VNIVERSITAT DE VALÈNCIA

Departament d'Astronomia i Astrofísica

Galaxy clustering: a point process

Autor:
Lluís Hurtado Gil

Supervisors:
Vicent Josep Martínez García
Pablo Arnalte Mur

Tesi Doctoral

Observatori Astronòmic
VNIVERSITAT DE VALÈNCIA

València, Maig 2016

Imatge de portada: fotografia processada per Vicent Peris a partir de dades d'ALHAMBRA.

DR. VICENT JOSEP MARTÍNEZ GARCÍA
Catedràtic d'Astronomia i Astrofisica a la Universitat de València

I

DR. PABLO ARNALTE MUR
Investigador Doctor d'Astronomia i Astrofisica a la Universitat de València

Que la present memòria, "Galaxy clustering: a point process", ha estat realitzada sota llur direcció per Lluís Hurtado Gil, i que constitueix la seua tesi doctoral per optar al grau de Doctor en Física.

I per que hi quede constancia i tinga els efectes oportuns, signem el present document en Paterna a 31 de Març de 2016.

Signat: Vicent J. Martínez García Pablo Arnalte Mur

A la família d'on vinc.
A la família on vaig. Clara.

Contents

Contents vi

List of Figures x

List of Tables xiii

1 Introduction 1
- 1.1 Cosmology . . . 2
  - 1.1.1 The standard cosmological model . . . 3
  - 1.1.2 The geometry of the Universe . . . 9
  - 1.1.3 Modern observational cosmology . . . 16
- 1.2 Spatial point processes in Astrophysics . . . 20
  - 1.2.1 Probability background . . . 23
  - 1.2.2 Definition of a Point Process . . . 25
  - 1.2.3 Marked processes . . . 27
  - 1.2.4 Examples of point processes . . . 29
  - 1.2.5 Moment measures . . . 31
- 1.3 Aim of this thesis . . . 34

I Summary statistics for the Galaxy Distribution 37

2 Counts-in-Cells Distribution 39
- 2.1 Introduction . . . 40
- 2.2 Estimation of the CiC distribution . . . 41
- 2.3 Data catalogs . . . 44
  - 2.3.1 The SDSS - New York University - Value Added Galaxy Catalog . . . 44
  - 2.3.2 LasDamas simulations . . . 45
- 2.4 Estimation of the errors: the jackknife method . . . 47
- 2.5 Best fit to $f_V(N)$ . . . 51
  - 2.5.1 Gravitational Quasi-Equilibrium Distribution . . . 51
  - 2.5.2 Negative Binomial Distribution . . . 53
  - 2.5.3 Log Normal Distribution with bias . . . 54
  - 2.5.4 Weibull Distribution . . . 57
- 2.6 Results . . . 59
  - 2.6.1 Fitting the results to a distribution function . . . 60
- 2.7 Analysis of the results . . . 66

3 Correlation functions **68**
3.1 Introduction . . . 68
3.2 Definition . . . 69
3.3 Estimation of $\xi(r)$ . . . 71
3.3.1 Correction of selection effects . . . 74
3.3.2 Correction of redshift distortions . . . 75
3.3.3 Estimation of the errors . . . 80
3.4 Evolution of galaxy spectral segregation in the ALHAMBRA Survey . . . 80
3.4.1 Data samples . . . 83
3.4.2 The correlation function at the smallest scales . . . 91
3.4.3 Results . . . 95
3.4.3.1 Power-law modeling . . . 97
3.4.3.2 Full samples . . . 99
3.4.3.3 Segregated samples . . . 102
3.4.3.4 Dependence of the bias on spectral type and redshift . . . 108
3.4.4 Conclusions . . . 111
3.5 Future work: new correlation function estimation for photometric surveys . . . 113

II Modeling the galaxy distribution **117**

4 Finite Gibbs processes **119**
4.1 Introduction . . . 119
4.2 Definition . . . 121
4.2.1 Probability density function of a Gibbs process . . . 122
4.2.2 Examples of Gibbs models . . . 124
4.2.3 The Papangelou conditional intensity . . . 131
4.2.4 The pseudolikelihood . . . 134
4.3 Point processes residuals analysis . . . 138
4.3.1 Local residuals . . . 139
4.4 Data catalogs . . . 144
4.4.1 Sloan Digital Sky Survey - DR8 . . . 144
4.4.2 LasDamas Simulation catalog . . . 145
4.4.3 Testing with Toy Models . . . 146
4.4.3.1 Generation of samples . . . 147
4.4.3.2 Fitting the sample . . . 148
4.4.4 SDSS populations . . . 151
4.4.5 LasDamas populations . . . 154
4.4.6 Conclusions and future work . . . 158

III Mining the galaxy distribution **165**

5 Mixture models **167**
5.1 Introduction . . . 167

5.2 Definition . . . 170
5.2.1 The surface density model . . . 171
5.2.2 Density profiles . . . 173
5.2.3 Fitting the parameters . . . 174
5.3 MultiDark simulated samples . . . 178
5.4 Mixture models for galaxy clusters . . . 180
5.4.1 Dark matter profiles . . . 180
5.4.2 Toy models . . . 184
5.4.3 MultiDark simulation . . . 201
5.5 Conclusions . . . 212

6 Conclusions **214**

IV Appendices **219**

A Agregació de galaxies: un procés puntual **221**
A.1 Introducció . . . 221
A.1.1 El model cosmlògic estàndard . . . 222
A.1.2 La distribució de galàxies com a procés puntual . . . 223
A.1.3 Objectius d'aquesta tesi . . . 224
A.2 Ajust de recomptes de galàxies per cel·les . . . 226
A.2.1 Funcions de distribució . . . 227
A.2.2 Tractaments del errors . . . 228
A.2.3 Resultats i conclusions . . . 229
A.3 Correlació de galàxies a escales curtes i segregació espectral . . . 229
A.3.1 El catàleg ALHAMBRA . . . 230
A.3.2 La funció de correlació projectada . . . 231
A.3.3 Resultats . . . 231
A.4 Modelització amb processos de Gibbs . . . 234
A.4.1 Models d'interacció . . . 234
A.4.2 Dades i ajust . . . 237
A.5 Models de mescla . . . 238
A.5.1 El model de densitat de superfície . . . 239
A.5.2 Aplicació i resultats . . . 239
A.5.3 Conclusions . . . 241
A.6 Conclusions . . . 241

Bibliography **242**

Acknowledgements **258**

List of Figures

1.1 The CfA2 and Las Campanas Redhisft survey 5
1.2 Abundance of light elements: theory and observations 7
1.3 Spectrum of the CMB radiation 8
1.4 Photometric redshift accuracy 20

2.1 Footprint of accepted cells for CiC 49
2.2 Jackknife error bars vs population variance in LasDamas 50
2.3 $f_V(N)$ distribution for the NYU-VAGC 60
2.4 Counts-in-cells fittings for population 1 and radius $24h^{-1}$ Mpc . 63
2.5 Counts-in-cells fittings for population 1 and radius $12h^{-1}$ Mpc . 64
2.6 Counts-in-cells fittings for population 1 and radius $6h^{-1}$ Mpc . . 64
2.7 Counts-in-cells fittings for population 2 and radius $24h^{-1}$ Mpc . 65
2.8 Counts-in-cells fittings for population 2 and radius $12h^{-1}$ Mpc . 65
2.9 Counts-in-cells fittings for population 2 and radius $6h^{-1}$ Mpc . . 66

3.1 2-dimensional correlation function distances decomposition . . . 77
3.2 2-dimensional correlation function 79
3.3 ALHAMBRA survey observed slice 83
3.4 Projected correlation functions for luminosity segregated galaxy samples 85
3.5 ALHAMBRA galaxy samples in template, redshift and luminosity 86
3.6 ALHAMBRA color-magnitude diagram 89
3.7 ALHAMBRA fields 2 and 4 with template segregation 90
3.8 Quartet of galaxies in ALHAMBRA survey 93
3.9 Near Neighbor distribution 94
3.10 HOD and projected correlation function 95
3.11 Projected correlation function for the full ALHAMBRA sample . 96
3.12 Power laws best fit parameters (full population) 100
3.13 Projected correlation function for the segregated ALHAMBRA sample 103
3.14 Power laws best fit parameters (segregated population, small scales) 104
3.15 Power laws best fit parameters (segregated populations, large scales) 106
3.16 Galaxy bias b for ALHAMBRA segregated populations 109

4.1 Strauss process: 2-dimensional sample 125
4.2 Geyer process: 2-dimensional sample 127
4.3 Fiksel process: 2-dimensional sample 129
4.4 Area Interaction process: 2- dimensional sample 131

4.5 SDSS-DR8 samples for Gibbs models testing 146
4.6 LasDamas samples for Gibbs models testing 147
4.7 Data and model of a Geyer process 149
4.8 Absolute and relative residuals of a Geyer process 150
4.9 Model, absolute and relative residuals of SDSS samples with a Geyer model 155
4.10 Model, absolute and relative residuals of SDSS samples with a Fiksel model 156
4.11 Model, absolute and relative residuals of SDSS samples with a Power Law model 157
4.12 Model, absolute and relative residuals of LasDamas samples with a Geyer model 159
4.13 Model, absolute and relative residuals of LasDamas samples with a Fiksel model 160
4.14 Model, absolute and relative residuals of LasDamas samples with a Power Law model 161

5.1 3D generated toy models of dark matter halos 185
5.2 MCMC parameters distribution of one cluster (Einasto's toy model) 186
5.3 Smoothed Mixture model (Einasto's toy model) 188
5.4 Absolute and relative residuals (Einasto's toy model) 189
5.5 Lurking plots (Einasto's toy model) 190
5.6 Smoothed Mixture model (Einasto's 3 clusters toy model) 192
5.7 Absolute and relative residuals (Einasto's 3 clusters toy model) . 193
5.8 3D generated toy models of dark matter halos with filamentary structure 195
5.9 Smoothed Mixture model (Einasto's toy model with filamentary structure) 196
5.10 Absolute and relative residuals (Einasto's toy model with filamentary structure) 197
5.11 Smoothed Mixture model (Einasto's 5 clusters toy model with filamentary structure) 198
5.12 Absolute and relative residuals (Einasto's 5 cluster toy model with filamentary structure) 199
5.13 Relative residuals (Einasto's overlapped toy model) 200
5.14 3D sample from MultiDark simulation 201
5.15 MCMC distribution of parameters from component 2 204
5.16 Data and model of the MultiDark sample 206
5.17 Absolute and relative residuals of a MultiDark sample 207
5.18 Cluster distribution and generation of a MultiDark sample (6 components) 209
5.19 Profile curves of components in a MultiDark sample 210
5.20 Cluster distribution and generation of the MultiDark sample (10 components) 211

List of Tables

2.1 SDSS selected samples 45
2.2 LasDamas selected mock catalogs 47
2.3 Counts-in-Cells best fit $f_V(N)$ 61
2.4 Expectancies 62

3.1 Characteristics of the galaxy samples used 88
3.2 Photometric Surveys 93
3.3 Results of the different fits to $w(r_p)$: power law and bias models 101

4.1 Fitted parameters for Gibbs models - SDSS-DR8 151
4.2 Fitted parameters for Gibbs models - LasDamas 158

5.1 Einasto toy model fitting 187
5.2 Density field local maximum points 202
5.3 MultiDark fitting for 6 components 205
5.4 MultiDark fitting for 10 components 212

Chapter 1

Introduction

'I know the desire of your minds that what ye have seen should verily be, not only in your thought, but even as ye yourselves are, and yet other. Therefore I say: Eä! Let these things Be! And I will send forth into the Void the Flame Imperishable, and it shall be at the heart of the World, and the World shall Be; and those of you that will may go down into it'.

Eru Ilúvatar

This PhD dissertation proposal is an application of point process statistics to the field of galaxy clustering in modern cosmology. Throughout this thesis work we will introduce and explain different methodologies classifiable into three main categories. These categories compound the three possible approximations to the analysis of point process: summary statistics (Chapters 2 and 3), modeling (Chapter 4) and data mining (Chapter 5). Main conclusions of this work can be found in Chapter 6. For an integral summary in catalan please visit the Appendix.

The multiple algorithms used include conventional statistics as well as new methods applied in the cosmological scenario for the first time, and even original contributions. Similarly, studied datasets include both known public access catalogs and recently published galaxy surveys.

Before entering the contributions of this thesis we will introduce in this chapter the necessary background of modern cosmology and spatial point processes.

1.1 Cosmology

Humans have always felt the necessity of explaining the origin, evolution and laws governing the known universe. Almost all cultures have engaged in this task with the elaboration of different cosmographies, usually as part of mythical or religious beliefs. In the construction of scientific knowledge, different philosophers of classic cultures proposed their visions of the universe content supported with rational arguments. Most notably, the greek Anaximandre (c.610 - c.546 BC), was the first to propose a demythified theory of celestial bodies mechanics, for which he is considered the father of cosmology and astronomy (Kahn, 1994). Many followed his work, including Aristotle, Aristarchus of Samos, Ptolemy, Hypatia or Hipparchus in the ancient greek culture (centuries VII BD to V AD). In ancient China numerous philosophers studied the cosmos, of which we could cite Gan De, Shi Shen or Su Song (centuries IV BD to XI AD). Most of them accompanied their models with observations, such as catalogs of stars or measurements of the movement of celestials objects. These efforts were continued in middle ages with the Arab and Persian contributions of authors like Muḥammad ibn Musa al-Khwarizmi (c.780-c.850) or Abu Yahya Zakariya' ibn Muhammad al-Qazwini (1203-1283).

In Europe, astronomy played a central role in the development of sciences after the XV century. Different authors largely contributed to a better understanding of our universe proposing visions of the cosmos supported with new observations and solid mathematical arguments. Authors like Nicolaus Copernicus (1473-1543), Tycho Brahe (1546-1601), Galileo Galilei (1564-1642) and Johannes Kepler (1571-1630) settled the basis of modern astronomy which are still fully relevant for today's science. But it was with Isaac Newton (1642-1726) that science went beyond compiling and describing observations to attempt to explain and discover the rules governing the skies. The path open by Newton lead scientists to a systematic study of the known world for two centuries. However, with the available technology and understanding of the physics, only the nearby universe was achievable to our observations and understanding. But that changed in the late XIX and first years of the XX century. The study of quantum mechanics, started with the work of Ludwig Boltzmann (1844-1906) and Max Planck (1858-1947), and relativity and cosmology, by Albert Einstein (1879-1955) and Henri Poincaré (1854-1912) among others, brought the necessary tools to observe, analyze and understand the contents of the cosmos at a greater scale (Harrison, 2000, Nussbaumer et al.,

2009) In addition, these advances led cosmologists to ask some of the deepest questions about the nature of our universe (Longair, 2006). It is in this moment that cosmology was born, the science that studies the origin and evolution of all the matter and energy, the universe.

1.1.1 The standard cosmological model

Today, the mainstream understanding of the cosmos is contained in the so called standard model, known as Λ - Cold Dark Matter (ΛCDM) model (Dodelson, 2003, Fukugita & Peebles, 2004, Lahav & Liddle, 2014). This thesis work is done under the assumption of this framework.

The construction of the standard model was possible thanks to the advancements made during the beginning of the XX century. On the theoretical side, Einstein's theory of General Relativity (Einstein, 1915) provided us with the necessary understanding of gravity and spacetime in the universe. This settled the base to develop the first cosmological models, solutions to Einstein's field equations. But these solutions ought to be consistent with the large-scale matter and energy distribution in the universe. The first models came assuming the easiest configuration of the matter in the universe, which is summarized in the *Cosmological Principle*: the mass distribution of the universe is homogeneous and isotropic, i.e, viewed on a sufficiently large scale, the properties of the universe are the same for all observers. Homogeneity implies that the matter density is constant at large scale, and thus doesn't vary between distant regions of the universe, while isotropy implies that this distribution is the same for every direction given an observer. In a classic geometrical language we might say that the universe mass distribution is invariant by translation and rotation.

These first insights date back to the same early decades, when Edwin Hubble measured the distance between the Milky Way and our neighbor galaxy Andromeda. This calculation was possible due to the contribution of Henrietta Swan Leavitt (Leavitt, 1908) with the period-luminosity relation for Cepheid stars. The result of a distance, far larger than the estimated size of the Milky Way, clearly proved the extragalactic character of Andromeda (Hernquist, 1990, Hubble, 1925). This put an end to the 'Great Debate' between Heber Curtis and Harlow Shapley (1920) (Trimble, 1995), and started the study of a universe compound by many galaxies

in a vastly larger space. The second major contribution of Edwin Hubble (Hubble, 1929) was the establishment of a relation between the radial velocity v of galaxies and their distances d. The relation proposed was the linear relation

$$cz = Hd \tag{1.1}$$

where $H = H(t)$ is the Hubble parameter, (in this work, for the present time value of $H(t)$ we use $H_0 = 100h$ km s^{-1} Mpc^{-1}, where h is a dimensionless constant depending on the estimated value of H_0). This relation, known as Hubble's law, implies a universe in expansion. Together with the Cosmological Principle and Einstein's general relativity, which describes how matter is driven by gravity, these are the base elements of the *Hot Big Bang* model.

The last decades of the XX century brought the first observations of significant parts of the universe, including distances far from neighbor galaxies. The Slice of the Universe survey (CfA2, De Lapparent et al. (1986)) showed clear structures at scales thought to be already homogeneous, and it was not until wider surveys (Las Campanas Redshift survey, see Fig. 1.1, Shandarin & Yess (1998)) that homogeneity was confirmed, signifying the 'end of greatness', where no structures appear to stand out above the others. Isotropy was also certified with later experiments, like the CMB measurements (Mather et al., 1994).

The standard model describes a universe started as a very hot and dense plasma formed by elementary particles. This is an homogeneous and isotropic universe, with no privileged points or directions, where its content evolves under the dynamics of gravity. In its original state after the initial singularity, density and energy in the universe were extraordinarily high.

The state of the universe is then drift by the thermodynamics of this expanding volume, decreasing temperature and density with time. This quenching determines the evolution of the contents of the universe and we can divide its timeline in several epochs. The first relevant epoch is the so called *Inflation*, a short epoch of around 10^{-34} seconds when the universe suffered a sudden and strong expansion. Despite the existence of this epoch has not been fully confirmed it is accepted and included in the cosmological standard model due to the reasonable explanation it gives to three underlying problems of the model. Two of these problems, the

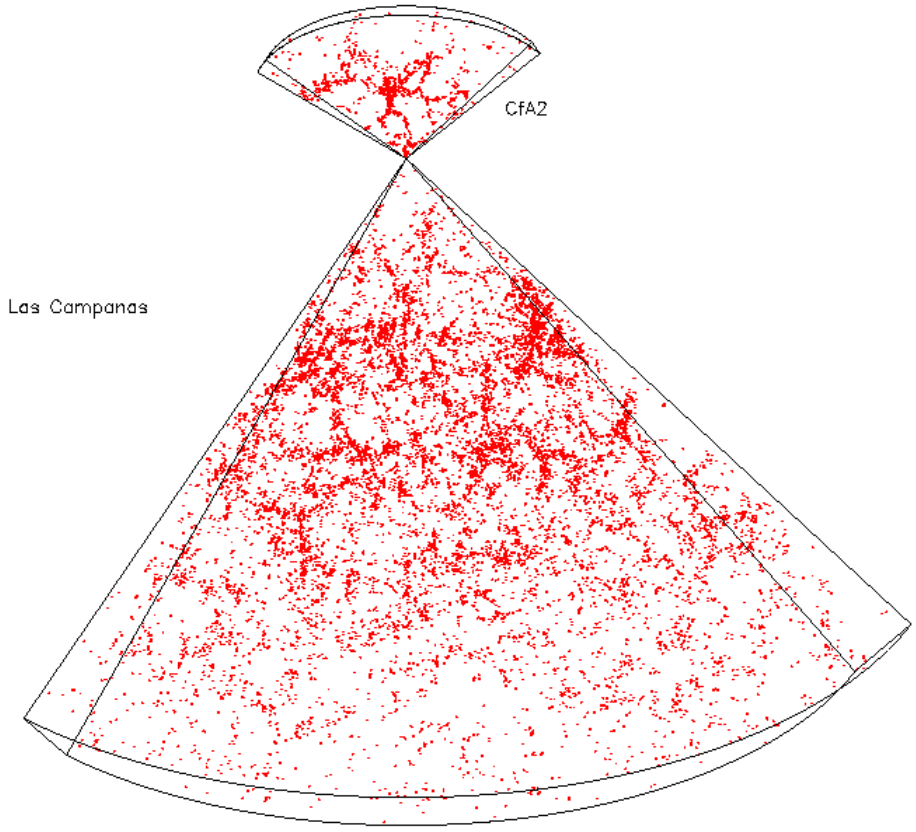

FIGURE 1.1: Comparison between the observed galaxies by the CfA2 and the Las Campanas Redshift survey. Image credit: Vicent Martínez.

flatness problem and the *horizon problem*, state that the flatness, homogeneity and isotropy found today in the universe were even more prefect in the past, a highly unlikely event. The third problem is the *monopole problem*, stating the absence of magnetic monopoles, a kind of particles predicted to be abundant otherwise.

Just after the Inflation and the creation of elementary particles, the following relevant epoch is the photon epoch. Pressure and temperature conditions were appropriate for the *primordial nucleosynthesis* to take part. In a lapse of three minutes, the first nuclei of light elements (hydrogen, helium and lithium) formed. This epoch lasted until radiation and matter density reach an equilibrium, moment in which we say the matter dominant epoch started. During this new epoch temperature keep dropping, and photons lost energy until they were not enough energetic to prevent nuclei from capturing electrons and create the first atoms.

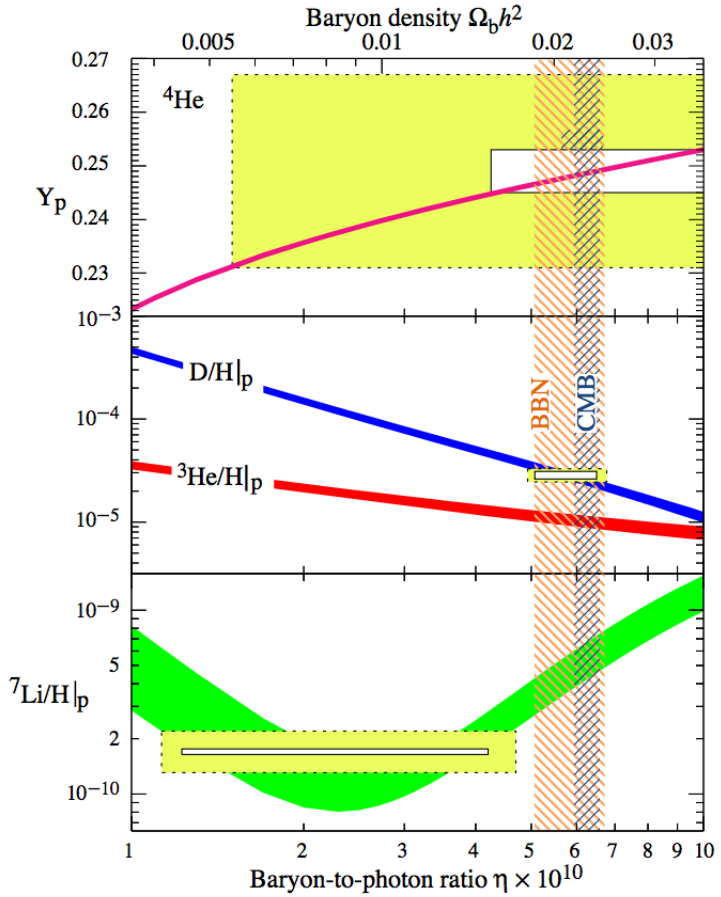

FIGURE 1.2: The abundances of ^4He, D, ^3He, and ^7Li as a function of the baryon-to-photon ratio η or, equivalently, baryon density Ω_b. The bands show the theoretical predictions with 95% confidence range, and the boxes the measured light element abundances. Data agrees with helium and deuterium predictions, but shows a tension with lithium. Figure from Nakamura et al. (2010)

Light scattered through the universe, creating the Cosmic Microwave Background (CMB) we observe today. That moment, known as decoupling, took place 380,000 years after the Big Bang when temperature in the universe was $\sim$ 3000K.

Dark matter and baryonic matter were now free to interact and create more complex structures drift by gravity. Galaxy structure was born and the universe shape was conformed in a wide range of scales, from the internal distribution of galaxies to the huge galaxy superclusters. This era lasted for $9.8 \cdot 10^9$ years.

The two main physical phenomena described in this timeline (nucleosynthesis and CMB) have been confirmed through observation and vitally contribute to confirm

the Hot Big Bang theory. Successive measurements of the light elements abundances have been made in objects where very little stellar nucleosynthesis has taken place (Peacock, 1999, Smith et al., 1993) and provide an overall satisfactory coincidence with the predictions (see Fig. 1.2). Almost the entire baryonic mass of the universe contributed to the formation of hydrogen nuclei (75 %) and ^4He helium (25 %). Residual quantities of deuterium (^2D), ^3He helium and lithium (^7Li) were also produced and detected. However, a discrepancy was found for this later case, too scarce (Burbidge et al., 1957, Clayton, 1968, Olive et al., 2000). Alternative theories have been proposed to explain this deviation including possible effects in the measurements over metal-poor population II stars (Korn et al., 2006, Meléndez & Ramírez, 2004) or the necessity of modification in the current standard model of particle physics (Dmitriev et al., 2004, Jedamzik, 2004).

The detection of the CMB by Penzias & Wilson (1965) has been followed by several new and more precise measurements performed with several spacial satellites: COBE (Mather et al., 1994), WMAP (Komatsu et al., 2011) and Planck (Planck Collaboration et al., 2015) experiments. This background radiation emitted after the matter-photon decoupling has been found to closely follow a black body spectrum with a temperature of $T = 2.73K$ (see Fig. 1.3). This radiation follows the expected properties, with a strong isotropy only broken by the small anisotropies produced by small density fluctuations in the last scattering surface from where photons were emitted. Other posterior effects can produce as well smaller anisotropies, but we will not consider them in this work. The CMB has resulted very useful to estimate the parameters of the cosmological model (Komatsu et al., 2011, Planck Collaboration et al., 2014).

The last main epoch of the universe timeline is the dark energy dominant epoch, where the universe started an accelerated expanding epoch lasting now for $4 \cdot 10^9$ years. Perlmutter et al. (1999) and Riess et al. (1998) found evidence of this acceleration in the expansion of the Universe, which is still of unclear origin. Different hypothesis have been presented to explain, or at least, model it. The *Cosmological constant* Λ is a parameter used to describe this dark energy responsible for the accelerated expansion. It is also referred as the vacuum energy density and is a fundamental element of the current cosmological standard model ΛCDM. The measurement of this expansion has allowed to obtain a more accurate estimation

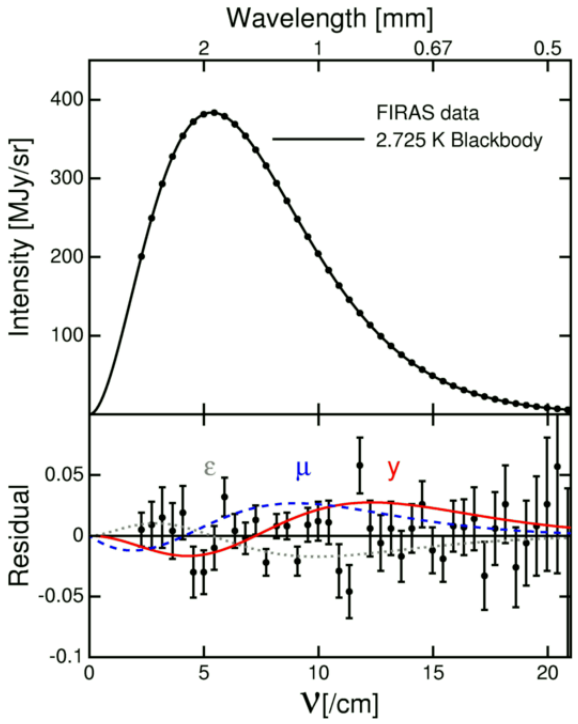

FIGURE 1.3: Spectrum of the CMB radiation observed by the FIRAS instrument on board of COBE (Mather et al., 1994), showing a remarkable agreement with a black body spectrum. The points with errors are the measurements of FIRAS, and the black solid line is the best-fit black body spectrum. The other lines in the bottom plot show the spectra for possible deviations from the black body case: that of a body with a reflectivity different from zero (dotted grey), and the effect of hot electrons adding an excess energy to the CMB either at $z \gtrsim 10^5$ (dashed blue) or $z \lesssim 10^5$ (solid red). In all cases, the curves shown are the maximum allowed deviations (at 95% confidence level) by FIRAS data. Figure by Ned Wright, http://www.astro.ucla.edu/∼wright/cosmolog.htm.

of the Hubble constant, which the Planck project measured in 67.3 ± 1.2 km s^{-1} Mpc^{-1} (Planck Collaboration et al., 2014).

The last element necessary to understand our model of the universe is the *Cold Dark Matter*. Since the density of radiation in the universe dropped, ending radiation dominating epoch, it is negligible compared with the matter (dark matter and baryonic) and dark energy. Therefore, we only need to include the contribution of the matter and the cosmological constant to properly describe the overall properties of the universe. The evolution and formation of matter structures during the matter domination epoch mainly depend on the dark matter, which represents 85 % of the total amount. This matter might be *cold* or *hot* depending on his energy levels after the decoupling, when the structures creation began. If dark matter

were cold, and therefore, these particles were nonrelativistic, they would form the kind of structures that have been observed in agreement with the modern calculations of the large scale structure estimators, like the power spectrum and the correlation function. For this reason the hot, relativistic, particles are considered to play a minor role in the evolution of the large scale structure of the universe.

1.1.2 The geometry of the Universe

Einstein's General Relativity Theory and the Cosmological Principle allow us to build a metric that satisfies Einstein's field equations and whose parameters can be fitted by observations. The geometrical description of our universe completes the ΛCDM model, constituting a theory capable of predictions and falsifications. From General Relativity we derive a universe where distance has to be calculated in the 4-dimensional space-time, while from the homogeneous and isotropic properties of the universe, we derive the Friedmann-Lemaître-Robertson-Walker metric:

$$ds^2 = -c^2dt^2 + a^2(t)\left[\frac{dr^2}{1-kr^2} + r^2d\theta^2 + r^2\sin^2\theta d\phi^2\right] \tag{1.2}$$

where r, θ and ϕ are comoving spherical coordinates and t is the cosmic time, the four dimensions of our relativistic spacetime, and c is the speed of light. The geometrical properties derived from the cosmological principle and the expansion of the universe are included in the metric through the rest of the elements. The *cosmic scale factor* $a(t)$ is a dimensionless function that describes the expansion of the universe depending on time. In a perfectly homogeneous and isotropic universe the comoving coordinates will remain constant and Hubble's Law can be generalized to $cz = H(t)d$ with $H = \dot{a}/a$. For a flat universe, in present time we take $a_0 = 1$.

The curvature constant k is a dimensionless number that can take three different values: $k = 0$ for a spatially flat universe, $k = -1$ if the universe is open (hyperbolic) and $k = 1$ for a closed universe (spherical). The first two cases build an expanding universe of infinite volume. As an analogy, a closed universe can be imagined as a spherical surface, where one could ideally circumnavigate the universe and reach his antipodal position by constantly travel.

The trigonometric expression of the spherical coordinates θ and ϕ ensures the isotropy in the metric.

It is common to express this metric in an alternative shape, where the curvature is included in the function $S_k(r)$:

$$ds^2 = -c^2dt^2 + a^2(t)\left[dr^2 + S_k^2(r)d\Omega^2\right] \tag{1.3}$$

where $d\Omega^2 = d\theta^2 + \sin^2\theta d\phi^2$ and

$$S_k(r) = \begin{cases} R(t)\sin\left(r/R(t)\right), & k = +1 \\ r, & k = 0 \\ R(t)\sinh\left(r/R(t)\right), & k = -1 \end{cases}$$

Function $R(t)$ is the radius of curvature, with dimensions of length, and for $t = t_0$ gives the radius of curvature of the universe at the present moment.

With this metric we can calculate how distances expand with time. For example, the wavelength of a photon varies between its emission and observation due to this expansion. For λ_e and λ_0 being the wavelength of this photon at emission and observing times, we have:

$$\frac{\lambda_e}{a(t_e)} = \frac{\lambda_0}{a(t_0)} \tag{1.4}$$

The difference between wavelengths produce the variation in the observed spectrum

$$z \equiv \frac{\lambda_0 - \lambda_e}{\lambda_e} = \frac{1}{a} - 1 \tag{1.5}$$

As the universe has been expanding, $a(t < t_0) < a_0$ and $z > 0$, which shifts the spectrum towards redder colors. The quantity z is known as the *cosmological redshift*. This observable is a key quantity in modern cosmology allowing us to study the evolution of the universe and the distribution of its content.

The dynamics of the universe can be described with Einstein's equations, which establish the connections between the geometry and the physical content. Assuming the FLRW metric and a perfect fluid with density ρ and pressure p, we can derive a solution, the Friedmann-Lemaître equations:

$$\begin{aligned} H^2 &= \frac{8\pi G\rho}{3} - \frac{kc^2}{a^2} + \frac{\Lambda}{3} \\ \frac{\ddot{a}}{a} &= -\frac{4\pi G}{3}(\rho + \frac{3p}{c^2}) + \frac{\Lambda}{3} \\ \dot{\rho} &= -3(\rho + \frac{p}{c^2})\frac{\dot{a}}{a} \end{aligned} \tag{1.6}$$

Where G is Newton's gravitational constant. The cosmological constant takes the form of the dark energy if we understand it as another component of the universe with the identification

$$\rho_\Lambda = \frac{\Lambda c^2}{8\pi G} \tag{1.7}$$

These equations have to be completed with an equation of state relating pressure and energy density: $p = p(\rho)$. This equation expresses the energy content of the universe and determines the evolution of the model, how it expands or contracts. In background conditions, the equation is.

$$p_i = w_i \rho_i \tag{1.8}$$

where w_i is a constant dimensionless quantity depending on each substance. In the ΛCDM model these quantities have been deduced to be $w_m = 0$ for matter (baryonic o dark), $w_r = 1/3$ for radiation and $w_\Lambda = -1$ for dark energy. So far, we have reduced our description of the universe to the correct estimation of the different species densities (the ρ's) and its curvature (k). These two quantities are related through the *critical density*, ρ_c, defined as the total density in a flat universe ($k = 0$). Its value is

$$\rho_c = \frac{3H_0^2}{8\pi G} \tag{1.9}$$

Therefore, for higher or lower values of this density we will have a closed or open universe respectively. This quantity is estimated to be $\rho_c = 2.775 \times 10^{11} h^2$ $M_\odot$ Mpc^{-3}. For each substance we usually normalize their densities by the critical density and operate with the values

$$\Omega_m = \frac{\rho_{m,0}}{\rho_c}, \Omega_r = \frac{\rho_{r,0}}{\rho_c}, \Omega_\Lambda = \frac{\rho_\Lambda}{\rho_c} \tag{1.10}$$

As observations do not demand a perfectly flat universe, we include as well the curvature of the universe with

$$\Omega_k = -\frac{k}{H_0^2} \tag{1.11}$$

However, we must note that this quantity is negligible in practice. Now,

$$\Omega_m + \Omega_r + \Omega_\Lambda + \Omega_k = 1 \tag{1.12}$$

If we express the first of the Friedmann's equations in 1.6 with these densities, we can isolate the Hubble function as

$$H(z) = H_0 \sqrt{\Omega_r (1+z)^4 + \Omega_m (1+z)^3 + \Omega_k (1+z)^2 + \Omega_\Lambda} \tag{1.13}$$

This is an interesting function that allows us to establish relations between time and distance with redshift. Cosmological time in a given redshift z can be obtained by integrating

$$t(z) = \int_z^\infty \frac{dz'}{(1+z')H(z')} \tag{1.14}$$

While the distance-redshift relation is

$$r(z) = c \int_0^z \frac{dz'}{H(z')} \tag{1.15}$$

We will widely use this later equation to determine the distance to a galaxy given its redshift value. $r(z)$ is known as the *comoving distance*, and can be understood

as a label, a value non dependent of the expansion. However, for a more intuitive idea of distance measurement, the separation between two objects measured with a rigid ruler, we must specify the effects of universe expansion. The expansion can be introduced in this quantity by just multiplying it by the scale factor $a(t)$, obtaining the *proper distance* $d(r) = a(t) \cdot r(z)$. This is a simplification of equation 1.2, where θ and ϕ are constant. This is unfortunately a non measurable distance, since we do not have rigid rulers for galaxies.

Another distance widely used in cosmology is the *luminosity distance* D_L. Given the emitted and the observed light of an object, one can derive the distance between the source and the observer using the inverse square relation, which expressed as the flux-luminosity relationship is:

$$D_L = \sqrt{\frac{L}{4\pi f}} \tag{1.16}$$

where f is the bolometric flux and L is the luminosity of the object. This distance requires *standard candles*, objects with known absolute luminosities, such as supernova of type Ia. For short distances, D_L is a good approximation to the natural notion of distance in Euclidean space. However, for non negligible redshifts, in an expanding and non flat universe, D_L greatly diverge from proper distance. We can generalize the relation between the f and L with

$$\begin{aligned} f &= \frac{L}{4\pi S_k(r)(1+z)} \\ D_L &= S_k(r) \cdot (1+z) \end{aligned} \tag{1.17}$$

where function $S_k(r)$ is defined as in eq. 1.3. Thus, in the expanding nearly flat universe of ΛCDM, curvature vanishes and $S_k(r) \simeq r$, but the expansion still should be taken into account:

$$D_L = r(1+z) \tag{1.18}$$

Fitting luminosity distances has been a very successful way of estimating the cosmological model parameters. Modern galaxy surveys will extend this kind of

analysis with a new type reference objects, the Baryonic Acoustic Oscillations (BAO). These are structures generated before the time of recombination which *froze* after decoupling and their distinctive shell-like shape and size can be still observed. Knowing this is a constant, we can use it an *standard yardstick* with the angular distance. If we know the length of an observable object, such as the shell diameter of a BAO, we can measure as well the subtended angle $\delta\theta$ of the yardstick and obtain the angular distance D_A:

$$D_A = \frac{l}{\delta\theta} \tag{1.19}$$

Where, for very distant objects, $\delta\theta \ll 1$. As for the luminosity distance we can generalize this quantity for the expanding curved universe. For a rigid distance l, from metric 1.3 we have

$$l = a(t_e) S_k(r) \delta\theta = \frac{S_k(r)\delta\theta}{1+z} \tag{1.20}$$

and the angular distance is

$$d_A = \frac{S_k(r)}{1+z} \tag{1.21}$$

For a flat expanding universe, this relates with the previous distances as

$$d_A = \frac{D_L}{(1+z)^2} = \frac{r}{1+z} \tag{1.22}$$

The estimation of these distances for the BAO's will significantly enhance our fitting of the cosmological parameters and our comprehension of the universe expansion.

One of the main success of modern cosmology and ΛCDM consist in having properly estimated the densities Ω_i. Many efforts have been dedicated to this aim through the study of different observables, such as the CMB, the accurate measurements of the distance-redshift relation with standard candles (such as the supernovae Ia), or the observation of large galaxy scale structures such as the BAOs. Recent measurements of these parameters can be found in Komatsu et al. (2011),

Kowalski et al. (2008), Tegmark et al. (2006) and Planck Collaboration et al. (2014) and the conclusions are that we live in a nearly flat universe ($\Omega_k \simeq 0$) with most of the energy content in the universe belonging to dark energy ($\Omega_\Lambda \simeq 0.73$) and matter ($\Omega_m \simeq 0.27$), the radiation energy density being negligible. The Hubble constant is estimated to be around 70km $s^{-1}Mpc^{-1}$, but in this work, as previously indicated, we will use $H_0 = 100h^{-1}$km $s^{-1}Mpc^{-1}$.

1.1.3 Modern observational cosmology

The structure of the universe can be observed and studied through many different probes. The most important ones are the Cosmic Microwave Background radiation and the matter distribution on large scales. In this thesis work we focus on the latter one, analyzing surveys of measured positions of galaxies in the universe. The main appreciable feature in these catalogs is the clustering of galaxies in clusters and other structures such as filaments, sheets and void-like structures. Modern surveys have grown in size and depth giving a vision where the homogeneity scale is achieved, and therefore allowing us to compare our models in quantitative detail. The study of these surveys acquires special interest since it can be assumed that galaxy positions trace the distribution of mass in the universe, however, this is only just an approximation since it can be shown that there are deviations between the galaxy and the dark matter distributions. This deviations will be measured through the so-called *galaxy bias*.

In order to be used as a catalog of 3-dimensional positions, or a spatial point process as we will introduce, a galaxy survey needs to include several quantities for every galaxy. In equatorial coordinates the position of a galaxy in the sky is determined by the distance and two angles, the right ascension and the declination. These two angles are estimated very accurately in the modern galaxy redshift surveys. It is in the estimation of the distance where most of the difficulty is found. Different physical properties can be used to estimate distances to distant objects, but only the distance-redshift relationship (equation 1.15) can be used for large amounts of distant galaxies. Our first task is then to obtain reliable *spectral energy distributions (SEDs)* of our galaxies, the light flux density as a function of wavelength. We can detect the presence of emission and absorption lines in these functions, features related to elements existing in the galaxies (or in the path

of their light to our detectiors), whose wavelength in rest frame are well known. Any deviation from this wavelength in the observed SED must be understood as a shift produced by a variation in the relative velocities between the observed galaxy and us. When this is exclusively due to the expansion of the universe, equation 1.15 provides us the exact comoving distance to that galaxy. However, galaxy redshifts are also contaminated by their peculiar velocities, independent of the expansion and caused by the gravitational drift of the surrounding matter. Peculiar velocities can greatly distort our measurements of the distances when their direction vector is colinear to and indistinguishable from the radial direction of the universe expansion. This is the case of galaxies infalling into clusters at a high velocity, we will explain a few ways to deal with these phenomena in section 3.3.2.

Meanwhile, here we introduce the two known methods of estimating the shift in the SEDs, spectroscopy and photometry. These methods give names to the galaxy surveys employing them.

Spectroscopic surveys

These surveys take great efforts in reliably obtaining the complete SED of every object in the survey. After a previous imaging of the sky area of study, objects of interest are targeted to be observed with detail by a spectrograph. If this process is repeated for a high number of galaxies, a high quality survey might be obtained (despite the redshift distortions explained above) where galaxy structure is easily identified, even with the naked eye.

Nevertheless, these kind of surveys have to cope with several difficulties that can greatly affect the final outcome. First, the observation of a high number of sources with precise spectrographs is a demanding task both in money and time, and the mapping of large areas of the sky might be hard to reach. In addition to this, the observation of a galaxy requires to isolate it from other galaxies, which can be difficult for close galaxy pairs. Even with modern spectrographs using optical fibers, large fractions of galaxies living in high density environments can be lost, difficulting the study of galaxy clustering at small scales.

Most of the spectroscopic surveys today are actually mixed surveys, where both the spectroscopic and the photometric methods are performed. It is the dedication devoted to each method which is used to classify them. Examples of successful surveys developed with these techniques are the several projects under the *Sloan*

Digital Sky Survey (York et al., 2000), the *2dF Galaxy Redshift* (Colless et al., 2001) survey, *VIPERS* (Guzzo et al., 2014), the *VIMOS-VLT Deep Survey* (Le Fèvre et al., 2005) and the *Baryon Oscillation Spectroscopic Survey* (BOSS, Dawson et al. (2013)). These surveys have provided the data used in abundant publications of the last 15 years and are responsible for relevant discoveries like the detection of the Baryonic Acoustic Oscillations. The proper description of large scale structure of the universe made by these surveys has greatly contributed to determine the parameters of the cosmological standard model and its validity.

In the near and medium term future new spectroscopic surveys will be performed, allowing us to perform new major advances. Some of these ongoing surveys include the EUCLID survey (Amendola et al., 2013), the *Dark Energy Spectroscopic Instrument* (DESI, Schlegel et al. (2011)), WAVE (Dalton et al., 2012) and *KMOS Redshift One Spectroscopic Survey* (KROSS, Stott et al. (2016)).

Photometric surveys

An alternative to the previous surveys are the photometric surveys. Instead of obtaining the whole SED, we are only interested in learning its general shape. After a deep study of near galaxies SEDs, we have realized that galaxies can be categorized in few types sharing common and regular shapes aside from wavelength shifts. Then, given the SED of a distant galaxy, even if it is on low resolution, it can be possible to include the galaxy in one of the previous categories after redshift correction. The amount of necessary correction is an indicative of the redshift of the observed galaxy and therefore its distance can be estimated. An example technique used for the estimation of photometric redshifts is the BPZ method (Benítez, 2000), used in the ALHAMBRA survey (see section 3.4.1).

This SED information is obtained using images taken through several filters, each one picturing the image of the galaxy in a different wavelength band. After a source detection process, the images can be used to measure the luminosity in each wavelength band and therefore, a discrete sampling of the SED values. Knowing that each galaxy should be close to one of the pre-established templates, the redshift estimation can be performed interpolating between different templates. The efficiency of this method depends on the accuracy of the templates, the resolution of the SED approximation and specially on the photometry precision.

The main advantages of this procedure are its cheapness, both in terms of money and observational time, and the acquisition of full images of the sky, where no bright enough source is lost and high density regions can be properly studied. On the disadvantages, as advanced, photometric redshifts may not be useful due to its approximate measurements, which imply higher uncertainties than the spectroscopic redshifts.

Among the photometric surveys we can found the *Cluster Lensing And Supernova survey with Hubble* (CLASH, Postman et al. (2012)), the *Advanced Large, Homogeneous Area Medium Band Redshift Astronomical survey* (ALHAMBRA, Moles et al. (2008), Molino et al. (2014)), used in this thesis work, the *Panoramic Survey Telescope & Rapid Response System* (PanStarrs, Kaiser et al. (2010)), the *Cosmological Evolution Survey* (COSMOS, Ilbert et al. (2009)), the *COMBO-17 Survey* (Wolf et al., 2001) and the *Canada-France-Hawaii Telescope Legacy Survey* (CFHTLS, Hoekstra et al. (2006)). Incoming surveys include the *Dark Energy Survey* (DES, Parkinson et al. (2012)), the *Large-aperture Synoptic Survey Telescope* (LSST, LSST Dark Energy Science Collaboration (2012)) and the *Javalambre Physics of the Accelerating Universe Astrophysical Survey* (J-PAS, Benítez et al. (2015)).

This thesis work makes use mainly of the SDSS and the ALHAMBRA surveys. For the latter, we show in Fig. 1.4 a comparison between the photometric and spectroscopic estimated redshifts of 3826 galaxies, certifying the general reliability of ALHAMBRA photometric redshifts.

1.2 Spatial point processes in Astrophysics

A galaxy spatial point process is a mathematical model that describes the arrangement of galaxies as they are distributed in the Universe. This description contains the geometrical and statistical information of the spatial patterns that galaxies form in their environment, and enclose most of the laws governing the spatial distribution of galaxies in the Universe. Gravity, and also other interaction phenomena at small scales, are the laws that shape the distribution of galaxies, creating the characteristic patterns of the Cosmic Web. Through the study of the

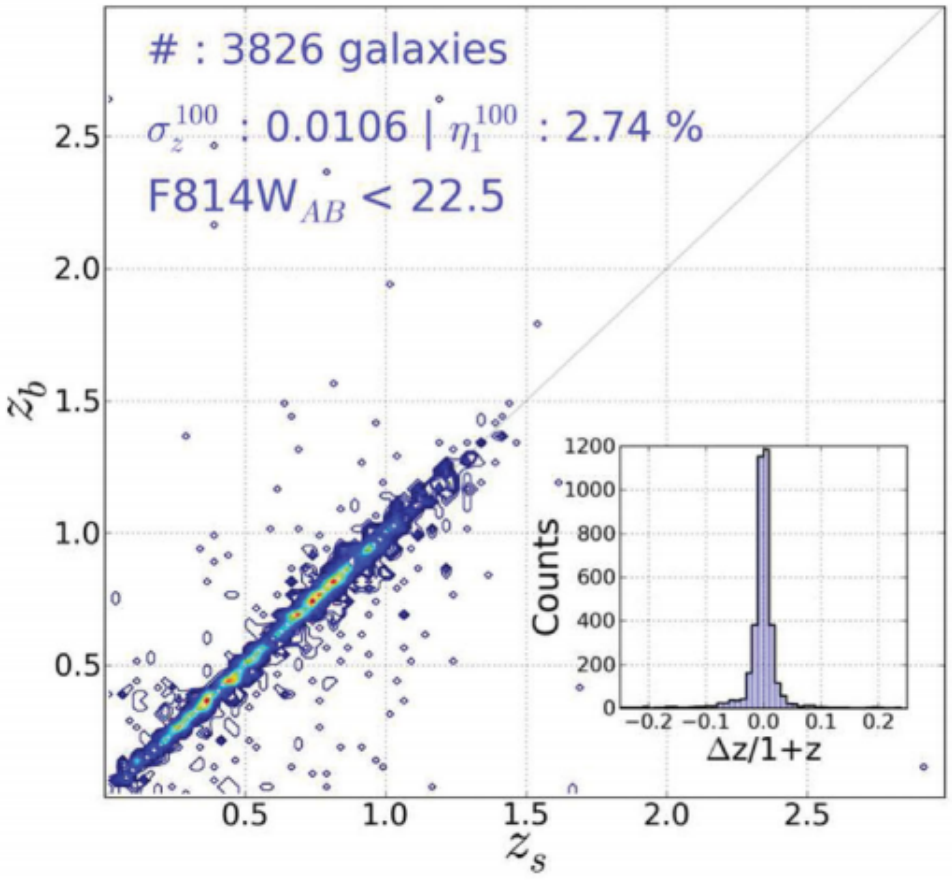

FIGURE 1.4: Comparison between the ALHAMBRA photometric redshift z_b and the spectroscopic redshift z_s, along with the error distribution $\Delta_z/(1+z)$ for a bright sample ($m_{F814W} < 22.5$) with $\sigma_z < 0.0106$ and a fraction of catastrophic outliers $\eta_1 \sim 2.7$ per cent. Figure from Molino et al. (2014)

observable galaxy distribution we may understand how these structures have been formed and, hopefully, understand the underlying laws that shape them.

Gravity produces a vast amount of diverse structures and patterns. Clusters, filaments, sheets and voids mixed in our studied galaxy samples, and versatile statistics are needed to detect, identify or model them. Through the general properties of galaxies in their environment and spatial relations between them, such as relative distances, we can infer the properties of the phenomena that produces such realization of points. As advanced in the beginning of this introduction, these statistics can be classified in three main approaches following Baddeley (2007), depending on the pursued results: **summarizing**, or expressing a general characteristic of the data statistically, which represents a "statistical summary" extracted from the data, such as the two-point correlation function; **mining**, or searching

relevant features in the data such as clusters, voids or filaments; and **modeling**, or formulating a probability model for the point pattern and fitting the model to the observed data.

One can make use of conventional methods that naturally arise from point process statistics, usually summary statistics, but often the most useful techniques at understanding a point process and answering our questions are created *ad hoc*. Multiple sciences make use of point processes. A non exhaustive list would include works in biology (Illian et al., 2008, Pfeifer et al., 1992), aesthetics (Penttinen & Ylitalo, 2015), criminology (Ang et al., 2012) or econometrics (Engle, 1982) among many more. Astrophysics is not an exception to this list, with contributions form Baddeley (2007) and Tempel et al. (2014). The effectiveness of our algorithms usually depends on going beyond general statistics and developing methodologies that incorporate specific knowledge from each study fields. This is specially necessary when modeling populations, since the dynamics ruling diverse phenomena are also peculiar. Nevertheless, the construction of such models is a necessary step to fully understand and describe a point process.

In a point process, all elements have two components: position and characteristics, which reflect the geometrical and physical reality. In astrophysics, every galaxy sample is embedded in a certain space. Since we are studying spatial point processes, we are interested in galaxy samples providing their locations in the Universe. Modern galaxy surveys usually contain the equatorial coordinates of galaxies plus their redshift estimation. These coordinates need to be converted into cartesian coordinates before applying the techniques. In addition, these surveys provide data from different epochs of the Universe, which conditions the distribution of our point process. As space evolves with redshift, galaxies must be studied separately depending on their location along the Universe history. Galaxies must be characterized from their observables, creating an effective classification regarding their main physical properties. This segregation tries to build data subsets made of elements having a certain common property or mark. This property would be usually of physical nature, like similar luminosities or spectral types. This classification will be specially useful when we try to decompose the overall galaxy pattern into the different sub patterns that coexist and constitute the parent galaxy distribution. When we have been able to disentangle these sub patterns, we also have been able to establish a correlation between the galaxy type

and the spatial behavior or distribution.

Galaxy catalogs are the result of an observational process, with its characteristic errors and inaccuracies. One of the major troubles is the presence of unobserved areas of the sky. Since galaxies are observed as projected points in the celestial sphere, we define the *angular selection function* as the area effectively observed by our telescopes. The rest of the sky is called in this thesis work the mask. In this mask we include not only regions not observed but also regions where the presence of a star or other bright objects does not allow us to properly observe the galaxies beyond them. The distance limits are defined by the properties of the survey and determine the closest and the furthest position observed where a galaxy could lie. These limits are usually defined in order to guarantee basic statistical properties of a point process, such as homogeneity. Together, angular mask and limits define the 3-dimensional *window*, which can be trimmed to shape it in more easily treatable geometries, such as a cuboid.

In this thesis we always make use of algorithms prepared to operate in three dimensions. This sometimes requires to develop or generalize previous algorithms, as we have done in chapters 4 and 5. This requires using 3-dimensional datasets with populations expressed in cartesian coordinates. As said, catalogues provide galaxies in equatorial coordinates, where the position of an object in the sky is defined by two angles, the right ascension (α) and the declination (δ). Once we obtain the comoving distance D between the galaxy and the observer with equation 1.15, we convert them into Cartesian coordinates with

$$\begin{aligned} x &= D \cdot \sin\theta \cdot \cos\varphi \\ y &= D \cdot \sin\theta \cdot \sin\varphi \\ z &= D \cdot \cos\theta \end{aligned} \tag{1.23}$$

where $\theta = (-\delta + 90) \cdot \pi/180$ and $\varphi = \alpha \cdot \pi/180$ when the catalog is provided in degrees.

1.2.1 Probability background

In this section we proceed to introduce and properly define the mathematical background of the point processes. Most of it can be consulted in (Illian et al., 2008). Further description of the used functions and objects will be given in subsequent sections, when necessary. We will need as well to introduce some basic definitions of Measure Theory to establish the needed probabilistic framework where we will develop the Point Process statistics.

Observation Window

An observation window is the measure space compound by the triplet $(W, \mathcal{B}, \nu)$, where

- $W \in \mathbb{R}^d$. It can be understood as the geometrical region introduced in the previous section and will be referred as *window* indistinctly.
- $\mathcal{B}$ is a Borel σ-algebra and
- $0 < \nu(W) < \infty$ is the Lebesgue measure.

A σ-algebra is a system of subsets χ of some set X satisfying:

- $X \in \chi$
- if $A \in \chi$, then $A^c \in \chi$
- if $A_1, A_2, \ldots \in \chi$, then $\bigcup_{i=1}^{\infty} A_i \in \chi$

We say that the family $\mathcal{B}^d$ of sets of $\mathbb{R}^d$ is a Borel σ-algebra if it is the smallest σ-algebra on $\mathbb{R}^d$ that contains all the open subsets of $\mathbb{R}^d$. In other words, $\mathcal{B}^d$ contains all the subsets of $\mathbb{R}^d$ that can be constructed from the open subsets by the basic operations and by limits.

The measure ν coincides in dimensions 1, 2 and 3 with the conventional measure of length, area and volume. Measures are crucial functions for the point processes analysis and require deeper definition:

The pair (X, χ) is called a measurable space and $A \in \chi$ is called a measurable set. The function $f : X \longrightarrow \mathbb{R}$ is said to be χ-measurable if for each $B \in \mathcal{B}$, the inverse image $f^{-1}(B) = \{x \in X | f(x) \in B\}$ belongs to χ (the σ-algebra associated to X).

Measure

A measure on (X, χ) is a function $\mu : \chi \longrightarrow [0, \infty[$ with the following properties

- $\mu(\emptyset) = 0$
- $\mu\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mu(A_i)$ for all the sets $\{A_i\}_{i=1}^{\infty} \in \chi$ with $A_i \cap A_j = \emptyset$ whenever $i \neq j$ (σ-additivity).
- if A, $B \in X$ and $B \subset A$, then $\mu(A \backslash B) = \mu(A) - \mu(B)$

In this work we use Radon measures, which in addition must be defined on $\mathcal{B}^d$ and are locally finite, this is, finite on bounded sets. In particular, the Lebesgue measure, which gives us the d-volume of $A \subset \mathbb{R}^d$, $\nu_d(A)$.

This is the appropriate topological definition of the universe where we proceed to study the point processes, but we still need to adapt it so we can analyze them with statistical and probabilistic techniques.

Probability space

A probability space is the points configuration space formed by the triplet $(\Omega, \mathcal{F}, \mathbf{P})$ where

- $\Omega = \bigcup_{i=1}^{\infty} W_i$ is the *state space*, and W_i is the subset of all n-tuples $\{w_1, \ldots, w_i\} \subset W$.
- $\mathcal{F}$ is the *events space*: the σ-algebra given by $\mathcal{F} = \sigma(\{\mathbf{w} = \{w_1, \ldots, w_i\} \in \Omega : n(\mathbf{w}_B) = n(\mathbf{w} \cap B) = m\})$ with $B \in \mathcal{B}$ and $m \in \mathbf{N}$
- $\mathbf{P}$ is the *probability measure*, a measure satisfying $\mathbf{P}(\Omega) = 1$ and model of our point process.

In this universe, $w \in \Omega$ is a sample of points, for instance, a fragment of a point process or a particular realization. $F \in \mathcal{F}$ is an event of our probabilistic phenomena. We can define as well *random variables*, which are $\mathcal{F}$-measurable functions

on $(\Omega, \mathcal{F}, \mathbf{P})$. The measurability condition ensures that for any random variable X[1] it is possible to define probabilities with a *Distribution function* F[2]:

$$\mathbf{P}(X \leq x) = \mathbf{P}(\{w \in \Omega : X(w) \leq x\}) = F(x) \tag{1.24}$$

1.2.2 Definition of a Point Process

Now we are in position to define

Point process
A point process is a measurable mapping in W from a probability space $(\mathcal{S}, \mathcal{A})$ in $(\Omega, \mathcal{F})$. Its definitions are given by

$$\mathbf{P}(X \in F) = \mathbf{P}\{w \in \mathcal{S} : X(w) \in F\} \tag{1.25}$$

with $F \in \mathcal{F}$ and X a random variable. The realization of a point process is a random set of points in W. We shall sometimes identify X with $\mathbf{P}(X \in F)$ and call them both a point process.

In a more graphical approximation, we may understand a point process as a random configuration of points $\mathbf{w}$ in a observation window W. This configuration of points is $\mathbf{w} = \{w_1, \ldots, w_n\}$, with n the corresponding number of points. Its properties include a) locally finite: the number of points $n(\mathbf{w})$ is finite for a finite volume W and b) simple: $w_i \neq w_j$ for $i \neq j$. In this work we assume $W = \mathbb{R}^3$ or a compact subset.

Some geometrical properties of the point processes are specially interesting in the cosmological case. We say that a point process X on W is

- *Stationary* if it has the same distribution as the translated process X_w, that is, $\mathbf{P}(\{w_1, \ldots, w_n\}) = \mathbf{P}(\{w_1 + w, \ldots, w_n + w\})$ for any $w \in W$.

[1] Notice the change in the notation: when defined, X might mean a parent universe or a random variable.

[2] Again, notation can be confusing: F might be an element of the events space or a distribution fucntion.

- *Isotropic* if it has the same distribution as the rotated process $\mathbf{r}X$, that is, $\mathbf{P}(\{w_1, \ldots, w_n\}) = \mathbf{P}(\{\mathbf{r}w_1, \ldots, \mathbf{r}w_n\})$ for any rotation matrix $\mathbf{r}$.

Now that the point process is fully defined, we can introduce the basic procedure for characterizing it.

The point process theory establishes, sometimes as a definition and sometimes as a theorem, that the distribution of a point process X is determined by the finite dimensional distribution of its count function[3], i.e. the joint distribution of $n(B_1)$, . . . , $n(B_m)$ for any B_1, . . . , $B_m \in \mathcal{B}$ and $m \in N$. Where $\mathcal{B}$ includes all the existing subsets in the X containing space. This can be understood as follows, if we are able to know how many elements inhabit any region (in size or shape) occupied by our sample, then we know everything that can be said about the distribution. This leads us to the definition of the probability of finding a region B containing N points $f(N, B)$. This distribution is more commonly used in its conditional shape, both $f_B(N)$ or $f_N(B)$. This will be the approach used in section 2.2 for random regions B with $\nu(B) = V$. An interesting case is when $N = 0$, the so called *Void function* or *void probability function*, where we study the probability of finding empty subregions of a given size or shape.

$$v(B) = P(n(B) = 0), B \in \mathcal{B} \tag{1.26}$$

This is a complementary statement, defining the occupancies of regions is equivalent to define the gaps between them. The information of the existing gaps will be used in section 4.2.3.

This characterization is sometimes done by means of the *intensity function*. For an infinitesimal region dB that contains the point x, we define the intensity function as

$$\lambda(x) = \lim_{\nu(dB)\to 0} \frac{\langle n(dB) \rangle}{\nu(dB)} \tag{1.27}$$

where $\langle \cdot \rangle$ denotes expected value of the random variable. Constant intensity is equivalent to stationarity in the process.

[3]We define the count function $n(B)$ has de number of elements $x \in X$ inside a volume or region B.

1.2.3 Marked processes

As already introduced, the elements that shape a point process usually have an attached information which is crucial to the proper understanding of the process and its distribution. In astrophysics, this additional characteristic might be, for example, the luminosity or the spectral type of a galaxy and are referred to as *marks*. When we study the process having this into account, we call it a marked point process. Any property of our measured elements could be determinant to the proper understanding of its nature, and therefore, the interactions and forces that drive their behavior and distribution. In many practical cases it has been proven that the correct classification of a population into two or more marks has been necessary to unveil the correlations beneath the distribution and satisfactorily evidenced by the right statistics analysis.

In this thesis work we will make use of the following mathematical interpretation:

Let be $X = \{x_i\}_{i=1}^{N}$ a point process and $m(x_i)$ has de corespondent mark of the point x_i, we denote M as the marked point process

$$M = \{[x_i; m(x_i)]\}_{i=1}^{N} \tag{1.28}$$

The mark $m(x_i)$ can adopt multiple forms, either quantitative (continuous) or qualitative (categorial, discrete) being the most common an integer or real number describing some property. In the cosmological case, where the elements of the process are galaxies, we can use as marks as many observable as we can obtain. These will be physical quantities such as luminosities, magnitudes, masses o information regarding their spectral distribution or morphological type.

Our approach to the mark analysis in this work is the construction of qualitative marks derived from continuous ones. In order to properly treat the data in the point process when applying our statistics, we select the marks of our interest and build a marked space where each galaxy locates in a point of the space. Notice that two different points, like two different galaxies, can share the same marks, and therefore, occupy a common coordinate in the marked space; we call this a *multiplicity*. Then, we perform a binning of the space creating two or more populations that satisfy a certain mark criterium. When the final number of populations is two we call it a *bivariate* process, or *multivariate* otherwise.

In addition, this technique can be performed independently of the spatial point process, for example when we use the luminosities in certain filters to segregate stars from galaxies.

When we proceed this way, we assume that segregated elements belonging to the same sample behave the same way. Elements that share some common properties can be hence treated equally in the statistical analysis. This assumption of marked processes is used in section 3.4.1, when we segregate galaxies depending on their spectral type.

1.2.4 Examples of point processes

Some point processes are specially interesting. We describe here some of them.

The Binomial point process

We say a point process follows a binomial distribution when the probability of finding in a region $B \in \mathcal{B}$ a random point uniformly distributed in a window W is, using Laplace formula,

$$\mathbf{P}(x \in B) = \frac{\nu(B)}{\nu(W)} = p \tag{1.29}$$

The random variable $n(B)$, the number of point contained in B, follows a Bernoulli distribution with probability p. If we generalize it to n independent points distributed uniformly, we obtain

$$\mathbf{P}(x_1 \in B_1, \ldots, x_n \in B_n) = \mathbf{P}(x_1 \in B_1) \cdot \ldots \cdot \mathbf{P}(x_n \in B_n) = \frac{\nu(B_1) \cdot \ldots \cdot \nu(B_n)}{\nu(W)^n} \tag{1.30}$$

for Borel subsets $B_1, \ldots, B_n$ of W.

Similarly, now $n(B)$ follows a binomial distribution with parameters $n = n(W)$ and $p = \nu(B)/\nu(W)$:

$$\mathbf{P}(n(B) = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{1.31}$$

The *intensity* of this process, or the mean number of points per unit volume is

$$\rho = \frac{n}{\nu(W)} \tag{1.32}$$

And the mean number of points in the set B is $\mathbb{E}(n(B)) = np = \rho\nu(B)$. For this process, the intensity function is constant and $\hat{\lambda} = \rho$.

