INVESTIGATING OBSERVATIONAL PROBES OF GALAXY EVOLUTION IN OBSERVATIONS AND SIMULATIONS

Dominic Bates

A Thesis Submitted for the Degree of PhD
at the
University of St Andrews

2022

Investigating observational probes of galaxy evolution in observations and simulations

Dominic Bates

University of
St Andrews

This thesis is submitted in partial fulfilment for the degree of

Doctor of Philosophy (PhD)

at the University of St Andrews

June 2019

Candidate's declaration

I, Dominic Bates, do hereby certify that this thesis, submitted for the degree of PhD, which is approximately 30,000 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for any degree.

I was admitted as a research student at the University of St Andrews in September 2015.

I received funding from an organisation or institution and have acknowledged the funder(s) in the full text of my thesis.

Date 28 June 2019 Signature of candidate

Supervisor's declaration

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date 28 June 2019 Signature of supervisor

Permission for publication

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand, unless exempt by an award of an embargo as requested below, that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that this thesis will be electronically accessible for personal or research use and that the library has the right to migrate this thesis into new electronic forms as required to ensure continued access to the thesis.

I, Dominic Bates, confirm that my thesis does not contain any third-party material that requires copyright clearance.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

Printed copy

No embargo on print copy.

Electronic copy

No embargo on electronic copy.

Date	29 June 2019	Signature of candidate
Date	28 June 2019	Signature of supervisor

Underpinning Research Data or Digital Outputs

Candidate's declaration

I, Dominic Bates, hereby certify that no requirements to deposit original research data or digital outputs apply to this thesis and that, where appropriate, secondary data used have been referenced in the full text of my thesis.

Date 29 June 2019 Signature of candidate

Abstract

In this thesis, we perform tests of galaxy evolution using a number of different observational probes. In chapter 2, we implement a method of obtaining redshift distributions for photometric samples of galaxies, known as clustering redshifts. We test this on real data from the Baryon Oscillation Spectroscopic Survey (BOSS), and simulated data from semi-analytic models, showing that assumptions about the bias evolution of the unknown sample can become important for some samples of galaxies, particularly at faint magnitudes. We also find that the choice of clustering scale makes a big difference to the noise in the recovered redshift distribution. In chapter 3, we apply the clustering redshifts method to data from the Sloan Digital Sky Survey (SDSS), recovering redshift distributions as a function of colour, allowing us to compute mass and luminosity functions over large volumes. Little evolution is seen in our recovered mass function between $0.2 < z < 0.8$, implying the most massive galaxies form most of their mass before $z = 0.8$. These mass functions are used to produce stellar mass completeness estimates for BOSS, giving a completeness of 80% above $M_\star > 10^{11.4} M_\odot$ between $0.2 < z < 0.7$, with completeness falling significantly at higher redshifts and lower masses. In chapter 4, we go on to investigate how well the formation history of a dark matter halo can be inferred in simulations from the observable properties of a galaxy, finding that applying a machine learning approach considering multiple properties performs significantly better than using individual properties. We add errors to parameters, finding that a machine learning approach still performs best, and finally use this approach to compute formation times for the GAMA survey. We investigate how formation time changes with environment at fixed mass, finding signs of assembly bias, with high mass halos in dense environments being younger than those in under-dense regions, and the trend reversing at lower halo masses, with halos in dense environments being older.

Acknowledgements

I would firstly like to thank my supervisor, Rita Tojeiro for her unparalleled guidance and support throughout my PhD. I am very grateful for her supervision, and am thankful that she always found time to answer my questions and provide help.

I would like to thank my office mates and fellow PhD students, for helping discuss problems, providing distraction from work (particularly during coffee breaks), and just generally helping me enjoy my time here.

I am very grateful to the university, and also to the other members of staff around the department for their help, and am very glad I had the chance to work with such a talented group of people.

I must also thank my family, who have always given encouragement (and financial help!) throughout my years at university.

I would lastly like to thank anyone else who has helped me enjoy my time here particularly my friends, who have made my time here unforgettable.

This work was supported by the Science and Technology Facilities Council (STFC), to which I am very greatful.

Contents

List of Figures

List of Tables

1 Introduction

Our current best-fit model of cosmology, the ΛCDM model, states that the universe consists mostly of dark energy, dark matter and baryonic matter. These invisible, dark matter particles collapse under gravity in to halos, and the baryonic component of galaxies (i.e. stars, gas and dust) form within the halo's gravitational well.

Understanding this link between galaxies and the underlying dark matter is vital, since this not only underpins our models of how galaxies form and evolve, but also, since dark matter and dark energy are not directly detectable, serves as the best way we can probe the underlying cosmology of our universe. Only by connecting these two areas can we build up a general picture of how our universe evolves.

The layout of this chapter is as follows: Section 1.1 gives an introduction to studies of galaxy evolution and outlines the ΛCDM model of cosmology. Section 1.2 gives some background on galaxy formation simulations and describes the simulations used in this thesis. Section 1.3 describes the different types of galaxy survey, and details some of the most important of these, along with describing how important observational probes, for example redshifts, mass functions and luminosity functions are computed. Finally, section 1.4 outlines the contents of this thesis.

1.1 Background

1.1.1 A history of the large-scale universe

The study of the large-scale universe arguably begins in the 1920s around the time of the so-called "great debate", when astronomers debated whether distant nebulae (i.e. galaxies) were small objects present at the edge of our own galaxy, or independent galaxies, far away from the Milky Way and large in physical size.

Hubble (1925) measured the absolute distance to the Andromeda galaxy using Cepheid variable stars, proving that these nebulae are situated at large distances from our galaxy, and that the Universe is much larger than the extent of our galaxy. Expanding on this, Hubble (1929) measured distances to more galaxies, and after combining this with data on recession velocity, showed that galaxies at larger distances are receding from us faster, implying a start time to the universe, and establishing the Big Bang model of cosmology.

The internal structure, or morphology, of galaxies has also been extensively studied, most notably in Hubble (1926), leading to the Hubble tuning fork classification for galaxies. In this system, all galaxies can be classified as either ellipticals or spirals, and spirals sub-divided in to galaxies with or without a bar in its centre. Further studies of these classifications, for example Holmberg (1958), concluded that a galaxy's physical properties are strongly correlated with its morphology, with elliptical galaxies appearing redder, and showing less star formation than spirals, which are typically bluer, with large amounts of ongoing star formation.

As astronomical instrumentation improved, the dynamics of galaxies were studied in detail in Rubin & Ford (1970) and Freeman (1970), showing that stars around the edge of a galaxy have a much higher velocity than is expected from purely the baryonic matter of the galaxy. Although similar effects were first observed in Zwicky (1933), this was the first convincing evidence for some other form of matter, suggesting a universe not made from purely baryonic matter, but also "dark matter".

More recently, Riess et al. (1998) and Perlmutter et al. (1999), after looking at large samples of supernovae at different redshifts found that the expansion of the universe is accelerating at late times, implying the existence of another form of energy, usually referred to as “dark energy”. Our current best-fit model of cosmology and galaxy evolution is therefore a model containing baryonic matter, dark matter, and dark energy, known as the ΛCDM model.

1.1.2 The ΛCDM model

In the ΛCDM model of cosmology, which is now well established (see Dodelson (2003); Liddle (2003) for a comprehensive outline of the model), after an initial big bang, an inflationary period occurs, growing quantum fluctuations quickly to cosmological scales. The universe initially consists of cold (i.e. non-relativistic) dark matter and a hot, dense plasma of photons and baryons. This is effectively opaque due to positively charged protons scattering these photons. As the universe expands, this plasma cools and electrons bind with protons in a period known as recombination, occurring roughly 400,000 years after the big bang. This forms hydrogen and small amounts of heavier elements, and allows photons to then travel freely, the afterglow of which is now observable as the cosmic microwave background (CMB).

The initial quantum fluctuations give rise to fluctuations in the density field, which grow over time under gravity and form the seeds of structure formation. Dark matter collapses in to halos around these density peaks, and gas falls to the centre. After cooling, this starts to form stars, forming the first protogalaxies at the centre of these halos. Halos are expected to grow and merge through time under gravity, with smaller halos merging to create lager ones, and groups of halos collapsing under gravity creating large bound structures. Since each galaxy follows the evolution of its host halo, the evolution of gas and stars within it is strongly affected by the merger history and environment of the halo (Matthee et al., 2017).

Until around 50,000 years after the big bang, the expansion of the universe is dominated by radiation. After this, until around 10 billion years, matter dominates the expansion, causing deceleration. From then until present day at roughly 13.7 billion years, dark energy dominates, pushing the universe apart, causing the accelerated expansion observed at late times. Figure 1.1 shows the best fit prediction of the expansion rate from CMB data, including the onset of accelerated expansion.

We can now fit the ΛCDM model to different observational signatures, for example, the CMB (Planck Collaboration et al., 2018), supernovae (Suzuki et al., 2012), Baryon Acoustic

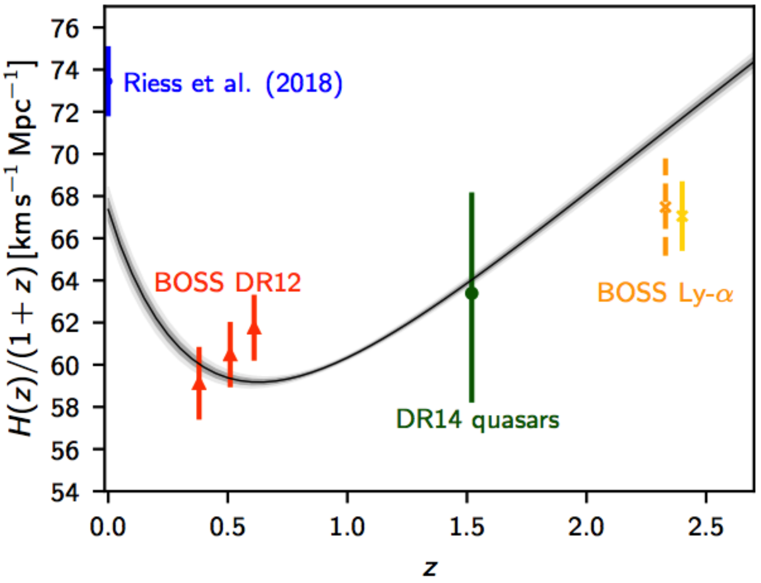

Figure 1.1: The comoving Hubble parameter (i.e. rate of expansion) as a function of redshift. The grey bands show the 68 % and 95 % confidence ranges of the best-fit Planck Collaboration et al. (2018) model. Data from other studies are shown as coloured points: BOSS DR12 (Alam et al., 2017) in red, BOSS DR14 (Zarrouk et al., 2018) in green, and BOSS Ly α (Bautista et al., 2017) in orange. The epoch of dark energy domination and onset accelerated expansion is clearly visible around $z = 0.6$. Figure taken from Planck Collaboration et al. (2018)

Oscillations (BAO) (Alam et al., 2017; Zarrouk et al., 2018), and weak lensing (Heymans et al., 2013). The data favours a universe currently containing roughly 4% baryons, 27% dark matter and 69% dark energy (Planck Collaboration et al., 2018). We can fairly simply model how the universe evolves on the largest of scales, however, moving to smaller, non-linear scales (e.g. intra-cluster scales), the physics involved is much more complex, and exactly how the ΛCDM model produces the galaxy population we observe is not yet fully understood.

1.1.3 Current consensus on galaxy evolution

One of the best ways to investigate how galaxies evolve in a cosmological context is by looking at their observational properties in galaxy surveys. We know from local galaxy surveys like the SDSS (York et al., 2000) that the galaxy population is highly bimodal in colour. The population of red, high mass, elliptical galaxies is known as the "red sequence", and the blue, lower mass, disk-like population is known as the "blue cloud". A small number of galaxies possess intermediate properties, in a region often referred to as the "green valley". The local galaxy bimodality in colour vs mass is shown in figure 1.2. Blue cloud galaxies in general tend to reside in under-dense regions of the universe, and have more cold gas and star formation occuring than in their red cloud counterparts, which tend to reside in highly dense environments, and tend to have quenched star formation. It is often therefore common to define these two groups by applying a dividing cut in the colour-mass plane, or specific star-formation rate ($sSFR = SFR/M_{\star}$) vs mass plane, although cuts based on morphology produce similar populations, as seen in figure 1.2.

Using deep surveys, this bimodality has been shown to be present up to high redshifts (Brammer et al., 2011; Muzzin et al., 2013), and the evolution of the number density of these two populations can be tracked. The number density of blue cloud galaxies seems to remain roughly constant or decrease towards low redshifts, whereas the number galaxies in the red cloud galaxies seems to increase at later times (Bell et al., 2004; Faber et al., 2007; Brammer et al., 2011; Muzzin et al., 2013). Since one might expect the number of star-forming blue-cloud galaxies to grow as new stars are formed, this implies some method of shutting off star formation and creating red-sequence galaxies.

With modern surveys, it is also possible to measure the stellar masses of galaxies, and how this property evolves with redshift. This is often represented by the galaxy stellar mass function (see also section 1.3.4), which measures the comoving number density of galaxies as

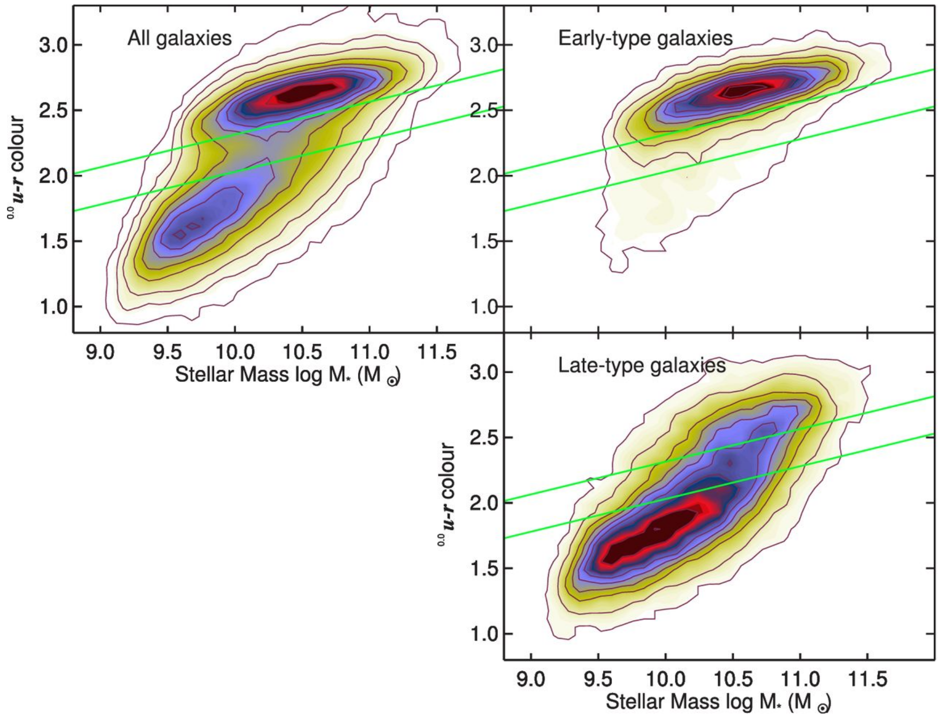

Figure 1.2: The galaxy bimodality in a sample of low redshift galaxies ($0.02 < z < 0.05$), plotted in u-r vs stellar mass. This is shown for the whole sample (top left), and separately for galaxies split by morphology in to early-type/ellipticals (top right) and late-type/spirals (bottom right). Figure taken from Schawinski et al. (2014).

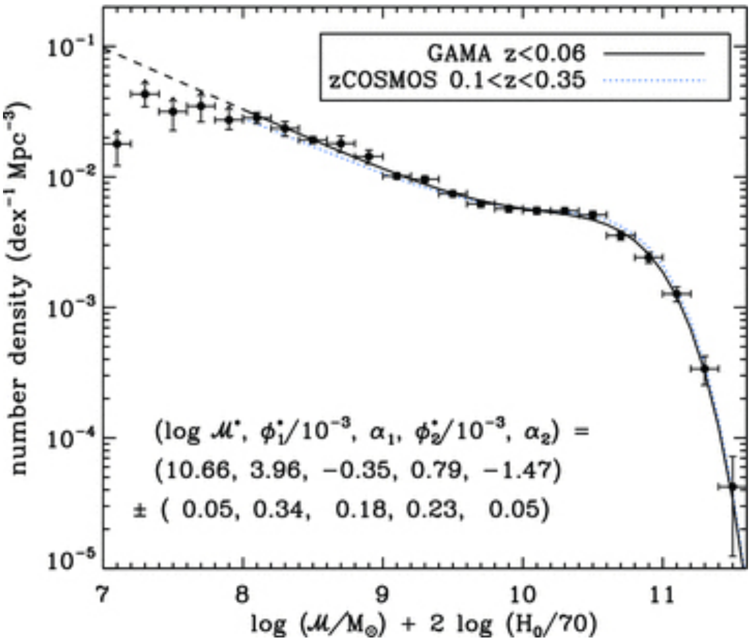

Figure 1.3: The local ($z < 0.06$) stellar mass function from the GAMA survey, with a double Schechter fit to the data points. Figure taken from Baldry et al. (2012).

a function of stellar mass. The local mass function is relatively well constrained down to low masses. The $z < 0.06$ mass function computed with the GAMA survey (Baldry et al., 2012) is shown in figure 1.3. It is easy to see that the number of galaxies drops off significantly above $M_\star \simeq 10^{10.5} M_\odot$ and rises towards lower masses. At higher redshifts, the mass function becomes more difficult to measure, as galaxies are much fainter, however, a number of surveys have investigated this, showing that the number of high mass galaxies increases dramatically from early times until roughly $z = 2$, then much more slowly after this, suggesting massive galaxies form most of their mass early on then evolve relatively passively (Marchesini et al., 2009; Moustakas et al., 2013; Muzzin et al., 2013). The number density of lower mass galaxies keeps increasing beyond $z = 2$, implying that at lower-masses, galaxies keep forming stars later.

The general idea is that blue cloud galaxies transition to the red cloud by some process or set of processes which disrupt morphology and quenches star formation. Galaxy mergers are thought to be important as these build mass, and can trigger processes like supernovae or AGN feedback, which can heat or blow out gas in the galaxy, quenching star formation (Sanders et al., 1988; Di Matteo et al., 2007; Jogee et al., 2009; Hopkins et al., 2013). Studies also suggest, however, that this may not be true for all masses, and environments, and that the likelihood of quenching, and quenching mechanism involved may vary significantly depending on other parameters (for example stellar mass or environment).

It has also long been known that the properties of galaxies are strongly correlated with both their clustering properties (Zehavi et al., 2005) and environment (Kimm et al., 2009; Grützbauch et al., 2011). Galaxy evolution is therefore strongly correlated the evolution of dark matter and cosmic structure (often referred to as the cosmic web). One idea is that the mass of the host halo predominantly dictates a galaxy's stellar mass, and in turn, its other properties, for example star formation history or morphology.

The difference in clustering between galaxies of different properties (e.g. mass, colour, SFR) is thought to be explainable by the fact that these properties are correlated with halo mass, and that more massive halos are more strongly clustered. In this regime, galaxies are thought to form via dissipative processes within the gravitational wells of the halos. Gas falls to the centre of the halo which ignites star formation and builds mass. It is known that the stellar mass to halo mass ratio varies for halos of different mass, peaking at stellar masses of $M_{halo} \sim 10^{12} M_{\odot}$ at low redshifts (Wang et al., 2013), implying that halos of this mass have been most efficient at forming stars. Above this, feedback from Active galactic nuclei (AGN) is thought to be be important for lowering this efficiency, and below this, supernovae feedback. There is also a significant scatter in this relation of roughly 0.2 dex in stellar mass at fixed halo mass (Tinker et al., 2017; Reddick et al., 2013). It is not certain, however, if this scatter is expected for a model with halo mass dictating all properties, or if other halo properties are correlated with stellar mass.

Looking at simulations, even for halos of the same mass, other parameters, for example the halo formation history or concentration, appear to affect the properties of the host galaxies (including their clustering), since a different halo formation history or concentration may correspond to a very different merger rate or star-formation history for the galaxy (Gao et al., 2005; Wechsler et al., 2006; Gao & M., 2007; Chaves-Montero et al., 2016; Matthee et al., 2017; Artale et al., 2018). This difference in galaxy properties depending on their halo formation history is often referred to as "assembly bias". Despite being difficult to measure observationally, this can potentially explain the distribution of galaxy parameters we see, and in particular, the relation between stellar mass and halo mass (the stellar-halo mass ratio), and the scatter in this measurement (Lehmann et al., 2017; Zentner et al., 2019).

1.1.4 Ongoing questions

Despite understanding many of the processes which govern how galaxies evolve, there are many questions which remain uncertain. From a galaxy evolution perspective: What is the dominant mechanism by which galaxies are quenched, and move to the red sequence, and does this change depending on galaxy mass, environment and redshift? How significant is the role of mergers, and what is the merger rate for different galaxy types? Evidence suggests both AGN, and supernovae feedback can release large amount of energy in to a galaxy causing gas heating, or even blowing out gas, but how important is each mechanism in quenching star formation?

From a cosmological perspective, there are also unsolved problems. Simulations appear to produce broadly the correct distribution of galaxies, but some differences remain (e.g. cusp/core problem, missing satellites problem, satalite plane problem) (Del Popolo & Le Delliou, 2017). Can these be fixed by improving the resolution and analytic descriptions in our simulations, or are there fundamental problems with ΛCDM? The properties of simulated galaxies of a certain halo mass appear to vary with halo assembly (i.e. assembly bias). How significant is this in the real universe? Also, although we have measures of how the expansion history of the universe evolves, as the precision and redshift range of these measurements improves, does ΛCDM still work well? Is dark energy explained by a cosmological constant, or does it vary with time? And does GR + dark matter work at all scales, or do we require some modification of gravity?

Answering these questions requires two things: Observational probes from the real universe, and theoretical models from which we are able to predict these observables. When investigating galaxy evolution and cosmology on very large scales, our best form of theoretical models comes in the form of cosmological galaxy simulations.

1.2 Simulations of galaxy evolution

Cosmological simulations have long been used to test our models of cosmology, e.g. Frenk et al. (1983); Davis et al. (1985). Advances in both computational power and our analytic descriptions of galaxy physics in recent years have allowed us to simulate the evolution of cosmic structure at high resolution over very large scales (e.g. Millennium (Springel et al., 2005a), EAGLE (Schaye et al., 2015), Illustris (Nelson et al., 2015)). The real universe relies on physical processes with an extremely large range of physical scales, from the atomic scales of nuclear fusion, to the megaparsec scales of cosmic structure, so simulating all processes directly is impossible. Because of this, we must develop analytic descriptions of physics to simulate things over large scales. This has lead to two broad classes of cosmological simulations: Semi-analytic models (SAMs) and cosmological hydro-dynamical simulations.

SAMs rely on firstly simulating the evolution of a universe with purely dark matter and dark energy (i.e. ignoring the evolution of baryons), and then using the resulting evolution of each individual dark matter halo to evolve galaxies using analytic descriptions. This has the advantage of being computationally cheap, so can be done over very large scales, and many times for different parameters, however relies heavily on analytic descriptions and ignores any feedback between baryons and dark matter. The second method, hydro-dynamical simulations, directly simulates the dynamics of gas and stars along with dark matter, with some analytic descriptions of star formation, feedback etc. This is more computationally expensive, however allows us to test our models of galaxy formation more directly, and allows us to investigate the formation processes leading to each individual galaxy in a more fundamental way.

Somerville & Davé (2015) present a detailed review of the different types of simulation, along with current simulations and their individual benefits. Here, we present a simplified view, outlining how both SAMs and hydrodynamical codes work, and detailing the simulations used in this thesis.

1.2.1 N-body simulations

Both SAMs and hydro-dynamical simulations rely on gravitationally simulating the evolution of a universe dominated by dark matter, (although baryonic physics is simulated along side this in hydro-dynamical codes). A simple approach would be to treat the universe as a box, split the universe in to a large number of (dark matter) particles, and compute the force from each particle to every other particle, then evolve these particles in small time steps (the force on particles at edges of the box can be computed using periodic boundary conditions). The problem with this method, however, is that this would become incredibly computationally expensive when working with large number of particles, since computational time would scale with $\mathcal{O}(n^2)$, where n represents the number of particles. Instead, sensible approximations need to be used, leading to two main methods: particle-based and mesh-based, although methods exist which are a hybrid between the two.

The most popular particle-based method is the tree code (Barnes & Hut, 1986), where rather than computing gravitational force due to all particles individually, distant particles are grouped together, and force is computed for the group. Alternatively, mesh-based methods compute the gravitational potential on a grid using Fourier transforms of the density field, and particles are moved along this potential (Hockney & Eastwood, 1988). Both these methods scale with particle number as $\mathcal{O}(nlog(n))$, so are significantly faster for large simulations, however each has advantages and disadvantages, discussed in more detail in Somerville & Davé (2015). Many modern simulations therefore use a hybrid of the two methods, using tree based methods for small scale forces, and a mesh for large-range forces (Springel et al., 2005a). Figure 1.4 shows density slices from one of these hybrid N-body simulations, the millennium-II simulation (Boylan-Kolchin et al., 2009), over different scales.

Upon simulating this evolution, dark matter halos must be identified from the resulting groups of particles, and also tracked across time-steps. This allows us to construct merger trees, describing the halo at an early time-step which corresponds to the same halo at later times, including which halos have merged to create it, and their masses. These merger trees are used as the input to SAMs. Finding halos is not trivial, particularly since, as is visible in figure 1.4, halos exist over very different scales, and are known to possess substructure (i.e. smaller halos which merged with a larger halo). Several methods of identifying these halos and quantifiying their properties exist, for example friends-of-friends (FOF) (Davis et al., 1985), where halos are defined as groups containing particles separated by some defined distance, spherical overdensity (SO) (Lacey & Cole, 1994) where particles within spheres are considered, and halos are defined as having some overdensity with respect to the background density (e.g. mean density of the universe), and also 6D phase-space based methods (Behroozi et al., 2013), where halos are identified by considering both particle position and velocity. Differences in recovered parameters (e.g. halo mass, radius) can vary between methods, however Knebe et al. (2013) find that these generally agree to within 10%.

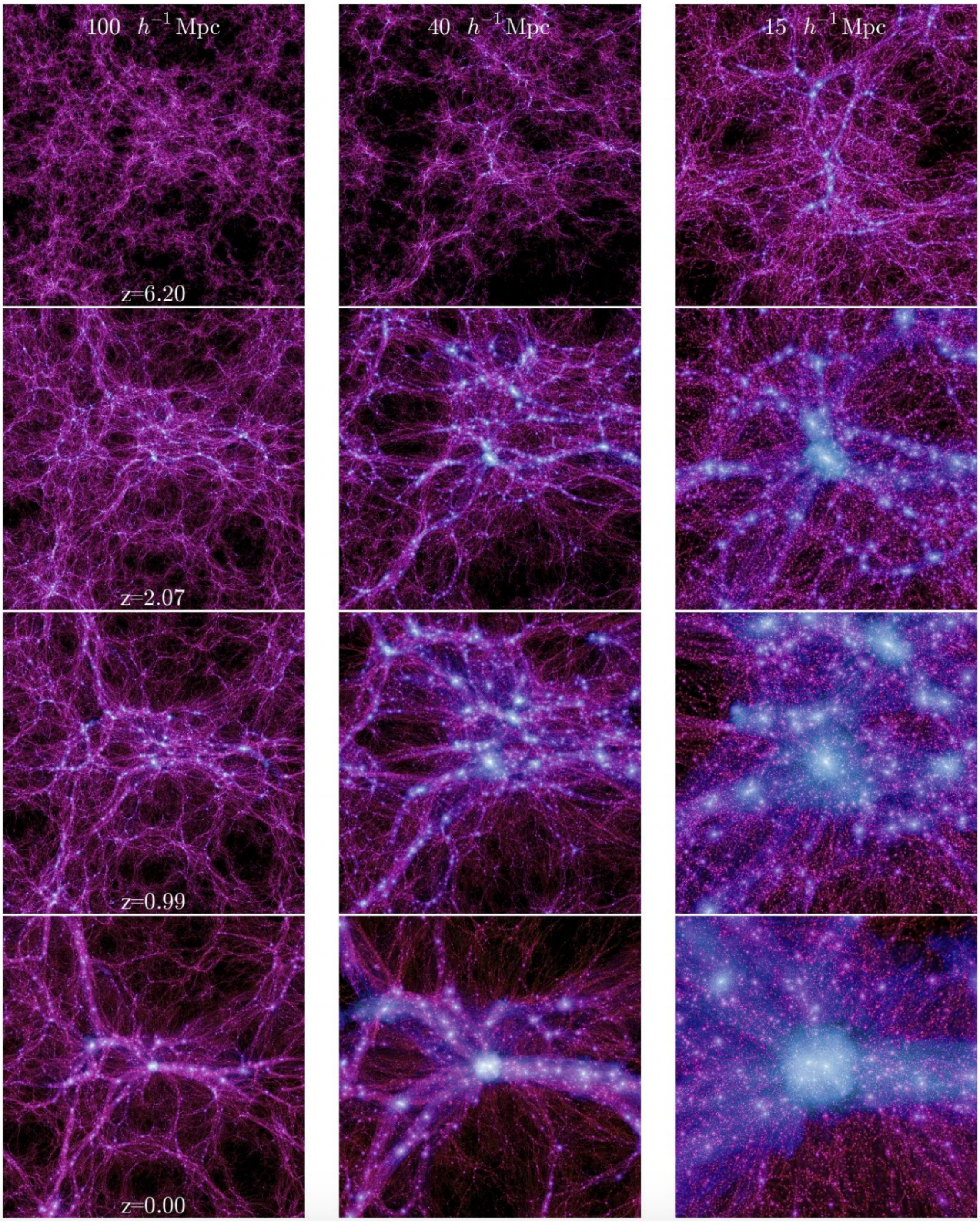

Figure 1.4: Density slices of the Millennium-II simulation (Boylan-Kolchin et al., 2009), showing the evolution of structure growth at different scales around the largest halo in the simulation ($M_{halo} = 8.3 \times 10^{14} h^{-1} M_{\odot}$ at $z = 0$). Figure taken from Boylan-Kolchin et al. (2009).

