, we add $2N_z$ parameters to the nuisance set. In addition to the nuisance parameters indicated previously, we must add to this set the parameter α which comes from the mass-richness scaling relation (see equation (5.11)). The normalization factor and the pivot point are fixed according to the values given by Simet et al. (2017) $M_0 = 10^{14.344} h^{-1} M_\odot$ and $\lambda_0 = 40$.

For our synthetic dataset, we consider a total sky area of 10 400 deg^2 which is in agreement with the survey area used in the RedMaPPer SDSS DR8 cluster catalog (Rykoff et al. (2014)), see section 4.1.3. To make a more realistic study, the limits of the richness and redshift of the used data are in agreement with the results obtained in chapter 4. We set the following ranges: for richness $10 \leq \lambda \leq 300$ and for redshift $0.05 \leq z \leq 0.40$. We employ 4 redshift bins and 5 richness bins in a logarithmic scale for our analysis. Therefore we have 8 nuisance parameters due to the photometric redshifts. To compute the cluster number counts, we use the Tinker mass function (5.2) for overdensity $\Delta = 200$ with respect to the background matter density.

The observed synthetic data of the cluster number counts are obtained from a random sampling of a Poisson distribution with the expected value computed by the

Tinker mass function. Table 5.1 shows the fiducial values for the parameters involved in this analysis. The cosmological parameters are taken from Planck Collaboration (2016). The parameters of the mass-richness relation are taken from Simet et al. (2017). The parameters of the observable-mass relation are taken from Aguena & Lima (2016). The parameters due to photometric redshifts are considered such that they are not biased and their scatter increases for high redshifts just as it is expected, see chapter 3. The considered priors are Gaussian distributions centered in the fiducial values with sigma values given by table 5.1. “None” label in the Prior column indicates that the parameter in question has a large flat prior distribution in all performed tests or the parameter is fixed in the analysis, we do not define a prior probability distribution. Figure 5.1 shows the fiducial model as well as the synthetic data obtained by Poisson sampling for every bin of redshift and richness. Note that for bins with high richness the cluster abundance is lower, then the Poisson error is larger compared with the number of clusters in the bin (see equation (5.29)).

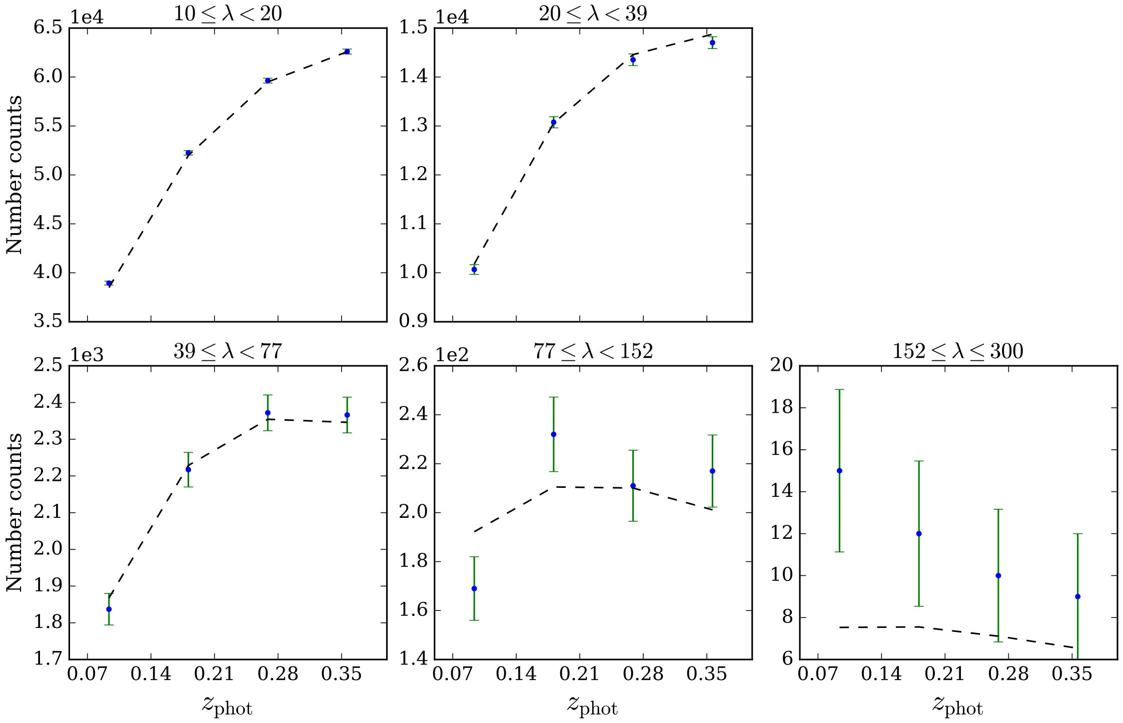

Figure 5.1: Cluster number counts for the fiducial model as well as for the synthetic data for every bin of redshift and richness. The error bars are computed as the root square of the variance according to equation (5.29).

Let n_α be the estimated number of galaxy clusters in a specific bin of redshift and richness (i.e., the expected number of galaxy clusters in the bin), then the probability

Table 5.1: Fiducial values used to obtain the observed synthetic data of the cluster abundance per bin. The values of parameters are taken from Planck Collaboration (2016); Simet et al. (2017); Aguena & Lima (2016). The photometric redshifts are considered without bias and a scatter increasing with the redshift (see chapter 3). The values in the Prior column are the sigma values for the prior probability distributions. The status column indicates whether the parameters is fixed or it is free in the analysis.

Parameter	Definition	Value	Prior	Status
	Cosmological parameters			
h	Hubble parameter	0.6774	±0.0096	Free
$\Omega_c h^2$	CDM density normalized by h	0.1188	None	Free
$\Omega_b h^2$	Baryon density normalized by h	0.02230	±0.00014	Free
w_0	Equation of state of DE	-1.0	None	Free
$\ln A_s$	Amplitude of the linear power spec.	-19.9615	±0.023	Free
n_s	Scalar spectral index	0.9667	±0.0040	Free
Ω_r	Radiation density	0.0	None	Fixed
Ω_k	Curvature density	0.0	None	Fixed
	Nuisance parameters			
α	Power-law index for mass-λ	1.33	±0.09	Free
$M_0/h^{-1}M_\odot$	Normalization factor for mass-λ	$10^{14.344}$	None	Fixed
λ_0	Richness pivot point	40	None	Fixed
A_b	Normalization factor of $M^{\rm bias}$	0.0	±0.09	Free
n_b	Redshift evolution of $M^{\rm bias}$	0.0	±0.05	Free
B_0	Normalization factor of $\sigma^2_{\ln M}$	0.0	±0.05	Free
B_z	Redshift evolution of $\sigma^2_{\ln M}$	0.0	±0.05	Free
B_M	Mass evolution of $\sigma^2_{\ln M}$	0.0	±0.05	Free
$M_s/h^{-1}M_\odot$	Mass pivot of $\sigma^2_{\ln M}$	$10^{14.2}$	None	Fixed
z_{b1}	$z_{\rm phot}$ bias for bin 1	0.0	±0.05	Free
z_{b2}	$z_{\rm phot}$ bias for bin 2	0.0	±0.05	Free
z_{b3}	$z_{\rm phot}$ bias for bin 3	0.0	±0.05	Free
z_{b4}	$z_{\rm phot}$ bias for bin 4	0.0	±0.05	Free
σ^2_{z1}	$z_{\rm phot}$ variance for bin 1	0.1	±0.05	Free
σ^2_{z2}	$z_{\rm phot}$ variance for bin 2	0.15	±0.05	Free
σ^2_{z3}	$z_{\rm phot}$ variance for bin 3	0.20	±0.05	Free
σ^2_{z4}	$z_{\rm phot}$ variance for bin 4	0.25	±0.05	Free

to observe N_α clusters is given by the Poisson distribution, see section 5.2.2. Here we assume that the measures taken in each bin are statistically independent, therefore the likelihood used for our analysis is given by

$$\mathbf{L}(\{N_\alpha\}|\{n_\alpha(\theta_i, s_i)\}) = \prod_{\alpha=1}^{N_{\rm bins}} \frac{n_\alpha^{N_\alpha} e^{-n_\alpha}}{N_\alpha!}, \tag{5.39}$$

where $N_{\rm bins}$ is the considered number of bins, $\{\theta_i\}$ correspond to the cosmological parameters and $\{s_i\}$ correspond to the nuisance parameters. In order to constrain the parameters involved in the model by using the `emcee` software package, we employ the Bayes' theorem to compute the posterior probability (see equation (5.26)). Hence, the logarithm of the posterior function is given by

$$\ln \mathbf{P}(\{n_\alpha(\theta_i, s_i)\}|\{N_\alpha\}) \propto \ln \mathbf{L}(\{N_\alpha\}|\{n_\alpha(\theta_i, s_i)\}) + \ln \mathbf{P}(\{n_\alpha(\theta_i, s_i)\}), \tag{5.40}$$

where

$$\ln \mathbf{L}(\{N_\alpha\}|\{n_\alpha(\theta_i, s_i)\}) = \sum_{\alpha=1}^{N_{\rm bins}} (N_\alpha \ln n_\alpha - n_\alpha - \ln N_\alpha!)\,. \tag{5.41}$$

and $\mathbf{P}(\{n_\alpha(\theta_i, s_i)\})$ is the prior probability of the model parameters.

5.4 Results and discussion

In order to understand the power of the cluster abundance as cosmological probes, we consider four cases for our analysis.

I Free cosmological parameters with fixed nuisance parameters. We use 6 parameters $\{h,\, \Omega_c h^2,\, \Omega_b h^2,\, w_0,\, \ln A_s,\, n_s\}$.

II Free nuisance parameters with fixed cosmological parameters. We use 14 parameters $\{\alpha,\, A_b,\, n_b,\, B_0,\, B_z,\, B_M,\, z_{b1},\, z_{b2},\, z_{b3},\, z_{b4},\, \sigma^2_{z1},\, \sigma^2_{z2},\, \sigma^2_{z3},\, \sigma^2_{z4}\}$.

III Free $\Omega_c\, h^2$ and w_0 parameters, the other parameters with Gaussian priors. We use 20 parameters $\{h,\, \Omega_c h^2,\, \Omega_b h^2,\, w_0,\, \ln A_s,\, n_s,\, \alpha,\, A_b,\, n_b,\, B_0,\, B_z,\, B_M,\, z_{b1},\, z_{b2},\, z_{b3},\, z_{b4},\, \sigma^2_{z1},\, \sigma^2_{z2},\, \sigma^2_{z3},\, \sigma^2_{z4}\}$.

IV Free all parameters in the model. We use 20 parameters $\{h,\, \Omega_c h^2,\, \Omega_b h^2,\, w_0,\, \ln A_s,\, n_s,\, \alpha,\, A_b,\, n_b,\, B_0,\, B_z,\, B_M,\, z_{b1},\, z_{b2},\, z_{b3},\, z_{b4},\, \sigma^2_{z1},\, \sigma^2_{z2},\, \sigma^2_{z3},\, \sigma^2_{z4}\}$.

These tests allow us to explore the impact in the constraint of the model parameters, by using self-calibration of the observable properties in the galaxy clusters (i.e, the nuisance parameters). On the other hand, the second case is focused on constraint of the nuisance parameters by fixing the parameters cosmological. This test allows us to look the sensibility of the number counts in the mass-richness relation, the observable-mass relation and the observed photometric redshift. To observe the effect produced by applying a thinning to the walkers, we compute the percentage relative error for the mean value $\langle p_k \rangle$ and σ_k in each parameter for every thinning test, then we have

$$\text{Percentage relative error} = \left| \frac{x_k - x_k^{th}}{x_k} \right| \times 100. \tag{5.42}$$

Here x_k is the value (either mean value or σ) in the test without thinning and x_k^{th} represents the value obtained by applying the thinning method.

Figure 5.2 and figure 5.3 show the walkers against the step number for the considered cases. The gray lines represent the fiducial value given in table 5.1. These plots allow us to find the step number in which the samples look to have a stable behavior. Note that in the case I, the stability is achieved in the step number ~ 2000, for the case II in the step number ~ 4000, for the case III in the step number ~ 3000 and for the case IV in the step number ~ 5000. The above values enable us to estimate the burn-in value in each test. To complement the visual decision of the burn-in performed through the figures of the walkers, we compute the ratio (5.36) to assess the remaining sample after the burn-in process. Figure 5.4 and figure 5.5 show the ratio for all parameters in each case. We observe that for the case I, case II and case III the remaining sample is reliable, because the threshold ratio is ~ 0.25 in the three cases. The above threshold ratio implies that the discrepancy between the mean value for every parameter obtained in each subsample is at most of one fourth of the estimated σ. Recall that for this calculation, we split the sample in two subsets. The case IV presents a reasonable threshold ratio of ~ 0.9. Although this value is large for having a high reliable sampling, we are satisfied. Because, we achieve low values for several parameters as well as the high values are lower than 1.0, considering that in this case, we are constraining 20 parameters without priors.