This is used as the most basic example of process, since all its points are independent and no structure is found on it. However, since the total amount of points in W is known, the number of points per subset is not independent:

$$n(B) = m \rightarrow n(W \backslash B) = n - m \tag{1.33}$$

Poisson point process

The homogeneous Poisson point process X, following Illian et al. (2008) and Baddeley et al. (2015a), is characterized by three fundamental properties:

(1) *Poisson distribution of point counts.* The number of points of X contained in any subregion B ($n(X \cap B)$) is a random variable with Poisson distribution.

(2) *Homogeneous intensity.* The expected number of points falling in B is $\mathbf{E} = \lambda \cdot \nu(B)$, where λ is constant.

(3) *Independent scattering.* For disjoint k sets $\{B_j\}$, $n(B_j)$ are independent random variables, for arbitrary k.

(4) *Conditional property.* Given that $n(X \cap B) = n$, the n points are independent and uniformly distributed in B.

The quantity λ, as before, is the intensity or *point density.* It describes the expected number of points to be found in a unit volume. Given λ we can apply properties (1) and (2) for the *1-dimensional distribution*

$$\mathbf{P}(n(B) = k) = \frac{(\lambda \cdot \nu(B))^k}{k!} \exp\left(-\lambda \cdot \nu(B)\right) \tag{1.34}$$

and therefore, $n(B)$ has a Poisson distribution with parameter $\lambda\nu(B)$. With property (2) we can generalize this to the *finite-dimensional distribution*

$$\mathbf{P}(n(B) = j_1, \ldots, n(B) = j_k) = \frac{\lambda^{j_1+\cdots+j_k} \cdot \nu(B_1)^{j_1} \cdots \nu(B_k)^{j_k}}{j_1! \cdots j_k!} \exp\left(-\sum_{i=1}^{k} \lambda\nu(B_i)\right) \tag{1.35}$$

We say this is a *completely random* process. This process satisfies stationarity ($\hat{\lambda} = \lambda = cnt$) and isotropy.

The Poisson process can yet be generalized to an *Inhomogeneous* Poisson process. This process has a non constant intensity function $\lambda(x)$ which may change from one position to another, showing variations in density. Now, the probability distribution of the number of points lying in a bounded region B will be given by

$$\mathbf{P}(n(B) = k) = \frac{(\Lambda(B))^k}{k!} \exp\left(-\Lambda(B)\right) \tag{1.36}$$

where $\Lambda(B)$ is the intensity measure defined as

$$\Lambda(B) = \int_B \lambda(x)dx \tag{1.37}$$

Finally, an interesting generalization of Inhomogeneous Poisson process must be introduced, the *Cox process* (Cox & Isham, 1980). In this case, the intensity function $\lambda(x)$ is itself a random function, and the intensity measure is generated from driving a random measure distribution. In this case we say the Cox process is two times random, since the intensity and the probability distribution of the number of points are random variables.

1.2.5 Moment measures

In general, point processes may not be as easy to characterize with few simple properties as the Poisson process. For this reason the moment measures might be the appropriate mathematical techniques to describe the probability distribution of the number of points lying in a bounded region B. Using these moments,

interesting functions might be created to obtain a quantification of their properties. These properties might include the level of aggregation or clustering and other functions based on local intensity. Such functions are very useful at detecting the presence of structure, discarding the possibility of a Poisson process.

We define the k-th order moment measure of a point process X as the measure $\mu^{(k)}$ defined by

$$\int_{\mathbb{R}^d} f(x_1, \ldots, x_k)\mu^{(k)}(d(x_1, \ldots, x_k)) = \mathbf{E}\left(\sum_{x_1,\ldots,x_k \in X} f(x_1, \ldots, x_k) \right) \tag{1.38}$$

where $f(x_1, \ldots, x_k)$ is any non-negative measurable function on $\mathbb{R}^d$. In particular, for the identity function

$$f(x_1, \ldots, x_k) = \mathbf{1}\{(x_1, \ldots, x_k) \in B_1 \times \ldots \times B_k\} \tag{1.39}$$

we get

$$\mu^{(k)}(B_1, \ldots, B_k) = \mathbf{E}(n(B_1) \cdots n(B_k)) \tag{1.40}$$

and, if $B_1 = \cdots = B_k = B$,

$$\mu^{(k)}(B^k) = \mathbf{E}(n(B)^k) \tag{1.41}$$

Thus $\mu^{(k)}$ yields the kth moment of the real-valued random variable $n(B)$, which is the number of points in B. This moment is directly related to the Counts-in-Cells distribution (see section 2.2), the probability of finding N points in a volume V.

Factorial moment measures

The k-th order moment measure $\alpha^{(k)}$ is

$$\int_{\mathbb{R}^d} f(x_1, \ldots, x_k)\alpha^{(k)}(d(x_1, \ldots, x_k)) = \mathbf{E}\left(\sum_{x_1,\ldots,x_k \in X}^{\neq} f(x_1, \ldots, x_k) \right) \tag{1.42}$$

where $\overset{\neq}{\sum}$ is the sum of all k-tuples of distinct points in X including all permutations of given points. This is the difference as respect to $\mu^{(k)}$, the sum omits all k-tuples in which 2 or more points are repeated. If $B_1, \ldots, B_k$ are disjoint sets, then

$$\mu^{(k)}(B_1 \times \ldots \times B_k) = \alpha^{(k)}(B_1 \times \ldots \times B_k) \tag{1.43}$$

and, for the case $k = 2$, we obtain

$$\alpha^{(2)}(B_1 \times B_2) = \mu^{(2)}(B_1 \times B_2) - \Lambda(B_1 \cap B_2) \tag{1.44}$$

with Λ is the intensity measure as defined in 1.37. For $k > 2$, the contribution of the intersections between sets must be removed one at a time from $\mu^{(k)}$ to keep the disjoint parts without omissions. Hence, for $B_1 = \cdots = B_k = B$,

$$\alpha^{(k)}(B^k) = \mathbf{E}(n(B)(n(B) - 1) \cdots (n(B) - n + 1)) \tag{1.45}$$

Product densities

A product density describes the frequency of possible configurations of n points. If we consider the spheres $b_1, \ldots, b_k$ with centers $x_1, \ldots, x_k$ and infinitesimal volumes $dV_1, \ldots, dV_k$, then $\varrho^{(k)}(x_1, \ldots, x_k)dV_1 \cdots dV_k$ is the probability that there is a point of X in each of the b_i spheres.

If $\alpha^{(k)}$ satisfies some continuity properties, then $\varrho^{(k)}$ is the kth product density of $\alpha^{(k)}$:

$$\alpha^{(k)}(B_1 \cdots B_k) = \int_{B_1} \cdots \int_{B_k} \varrho^{(k)}(x_1, \ldots, x_k)dx_1 \cdots dx_k \tag{1.46}$$

and, therefore,

$$\int_{\mathbb{R}^d} f(x_1,\ldots,x_k)\varrho^{(k)}(x_1,\ldots,x_k)(d(x_1,\ldots,x_k)) = \mathbf{E}\left(\sum_{x_1,\ldots,x_k\in X}^{\neq} f(x_1,\ldots,x_k)\right) \tag{1.47}$$

This is a more intuitive descriptor with wide applications in its first orders. For $k = 1$, the product density is $\varrho^{(1)}(x)dx = \lambda(x)dx$, the intensity function, and for $k = 2$ under conditions of stationarity and isotropy we derive the pair correlation function. We will discuss this further in section 3.2

1.3 Aim of this thesis

In this thesis we approach the study of galaxy large-scale structure distribution from a point process point of view. The point process analysis understands probabilistic events as points in a framework. As introduced in section 1.2, three strategies can be followed depending on the used methodologies: the summary statistics, the data mining and the modeling. We perform all this approaches in different cases, proving that the point process approximation is an excellent strategy for the analysis of galaxy distribution.

The main aim of our studies is, therefore, to effectively map the found galaxy patterns with statistically based methods. The methods and models used try to highlight different properties of the galaxy distribution, which respond to our questions and help us to increase our comprehension of the Universe. This branch of point processes includes a wide range of different techniques and methodologies and still produces new applications that enhance our capacity of deducing the nature of the phenomena we study. The selection of techniques we have made use of include some of the most well known functions in astrophysics today, mainly summary statistics. We will use these statistics over datasets selected from some of the most recent galaxy surveys, allowing us to obtain significant results that increase our comprehension of galaxy distribution. For our data mining and modeling analysis we proceed the opposite way, testing original point process based methodologies over well known datasets. We include as well specific adaptations introduced by ourselves. These algorithms could constitute relevant contributions for the analysis of the galaxy distribution.

The thesis is divided in 6 chapters. Chapter 1, this one, broadly introduces the cosmological and probabilistic theoretical backgrounds. Part II contains chapters 2 and 3, which constitutes the summary statistics work. In chapter 2 we make use of the Counts-in-Cells (CiC) algorithm, that will allow us to fit and discriminate several probability density functions of the galaxy abundance distribution. These distributions include newly proposed functions, based in the cosmological model. We use data from the Sloan Digital Sky Survey and LasDamas simulations. In chapter 3 we analyze clustering with the correlation function. The existence of redshift distortions forces us to include modifications and use the projected correlation function in our analysis. Using the recently completed ALHAMBRA survey catalog we are able to infer the behavior of galaxy clustering at scales never observed before. In chapter 4 (part II) we attempt the modeling of a galaxy sample. The Gibbs models describe the distribution of a point process regarding the interactions between points, a technique that we will apply over both the SDSS and LasDamas simulations again. As a complete model we obtain a full parametric description of our galaxy distribution, with a characterization of the clustering levels. Finally, in chapter 4 (part III), a combination of modeling and data mining is presented: the Mixture models, a statistic based on multivariate analysis and cluster finding algorithms. Assuming models for the different structures present in a galaxy sample, we can identify them and model the whole set of structures at the same time. With such a model properly fitted, we can study the pattern in detail, measuring the shapes and sizes of clusters or their interaction with other structures. A Mixture model is hence, also a data mining technique. We will use data from the MultiDark simulation, where several clusters and structures can be identified and modeled with dark matter profiles. With the Mixture model approach we were able to obtain an integral vision of the sample and an effective segregation of each cluster from its neighbor. Both chapters 4 and 5 are finished with a complete error analysis based in modern residual analysis for point processes. Chapter 6 is a conclusion of the obtained results.

Altogether, in this thesis work we aim to show how the point process strategy is an excellent choice for engaging the analysis of galaxy distribution, providing us with powerful methodologies that reveal the nature and behaviors that constitutes the phenomenon of galaxy clustering.

Part I

Summary Statistics for the Galaxy Distribution

'What!' cried Bilbo. 'You can't tell which parts were mine, and which were the Dúnadan's?'

'It is not easy for us to tell the difference between two mortals,' said the Elf.

'Nonsense, Lindir,' snorted Bilbo. 'If you can't distinguish between a Man and a Hobbit, your judgement is poorer than I imagined. They're as different as peas and apples.'

'Maybe. To sheep other sheep no doubt appear different,' laughed Lindir. 'Or to shepherds. But Mortals have not been our study. We have other business.'

Lindir and Bilbo

Chapter 2

Counts-in-Cells Distribution

Summary statistics include those data analyses that express a general trend or property of a point process in a simplified expression, such as a value or a curve. Basic statistics, such as the mean intensity of a process, can be included in this category, but they are considered not enough informative to extract significant conclusions. More complex procedures, the so called first and second order characteristics, have been proved to be extremely powerful. These include the Counts-in-Cells (CiC, this chapter), and the pair correlation function (chapter 3). The interested researcher can also make us of higher order statistics, for more information please read Illian et al. (2008). However, it is always important to remember that despite their remarkable contribution to the understanding of galaxy distribution they are not sufficient to fully describe a point process. Summary statistics used in part I are widely used statistics in modern cosmology, and we will apply them to obtain new relevant conclusions about the galaxy distribution, making use of data from recently published galaxy surveys as well.

In this first work we introduce in detail the Counts-in-Cells methodology and apply it over data from the NYU-VAGC galaxy survey (Blanton et al., 2005). This statistic provides a easily interpretable and yet highly informative descriptor of the galaxy distribution. Applying it over two populations we will describe the evolution of clustering both in terms of redshift and magnitude limits. For a deeper understanding, this characterization can be fitted with different distribution functions. Some of these functions have been widely used in cosmology, such as the Gravitational Quasi Equilibrium Distribution (Saslaw & Hamilton, 1984), the

Negative Binomial Distribution (Fry, 1986) or the Log Normal distribution (Coles & Jones, 1991). As an original contribution to the galaxy distribution analysis we will make use of the Weibull distribution (Weibull, 1951). With this analysis we expect to obtain a reliable description of the galaxy distribution as seen in the Counts-in-Cells.

In 2.2 we explain how to properly estimate the CiC distribution, dealing with edge correction problems. Section 2.3 presents the datasets in use, the mentioned NYU-VAGC galaxy survey and the LasDamas simulation catalog. Uncertainties in the estimation of our distribution are analyzed in section 2.4 using the Jackknife method, which we will use again in section 3.3.3. After obtaining our observational CiC distribution we proceed to fit it with several different probability distributions introduced in 2.5 in order to obtain the best descriptor. Results are shown in section 2.6 and conclusions are summarized in section 2.7

2.1 Introduction

One of the first statistics applied to study the galaxy clustering in the very early galaxy catalogs (that were built on the projected celestial sphere) was the Counts-in-Cells method. Hubble (1934) was the first to notice that the distribution of galaxy counts in 2-dimensional cells could be well approximated by a lognormal distribution. This technique permits to describe the spatial distribution of galaxies in a way that its is complementary to other descriptors of the galaxy clustering such as the correlation function or the power spectrum (Peebles, 1980, White, 1979). The counts probability distribution function $f_V(N)$ gives the probability that a randomly placed volume in the universe will contain exactly N galaxies. For $N = 0$, this function is known as the void-probability function (Maurogordato & Lachieze-Rey, 1987) and it is of particular interest, since it is related with higher order correlation functions (White, 1979) and provides a simple approach to establish hierarchical scaling relations (Balian & Schaeffer, 1989, Croton et al., 2004). Fry & Gaztanaga (1994) have shown that correlation functions can be measured from the moments of $f_V(N)$. Despite its clear interest, as Yang & Saslaw (2011) have pointed out, this count probability distribution function has not received as much attention as other more commonly used descriptors of galaxy clustering (Martínez & Saar, 2002).

The Counts-in-Cells process consists in counting the number of existing elements of our sample lying inside cells. The cells are regions or volumes contained in our geometry, fragments of the window. Depending on the shape, size and distribution of these regions our counts will reflect different aspects of the studied galaxy sample. Therefore, the CiC is itself a spatial point process where its geometry is inherited from the original sample and the spatial locations (centers of the cells) are defined by the user. The counts can be used as marks associated to each point, containing the descriptive information of the distribution.

Although a first order characteristic, Counts-in-Cells is a highly informative statistic, containing information from higher orders. Being able to effectively describe the CiC distribution of galaxy populations is a necessary and useful aim (Baugh et al., 1995, Colombi, 1994, Gaztañaga et al., 2000, Ueda & Yokoyama, 1996, Yang & Saslaw, 2011). A reliable fit of the CiC distribution allows us to infer properties of the galaxy distribution physics (Saslaw & Fang, 1996), or properly produce galaxy mocks (Labatie et al., 2010), abundantly used in cosmology (this thesis work, for example). Among other diverse applications, CiC can test the quality of N-body simulations or infer properties from the galaxy distribution such as the galaxy-dark matter bias (López-Sanjuan et al., 2015).

2.2 Estimation of the CiC distribution

For our Counts-in-Cells process we use spherical cells in 3-dimensional redshift space with constant radius. These cells are described by their location (3 coordinates) plus a radius r, and distributed uniformly in space allowing them to intersect. The higher the number of cells used in the calculation, the most precise will be our calculation of the Counts-in-Cells process.

Due to the irregular geometry containing the samples, the creation of the cells catalogs starts defining a uniformly distributed population of α,δ coordinates in the window of the galaxy sample. These equatorial coordinates will determine the cells location with a third distance coordinate. In order to satisfy a uniform distribution within the redshift space, we distribute the distances following the function

$$D(\xi) = ((D_{max}^3 - D_{min}^3) \cdot \xi + D_{min}^3)^{1/3} \tag{2.1}$$

where D_{min} and D_{max} are the distances corresponding to the redshift limits of our population and $\xi \sim U(0,1)$. Galaxy surveys geometry is usually slice shaped, like a cone with its vertex in the observer. That requires distributing the distances D accordingly to ensure a uniform distribution along the line of sight. Once we have our cell centers population (α, δ, D), we convert them into cartesian coordinates with equation 1.23.

Now we can proceed to calculate the percentage of our cells not included in the window, this is, covered by the mask. For this purpose was created the software MANGLE (Hamilton & Tegmark, 2004, Swanson et al., 2008). MANGLE performs several operations with complex angular windows, mapping them with sky-projected polygons and allowing us to find the polygons containing a given point on the sphere.

A galaxy survey window can be highly irregular, masked by stars and other objects. This affects not only the distribution of the cells but also their effective volume, since even if the center of the cell is inside the window, a significant part of the sphere might be outside. Once the centers of the cells are determined, we estimate the effective volume of the cell contained in the window.

This can be done by Monte Carlo integration, populating a large number of Poisson distributed points in our window and counting the number of points that lie inside our cells. This is exactly a CiC calculation over a Poisson process. Given a window of volume $\nu(W) = V$ and a cell radius r we determine the number of necessary random points with

$$N = 1/(e^2 \cdot t) \tag{2.2}$$

where $t = 4\pi r^3/3V$ is the ratio between the volume of the cells and the total volume, and e is the uncertainty level. The number of points in a non masked cell has expectancy $N \cdot t$, and only cells containing at least a 95% of this quantity will be accepted. This percentage of cell volume into the window will be estimated with uncertainty e.

For accepted cells we define the relative volume of the cell as the fraction of counted points with respect to $N \cdot t$. The accepted cells in right ascension and declination coordinates create an avoidance area keeping the centers of the cells away from the borders of the survey and other internal big gaps, but still covering small mask regions (see Fig. 2.1).

With our accepted cells we can proceed to perform the Counts-in-Cells. The counting is done using the Euclidian metric and we are interested in knowing how many galaxies are found inside every cell of radius r.

$$(x_1 - g_1)^2 + (x_2 - g_2)^2 + (x_3 - g_3)^2 < r^2 \tag{2.3}$$

where $\mathbf{g} = (g_1, g_2, g_3)$ is the comoving position of a galaxy, and the cell is centered in $\mathbf{x} = (x_1, x_2, x_3)$.

Then we multiply the number of counts by the inverse of the relative volume to obtain our Counts-in-Cells distribution. With this correction we estimate the number of galaxies unseen by the masking effects or redshift limits assuming a uniform distribution of galaxies for cells dimensions.

The Counts-in-Cells distribution can be binned in a histogram of frequencies, each bin containing the number of cells with N galaxies. If we normalize it we obtain the probability density function $f_V(N)$ of finding N galaxies in a cell of volume V. In our case, we will compute the CiC distribution $f_V(N)$ for our galaxy samples using three different cell radii: 6, 12 and $24h^{-1}$ Mpc.

2.3 Data catalogs

The data analyzed in this section is a galaxy sample from the Main catalog of the Sloan Digital Sky Survey (SDSS) Data Release 7 (DR7, Abazajian et al. (2009)). This catalog is provided by The New York University - Value Added Galaxy Catalog (Blanton et al., 2005), presented in the next section. In addition to this, we also make use of LasDamas simulation catalog (McBride et al., 2011) to test our error analysis estimations.

2.3.1 The SDSS - New York University - Value Added Galaxy Catalog

The NYU-VAGC (Blanton et al., 2005) is composed by data from the Sloan Sky Digital Survey DR7 Abazajian et al. (2009) and the 2-Micron All-Sky Survey (2MASS) (Skrutskie et al., 1997), although we only make use of the former one. The SDSS-DR7 mapped one quarter of the entire sky and performed a redshift survey of galaxies, quasars and stars. It consists of a series of three interlocking imaging and spectroscopic surveys, carried out over an eight-year period with a dedicated 2.5m telescope located at Apache Point Observatory in Southern New Mexico.

The NYU-VAGC survey provides us with the position and redshift of more than 550.000 galaxies with corrected extinction and K-corrected absolute magnitudes for 8 bands, of which the u, g, r, i, and z bands come from the SDSS. In addition, NYU-VAGC catalog also contains a survey geometry catalog, which define the window of the galaxy population and can be operated using the software MANGLE. This survey includes carefully constructed large-scale structure samples useful for calculating power spectra, correlation functions, etc.

In this work we have used two different samples selected in the r band, corresponding to the DR72 catalogs of the SDSS (Abazajian et al., 2009) included in NYU-VAGC with r band apparent magnitude limit of 17.6. These two samples are located in the same angular region.

In this work we reproduce and compare our analysis with data from the Las-Damas simulations (McBride et al., 2011), and for this reason we select comparable datasets. Our first population is at low redshift, between 0.05 and 0.106. The low redshift limit ensures that the sample is not greatly affected by the peculiar velocities, and excludes the Coma and Virgo clusters. We take as well a second sample with redshifts between 0.075 and 0.165. To determine a suitable absolute magnitude cut, we define a faint limit M_r where the limiting magnitude has been reached to ensure luminosity completness. In this way, we can assure a similar comoving number density of galaxies within the redshift limits. These limits are $M_r < -20$ for the first population and $M_r < -21$ for the second population. The samples are summarized in Table 2.1.

TABLE 2.1: SDSS selected samples

Sample	Redshift	Magnitude $M - 5\log(h)$	Density $\bar{n}$ h^3 Mpc^{-3}	Galaxies
Pop1	0.05 – 0.106	$M_r < -20$	5.7×10^{-3}	113483
Pop2	0.075 – 0.165	$M_r < -21$	1.04×10^{-3}	76688

Regarding the cosmological parameters used, we work with $\Omega_m = 0.25$, $\Omega_k = 0.0$, $\Omega_\Lambda = 0.75$ and $H_0 = 100h$ km s^{-1}Mpc^{-1}. We calculate comoving distances integrating the Friedmann eq. 1.15.

2.3.2 LasDamas simulations

As we will see in section 2.4, our error estimation method might be affected by the selection procedure explained in section 2.2. In order to test any possible systematic effects, we will make use of the multiple realizations of the Large Suite of Dark Matter Simulations (LasDamas) McBride et al. (2011), Swanson (Accessed: 2016-03-25), a project that ran a large suite of cosmological N-body simulations that follow the evolution of dark matter in the universe. Results provide us with an adequate resolution in many large boxes, rather than a single realization at high resolution. The enormous volume of generated data is appropriate for statistical studies of the distribution of galaxies and halos. The LasDamas simulations are designed to model the clustering of Sloan Digital Sky Survey (SDSS) galaxies in a wide luminosity range, with the goal of assisting in the modeling of galaxy clustering measurements. Specifically, the simulations are used to construct detailed mock galaxy catalogs by placing artificial galaxies inside dark matter halos using the Halo Occupation Distribution (HOD) with parameters fitted to reproduce the galaxy number density and projected correlation function of the respective SDSS galaxy samples. The HOD describes the distribution of galaxies within the dark matter halos. It uses three related properties of the halo model: the probability distribution relating the mass of a dark matter halo to the number of galaxies that form within that halo, the distribution in space of galactic matter within a dark matter halo and the distribution of velocities of galaxies relative to the dark matter within the halo. LasDamas is also designed to reproduce the SDSS

DR7 geometry. Alltogether, we can see that LasDamas simulations are specially adequate for NYU-VAGC comparison.

Simulations follow the same cosmological model assumed in this work, with $\Omega_m = 0.25$, $\Omega_\Lambda = 0.75$ and $H_0/100 = 0.7$ km $s^{-1}Mpc^{-1}$ as cosmological parameters. Data used consist on two different sets of mocks with with different properties, matched to our SDSS samples. In each case, we use 20 mock realizations. One can see them summarized on Table 2.2. Simulation *Esmeralda* is started at an initial redshift of $z = 99$; while *Carmen* is started at $z = 49$. Mocks are generated with the same initial power spectrum, but a different random seed.

TABLE 2.2: LasDamas selected mock catalogs

Parent Simulation	Redshift	Magnitude $M - 5\log(h)$	Density h^3 Mpc^{-3}	Number of particles
Esmeralda	0.05 – 0.106	$M_r < -20$	6.01×10^{-3}	121838.9
Carmen	0.075 – 0.165	$M_r < -21$	1.11×10^{-3}	81733.8

Densities and number of particles are shown for the average of the used LasDamas samples.

As explained in section 2.4, we use LasDamas simulations not only for the estimation of error bars in the CiC distribution, but in addition, we test the jackknife estimation method against LasDamas sample-to-sample variations. This will allow us to extend these results to the jackknife analysis performed with SDSS and test the reliability of its error estimations.

2.4 Estimation of the errors: the jackknife method

The Jackknife resampling method, together with the Bootstrap method (Efron & Tibshirani, 1994), allows us to estimate the errors in multiple statistics. It can be used when we make use of summary statistics, like the Counts-in-Cells or the correlation function. With these statistics, we perform an analysis over an entire population, shrinking its information into a single quantity, such as the first

and second characteristics. These methods are meant to be used when the limited volume of data in use does not allow us to compare our calculations among different populations, and therefore, we do not have a direct way to obtain its variation. As we only have one single galaxy population for each sample, we have to apply an internal error estimate method over the entire population.

We chose to apply the 'delete one jackknife' method (Norberg et al., 2009) over our populations. Our data consist on coordinates in a certain space contained by a window. Dividing this window into different regions we generate N_{sub} subvolumes of different disjoint areas of the sky with the same redshift limits. Each subvolume contains a subsample of cells and deleting one of these subsamples from the parent population we produce a new sample. Cells are included in the corresponding subvolume using their center coordinates, no further calculation is made regarding their volume overlapping with neighbor subsamples. These resampled data shares its cells and galaxies with the original one but lacks one of the subsamples. We can repeat this procedure N_{sub} times by systematically omitting, in turn, each of the subsamples in which the data has been split. The resampling of the data set consists on $N_{sub} - 1$ remaining subsamples, with volume $\nu(W) - \nu(V_i)$, where V_i is the volume occupied by the ith subvolume.

Our parent population of cells C occupies an irregular area of the sky due to the rejection process explained in section 2.2, creating a characteristic footprint (see Fig. 2.1). In order to ensure that all jackknife samples contain the same amount of cells, subsamples are selected in the following way: we sort our cells by declination and create 7 (or other desired number) subgroups of equal amount, in such a way that the groups are also sorted by declination. Then, we perform the same division for each subgroup with the right ascension, sorting each of them and dividing into another 7 subgroups, 49 in total. Now we have subgroups of cells located in differently shaped rectangles but containing the same amount of cells. The appearance of the resulting subsamples can be seen in Fig. 2.1.

Let f be the Counts-in-Cells distribution of our parent population. Once we have generated the new subsamples from the original population, we proceed to calculate the covariance matrix as

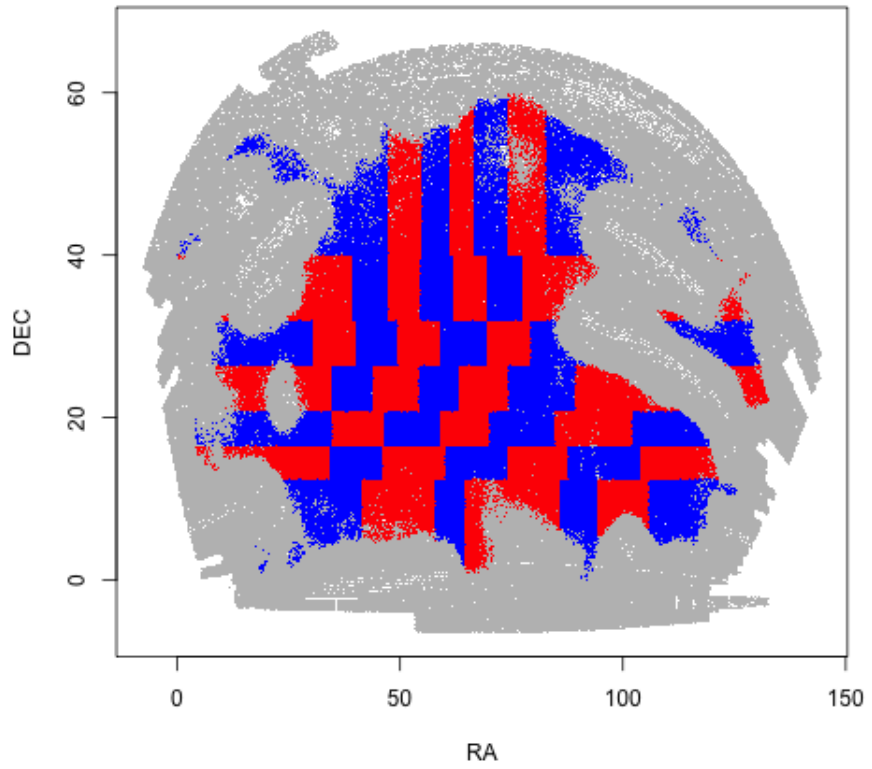

FIGURE 2.1: Footprint of our accepted cells centers for $r = 24h^{-1}$ in Pop1. Alternating red and blue colors, the different jackknife subsamples used for the error estimation. Footprint of the SDSS mask in grey. Right ascension and declination (in degrees) are relocated for calculation.

$$\Sigma_{ij} = \frac{N_{sub} - 1}{N_{sub}} \sum_{k=1}^{N_{sub}} (f^k(i) - \bar{f}(i))(f^k(j) - \bar{f}(j)) \tag{2.4}$$

where N_{sub} is the number of used subsamples, f^k is the Counts-in-Cells distribution when we omit subsample k, and $\bar{f}$ is the mean of distributions f^k for every subsample $N_{sub} - k$. We evaluate these functions in the values i and j corresponding to all the considered values of N, this is, the values of the CiC frequency histogram, from 0 to the maximum number of galaxies contained into a cell. We finally calculate our standard deviations as $\sigma_i = \sqrt{\Sigma_{ii}}$.

Once this analysis is done, it is recommendable to test the error estimation of the Jackknife method. To do so, we can use the numerous realizations the LasDamas simulations provides us. We apply the 'delete one jackknife' method over one

of the LasDamas realizations and compare the obtained σ_i with the standard deviations of many different LasDamas realizations. To generate this standard deviations we perform the Counts-in-Cells process described in section 2.2 over 20 realizations. When we obtain the CiC distribution over 20 different realizations, we have enough data to calculate the mean and variance of this distribution. We compute the mean and the variance with

$$\begin{aligned} \bar{x}^i_{20} &= \frac{1}{20}\sum_{j=1}^{20} f^i_V(j) \\ S^i_{20} &= \sqrt{\frac{1}{20-1}\sum_{j=1}^{20}(f^i_V(j) - \bar{x}^i_{20}(j))^2} \end{aligned} \tag{2.5}$$

where $f^i_V(j)$ is the ith realization distribution evaluated at $N = j$.

Now, we can compare the jackknife error estimates with the standard deviations of the LasDamas simulations and check if jackknife errors are over- or underestimated (Fig. 2.2).

After these results, we see that the jackknife errors for the SDSS samples are typically larger than the errors obtained from the sample-to-sample variances in the catalogs. This indicates that the same systematic error should be found for the SDSS samples, with jackknife estimated error bars overestimating the real uncertainty. This proves that the jackknife estimations are unreliable in strong mask conditions, even after our cell rejection process. Since we assume equation 2.5 as a more robust method of variance estimation than the Jackknife method, we will make use of the variances estimated from the LasDamas simulations when fitting our probability functions.

2.5 Best fit to $f_V(N)$

The Counts-in-Cells (CiC) methodologies introduced in section 2.2 have been applied over the data presented in sections 2.3.1 and 2.3.2. As previously explained

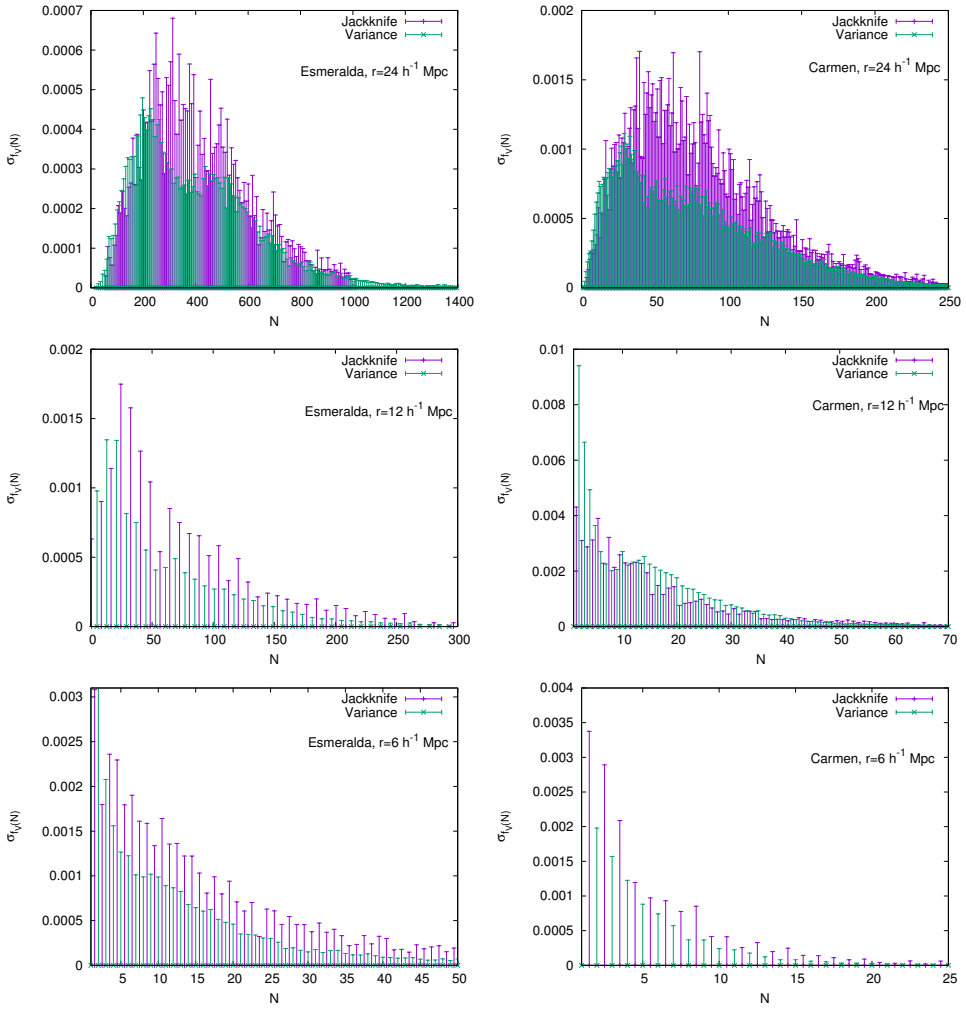

FIGURE 2.2: Comparison between 'delete one Jackknife' method error for $f_V(N)$ (pink) and its population variance (green) in LasDamas simulation. X axis is number of galaxies per cell, and Y axis is the normalized found variance. Left: Esmeralda mocks, from top to bottom, radii 24, 12 and 6 h^{-1} Mpc. Right: Carmen mocks, from top to bottom, cell radii 24, 12 and 6 h^{-1} Mpc.

this creates an observed probability density function that we will attempt to fit to different theoretical distributions. Some of these functions have been commonly used in the cosmological literature to fit the CiC distribution, and some are an original contribution of this work. It is therefore of high interest to discriminate the best fitting model and their best fitting parameter values. Despite the information contained in a CiC distribution does not fully characterize the galaxy distribution, and it can be considered as a constraint for models or N-body simulations, if we expect them to be reliable. Now we present the used probability density functions.

2.5.1 Gravitational Quasi-Equilibrium Distribution

The GQED (Saslaw & Hamilton, 1984) is derived from a thermodynamical description of the galaxy fluid, and it can be also derived from the statistical mechanics (Ahmad et al., 2002). One of its principal assumptions is to accept that the galaxy accretion evolves through a sequence of quasi-equilibrium states, a basic condition to start the thermodynamical approximation. This is possible if we assume an infinite quantity of galaxies in the universe. With this assumption we say clusters and large structures remain stable along large cosmological times before changing into a new quasi-equilibrium state.

We will assume as well that galaxies are formed without dynamical interactions with the exterior and then they interact gravitationally as punctual masses with the rest of the universe. As the clusters dimensions are smaller than the curvature radius of the universe and the velocities involved are much lower than the speed of light, we can assume Newtonian gravity with potential $\phi = r^{-1}$.

After these assumptions, the GQED can be derived as the pdf

$$f_V(N) = \frac{\bar{N}(1-b)}{N!}[\bar{N}(1-b)+Nb]^{N-1}e^{-[\bar{N}(1-b)+Nb]} \tag{2.6}$$

which gives us the probability of finding N galaxies in a volume V where the expected value of N is the product of the cell volume by the mean density of the galaxies, $\bar{N} = \bar{n}V$. This is the so-called Gravitational Quasi-Equilibrium Distribution (GQED). Details in the deduction and properties of this distribution can be fully consulted in Saslaw & Hamilton (1984), Sheth & Saslaw (1996) and Sheth (1995).

Parameter b allows us to study both physical and statistical properties of this distribution. For the limit $b = 0$ we have a Poisson distribution, where galaxies are uniformly distributed. For big scales, where fluctuations are small, the density function becomes Gaussian. As b value can be determined from physical magnitudes of the galaxies, we have a free parameter model of the galaxy distribution. As we find in Ahmad et al. (2002), b represents a measure of the state of aggregation and we can express it as

$$b = \frac{3/2(Gm^2)^3\bar{n}T^{-3}}{1 + 3/2(Gm^2)^3\bar{n}T^{-3}} \tag{2.7}$$

This expression relates b to galaxy mass m, mean density $\bar{n}$, galaxies kinetic temperature T and gravitational constant G. Dependence of b on V can be obtained empirically from the variance of the number of counts in a cell of volum V. We use our density function to calculate this variance:

$$\langle (\Delta N)^2_V \rangle = \frac{\bar{N}}{(1 - b(V))^2} \tag{2.8}$$

This result allows us to describe aggregation of galaxies and, given one more step, relate b with the volume integral of the two point correlation function

$$b = 1 - (\bar{N}\bar{\xi}_2(V) + 1)^{-1/2} \tag{2.9}$$

This is how b depends on the correlation function $\bar{\xi}_2(V)$ (see section 3.2), which is defined as

$$\bar{\xi}_2(V) = \frac{1}{V^2}\int_V \xi_2(\mathbf{r}_1, \mathbf{r}_2) d^3\mathbf{r}_1 d^3\mathbf{r}_2 \tag{2.10}$$

However, in the fitting process, we will use this model as a two free parameter function over $\bar{N}$ and b.

2.5.2 Negative Binomial Distribution

The Negative Binomial Distribution (NBD) was firstly proposed in the cosmological context by Carruthers & Duong-van (1983) and derived later by Elizalde & Gaztanaga (1992). With this model we study the probability of distributing N galaxies in m disconnected boxes. The probability of finding a galaxy in one of these boxes is proportional to the number of galaxies already located in the box. Expressing the probability function in terms of galaxies per cell instead of its conventional form, we have:

$$f_V(N) = \frac{\Gamma(N + \frac{1}{g})}{\Gamma(\frac{1}{g})N!} \frac{\bar{N}^N (\frac{1}{g})^{\frac{1}{g}}}{(\bar{N} + \frac{1}{g})^{N+\frac{1}{g}}} \tag{2.11}$$

where Γ is the gamma function and $\bar{N} = \bar{n}V$, the expectation value of N. Similarly as we had with the value b for the GQED, g is also a aggregation parameter. We can obtain it theoretically for the NBD with

$$g = \bar{\xi}_2(V) = \frac{\langle (\Delta N)^2 \rangle - \bar{N}}{\bar{N}^2}, \tag{2.12}$$

as we can see, g depends on the volume of the cell. $\bar{\xi}_2(V)$ is defined as in eq. 2.10.

This function is widely used in statistics and was firstly introduced in the cosmological context as a phenomenological model without physical explanation, which, in addition, is found to violate the second law of thermodynamics (Saslaw & Fang, 1996, Yang & Saslaw, 2011). However, it provides a fair agreement with the observational distribution and it is thought to be related with the hierarchical universes properties. A deeper study of the implications of using such distribution suggest that the NBD assumes galaxies to be formed where there is already a galaxy cluster. Hence, this model does not take infall into account, but can describe the case where galaxies form from the merger of less massive objects. As before, this function will be fitted over parameters $\bar{N}$ and g.

2.5.3 Log Normal Distribution with bias

The Log Normal Distribution (LND), was firstly used by Hubble (1934) but it was not formally proposed until Coles & Jones (1991). It was one of the first fully described stochastic models used to model the distribution of matter density. This model allows us to calculate many complex properties of the CiC distribution. Despite being probably the most commonly used function for modeling the CiC distribution of galaxies, it has proved to be insufficient, and we will fit instead a modification of this function by a galaxy bias parameter (Arnalte-Mur et al., 2016, Dekel & Lahav, 1999).

First, we will introduce the Log Normal distribution. We will start with a Gaussian random field $X(r)$, where we will study the probability of finding a certain density

from our field in a given position. Such field will have a one-point probability density function $f_1(x)$ given by a normal distribution $X \sim N(\mu, \sigma^2)$:

$$f_1(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-(x-\mu)^2/2\sigma^2\right) \tag{2.13}$$

Furthermore, all the higher order n-point pdf's, $f_n(\mathbf{x})$, of field values at different positions r_i are multivariate normal:

$$f_n(\mathbf{x}) = (2\pi)^{-n/2}|\mathbf{M}|^{-1/2} \exp\Big(-\frac{1}{2}\sum_{i,j}\mathbf{M}^{-1}_{i,j}x_i x_j\Big) \tag{2.14}$$

where $\mathbf{x} = (x_1, \ldots, x_n)$, $x_i = x(r_i)$ and $\mathbf{M}$ is the covariance matrix $\mathbf{M}_{ij} = \langle (X_i - \mu)(X_j - \mu)\rangle$ where $M_{ii} = \sigma^2$. In addition, we assume the field is statistically homogeneous, i.e., $\mu_i = \mu$. $\mathbf{M}_{ij}$ is determined by the covariance function, $\Xi(r)$, which depends only on $r_{ij} = |r_i - r_j|$ if the field is statistically isotropic. Thus

$$\Xi(r_{ij}) = \{[X(r_i) - \mu][X(r_j) - \mu]\} = \mathbf{M}_{ij} \tag{2.15}$$

Functions $\Xi(r)$ and f_n allows us to specify all the finite dimensional pdf's for a Gaussian random field and obtain exact solutions for many of the properties of such fields. Now, we will construct a non-Gaussian field by applying a non-linear transformation of Gaussian fields:

$$Y(\mathbf{r}) = \exp\left(X(\mathbf{r})\right) \tag{2.16}$$

where the new one-point probability function in the field Y is

$$f_1(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{(\log y - \mu)^2}{2\sigma^2}\Big)\frac{1}{y} \tag{2.17}$$

This is the Log Normal variable $Y \sim \Lambda(\mu, \sigma^2)$ pdf, with the same parameters than the original Gaussian variable. Thus, in the multivariate case we have the pdf

$$f_n(y_1,\ldots,y_n) = (2\pi)^{-n/2}|\mathbf{M}|^{-1/2} \times \exp\left(-\frac{1}{2}\sum_{i,j}\mathbf{M}_{ij}^{-1}\log(y_i)\log(y_j)\right)\prod_{i=1}^{n}\frac{1}{y_i} \tag{2.18}$$

where $\mathbf{M}$ is the covariance matrix of the variable X.

As the Gaussian distribution provides us with a valid description of linear and weak density perturbation fields, the Log Normal distribution represents the same case for the non-linear case and is one of the few non-Gaussian random fields for which interesting properties are calculable analytically. In Coles & Jones (1991) one may find further motivation on this model, such as agreements with observational data. This has turned the Log Normal distribution into one of the most well known and widely applied models (Kayo et al., 2001, Kitaura et al., 2010, Wild et al., 2005).

However, we will modify this distribution in order to improve its efficiency as a galaxy distribution descriptor. The introduction of the bias parameter into the Log Normal distribution has proved useful (Arnalte-Mur et al., 2016). This method relies on the assumption that the dark matter density fluctuations follow a local non-linear transformation of the initial energy density perturbation. Therefore, we model the galaxy distribution as a Cox model with the galaxy density field $\Delta = N/\bar{N}$, where $\bar{N}$ is the mean of the CiC distribution.

That will make the Log Normal distribution a good candidate to model this field, but it should be conveniently modified by the galaxy bias in order to model the galaxy distribution. We expect however difficulties for small N or r, where shot noise can be more relevant. In this case, we assume the simplest case of a scale-independent and linear bias, so $\delta_g = b\delta$, where δ_g and δ are the contrasts of the galaxy number density and the matter density, respectively. Introducing this bias in the expression 2.17 for the matter density, we obtain (Arnalte-Mur et al., 2016)

$$f(\Delta) = \frac{1}{\sqrt{2\pi H_0}}\frac{\exp\left(-\frac{1}{2}\frac{y^2}{H_0}\right)}{\Delta + b - 1} \tag{2.19}$$

where

$$\begin{aligned} H_0 &= \ln(1+C) \\ y &= \ln\left(\frac{\sqrt{1+C}}{b}(\Delta + b - 1)\right) \end{aligned} \tag{2.20}$$

and C is the variance in the matter distribution, which we expect to be roughly $C = \sigma_8^2$ for cell radius $r = 8h^{-1}$ Mpc at $z = 0$. After normalizing N into Δ, the free parameters of this distribution are C and the bias b.

2.5.4 Weibull Distribution

The Weibull distribution was originally proposed by Weibull (1951) in the description of compound bodies where an event over one of its parts should be considered an event over the entire body. Complementarily, we can say that no event had happened over the body if, and only if, it has not happened over any of its parts. The example used by professor Weibull in his paper in 1951 consisted in the probability of breaking a chain. We only need one single broken link to say that the chain is broken. This distribution is also thought to model an evolving process where series of units are aggregated into a bigger body over time. Weibull distribution was found to succesfully fit the size distribution of certain particles (Rosin & Rammler, 1933).

Weibull distribution have been successful at describing growth models which are based on natural deterministic growth models and get their random nature by random stopping times (Ghorbani et al., 2006). Without deeper physical motivations for this distribution we proceed as with the Negative Binomial Distribution and include it in our fitting analysis.

We can follow the idea of professor Weibull in order to deduce the expression of this distribution. If p_i is the probability of the ith galaxy to be the last one to be considered a member of the cluster, and then, to complete it, and N galaxies have been accreted before the cluster is complete, we call P the probability of having the cluster complete when the nth galaxy is accreted. P can be calculated as

$$(1-P) = \prod_{i=1}^{N}(1-p_i) \tag{2.21}$$

For a Weibull distribution, we consider $\forall i,j : p_i = p_j$ and p_i constant in time, then

$$(1-P) = (1-p)^N \tag{2.22}$$

where, $1-P$ is the probability of having a non complete cluster after N galaxy accretions and $1-p$ is the probability of non completing the cluster with a single accretion.

Weibull propose an generalized exponential distribution function for p random variable, where events (galaxy arrivals to the cluster) are independent in time:

$$F(x) = 1 - e^{-\phi(x)} \tag{2.23}$$

for a certain $\phi(x)$ function. If we consider $p = 1 - e^{-\phi(x)}$ then, $P = 1 - e^{-N\phi(x)}$. Now we have to specify the function $\phi(x)$. The only necessary general condition this function has to satisfy is to be a positive, nondecreasing function, vanishing at a certain value. The most simple function satisfying this condition is

$$\phi(x) = (\frac{x-\theta}{\lambda})^k \tag{2.24}$$

thus,

$$F(x) = 1 - e^{-(\frac{x-\theta}{\lambda})^k} \tag{2.25}$$

And, differentiating, the Weibull probability density function is

$$f(x) = \frac{k}{\lambda}(\frac{x-\theta}{\lambda})^{k-1} e^{-(\frac{x-\theta}{\lambda})^k} \tag{2.26}$$

Parameter $\lambda > 0$ is the scale parameter and in our case it is related with the number of galaxies per cell, the galaxy density. Parameter $k > 0$ is the shape parameter and determines the pattern of the distribution. k is dimensionless, related with the shape, not the size, which make the Weibull distribution interesting for describing galaxy large scale structures, due to its self-similar morphology. As we can only find nonnegative numbers of galaxies per cells, we will take θ as 0. The use of this function in cosmology is an original contribution of this work.

2.6 Results

Calculation tasks of Counts-in-Cells involve long times and heavy calculations. To achieve an appropriate reliability in our data, it is necessary to use at least as many cells as galaxies in our populations. In addition, nearly 50% of cells are rejected by mask effects, so we have to double the number of tested cells. In Table 2.3 one can find the numbers of used cells for each SDSS sample and LasDamas realizations.

In Fig. 2.3 we show the Counts-in-Cells probability density function $f_V(N)$ for populations 1 and 2 in the three used radii: 6, 12 and $24h^{-1}$ Mpc. The curves show the expected probabilities for these samples and radii (Yang & Saslaw, 2011), with higher values of $\hat{N}$ (the average of the CiC distribution) for bigger cells. Used binning for galaxy abundance are the natural numbers with frequencies centered between the bins.

The CiC distribution of cells with radius $24h^{-1}$ Mpc show no void function, with all of its cells occupied by more than one galaxy. The opposite is found for the minimum cell radius considered, $6h^{-1}$ Mpc, where the CiC distribution is dominated by the void function. We may consider an intermediate case. From a differential point of view, we can think in a pdf starting with $f_V(0) > 0$ and $\mathrm{d}f_V(N)/\mathrm{d}N = 0$ for $N = 0$. The required cell size to obtain such distribution is an indicative of the maximum void size. We seem to be close to it in Pop2 with cell radius $12h^{-1}$ Mpc.

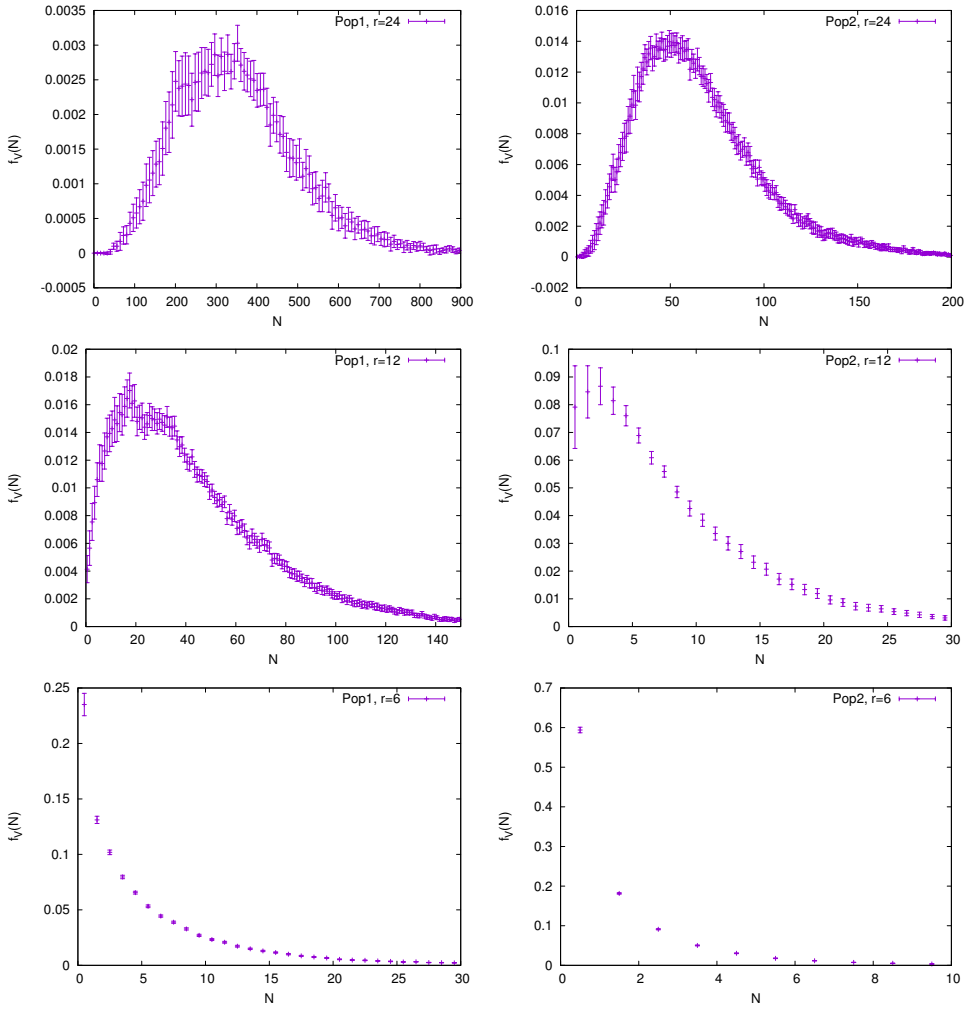

FIGURE 2.3: Counts-in-cells probability distribution of populations 1 and 2 (left to right) for cells of radii 24, 12 and 6 h^{-1} Mpc (top to bottom) with bins of width 1. Error bars obtained from variances of the corresponding LasDamas realizations.

2.6.1 Fitting the results to a distribution function

In this section we proceed to fit the probability distribution functions defined in section 2.5 to our Counts-in-Cells observed distributions. We calculate the χ^2 of the fit

$$\chi^2 = \sum_{N=0}^{N_{max}} \frac{(f_V^{obs}(N) - f_V(N,\theta))^2}{\sigma_N^2} \tag{2.27}$$

between the observed distribution and the theoretical distribution as a qualitative measure of goodness of fit where N_{max} is the largest number of galaxies in a cell. The fitting will make use of the diagonal values σ^2, ignoring correlations between bins. The vector θ contains the parameters of the theoretical distribution. Values with $f_V(N) = 0$ are excluded from the fitting.

We summarize the found θ and χ^2 values in Table 2.3. For samples Pop1 and Pop2 and radii 6, 12 and $24h^{-1}$ Mpc, we present the number of used cells and the distributions best fit parameters with 1σ error bars obtained from the χ^2 distribution. The estimation of the χ^2 distribution neglecting correlations between different bins of the observed pdf might have underestimated the obtained best fit χ^2 values and the parameters 1σ error bars.

TABLE 2.3: Counts-in-Cells best fit $f_V(N)$

Sample			GQED			NBD		
Population	Cells	r	$\bar{N}$	b	χ^2	$\bar{N}$	g	χ^2
Pop1	200167	24	361.7 ± 0.8	$0.8768\pm^{0.0017}_{0.0008}$	588.574	$357.8\pm^{2}_{1.4}$	$0.1785\pm^{0.002}_{0.0014}$	278.181
	151661	12	$45.8\pm^{0.8}_{0.7}$	0.8171 ± 0.0018	623.136	$44.87\pm^{0.2}_{0.1}$	$0.611\pm^{0.005}_{0.007}$	501.702
	191365	6	$5.72\pm^{0.19}_{0.09}$	$0.708\pm^{0.005}_{0.009}$	364.385	$5.64\pm^{0.05}_{0.09}$	$1.69\pm^{0.06}_{0.03}$	400.83
Pop2	134136	24	$65.8\pm^{0.4}_{0.3}$	0.7588 ± 0.003	141.303	$65\pm^{0.14}_{0.3}$	0.249 ± 0.003	141.743
	116718	12	$8.18\pm^{0.3}_{0.18}$	$0.647\pm^{0.004}_{0.009}$	31.6512	$8.18\pm^{0.11}_{0.13}$	$0.83\pm^{0.03}_{0.02}$	28.1752
	203882	6	$0.991\pm^{0.011}_{0.009}$	$0.472\pm^{0.002}_{0.003}$	83.6695	1.018 ± 0.009	$2.37\pm^{0.03}_{0.04}$	43.5573
			Weibull			Log Normal + bias		
			k	λ	χ^2	C	b	χ^2
Pop1	200167	24	$2.74\pm^{0.05}_{0.06}$	$397\pm^{4}_{5}$	973.033	0.083 ± 0.003	1.48 ± 0.03	274.117
	151661	12	$1.28\pm^{0.02}_{0.03}$	$48.3\pm^{0.8}_{1}$	740.859	$0.375\pm^{0.013}_{0.012}$	$1.35\pm^{0.03}_{0.02}$	296.922
	191365	6	0.769 ± 0.01	$5.22\pm^{0.13}_{0.1}$	211.718	1.12 ± 0.015	1.362 ± 0.008	329.396
Pop2	134136	24	$2.21\pm^{0.06}_{0.07}$	71.6 ± 1.9	929.812	$0.171\pm^{0.006}_{0.004}$	$1.25\pm^{0.04}_{0.02}$	122.3
	116718	12	$1.08\pm^{0.017}_{0.02}$	$8.9\pm^{0.18}_{0.2}$	39.6672	$0.436\pm^{0.015}_{0.014}$	1.48 ± 0.05	23.1451
	203882	6	0.752 ± 0.003	1.177 ± 0.009	347.899	1.27 ± 0.04	1.38 ± 0.017	1264.05

The number of used bins for the χ^2 fittings are: for Pop1, $N_{max} = 1046$, 357 and 114 for cell radii 24, 12 and $6h^{-1}$ Mpc respectively, and for Pop2, $N_{max} = 257$, 91 and 28 for cell radii 24, 12 and $6h^{-1}$ Mpc respectively.

We use the χ^2 values to discriminate the best fitting distribution. These values can be compared with results in Figs. 2.4 to 2.9. Top box shows the observational and best fit distributions. In the bottom box we can see the residuals, the difference between the best-fit distributions and the observed one (X axis). As expected, curves outside the error bars have higher values for χ^2.

For radii 12 and 24 h^{-1} Mpc, we appreciate that the Log Normal distribution with bias perform the best fit, with lower χ^2 values and residuals inside the error bars, except for small N. The results found for the NBD are close in goodness of fit, although always slightly away from observations. For $r = 6h^{-1}$ Mpc, the Log Normal is unable to perform the best fit, being overcome by the other distributions. As we mentioned in section 2.5.3, this is expected, since shot noise is not negligible. In the smallest radius, the Weibull distribution gives the best fit for sample Pop1 (lower redshift and magnitude), while we find it with the NBD for sample Pop2 (higher redshift and magnitude).

Regarding the found parameters, the fitted $\bar{N}$ are all close to the expectancies. For comparison, we include these values in Table 2.4 left.

TABLE 2.4: Expectancies

Population	$\bar{N}$			$\hat{N}$		
	24	12	6	24	12	6
Pop1	363.36	45.42	5.68	354.75	45.44	5.83
Pop2	66.83	8.35	1.04	65.75	8.4	1.05

Expectancies found for the samples and cells radii. $\bar{N} = \bar{n}V$: number density of the samples times the volume of the cell. $\hat{N}$: mean of the Counts-in-Cells distribution.

Parameter b in the GQED is related with the correlation function as defined in eq. 2.10, in such a way that b is smaller for higher amplitudes of $\bar{\xi}_2(V)$. A monotonic trend is found between b and r, indicating stronger correlations inside bigger cells, which contain more structure. For the NBD we have $g = \bar{\xi}_2(V)$, and therefore the opposite trend is found.

As the NBD, the Weibull distribution was used without physical motivation, but merely as a common distribution for modeling growth processes. However, it seems that the growing process of galaxy clusters does not follow a Weibull distribution. An exception is found for Pop1 and cell radius $6h^{-1}$ Mpc, where we have the best fit, although the distribution is more than 1σ away from the observed distribution for a significant part of the curve.

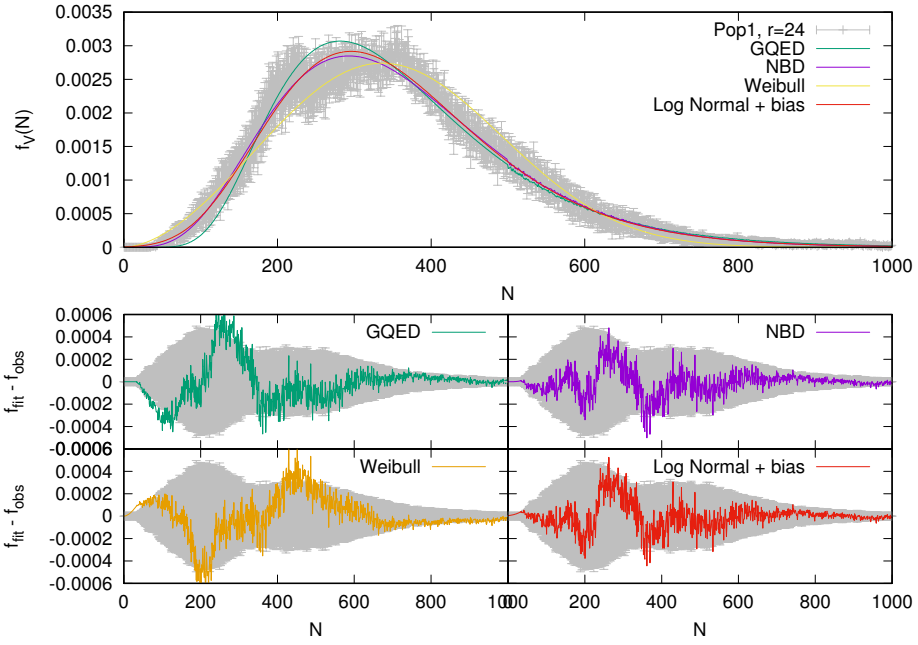

FIGURE 2.4: Counts-in-cells results for population 1 and radius $24h^{-1}$ Mpc. Top: $f_V(N)$ CiC distribution function with best fit models and error bars. Bottom: probability distribution residuals, model minus observations.

The Log Normal distribution has been normalized by the CiC distribution mean before being fitted. These quantities are summarized in Table 2.4, right. Parameter C, the variance of the matter distribution strongly varies with r. This quantity is roughly related with the cosmological parameter σ_8, which at $z = 0$ is measured $\sigma_8 = 0.828 \pm 0.012$ (Planck Collaboration et al., 2014). Therefore, for the case of cells of radius r=8 Mpc/h, we would expect the value of C to be close to $C_8 = \sigma_8^2 = 0.686$. Despite the strong variations of C, we can see its monotonic evolution with the cell radius with values in agreement with σ_8. The same cannot be said for the bias parameter b, with values higher than expected for Pop1, where for the lower magnitude limit, smaller bias was expected.

2.7 Analysis of the results

In this first work of this thesis we proceeded with a blind fit of four different probability functions. A blind fit allows us to discriminate the best model of the observed Counts-in-Cells distribution. This work provides a very necessary tool

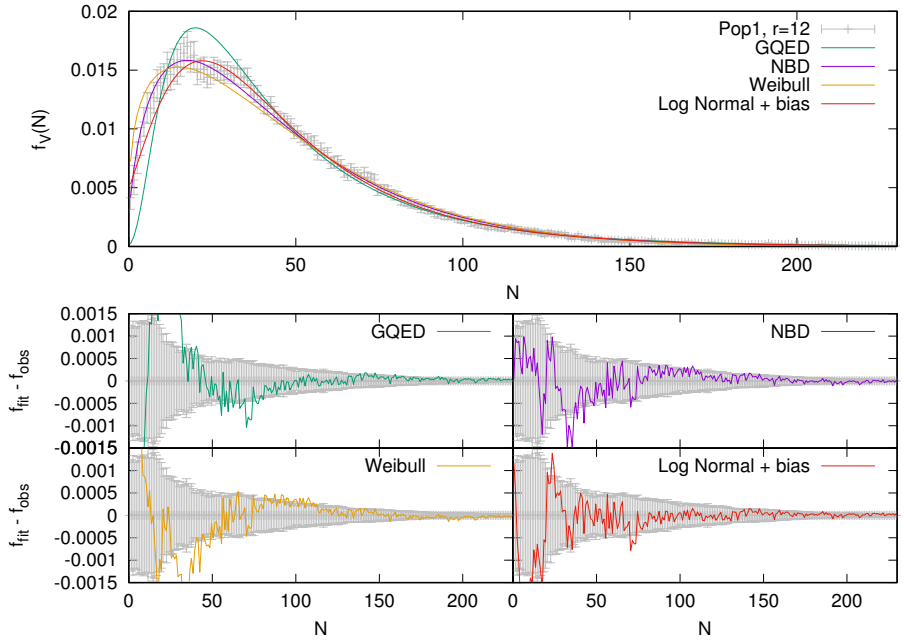

FIGURE 2.5: Counts-in-cells results for population 1 and radius $12h^{-1}$ Mpc. Top: $f_V(N)$ CiC distribution function with best fit models and error bars. Bottom: probability distribution residuals, model minus observations.