1.2.2 Hydro-dynamical simulations

Hydro-dynamical simulations require modelling gas physics alongside the evolution of dark matter. This allows us to directly simulate galaxy evolution, although due to the computational complexity, a number of assumptions must be made, and analytic models chosen. Two main methods exist for hydrodynamical simulations: Firsly, Smoothed Particle Hydrodynamics (SPH), where each particle, rather than being treated as a point source, is represented as smooth density distributions, peaking in the centre. These particles track the evolution of the gas and stars, and global properties (e.g. pressure, density) are computed from the weighted sums of the properties of neighboring particles. Secondly, we have mesh codes, where baryons are modelled as grid cells, and properties are tracked across cell boundaries. Typically, mesh codes are better at handling shocks and surface instabilities, while SPH is more adaptive, providing a better dynamic range for the same amount of CPU time. Recently, however, the difference between both methods is closing (Somerville & Davé, 2015).

Hydro codes require a set of analytic descriptions for certain areas of physics which it is not possible to model directly, often referred to as "sub-grid" physics. For example, simulating star formation directly would require simulations accurate down to atomic scales, so instead, stars are created when gas reaches some specific density and temperature. Accretion on to black holes is also difficult to model, since this is dictated by very small scale physics, so instead, black holes are evolved using analytic formulae. Feedback processes are extremely important in quenching galaxies, both from supernovae in star forming regions and also AGN feedback. These are both modelled analytically, the specifics of which can be very important for the resulting distribution of galaxies

Two of the most widely used cosmological hydro simulations are the EAGLE (Schaye et al., 2015), and Illustris (Nelson et al., 2015) projects. EAGLE is run using 7 billion particles in a 100 Mpc^3 box, designed to create large populations of galaxies, producing 10,000 galaxies of mass greater than the Milky Way. Illustris is a project of comparable scale, containing 18 billion particles in a 106 Mpc^3 box, also producing large samples of realistic galaxies. Illustris-TNG (Pillepich et al., 2018) has simulated this evolution for three different boxes of 50, 100 and 300 Mpc^3 in volume. We refer the reader to Somerville & Davé (2015) for a detailed comparison of these models.

1.2.3 Semi-analytic models

A popular alternative to running full hydro-dynamical simulations is using SAMs instead. Here, rather than simulating the evolution of gas and stars alongside dark matter, these things are done separately. The merger tree (see section 1.2.1) is taken from a cosmological-volume dark matter only simulation, and galaxies are created in the centre of dark matter halos. These are then evolved analytically relative to the merger tree through a number of physical descriptions of, for example, star formation, merging, gas accretion, feedback, morphological transformations etc., the specifics of which can vary significantly between models. Despite requiring more assumptions and analytic prescriptions than hydro-dynamical codes, due to their computational simplicity, SAMs can be run over much larger volumes, and can be run many times for different parameters in order that the closest match to the real universe can be obtained.

Some popular SAMs include *GALFORM* (Cole et al., 2000; Gonzalez-Perez et al., 2014), *DLB07* (De Lucia & Blaizot, 2007), *LGalaxies* (Henriques et al., 2013, 2015), *SAG* (Cora et al., 2018), and *SAGE* (Croton et al., 2006, 2016). For papers describing the most widely used SAMs and comparing their outputs, see De Lucia et al. (2010); Knebe et al. (2015); Asquith et al. (2018). In chapters 2, 3, and 4 we require mock galaxy catalogues from SAMs (i.e. simulated survey observations, for which we know all the information). We choose to use LGalaxies for this purpose, since it provides star formation histories with little effort, and also appears to best reproduce the stellar mass function and observed distribution of red and blue galaxies (Asquith et al., 2018). In chapter 3, we also compare this to results using SAGE.

LGalaxies (Henriques et al., 2013, 2015) is a SAM aimed at better explaining the observed evolution of stellar mass with redshift. It is an updated version of the Guo et al. (2011) SAM, with updated analytic descriptions of reincorporation of gas ejected from supernova feedback, density of cold gas required for star formation, ram-pressure stripping, and AGN radio-mode feedback. Henriques et al. (2013) is updated to Planck cosmology in Henriques et al. (2015), which is the model used in this thesis. One of the main advantages of LGalaxies is that rather than simply choosing reasonable values of parameters describing sub-grid physics, the parameter space is fully sampled with a Markov chain Monte Carlo (MCMC) method in order to find the best fit set of parameters, along with the errors. The SAM was fit to the stellar mass function, and K-band and B-band luminosity functions at $z = 0, 1, 2$, and 3.

We will also compare this model to the SAGE semi analytic model Croton et al. (2016)

in chapter 3, which is an updated version of the model presented in Croton et al. (2006), obtainable through the Theoretical Astrophysical Observatory (TAO) (Bernyk et al., 2016), a publicly available online tool used to obtain magnitudes, spectra, and light-cones from the output of several different SAMs. This model, in comparison to Croton et al. (2006), contains updates to gas accretion, supernovae feedback, AGN feedback, treatment of gas in satellite galaxies, galaxy mergers, and build up of intra-cluster stars. The model is fit to the $z = 0$ stellar mass function, cold gas fraction, stellar metallicity-stellar mass relation, baryonic Tully-Fisher relation, and black hole-bulge mass relation.

1.2.4 Insights from simulations

A useful byproduct of simulations is that we can produce mock data, i.e. translate simulated data in to fake observations for which we know both observable parameters (e.g. magnitudes, images, spectra), and also the true parameters we are interested in (e.g. stellar mass, star formation history, halo mass). If we want to recover these parameters for the real universe, we can test how well our methods work on mock data and gain insight as to how accurate our methods are. We describe this in section 3.3.2.

In addition to being useful for testing our methods, we can learn much about the specifics of galaxy evolution from simulations. In particular, by running models with different initial conditions or physical prescriptions, and comparing the resulting measurements to the real universe, we can constrain parameters that are difficult to extract from purely observational data, for example, the methods of star formation quenching and their importance (Wild et al., 2009; Gabor et al., 2010; Zolotov et al., 2015; Pallero et al., 2018), the importance of stellar and AGN feedback (Aumer et al., 2013; Agertz et al., 2013; Stinson et al., 2013; Hopkins et al., 2014), the feedback between barons and dark matter (Duffy et al., 2010; Governato et al., 2012), the formation of galactic structure (Guedes et al., 2011; Aumer et al., 2014; Crain et al., 2015), the rate and effects of galaxy mergers (White, 1978; Lacey & Cole, 1994; Springel et al., 2005b; Qu et al., 2017), and importance of assembly bias (see section 1.1.3).

1.3 Galaxy surveys and measuring observables

Galaxy surveys have improved dramatically since Hubble (1926), and we can now collect data for millions of galaxies, over a broad range of physical parameters. Galaxy surveys can be broadly split in to two categories: *photometric surveys*, where images are taken in a number of different bands, and *spectroscopic surveys*, where optical fibres are allocated to galaxies, and a spectrum is taken of the central region of the galaxy. More recently, another type of survey known as Integral Field Unit (IFU) surveys has become possible, for example the MANGA survey (Bundy et al., 2015), where multiple spectra are taken at different points across a galaxy. This allows for a detailed look at how properties vary across a galaxy, but is currently not possible for large numbers of galaxies out to high redshifts. In this section, we outline both photometric and spectrosopic surveys, including the specific surveys used in this study, and describe how some common observables that surveys seek to measure are computed.

1.3.1 Photometric surveys

Photometric surveys are surveys where entire regions of sky are imaged in a number of different photometric bands. Surveys are generally categorised by the wavelengths they observe in, for example, ultra-violet (UV), optical, infra-red (IR), or radio. With photometry, we can investigate galaxy morphology, angular clustering, and also analyse colour and magnitudes in order to gain insight in to the physical processes occurring within. More complex measurements, for example redshifts, stellar masses, or star formation histories are more difficult to compute from photometry, however perform better with narrow photometric bands and a large wavelegth range of observations (Abdalla et al., 2011; Mitchell et al., 2013). A big advantage of photometric surveys is that we can collect data for all galaxies down to a survey magnitude limit, so can analyse galaxy populations with well understood completeness.

The Sloan Digital Sky Survey (York et al., 2000; Gunn et al., 1998) is one of the most widely used photometric surveys, imaging over 14,000 deg^2 of sky in both the north and south galactic cap. It has produced photometry for $\sim$ 260,000,000 stars and $\sim$ 208,000,000 galaxies in the u, g, r, i, and z bands (near UV to near IR). The magnitude limits in each band, defined as the 95% completeness limit for point sources (i.e. the magnitude at which 95% of objects will be detected) are as follows: $u = 22.0$, $g = 22.2$, $r = 22.2$, $i = 21.3$, $z = 20.5$. This allows for photometry to be measured for the brightest galaxies out to $z \approx 0.7$. In this thesis, we use data from data release 8 (Aihara et al., 2011) and describe our specific sample selection in section 3.2.2.

Multiple current and future surveys will extend the work of the SDSS to deeper magnitdes, and high image quality. One example of this is the DESI Legacy Imaging Survey (Dey et al., 2019), comprising of three public projects: the Dark Energy Camera Legacy Survey (DECaLS), the BeijingâĂŞArizona Sky Survey (BASS), and the Mayall z-band Legacy Survey. These surveys will jointly image ∼ 14,000 degrees of sky in the northern hemesphere in the g, r and z bands. Some other examples include the dark energy survey (DES) (DES Collaboration et al., 2017), imaging ∼ 5,000 degrees of sky in g, r, i, z, and Y, the Hyper Suprime-Cam Subaru Strategic Program (HSC) (Aihara et al., 2018), measuring deeper photometry in similar photometric bands, but over a smaller area of sky, and also EUCLID (Laureijs et al., 2011), which will obtain spectra alongside high quality single-band photometry from space.

1.3.2 Spectroscopic surveys

The second broad class of surveys are spectroscopic surveys. Here, galaxies are selected with some criteria from photometry and allocated fibers, which are fed in to a spectrograph to produce spectra, with modern surveys obtaining spectra for hundreds of thousands, or even millions of galaxies. By fitting models to these spectra we can measure the quantities of gas, stars and dust in the galaxy, the stellar mass, star formation history, stellar populations, dynamics of stars and gas, AGN activity, and importantly, measure an accurate redshift for the galaxy. Here, we outline some of the most widely used spectroscopic surveys, including the data used in this thesis.

Building on the 2dF Galaxy Redshift Survey (Colless et al., 2003), which measured 220,000 galaxy spectra over 1500 deg^2, one of the most widely used spectroscopic surveys is the SDSS spectroscopic survey (York et al., 2000). This survey consists of two samples of galaxies containing more than 1 million spectra over roughly 8,000 deg^2 (a sub-sample of the area of the SDSS photometric survey described in secion 1.3.1): the main sample (Strauss et al., 2002) and Luminous Red Galaxy (LRG) sample (Eisenstein et al., 2001). In the main sample, spectra are measured for all galaxies down to a magnitude limit of r=17.77, producing a complete sample of galaxies below this limit. The LRG sample is designed to produce a sample of luminous intrinsically red galaxies, extending fainter and farther than the main sample, however these are selected using a much more complex set of magnitude cuts.

GAMA (Driver et al., 2009) is a survey completed more recently, which despite covering a smaller area of 286 deg^2, extends much deeper to $r < 19.8$ mag, imaging roughly 300,000 galaxies. GAMA is designed primarily to study structure on scales of 1 kpc to 1 Mpc, including galaxy clusters, groups, and mergers, along with fundamental observables like the merger rate and stellar mass function. Due to the fact that regions of sky were revisited many times, the sample of galaxies is highly complete. We use data from the GAMA survey in chapters 3 and 4.

The BOSS (Eisenstein et al., 2011; Dawson et al., 2013; Gunn et al., 2006; Smee et al., 2013) and eBOSS (Blanton et al., 2017; Dawson et al., 2016; Gunn et al., 2006; Smee et al., 2013) surveys are more recent iterations of the SDSS. Differently to the SDSS main and GAMA surveys, rather than being primarily interested in measuring galaxy properties for magnitude limited samples, these surveys focus on obtaining spectra (and redshifts) for large numbers of massive, bright galaxies. Because of this, these surveys rely on a number of complex colour and magnitude cuts, leading to a sample which is not complete (i.e. a sample which does not target all galaxies of a given magnitude or stellar mass). These bright, highly biased (i.e. strongly clustered), samples are chosen for for the purpose of measuring the baryon-acoustic oscillation length scale in order to constrain the expansion history of the universe.

BOSS has measured spectra for ~ 1.5 million luminous red galaxies (LRGs) out to $z = 0.7$, over roughly $10,000$ deg^2 of sky. eBOSS is the extension of this program, and has measured $\sim 375,000$ LRGs (Prakash et al., 2016) at $0.7 < z < 0.8$, and $\sim 740,000$ quasars (QSOs) (Myers et al., 2015) over the range $0.9 < z < 3.5$ both over $7,500$ deg^2 of sky. More recently, it has also measured spectra for roughly $200,000$ emission line galaxies (ELGs). In this thesis, we use data from BOSS DR12 (Alam et al., 2015) LRGs in the North Galactic Cap (NGC) covering 6851 deg^2 and eBOSS DR14 data (Abolfathi et al., 2018), containing both the LRG and quasar samples, covering 1011 and 1214 deg^2 respectively, within the BOSS NGC area. Our exact selection from this data is described in section 3.2.1.

Two future large-volume spectroscopic surveys are DESI (DESI Collaboration et al., 2016) and Euclid (Laureijs et al., 2011). Extending the work of the BOSS and eBOSS surveys, DESI (DESI Collaboration et al., 2016) will obtain spectra for LRGs up to $z = 1.0$, ELGs up to $z = 1.7$, and QSOs up to z = 3.5 for the purposes of measuring BAO and the growth of structure. To maximise efficiency during bright time, DESI will also conduct a magnitude-limited Bright Galaxy Survey(BGS) obtaining spectra for an additional ~ 10 million galaxies, leading to $\sim$ 30 million total galaxy and quasar redshifts. Euclid, will, alongside measuring photometry, produce grism spectroscopy for ~ 50 million galaxies.

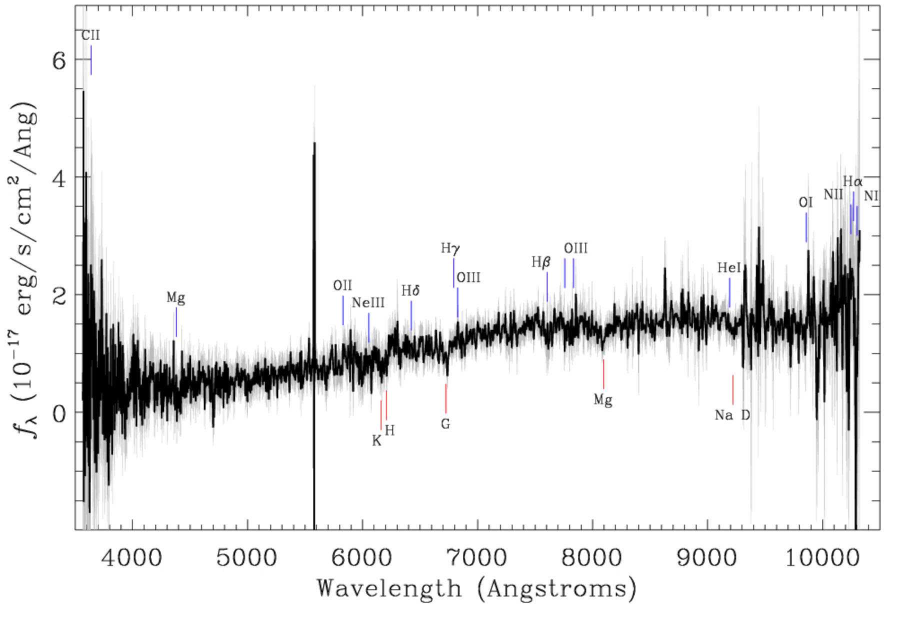

Figure 1.5: The spectrum of a randomly selected galaxy from the BOSS survey Dawson et al. (2013), showing absorption (red) and emission (blue) lines. The galaxy redshift of $z = 0.56429 + / - 0.00021$ is computed from the shifted lines. Figure created in the SDSS SkyServer tool.

1.3.3 Redshift estimates

One of the most fundamental parameters we can measure for a galaxy is its distance from us. This measurement allows us to understand the physical size and environment of the galaxy, compare galaxy populations across cosmic time, and analyse galaxy magnitudes and spectra in their rest-frame wavelength. Since distances are difficult to measure for large numbers of galaxies, a galaxy's redshift is normally considered instead, computed by measuring the shift in emission or absorption lines in galaxy spectra. An example spectrum is shown for a BOSS LRG at $z = 0.56$, showing a number of shifted emission and absorption lines. Spectroscopic follow up of photometric objects, however, is only possible for a small fraction of galaxies, so when analysing purely photometric data, photometric redshifts must be estimated instead.

Sánchez et al. (2014) present a detailed comparison of a number of different photometric redshift codes. Here, we summarise the main methods. Most methods apply one of two techniques: One can compare a galaxy's broad-band colours to a library of spectral energy distributions (SEDs), containing a representative sample of galaxy templates, and convolve these with filters at different redshifts. The best fit can then be computed to give the galaxy redshift. Some popular implementations of this are *hyperz* (Bolzonella et al., 2000), *BPZ* (Benítez, 2000) and *LePhare* (Ilbert et al., 2006). Alternatively one can use a spectroscopic sample of galaxies to train a machine learning algorithm to find relations between broad-band colours and redshift, which can then be used to estimate the redshift of the object, for example in *ANNz* (Collister & Lahav, 2004), and *ArborZ* (Gerdes et al., 2010a) .

For high quality photometric data, photometric redshifts are often very accurate, however, for particularly dim or high redshift objects, the quality of redshift estimation descreases. Fitting based methods are reliant on SED libraries, so if only a few photometric bands are present, or if the templates are not representative of the galaxy in question, an incorrect galaxy SED at the wrong redshift may produce the best fit. Similarly, machine learning methods are reliant on the training sample, and if the sample does not cover the entire range of magnitudes and redshifts, then the algorithm may also incorrectly allocate a redshift. Both methods are also reliant on the quality of photometry, which at faint magnitudes can have large corresponding errors. In chapter 2, we discuss an alternative method of computing redshifts, or redshift distributions, from clustering information rather than photometry.

1.3.4 Mass functions

The stellar mass of a galaxy is another incredibly useful property since it represents the product of the star formation and merger history of a galaxy. For this reason, the stellar mass function, $\Phi(M_\star)$, is a property which many surveys seek to measure. It represents the comoving number density of galaxies as a function of mass at a particular redshift, hence is useful for testing our models of galaxy evolution (e.g. how does the number density of galaxies of different masses evolve? How do galaxies build up mass? And how important are mergers for different types of galaxies?). Furthermore, it is also important as a constraint which our simulations must reproduce.

Stellar mass can be derived from fitting models to either spectra or high quality multi-band photometry. This is done by summing up the SEDs of different aged stellar populations, known as Simple Stellar Populations (SSPs) (e.g. Bruzual & Charlot (2003); Maraston & Strömbäck (2011)), and comparing this to the observed SED. The best-fit combination of SSPs, along with redshift, dust extinction, metalicity etc., can be then used to derive galaxy parameters, for example SFH or stellar mass. Due to the large number of possible solutions, parameterisations are are often used for some properties, particularly if data is of low-quality. So, for example, rather than fitting for any possible SFH, a set of reasonable models are often chosen before fitting. The choice of SSPs, parameterisations, and fitting method can have a large effect on the recovered parameters. We investigate these offsets for different stellar mass estimates of BOSS galaxies in section 3.6.

As discussed in section 1.1.3, the local mass function is relatively well constrained even down to low masses (Cole et al., 2001; Bell et al., 2003; Panter et al., 2007; Baldry et al., 2008, 2012). At higher redshifts, measuring the mass function is more difficult, since galaxies are fainter, however many surveys have measured the evolution of the mass function across wide redshift ranges (Pérez-González et al., 2008; Moustakas et al., 2013; Muzzin et al., 2013; Leauthaud et al., 2016; Capozzi et al., 2017; Guo et al., 2018). The number of massive galaxies is typically found to be fairly constant between $0 < z < 1$, and drops towards higher redshifts, implying these galaxies form early, and don't evolve in mass significantly at late times. Similar trends are true for lower mass galaxies, however the number density of galaxies appears to keep increasing even at lower redshifts, implying these galaxies keep building mass later on. Due to the difficulty of obtaining high quality spectra or photometry out to high redshift, the surveys used for these studies are typically narrow, so can be strongly affected by sample variance. In chapter 3 we investigate a technique for computing mass functions using SDSS photometric data (covering $\sim \frac{1}{4}$ of the sky), allowing us to produce mass functions with little sample variance.

1.3.5 Luminosity functions

Another commonly measured observable quantity is the luminosity function. Similarly to the stellar mass function, this represents the comoving number density of galaxies as a function of luminosity in a particular photometric band (i.e. absolute magnitude). An advantage of using luminosity functions rather than stellar mass functions is that the computed luminosities are generally less model dependent than stellar masses. This is because the luminosity in a particular band is simply the rest-frame SED of the galaxy integrated over the extent of the particular photometric bands, and little assumptions are needed as to the stellar populations, gas, dust, etc. in a galaxy.

Due to their relative simplicity, luminosity functions are often computed using photometric surveys. Redshifts can be computed photometrically, and then the absolute magnitude in a particular band can be computed given a cosmological model. When comparing galaxies at different redshifts however, the computed absolute magnitudes in the same band will be covering different parts of the rest-frame SED. To make sure the same wavelengths are being considered, a correction to the absolute magnitude must be made, known as a k-correction. If fitting a galaxy SED to photometry, the correction can be computed by extrapolating the best-fit rest-frame SED to the redshift of the observed band, although often, tables of these corrections are computed for samples of low redshift spectroscopic data which can then be applied to galaxies as a function of observed magnitudes and redshift. These methods can be sometimes problematic however, for galaxies at high redshifts, where the rest-frame magnitude is shifted by large amounts from the observed bands, and the k-correction is more significant.

1.4 This thesis

In this thesis, we investigate a number of different observational probes of galaxy evolution in large volume surveys. In chapter 2, we outline a method of obtaining redshift distributions for photometric data using galaxy clustering known as clustering redshifts. We outline the theory and test this in both observed and simulated data. In chapter 3, we extend this work, applying the method to photometric data from the SDSS, and after recovering redshift distributions, we compute mass and luminosity functions for the sample, and use these mass functions to estimate the stellar mass completeness of the BOSS survey. In chapter 4, we investigate how correlated the observable properties of galaxies are with the formation time of their dark matter halo. We train a machine learning algorithm to predict formation time based on observable properties of galaxies. We then apply this to real data from the GAMA survey to investigate if formation time is correlated with environment at fixed halo mass. Finally, in chapter 5, we summarise the findings of this thesis, and outline possible extensions of this work.

2

Clustering Redshifts

The redshift of a particular galaxy is an extremely important property to understand. Once computed, it allows us to compute rest-frame magnitudes and colours, and is needed if we seek to study more complex parameters, for example stellar mass or star formation rate. Additionally, assuming a cosmological model, we can compute the distance to the galaxy, and age of the universe for the observation, allowing us to test our models of galaxy evolution and cosmology across cosmic time.

Measuring redshifts with spectra is fairly straightforward, however if spectra are not available, which is the case for the majority of galaxy data we have, this is significantly more difficult, particularly if photometric data is noisy or consists of few photometric bands. In this chapter, we investigate and implement a method of computing redshift distributions for galaxies, requiring only angular positions, based on the clustering properties of the galaxy sample, known as clustering redshifts.

The layout of this chapter is as follows: In section 2.1 we outline the basics of galaxy clustering and discuss how this can be computed for real data. In section 2.2, we outline the clustering redshifts method used in this thesis. In section 2.3 we present the data samples we use to test our implementation. In section 2.4 we investigate the performance of the technique on spectroscopic data. In section 2.5 we extend this to mock data, also investigating the effect of bias evolution on recovered distributions. Finally, in section 2.6 we summarise our findings. Part of this chapter is presented in Bates et al. (2019).

2.1 Galaxy clustering

The positions of galaxies are not randomly distributed in the sky, but clustered together around highly dense regions of the universe. In the ΛCDM model, this clustering is primarily dictated by the clustering of dark matter. Since the galaxies in halos of different environments undergo very different physical processes, understanding the clustering of a particular population of galaxies can be very useful for understanding these populations. In order to properly analyse galaxy clustering, we must be able to measure this quantitatively; this is most often done using the galaxy correlation function.

2.1.1 Correlation functions overview

The correlation function computes the clustering amplitude of a population of galaxies as a function of separation. Most frequently, the two-point correlation function is used, which considers separation between pairs of galaxies, and can be computed for a single galaxy sample as the *autocorrelation* function, or for two galaxy samples with respect to each other as the *crosscorrelation* function

The *autocorrelation* function describes the clustering amplitude of a single population of galaxies, and is denoted as $\xi(r)$ in Euclidian co-ordinates (or $w(\theta)$ in angular space). More rigorously, the function, $\xi(r)$, describes the probability dP of finding a galaxy in a volume dV, at a separation of r from another galaxy, when compared with a set of randomly distributed points,

$$dP = n(1+\xi(r))dV \tag{2.1}$$

where n is the number density of the population (Peebles et al., 1980). One can take a sample of galaxies (often referred to as the “data catalogue”), and measure the distance from every galaxy to every other galaxy, to get a distribution of galaxy pair separations, $DD(r)$. Then the equivalent statistic, $RR(r)$, can be computed for a set of randomly distributed points over the same area, known as the "random catalogue". The simplest estimator for the correlation function is then the Peebles & Hauser (1974) estimator,

$$\xi(r) = \frac{DD(r)}{RR(r)} - 1 \tag{2.2}$$

where DD and RR are normalised with regards to the number of objects in the catalogues, accounting for different sized data and random catalogues, following $\xi(r) = (\frac{n_r}{n_d})^2 \cdot (\frac{DD}{RR}) - 1$, where n_d and n_r are the number of objects in the data and random catalogues respectively. It can be seen from this estimator that the correlation function represents how many more data-data pairs there are than random-random pairs as a function of distance.

Another popular estimator is the Davis & Peebles (1983) estimator, defined as, $\xi(r) = \frac{DD(r)}{DR(r)} - 1$, where DR is the distribution of data-random pair separations. Probably the most widely used estimator currently is the Landy & Szalay (1993) estimator,

$$\xi(r) = \frac{DD(r) - 2DR(r) + RR(r)}{RR(r)} \tag{2.3}$$

and although little difference is found between different estimators, the Landy & Szalay estimator has been shown to produce smaller variance in many cases (Landy & Szalay, 1993; Vargas-Magaña et al., 2013).