Figure 5.6 and figure 5.7 show the percentage error relative computed for the thinning tests as function of the parameters according to equation (5.42). In generally, the thinning method affects more the mean value than the sigma one, in all cases. For the case I, we observe that the relative error does not exceed the value of 0.05% for all parameters, except in $\Omega_b h^2$, wherein the error reaches $\sim 0.2\%$. The case II shows that the relative error reaches the largest value in $\sim 1.2\%$. Here n_s, B_0 and B_z are the parameters whose error is higher. For the other parameters the relative error does not exceed the $\sim 0.42\%$. We observe that the relative error reaches a value of $\sim 13\%$ for the parameter B_0 (Thinning = 100) and a value of $\sim 9\%$ for the parameter A_b (Thinning = 50) in the case III. Nevertheless, for the other parameters the relative error does not exceed a value of $\sim 3\%$. The case IV shows that for all parameters, the relative error does not exceed a value of $\sim 0.2\%$, except to parameter B_0, wherein the error reaches a value of $\sim 6.2\%$. We can assert that for all cases the results are little affected by a thinning of 100. Table 5.2 shows the walkers, steps by walker, burn-in and thinning for every performed case. For the thinning process, we reduce the burn-in in agreement with the chosen thinning values.

Table 5.2: Walkers, steps by walker, burn-in and thinning for every performed case.

Case	Walkers	Steps	Burn-in without thinning	Thinning value
I	120	13 000	2 000	100
II	200	13 000	4 000	100
III	120	10 000	3 000	100
IV	100	16 000	5 000	100

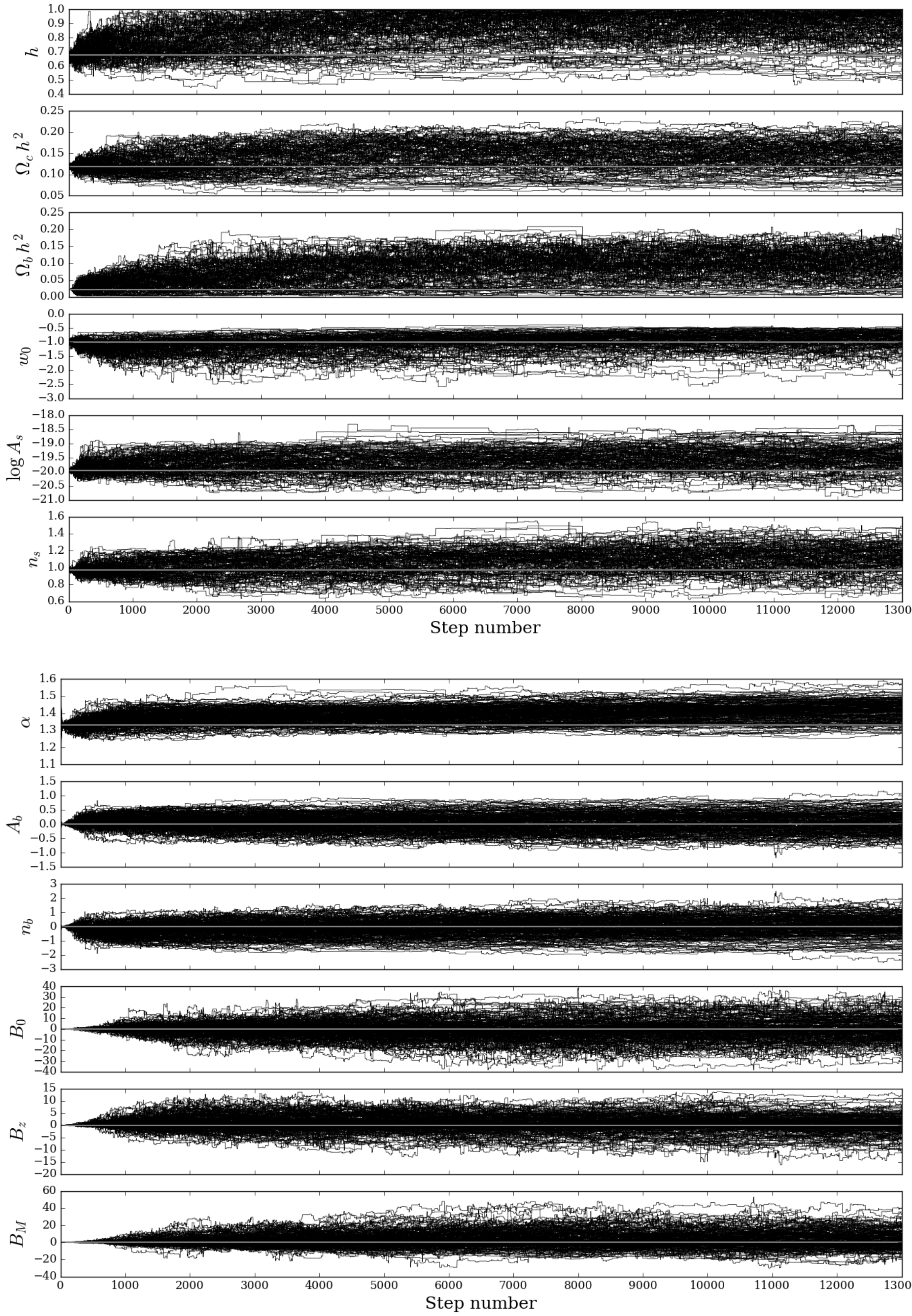

Figure 5.2: Position of each walker as function of the number step with fixed nuisance parameters (**Top**) and with fixed cosmological parameters (**Bottom**). The gray lines represent the true value of the parameters.

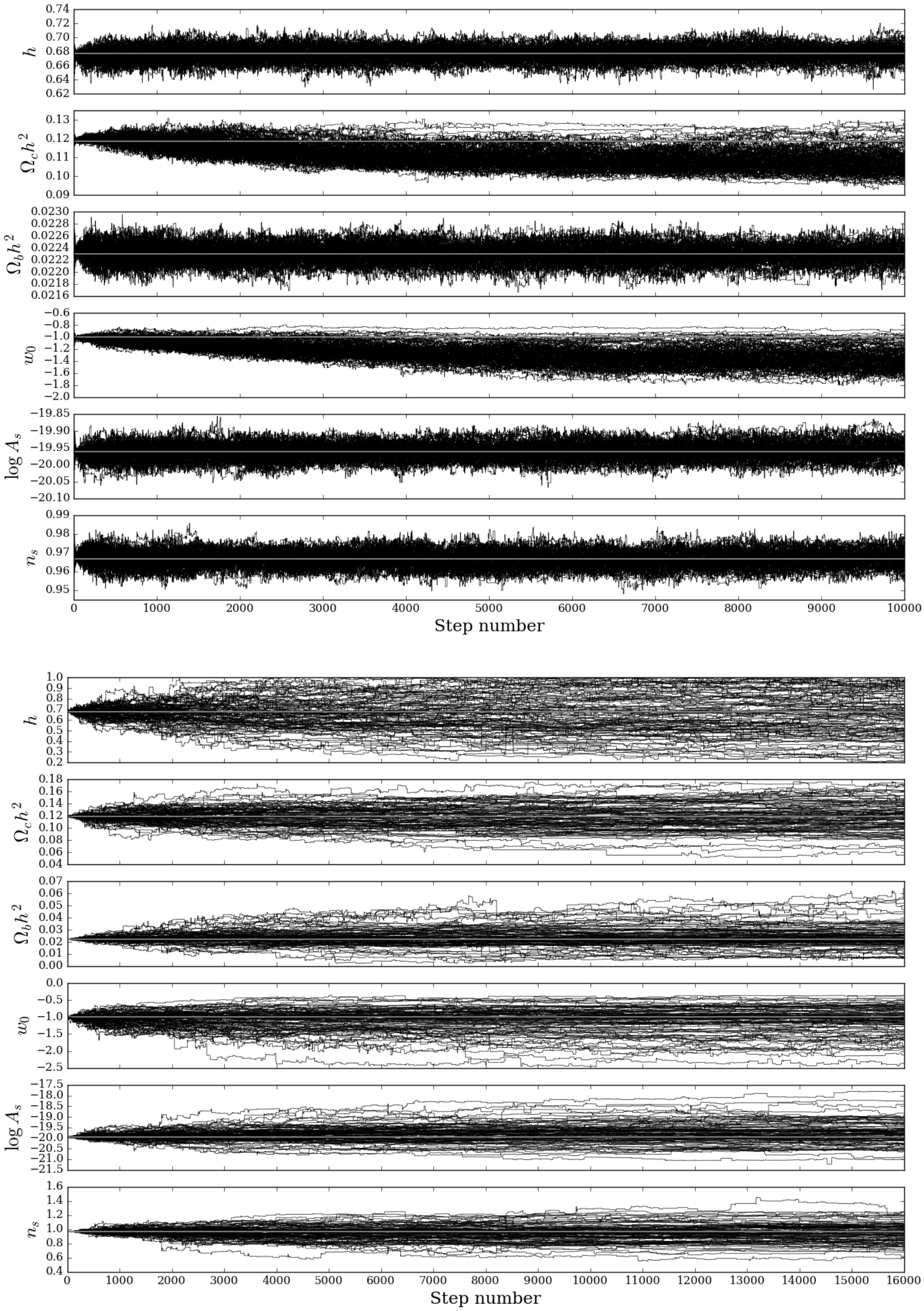

Figure 5.3: Position of each walker as function of the number step with Gaussian priors (**Top**) and without priors (**Bottom**). The gray lines represent the true value of the parameters.

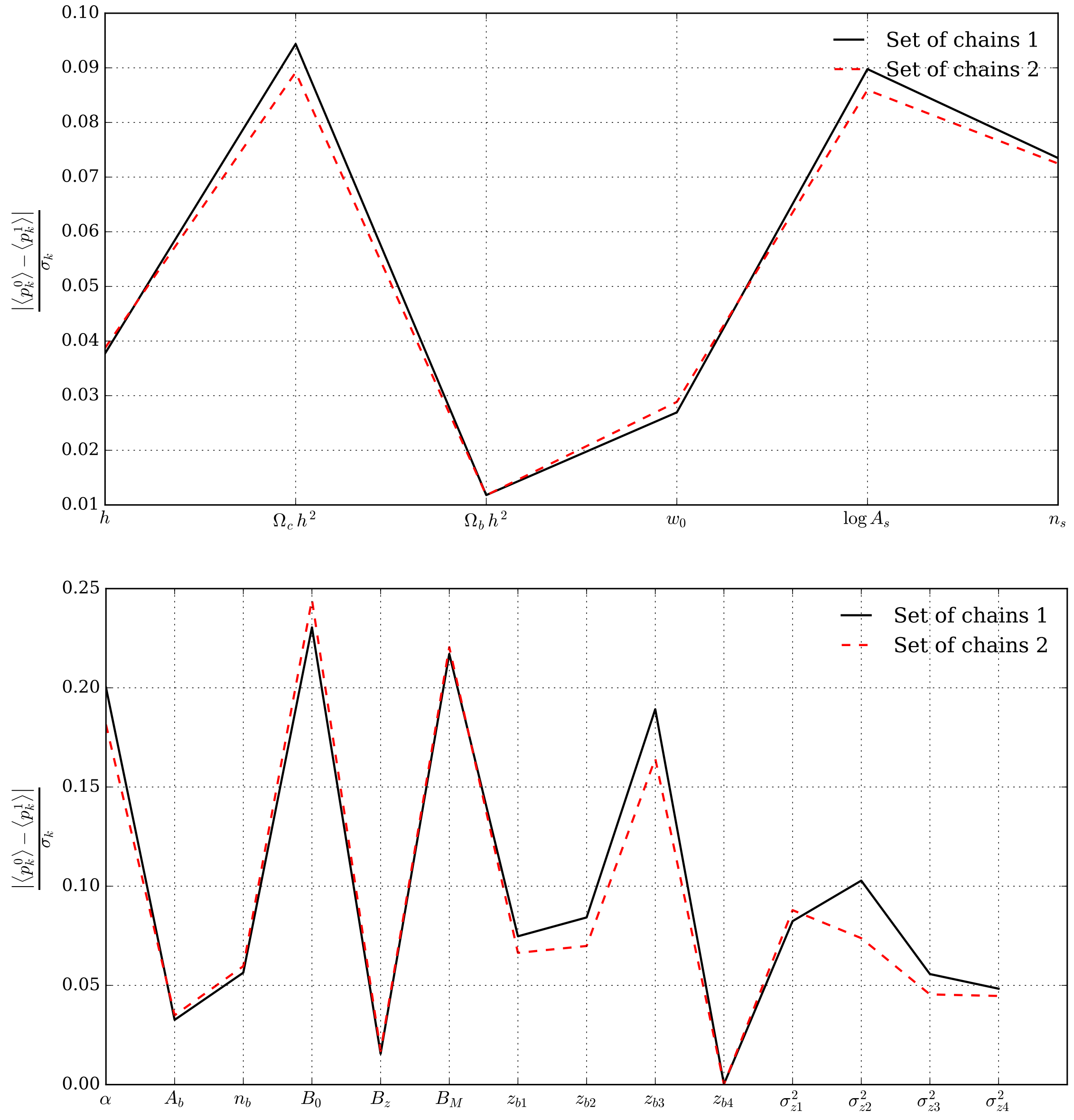

Figure 5.4: Convergence test in each parameter with fixed nuisance parameters (**Top**) and with fixed cosmological parameters (**Bottom**). Black solid line indicates the convergence ratio by using the sigma value for the first set of chains, red dashed line indicates the convergence ratio by using the sigma value for the second set of chains. The computed ratio (5.36) in all parameters for both panels show us that the remaining sample after the burn-in process is reliable to perform the statistical analysis. Note that the ratio does not exceed the value ~ 0.1 for the top panel and the value ~ 0.25 for the bottom panel.