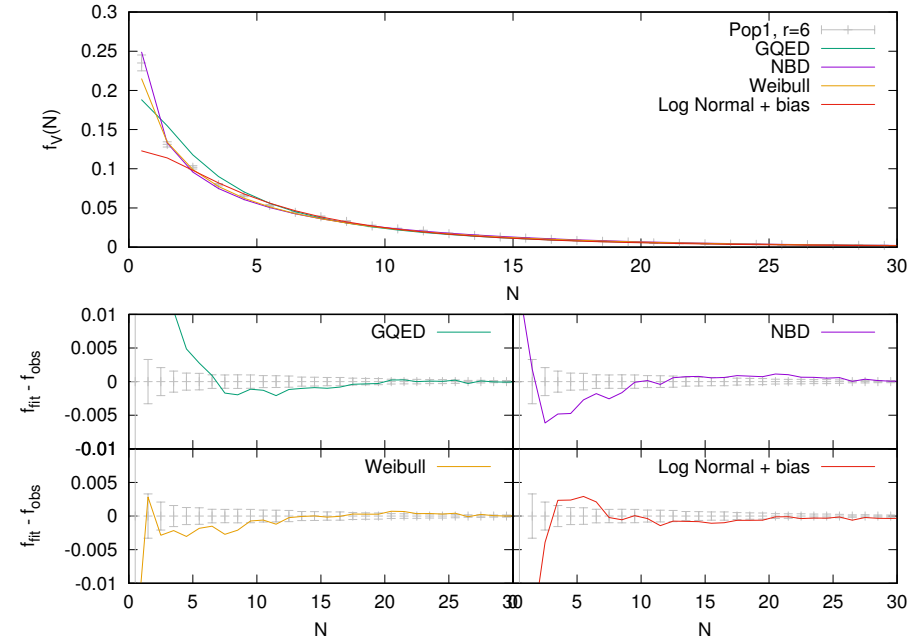

FIGURE 2.6: Counts-in-cells results for population 1 and radius $6h^{-1}$ Mpc. Top: $f_V(N)$ CiC distribution function with best fit models and error bars. Bottom: probability distribution residuals, model minus observations.

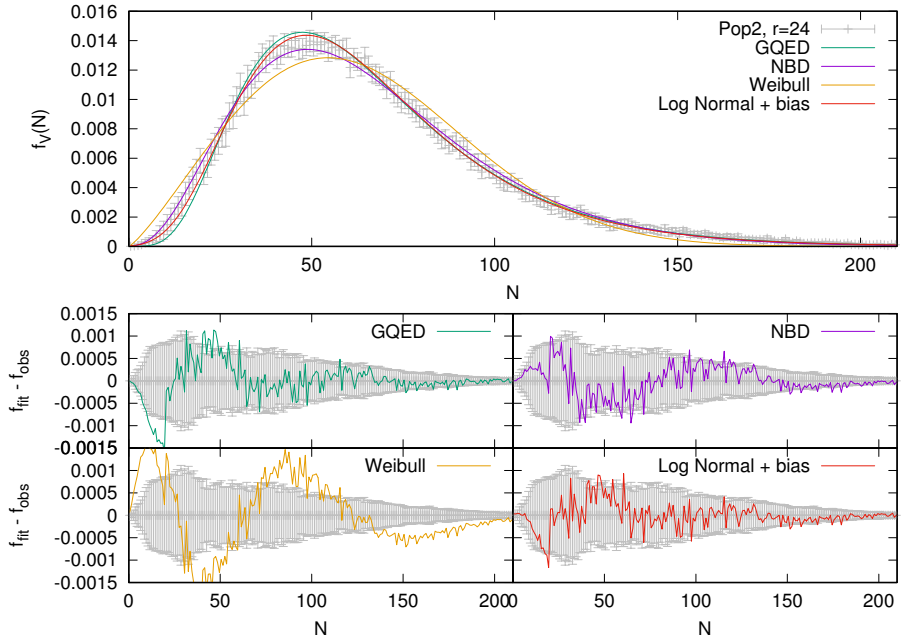

FIGURE 2.7: Counts-in-cells results for population 2 and radius $24h^{-1}$ Mpc. Top: $f_V(N)$ CiC distribution function with best fit models and error bars. Bottom: probability distribution residuals, model minus observations.

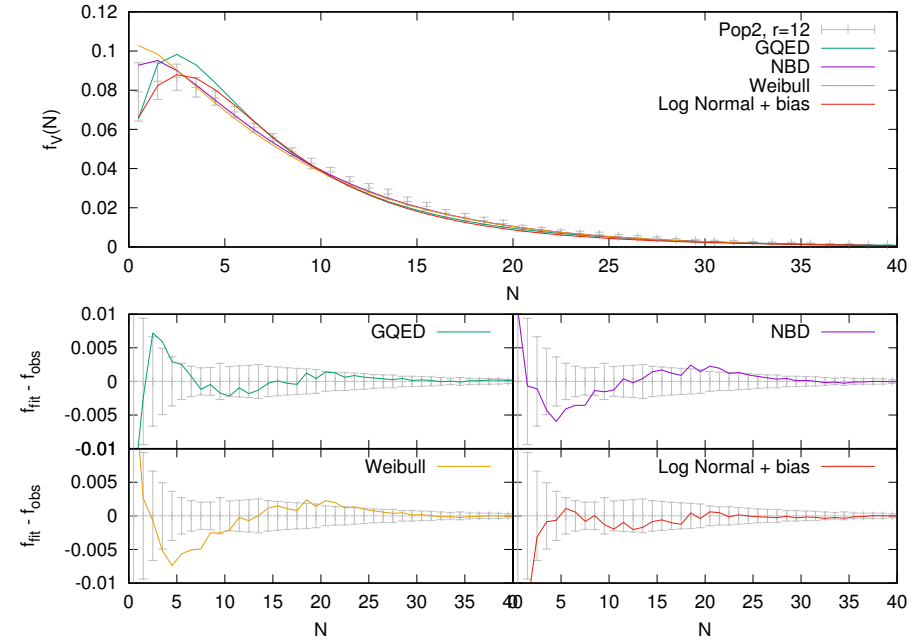

FIGURE 2.8: Counts-in-cells results for population 2 and radius $12h^{-1}$ Mpc. Top: $f_V(N)$ CiC distribution function with best fit models and error bars. Bottom: probability distribution residuals, model minus observations..

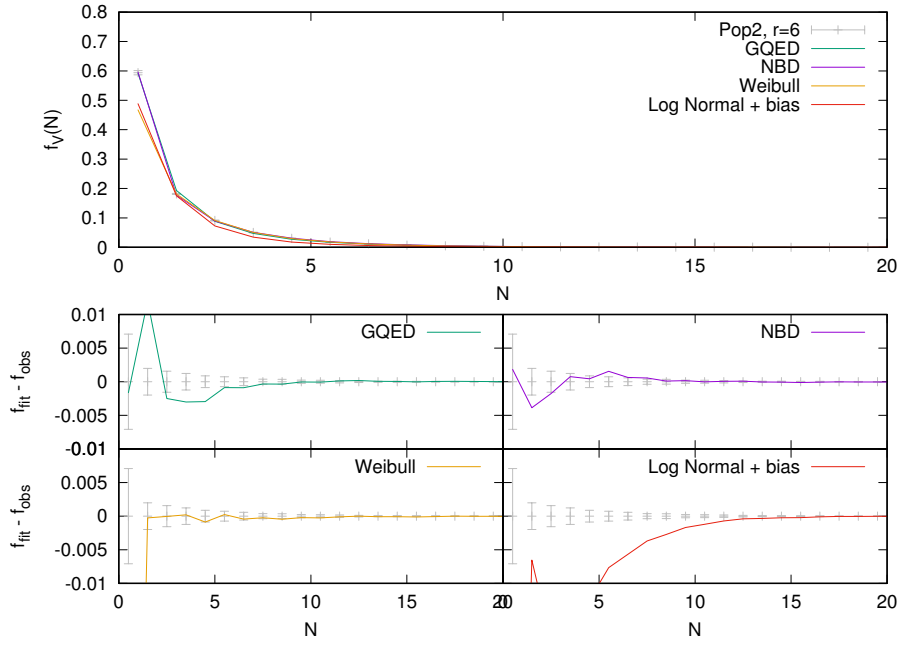

FIGURE 2.9: Counts-in-cells results for population 2 and radius $6h^{-1}$ Mpc. Top: $f_V(N)$ CiC distribution function with best fit models and error bars. Bottom: probability distribution residuals, model minus observations.

for many different applications in modern cosmology, such as the generation of galaxy mocks or the testing of N-body simulations.

Regarding the χ^2 results, we can state that the Log Normal distribution with bias is the best model for radii 12 and $24h^{-1}$ Mpc. This probability distribution has three free parameters, despite we have normalized the CiC distribution to its expectancy before fitting, reducing the model to two fitted parameters. This could explain its better performance, something that we will try to analyze in detail in a future work with the help of hypothesis tests. The estimation of parameters C and b incorporates an interesting asset and allows its direct comparison with other methods of bias estimation(López-Sanjuan et al. (2015), section 3.4.3.4). Nevertheless, we might have obtained results for the bias parameter in tension with other works (Meneux et al., 2006, Zehavi et al., 2011).

Regarding the GQED and the Negative Binomial distributions, we have obtained similar results to Yang & Saslaw (2011), with generally good fittings (inside the 1σ values). The Negative Binomial distribution obtains smaller χ^2 values than the GQED despite its non physical motivation. We must take into account here that,

nevertheless, this distribution is commonly understood as a descriptor of Counts-in-Cells events, independently of the nature behind the point process. This could explain the quality of the found fits.

The Weibull distribution is found to be the less capable of describing the CiC distribution, despite the best fit found for cells of radius $6h^{-1}$ Mpc in Population 1. A possible way of improving this distribution could be a modification of function 2.24. Remember that this function is chosen for being the simplest one satisfying the required conditions of a probability density function. With a physically motivated function we might obtain different results.

We have also fitted the calculated Counts-in-Cells distribution to the standard Log Normal distribution (eq 2.19 with $b = 1$), and the fits are systematically worse than the fits obtained including the bias parameter. This is obviously expected, since the inclusion of an extra parameter helps for fitting the counts, nevertheless, we conclude that this is an important and original result of this chapter: the best fit for the galaxy Counts-in-Cells distribution function is a Log Normal distribution modified by the inclusion of the bias term.

The results presented in this section will be published as Hurtado-Gil et al. (in prep.b).

Chapter 3

Correlation functions

3.1 Introduction

The correlation functions are among the so called 'second-order characteristics' and are one of the most widely used estimators for point process characterization. It has proved to be one of the most productive statistics in cosmology since it was firstly used (Peebles, 1980, Totsuji & Kihara, 1969). With every new published galaxy survey, new relevant conclusions of the galaxy distribution have been obtained with this statistic.

They are easy to perform while still very informative, and can provide relevant information about all scales in the population. This has made them a necessary estimator of the galaxy distribution for any galaxy survey. We have introduced the mathematical deduction of this estimator in section 1.2.5 to properly understand its power and situation in the point process theory.

We introduce the historical development of this function and the necessary corrections in astrophysics. In sections 3.2 and 3.3 we introduce the historical development of this function and the necessary corrections in astrophysics. These corrections are motivated by multiple uncertainties introduced in the estimation of the correlation function by both instrumental and physical reasons. The resulting is the projected correlation function, widely used in astrophysics for its reliability and its analysis capacity.

In section 3.4 we introduce the ALHAMBRA survey (Moles et al., 2008, Molino et al., 2014), a recently published galaxy survey from which we will obtain our conclusions of the galaxy distribution. Using the projected correlation function and additional corrections, we are able to clearly describe the nature of galaxy clustering in conditions never studied before with the ALHAMBRA detail. The high quality photometry of the ALHAMBRA survey makes it an optimal choice for the analysis of clustering in the small scales of galaxy distribution, while comparing with clustering from field galaxies. This analysis is completed with the spectral segregation of samples, which greatly helps our understanding of clusters nature. As it is well known, different galaxies cluster at different levels, and a segregated analysis could clarify the observed trends. With the ALHAMBRA spectral classification we can perform our analysis for 'quiescent' and 'star-forming' galaxies, and understand variations in clustering based on spectral types. In addition to this, we include the calculations of the galaxy bias (see section 1.1.3) for a large number of redshift intervals with segregation.

The present chapter is divided as follows. Sections 3.2 and 3.3 introduce the pair correlation function and how to use it, including several estimators, necessary corrections and uncertainties. Section 3.4 presents the ALHAMBRA survey (Moles et al., 2008, Molino et al., 2014), the samples used in this work, and analyse its suitability for this work. In section 3.4.3 we perform and analyze the calculations, including the galaxy dark-matter bias. Conclusions are summarized in section 3.4.4. Contents included in section 3.4 are published in Hurtado-Gil et al. (2016).

This chapter includes an introduction to a future work in section 3.5, related with the calculation of the pair correlation function with photometric surveys.

3.2 Definition

For the sake of clarity, we will reproduce the interpretation of the product densities defined in section 1.2.5 for order $k = 2$. In a process of intensity λ, the probability of finding a point in a infinitesimally small sphere $b(x)$ of volume $\mathrm{d}x$ centered at x is $\lambda(x)\mathrm{d}x$. If we consider now a second point y at distance r from x, we can calculate the probability that there is a point both in the spheres $b(x)$ and $b(y)$.

As seen, this is the second-order product density $\varrho^{(2)}(x,y)dxdy$. We can normalize these values by the intensities at locations x and y to obtain the *pair correlation function*

$$g(x,y) = \frac{\varrho^{(2)}(x,y)}{\varrho^{(1)}(x) \cdot \varrho^{(1)}(y)} = \frac{\varrho^{(2)}(x,y)}{\lambda(x) \cdot \lambda(y)} \tag{3.1}$$

In the homogeneous and isotropic case, this can be simplified since intensity is constant and our functions only depend on the distance r between x and y: $\varrho^{(2)}(x,y) = \varrho^{(2)}(r)$. The expression simplifies in

$$g(r) = \frac{\varrho^{(2)}(r)}{\lambda^2} \tag{3.2}$$

where *g(r)* is the isotropic and homogeneous pair correlation function, which is equally 1 if the distribution is Poisson.

Generally, point interactions are effective below a given range of correlation r_c, and therefore, for any $r > r_c$ the distribution is approximately Poisson. This kind of distributions are called *Markov* distributions, and are expected to have $g(r) = 1$ above this range. For $r < r_c$ the behavior of $g(r)$ can be complex but one out of four main trends is usually found. First, that of a **Poisson** process, where $g(r) = 1$. Second, we say we have a **cluster** process when $g(r) \geq 1$. Third, the opposite case, a **regular** process, where $g(r) \leq 1$. Finally, if $g(r) = 0$ for $r < r_c$ we have a **hardcore** process where no interactions are allowed bellow this range. These four cases can coexist in the same function at different ranges, showing different kinds of interactions depending on the scale.

However, the galaxy distribution shows a different behavior, where gravity extends its attraction infinitely. Hence, no finite correlation range is found for the pair correlation function, but an asymptotic approximation to an uncorrelated pattern. For scales below $120h^{-1}$ Mpc, the correlation functions shows a clustered pattern, with a bump around $105h^{-1}$ Mpc, corresponding to the shell radius of the BAOs (see section 1.1.2). After this, galaxies separated by greater distances describe a regular pattern, which tends asymptotically to 1.

As it is customary in astrophysics, the pair correlation function is redefined as

$$\xi(r) = g(r) - 1 \tag{3.3}$$

This is a useful way to subtract the "Poisson component", leaving in $\xi(r)$ the purely non-Poisson behavior of the distribution. It is clearly seen in the probability density expression

$$\varrho^{(2)}(r)dr = (\xi(r) + 1) \cdot \lambda^2 dr \tag{3.4}$$

3.3 Estimation of $\xi(r)$

The estimation of the pair correlation function is similar to the estimation of a Counts-in-Cells distribution, but it is estimated for a wide range of distances and normalized by the volumes involved. The galaxy counts are performed in shells or concentric cells centered on each galaxy. Ideally, the shells thickness would be infinitesimally thin, but in a discreet sample, such as a galaxy population, we require not null volumes. We are then interested in counting the number of galaxy pairs at an interesting range of several distances.

In order to perform this calculation we will refer as separated by distance r, those pairs whose separation distance lies between r and $r + dr$. Therefore, the volume created by this shell-shaped cell is

$$V_{sh} = \frac{4\pi}{3}[(r + dr)^3 - r^3] \tag{3.5}$$

Given a 3-dimensional region W containing N galaxies, one could estimate the pair correlation function with

$$\hat{\xi}(r) = \frac{V(W)}{N^2} \sum_{i=1}^{N} \frac{n_i(r)}{V_{sh}} - 1 \tag{3.6}$$

Where $n_i(r)$ is the number of galaxies lying inside the volume V_{sh} centered at the galaxy i. This estimation can be found at Rivolo (1986).

However, this estimator suffers from a severe underestimation of pairs at long distances. If the distances we are interested in are comparable to the dimensions of our population geometry, we will underestimate the number of real galaxy pairs when one of our galaxies is too close to the window limits. This problem is similar to that of the cells lying too close to the mask in the CiC calculations and its solution appears to be similar as well: we will use as centers for counting neighbors only galaxies lying within an inner region W_{in} where no galaxy is closer to the limits of W than a distance r_{max}. With this modification, we have the minus estimator

$$\hat{\xi}_{min}(r) = \frac{V(W)}{N \cdot N_{in}} \sum_{i=1}^{N_{in}} \frac{n_i(r)}{V_{sh}} - 1 \tag{3.7}$$

where N_{in} is the number of galaxies contained in the inner region.

Despite the reliability of this method and its wide applicability, it has a clear handicap since a significant amount of data remains unused, which increases biases and uncertainties. To avoid this problem we can proceed with an edge correction that compensates the loss of information normalizing the pair counts by the volume V_i of the intersection of V_{sh} and W. This edge-correction estimator can be calculated as

$$\hat{\xi}_{min}(r) = \frac{V(W)}{N^2} \sum_{i=1}^{N} \frac{n_i(r)}{V_i} - 1 \tag{3.8}$$

However, this estimator requires of complex integration methods to calculate V_i for every galaxy.

Instead, Poisson distributed samples of points in W can be used to approximate this integral as in a Monte Carlo volume integration. Again, this procedure is similar to the one we used for the estimation of the relative volume of cells in the CiC analysis. Once the galaxy population and its containing geometry is perfectly known, we can generate a Poisson distributed sample of points in the same geometry (i.e., a Poisson point process with N_R points). As said, for this kind of population, the pair correlation function $\xi(r)$ is equal to 0 and any divergence due to border effects can be used to correct the same underestimation in the galaxy

population. The function $\sum n_i(r)$ is now substituted by the total number of pairs in the sample at separated by a distance r. This will be calculated for pairs of galaxies ($DD(r)$), pairs of random points from the Poisson sample ($RR(r)$) and mixed pairs formed by a galaxy and a random point ($DR(r)$). Different estimators were proposed by (Davis & Peebles, 1983, Hamilton, 1993, Landy & Szalay, 1993, Peebles & Hauser, 1974), and are given by

$$\hat{\xi}_{PH}(r) = \left(\frac{N_R}{N_D}\right)^2 \frac{DD(r)}{RR(r)} - 1$$
$$\hat{\xi}_{DP}(r) = \frac{N_R}{N_D}\frac{DD(r)}{DR(r)} - 1$$
$$\hat{\xi}_{HA}(r) = \frac{DD(r)\cdot RR(r)}{(DR(r))^2} - 1$$
$$\hat{\xi}_{LS}(r) = 1 + \left(\frac{N_R}{N_D}\right)^2 \frac{DD(r)}{RR(r)} - 2\frac{N_R}{N_D}\frac{DR(r)}{RR(r)}$$

where N_D is the number of galaxies in the population and N_R is the number of points in the auxiliary Poisson sample. Since this is a Monte Carlo procedure, the larger N_R is, the more accurate the result. In this thesis we use $N_R = 20N_D$. Several works compared the performance of these estimators, in terms of bias and variance, in the cosmological scenario (Kerscher, 1999, Labatie et al., 2010, Pons-Bordería et al., 1999). The general conclusion, as first noted by Hamilton (1993), is that the bias is lower for $\hat{\xi}_{HA}(r)$ and $\hat{\xi}_{LS}(r)$, especially at large scales. As usual in cosmology, we use the Landy-Szalay (LS) estimator through this thesis.

3.3.1 Correction of selection effects

Galaxy surveys are usually affected by selection effects that alter our estimations of the true interaction between points. This selection is due to observational difficulties that introduce inhomogeneities in the population. The first case is explained in the previous section, when the limits of the window introduce a border effect that must be corrected. The edge correction is satisfactory even if the window adopts very irregular shapes due to the adaptability of the Monte Carlo volume integrations. However, other effects must be taken into account to perform a reliable estimation of $\xi(r)$.

Even regions lying in the window can present different levels of completeness due to different conditions during the observing nights or to masking objects like bright stars that do not allow to observe the faint galaxies around them. This selection effect can only be poorly corrected aside from reducing the window area. However, we must deal with it if we want to study the cosmic variance, the variance of observations of the universe at extreme distances.

In addition, selection effects can appear not only in the observation of the sky window of the survey, but also in the depth of the observations, aside from other more subtile effects. At large distances the faintest galaxies are difficult to observe properly and its detection is non reliable or inexistent. That creates an inhomogeneity, showing higher galaxy densities at closer distances. In addition, this selection is a strong bias, since the lost galaxies are the faintest ones, which behave differently to the brighter ones in their clustering interactions (Arnalte-Mur et al., 2014). Two corrections can be introduced to address these effects. First, we must introduce an appropriate selection in absolute magnitude and redshift to produce a "volume limited" sample. This way we eliminate the faintest galaxies at closer distances to reach a luminosity completeness at all distances, although, this implies losing a large amount of valid data. And second, strong inhomogeneities of the galaxy density in depth must be introduced in the generation of the auxiliary Poisson population. This density of galaxies can be modeled by cubic splines or other smoothing estimators and can be effectively used to generate the right amount of random points at each distance, producing less points where the selection function or any observational effects produce an underdensity.

An additional correction that we introduce in our calculations is the integral constraint correction (Peebles, 1980). There is a bias in the estimation due to the use of a finite volume. Correlations are measured with respect to the mean density of the selected sample (our galaxy survey) instead of with respect to the global mean (that of the parent population). As explained in Bernardeau et al. (2002) and Labatie et al. (2010), at first order the bias introduced by the integral constraint is given by

$$\xi(\mathbf{r}) = \xi^{true}(\mathbf{r}) - K \tag{3.9}$$

where,

$$K = \frac{1}{V^2} \int_V \int_V d^3\mathbf{r}_1 d^3\mathbf{r}_2 \xi^{true}(\mathbf{r}_2 - \mathbf{r}_1) \tag{3.10}$$

and V is the volume occupied by the galaxy population. In practice, the estimation of K can be done following Roche et al. (1999), making use of the auxiliary Poisson catalogue to compute numerically the double integral as

$$K \simeq \frac{\sum_i RR(r_i)\xi^{model}(r_i)}{\sum_i RR(r_i)} = \frac{\sum_i RR(r_i)\xi^{model}(r_i)}{N_R(N_R - 1)} \tag{3.11}$$

In this thesis, we use a power law as a parametric model for ξ^{model}, evaluated in the same distance ranges r_i where the ξ function is evaluated. Nevertheless, the correction introduced by the integral constraint is generally smaller than the error bars calculated for our estimations.

3.3.2 Correction of redshift distortions

The estimation of the pair correlation function in three dimensions demands the 3-dimensional position of each object, which, as seen in eq. 1.23, implies knowing the right ascension α, declination δ and distance of the galaxy with respect to us. The position of the galaxy in the sky is easily obtained, but the distance needs a good estimation of the cosmological redshift to be properly calculated.

As seen in eq. 1.15, when we obtain our distances from the redshift of a galaxy, it is affected by the double component of the redshift: the cosmological redshift due to the Hubble flow, which in principle is directly linked to the position, and any additional peculiar velocity of the object. These peculiar velocities are produced by the infalling and other galaxy movements due to strong and close gravitational interactions, specially in clusters. If the galaxy is moving coherently to the Hubble flow, its redshift will appear larger, implying a larger distance, while, in the other sense, if the galaxy is following towards us, the redshift will be smaller and the galaxy will appear closer than it really is. Together, this makes clusters appear as an elongated shape called Fingers-of-God. Several other redshift distortions are found when measuring galaxy redshifts, but we will not try to introduce further corrections in our estimations of $\xi(r)$

In addition to this natural distortion, observations can introduce new errors in the redshift measurements that must be treated. In the spectroscopic surveys, the redshift is measured with great accuracy (typically, $\sigma(z) \sim 10^{-4}$) and the derived distances are generally reliable. However, photometric surveys generally present greater inexactitudes due to the low resolution spectral energy distributions and the difficulty to properly locate the emission and absorption lines. In the best cases, when the number of used filters is large enough, this uncertainty can be reduced to $\sigma(z) \sim 0.01(1+z)$ (Moles et al., 2008, Molino et al., 2014).

In order to deal with this non-isotropic redshift-space galaxy distribution, Davis & Peebles (1983) started using the 2-dimensional correlation function $\xi(r_{||}, r_{\perp})$, where the dependence on scale is split into the line of sight separation $r_{||}$, and the transverse separation $r_{\perp}$[1]. This is a decomposition of the galaxy pairs distances into the plane that contains the galaxy pair and the observer. It can be seen in the scheme of Fig. 3.1. Since the angle of separation of galaxies is generally small, the parallel component is barely radial with respect to the observer, the direction where most of the explained redshift distortions appear.

The formal definition of $r_{\perp}$ and $r_{||}$ starts with the locations $\mathbf{s}_1$, $\mathbf{s}_2$ of two galaxies in the observed redshift space. We define the separation vector, $\mathbf{s} \equiv \mathbf{s}_2 - \mathbf{s}_1$, and the line-of-sight vector, $\mathbf{l} \equiv \mathbf{s}_1 + \mathbf{s}_2$. From these, we now define the parallel and perpendicular distances of the pair as

$$r_{||} \equiv \frac{|\mathbf{s} \cdot \mathbf{l}|}{|\mathbf{l}|}, r_{\perp} \equiv \sqrt{\mathbf{s} \cdot \mathbf{s} - r_{||}^2} \tag{3.12}$$

Redshift errors will influence the result in the apparent line-of-sight direction $\mathbf{l}/|\mathbf{l}|$, and through that, the perpendicular component $r_{\perp}$, and mainly, the parallel one $r_{||}$. These errors will grow with the redshift uncertainties and with the angular separation of the galaxy pair.

The adaptation of the point correlation function to this decomposition substitutes the dependence on r with a dependence on $(r_{\perp}, r_{||})$, allowing us to deal with the non-isotropy of the survey in the radial dimension. The 2-dimensional correlation function is then estimated using the Landy and Szalay estimator:

[1]In the literature , the symbol π is typically used for $r_{||}$, and σ or r_p for $r_{\perp}$. We use $r_{||}$ and $r_{\perp}$ or r_p indistinctly.

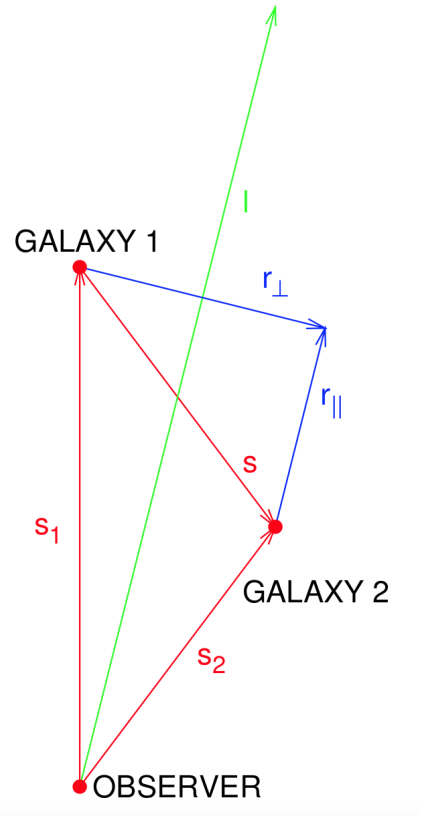

FIGURE 3.1: Transversal ($r_\perp$) and longitudinal ($r_{||}$) decomposition of the galaxy pairs distances into the plane. Credit: Pablo Arnalte-Mur.

$$\hat{\xi}_{LS}(r_\perp, r_{||}) = 1 + \left(\frac{N_R}{N_D}\right)^2 \frac{DD(r_\perp, r_{||})}{RR(r_\perp, r_{||})} - 2\frac{N_R}{N_D}\frac{DR(r_\perp, r_{||})}{RR(r_\perp, r_{||})} \tag{3.13}$$

where pairs are calculated as above and every pair is decomposed in two separation distances: the transverse $[r_\perp, r_\perp + dr_\perp]$ and the paralel $[r_{||}, r_{||} + dr_{||}]$.

Here we can see how an aim for this decomposition is the study of the redshift distortions (see Fig. 3.2). These redshift space distortions on the correlation function can be modeled (Hamilton, 1998, Kaiser, 1987) and used to constrain the cosmological parameters, in particular Ω_m (Cabré & Gaztañaga, 2009, Hawkins et al., 2003, Peacock et al., 2001).

Now, the 2-dimensional pair correlation function can be used as well to neutralize the redshift distortions by projecting the correlation function into the transverse direction (Davis & Peebles, 1983). This projection is done integrating along the line-of-sight, where uncertainties are mainly found:

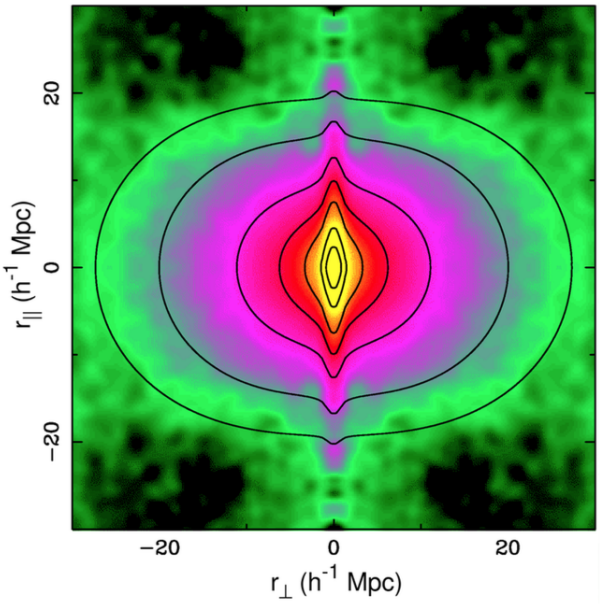

FIGURE 3.2: 2-dimensional correlation function for the 2dFGRS survey, showing the effects of redshift space distortions. The cigar-like shape at $r_\perp \sim 0$ is due to random peculiar velocities in virialised groups, while the oval-like shape at larger transverse separations is the signature of coherent infall. Image from Peacock et al. (2001).

$$w(r_\perp) \equiv 2 \int_0^\infty \xi_{LS}(r_\perp, r_{||}) dr_{||} \tag{3.14}$$

This is the projected correlation function. Assuming that the real-space distribution is isotropic, we can relate w to the real space correlation function, ξ_r, as

$$w(r_\perp) = 2 \int_{r_\perp}^\infty \xi_r(r) \frac{r dr}{(r^2 - r_\perp^2)^{1/2}} \tag{3.15}$$

This relation can be inverted obtaining ξ_r, in terms of w,s as the Abel integral:

$$\xi_r(r) = -\frac{1}{\pi} \int_r^\infty \frac{dw(r_\perp)}{dr_\perp} \frac{dr_\perp}{(r_\perp^2 - r^2)^{1/2}} \tag{3.16}$$

In practice, we have to set finite upper limits in our integrals. In equation 3.15 the upper limit $r_{||,max}$ must be fixed instead of ∞. In principle it should be large enough to include all the correlated pairs, but if it is to large, it will introduce extra noise in the calculation. Arnalte-Mur et al. (2014) recommend to use a value of

$r_{\perp,max} = 200h^{-1}$ Mpc in the photometric ALHAMBRA survey (see section 3.4.1), a value that depends on the typical redshift error Δz. In equation 3.16 an upper limit $r_{\perp,max}$ must be also fixed, being the maximum transverse separation allowed by the geometry of the survey.

3.3.3 Estimation of the errors

The error bars depicted for the projected correlation function are calculated following the 'delete one jackknife' method explained in section 2.4 for the Counts-in-Cells case. Its application to correlation functions is direct, with the minor difference of the subsample selection criterium. Our data from ALHAMBRA survey (see next section 3.4.1) contains images from 47 different CCDs, covering each one around 0.05 deg^2 of the sky. We take these areas to build the subsamples, extending them in redshift. Therefore, each jackknife subsample contains the galaxies of 46 CCDs.

3.4 Evolution of galaxy spectral segregation in the ALHAMBRA Survey

It has been well established that different types of galaxies cluster in different ways (Davis et al., 1988, Domínguez-Tenreiro & Martínez, 1989, Einasto, 1991, Guzzo et al., 1997, Hamilton, 1988, Li et al., 2006, Loveday et al., 1995, Martínez et al., 2010, Phleps et al., 2006). Elliptical galaxies are preferentially located at the cores of rich galaxy clusters, i.e, in high density environments, while spiral galaxies are the dominant population in the field (Cucciati et al., 2006, Davis & Geller, 1976, Dressler, 1980, Giovanelli et al., 1986). This phenomenon, called galaxy segregation, has been confirmed in the largest galaxy redshift surveys available up to date, the 2dF Galaxy redshift survey (2dFGRS, Madgwick et al., 2003), the Sloan Digital Sky Survey (SDSS, Abbas & Sheth, 2006, Zehavi et al., 2011) and the Baryonic Oscillation Spectroscopic Survey (BOSS, Guo et al., 2013). The

dependence of clustering on different galaxy properties such as stellar mass, concentration index, or the strength of the 4000 Å-break has been studied by Li et al. (2006).

Since segregation is a consequence of the process of structure formation in the universe, it is therefore very important to understand its evolution with redshift or cosmic time. Several works have extended the analysis of segregation by colour or spectral type to redshifts in the range $z \sim 0.3 - 1.2$ using recent spectroscopic surveys such as the VIMOS-VLT Deep Survey (VVDS, Meneux et al., 2006), the Deep Extragalactic Evolutionary Probe 2 survey (DEEP2, Coil et al., 2008), or the PRIsm Multi-object Survey (PRIMUS, Skibba et al., 2014). De la Torre et al. (2011), instead, used the zCOSMOS survey to study segregation by morphological type at $z \sim 0.8$. All these studies show that segregation by colour or spectral type was already present at $z \sim 1$. In particular, Meneux et al. (2006), using a sample of 6,500 VVDS galaxies covering half a square degree, have unambiguously established that early-type galaxies are more strongly clustered than late-type galaxies at least since redshift $z \sim 1.2$. The correlation length obtained by these authors for late-type galaxies is $r_0 \sim 2.5h^{-1}$ Mpc at $z \sim 0.8$ and roughly twice this value for early-type galaxies. They have also calculated the relative bias between the two types of galaxies obtaining an approximately constant value $b_{\rm rel} \sim 1.3 - 1.6$ for $0.2 \leq z \leq 1.2$ depending on the sample. This value is slightly larger than the one obtained by Madgwick et al. (2003) $b_{\rm rel} \sim 1.45 \pm 0.14$ for the 2dF Galaxy Redshift Survey with median redshift $z = 0.1$. The results obtained by Coil et al. (2008) for DEEP2 reinforced those outlined above, although the measured correlation lengths for DEEP2 galaxies are systematically slightly larger than the values reported for the VVDS sample by Meneux et al. (2006). In addition, Coil et al. (2008) have detected a significant rise of the correlation function at small scales $r_p \leq 0.2h^{-1}$ Mpc for their brighter samples. For the zCOSMOS-Bright redshift survey, de la Torre et al. (2011) found also that early-type galaxies exhibit stronger clustering than late-type galaxies on scales from 0.1 to $10h^{-1}$ Mpc already at $z \simeq 0.8$, and the relative difference increases with cosmic time on small scales, but does not significantly evolve from $z = 0.8$ to $z = 0$ on large scales. A similar result is reported by Skibba et al. (2014). These authors show that the clustering amplitude for the PRIMUS sample increases with color, with redder galaxies displaying stronger clustering at scales $r_p \leq 1h^{-1}$ Mpc. They have also detected a color dependence

within the red sequence, with the reddest galaxies being more strongly correlated than their less red counterparts. This effect is absent in the blue cloud.

Several broad-band photometric surveys have extended these studies to even larger redshift (e.g. Hartley et al., 2010, McCracken et al., 2015). Hartley et al., using data from the UKIDSS Ultra Deep Survey, find segregation between passive and star-forming galaxies at $z \lesssim 1.5$, but find consistent clustering properties for both galaxy types at $z \sim 2$.

In the present work we use the high-quality data of the Advanced Large Homogeneous Area Medium-Band Redshift Astronomical survey (ALHAMBRA) (Moles et al., 2008, Molino et al., 2014)[2] to study the clustering segregation of quiescent and star-forming galaxies. ALHAMBRA is very well suited for the analysis of galaxy clustering and segregation studies at very small scales. With a reliable calculation of the projected correlation function we find a clear steepening of the correlation at scales between 0.03 to $0.2h^{-1}$ Mpc (Coil et al., 2008, Phleps et al., 2006), specially for the star-forming galaxies. Moreover, its continuous selection function over a large redshift range makes ALHAMBRA an ideal survey for evolution studies (see Fig. 3.3). In Arnalte-Mur et al. (2014) (hereafter AM14) the authors presented the results of the evolution of galaxy clustering on scales $r_p < 10h^{-1}$ Mpc for samples selected in luminosity and redshift over ~ 5 Gyr by means of the projected correlation function $w_p(r_p)$. In this work we use the same statistic to study the evolution of galaxy segregation by spectral type at $0.35 < z < 1.1$.

Details on the samples used in this analysis are described in Section 3.4.1. In section 3.4.2 we compare the ALHAMBRA survey with other photometric surveys and justify its suitability for the analysis of correlations at small scales. Section 3.4.3 we present the calculations, including the power-law modeling (section 3.4.3.1), and the galaxy bias (section 3.4.3.4). Conclusions are summarized in section 3.4.4. Throughout this work we use a fiducial flat ΛCDM cosmological model with parameters $\Omega_M = 0.27$, $\Omega_\Lambda = 0.73$, $\Omega_b = 0.0458$ and $\sigma_8 = 0.816$ based on the 7-year *Wilkinson Microwave Anisotropy Probe (WMAP)* results (Komatsu et al., 2011). All the distances used are comoving, and are expressed in terms of the Hubble parameter $h \equiv H_0/100$ km s^{-1} Mpc^{-1}. Absolute magnitudes are given as $M - 5\log_{10}(h)$.

[2]http://alhambrasurvey.com

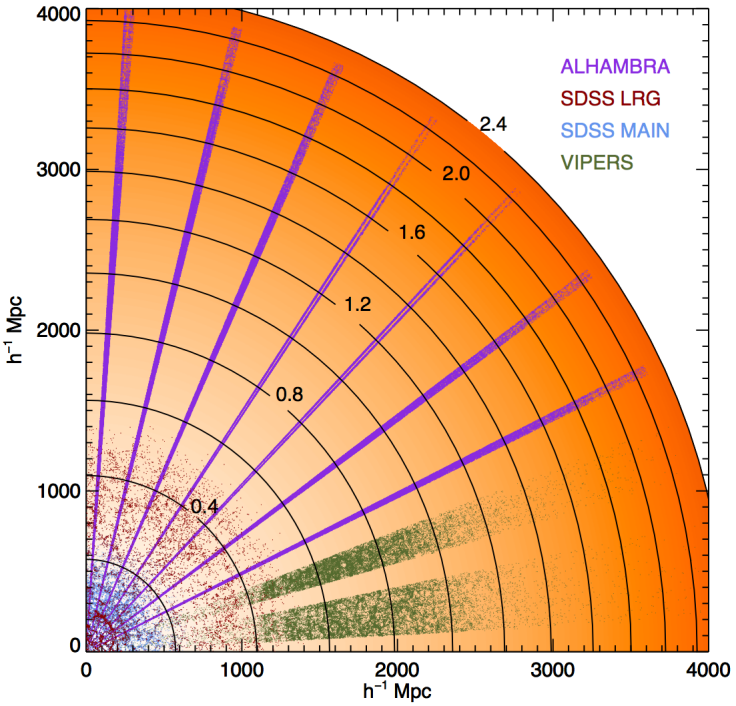

FIGURE 3.3: Comparison between the galaxies observed by the ALHAMBRA survey, SDSS and VIPERS. Axis in comoving distance, spheres in redshift. Image credit: Vicent Martínez.

3.4.1 Data samples

The ALHAMBRA survey (Moles et al., 2008, Molino et al., 2014) is a project that has imaged seven different areas in the sky through a purposedly-built set of 20 contiguous, non-overlapping, 310 Å-wide filters covering the whole visible range from 3500 to 9700 Å, plus the standard near-infrared JHK_s filters. The nominal depth (5σ, $3''$ aperture) is $I_{AB} \sim 24.5$ and the total sky coverage after masking is 2.381deg^2. The final catalogue, described in Molino et al. (2014), includes over 400,000 galaxies, with a photometric redshift accuracy better than $\sigma_z/(1+z) = 0.014$. Full details on how the accuracy depends on the sample magnitude, galaxy type, and Bayesian odds selection limits are given in that work. For the characteristics of the sample that we use in this paper the authors quote a dispersion $\sigma_{\text{NMAD}} < 0.014$ and a catastrophic rate $\eta_1 = 0.04\%$ [3]. There is no evidence of significantly different behavior for galaxies with spectral energy distributions corresponding to quiescent or star-forming types. Contamination

[3]Where σ_{NMAD} is the normalized median absolute deviation, and η_2 is defined as the proportion of objects with absolute deviation $|\delta z|/(1+z) > 0.2$.

by AGNs is minor (approximately 0.1% of the sources could correspond to this class, which has not been purged from the ALHAMBRA catalogues) and should be dominated by low-luminosity AGN, which are in many cases fit by strong emission-line galaxies with an approximately correct redshift.

Object detection is performed over a synthetic image, created via a combination of ALHAMBRA filters, that mimics the Hubble Space Telescope F814W filter (hereafter denoted by I) so that the reference magnitude is directly comparable to other surveys. Photometric redshifts were obtained using the template-fitting code BPZ (Benítez, 2000), with an updated set of 11 Spectral Energy Distribution (SED) templates, as described in Molino et al. (2014). The Bayesian approach to photometric data has proved to be successful in the ALHAMBRA survey (Ascaso et al., 2015). Although a full posterior probability distribution function in redshift z and spectral type T is produced for each object, in this work we take a simpler approach and assign to each galaxy the redshift z_b and type T_b corresponding to the best fit to its observed photometry. We have checked that the errors induced by the redshift uncertainties, which are partly absorbed by the deprojection technique, are under control as long as we use relatively bright galaxies with good quality photometric redshift determinations. This makes ALHAMBRA a very well suited catalogue: together with the high resolution photometric redshifts, the abundant imaging allows us both a reliable color segregation, used in this work, and a high completeness in the galaxy population at small scale separations, which will be the specific object of a future work.

The ALHAMBRA data was already used for clustering studies by Arnalte-Mur et al. (2014), who focused on the study of galaxy segregation by luminosity. In this work the authors find a clear evolution of the projected correlation amplitude with luminosity (see Fig. 3.4). These results confirmed that different types of galaxies cluster in different ways, and we will extend this analysis to a new type of galaxy segregation. Clustering is also dependent on redshift, showing a hint of evolution with lower amplitudes for higher redshifts. Analogous results are found for the galaxy bias, with dependence of luminosity and redshift, where high redshift bright galaxies present higher values of bias than faint or low redshift galaxies.

We have drawn different samples from the ALHAMBRA survey to perform our analysis in a similar way as was done in Arnalte-Mur et al. (2014). First, we cut the magnitude range at $I < 24$, where the catalogue is photometrically complete

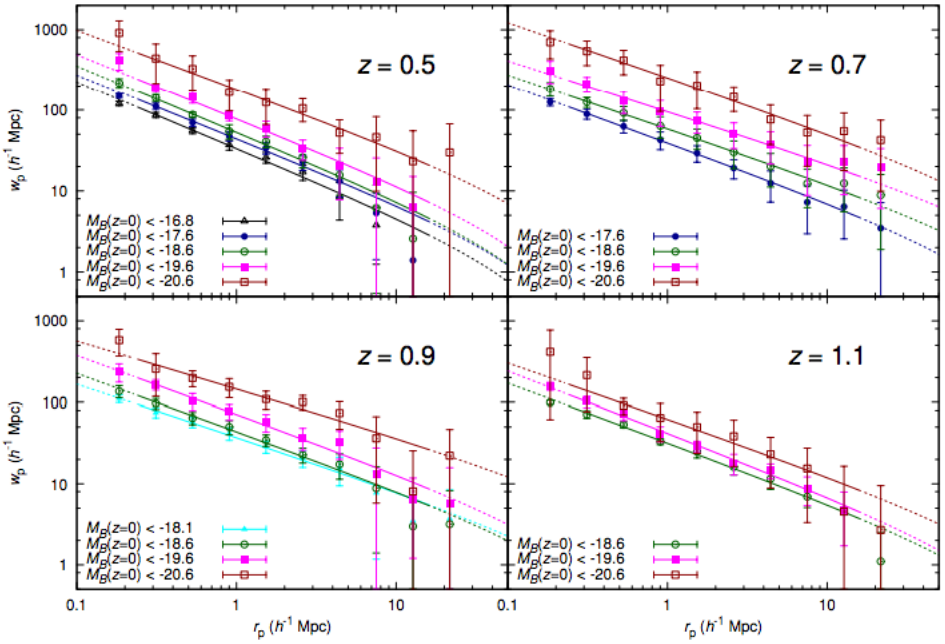

FIGURE 3.4: Projected correlation functions for luminosity segregated galaxy samples from Arnalte-Mur et al. (2014). The solid lines show the corresponding best-fit power laws in the range in which the fit was done. Dashed lines show the extrapolation of these models to larger or smaller scales.

(Molino et al., 2014) and we do not expect any significant field-to-field variation in depth. Second, stars are eliminated using the star-galaxy separation method described in Molino et al. (2014). As explained in Arnalte-Mur et al. (2014), the expected contamination by stars in the resulting samples is less than 1 per cent. Finally, we cleanse the catalogue using the angular masks defined in Arnalte-Mur et al. (2014), which eliminate regions with less reliable photometry around bright stars or image defects, or very close to the image borders. The sample selected in this way contains 174,633 galaxies over an area of $2.381\,\mathrm{deg}^2$, *i.e.*, with an approximate source density of 7.3×10^4 galaxies per square degree.

Given the ALHAMBRA depth, we divide our sample in 5 non-overlapping redshift bins. These redshift bins are $[0.35, 0.5[$, $[0.5, 0.65[$, $[0.65, 0.8[$, $[0.8, 0.95[$, and $[0.95, 1.1]$ [4]. As in this work we focus on the galaxy spatial segregation by spectral type we use a luminosity selection to obtain a fixed number density. In this way we guarantee that we are comparing similar populations at different redshifts. In

[4]Note that the redshift bins used here are different to those in Arnalte-Mur et al. (2014), where overlapping bins were allowed.

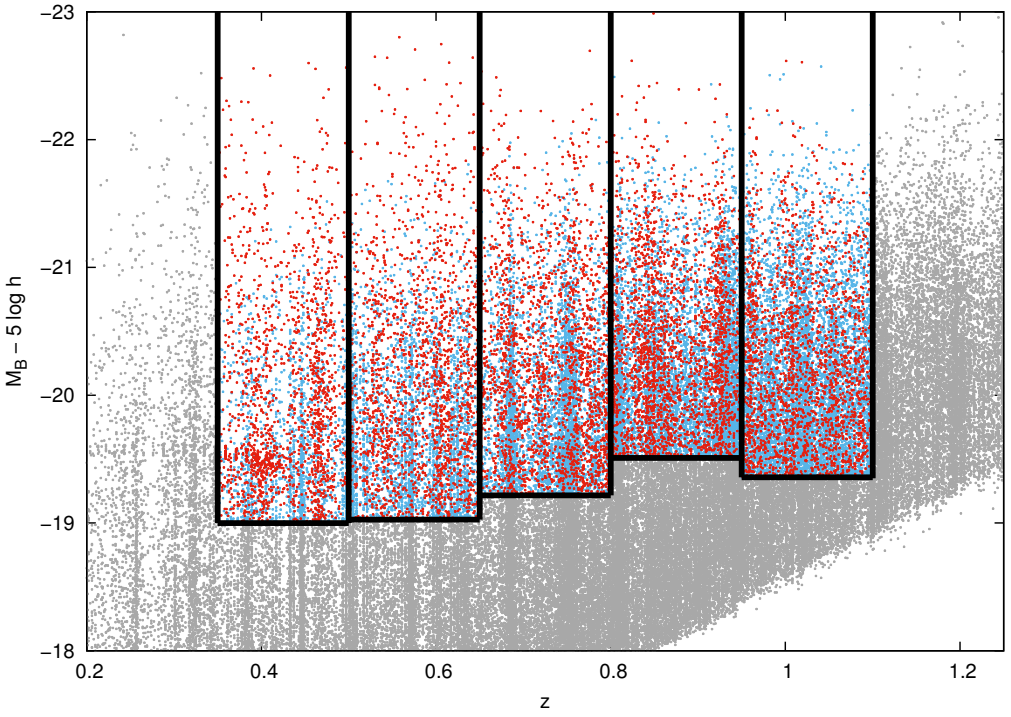

FIGURE 3.5: Selected samples with fixed number density in the photometric redshift vs. absolute B-band magnitude diagram. The quiescent and star-forming galaxy samples are plotted in red and blue color, respectively. The solid lines mark the boundaries of our selected samples described in Table 3.1.

order to select a sample that is complete up to $z = 1.1$ we define a threshold magnitude of $M_B^{th}(0) - 5\log(h) = -19.36$ for the highest redshift bin. This limit determines the galaxy number density ($\bar{n} = 9.35 \times 10^{-3}\, h^3$ Mpc^{-3}) that we will keep constant for the remaining redshift bins. This way, our results do not rely on measurements of the luminosity function. This will allow us to study the evolution with redshift of the galaxy spectral segregation. Fig. 3.5 shows the luminosity and redshift selections used in this work. We should remark on the non-monotonic evolution of the faint limit of our samples with redshift. This effect is not unexpected, as a combination of cosmic variance in the large-scale structure and the artificial redshift peaks that are induced by the photometric redshift methods produce density changes that are observable at the scales we are using. In any case the effect is very small, representing a variation of only 0.1 magnitudes per bin over a monotonic evolution.

We classify our galaxies as 'quiescent' and 'star-forming' according to the best-fitting template, Tb, obtained from the BPZ analysis. Templates 1 to 5 correspond

to quiescent galaxies, 6 and 7 correspond to star-forming galaxies, and 8 to 11 correspond to starburst galaxies. We consider as quiescent galaxies those with a template value smaller than 5.5, and star-forming those with a value bigger than 5.5. Therefore, we include in the star-forming category also those galaxies classified as starbursts. Note that in the fitting process interpolation between templates is performed.

In a previous work, Pović et al. (2013) built a morphological catalogue of 22,051 galaxies in ALHAMBRA. We cannot, however, use this catalogue as the basis for our analysis as it includes only a small subset of the galaxies in our sample: in its cleanest version it is limited to $AB(\mathrm{F613W}) < 22$ and redshift $z < 0.5$ for ellipticals. A cross-check showed that, if we identify quiescent galaxies as early-type and star-forming galaxies as late-type, our SED-based classification agrees with the morphological one for over 65% of the sample. Taking into account that the nominal accuracy of the morphological catalogue is 90%, that we are actually using only the objects close to its detection limit and that, as noted in Pović et al. (2013), the relationship between morphological- and colour-based classifications is far from being as direct as could naïvely be expected, we consider that these figures prove that the classification is accurate within the expected limits.

TABLE 3.1: Characteristics of the galaxy samples used

			Quiescent galaxies				Star-forming galaxies				
Sample	z range	$V(h^{-3}Mpc^3)$	N_{Q}	$\bar{n}(h^3Mpc^{-3})$	M_B^{med}	$\bar{z}$	N_{Sf}	$\bar{n}(h^3Mpc^{-3})$	M_B^{med}	$\bar{z}$	$\frac{N_{\mathrm{Q}}}{N_{\mathrm{Q}}+N_{\mathrm{Sf}}}$
z0.43	$0.35-0.5$	3.48×10^5	1650	4.74×10^{-3}	-20.53	0.43	1605	4.61×10^{-3}	-20.26	0.43	0.51
z0.57	$0.5-0.65$	5.42×10^5	1818	3.35×10^{-3}	-20.77	0.58	3258	6.01×10^{-3}	-20.35	0.57	0.36
z0.73	$0.65-0.8$	7.33×10^5	2291	3.12×10^{-3}	-20.87	0.73	4570	6.23×10^{-3}	-20.56	0.73	0.33
z0.88	$0.8-0.95$	9.09×10^5	2509	2.75×10^{-3}	-21.06	0.87	6002	6.6×10^{-3}	-20.82	0.88	0.29
z1.00	$0.95-1.1$	1.06×10^6	2182	2.05×10^{-3}	-20.91	1.02	7768	7.30×10^{-3}	-20.74	1.03	0.22

V is the volume covered by ALHAMBRA in each redshift bin. For each of the samples selected by spectral type we show the number of galaxies N, the mean number density $\bar{n}$, the median B-band absolute magnitude M_B^{med} and the mean redshift $\bar{z}$. The last column gives the fraction of early-type galaxies in the bin. NW ALH-4 frame is not included.

In Fig. 3.6 we show how our classification of quiescent and star-forming galaxies performs on a color-luminosity diagram. We plot M_r and M_u, which correspond

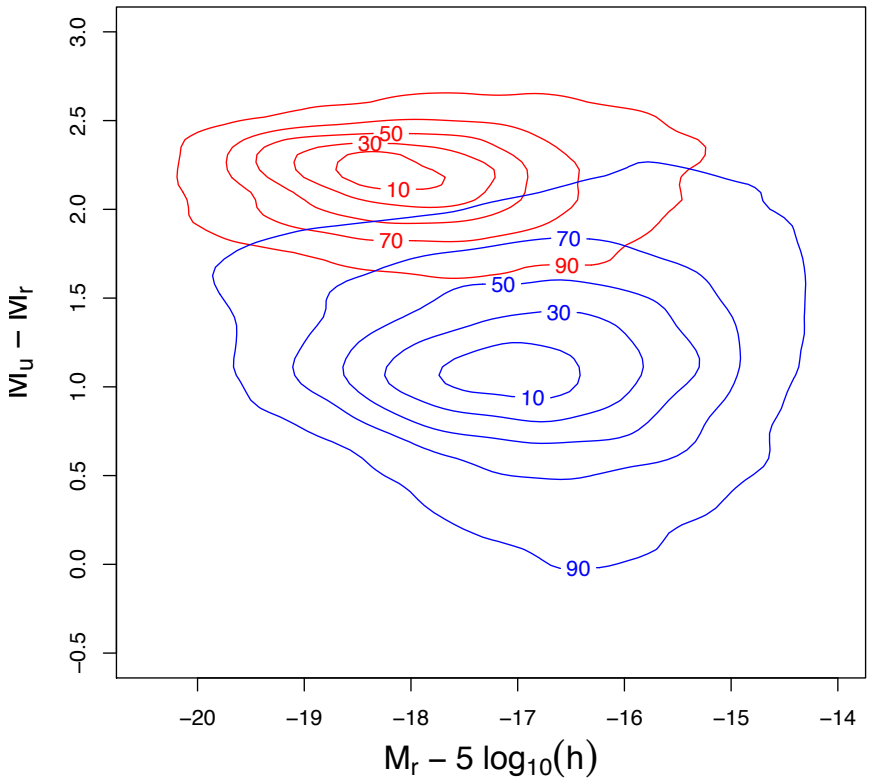

FIGURE 3.6: Absolute rest-frame broad-band color-magnitude diagram for galaxies with redshifts between 0.35 and 0.75. Quiescent and star-forming galaxies (selected by their best-fit spectral type) are shown, respectively, as red and blue percentile contours. We have used SDSS absolute magnitudes derived from the ALHAMBRA photometry as described in the text. Our classification by (photometric) spectral type closely matches the usual broad-band colour selection.

to the absolute magnitudes in the SDSS rest-frame broad-band filters r and u, and were estimated from ALHAMBRA data by Stefanon (2011) for galaxies with redshift $0.35 < z < 0.75$ and good quality photometric redshifts. We see how well the ALHAMBRA spectral-type classification reproduces the expected behavior (Bell et al., 2004): quiescent galaxies correspond to the 'red sequence' in the diagram, while star-forming galaxies form the 'blue cloud'. In addition to the clear segregation in color, we see that quiescent galaxies show, on average, slightly brighter luminosities than star-forming ones. This shows that our selection by (photometric) spectral type is almost equivalent to a selection in broad-band color.

In Fig. 3.7 we show the projection of two fields, ALH-2 and ALH-4/COSMOS,

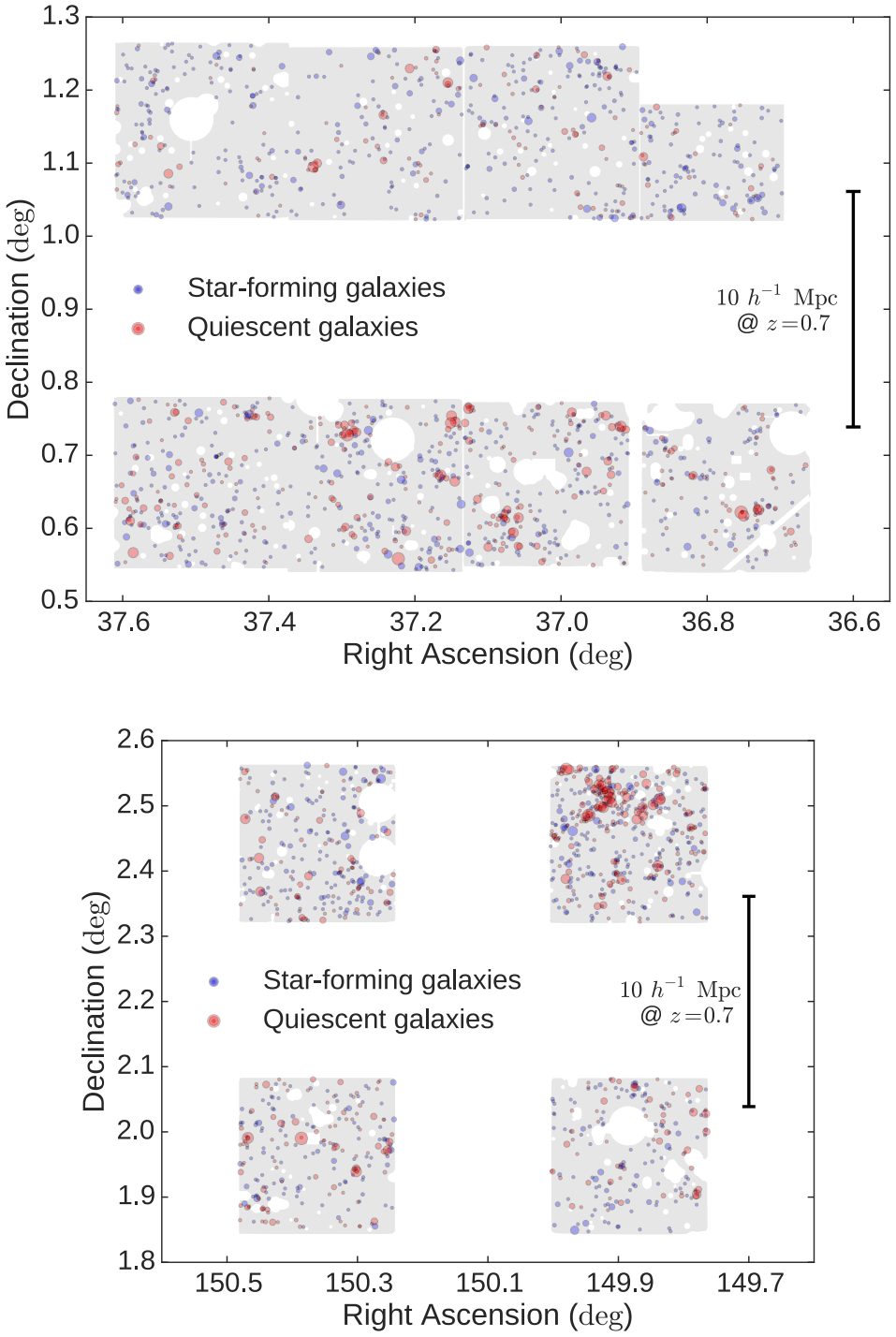

FIGURE 3.7: Projection onto the sky of the z0.73 sample ($0.65 < z < 0.8$) for two ALHAMBRA fields: ALH-2 (top) and ALH-4/COSMOS (bottom). Galaxies have been colored according to their type: blue circles correspond to star-forming and red circles to quiescent galaxies, and the size of each circle is proportional to the luminosity of the corresponding galaxy. North is to the top and East is to the left. The diagram shows the geometry of the ALHAMBRA fields, with the angular mask described in the text displayed as a light-grey background. The scale of 10 h^{-1} Mpc at $z = 0.7$ is indicated as a vertical bar. A heavy concentration of red circles (quiescent galaxies), corresponding to the big coherent structure described in the text, is patent in the NW quadrant of the ALH-4/COSMOS field.

onto the plane of the sky. The coherent superstructure in the ALH-4/COSMOS field at $0.6 < z < 0.8$ is well appreciated. The skeleton of the structures, that form the cosmic web, is perfectly delineated by the red quiescent galaxies, while blue star-forming ones tend to populate the field or lower density regions. A similar trend was also visible in the redshift versus right ascension diagram of Guzzo et al. (2014). We will further study this color-density relation (Cucciati et al., 2006) in the following sections by means of the projected correlation function.

Finally, we remove the North-West ALH-4 frame from the analysis (the top-right section on the bottom panel of Fig. 3.7). As seen in Arnalte-Mur et al. (2014), there exists an anomalous clustering in the ALH-4 field, which overlaps with the Cosmic Evolution Survey (COSMOS, Scoville et al., 2007b). It is well known that the COSMOS survey presents higher clustering amplitude than similar surveys (de la Torre et al., 2010, McCracken et al., 2007) due to the presence of large overdense structures in the field (Guzzo et al., 2007, Scoville et al., 2007a). This overdensity of structures is also observed in ALHAMBRA when comparing this peculiar region of the ALH-4 field with the rest of the fields Ascaso et al. (2015), Molino et al. (2014). In Arnalte-Mur et al. (2014) the authors showed that ALH-4/COSMOS is an outlier in terms of clustering. We have seen that not only does this region introduce anomalies in the measurement of the clustering statistics, but it also affects the error estimation of these statistics. In Arnalte-Mur et al. (2014) the authors also identified ALH-7/ELAIS-N1 as an outlier field (although the significance of the anomaly was smaller in this case). However, here we do not find any significant change in our results when removing the ALH-7/ELAIS-N1 field, so we keep it in for all our calculations.

For each redshift bin we will analyze the clustering for the full selected population and separately for quiescent and star-forming galaxies. Table 3.1 summarizes the different samples in each redshift bin. Columns $N_{\rm Q}$ and $N_{\rm Sf}$ are the number of quiescent and star-forming galaxies respectively, and the last column is the fraction of quiescent galaxies in each of the redshift subsamples.

We see that the fraction of quiescent galaxies decreases with redshift. This behavior is expected qualitatively, as blue star-forming galaxies are dominant at earlier cosmic times, while red quiescent galaxies appear late, once star formation stops. This trend was also observed in a similar redshift range by e.g. Zucca et al. (2009) for the zCOSMOS 10k bright sample. They found that the population of bright

late-type galaxies becomes dominant at higher redshifts, and therefore the fraction of early-type galaxies decreases with redshift accordingly.

3.4.2 The correlation function at the smallest scales

It has been argued that the fact that the measured two-point correlation function has been consistent with a power law spanning a vast range of scales could be just a cosmic coincidence (Watson et al., 2011). As already introduced in section 1.1.3 there is an intrinsic difficulty in the measurement of the small scale correlation function in spectroscopic redshift surveys, mainly due to fiber collisions in multi-fiber spectrographs. For example, the physical size of the fibers in the SDSS spectrograph does not allow to take spectra of galaxies that are separated less than $55''$ in the same plate, which corresponds to about $0.5h^{-1}$ Mpc at $z = 0.7$. There are methods for correcting this problem (Guo et al., 2012a), however considering the complex dissipative processes (dynamical friction, tidal interaction, etc) involved in the clustering of galaxies at the smallest scales, it would be desirable to measure the correlation function directly on samples not affected by the fiber-collision problem (in fact, the correction is based on performing a cross-correlation between galaxies with measured spectroscopic redshift and all targets from the image catalog, including therefore a bias from physically uncorrelated pairs).

Due to the ALHAMBRA high quality deep photometry, we are able to obtain reliable redshift measurements for galaxies in the crowded central regions of clusters and groups. Fig. 3.8 shows a small patch of the sky where we can appreciate how very close pairs of galaxies are well represented and observed in the ALHAMBRA survey. Three out of the four objects with marked photometric redshift lie within the same group. Their angular separations are about ~ 8" which at their common redshift of about $z \sim 0.19$ correspond to a projected distance of $r_p \sim 0.025h^{-1}$ Mpc.

The ALHAMBRA survey allows us to obtain measurements of the correlation function at very small scales that are more reliable than those from most previous surveys. This can be illustrated by plotting the frequency distribution of the r_p-distance to the nearest neighbor. This is the projected separation distance as described in Fig. 3.1.