While the *autocorrelation* function describes the clustering of a single galaxy population, it is also possible to define a *crosscorrelation* function, which instead considers how two populations cluster with respect to each other. So instead of computing $DD(r)$, one would compute $D_1D_2(r)$, the distribution of cross-pair distances between galaxies in dataset 1, and galaxies in the dataset 2. The Landy & Szalay (1993) estimator can then be modified as,

$$\xi(r) = \frac{1}{R_1R_2(r)}(D_1D_2(r) - D_1R_2(r) - R_1D_2(r) + R_1R_2(r)) \tag{2.4}$$

The crosscorrelation then describes how likely you are to find pairs of galaxies across two samples, compared with two sets of randomly distributed points.

2.1.2 Types of correlation function

There are a number of different commonly used types of correlation function. The simplest of these is the angular correlation function, $\omega(\theta)$, which considers only angular separation of galaxies, and hence only requires angular positions. This can therefore be done for any sample of galaxies taken from photometry, however all clustering information is integrated along the line of sight, so clustering along the line-of-sight is ignored.

Another commonly used correlation function is the real space correlation function, $\xi(r)$, which instead considers 3-d distance between galaxy pairs. A redshift, and also angular positions, are needed for this, however a measurement in 3-d means that unassociated structure along the line of sight is accounted for (e.g. galaxy pairs with small angular separation, but large redshift separation). Because of this, the resulting measurement considers clustering over only the physical scales chosen.

The redshift of a galaxy is dependent on both its recession velocity due to the expansion of the universe, and also its peculiar motion due to gravitational effects due to its local environment. The combined velocity is often represented at low redshift as, $v_r = H_0 D + v_p$, where H_0 is the $z = 0$ Hubble constant, D is the distance to the object, and v_p is the peculiar velocity.

When considering galaxies in redshift-space, these peculiar velocities cause a squashing or elongation along the line-of-sight, known as Redshift-Space Distortions (RSD), which can manifest itself in two ways: one of those is the Fingers of God (FOG) effect. This is observable on small scales, and is due to the fact that galaxies are often bound in structures, for example galaxy clusters, so will have large peculiar velocities due to their motion around the cluster centre. This causes an elongation of the galaxy distribution in redshift space. The second effect is known as the Kaiser effect, and is due to the in-fall of galaxies towards high density regions. This occurs on larger scales, and is generally smaller than the FOG effect (see Coil (2013) for further discussion of these effects).

Since these effects only occur in the line-of-sight, instead of measuring $\xi(r)$, the correlation function is often measured as a function of both line-of-sight and perpendicular velocity separately, $\xi(r_p, \pi)$. This two-dimentional correlation function is shown for the BOSS survey in figure 2.1. The FOG and Kaiser effect are both visible. RSDs can provide important cos-

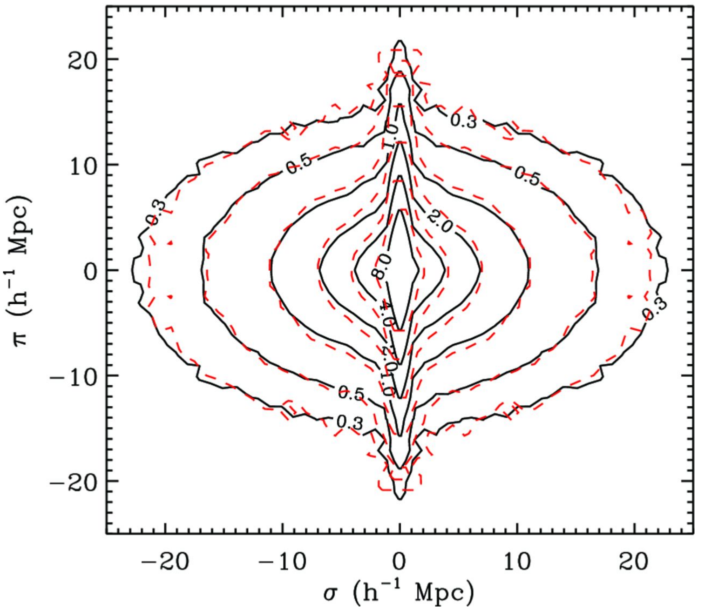

Figure 2.1: Contours of the two-dimensional autocorrelation function $\xi(\sigma, \pi)$ for DR9 BOSS-CMASS north galaxies (dashed red contours) at $0.4 < z < 0.7$. The same measurement for mock data is shown in black. The FOG effect is clearly visible near small values of σ ($\lesssim 2\ h^{-1}Mpc$), and Kaiser effect can be seen stretching the contours along the x-axis at larger scales ($\gtrsim 2\ h^{-1}Mpc$). Figure taken from Nuza et al. (2013).

mological constraints (e.g. Beutler et al. (2014); White et al. (2015)), however, often these effects are undesired. In this case, the correlation function can be integrated between some line-of-sight limits, i.e. $\int_{\pi_1}^{\pi_2} \xi(r_p, \pi) d\pi$, chosen to limit the effects of RSDs. This leads to the projected correlation function, $\omega(r_p)$.

Since the spatial clustering of observable galaxies does not precisely match the clustering of all matter in the Universe, galaxies (and also dark matter halos) are often described as being "biased" tracers of the overall matter distribution. Because of this, the *linear bias*, $b = (\xi_{gal}/\xi_{dark\,matter})^{1/2}$, of a sample of galaxies is often computed (Kaiser, 1984; Bardeen et al., 1986), effectively representing the relative amplitude of the galaxy correlation function compared with the overall matter distribution. Bias is therefore a function of scale, however is shown to be fairly scale-dependent on large scales ($> 1Mpc$) (Mann et al., 1998). It is also shown to depend strongly on the mass of the dark matter halo, as well as the epoch of galaxy formation (Mo & White, 1996), and is known to be larger at early times, but will tend to 1 as time progresses (Fry, 1996; Tegmark & Peebles, 1998). Often, the relative linear bias b_1/b_2 is computed between two samples of galaxies.

2.1.3 Additional considerations

Due to the imperfect nature of real data, there are a number of considerations when making clustering measurements. One such consideration is that the sky footprint of a galaxy survey is normally not a perfectly regular shape, and regions are frequently obscured by stars and other objects. Furthermore, the completeness (i.e. the expected number of observed galaxies over a region of the sky) can vary throughout the survey footprint due to imaging depth, target selection, or overlap of spectroscopic plates. Because of this, when computing correlation functions, the random catalogue (producing the RR(r) term) must possess these features too. The catalogue is therefore produced by creating random points with the same footprint, completeness, redshift distribution, and obscured regions as the survey itself. These random catalogues contain many more times (5-20x) the number of points of the data sample to ensure that when computing correlation functions, most of the statistical variance is coming from the data.

For creating random catalogues from both real photometric data, and simulated photometric and spectroscopic data, we use the MANGLE software (Swanson et al., 2008). This software allows us to input a mask describing the survey geometry and completeness, along with additional masks describing excluded regions, and create random points (in angular space) accordingly. When computing random catalogues for spectroscopic data, we must also allocate redshifts to the random points. To do this, we allocate random redshifts to these galaxies from the data sample, such that the resulting random catalogue will have the same redshift distribution as the data. This is shown to produce less biases than other methods of allocating redshifts (for example fitting a spline to the observed redshift distribution and drawing redshifts from this) (Ross et al., 2012).

In order to compute a correlation function, the $DD(r)$, $DR(r)$ and $RR(r)$ terms must be computed as in section 2.1.1. This relies upon computing the distance (either angular or physical) from every point to every other point, meaning computational time scales with $\mathcal{O}(n^2)$, where n represents the number of data points in the sample. Correlation functions are often computed for many 100s of thousands of galaxies, with random catalogues 10 or more times larger than this, so this can be extremely computationally expensive, particularly when the $RR(r)$ term.

If the correlation function is being computed over small scales, then computing the distance to every far away galaxy is unnecessary. It is then possible to apply a tree-based method (for example a k-d tree), where galaxies are grouped in to different regions known as branches or leafs, based on their positions. When computing distances, regions which are further away than the scales of interest can be ignored, meaning correlation functions can be sped up by orders of magnitude. For studies where large scales are of interest (for example BAO studies), the speed-up is not significant, since the majority of pair distances must be computed, however in this study we are mostly interested in smaller scales, so apply a tree based method, showing orders of magnitude of speed-up.

One additional thing to consider when computing clustering measurements are errorbars. The simplest form of errorbar is a Poisson errorbar. This can be computed by assuming that the error in $DD(r)$ follows $\sqrt{DD(r)}$. In reality, errors do not follow a Poisson distribution since, if a galaxy lies a certain distance away from another, it is quite likely that another galaxy will be at the same distance (i.e. if those galaxies are part of the same cluster). For this reason, Poisson errorbars underestimate the true error, particularly at large scales. An alternative to this is measuring a bootstrap or jackknife error (see Nikoloudakis et al. (2013) for a more detailed outline of these methods).

Bootstrap errors are computed by dividing the galaxy sample in to a number of similar sized areas of sky (e.g. 16), and computing the correlation function, $\omega(\theta)_i$, for each subsample. The error is then computed from the variance of these measurements, (e.g. $\sigma^2_{boot} = \frac{1}{N-1}\sum_{i=1}^{N}[\omega(\theta)_i - \omega(\theta)_{tot}]^2$, where N is the number of subfields and ω_{tot} is the correlation function across the full field). Jackknife errors are computed similarly, however samples are removed in turn, and the correlation function, $\omega(\theta)_i$, of the rest of the sample is computed (e.g. for 15/16 of the sample). This is done for all subsamples and the error is again computed from the variance of these measurements (equivalently, $\sigma^2_{jack} = \sum_{i=1}^{N}[\omega(\theta)_i - \omega(\theta)_{tot}]^2$).

In our preliminary tests (for the samples and scales described in section 2.4), we find that jackknife and bootstrap methods produce similar sized errors, which are both larger than those produced through Poisson estimates (particularly at large scales), as also seen in Nikoloudakis et al. (2013). We opt to use a bootstrap method for this thesis, since jackknife errors require computing the correlation function many times for a sample of similar size to the full sample. Bootstrapping, conversely computes the correlation function for several small samples so performs significantly quicker than jackknife for our simple implementation (although more complex implementations may increase the speed of jackknife methods).

2.1.4 Weights for spectroscopic data

When dealing with spectroscopic data, additional considerations must be taken in order to remove biases in clustering measurements. We outline the main considerations for the BOSS and eBOSS surveys, as this is the spectroscopic data we are using in this chapter and chapter 3, however many of these biases are present in other spectroscopic surveys. A detailed discussion of these different biases is presented in Anderson et al. (2012), however, here, we outline the information relevant for this study.

Spectra for BOSS and eBOSS are obtained using fibres, drilled in to metal plates to match the corresponding galaxy position in the sky. The targets are chosen to maximise the number of galaxies that are allocated a fibre, however, since the diameter of the fibres and holes extend beyond their collecting area, when two galaxies are close in the sky ($\Delta\theta < 0.017$ degrees), it is not possible to allocate both galaxies a fibre, and only one galaxy can be observed. This is known as a fibre collision. Although some areas of sky are observed more than once to minimise this, a number of galaxies are never observed, particularly in highly dense regions. If ignored, this would create a bias in clustering measurements, meaning one would measure too few pairs at small scales. To get around this problem, weights must be added to galaxies. In the BOSS and eBOSS surveys, these are referred to as close pair weights, w_{cp}, which are set to 1 for all galaxies, and when a galaxy cannot be observed due to fibre collisions, the nearest galaxy is up-weighted by 1.

Ross et al. (2011) examine of the large scale angular clustering of CMASS galaxies, finding that the density of stars has a significant effect on the observed density of galaxies. This can introduce additional biases in clustering measurements, which again, must be corrected for. Additional parameters such as galactic extinction, seeing, airmass, and sky background can also produce fluctuations. These can be corrected for by weighting galaxies as a function of the given systematic effect. In BOSS and eBOSS, this is referred to as w_{sys}.

Another problem is the fact that, although the vast majority of galaxies are allocated an object classification and redshift, for some galaxies, spectra are too noisy, and the fitting method fails. These failures are dependent on the particular fibre used, which may have degraded transmission, and also on the region of the CCD, which may vary in optical quality. These fibres are generally allocated to similar positions in the plate and CCD, so the failure rate is a strong function of angular position for each observation. Since redshifts are computed successfully for the majority of galaxies, this effect is fairly small, so is corrected for by up-weighting the nearest object with a successfully computed redshift, similarly to the fibre collision weights. This is referred to as a redshift failure weight, w_{rf}.

In order to optimize clustering measurements in the face of shot-noise and cosmic variance, additional weights can be applied (Feldman et al., 1994), labelled w_{fkp} weights (Feldman et al., 1994). These are computed as, $w_{fkp} = \frac{1}{1+\bar{n}(z_i)P_0}$, where $\bar{n}(z_i)$ is the number density of the sample and P_0 is a constant. Ross et al. (2012) find that this reduces the variance in the BOSS DR9 correlation function by 20%. Full details of this are presented in Anderson et al. (2012).

To take in to account all weights simultaneously, weights are combined using the following formulae:

$$w_{tot} = w_{fkp} \cdot w_{sys} \cdot (w_{rf} + w_{cp} - 1) \tag{2.5}$$

After allocating weights to all galaxies, when computing galaxy pairs (e.g. for the $DD(r)$

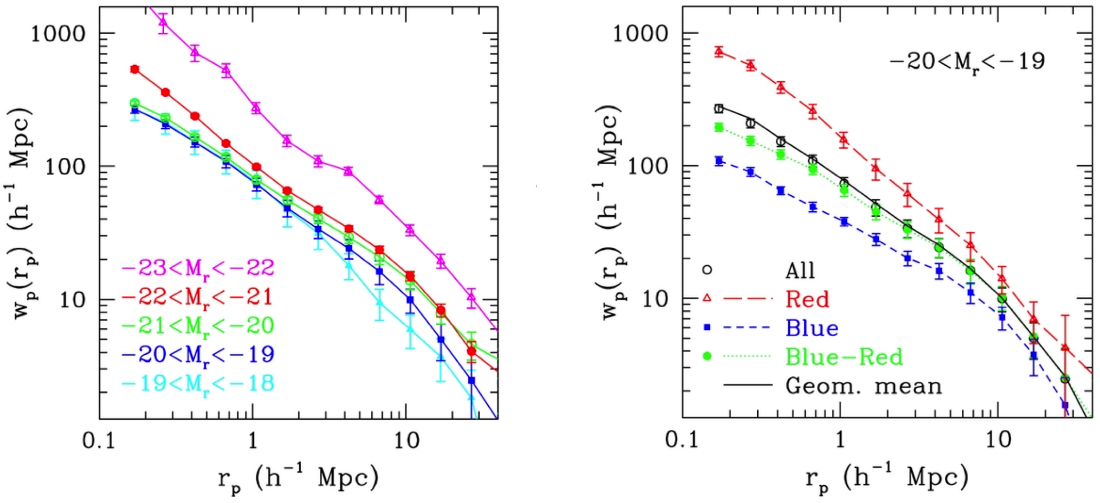

Figure 2.2: Left: Projected autocorrelation functions for SDSS DR7 data binned by k-corrected r-band absolute magnitude. Right: Projected autocorrelation functions in the $20 < M_r < 19$ for blue galaxies, red galaxies, and a sample containing both (all), along with the crosscorrelation between the two samples (blue-red). Both figures are from Zehavi et al. (2005)

or $RR(r)$ term in equation 2.2), pairs are weighted by the product of w_{tot} for both galaxies. Since the random catalogue is already defined to match the observed data, w_{sys}, w_{rf}, and w_{cp} are accounted for automatically, however w_{fkp} must be computed for the random catalogue also.

2.1.5 Clustering properties of galaxies

The clustering properties of galaxies have long been known to depend on their observable properties, such as magnitude, colour, and morphology (Hubble, 1936; Davis & Geller, 1999; Davis et al., 1988; Norberg et al., 1988; Zehavi et al., 2005; Swanson et al., 2008). Figure 2.2 (left) shows the correlation function measured for galaxies in a number of bins of r-band absolute magnitude. It can be seen that the brighter magnitude bins possess a higher correlation function amplitude, reflecting the fact that these galaxies are preferentially present in higher density regions. This is also true for galaxies with bulge-like morphologies, old stellar populations and red colours, as shown in figure 2.2 (right). This follows the galaxy bimodality (discussed in section 1.1.3), showing that red, bulge-dominated galaxies are higher mass, have less star formation, and are generally found in dense environments.

It is often common to think of this trend as an effect of halo mass. Galaxies of high luminosity or stellar mass are more common in higher mass halos, which are known to be more strongly clustered (Kaiser, 1984; Cole & Kaiser, 1989; Kauffmann et al., 1997; Mo & White, 1996; Sheth & Tormen, 1999). Similar trends are also true for morphology and colour. Understating how galaxy properties affect clustering is important, since this underpins both our models of cosmology (i.e. how dark matter halos and cosmic structure forms and evolve) and galaxy evolution (i.e. how do baryons trace the dark matter distribution, and do parameters other than halo mass dictate this?). In section 2.5.1, we investigate how the clustering properties of samples of galaxies in different colour and magnitude bins evolve with redshift. We do this in order that we can use the resulting evolution to account for these effects when computing redshifts from clustering measurements.

2.2 Clustering Redshifts

As discussed in section 1.3.3, accurate redshift estimates are extremely useful for galaxy surveys. With high-quality spectra, redshifts can be accurately measured, however if only broadband photometry is available, photometric redshift estimates must be used instead. These are typically much less reliable, and can present biases dependent on the choice of model SEDs or training samples. In recent years an alternative method of obtaining redshifts has become popular, known as clustering redshifts. This is a method of computing the redshift distribution of a sample of galaxies from its clustering relative to a spectroscopic sample. This has the advantage of being significantly less biased (Rahman et al., 2016a; Scottez et al., 2018), and since it only requires angular positions, is significantly less reliant on the quality of photometry, and can be done for any sample of galaxies. In this section, we outline the clustering redshifts technique.

2.2.1 Background

Crosscorrelations have long been used to test for physical association (Seldner & Peebles, 1976); however, the idea of using crosscorrelations to produce accurate redshift distributions has only become common over the last decade, partly due to the increase in data from large volume spectroscopic and photometric surveys.

Phillipps et al. (1985) investigated determining correlation functions from samples with only partial redshift information; later, in Phillipps & Shanks (1987), luminosity functions are computed given the assumption that galaxies close in the sky are likely at the same redshift.

Schneider et al. (2006) more generally investigate this technique by measuring crosscorrelations with galaxies binned by photometric redshift. This approach is built on more formally in Newman (2008), and later Matthews & Newman (2010) and Matthews & Newman (2012), where a method is outlined for computing redshift distributions by measuring the angular crosscorrelation between a photometric sample and different redshift bins of a spectroscopic sample. The amplitude of the crosscorrelation is fitted by an analytical form; since the crosscorrelation depends not only on the redshift distribution of the photometric sample, but also on the evolution in bias of the two samples, an iterative technique is employed to correct for this, assuming that the evolution of clustering amplitude is proportional in both the spectroscopic and photometric sample.

Some variants on this method have subsequently appeared. For example, Schmidt et al. (2013) and Ménard et al. (2013) propose a similar technique, measuring angular crosscorrelations with a spectroscopic sample, but over constant physical scale. Furthermore, bias evolution is corrected for by assuming a bias evolution law, and the effect of this assumption is tested, down to non-linear scales. More recent studies applying these methods include Rahman et al. (2015), Rahman et al. (2016b), Rahman et al. (2016a), Scottez et al. (2016), and Scottez et al. (2018). van Daalen & White (2018) present a method for computing luminosity functions using clustering information and apparent magnitudes as input.

The biggest difference in methods comes down to how a correction for the bias of the unknown sample of galaxies is applied. Most studies rely on some variant of either the Newman (2008) or Ménard et al. (2013) methods, so we outline the basics of each method in the following sections.

2.2.2 Different methods

Both the Newman (2008) (hereon N08) and Ménard et al. (2013) (hereon M13) methods rely on crosscorrelating a photometric sample of unknown redshifts, i.e. the “unknown” sample, with a sample of galaxies with spectroscopic redshifts, the “reference” sample. This crosscorrelation is performed over many different redshift bins of the reference sample

If the unknown sample overlaps in redshift with one of the bins of the reference sample, they will occupy the same density field, and their positions will be correlated. This crosscorrelation in this bin will therefore have a positive amplitude. The amplitude of crosscorrelation in each bin can be then used to estimate how many galaxies are at each redshift, and hence can produce a redshift distribution for the unknown sample.

The measured crosscorrelation amplitude in each bin, however, depends not only on the number of unknown sample objects at that redshift, but also the intrinsic clustering strength (or bias) of the two samples. The difference between N08 and M13 is how this bias evolution is corrected for in the two samples. We summarise these two methods in the following sections.

2.2.3 Newman method

This method is based around N08, but is expanded upon in Matthews & Newman (2010) (hereon M10) and Matthews & Newman (2012). The crosscorrelation amplitude between the two samples is used to produce a redshift distribution, but bias corrections are computed firstly by measuring the autocorrelations of the unknown and reference samples. We outline these steps in the following sections, however see M10 for full details.

Unknown Sample Autocorrelation

In N08, the intrinsic clustering of the unknown sample is computed by measuring its angular autocorrelation, $\omega_u(\theta)$. Following M10, assuming the real-space correlation function can be fit by the power law, $\xi_u(r) = (\frac{r}{r_{0,u}})^{-\gamma_u}$, where $r_{0,u}$ and γ_u are constants, it is possible to show that the angular autocorrelation will then follow the power law, $\omega_u(\theta) = A_u\theta^{1-\gamma_u} - C_u$, where A_u and C_u are constants. This power law is fit to the measured autocorrelation in order to recover γ_u for the unknown sample.

Reference Autocorrelation

We then need to know how the clustering of the reference sample evolves with redshift. Firstly, after defining redshift bins over which to measure these parameters, the projected correlation function, $\omega_p(r_p)$, is computed in each redshift bin (since this limits the effect of redshift space distortions compared to the 3D version). Following M10 this is computed over projected distances of $1.5 < r_p < 15\ h^{-1}$Mpc, and limit line-of-sight pair distances to $0 < \pi < 30\ h^{-1}$Mpc.

Again modeling $\xi(r)$ as a power law and solving for the projected correlation function $w_p(r_p)$ gives,

$$w_p(r_p) = r_p \left(\frac{r_{0,r}}{r_p} \right)^{\gamma_r} H(\gamma_r)$$

where $H(\gamma)$ is defined as $H(\gamma) = \Gamma(1/2)\ \Gamma((\gamma-1)/2)\ /\ \Gamma(\gamma/2)$, and $\Gamma(x)$ is the standard gamma function. This is then fit to the computed $w_p(r_p)$ for each redshift bin of the reference sample, allowing us to recover $r_{0,rr}$ and γ_{rr} as a function of redshift.

Since we are only computing the correlation function out to $\pi < 30\ h^{-1}$Mpc, and not to infinite line-of-sight separation, a correction must be applied for the fraction of pairs missed. This correction is equal to $\int_0^{\pi_{max}} \xi(r_p, \pi) d\pi\ /\ \int_0^{\infty} \xi(r_p, \pi) d\pi$, which is therefore a function of r_p.

The denominator can simply be computed by substituting the fitted values of $r_{0,rr}$ and γ_{rr} in to the equation $w_p(r_p) = r_p(r_{0,rr}/r_p)^{\gamma_{rr}} H(\gamma_{rr})$, however the numerator must be calculated by numerically integrating $\int_0^{\pi_{max}} \xi[(r_p^2 + \pi^2)^{1/2}] d\pi$ for the previously fitted values of $r_{0,uu}$ and γ_{uu}. The correlation function is divided by this correction, and this fitting and correction is repeated iteratively until a convergence is reached, giving the true $r_{0,uu}$ and γ_{uu} for all redshift bins.

Unknown-Reference Crosscorrelation

Since we now have estimates of the individual clustering properties of the two samples, the angular crosscorrelation, $\omega_{ur}(\theta)$, is measured between the unknown sample, and different redshift bins of the spectroscopic sample. This can be computed over a constant angular scale as in M10, or computed over an angular scale such that θ_{min} and θ_{max} correspond to a particular physical scale at the redshift of the bin in question.

As when computing the unknown sample autocorrelation, the crosscorrelation in each bin can be fitted by the power law $w_{ur}(\theta) = A_{ur}\theta^{1-\gamma_{ur}} - C_{ur}$ to recover the amplitude of the crosscorrelation A_{ur}. M10 find a significant degeneracy between A_{ur} and γ_{ur}, so γ_{ur} is fixed in each bin. Since the clustering of the samples with respect to each other is expected to be intermediate to the clustering of each sample separately, γ_{ur} is estimated as the arithmetic mean of the values of γ_u and γ_r measured previously, and only fit for A_{ur} and C_{ur}.

Computing the Redshift Distribution

In this formalism, the redshift distribution, $\phi(z)$, can be estimated as,

$$\phi(z) = \frac{dl/dz}{D(z)^{1-\gamma_{ur}} H(\gamma_{ur}) r_{0,ur}^{\gamma_{ur}}} A_{ur}(z)$$

where $l(z)$ is the comoving distance to redshift z, $D(z)$ is the angular size distance, and γ_{ur}, A_{ur} and $H(\gamma)$ are defined as before.

In order to compute this, some approximations must be made about the redshift dependence of, $r_{0,u}$, the scale length of the unknown sample (representing the uncertainty in bias evolution of the unknown sample). It is worth noting that the amplitude of this does not affect the shape the recovered redshift distribution, only its amplitude. With this assumed $r_{0,u}$, the linear biasing assumption that $r_{0,ur}^{\gamma_{ur}} = (r_{0,u}^{\gamma_u} r_{0,r}^{\gamma_r})^{1/2}$ is employed, allowing us to compute the redshift distribution, $\phi(z)$, of the unknown sample. This is then normalised by the total number of galaxies in the sample to give the true distribution.

2.2.4 Menard method

The M13 method, similarly, relies upon measuring the unknown-reference sample crosscorrelation in many redshift bins, however, the bias evolution is corrected for differently. The method is detailed in Ménard et al. (2013), however we summarise the important points here.

To produce an $\phi(z)$ measurement (represented as dN/dz in M13), we measure the angular crosscorrelation, $\omega_{ur}(\theta, z)$, between an unknown sample, and different redshift bins of a reference sample. Since we are interested in how the amplitude of this quantity evolves with redshift, we integrate over θ to produce

$$\bar{\omega}_{ur}(z) = \int_{\theta_{min}}^{\theta_{max}} d\theta W(\theta) \omega_{ur}(\theta, z) \tag{2.6}$$

where $W(\theta)$ is the weight function, $W(\theta) = \theta^{-1}$, designed to optimise the signal-to-noise ratio. In order to probe the same physical scale at all redshifts, θ_{min} and θ_{max} are computed differently for each redshift, such that they correspond to the same physical scales $r_{p,min}$ and $r_{p,max}$.

From M13, the integrated crosscorrelation is,

$$\bar{\omega}_{ur}(z) \propto \frac{dN_u}{dz}(z) \bar{b}_u(z) \bar{b}_r(z) \bar{\omega}_{DM}(z) \tag{2.7}$$

where $\frac{dN_u}{dz}(z)$ is the redshift distribution of the unknown sample, $\bar{b}_u(z)$ and $\bar{b}_r(z)$ are the evolution in bias of the unknown and reference samples, respectively, over the same scales, and $\bar{\omega}_{DM}(z)$ is the equivalent evolution in the integrated dark matter correlation function.

The bias evolution of the unknown sample

In order to compute a redshift distribution, we need an estimate of $\bar{b}_u(z)$, $\bar{b}_r(z)$, and $\bar{\omega}_{DM}(z)$. Assuming linear biasing, the integrated autocorrelations of the unknown and reference samples as a function of redshift can be written as $\bar{\omega}_{uu}(z) = \bar{b}_u^2(z)\bar{\omega}_{DM}(z)$ and $\bar{\omega}_{rr}(z) = \bar{b}_r^2(z)\bar{\omega}_{DM}(z)$ respectively. We are able to measure both $\omega_{uu}(z)$ and $\omega_{rr}(z)$, so we can substitute these in to equation 2.7, producing,

$$\frac{dN_u}{dz}(z) \propto \frac{\bar{\omega}_{ur}(z)}{\sqrt{\bar{\omega}_{uu}(z)\bar{\omega}_{rr}(z)}} \tag{2.8}$$

We can measure $\bar{\omega}_{ur}(z)$, the integrated crosscorrelation between the unknown sample and each bin in redshift of the reference sample, and also $\bar{\omega}_{rr}(z)$, the integrated autocorrelation of the reference sample over the same redshift bins and physical scale. We can also remove the constant of proportionality by normalising $\frac{dN_u}{dz}(z)$ to the number of galaxies in the unknown sample. The only parameter we cannot compute is $\bar{\omega}_{uu}(z)$ (again representing the uncertainty in the bias evolution of the unknown sample), since we have no redshift information for the unknown sample.