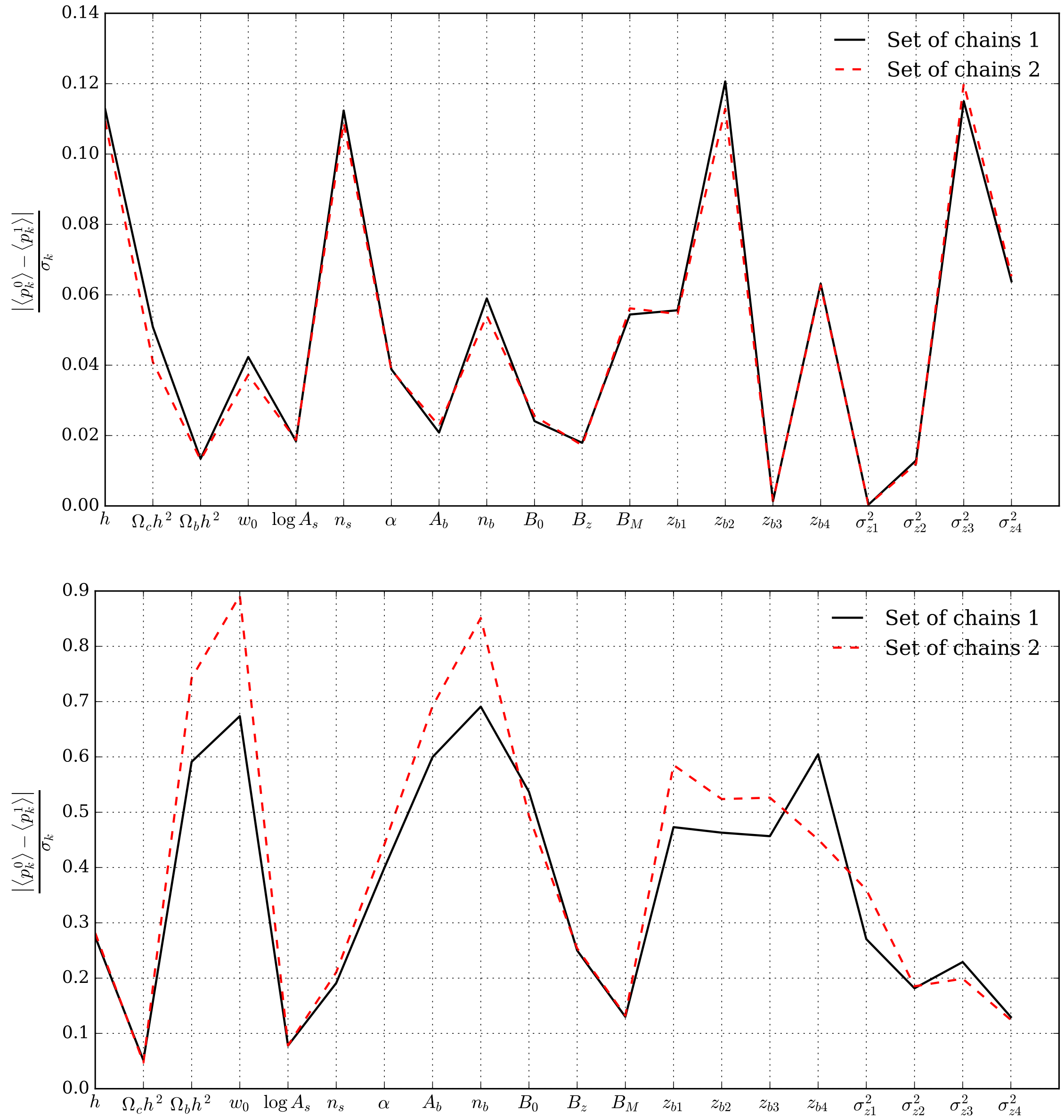

Figure 5.5: Convergence test in each parameter with Gaussian priors (**Top**) and without priors (**Bottom**), see equation (5.36). Black solid line indicates the convergence ratio by using the sigma value for the first set of chains, red dashed line indicates the convergence ratio by using the sigma value for the second set of chains. The top panel shows that the sample obtained after the burn-in process is reliable. Observe that in this case the ratio (5.36) threshold is ~ 0.12, which is a good value. On the other hand, in the bottom panel, we have a reasonable threshold ratio of ~ 0.9, considering that we are constraining 20 model parameters without priors.

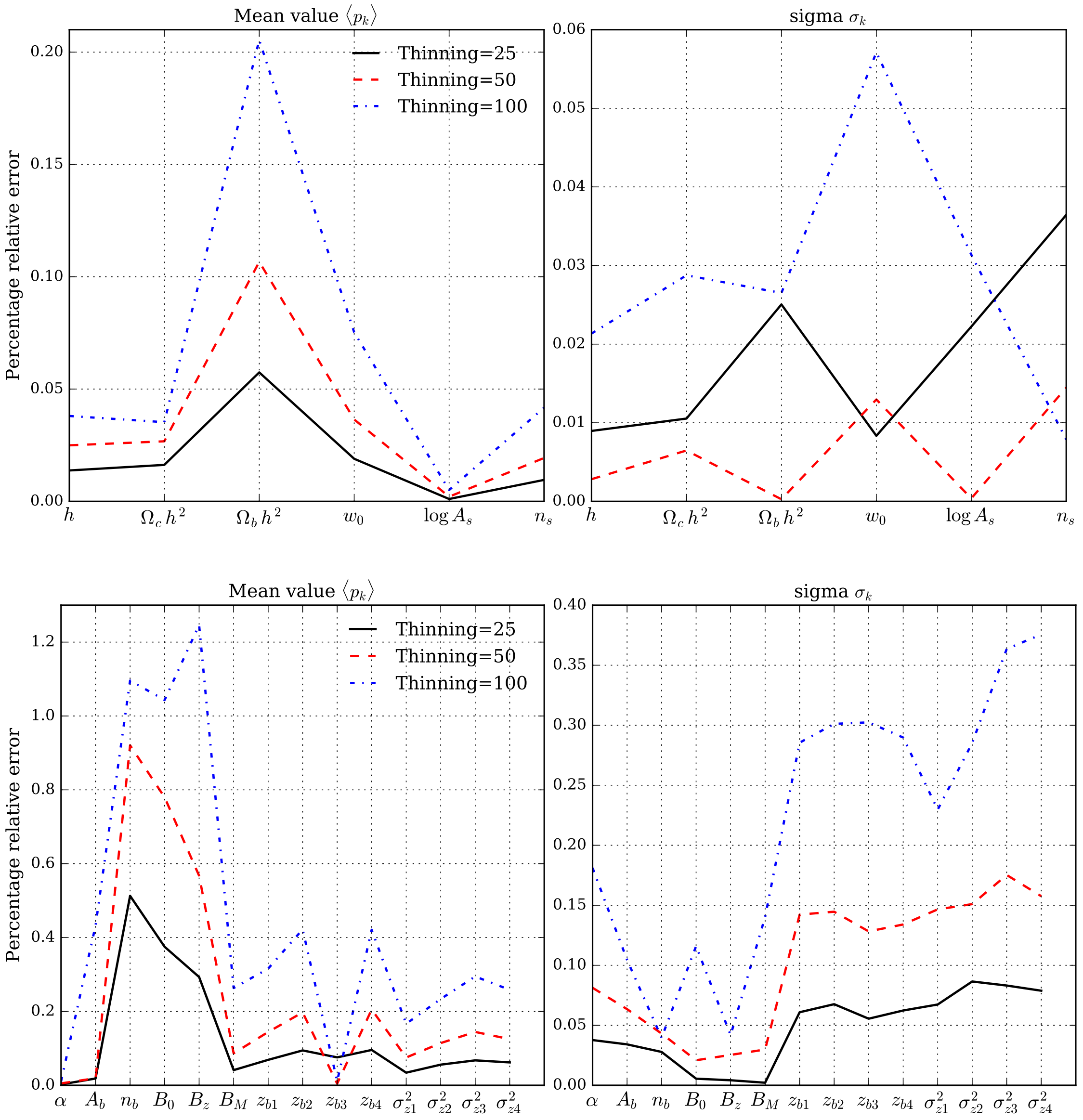

Figure 5.6: Percentage relative error for three thinning tests in each parameter with fixed nuisance parameters (**Top**) and with fixed cosmological parameters (**Bottom**), see equation (5.36). Note that in both cases, the thinning method affects more the mean value than the sigma one. Nonetheless for the left-top plot, we observe that the percentage relative error in all parameters is less than 0.05%, except in $\Omega_b h^2$ in which the error reaches $\sim 0.2\%$ in the test with Thinning $= 100$. On the other hand, the percentage relative error reaches a value of $\sim 1.2\%$ in the test with Thinning $= 100$ for the left-bottom plot, wherein n_s, B_0 and B_z are the parameters with larger error. The relative error does not exceed the $\sim 0.42\%$ in the other parameters.

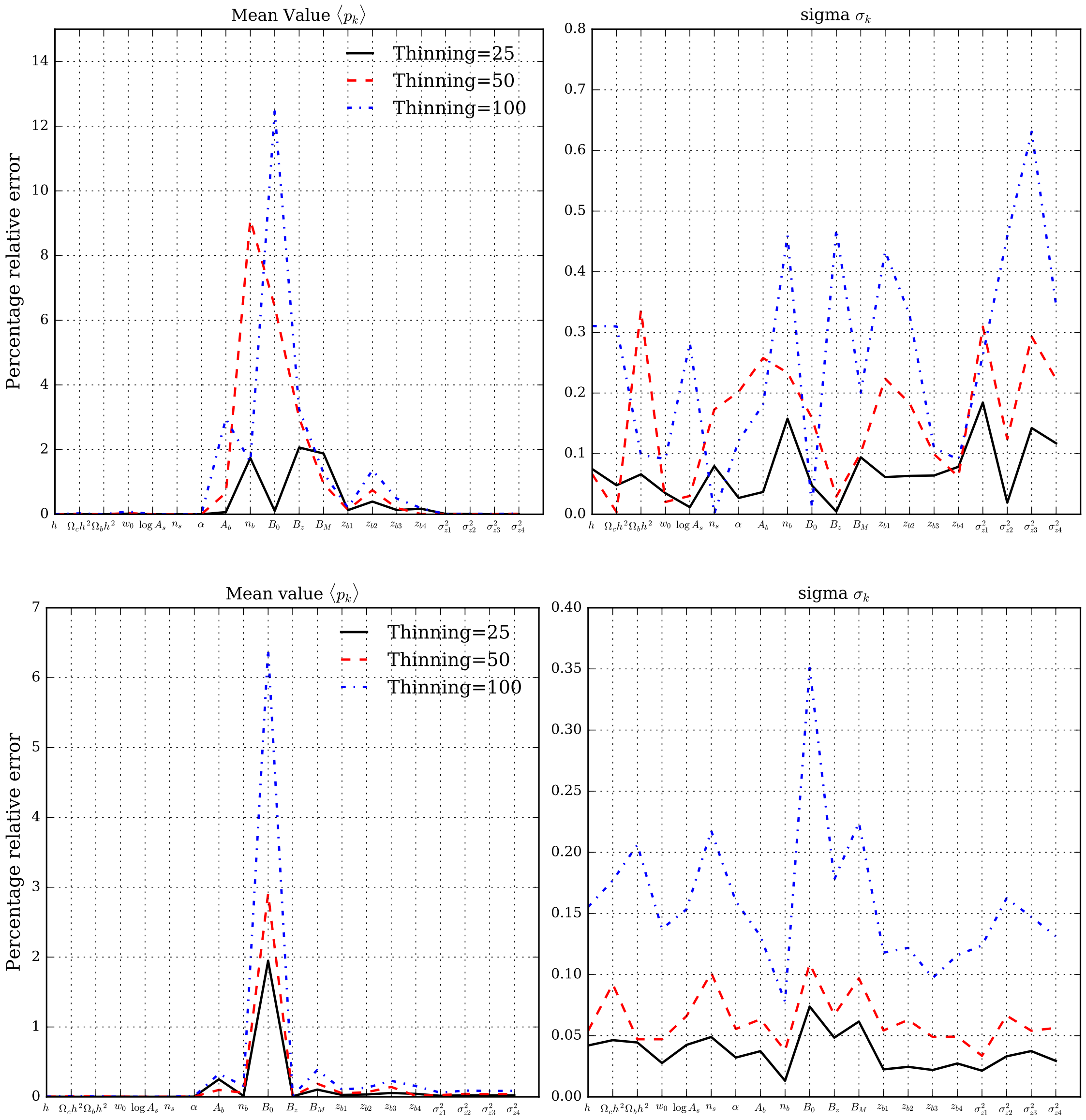

Figure 5.7: Percentage relative error for three thinning tests in each parameter with Gaussian priors (**Top**) and without priors (**Bottom**). The thinning method affects more the mean value than the sigma one, in both cases. The left-top plot shows that the relative error reaches a value of $\sim 13\%$ for the parameter B_0 (Thinning $= 100$) and a value of $\sim 9\%$ for the parameter A_b (Thinning $= 50$). However, for the other parameters the relative error does not exceed a value of $\sim 3\%$. The left-bottom plot shows that for all parameters, the relative error does not exceed a value of $\sim 0.2\%$, except to parameter B_0, wherein the error reaches a value of $\sim 6.2\%$.