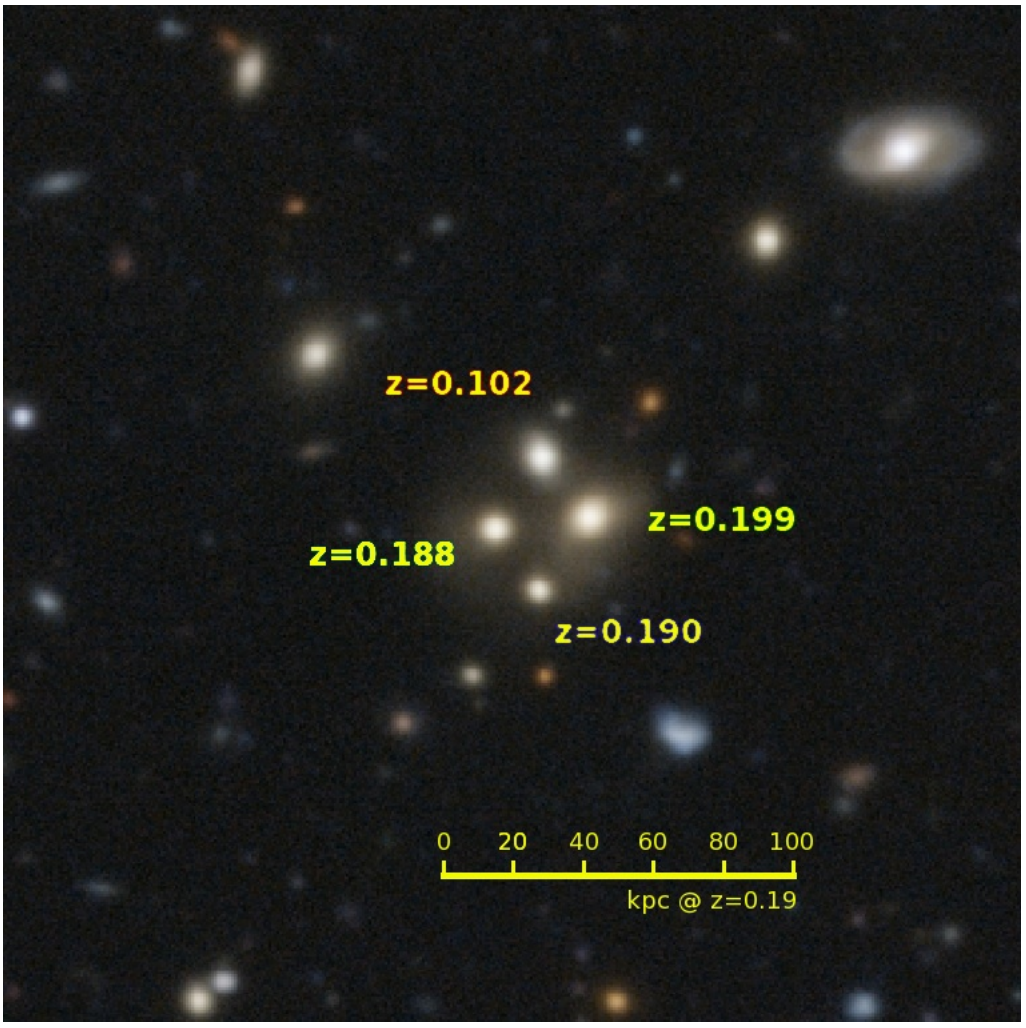

FIGURE 3.8: Image from the ALHAMBRA galaxy survey. Three galaxies lie separated by distances around 40 kpc, an example of the many close pairs and galaxy groups observed by ALHAMBRA. Image processed by Vicent Peris, indications added by Alberto Fernández-Soto.

We perform this calculation comparing the ALHAMBRA survey with data from three photometric galaxy surveys: the DEEP2 survey (Newman et al., 2013), the PRIMUS survey (Coil et al., 2011) and the COSMOS survey (Ilbert et al., 2008). All samples have been selected as in section 3.4.1, with redshift $0.35 < z < 1.0$ and a luminosity threshold of $M \leq -18.6 - 0.6 \cdot z$ for the B band in ALHAMBRA and DEEP2, the g band from *ugriz* for PRIMUS and the V-subaru band for COSMOS. The quality Q of redshifts for DEEP2 and PRIMUS is defined in Newman et al. (2013) and Cool et al. (2013) respectively. Both use spectroscopic redshifts, but PRIMUS quality is comparable to ALHAMBRA and COSMOS. For comparison, we show in Table 3.2 the main features for each survey.

TABLE 3.2: Photometric Surveys

	ALHAMBRA	DEEP2	PRIMUS	COSMOS
N (in sample)	33776	8735	55901	44839
Area (deg^2)	2.38	2.9	9.1	2
Depth	$I_{AB} < 24$	$R_{AB} < 24.15$	$i_{AB} < 23.5$	$i^{+}_{AB} < 24$
$\sigma_z/(1+z)$	~ 0.014	$Q > 2$	$Q > 2$	~ 0.007

Comparison between redshift surveys. For a detailed definition of value Q please read Newman et al. (2013) for DEEP2 survey and (Cool et al., 2013) for PRIMUS.

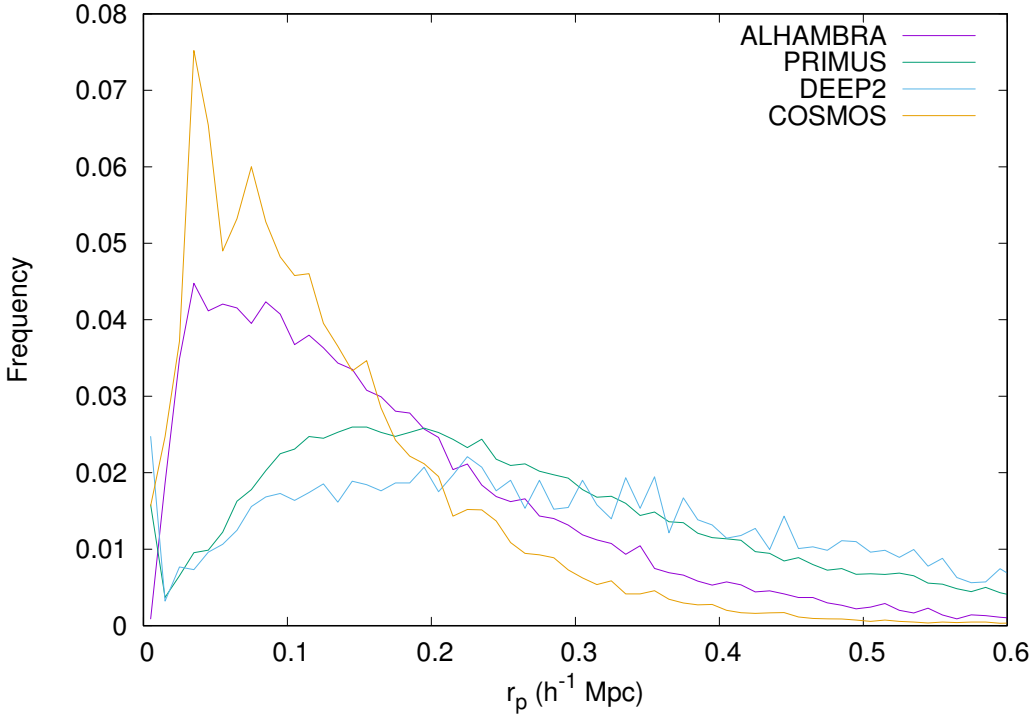

FIGURE 3.9: Near neighbor distribution for the ALHAMBRA, DEEP2, PRIMUS and COSMOS photometric surveys.

The near neighbor distributions are shown in Fig. 3.9. Distance to the nearest neighbor is calculated in projected distances (r_p) with $\pi_{max} = 200h^{-1}$ Mpc and lineal bins. While for ALHAMBRA and COSMOS, the nearest-neighbor distance peaks at scale $r_p < 50h^{-1}$ kpc, for the other two samples the peak occurs at scales $\sim 150h^{-1}kpc$.

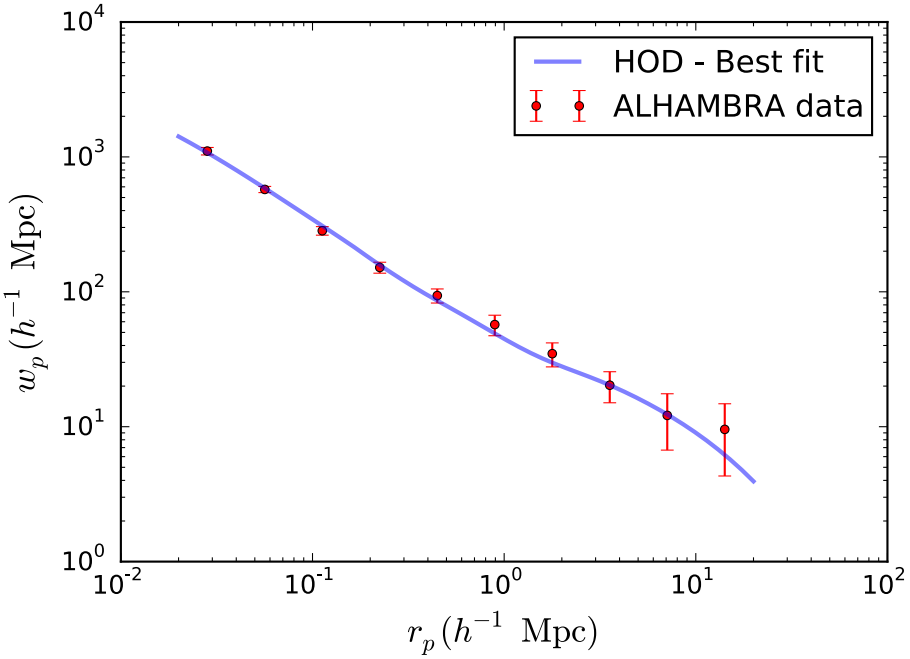

FIGURE 3.10: Halo Occupation Distribution with NFW profile and the projected correlation function of the ALHAMBRA survey. The agreement between both curves also at the small scales reinforces the NFW profile as a good model for galaxy clustering.

Measurements of the projected correlation function for large scales have been performed for many galaxy surveys obtaining very similar results and a general agreement with models. However, it is at the small scales we are studying, where the departures between data and models might occur. Using the Halo Occupation Distribution (HOD, Kravtsov et al. (2004)) and the Navarro-Frenk-White profile (NFW, Navarro et al. (1997)) one can predict the shape of the projected correlation function. However, it is unclear if the NFW profile is actually able to model the observed galaxy distribution (Merritt et al., 2006, Piscionere et al., 2014, Watson et al., 2012). In Arnalte-Mur et al. (in prep.), using the presented calculations with the ALHAMBRA survey, we expect to contribute to the solution this problem. In Fig. 3.10 we can see how the Halo Occupation Distribution together with the NFW profiles can satisfactorily reproduce the observed values of the projected correlation function for ALHAMBRA. For this calculation we have used a galaxy sample with redshift $0.55 < z < 0.85$ and $M_B^{th} < -18.6$, containing 16979 galaxies, as selected in Arnalte-Mur et al. (2014). The HOD is calculated as in Coupon et al. (2012).

3.4.3 Results

In this section we make use of the projected correlation function $w_p(r_p)$, introduced in section 3.1. First we present the results for the different samples described in section 3.4.1. This is done in subsection 3.4.3.1. We also present the analysis of the bias (subsection 3.4.3.4). The calculation has been performed for scales from 0.03 to $10.0\,h^{-1}$ Mpc for the projected correlation function and from 1.0 to $10.0\,h^{-1}$ Mpc for the bias. Fig. 3.11 shows the projected correlation function for the full samples. The first remarkable result that deserves to be pointed out is a clear change of the slope of the $w_p(r_p)$ functions around $r_p \sim 0.2\,h^{-1}$ Mpc, as already mentioned by Coil et al. (2006) and Coil et al. (2008).

In this section, we compare our results with previous works that studied the galaxy clustering and its dependence on spectral type or color in the redshift range $z \in [0, 1]$, as mentioned in the introduction. Given the luminosity selection of our sample (see section 3.4.1), in each case we use for comparison the published results for volume-limited samples with number density closest to $n = 10^{-2}\,h^3\,\mathrm{Mpc}^{-3}$. The number density of the samples shown in our comparisons are within 20% of this figure with two exceptions: the PRIMUS sample at $z \simeq 0.4$ (with number density of $n = 1.6 \times 10^{-2}\,h^3\,\mathrm{Mpc}^{-3}$, Skibba et al., 2014), and the VIPERS sample at $z \simeq 0.6$ (with number density of $n = 0.33 \times 10^{-2}\,h^3\,\mathrm{Mpc}^{-3}$, Marulli et al., 2013). In the case of Meneux et al. (2006), they use a flux-limited sample resulting in a evolving number density with redshift in the range $n = 0.33 - 1.2 \times 10^{-2}\,h^3\,\mathrm{Mpc}^{-3}$.

3.4.3.1 Power-law modeling

Power laws are simple and widely used models to describe the correlation function of the galaxy distributions, as they provide a very good approximation over a large range of scales with only two free parameters. The observed change of the slope mentioned above forced us to model the projected correlation function w_p by means of two power laws, one that fits the function at small scales and the other one at large scales. A similar treatment was done by Coil et al. (2006) in their analysis of the clustering in the DEEP2 survey at $z = 1$. The departure from power-law behavior at small scales can be explained naturally in the framework of the halo occupation distribution (HOD) model that considers the contribution to the correlation function of pairs within the same halo (one-halo term), which is

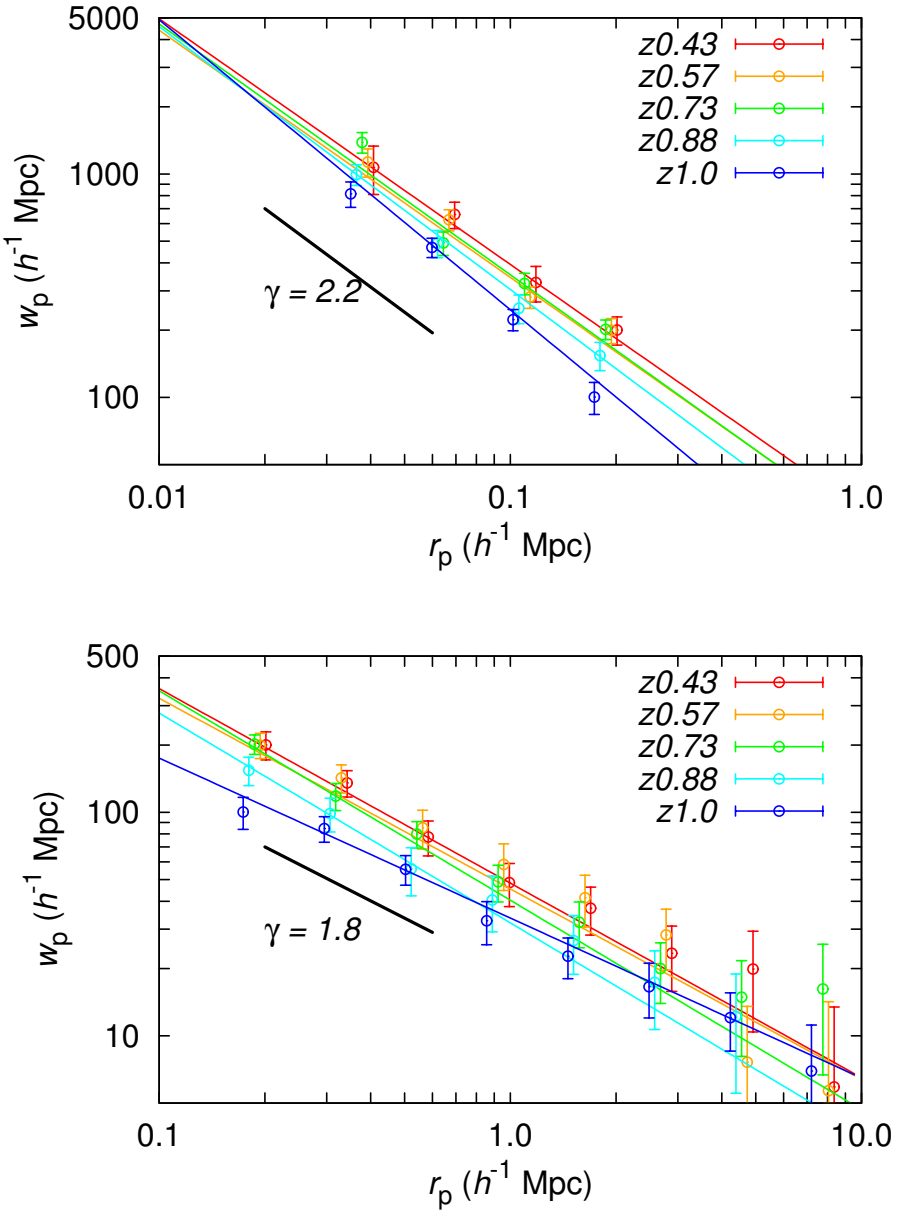

FIGURE 3.11: Projected correlation function for the full population sample in each of the redshift bins (points with error bars). Top: small scales ($0.03 < r_p < 0.2$). Bottom: large scales ($0.2 < r_p < 10.0$). Error bars are calculated with the delete-one jackknife method and values at the same r_p are shifted for clarity. Solid lines with matching colors show the best-fit power law in each case. The black segment represents the mean slope of the curves.

dominant at short scales, and the transition to the regime where the function is dominated by pairs from different halos (two-halo term), at large scales. We will present HOD fits to the ALHAMBRA data in a separate paper. Therefore, we fit two power laws as:

$$w_p^s(r_p) = Ar_p^{\beta}, \text{if } r_p \leq r_s \tag{3.17}$$

for the small scales, and

$$w_p^l(r_p) = Cr_p^{\delta}, \text{if } r_p \geq r_s \tag{3.18}$$

for the large ones. We fix value $r_s \simeq 0.2h^{-1}$ Mpc. An abrupt change in the projected correlation function has also been detected at this scale by Phleps et al. (2006) for the blue galaxies of the COMBO-17 sample. A, β, C and δ are the free parameters. We treat each power law independently and express them in terms of the equivalent model for the 3-dimensional correlation function ξ.

$$\xi^{\rm pl}(r) = \left(\frac{r}{r_0}\right)^{-\gamma} . \tag{3.19}$$

A and β (analogously, C and δ) can be related to the parameters γ (power-law index) and r_0 (correlation length) as shown in Davis & Peebles (1983):

$$A = r_0^{\gamma}\frac{\Gamma(0.5)\Gamma\left[0.5(\gamma-1)\right]}{\Gamma(0.5\gamma)} \ , \ \beta = 1-\gamma\,. \tag{3.20}$$

We have performed the fitting of this model to our data using a standard χ^2 method, by minimizing the quantity

$$\chi^2(r_0,\gamma) = \sum_{i=1}^{N_{bins}}\sum_{j=1}^{N_{bins}}(w_p(r_i) - w_p^{pw}(r_i))\cdot\Sigma_{ij}^{-1}\cdot(w_p(r_j) - w_p^{pw}(r_j))\,, \tag{3.21}$$

where Σ is the covariance matrix. We fit this model to our data at scales $0.03 \leq r_p \leq 0.2h^{-1}$ Mpc and $0.2 \leq r_p \leq 10.0h^{-1}$ Mpc for each sample using the covariance matrix computed from eq. 2.4, to obtain the best-fit values of r_0, γ and

their uncertainties (see Table 3.3). This fitting has been performed using the POWERFIT code developed by Matthews & Newman (2012).

We must remark that the statistical errors of the correlation function at different separations r_p are heavily correlated because a given large-scale structure adds pairs at many different distances. The higher the values of the off-diagonal terms of the covariance matrix, the stronger the correlations between the errors. When this happens the best fit parameters r_0 and γ might be affected as has been illustrated by Zehavi et al. (2004) for the SDSS survey. One could ignore the error correlations and use only the diagonal terms, but this is not justified if these terms are dominant. The parameters of the fits are listed in Table 3.3.

3.4.3.2 Full samples

Fig. 3.11 (top panel) shows the measurements of the projected correlation function $w_p(r_p)$ for the full samples at the small scales ($0.03 \leq r_p \leq 0.2h^{-1}$ Mpc). The bottom panel shows the same function for large scales ($0.2 \leq r_p \leq 10.0h^{-1}$ Mpc). Looking at both diagrams, we confirm the rise of the correlation function at small scales already detected by Coil et al. (2008) with values of $\gamma \sim 2.2$ (for the slope of the 3-dimensional correlation function). We can also appreciate in the top panel of Fig. 3.11 that the correlation functions are steeper for the high-redshift samples with values of γ increasing from ~ 2.1 for the closest redshift bin ($z \sim 0.4$) to ~ 2.3 for the farthest ($z \sim 1$). The correlation length significantly decreases with increasing redshift (see also Table 3.3).

For the large scales $0.2 \leq r_p \leq 10.0h^{-1}$ Mpc, the slope of the correlation function is rather constant for all samples with values around $\gamma = 1.8$, while again the correlation length decreases with redshift from $r_0 = 4.1 \pm 0.5$ for $z \sim 0.4$ to $r_0 = 3.5 \pm 0.3$ for $z \sim 1$. The evolution of the amplitude indicates that the change in clustering is mainly driven by the overall growth of structure in the matter density field. As we use for the fits the scales $0.2 < r_p < 10.0h^{-1}$ Mpc (the 2-halo term becoming important at scales $r_p > 1.0\,h^{-1}$ Mpc) the fact that the slope γ does not significantly change also implies that the 2-halo contribution for this population does not significantly change its profile over this redshift interval. All these effects were studied in detail in Arnalte-Mur et al. (2014) and extended to

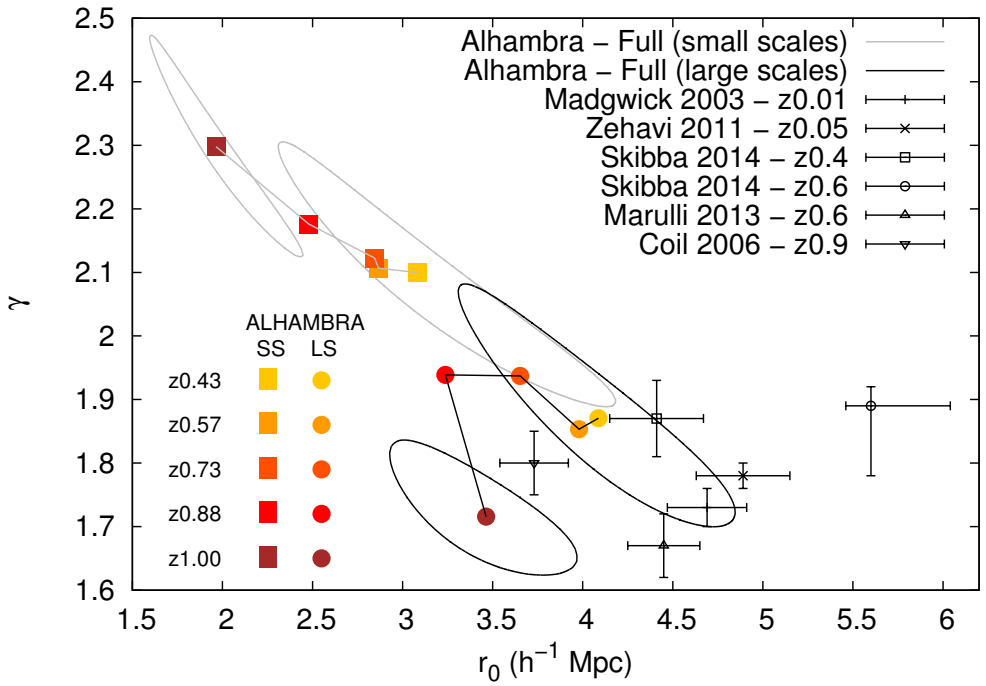

FIGURE 3.12: Parameters r_0, γ obtained from the power-law fit to the projected correlation functions of our full population samples. In black, the 1σ confidence regions of the large scales fit ($0.2 < r_p < 10.0h^{-1}$ Mpc) and in grey, the 1σ confidence regions of the small scales fit ($0.03 < r_p < 0.2h^{-1}$ Mpc). For clarity, we show only the regions for the first and last redshift bin. Lines link the best-fit results for each sample accross different redshift bins. For comparison, we show as points with error bars the results of Madgwick et al. (2003) (2dF), Zehavi et al. (2011) (SDSS), Coil et al. (2006) (DEEP2), Marulli et al. (2013) (VIPERS) and Skibba et al. (2014) (PRIMUS) (see the text for details). The parameters and their 1-sigma variation have been calculated using the method described in Sect. 3.4.3.1.

samples with different luminosities (see e.g. their Fig. 7). We have seen that this is only broken at shorter scales, where the curve presents slightly higher values.

The overall trend can be visualized in Fig. 3.12, where we show the evolution of the best-fit parameters of the 3-dimensional correlation function $\xi(r)$ for small and large scales in the full population samples. Despite the great uncertainties, the diagram shows evolution with r_0 decreasing for both scale ranges as redshift grows. In addition, at small scales, the slope γ also increases with redshift. The evolution, at large scales, of the correlation length extrapolates well to lower redshift with the value reported by Zehavi et al. (2011) for the SDSS and by Madgwick et al. (2003) for the 2dF galaxy redshift survey. Zehavi et al. (2011) analyzed the SDSS Main catalogue by means of the projected correlation function. They obtained

TABLE 3.3: Results of the different fits to $w(r_p)$: power law and bias models

Sample	Full population				
	r_0^s	γ^s	r_0^l	γ^l	b
z0.43	3.1 ± 0.6	2.1 ± 0.13	4.1 ± 0.5	1.87 ± 0.12	1.21 ± 0.14
z0.57	2.9 ± 0.4	2.11 ± 0.1	4 ± 0.5	1.85 ± 0.1	1.23 ± 0.17
z0.73	2.8 ± 0.5	2.12 ± 0.12	3.7 ± 0.4	1.94 ± 0.1	1.25 ± 0.14
z0.88	2.5 ± 0.4	2.18 ± 0.1	3.2 ± 0.7	1.94 ± 0.15	1.2 ± 0.5
z1.00	2 ± 0.3	2.3 ± 0.11	3.5 ± 0.3	1.72 ± 0.06	1.3 ± 0.13
	Quiescent galaxies				
z0.43	4 ± 1.2	2.11 ± 0.17	4.9 ± 0.7	1.89 ± 0.16	1.26 ± 0.19
z0.57	2.3 ± 0.5	2.62 ± 0.18	5.4 ± 0.8	1.85 ± 0.11	1.8 ± 0.2
z0.73	3.6 ± 0.7	2.29 ± 0.14	4.3 ± 0.7	2.15 ± 0.16	1.4 ± 0.2
z0.88	4 ± 0.9	2.25 ± 0.13	4.2 ± 0.8	2.14 ± 0.17	1.6 ± 0.3
z1.00	3.5 ± 0.9	2.28 ± 0.16	4.8 ± 0.8	1.8 ± 0.13	1.9 ± 0.3
	Star-forming galaxies				
z0.43	2.4 ± 1.7	2 ± 0.5	4.3 ± 0.5	1.66 ± 0.13	1.33 ± 0.18
z0.57	2.2 ± 0.7	2.1 ± 0.2	3.6 ± 0.4	1.73 ± 0.12	1.21 ± 0.17
z0.73	2.8 ± 0.9	2.1 ± 0.2	3.5 ± 0.4	1.86 ± 0.14	1.18 ± 0.14
z0.88	1.8 ± 0.5	2.3 ± 0.2	3 ± 0.4	1.7 ± 0.12	1.2 ± 0.4
z1.00	1.7 ± 0.3	2.34 ± 0.13	3.2 ± 0.3	1.69 ± 0.09	1.25 ± 0.13

Results of the fits of the power law model and the bias model to the data for each of our samples. r_0^s and γ^s correspond to the scales $0.03 < r_p < 0.2h^{-1}$ Mpc, and r_0^l and γ^l to the scales $0.2 < r_p < 10.0h^{-1}$ Mpc. These parameters have been calculated using the methods described in Sections 3.4.3.1 and 3.4.3.4.

values for the parameters r_0 and γ by the same method used here, over the scale range $0.1 < r_p < 50h^{-1}$ Mpc.

The values correspond to the galaxies selected in the luminosity bin $-20 < M_r < -19$ and $0.027 < z < 0.064$, with a number density ($n = 10.04 \times 10^{-3}\, h^3\, \mathrm{Mpc}^{-3}$) and typical luminosity ($L^{\mathrm{med}}/L^{\star} = 0.4$) similar to the ALHAMBRA sample used in this work, so this is, qualitatively, a valid comparison. As we can see in Fig. 3.12, the slope of the correlation function for the full SSDS main sample is $\gamma = 1.78 \pm 0.02$ compatible within one σ with the values obtained for the ALHAMBRA survey at higher redshift within the range of large scales analyzed here, and the correlation length $r_0 = 4.89 \pm 0.26$ follows the evolutionary trend delineated by the ALHAMBRA higher redshift samples: r_0 increases at lower redshifts. Very similar results

have been obtained by Madgwick et al. (2003) for the 2dFGRS with $\gamma = 1.73 \pm 0.03$ and $r_0 = 4.69 \pm 0.22$ within the range $0.2 < r_p < 20.0h^{-1}$ Mpc in the redshift interval $0.01 < z < 0.015$. At larger redshift our results can be compared with the ones reported by Skibba et al. (2014) for the PRIMUS survey. They have analyzed two bins of redshift $0.2 < z < 0.5$ and $0.5 < z < 1$ with $M_g < -19$. In Fig. 3.12 we have displayed their results for the correlation function parameters. We also plot a point corresponding to the VIMOS Public Extragalactic Redshift Survey (VIPERS) from Marulli et al. (2013) and another point corresponding to the DEEP2 survey from Coil et al. (2006). All these results, for the three high redshift surveys, show perfect agreement with our own ALHAMBRA results.

It is important to understand the correlation between parameters γ and r_0, as its interpretation can be delicate. If, for instance, γ grows with the redshift of the sample, r_0 will tend to reduce its value, as $\xi(r) = 1$ for shorter distances, as it can be appreciated in the top-left points (short r_p scales) displayed in Fig. 3.12. We must have this in mind for a proper understanding of our results. On the other hand, the decrease of r_0 with increasing redshift when γ does not change, as we find in bottom-right points (large scales) in Fig. 3.12, can be interpreted as a self-similar growth of the structure at the calculated scales. This effect is specially reflected in the tilt of the confidence ellipses in Fig. 3.12, which shows the negative correlation between r_0 and γ.

3.4.3.3 Segregated samples

Fig. 3.13 shows the projected correlation function $w_p(r_p)$ for the quiescent and star-forming galaxies at the five redshift bins, compared to the full population. As expected, the full population result occupies an intermediate position at low redshift, but evolves with redshift towards star-forming positions. This is expected due to the higher abundance of the latter in our samples, specially at high redshift. A visual inspection of Fig. 3.13 suggests that the projected correlation function shows the double slope corresponding to the 1-halo and the 2-halo terms, specially for the star-forming galaxies, due to their tendency to cluster in lower mass halos with smaller virial radii Seljak (2000).

Quiescent galaxies show a higher clustering at every redshift bin. In order to study the change of the clustering properties with redshift and spectral type, we fit the

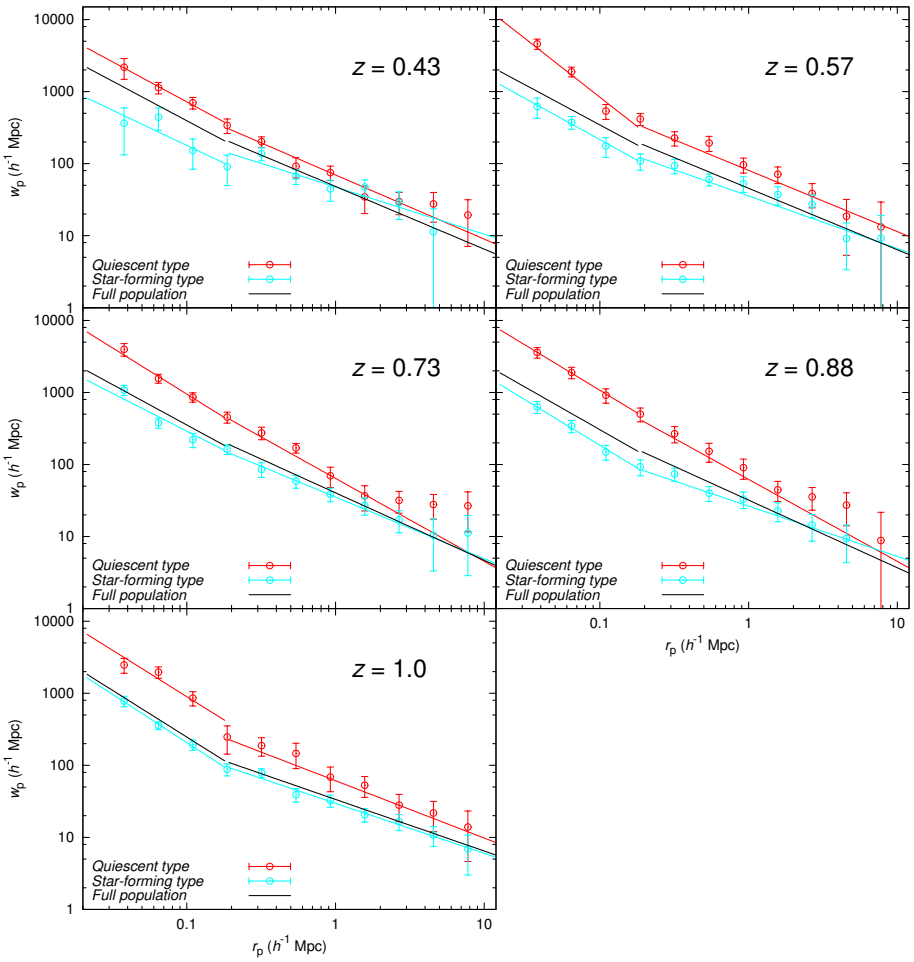

FIGURE 3.13: Projected correlation functions for quiescent (red) and star-forming (blue) galaxies (points with error bars). Solid lines with matching colors show the best-fit power law in each case. For reference, we also show the results for the full population with the continuous black line. From top to bottom, left to right, the five redshift bins: $(0.35 < z < 0.5)$, $(0.5 < z < 0.65)$, $(0.65 < z < 0.8)$, $(0.8 < z < 0.95)$ and $(0.95 < z < 1.1)$. Error bars are calculated with the delete-one jackknife method.

projected correlation function $w_p(r_p)$ of each sample with a power law model, using the method described above.

The amplitude of their correlation functions, as well as their slope, is higher than that for the star-forming galaxies in all cases. As for the full population we have modeled the correlation function with two different power laws at scales larger and smaller than $r_p = 0.2h^{-1}$ Mpc. Star-forming galaxies show for all redshift bins a clear rise in their correlation function at small separations. As mentioned in the introduction, Coil et al. (2008) found the same result for the bright blue

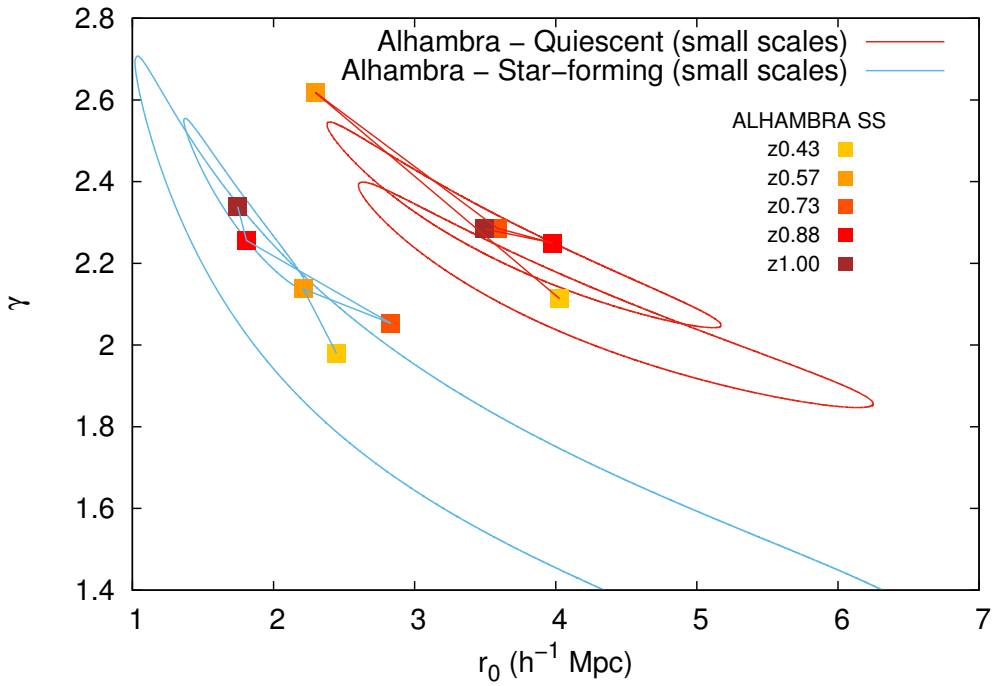

FIGURE 3.14: Parameters r_0, γ obtained from the power-law fit to the projected correlation functions of our spectral segregated samples for the small scales ($r_p \leq 0.2h^{-1}$). In red, the 1σ confidence regions of the quiescent galaxies fit and in blue, the 1σ confidence regions of the star-forming galaxies fit. For clarity, we show only the regions for the first and last redshift bin. Lines link the best-fit results for each sample across different redshift bins. The parameters and their 1-sigma variation have been calculated using the method described in Section 3.4.3.1

galaxies of the DEEP2 galaxy redshift survey. They found that the effect is more pronounced at higher redshift corresponding to brighter galaxies. For the quiescent galaxies we would have fitted a single power law for the whole range, in particular for some redshift bins. However we have proceeded in the same way for the two galaxy types in order to simplify the analysis of the segregation. The comparison of the best-fit model to the data in each case is shown in Fig. 3.13, and we see an excellent agreement in all cases. The parameters obtained from the fits are listed in Table 3.3.

As we have done for the full population, to visualize if there is any evolution of the correlation function parameters we show the diagram of γ vs. r_0 in Figs. 3.14 and 3.15 for the segregated populations with the corresponding confidence regions, separated in the two scale regimes. In both cases (short and large scales) the parameter space occupied by quiescent galaxies can be clearly distinguished from the

space occupied by star-forming galaxies, the first ones showing larger correlation length for both scaling ranges with the difference between both types well over 3σ for small scales and about 2σ for large scales.

At short scales the exponent of the correlation function γ is similar for both galaxy types with values around $\gamma \sim 2.2$ (Fig. 3.14). The correlation length for star-forming galaxies varies roughly in the range $r_0 = [2, 3]\, h^{-1}$ Mpc. A visual hint of evolution could be appreciated for the star-forming galaxies, with steeper correlation functions (and lower correlation lengths) for higher redshifts, nevertheless, given the large error bars (see Table 3.3), this trend is not really significant. The parameters of the correlation function for quiescent galaxies at small scales do not show any evolution at all with values for γ in the range $[2.1, 2.3]$ and correlation lengths in the range $r_0 = [3.5, 4.0]\, h^{-1}$ Mpc for all redshift bins except for the second bin ($z = 0.57$), which displays a higher value of the exponent γ and smaller r_0. The ALHAMBRA survey has allowed us to measure the behavior of the clustering properties of the segregated samples at these very short scales. These scales had not been previously studied with the detail that we are showing here because other samples cannot reliably estimate the correlation function at $r_p < 0.1 h^{-1}$ Mpc because they are not deep enough, or dense enough at these distance, due for example to fiber collisions in the case of spectroscopic surveys Guo et al. (2012b).

Fig. 3.15 shows the same results at scales $0.2 < r_p < 10.0 h^{-1}$ Mpc. For this scale range, we can compare with the results from other authors. We see again that the regions of parameter space occupied in the diagram for quiescent and star-forming galaxies are different. Star-forming galaxies present both lower exponent γ and lower correlation length r_0 than quiescent galaxies. These differences are significant at the 2σ level. The value of γ, exponent of the correlation function, is roughly constant for all redshift bins and is ~ 1.7. A hint of evolution can be seen in the correlation length, since r_0 decreases from $r_0 \sim 4.3$ to $r_0 \sim 3\, h^{-1}$ Mpc with increasing redshift, which corresponds to a $\sim 2\sigma$ change in r_0. The values of the correlation function parameters reported by other authors for different samples at lower and similar redshift are compatible with the ALHAMBRA results shown here.

In the diagram we see that our fits are consistent with the points corresponding to the correlation function parameters of the active galaxies from the 2dFGRS

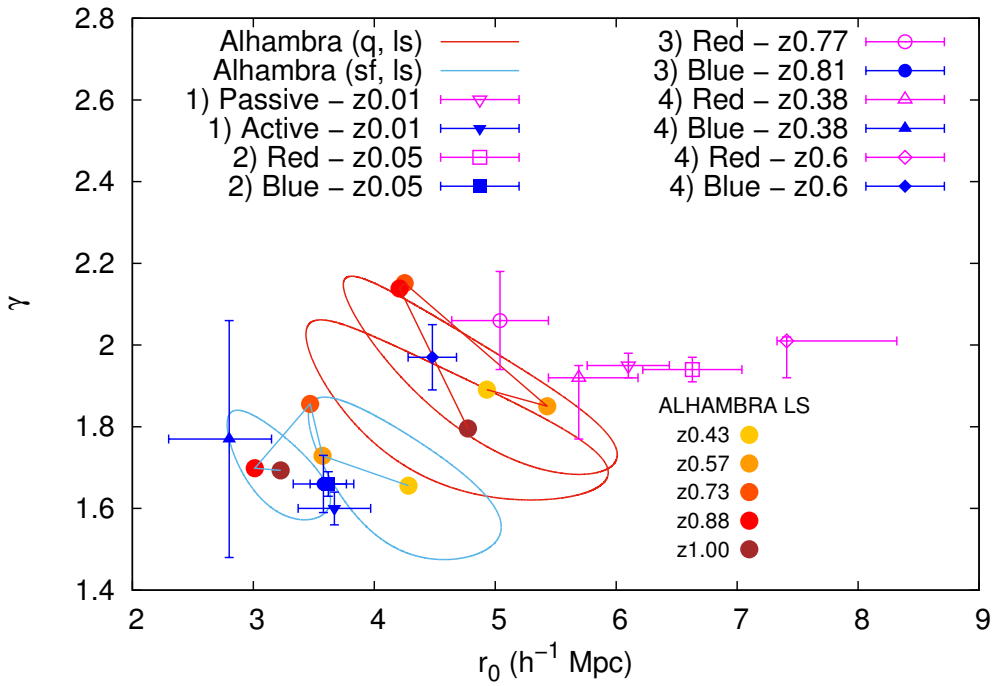

FIGURE 3.15: Parameters r_0, γ obtained from the power-law fit to the projected correlation functions of our spectral segregated samples. In red, the 1σ confidence regions of the quiescent galaxies fit and in blue, the 1σ confidence regions of the star-forming galaxies fit. For clarity, we show only the regions for the first and last redshift bin. Lines link the best-fit results for each sample across different redshift bins. The parameters and their 1-sigma variation have been calculated using the method described in Section 3.4.3.1. For comparison, we plot the results obtained by 1) Madgwick et al. (2003) (2dF) at $z \sim 0.01$, 2) Zehavi et al. (2011) (SDSS) at $z \sim 0.05$, 3) Coil et al. (2008) (DEEP2) at $z \sim 0.9$ and 4) Skibba et al. (2014) (PRIMUS) at $z \sim 0.38$ and $z \sim 0.6$ (see the text for details).

(Madgwick et al., 2003) at $z \sim 0.01$, a blue subsample drawn from the SDSS-main (Zehavi et al., 2011) at $z \sim 0.05$, a blue population of the DEEP2 redshift survey (Coil et al., 2008) at $z \sim 0.9$, and the blue sample from the PRIMUS survey (Skibba et al., 2014) at $z \sim 0.4$. This result also agrees with the qualitative behavior of the evolution of the correlation length reported by Meneux et al. (2006) from the VVDS sample where they conclude that the clustering amplitude of the late-type star-forming galaxies remains roughly constant since $z \sim 1.5$, although they found a slight rise of this amplitude at their larger redshift bin $1.2 < z < 2.0$. However, one should bear in mind that Meneux et al. (2006) use a flux-limited sample, so this evolution may be affected by the change in luminosity of the samples.The values of the correlation length reported by Meneux et al. (2006) are

slightly smaller than the values calculated here for the ALHAMBRA survey.

Quiescent galaxies show stronger clustering than late-type star-forming galaxies. Their correlation function parameters at large scales are nearly compatible within the errors with fixed values around $r_0 = 5\,h^{-1}$ Mpc and $\gamma = 2$, but the clustering length is smaller than the one calculated at low redshift by Madgwick et al. (2003) for passive galaxies in the 2dFGRS and by Zehavi et al. (2011) for the red galaxies in SDSS. Instead, the values of the amplitude of the correlation function reported by Skibba et al. (2014) at $z \sim 0.4$ for PRIMUS, by Coil et al. (2008) at $z \sim 0.9$ for the DEEP2 and by Meneux et al. (2006) agree with our results within the errors.

For both star-forming and passive galaxies, the only discrepant measurement in Fig. 3.15 is that corresponding to the PRIMUS samples at $z \sim 0.6$, which show values of r_0 significantly larger than those obtained by ALHAMBRA at similar redshifts (and also by DEEP2 at $z \sim 0.9$). This difference may be due to the fact that the PRIMUS survey includes the COSMOS field, which contains a large overdensity at this redshift affecting the clustering measurements (see the discussion in Section 3.4.1).

This segregation is generally explained by the tendency of red, quiescent or early-type galaxies to form in dense environments, while blue, star-forming or late-type galaxies typically form in the field or in low mass haloes (Bell et al., 2004, Dressler, 1980, Goto et al., 2003, McNaught-Roberts et al., 2014, Thomas et al., 2005, Zucca et al., 2009).

3.4.3.4 Dependence of the bias on spectral type and redshift

In order to disentangle the evolution of the galaxy clustering of different populations from the overall growth of structure, we study the bias b of our samples based on the projected correlation function measurements. We use a simple linear model, with a constant and scale-independent bias. In this model, the galaxy projected correlation function is given by

$$w_p(r_p) = b^2 w_p^m(r_p)\,, \tag{3.22}$$

where b is the bias, and $w_p^m(r_p)$ is the theoretical prediction for the projected correlation function of the matter distribution. Our model for w_p^m is based on ΛCDM with cosmological parameters consistent with the WMAP7 results (Komatsu et al., 2011), including a normalization of the power spectrum $\sigma_8 = 0.816$. The matter power spectrum at the median redshift of each sample is obtained using the CAMB software (Lewis et al., 2000), including the non-linear HALOFIT corrections (Smith et al., 2003). We obtain the real-space correlation function $\xi(r)$ by a Fourier transform of the matter power spectrum and the final projected correlation function w_p using eq. 3.15.

We fit this model to our data in the range $1.0 < r_p < 10.0\,h^{-1}$ Mpc, corresponding mainly to the two-halo term of the correlation function. The best fit value and uncertainty of the bias is obtained by the same method as described in Section 3.4.3.1 for the parameters of the power-law model. The results of these fits for each of our samples are listed in Table 3.3.

We show the evolution of the bias as a function of redshift for our different populations in the top panel of Fig. 3.16 (red and blue squares). As expected, we also see the effect of spectral segregation in this case, as the bias of early-type quiescent galaxies is consistently larger than that of late-type star-forming galaxies. The bias observed for the full population, not shown, is similar to that of the star-forming galaxies. For comparison we show, as solid lines, the bias for dark matter haloes of a fixed mass, according to the model of Tinker et al. (2005), and the values obtained by previous works for samples at similar redshift ranges and number densities. The bias values in those cases were obtained by a similar method as here, and using compatible scale ranges[5].

For the star-forming galaxies, we obtain that their bias is approximately constant, with values $b \simeq 1.25$, over the range we explore. This explains the evolution of r_0 at large scales observed in Fig. 3.15. If the bias is constant, the main driver for the evolution of the galaxy clustering amplitude is the growth factor, therefore r_0 grows with cosmic time, as observed. Given the uncertainties, and the relatively slow evolution of the halo bias, the measured bias in this case is also consistent with the evolution of the bias of haloes with mass in the range $M_{\rm h} \simeq 10^{11.5} - 10^{12}\,h^{-1}\,{\rm M}_\odot$. As shown in Fig. 3.16, our results for the star-forming population

[5]In the case of Madgwick et al. (2003), as the bias values are not given explicitly, we derived them using their power law best fit at a scale of $5h^{-1}$ Mpc.

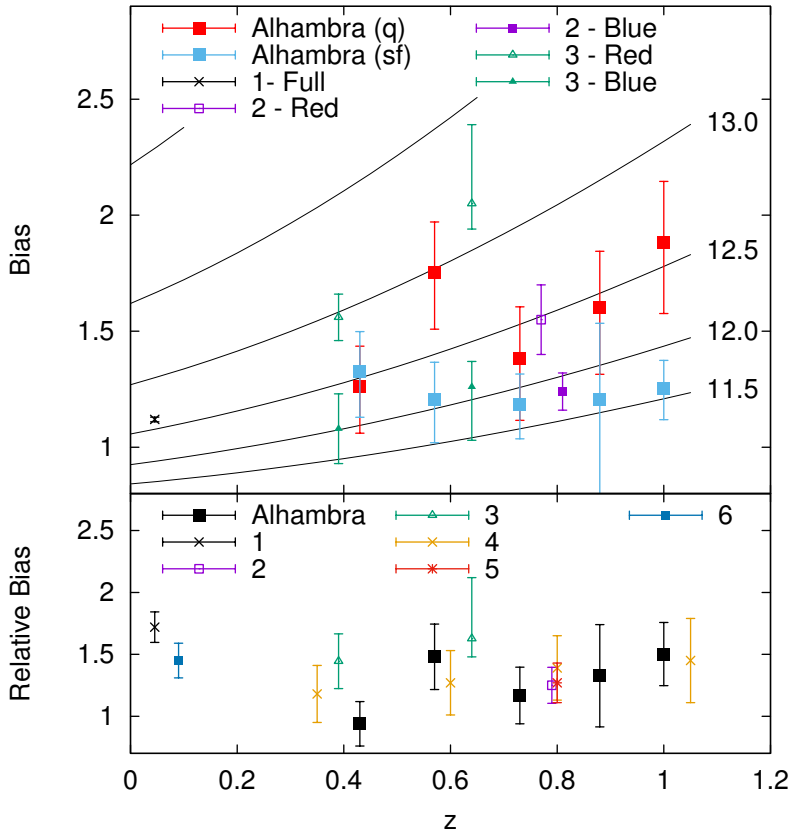

FIGURE 3.16: Galaxy bias of our quiescent (q) and star-forming (sf) galaxies as function of their median redshift. Bias is estimated by a fit to eq. (3.22), as described in Sect. 3.4.3.4. The solid lines correspond to the bias of haloes of a fixed mass, according to the model of Tinker et al. (2005). These lines are labelled with the corresponding halo mass in terms of $\log_{10}[M_{\rm h}/(h^{-1}\,{\rm M}_{\odot})]$. For comparison, we plot the results obtained by 1) Zehavi et al. (2011), 2) Coil et al. (2008), 3) Skibba et al. (2014), 4) Madgwick et al. (2003), 5) Marulli et al. (2013), 6) Meneux et al. (2006) and 7) de la Torre et al. (2011). Bias values from different authors have been adapted for the assumed cosmological parameters used in this paper.

are fully consistent with those obtained for similar populations of blue galaxies in the the DEEP2 (Coil et al., 2008) and PRIMUS (Skibba et al., 2014) surveys.

The bias of quiescent galaxies shows a clear evolution, increasing with redshift, which is remarkably similar to the expected evolution of the bias for haloes of mass $M_{\rm h} \simeq 10^{12.5}\,h^{-1}\,{\rm M}_{\odot}$. This clear evolution of the bias in this case compensates the clustering evolution due to the growth factor, resulting in an approximately constant value of r_0, as shown above in Fig. 3.15. Our results for this population are consistent with the observed bias for red galaxies in the DEEP2 and PRIMUS

surveys. We note that the bias measurement for the PRIMUS sample at $z \simeq 0.6$ may be affected by the presence of a large overdensity in the COSMOS field, as noted above. The larger number of bins in redshift used in this ALHAMBRA analysis allows us to see more clearly this evolutionary trend.

In the bottom panel of Fig. 3.16 we show the relative bias, defined as the ratio of the bias of quiescent galaxies over the bias of the star-forming ones, as function of redshift. We also show for comparison the results from previous surveys at similar redshifts including, in addition to those shown in the top panel, the VVDS survey (Meneux et al., 2006), and the zCOSMOS-Bright survey (de la Torre et al., 2011). Meneux et al. (2006) used flux-limited samples with evolving galaxy density so their absolute bias measurements are not comparable to ours. The analysis in de la Torre et al. (2011) only provided values of the relative bias of their samples.

At $z \sim 0$ we show the results from the 2dFGRS survey (Madgwick et al., 2003) and the SDSS Zehavi et al. (2011). We note that in the two low redshift cases, the relative bias is calculated using a slightly different method: Madgwick et al. (2003) calculate it using the ratio of the galaxy variances $\sigma_{8,\mathrm{gal}}$ of the samples, while we have calculated the relative bias for SDSS as the ratio of the best-fit power laws of the two samples at a scale $r = 5\,h^{-1}\,\mathrm{Mpc}$ (see, e.g. Eq. 9 of Norberg et al. (2002)).

We obtain values of the relative bias in the range $b_{\mathrm{rel}} \simeq 1 - 1.5$, consistent with all previous results at similar redshifts. The relative bias shows a very faint evolution, slightly increasing with redshift. However, given the errors, our results are also consistent with being constant. A similar, faint trend is also seen for the VVDS results of Meneux et al. (2006). However, if we include the $z \sim 0$ values, this evolution is broken, and the best description of the results is a constant relative bias with redshift.

Overall, our results indicate that, for samples selected by the same B-band luminosity and redshift, passive galaxies reside in haloes up to 10 times more massive than those hosting active galaxies. When studying the evolution with redshift, the observed bias suggests that quiescent galaxies (following a constant number density selection) reside in haloes of constant mass, while this is not clear in the case of star-forming galaxies. In the latter case, there seems to be an indication that they populate slightly more massive haloes at lower redshift. We will study

the relation between galaxies and dark matter haloes using a more detailed HOD modeling in a future work.

3.4.4 Conclusions

The ALHAMBRA survey allows us to perform accurate clustering calculations with different segregation criteria. Its 23-filter photometry provides reliable galaxy parameters with good completeness out to very high redshifts ($z \sim 1.10$), opening the possibility to analyze the galaxy clustering of different galaxy populations. In AM14 the authors chose to select galaxy samples using different luminosity thresholds, while in this work we made a selection by spectral type. This selection follows the spectral classification of the ALHAMBRA photometric templates and has been proved to match remarkably well the usual selection by broad-band color.

A rise of the correlation function at small scales is found as already noticed by Coil et al. (2008). We have been able to show that this trend holds at smaller scales and to characterize its redshift evolution.

Our sample allows us to measure the clustering properties of galaxy populations segregated by spectral type and their redshift evolution in an homogeneous way. At scales larger than $0.2\,h^{-1}$ Mpc, quiescent galaxies cluster with a higher amplitude than star-forming ones. The difference is significant at the 2σ level. There is also a significant hint of evolution (2σ) in the clustering amplitude of active galaxies, while the clustering of the passive ones remains constant. These results are compatible with previous works in the literature, but in the present work we have increased the redshift resolution.

Regarding the small scales ($r_p < 0.2\,h^{-1}$ Mpc) we find almost no change in the correlation function compared with the large scales in the quiescent population. On the other hand, the star-forming galaxies show a clear variation in the slope between small and large scales, which is possibly decreasing towards low redshifts.

Our measurements of the bias value for the different populations show strong segregation between them. The bias of the quiescent population clearly evolves with redshift following the expected behavior for haloes of approximate mass $10^{12.5}\,h^{-1}M_{\odot}$. The star-forming population bias remains basically constant over our observed redshift range, but it can still be compatible with the theoretical

evolution of lower mass haloes ($M_h \sim 10^{11.5-12}\, h^{-1} M_{\odot}$). As a consequence, the relative bias hints at a slow evolution, which would not be completely consistent with observations at $z \sim 0$.

These results have been published in Hurtado-Gil et al. (2016).

3.5 Future work: new correlation function estimation for photometric surveys

In the next years multiple galaxy surveys will be produced, providing us with an extremely extensive and complete dataset for cosmological research. Most of this information will be photometric, which implies a special dedication in terms of methodology and software development. These new methodologies must be focused to make full use of photometric information included in the *photo-redshift distribution function (zPDZ)*, this is, the posterior probability function describing the quality fit of our measurements to our galaxy templates. Generally, catalogs only publish the best fit plus the redshift uncertainty $\sigma_z/(1+z)$, together with a quality flag, which vary from project to project. In addition, conventional statistics, like the projected correlation function, are not prepared to deal with the complete information contained in a zPDZ. This restrains our use of galaxy data, limiting us the access to relevant science. In an attempt to correct this situation, we present a new methodology that makes full use of the zPDZ information for galaxy correlation calculations. This procedure follows a common philosophy with other photometric-redshift works, e.g. López-Sanjuan et al. (2015), in the ALHAMBRA collaboration.

The idea is, basically, to perform the calculations of functions $DD(r)$ and $DR(r)$ when the redshift of the galaxy is determined by a redshift probability function p (the posterior zPDZ) instead of using a single value. This distribution is generated collapsing the zPDZ values over the redshift dimension, summing the values for every possible template. The location of the galaxy is complete with the right ascension and declination of the galaxy. Using equation 1.23 we obtain the director vector (u, v, w) for $Dist = 1$.

Now, given 2 galaxies described as above, we want to know which is the probability of having them separated by a distance r. Using equation 1.15, for a

given redshift, we know that the distance at which a galaxy is located is a value $d \in [D_{min}, D_{max}]$, where D_{min} and D_{max} are the distances corresponding to the minimum and maximum redshift in our galaxy sample. Hence, we define A as the subset satisfying

$$(d_1 u_1 - d_2 u_2)^2 + (d_1 v_1 - d_2 v_2)^2 + (d_1 w_1 - d_2 w_2)^2 = r^2 \tag{3.23}$$

for pairs $(d_1, d_2) \in [D_{min}, D_{max}]^2$, where d_1 and d_2 correspond to the distances of our two galaxies. Since D_{min} and D_{max} are constant quantities for the galaxy population and r is chosen by the user, subset A is defined by the (α, δ) coordinates of each galaxy pair. Given the redshift probability functions p_1 and p_2 of our galaxies, we can compute the probability of having these galaxies in our range of interest with

$$I(p_1, p_2) = \int_A p_1(z') \cdot p_2(z'') dz' dz'' \tag{3.24}$$

This integral can be numerically calculated randomizing the $p_i(z)$ distributions with m elements, which turns the computation of $DD(r)$ into an $N_D^2 \cdot m^2$ calculation with four nested loops.

Notice that this can be easily adapted when, as in this thesis work, we study galaxy clustering as a marked point process. If we are only interested in the template types T_1 for the first galaxy, and T_2 for the second one, we only have to limit the integral with

$$I(P_1, P_2; T_1, T_2) = \int_{T_1 \times T_2} \int_A P_1(z', t') \cdot P_2(z'', t'') dz' dz'' dt' dt'' \tag{3.25}$$

where P_1 and P_2 are the zPDZ distributions without collapsing.

Finally, we can obtain the number of galaxy pairs in our sample at distance r just summing the probabilities I

$$DD(r) = \sum_{i=1}^{N_D} \sum_{j=1; i \neq j}^{N_D} I(p_i, p_j) \tag{3.26}$$

where N_D is the total number of galaxies. A template segregation version of this equation is also direct. Similarly, the calculation of $DR(r)$ can be deduced from the above formulae replacing one of the pdfs by a fixed location.

With this function, we would improve the reliability our estimation of the correlation function, but it will not eliminate the effects due to redshift distortions, making it necessary to adapt this zPDZ strategy to corrected estimators, such as the projected correlation function. We expect to apply this new statistic over coming galaxy surveys, such as the *Javalambre-Physics of the Accelerating Universe* (Benitez et al., 2014).

Part II

Modeling the Galaxy Distribution

'One interaction to rule them all
and into the probability density bind them'
Paraphrasing Sauron of Mordor

Chapter 4

Finite Gibbs processes

4.1 Introduction

Well known statistics like kernel density estimators provide valuable descriptions of the galaxy field, but might be considered excessively general, blind to the real nature of the structures behind the galaxy distribution. Summary statistics, like correlation functions, are capable of providing reliable and relevant information regarding isotropic structures, as clusters and BAOs, but shrink all information, losing the detail of particular structures.

Solving this problem implies formulating a model, a map of the process that determines the abundance of points for every given region. As said in section 1.2, we aim to build a probabilistic model, which understands this abundance as a random variable and can be evaluated for every location in the process, point or not. This model needs to be fitted to the observed data. This involves simplification, because a model must be tractable and comprehensible (Baddeley, 2007). Modeling is probably the most difficult of the three approaches of this thesis work, and yet, a conclusive analysis of a dataset is usually possible only by modeling. This might be a pending task in modern cosmology, to provide an efficient probabilistic model of galaxy distribution.

The development of point process theory in the late years has brought a whole new branch of statistical analysis (Chiu et al., 2013, Moller & Waagepetersen, 2003, Van Lieshout, 2000). As seen in section 1.2.1, the well known methodologies

appear integrated together with theorems already applied in different sciences with success (Baddeley & Turner, 2005). However, many methodological problems still need to be solved in order to numerically test the full capabilities of new models, and therefore, deeper efforts on statistics would be needed to obtain more versatile and adaptive algorithms.

As a first approximative step in this direction, we test several point process models over galaxy catalogs. This work consists on making the first steps on galaxy distribution modeling. For this reason we focus on original techniques, applied in cosmology for the first time, while using well known data catalogs from highly manipulated data sources. The chosen models are examples of Gibbs models, parametric field estimators based on point interactions. Unlike kernel density estimators, in this new functions we do not only have *irregular* parameters, but also fitted parameters which provide real information of the galaxy sample.

With these models we expect to provide both a characterization of the studied point process and a tractable expression of the points expectancy at every location. The parameters involved in our probabilistic model adopt certain values depending on the nature of the process, quantifying some of its most notable properties, such as being clustered or the profile of these clusters. Once the model is properly fitted, we have solved the point process problem. As introduced in section 1.2.2 we can say a process is modeled when we are able to predict the expected number of points in any given region. This can be extended to these locations not occupied by original points of the data, picturing the nature of the process beyond the particular realization of our studied dataset and allowing us to simulate and reproduce data-like patterns. The advantages are diverse and will be explained in the conclusions (section 4.4.6).

The validation of a point process model is another challenging task, specially for 3-dimensional datasets. The residual analysis methodologies published in Baddeley et al. (2005) and other works of the same authors (Baddeley & Turner, 2005) are the most suitable techniques to perform this validation. New code has been developed by the author of this thesis, adapting the residual analysis algorithms to the 3-dimensional case. With these methods we expect to detect possible disagreements between data and model as well as to understand the weaknesses of these models.

In section 4.2 we introduce the Finite Gibbs models, probabilistic models that describe the distribution of points in a pattern based in different user defined interactions. This will require a deep introduction to these models with examples and functions used in the field (sections 4.2.1 and 4.2.2). That includes the definition of conditional probabilities and fitting techniques (sections 4.2.3 and 4.2.4). The analysis of residuals is also an important contribution of point process methodologies that will be used again in Chapter 5. In section 4.4 we introduce the data used from SDSS-DR8 catalog and new samples extracted from LasDamas simulations. Finally, in section 4.4.3 we test our models over generated toy models before using them over the presented data sets (sections 4.4.4 and 4.4.5). Conclusions can be found in section 4.4.6.

4.2 Definition

Finite Gibbs models are probabilistic descriptors of point processes that base their modeling in trend or field density and point interaction. Processes described this way include a wide variety of different patterns, although we are only going to focus on a few of them. However, they all follow the same motivation: the detection and characterization of interactions between points. The analysis of the interactions or forces between points is a constant in this work and a central issue in any clustering study. An introduction to the point process statistics of the Gibbs models can be found in Illian et al. (2008) and, for an advanced discussion, consult Baddeley et al. (2015b), Baddeley & Turner (2005), Baddeley et al. (2013).

The previous methods, the Counts-in-Cells and the correlation function, allow us to characterize the intensity and nature of the interaction between points from an independent point of view that does not require to know the forces acting in the pattern. But with a model based on a Gibbs process, we can attempt to model the points distributions from their interactions, directly modeling its properties and validating it.

These kind of processes can be understood as a generalization of the independent processes. For a Poisson process the existence or not of a point in a certain volume of the space is independent of the distribution of the rest of the population. When this is not the case and points interact with each other, we can model it with a

Gibbs process. This is clearly the case of the galaxy distribution, where galaxies interact with each other through gravity and other processes, and therefore the probability of finding a galaxy is higher in a high density region, like a cluster, or lower in the voids. It is through probability densities and conditional probabilities that Gibbs processes are usually studied.