M13 show in their figure 1 that for a range of assumed bias evolutions of the unknown sample, if the redshift distribution is narrow, $\sigma_z < 0.2$, the effects of bias evolution on the recovered distribution are small, and therefore the distribution can be estimated as $\frac{dN_u}{dz}(z) \propto \bar{\omega}_{ur}(z)$. Some studies, however choose to assume analytic forms for the bias evolution, e.g. M13, Schmidt et al. (2013), and some assume no evolution, and factor any deviation from this into their error budgets (Gatti et al., 2018).

2.2.5 Our method

In Gatti et al. (2018) the performance of three of these methods are investigated: N08, M13, and Schmidt et al. (2013) (based on M13). They apply all methods to simulated Dark Energy Survey (DES) data, finding that the N08 method produces slightly noisier redshift distributions due to having two extra degrees of freedom when fitting the crosscorrelation amplitude;

furthermore, they report that in their tests, the proportional bias assumption is not always accurate for realistic data.

In our preliminary tests, the noise of all techniques was largely due to noise in the cross-correlation functions, and the choice of method made only small differences to the noise of the recovered $n(z)s$. The main difference between methods is how the bias evolution correction is applied. Since, firstly, Gatti et al. (2018) find that the Menard method appears to produce slightly less noisy distributions, and secondly, we will be investigating methods of correcting for bias in later sections, we choose to adopt a method based on Ménard et al. (2013) and apply our own bias correction.

2.3 Data sample

In sections 2.4 and 2.5, we perform two tests of the clustering redshifts method: one on spectroscopic data, and one on mock catalogues from simulations. We describe both the real and mock data used in these tests in this section.

2.3.1 The BOSS sample

Our first test is performed on spectroscopic data. For this we use data from the BOSS survey (Eisenstein et al., 2011; Dawson et al., 2013; Gunn et al., 2006; Smee et al., 2013). The survey is described in more detail in section 1.3.2, obtaining spectra for ~ 1.5 million luminous red galaxies (LRGs) out to $z = 0.7$. Since this sample covers a large sky area and provides continuous, large numbers of redshifts over the range $0 < z < 0.8$, it is ideal to use as a reference sample. Our sample constsists of BOSS DR12 (Alam et al., 2015) LRGs in the North Galactic Cap (NGC) covering 6851 deg^2, and containing 812,000 galaxies. We later remove a sub-sample of this data, recovering its distribution using the remaining galaxies. When computing correlation functions, we use large scale structure catalogues from Reid et al. (2012), which are described in more detail in section 2.1.4.

2.3.2 Mock samples

We also perform tests on mock data, requiring both mock unknown samples, and a mock reference sample. In chapter 3, we apply the clustering redshifts technique to the SDSS photometric survey (York et al., 2000; Gunn et al., 1998) (see section 1.3.1), using the BOSS survey as a reference sample, so here we produce mock surveys with comparable photometry and sample selection to these, such that we can test our method on realistic data.

To produce the samples, we make use of data from the semi-analytic model LGalaxies (Henriques et al., 2015), described in more detail in section 1.2.3, run on the Millennium simulation (Springel et al., 2005a) and rescaled to Planck cosmology (Planck Collaboration et al., 2014). We create mock catalogues by selecting galaxies from a light-cone covering 1/8th of the sky. Galaxy SEDs are computed using Maraston & Strömbäck (2011) stellar population models (SSPs), using a Kroupa et al. (2001) initial mass function (IMF), and convolved with SDSS photometric filters to produce SDSS-like magnitudes. The catalogue produced has angular positions, redshifts, and SDSS magnitudes (u, g, r, i, z) with reddening applied (Calzetti et al., 2000).

We add photometric errors to the magnitudes of the SAM by looking at how the error on a fitted magnitude in the SDSS varies as a function of that magnitude (i.e. g vs σ_g, r vs σ_r, i vs σ_i). For every mock galaxy, we use its magnitude to compute the mean error at this magnitude in SDSS, then draw a Gaussian random error using this value as the standard deviation. We compute errors for all mock galaxies in the g, r and i bands, and add these errors to our mock galaxy magnitudes.

From this simulated galaxy catalogue, we define a mock reference sample and photometric survey. We define our reference sample by applying the colour and magnitude cuts of the BOSS survey described in Dawson et al. (2013). This procedure produces samples with comparable redshift distribution to the BOSS survey (and is further discussed in section 3.5). We refer to this sample as $BOSS_{LGalaxies}$. To create a mock SDSS photometric survey from which we can draw unknown samples, we cut both catalogues to $i < 21$, and also remove all galaxies present in our mock reference sample. We refer to this mock survey as $SDSS_{LGalaxies}$. Random samples (with a size 10x the data) are created for both these mock surveys as described in 2.1.3.

2.4 Tests on spectroscopic data

Firstly, we test the clustering redshifts method on spectroscopic data, where we have accurate real-world redshifts estimates. We use data from the BOSS survey, described in section 2.3.1. Following the sample selection in Maraston et al. (2013), we split the sample in to a “blue” and “red” sample with the magnitude cuts $g - i \leq 2.35$, and $g - i > 2.35$ respectively, leaving us with 140,000 blue sample and 672,000 red sample galaxies. This cut is designed to produce

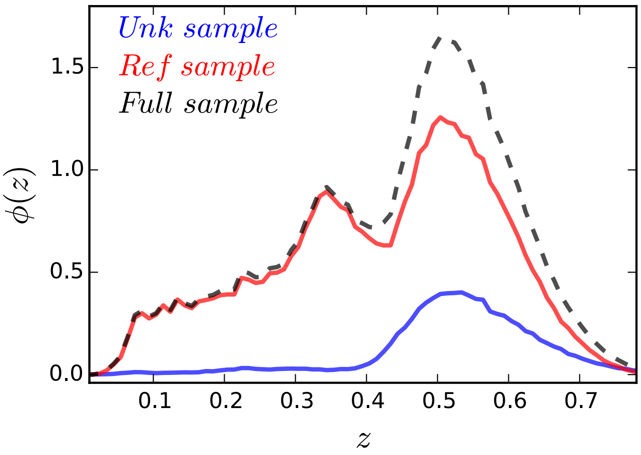

Figure 2.3: The redshift distribution of the different samples used in this test: The unknown sample, consisting of BOSS galaxies with blue colour cuts applied (blue), the reference sample, consisting of the remaining BOSS galaxies (red), and the full BOSS sample (black). Distributions are normalised such that the integral of the distribution of the full sample equals 1.

two samples of galaxies of different star-formation rates, and hence redshift distributions and bias. This allows us to throw away the redshift information for one sample, and recover the redshift distribution using the remaining galaxies as the reference sample. The resulting red, blue, and full samples are shown in figure 2.3.

The blue sample is clearly much smaller than the red sample, with the majority of the sample situated between $0.4 < z < 0.6$. Since the red sample contains large numbers of galaxies across all redshifts, we use this as our test reference sample, and attempt to recover the redshift distribution of the blue sample, our test unknown sample. We apply the clustering redshifts method detailed in section 2.2.4, measuring the bias evolution of the refernce sample through its autocorrelation in a number of redshift bins. We choose redshift bins of size $\Delta z = 0.2$ between $z = 0.02$ and $z = 0.72$ to cover the range of the reference sample. Since in the real universe, we do not know the bias evolution of the unknown sample, we assume a flat bias evolution ($db/dz = 0$) to see how well this performs.

Crosscorrelations are measured between the unknown and reference samples across the same redshift bins of size $\Delta z = 0.2$. We measure these crossorrelations over angles corresponding to $1.5\ Mpc < r_p < 5\ Mpc$ at each redshift. This scale is chosen because, although large-scale clustering is thought to be less dependent on assumptions about the bias evolution of the unknown sample than small scales (Ménard et al., 2013), recovered $\phi(z)$s are significantly less noisy at small scales, mostly due to the fact that there is significantly more clustering power (i.e. a stronger correlation function amplitude). Furthermore, large scales are more susceptible to spurious correlations due to large-scale structure (e.g., chance alignment of structure at different redshifts). We investigate this issue later, in section 2.5.3, however, for now, this scale is chosen since it is the smallest scales we can measure, while ensuring that all scales are above the SDSS fibre-collision limit of $55''$. Correlation function errors are computed using a bootstrap method with 16 subsamples, as described in section 2.1.3.

The measured crosscorrelation functions in each bin are presented in figure 2.4. It can be seen that below $z = 0.4$, crosscorrelations are noisy, and generally of very low amplitude. This follows from the fact that blue sample galaxies are mostly situated above $z = 0.4$ (fig 2.3), hence their positions are uncorrelated with reference sample galaxies at redshifts below this. Above these redshifts, the crosscorrelation amplitude increases, reaching a maximum around roughly $z = 0.53$, where the redshift distribution of the unknown sample peaks. After computing the integrated crosscorrelation using the M13 method, we produce a clustering redshift estimate of the distribution. This is presented in figure 2.5

The clustering redshift estimate follows the spectroscopic distribution very closely. At low redshift, where there are very few unknown sample galaxies, the estimate mostly corresponds to noise in the correlation function, centered around $\phi(z) = 0$. At higher redshifts, where the unknown sample number density is higher, the clustering redshift estimate successfully manages to reproduce the distribution, again with some scatter around the true value.

To demonstrate the effect of unknown sample number density on the recovered redshift distribution, we compute the redshift distribution for a random sub-sample of the blue sample, containing 35,000 galaxies (i.e. 1/4 of the sample), computing the redshift distribution as before, by crosscorrelating with the red sample. The resulting distribution is shown in figure 2.6. The redshift distribution is recovered similarly well, however, scatter around the true value is visibly larger, due to larger clustering errors.

We investigate performing a bootstrapping method for the entire process, rather than just for computing correlation function errors, however, we find, as seen in figure 2.6, that subsampling the data produces very noisy values of $\bar{\omega}_{ur}(z)$. Since the next step requires normalising this integrated correlation function to the number of galaxies in the sample, if $\bar{\omega}_{ur}(z)$ is

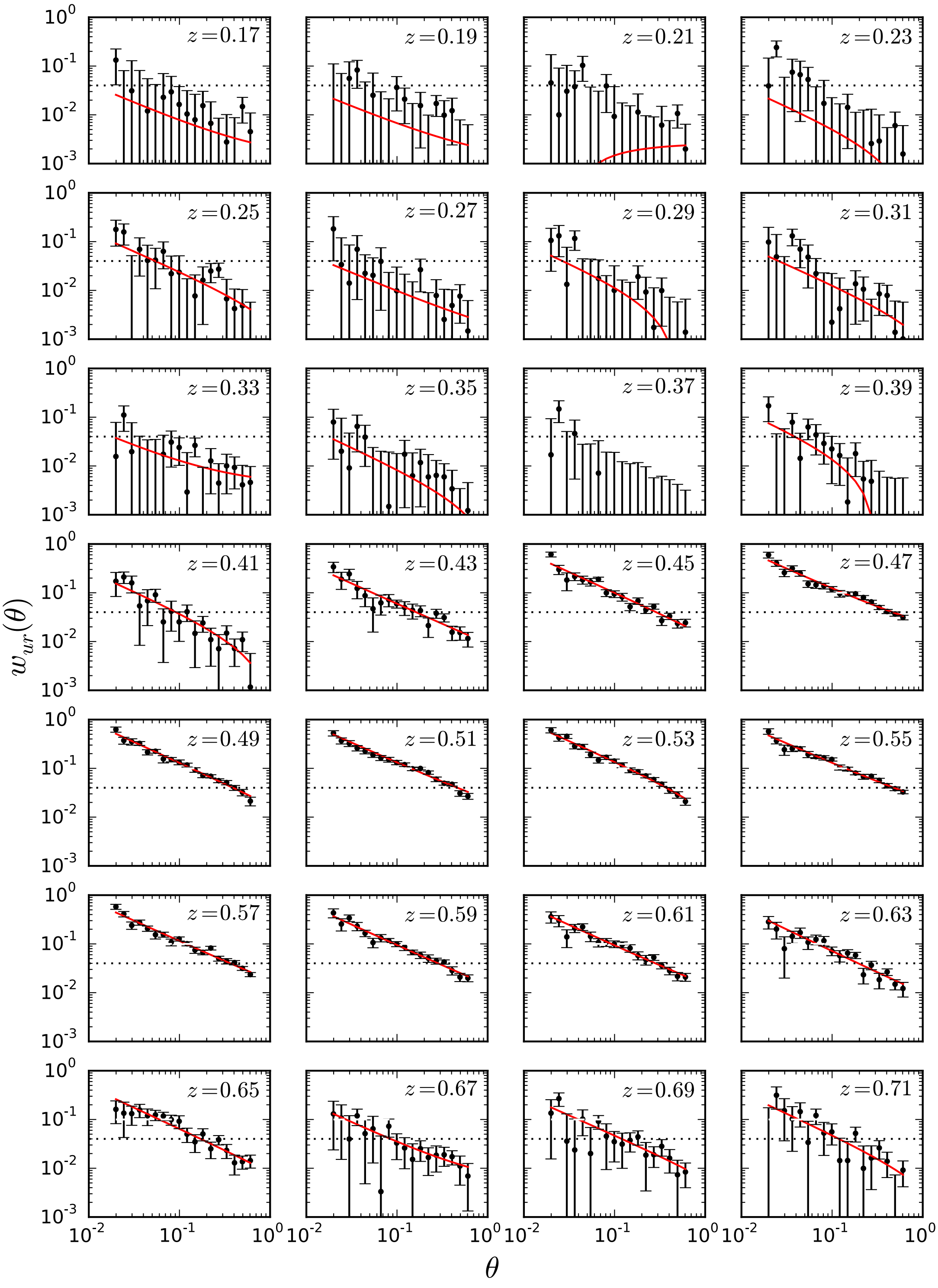

Figure 2.4: The angular crosscorrelation between the unknown sample (a blue subsample of BOSS galaxies), and a number of different redshift bins of the reference sample (the remaining red sample of BOSS), shown as black points. The measured crosscorrelation is shown as black points. For reference, each measurement is fitted with a power law, $w_{ur}(\theta) = A_{ur}\theta^{1-\gamma_{ur}} - C_{ur}$, in red, and the constant value of $w_{ur}(\theta) = 0.05$ is shown as the dotted black line. In the M13 method, this crosscorrelation function is integrated to produce an amplitude, which is used to produce the redshift distribution.

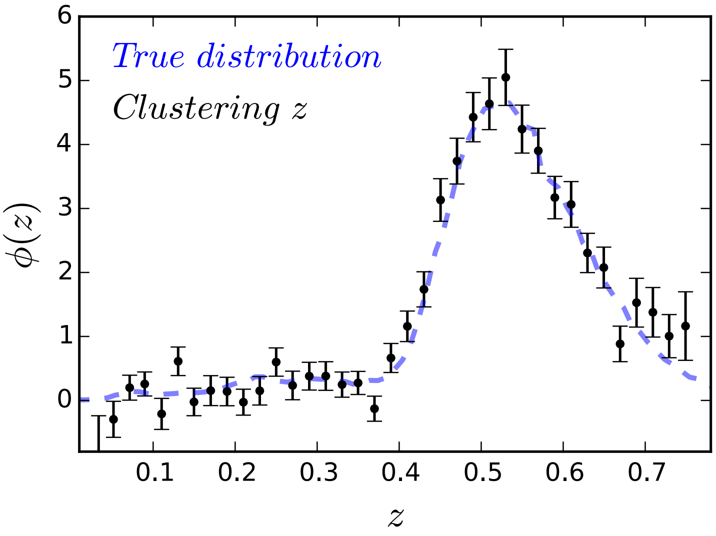

Figure 2.5: The clustering redshifts recovery of a blue sample of BOSS galaxies (black points), using the remaining galaxies as a reference sample, normalised such that the integral of the distribution equals 1. The equivalent "true" spectroscopic redshift distribution of this sample is shown in blue.

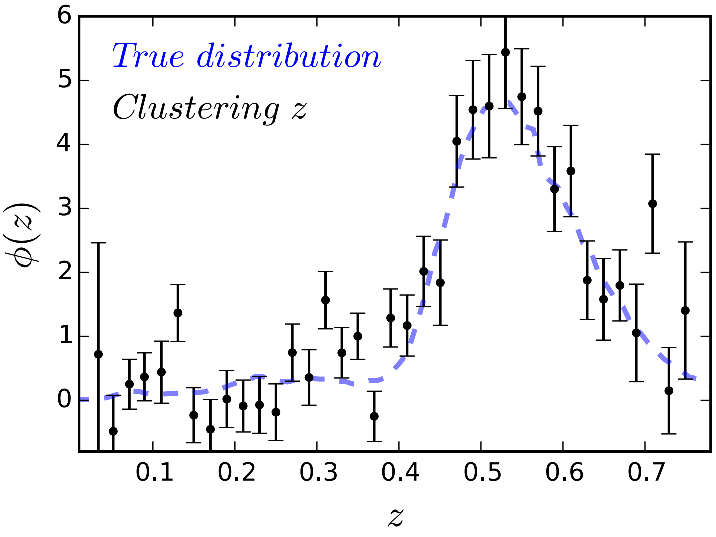

Figure 2.6: The clustering redshifts recovery of 1/4 of the blue sample of BOSS galaxies (black points), as plotted in figure 2.5. Noise in the recovered distribution is visibly larger than for the full sample.

dominated by noise, then its integral can become close to zero or even negative. This means that distributions are often inflated to extremely very large values, and bootstrap errors become very difficult to calculate. For this reason we stick to performing the bootstrapping when computing correlation function errors and propagate these errors through.

In these tests, we are assuming no bias evolution in the unknown sample. This assumption appears to produce an accurate redshift distribution in this test, implying the bias correction is not significant here. Narrow redshift distributions are known to be less affected by the bias evolution assumption (Newman, 2008; Ménard et al., 2013), and from Ménard et al. (2013) fig. 1, it can be seen that for samples of comparable redshift width to the blue sample, one should expect only small errors in the mean recovered redshift ($\Delta z_{mean} \simeq 0.01 - 0.1$), with a steeper bias evolution equating to a larger error. The fact that we observe no significant offset implies the bias evolution of the sample between $z = 0.4$ and $z = 0.7$ is shallow. Since photometric samples of interest in the real universe may possess a very different evolution in bias, or much wider redshift distribution, we next investigate how well redshift distributions can be recovered in simulations.

2.5 Tests on simulations

The advantage of testing our method in simulations is that here we can select many samples of very different magnitude, colour, and bias, for which we know the true redshifts. This allows us to test how the method recovers redshift distributions for very different populations of galaxies. Since in chapter 3, we apply the technique of clustering redshifts to photometric data from the SDSS survey in order to recover useful parameters, we create mock surveys with the same photometry, photometric errors, survey volume and magnitude limit as the SDSS, along with random catalogues. We also create a mock reference sample by applying the colour cuts of the BOSS survey to the data. This process is described in section 2.3.2. We are left with mock catalogues, $SDSS_{LGalaxies}$, and $BOSS_{LGalaxies}$ containing angular positions, magnitudes, and redshifts for all galaxies. We use these two samples to test our method.

2.5.1 Bias evolution in simulations

We plan to apply the methodology described in previous sections to many different magnitude and colour bins of galaxies of the $SDSS_{LGalaxies}$ sample. From Newman (2008); Ménard et al. (2013), when the unknown sample redshift distribution is wide, the bias evolution correction becomes more significant. As seen in later sections (e.g. 2.5.2), although distributions are generally narrow, they can often be wider than $\sigma_z = 0.2$, meaning the choice of a flat bias evolution correction may not be sufficcient.

Since we are applying the clustering redshifts method to data from the LGalaxies SAM, we choose to produce a bias correction from the simulation in order that we can test its significance. We do this by measuring the evolution in bias for many different samples of galaxies in the model to produce a correction. This approach has the advantage of being applicable to any sample of galaxies since we can look at the same bin of galaxies in our model and compute a correction (rather than applying one correction for all galaxies). We will see, however, that while it should be correct for mock data, the derived correction may not hold for real data.

We first measure the integrated autocorrelation, $\bar{\omega}_{uu}(z)$ as a function of redshift in L-Galaxies data. We measure $\bar{\omega}_{uu}(z)$ for 10 bins of i-band magnitude of width $\Delta i = 0.25$ between $17 < i < 20$ and $\Delta i = 0.125$ between $20 < i < 21$ (since we have significantly more galaxies at fainter magnitudes). For each magnitude bin, we measure $\bar{\omega}_{uu}(z)$ in redshift bins of width $\Delta z = 0.1/3$, as this is the binning we will apply to test our data in section 2.5.2. Measuring $\bar{\omega}_{uu}(z)$ for galaxies binned by magnitude accounts for any dependence of the bias correction on luminosity. Figure 2.7 shows $\bar{\omega}_{uu}(z)$ computed in three different magnitude bins, with their redshift distributions, for reference. Two separate panels show the integrated correlation function amplitude over either small scales (0.5<r_p<1.5 Mpc, middle panel) or large scales (5<r_p<15 Mpc, bottom panel).

Looking at both small scale and large scale clustering, in all bins of magnitude, there is a significant increase in the clustering amplitude towards lower redshifts. This behavior is particularly noticeable in the two faintest magnitude bins. There is also, in all magnitude bins, an increase in clustering amplitude towards higher redshifts, although in general this evolution is smaller at larger scales.

Some of the evolution in integrated clustering amplitude seen in figure 2.7 is expected from the fact that we have magnitude-limited samples. For a given magnitude bin, galaxies at high redshift are, on average, intrinsically more luminous and therefore more massive and more strongly biased. This will cause an increase in clustering amplitude towards high redshifts, as seen in Figure 2.7 in all magnitude bins.

The reason for the increase towards low redshifts is less obvious, however several things

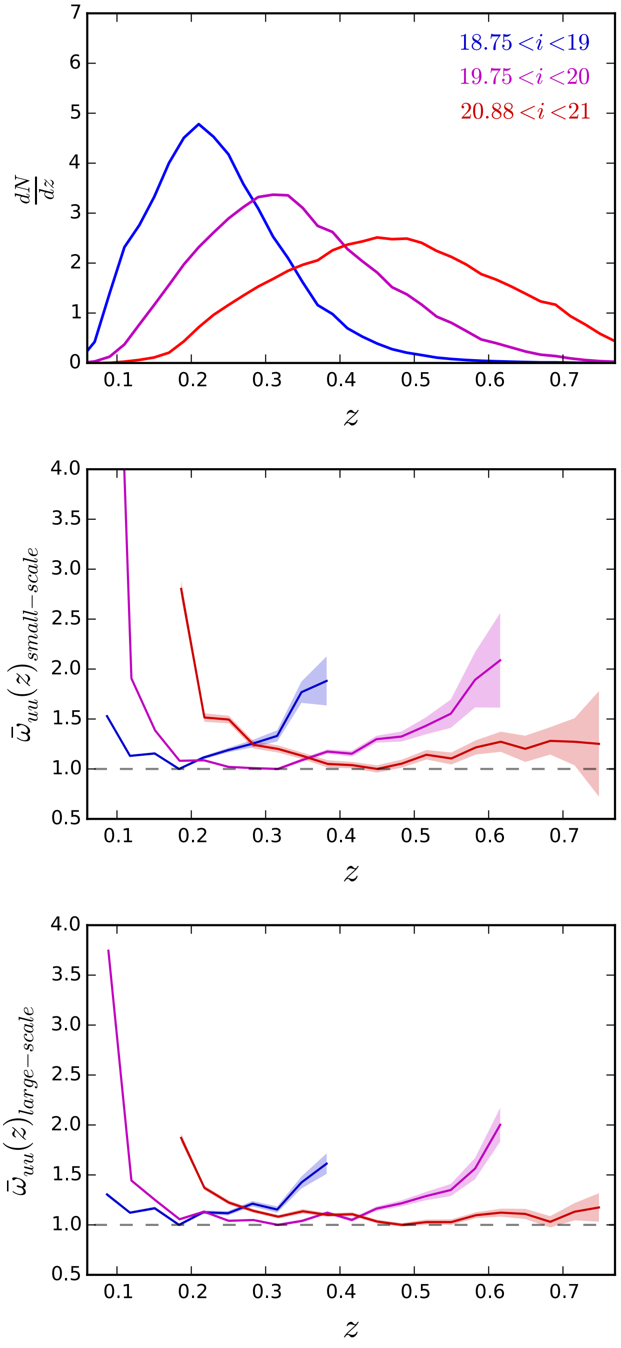

Figure 2.7: (top) The normalised redshift distributions of three different magnitude bins of LGalaxies data: $18.75 < i < 19$ (blue), $19.75 < i < 20$ (magenta), and $20.875 < i < 21$ (red). The centre and bottom plots show the integrated angular autocorrelation function, $\bar{\omega}_{uu}(z) = \bar{b}_u^2(z)\bar{\omega}_{DM}(z)$, as a function of redshift for these three bins. The distributions are shown for small scales, $0.5 < r_p < 1.5$ Mpc (centre) and large scales, $5 < r_p < 15$ Mpc, (bottom). Since the clustering redshifts method only depends on the evolution of $\bar{\omega}_{uu}(z)$, not the overall amplitude, integrated correlation functions are normalised such that the minimum value is 1. Error bars are computed from correlation function errors.

may be contributing to this. Firstly we are measuring the evolution of $\bar{\omega}_{uu}(z)$, which captures the evolution of both the bias of the unknown sample and of the dark matter power spectrum (i.e. $\bar{\omega}_{DM}(z)$), the latter of which increases in amplitude towards low redshift. However, the change in $\bar{\omega}_{DM}(z)$ would not fully explain the change in amplitude that we observe. Secondly, satellite fraction is known to increase towards lower galaxy luminosity and stellar mass (Mandelbaum et al., 2006; Reddick et al., 2013), which would also contribute towards a larger $\bar{\omega}_{uu}(z)$ with decreasing redshift, as our magnitude-limited samples will be increasingly dominated by lower stellar mass objects at low redshift. Thirdly, we ought to consider whether the L-Galaxies model produces realistic galaxy clustering. van Daalen et al. (2016) present an exploration of how the clustering of galaxies can aid the constraint of semi-analytic models. In their Fig. 5 they show that without explicitly using clustering as a model constraint, several flavours of L-Galaxies models fail to reproduce the clustering of low-mass galaxies ($M_\star \lesssim 10^{9.5} M_\odot$), even if the clustering of high mass galaxies matches the SDSS-measured correlation functions very well. The tendency for L-Galaxies to over-predict the clustering amplitude of low-mass galaxies (the model that we use here is not calibrated using clustering) will also contribute towards the behaviour seen in Figure 2.7.

This difference between clustering in LGalaxies and the real universe is not a problem in this chapter, since we are using these measurements to correct for bias evolution in our mocks (which are created from the same simulation). In chapter 3, however, we apply the clustering redshifts method to real data, so as well as this method of bias correction, we apply several other analytic bias corrections measured from real data (see section 3.4).

When applying a bias correction from L-Galaxies in future sections, we use the measured evolution of $\bar{\omega}_{uu}(z)$ in $SDSS_{LGalaxies}$ from Figure 2.7 as an estimate of the bias evolution of the unknown sample, following equation 2.8. This correction is computed in the same magnitude bin and over the same physical scales as the cross-correlation is measured. We will test the significance of this correction by comparing the resulting redshift distributions to those where no correction is applied.