5.4.1 Constraining the parameters

For the determination of cosmological parameters via cluster number counts, we use the `emcee` software package described in section 5.2.3. We will focus the discussion in the density parameter of cold dark matter normalized with the Hubble parameter $\Omega_c h^2$ and the parameter which characterizes the equation of state of the dark energy w_0.

Figure 5.8 shows the 1σ and 2σ confidence region for the constraint of cosmological parameters with fixed nuisance parameters and for the constraint of nuisance parameters with fixed cosmological parameters, i.e, case I and case II. We observe that for both cases, the true values are recovered at the 2σ level. However, the constraints on cosmological parameters are weak unlike the constraints on nuisance parameters. The marginalized constraints on cosmological parameters show that we achieve to estimate the $\Omega_c h^2$ and w_0 in 1σ confidence interval. The parameter h is that one with the worst constraint in this test.

Figure 5.9 shows the constraints of the cosmological parameters with priors and without priors, i.e., case III and case IV. The main conclusion in this figure is that the priors in the parameters without interest improve the constraints. In the case III, we observe a positive covariance between $\Omega_c h^2$ and w_0 which is not present in the other cases. We see the best fit point within the 1σ error contours in most cases and within the 2σ error contours in all cases which indicate that the error model we are using is good and viable. We compute the cluster number counts for every bin of redshift and richness comparing the fiducial model with the best-fitting at 1σ level, for the case III (with priors) and the case IV (without priors). Note that the values estimated by MCMC sampling at 1σ level achieve to recover the number counts predicted by the fiducial model. Furthermore, we observe that the application of priors in the parameters improves the confidence region in the number counts.

Table 5.3 shows the marginalized results, the mean value and the standard deviation for the model parameters in the four cases analyzed. We can see that the priors increase the precision in the estimation of the parameters, in other words the case III shows lower values of the standard deviation than in the other cases. E.g., for the parameter $\Omega_c h^2$ the standard deviation in the case III is $\sigma_{\Omega_c h^2} = 0.0059$ whereas in the other cases, the standard deviation is an order of magnitude greater. We note that the self-calibration method using only number counts allows us to recover the power-law index α for the case IV (no priors). This method can complement other calibration methods for the observable-mass relation, thereby obtaining more reliable results.

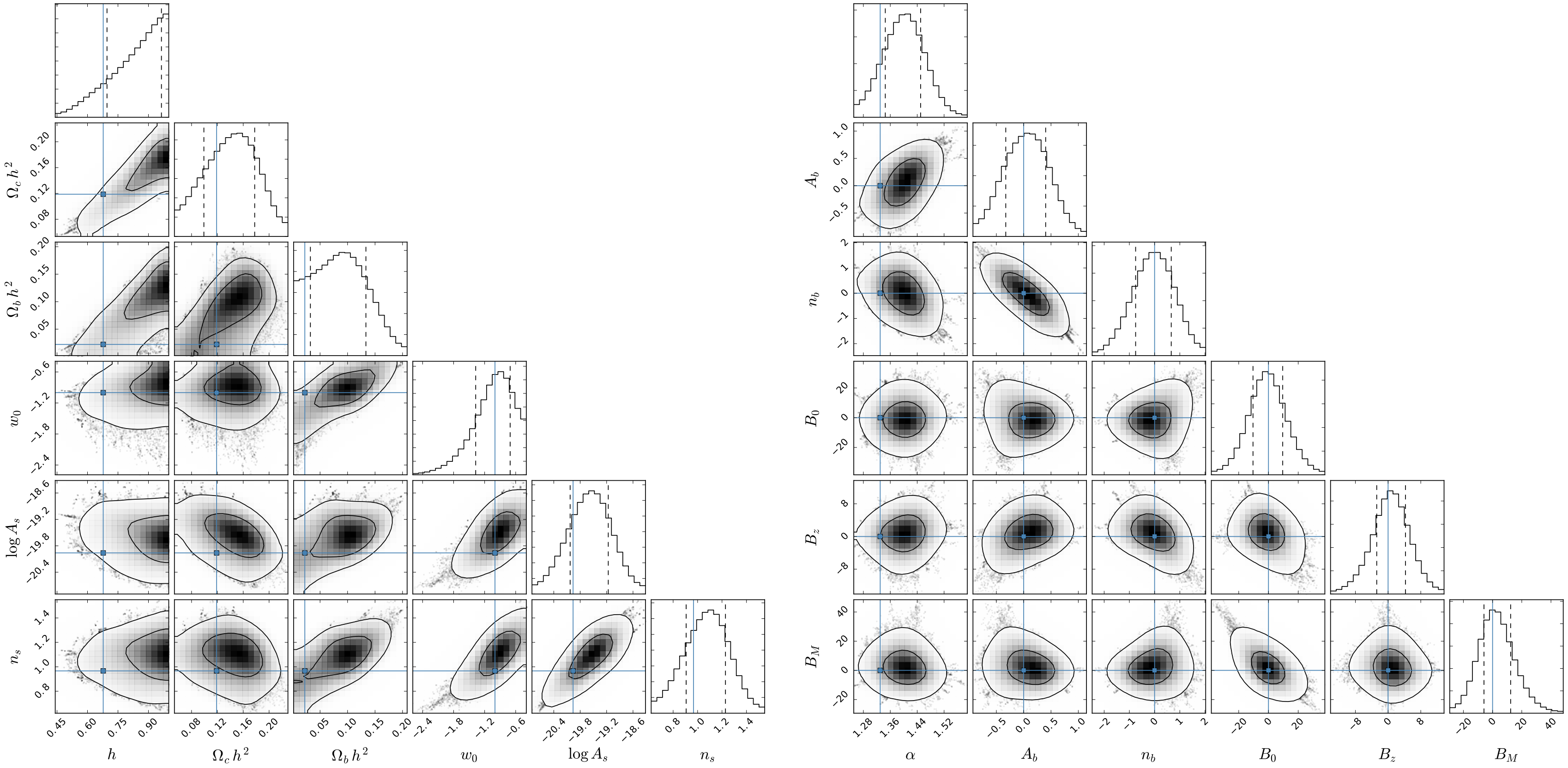

Figure 5.8: Constraint of model parameters for both 1σ and 2σ confidence regions with fixed nuisance parameters (**Top**) and with fixed cosmological parameters (**Bottom**). Note that for both cases, the true values are recovered until the 2σ level. However, the constraints on cosmological parameters in the top plot are weak unlike the constraints on nuisance parameters in the bottom plot.

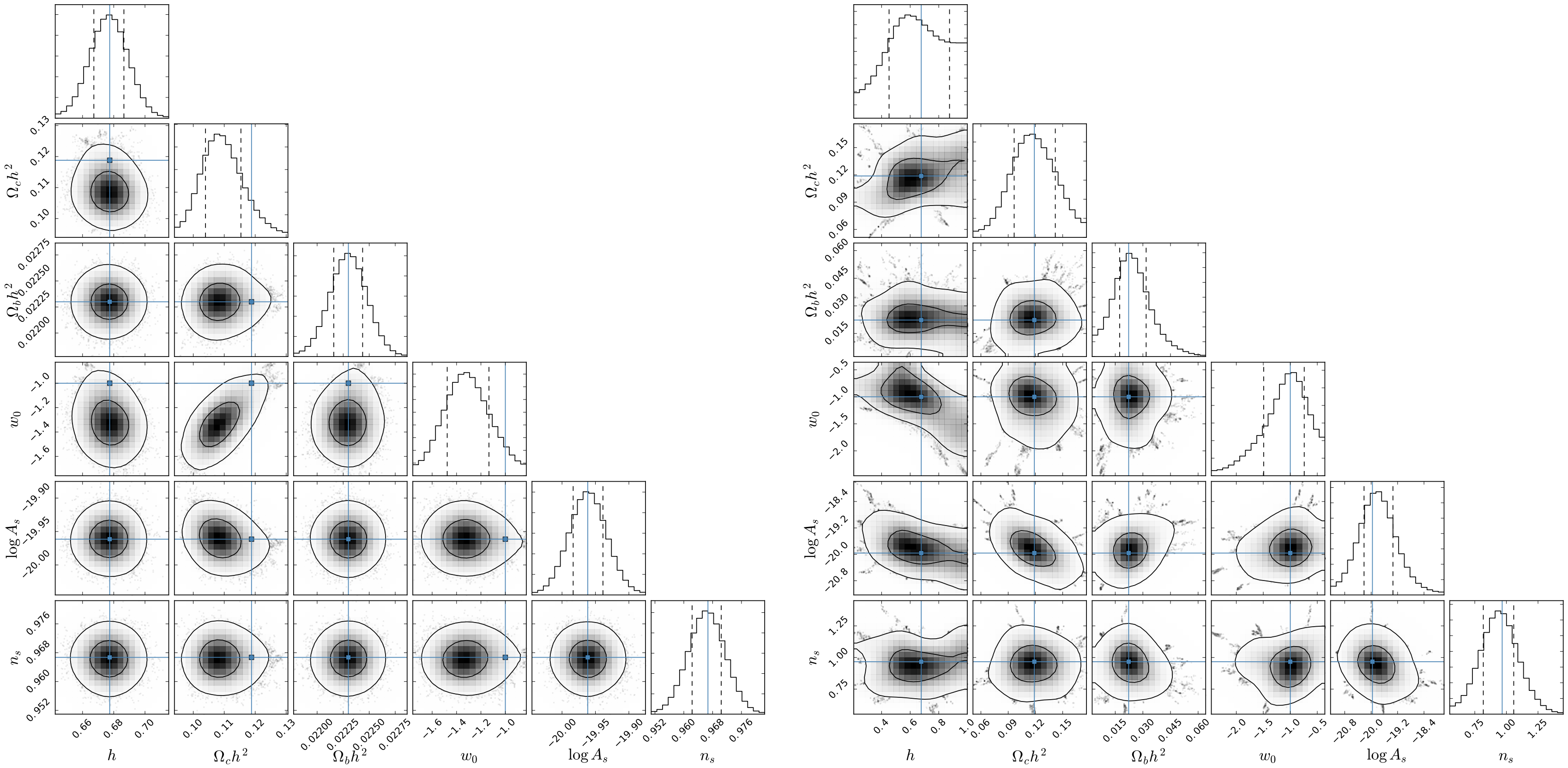

Figure 5.9: Constraint of cosmological parameters for both 1σ and 2σ confidence regions with Gaussian priors (**Top**) and without priors (**Bottom**). The top plot shows as the constraints improve with priors in the parameters without interest. Nonetheless, the true values of the $\Omega_c h^2$ and w_0 parameters are outside 1σ confidence region. In the case with free parameters, bottom plot, the true values are recovered at 1σ level, but their constraints get worse.