This branch of statistical analysis has already been tried satisfactory in cosmology (Tempel et al., 2014) in the modeling of complex structures like the galaxy filaments of the SDSS. In this work, we will only deal with the most conventional Gibbs models already tested in other fields, like biology or particle physics, and the battery of functions, methods and tests used to analyze and fit a point process.

4.2.1 Probability density function of a Gibbs process

Let be $X = \{x_i\}_{i=1}^{N}$ a spatial point process consisting of unordered location of points in a bounded region W. The process can be defined by a multivariate probability density $f_n(x_1, ..., x_n)$ for $x_1, ..., x_n \in W$, which expresses the probability of configuration X.

Notice that this function can be evaluated for any subset of points belonging to the region W. Generally in the construction of the patterns it is assumed that f_n is a probability density with respect to the unit rate Poisson process on W. This can be understood with the following interpretation: generate a realization $\{x_1, ..., x_n\}$ from a binomial process in W with n points. Accept each point x_i with a probability proportional to f_n. The accepted point pattern exactly follows the density function f_n.

However, this is a general definition difficult to manipulate and we will progressively introduce assumptions and simplifications until we define the Gibbs models used in this work. The first basic assumption introduced to avoid degenerate cases in our processes is the positivity condition, which states:

$$\text{if } f(X) > 0 \text{ and } Y \subset X, \text{ then } f(Y) > 0$$

Now we can start to introduce different known processes in a Gibbs shape. An homogeneous Poisson process probability density could be defined through the

function $f(x_1, ..., x_n) = \alpha\beta^n$, where α represents the normalizing constant and β is the intensity of the process. We generalize this definition and define an inhomogeneous Poisson process as

$$f(x_1, ..., x_n) = \alpha \prod_{i=1}^{n} b(x_i) \tag{4.1}$$

where $b(u)$ is the process intensity function defined in $u \in W$. This function is also called the 'activity' or the 'trend' of the process and it might be used to model any external influence in the distribution, containing information regarding the density of points. In a terrestrial example, imagine a forest in a valley with a variable soil composition that alters the growing of the trees, then this different soil should be mapped by the trend.

However, these models do not include interactions between points. When limited to the study of pairs of points with spatial locations, we can define the Gibbs model for *pairwise interaction points*

$$f(x_1, ..., x_n) = \alpha \prod_{i=1}^{n} b(x_i) \prod_{i<j} h(x_i, x_j) \tag{4.2}$$

where $h(x_i, x_j)$ is the 'interaction' function determining which is the probability for a pair of points to coexist in the locations (x_i, x_j), containing information regarding the association of points.

Generally, and specially in physics, the forces that drive the interactions do not have a limited range of action. But, as in gravity, some of these forces can be negligible at large distances, creating a range at which we can assume no interactions. This range is similar to the correlation range of the correlation function, and it could be introduced in the Gibbs model, simplifying its use and expression. Therefore, any Gibbs process with a range of interaction r will behave as a Poisson process for pairs separated by a distance larger than r. When this is done we say we have a *Markov point process*. This is an important property that states that the probability of finding a set of points in a particular set of locations $\{x_i\}_{i=1}^{n} \in W$ depends only on the already existing points around these location. For a single location $x \in W$ we can say

$$f_1(x|X) = f_1(x|\zeta \cap X) \tag{4.3}$$

where ζ is the neighborhood of the location x. This can be easily generalized changing x by $\{x_i\}_{i=1}^n$ and ζ by $\bigcup_{i=1}^n \zeta_i$.

4.2.2 Examples of Gibbs models

For the sake of clarity we present some common examples of Markov point processes with interactions. Processes defined over the relative distances between points are called Pairwise Gibbs models. The Strauss, Geyer and BadGey models are examples of discrete pairwise models, while the Fiksel and Power law models are continuous pairwise models. As an example of non pairwise model we introduce the Area interaction model.

When possible, we include examples of 2-dimensional samples of the mentioned models generated with Metropolis-Hastings algorithm using *spatstat* (Baddeley & Turner, 2005).

Strauss model

The Strauss model (Baddeley & Turner, 2000a) is a process which produces regular patterns where pairs are prevented to appear when they are separated by a distance closer than r with a probability γ. Its interaction function is therefore,

$$h_\gamma(u,v) = \begin{cases} \gamma, & \text{if } ||u-v|| < r \\ 1, & \text{otherwise} \end{cases}$$

The probability density function simplifies in

$$f(x) = \alpha\beta(x)\gamma^{s(x|X)} \tag{4.4}$$

where $s(x|X) = |\{(x,u) : ||x-u|| < r\}|$ is the number of points with distance to x below r. In the extreme case, when $\gamma = 1$, we have a perfect Poisson process with intensity function β. On the other hand, if $\gamma = 0$, no pairs will be allowed to exist if they are separated by a distance smaller than r, this is called a *hardcore* process

(already mentioned in section 3.1). Strauss processes can be used for example to model tree patterns where due to competition for the soil resources it is unlikely for a tree to grow close to another one. Due to its regular nature we will not use this model in the cosmological scenario, but it is always interesting to introduce it as a basic case. A 2-dimensional representation can be seen in Fig. 4.1.

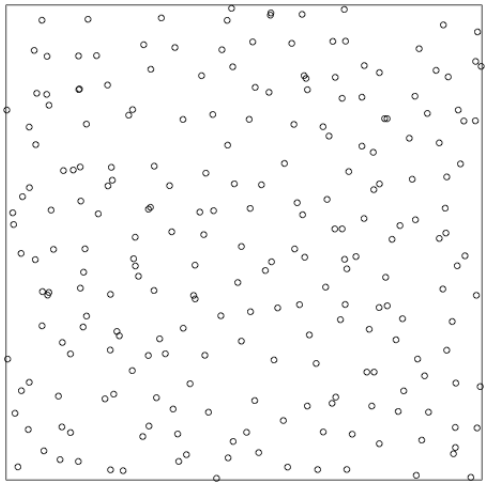

FIGURE 4.1: Realization of a Strauss process in two dimensions with $\beta = 2$, $\gamma = 0.2$ and $r = 3$. Box $= [0, 50]^2$.

Geyer model

A generalization of the Strauss model was made in order to include clustered patterns. Geyer (1999) developed the *Geyer* model, which includes the possibility of a $\gamma > 1$, turning the Strauss process into a clustering process. Since in this case the probability density is not integrable, a saturation threshold must be imposed, and the probability density function is written as

$$f(x) = \alpha\beta(x)\gamma^{min(sat, s(x|X))} \tag{4.5}$$

where *sat* is the saturation threshold chosen for our population. The value γ is a measure of the level of clustering or rejection existing on a given range r, but we must take into account that all pairs inside this range are counted equally. If the range is larger than the size of a structure, and therefore it includes uncorrelated

pairs, our measure of γ will be affected. A graphical 3-dimensional example of this model can be found in section 4.4.3.2. In Fig. 4.2 we present two examples in two dimensions, showing some of the diverse structures that this model can reproduce.

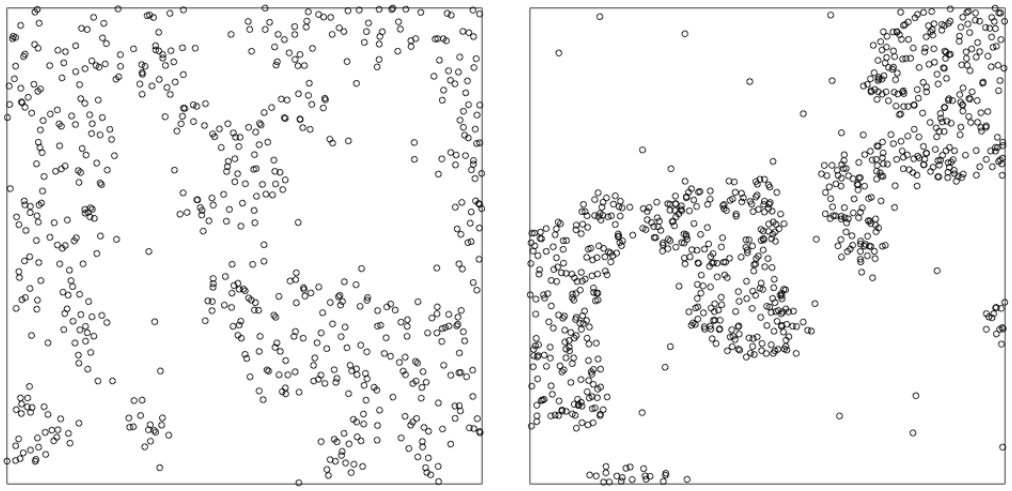

FIGURE 4.2: Realization of a Geyer process in two dimensions. Left: $\beta = 1.25$, $\gamma = 1.4$, $r = 3$ and $sat = 10$. Right: $\beta = 0.75$, $\gamma = 1.1$, $r = 3$ and $sat = 10$. Box $= [0, 50]^2$.

This model can be as well generalized in the multiradial expression of the *BadGey* model (Baddeley et al., 2015b). This generalization consists in considering several superimposed profiles with different correlation ranges and specific saturation thresholds for each one of them. The new interaction function is a vector of m Geyer interaction functions. As in the Strauss model:

$$h_{\gamma_i}(u, v) = \begin{cases} \gamma_i, & \text{if } ||u - v|| < r_i \\ 1, & \text{otherwise} \end{cases}$$

with $i \in \{1, ..., m\}$. The joint probability density function is written as the product of the m functions:

$$f(x) = \alpha\beta(x) \prod_{i=1}^{m} \gamma_i^{min(sat_i, s_i(x|X))} \tag{4.6}$$

Notice how, if the radii are ordered $r_1 < ... < r_m$, any radius r_i is included in the correlation range of the following profiles with $r_j > r_i$. This implies that the γ_i values are not independent. The BadGey model is specially interesting when our structure presents an outskirt distribution, with several levels of clustering (or rejection) growing from the center. Among the spherically symmetric structures, this model is limited to cluster-like structures, where this monotony in the profile can be found.

Even if it is interesting to separate the different clustering levels of a structure, clusters present a continuous profile. For this reason continuous pairwise Gibbs models were developed.

Fiksel model

A continuous example of interaction is the Fiksel model (Fiksel, 1984). This model produces a clustered pattern where interactions appear concentrated in overdensities. For a pair with separation distance $||u - v|| = d$, its interaction function is the following:

$$h_{a,\kappa}(u,v) = \begin{cases} 0, & \text{if } d < r_0 \\ \exp\left(a \cdot \exp\left(-\kappa \cdot d\right)\right), & \text{if } r_0 < d < r_1 \\ 1, & \text{if } r_1 < d \end{cases}$$

This model counts with two interaction distances: r_0 works as a hardcore distance, not allowing any closer interaction, and r_1, which works as a Markov distance, limiting the effects of the interaction. In the middle range an exponential function evaluates higher when closer to 0, increasing the strength of the interaction for the closer pairs. For pairwise models we need to calculate the probability for each pair of points using equation 4.2. No simplification through functions like the $s(x|X)$ function in the Strauss model is available:

$$f(x) = \alpha\beta(x) \prod_{i=1}^{N} h_{a,\kappa}(x, x_i) \tag{4.7}$$

Parameters a and κ describe the amplitude and the slope of the interaction. For high values of a the interaction will be stronger, while for high values of κ, the slope will be steeper. An example of this process is showed in Fig. 4.3.

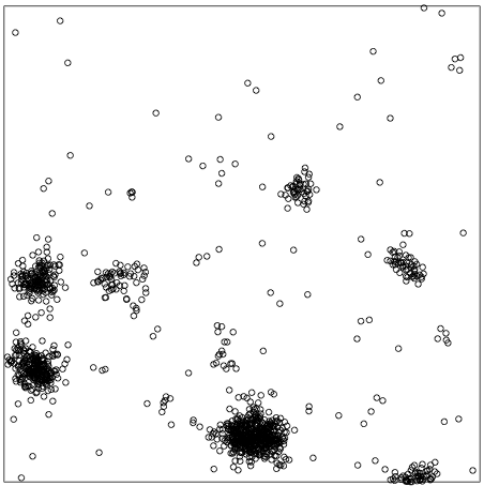

FIGURE 4.3: Realization of a Fiksel process in two dimensions with $\beta = 2.5$, $\kappa = 0.2$, $a = 1.6$, $r_0 = 0.07$ and $r_1 = 3$. Box $= [0, 50]^2$.

Power-law model

Power laws have been longly used in cosmology to fit the pair correlation function of a galaxy distribution. Motivated by the good agreement generally found in bibliography, we develop a model following the same profile. However, as seen in the previous models, several requirements have to be fulfilled. Markov models assume no correlations beyond a certain range, which implies that the profile function $h(u, v)$ must be equal 1 for distances $d = ||u - v||$ greater than the correlation range. For this reason we will exponentiate the power law. In addition, since decreasing power laws have a pole at $d = 0$, another caution must be taken: a hardcore distance must be imposed for distances $d < r_0$. Below this range the interaction function will remain undefined and pairs will be ignored.

$$h_{a,b}(u,v) = \begin{cases} 0, & \text{if } d < r_0 \\ \exp\left(a \cdot d^{-b}\right), & \text{if } r_0 < d < r_1 \\ 1, & \text{if } r_1 < d \end{cases}$$

with r_1 being the correlation threshold as usual and b and a our free parameters. As before, the resulting distribution is

$$f(x) = \alpha\beta(x) \prod_{i=1}^{N} h_{a,b}(x, x_i) \tag{4.8}$$

Similarly to the Fiksel model, parameters a and b describe the amplitude and the slope of the interaction. For high values of a the interaction will be stronger, while for high values of γ, the slope will be steeper. This model is an original contribution of this thesis work and the PowerLaw model is not completely usable in the *spatstat* code. For this reason we will not show a realization of this process as we have done for the rest of models.

Area-interaction model

Beyond the pairwise models, when only two points participate in each interaction, some models are able to describe higher-order interactions, were the number of involved points is higher than two. One example is the area-interaction process, with probability density

$$f(x_1, ..., x_n) = \alpha \prod_{i=1}^{n} \beta(x_i)\gamma^{-A(x_1,...,x_n)} \tag{4.9}$$

In this model, the same structure is followed as in the Geyer model, but the function $A(x)$ denotes the volume of the region obtained by drawing a sphere of radius r centered at each point x_i of the distribution, and taking the total volume of the union of these spheres. The more regular the pattern is, the higher this volume would be, since the sphere will overlap less.

It has an interesting difference with respect to the Geyer process: as before, a $\gamma < 1$ indicates a regular processes and $\gamma = 1$ a Poisson process. But this time the clustering case of $\gamma > 1$ is integrable and can be perfectly used. Despite these

promising properties, we leave the use of this distribution for a future work. As an example, we show in Fig. 4.4 a realization of this model.

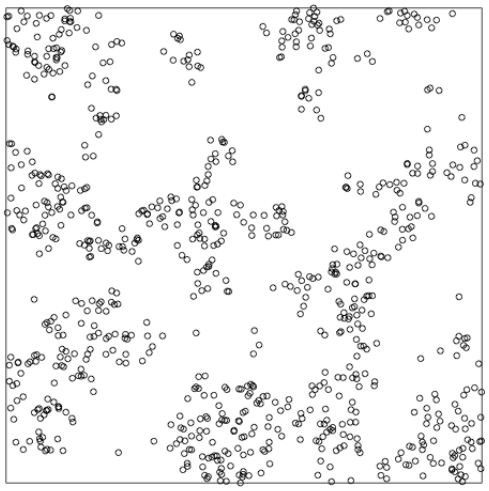

FIGURE 4.4: Realization of an Area-interaction process in two dimensions with $\beta = 4$, $\gamma = 5.6$ and $r = 3$. Box $= [0, 50]^2$.

4.2.3 The Papangelou conditional intensity

The above presented are only a few simple processes among the multiple used Gibbs models in different sciences, specially biology. But for many of the Gibbs models, the probability density function could be too complicated or difficult to treat. Interactions between points can be much more easily treated using conditional probabilities. For this purpose the *Papangelou conditional intensity* was elaborated (Baddeley et al., 2005, Papangelou, 1974), which evaluates the probability of finding a point in a given location given the rest of the distribution. For a spatiotemporal process, this corresponds to basing the prediction of a future event in the known past or 'history'. In an unordered space of two or higher dimensions, the history must be the existing pattern.

Hence, for a point process X contained in a region W, we define the Papangelou conditional intensity $\lambda(u, X)$ in the location $u \notin X$ as

$$\lambda(u, X) = f(\{u\} \cup X)/f(X) \tag{4.10}$$

taken $0/0 = 0$. This probability can be defined as well for locations already occupied by the pattern:

$$\lambda(x, X) = f(X)/f(X \setminus \{x\}) \tag{4.11}$$

This notation greatly simplifies the expressions, since most of times the difference between the numerator and the denominator is a single factor in the product and the rest gets cancelled. In addition, the normalization constant α vanishes as well, which facilitates many numerical simulation procedures. For interaction free models, like the Poisson process, the conditional intensity is $\lambda(u, X) = b(u)$, and for pairwise interactions we have

$$\lambda(u, X) = b(u) \prod_{i=1}^{n} c(u, x_i) \tag{4.12}$$

where $c(u, x_i)$ is the corresponding interaction function $h(u, x)$ at location u over the whole set X. Thanks to the positivity condition it is assured that the Papangelou conditional intensity λ of a finite point process uniquely determines its probability density f and vice versa.

As we will see in the next section, these conditional intensities can be fitted in their log-linear shape. We introduce this log-linear expression in general and for the above introduced models. This allows us to separate the conditional intensity into first-order and higher-order terms:

$$\log\lambda_\theta(u; X) = \eta \cdot T(u) + \phi \cdot V(u; X) \tag{4.13}$$

The first term corresponds to the intensity or trend term, $\beta(u)$ in eq 4.2, and the second one to the interaction function. η and ϕ are the logarithm of the parameters that linearize with the logarithm. Real function $T(u)$ is a non parametric statistic describing the trend of the pattern, while $V(u; X)$ is a vector function evaluating the quantities needed to determine the interaction of u with X. In the case of pairwise models, these quantities are based on relative distances of u with the rest of points in the process. Functions $T(u)$ and $V(u; X)$ are usually known as the sufficient statistics.

The following are the expressions for the conditional intensity and its logarithm for our models from section 4.2.2. For the Strauss model (eq. 4.4):

$$\lambda(u,X) = \beta(u)\gamma^{t(u;X)}, \qquad \log\lambda(u;X) = \log\beta + (\log\gamma)t(u;X) \tag{4.14}$$

where $t(u;X) = s(X \cup \{u\}) - s(X)$ is the number of points of X that lie within a distance r of location u. Here $\eta = \log\beta$ and $\phi = \log\gamma$. The Geyer model (eq. 4.5) has a similar expression, with $t(u;X) = min(sat, s(X \cup \{u\}) - s(X))$.

For the Fiksel (eq. 4.7) and the Power-law model (eq. 4.8):

$$\lambda(u,X) = \beta(u)\prod_{i=1}^{N} e^{a\cdot e^{-\kappa\cdot d_i}}, \qquad \log\lambda(u;X) = \log\beta + a\cdot\sum_{i=1}^{N} e^{-\kappa\cdot d_i} \tag{4.15}$$

$$\lambda(u,X) = \beta(u)\prod_{i=1}^{N} e^{a\cdot d_i^{-b}}, \qquad \log\lambda(u;X) = \log\beta + a\cdot\sum_{i=1}^{N} d_i^{-b} \tag{4.16}$$

where d_i is the distance between u and the rest of data points in the pattern. Here $\eta = \log\beta$ and $\phi = a$, κ and b are not linearized.

And for the area-interaction model, the expression linearizes similarly to the Strauss model:

$$\lambda(u,X) = \beta(u)\gamma^{B(u;X)}, \qquad \log\lambda(u;X) = \log\beta + (\log\gamma)B(u;X) \tag{4.17}$$

where $B(u;X) = A(X \cup \{u\}) - A(X)$ is the area of that part of the disc of radius r centered on u that is not covered by the discs of radius r centered at the other points $x_i \in X$. The area of $B(u)$ can be understood as the area that only belongs to u, if $\gamma < 1$ a new point will be less likely to exist in the area if it is small, and the other way round for $\gamma > 1$.

4.2.4 The pseudolikelihood

Now that we have the basics of the Gibbs processes theory that we are going to use in this thesis work, it is time to introduce the statistics analysis machinery

needed to estimate the parameters employed in our models. Given a model describing a spatial population, the most generally used technique for fitting the parameters is the maximum likelihood. However, maximum likelihood is computationally intensive, and employs simulation algorithms that are specific to the chosen model. This is specially costly for inhomogeneous and interaction patterns due to increased parameter dimensionality and the complexity of simulation. Generally, the estimation of the normalization constant α is an intractable function of the parameters θ, due to discontinuities in the irregular parameters (such as r). Nevertheless, as previously said, it disappears with the Papangelou conditional function, and we can try an estimation method based on this conditional probability. Despite a likelihood estimation is intractable due to the discontinuity of its variables (mainly r), an alternative approximation was found by Besag (1975), who proposed the pseudolikelihood estimator, which satisfies unbiased estimating equations and is consistent and asymptotically normal under suitable conditions.

Originally, it was defined for a finite set of random variables $X_1, ..., X_n$ as the product of the conditional likelihoods of each X_i given the other variables $\{X_j, j \neq i\}$. But Besag et al. (1982) extended the pseudolikelihood to point processes, for which it can be viewed as an infinite product of infinitesimal condition probabilities. Given a point process with conditional intensity $\lambda_\theta(u, X)$ over a subset $A \subseteq W$, the pseudolikelihood is defined as

$$PL_A(\theta, X) = \left(\prod_{x_i \in A} \lambda_\theta(x_i, X) \right) \exp\left(- \int_A \lambda_\theta(u, X) du \right) \tag{4.18}$$

If the process is Poisson, the pseudolikelihood coincides with the likelihood. For 'weak interactions', in the sense that $\lambda_\theta(u; X)$ can be approximated well by a function of u only, the process is approximately Poisson and the pseudolikelihood is a good approximation to the likelihood. Strong interactions may produce incorrect results (Baddeley & Turner, 2000a).

$$PL(\theta, X) = \left(\prod_{i=1}^{N} b_\theta(x_i) \prod_{i \neq j} h_\theta(x_i, x_j) \right) \exp\left(- \int_W b_\theta(u) \prod_{i=1}^{N} h_\theta(u, x_i) du \right) \tag{4.19}$$

In practice, the pseudolikelihood requires a numerical device to compute its estimations. This method, the Baddeley-Turner device, can be fully consulted in (Baddeley & Turner, 2000b):

The first step of this estimation consists in the approximation of the integral of the conditional function, the exponentiated term in the previous expression. Let X be a Gibbs point process with conditional intensity $\lambda_\theta(u; X)$, we can approximate this integral with a quadrature rule

$$\int_W \lambda_\theta(u; X)du \approx \sum_{j=1}^{m} \lambda_\theta(u_j; X) \cdot w_j \tag{4.20}$$

and we can express our numerical approximation of the pseudolikelihood, here in logarithmical shape

$$\log PL(\theta; X) \approx \sum_{i=1}^{N} log\lambda_\theta(x_i, X) - \sum_{j=1}^{m} \lambda_\theta(u_j; X) \cdot w_j \tag{4.21}$$

where $\{u_j\}_{j=1}^m$, are locations in W and $w_j > 0$ quadrature weights summing up to $|W|$. The $\{u_j\}_{j=1}^m$ can be understood as a grid of points covering the entire region W. This grid is used to perform a numerical integration, generally by the midpoint law, and one can distribute it equally spaced in the region W or adaptively if we already know the subregions where most of the intensity of λ is concentrated.

This is more compactly expressed if the $\{u_j\}_{j=1}^m$ set includes the N elements of X:

$$\log PL(\theta; X) \approx \sum_{j=1}^{m} (y_j log\lambda_\theta - \lambda_\theta) \cdot w_j \tag{4.22}$$

where λ is evaluated both in the data points x_i and in the grid or dummy points u_j. We define $y_j = z_j/w_j$, and

$$z_j = \begin{cases} 1, & \text{if } u_j \text{ is a data point, } u_j \in X \\ 0, & \text{if } u_j \text{ is a dummy point, } u_j \notin X \end{cases}$$

where w_j are the weights. These weights correspond to the cell occupied by a data or dummy point in the window W, as is customary in the quadrature integrations. However, when this point is close to a data point, the shape of the cell should be affected and the weight must be altered. Baddeley & Turner (2000b) propose several ways to calculate these weights. Through this thesis work, we have opted for the Voronoi tessellation. Once the data and the dummy points have been put together in the quadrature scheme, a Voronoi tessellation is created in the region W, being the volume of each tile the weight w_j in the integration. If the dummy points in the quadrature form a perfect grid, the tile of these points will be a perfect cube if they are away of any data point, but if a data point lies close to them the Voronoi tile will adapt its shape accordingly.

Expression 4.22 can be maximized using standard software for fitting generalized linear models (McCullagh & Nelder, 1989) like the function *glm.fit* found in the general CRAN package *stats*. The whole machinery for fitting Gibbs models can be found as well in the CRAN packages *spatstat* (Baddeley et al., 2015b, Baddeley & Turner, 2005), including the pseudolikelihood estimation. This latter package is only fully implemented for 2-dimensional processes. We have developed the necessary code generalizing it for the fitting of 3-dimensional point processes.

However, as introduced in section 4.2.3, if we want to fit the parameters of the conditional intensity model $\lambda_\theta(u; X)$ with the pseudolikelihood, we need it to be loglinear in the parameters θ. We can reexpress: equation 4.13 in a more compact way:

$$\log \lambda_\theta(u; X) = \theta \cdot S(u; X) \tag{4.23}$$

where $S(u; X) = (T(u), V(u; X))$ is a real-valued (if the model has no interactions) or vector-valued function at location u for a distribution X. Parameters $\theta = (\eta, \phi)$ appearing in the loglinear form are called 'regular', while all other parameters are 'irregular'. As seen, parameters β and γ are generally regular and any parameter involved in the functions of the γ exponent is irregular, like the range of interaction r. For example, for the Strauss model, parameters $\log \beta$ and $\log \gamma$ are regular parameters, but its range of interaction is irregular. For the Fiksel and the Power-law model, only the parameter a is regular, while κ, b and the two ranges r_0 and r_1 are irregular again.

The fitting of these irregular parameters is unclear and there is not a general method for their estimation. A possible strategy consists on fitting the regular parameters for a grid of irregular parameters and choosing the best option, but for high number of irregular parameters it could be highly expensive.

4.3 Point processes residuals analysis

The fitting of a model for a point process requires a sufficient battery of tests to check the quality of our estimation. In this section we present several tests intended to test the overall quality of the fit over the entire pattern, identifying anomalies or departures in the fitting with respect to the model, which allows us to analyze the quality of the test for the different structures present in the pattern. A comparison between the quality fit of the different models is also interesting.

The main tool for the model validation is the residual, a quantity comparing the observation with the model. Residuals can be used in different ways depending on which aspect of the model or the process we wish to highlight.

Pearson's χ^2 estimator

This statistic is not based on the analysis of residuals, but the function involved (the estimation of the number of points in a given volume) is essentially the same function used in the residual analysis. We decide to include this estimator here.

Once our probability density function or conditional intensity function has been estimated, it can be normalized using the total number of points in the pattern so it predicts the number of points in a given region of the space. Then we have to integrate the conditional intensity function in the proper set of regions and compare with the real number of points using a Pearson's χ^2 estimator.

In Kuhn et al. (2014), the authors integrate the model in n regions with approximately equal number of counts. These regions are obtained by tesselating the window using adaptive polygonal cells. However, for this thesis work, we would require of an advanced Voronoi programming for three dimensions and this is beyond our scope. We just propose this goodness of fit test here for future applications.

If N_i is the number of points for each cell $\{A_i\}_{i=1}^m$ and $\mathbb{E}(A_i)$ is the expected number of points, then

$$X^2 = \sum_{i=1}^{m} \frac{(N_i - \mathbb{E}(A_i))^2}{N_i} \tag{4.24}$$

The estimation of the number of galaxies in a given region can be obtained by simply integrating $\mathbb{E}(A_i) = \int_{A_i} \lambda(u)du$. We calculate p-values assuming X^2 is χ^2 distributed with m degrees of freedom.

4.3.1 Local residuals

The residuals analysis strategy in this work follows that of Baddeley et al. (2005) and is developed in code in the CRAN package *spatstat*. Residuals can be defined in multiple ways and they require appropriate plots and transformations for assessing each component of the fitted model. This allows us to identify peculiarities in the residual map related to the structures in the pattern. The behavior of the fit at different locations can be also inferred from this map.

The calculation of the residuals keeps the idea behind the *observation minus predicted* comparison used in the Pearson's χ^2 estimator. This is the natural generalization of the residuals for point process in time, exactly as we generalized the conditional intensity function λ using the known points in a process as history for the prediction of the location of a new one.

Three different kinds of residuals can be estimated for a given location $u \in W$ in a point process X with N data points:

Raw residuals are defined as the absolute difference between the real number of points in a region B and our estimation for the same region. These are the residuals used in the χ^2 estimator. For a model $\hat{\lambda}_\theta$

$$R_{\hat{\theta}}(B) = n(X \cap B) - \int_B \hat{\lambda}_\theta(u; X)\mathrm{d}u \tag{4.25}$$

where $n(X \cap B)$ is N_i, the number of data points in the region B. These residuals create an atomized distribution when regions B are small enough, being close to 1

at the data points and $-\hat{\lambda}_\theta(u;X)$ at all other locations u in W. For exact models, the average of the raw residuals should be zero.

Following the general strategy of the point processes analysis, the residuals can be evaluated as well at the empty location u. The *absence* of points at these locations is also informative. Through this work we will only use the raw residuals, but it is also interesting to introduce other alternatives.

Inverse residuals are based in the Stoyan-Grabarnik (Stoyan & Grabarnik, 1991) exponential energy marks (eem), a diagnostic tool that gives the 'mark' $m_i = 1/\hat{\lambda}_\theta(u;X)$ to every point x_i in the data pattern. These marks have the property that, for well fitted models, the sum of all marks in a region B has expected value equal to the area of B. Using this, we can define the inverse residuals $I(B)$ as

$$I_{\hat{\theta}}(B) = \sum_{x_i \in X \cap B} \frac{1}{\hat{\lambda}_\theta(x_i;X)} - \int_B \mathbf{1}\{\hat{\lambda}_\theta(u;X) > 0\}\mathrm{d}u \tag{4.26}$$

The first term is the sum over the exponential energy marks for every point in a region B and the second term is the area of the region B where $\hat{\lambda}_\theta$ is defined positive. The conditional intensity function is always non negative but, if existing, the zero points must be avoided in the eem. These residuals can be obtained as well dividing the raw residuals by $\hat{\lambda}_\theta$.

But Baddeley & Turner still propose a third kind of residuals, the *Pearson residuals*. Now the residuals will be calculated using the square root of the conditional intensity function

$$P_{\hat{\theta}}(B) = \sum_{x_i \in X \cap B} \frac{1}{\sqrt{\hat{\lambda}_\theta(x_i;X)}} - \int_B \frac{1}{\sqrt{\hat{\lambda}_\theta(u;X)}}\mathrm{d}u \tag{4.27}$$

Again $\hat{\lambda}_\theta(u;X) > 0$ is necessary, but it can be zero at $u \notin X$

Smoothed residuals plots

The above defined residuals can be calculated for all $u \in W$ points and therefore we can create plots to visualize the quality of our fitting. If the model is well fitted, one expects to find an uncorrelated distribution in the residual map, with no correlations between the values of the residuals and the locations of the data

points. Residuals must be of low amplitude compared with the conditional surface density function and must oscillate symmetrically around zero. This thesis work performs all its calculations in 3-dimensional situations, and the plot of any kind of map must be done after integrating the values of the non projected dimension over the projected ones. Maps of this kind can be dealt as matrices with values corresponding to the integrals of the non projected dimension and plotted assigning an intensity color scale for the range of values present in the matrix.

However, for the proper visualization of the residual map a smoothing should be done before the integration. As said with the raw residuals, when evaluated for a large amount of u points in W the map may appear too atomized, showing values close to 1 when we are close to a data point, and close to 0 for the rest. As we know that the sum of the residuals should be approximately zero, we can proceed with a smoothing of the map, understanding that the shape of the structures in the pattern and its general trend are bigger than the area occupied by our numerical integration cells.

Hence, we proceed with the smoothing of our raw residual map. A spatial distribution X in W with N points can be smoothed by

$$\lambda^*(u) = \sum_{i=1}^{N} \kappa_\omega(u - x_i), \qquad u \in W \tag{4.28}$$

as found in Martínez & Saar (2002). The density field λ^* is an estimation of the local density of the distribution for every location u that involves all the existing points x_i in X. The kernel function used in this work is the Gaussian filter, standardly used in cosmology,

$$\kappa_\omega(\mathbf{y}) = \frac{1}{(2\pi)^{3/2}\omega^3} \exp\left(-\frac{|\mathbf{y}|^2}{2\omega^2} \right) \tag{4.29}$$

where ω is the smoothing radius or bandwidth. The result of the smoothing strongly depends on the choice of this quantity. The choice of this quantity might be a complex calculation or can be estimated by more heuristic means until the choice reveals the desired patterns. As an orientation, a small bandwidth will create spurious structures smaller that the real structures present in the map,

adding noise. In the other hand, if the bandwidth is too big, it will blur the map eliminating interesting features that we need to properly analyze the residuals.

Now we are ready to define the smooth residual map used in our thesis work. As found in Baddeley et al. (2005):

$$s(u) = \lambda^*(u) - \lambda^\dagger(u) \tag{4.30}$$

is the smoothing map of the residuals where $\lambda^\dagger(u)$ is a smoothed version of the parametric estimate of the intensity according to the fitted model,

$$\lambda^\dagger(u) = \int_W \kappa_\omega(u-v)\hat{\lambda}_\theta(v)\mathrm{d}v \tag{4.31}$$

This definition allows us to study our model in detail, analyzing the quality of the fit for every desired region, checking if it has overfitted or underfitted the model or if the shape of the conditional intensity is not adequate. Using $\lambda^*(u)$ and $\lambda^\dagger(u)$, new quantities can be calculated to obtain new interesting insights. One is the relative error $e(u)$, which tests the quality of the fit at u normalizing by the amplitude of the model.

$$e(u) = s(u)/\lambda^\dagger(u) \tag{4.32}$$

For any structure properly mapped by the model, $s(u)$ will be a small quantity and the values of $e(u)$ in its region will be low, but if a structure is not included in the model, the lack of fitting will be included in $s(u)$ and amplified by $\lambda^\dagger(u)$. This function highlights regions where the content of the process is poorly modeled with respect to the average fitting quality of the whole process. Therefore, with function $e(u)$ we can detect structures that our model cannot properly model. For example, given a sample of a galaxy population with clusters and filaments, the relative error $e(u)$ of a model meant to describe clusters but not filaments, should show high values at locations where the filaments are present.

Fitting a point process is not just about finding a model that produces 'small' residuals, but also verify that no structures, as small as they could be, are outside the model.

The values of functions $\lambda^\dagger$, $s(u)$ and $e(u)$ will be usually presented in plots summed along the dimension Z. These residual plots use a chromatic scale from dark blue to dark red, with white for zero values. Red color indicates an underestimating, while blue should be understood as overestimating.

The general quality of the fitting can also be tested comparing the amount of not modeled mass with the total mass. We call this quantity the amplitude reduction (AR), a real function that gives us the proportion of the data included in the model. For a flat model $\lambda^\dagger(u) = 0$, $AR = 0$, and for a perfect model $\lambda^\dagger(u) = \lambda^*(u)$, $AR = 1$.

$$AR = 1 - \frac{\int_W |s(u)|du}{\int_W \lambda^*(u)du} \tag{4.33}$$

Lurking plots

Finally, Baddeley et al. (2005) provide us with a versatile tool to detect poor fits in user defined regions. This can be useful is we suspect that the data may depend on a covariate not included in the model. If the covariate is properly included in the model, the lurking plots should be uncorrelated (with values around zero), and a significant departure should be understood as a correlation, and therefore, these covariates still have to be included in the model. Let $Z(u)$ be a spatial covariate of our interest defined in W, we define the region

$$B(z) = \{u \in W : Z(u) \leq z\} \tag{4.34}$$

This region contains all locations u in W with values of our covariate Z smaller or equal to z. For different ordered values of z we create several regions $B(z)$ and evaluate the residual map for each one. For example, for the raw residuals we have

$$A(z) = n(X \cap B(z)) - \int_{B(z)} \hat{\lambda}(u, X)\ \mathrm{d}u \tag{4.35}$$

And now we can plot $A(z)$ against z to obtain the dependence of our residuals with the covariate $Z(u)$. The most common lurking plot takes $Z(u)$ as one of the axis, for example $Z(x, y, z) = x$, which plot the evolution of the residuals along the x dimension. In our works, we define sets $B(z)$ including regions between pairs of

consecutive z values. Therefore, the resulting quantities are the differentiation of function $A(z)$.

This statistic will be used only in Chapter 5.

4.4 Data catalogs

Our Gibbs models will be tested in both the SDSS real catalog and LasDamas simulation. We select several subsamples following the same criterium, making them comparable populations.

4.4.1 Sloan Digital Sky Survey - DR8

This survey has already been used in section 2.1, and we will make use of it again in this section.

In section 4.4.3 we apply the discussed Gibbs models over galaxy populations. That demands a galaxy survey with certain properties. The geometry of the survey should be extensive enough to select a comfortable sample where analysis could be made with minimum mask effects. Density population is also important since we need to select a comparable population where all galaxies belong to the same distribution.

For these reasons we opted to use the SDSS-DR8 data set (Aihara et al., 2011). This catalog was used by Tempel et al. (2012) to construct a prepared dataset. This data is a selection of the main contiguous area of the survey (the Legacy Survey). The survey is complete up to the Petrosian magnitude $m_r = 17.77$ (Strauss et al., 2002), which is used as lower magnitude limit of the sample, a limit applied after the Galactic extinction correction. Around a 6% of the galaxies in this survey are without observed spectra due to fibre collisions (the minimum separation between spectroscopic fibers is 55"). However, this only affects to close separation pairs and we meant to study bigger structures. The population provided by Tempel et al. (2012) is already cleaned of duplicated entries, stars or other artifacts, creating an appropriate sample for our point process algorithms. SDSS-DR8 is a very suitable dataset for point process analysis of the large scale structure (Einasto et al., 2016).

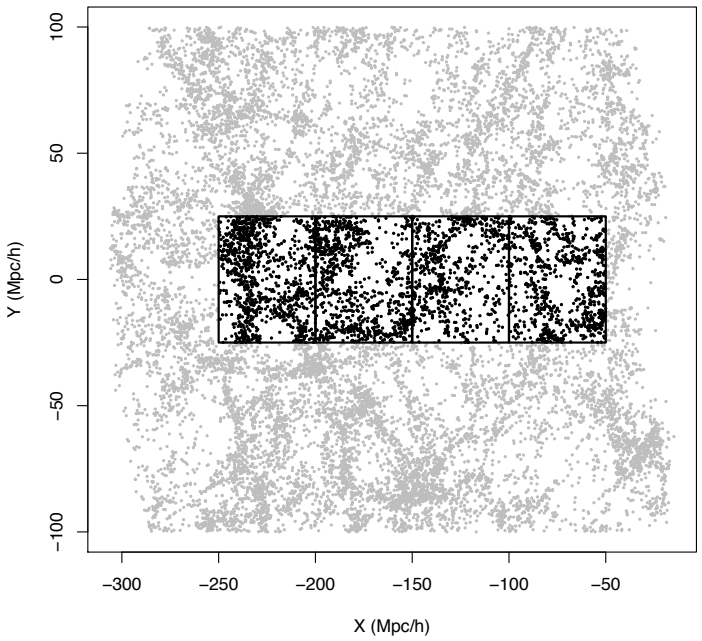

FIGURE 4.5: In black, samples from the SDSS-DR8 used in the Gibbs models testing. Observer is at (0,0).

In order to ensure comparable data in our galaxy samples, we will impose a luminosity threshold on our galaxies with $M_r < -20$ on SDSS r band. This limit keeps completeness for redshift $z < 0.1$.

From this catalog we select four contiguous boxes of side size $50h^{-1}$ Mpc, occupying the $200h^{-1}$ Mpc between redshifts 0.02 and 0.085. This allow us to test our models in a variety of samples and compare. We named these samples $DR8_1$, $DR8_2$, $DR8_3$ and $DR8_4$ from closest to furthest. Number of galaxies and densities per cube are summarized in Table 4.1.

4.4.2 LasDamas Simulation catalog

For comparison with simulated datasets, we will make use of the LasDamas simulation catalog (McBride et al., 2011) presented already in section 2.3.2. The used mock catalog was obtained from the *Esmeralda* simulation, contained in the LasDamas Gamma Release, with $M_r < -20.0$ and spanning from $z = 0.02$ to

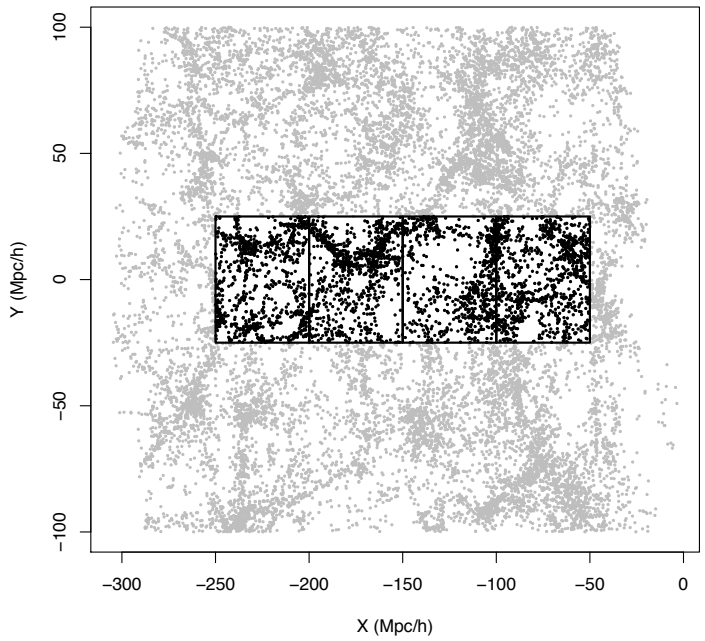

FIGURE 4.6: Samples from LasDamas simulations used in the Gibbs models testing.

$z = 0.106$. We have only made use of one of the multiple realizations, containing 1250^3 particles of mass $0.931 \times 10^{10} h^{-1} M_{\odot}$ in a box of size $640 h^{-1}$ Mpc.

As with the SDSS samples, from this realization we have selected four contiguous boxes of side size $50 h^{-1}$ Mpc, occupying the $200 h^{-1}$ Mpc between redshifts 0.02 and 0.085. We named the samples LD_1, LD_2, LD_3 and LD_4 from closest to furthest. Number of particles and densities per cube are summarized in Table 4.2.

4.4.3 Testing with Toy Models

In this section, we test the techniques presented in section 4.2 over galaxy populations. We apply the Geyer, Fiksel and Power law models over the described populations, obtaining conclusions from the fitted parameters. The residuals analysis also provide us with a powerful mechanism for testing the results.

The simulation of point processes following a given distribution is a key procedure in point process analysis. We start with a toy model example for the Geyer model, where we can see the capabilities of our methods.

4.4.3.1 Generation of samples

In order to test our models and analyze their effectiveness in data modeling, we generate several data samples following known distributions with known parameters. As found in Illian et al. (2008) (page 146) several Gibbs processes can be easily generated using the Birth & Death algorithm. We start with an initial point configuration (generally a poisson process). Then we delete one of these points at random (death) and substitute it with a new point (birth) generated accordingly to a conditional density function,

$$\phi(x) = \exp\Big(- \sum_{j=1, j\neq k}^{n} \Phi(||x - x_j||)\Big) \tag{4.36}$$

where x is the new point and Φ is our sufficient statistic (eq. 4.13), which depends on the Gibbs model. Point x is generated uniformly in W along with an independent uniform random number $\xi \in [0, 1]$. However, we will only accept x as a member of the point process after a rejection sampling, i.e., if $\xi \leq \phi(x)$. We have used this algorithm for generating different patterns.

For a $\gamma-$Geyer process (eq. 4.14), the non-normalized conditional density function is

$$\phi(x) = \exp\Big(- \alpha \sum_{j=1, j\neq k}^{n} \mathbf{1}(0 < ||x - x_j|| \leq r)\Big) = \exp\Big(\alpha \cdot t(x; r)\Big) \tag{4.37}$$

where $\alpha = \log \gamma$. This means that, given the process X and a candidate element x, we accept it in the γ-Geyer process if $\xi \leq \phi(x) = \gamma^{min(sat, t(x,r))}$ where $min(sat, t(x, r))$ is the minimum between the saturation value sat and $t(u; X)$, the number of elements from X closer to x than r (the above summation expression).

4.4.3.2 Fitting the sample

As a first test of these point process distributions, we have generated a sample following the Geyer method. The geometry of the process is a cubic box $[0, 50]^3$. We will fit the parameters of the distributions knowing the interaction radii through a pseudo-likelihood estimation (see section 4.2.4). A few concerns must be taken

into account. The pseudo-likelihood is a delicate estimator and no confidence intervals are included in this calculation. In addition, the Birth-Death generation method requires a high number of iterations before convergence into the desired pattern, which makes it difficult to generate clustered processes.

After the fitting, we can calculate over a grid of thickness $h = 1$, the conditional intensity $\lambda(u, X)$ (and its smoothed version $\lambda^{\dagger}(u)$), the smoothed raw residuals $s(u)$ and the relative errors $e(u)$ (see section 4.3). After evaluating these functions on our cubic grid, we project the values summing each row into a 2D map. I present here the resulting plots when we project the dimension Z.

The generated sample includes 500 points following a Geyer point process with $\gamma = 1.5$, interaction radius $r = 5$ and saturation $sat = 10$. The fitted results obtained by the pseudo-likelihood are $\beta = 0.001$ and $\gamma = 1.38$, confirming the clustered pattern. These parameters give us an amplitude reduction of $AR = 0.83$ (see eq. 4.33).

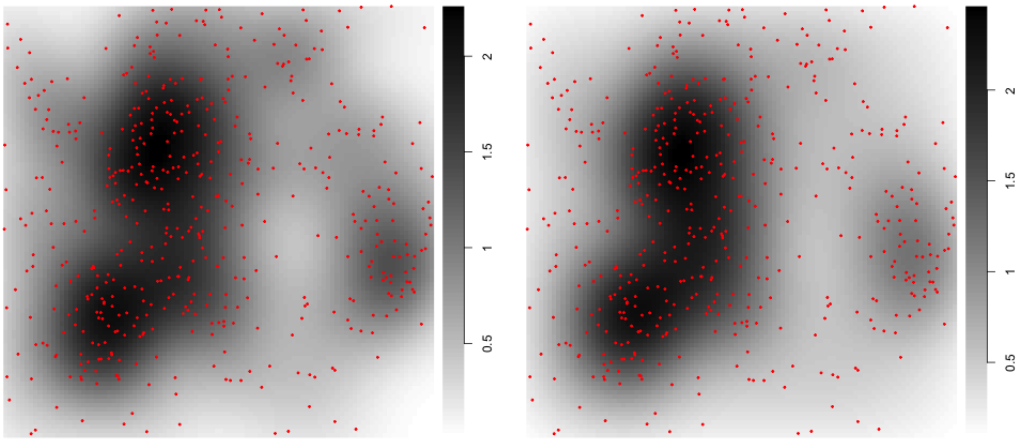

FIGURE 4.7: Left: $\lambda^{*}(u, X)$, density field estimation. Right: $\lambda^{\dagger}(u, X)$, smoothed fitted model of the generated Geyer process with pseudo-likelihood estimation. Red dots indicate the locations of the generated points. Dark area indicates higher density field or conditional probability of being occupied by a point, while lighter areas are unlikely of being occupied by a point. Smoothing bandwidth is $\omega = 1$. Values summed over the dimension Z.

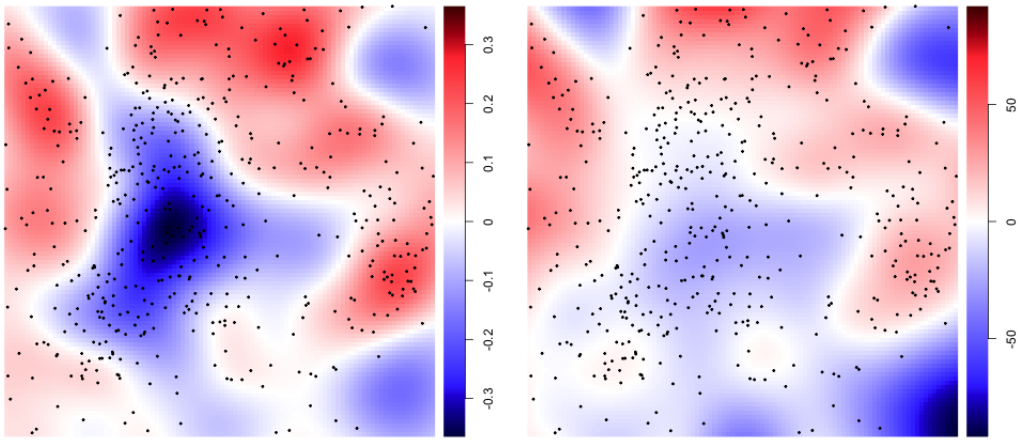

FIGURE 4.8: Smoothed raw residuals $s(u)$ (left) and relative residuals $e(u)$ (right) of the generated Geyer process with pseudo-likelihood estimation. Black dots indicate the locations of the generated points. Smoothing bandwidth is $\omega = 1$. Values summed over the dimension Z.

Fig. 4.7 is a comparison between the data density field (λ^*, left) and the smoothed fitted model ($\lambda^\dagger$, right). The similar shape and amplitude of the functions indicate a good agreement between data and model. Fig. 4.8 shows the found residuals ($s(u)$, left). Despite any possible correlation between the residuals and the points, the low amplitude of these residuals compared to the amplitude of the data density field indicates a satisfactory modeling of the process. In addition, large values of the data or model density field are expected to have bigger absolute residuals. This is summarized by the quantity $AR = 0.83$, i.e., around the 83% of the data is correctly modeled. As expected from a good fit, average of residuals is close to zero. Right image, $e(u)$ shows the relative residuals. Areas with higher values are those closer to the edges of the window, where the estimation of the model is biased, the inner regions of the volume show no lack of fitting.

4.4.4 SDSS populations

In this section we proceed to show and explain the results obtained from the fitting and residual analysis of SDSS samples. The models chosen for these populations are the Geyer model, the Fiksel model and the Power Law model. These models are used in the description of clustered processes, and are easy to implement in three dimensions.

As explained, we must pick values for the irregular parameters before proceeding with the pseudo likelihood estimation of the regular parameters. After some attempts and a brief analysis with the pair correlation function, we have chosen some motivated values. As radius of interaction we will use $r_1 = 3h^{-1}$ Mpc, and for the Fiksel and the Power Law model, the hardcore interaction radius will be $r_0 = 0.07h^{-1}$ Mpc, small enough to guarantee that no pair of galaxies in our samples lay closer than this distance. For the Geyer model we have decided a saturation value $sat = 100$, big enough to avoid any possible underestimation of the galaxy concentration. The slope parameter for the Fiksel model will be $\kappa = 0.2$ and for the Power Law model $b = 0.4$. These values are fixed for all calculations. We present the fitted results and the amplitude reduction values (AR, see equation 4.33) in Table 4.1.

TABLE 4.1: Fitted parameters for Gibbs models - SDSS-DR8

	Number of	Density	Geyer			Fiksel			Power Law		
Sample	Particles	h^3Mpc^{-3}	β	γ	AR	β	a	AR	β	a	AR
$DR8_1$	841	6.73×10^{-3}	0.003	1.47	0.55	0.003	1.71	0.56	0.003	1.59	0.55
$DR8_2$	697	5.58×10^{-3}	0.003	1.43	0	0.004	1.65	0.12	0.004	1.55	0.07
$DR8_3$	941	7.53×10^{-3}	0.004	1.32	0.43	0.004	1.48	0.42	0.004	1.40	0.42
$DR8_4$	1119	8.95×10^{-3}	0.004	1.31	0.58	0.005	1.47	0.57	0.005	1.40	0.58

In Figs. 4.9 (Geyer), 4.10 (Fiksel) and 4.11 (Power law) we show in three columns the smoothed Papangelou conditional probability ($\lambda^\dagger(u, X)$, eq. 4.31), the smoothed raw residual map ($s(u)$, eq. 4.30) and the relative errors map ($e(u)$, eq. 4.32) for the four samples and the Geyer model. These functions are defined at every location in the studied window and give us different informations of interest. This is detailed in sections 4.2.3 and 4.3. We do not consider necessary to include plots for the data density field λ^*.

Regarding the fitted parameters, for the trend function we have assumed a constant function, with β fitted around 0.004. This value fluctuates inversely to the interaction parameter, showing a stronger trend when the interaction is weaker.

As expected for a galaxy sample, the modeling of clustered populations produces values of $\gamma > 1$, indicating an aggregation of points inside our correlation range.

The interpretation of the continuous models are more complex. For the correct irregular and fitted values the models are expected to transit from values above 1 to exactly 1 after distance r_1, when the interaction between points is no more. This transit should be as smooth as possible, but instead we find that both models present values higher than 1 at distance r_1, specially the Power law. This creates a step in a continuous function, underestimating correlations for pairs separated more than $3h^{-1}$ Mpc. This indicates us that the model should be modified in order to include large range correlations.

Despite the absence of error bars, the estimated parameters γ and a might help us to detect differences in the levels of clustering and compare the samples. We can understand that samples obtaining higher values for these quantities present stronger aggregations. Interestingly, all used models respond the same way to the cosmic variance present in the four different samples. If we compare the values obtained for γ and a we can see a correlation between them, presenting higher and lower values for the same samples. In addition, these parameters show a correlation with distance, with higher values for closer samples, indicating a stronger clustering in samples with lower redshift. The limited range of redshift used and the absence of error bars in the estimated parameters do not allow us to deduce stronger conclusions, but at least, the evolution of these clustering amplitude parameters is in the expected direction.

Images reveal the capacity of this model to detect clustered regions. Denser regions are modeled with higher values of $\lambda^{\dagger}(u, X)$. However, the residual analysis, column two of Figs 4.9 to 4.11, clearly shows an overestimation (blue spots) of these regions, while the rest of the sample is mostly underestimated. The model detects the domination of these structures over the field galaxies, less clustered and appearing in red. In this situation it can be interesting to observe the values of the relative errors (third column). Despite the overestimation of some structures, colors present in these plots classify the simulated galaxies in three main groups. Red areas are usually occupied by field galaxies, where the residual error function $e(u)$ detects an important lack of fitting. Blue areas correspond to voids, where the model is not able to map the extremely low density of these regions. Two different kind of regions appear in white, the transition between the galaxy field and voids and the denser regions previously mentioned. This implies that, although these structures are overestimation, the relative error is lower than the average of

the sample. This is the capacity of the residual error function of detecting real non-modeled structures.

A solution for this problem might consist in a reduction of the interaction radius, allowing the model to describe finer structures, instead of focusing on the dominant ones. A smaller radius, comparable to the thickness of the filaments, will help to model these abundant unisotropic structures. Nevertheless, as we just previous mentioned, these models may also need to include correlations in large distances. This creates a double interaction range, one being strong and close and the other weak and distant. For future works, we propose a steeper Power law model, with higher values for small distances and a large asymptotic tail.

Is interesting to detect the odd case of $DR8_2$, where a strong concentration in the upper part of the sample (second row in Fig. 4.9 and followings) reaches amplitude values around 15, several times above the peaks of the rest of samples. These high estimated model densities worsen the general performance of the model, as we can see from the AR value of 0% and the residual values around the peak. This case was not found for LasDamas samples. This effect is a consequence of the presence of a strong cluster located in the less dense SDSS sample. The contrast between this structure and the rest of the sample is a challenge for our models and fails to map the galaxy distribution. In Tempel et al. (2012) this structure is detected as a Finger-of-God and catalogued as a cluster of 113 galaxies.

This is an example of a strong interaction process, where the pseudolikelihood is not recommended. An improved fitting algorithm might be necessary to prevent from such catastrophic fittings. Nevertheless, in the relative error plots (column in the right) we can see how the performance of describing the rest of the sample has been similar to that of the other samples. This shows how despite the bad fitting of an identifiable structure, the model is still capable of describing the rest of the population relatively unaffected.

Regarding the obtained values of the AR, all models show values of an amplitude reduction around 0.55, with the exception of $DR8_2$. This implies that more than half of the data content is included in the model, suggesting a reasonable, though indicative, description of the galaxy distribution. A comparison between the middle column of all three models reveals smaller fluctuations of the residual values for the continuous models, specially the Fiksel model. This might be due to two

different reasons. One, a higher capacity of modeling the multiscalar nature of galaxy clustering, where the Geyer flat profile will always over- or underestimate the real amount of clustering. And two, compared with the Power law, the Fiksel model has a smoother transition between the correlated and the uncorrelated range.

4.4.5 LasDamas populations

In this section we will repeat the analysis made over the samples from the LasDamas simulations. We make use of the same models with the chosen values for the irregular parameters. The fitted results with the amplitude reduction values (AR, see equation 4.33) are presented in Table 4.2

TABLE 4.2: Fitted parameters for Gibbs models - LasDamas

	Number of	Density	Geyer			Fiksel			Power Law		
Sample	Particles	$h^3 Mpc^{-3}$	β	γ	AR	β	a	AR	β	a	AR
LD_1	1024	8.19×10^{-3}	0.004	1.40	0.58	0.005	1.60	0.58	0.005	1.50	0.58
LD_2	751	6.01×10^{-3}	0.003	1.45	0.5	0.003	1.68	0.49	0.003	1.56	0.48
LD_3	943	7.54×10^{-3}	0.004	1.30	0.52	0.005	1.46	0.51	0.005	1.38	0.51
LD_4	932	7.47×10^{-3}	0.003	1.49	0.58	0.004	1.74	0.59	0.004	1.61	0.59

Again, in Figs. 4.12 (Geyer), 4.13 (Fiksel) and 4.14 (Power law) we plot the results for the functions $\lambda^{\dagger}(u, X)$, $s(u)$ and $e(u)$ and the four LasDamas samples.

The results obtained with the samples from the LasDamas simulations are similar to those of the SDSS-DR8, with fitted γ values indicating a clustered pattern. Parameters ranges are coincident in both catalogs, certifying that both populations correspond to point processes of the same nature. The previous conclusions from SDSS-DR8 samples are generally valid for LasDamas, with a few exceptions.

No monotonic evolution of the fitted parameters is found, although the correlation between models for the same samples is preserved. As we said in the previous section, this trend needs deeper analysis over wider redshift ranges before being confirmed.

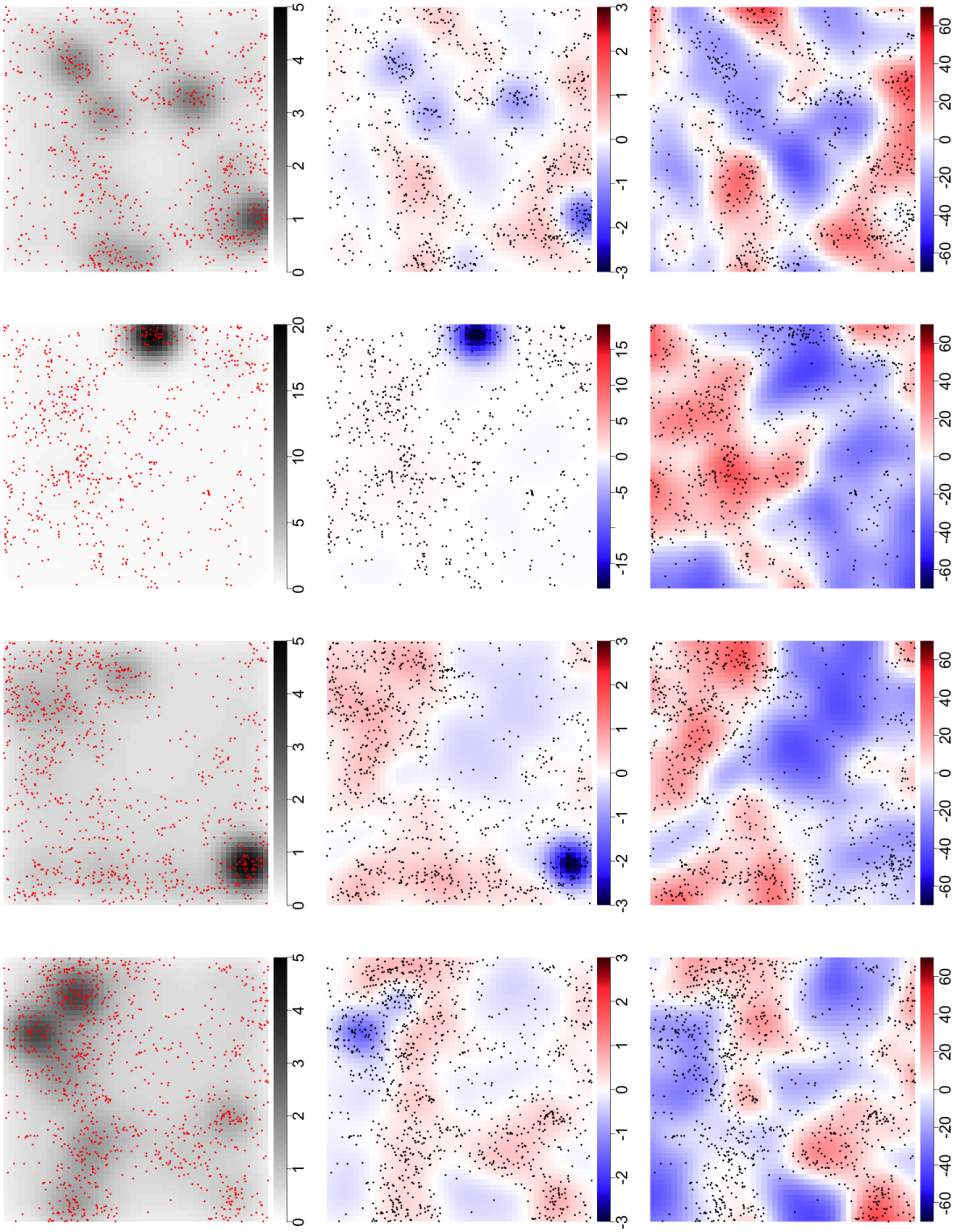

FIGURE 4.9: Z-projection of the evaluated functions for the four SDSS-DR8 samples (black dots) with fitted Geyer model. Top to bottom: sample $DR8_1$ to sample $DR8_4$. Left to right: functions $\lambda^\dagger(u, X)$, $s(u)$, $e(u)$. Scaling in $DR8_2$ has an independent scale due to its outlier values.

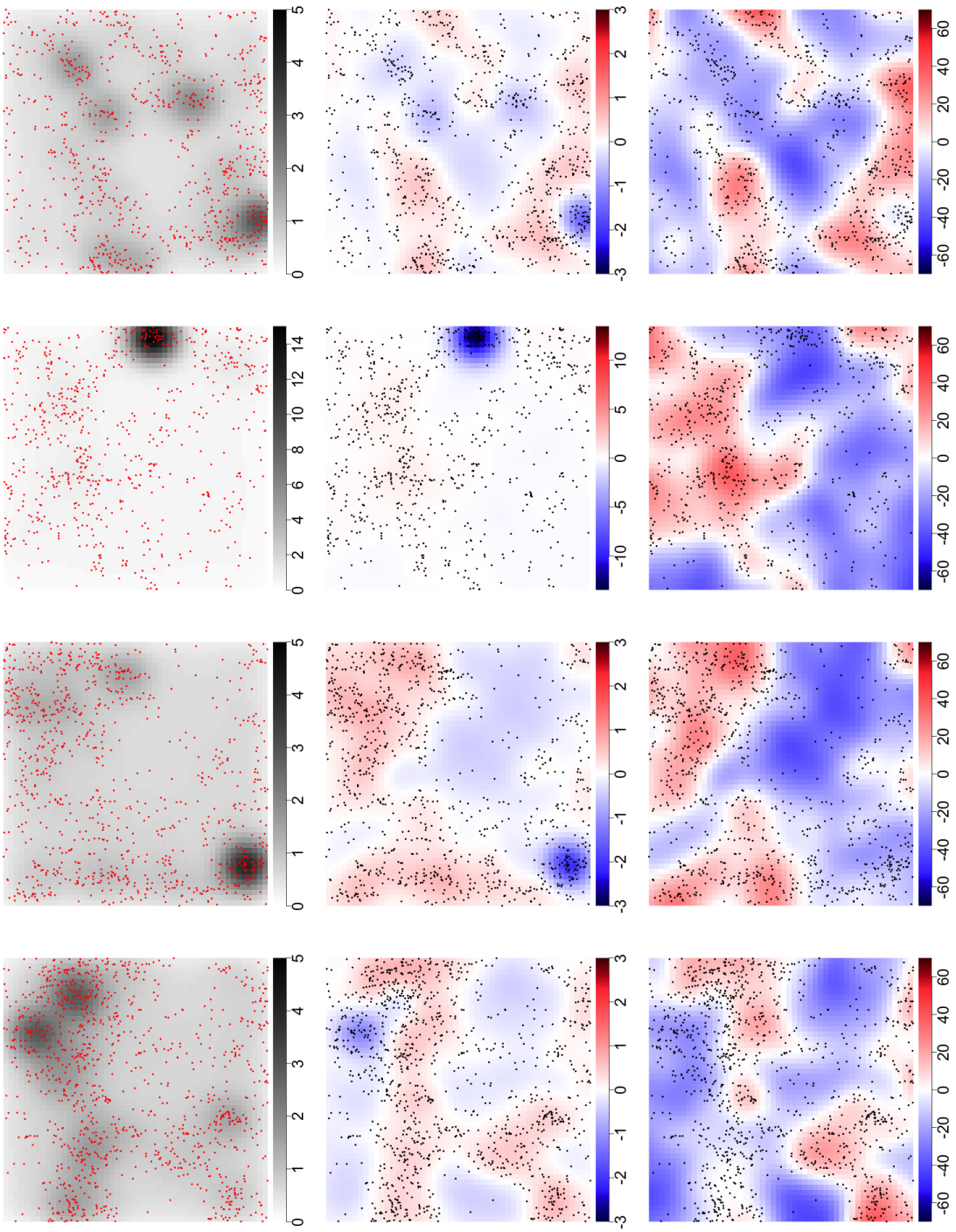

FIGURE 4.10: Z-projection of the evaluated functions for the four SDSS-DR8 samples (black dots) with fitted Fiksel model. Top to bottom: sample $DR8_1$ to sample $DR8_4$. Left to right: functions $\lambda^\dagger(u, X)$, $s(u)$, $e(u)$. Scaling in $DR8_2$ has an independent scale due to its outlier values.