2.5.2 Recovering distributions of mock data

In order to test the clustering redshifts method, we bin our mock photometric survey, $SDSS_{LGalaxies}$ by magnitude and colour, such that we have many samples of galaxies to test our method on of different magnitude and colour. To do this, we firstly bin by i-band magnitude in bins of

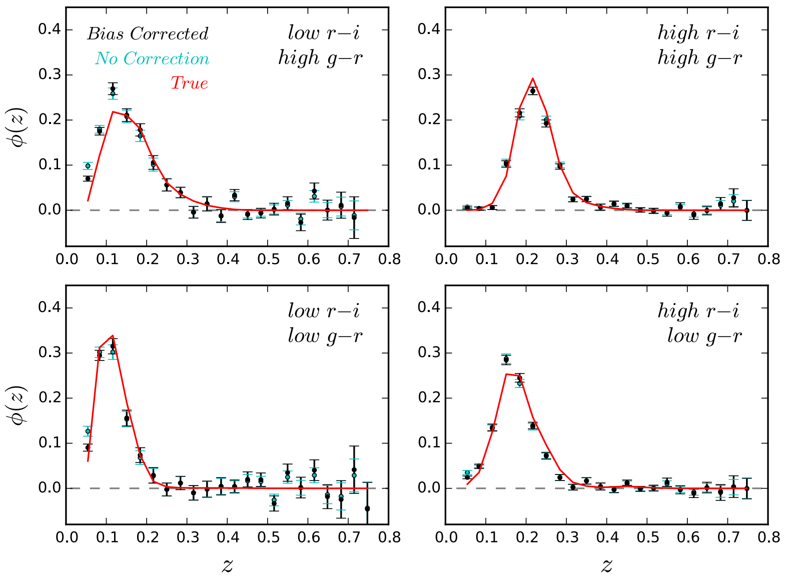

Figure 2.8: The recovered redshift distributions of different bins of colour of $\text{SDSS}_{LGalaxies}$ data, with both no bias correction (cyan), and the bias correction computed in section 2.5.1 (black). The true distribution is given by the red line. We choose galaxies from a bright magnitude bin ($18 < i < 18.25$), and show four colour bins covering the extent of the colour space.

width $\Delta i = 0.25$ between $17 < i < 20$ and $\Delta i = 0.125$ between $20 < i < 21$ where the large number of galaxies allows us to bin more finely. Within each of these magnitude bins, we then bin by $r-i$, and then by $g-r$. We choose a number of bins such that each contains $> 100,000$ galaxies, as we found this to be roughly the minimum number of galaxies required to recover a moderately noise-free $n(z)$. At fainter magnitudes, the size of bins is comparable to the photometric error in the SDSS, so smaller bins would not provide significantly more information as galaxies are already scattered between bins. Binning by i, $r-i$ and $g-r$ produces 492 bins.

We then recover the redshift distributions of all bins of $SDSS_{LGalaxies}$ by crosscorrelating with a reference sample, $BOSS_{LGalaxies}$, as in section 2.4. We integrate over the scales 0.5 to 1.5 Mpc, and when computing crosscorrelations, split our reference sample in to bins of width $\Delta z = 0.1/3$ - the same redshift bins that our bias evolution correction is measured over. We correct for bias evolution using the computed evolution in LGalaxies detailed in section 2.5.1. Correlation function errors are again computed using a bootstrap method, which in turn is used to compute errors on the final $\phi(z)$ following equations 2.6 and 2.8. Figures 2.8 and 2.9 show the recovered and true redshift distributions of a selection of these colour bins, in bright and faint magnitude bins, respectively.

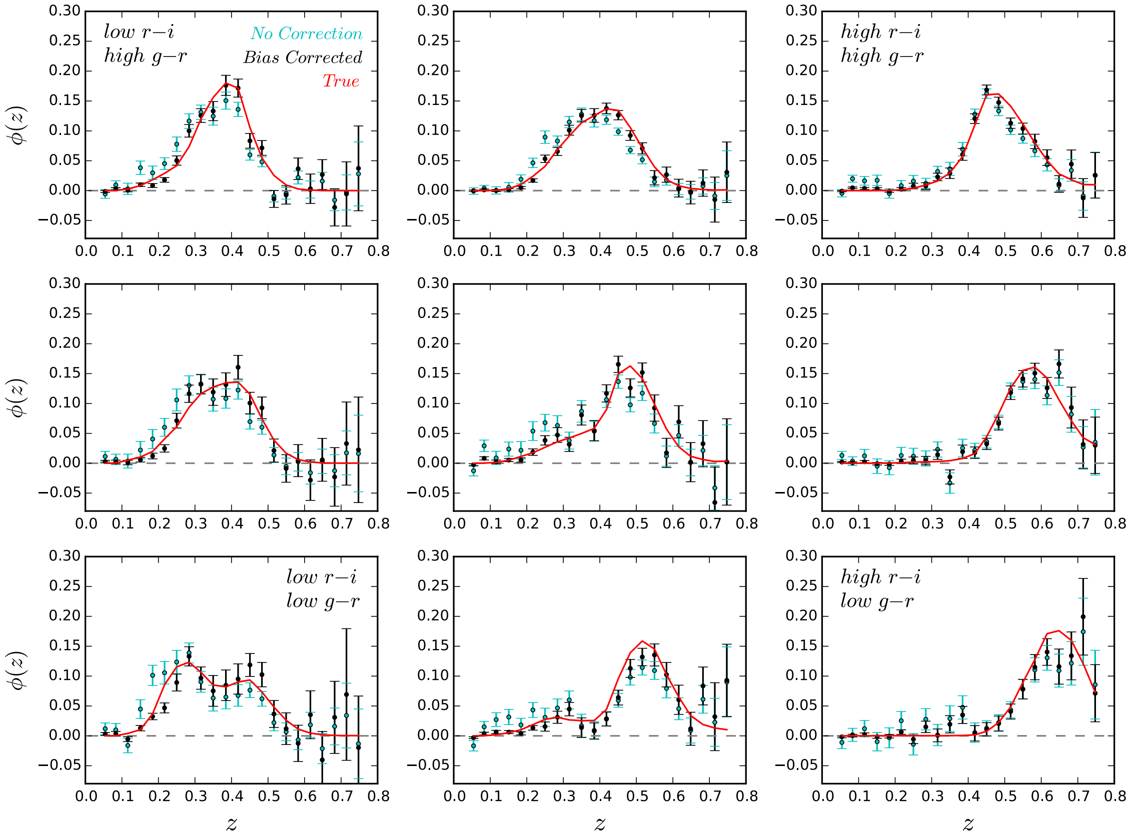

Figure 2.9: The recovered redshift distributions of different bins of colour of SDSS$_{LGalaxies}$ data, with no bias correction (cyan), and the bias correction computed in section 2.5.1 (black). The true distribution is given by the red line. Here we choose galaxies from the faintest magnitude bin ($20.875 < i < 21$), containing 7x7 bins in $r - i$ and $g - r$. A selection of bins is presented throughout the colour space (bins 2, 4 and 6 in both dimensions). When computing redshift distributions, the crosscorrelation is integrated over small scales ($0.5 < r_p < 5$ Mpc).

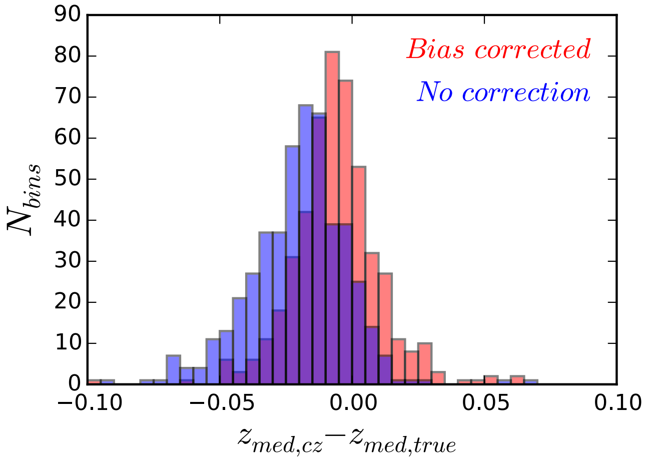

Figure 2.10: Testing the recovery of redshift distributions in L-Galaxies. The distribution of errors in the median redshift, $z_{med,true} - z_{med,cz}$, for all bins of colour and magnitude, both with and without the bias correction.

Figure 2.8 shows four bins of $r-i$ and $g-r$ covering the extent of the colour space of a bright magnitude bin. It can be seen that the redshift distribution, $\phi(z)$, is recovered well for a range of different values of $g-r$ and $r-i$. Adding a bias correction does not significantly affect the recovered distribution, likely because distributions are narrow, and because the correction, computed in section 2.5.1, is fairly small at bright magnitudes. Examining the faintest magnitude bin, presented in figure 2.9, redshift distributions are again recovered well for a range of different values of $g-r$ and $r-i$, however the bias evolution correction becomes more important, pushing our $\phi(z)s$ closer to the correct values. This effect appears to be particularly true for wider distributions, where the bias is likely changing between low and high redshift ass seen previously in figure 2.7.

If using a photometric survey with smaller photometric error, for example DECaLS or DES, redshift distributions for a given colour bin would be much narrower since galaxies will be less scattered between neighbouring bins. The correction therefore becomes less significant, particularly at faint magnitudes, where SDSS errors are large. Errors are visibly larger at higher redshift ($z > 0.65$), where the number density of objects in the reference sample is low, which can sometimes cause an error in normalisation. This effect should average out over many bins, however, and will be less of a problem when using real data since the true BOSS sample has a larger area, and there are additional eBOSS galaxies and quasars above this redshift.

As a further check of our bias correction, we compute the true median redshift, $z_{med,true}$, for all colour bins, along with the median redshift using clustering redshifts, $z_{med,cz}$, both with and without a bias correction. We compute the error in the median redshift, $z_{med,true} - z_{med,cz}$, for all bins, and present the distribution of errors in figure 2.10. Without a bias correction, median redshifts are almost always slightly below the true value. The bias correction shifts the median to higher redshifts, although there remains a similar amount of scatter around the correct value. These errors in the median redshift are fairly small however, relative to the size of our redshift bins (width $\Delta z = 0.033$). The scatter is partly due to noise in the recovered redshift distribution, but also may arise because we compute a bias correction for an entire magnitude bin, and this approach may not necessarily describe the bias evolution of all bins of colour within this.

2.5.3 Testing the fitting scale

Here we investigate how the choice of fitting scale affects the recovered $\phi(z)$. We compute redshift distributions of mock data for the same bins of magnitude and colour, but over larger scales. Figure 2.11 shows the recovery of several colour bins within the faintest magnitude bin (equivalent to figure 2.9), but integrated over large scales ($15 < r_p < 50$ Mpc).

While redshift distributions are generally recovered successfully, there is a significant amount of extra noise when compared with the small scale recovery (figure 2.9). We compute the average error in $\phi(z)$ (i.e. the error due to errors in the correlation functions) for both small scales and large scales. We average this error across all colour bins, and all redshifts; when recovering redshift distributions over large scales, the error is on average 2.4x larger. When noise becomes large, a significant error in normalisation can appear, as seen in, for example, the bin of lowest $r - i$ and highest $g - r$ of figure 2.11. For this reason, we conclude that measuring small scale clustering is best when applying to real data.

2.6 Conclusions

We have implemented a method of clustering redshifts based on the method of Ménard et al. (2013). We have performed tests on the BOSS survey, splitting galaxies by colour in to a blue and red sample. Recovering the redshift distribution of the blue sample by crosscorrelation with the red sample is very successful, even when assuming no evolution in bias for the unknown sample. We test the recovery of a sub-sample of the unknown sample, showing that

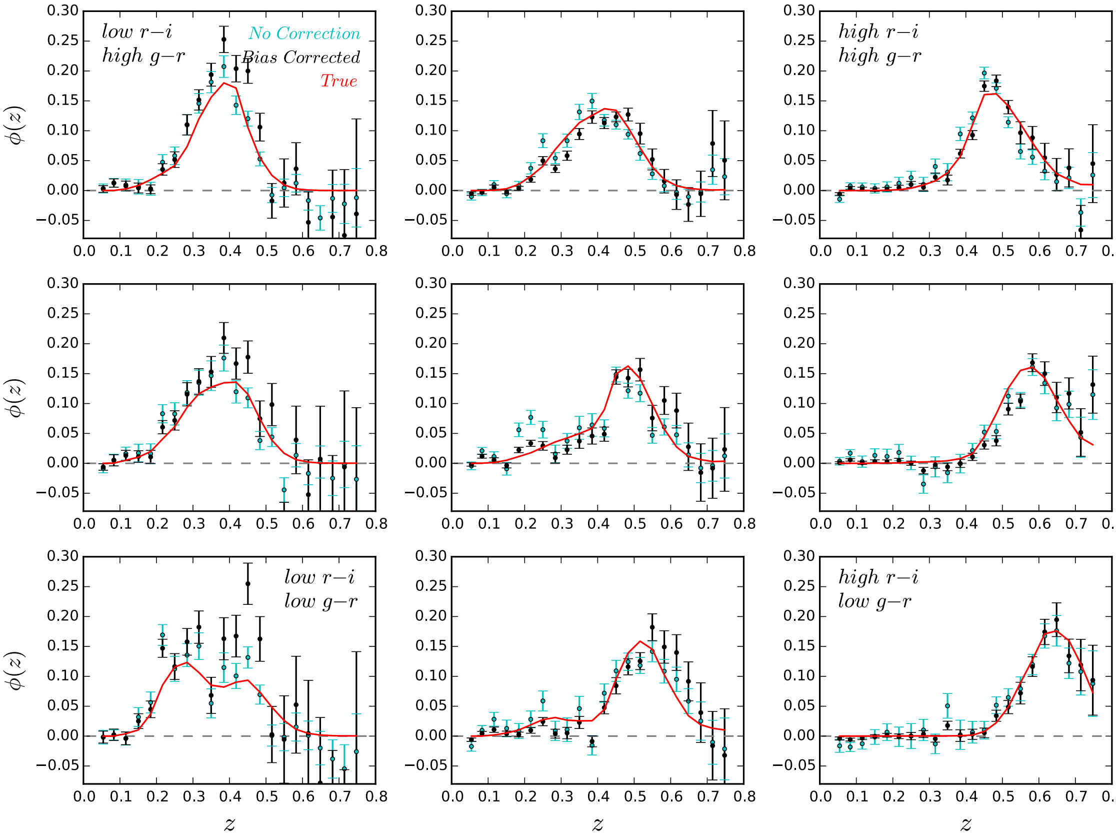

Figure 2.11: Same as in figure 2.9, except the crosscorrelation is integrated over large scales (15 < r_p < 50 Mpc), rather than smaller scales.

noise in the recovered $\phi(z)$ increases significantly as the number of galaxies in the unknown sample decreases.

We have also investigated how the method performs using mock galaxy catalogues defined from semi-analytic models. We define a mock SDSS survey and mock BOSS survey with comparable magnitudes and photometric errors to the real surveys. We firstly investigate how the bias evolves in the mock SDSS, finding that clustering amplitude for magnitude limited samples increases at high redshifts, likely due to the fact that these galaxies are intrinsically luminous and hence strongly biased. Clustering amplitude also increases at low redshifts in these samples, implying that, although an increasing satellite fraction may have some effect on this, amplitude in the model is too strong here.

We apply our implementation of clustering redshifts to recover redshift distributions of mock SDSS galaxies in small bins of i, $g-r$, and $r-i$, using a bias correction defined from how clustering amplitude evolves in the model. We find that at bright magnitudes, redshift distributions are recovered well, and the choice of bias correction has little effect. At fainter magnitudes, distributions are recovered well with a bias correction, however, with no correction applied, redshift distributions are biased towards low redshifts. The significance of the bias correction here is likely because redshift distributions are much wider at faint magnitudes and hence clustering is more affected by the bias evolution. This increased width is likely due to the fact that photometric error is higher at fainter magnitudes, scattering galaxies across colour-bins, and hence redshift is less well constrained in each bin. We note that this may not be as significant a problem for future photometric surveys, for example DES or DECaLS, with better photometry and smaller photometric error.

We finally investigate how the choice of clustering scale affects the recovered distributions. We find that crosscorrelations over larger scales are noisier, leading to more error in recovered $\phi(z)s$. We therefore suggest using the smallest scale possible when recovering redshift distributions of real data.

3

Mass Functions, Luminosity Functions and Completeness

In this chapter, we apply the clustering redshifts method to real data, with the view of using these redshift estimates to measure how parameters such as stellar mass and luminosity evolve through time. To do this we bin the SDSS photometric survey in to small bins of colour and magnitude, and then compute redshift distributions of all bins. With these recovered redshifts, we compute masses and luminosities, allowing us to produce mass and luminosity functions in magnitude limited samples. From the recovered mass functions we also measure the stellar mass completeness of the BOSS survey. Much of the work of this chapter is presented in Bates et al. (2019).

3.1 Introduction

Large spectroscopic galaxy surveys are useful tools for studying galaxy evolution. They allow us to determine stellar masses, star formation histories, and dynamics for large numbers of galaxies. In particular, deep, small area surveys such as PRIMUS (Coil et al., 2011), DEEP2 (Newman et al., 2013), and VIPERS (Guzzo et al., 2014) contain data for galaxies over a broad range of masses, colours, morphologies, and redshifts, allowing tests of galaxy evolution on very different objects. These surveys, however, are not ideal for investigating galaxy evolution at the highest masses, since the number density of galaxies above, for example, $M_\star > 10^{11.5} M_\odot$, is extremely low. Due to their small area, these pencil-beam surveys typically only target tens or hundreds of galaxies above this mass, and are strongly affected by sample variance.

An ideal approach to study galaxy evolution at these masses is using large-volume cosmological redshift surveys, which typically target the highest mass galaxies over very large regions of the sky. The Baryon Oscillation Spectroscopic Survey (BOSS) (Eisenstein et al., 2011; Dawson et al., 2013) is the most extensive of these to date, measuring spectra for roughly 1.5 million luminous red galaxies (LRGs) over 10,000 deg^2 of sky at $z < 0.7$. BOSS contains over 100,000 galaxies with stellar masses $M_\star > 10^{11.5} M_\odot$ (Maraston et al., 2013), so it is able to study this end of the mass function with very little shot noise. Ongoing and future surveys such as eBOSS (Blanton et al., 2017; Dawson et al., 2016) and DESI (DESI Collaboration et al., 2016) will extend this study to higher redshifts and larger numbers of galaxies, providing additional data to better probe these masses.

One limitation of these surveys, however, is that they are optimised for cosmology, not galaxy science. Their target selection therefore involves a number of complex colour cuts, leading to samples of galaxies that are incomplete in both stellar mass and colour. To study galaxy evolution at these masses, we must quantify this incompleteness.

One method of determining incompleteness is by comparing the distribution of galaxies as a function of mass in one sample, to that of another sample which is complete in stellar mass. In Leauthaud et al. (2016), they chararacterise the stellar mass completeness of BOSS using Stripe 82, a narrow region of the SDSS with deeper $ugriz$ photometry, as well as near-IR photometry from the UKIRT Infrared Deep Sky Survey (UKIDSS) (Lawrence et al., 2007), allowing for more accurate photometric redshifts and stellar masses.

Large-area broad-band photometric surveys such as the Sloan Digital Sky Survey (SDSS) (York et al., 2000) are complete down to some magnitude, provide data over a large area ($\simeq$ 14000 deg^2) for galaxies over a range of magnitudes and colours, so would be ideal for this purpose. One disadvantage, however, is that for SDSS-like data, photometric redshifts can be unreliable (Rahman et al., 2016a). In this paper, we outline a method of computing luminosity and mass functions (and hence completeness) from broad-band surveys using a technique known as clustering redshifts. This technique is described in detail in chapter 2, but to summarise, the method obtains the redshift distribution for a set of galaxies by crosscorrelating their positions with those of a spectroscopic sample in different redshift bins.

In this paper we use clustering redshifts to recover the redshift distributions of samples of galaxies from the SDSS photometric survey in small bins of magnitude and colour, isolating galaxies of similar type. After recovering redshift distributions of bins, we use these colours and redshifts to compute stellar masses and luminosities by examining simulated galaxies in the same bins of colour-redshift space. Finally, we compute targeting completeness for the BOSS spectroscopic sample.

The layout of this chapter is as follows: Section 3.2 describes both the real and mock data used in this study. Section 3.3 describes how we correct for the bias evolution on real data, and how we compute stellar masses and luminosities using semi-analytic models. In section 3.4, we apply the clustering redshifts technique to real photometric data. In section 3.5 we determine how accurately mass and luminosity functions can be recovered using our method. In section 3.6, this technique is applied to real SDSS photometry to produce real mass and luminosity functions. Section 3.7 presents completeness measurements for BOSS using these computed mass and luminosity functions. Finally, in section 3.8, we discuss these results, and outline possible extensions of this work.

3.2 Data

In this chapter, we apply the clustering redshifts technique to real data, so require both photometric and spectroscopic data. We also make use of mock catalogues, both to test our method, and also to compute masses and luminosities from the recovered redshifts and colours of photometric data.

3.2.1 BOSS and eBOSS spectroscopic surveys survey

Building on the work presented in chapter 2, as a reference sample, we use data from both the SDSS-III: BOSS (Eisenstein et al., 2011; Dawson et al., 2013; Gunn et al., 2006; Smee et al., 2013) and also SDSS-IV: eBOSS (Blanton et al., 2017; Dawson et al., 2016; Gunn et al.,

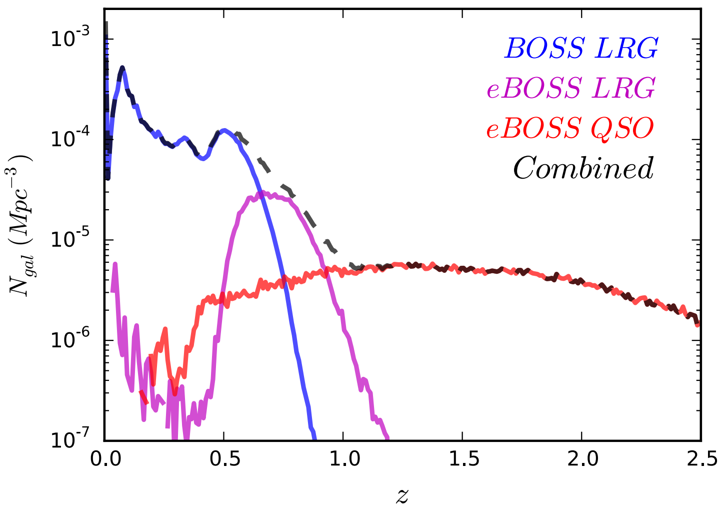

Figure 3.1: The comoving number density of three different spectroscopic galaxy samples described in section 3.2.1: At low redshift, BOSS DR12 LRGs (blue), intermediate redshifts, eBOSS DR14 LRGs (magenta), and higher redshifts, eBOSS DR14 quasars (red). The number density for the combination of these three samples is shown as the black dotted line.

2006; Smee et al., 2013) spectroscopic surveys. Both of these surveys are described in detail in 1.3.2. These samples are ideal for a reference sample, as they cover a large area and provide continuous, large numbers of redshifts over the range $0 < z < 3$. The reference sample is therefore a combination of BOSS DR12 (Alam et al., 2015) LRGs in the North Galactic Cap (NGC) covering 6851 deg^2 and eBOSS DR14 data (Abolfathi et al., 2018), containing both the LRG and quasar samples, covering 1011 and 1214 deg^2 respectively, within the BOSS NGC area. The total sample is then 1.1 million galaxies. The number density of both the individual and combined samples are shown in figure 3.1. When computing correlation functions in later sections, we use large scale structure catalogues from Reid et al. (2012) for the BOSS LRG sample, Bautista et al. (2017) for eBOSS LRGs, and Ata et al. (2018) for eBOSS quasars, using 10x randoms for all samples - note that the random catalogues account for the changing area of the different samples.

3.2.2 SDSS photometric survey

Our photometric survey (the sample for which we wish to compute redshift distributions, along with masses and luminosities) consists of data from the SDSS photometric survey (York et al., 2000; Gunn et al., 1998), described in detail in section 1.3.1. We use photometry from

DR8 (Aihara et al., 2011), selecting only objects morphologically classified as galaxies, and only data from the primary survey (i.e., the best observation for each object). To create our catalogues, we use g, r and i band modelMag magnitudes (see Stoughton et al. (2002)), and constrain the sample to $i < 21$ to avoid incompleteness (York et al., 2000). We only include galaxies in the same region as the BOSS NGC DR12 footprint using the following masks detailed in Anderson et al. (2012): The survey geometry mask, and veto masks for bright stars, unphotometric seeing, and bright objects. Finally, we remove all galaxies that are also in our reference sample, leaving 53 million galaxies over ~ 7000 deg^2. We create random catalogues for this sample as in section 2.1.3, using 10x randoms.

3.2.3 Mock catalogues

Following chapter 2, we create mock catalogues to test our method on. Here, however, we use data from both the LGalaxies and SAGE semi-analyic models. These models are described in detail in section 1.2.3. Along with the LGalaxies lightcone created in section 2.3.2, we also use a smaller lightcone from SAGE covering a 100 deg^2 area, run on the MultiDark MDPL2 simulation (Klypin et al., 2016; Knebe et al., 2018), with SEDs and magnitudes also computed using Maraston & Strömbäck (2011) SSPs and a Kroupa et al. (2001) IMF. Both catalogues have angular positions, redshifts, and SDSS magnitudes with reddening and photometric errors applied following section 2.3.2. In section 3.3.2, we also require present day stellar masses and absolute magnitudes for our mock galaxies, so add these to both catalogues.

We define the same reference and photometric samples as in section 2.3.2 by applying the BOSS colour cuts, and $i < 21$ limit for SDSS data, producing the catalogues $SDSS_{LGalaxies}$, $SDSS_{SAGE}$, $BOSS_{LGalaxies}$ and $BOSS_{SAGE}$. This procedure produces samples with comparable redshift distribution to the BOSS survey which is further discussed in section 3.5.

3.3 Method

Chapter 2 outlines the clustering redshifts method, and shows that we can successfully bin mock galaxies by magnitude and colour, and recover their distributions. Here, we extend this work, applying this method to real data, going on to produce mass and luminosity functions from the recovered distributions. To perform this analysis on real data, however, a number of additional considerations must be made.

3.3.1 Bias corrections for real data

We previously computed a bias correction by measuring how clustering evolves in the LGalaxies model, showing that although this is successful when applied to mock data, clustering strength appears to be higher at low masses than in the real universe, so the computed bias correction may not be fully applicable. For this reason we now also consider analytic forms of the bias when investigating the effect of the bias evolution correction on our recovered redshift distributions, and mass and luminosity functions.

Rahman et al. (2015) compute bias corrections fit to SDSS main spectroscopic sample clustering: one that evolves as $db/dz = 1$ and one that evolves as $db/dz = 2$. For our study we choose the more extreme evolution, $b_1(z) = 0.7 + 2z$, in an effort to bound the effect of this uncertainty. We also investigate a correction from Rahman et al. (2016a), which takes the form $b_2(z) = 1$ for $z < 0.1$, and $b_2(z) = 0.9 + z$ for $z \geq 0.1$. This leaves us with three different bias models to investigate: one from LGalaxies and two analytic laws, fitted to spectroscopic data.

In Sections 3.4 and 3.6 we will compute redshift distributions, luminosity functions, stellar mass functions, with different assumptions about the bias evolution of the unknown sample. We firstly test our method on mock data, where we will apply the bias evolution derived from simulations, and quantify the effect of not correcting for this evolution at all. When applying this to real data we also consider the two analytic forms. Ultimately, we use these models to quantify the effect of the bias correction in our final measurements, and add this systematic error to our estimates of the statistical error.

3.3.2 Computing masses and luminosities

After binning our SDSS sample in small bins of magnitude and colour, and recovering their redshift distributions as in 2.5.2, for each bin we know the value of i, $r - i$ and $g - r$, along with the number of galaxies at each redshift. At all redshifts, we therefore have a measure of the rest-frame spectral energy distribution (SED) of galaxies in this bin. We can compute parameters from this SED (e.g. mass, luminosity) and allocate these to photometric galaxies in the correct quantities.

To compute masses and luminosities for our bins of colour, we choose to use semi-analytic models. After recovering redshift distributions of all bins of colour of our photometric survey, we compute the distribution of mass or luminosity within the same colour-redshift bin of the SAM, and apply this probability density function (PDF) to the real data (i.e. multiply this PDF by the number of galaxies in the same bin of the real data). After recovering masses and luminosities at each redshift, for all bins of colour, these distributions are summed to produce mass and luminosity functions (see sections 3.5 and 3.6).

Although SAMs do not predict precisely the correct number density of galaxies of given colours, for a given set of colours and redshift, the type of galaxy (i.e. star formation history (SFH), mass, luminosity) is assumed to be representative of those in the real universe. This method allows us to account for photometric errors by adding errors to our SAM, scattering galaxies across magnitude bins. Using the PDF of mass and luminosity in each bin rather than just a best fit ensures that the correct distribution of mass is allocated in each bin. This technique is, in essence, similar to Pacifici et al. (2012), where a library of physically motivated SFHs is computed from SAMs, and then used to fit individual galaxy SEDs. We test how well our method performs, including the important on the choice of SAM in section 3.5.

3.4 Recovering redshift distributions of real data

Here we recover redshift distributions of galaxies from the SDSS, which is described fully in section 3.2.2. When recovering redshift distributions of samples, we cross-correlate the sample with our reference sample described in section 3.2.1, consisting of BOSS and eBOSS LRGs and quasars.