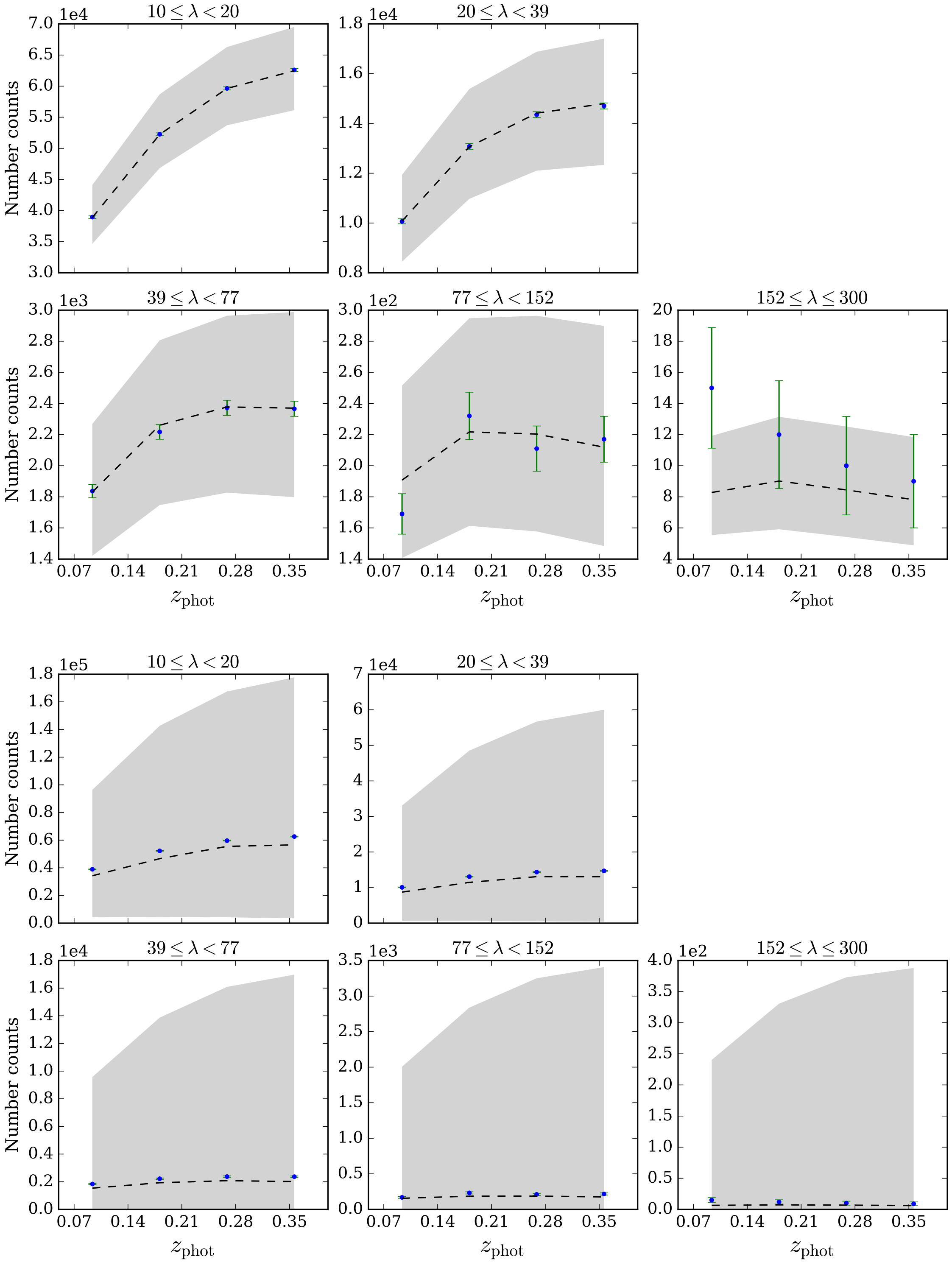

Figure 5.10: Cluster number counts for every bin of redshift and richness comparing the fiducial model with the best-fitting at 1σ level (black dashed line), for the case with priors (**Top**) and the case without priors (**Bottom**). The gray areas indicate the 1σ confidence regions of the best-fitting of the cluster abundance model.

Table 5.3: Best-fitting values at 1σ confidence level, the mean value and the standard deviation for the model parameters in the four cases analyzed.

Parameter	Best fit (I)	Mean (I)	Best fit (II)	Mean (II)	Best fit (III)	Mean (III)	Best fit (IV)	Mean (IV)
h	$0.86^{+0.10}_{-0.17}$	0.84 ± 0.12	—	—	$0.6769^{+0.0095}_{-0.0097}$	0.6769 ± 0.0097	$0.65^{+0.23}_{-0.20}$	0.65 ± 0.19
$\Omega_c h^2$	$0.143^{+0.034}_{-0.044}$	0.140 ± 0.036	—	—	$0.1089^{+0.0065}_{-0.0050}$	0.1096 ± 0.0059	$0.118^{+0.024}_{-0.021}$	0.118 ± 0.021
$\Omega_b h^2$	$0.086^{+0.047}_{-0.054}$	0.085 ± 0.045	—	—	0.0223 ± 0.0001	0.0223 ± 0.0001	$0.023^{+0.009}_{-0.006}$	0.024 ± 0.009
w_0	$-0.97^{+0.26}_{-0.41}$	-1.03 ± 0.35	—	—	$-1.32^{+0.18}_{-0.16}$	-1.31 ± 0.17	$-1.03^{+0.29}_{-0.46}$	-1.10 ± 0.38
$\ln A_s$	$-19.57^{+0.41}_{-0.46}$	-19.59 ± 0.44	—	—	$-19.961^{+0.023}_{-0.022}$	-19.961 ± 0.023	$-19.82^{+0.48}_{-0.40}$	-19.79 ± 0.48
n_s	$1.09^{+0.14}_{-0.18}$	1.07 ± 0.16	—	—	0.9663 ± 0.0040	0.9663 ± 0.0041	$0.95^{+0.11}_{-0.13}$	0.95 ± 0.12
α	—	—	$1.403^{+0.047}_{-0.058}$	1.400 ± 0.052	$1.307^{+0.014}_{-0.013}$	1.31 ± 0.013	$1.40^{+0.15}_{-0.12}$	1.42 ± 0.14
A_b	—	—	$0.06^{+0.34}_{-0.39}$	0.05 ± 0.35	$0.004^{+0.051}_{-0.046}$	0.005 ± 0.048	$0.09^{+0.40}_{-0.59}$	0.03 ± 0.53
n_b	—	—	$-0.03^{+0.69}_{-0.72}$	-0.05 ± 0.70	$0.0004^{+0.0472}_{-0.0510}$	-0.001 ± 0.049	$-0.48^{+1.03}_{-0.70}$	-0.34 ± 0.97
B_0	—	—	$-0.85^{+10.43}_{-9.56}$	-0.55 ± 10.54	0.001 ± 0.051	0.001 ± 0.051	$-0.10^{+1.41}_{-1.42}$	-0.02 ± 1.65
B_z	—	—	$0.67^{+3.59}_{-3.45}$	0.63 ± 3.84	$0.001^{+0.051}_{-0.048}$	0.002 ± 0.050	$0.15^{+1.33}_{-0.96}$	0.33 ± 1.34
B_M	—	—	$1.46^{+10.91}_{-7.35}$	3.04 ± 10.08	$0.002^{+0.049}_{-0.048}$	0.002 ± 0.049	$0.16^{+1.45}_{-1.05}$	0.50 ± 1.78
z_{b1}	—	—	$0.21^{+0.14}_{-0.16}$	0.20 ± 0.15	$0.019^{+0.020}_{-0.018}$	0.019 ± 0.019	$0.21^{+0.17}_{-0.22}$	0.20 ± 0.18
z_{b2}	—	—	$0.17^{+0.16}_{-0.18}$	0.16 ± 0.16	-0.004 ± 0.020	-0.004 ± 0.020	$0.15^{+0.17}_{-0.18}$	0.14 ± 0.17
z_{b3}	—	—	$-0.05^{+0.24}_{-0.21}$	-0.04 ± 0.23	$-0.013^{+0.027}_{-0.030}$	-0.014 ± 0.029	$0.04^{+0.21}_{-0.19}$	0.04 ± 0.19
z_{b4}	—	—	$-0.25^{+0.29}_{-0.27}$	-0.24 ± 0.28	$-0.013^{+0.036}_{-0.043}$	-0.015 ± 0.040	$-0.05^{+0.24}_{-0.23}$	-0.05 ± 0.21
σ^2_{z1}	—	—	$0.25^{+0.11}_{-0.09}$	0.27 ± 0.11	$0.135^{+0.019}_{-0.018}$	0.136 ± 0.019	$0.26^{+0.13}_{-0.11}$	0.27 ± 0.13
σ^2_{z2}	—	—	$0.37^{+0.18}_{-0.16}$	0.40 ± 0.21	$0.158^{+0.023}_{-0.021}$	0.160 ± 0.022	$0.31^{+0.26}_{-0.15}$	0.35 ± 0.19
σ^2_{z3}	—	—	$0.64^{+0.40}_{-0.33}$	0.69 ± 0.37	$0.211^{+0.030}_{-0.027}$	0.212 ± 0.028	$0.44^{+0.32}_{-0.22}$	0.51 ± 0.29
σ^2_{z4}	—	—	$0.68^{+0.42}_{-0.36}$	0.73 ± 0.38	$0.254^{+0.032}_{-0.029}$	0.255 ± 0.030	$0.55^{+0.34}_{-0.30}$	0.58 ± 0.31

5.5 Summary and conclusions

Here we constrain the cosmological parameters through galaxy cluster abundance. In order to include the effects due to the photometric redshifts and the estimated observable mass, we employ the self-calibration model proposed by Lima & Hu (2007). The Tinker mass function is used to estimate the number counts in a mass range and a redshift bin, see Tinker et al. (2008). We assume that the richness and the cluster mass are related through a power law according to Rozo et al. (2010); Mana et al. (2013); Simet et al. (2017). We introduce the concept of Bayesian statistics and of Monte Carlo sampling via Markov chains (i.e., MCMC methods). We use the `emcee` software package in our analysis. For the cosmological forecasting, we assume a flat ΛCDM as fiducial model. The synthetic data are obtained from a Poisson sampling. We set the likelihood for this problem to be a product of independent Poisson distributions, one for each bin.

For our analysis, we have performed four tests: an error analysis of free cosmological parameters with fixed nuisance parameters; free nuisance parameters with fixed cosmological parameters; flat prior on the density of dark matter and on the equation of dark energy with Gaussian priors in the other parameters; and a flat prior on all parameters. We set the burn-in in each case through visual analysis of the walkers against step number. We note that the tests with more free parameters need a high burn-in range. The case with all free parameters presents more difficulty to obtain a successful convergence test. We observe that the thinning process applied to the chains does not produce significant changes in the mean value and variance for every parameter involved in the study. We observe that the self-calibration method via cluster abundance allows us to recover the fiducial values in all considered cases. On the other hand, Lima & Hu (2004); Lima & Hu (2005); Lima & Hu (2007); Aguena & Lima (2016) show that the inclusion of clustering of clusters as well as the purity and completeness in the analysis could solve the degeneracy problems and improve the results. In fact, the implementation of these observables would be the next step to follow in this work. We observe that the self-calibration method along with other methods, e.g., weak gravitational lensing Simet et al. (2017), allow us to improve the mass-richness relation according to the results obtained for the power law index α.

Chapter 6

Conclusions

This work is focused on the study of optical galaxy clusters in three steps: analysis and estimation of photometric redshifts for galaxies, detection of galaxy clusters from galaxy survey and constraining of parameters from a given cosmological model via cluster abundance. We perform a review of the basic theoretical framework for understanding the formation of galaxy clusters. We present the fundamental concepts of standard cosmology in the background as well as beyond of the background (linear theory). The conclusions are:

I. **Analysis and estimation of photometric redshifts:**

   The detection of the optical galaxy cluster needs accurate and precise measurements of redshift for the galaxies. Photometric redshifts allow us to probe much larger volumes of the Universe than possible with spectroscopic redshifts, but they have large measurement uncertainties. We have approached this problem by using the `ANNz2` and `GPz` machine learning codes. We have investigated the degradation in the accuracy and precision of the recovered of photometric redshifts applying the machine learning methods to deep photometric datasets, which are trained using much shallower and brighter spectroscopic samples. For this analysis, we utilize the Monte-Carlo random sampling for defining a photometric redshift estimator based on the cumulative distribution function (CDF). We note that the distribution of the z-phot estimators based on the CDF fits better the PDF stacking of all galaxies in the data set. Nonetheless, these estimators have a greater scatter than their counterparts. We estimate the photometric redshift for the samples GAMA DEEP and GAMA MAIN (subsets of the SDSS DR12), which are trained by the spectroscopic GAMA survey.

In the degradation analysis, we observe that comparatively, the CDF-ANNz2 estimator shows better performance at higher redshifts, albeit with larger scatter. We show that the estimated photometric redshift loses quality for deeper cuts. We have problems estimating higher redshifts. This carries unwanted effects in the number density of galaxies and the increase of impurities in the detection of galaxy clusters. The density of galaxies given by CDF-ANNz2 estimator has the least error according the number density of galaxies given by z-spec data in deeper cuts and high redshifts.

II. **Detection of galaxy clusters:**

We present the VT-FOFz cluster finder, which combines two techniques Voronoi Tessellation (VT) and Friends of Friends (FOF). We employ the CDF-ANNz2 estimator as single value of the photometric redshift for the galaxies. By using the mock catalogs, we observe that the cut in r-band magnitude as well as the quality of estimated photometric redshifts play an important role in the detection of galaxy clusters and in the redshift depth of the cluster catalog. We note that approximately the 40% of densest galaxies are candidates to form galaxy clusters, for all r-band cuts. We compute the completeness and purity to assess the performance of the cluster finder. At high redshift we detect less clusters, hence the completeness worsens. However, a large number of the few detected galaxy clusters are matched with the haloes in that redshift region, thus implying high purity. The massive haloes tend to match clusters with high richness as it is expected. There is a large scatter in the observable-mass relation which obtain from the mocks.