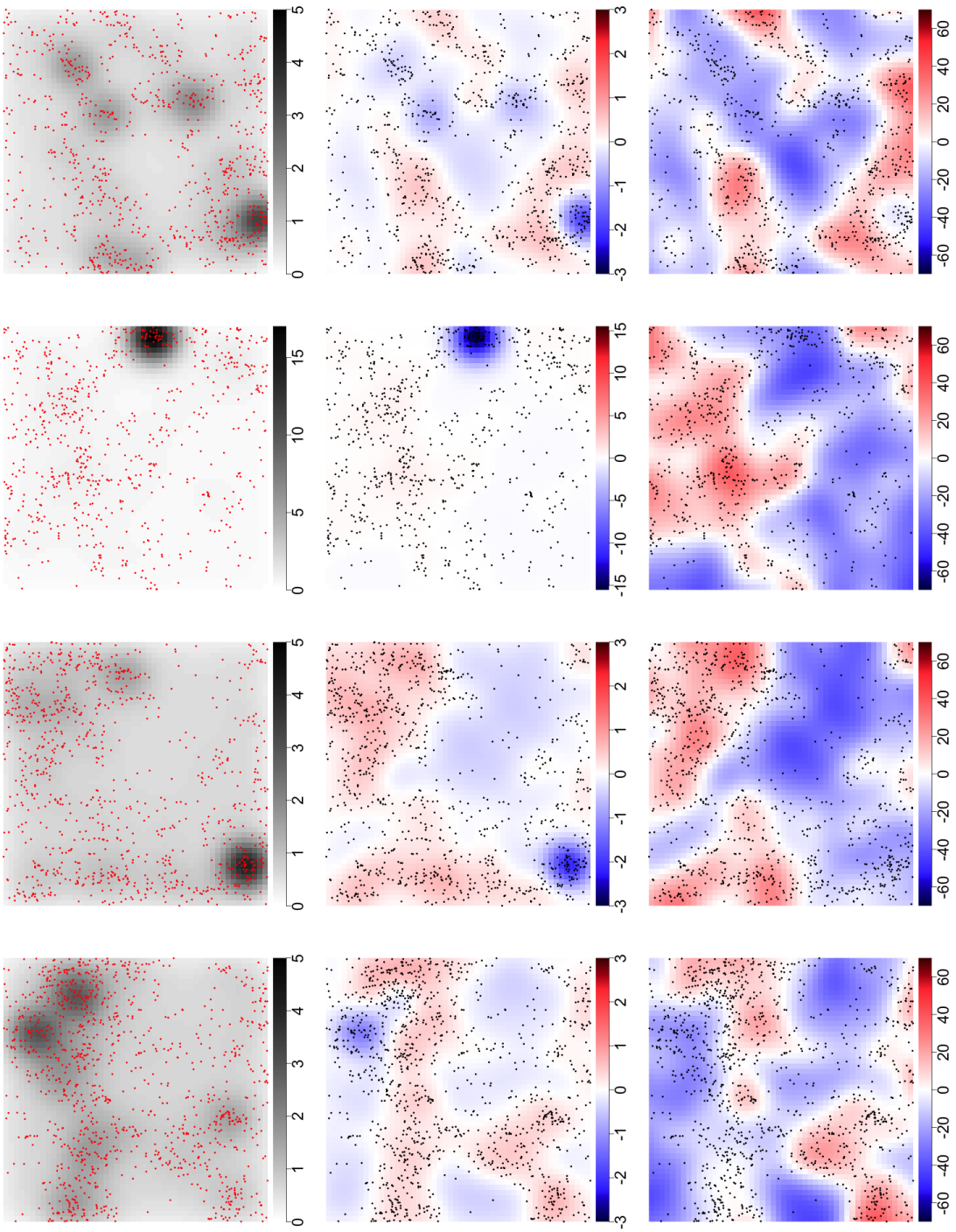

FIGURE 4.11: Z-projection of the evaluated functions for the four SDSS-DR8 samples (black dots) with fitted Power Law model. Top to bottom: sample $DR8_1$ to sample $DR8_4$. Left to right: functions $\lambda^\dagger(u, X)$, $s(u)$, $e(u)$. Scaling in $DR8_2$ has an independent scale due to its outlier values.

Due to the absence of strong overdensities in the LasDamas surveys, such as the cluster found in sample $DR8_2$, we have no catastrophic fittings with $AR = 0$, and we should understand these events as the result of expectable stochastic fluctuations in the galaxy density.

4.4.6 Conclusions and future work

Gibbs models introduce very useful utilities in the statistical analysis of point processes. We can calculate the conditional probability of a galaxy to belong to a certain process (given the rest of the galaxies). With this preliminary work we have shown that the effective calculation of this probabilistic model is an achievable task but still demands strong improvements.

We expect interaction models to be more successful at scales where interaction dynamics are dominant over any trend or anisotropy. These scales, where we will focus future works, might be the inner parts of galaxy clusters, where the host galaxy interact with their satellites, or large scale galaxy fields, where forces have a large range of interaction and no high density structures domain the distribution.

Correct characterization of a galaxy population by means of a Gibbs model will give us relevant and complete information of its distribution. The parametric nature of these models summarizes the nature of the process with very precise information of its structure and, at the same time, we can evaluate the model at any desired location. This localized analysis is one of the most relevant advantages of modeling the point process, since we are able to study the content for specific structures and environments. Effective modeling opens the door to powerful data mining.

Even more, if the probability of a galaxy of belonging to a certain process can be effectively estimated, this can be naturally extended to any location in the window, allowing us to predict locations for points. This is an extremely useful ability in cosmology, since missing data is a constant issue in the study of galaxy surveys. With an efficient model for galaxy populations, we would be able to recreate or complete masked information.

The uncertainty of galaxy redshift measurements due to fiber collisions or the masking of fragments of the sky by closer objects might be solved with a Gibbs

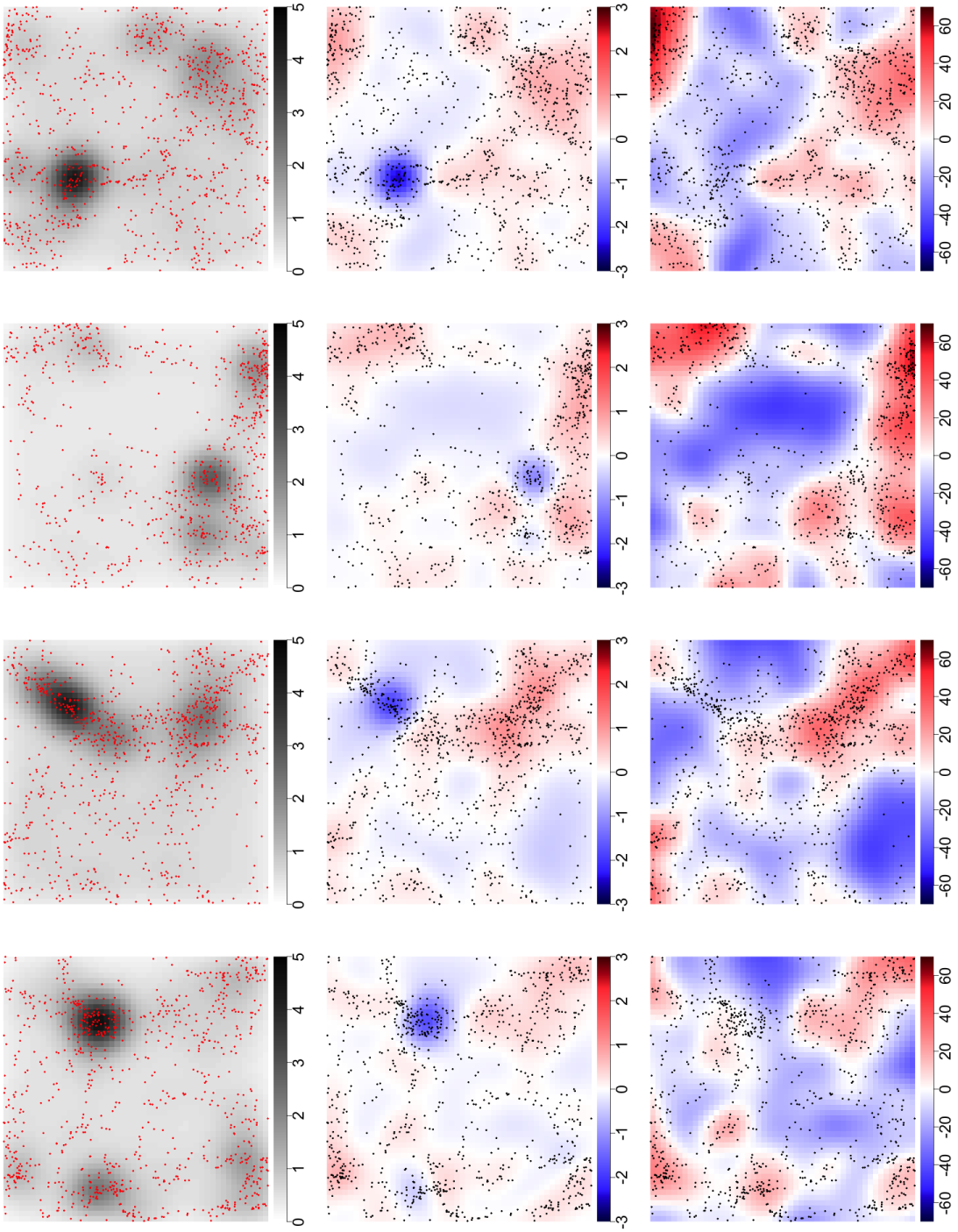

FIGURE 4.12: Z-projection of the evaluated functions for the four Las-Damas samples (black dots) with fitted Geyer model. Top to bottom: sample LD_1 to sample LD_4. Left to right: functions $\lambda^{\dagger}(u, X)$, $s(u)$, $e(u)$.

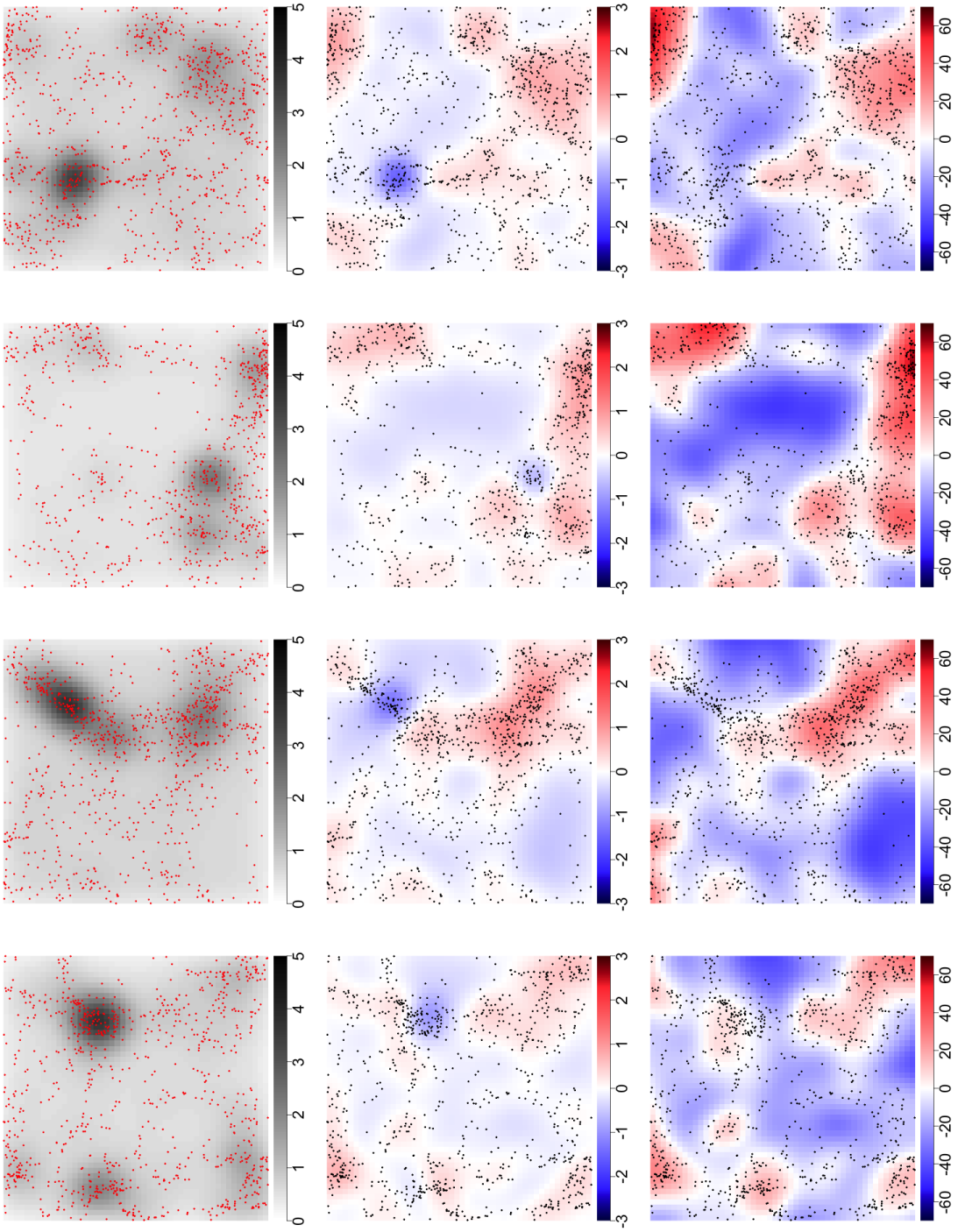

FIGURE 4.13: Z-projection of the evaluated functions for the four Las-Damas samples (black dots) with fitted Fiksel model. Top to bottom: sample LD_1 to sample LD_4. Left to right: functions $\lambda^{\dagger}(u, X)$, $s(u)$, $e(u)$.

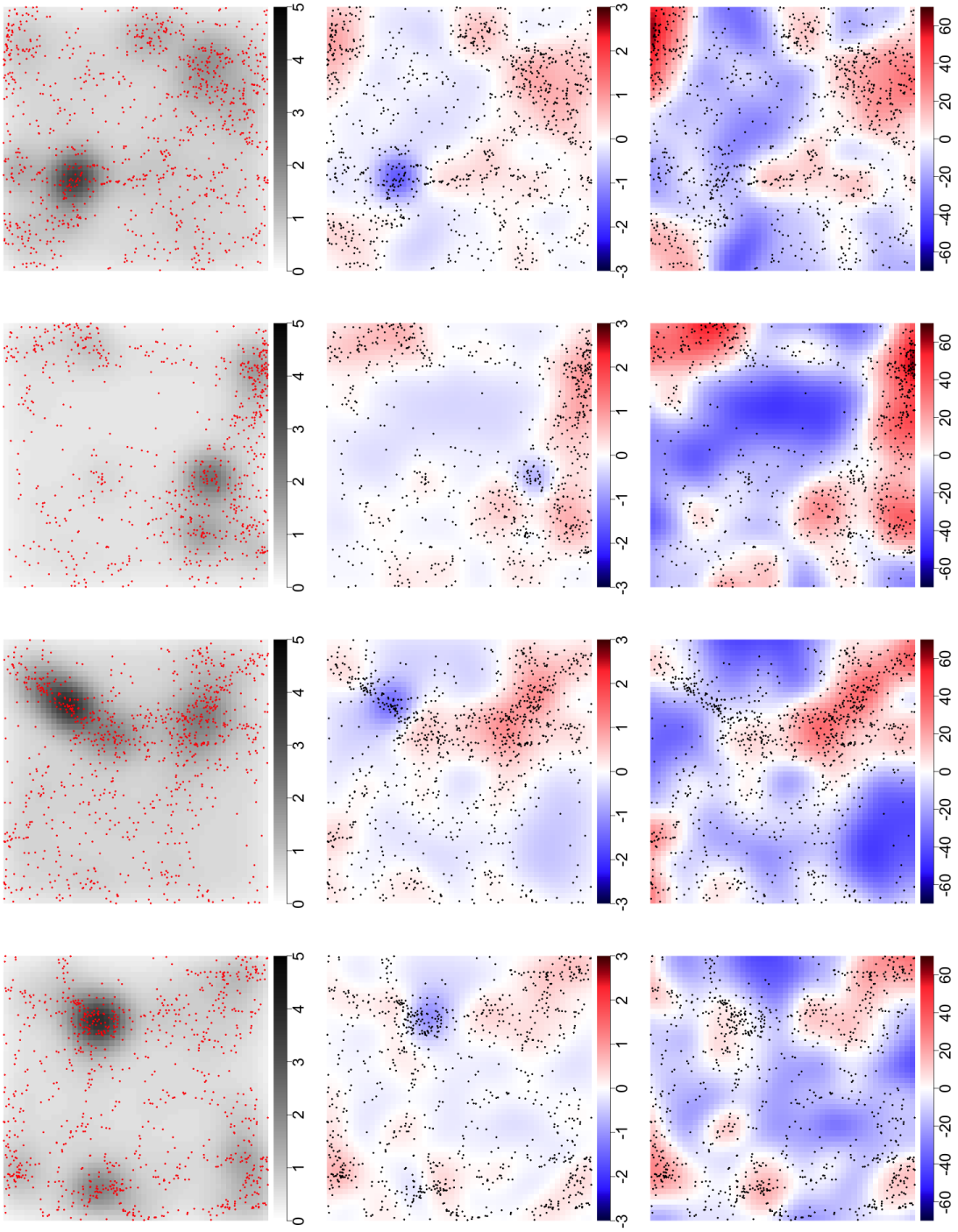

FIGURE 4.14: Z-projection of the evaluated functions for the four Las-Damas samples (black dots) with fitted Power Law model. Top to bottom: sample LD_1 to sample LD_4. Left to right: functions $\lambda^{\dagger}(u, X)$, $s(u)$, $e(u)$.

based technique. Given a galaxy sample, we can fit a Gibbs model for those galaxies whose redshift is reliably estimated. The resulting probability function $\lambda(u, X)$ will serve us to correct the position of the remaining galaxies using both their original distance estimation and the most likely positions derived from the Gibbs model. Similarly, this information can be used as a prior for Bayesian photometric redshift estimation or for correct cluster membership classification.

In addition, this method could be extended to the problem of inpainting, reconstructing areas of the sky where the position of galaxies cannot be observed and placing new galaxies accordingly to the intensity and interaction distribution of the fitted model. Although this method does not allow us to recover the true locations of unobserved galaxies, we can reproduce a sample following the modeled distribution, and therefore keeping its main properties.

For these future works, we plan to develop this new area of research with the introduction of more advanced methods of parameter estimation and more complex models. The limitations of pseudolikelihood can be overcome through more advanced estimation techniques, such as the ABC method (Stoica et al., 2015) or the Monte Carlo Maximum Likelihood. New methods will be tested as well, like the mentioned Area Interaction method, and irregular parameters will receive a more systematic analysis.

Part III

Mining the Galaxy Distribution

'The Dwarves delved deep at that time, seeking beneath Barazinbar for mithril*,*
the metal beyond price that was becoming yearly ever harder to win'.
The Lord of the Rings - Appendices

'they delved too greedily and too deep,
and disturbed that from which they fled'
Gandalf the Grey

Chapter 5

Mixture models

5.1 Introduction

The point process analysis widely extends to a multitude of algorithms and methodologies that help us to study point patterns from very different points of view, adapting the methods to our interests. Despite their strong potentialities, Finite Gibbs models are a vast family of point processes that demand abundant work to develop truly effective models of the galaxy distribution. This is a necessary task in modern cosmology if we want to fully understand the nature of the Cosmic Web.

In this chapter, instead of modeling the galaxy distribution through the intrinsic properties of galaxy interaction, that determines the morphology and evolution of galaxy structures, we will model the final results, the clusters and other overdensities created by gravity. For this reason, in the present chapter we will attempt the modeling of our galaxy population granting a differentiated role for every structure present in the sample. This has the advantage of a more versatile adaptability but the disadvantage of a higher number of assumptions. We move from galaxy level models to galaxy structure models.

A structure modeling technique can be used as well as a data mining technique, where these structures are searched and extended. As we introduced in Chapter 1, this chapter will be devoted to searching relevant features in the data such as clusters, voids or filaments. Abundant work has be done in this direction in cosmology

recently. Examples of these contributions can be found in Cautun et al. (2013), Einasto et al. (2016), Forero-Romero et al. (2009), Hahn et al. (2007), Lee & Lee (2008) and Ascaso et al. (2015). This later made use of Bayesian techniques for the identification of groups and clusters in the ALHAMBRA survey (Moles et al., 2008, Molino et al., 2014).

An example of a point process based data mining technique can be found in Tempel et al. (2014), where the authors built an algorithm for filament detection, the Bisous model. This is a precise and highly effective data mining tool with a very specific aim: the mapping of the filamentary network of a galaxy sample. The benefits of obtaining a reliable catalog of filaments with abundant information are remarkable and suppose a clear example of the capabilities and results that modern point process techniques can produce.

If we are interested in studying the peculiar structures of a sample and its physical properties in the global context of its parent survey, an interesting choice is the cluster analysis. Cluster analysis consists on identifying and characterizing groups of points in a dataset when represented under certain variables. If these variables are chosen wisely, points that aggregate trend to share common properties. In our case, points are galaxies represented by their spatial locations, and aggregation implies gravity binding and membership to a common structure. These methodologies include a wide variety of statistics and their ability to describe complex structures over modern surveys is leading us to a more detailed cosmography and understanding of the Cosmic Web.

Many of these are non-parametric methods, like the friend-of-friends algorithm, the k-means estimator, the kernel density estimator, etc. However, non-parametric methods are usually strongly dependent on arbitrary choices in the analysis. Instead, parametric clustering solves all of these problems, as we will show. For this work we opt for the Finite Mixture Modeling (Everitt et al., 2011). This technique allows us to identify peculiarities in our sample and characterize them through different kinds of models. The way Mixture models are used combines the *bump hunting* task of differentiating between white noise and peaked distributions with the analysis of the *multimodality* of a distribution, this is, the number of different coexisting peaks. To this we should add the morphological characterization of our peaks. This is therefore a double technique, both model and data mining algorithm.

With Mixture models we can still make use of the battery of functions introduced in sections 4.2 and 4.3. The modeling of a point process through a probability density function, its Papangelou conditional function and the associated residual analysis are perfectly valid for any probabilistic approach. In addition, with Mixture models we are able not only to model the distribution but to perform highly robust and versatile data mining techniques.

The aims of this work are to test the Mixture models over a data catalog and evaluate their capacity of describing its content. As with the Gibbs models in section 4.2, we test the usage and potentialities of Mixture models in large scale structure analysis for the first time. For this reason we test our new methodology while using well known datasets.

We expect to be able to classify our data in two simultaneous ways. First, the mining of the sample, identifying the location and characterizing the shape and main physical properties of galaxy clusters. And second, the modeling of the population, assigning for each galaxy a membership in the found clusters. Altogether, this consists in a model that allow us to obtain an original insight on the clusters morphology while describing their interactions with neighbor structures.

This chapter is organized as follows. Section 5.2 introduces the Mixture models, with the fitting algorithms used in this work and their main applications. In section 5.3 we describe the modeled data sample from the MultiDark simulation (Klypin et al., 2011). Section 5.4 describes the physics necessary to complete our model and test it against several populations, first some generated toy models and finally the samples extracted from the MultiDark simulation. Conclusions are presented in section 5.5.

5.2 Definition

Mixture models are probabilistic and parametric models describing the probability of finding a point in a given location. They are always composed by two or more components used to map a population. These components correspond to differentiated structures and we should find the right parameters to properly adapt each shape to the given distribution. This technique dates back to K. Pearson in the 1890s and is commonly used with k Gaussian distributions over a multimodal

population. We will conveniently modify it to our interest with astrophysically motivated functions. A correct choice of our model will allow us to identify and properly describe not only the main structures in the sample but their internal substructures, as these components are allowed to overlap. The estimation of the number of needed components or bumps also requires attention, since the final result is highly dependent on this quantity.

The number of final parameters of the model is also relevant: in a 3-dimensional situation, as ours, all bump models need three parameters to set the center of the bump, plus other p parameters to describe the shape of the bump. In addition, Mixture models include mixture coefficients to establish the relative intensity of each bump. These coefficients work as a normalization factor that gives each bump its right mass. If we use as well another parameter for a background component we obtain $k \cdot (3 + p + 1) + 1$ parameters.

Due to its bump hunting motivated architecture, Mixture models can effectively detect and characterize galaxy overdensities. The discrimination of the galaxies in different structures is softer than in other component detection algorithms like the k-means, where objects are assigned to only one cluster, we call this a hard classification algorithm. With Mixture models instead, we assign a membership probability to each galaxy, describing the probability of belonging to each cluster.

This tool has been already applied by Kuhn et al. (2014) in the modeling and identification of star clusters in 2-dimensional images. We will apply it to the 3-dimensional galaxy scenario. As in Kuhn et al. (2014) we will add as well a background component, meant to include galaxies not belonging to any of the bumps but in the rest of the distribution. As a final motivation, if we are able to properly map the clusters and field galaxies of a population but not the filaments or other more complex structures, we might expect to detect them using estimators such as the relative error function $e(u)$ in equation 4.32.

Now, we proceed to define our mixture model and the fitting of its parameters.

5.2.1 The surface density model

In this section we proceed to build the probability density function, or surface density model, of our galaxy cluster population. As found in Everitt et al. (2011),

Finite Mixture densities are a family of probability density functions of the form

$$f(\mathbf{x};\alpha,\theta) = \sum_{i=1}^{c} \alpha_i \cdot g_i(\mathbf{x};\theta) \tag{5.1}$$

where $\mathbf{x}$ is a k-dimensional random variable, the family of g_i are the component densities parametrized by θ and α are the nonnegative mixing proportions or mixture coefficients. The number of components is c.

In the spatial point process application, the model is defined over a window W containing N points. Since this is a density model, f can be evaluated at any location of the window, not only where a point lies.

Our implementation of the Mixture model for the galaxy clusters structure follows the strategy of Kuhn et al. (2014). The first needed element is the profile of the cluster component. These component can be as irregular as we are able to model them, but since this method is generally used as a cluster classification method, we will model two different types of components, the clusters and the background.

In this work, the cluster profile $\rho(\mathbf{r} - \mathbf{r_0}; \mathbf{s})$ assumes spherical symmetry for our fitted structures, and describes the density of the cluster at a given distance from the center $\mathbf{r_0}$. As a bump hunting algorithm, this is a general condition that allows us to easily locate the structures in the space. Parameters $\mathbf{s}$ are used to describe the shape of the profile. These parameters are both included in the profile parameter set $\theta = (\mathbf{r_0}, \mathbf{s})$.

The background component is treated as another cluster but no parameters are necessary. Therefore, our background will be approximated with an homogeneous Poisson process, i.e., $\rho_b(\mathbf{r} - \mathbf{r_0}) = 1$. Galaxies located in the field outside any cluster will be mapped by this component.

The next necessary element to build our Mixture model is the number of components. In this case we use $c-1$ cluster components plus the background component.

$$\Sigma(\mathbf{r};\alpha,\theta) = \sum_{i=1}^{c-1} \alpha_i \cdot \rho_i(\mathbf{r} - \mathbf{r_0}; \mathbf{s}) + \alpha_c \cdot \rho_c(\mathbf{r}) \tag{5.2}$$

It may be convenient to fix $\alpha_1 = 1$. This way we eliminate one of the fitted parameters and express the relative mass of each component with respect to the first one, chosen by the user. Therefore, for $\alpha_2 = 2$, the mass of the second component will be twice that of the first component. The last necessary step to build our Mixture model is to normalize dividing by the total mass of the model

$$M = \int_W \Sigma(\mathbf{r}; \alpha, \theta) d\mathbf{r} \tag{5.3}$$

However, this can only be done if the profiles are integrable, $\int_W \rho(\mathbf{r}, \theta) d\mathbf{r} < +\infty$ for all θ. Finally, we define our Mixture model as

$$\lambda(\mathbf{r}; \alpha, \theta) = \frac{1}{M} \Sigma(\mathbf{r}; \alpha, \theta) \tag{5.4}$$

As we will see in the following sections, it may be also interesting to normalize the model to the total number of objects N. This way, the integral of $\lambda(\mathbf{r})$ in a region $A \subset W$ is the estimated number of objects in A. And the other way round, we can integrate only one of the fitted components over the entire volume W to estimate the number of objects belonging to this single component. When λ is defined like this, our studied point process can be understood as an inhomogeneous Poisson distribution with function λ or as a generalization of a Neyman-Scott process (Neyman & Scott, 1958) where the daughter points are generated accordingly to our components profiles.

5.2.2 Density profiles

Once the Mixture Model is fitted, individualized analysis of each component may be of interest. This allows us to segregate the contribution of each component at every location in the window, having a clear vision of how components relates each other.

Given a component, we already have its profile density as expressed in the surface density model $p_i \rho_i(\mathbf{r} - \mathbf{r_0}; \mathbf{s})$. We can differentiate this profile with respect to distance, evaluating the component around the center of the cluster ($\hat{\mathbf{r}}_\mathbf{0}$) and normalizing by the volume.

$$\bar{\rho}_i(\mathbf{r}-\mathbf{r_0};s) = \frac{\alpha_i}{M} \cdot \frac{\int_{\mathbf{r}}^{\mathbf{r}+d\mathbf{r}} \rho_i(\mathbf{t}-\mathbf{r_0};\mathbf{s})d\mathbf{t}}{V(\mathbf{r}+d\mathbf{r})-V(\mathbf{r})} \tag{5.5}$$

where $V(\mathbf{r})$ is the volume of the sphere with radius $||\mathbf{r}||$. This is the 1-dimensional profile of component i. If we now substitute ρ by the full mixture model, and evaluate it as before, centered in $\mathbf{r_0}$, we will obtain our estimation of the density profile of the whole sample as seen from $\mathbf{r_0}$, i.e. the 1-dimensional profile of the Mixture model centered in the component i.

$$\bar{P}_i(\mathbf{r}-\mathbf{r_0};\mathbf{s}) = \frac{\alpha_i}{M} \cdot \frac{\int_{\mathbf{r}}^{\mathbf{r}+d\mathbf{r}} \Sigma(\mathbf{t}-\mathbf{r_0};\mathbf{s})d\mathbf{t}}{V(\mathbf{r}+d\mathbf{r})-V(\mathbf{r})} \tag{5.6}$$

where $\mathbf{r_0}$ is the center of component i. For short distances, $\bar{\rho}_i$ and $\bar{P}_i$ should have similar values, but for distances at which our component i starts to increasingly overlap with other components, these function will start to diverge, with $\bar{P}_i$ depicting the profile of other structures. Having the ability to segregate the contribution of each component to the overall density profile allows us to measure the level of interaction between structures and obtain density profiles without contamination from close sources.

One of the advantages of models over classification methods is the ability of generating new samples equally distributed. For a density model this can be easily done simply using rejection sampling techniques for each function $\rho_i(\mathbf{r}-\mathbf{r_0},\mathbf{s})$ for the estimated number of objects per component. As said, these numbers are obtained integrating each component over W and normalizing by the total amount of points.

5.2.3 Fitting the parameters

The best fit parameters will be found using the complementary approaches of different techniques. The first one and most important in the Posterior probability function. We start defining the maximum likelihood estimation (MLE), which seeks the model that maximizes the likelihood of the data. Given the parameters (α,θ), the log-likelihood is given by

$$\ln L(\mathbf{p}, \theta; X) = \sum_{i=1}^{N} \log \Sigma(\mathbf{r}_i; \mathbf{p}, \theta) \tag{5.7}$$

Where $\mathbf{r_i}$ are the locations of the points in X. If the model is normalized to the total number of objects N, we have the model $\lambda(\mathbf{r}; \alpha, \theta) = \frac{N}{M}\Sigma(\mathbf{r}; \alpha, \theta)$ and its log-likelihood can be expressed as

$$\log L(\alpha, \theta; X) = \sum_{i=1}^{N} \log \lambda(\mathbf{r_i}; \alpha, \theta) - \int_W \lambda(\mathbf{r}'; \alpha, \theta) dr' \tag{5.8}$$

which we will use for this work. The first term is evaluated for all the N galaxies of the population and the integral in the second term is the expected number of galaxies in the window.

Now we are interested in finding the highest value for the function above. We find it using the Posterior probabilistic function as obtained by the MCMC routine with a Bayesian approximation. Bayesian inference derives the posterior probability as a consequence of two antecedents, a prior probability and the likelihood function 5.8. Bayesian inference computes the posterior probability according to Bayes' theorem:

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)} \tag{5.9}$$

where H is the hypothesis we want to test based on the evidence or data D. Hence, $P(H)$ is the *prior probability*, the probability of H before D is observed. This indicates one's previous estimate of the probability that the hypothesis is true. $P(H|D)$ is the *posterior probability*, the probability of H given D. This tells us what we want to know: the probability of a hypothesis given the observed evidence. $P(D|H)$ is the probability of observing D given H, this is the *likelihood*. It indicates the compatibility of the hypothesis with the data. $P(D)$ is the normalizing constant, sometimes termed the *marginal likelihood*. This factor is the same for all possible hypotheses being considered. This means that this factor does not enter into determining the relative probabilities of different hypotheses, and is usually ignored.

This law can be used in Bayesian inference to fit a model H over a set of observations D when we already have an idea of the approximate behavior of $P(H)$. Substituting the data by the spatial point process X and the hypothesis by the Mixture model, represented by its parameters $\Theta = (\alpha, \theta)$, we have

$$P(\Theta|X) = \frac{P(X|\Theta) \cdot P(\Theta)}{P(X)} \tag{5.10}$$

The difficulty in the calculations of this expression resides in the term $P(X)$, which acts as a normalization constant. This is solved using the proportionality between the posterior likelihood with the likelihood and the prior probability, which expressed in a logarithmic shape is

$$\log P(\Theta|X) \propto \log P(X|\Theta) + \log P(\Theta) \tag{5.11}$$

Where $\log P(X|\Theta)$ is our log-likelihood function in 5.7. In this work we use a non informative quasi flat prior, with $P(\Theta)$ a Gaussian distribution centered in 0 with a large standard deviation. Then, given two sets of values Θ_0 and Θ_1, we can test which one fits better the data calculating the correspondent posteriors.

MCMC uses this to map the overall distribution of the log-likelihood posterior strategically evaluating different combinations of parameters. This gives us valuable information, mapping the region of the parametric space where the best fit Θ values live. To start this routine, the procedure demands an initial guess of Θ, somewhere to start evaluating $\log P(X|\Theta)$. It is more effective when this guess is accurate and close to the areas of higher likelihood values, and for this reason we previously estimate it. The obtained distribution contains the best fit set of parameters, which can be found maximizing the Posterior.

The goodness of fit of any mixture model is highly dependent of the locations of the clusters $\mathbf{r_0}$, only once these parameters are correctly fitted it makes sense to fit the shape parameters or the mixture coefficients. Hence, we evaluate the density field of the galaxy distribution using the kernel estimator of equation 4.28 and then select the local maxima on the resulting function. These maxima are the first candidates to become centers of clusters, as peaks of the overdensities in the distribution. This is just an indicative result, and highly depends of the

assumed bandwidth. The MCMC is a costly routine but necessary to obtain a reliable map of the log-likelihood posterior function and the standard deviation of each parameter. The software used in these calculations is included in the CRAN package *LaplacesDemon* (Statisticat & LLC., 2013a,b,c,d), which provides more than forty different MCMC algorithms. These algorithms decide the next combination of parameters θ to be tested in order to correctly map the normalized log-likelihood function 5.8. Some of these algorithms, the so called *adaptive*, use the previous evaluations to choose the next θ. These are usually more efficient in finding the overall distribution of the function, but we must always finish with a long run of a non adaptive algorithm to ensure a proper convergence.

Additional algorithms can be used complementarily to the MCMC evaluations, for example, helping us to obtain a better guess of the starting points. On the other hand, once a good approximation of the Maximum Posterior is found, we can polish it with additional evaluations around it.

One of these algorithms is the Laplace Method, also included in *Laplace's Demon* package. This method approximates the gradient of an unnormalized joint posterior density estimating the global maximum and the variance of each parameter involved. This is a family of asymptotic techniques used to approximate integrals (Azevedo-Filho & Shachter, 1994).

The other one is the well known Nelder-Mead algorithm (Nelder & Mead, 1965), which minimizes the result of a function using only its values. This algorithm can be found in the CRAN package *stats* (Team, 2015) under the name *optim*. This algorithm may help us to find initial values for the MCMC iterations. In cluster analysis the locations of the clusters are generally independent from other parameters related with shape or size. Fitting all these parameters at the same time might be an unnecessarily costing routine. Instead, we could use the Nelder-Mead algorithm to approach the first subset of parameter while the rest remain *frozen*.

Similarly, we can use the *frozen* Nelder-Mead fitting algorithm to optimize the Maximum Posterior parameters found in the MCMC routine. Is a slow but robust method, and a good choice for final values fitting.

In both applications, improving the best fit and approaching the initial values, we use the *optim* code several times freezing each time a different set of parameters for each model component.

Finally, we must discuss the right number of model components. The best fit log-likelihood value of a model will always improve with a higher number of components. However, we must identify the right number of structures that we should model with cluster profiles. This can be done by means of the the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC). These tools are used to compare the quality of two fitted models over the same data. These models need to be nested as well, which, in this case, means that one of them is using the same components of the other one plus extra components. These quantities are related with the value of the log-likelihood, but incorporate as well the number of used parameters. This way, the information criteria penalize those models where an excessive number of parameters does not imply the same improvement in the log-likelihood. We intend to find the best fit without over populating the galaxy distribution with spurious clusters. It is then recommendable to perform the above fitting procedure with different numbers of clusters and test the results with

$$\mathrm{BIC} = -2\log L + (|\theta| \cdot k + 1)\log N \tag{5.12}$$

$$\mathrm{AIC} = -2\log L + 2(|\theta| \cdot k + 1) \tag{5.13}$$

where $|\theta|$ is the number of parameters in θ, k is the number of clusters, N is the number of points in the dataset, and $\log L$ is the log-likelihood for the best fit parameters θ. These information criteria give the smaller result for the best model configuration of k and θ. The choice of AIC or BIC is widely debated (Burnham & Anderson, 2002, Everitt et al., 2011, Kass & Wasserman, 1995, Konishi & Kitagawa, 2008, Lahiri, 2001) with arguments for both options; as it is common, we use both.

5.3 MultiDark simulated samples

In section 5.4 we will make use of samples from the MultiDark simulation project (Klypin et al., 2011) in order to test the Mixture Model approach. This model will be tried first over self-generated toy-models but a realistic galaxy-like sample is necessary to fully calibrate the capacities of our galaxy cluster profile fitting algorithm. A real galaxy sample extracted from a galaxy survey would have been the optimal choice, but an N-body simulation offers simplicity and a much higher resolution that allows us to easily apply our methodologies over a sample of scientific interest.

As described by the collaboration, the Spanish MultiDark Consolider project supports a variety of efforts to identify and detect dark matter. As a result of the collaboration between the Leibniz-Institute for Astrophyics Potsdam and this project, the MultiDark Database was created to publish cosmological simulations, which allows scientists worldwide to explore these data for studying the large scale structure of the universe as well as the properties of dark matter halos. We focus on the study of dark matter halos, which agreed with the assumptions made for our model, selecting a small region of interest to devote our detailed analysis and modeling. With a high density sample of these dark matter particles we simulate a galaxy distribution.

These data are extracted from the Bolshoi simulation (from Russian "big"), a simulation of volume $(250h^{-1}$ Mpc$)^3$ and a mass resolution of $1.35 \cdot 10^8 h^{-1}$ M$_\odot$, which is more than a factor of 6 better than the mass resolution of the Millennium simulation. Its details are described in Klypin et al. (2011). The cosmological parameters used for this simulation are $\Omega_m = 0.27$, $\Omega_b = 0.0469$, $\Omega_\Lambda = 0.73$ and $H_0 = 100h$ km s^{-1}Mpc^{-1}. The used snapshot is at redshift $z = 0$ from which the 0.001% of the particles have been extracted.

Since our study is related with dark matter halos and overdensities, it is of great interest to count with a halo catalog previously obtained from the used simulations. This is provided by the BDM tables, which identify these halos using the Bound Density Maximum (BDM) algorithm. The MultiDark simulation provides two different versions: one using the standard overdensity criterion with $360 \cdot \rho_{back}$ (background density; BDMV) to define halos, and the other with halos cut-off at

$200 \cdot \rho_{crit}$ (critical density; BDMW). In addition, halos are divided into two groups: distinct halos and subhalos. We make use of the first version.

Despite further specifications on halos and subhalos structuring, we will limit our use of this information to the identification of bigger halos in our selected sample. Using this halos catalog we have selected a flat cuboid containing different halos, 3 of which are among the 100 most massive ones in the full Bolshoi simulation box plus several other halos of smaller size. The cuboid has a volume of $4375h^{-3}$ Mpc^3 and contains 2081 particles. Its dimensions are $25 \times 25 \times 7h^{-1}$ Mpc, where the squared face facilitates the 2-dimensional examination. The final particle density is $0.4757h^{-3}$ Mpc^3, higher than other used galaxy surveys densities from this work but necessary to properly characterize the dark matter halos.

5.4 Mixture models for galaxy clusters

In the application of the Mixture models for galaxy cluster characterization, we make use of different dark matter profiles to model our galaxy groups. These functions will be tested against generated samples following the different profiles (toy models) and against data from the MultiDark simulation.

5.4.1 Dark matter profiles

The presented models correspond to particular cases of two different families of density profiles: the Sérsic and the Hernquist profiles. We make use of the well known Einasto (Einasto, 1965, 1968, 1969) and Navarro-Frenk-White (Navarro et al., 1995) profiles, as derived from these families. These models show a steepening of the logarithmic slope with increasing radius, as seen with the N-body simulations of cold dark matter (CDM) halos (Efstathiou et al., 1988, Frenk et al., 1988, West et al., 1987). Dark Matter halos in simulations (or galaxy clusters and groups) tend to present "universal profiles" that can be modeled with a simple function form. This well known property of clustered matter may help us to easily model the content and structure of a data sample, such as the introduced MultiDark sample.

Einasto model

The Sérsic profile (Sérsic, 1963, 1968) is an empirical fitting function originally used for describing the luminosity profiles of early-type galaxies and bulges, and therefore, is used on projected images (Caon et al., 1993, Graham & Guzmán, 2003). It is a generalization of the $R^{1/4}$ profiles of de Vaucouleurs (1948) with an $R^{1/n}$ profile:

$$I(R) = I_e \exp\Big(- b_n \Big[(R/R_e)^{1/n} - 1 \Big] \Big) \tag{5.14}$$

where n is monotonically related to how centrally concentrated a galaxy's light profile is, while R is the projected radius. I_e is the intensity at the projected effective radius R_e. The term b_n is a function of n and must satisfy that R_e encloses half of the total galaxy light (Caon et al., 1993, Ciotti, 1991). This function can be approximated when $n \gtrsim 0.5$ by the equation given in Prugniel & Simien (1997):

$$b_n \approx 2n - 1/3 + 0.009876/n \tag{5.15}$$

However, Sérsic's model is traditionally applied to the projected (surface) densities of galaxies, not to 3-dimensional density. Einasto (1965, 1968, 1969) independently developed a model for 3-dimensional density profiles that follows Sérsic's functional form. It is usually given as

$$\rho(r) = \rho_e \exp\Big(- d_n \Big[(r/r_e)^{1/n} - 1 \Big] \Big) \tag{5.16}$$

where r is the 3-dimensional (i.e. not projected) radius. Similarly to b_n, d_n is a function of n satisfying that ρ_e is the density at the radius r_e that defines a volume containing half of the total mass. Despite it can be also approximated for values of n above 0.5, we will obtain it solving the equation $\Gamma(3n) = 2\gamma(3n, d_n)$, where Γ is the complete gamma function and γ is the incomplete gamma function. Since it mixes with the mixture coefficients, parameter ρ_e will not be fitted. The Einasto profile has been successfully used at modeling galaxy clusters (Aceves et al., 2006, Graham & Guzmán, 2003, Merritt et al., 2006, Navarro et al., 2004) and references therein.

Navarro et al. (2004) wrote "adjusting the parameter [n] allows the profile to be tailored to each individual halo, resulting in improved fits". Such a breaking of structural homology (see Graham & Colless (1997) for an analogy with projected luminosity profiles) replaces the notion that a universal density profile may exist (Merritt et al., 2006).

Hence, this model rejects the existence of a universal density profile, breaking the structural homology.

In our fitting process we will need to normalize the mixture model integrating the different cluster components. This can be done analytically as explained in Retana-Montenegro et al. (2012). We know that the total mass enclosed by the halo is

$$M = 4\pi\rho_e h^3 n\Gamma(3n) \tag{5.17}$$

where $h = r_e/d_n^n$. If we are interested in integrating the mass contained within a radius r, we can obtain it through the variable $s = d_n^n r/r_e$ and

$$M_r = M\Big(1 - \frac{\Gamma(3n, s^{1/n})}{\gamma(3n)}\Big) \tag{5.18}$$

Using this value is a key element to perform a satisfactory fitting of our model.

Hernquist model

The other family of models we considered is the Hernquist family (Hernquist, 1990), a generalization of the Jaffe's profile (Jaffe, 1983). The shape of this model is determined through three parameters (α, β, γ) which determine the shape and slope of the profile. The density curve is described here by a double power-law model with variable slopes:

$$\rho(r) = \rho_s 2^{(\beta-\gamma)/\alpha}\Big(\frac{r}{r_s}\Big)^{-\gamma}\Big[1 + \Big(\frac{r}{r_s}\Big)^{\alpha}\Big]^{(\gamma-\beta)/\alpha} \tag{5.19}$$

where ρ_s is the density at the scale radius r_s. These marks the center of the transition region between the inner and the outer power law having slopes of $-\gamma$ and $-\beta$, respectively. The parameter α controls the sharpness of the transition.

Certain works (Graham et al., 2003, Klypin et al., 2001) found degeneracies when all five parameters are allowed to vary. Merritt et al. (2006) recommend instead to fix $\alpha = 1$ and $\beta = 3$, which leaves

$$\rho(r) = \frac{\rho_s 2^{3-\gamma}}{(r/r_s)^{\gamma}(1 + r/r_s)^{3-\gamma}} \tag{5.20}$$

Depending on the values given to the γ parameter we obtain a different inner profile. For $\gamma = 1$ we have the Navarro-Frenk-White (NFW) profile (Navarro et al., 1995). This is one of the most commonly used profiles due to its efficiency at modeling dark matter halos and galaxy clusters (Bertone, 2010, Jing, 2000, Klypin et al., 2002), despite its steep profile for small distances, which could be in contradiction with the expected flat distribution (Navarro et al., 1996, Schaller et al., 2015).

However, while the Einasto profile has a finite value at $\rho(0)$ and finite mass for infinite radius, the Hernquist models do not satisfy any of these conditions. The non integrability of the NFW profile prevents it from being used as a probability density function, as we need in the construction of Mixture models. In addition, for a $(1, 3, \gamma)$ profile, we do not have a analytic expression of the halo mass contained within a given radius, which forces us to numerically integrate the mass under the profile. The high computation costs greatly hinders the fitting of the model, not allowing us to fulfill our fitting and residual analysis procedure.

For this reason, after trying a $(1, 3, \gamma)$ profile, we used NFW profiles ($\gamma = 1$), which still inherit the problems of the general family but provide an analytical integrable expression for the halo mass at its virial radius (Cooray & Sheth, 2002). However, the obtained Mixture model doesn't satisfy the residuals tests (section 4.3), indicating a clear overestimation for the halos and an underestimation for the background galaxies. The reason why this truncated version of the profile is not yet satisfactory may reside in the pole around the center. The Maximum Likelihood estimator diverges when numerically estimating the parameters of such a profile. In this case the obtained parameters concentrate the mass of the profile around the pole, maximizing the likelihood function. A profile truncated around the pole failed for similar reasons.

Due to these problems, in this thesis work we use only models based on the Einasto profile.

5.4.2 Toy models

We will test our fitting procedures and the whole battery of residual analysis techniques presented in section 4.3 over several self-generated toy models before trying real data samples. This preliminary analysis will help us to understand the weakness of our Mixture model algorithm and prevent major uncertainties in our results and conclusions.

In a cube of side 60 we have placed four different clusters generated following Einasto profiles of known parameters through a rejection sampling method. The input parameters of the samples are summarized in Table 5.1. The relative density of the clusters is determined by the generation process and the mixture coefficients are derived from the number of points per cluster N, the Einasto radius r_e and the Sérsic index n. We add as well a population of Poisson distributed points in the box to be fitted by the background component. Together with this component, our sample contains 2782 points, with number density 1.29×10^{-2}. We show the toy model sample generated this way in Fig. 5.1.

All calculations of this section are done in three dimensions, but we only show the X vs Y projection to avoid prolixity.

Einasto's box

We start our MCMC routine sampling the distribution of our parameters. Knowing that the sample contains four clusters, an Einasto's model include 3 location parameters (x_0, y_0, z_0), plus 2 shape parameters (r_e,n) per cluster, plus 4 mixture coefficients α_i ($\alpha_1 = 1$ is fixed but we add the background free mixture parameter α_c), there are 24 free parameters in total. We run a sampling of 50000 iterations, enough to picture the distribution of the parametric set. As an example, the Posterior probability functions of the parameters of component 2 are summarized in Fig. 5.2.

As explained in section 5.2.3, once the MCMC sampling is complete, it provides us the standard deviation of each parameter. The parameters values shown in

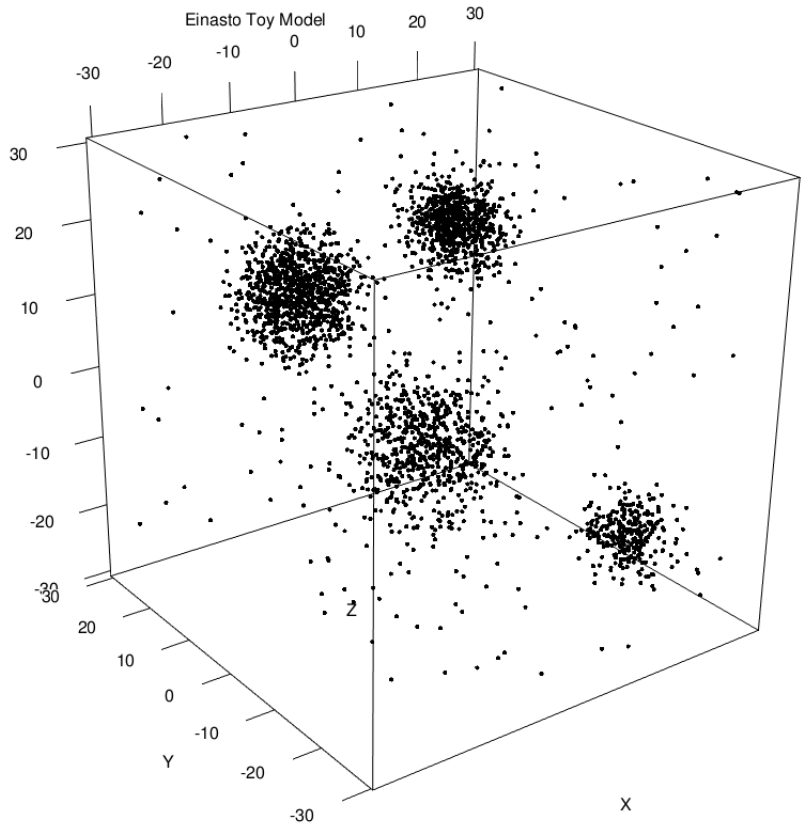

FIGURE 5.1: 3D samples of four Einastos' galaxy clusters with background.

Fig. 5.2 correspond to the mean of the MCMC distribution, and might differ from the best fit values. However, it is important to sufficiently map the behavior of the parameters in order to guarantee the robustness of the best fit and its variation.

Finally, we proceed with the *optim* routine as explained in section 5.2.3, freezing different sets of non correlated values each time and improving the best fit sequentially. The final results are included in Table 5.1. We show the true parameters values used to generate the sample together with the means and standard variations obtained from the MCMC routine. We add as well the values corresponding to the Maximum Posterior Estimation (MPE). As expected, the obtained fitted parameters show a good agreement for the centers of the clusters, but bigger uncertainties are found for the rest of parameters, and even tensions with the true values. The errors associated with the estimated number of points per cluster are

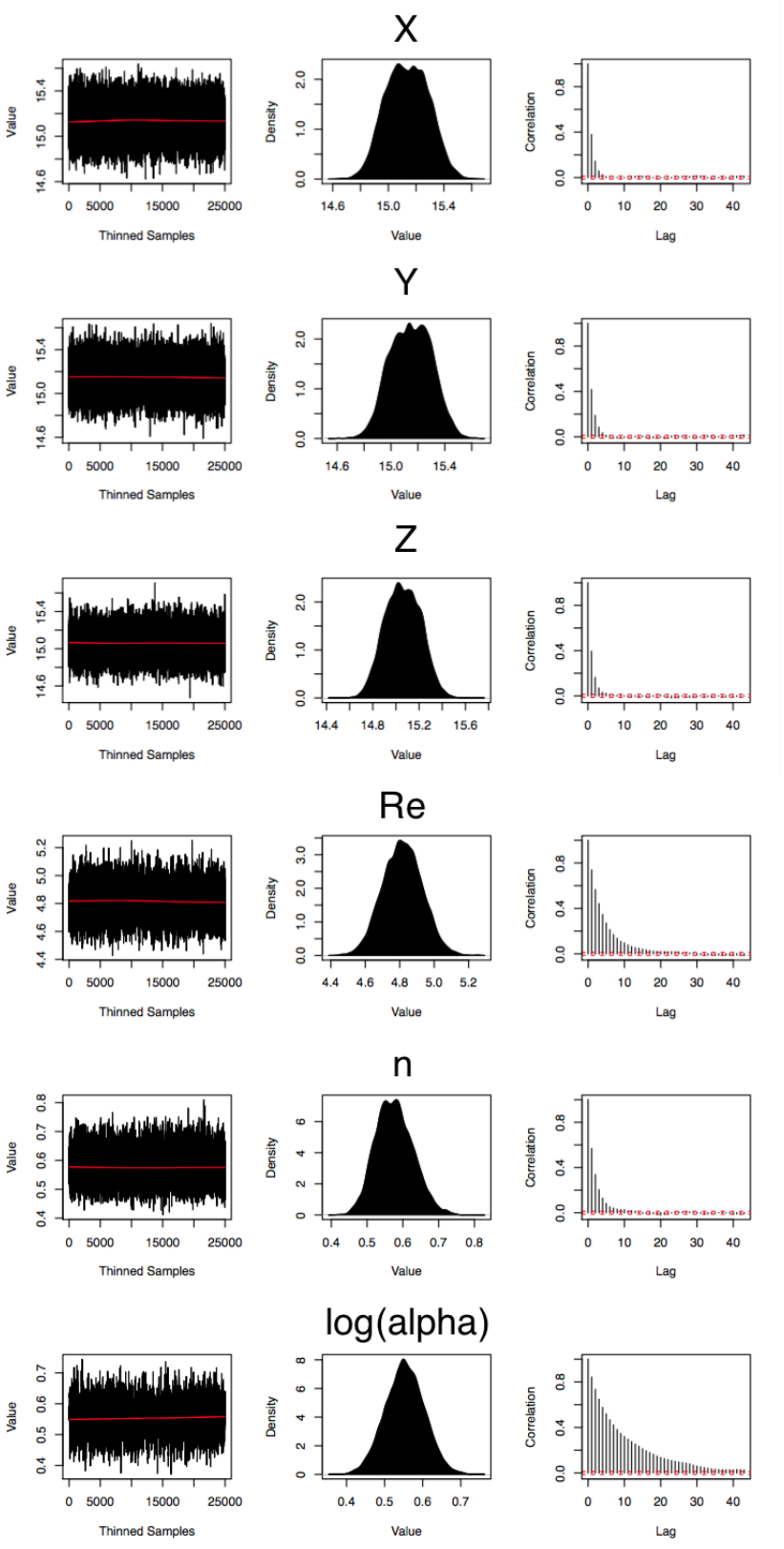

FIGURE 5.2: Results of the MCMC sampling for our Einasto toy model. We show the MCMC distribution of the 6 fitted parameters of a Einasto's cluster. For simplicity, we show only the distributions corresponding to the second cluster. The mixture coefficient α is showed in logarithmic scale. The left column show the values of the evaluated MCMC chain, in the center is the MCMC distribution of each parameter and the third column is the autocorrelation of the chain values.

calculated as the standard deviation of the estimation of number of points for each MCMC evaluated parameters set.

TABLE 5.1: Einasto toy model fitting

	$True^1$	$Mean^1$	MPE^1	$True^2$	$Mean^2$	MPE^2	$True^3$	$Mean^3$	MPE^3	$True^4$	$Mean^4$	MPE^4
x_0	-5	-4.6 ± 0.2	-4.5	15	15.14 ± 0.15	15.21	-12	-12.14 ± 0.14	-12.17	18	18.2 ± 0.2	18.1
y_0	-5	-4.9 ± 0.2	-4.8	15	15.15 ± 0.15	15.29	12	12.00 ± 0.14	12.06	-18	-18.00 ± 0.2	-17.95
z_0	-5	-4.9 ± 0.2	-4.9	15	15.06 ± 0.15	15.06	12	12.00 ± 0.15	11.88	-18	-18.3 ± 0.2	-18.24
r_e	7	7.08 ± 0.18	7.19	5	4.81 ± 0.11	4.9	6	5.86 ± 0.1	5.87	4	4.02 ± 0.15	4.18
n	0.7	0.58 ± 0.06	0.6	0.5	0.57 ± 0.05	0.55	0.3	0.32 ± 0.03	0.32	0.6	0.59 ± 0.09	0.57
α	1	1	1	3.48	3.6 ± 0.4	3.6	3.24	3.3 ± 0.3	3.3	2.53	2.4 ± 0.4	2.4
N	651	647 ± 28	652	717	713 ± 28	705	950	945 ± 30	934	288	287 ± 19	305

Summary of the MCMC Posterior probability results. Columns 1 shows the true values used for generating the sample. Columns 2 shows the mean and standard deviation of the Posterior distribution deduced from the MCMC calculations. Column 3 shows the MPE parameters. Each group of columns corresponds to a component. The background component mixture coefficient has mean $\alpha_c = 0.007 \pm 0.004$ (same results are found for the MPE) and number of estimated points $N = 191 \pm 15$ ($N = 185$ for the MPE, true value was 176).

After fitting the model $\lambda(\mathbf{r}; \alpha, \theta) = \frac{N}{M}\Sigma(\mathbf{r}; \alpha, \theta)$ as explained in section 5.2.3, we can evaluate the probability of finding a galaxy in a given point given the original distribution. In Fig. 5.3 we show the projection of the smoothed function $\lambda^\dagger(u, X)$ as introduced in equation 4.31, where we have used a smoothing bandwidth of $\omega = 3$. We can see how regions occupied by clusters are darker than the rest of the window, since these are the most likely locations for a galaxy to be found.

Now we can perform the different goodness of fit and residual analysis tests. The best fit produced information criteria $BIC = 16004.2$ and $AIC = 15861.85$. The amplitude reduction is $AR = 0.91$, which could be understood as a correct modeling of the 91% of the total data.

The raw residuals are shown in Fig. 5.4 (left). The amplitude of the errors has been greatly reduced and symmetrically oscillates around 0. Althugh biggest errors are still correlated with clusters locations, they always show a bimodal pattern between overestimated (blue) and underestimated (red) values.

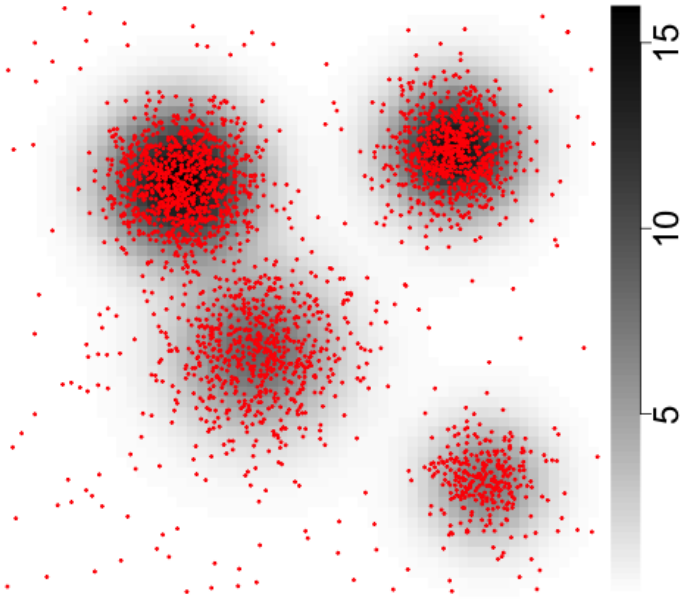

FIGURE 5.3: $\lambda^{\dagger}(u, X)$ smoothed probability function of the Einasto's toy model (in grey). In red, points from the generated halos plus background. Projection over the Z axis. Bandwidth $\omega = 3$.

The relative errors function $e(u)$ shows the non modeled data (Fig. 5.4). None of this data is inside the clusters. The biggest non modeled structures are points from the background. It is known that a Poisson process also shows over- and underdensities at the right scales. With a constant model, such as the background component, shotnoise peaks and voids will appear as non modeled structures. This is the maximum efficiency we can expect. Once we reach this level, we can say that no structures of interest are left to model. The difficulty resides in realizing we have reached this level.

The lurking plots, as introduced in 4.3.1, are shown in Fig. 5.5, integrating the raw residuals over disjoint sections of the X coordinate. The peaks that this image is showing us coincide with the locations of the clusters, with much lower values for regions only occupied by the background.

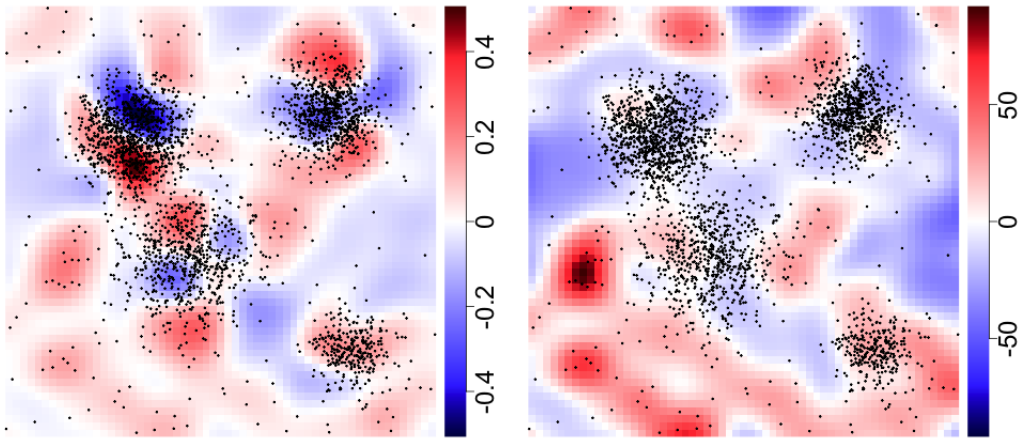

FIGURE 5.4: Smoothed raw residuals $s(u)$ (left, eq. 4.30) and relative smoothed raw residuals (right, eq. 4.32) of the Einasto's best fit parameters for our toy model. Bandwidth $\omega = 3$.