As a demonstration of the technique applied to real data, we firstly split the SDSS sample in to small bins of $g-r$ of width $\Delta(g-r) = 0.033$, and recover redshift distributions of all bins as described in section 2.5.2, with no unknown sample bias correction applied. The resulting redshift distribution of all $g-r$ bins is shown in figure 3.2. The colour bimodality in the galaxy population is clearly visible, with two discrete populations visible at all redshifts, and $g-r$ values of these populations increasing at higher redshifts. Horizontal structure is visible in the distribution, showing correlations in the redshift distributions of different bins. This likely represents correlated structure between colour bins (i.e. from the fact that these samples all occupy the same density field). The number density of the sample decreases after $z \simeq 0.4$ when the $i < 21$ magnitude cut of the SDSS sample becomes significant.

With the view of computing galaxy parameters, we now bin by three parameters simulta-

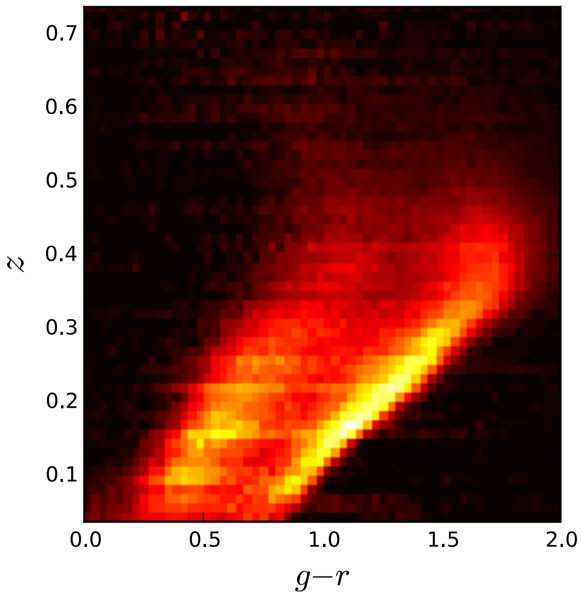

Figure 3.2: Redshift distributions of SDSS galaxies binned by $g - r$. These distributions are plotted vertically for each $g - r$ bins, producing the overall distribution of galaxies in $g - r$ vs redshift.

neously: i, $r - i$, and $g - r$. We follow the same method as section 2.5.2, firstly binning galaxies i-band magnitude in bins of width $\Delta i = 0.25$ between $17 < i < 20$ and $\Delta i = 0.125$ between $20 < i < 21$, then within each of these bins, binning by $r - i$, and then by $g - r$ such that each bin contains $> 100,000$ galaxies (the $i < 17$ magnitude limit comes from the fact that brighter than this, we do not have enough galaxies to split in to colour bins). We again cross-correlate each of these bins with our combined reference sample in order to recover redshift distributions.

In order that we can test if our method is working well on real data, we compare our redshift distributions to data from the GAlaxy and Mass Assembly survey (GAMA) (Driver et al., 2009), described in section 1.3.2. GAMA is a spectroscopic survey, magnitude limited to $r < 19.8$, targeted over ~286 deg^2 of sky. Below $r < 19.8$, GAMA is highly complete ($> 95\%$), although completeness drops for fainter magnitudes. For $r \lesssim 19.8$, GAMA redshift distributions can be compared to the distributions we recover for SDSS data.

An example of some recovered distributions is presented in Figure 3.3, alongside the redshift distributions measured from the GAMA survey (Baldry et al., 2018) in the same bins of magnitude and colour. We choose an intermediate magnitude bin, $19 < i < 19.25$, in order

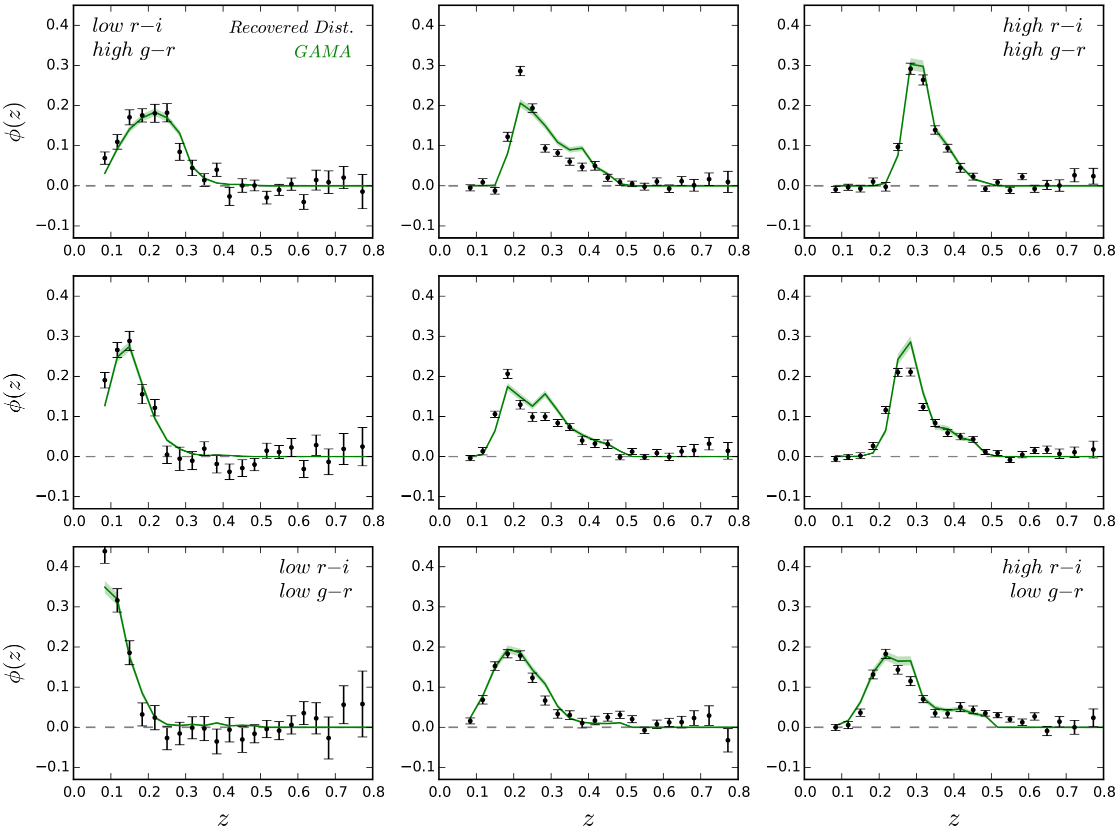

Figure 3.3: The recovered redshift distributions of different bins of colour of real data (black points). The spectroscopic redshift distribution of GAMA galaxies is indicated in the same colour bin (green line). We choose the magnitude bin ($19 < i < 19.25$), and only show bins with small values of $r - i$ ($\lesssim 0.6$) in order to avoid the $r < 19.8$ magnitude cut in GAMA.

that we have galaxies over a range of redshifts. In order to lessen the effect of the r-band magnitude cut in GAMA, we only show the bluest bins such that bins have $r \lesssim 19.8$, where incompleteness is not significant. Clustering redshifts recovery is shown without any bias correction, since in our tests, the correction is not significant in these bins, and we will be investigating the effect of this in later sections.

Recoveries of SDSS redshift distributions generally match the corresponding GAMA colour bin well. Some small differences are visible; however, this was also true for the simulated data in figure 2.9, for which the mass function is recovered successfully. If we take only bins below i < 19.25, we can use GAMA to compute the error in the median redshift of each colour bin, $z_{med,GAMA} - z_{med,cz}$, as in section 2.5.2. After computing this for all bins, the average error is $\delta z_{med} = -0.01$, indicating no significant offset with the spectroscopic redshift distribution.

3.5 Mass and Luminosity functions of mock data

We now test our method of computing masses and luminosities, described in section 3.3.2. We choose to firsly test this on mock data, such that we can investigate the significance of both the choice of bias correction, and also the choice of SAM used to compute masses and luminosities. As in the previous section, (and section 2.5.2), we bin our mock photometric survey, $SDSS_{LGalaxies}$, in to the same bins of colour and magnitude, recovering redshift distribution for all bins. We then take each of these bins at a given redshift and allocate masses and luminosities by looking in both LGalaxies (the same model, but with different photometric noise applied), and smaller lightcones from SAGE (a different model, also with photometric noise applied). This approach tests how much the choice of model affects the estimated stellar masses and luminosities. After summing the mass and luminosity distributions for all bins of colour and redshift, we produce mass and luminosity functions between $17 < i < 21$. Errors are computed using the error in $\phi(z)$ from the clustering redshifts method. The recovered luminosity and mass function are shown in figures 3.4 and 3.5.

Figure 3.4 displays the recovery of luminosity functions of our $SDSS_{LGalaxies}$ survey, in different redshift bins. Since the luminosities allocated to our galaxies are in the rest-frame, the recovered luminosity functions are by definition k-corrected. We use both LGalaxies and SAGE to compute luminosities. The true luminosity function is recovered well at all redshifts, independent of whether LGalaxies or SAGE is used to compute an absolute magnitude. This

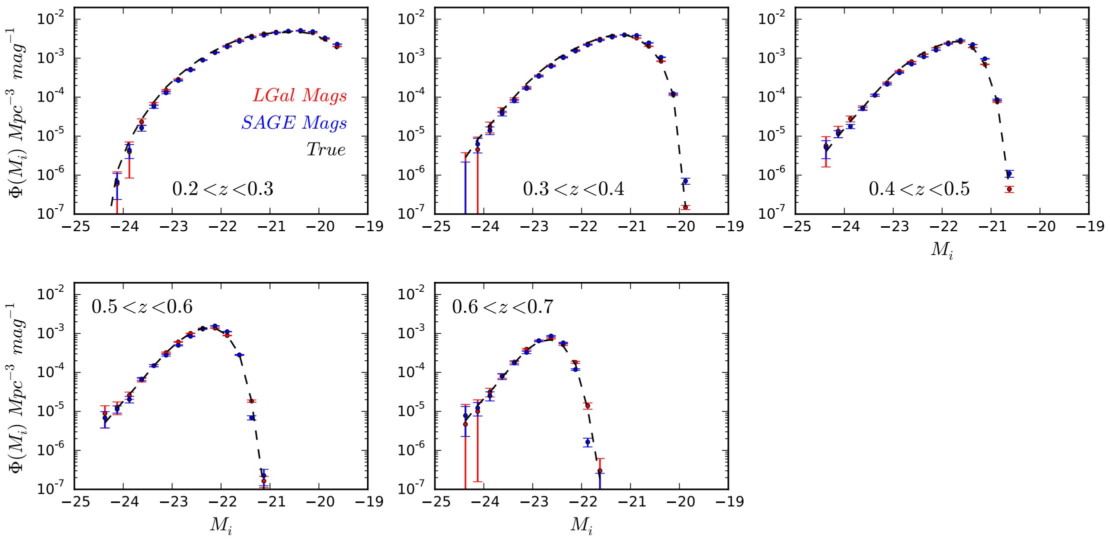

Figure 3.4: The recovered i-band k-corrected luminosity function of SDSS$_{LGalaxies}$ data for a number of different redshift bins between 0.2 and 0.7. Absolute magnitudes are recovered using colour-luminosity relations from both LGalaxies (red) and SAGE (blue) as described in section 3.5, and the true luminosity function is shown as the black dotted line. The fall-off in the luminosity functions towards faint magnitudes is due to the i<21 cut in our sample.

result makes sense, since an absolute magnitude depends only on the redshift, cosmological model, and galaxy SED. Since we have accurate recovered redshifts and i, $r-i$ and $g-r$, we have effectively a rest frame SED, so the computed magnitude from this should not be particularly dependent on the SAM chosen.

Figure 3.5 shows mass functions, again recovered at different redshifts for the two different models. Using LGalaxies to recover masses works very well (i.e., the same model to convert colours and redshifts to masses), with the recovered mass functions almost exactly matching the true values at all redshifts and masses. Examining the SAGE results, at $M_\star < 10^{11.25} M_\odot$, mass functions are recovered well; however, above these masses, the number of high mass galaxies is under-predicted.

In order to understand this difference, we compare the distribution of colours in both models as a function of mass in figure 3.6. In the two lowest mass bins ($M_\star = 10^{9.25} M_\odot$ and $10^{10.25} M_\odot$), both SAGE and LGalaxies cover roughly the same colour space at both redshifts ($z = 0.25$ and 0.5). This result implies that colours of low mass galaxies ($M_\star \lesssim 10^{11} M_\odot$) are fairly independent of the semi-analytic model chosen, and explains why the mass function is

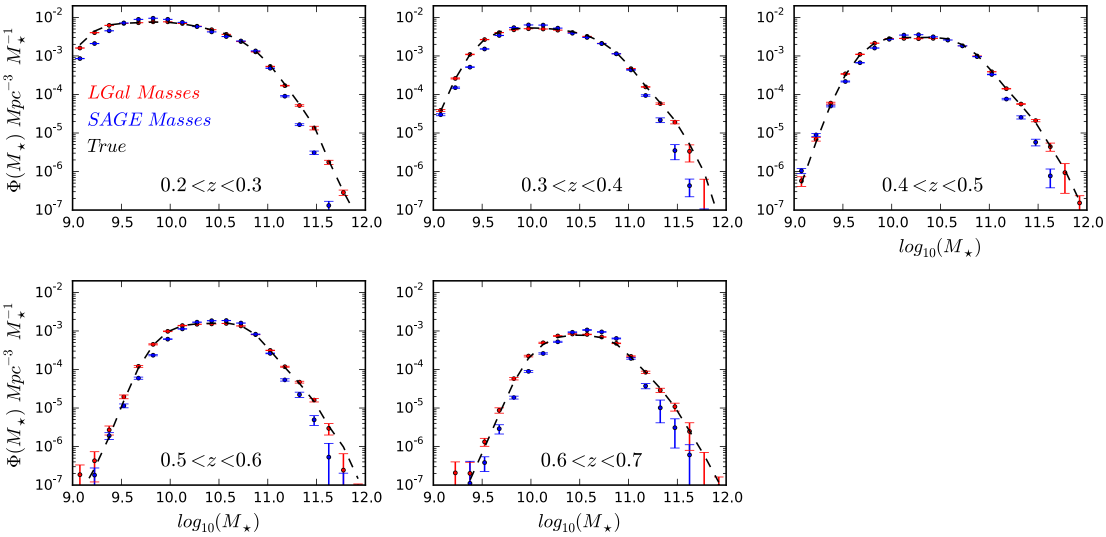

Figure 3.5: The recovered stellar mass function of SDSS$_{LGalaxies}$ data for a number of different redshift bins between 0.2 and 0.7. Masses are recovered using colour-mass relations from both LGalaxies (red) and SAGE (blue) as described in section 3.5, and the true stellar mass function is shown as the black dotted line. The fall-off in the mass functions towards lower masses is due to the i<21 cut in our sample.

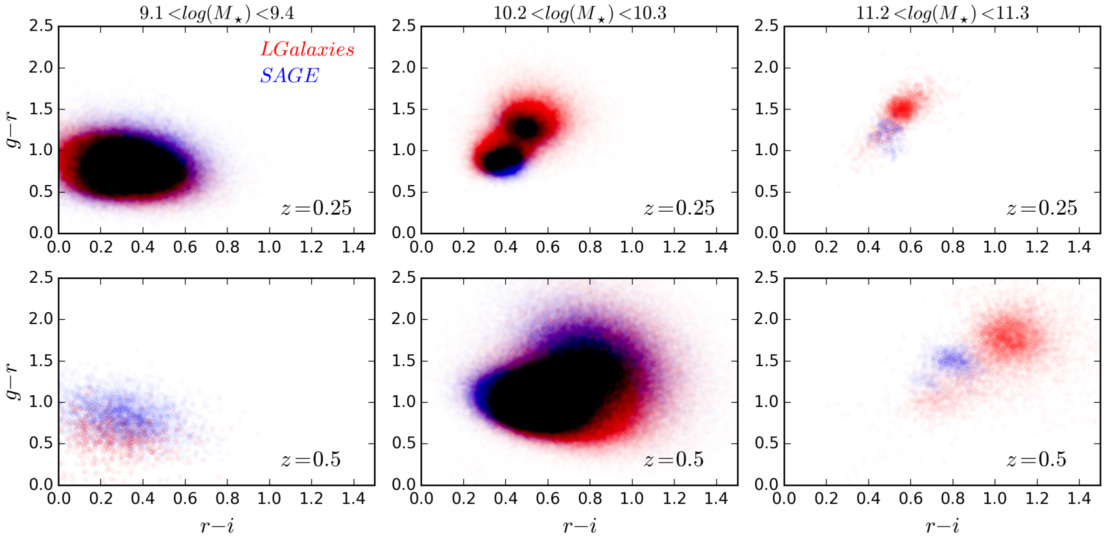

Figure 3.6: The $r-i$ and $g-r$ colours of galaxies in LGalaxies (red) and SAGE (blue). This distribution is shown for two different redshifts, 0.25 and 0.5, and for three bins of mass centered around 9.25, 10.25 and 11.25 log($M_\odot$).

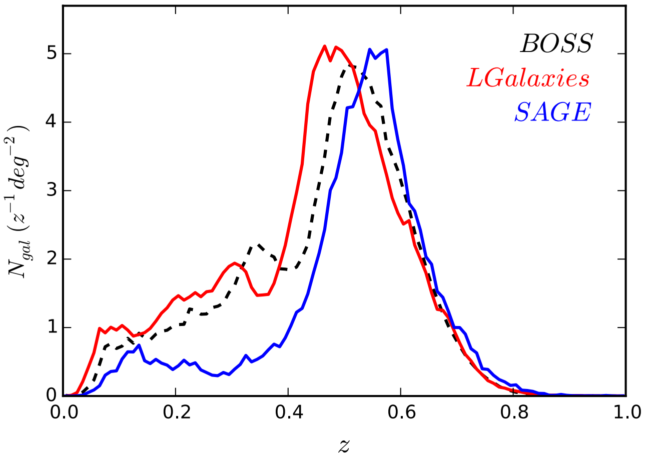

Figure 3.7: The redshift distribution (number of galaxies deg^{-2} z^{-1}) for the BOSS survey (black dotted line) compared with redshift distributions of LGalaxies (red) and SAGE (blue) with the BOSS colour cuts.

recovered well at lower masses. In the high mass bin ($M_\star = 10^{11.25} M_\odot$), colours are visibly different in the two models. This behavior implies that high mass galaxies likely have different formation processes in the two models, and explains why mass functions are not recovered as well.

Since we do not know exactly which model best describes the real universe at high masses, we investigate how well both can reproduce the BOSS survey (containing large numbers of massive galaxies). We apply the colour cuts of BOSS to both samples as described in section 3.2.3, and compare the redshift distributions of these samples and the real BOSS survey in figure 3.7. It can be seen that LGalaxies reproduces both samples within the BOSS survey: the LOWZ sample at $0 < z < 0.4$, and CMASS sample at $0.4 < z < 0.8$. These are recovered with broadly the same number density, and although there is a slight offset in the peak of the sample, the overall shape of both samples is recovered well. SAGE manages to select some galaxies with a distribution similar to CMASS; however, at low redshift most galaxies are missing, and the overall shape is significantly different.

For this reason we trust that the colours of galaxies are significantly closer to those of the real universe in LGalaxies. We therefore opt to use LGalaxies when computing masses of real data in section 3.6.

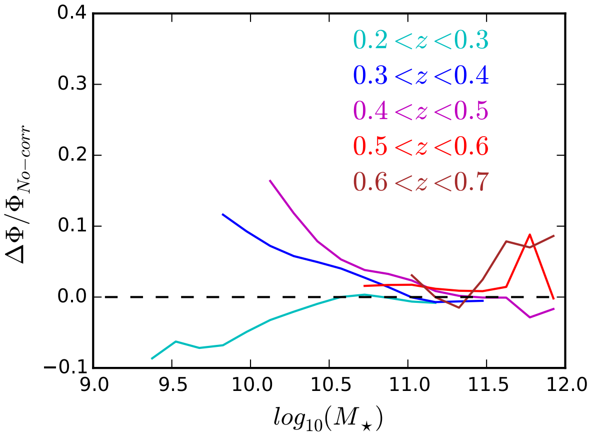

Figure 3.8: The fractional difference in the mass function between using an unknown sample bias correction from SAMs and no-correction, plotted for masses where mass functions are less than 95% complete as in section 3.6.1

3.5.1 The effect of a bias correction on stellar mass functions from simulated data

We now test the importance of the bias correction on the recovered mock mass functions. To do this we compute mass functions with and without the bias correction detailed in section 2.5.1. We then compute the ratio of these mass functions $\frac{\Phi(M_\star)}{\Phi(M_\star)_{nc}} - 1$, where $\Phi(M_\star)_{nc}$ represents the mass function without bias correction. This quantity shows the fractional change in mass function when using a bias correction compared with no correction. We show this quantity for different redshift bins in figure 3.8.

The effect of the bias correction is more pronounced at lower masses $M_\star < 10^{10.5} M_\odot$; at larger masses the change is only of the order of a few percent. We will see later, in Section 3.6.3, that this matches well with similar tests in the data, and that these uncertainties are comparable to our statistical error.

3.6 Mass and Luminosity functions of real data

We now apply the technique to real SDSS data to produce stellar mass and luminosity functions. As described in section 3.4, we recover redshift distributions of SDSS galaxies, split in to many bins of colour and magnitude between $17 < i < 21$. Note that we do not apply any bias correction for the unknown sample since we will test the effect of this later. We compute stellar

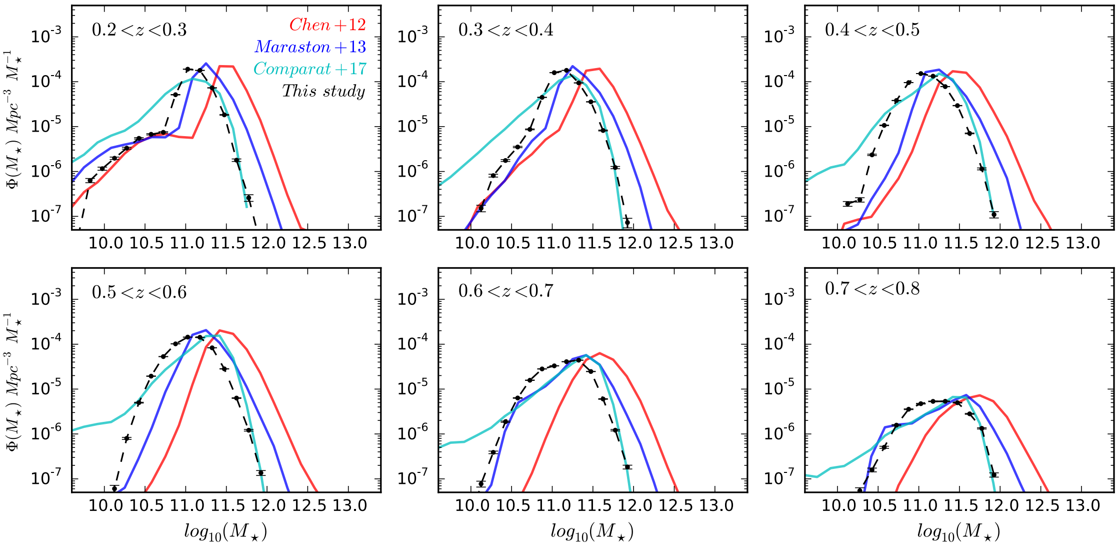

Figure 3.9: Mass functions of BOSS galaxies using four different methods, shown for six different bins of redshift. Our method described in section 3.5 is shown in black, along with Che12 in red, Mar13 in blue, and Com17 in cyan.

mass and luminosity distributions for each bin using the colour-mass/luminosity relations of LGalaxies following section 3.5.

Since our reference sample was originally removed from the SDSS sample, we compute stellar masses and luminosities for these galaxies in the same way as for the unknown sample: i.e. for a given colour-redshift bin of our reference sample, we compute the stellar mass or luminosity distribution within the same bin of L-Galaxies. We use spectroscopic redshift distributions instead of clustering redshifts for our reference sample.

After adding together the stellar-mass and luminosity distributions of our photometric sample and reference sample in all colour-magnitude bins, we can produce global stellar mass and luminosity functions, but firstly, we investigate the agreement between this method and the different published stellar mass estimates for BOSS.

We compute mass functions for just BOSS galaxies using our method, and compare this to those computed using three other methods: 1) Chen et al. (2012), hereon Che12, where galaxy parameters are modelled based on a library of model spectra for which principal components have been identified. 2) Maraston et al. (2013), hereon Mar13, where stellar population models are fit to the observed $ugriz$ magnitudes, as well as the spectroscopic redshift of each galaxy. 3) Comparat et al. (2017), hereon Com17, which for a given spectra finds the best-fit combination of single-burst SSPs. All three methods use Maraston & Strömbäck (2011) SSPs and a Kroupa et al. (2001) IMF. The four mass functions are presented in figure 3.9.

Although all methods generally agree on the shape of the mass function, there is a clear offset between methods. In particular, Che12 predicts the highest masses. Both Mar13 and our method predict broadly the same shape as Che12 at all redshifts, but this is offset towards slightly lower masses. The shape of the Com17 mass function appears slightly different, and predicts a larger number of low mass ($M_\star < 10^{10} M_\odot$) galaxies; however, the number of high mass galaxies is similar to our method.

Since all methods apply the same SSPs and IMF, the offsets between methods are likely due to assumptions about the SFH and dust. Maraston et al. (2013) also compare BOSS mass estimates from Che12 and Mar13 in their Appendix A., finding a similar constant offset of roughly 0.2 dex. They argue that while their use of a mostly passive template should increase stellar mass estimates (due to the mass-light ratio of stellar population models increasing with age), the dominant effect is likely the inclusion of dust in Che12 which may force their model to fit for a larger old component, hence increasing the global M/L ratio and producing a higher stellar mass. This likely explains the offsets between these methods, seen in figure 3.9. Applying the same argument, our estimates can then be explained to be systematically slightly lower due being computed using a range of different SFHs, rather than passive estimates as in Mar13.

These results appear to show that our method is broadly consistent with other published mass estimates, motivating its application in future sections. Furthermore, when comparing mass functions or completeness estimates across methods, these offsets clearly must firstly be taken in to account.

3.6.1 Galaxy Stellar Mass Functions from SDSS

Computed stellar mass functions after summing the mass distributions across all colour bins of our photometric and reference sample are presented in Figure 3.10. The 95% completeness limits are shown in grey, computed as regions where LGalaxies becomes less than 95% complete due to the magnitude cuts. The bright magnitude cut ($i > 17$) is significant in the two lowest redshift bins; however, the impact of this becomes less significant at higher redshifts.

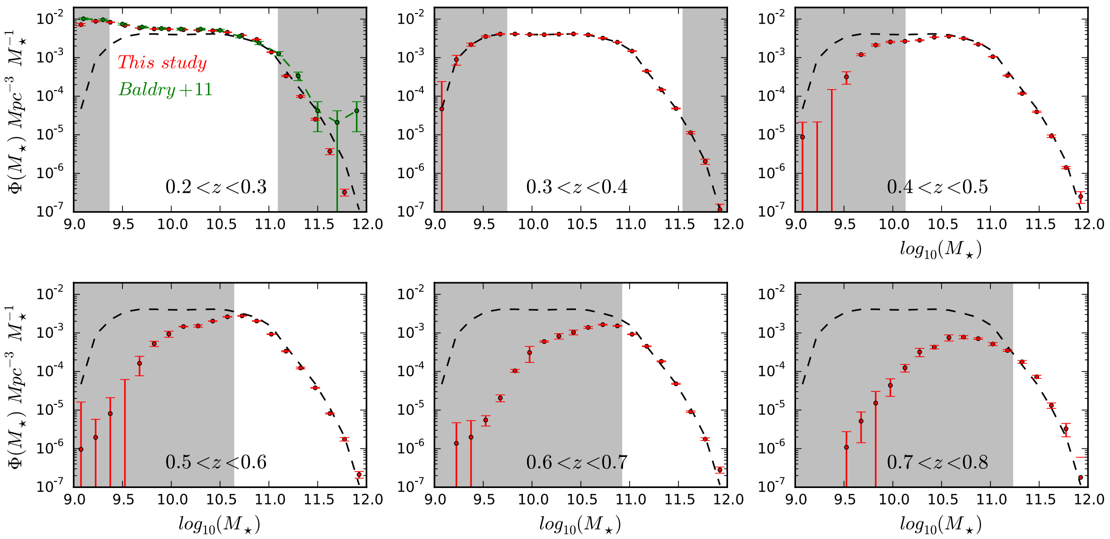

Figure 3.10: Recovered mass functions for real SDSS data in a number of different redshift bins (red). The green points in the redshift bin 0.2 < z < 0.3 are the GAMA mass function (z < 0.06) (Baldry et al., 2012). For reference, the mass function computed using our method in the 0.3 < z < 0.4 bin is shown in all bins as the black dashed line. Regions where our mass functions are less than 95% complete (due to the photometric sample magnitude cut) are shown in grey.