We run the VT-FOFz cluster finder in a sample of the GAMA DEEP survey. We compare the obtained cluster catalog with the redMaPPer SDSS DR8 cluster catalog. We observe that the Voronoi selection as well as the transverse linking length affect in the same way the detection of galaxy clusters. The selection of low values of P_{cut} and b_r parameters allows us to avoid the detection of large and/or rich galaxy clusters, which do not match with redMaPPer clusters. Low values of these parameters imply lost in the detection of true galaxy cluster. For $N_{g,\text{GD}} \geq 5$ we recover a large number of redMaPPer cluster ($\mathbf{C} > 0.9$) until $z \approx 0.33$, but with low purity. The results obtained for $z > 0.33$ are less reliable. The Voronoi selection allows us to reduce the scatter in the redshift relation between the GAMA DEEP clusters and redMaPPer clusters.

Comparing the redshift distribution for $N_{g,\mathrm{GD}} \geq 15$ and $N_{g,\mathrm{GD}} \geq 20$ in the cases in which we recover a large number of redMaPPer clusters, we show that the case with Voronoi selection allows us to obtain the best results.

III. **Constraint of the cosmological parameters:**

We forecast constraints of the cosmological parameters through galaxy cluster abundance through a MCMC method. We use the self-calibration model proposed by Lima & Hu (2007) to include the effects due to the photometric redshifts and the estimated observable mass. We employ the Tinker's mass function to estimate the number counts in a range of mass and a redshift bin. We assume that the richness and the cluster mass are related through a power law according to Rozo et al. (2010); Mana et al. (2013); Simet et al. (2017).

For our analysis, we propose four tests: flat priors for cosmological parameters with fixed nuisance parameters; flat priors for nuisance parameters with fixed cosmological parameters; flat priors for density of dark matter and equation of dark energy with Gaussian priors in the other parameters; and flat priors for all parameters. We note that the self-calibration method via cluster abundance allows us to recover the fiducial values in all considered cases. We observe that the self-calibration method along with other methods, e.g., weak gravitational lensing (Simet et al. (2017)), allow us to improve the mass-richness relation according to the results obtained for the power law index α.

We set a list of the possible future works based on the obtained results presented in this thesis.

- Use the CDF-ANNz2 estimator for determining the single value of the photometric redshift in deeper and larger photometric surveys, which are currently being developed, e.g., DES, LSST, Euclid, WFIRST, J-PAS.

- Implement the full probability distribution function of the photometric redshift in the VT-FOFz cluster finder.

- Determine the mass-richness relation for the galaxy clusters detected by the VT-FOFz cluster finder through weak gravitational lensing analysis, see Simet et al. (2017).

- Introduce clustering of clusters as well as completeness and purity in the constraint of cosmological parameters through galaxy cluster number counts.

- Employ the detected galaxy clusters from the VT-FOFz cluster finder to constrain cosmological parameters.

Appendix A

General relativity for cosmology

In the beginning of the 20th century Albert Einstein proposed the concept that gravity is not a force, but is rather the local manifestation of spacetime geometry. This gravity theory is called general relativity (GR). He used the Riemannian geometry to explain his theory. In order to understand the concepts involved in general relativity, we are going to explain the most important concepts of differential geometry used here.

Definition A.0.1 (Manifold)**.** Set of pieces which can be "sewed" smoothly. More precisely, a real manifold $\mathcal{M}$, n-dimensional and c^{∞} is a set with a collection of open sets $\{O_\alpha\}$ which satisfies the following properties:

1. $\forall\, p \in \mathcal{M}$ is at least in a O_α (i.e. $\{O_\alpha\}$ cover $\mathcal{M}$).
2. $\forall\, \alpha$ there is an one to one function $\varphi_\alpha : O_\alpha \to U_\alpha$ called coordinate chart, where U_α is an open subset which belongs to $\mathbb{R}^n$.
3. If two open set O_α and O_β are overlapped $O_\alpha \cap O_\beta = \emptyset$ we can consider the function $\varphi_\beta \circ \varphi_\alpha^{-1}$ which takes points in $\varphi_\alpha(O_\alpha \cap O_\beta) \subset U_\alpha \subset \mathbb{R}^n$ to $\varphi_\beta(O_\alpha \cap O_\beta) \subset U_\beta \subset \mathbb{R}^n$.

The figure A.1 shows a representation of the manifold concept.

Definition A.0.2 (Metric)**.** The metric associated to the manifold $\mathcal{M}$ is a non-degenerate symmetric tensor of type (0,2), which satisfies the following properties

1. $g(v_1, v_2) = g(v_2, v_1) \;\; \forall\, v_1, v_2 \in V_p$. (Symmetry)
2. $g(v, v_1) = 0 \;\; \forall\, v \in V_p$ iff $v_1 = 0$. (Non-degeneration)

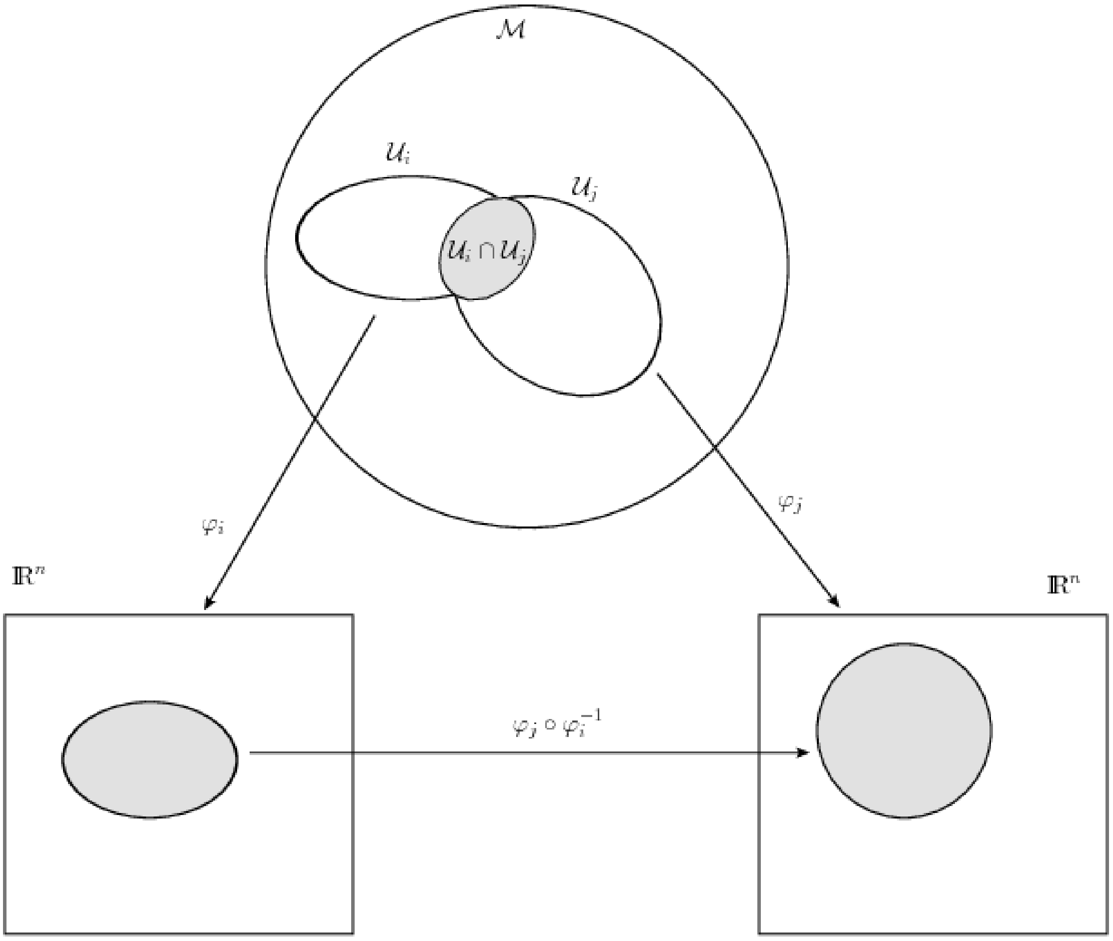

Figure A.1: Representation of a manifold. Adapted from Coutant (2012)

Here V_p is the tangent space at p point of $\mathcal{M}$. The metric defines the inner product in the tangent space V_p. Given a coordinate basis we can write the metric as

$$\mathrm{d}s^2 = g_{ab}\mathrm{d}x^a\mathrm{d}x^b, \text{ thus } g(u,v) = g_{ab}u^a v^b, \ \forall\, u,v \in V_p. \tag{A.1}$$

The metric can be interpreted as the measure of an infinitesimal distance in the manifold. Therefore, the metric allows the computation of path length proper time, in addition it allows us to determine the shortest distance between two points.

Spacetime is modeled as a 4-dimensional Lorentzian manifold, where the metric signature is (3,1) or (-,+,+,+). This allows us to classify the tangential vectors in three types:

- timelike if $g(v,v) < 0$,

- null or lightlike if $g(v,v) = 0$,
- spacelike if $g(v,v) > 0$,

where $v \in V_p$.

Definition A.0.3 (Covariant derivative)**.** Just as in differential calculus, we define the concept of a manifolds here. The covariant derivative is a map from (k,l) tensor fields to $(k,l+1)$ tensor fields, which acts linearly on its arguments and obeys the Leibniz rule on tensor products. In a coordinate basis the covariant derivative is given by

$$\nabla_a T^{b_1\ldots b_k}_{c_1\ldots c_l} = \partial_a T^{b_1\ldots b_k}_{c_1\ldots c_l} + \sum_i \Gamma^{b_i}_{ad} T^{b_1\ldots d\ldots b_k}_{c_1\ldots c_l} - \sum_j \Gamma^{d}_{ac_j} T^{b_1\ldots b_k}_{c_1\ldots d\ldots c_l}, \tag{A.2}$$

where Γ is a tensorial density called connection. This geometric object allows us to keep the covariance in the derivative.

The torsion tensor T^c_{ab} for a manifold may be defined as

$$\mathrm{T}^c_{ab} = \Gamma^c{}_{ab} - \Gamma^c_{ba}. \tag{A.3}$$

We say that a connection is metric compatible if the covariant derivative of the metric with respect to that connection is everywhere zero, see Carroll (2004). The fundamental theorem of Riemannian geometry states that on any Riemannian manifold with metric g_{ab} there is a unique metric compatible connection with torsion free. In other words, it is satisfied the following properties:

- **Torsion free:**

$$\mathrm{T}^c_{ab} = 0 \;\text{ thus }\; \Gamma^c{}_{ab} = \Gamma^c_{ba}. \tag{A.4}$$

- **Metric compatible:**

$$\nabla_a g_{bc} = 0. \tag{A.5}$$

The connection coefficients are called Christoffel symbols and these are given by

$$\Gamma^a{}_{ab} = \frac{1}{2} g^{cd} \left(\partial_a g_{bd} + \partial_b g_{ad} - \partial_d g_{ab}\right). \tag{A.6}$$

Definition A.0.4 (Riemann tensor)**.** The geometric object which quantifies the intrinsic curvature is called Riemann tensor. This is define as

$$\nabla_a \nabla_b w_c - \nabla_b \nabla_a w_c = R_{abc}{}^{d} w_d, \quad \forall\, w \in V_p^*, \tag{A.7}$$

where V_p^* is the dual space at p. The Riemann tensor is related with the Christoffel symbols through the following expression

$$R_{abc}{}^{d} = \partial_b \Gamma^d_{\ ac} - \partial_a \Gamma^d_{\ bc} + \Gamma^e_{\ ac} \Gamma^d_{\ eb} - \Gamma^e_{\ bc} \Gamma^d_{\ ea}. \tag{A.8}$$

The Riemman tensor satisfies the following properties:

1. $R_{abc}{}^{d} = -R_{bac}{}^{d}$,
2. $R_{[abc]}{}^{d} = 0$,
3. $R_{abcd} = -R_{abdc}$,
4. $\nabla_{[a} R_{bc]d}{}^{e} = 0$, Bianchi identities,
5. $R_{abcd} = R_{cdab}$.