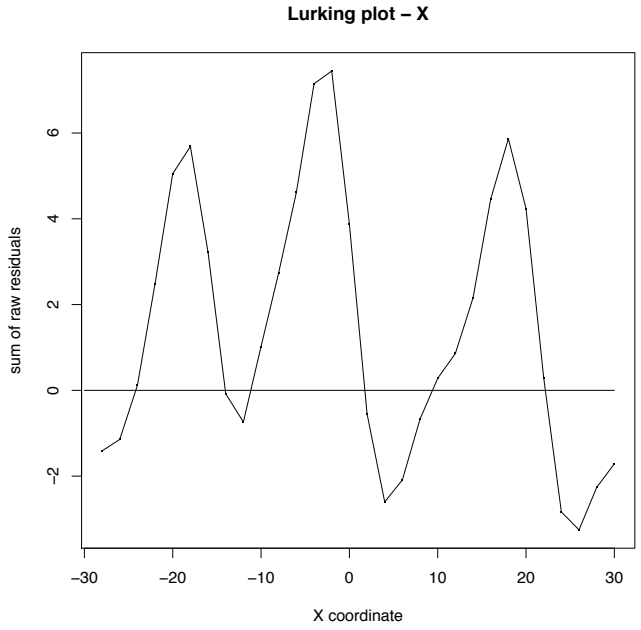

FIGURE 5.5: Lurking plot of the Einasto's best fit parameters for our toy model. Bandwidth $\omega = 3$.

Further tests

As expected, this example is satisfactorily fitted and the best fit values are reasonably close to the true values. We can enrich the study of the mixture modeling with variations over the original case. We know beforehand that our sample contains four well defined clusters. However, in the cosmological scenario, deciding the number of components to be included in the model is a delicate task with decisive consequences in our results. Many different techniques can be used to identify overdensities in populations, which indeed are non parametric cluster finding algorithms, such as the friends-of-friends algorithm or k-means. Density field estimators, like a Gaussian kernel function, can be used as well. However, it might be necessary to make use of posterior confirmations.

As introduced in section 5.2.3 The Bayesian and Akaike Information Criteriums (BIC and AIC) might be used to check the quality of a mixture model fitting depending on the number of used components. This technique requires to fit as many models as different numbers of components are going to be used. For this reason we calculated these values in the previous examples and we are going to repeat the analysis over different modified versions so we can compare. For fitting the model in these samples we will make use of the Nelder-Mead algorithm only.

Insufficient cluster components

If we only consider three cluster components in our sample of Einasto's clusters, we will obtain a mixture model where one of the clusters is unmapped. In order to cover the extra density left by the non-existent cluster, the background component or the closer clusters will increase their density. Since in our example clusters are clearly separated, it is the background component which has increased its value from $\alpha_c = 0.007$ to $\alpha_c = 0.018$, more than twice the original value, which implies doubling the density of galaxies.

The detection of this lack of fitting due to an underestimation of the cluster components can be done using the residuals plots, which clearly show the lack of fitting as shown in Figs. 5.6 and 5.7. The relative errors function $e(u)$ is specially effective to separate unmapped components from noise.

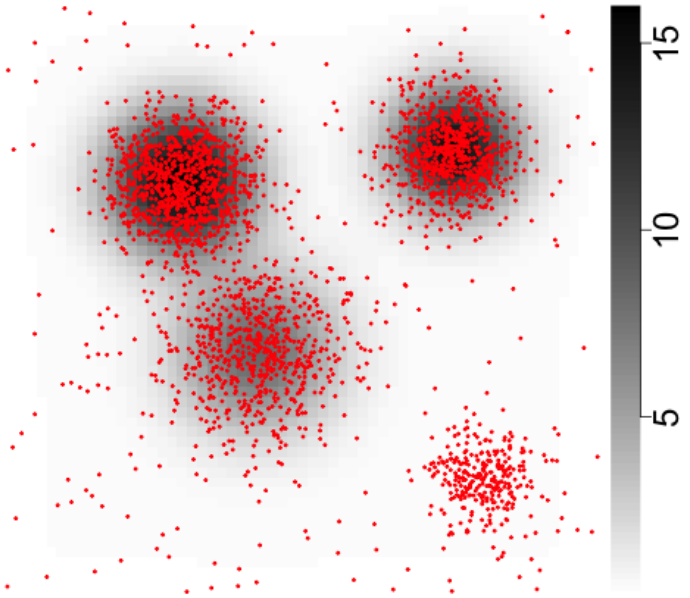

FIGURE 5.6: $\lambda^{\dagger}(u, X)$ smoothed probability function of the Einasto's toy model (in grey). Points from the generated halos plus background in red. Only three clusters have been used in this model. Bandwidth $\omega = 3$.

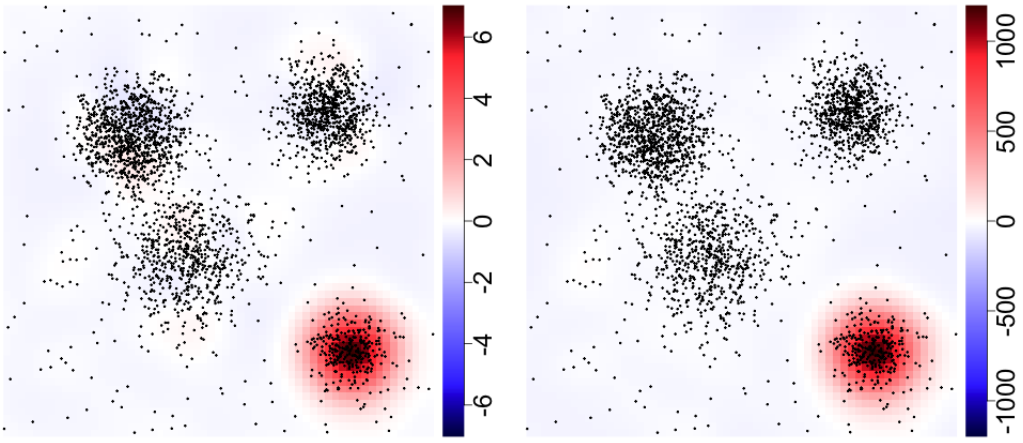

FIGURE 5.7: Smoothed raw residuals $s(u)$ (left, eq. 4.30) and relative smoothed raw residuals (right, eq. 4.32) of the Einasto best fit parameters for our toy model. Only three clusters have been used in this model. Bandwidth $\omega = 3$.

Finally, the information criteria show worse values than in the previous example due to a smaller likelihood value, with BIC = 18102.87 and AIC = 18209.62, confirming that the presence of a fourth cluster component was needed. Properly understanding this trivial example with a toy model will be of great help with more complex cases.

Excessive cluster components

Alternatively, we can fit the data sample with an excessive number of cluster components. In this case, unnecessary components will show parameters with higher error bars and a much smaller mixture coefficient compared with the rest of components. The contribution of the extra component would be probably confused with that of the background component, a flat and wide profile shape. The center of this component highly dependent of the initial parameters. After trying several tryouts, we sometimes found this component located around an equidistant point to the rest of the clusters, and sometimes cornered against the borders of the window.

While the BIC has a bigger value when an unnecessary component is introduced, the AIC is slightly smaller. In principle, it should be understood that adding an extra component was an appropriate decision, however, we know this is not the case, which make us realize the importance of never relying on a single goodness of fit criterium but to test instead as many tools and residual analysis we may have. The low mass of the extra component made the visualization of functions $\lambda^{\dagger}(u, X)$, $s(u)$ and $e(u)$ very similar to those of the original example with 4 clusters and 4 components. For this reason, no images are shown for this case.

Unpredicted structures

The irregularity of galaxy samples demands a preliminary analysis of what happens when the galaxy distribution includes structures not meant to be mapped by our dark matter halo profiles. For example, a highly anisotropic elongated structure. We add around 300 extra galaxies describing a scattered line crossing the window (Fig. 5.8).

The mixture model of four components might suffer a perturbation of the cluster component parameters, specially those situated closer to the new structure. Again,

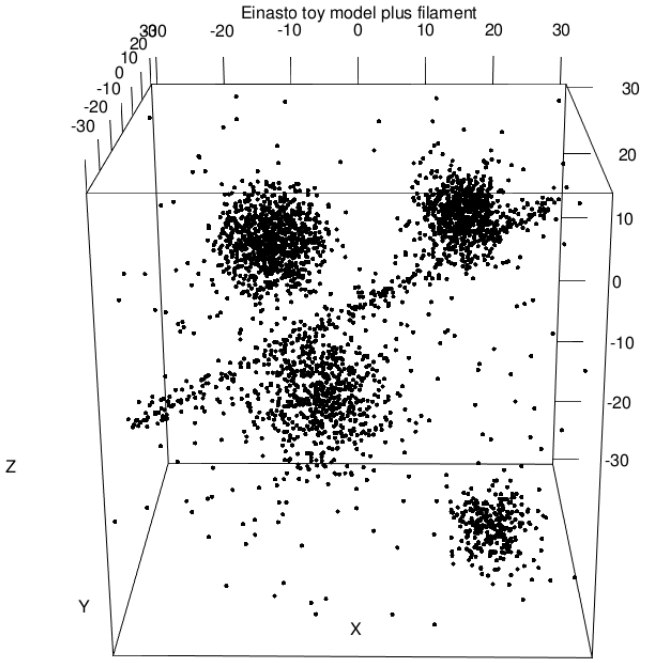

FIGURE 5.8: 3D samples of four Einasto galaxy clusters with background and a filamentary structure.

it is with the residual analysis that we obtain a clear vision of a lack of fitting (Fig. 5.10).

The detection of these unmapped structures can be seen in a complementary way. If the mixture model is properly fitted and the number of cluster components is the right one, the residuals and the function $e(u)$ should show only noise and non clustered structures. If these structures are significant, they should prevail over the noise, and therefore, detectable. This is the same strategy used to detect a lack of cluster components, but applied on unknown structures. In this case, the information criteria obtained values BIC = 18653.15 and AIC = 18487.15.

Just in order to see what happens when we add extra components to this sample, we fit a mixture model of five cluster components. We located the initial values of parameters corresponding to the new component in the center of the window, which, after fitting, drift to an intermediate position between the clusters and over the filamentary structure. As seen in Fig. 5.11, we see how new components are incapable of properly describe an elongated shape and instead have to adapt its radius and other shape parameters to the structure thickness.

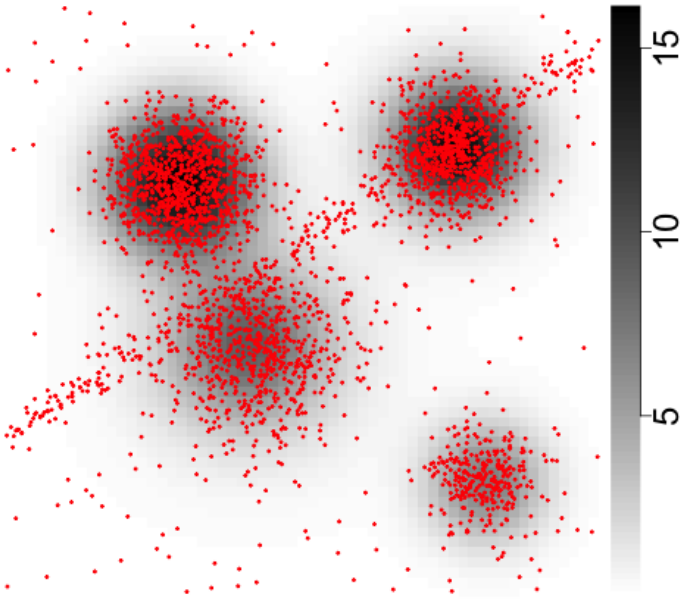

FIGURE 5.9: $\lambda(u, X)$ conditional probability function of the Einasto toy model with a filamentary structure (in grey). Points from the generated halos with background and a filamentary structure in red. Bandwidth $\omega = 3$.

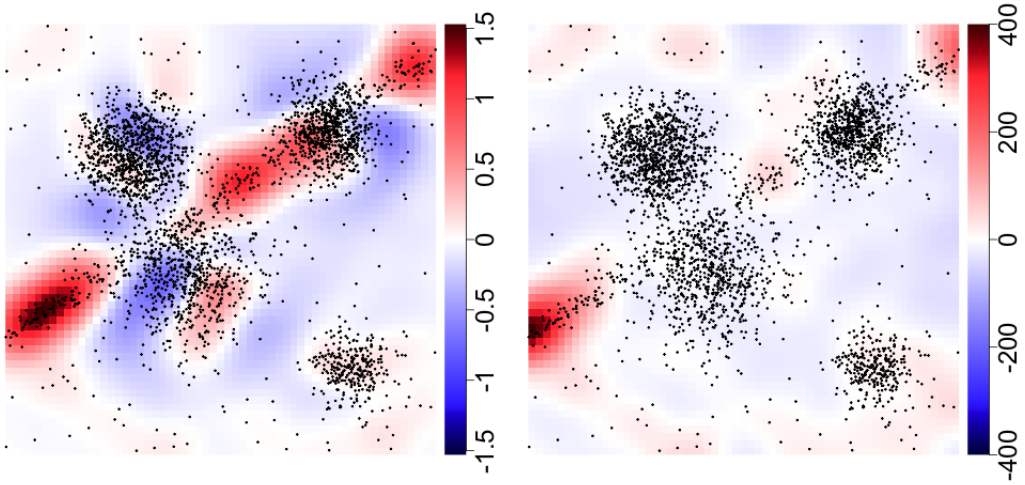

FIGURE 5.10: Smoothed raw residuals $s(u)$ (left, eq. 4.30) and relative smoothed raw residuals (right, eq. 4.32) of the Einasto best fit parameters for our toy model. An elongated structure has been added to this sample. Bandwidth $\omega = 3$.

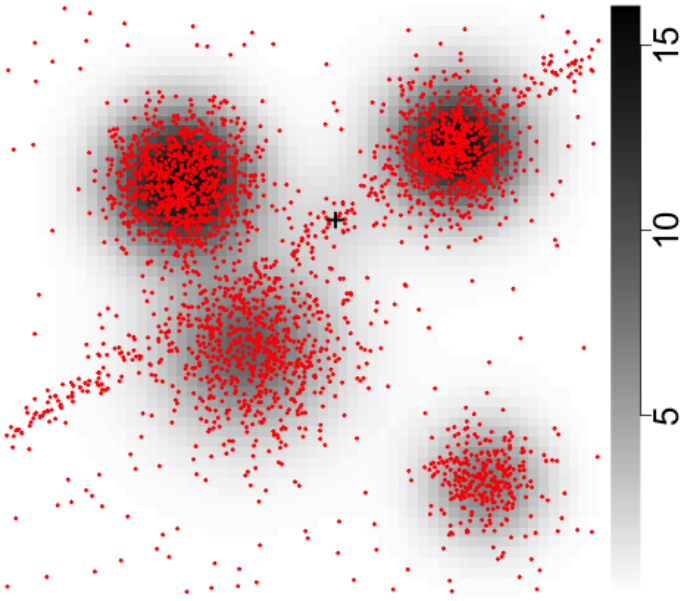

FIGURE 5.11: $\lambda^{\dagger}(u, X)$ smoothed probability function of the Einasto toy model (in grey). Points from the generated halos with background and a filamentary structure in red. Five cluster components have been used in this model. Black cross represents the center of the fifth component.

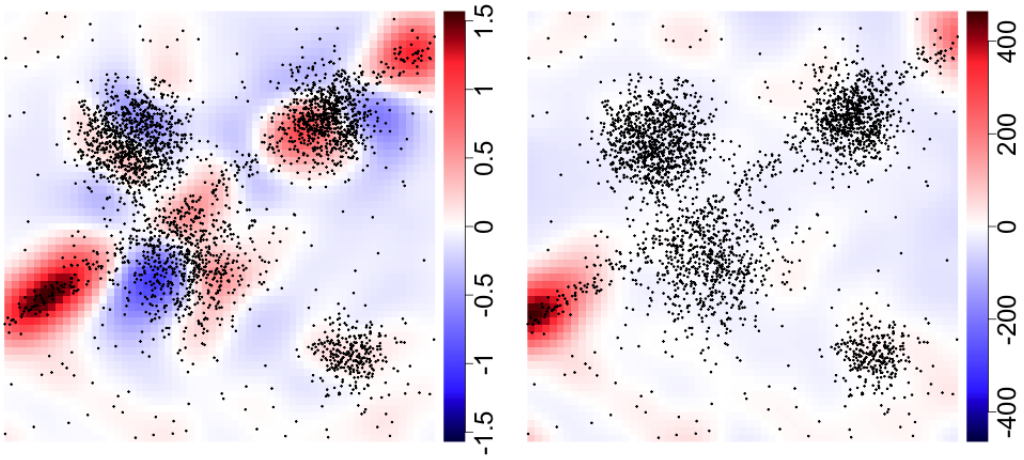

FIGURE 5.12: Smoothed raw residuals $s(u)$ (left, eq. 4.30) and relative smoothed raw residuals (right, eq. 4.32) of the Einasto best fit parameters for our toy model. An elongated structure has been added to this sample and five cluster components have been used in this model.

The residual plots (Fig. 5.12) show how the errors have been reduced around the new 'cluster'. The BIC and AIC also showed an improvement in their values, slightly smaller when no attempt to model the filament is done: BIC = 18470.6 and AIC = 18252.6.

However, it is clear that this kind of structures demand proper models.

Overlapping clusters

Mixture models are specially interesting when used to disentangle different distributions mixed together in the same region, where a disjoint partition of the window (like a hard classification method) would not respect the shape of the distribution tails. In the above examples, clusters are separated enough to recognize most of their population. However, Mixture models show the same efficiency when two or more clusters share a region of the window with a significant amount of their volume distribution.

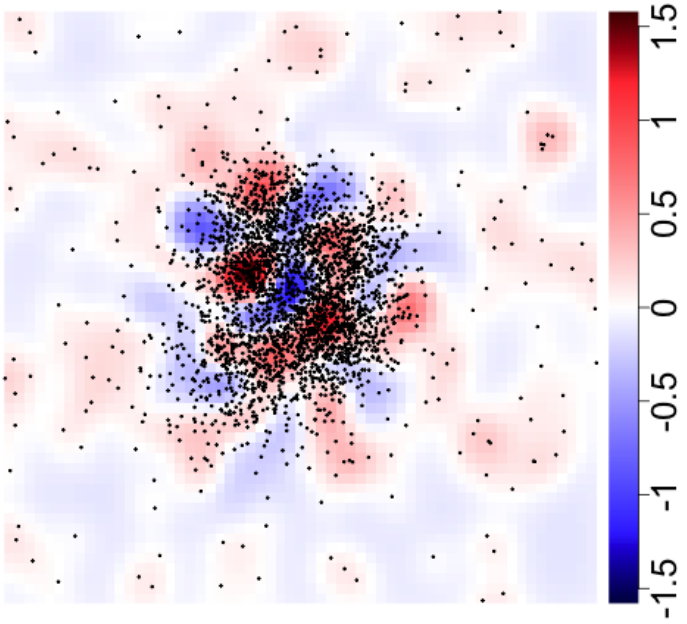

FIGURE 5.13: Smoothed raw residuals $s(u)$ (eq. 4.30) of the Einasto best fit parameters for a group of 4 overlapping clusters. Bandwidth $\omega = 2$.

In order to show this, we have generated four new clusters with close centers, in such a way that they overlap in the center of the window. The fitting and posterior residuals analysis show the expected good results. Even when it is hard to determine to which cluster each point belongs, the volume estimation is correct and the resulting estimated number of points per cluster is satisfactory. The results with true values in brackets for the four clusters and the background are: 675 (660), 728 (756), 989 (959), 296 (299) and 187 (200).

In Fig. 5.13 we show the raw residuals plot.

5.4.3 MultiDark simulation

We finally are in position to apply the presented methods on a more complex dataset. The MultiDark sample presented in section 5.3 will be deeply analyzed in this section. It can be seen in Fig. 5.14 with abundant structure and variations in the shape and size of its clusters.

We start with the calculation of the density field, using a Gaussian filter function, obtaining a list of local maximum points, candidates to become the initial values of the clusters centers ($\mathbf{r_0}$). With bandwidth $\omega = 1$, our examination gives us the values summarized in Table 5.2. We show only the 11 densest points.

We performed the fitting of a mixture model with number of cluster components from $k = 3$ to $k = 11$, including each time less dens positions. From 3 to 10 we observed a monotonic improvement of both the Maximum Likelihood and the Bayesian and Akaike Information Criteria. However, components 7 to 11 degenerated their best fit parameters n to unphysical values, probably due to the small number or points around these positions, their proximity to the edge of the window or an irregular shape of the overdensity. For cluster sized halos, this sample, n is expected to be below 8 (Merritt et al., 2006), however, we found how in these low populated overdensities the Sérsic index grows far above that limit, turning the Einasto profile into a power-law shaped curve. A power-law profile models the existence of a few points in a small region, concentrating the mass of the component in a single location. This does not correspond to the galaxy clusters we aim to study and we decide to limit the index to $n < 30$ in our fitting process, avoiding numerical degeneracies. For this reason, and even if the information criteria

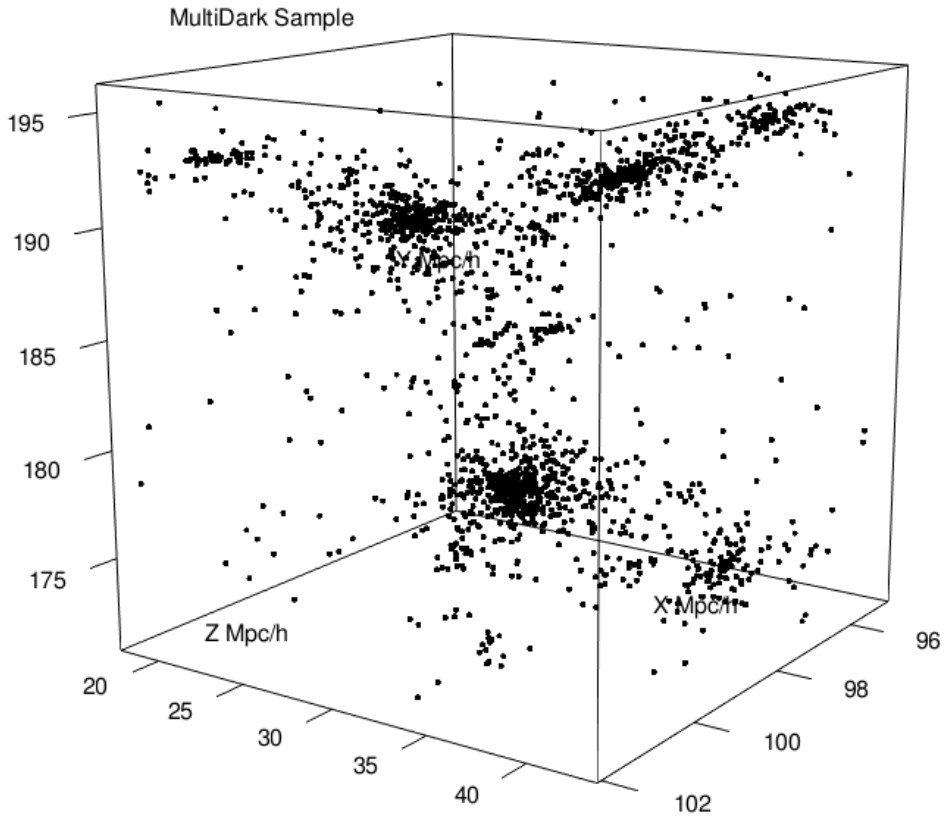

FIGURE 5.14: 3D sample of dark matter particles from the MultiDark simulation. Units in h^{-1} Mpc. We use it as a simulation of a galaxy point process.

supports a higher number of components, we have decided to perform our analysis with 6 components. For comparison, at the end of this work, we present as well the same results for 10 components.

The initial values of our parametric set Θ includes the local maximum points of the density field, and starting values $r_e = 1$, $n = 1$ and $\alpha = 1$. This gives us 36 parameters to be fitted. As explained in section 5.2.3, it is recommended to perform an MCMC analysis of the parameters distribution. This allows us to check of possible multimodalities in our parametric space indicating a lack of components, ensures a non-pathological distribution of our parameters and gives us an estimation of the confidence limits.

As an example, we show in Fig. 5.15 the parametric distribution of component 2 (corresponding to point 2 in Table 5.2). For 3.600.000 iterations, we found a good

TABLE 5.2: Density field local maximum points

C	x h^{-1} Mpc	y h^{-1} Mpc	z h^{-1} Mpc	ρ h^3 Mpc^{-3}	$\log L$	BIC	AIC
1	33.52	178.76	99.68	12.71			
2	36.74	192.38	98.77	8.92			
3	25.98	189.42	98.89	7.78	446.52	−793.7	−880.02
4	38.59	193.83	96.38	3.19	653.91	−1124.44	−1283.81
5	39.36	174.26	97.53	2.77	1022.55	−1815.88	−2015.1
6	20.59	192.61	100.77	1.82	1277.72	−2272.75	−2518.45
7	32.87	185.54	99.919	1.22	1427.59	−2534.27	−2771.17
8	25.21	181.97	95.41	1.06	1566.72	−2766.69	−3037.44
9	33.26	189.79	99.13	1.04	1613.93	−2815.26	−3119.86
10	33.43	172.23	100.21	0.31	1661.33	−2864.22	−3202.66
11	27.22	173.98	96.93	0.30	1673.2	−2842.12	−3214.4

Column C is the number of maxima from the density field (bandwidth $\omega = 1$) modeled by the Mixture model. Columns x, y and z include the local maxima points coordinates in the density field sorted decreasingly in density ρ. Column $\log L$ shows the obtained log-likelihood for the best fit model with C components, using the local maxima as initial parameters. Columns BIC and AIC are the corresponding information criteria for the used number of components and $\log L$.

chain mixing, despite the strong autocorrelation (column 3) in parameters r_e and $\log_{10}(\alpha)$. Only for $\log_{10}(\alpha)$ (bottom) we reached no stationarity. This parameter might need a higher number of iterations, since it is more strongly correlated with the rest of components, or the problem is of a different nature.

The mean of the parameters Posterior distribution obtained by the MCMC routine can be consulted in Table 5.3 with error bars. We also include the Maximum Posterior Estimation.

The background component was fitted with a mixture coefficient of $\alpha_c = 0.008 \pm 0.004$, which means that around 442 galaxies do not belong to any cluster. However, the big uncertainty of this parameter gives us little confidence on it. After the analysis performed with the toy models, we have decided to trust the parameters obtained as the average of the posterior distributions. Hereafter, all analysis performed will make use of these values.

The amplitude reduction obtained with this parametric set is $AR = 0.72$, i.e. a

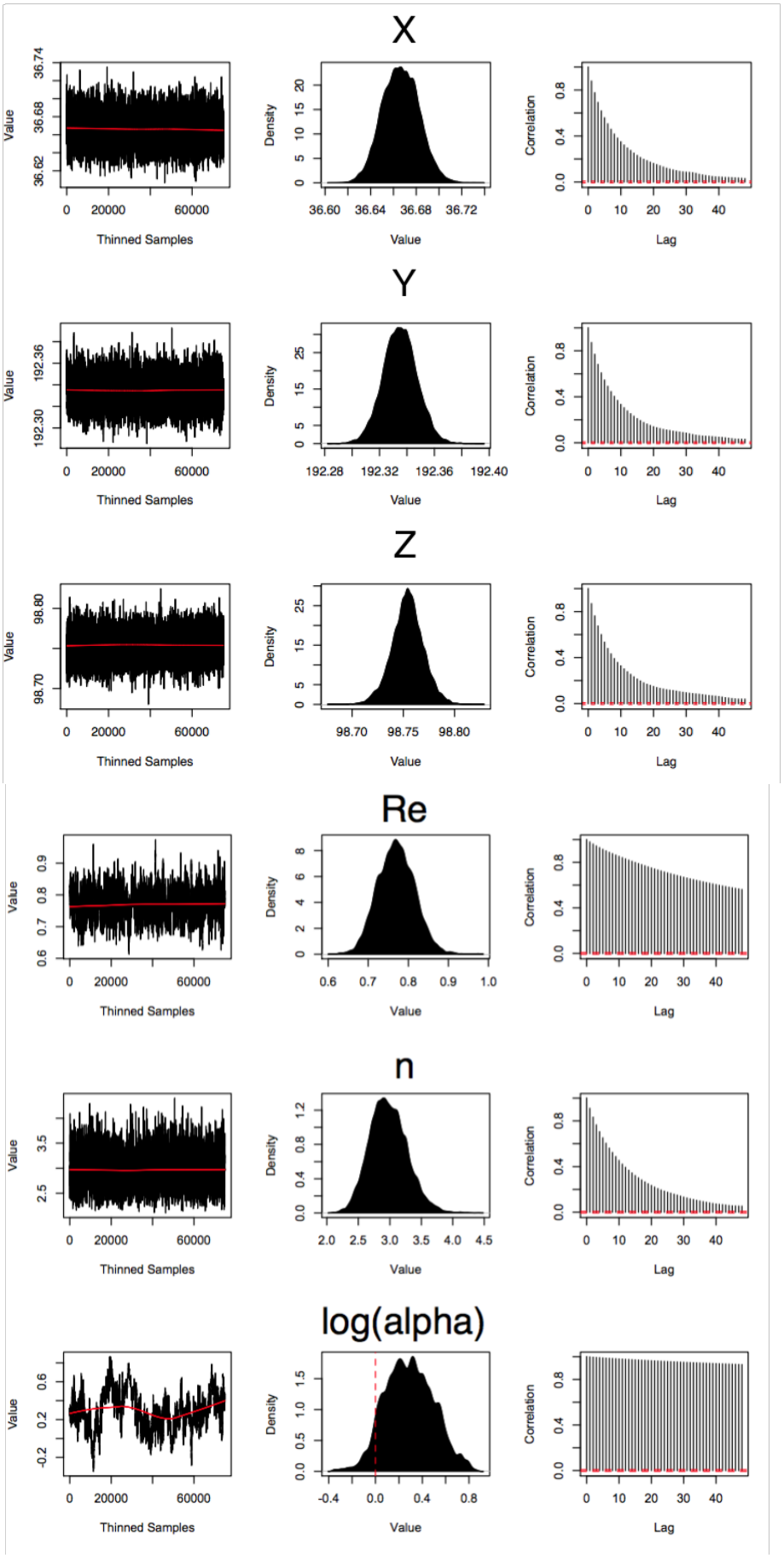

FIGURE 5.15: Parametric distribution of component 2. From top to bottom: x_0, y_0, z_0, r_e, n and $\log_{10}\alpha$.

TABLE 5.3: MultiDark fitting for 6 components

	Mean and standard deviation					
	C_1	C_2	C_3	C_4	C_5	C_6
x_0	33.46 ± 0.014	36.67 ± 0.016	26.00 ± 0.03	38.60 ± 0.04	39.47 ± 0.02	20.40 ± 0.02
y_0	178.8 ± 0.018	192.33 ± 0.012	189.35 ± 0.03	193.88 ± 0.03	174.22 ± 0.02	192.72 ± 0.019
z_0	99.7 ± 0.016	98.75 ± 0.015	98.87 ± 0.02	96.18 ± 0.03	97.44 ± 0.03	100.81 ± 0.018
r_e	1.05 ± 0.04	0.77 ± 0.05	1.11 ± 0.07	0.75 ± 0.08	0.90 ± 0.1	0.56 ± 0.08
n	2.4 ± 0.19	3 ± 0.3	2.5 ± 0.25	2.2 ± 0.4	3.2 ± 0.6	3.2 ± 0.9
α	1.5 ± 0.9	2 ± 1	0.75 ± 0.4	0.7 ± 0.5	0.4 ± 0.3	1
N	615 ± 22	351 ± 19	373 ± 19	100 ± 11	123 ± 12	71 ± 8
	Maximum Posterior Estimation					
x_0	33.51	36.7	26.05	38.59	39.43	20.36
y_0	178.81	192.33	189.34	193.86	174.24	192.71
z_0	99.69	98.72	98.87	96.17	97.48	100.85
r_e	1.09	0.8	1.05	0.79	0.73	0.38
n	2.7	3	2.8	3.3	2.0	3.5
α	0.52	0.79	0.36	0.23	0.31	1
N	598	376	391	106	100	56

Best fit estimation of the 6 component Mixture model parameters for the MultiDark sample. Component 6 has been used as the first component in the model, and therefore $\alpha_6 = 1$. Top: Mean and standard variation of the posterior distribution. Background component has $\alpha_c = 0.008 \pm 0.004$, 442 ± 26 galaxies. Bottom: maximum posterior estimation. Background component has $\alpha_c = 0.008$, 453 galaxies.

72% of the data is properly modeled. In Table 5.3 we see that r_e and n values are around the expected physical values, with Einasto radii around $1h^{-1}$ Mpc and Sérsic index around 3. Merritt et al. (2006) found similar values for cluster size halos. We appreciate how the mixture coefficients α_i and the estimated number of particles per cluster N are generally correlated with the densities in Table 5.2. The radii r_e and the Sérsic index n, the shape parameters, are uncorrelated with this table.

In Fig. 5.16 we compare the data density field ($\lambda^*(u, X)$, left) with our smoothed model ($\lambda^\dagger(u, X)$, right) where we can see a clear agreement, both in the distribution and size of the halos. The main structures are mapped and the amplitude of the model is similar to that of the data density field. The main differences between data and model rely on the asymmetry of the real structures, difficult to describe with spherical halos. In addition, the data density field maps the shot noise corresponding to the background, while our constant model can only reach

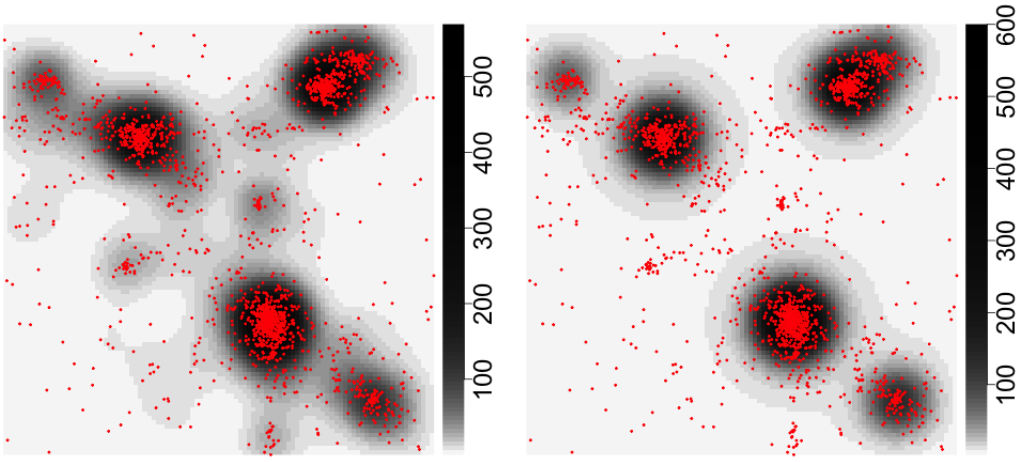

FIGURE 5.16: Left: $\lambda^*(u, X)$ density field estimation. Right: $\lambda^\dagger(u, X)$ of the MultiDark sample with our Mixture model (6 components). Dark area indicate higher density field or conditional probability of being occupied by a point. Densities showed in logarithmic scale for clarity. Smoothing bandwidth is $\omega = 1$. Values summed over dimension Z.

this level of detail.

The residuals plot (Fig. 5.17, left) shows a significant reduction of the data amplitude, satisfactorily modeling the cluster components. Errors are distributed with mean 0, and most clusters have underfitted and overfitted regions, as analyzed in section 5.4.2. Bad fitting happens with the biggest clusters (components 1 and 3), which show the characteristic double (red and blue) pattern. However, this is an indicative of a good fit, where the model has found and mapped the structures, minimizing the errors around the component. Nevertheless, the relative errors (Fig. 5.17, right) show a lack of fitting for those structures which count with no component in the model to be described by, mainly the overdensities corresponding to peaks 7, 8 and 10 in Table 5.2 (big red spots). As explained, we tried to include these structures in the model but the Einasto profiles were not able to describe them satisfactorily. The $e(u)$ function is a very useful tool to discover those structures that have not been described by the model.

The presence of each cluster component and their fitted radii can be clearly seen in Fig. 5.18. This result can be reproduced using rejection sampling techniques, as we did to generate the toy models. Using the fitted parameters, we can generate spherical dark matter halos centered at the same place that the original halos, with

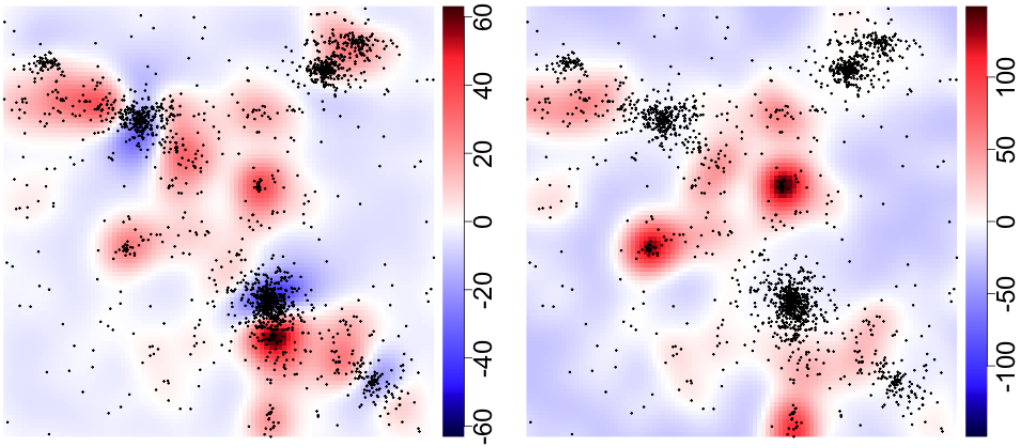

FIGURE 5.17: Smoothed raw residuals $s(u)$ (left) and relative residuals $e(u)$ (right) of the MultiDark sample with our Mixture Model. Smoothing bandwidth is $\omega = 1$. Values summed over dimension Z.

shapes following the estimated r_e and n and populated with the deduced numbers of particles N. We show this in Fig. 5.18 together with the real sample (left) so they can be compared. The overall agreement is very satisfactory, reproducing the clusters with close shapes and sizes. It is with the background component where we detect an excess of galaxies. The absence of peaks 7, 8 and 10 in our model might be the cause of overpopulating the background with galaxies that do not belong to any component.

The quality fit of each cluster can be checked using our estimations of their density profiles. As an interesting example, we compare the profile density curves of clusters 2 and 5. These clusters are located close of one another in the top right corner of the image, and their profile curves rapidly overlap. The contribution of each cluster to the profile curve of its neighbor can be easily inferred and isolated in Figs. 5.19. As explained in section 5.2.2, given a Mixture model and one of its components, we can calculate the fitted parametric density profile of a single component ($\bar{\rho}_i(\mathbf{r} - \mathbf{r_0}; \theta)$, in green) or the same density profile when the whole model is taken into account ($\bar{P}_i(\mathbf{r} - \mathbf{r_0}; \theta)$, in blue). This calculation does not require any additional fitting.

The red line ($\bar{P}_i$) follows remarkably well the empiric profile, coinciding with the green one for short distances. As can be seen, the empiric profile (black circles) shows a bump in this density around $r = 4h^{-1}$ Mpc, the distance that separates

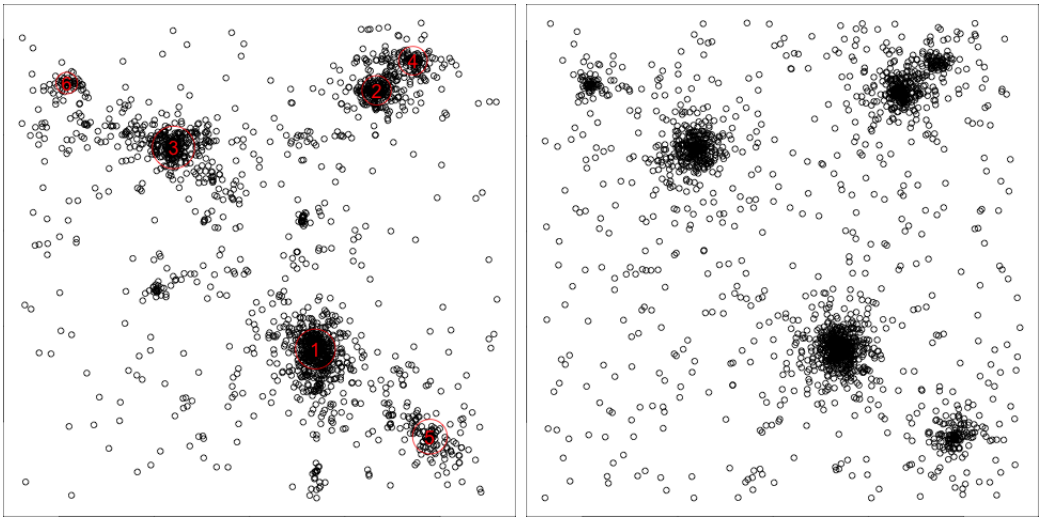

FIGURE 5.18: Left: Clusters and fitted Einasto radii of the MultiDark sample with our Mixture Model. Cluster components are sorted as in Tables 5.2 and 5.3. Right: Particle sample generated with rejection sampling method following the best fit values of Table 5.3.

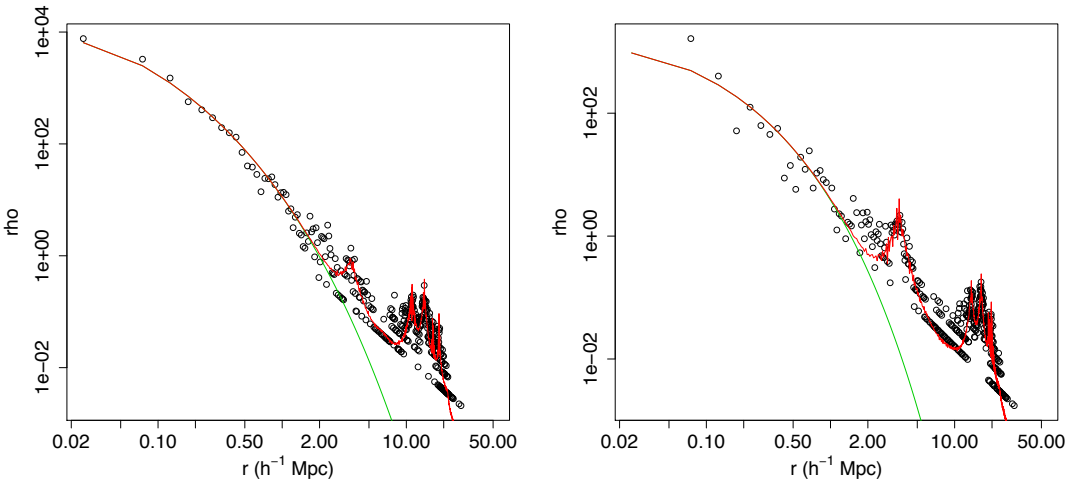

FIGURE 5.19: Density Profile curves of components 2 (left) and 4 (right). Black circles represent the densities at the galaxies locations, the red line is the density profile of the whole Mixture Model ($\bar{P}_i(\mathbf{r}-\mathbf{r_0};\theta)$) and the green line is the density profile of each isolated component ($\bar{\rho}_i(\mathbf{r}-\mathbf{r_0};\theta)$).

the center of components 4 and 5. This bump is the contribution of each cluster to the other cluster density profile, and it can be separated making use of our best fit Mixture model. The green line ($\bar{p}_i$) separates from the total profile $\bar{P}_i$ after the contribution of neighboring structures is significant. Mixture models allow us to calculate the real profile of clusters beyond the limits of their empirical profile. This can be used to obtain a better understanding of objects with strong interactions with neighbors.

Regarding the initial decision of fitting a 6 or more than 6 cluster component model, we state that this is the best option, since all fitted parameters were inside the expected values for halos of this size. However, 10 components could be an interesting choice as well, based on the minimum reached by the Bayesian Information Criteria, despite the odd behavior of the Sérsic index n. In Table 5.4 and Fig. 5.20 we show the results for a 10 components Mixture model.

TABLE 5.4: MultiDark fitting for 10 components

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
x_0	33.46	36.66	26.00	38.59	39.47	20.41	32.80	25.08	32.69	35.48
y_0	178.8	192.33	189.35	193.87	174.22	192.72	185.50	181.84	189.95	174.64
z_0	99.70	98.76	98.88	96.18	97.44	100.81	100.05	95.20	99.04	99.09
r_e	1.12	0.83	1.25	0.80	1.03	0.60	0.74	0.71	0.46	3.00
n	2.54	2.97	2.52	2.04	3.13	8.76	26.70	25.17	26.82	0.09
α	1.91	2.34	0.83	0.93	0.48	1	0.15	0.14	0.25	0.04
N	621	336	380	100	135	87	37	30	15	57

Best fit estimation of the 10 components mixture model parameters for the MultiDark sample. Background component has $\alpha_c = 0.08$, 284 galaxies.

The similarity of all values for components 1 to 6 with those from Table 5.3 shows that the new clusters do not significantly affect the rest of the structures. As said, notice the odd values for n in components 7 to 10.

The simulation of our data (Fig. 5.20, bottom right) is now much closer to the MultiDark pattern. The 10th component (number 0 in Fig. 5.20, bottom-left) was meant to model the concentration of points close to the bottom border. However,

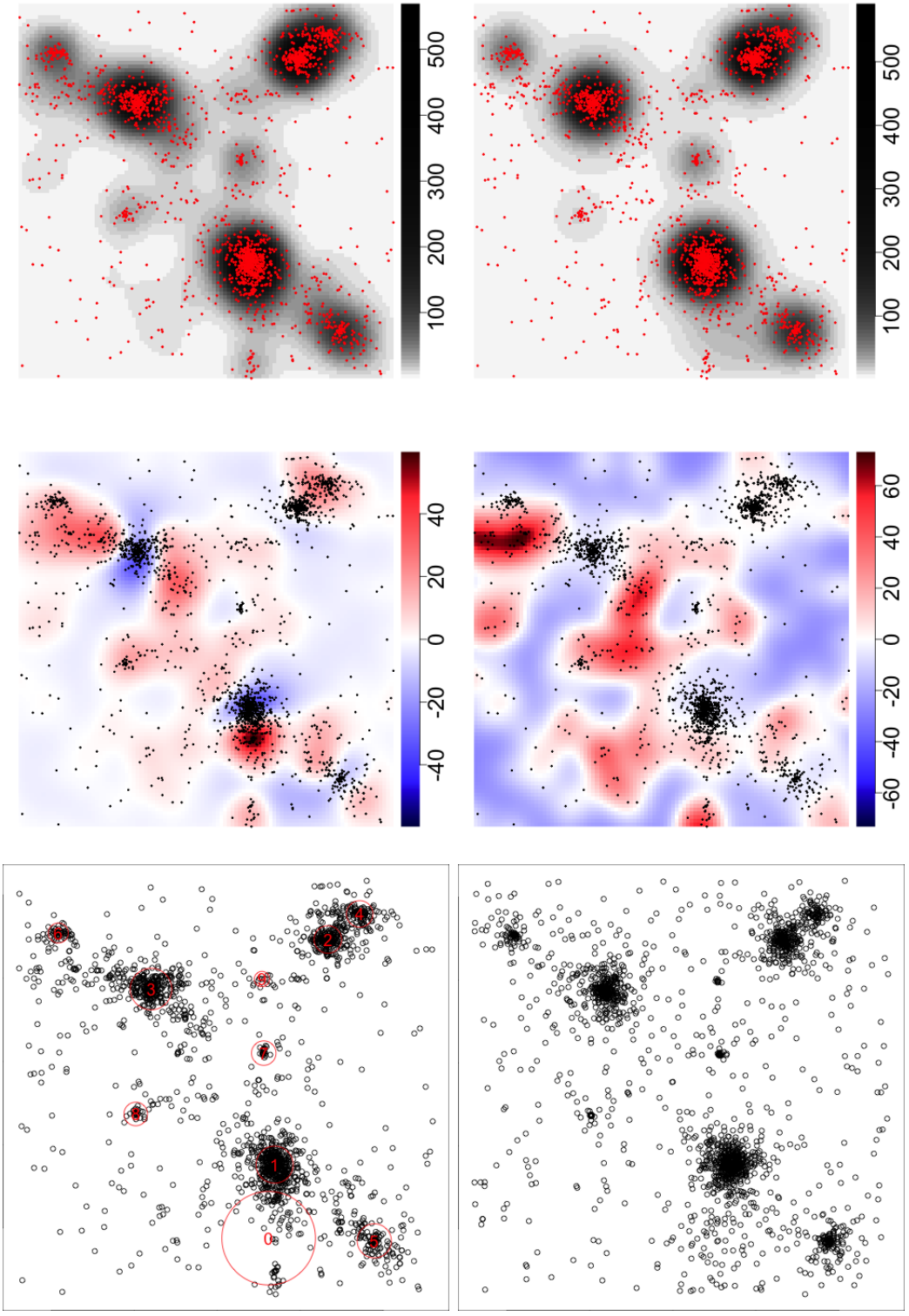

FIGURE 5.20: Top left: $\lambda^*(u, X)$ density field estimation (same as in Fig. 5.16). Top right: $\lambda^\dagger(u, X)$ of the MultiDark sample with our Mixture model (10 components). Dark areas indicate higher density field or conditional probability of being occupied by a point. Densities showed in logarithmic scale for clarity. Middle left: smoothed raw residuals $s(u)$. Middle right: relative residuals $e(u)$ (smoothing bandwidth is $\omega = 1$, values summed over dimension Z). Bottom left: Clusters and fitted Einasto radii of the MultiDark sample with our 10-component Mixture model. Cluster components are sorted as in Table 5.4. Bottom right: Particle sample generated with rejection sampling method following the best fit values of Table 5.4.

due to its irregular shape and the proximity to the border, this component degenerated into an intermediate position between the group of particles and component 1, reducing its contribution to the mass of the model with a low mixture coefficient ($\alpha_{10} = 0.22$) and a wide radius ($r_e = 0.42$). That clearly states that component 10 should not be considered despite the information criteria.

5.5 Conclusions

In this chapter we have presented and tested the Mixture Modeling technique, a modeling and data mining method capable to classify structures which are not clearly separated. Such a tool is of major interest to the analysis of galaxy structures at a wide range of scales, where different objects mix and it is crucial to understand their coupling. From galaxy satellites to subhalos, these structures extend their limits beyond the limits of their neighbors, deeply affecting the nature of other objects.

Hard classification methods, such as friends-of-friends and k-means are unable to realistically classify this kind of configurations, forcing every element (particle or galaxy) to belong to one and only one component. With Mixture modeling we classify the points in a soft way, which means that the membership of a particle to a given component is a probability. This opens the door to the application of the abundant point process methodologies and results, the probabilistic analysis of point patterns.

In this initial work we have chosen to model the galaxy clusters making use of a high density dark matter simulation, MultiDark, which contains the perfect samples both for testing and proving that results of physical interest can be obtained. Galaxy clusters have been modeled using the dark matter Einasto profile. This introduces the parametric modeling, and therefore, a direct method of synthesizing the relevant properties of our studied clusters.

The parametric soft classification method of Mixture models includes further applications, such as the generation of probabilistically equivalent data samples or the isolated analysis of components, which allows us to understand the nature of our objects in mixing environments.

In order to continue this work, we plan to classify the distribution of galaxies in the local universe. Reliable results require high quality data, with precise redshift estimation and a high density of particles. We expect to find these conditions in the recently published catalogs of the nearby universe (Tempel et al., 2016). The study of satellite galaxy and their interaction with the central galaxy can be also dealt by Mixture models. As an improvement of the model, we plan to include non spherical profiles, for example with the inclusion of elongated ellipsoids (Kuhn et al., 2014).

The results shown in this chapter will be published in Hurtado-Gil et al. (in prep.a).

Chapter 6

Conclusions

In this thesis work we have studied the galaxy distribution using point process analysis techniques. Making use of three possible approaches, summary statistics, data mining and modeling, we have been able to obtain relevant conclusions about the galaxy distribution in the universe, as well as to develop new methodologies that allow us to analyze in detail the galaxy structures and interaction. The main conclusions of this work are as follows.

Using data from the Sloan Digital Sky Survey (Blanton et al., 2005), we performed a blind analysis of the best fit of the observational distribution of Counts-in-Cells (Chapter 2). The analysis found that the modeling of such distribution has no clear answer. For cells of radii 12 and $24h^{-1}$ Mpc, the Log Normal distribution plus a bias parameter is the best known distribution at representing this statistic of the galaxy distribution. The Log Normal distribution, proposed by Coles & Jones (1991), is a non linear modification of the Gaussian distribution. However, it was found that its fit to the galaxy distribution was inadequate, overcome by other proposals such as the gravitational quasi-equilibrium distribution (GQED, Saslaw & Hamilton (1984)) or the negative binomial distribution (NBD, Carruthers & Duong-van (1983)). In the recently published work by Arnalte-Mur et al. (2016), the authors use the Log Normal distribution for modeling the Dark Matter distribution in the universe and modify it with an additional bias parameter to model the galaxy distribution. However, for cells of radius $6h^{-1}$ Mpc, our χ^2 analysis found no clear answer, since the best fit was obtained by the NBD and the Weibull distribution (an original contribution of this work) depending on redshift

and limit magnitude. In a future work we will include an hypothesis test in order to discriminate the best fitting distribution in terms of information criteria.

The second summary statistic used in this work is the projected correlation function (Chapter 3). Using data from the ALHAMBRA survey (Moles et al., 2008, Molino et al., 2014), we have been able to discover new aspects of the behavior of galaxies at scales never observed before with these data quality conditions. Its powerful photometry provides a reliable estimation of the galaxy redshift for objects occupying the inner parts of galaxy clusters and groups. The analysis of these populations by means of the projected correlation function reveals stronger clustering at scales smaller than $r_p \sim 0.2h^{-1}$ Mpc than the expected from the power law trend. This distance indicates a separation region, where physics from the inside of galaxy clusters strongly affects aggregation. Regarding the redshift evolution, we have been able to state that the projected correlation function decreases its amplitude with redshift, breaking the power law trend found at low redshift into the two mentioned regimes at higher redshifts. Using the template classification of the ALHAMBRA survey, we segregate our population into 'star-forming' and 'quiescent' galaxies, reproducing the same calculations. This showed how quiescent galaxies present higher levels of clustering at every scale range, but it is among the star-forming ones that the double slope trend is maintained. This is a new indication of a different clustering behavior depending on galaxy spectral type. The galaxy bias presents similar results, with higher values for quiescent galaxies and increasing with redshift, while star-forming galaxies show no evolution of its lower bias.

As a future work we expect to develop a modification of the correlation function that could make full use of the photometric redshifts. These redshifts measurements, the zPDZ, contain the posterior information of every estimated galaxy redshift. Coming photometric galaxy surveys, such as JPAS (Benitez et al., 2014), will allow us to extract more reliable measurements of the galaxy correlation with this probabilistic approach.

The second and third parts of this research work deepen in the field of point process statistics, experimenting with methodologies applied for the first time in the cosmological scenario that are allowing us to reveal unknown features of the galaxy distribution. We have developed the methods in more complex situations than in previous works, such as three dimensions geometries, or created new functions of

interest to understand the point processes underlying the galaxy structure. These tools have been mainly tested in well known environments, such as highly reliable galaxy surveys or galaxy simulations.

In chapter 4, the first of these new methodologies is the Gibbs models, probability functions that describe the behavior of a point process in terms of the interactions between its elements. Three different models have been tried: Geyer, Fiksel and the Power Law model. These distributions create clustered patterns that can be of interest for our galaxy populations. After applying them over samples from the SDSS-DR8 and the LasDamas simulations, we observed how around 50% of the data content was described by the model. Gibbs modeling describes a point process through the intrinsic behavior of its elements, building the structures as a consequence. In addition, the information contained in a Gibbs model is general and local at the same time. The estimated parameters of these models provide us with a summary of the pattern general properties, while a map of the whole process makes the analysis of peculiar structures possible. In the future we expect to improve this modeling technique with more advanced models and fitting algorithms to obtain more accurate descriptions of the galaxy large-scale structure. The development of an efficient model for the galaxy distribution based on intrinsic properties, such as interaction, will allow us to build realizations of the galaxy point process. This could be used to increase our understanding of the galaxy distribution as well as to complete masked surveys or generate simulations.

The opposite kind of modeling was attempted in chapter 5, where Mixture models used describe broad structures. These models are a combination of different research fields in statistics and can be used for mining our data as well. They gather the *bump hunting* techniques with the multivariate analysis and characterization of spatial structures. Mixture models were designed to simultaneously model different structures sharing a common environment. This alone already implies localizing these structures, but it also implies characterizing the found structures with parametric models. Once a Mixture models is fitted, we have a soft classification algorithm, which assigns probabilities for every particle to belong to a certain structure.

This is of great interest for astrophysics, where, usually, structures and datasets interact and cannot be separated cleanly. If, in addition, we can model these objects while classifying them, the Mixture models can be a tool of high interest.

We tried it with success over a particle sample from the MultiDark simulation and we have been able to identify its content and modeling the structure. Again, modeling allows a localized analysis and separation of overlapped structures where conventional non-parametric techniques are inefficient. We plan to continue this work applying it over real data from nearby universe surveys, where the redshifts quality and completeness are favorable.

Through this thesis work, we have been able to state that point process analysis is an excellent strategy for the analysis of the galaxy distribution. Using different methodologies of summary statistics, modeling and data mining, we start a work conducting to a better description and understanding of the structures formed by galaxies.

Part IV

Appendices

'per açò fem aquests Mil proverbis *ab que donem doctrina com hom se sapia haver en la fi a la qual es creat'*

Ramon Llull

Appendix A

Agregació de galaxies: un procés puntual

Aquest és un resum dels continguts de la present tesi doctoral. Les referències a taules i figures corresponen a les mostrades en el cos del text.

A.1 Introducció

La cosmologia és la ciència que estudia l'origen i la construcció de l'Univers en el seu conjunt. Aquesta disciplina compta amb antecedents històrics remotíssims que evidencien la seua profunda vessant filosòfica. La cosmologia moderna però, situa els seus origens a principis del segle XX, gràcies a diversos avanços tècnics i teòrics que ens han permès estudiar els objectes que constitueixen l'Univers a grans escales.

L'elaboració dels grans cartografiats de galàxies, mapes que contenen les localitzacions espacials i les principals propietats observables dels astres, ens han proporcionat una visió de l'estructura a gran escala, o *Large Scale Structure* en anglès, del contingut de l'Univers contrastable amb els models teòrics. Un dels enfocaments més efectius a l'hora d'estudiar aquestes cartografies és l'estadístic, en particular, l'anàlisi de processos puntuals.

Aquest procediment entén una distribució de punts aïllats en l'espai o el temps, per exemple, galàxies en l'Univers, com la realització d'una variable aleatoria. Poderoses ferramentes s'han desenvolupat en aquest camp, permetent obtindre una descripció completa de molts dels fenòmens que hi tenen lloc a aquestes escales.

A.1.1 El model cosmlògic estàndard

El model cosmològic estàndard, anomenat 'Λ- Matèria Fosca Freda', o Λ-CDM en anglès, és el model que generalment s'accepta en l'actualitat per a entendre el contingut i l'evolució de l'Univers. Aquest model s'ha anat desenvolupant al llarg del segle XX però és apartir del 2000 quan es dóna la versió vigent.

Aquesta teoria estableix que una 'Gran Explosió Calenta' (o *Hot Big Bang*) inicial suposà l'inici de l'Univers que coneguem, alliberant tota la matèria i energia que en forma part. Aquesta teoria es construeix sobre tres punts: el *Principi Cosmològic*, que estableix que la distribució de matèria de l'Univers és uniforme i isòtropa, la teoria de la *Relativitat General*, que descriu la dinàmica gravitatoria, i per últim, l'observada expansió de l'Univers des d'un origen més dens i calent. Aquest model ha resistit tres evidències observacionals que li donen força: la llei de Hubble, que relaciona la distància de les galaxies amb la seua velocitat de recessió (la velocitat amb que s'allunyen les unes de les altres), la detecció del Fons Còsmic de Microones, o CMB en anglès, i la mesura de les abundàncies d'elements lleugers.

Amb la teoria de la Relativitat General d'Einstein i el Principi Cosmològic es pot construir la mètrica de Friedmann-Lemaître-Robertson-Walker (FLRW, equació 1.2), una manera de calcular distàncies en l'espai-temps. Aquesta ferramenta matemàtica ens permet modelitzar l'expansió de l'Univers a través del factor d'escala $a(t)$, que descriu l'expansió de l'espai amb el temps. Açò afecta com percebem les distàncies entre els objectes o com augmenta la longitud d'ona d'un fotó. En aquest últim cas parlem del *redshift*, la reducció de la freqüència d'un fotó rebut en referència a la que tenia en el moment de ser emès.

Quan forcem aquesta geometria a ajustar-se a les observacions trobem un Univers caracteritzat per una topologia quasi plana i amb només un $\sim 5\%$ de matèria bariònica o ordinaria. El $\sim 95\%$ restant del contingut de l'Univers està compost per les components fosques. Un $\sim 70\%$ correspon a l''energia fosca', una espècie

d'energia del buit que empeny l'Univers a expandirse acceleradament. L'altre $\sim 25\%$ és 'matèria fosca', aquesta component és necessaria per tal d'entendre el comportament dels astres en escales superiors a les d'una galaxia, ja que sense ella entrariem en conflicte amb les lleis de la Relativitat General.

A.1.2 La distribució de galàxies com a procés puntual

Una simple anàlisi ocular de la distribució de les galàxies en l'espai revela com aquesta està lluny de ser uniforme fins i tot a escales de l'ordre dels 150 h^{-1} Mpc, quan comença a ser verificable el Principi Cosmològic. A escales menors la gravetat actua com a principal motor de la dinàmica de les galàxies, que les agrupa en estructures o pertorbacions. És en aquest context que fem servir l'estadística de processos puntuals per a descriure els patrons formats.

La teoria de processos puntuals estableix que coneguem tota la informació continguda en un procés puntual quan som capaços de predir el nombre de punts que ocupen qualsevol subconjunt obert de l'espai on habiten. No obstant això, assolir aquest grau de coneixement sobre una població pot ser quasi impossible, de manera que s'utilitzen multitud d'estadístics que incrementen la nostra comprehensió de les dades i els seus patrons des de diversos angles. Considerem tres possibles aproximacions o estratègies. L'estadística de resums descriu amb generalitat la distribució formada per les galàxies, donant-nos informació corresponent a trets de conjunt, atribuibles a tota la població i no a un subconjunt dels punts. La mineria de dades ens permet aïllar i identificar membres de la mostra poblacional que satisfan unes determinades propietats, sovint una combinació de la seua distribució geomètrica o altres trets. I el modelatge, la definició d'un model que efectivament sintetitza i justifica la localització de cada galàxia mitjançant l'ús de funcions de probabilitat.

A.1.3 Objectius d'aquesta tesi

En aquesta tesi ens proposem aplicar diverses tècniques d'anàlisi de processos puntuals sobre la distribució de galàxies. Farem servir mètodes corresponents a totes tres aproximacions, alguns dels quals desenvolupats per a ser aplicats per primera vegada en cosmologia.

Les oportunitats que ofereixen aquestos anàlisis són abundants. La parametrització d'aquestes quantitats ens ajuda a comparar els resultats dels estimadors que fem servir, dibuixant universos preferibles. Aquestos resultats poden tenir implicacions de pes en la construcció de simulacions de distribució de matèria o ajudar-nos a realitzar prediccions sobre les estructures que trobarem amb les noves generacions de cartografiats de galàxies. Entendre correctament i poder manipular expresions que descriuen la distribució de galàxies és fonamental per a descobrir allò que encara no hem vist. Tots els anàlisis realitzats en aquesta tesi s'han efectuat amb metodologies adaptades a tres dimensions espacials.

En els capítols 2 a 5 presentem quatre treballs integrats en els enfocaments estadístics abans descrits. Cada capítol conté les metodologies i dades que fem servir en la nostra recerca. En el capítol 6 presentem les principals conclusions.

En el treball sobre **recomptes en cel·les** sobre dades del catàleg NYU-VAGC (capitol 2) obtenim una descripció de la freqüència de galàxies per volum, un descriptor bàsic però central en la teoria de processos puntuals. Tot i que el càlcul *per se* d'aquest estadístic ja és informatiu, es poden obtindre més conclusions a través de l'ajust de diverses distribucions de probabilitat a les freqüències observades. Els parametres obtinguts constitueixen indicadors físics de la naturalesa del procés, i l'obtenció d'un ajust de qualitat ens pot permetre generar simulacions de galaxies més fiables en el futur.

Una de les tècniques més emprades i més efectives en la cosmologia d'estructures a gran escala és la **funció de correlació**, amb les seues diverses variants. En la secció 3.1 farem servir la funció de correlació projectada, concebuda per a treballar sobre dades de catàlegs de galàxies on trobem problemes de distorsions de *redshift*. Amb les dades del catàleg ALHAMBRA però, tenim l'oportunitat d'estudiar l'agregació de les galàxies a escales molt menors a les estudiades amb anterioritat, obtenint una clara imatge del seu comportament a distàncies de separació inferiors als $\sim 0.2h^{-1}$ Mpc, quan les galaxies interaccionen físicament i no només gravitacionalment. A més, la fotometria d'ALHAMBRA ens permet fer aquest anàlisi amb segreació espectral, analitzant les diferencies entre galàxies amb alta i baixa formació estelar.