The faint magnitude cut becomes more significant at higher redshifts; however, we are still mostly complete at the very high mass end ($\gtrsim 10^{11}M_\odot$) across the range ($0.4 < z < 0.8$). Tabulated versions of these mass functions are presented in Appendix 3.A.

Mass functions for the lowest redshift bins match closely with GAMA mass functions ($z < 0.06$) over the complete regions, indicating no significant offsets between our masses and GAMA. At the high mass end ($M_\star > 10^{11}M_\odot$), little evolution is evident over the redshift range ($0.4 < z < 0.8$), and the mass function is broadly consistent with GAMA ($z < 0.06$), implying there is no significant enhancement of the high mass end of the mass function after $z = 0.8$.

3.6.2 Luminosity Functions from SDSS

Our computed luminosity functions are shown in Figure 3.11. Magnitudes shown are absolute, dust-corrected magnitudes. Incompleteness is again visible for bright galaxies at low redshifts (due to the $i > 17$ cut); however, beyond redshift 0.4 we are complete for $M_i \lesssim -23.5$, allowing us to compare the evolution of the brightest galaxies across multiple bins.

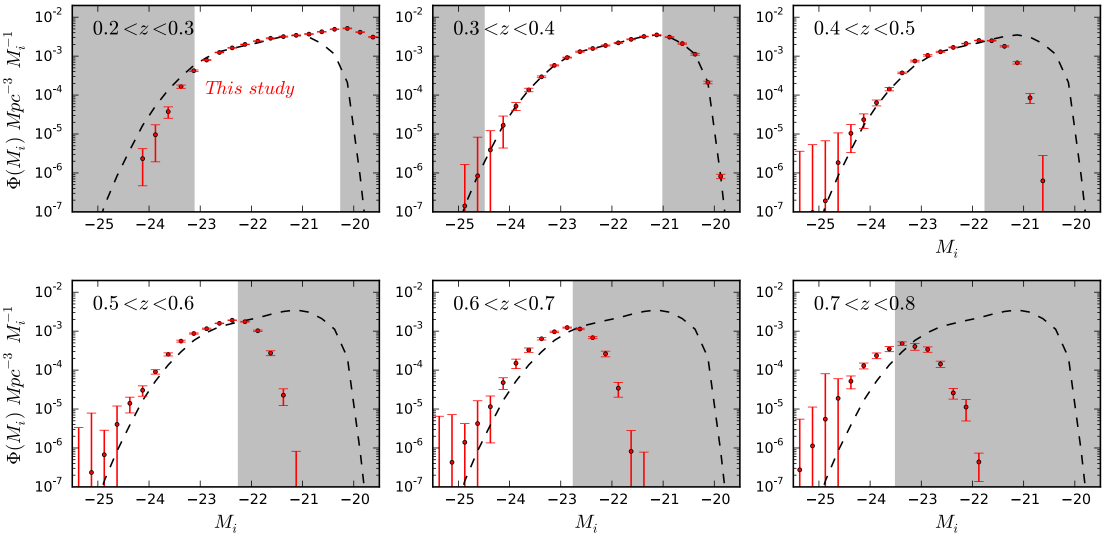

Figure 3.11: Recovered luminosity functions for real SDSS data in a number of different redshift bins (red). The luminosity function for (0.3 < z < 0.4) is shown as the black dashed line in all redshift bins for reference. Regions where our mass functions are less than 95% complete (due to the photometric sample magnitude cut) are shown in grey.

There appears to be a significant amount of evolution over the range ($0.3 < z < 0.8$), with significantly more luminous galaxies present at higher redshifts. If these luminous galaxies are evolving passively, with little ongoing star formation, we would expect their stellar populations to decrease in brightness as young stars die out. Wake et al. (2008) find similar evolution, and find that this is inconsistent with purely passive evolution. Analysis of these and similar luminosity functions as a test of passive evolution may be of interest for future studies.

Notably, van Daalen & White (2018) present a similar method of obtaining a redshift-dependent luminosity functions, using only clustering information and the apparent magnitudes of the galaxies as input. Compared to the method used in this thesis, this has little dependence on assumptions made about the bias evolution of samples of galaxies, however conversely relies on modelling the shape of the luminosity function.

3.6.3 The effect of a bias correction on stellar mass functions

We test how dependent our results are on the choice of unknown sample bias correction as in section 3.5.1. Figure 3.12 shows the fractional change in the mass function after applying three different unknown sample bias corrections (relative to no correction). We use the correction

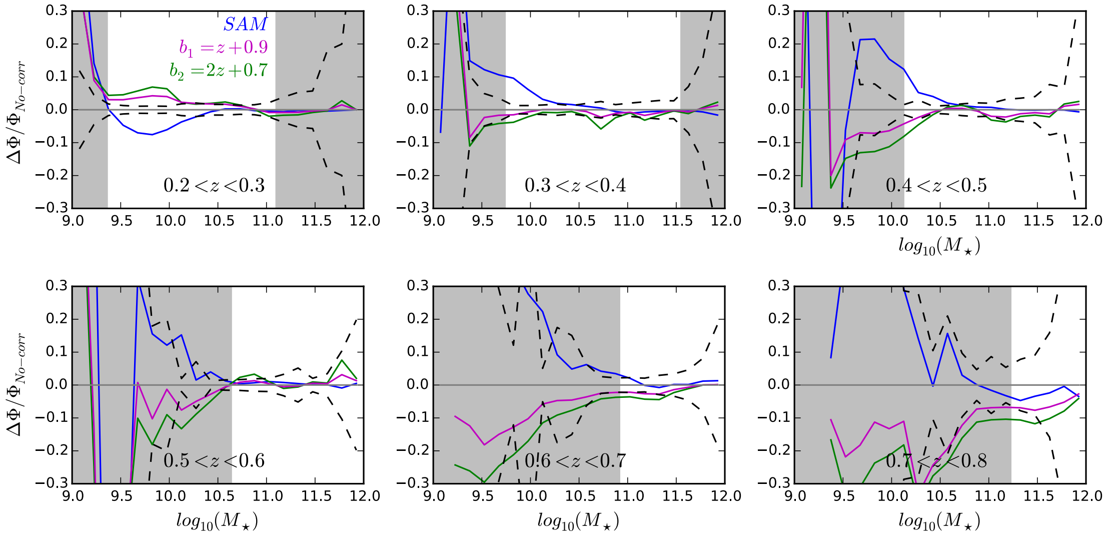

Figure 3.12: The fractional change in the mass function after applying three different unknown sample bias corrections, shown between $0.2 < z < 0.8$. Our correction from L-Galaxies is shown in blue, and two analytic bias corrections are shown in magenta and green. For reference, the size of the error in the mass function is shown as the black dashed line. Regions where mass functions are less than 95% complete (due to the photometric sample magnitude cut) are shown in grey.

computed from L-Galaxies in section 2.5.1, and two different analytic bias laws outlined in section 3.3.1

Some differences are seen in how the different bias laws affect the mass functions, particularly at lower masses, with the two analytic laws predicting fewer low mass galaxies at high redshifts. At higher masses, however, both the SAM and analytic bias corrections only change the mass function by a few percent, which is normally smaller than, or comparable to the size of our mass function errors. In the analysis of future surveys, where clustering errors will be significantly smaller, the choice of bias correction might play a more significant role. For the data presented here, however, the effect is minimal. When tabulating our mass functions, luminosity and completeness estimates in tables 3.A.1, 3.A.2, 3.A.3, we apply no bias correction, but use the maximum offset in the mass function from the three bias laws an estimate of the systematic error due to the unknown sample bias, which can be added to our errors in quadrature.

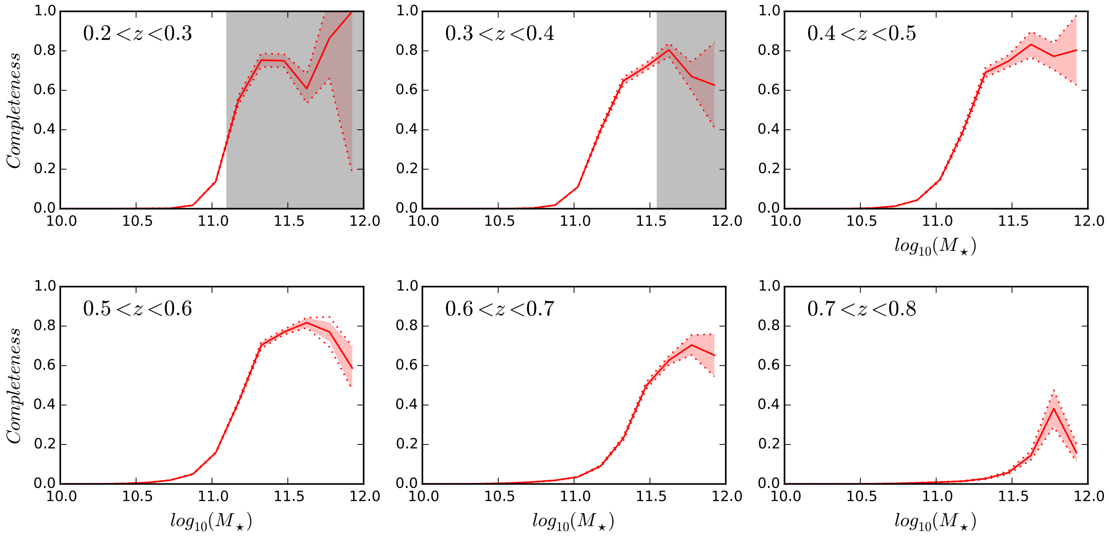

Figure 3.13: Stellar mass completeness estimates for BOSS between $0.2 < z < 0.8$, computed using the SDSS mass functions recovered in section 3.6.1. Completeness estimates are shown as the solid red line. The shaded red region represents the errors due to the clustering redshifts method, and the dotted red line represents the same error, but with our systematic correction added in quadrature. Regions where the mass functions are less than 95% complete (due to the photometric sample magnitude cut) are the grey regions.

3.7 Stellar mass completeness of BOSS

Having computed stellar mass functions out to $z = 0.8$, we can now measure the stellar mass completeness of the BOSS spectroscopic sample. We first take both the SDSS and BOSS masses computed in section 3.6.1. The completeness at a particular redshift is therefore just the mass function of BOSS at that redshift divided by the SDSS mass function. The resulting completeness is displayed in figure 3.13 for 6 bins of redshift between $0.2 < z < 0.8$.

At low redshift $z < 0.4$, our SDSS mass functions are not complete at higher masses due to the bright magnitude cut ($i > 17$). This effect is also true for low masses at higher redshifts due to the faint ($i < 21$) cut. Completeness estimates of BOSS in these regions may not be fully representative and is shown in grey in figure 3.13. Between $0.4 < z < 0.8$, however, we are not affected by these cuts over the mass range of BOSS galaxies.

Between $0.2 < z < 0.7$, the stellar mass completeness of BOSS appears similar across all redshifts. Over this redshift range, above $M_\star \simeq 10^{11.4} M_\odot$, BOSS is roughly 80% complete, with completeness falling to roughly zero at masses lower than $M_\star \simeq 10^{11} M_\odot$. In the $0.6 < z < 0.7$ bin, incompleteness appears at slightly higher masses than in the lower redshift bins. This decrease in completeness mirrors the decrease in number density of the sample shown in figure 3.1, which peaks just above $z = 0.5$ and falls off at higher redshifts. Looking in the highest redshift bin, BOSS is around 30% complete, only at the highest masses ($M_\star \gtrsim 10^{11.6}$).

Guo et al. (2018) incorporate a missing fraction (incompleteness) component into their conditional stellar mass function model, and analyse the clustering of BOSS galaxies to produce completeness estimates for BOSS. They find that BOSS is around 80% complete above $M_\star \gtrsim 10^{11.3} M_\odot$ between $0.2 < z < 0.6$, with completeness falling off significantly at higher redshifts. This analysis is in good agreement with our results, showing very similar evolution with redshift and mass, although some offsets may be present due to using different mass estimates. Leauthaud et al. (2016), discussed in section 3.1, report similar completeness estimates at most redshifts and masses, however predict close to 100% completeness at the highest masses, which is not shown in Guo et al. (2018) or our estimates.

3.8 Conclusions

We show that the clustering redshifts method can be applied successfully to real photometric data from the SDSS, recovering redshift distributions out to $z = 0.8$, as a function of magnitude and colour. We show that these redshift distributions are broadly consistent with distributions from the GAMA survey over the same bins.

We show that stellar masses and luminosities can be computed by binning semi-analytic models in colour and redshift space, and applying these masses and luminosities to galaxies over the same bins of magnitude/colour and redshift. We test how well mass functions are recovered using this method, considering two different semi-analytic models (LGalaxies and SAGE). We show that both models produce similar mass estimates for low to intermediate mass galaxies ($M_\star < 10^{11.25} M_\odot$), however estimates differ at higher masses. After testing how well both models re-produce galaxies in the BOSS survey, we conclude that high mass galaxies in LGalaxies best represent the real universe, so opt to compute masses using this model. We also test how well luminosity functions can be recovered using the same method, finding that both models well recover the luminosity function of the mock data.

We apply this method to real data, recovering mass function and luminosity functions for a large ($\sim 7000\ deg^2$) sample of galaxies from the SDSS ($i < 21$), allowing us to understand their evolution with little sample variance. We find little evolution at high masses between $0.2 < z < 0.8$, suggesting that the most massive galaxies form most of their mass before this time, and do not evolve significantly in mass afterwards. The lack of evolution over these redshifts agrees well with other studies, for example, Pérez-González et al. (2008); Moustakas et al. (2013); Leauthaud et al. (2016); Guo et al. (2018). In our study, the effect of a bias correction on the recovered mass functions is generally comparable to, or smaller than, the error, however this may not be the case for future large-volume surveys. Our luminosity functions show some evolution with redshift, possibly due to passive evolution.

We also produce targeting completeness measurements for BOSS using these mass functions, suggesting that over the redshift range $0.2 < z < 0.7$, BOSS is around 80% complete at high masses ($M_\star > 10^{11.4} M_\odot$), and falling to almost zero below $M_\star < 10^{11} M_\odot$. In our highest redshift bin ($0.7 < z < 0.8$) BOSS is strongly affected by incompleteness, and is only about 30% complete at the highest masses $M_\star \gtrsim 10^{11.6} M_\odot$. We also demonstrate that when comparing mass functions or completeness estimates between methods, significant offsets can be present, which require correction.

Our completeness estimates are in good agreement with Guo et al. (2018), finding that BOSS is around 80% complete above $M_\star \gtrsim 10^{11.3} M_\odot$ between $0.2 < z < 0.6$, with completeness falling off significantly at higher redshifts. Similar completeness estimates are also found in Leauthaud et al. (2016), however higher completeness is found at the highest masses when compared with our estimates or Guo et al. (2018).

Ongoing and future large-volume spectroscopic surveys, for example eBOSS, DESI and EUCLID (Laureijs et al., 2011), will produce large number of spectra out to higher redshifts. This will firstly allow for better clustering redshifts estimates due to having a larger reference sample, but also produce large spectroscopic galaxy samples, for which incompleteness must be understood. Combining these data with ongoing and future photometric surveys, for example, The Dark Energy Camera Legacy Survey (DECaLS) (Dey et al., 2018), and The Dark Energy Survey (DES) (DES Collaboration et al., 2017), will allow for redshift distributions to be computed out to higher redshifts, and in much smaller bins of colour, due to these new surveys reaching much deeper and having much smaller photometric error. This will allow us to not only to understand the completeness of these spectroscopic samples, but also compute stellar mass and luminosity functions over the largest volumes possible.

The methods used in this study, and similar techniques, will therefore be important tools for the next generation of galaxy surveys in order to utilise these large databases, and to understand the galaxy populations present.

Appendices

3.A Tabulated mass functions, luminosity function and completeness

Here we present tabulated versions of our stellar mass functions and i-band luminosity functions in tables 3.A.1 and 3.A.2 respectively, and our completeness estimates in table 3.A.3. In each table, we also present the error in our mass functions due to the clustering redshifts method, and the systematic error due to the bias correction, which can be added together in quadrature.

Table 3.A.1: Tabulated stellar mass functions computed as in section 3.6.1

0.2 < z < 0.3				0.3 < z < 0.4			
$\log(M\star)$ $(M_\odot)$	Φ $(10^{-3}Mpc^{-3})$	Φ_{err}	Φ_{sys}	$\log(M\star)$ $(M_\odot)$	Φ $(10^{-3}Mpc^{-3})$	Φ_{err}	Φ_{sys}
9.375	8.270	0.143	0.362	...	...	...	...
9.525	6.849	0.092	0.322	...	...	...	...
9.675	5.866	0.068	0.411	...	...	...	...
9.825	5.557	0.067	0.420	9.825	4.113	0.048	0.394
9.975	5.476	0.062	0.349	9.975	3.985	0.048	0.246
10.125	5.321	0.099	0.197	10.125	3.946	0.055	0.125
10.275	5.167	0.108	0.110	10.275	4.061	0.068	0.078
10.425	5.001	0.081	0.097	10.425	4.132	0.051	0.065
10.575	4.555	0.074	0.107	10.575	3.896	0.074	0.069
10.725	3.817	0.053	0.055	10.725	3.209	0.080	0.187
10.875	2.965	0.045	0.008	10.875	2.513	0.039	0.060
11.025	1.395	0.030	0.027	11.025	1.479	0.030	0.015
...	...	...	...	11.175	0.448	0.011	0.014
...	...	...	...	11.325	0.147	0.004	0.002
...	...	...	...	11.475	0.048	0.001	0.000

0.4 < z < 0.5				0.5 < z < 0.6			
$\log(M\star)$ $(M_\odot)$	Φ $(10^{-3}Mpc^{-3})$	Φ_{err}	Φ_{sys}	$\log(M\star)$ $(M_\odot)$	Φ $(10^{-3}Mpc^{-3})$	Φ_{err}	Φ_{sys}
10.275	2.832	0.079	0.148	...	...	...	...
10.425	3.391	0.035	0.136	...	...	...	...
10.575	3.577	0.046	0.072	...	...	...	...
10.725	3.162	0.056	0.042	10.725	2.768	0.045	0.064
10.875	2.221	0.046	0.018	10.875	2.038	0.038	0.067
11.025	1.059	0.021	0.033	11.025	0.929	0.018	0.009
11.175	0.344	0.011	0.012	11.175	0.336	0.011	0.003
11.325	0.118	0.004	0.002	11.325	0.123	0.006	0.000
11.475	0.039	0.002	0.001	11.475	0.038	0.001	0.000
11.625	0.0092	0.0007	0.0000	11.625	0.0082	0.0003	0.0000
11.775	0.0014	0.0001	0.0000	11.775	0.0018	0.0002	0.0000
11.925	0.00025	0.00008	0.00000	11.925	0.00021	0.00005	0.00000

0.6 < z < 0.7				0.7 < z < 0.8			
$\log(M\star)$ $(M_\odot)$	Φ $(10^{-3}Mpc^{-3})$	Φ_{err}	Φ_{sys}	$\log(M\star)$ $(M_\odot)$	Φ $(10^{-3}Mpc^{-3})$	Φ_{err}	Φ_{sys}
11.025	0.924	0.020	0.068	...	...	...	...
11.175	0.452	0.009	0.056	...	...	...	...
11.325	0.184	0.005	0.033	11.325	0.178	0.015	0.018
11.475	0.048	0.002	0.019	11.475	0.073	0.007	0.008
11.625	0.0091	0.0004	0.008	11.625	0.013	0.002	0.001
11.775	0.0018	0.0001	0.0001	11.775	0.0033	0.0012	0.0002
11.925	0.00028	0.00005	0.00001	11.925	0.00018	0.00042	0.00007

Table 3.A.2: Tabulated i-band luminosity functions computed as in section 3.6.2

0.2 < z < 0.3				0.3 < z < 0.4			
M_i (mag)	Φ $(10^{-3}Mpc^{-3})$	Φ_{err}	Φ_{sys}	M_i (mag)	Φ $(10^{-3}Mpc^{-3})$	Φ_{err}	Φ_{sys}
...	...	...	...	-24.375	0.0037	0.0063	0.0000
...	...	...	...	-24.125	0.015	0.014	0.001
...	...	...	...	-23.875	0.053	0.013	0.004
...	...	...	...	-23.625	0.131	0.013	0.010
...	...	...	...	-23.375	0.296	0.015	0.026
-23.125	0.417	0.018	0.024	-23.125	0.577	0.018	0.041
-22.875	0.789	0.024	0.036	-22.875	0.907	0.024	0.0414
-22.625	1.213	0.025	0.038	-22.625	1.276	0.043	0.011
-22.375	1.612	0.029	0.030	-22.375	1.513	0.043	0.132
-22.125	1.972	0.037	0.067	-22.125	1.886	0.048	0.143
-21.875	2.435	0.046	0.109	-21.875	2.302	0.049	0.118
-21.625	2.905	0.071	0.122	-21.625	2.804	0.049	0.149
-21.375	3.221	0.078	0.084	-21.375	3.26	0.053	0.197
-21.125	3.445	0.073	0.044	-21.125	3.417	0.107	0.123
-20.875	3.778	0.077	0.080	...	...	...	...
-20.625	4.216	0.071	0.173	...	...	...	...
-20.375	4.968	0.121	0.253	...	...	...	...

0.4 < z < 0.5				0.5 < z < 0.6			
M_i (mag)	Φ $(10^{-3}Mpc^{-3})$	Φ_{err}	Φ_{sys}	M_i (mag)	Φ $(10^{-3}Mpc^{-3})$	Φ_{err}	Φ_{sys}
-25.125	0.000	0.007	0.000	-25.125	0.000	0.007	0.000
-24.875	0.000	0.005	0.000	-24.875	0.001	0.002	0.000
-24.625	0.002	0.009	0.000	-24.625	0.005	0.006	0.001
-24.375	0.010	0.009	0.001	-24.375	0.015	0.008	0.002
-24.125	0.025	0.008	0.002	-24.125	0.029	0.008	0.002
-23.875	0.064	0.011	0.007	-23.875	0.089	0.012	0.004
-23.625	0.147	0.014	0.015	-23.625	0.254	0.024	0.037
-23.375	0.383	0.018	0.024	-23.375	0.534	0.029	0.036
-23.125	0.735	0.035	0.024	-23.125	0.864	0.025	0.037
-22.875	1.053	0.040	0.023	-22.875	1.174	0.025	0.049
-22.625	1.306	0.031	0.064	-22.625	1.567	0.022	0.065
-22.375	1.677	0.030	0.081	-22.375	1.842	0.028	0.082
-22.125	2.132	0.029	0.133	..	..	...	...
-21.875	2.490	0.045	0.160	...	...	...	...
-21.625	2.437	0.078	0.144	...	...	...	...

0.6 < z < 0.7				0.7 < z < 0.8			
M_i (mag)	Φ $(10^{-3}Mpc^{-3})$	Φ_{err}	Φ_{sys}	M_i (mag)	Φ $(10^{-3}Mpc^{-3})$	Φ_{err}	Φ_{sys}
-25.125	0.000	0.007	0.000	-25.125	0.001	0.014	0.000
-24.875	0.001	0.001	0.000	-24.875	0.005	0.015	0.000
-24.625	0.004	0.013	0.000	-24.625	0.018	0.024	0.000
-24.375	0.011	0.009	0.001	-24.375	0.033	0.018	0.000
-24.125	0.044	0.014	0.001	-24.125	0.124	0.026	0.001
-23.875	0.148	0.014	0.007	-23.875	0.229	0.023	0.007
-23.625	0.341	0.036	0.016	-23.625	0.388	0.036	0.016
-23.375	0.579	0.028	0.025	...	...	...	...
-23.125	0.949	0.025	0.054	...	...	...	...
-22.875	1.243	0.050	0.085	...	...	...	...

Table 3.A.3: Tabulated stellar mass completeness for BOSS computed as in section 3.7

0.2 < z < 0.3				0.3 < z < 0.4			
log($M_\star$) ($M_\odot$)	*Completeness*	σ_{comp}	σ_{sys}	log($M_\star$) ($M_\odot$)	*Completeness*	σ_{comp}	σ_{sys}
10.575	0.0015	0.0001	0.0000	10.575	0.0009	0.0001	0.0000
10.725	0.0020	0.0001	0.0000	10.725	0.0027	0.0001	0.0001
10.875	0.0178	0.0003	0.0000	10.875	0.0180	0.0003	0.0002
11.025	0.1380	0.0028	0.0011	11.025	0.1104	0.0016	0.0004
...	...	...	...	11.175	0.3997	0.0085	0.0068
...	...	...	...	11.325	0.6478	0.0158	0.0035
...	...	...	...	11.475	0.7198	0.0183	0.0022

0.4 < z < 0.5				0.5 < z < 0.6			
log($M_\star$) ($M_\odot$)	*Completeness*	σ_{comp}	σ_{sys}	log($M_\star$) ($M_\odot$)	*Completeness*	σ_{comp}	σ_{sys}
10.575	0.0031	0.0001	0.0001	...	...	...	...
10.725	0.0124	0.0002	0.0001	10.725	0.0190	0.0004	0.0002
10.875	0.0438	0.0009	0.0001	10.875	0.0501	0.0010	0.0009
11.025	0.1473	0.0025	0.0028	11.025	0.1586	0.0024	0.0005
11.175	0.3893	0.0095	0.0078	11.175	0.4146	0.0095	0.0009
11.325	0.6904	0.0204	0.0044	11.325	0.7030	0.0162	0.0005
11.475	0.7469	0.0281	0.0025	11.475	0.7699	0.0111	0.0021
11.625	0.8322	0.0637	0.0025	11.625	0.8172	0.0251	0.0005
11.775	0.7719	0.0660	0.0013	11.775	0.7703	0.0484	0.0274
11.925	0.8036	0.1761	0.0012	11.925	0.5874	0.1085	0.00066

0.6 < z < 0.7				0.7 < z < 0.8			
log($M_\star$) ($M_\odot$)	*Completeness*	σ_{comp}	σ_{sys}	log($M_\star$) ($M_\odot$)	*Completeness*	σ_{comp}	σ_{sys}
11.025	0.0357	0.0007	0.0007	...	...	...	...
11.175	0.0909	0.0018	0.0025	...	...	...	...
11.325	0.2318	0.0061	0.0058	11.325	0.0270	0.0019	0.0016
11.475	0.4971	0.0137	0.0042	11.475	0.0576	0.0046	0.0038
11.625	0.6266	0.0212	0.0011	11.625	0.1451	0.0177	0.0054
11.775	0.7042	0.0499	0.0007	11.775	0.3818	0.0888	0.0050
11.925	0.6521	0.1070	0.0003	11.925	0.1583	0.0458	0.0005

4

Halo formation from observables

In the following chapter, we investigate how the formation history of a dark matter halo is correlated with the observable properties of a galaxy. We investigate these relations in SAMs before training a machine learning algorithm to infer formation time based on observable properties. We then apply this trained algorithm to real data, inferring halo formation times for the GAMA survey in order to investigate whether the formation times of halos of fixed mass change with environment.

4.1 Introduction

Hierarchical structure growth is a basic prediction of the ΛCDM model (see section 1.1.2), where dark matter particles collapse under gravity forming halos, and halos merge with each other through time. In our Universe, the clustering strength of the majority of these halos is significantly higher than the overall mass distribution (Kaiser, 1984; Cole & Kaiser, 1989; Kauffmann et al., 1997; Mo & White, 1996; Sheth & Tormen, 1999). This increased clustering is often referred to as *halo bias* (see section 2.1.2). The Press & Schechter (1974) model has been long used to explain halo formation, and predicts that the bias of halos depends only on their mass. It is possible to test this in simulations, which show a wide variety of assembly histories for halos of the same mass, and many studies find small dependencies on halo clustering (at constant mass) with other properties, for example halo formation history, structure, or spin (Gao et al., 2005; Wechsler et al., 2006; Gao & M., 2007; Li et al., 2008; Chaves-Montero et al., 2016). When looking at the highest mass halos, the older of these (i.e. formed earliest) appear to be less biased, and at lower masses the trend is reversed with older haloes being more biased (Gao et al., 2005; Wechsler et al., 2006; Hahn et al., 2009). This is referred to as *halo assembly bias* (as described in section 1.1.3).