The Ricci tensor is defined as

$$R_{ac} = R_{abc}{}^{b}. \tag{A.9}$$

The scalar curvature or the Ricci scalar is given by

$$R = R^a_{\ a} = g^{ab} R_{ab}. \tag{A.10}$$

Definition A.0.5 (Geodesic equation)**.** In Euclidean geometry the shortest distance between two points is the straight line; in Riemannian geometry this concept is generalized. We may think that a particle is moving on a sphere and its movement is restricted to the surface. The path that the particle must follow to travel the shortest distance is not the straight line, it is actually a curved path. In order to find this path in the manifold, we consider the coordinate chart φ, in which the geodesic can be seen as path x^μ on $\mathbb{R}^n$. The geodesic equation in $\mathbb{R}^n$ is given by

$$\frac{\mathrm{d}x^\nu}{\mathrm{d}\lambda} + \Gamma^\nu_{\ \mu\gamma} \frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda} \frac{\mathrm{d}x^\gamma}{\mathrm{d}\lambda} = 0, \tag{A.11}$$

where λ is called parameter affine.

As the geodesic is the shortest path between two points, we can find the equations (A.11) by using the Euler-Lagrange equations

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{\partial\mathcal{L}}{\partial\dot{x}^{\alpha}}\right)-\frac{\partial\mathcal{L}}{\partial x^{\alpha}}=0, \tag{A.12}$$

for the Lagrangian

$$\mathcal{L}=g_{\mu\nu}\dot{x}^{\mu}\dot{x}^{\nu}. \tag{A.13}$$

The above equations allow us to find the Christoffel symbols without using equation (A.6).

In order to find the field equations which describe the gravity theory proposed by Einstein, David Hilbert in 1915 proposed the following action

$$S_{\mathrm{EH}}=\int_{\nu}\mathrm{d}^{4}x\sqrt{-g}R, \tag{A.14}$$

where g is the determinant of the metric tensor, R is the scalar curvature and ν is a hypervolume of the spacetime. The above expression is known as Einstein-Hilbert action. Actually to find the field equations, we need the following general action

$$S=\frac{1}{2\kappa}\left(S_{\mathrm{EH}}+S_{\mathrm{GYH}}\right)+S_{\mathrm{M}}, \tag{A.15}$$

where κ is a constant, S_{GYH} is a boundary term to relax the boundary conditions and S_{M} is the action which involves the matter field. The term S_{GYH} is proposed by Gibbons-York-Hawking in 1977, see Gibbons & Hawking (1977); Hawking & Horowitz (1996). This term is given by

$$S_{\mathrm{GYH}}=2\oint_{\partial\nu}\mathrm{d}^{3}y\epsilon\sqrt{|h|}k. \tag{A.16}$$

Here $\partial\nu$ is the boundary of ν, h is the determinant of the induced metric, k is the trace of the extrinsic curvature and ϵ is ±1 depending whether $\partial\nu$ is timelike or spacelike (we assume that $\partial\nu$ is everywhere non-null). The x^{α} coordinates are used for the region ν and y^{α} coordinates for $\partial\nu$. The matter action is given by

$$S_{\mathrm{M}}=\int_{\nu}\mathrm{d}^{4}x\sqrt{-g}\,\mathcal{L}_{\mathrm{M}}\left(g_{\alpha\beta},\psi\right). \tag{A.17}$$

By using the properties of variational calculus, the Gauss-Stokes theorem and supposing that $\delta g_{\mu\nu} = 0$ in $\partial\nu$ as boundary condition, we have that the variation of Einstein-Hilbert action is given by

$$\delta S_{\mathrm{EH}} = \int_{\nu} \mathrm{d}^4x\sqrt{-g}\left(R_{\alpha\beta} - \frac{1}{2}Rg_{\alpha\beta}\right)\delta g^{\alpha\beta} - \oint_{\partial\nu} \mathrm{d}^3y\,\epsilon\sqrt{|h|}h^{\alpha\beta}n^{\sigma}\partial_{\sigma}\delta g_{\alpha\beta}, \tag{A.18}$$

where n^{μ} is a normal vector to $\partial\nu$ such that $n^{\mu}n_{\mu} = \epsilon = \pm 1$. The boundary term in the above expression can be avoided if we consider that term $\partial_{\sigma}\delta g_{\alpha\beta} = 0$ as other boundary condition. Although this argument carries to Einstein field equations we would be fixing two boundary conditions. To avoid the above, Hawking, York and Gibbons introduced the term S_{GYH} in the action which allows us to have a well defined variational problem by using only one boundary condition $\delta g_{\alpha\beta} = 0$. Therefore, by using the definition of extrinsic curvature we have

$$\delta S_{\mathrm{GYH}} = \oint_{\partial\nu} \mathrm{d}^3y\,\epsilon\sqrt{|h|}n^{\alpha}h^{\nu\gamma}\partial_{\alpha}\delta g_{\gamma\nu}. \tag{A.19}$$

The above result cancels the boundary term in the variation of Einstein-Hilbert action.

The variation of the matter action is given by

$$\delta S_{\mathrm{M}} = \int_{\nu} \mathrm{d}^4x\sqrt{-g}\left(\frac{\partial\mathcal{L}_{\mathrm{M}}}{\partial g^{\alpha\beta}} - \frac{1}{2}g_{\alpha\beta}\mathcal{L}_{\mathrm{M}}\right)\delta g^{\alpha\beta}. \tag{A.20}$$

We define the energy-momentum tensor as

$$T_{\alpha\beta} \equiv -2\frac{\partial\mathcal{L}_{\mathrm{M}}}{\partial g^{\alpha\beta}} + g_{\alpha\beta}\mathcal{L}_{\mathrm{M}}, \tag{A.21}$$

which is symmetric and satisfies

$$\nabla_{\beta}T^{\alpha\beta} = 0. \tag{A.22}$$

This equation is known as the energy conservation equation. Therefore, equation (A.20) can be rewritten as

$$\delta S_{\mathrm{M}} = -\frac{1}{2}\int_{\nu} \mathrm{d}^4x\sqrt{-g}\,T_{\alpha\beta}\delta g^{\alpha\beta}. \tag{A.23}$$

Thus, by combining equation (A.18), equation (A.19) and equation (A.23) we obtain

$$\delta S = \frac{1}{2}\left[\int_{\nu} \mathrm{d}^4x\sqrt{-g}\left(\frac{1}{\kappa}\left(R_{\alpha\beta} - \frac{1}{2}Rg_{\alpha\beta}\right) - T_{\alpha\beta}\right)\delta g^{\alpha\beta}\right]. \tag{A.24}$$

If we impose that the variation in the total action is invariant with respect to $\delta g^{\alpha\beta}$ then $\delta S = 0$, and replacing $\kappa = 8\pi G$, we obtain the Einstein field equations

$$R_{\alpha\beta} - \frac{1}{2}g_{\alpha\beta}R = 8\pi G T_{\alpha\beta}. \tag{A.25}$$

To consider the cosmological constant Λ we must modify the Einstein-Hilbert action in the following way

$$\tilde{S}_{\text{EH}} = \int_{\nu} \mathrm{d}^4x\sqrt{-g}R - 2\Lambda. \tag{A.26}$$

Then, the Einstein field equations with cosmological constant are given by

$$R_{\alpha\beta} - \frac{1}{2}g_{\alpha\beta}R + \Lambda g_{\alpha\beta} = 8\pi G T_{\alpha\beta}. \tag{A.27}$$

In conclusion, the general relativity can be summarized in the following postulates:

1. The spacetime is decribed by a 4-dimensional manifold $\mathcal{M}$ and a Lorentzian metric g on $\mathcal{M}$.

2. Local conservation of energy: It does exist a symmetric tensor $T_{\alpha\beta}(\psi) = T_{\beta\alpha}$ which is function of matter fields and satisfies

$$\begin{aligned} T_{\alpha\beta} &= 0 \text{ on } u \subset \mathcal{M} \text{ iff } \psi_{\text{i}} = 0 \ \forall\, i, && \text{(A.28)} \\ \nabla_{\beta}T^{\alpha\beta} &= 0. && \text{(A.29)} \end{aligned}$$

3. The metric g on $\mathcal{M}$ is determined by the Einstein field equations

$$R_{\alpha\beta} - \frac{1}{2}g_{\alpha\beta}R = 8\pi G T_{\alpha\beta}, \tag{A.30}$$

where $R_{\alpha\beta}$ is the Ricci tensor, R is the scalar curvature and $T_{\alpha\beta}$ is the energy-momentum tensor. Here we consider $c = 1$.

Appendix B

Newtonian treatment for matter perturbations

Newtonian gravitation is an adequate description of general relativity in the sub-horizon limit (i.e., $k\eta \gg 1$) for non-relativistic matter perturbations. We consider a non-relativistic fluid with density ρ, pressure $P \ll \rho$ and velocity $\vec{v}$. The position of a fluid element is given by $\vec{r}$ and the time is t. Equations of motion are given by the dynamics fluid theory

$$\begin{aligned} \frac{\partial \rho}{\partial t} + \nabla_r(\rho\vec{v}) &= 0, \text{ Continuity equation,} \qquad (B.1) \\ \left(\frac{\partial}{\partial t} + \vec{v}\cdot\nabla_r\right)\vec{v} &= \frac{\nabla_r P}{\rho} - \nabla_r\Phi, \text{ Euler's equation,} \\ \nabla_r^2\Phi &= 4\pi G\rho, \text{ Poisson's equation.} \end{aligned}$$

Disturbing the fluid properties, we have:

$$\begin{aligned} \rho = \rho_0 + \delta\rho(t,\vec{r}), &\qquad \vec{v} = \vec{v}_0 + \delta\vec{v}(t,\vec{r}), \qquad (B.2) \\ \Phi = \Phi_0 + \delta\Phi(t,\vec{r}), &\qquad P = P_0 + \delta P(t,\vec{r}), \end{aligned}$$

where the zero index denotes the background value. Here we assume adiabatic perturbations (entropy perturbations are null, see Mukhanov (2005)), therefore we have the following relation

$$\frac{\delta P}{\delta\rho} = \frac{\partial P}{\partial\rho} = c_s^2, \qquad (B.3)$$

where c_s is called sound speed.

Introducing expressions (B.2) in the equations (B.1) and using the results obtained in the section 2.2 we obtain the equations for first order perturbations:

$$\begin{aligned}\frac{\partial\delta\rho}{\partial t}+\rho_0\nabla_r\delta\vec{v}+\nabla_r\left(\delta\rho\,\vec{v}_0\right) &= 0, \\ \frac{\partial\delta\vec{v}}{\partial t}+\left(\delta\vec{v}\cdot\nabla_r\right)\vec{v}_0+\left(\vec{v}_0\cdot\nabla_r\right)\delta\vec{v} &= -\frac{c_s^2\nabla_r\delta\rho}{\rho_0}-\nabla_r\delta\Phi, \\ \nabla_r^2\delta\Phi &= 4\pi G\delta\rho.\end{aligned} \tag{B.4}$$

Until now, we have worked in the Eulerian coordinate system. In order to find the solution for the system of partial differential equations (B.4) we will change to the Lagrangian coordinate system (it is the analogous to comoving coordinates). The Eulerian and Lagrangian coordinate systems are related by the following transformation law:

$$t=t',\;\; \vec{r}=a(t')\vec{x},\;\; \text{then}\;\; \frac{\partial}{\partial t}=\frac{\partial}{\partial t'}-\vec{v}_0\cdot\nabla_r;\quad \nabla_r=\frac{1}{a}\nabla_x, \tag{B.5}$$

where $(t',\vec{x})$ are the Lagrangian coordinates and $(t,\vec{r})$ are the Eulerian coordinates. These expressions allow us to rewrite the equations (B.4) in the Lagrangian coordinate system. To simplify the notation we use the following convention $t'\rightarrow t$ and $\nabla_x\rightarrow\nabla$. The system of partial differential equations for perturbations in Lagrangian coordinate is given by:

$$\begin{aligned}\frac{\partial\delta}{\partial t}+\frac{1}{a}\nabla\delta\vec{v} &= 0, \\ \frac{\partial\delta\vec{v}}{\partial t}+H\delta\vec{v}+\frac{c_s^2}{a}\nabla\delta+\frac{1}{a}\nabla\delta\Phi &= 0, \\ \nabla^2\delta\Phi &= 4\pi Ga^2\rho_0\delta.\end{aligned} \tag{B.6}$$

Here δ is called contrast density and it is defined as $\delta\equiv\delta\rho/\rho_0$. We combine the equations (B.6) to obtain the equation which describes the growth of perturbations in the non-relativistic matter distribution for a universe in expansion

$$\frac{\partial^2\delta}{\partial t^2}+2H\frac{\partial\delta}{\partial t}-\left(4\pi G\rho_0\delta+\frac{c_s^2}{a^2}\nabla^2\delta\right)=0. \tag{B.7}$$

The evolution of large structure is due to gravitational instability caused by the small initial irregularities in the distribution of matter. Those regions with more matter exert a greater gravitational force to their neighboring regions causing an increase of matter in these regions. The fluctuations in the density field grow to produce the structures observed now. This simple picture does not explain all process involved in galaxy formation (actually, there are also complicated astrophysical processes involved here). However, this model allows us to understand how the linear perturbations evolve inside the horizon.