En la segona meitat de la tesi iniciarem l'estudi de nous mètodes per a la cosmologia, amb la intenció d'anar més enllà dels estadístics de resum presentats fins

ara. Són models complets de la distribució de galàxies, que permeten descriure al complet una població. Els **processos de Gibbs** descriuen la distribució de punts a partir d'una funció d'intensitat i una funció d'interacció entre punts. Aquesta funció d'interacció descriu el comportament d'un punt o galàxia donat el seu entorn, cohesionant tota la població sota un mateix model. En el nostre cas, aquesta interacció és la gravitatòria, de tipus atractiu, cosa que hauria de quedar reflectida en l'ajust paramètric.

Finalment, amb l'ús dels **models de mescla**, farem servir un mètode que uneix el modelatge i la mineria de dades. Aquestos models conceben la distribució de les galàxies com un conjunt de components amb identitat pròpia però amb interacció. Aquestos components corresponen a modelatges previs d'estructures de galàxies, tals com els cúmuls de galàxies. Amb l'ajust adequat d'aquestos models, obtenim una classificació de les galàxies en els seus cúmuls, que caracteritzem paramètricament. Açò ens obrirà les portes a abundants aplicacions.

Amb aquestos quatre treballs demostrem que l'anàlisi de la distribució de galàxies com a procés puntual no només és possible sinó molt efectiu, proporcionant-nos abundant informació de rellevància.

A.2 Ajust de recomptes de galàxies per cel·les

El recompte de cel·les consisteix a contar el nombre de galàxies que es troben a l'interior d'una cel·la definida i situada per l'usuari. Les dades que farem servir en el nostre anàlisi corresponen al NYU-VAGC (Blanton et al. (2005), veure secció 2.3.1), un catàleg que unifica observacions del *Sloan Digital Sky Survey* (SDSS) i el *2 Micron All Sky Survey* (2MASS). Aquest catàleg proporciona observacions molt fiables del *redshift* que permeten una bona estimació de la distància a que es troba cada galàxia. Amb aquesta informació podem construir una representació en tres dimensions de la distribució puntual de galàxies.

Les dos poblacions sel·leccionades cobreixen diferents rangs de *redshift*. La primera inclou 113483 galaxies amb *redshifts* en $0.05 < z < 0.106$ i magnitud absolutes $M_r < -20$. La segona, 76688 galaxies amb $0.075 < z < 0.165$ i $M_r < -21$.

Les cel·les que farem servir seran esferes de radis 6, 12 i 24 h^{-1} Mpc distribuides uniformement sobre el mateix volum que ocupa el catàleg. El recompte es realitza

senzillament calculant quantes galàxies estan a una distància del centre de la cel·la menor que r.

Per a realitzar aquest anàlisi correctamente cal tenir en compte diversos aspectes. Els cartografiats sovint venen acompanyats d'una finestra, una regió del cel definida geomè- tricament on s'inclouen aquelles regions que han pogut ser observades eficaçment, sense interferències d'estrelles brillants o problemes instrumentals, de manera que desconeguem allò que s'hi troba a fora. Quan contem el nombre de galàxies situades a l'interior d'una cel·la hem de tenir en compte el seu volum efectiu dins la màscara, i modificar adecuadament el nostre resultat. Acceptarem només aquelles cel·les que hagen perdut per efectes de finestra un màxim del 5% del seu volum, que serà compensat multiplicant el nombre de galàxies que conté per la inversa d'aquest volum faltant.

Una vegada corregit el recompte de galàxies pels volums efectius de les cel·les obtenim la funció de densitat de probabilitat del recompte $f_V(N)$ normalitzant l'histograma de freqüències.

A.2.1 Funcions de distribució

Amb els ajustos de les funcions $f_V(N)$ obtenim una caracterització de la distribució del recompte de cel·les que ens ajuda a entendre el seu comportament. L'ajust es realitzarà amb la χ^2, minimitzant l'expressió $\sum_{i=1}^{N}(f_V(N)-f_\theta(N))^2/\sigma^2$ sobre els paràmetres θ. Les nostres funcions de distribució són:

Distribució Gravitacional de Quasi-Equilibri (GQED)

Aquesta distribució, proposada per Saslaw & Hamilton (1984) és una descripció termodinàmica del fluid de galàxies. Els seus paràmetres lliures són $\bar{N} = \bar{n}V$ on $\bar{n}$ és la densitat de galàxies de la mostra i V és el volum de la cel·la, i b, un paràmetre que descriu el nivell d'agregament de la mostra. La seua funció de densitat de probabilitat és

$$f_V(N) = \frac{\bar{N}(1-b)}{N!}[\bar{N}(1-b)+Nb]^{N-1}e^{-[\bar{N}(1-b)+Nb]} \qquad \text{(A.1)}$$

Binomial Negativa

Aquesta funció, proposada per Carruthers & Duong-van (1983), s'utilitza comunment en estadística per a descriure la distribució de punts en caixes. És per tant un candidat natural per a descriure un recompte per cel·les, tot i que s'ha demostrat que no és consistent amb la física de les galàxies. De manera similar a l'anterior, els paràmetres són $\bar{N}$ i g, que fa un paper similar a b.

$$f_V(N) = \frac{\Gamma(N + \frac{1}{g})}{\Gamma(\frac{1}{g})N!} \frac{\bar{N}^N (\frac{1}{g})^{\frac{1}{g}}}{(\bar{N} + \frac{1}{g})^{N+\frac{1}{g}}} \tag{A.2}$$

Log Normal amb biaix

La distribució Log Normal pot utilitzar-se amb éxit per a descriure la distribució no linial de les fluctuacions de densitat de la matèria fosca (Arnalte-Mur et al., 2016). Per tal de descriure amb éxit la distribució de galàxies caldrà introduir el biaix b entre aquestes dos distribucions de matèria. Normalitzant la distribució sobre l'esperança del recompte de cel·les $\bar{N}$ obtenim una distribució amb dos paràmetres, la variància de la distribució de matèria C i el biaix b.

$$f_V(\Delta) = \frac{1}{\sqrt{2\pi H_0}} \frac{\exp\left(-\frac{1}{2}\frac{y^2}{H_0}\right)}{\Delta + b - 1} \tag{A.3}$$

on $\Delta = N/\bar{N}$, $H_0 = \ln(1 + C)$ i

$$y = \ln\left(\frac{\sqrt{1+C}}{b}(\Delta + b - 1)\right) \tag{A.4}$$

Weibull

La distribució Weibull (Weibull, 1951) descriu fenòmens tals com l'esperança de processos que acumulen probabilitats de mort creixent al llarg de la seua vida útil. També s'ha utilitzat amb éxit per a distribuir el tamany de partícules formades per agregació fins a completar el seu tamany final. Nosaltres la farem servir per primer cop en el context del recompte de cel·les. Els seus paràmetres són l'escala

λ, relacionat amb el nombre de galàxies esperat per cel·la, i la pendent k, que determina la forma de la distribució.

$$f_V(N) = \frac{k}{\lambda}\Big(\frac{N-\theta}{\lambda}\Big)^{k-1} e^{-(\frac{N-\theta}{\lambda})^k} \tag{A.5}$$

A.2.2 Tractaments del errors

Per tal de contrastar la qualitat del nostre ajust és necessari disposar d'una estimació de l'error esperable en les nostres observacions. La variància còsmica que podem trobar en observar diferents regions del cel corresṕon una font d'incertessa dins la qual s'han de donar les divergències acceptables entre el nostre model i les observacions.

Degut a la presència de la finestra farem servir simulacions adaptades a les nostres dades, com és el cas dels catàlegs *LasDamas* (McBride et al., 2011). Aquestes són simulacions de la distribució real de les galàxies però generades computacionalment respectant les condicions cosmològiques del cartografiat. Imiten les propietats de les dades del catàleg SDSS, amb l'avantatge que no pateixen els efectes d'una finestra.

Les variancies trobades en el recompte de cel·les de les realitzacions de *LasDamas* ens serviran com a estimació dels errors en la funció de densitat observacional $f_V(N)$.

A.2.3 Resultats i conclusions

Després de calcular les distribucions del recompte de cel·les sobre les dos poblacions de galàxies amb els tres radis hem procedit a ajustar les funcions anteriors. Aquest és un ajust cec que busca el millor ajust per a la distribució observacional. Un model empíric que ens done una manera senzilla de descriure la distribució de galàxies.

El resultats poden consultar-se a la secció 2.6, on mostrem les funcions per al seu millor ajust i el residu amb la distribució observada.

L'ajust amb χ^2 obté els millors resultats per a la distribució Log Normal amb el biaix, tret del cas de les cel·les de radi 6. Aquest resultat reforça l'interés d'aquesta distribució per al context cosmòlogic, ja que incorpora la física pertorbativa del camp de densitat de la matèria fosca i el biaix amb la distribució de galàxies.

A.3 Correlació de galàxies a escales curtes i segregació espectral

El catàleg ALHAMBRA (*Advanced Large Homogeneous Area Medium-Band Redshift Astronomical survey*, en anglès) (Moles et al., 2008, Molino et al., 2014) és un cartografiat fotomètric que fa servir 20 filtres de banda estreta més 3 filtres de banda ampla en l'infrarroig. Aquesta abundant informació ens permet obtenir imatges de tot el cel alhora que estimar amb alta fidelitat el *redshift* de cada font detectada. Aquest és un avantatge considerable sobre els catàlegs espectroscòpics, on l'estimació del *redshift* d'objectes angularment propers està limitada. Açò suposa que ALHAMBRA és un catàleg d'interés per a l'anàlisi del comportament de les galàxies a l'interior dels cúmuls.

A més, deduida d'aquesta mateixa fotometria, podem classificar les nostres galàxies depenent de diverses propietats físiques, i estudiar com canvien les seues distribucions. Quan diferents tipus de punts conviuen en una mateixa població és interessant analitzar-los des l'òptica dels processos puntuals amb marques. Aquestes marques són propietats no espacials dels punts que ens permeten segregar els punts en aquelles categories que segueixen un mateix patró o model. És conegut que les galàxies tendeixen a agregar-se de diferents maneres depenent de diverses propietats, tal i com són la massa, la lluminositat o el color. Nosaltres, però, analitzarem les diferencies que s'observen entre galaxies de diferents tipus espectrals, una característica que ve marcada principalment per la seua composició i els seus nivells de formació estelar.

A.3.1 El catàleg ALHAMBRA

El projecte ALHAMBRA ha produit imatges d'una secció de 2.381graus^2 del cel. Aquest cartografiat conté de l'ordre de 400000 galàxies, amb una precisió de *redshift* superior a $\sigma_z/(1+z) = 0.014$. Amb aquestes dades podem construir una sel·lecció de galàxies que ens permetrà realitzar els anàlisis d'interès. En primer lloc realitzem un tall en magnitud $I < 24$ per a garantir la completitud fotomètrica (I representa al filtre F814W del Hubble Space Telescope) i eliminem els estels. Una finestra ha sigut especialment dissenyada (Arnalte-Mur et al., 2014) per a exclure objectes dubtosos. Amb les dades resultants construim 5 subconjunts, cadascún corresponent a una banda disjunta de grossor 0.15 en *redshift* des de 0.35 fins a 1.1. Per tal d'assegurar que les 5 mostres són comparables definim un tall en magnitud superior a $M_B = -19.36$ per al fragment més llunyà i calculem els talls necessaris en la resta per a igualar la densitat d'objectes. Les dades d'ALHAMBRA contenen també una classificació dels objectes en típus espectrals, incloent galàxies el·líptiques, espirals o irregulars. Aquest paràmetre ens permet separar-les en galàxies amb baixa formació estelar i galàxies amb alta formació estelar. La classificació final d'objectes pot consultar-se a la Taula 3.1.

A.3.2 La funció de correlació projectada

La funció de correlació de dos punts descriu l'excés o el defecte de correlació entre les distàncies relatives dels punts. Quan els punts formen grups o cúmuls amb una densitat superior a la mitjana diem que estan agregats i les seues distàncies relatives tendeixen a ser menudes. Açò provoca que, comparat amb una població distribuida uniformement, aquestes distàncies menudes siguen molt més abundants. La funció de correlació de dos punts en processos isòtrops pot definir-se implícitament com la funció $\xi(r)$ en l'equació 3.4.

Malauradamente, degut a les distorsions del *redshift* i a possibles inexactituds en la seua estimació no podem garantir la isotropia i ens veiem forçats a introduir una projecció. Açò consisteix en descompondre la distància de separació entre galàxies en una component paral·lela ($r_{||}$) i un altra perpendicular ($r_\perp$) a la línia de visió de l'observador. Després d'estimar $\xi(r_{||}, r_\perp)$ segons l'equació 3.13 podem integrar sobre $r_{||}$ per tal d'obtenir la nostra funció de correlació projectada $w(r_\perp)$ (veure equació 3.15).

El tractament dels errors d'aquesta funció es farà seguint el procediment *Jackknife*, on generem noves realitzacions de la nostra població tot eliminant per torns regions disjuntes de la població original. Aquestes noves mostres s'utilitzaran per a estimar la variància de les mesures.

A.3.3 Resultats

Els resultats del càlcul de la funció $w(r_\perp)$ sobre les nostres dades ens mostren una profunda correlació, especialment a escales curtes ($r_\perp < 0.2h^{-1}$ Mpc), on la pendent de la corba canvia bruscament, sobretot per a la població de major *redshift*. Aquest efecte es pot comprovar adequadament mitjançant un anàlisi de l'ajust de dues lleis de potència $w(r_\perp) = Ar_\perp^b$ per a cada tram de la corba. Després d'estimar el millor ajust per als paràmetres A i b comprovem com les regions de confiança 1σ són clarament independents per al trams de *redshift* superiors però convergeixen als inferiors. Açò indica una evolució en la distribució de les galàxies, des de dos règims d'agregament diferenciats cap a un amb propietats comunes (veure Fig. 3.12).

És interessant estudiar aquesta dualitat en l'agregament a través de la segregació de les nostres poblacions (Fig. 3.13). El primer tret que s'observa és la pronunciada separació de les corbes. Les galàxies amb baixa formació estelar, més velles i passives, tendeixen a ocupar regions més pròximes al centre dels halos, on tenim un major nivell de correlació. Les galàxies amb alta formació estelar, més joves i actives, són dominants en amples regions de baixa densitat. Així i tot, són aquestes galàxies, als grups de galàxies blaves, les que contribueixen més intensament a crear la dualitat d'agregació amb l'escala.

Quan ha sigut possible, hem acompanyat aquestos resultats dels valors obtesos per altres autors, comparant-los amb els nostres i obtenint una equivalència general. Açò ha sigut possible només per a escales superiors a $0.1h^{-1}$ Mpc, ja que no hi ha a la bibliografia treballs comparables als nostres resultats en escales curtes.

Dependència del biaix

Un resultat interresant que es pot obtenir a partir de les corbes de la funció de correlació projectada és la dependència del biaix entre matèria fosca i bariònica.

S'ha comprovat una desviació entre la distribució teòrica de la matèria fosca i la distribució observada de les galàxies. Aquest biaix b es defineix com

$$b = \sqrt{\frac{w_p(r_\perp)}{w_p^m(r_\perp)}} \tag{A.6}$$

El nostres models actuals de la distribució de matèria fosca estan basats en les mesures de les constants cosmològiques de l'experiment WMAP7 (Komatsu et al., 2011), i els podem generar per a diferents valors del *redshift* amb el codi `CAMB`. Aquestos models però, només són fiables per a distàncies superiors a $1h^{-1}$ Mpc, així que ajustarem b en el rang $1.0 < r_\perp < 10.0h^{-1}$ Mpc. Aquestes quantitats es mostren en la Fig. 3.16, tant per a galàxies fredes com per a galàxies amb alta formació estelar.

El resultats mostren una clara evolució per al biaix de les galàxies fredes, creixent amb el *redshift* de manera similar a com ho farien halos de massa $M_h \sim 10^{12.5}h^{-1}M_\odot$. En canvi, per a les galàxies amb alta formació estelar, és quasi constant, el que enllaça amb l'evolució observada de r_0.

Les escales més petites

Degut a l'interès que suposa la posibilitat d'estudiar les correlacions de galàxies a escales molt petites, hem comparat ALHAMBRA amb altres cartografiats de galàxies similars. Els projectes COSMOS (Ilbert et al., 2008), DEEP2 (Newman et al., 2013) i PRIMUS (Coil et al., 2011) són cartografiats fotomètrics que cobreixen diverses regions del cel. Els catàlegs produits són capaços d'estimar el *redshift* d'una gran quantitat d'objectes sense els problemes de l'espectrografia, que no permeten observar amb presició objectes molt propers angularment. A aquesta profunditat d'imatge cal afegir la presició amb que mesurem la distància a les galàxies, on ALHAMBRA ha demostrat ser altament satisfactori.

Tot plegat, comptem amb uns catàlegs que ens permeten estudiar com mai abans la distribució de les galàxies en aquelles regions on es troben més aprop les unes de les altres, és a dir, l'interior dels cúmuls de galàxies. A la Fig. 3.9 trobem una comparació de la distribució del veí més pròxim per als esmentats catàlegs. La conclusió que hi podem obtenir és que ALHAMBRA i COSMOS contenen dades

especialment útils per a l'estudi de les correlacions a curta escala. Per al cas d'ALHAMBRA, ho hem reproduit a les Figs. 3.11 i 3.13, on es recullen els valors de la funció de correlació per a distàncies de separació inferiors als 200 kpc.

A.4 Modelització amb processos de Gibbs

Els models d'interaccions de partícules són unes ferramentes estadístiques que comencen a aplicar-se per primer cop en la investigació en cosmologia. Aquestos models, anomenats processos finits de Gibbs, descriuen un procés puntual com una funció de probabilitat on la posició d'un element dependrà d'una funció de densitat, que determina l'abundància de punts en una localització de l'espai, i d'una funció d'interaccions, que determina la probabilitat de trobar un punt a partir de la preséncia dels punts veïns.

Aquesta propietat de determinar la preséncia d'un punt en funció de la distribució dels voltants, i només d'allò que hi trobem dins d'una determinada àrea, s'anomena de *Markov*, i fa referència a una dependència limitada en l'espai en aquells events que s'han produit dins d'una regió finita, com per exemple una esfera de radi r centrada en la nostra localització d'interès.

A.4.1 Models d'interacció

Tot plegat, la funció de densitat de probabilitat d'un model finit de Gibbs pren la forma de l'equació 4.2

$$f(x_1, ..., x_n) = \alpha \prod_{i=1}^{n} b(x_i) \prod_{i<j} h(x_i, x_j) \tag{A.7}$$

on $b(x)$ és la funció de densitat i $h(x_i, x_j)$ és la funció d'interacció, que determina la probabilitat que se'n deriva de la parella de punts x_i, x_j. En aquest treball utilitzarem distribucions que determinen les seues probabilitats a partir de la distància relativa entre dos punts $d = ||x_i - x_j||$. Aquesta funció h permet classificar els processos en tres grans tipus. En primer lloc, els processos de Poisson, on $h(u, v) = 1$ per a tota parella u, v. Aquest procés no presenta interaccions de cap tipus, i

els seus punts són independents. Donat que només treballarem amb processos de Markov, assumirem que $h(u, v) = 1$ sempre que la distància entre els dos punts siga major a un radi de correlació. En segon lloc, quan $h(u, v) > 1$ trobem que el procés presenta una agregació de punts, i aquestos tendeixen a estar més aprop els uns dels altres que en el cas Poisson. I en tercer lloc, quan $h(u, v) < 1$ trobem l'efecte contrari, on els punts tendeixen a guardar una distància mínima de separació entre ells, el que es coneix com a patró regular.

Per tal de modelitzar la distribució de galàxies hem assumit una funció de densitat $b(x)$ constant que no privilegie cap regió de l'espai i fins a tres funcions d'interacció diferents:

Per a una parella de punts u i v, el model de **Geyer** es defineix amb la funció

$$h_\gamma(u, v) = \begin{cases} \gamma, & \text{if } ||u - v|| < r \\ 1, & \text{otherwise} \end{cases}$$

on $\gamma > 1$ és el paràmetre d'agregació. Per tal de garantir la integrabilitat d'aquest model cal imposar un límit superior *sat* al nombre de parelles d'un punt determinat $s(x|X)$. El valor de la funció d'interacció per a u en el procés X serà $\gamma^{min(sat, s(u|X))}$.

El model de **Fiksel** és un model continu que es regeix amb la funció

$$h_{a,\kappa}(u, v) = \begin{cases} 0, & \text{si } d < r_0 \\ \exp\left(a \cdot \exp\left(-\kappa \cdot d\right)\right), & \text{si } r_0 < d < r_1 \\ 1, & \text{si } r_1 < d \end{cases}$$

Aquest model compta amb dos distàncies d'interacció: r_0 elimina la posibilitat de formar parelles de punts per sota d'aquesta distància i permet la integrabilitat del model, r_1 funciona com a distància de Markov. Per a distàncies intermitges, una funció exponencial pren valors més alts quan més pròximes són les parelles de punts. Els paràmetres a i κ determinen la forma del perfil d'interacció.

Finalment, el model de **Llei de potència** funciona de manera similar al model de Fiksel però amb un perfil potencial enlloc d'exponencial.

$$h_{a,\gamma}(u,v) = \begin{cases} 0, & \text{si } d < r_0 \\ \exp\left(a \cdot d^{-b}\right), & \text{si } r_0 < d < r_1 \\ 1, & \text{si } r_1 < d \end{cases}$$

amb r_0 i r_1 com abans i a i b els paràmetres lliures.

Per treballar amb aquestos models però, no farem servir l'expressió A.7 de la funció de probabilitat, sinó que farem servir la funció condicional de *Papangelou*, que es defineix com $\lambda(u, X) = f(\{u\} \cup X)/f(X)$, és a dir, la probabilitat condicional de trobar un punt en u donat X. Aquesta expressió simplifica notablement les funcions anteriors i elimina la necessitat de calcular la constant de normalització α. L'expressió resultant d'aquesta probabilitat condicional pot resumir-se en

$$\lambda(u, X) = b(u) \prod_{i=1}^{n} c(u, x_i) \tag{A.8}$$

on $c(u, x_i)$ equival a la funció d'interacció evaluada sobre el punt u respecte la resta de punts d'X.

L'ajust d'aquestes funcions és una tasca complicada que involucra una gran quantitat d'evaluacions dels estadístics corresponents a cada model. Malauradament, la constant de normalització α és generalment intractable, i la impossibilitat de calcular-la eficaçment ens impossibilita ajustar els nostres models mitjançant la funció de màxima versemblança. L'alternativa que farem servir és una funció coneguda com *pseudolikelihood* (Baddeley & Turner, 2000b, Besag, 1975), una aproximació de la funció de versemblança. Aquesta funció es pot calcular, en la seua form logarítmica, com

$$logPL(\theta; X) \approx \sum_{j=1}^{m} (y_j log\lambda_\theta - \lambda_\theta) \cdot w_j \tag{A.9}$$

on λ es evaluada sobre tot punt de W, tant punts de les dades com localitzacions de l'espai. Els pesos w_j corresponen a cadascún d'aquestos punts i definim $y_j = z_j/w_j$ on $z_j = 1$ si correspon a un punt de les dades o 0 en cas contrari.

Per a l'anàlisi de la qualitat dels ajustos tampoc comptem a un mètode convencional i suficient, sinó que farem servir tota una batèria de resultats que ens ajuden a entendre, sobretot mitjançant visualitzacions dels resultats, la bondat de l'ajust. D'entre aquestes tècniques destaquem el càlcul dels residus, una estimació de la diferència entre el nombre de punts en una regió del volum i la seua estimació amb el model. Quan aquest procediment es realitza sobre regions molt petites és convenient convolucionar el resultat per un filtre que suavitze la funció resultant:

$$s(u) = \sum_{i=1}^{N} \kappa_\omega(u - x_i) - \int_W \kappa_\omega(u - v)\hat{\lambda}_\theta(v)\mathrm{d}v \tag{A.10}$$

on u és una localització de W. A partir d'aquesta funció poden obtenir-se diversos test que ens mostren la qualitat de l'ajust des de diverses òptiques.

A.4.2 Dades i ajust

Per a comprovar les capacitats dels nostres models i les diverses metodologies que acompanyen, els aplicarem sobre dos conjunts de dades diferents. Un serà el catàleg SDSS que ja hem introduit anteriorment però sobre una versió més recent, el DR8 (Tempel et al., 2012) . L'altre correspon de nou a les simulacions LasDamas, que ens serviran, com abans, per a estudiar possibles efectes sistemàtics entre totes dues mostres.

Les dades sel·leccionades corresponen a quatre cubs de $50h^{-1}$ Mpc de costat, amb *redshifts* entre 0.02 i 0.085. D'aquesta manera ens assegurem que les dades siguen comparables.

L'ajust d'aquestes poblacions està descrit amb detall a les seccions 4.4.5 i 4.4.4. A grans trets, el que trobem és una caracterització local de l'agregació de punts. Açò es pot entendre com un camp de densitat paramètric, que no només ens proporciona informació detallada de cada punt sinó que ens permet estudiar les propietats d'aquestes poblacions mitjançant els parametres ajustats.

Els resulats indiquen, tal i com esperàvem, valors paramètrics corresponents a processos d'agregació, amb $\gamma \sim 1.4$ per al model de Geyer. Els parametres ajustats de tots tres models mostren certa volatilitat entre les diverses poblacions, una

variació que requereix un anàlisi més detallat per tal de comprovar la possible correlació d'aquestos valors amb les estructures de les mostres.

L'anàlisi dels residus també resulta molt efectiu, proporcionant-nos una informació detallada que ens permet identificar ràpidament aquelles estructures de les dades que destaquen com sobredensitats de punts o buits. En particular, per a les poblacions del catèleg SDSS, comprovem com la segona mostra presenta una major variància en la distribució de galàxies per als tres models, amb una forta agregació de punts molt localitzada. Aquest resultat suposa també una divergència entre SDSS i LasDamas.

A.5 Models de mescla

Els models de mescla (o *Mixture models* en anglès) són una ferramenta estadística utilitzada per primer cop per Karl Pearson l'any 1890. Amb aquest mètode es pretén obtenir una distribució paramètrica conjunta d'una població composta per diverses components no distribuides idènticament. D'aquesta manera ajustem cadascuna de les components amb diverses funcions de probabilitat condicionades, el qual ens permet aïllar i obtenir un model per a la distribució de cadascuna de les components.

Aquesta tècnica l'hem aplicada sobre la distribució de matèria, com a part d'un estudi sobre modelització de la distribució de galàxies. Aquest treball ens permet desenvolupar una nova metodologia per a l'extracció de mostres d'interès en una població de galàxies (tècniques de *data mining*). Fent servir dades de la simulació *MultiDark* (Klypin et al., 2011) hem ajustat el perfil d'Einasto per a la matèria fosca sobre diversos cúmuls.

A.5.1 El model de densitat de superfície

En primer lloc hem definit un model sobre la densitat de partícules en la mostra. Aquest model compta amb un perfil d'Einasto ρ per cada cúmul en la població més una component plana ρ_c per a aquelles partícules que no pertanyen a cap cúmul.

$$\Sigma(\mathbf{r};\alpha,\theta) = \sum_{i=1}^{c-1} \alpha_i \cdot \rho(\mathbf{r}-\mathbf{r_0};\mathbf{s}) + \alpha_c \cdot \rho_c(\mathbf{r}) \tag{A.11}$$

Els coeficients α_i són parametres lliures que ens permeten normalitzar la funció. Donat que cada component pot tenir una constant de normalització diferent cal ajustar tants coeficients com components té el model. Com hem dit abans, aquest model ens permet extraure cada cúmul amb la seua funció de densitat $\alpha_i \cdot \rho(\mathbf{r}-\mathbf{r_0};\mathbf{s})$ encara que en l'espai físic estiga contaminat per la presència d'altres cúmuls i no siga fàcil la seua identificació.

Pel que fa a l'ajust del model, en aquesta ocasió el càlcul numèric de la constant de normalització conjunta permet l'ajust dels paràmetres mitjançant la maximització de la funció de versemblança. Hem fet servir un procediment de MCMC amb *priors* quasiplans, el que ens proporciona no només el millor ajust per als nombrosos paràmetres sinó també una estimació del seu error. La idoneïtat del model serà testada mitjançant criteris informatius com el *Bayesian* i el *Akaike information criterium*, que tenen en compte no només el valor de la likelihood sinó també el nombre de paràmetres (i components) fets servir.

A.5.2 Aplicació i resultats

Abans de treballar sobre dades de la simulació *MultiDark* hem realitzat una serie de proves amb models de joguet autogenerats (secció 5.4.2). Aquestos experiments ens permeten obtenir una idea més completa del comportament esperable del model en diverses situacions i detectar així posibles errors.

Sobre una mostra amb quatre cúmuls generats per reproduir un perfil conegut d'Einasto (Einasto, 1965, 1968, 1969), realitzem l'ajust del model de mescla. Açò ens permet obtenir una primera impressió de la qualitat dels valors recuperats. Posteriorment, afegim nous casos, tals com ajustar un model amb un nombre incorrecte de components o ajustar estructures que no corresponen a un perfil de cúmul de matèria fosca, com per exemple, un filament. Aquestes són situacions reals amb les que caldrà enfrontar-se quan tractem dades de simulacions o catàlegs de galàxies.

L'anàlisi de la qualitat dels ajustos la realitzem fent servir la mateixa metodologia que hem presentat a la secció anterior per als models de Gibbs, en particular els residus de l'ajust.

Finalment, optem per una mostra sel·leccionada de les dades de *MultiDark* (Klypin et al., 2011), un ortoedre que conté 2081 partícules distribuides en diversos cúmuls de diferent tamany. El primer càlcul necessari per a l'ajust del model és el nombre de components que conté la mostra. Aquesta és una elecció delicada que ha de ser contrastada amb posterioritat a l'ajust del model mitjançant els criteris informatius. Una bona aproximació a aquest problema consisteix a calcular el camp de densitat de la població i prendre els màxims locals del camp com a candidats per a centres de cada component (parametres $\mathbf{r_0}$). D'aquesta manera podem iniciar el càlcul MCMC amb una estimació del nombre de cúmuls i de la meitat dels paràmetres.

Els càlculs amb MCMC poden ser costosos, especialment per a funcions amb un alt nombre de paràmetres lliures. Però és un cost necessari per a obtenir un ajust robust. Aquestos resultats poden consultar-se a la Taula 5.3, on podem comprovar que els parametres obtinguts no només reprodueixen fidelment la distribució de les partícules, sinó que són valors esperables per a aquesta classe d'estructures.

L'anàlisi de residus mostra un ajust satisfactori que aconsegueix reproduir fins al 75% de la densitat de superfície de les dades. A més, aquesta tècnica és especialment últil per tal de descobrir la necessitat d'afegir noves components al nostre model. Inicialment hem realitzat un ajust amb 6 components per raons de sentit físic, però posteriorment, després de l'anàlisi dels residus, optem per un ajust amb 10 components. Aquest exercici ens permet comprovar com certes estructures de menor tamany degeneren els paràmetres del perfil d'Einasto cap a una funció més similar a una llei de potències. En aquestes situacions cal optar per un criteri físic més realista i assumir el model amb 6 components.

Dos resultats adicionals poden obtindre's de l'ajust realitzat. Un és la reproducció dels perfils dels cúmuls més enllà de les distàncies en que comencen a interaccionar fortament amb altres estructures. Amb aquesta tècnica aconseguim reproduir els perfils de densitat de cada cúmul en rangs de distàncies intractables mitjançant

altres mètodes. I en segon lloc, una vegada conegut el nostre model de mescla podem reproduir fàcilment realitzacions d'aquest procés puntual i generar poblacions independents i idènticament distribuides a la original.

A.5.3 Conclusions

Els models de mescla ens permeten descriure molt eficaçment una població de punts (ja siguen partícules de matèria fosca o galàxies) amb el detall necessari per a deduir nombroses propietats. El coneixement de la distribució conjunta i separable de les diverses components d'una població pot tenir aplicacions en astrofísica més enllà de l'estudi dels cúmuls de galàxies, com per exemple les interaccions de galàxies amb els seus satèl·lits.

Per a un treball futur esperem poder aplicar aquesta tècnica sobre dades de l'univers pròxim, mostres amb major densitat d'objectes i estimacions fiables de la distància de cada galàxia.

A.6 Conclusions

Aquesta tesi recull diversos treballs realitzats en el marc de l'anàlisi de processos puntuals, cobrint les tres principals aproximacions de l'estudi d'aquestos objectes: l'estadística de resums, el modelatge i la mineria de dades. Hem demostrat com les diverses tècniques que s'enmarquen en aquestes categories són òptimes per a l'estudi de la distribució de galàxies, obtenint resultats que suposen un avanç tant en la comprensió de les propietats de la distribució de galàxies com en el desenvolupament de noves tècniques especialment adaptades al context cosmològic.

D'aquestos resultats destaquem l'ajust d'una nova funció de distibució, la distribució Log Normal amb biaix, que permet descriure fidelment la distribució del recompte per cel·les d'una mostra de galàxies. També com a exemple de l'ús d'estadístics de resum, amb la funció de correlació de dos punts, i fent servir l'excel·lent fotometria d'ALHAMBRA, hem aconseguit obtenir les primeres mesures d'aquesta funció per a distàncies molt inferiors als $0.2h^{-1}$ Mpc, mesures que descriuen el comportament de l'interior dels cúmuls de galàxies.

Amb els models de Gibbs hem modelat la distribució de les galàxies alhora que la caracteritzem amb paràmetres que descriuen les interaccions. El desenvolupament de mètodes més eficaços per a modelar una població de galàxies ens proporciona no només una caracterització global sinó informació corresponent a cada punt del volum.

I per últim, els models de mescla ens han permès extraure subconjunts de dades on altres mètodes poden no ser suficients per a separar diverses estructures que interaccionen en una mostra. Amb aquestos models podem estudiar amb més detall la natura dels diversos objectes que trobem als cartografiats de galàxies, i entendre així la seua morfologia.

Les coordenades i les propietats que conformen un catàleg cosmològic creen un extraordinari procés puntual en termes tant d'extensió com de complexitat. La comprensió d'aquest fenòmen s'assolirà mitjançant l'adecuada descripció dels fenòmens físics que regeixen l'univers i la correcta translació d'aquestes lleis als models probabilístics.

Bibliography

Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2009, ApJS, 182, 543

Abbas, U., & Sheth, R. K. 2006, MNRAS, 372, 1749

Aceves, H., Velázquez, H., & Cruz, F. 2006, MNRAS, 373, 632

Ahmad, F., Saslaw, W. C., & Bhat, N. I. 2002, ApJ, 571, 576

Aihara, H., Allende Prieto, C., An, D., et al. 2011, ApJS, 193, 29

Amendola, L., Appleby, S., Bacon, D., et al. 2013, Living Reviews in Relativity, 16, arXiv:1206.1225

Ang, Q. W., Baddeley, A., & Nair, G. 2012, Scandinavian Journal of Statistics, 39, 591

Arnalte-Mur, P., Hurtado-Gil, L., Martínez, V. J., & Peacock, J. A. in prep.

Arnalte-Mur, P., Vielva, P., Martínez, V. J., et al. 2016, JCAP, 3, 005

Arnalte-Mur, P., Martínez, V. J., Norberg, P., et al. 2014, MNRAS, 441, 1783. Including Hurtado-Gil, L.

Ascaso, B., Benítez, N., Fernández-Soto, A., et al. 2015, MNRAS, 452, 549. Including Hurtado-Gil, L.

Azevedo-Filho, A., & Shachter, R. D. 1994, in Proceedings of the Tenth international conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc., 28–36

Baddeley, A. 2007, in Statistical Challenges in Modern Astronomy IV, Vol. 371, 22

Baddeley, A., Rubak, E., & Turner, R. 2015a, Spatial Point Patterns: Methodology and Applications with R (CRC Press)

—. 2015b, Spatial Point Patterns: Methodology and Applications with R (London: Chapman and Hall/CRC Press), in press

Baddeley, A., & Turner, R. 2000a, Australian & New Zealand Journal of Statistics, 42, 283

—. 2000b, Australian & New Zealand Journal of Statistics, 42, 283

—. 2005, Journal of Statistical Software, 12, 1

Baddeley, A., Turner, R., Mateu, J., & Bevan, A. 2013, Journal of Statistical Software, 55, 1

Baddeley, A., Turner, R., Müller, J., & Hazelton, M. 2005, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67, 617

Balian, R., & Schaeffer, R. 1989, A&A, 220, 1

Baugh, C. M., Gaztanaga, E., & Efstathiou, G. 1995, MNRAS, 274, 1049

Bell, E. F., Wolf, C., Meisenheimer, K., et al. 2004, ApJ, 608, 752

Benítez, N. 2000, ApJ, 536, 571

Benitez, N., Dupke, R., Moles, M., et al. 2014, ArXiv e-prints, arXiv:1403.5237

Benítez, N., Dupke, R., Moles, M., et al. 2015, in Highlights of Spanish Astrophysics VIII, ed. A. J. Cenarro, F. Figueras, C. Hernández-Monteagudo, J. Trujillo Bueno, & L. Valdivielso, 148–153

Bernardeau, F., Colombi, S., Gaztañaga, E., & Scoccimarro, R. 2002, PR, 367, 1

Bertone, G. 2010, Particle dark matter: Observations, models and searches (Cambridge University Press)

Besag, J. 1975, The statistician, 179

Besag, J., Milne, R., & Zachary, S. 1982, Journal of Applied Probability, 210

Blanton, M. R., Schlegel, D. J., Strauss, M. A., et al. 2005, AJ, 129, 2562

Burbidge, E. M., Burbidge, G. R., Fowler, W. A., & Hoyle, F. 1957, Reviews of modern physics, 29, 547

Burnham, K. P., & Anderson, D. R. 2002, Model selection and multimodel inference: a practical information-theoretic approach (Springer Science & Business Media)

Cabré, A., & Gaztañaga, E. 2009, MNRAS, 393, 1183

Caon, N., Capaccioli, M., & D'Onofrio, M. 1993, MNRAS, 265, 1013

Carruthers, P., & Duong-van, M. 1983, Physics Letters B, 131, 116

Cautun, M., van de Weygaert, R., & Jones, B. J. T. 2013, MNRAS, 429, 1286

Chiu, S. N., Stoyan, D., Kendall, W. S., & Mecke, J. 2013, Stochastic geometry and its applications (John Wiley & Sons)

Ciotti, L. 1991, A&A, 249, 99

Clayton, D. D. 1968, Principles of stellar evolution and nucleosynthesis (University of Chicago press)

Coil, A. L., Newman, J. A., Cooper, M. C., et al. 2006, ApJ, 644, 671

Coil, A. L., Newman, J. A., Croton, D., et al. 2008, ApJ, 672, 153

Coil, A. L., Blanton, M. R., Burles, S. M., et al. 2011, ApJ, 741, 8

Coles, P., & Jones, B. 1991, MNRAS, 248, 1

Colless, M., Dalton, G., Maddox, S., et al. 2001, MNRAS, 328, 1039

Colombi, S. 1994, ApJ, 435, 536

Cool, R. J., Moustakas, J., Blanton, M. R., et al. 2013, ApJ, 767, 118

Cooray, A., & Sheth, R. 2002, PR, 372, 1

Coupon, J., Kilbinger, M., McCracken, H. J., et al. 2012, A&A, 542, A5

Cox, D. R., & Isham, V. 1980, Point processes, Vol. 12 (CRC Press)

Croton, D. J., Colless, M., Gaztañaga, E., et al. 2004, MNRAS, 352, 828

Cucciati, O., Iovino, A., Marinoni, C., et al. 2006, A&A, 458, 39

Dalton, G., Trager, S. C., Abrams, D. C., et al. 2012, in SPIE Proc., Vol. 8446, Ground-based and Airborne Instrumentation for Astronomy IV, 84460P

Davis, M., & Geller, M. J. 1976, ApJ, 208, 13

Davis, M., Meiksin, A., Strauss, M. A., da Costa, L. N., & Yahil, A. 1988, ApJL, 333, L9

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

Dawson, K. S., Schlegel, D. J., Ahn, C. P., et al. 2013, AJ, 145, 10

de la Torre, S., Guzzo, L., Kovač, K., et al. 2010, MNRAS, 409, 867

de la Torre, S., Le Fèvre, O., Porciani, C., et al. 2011, MNRAS, 412, 825

de la Torre, S., Le Fèvre, O., Porciani, C., et al. 2011, MNRAS, 412, 825

De Lapparent, V., Geller, M. J., & Huchra, J. P. 1986, The Astrophysical Journal, 302, L1

de Vaucouleurs, G. 1948, Annales d'Astrophysique, 11, 247

Dekel, A., & Lahav, O. 1999, ApJ, 520, 24

Dmitriev, V. F., Flambaum, V. V., & Webb, J. K. 2004, PRD, 69, 063506

Dodelson, S. 2003, Modern cosmology (Academic press)

Domínguez-Tenreiro, R., & Martínez, V. J. 1989, ApJL, 339, L9

Dressler, A. 1980, ApJ, 236, 351

Efron, B., & Tibshirani, R. J. 1994, An introduction to the bootstrap (CRC press)

Efstathiou, G., Frenk, C. S., White, S. D. M., & Davis, M. 1988, MNRAS, 235, 715

Einasto, J. 1965, Trudy Astrofizicheskogo Instituta Alma-Ata, 5, 87

—. 1968, Publications of the Tartu Astrofizica Observatory, 36, 414

—. 1969, Astrofizika, 5, 137

Einasto, M. 1991, MNRAS, 252, 261

Einasto, M., Heinämäki, P., Liivamägi, L. J., et al. 2016, A&A, 587, A116. Including Hurtado-Gil, L.

Einstein, A. 1915, Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin), Seite 778-786., 778

Elizalde, E., & Gaztanaga, E. 1992, MNRAS, 254, 247

Engle, R. F. 1982, Econometrica: Journal of the Econometric Society, 987

Everitt, B., Landau, S., Leese, M., & Stahl, D. 2011, Cluster Analysis (¿ Wiley Series in Probability and Statistics)

Fiksel, T. 1984, Elektronische Informationsverarbeitung und Kybernetik, 20, 270

Forero-Romero, J. E., Hoffman, Y., Gottlöber, S., Klypin, A., & Yepes, G. 2009, MNRAS, 396, 1815

Frenk, C. S., White, S. D. M., Davis, M., & Efstathiou, G. 1988, ApJ, 327, 507

Fry, J. N. 1986, ApJ, 306, 358

Fry, J. N., & Gaztanaga, E. 1994, ApJ, 425, 1

Fukugita, M., & Peebles, P. J. E. 2004, ApJ, 616, 643

Gaztañaga, E., Fosalba, P., & Elizalde, E. 2000, ApJ, 539, 522

Geyer, C. J. 1999, Stochastic geometry: likelihood and computation, 80, 79

Ghorbani, H., Moller, H., & Stoyan, D. 2006, South African Statistical Journal, 40, 75

Giovanelli, R., Haynes, M. P., & Chincarini, G. L. 1986, ApJ, 300, 77

Goto, T., Yamauchi, C., Fujita, Y., et al. 2003, MNRAS, 346, 601

Graham, A., & Colless, M. 1997, MNRAS, 287, 221

Graham, A. W., Erwin, P., Trujillo, I., & Asensio Ramos, A. 2003, AJ, 125, 2951

Graham, A. W., & Guzmán, R. 2003, AJ, 125, 2936

Guo, H., Zehavi, I., & Zheng, Z. 2012a, ApJ, 756, 127

—. 2012b, ApJ, 756, 127

Guo, H., Zehavi, I., Zheng, Z., et al. 2013, ApJ, 767, 122

Guzzo, L., Strauss, M. A., Fisher, K. B., Giovanelli, R., & Haynes, M. P. 1997, ApJ, 489, 37

Guzzo, L., Cassata, P., Finoguenov, A., et al. 2007, ApJS, 172, 254

Guzzo, L., Scodeggio, M., Garilli, B., et al. 2014, A&A, 566, A108

Hahn, O., Porciani, C., Carollo, C. M., & Dekel, A. 2007, MNRAS, 375, 489

Hamilton, A. J. S. 1988, ApJL, 331, L59

—. 1993, ApJ, 417, 19

Hamilton, A. J. S. 1998, in Astrophysics and Space Science Library, Vol. 231, The Evolving Universe, ed. D. Hamilton, 185

Hamilton, A. J. S., & Tegmark, M. 2004, MNRAS, 349, 115

Harrison, E. 2000, Cosmology: the science of the universe (Cambridge University Press)

Hartley, W. G., Almaini, O., Cirasuolo, M., et al. 2010, MNRAS, 407, 1212

Hawkins, E., Maddox, S., Cole, S., et al. 2003, MNRAS, 346, 78

Hernquist, L. 1990, ApJ, 356, 359

Hoekstra, H., Mellier, Y., van Waerbeke, L., et al. 2006, ApJ, 647, 116

Hubble, E. 1929, Proceedings of the National Academy of Science, 15, 168

—. 1934, ApJ, 79, 8

Hubble, E. P. 1925, Popular Astronomy, 33, 252

Hurtado-Gil, L., Kuhn, M. A., Arnalte-Mur, P., Feigelson, E., & Martínez, V. J. in prep.a

Hurtado-Gil, L., Martínez, V. J., Arnalte-Mur, P., Pons-Borderia, M. J., & Pareja, C. in prep.b

Hurtado-Gil, L., Arnalte-Mur, P., Martínez, V. J., et al. 2016, ApJ, 818, 174

Ilbert, O., Salvato, M., Capak, P., et al. 2008, in Astronomical Society of the Pacific Conference Series, Vol. 399, Panoramic Views of Galaxy Formation and Evolution, ed. T. Kodama, T. Yamada, & K. Aoki, 169

Ilbert, O., Capak, P., Salvato, M., et al. 2009, ApJ, 690, 1236

Illian, J., Penttinen, A., Stoyan, H., & Stoyan, D. 2008, Statistical analysis and modelling of spatial point patterns, Vol. 70 (John Wiley & Sons)

Jaffe, W. 1983, MNRAS, 202, 995

Jedamzik, K. 2004, PRD, 70, 063524

Jing, Y. P. 2000, ApJ, 535, 30

Kahn, C. H. 1994, Anaximander and the origins of Greek cosmology (Hackett Publishing)

Kaiser, N. 1987, MNRAS, 227, 1

Kaiser, N., Burgett, W., Chambers, K., et al. 2010, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7733, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 0

Kass, R. E., & Wasserman, L. 1995, Journal of the american statistical association, 90, 928

Kayo, I., Taruya, A., & Suto, Y. 2001, ApJ, 561, 22

Kerscher, M. 1999, A&A, 343, 333

Kitaura, F.-S., Jasche, J., & Metcalf, R. B. 2010, MNRAS, 403, 589

Klypin, A., Kravtsov, A. V., Bullock, J. S., & Primack, J. R. 2001, ApJ, 554, 903

Klypin, A., Zhao, H., & Somerville, R. S. 2002, ApJ, 573, 597

Klypin, A. A., Trujillo-Gomez, S., & Primack, J. 2011, ApJ, 740, 102

Komatsu, E., Smith, K. M., Dunkley, J., et al. 2011, ApJS, 192, 18

Konishi, S., & Kitagawa, G. 2008, Information criteria and statistical modeling (Springer Science & Business Media)

Korn, A. J., Grundahl, F., Richard, O., et al. 2006, Nat, 442, 657

Kowalski, M., Rubin, D., Aldering, G., et al. 2008, ApJ, 686, 749

Kravtsov, A. V., Berlind, A. A., Wechsler, R. H., et al. 2004, ApJ, 609, 35

Kuhn, M. A., Feigelson, E. D., Getman, K. V., et al. 2014, ApJ, 787, 107

Labatie, A., Starck, J.-L., Lachièze-Rey, M., & Arnalte-Mur, P. 2010, ArXiv e-prints, arXiv:1009.1232

Lahav, O., & Liddle, A. R. 2014, ArXiv e-prints, arXiv:1401.1389

Lahiri, P. 2001, in Model selection, IMS

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

Le Fèvre, O., Vettolani, G., Garilli, B., et al. 2005, A&A, 439, 845

Leavitt, H. S. 1908, Annals of Harvard College Observatory, 60, 87

Lee, J., & Lee, B. 2008, ApJ, 688, 78

Lewis, A., Challinor, A., & Lasenby, A. 2000, ApJ, 538, 473

Li, C., Kauffmann, G., Jing, Y. P., et al. 2006, MNRAS, 368, 21

Longair, M. S. 2006, The cosmic century: a history of astrophysics and cosmology (Cambridge University Press)

López-Sanjuan, C., Cenarro, A. J., Hernández-Monteagudo, C., et al. 2015, A&A, 582, A16. Including Hurtado-Gil, L.

Loveday, J., Maddox, S. J., Efstathiou, G., & Peterson, B. A. 1995, ApJ, 442, 457

LSST Dark Energy Science Collaboration. 2012, ArXiv e-prints, arXiv:1211.0310

Madgwick, D. S., Hawkins, E., Lahav, O., et al. 2003, MNRAS, 344, 847

Martínez, V. J., Arnalte-Mur, P., & Stoyan, D. 2010, A&A, 513, A22

Martínez, V. J., & Saar, E. 2002, Statistics of the Galaxy Distribution (Chapman &amp)

Marulli, F., Bolzonella, M., Branchini, E., et al. 2013, A&A, 557, A17

Mather, J. C., Cheng, E. S., Cottingham, D. A., et al. 1994, ApJ, 420, 439

Matthews, D. J., & Newman, J. A. 2012, ApJ, 745, 180

Maurogordato, S., & Lachieze-Rey, M. 1987, ApJ, 320, 13

McBride, C., Berlind, A. A., Scoccimarro, R., et al. 2011, in Bulletin of the American Astronomical Society, Vol. 43, American Astronomical Society Meeting Abstracts, 249.07

McCracken, H. J., Peacock, J. A., Guzzo, L., et al. 2007, ApJS, 172, 314

McCracken, H. J., Wolk, M., Colombi, S., et al. 2015, MNRAS, 449, 901

McCullagh, P., & Nelder, J. A. 1989, Generalized linear models, Vol. 37 (CRC press)

McNaught-Roberts, T., Norberg, P., Baugh, C., et al. 2014, MNRAS, 445, 2125

Meléndez, J., & Ramírez, I. 2004, ApJL, 615, L33

Meneux, B., Le Fèvre, O., Guzzo, L., et al. 2006, A&A, 452, 387

Merritt, D., Graham, A. W., Moore, B., Diemand, J., & Terzić, B. 2006, AJ, 132, 2685

Moles, M., Benítez, N., Aguerri, J. A. L., et al. 2008, AJ, 136, 1325

Molino, A., Benítez, N., Moles, M., et al. 2014, MNRAS, 441, 2891

Moller, J., & Waagepetersen, R. P. 2003, Statistical inference and simulation for spatial point processes (CRC Press)

Nakamura, K., Group, P. D., et al. 2010, Journal of Physics G: Nuclear and Particle Physics, 37, 075021

Navarro, J. F., Eke, V. R., & Frenk, C. S. 1996, MNRAS, 283, L72

Navarro, J. F., Frenk, C. S., & White, S. D. M. 1995, MNRAS, 275, 720

—. 1997, ApJ, 490, 493

Navarro, J. F., Hayashi, E., Power, C., et al. 2004, MNRAS, 349, 1039

Nelder, J. A., & Mead, R. 1965, The computer journal, 7, 308

Newman, J. A., Cooper, M. C., Davis, M., et al. 2013, ApJS, 208, 5

Neyman, J., & Scott, E. L. 1958, Journal of the Royal Statistical Society. Series B (Methodological), 1

Norberg, P., Baugh, C. M., Gaztañaga, E., & Croton, D. J. 2009, MNRAS, 396, 19

Norberg, P., Baugh, C. M., Hawkins, E., et al. 2002, MNRAS, 332, 827

Nussbaumer, H., Bieri, L., et al. 2009, Discovering the Expanding Universe, by Harry Nussbaumer, Lydia Bieri, Foreword by Allan Sandage, Cambridge, UK: Cambridge University Press, 2009, 1

Olive, K. A., Steigman, G., & Walker, T. P. 2000, PR, 333, 389

Papangelou, F. 1974, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 28, 207

Parkinson, D., Riemer-Sørensen, S., Blake, C., et al. 2012, PRD, 86, 103518

Peacock, J. A. 1999, Cosmological Physics (Cambridge university press)

Peacock, J. A., Cole, S., Norberg, P., et al. 2001, Nat, 410, 169

Peebles, P. J. E. 1980, The large-scale structure of the universe (Princeton university press)

Peebles, P. J. E., & Hauser, M. G. 1974, ApJS, 28, 19

Penttinen, A., & Ylitalo, A.-K. 2015, arXiv preprint arXiv:1506.07800

Penzias, A. A., & Wilson, R. W. 1965, ApJ, 142, 419

Perlmutter, S., Aldering, G., Goldhaber, G., et al. 1999, ApJ, 517, 565

Pfeifer, D., Albrecht, M., & Bäumer, H. 1992, Spatial point processes and their applications to biology and ecology (Carl von Ossietzky Universität)

Phleps, S., Peacock, J. A., Meisenheimer, K., & Wolf, C. 2006, A&A, 457, 145

Piscionere, J. A., Berlind, A. A., McBride, C. K., & Scoccimarro, R. 2014, ApJ, 806, 125

Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2014, A&A, 571, A16

Planck Collaboration, Aghanim, N., Arnaud, M., et al. 2015, ArXiv e-prints, arXiv:1507.02704

Pons-Bordería, M.-J., Martínez, V. J., Stoyan, D., Stoyan, H., & Saar, E. 1999, ApJ, 523, 480

Postman, M., Coe, D., Benítez, N., et al. 2012, ApJS, 199, 25

Pović, M., Huertas-Company, M., Aguerri, J. A. L., et al. 2013, MNRAS, 435, 3444

Prugniel, P., & Simien, F. 1997, A&A, 321, 111

Retana-Montenegro, E., van Hese, E., Gentile, G., Baes, M., & Frutos-Alfaro, F. 2012, A&A, 540, A70

Riess, A. G., Filippenko, A. V., Challis, P., et al. 1998, AJ, 116, 1009

Rivolo, A. R. 1986, ApJ, 301, 70

Roche, N., Eales, S. A., Hippelein, H., & Willott, C. J. 1999, MNRAS, 306, 538

Rosin, P., & Rammler, E. 1933, Zement, 31, 427

Saslaw, W. C., & Fang, F. 1996, ApJ, 460, 16

Saslaw, W. C., & Hamilton, A. J. S. 1984, ApJ, 276, 13

Schaller, M., Frenk, C. S., Bower, R. G., et al. 2015, MNRAS, 451, 1247

Schlegel, D., Abdalla, F., Abraham, T., et al. 2011, ArXiv e-prints, arXiv:1106.1706

Scoville, N., Aussel, H., Benson, A., et al. 2007a, ApJS, 172, 150

Scoville, N., Aussel, H., Brusa, M., et al. 2007b, ApJS, 172, 1

Seljak, U. 2000, MNRAS, 318, 203

Sérsic, J. L. 1963, Boletin de la Asociacion Argentina de Astronomia La Plata Argentina, 6, 41

—. 1968, Atlas de galaxias australes, Vol. 1 (Observatorio Astronomico, Universidad Nacional de Cordoba)

Shandarin, S. F., & Yess, C. 1998, The Astrophysical Journal, 505, 12

Sheth, R. K. 1995, MNRAS, 274, 213

Sheth, R. K., & Saslaw, W. C. 1996, ApJ, 470, 78

Skibba, R. A., Smith, M. S. M., Coil, A. L., et al. 2014, ApJ, 784, 128

Skrutskie, M. F., Schneider, S. E., Stiening, R., et al. 1997, in Astrophysics and Space Science Library, Vol. 210, The Impact of Large Scale Near-IR Sky Surveys, ed. F. Garzon, N. Epchtein, A. Omont, B. Burton, & P. Persi, 25

Smith, M. S., Kawano, L. H., & Malaney, R. A. 1993, ApJS, 85, 219

Smith, R. E., Peacock, J. A., Jenkins, A., et al. 2003, MNRAS, 341, 1311

Statisticat, & LLC. 2013a, Bayesian Inference, r package version 13.03.04

—. 2013b, LaplacesDemon: Complete Environment for Bayesian Inference, r package version 13.03.04

—. 2013c, LaplacesDemon Examples, r package version 13.03.04

—. 2013d, LaplacesDemon Tutorial, r package version 13.03.04

Stefanon, M. 2011, PhD Thesis, Universitat de Valencia, 1, 1

Stoica, R. S., Philippe, A., Gregori, P., & Mateu, J. 2015, ArXiv e-prints, arXiv:1507.04228

Stott, J. P., Swinbank, A. M., Johnson, H. L., et al. 2016, ArXiv e-prints, arXiv:1601.03400

Stoyan, D., & Grabarnik, P. 1991, Mathematische Nachrichten, 151, 95

Strauss, M. A., Weinberg, D. H., Lupton, R. H., et al. 2002, AJ, 124, 1810

Swanson, M. Accessed: 2016-03-25, LasDamas webpage, `http://lss.phy.vanderbilt.edu/lasdamas/`

Swanson, M. E. C., Tegmark, M., Hamilton, A. J. S., & Hill, J. C. 2008, MNRAS, 387, 1391

Team, R. C. 2015, URL http://www. R-project. org

Tegmark, M., Eisenstein, D. J., Strauss, M. A., et al. 2006, PRD, 74, 123507

Tempel, E., Kipper, R., Tamm, A., et al. 2016, A&A, 588, A14

Tempel, E., Stoica, R. S., Martínez, V. J., et al. 2014, MNRAS, 438, 3465

Tempel, E., Tago, E., & Liivamägi, L. J. 2012, A&A, 540, A106

Thomas, D., Maraston, C., Bender, R., & Mendes de Oliveira, C. 2005, ApJ, 621, 673

Tinker, J. L., Weinberg, D. H., Zheng, Z., & Zehavi, I. 2005, ApJ, 631, 41

Totsuji, H., & Kihara, T. 1969, PASJ, 21, 221

Trimble, V. 1995, Publications of the Astronomical Society of the Pacific, 1133

Ueda, H., & Yokoyama, J. 1996, MNRAS, 280, 754

Van Lieshout, M. 2000, Markov point processes and their applications (World Scientific)

Watson, D. F., Berlind, A. A., McBride, C. K., Hogg, D. W., & Jiang, T. 2012, ApJ, 749, 83

Watson, D. F., Berlind, A. A., & Zentner, A. R. 2011, ApJ, 738, 22

Weibull, W. 1951, Journal of applied mechanics, 103

West, M. J., Dekel, A., & Oemler, Jr., A. 1987, ApJ, 316, 1

White, S. D. M. 1979, MNRAS, 189, 831

Wild, V., Peacock, J. A., Lahav, O., et al. 2005, MNRAS, 356, 247

Wolf, C., Dye, S., Kleinheinrich, M., et al. 2001, A&A, 377, 442

Yang, A., & Saslaw, W. C. 2011, ApJ, 729, 123

York, D. G., Adelman, J., Anderson, Jr., J. E., et al. 2000, AJ, 120, 1579

Zehavi, I., Weinberg, D. H., Zheng, Z., et al. 2004, ApJ, 608, 16

Zehavi, I., Zheng, Z., Weinberg, D. H., et al. 2011, ApJ, 736, 59

Zucca, E., Bardelli, S., Bolzonella, M., et al. 2009, A&A, 508, 1217

Acknowledgements I

ALHAMBRA is based on observations collected at the German-Spanish Astronomical Center at Calar Alto, which is jointly operated by the Max-Planck-Institut für Astronomie (MPIA) and the Instituto de Astrofísica de Andalucía (CSIC). This work was mainly supported by the Spanish Ministry for Economy and Competitiveness and FEDER funds through grants AYA2010-22111-C03-02 and AYA2013-48623-C2-2, and by the Generalitat Valenciana through project PrometeoII 2014/060. We also acknowledge support from the Spanish Ministry for Economy and Competitiveness and FEDER funds through grants AYA2012-39620, AYA2013-40611-P, AYA2013-42227-P, AYA2013-43188-P, AYA2013-48623-C2-1, ESP2013-48274, AYA2014-58861-C3-1, Junta de Andalucía grants TIC114, JA2828, P10-FQM-6444, and Generalitat de Catalunya project SGR-1398.

Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III web site is http://www.sdss3.org/.

SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofísica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington, and Yale University.

We thank LasDamas project for the provided simulations used in this work.

Funding for PRIMUS is provided by NSF (AST-0607701, AST-0908246, AST-0908442, AST-0908354) and NASA (Spitzer-1356708, 08-ADP08-0019, NNX-09AC95G).

Funding for the DEEP2 Galaxy Redshift Survey has been provided by NSF grants AST-95-09298, AST-0071048, AST-0507428, and AST-0507483 as well as NASA LTSA grant NNG04GC89G.

We thank COSMOS project for the provided galaxy catalog used in this work.

The CosmoSim database used in this paper is a service by the Leibniz-Institute for Astrophysics Potsdam (AIP). The MultiDark database was developed in cooperation with the Spanish MultiDark Consolider Project CSD2009-00064.

The Bolshoi and MultiDark simulations have been performed within the Bolshoi project of the University of California High-Performance AstroComputing Center (UC-HiPACC) and were run at the NASA Ames Research Center. The MultiDark-Planck (MDPL) and the BigMD simulation suite have been performed in the Supermuc supercomputer at LRZ using time granted by PRACE.

Acknowledgements II

Aquesta secció ha de començar necessariament amb l'agraïment a aquell que va fer possible que ara haja d'estar agraït també amb molta altra gent. De fet, aquest és un dels majors mérits atribuibles a algú. La rapidessa amb que Vicent Martínez va acceptar dirigir-me una beca de col·laboració departamental em va sorprendre per la facilitat amb que vaig entrar al món de la investigacsió científica. D'ahí, cap a un máster i quatre anys d'investigació predoctoral, s'han succeït episodis i experiéncies que per extensió no podria, i no sempre convé, incloure ací.

Qui si convé anomenar són tots els *sputniks* que han fet possible i han posat color a aquestos anys. Pablo Arnalte, qui de company predoctoral passà a codirector de tesi i imprescindible mentor; Elisa Nespoli, de qui amb silenci observador i respecte aprenguí com treballen els científics de veritat; Alberto Fernández-Soto, font inesgotable de coneiximents, útils consells i solucions laborals; i en general als companys de l'Observatori Astronómic de la Universitat de València: Fernando Ballesteros, Amelia Ortiz, Juan Fabregat, Juan Carlos Guirado, Julia Suso, Lorena Nieves, Leonardo Gouvelis, Rebecca Azulay, Sofía Fuentes, el *Pix Insight team* Vicent Peris, Juan Conejero i Javio Sanchis, i també a Miquel Gómez i Xusa Moya, i com no, a l'increïble Javier Díez. I al sosteniment acadèmic i laboral, cal incloure el del *Instituto de Física de Cantabria*, per la confiança quasi cega dipositada en mi per a honorar la seua llista de membres.

Regarding the planet, I must not forget the wise man of *Tartu Observatoorium*, Enn Saar, who offered the most welcoming and inspiring symposiums I have ever been and enjoyed together with Radu Stoica, from whom I wish to learn so much. Elmo, Juhan, Maarja, Jaan, Pekka and many more are among the people that made my world bigger. At the other side of the Atlantic, Eric, Jogesh, Murali and the Center for Astrostatistics at Penn State made me get a glance of how the boundaries of sciences look like. Conferences and summer schools bring you friends to share the universe, and make it less empty. Dory, Natasha, Daniel, Nathan, Dillon, Alex, Adrew, Mateja and many more fill these awesome experiences. Special mention to Rolf Turner for writing the `spatstat` code and solving my code problems selflessly.

Altres treballaren abans per a que açò fora possible, com aquells docents que aconseguien que parares d'escriure i començares a parar atenció: Josep Guia, Juan Climent, José Vicente Beltrán, Francisca Mascaró, Pep Mulet, Pablo Galindo, Miquel Portilla i Ramon Lapiedra. A En David, pels anys de campus, ciència i rol. I parlant dels anys de les aules, a Tere, pel temps compartit, i aquells de qui aprenguí a jugar al truc: Juan, Juanjo, Anna, Perfecto, Celest, els dos Javis, Txus, Marina, el Dr. Cosme i el futur Dr. Folch. I parlant de *bergantes*, els socis de l'Alqueria: Rubén, Carlos, Vicente i Enric, i Laia, María, Almudena, Javi, Alex i Lucía, per tantes nits.

I molts més han treballat per a fer açò una realitat, com són tots els valencians, càntabres i altres gents que, sense saber-ho ni donar permís, han pagat este llibret. I a Google, YouTube, Wagner i el Heavy Metal, i desenes de serveis gratuits i de gran qualitat que el mercat posa a la nostra disposició i que fan possible un treball més ràpid i agradable.

I finalment distingir a aquells qui des del principi entregaren el seus esforços vitals i la seua companyia: els meus pares, mon germà Vicent i els meus àvis i familiars, sense els quals res de tot açò tindria valor.

A Aquell qui És per haver-nos donat un objecte d'estudi, i a Clara per haver-me donat un motiu per a estudiar-lo.