To solve equation (B.7) we use the Fourier transformation, so we can write

$$\ddot{\delta}_k + 2H\dot{\delta}_k + 4\pi G\rho_0 \left(\frac{c_s^2 k^2}{4\pi G a^2 \rho_0} - 1 \right) \delta_k = 0, \tag{B.8}$$

where δ_k is the Fourier mode of the contrast density. The critical wavenumber k_J is called the Jeans wavenumber and it is defined as

$$k_J^2 \equiv \frac{4\pi G\rho_0 a^2}{c_s^2}, \quad \text{then} \quad \lambda_J = \frac{2\pi}{k_J}, \tag{B.9}$$

so equation (B.8) is rewritten as

$$\ddot{\delta}_k + 2H\dot{\delta}_k + 4\pi G\rho_0 \left(\frac{\lambda_J^2}{\lambda^2} - 1 \right) \delta_k = 0. \tag{B.10}$$

For a static universe ($H = 0$, so a is a constant), we have

$$\ddot{\delta}_k + 4\pi G\rho_0 \left(\frac{\lambda_J^2}{\lambda^2} - 1 \right) \delta_k = 0. \tag{B.11}$$

Here we have a harmonic oscillator equation, then the solution is given by

$$\delta_k \propto \mathrm{e}^{\pm i\omega t} \quad \text{where} \quad \omega = 2\sqrt{\pi G\rho_0} \left[\left(\frac{\lambda_J}{\lambda} \right)^2 - 1 \right]^{1/2}. \tag{B.12}$$

In the limit $\lambda_J \gg \lambda$ the solution is periodic. The gravity is negligible in comparison with the pressure. The term $(c_s k)^2$ dominates over $4\pi G\rho_0$, where the first term is related with the pressure perturbations and the second term is related with the gravitational potential. For the case $\lambda_J < \lambda$, we have unstable modes in which the pressure cannot hold the collapse or expansion of the perturbations. These are known as decay modes or growth modes. In contrast to the static case, for the expansion case we have the damping term $2H\dot{\delta}_k$ in the differential equation. According to the structures observed in the Universe, the perturbations in the matter distribution have modes that grow. Therefore we require that $\lambda \gg \lambda_J$ in which the scales are much larger than the Jeans' scale. Here gravity dominates and the $\vec{k}$

dependence in each Fourier mode can be neglected then all modes grow equally. The perturbations can be rewritten as $\delta_k(a) = D(a)\delta_k$, where $D(a)$ is called growth factor. Expression (B.8) allows us to obtain a differential equation for $D(a)$

$$\ddot{D} + 2H\dot{D} = 4\pi G\rho_0 D. \tag{B.13}$$

By using the relation $\mathrm{d}a = aH(a)\mathrm{d}t$ we rewrite equation (B.13) as

$$\frac{\mathrm{d}^2 D}{\mathrm{d}a^2} + \left(\frac{\mathrm{d}\ln H}{\mathrm{d}a} + \frac{3}{a}\right)\frac{\mathrm{d}D}{\mathrm{d}a} = \frac{3\Omega_m}{2a^5}\left(\frac{H_0}{H}\right)^2 D. \tag{B.14}$$

The above equation was computed in the chapter 2 for matter perturbations in the sub-horizon limit. Here we show that the Newtonian theory is sufficient to describe the matter perturbation at late time for scales which are well within the horizon.

References

Abdalla F. B., Amara A., Capak P., Cypriano E. S., Lahav O., Rhodes J., 2008, MNRAS, 387, 969

Abdalla F. B. et al., 2011, MNRAS, 417, 1891

Abell G. O., 1958, ApJS, 3, 211

Abell G. O. et al., 1989, ApJS, 70, 1

Aguena M. and Lima M., 2016, arXiv:1611.05468v1 [astro-ph.CO]

Allen S. W., Evrard A. E., Mantz A. B., 2011, ARA&A, 49, 409-70

Almosallam I. A. et al., 2016, MNRAS, 455, 2387

Almosallam I. A. et al., 2016, MNRAS, 462, 726

Amendola L. et al., 2007, Phys. Rev. D, 75, 083504

Anderson L et al., 2014, MNRAS, 441, 24

Arnouts S. et al., 1999, MNRAS, 310, 540

Baldry I. K. et al., 2010, MNRAS, 404, 86

Ball N. M. et al., 2008, ApJ, 683, 12

Bardeen J., 1980, Phys. Rev. D, 22, 1882

Bardeen J. M., Bond, J.R., Kaiser, N. and Szalay, A. S., 1986, ApJ, 304, 15

Bartelmann M., 2010, Classical Quantum Gravity, 27, 233001

Battye R. A. & Weller J., 2003, Phys. Rev. D, 68, 083506

Benitez N., 2000, ApJ, 536, 571

Berlind A. A. et al., 2006, ApJS, 167, 1

Bharadwaj S., Bhavsar S. P. and Sheth J. V., 2004, ApJ, 606, 25-31

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

Blumenthal G. R. et al., 1984, Nature, 311, 517-525

Bolzonella M. et al., 2000, A&A, 363, 476

Bonfield D. G. et al., 2010, MNRAS, 405, 987

Borgani S. & Guzzo L., 2001, Nature, 409, 39-45

Borgani S. et al., 2001, ApJ, 561, 13-21

Botzler C. S. et al., 2004, MNRAS, 349, 425

Brammer G. B. et al., 2008, ApJ, 686, 1503

Bryan G. L. & Norman M. L., 1998, ApJ, 495, 80-99

Candela J. Q. & Rasmussen C. E., 2005, J. Mach. Learn. Res., 6, 1939

Carrasco Kind M. & Brunner R. J., 2013, MNRAS, 432, 1483

Carroll S. M., 2004, *Spacetime and Geometry An Introdution to General Relativity.* Addison Wesley. ISBN 0-8053-8732-3

Chevallier M. & Polarski D., 2001, Int. J. Mod. Phys. D, 10, 213

Christodoulou L., et al., 2012, MNRAS, 425, 1527

Cole S. et al., 2000, MNRAS, 319, 168

Collister A. A. & Lahav O., 2004, PASP, 116, 345

Copeland E. J. et al., 2006, Int. J. Mod. Phys. D, 15, 1753

Courtin J. et al., 2011, MNRAS, 410, 1911-31

Coutant M., 2012, arXiv:1405.3466 [hep-th]

Crocce M. et al., 2010., MNRAS, 403, 1353-67

Cui W. et al., 2012, MNRAS, 423(3), 2279-87

Dodelson S., 2003, *Modern Cosmology.* Amsterdam (Netherlands): Academic press. ISBN 0-12-219141-2

Driver S. P. et al., 2009, Astronomy and Geophysics, 50, 5.12

Driver S. P. et al., 2011, MNRAS, 413, 971

Efstathiou G. et al., 1988, MNRAS, 235, 715

Eisenstein D. J. et al., 2005, ApJ, 633, 560

Erickson B. M. S. et al., 2011, Phys. Rev. D, 84, 103506

Evrard, A. E. et al., 2002, ApJ, 573, 7

Farrens S. et al., 2011, MNRAS, 417, 1402

Feldmann R. et al., 2006, MNRAS, 372, 565

Firth A. E. et al., 2003, MNRAS, 339, 1195

Foreman-Mackey D. et al., 2013, PASP, 125, 925 (306pp)

Freedman W. L. et al., 2001, ApJ, 553, 47

Freeman K. C., 1970, ApJ, 160, 881

Gal R. R. et al., 2009, ApJ, 137, 2981

Gibbons G. W. & Hawking, S. W., 1977, Phys. Rev. D, 15, 2752

Gladders M. D. et al., 2007, ApJ, 655, 128

Goodman J. & Weare J., 2010, Comm. App. Math. Comp. Sci., 5, 65

Hao J. et al., 2010, ApJS, 191, 254

Hawking S. W. & Horowitz, G. T., 1996, Class. Quant. Grav., 13, 1487-1498

Hildebrandt H. et al., 2010, A&A, 523, A31

Hoecker A. et al., 2007, PoS, ACAT, 040

Hu, W. & Dodelson, S. 2002, ARA&A, 40, 171

Hubble E., 1929, N. A. S., 15, 168-173

Huchra J. P. & Geller M. J., 1982, ApJ, 257, 423

Hung-Yu J. et al., 2014, ApJ, 788, 109

Ilbert O. et al. 2006, A&A, 457, 841

Jenkins, A., et al., 2001, MNRAS, 321, 372

Jouvel, S. et al. 2009, A&A, 504, 359

Kilbinger M. et al., 2013, MNRAS, 430, 2200

Kim R. S. J. et al., 2002, AJ, 123, 20

Kravtsov A. V. et al., 1997, ApJS, 111, 73

Kravtsov A. V. & Borgani S., 2012, ARA&A, 50, 353-409.

Kuhlen M. et al., 2005, MNRAS, 357, 387

Lagos C. d. P. et al., 2012, MNRAS, 426, 2142-2165

Lewis A. et al., 2000, ApJ, 538, 473

Lewis A. & Bridle S., 2002, Phys. Rev. D, 66, 103511

Lima M. & Hu W., 2004, Phys. Rev. D, 70, 043504

Lima M. & Hu W., 2005, Phys. Rev. D, 72, 043006

Lima M. & Hu W., 2007, Phys. Rev. D, 76, 123013

Lima M. et al., 2008, MNRAS, 390, 118

Linder E. V., 2003, Phys. Rev. Lett., 90, 091310

Linder E. V. & Jenkins A., 2003, MNRAS, 346, 573

Liske J. et al., 2015, MNRAS, 452, 2087

Lopes P. A. A. et al., 2004, AJ, 128, 1017

MacKay D., 2003, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press

Mana A. et al., 2013, MNRAS, 434, 684-695

Mantz A. et al., 2008, MNRAS, 387, 1179-1192

Mantz A. et al., 2010, MNRAS, 406, 1759

Massey R., Kitching T., Richard J., 2010, Rep. Progress Phys., 73, 086901

Merson A. I. et al., 2013, MNRAS, 342

Mukhanov V., 2005, *Physical Fundations of Cosmology.* Cambridge University Press

Owen A. B., 2017, Journal of Computational and Graphical Statistics, 26(3), 738-744.

Oyaizu, H., et al., 2008, ApJ, 689, 709

Percival W. J., Cole S., Eisenstein D. J., Nichol R. C., Peacock J. A., Pope A. C., Szalay A. S., 2007, MNRAS, 381, 1053

Perlmutter S. et al., 1999, ApJ, 517, 565

Planck Collaboration, 2016, A&A, 594, A20

Planck Collaboration, 2016, A&A, 594, A13

Press, W. H. and Schechter, P., 1974, ApJ, 187, 425-438

Press W. H. et al., 2007, Numerical Recipes The Art of Scientific Computing., Cambrigde-USA, 3rd ed. ISBN 0521884071

Raichoor A. et al., 2016, A&A, 585, A50

Ramella M. et al., 2001, A&A, 368, 776

Riess A. G. et al., 1998, AJ, 116, 1009-1038

Rozo E. et al., 2010, ApJ, 708, 645-660

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

Rycroft C. H., 2009, Chaos, 19, 041111

Rykoff E. S. et al., 2012, ApJ, 746, 178

Rykoff E. S. et al., 2014, ApJ, 785, 104 (33pp)

Rykoff E. S. et al., 2016, ApJS, 224, 1 (19pp)

Sadeh I. et al., 2016, PASP, 128, 104502 (18pp)

Sánchez C. et al., 2014, MNRAS, 445, 1482

Sheth R. K. & Tormen G., 1999, MNRAS, 308, 119

Sheth R. K. et al., 2001, MNRAS, 323, 1

Simet M. et al., 2017, MNRAS, 466, 3103

Sivia D. S. & Skilling J., 2006, *Data Analysis A Bayesian Tutorial.* New York (United State): Oxford university press. ISBN 0-19-856831-2

Soares-Santos M. et al., 2011, ApJ, 727, 45

Spergel D. N. et al., 2003, ApJS, 148, 175-194

Springel V. et al., 2005, Nature, 435, 629

Springel V., 2005, MNRAS, 364, 1105

Stanek R. et al., 2009, MNRAS, 394, L11-15

Tinker J. et al., 2008, ApJ, 688, 709

Tsiligkaridis T., Hero A., 2013, IEEE Trans. Signal Process., 61, 5347

Vanzella E. et al., 2004, A&A, 423, 761

Vikhlinin A. et al., 2009, ApJ, 692, 1060-1074

Voit G. M., 2005, Rev. Mod. Phys., 77, 207

Voronoi G., 1907, Journal für die Reine und Angewandte Mathematik , 133, 97

Warren M. S. & Salmon J. K., 1993, in Supercomputing 93, IEEE Comp. Soc.

Way M. J. et al., 2009, ApJ, 706, 623

Wittman D., 2009, ApJ, 700, L174

Zhang Y. et al., 2005, Time Series Gaussian Process Regression Based on Toeplitz Computation of O(N2) Operations and O(N)-level Storage. IEEE, Seville, Spain, p. 3711

Zou H. et al., 2015, PASP, 127, 94-101

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