Since galaxies form and evolve within these halos, their properties (including bias) are therefore dependent on the type of halo in which they form, and observed galaxy properties (e.g. colour, mass, star formation rate) are seen to be strongly correlated with clustering (Norberg et al., 1988; Cresswell & Percival, 2009; Zehavi et al., 2005; Christodoulou et al., 2012). Assuming no halo assembly bias, these differences are fully explainable by the fact that these properties are correlated with halo mass, which in turn is correlated with clustering strength. If at fixed halo mass, galaxy properties depend on secondary halo properties, then this implies assembly bias. This manifestation of assembly bias in galaxy properties is usually referred to as *galaxy assembly bias* (Zentner et al., 2014; Wechsler & Tinker, 2018), which has important implications for halo occupation distribution and abundance matching methods, since these usually assume simple relations between halo mass and galaxy observables.

Given that halo assembly bias is seen in simulations, an important question is whether it is important in the real Universe too. Some studies appear to have detected signals of this Yang et al. (2006); Wang et al. (2013); Tojeiro et al. (2017); Montero-Dorta et al. (2017), although other studies find no dependence on parameters other than halo mass (Skibba et al., 2006; Abbas & Sheth, 2006; Zu & Mandelbaum, 2016). Investigating assembly bias in the real Universe is difficult, particularly since halo properties are difficult to measure in the real Universe, and we must therefore rely heavily on the directly observable properties of galaxies to test this.

In this chapter, we seek to investigate how one of these features in particular, the formation history of a halo, depends on both the environment of the halo, and also the properties of the galaxies in the halo. If a halo of a given mass assembles its mass very early, it follows that this may affect the properties of the galaxy (e.g. the star formation history may be affected due to lots of early-time mergers). We investigate these correlations in simulations, before investigating how well halo formation time can be predicted using properties of the galaxy. This is done with the view of attempting to compute formation times of real galaxies in different environments to look for a difference, and hence a signal of halo assembly bias.

Tojeiro et al. (2017) investigate in simulations how different observable parameters correlate with formation time, before using these relations as an estimate of formation time for real data. They find that stellar to halo mass ratio correlates most strongly with formation time, however correlations are seen with formation time for other parameters. Since these other parameters may possess additional information, it follows that formation time may be better predicted by considering several parameters at once. In this case, the relationship between parameters will become much more complex, and qualitative analysis or simple fits to the data cannot be used.

In recent years, machine learning has become increasingly popular in astronomy for problems such classification or regression problems. Studies have used machine learning algorithms for, amongst other things, photometric redshift (Csabai et al., 2003; Ball et al., 2008; Carliles et al., 2010; Gerdes et al., 2010b) and halo mass prediction (Calderon & Berlind, 2019), image classification (Banerji et al., 2010; Hocking et al., 2018; Pasquet et al., 2019), and detection of transients (Mahabal et al., 2008). Here, we apply machine learning techniques to simulations in order to both investigate the relationships between these parameters, and also train a model capable of computing formation times for real galaxies based on a number of observable parameters.

4.2 Machine learning

4.2.1 Overview

Machine learning is the name given to a broad class of algorithms which, after being applied to data, perform some task without explicitly being told how to achieve this. These algorithms "learn", recognising patterns without specific instructions. This can be particularly useful in applications where patterns are very complex, or in cases where there are vast amounts of data, too large for a human to process. These algorithms have a variety of uses outside of astronomy, including, amongst other things, natural language processing, recommendation systems, trading, and image classification. These techniques have become increasingly more commonly used in the last decade, both inside and outside of astronomy.

Machine learning tasks can be broadly classified in to two categories: supervised and unsupervised learning. Supervised learning involved passing an algorithm data with some desired known output (e.g. images with labels describing what the image contains: dog, cat etc.). The algorithm learns to recognise patterns in the data and then, once trained, can be applied to new data (e.g. raw images in order to generate labels). Supervised learning is well suited to classification and regression problems, and is now often used in astronomy for many different tasks. The other broad category is unsupervised learning, where only input data is given to the algorithm (e.g. images without labels). These algorithms again recognise patterns or structure in the data, however since no specific output is given, they will group the data into characteristically similar groups (e.g. groups all images containing small animals together). These algorithms can be good for reducing dimensionality, or classifying data when specific categories are not needed or known. These have various applications advertising, document clustering, fraud detection etc.

Due to the large amounts of data required in modern astronomical studies, and the complexity of the problems in question, astronomers have been increasingly turning to machine learning. Some widely used applications of machine learning in astronomy (predominantly applying supervised learning) include: estimating photometric redshifts (Csabai et al., 2003; Ball et al., 2008; Carliles et al., 2010; Gerdes et al., 2010b) (see section 1.3.3 for a detailed description of this), classifying galaxy morphology (Banerji et al., 2010; Hocking et al., 2018; Pasquet et al., 2019), cluster mass estimates (Ntampaka et al., 2016, 2018; Armitage et al., 2019), halo mass estimates (Calderon & Berlind, 2019), and classificaiton of transient events (Mahabal et al., 2008).

In this study, we apply a method similar to Calderon & Berlind (2019), applying machine learning to simulated observations to predict the formation time of dark matter halos. To do this, we opt to use random forests.

4.2.2 Random forests

Random forests are a machine learning algorithm used for classification and regression tasks. The idea is that multiple decision trees are constructed on training data, and then the average classification is outputted as the classification of the forest.

A decision tree is a class of algorithm designed to perform either *classification* (classifying an input in to one of several discrete outputs), or *regression* (where the output can instead take continuous values). Decision trees work by taking data with some input values (e.g. size and weight of an animal), and making a series of random splits in this parameter space (e.g. is the animal taller than 1m?). The data is sorted down these *branches* depending on its input value, and after many different splits have been made, the data will have been sorted in to many different groups (or *leaves*), the minimum size of which can be specified. The value of the output (e.g. the type of animal) in each of these leaves describes the prediction of the decision tree. After being trained, this decision tree can be applied to new data, which after being passed through the tree, will be classified in to the outputs of the trained decision tree.

Decision trees that have a very large number of branches tend to learn irregular patterns (i.e. they overfit the patterns in the training sets), so may not perform well on new data. Random forests, first outlined in Ho (1995), are a way of using multiple decision trees, trained on different sub-samples of the training data in order to better improve the prediction. Random forests therefore rely on bootstrapping, where the output for new data is computed from the average of the trained trees. Random forests further improve on this, since for each random subsample of the training data, a random subset of the total features are chosen (often referred to as feature bagging) in order to lessen the reliance on particular parameters. A more detailed discussion of this method is presented in Ho (2002). A somewhat related type of algorithm are boosting algorithms, where trees are assigned weights based on the accuracy of their predictions, and the weighted averages of each of their outputs are used to compute the final predictions.

In this study, we opt to use random forests as our supervised learning method because, firstly, they are simple to implement and optimise (with only a few free parameters, e.g. the number of trees, depth of trees), and secondly, as shown in later sections, when adding observational errors to data, correlations between parameters are weakened, so more complex methods (e.g. neural networks) will likely perform similarly.

4.3 Data

Since we are seeking to investigate correlation between galaxy observables and halo formation time, and train a machine capable of predicting this, we require a mock data set in which to investigate this. In order to then invesitage formation time in the real Universe, we also require an equivalent real data sample. We outline simulated and real samples used in the following sections.

4.3.1 Simulated data

To investigate how correlated galaxy parameters are with halo formation time, and also to train our machine learning algorithm, we use data from the LGalaxies SAM (Henriques et al., 2015), described in detail in section 1.2.3. We define a sample of halos in the simulation with mass $M_{halo} > 10^{10.5}$ at $z = 0.15$ (comparable to the median redshift of the GAMA survey) which will then apply this to. We also restrict the sample to galaxies that are classified as centrals (i.e. the brightest galaxy in the cluster). This produces a sample comparable to that used in Tojeiro et al. (2017).

We also obtain a number of observable properties for the galaxies in each halo: stellar mass, star-formation rate, mass-weighted-age, and star-formation history. The details of these properties are described in full in this section. In later sections we also compute realistic errors for these parameters (see section 4.4.4), using the output of the spectral fitting code VESPA (Tojeiro et al., 2007) applied to mock galaxy SEDs in the simulation. These SEDs are computed using Maraston & Strömbäck (2011) stellar population models (SSPs), using a Kroupa et al. (2001) initial mass function (IMF), and convolved with SDSS photometric filters to produce SDSS magnitudes.

Properties

Here we outline the properties considered in this study. Much of this work is based upon the work done by Tojeiro et al. (2017), and we choose to consider a number of similar galaxy and halo properties. We obtain halo masses for the simulation and compute halo formation times for all galaxies. Since there is no single way to define when a halo has "formed", we investigate three different definitions of this: f_{half}, f_{vmax}, and f_{core}, defined as follows:

- f_{half}: The time at which the halo formed half of its present day mass.

- f_{vmax}: The time at which the halo reached its maximum virial velocity.
- f_{core}: the earliest time at which the halo's mass reached $M_{halo} > 10^{11} M_{\odot}$.

These formation time definitions therefore describe different things and hence will be correlated differently with observable galaxy parameters. Li et al. (2008) present a detailed comparison of these definitions, and 5 others, however we summarise the motivation of our chosen parameters here: f_{half} conveys the hierarchical formation of halos, showing the half-way point of mass accretion, f_{vmax} shows the time which the halo reached maximum virial velocity, hence represents the time at which halo accretion begins to slow down. f_{core} represents roughly the time at which a halo is massive enough to host a bright central galaxy.

In order to investigate how these correlate with galaxy observables, we define a number of different parameters for each galaxy in our sample:

- $M_{\star}$: The total stellar mass of the galaxy in units of solar mass, $M_{\odot}$
- SFR: The star-formation rate of the galaxy in units of $M_{\odot} Yr^{-1}$
- $sSFR$: The specific star-formation rate; i.e. the star formation rate of the galaxy divided by its stellar mass
- $s-h\ ratio$: The ratio of stellar mass to halo mass for the galaxy
- MWA: The stellar mass weighted age of the galaxy.
- SFH: The star-formation history of the galaxy in 13 logarithmically spaced bins, measuring the mass of stars formed across each time bin of the simulation.

4.3.2 GAMA

When applying our method to real data, we make use of data from the GAMA survey (described in section 1.3.2, which is a highly complete spectroscopic survey, magnitude limited to $r < 19.8$, targeted over ∼286 deg^2 of sky. Following Tojeiro et al. (2017), we select grouped central galaxies to ensure that our selection contains only the most massive galaxy in the group, and hence its evolution is not dominated by a nearby, more massive galaxy. We do this using the (Robotham et al., 2011) group catalogue, computed using a friends-of-friends (FoF) algorithm, applying the cut $0.04 < z < 0.263$ to ensure the robustness of the geometric environment classifications. We use halo mass estimates from Han et al. (2015), where power-law combinations of six physical parameters are investigated as predictors of group halo mass when matched to weak lensing masses. We use halo masses computed from group luminosity, which are computed in the r-band, and are corrected for the fraction of light in galaxies below the survey flux limit using the GAMA luminosity function (Robotham et al., 2011).

It is worth noting that it is possible that if halo masses are significantly correlated with some additional galaxy properties beyond group luminosity (for example galaxy colour), then this could introduce systematic dependencies when comparing halo masses across different samples (if for example the two samples have very different colours). A full investigation in to the significance of this effect is not considered in this thesis, however we not that this could introduce some degeneracies when weighting galaxies by halo mass in section 4.5 across different bins of geometric environment.

To obtain physical parameters for our galaxies, we obtain catalogues containing the output of the VESPA spectral fitting code (Tojeiro et al., 2007) applied to GAMA spectra, giving us $M_\star$, MWA, SFR, sSFR, and SFH in 16 logarithmically spaced bins of lookback time. Since our simulated data contains SFHs, instead, over 13 bins, we linearly interpolate GAMA SFHs in to the same bins as the simulated data (i.e. 13 logarithmically spaced bins). When comparing masses of galaxies from VESPA to those from the simulation, we notice an offset. A similar offset is also present when comparing VESPA masses to the estimates of Taylor et al. (2011), computed using photometry and a library of SFHs. These offsets are likely due to differences in the fitting method, or the fact that spectroscopic masses are computed only from regions of the galaxy overlapping the fiber which may have a different mass-light ratio. In order to ensure that stellar masses are comparable in our real data and training data, we apply offsets to our spectroscopic masses such that they match the simulation. To do this, we define a sample of galaxies comparable to the GAMA survey in our simulation ($r < 19.8$), and apply an offset to VESPA masses until the high mass tail of both distributions overlap.

In order to investigate how environment affects the formation time of a halo, we make use of two different environmental measures. We firstly consider the geometric environment classifications from Eardley et al. (2015), who compute the tidal tensor from a smoothed galaxy density field. Classifications of environment are then defined for each galaxy by comparing the the computed eigenvalues to some threshold value. The four definitions describe the dimensionality of collapse: a void if all eigenvalues are below the threshold (no collapse), a sheet if only one eigenvalue is above this (one dimensional collapse), a filament if two eigenvalues are greater than this (two dimensional collapse), and a knot if all eigenvalues pass the threshold (collapse in all dimensions). Following Tojeiro et al. (2017), we use a smoothing scale of $4h^{-1}Mpc$ for the density field, and a threshold value for the eigenvalues of 0.4.

Our second definition of environment is from Kraljic et al. (2018) where the DisPerSE algorithm (Sousbie, 2011; Sousbie et al., 2011) is applied to GAMA data to quantify their distance to the nearest node, filament, or wall. A full description of this process applied to GAMA data is presented in Kraljic et al. (2018), however we present a summary here. DisPerSE (Sousbie, 2011; Sousbie et al., 2011) uses Delaunay tessellation, a way of defining multiple triangular volumes between a set of points, in order that cells, faces, edges and vertices are defined across the survey volume. The density field is extracted from the critical points (e.g. maxima and minima) and the different areas of the cosmic web (voids, walls, filaments, nodes) can then be defined from this field. In our studies we primarily use the the distance to nearest node d_{node}. The accuracy of the reconstruction of the cosmic web features is sensitive to the sampling of the dataset, so, following Kraljic et al. (2018), all the distances used in this chapter are normalised by the redshift dependent mean inter-galaxy separation $\langle D_z \rangle$, where $\langle D_z \rangle = n(z)^{1/3}$, and $n(z)$ represents the number density of galaxies at a given redshift z. For GAMA, $\langle D_z \rangle$ varies from 3.5 to 7.7 Mpc, with a mean value of ~ 5.6 Mpc.

4.4 Formation time from galaxy parameters

4.4.1 Formation times

We firstly investigate how halo mass is correlated with formation time for each of the three definitions. We compute the three definitions of formation times f_{half}, f_{vmax}, and f_{core}, for all halos in our sample, defined as the lookback time from $z = 0$. We present plots of halo mass vs formation time for each definition in figure 4.1.

Looking at both f_{half} and f_{vmax}, higher mass halos appear to form later. This seemingly describes the hierarchical nature of structure growth where halos gradually build up mass, and more massive halos keep accreting later in a “bottom-up” scenario. Looking at f_{core}, conversely, there is a strong positive correlation of formation time with halo mass, with high mass halos forming earlier. Since f_{core} equates to the halo mass needed to host a bright central galaxy, massive galaxies then appear to form much earlier in massive haloes. This reversal of trend compared with f_{half} and f_{vmax} is likely due to the constant mass scale considered for f_{core}, showing that since high mass halos passed this threshold earlier, they must have begun

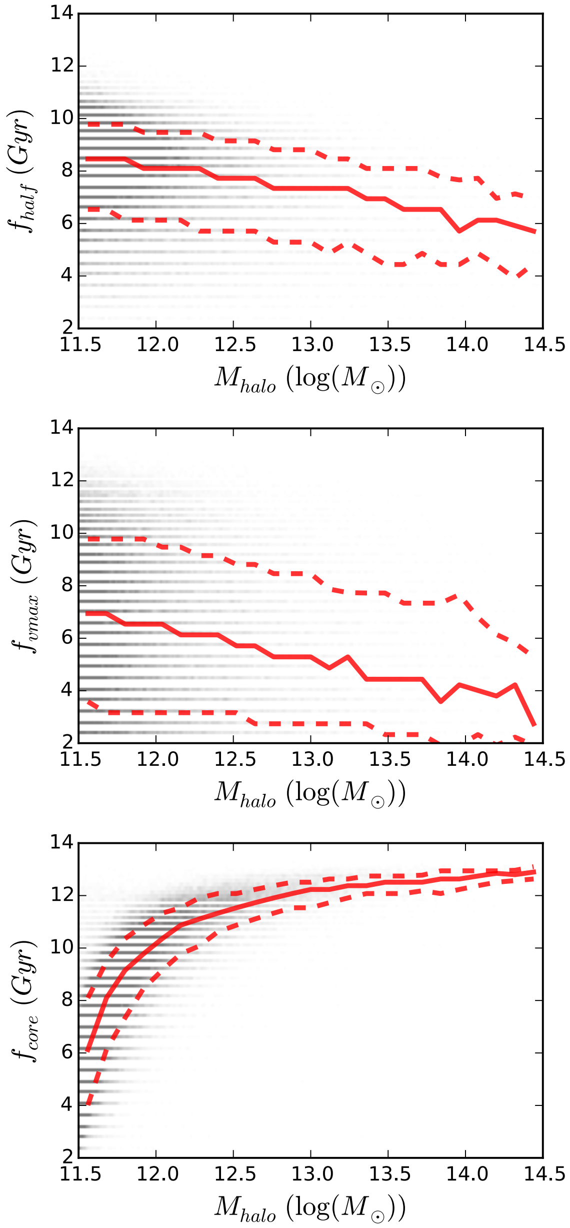

Figure 4.1: Halo mass vs three different definitions of halo formation time: f_{half} (top), f_{vmax} (middle), and f_{core} (bottom). Formation times are defined as a lookback time from $z = 0$. Individual points are shown in black, and a running median and standard deviation (as a function of halo mass) is shown in red. Halo mass is clearly correlated with all three definitions of formation time.

assembling earlier. These differences outline the fact that care must be made when choosing a definition of halo formation, since each definition captures different elements and time-scales of the process.

We also investigate how halo mass is correlated with each of the galaxy parameters described in section 4.3.1. This is shown in figure 4.2 for our three definitions of halo mass. All parameters seem to show strong trends with halo mass. Stellar mass increases with increasing halo mass as expected, since galaxies in high mass halos are generally massive and bright. MWA, SFR and sSFR all show similar trends, following from the fact that galaxies in more massive halos generally have older stellar populations and little ongoing star formation.

4.4.2 One parameter fits

Here we investigate which individual galaxy observables are most correlated with halo formation time, and hence, which parameters will be most useful when attempting to recover formation times. We plot all of our observable parameters against the three definitions of halo formation time in figure 4.3. Note that we do not plot the full star formation history, since this is an array of values, however some of this information will be captured by the sSFR (recent star formation) and MWA (the average age of the stellar population).

Almost all parameters show some correlation with formation time, for all three definitions of formation time, however SFR in particular shows almost no correlation except with f_{core}. f_{vmax} appears the most weakly correlated with the five observable, whereas f_{core} shows the strongest. Looking at figures 4.2 and 4.1 it is possible that much of this correlation is due to correlations between the observables and halo mass (which is in turn correlated with formation time).

If we are looking for signs of assembly bias, we are interested in how useful each of these parameters are in describing halo formation history at fixed halo mass. Therefore, we quantify how good a predictor each parameter is, in a number of different bins of halo mass. We split the data into bins of size $\Delta log(M_{halo}) = 0.2$, and then further split data from each bin in to a “training” and “test” sample, of size 80% and 20% respectively. We do this so we can fit a relation between each parameter and formation time in the training data, and then apply this relation to the same parameter in the test data to see how well formation times can be inferred. We take each parameter in turn, and for every halo mass bin, fit a 2nd order polynomial (e.g. to

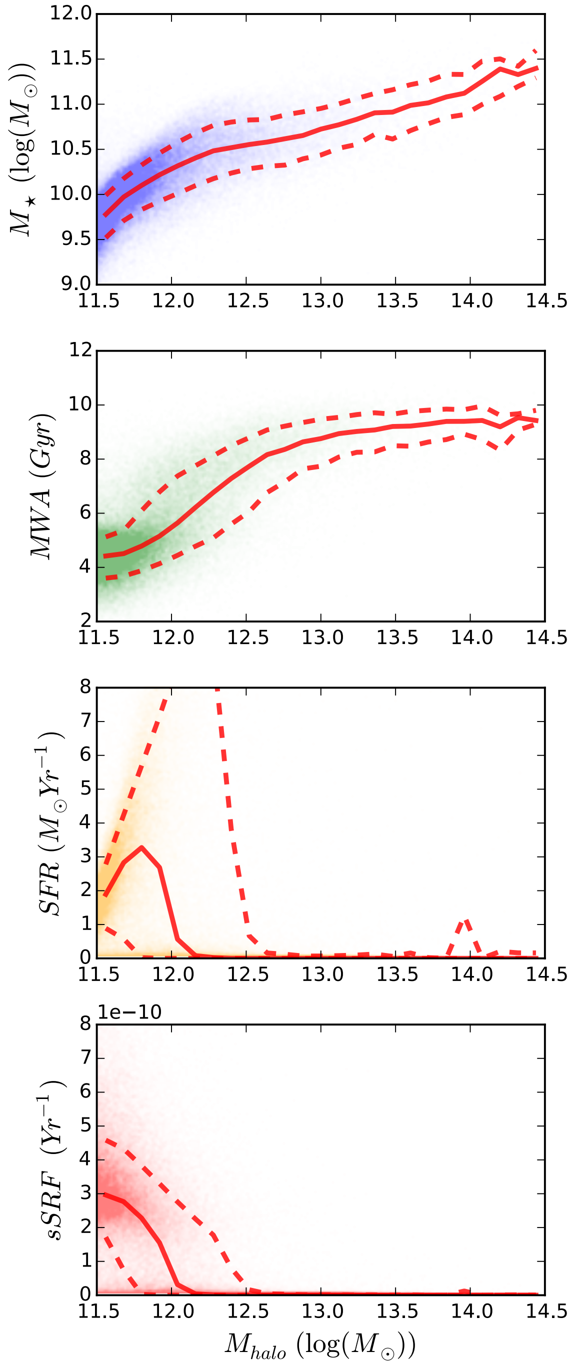

Figure 4.2: Halo mass vs each of the parameters considered in this study: Stellar mass, mass-weighted age, star formation rate, and specific star formation rate. Individual galaxies are shown as points, and a running median and standard deviation (as a function of halo mass) is shown in red. Clear correlations are shown with all parameters.

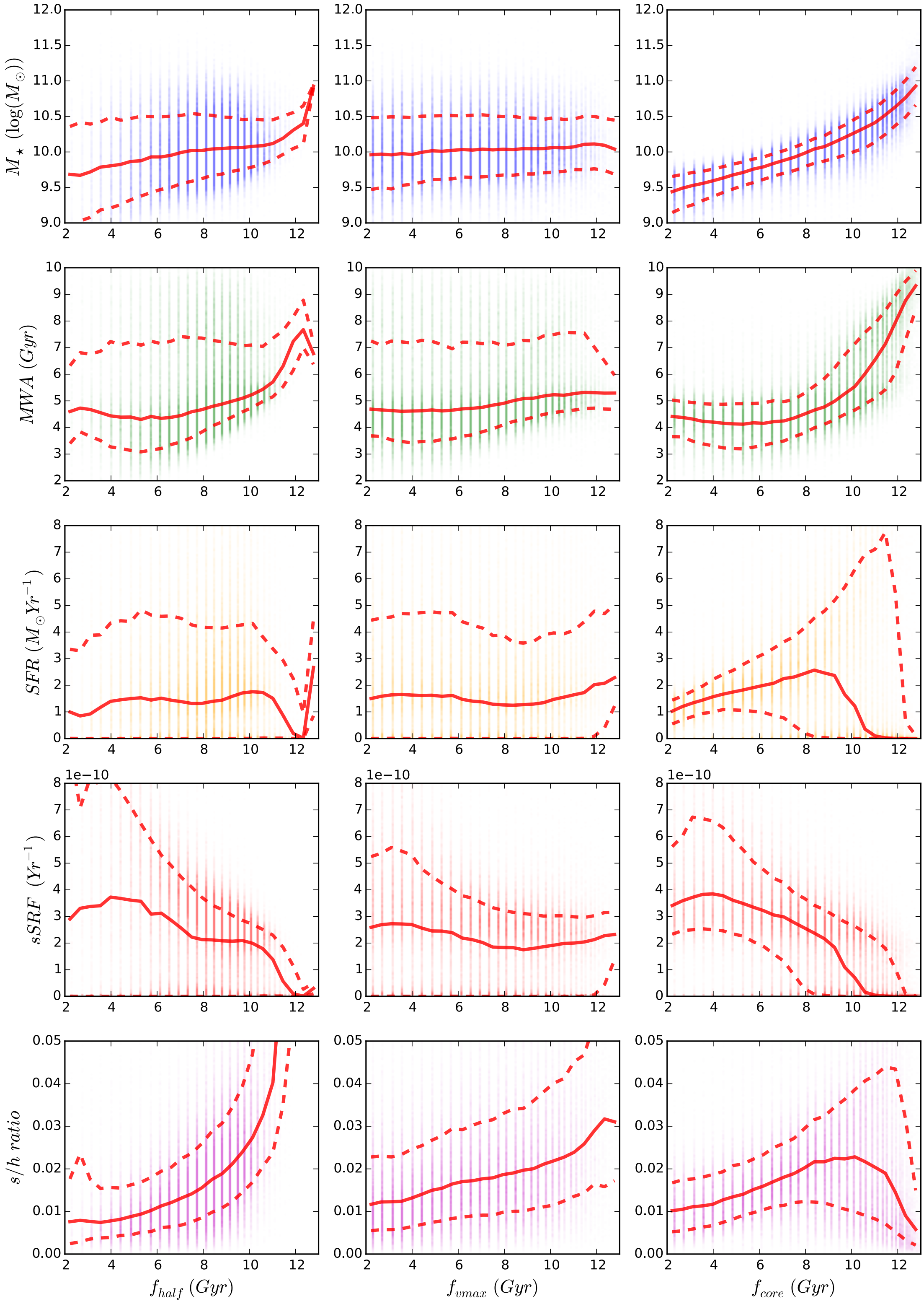

Figure 4.3: Halo formation times (f_{half} (left), f_{vmax} (middle), and f_{core} (right)) vs a number of different observational paramaters: Stellar mass, MWA, SFR, sSFR, and stellar-halo mass ratio. Individual galaxies are plotted as translucent points, and a running median and standard deviation (as a function of halo formation time) is shown in red.

$M_\star$ vs formation time). We then use this fit to infer formation times for the test data such that we can compare this to the true value. This is done for all bins of halo mass, all parameters, and also done for our three different definitions of formation time.

In order to test how effective each parameter is at reproducing halo formation time in each bin, we compute the Root Mean Squared (RMS) error, σ_{RMS}, for each of our test samples:

$$\sigma_{RMS} = \sqrt{\frac{\sum_{n=0}^{N}(f_{tr,n} - \hat{f}_n)^2}{N}}$$

where N is the number of objects in the test sample, $f_{tr,n}$ is the true formation time, $\hat{f}_n$ is the formation time inferred from the parameter. This therefore describes the square root of the average difference between the true and inferred values squared, hence is therefore comparable to a standard deviation. We present the resulting RMS errors as a function of halo mass (for all parameters and formation time measures) in figure 4.4.

It is firstly apparent that, for all three formation time measures, it is easier to infer formation time at higher halo masses, particularly for f_{core}. This is likely due to the fact that the most massive halos form early (halo downsizing), with less massive halos forming later, and with a greater distribution of formation times. This follows figure 4.1, our plot of halo mass vs formation time, showing a higher standard deviation in formation time at lower masses, particularly for f_{core}.

It is also apparent that observational parameters are only slightly better at inferring formation time than halo mass alone. SFR appears to perform particularly badly, showing no increase in accuracy over simply considering halo mass. This follows from figure 4.3, where SFR is shown to be highly uncorrelated with formation time. Other parameters perform slightly better (again agreeing with figure 4.3), constraining formation time better at lower halo masses, where halo mass alone is not a good predictor. This implies that formation time has a greater effect on observed galaxy parameters at lower halo masses, with high mass halos of different formation time appearing observationally similar.

4.4.3 Random forest predictions

Since we have quantified how well formation time can be inferred by individual parameters, we now investigate how well it can be inferred by considering multiple parameters at once. To

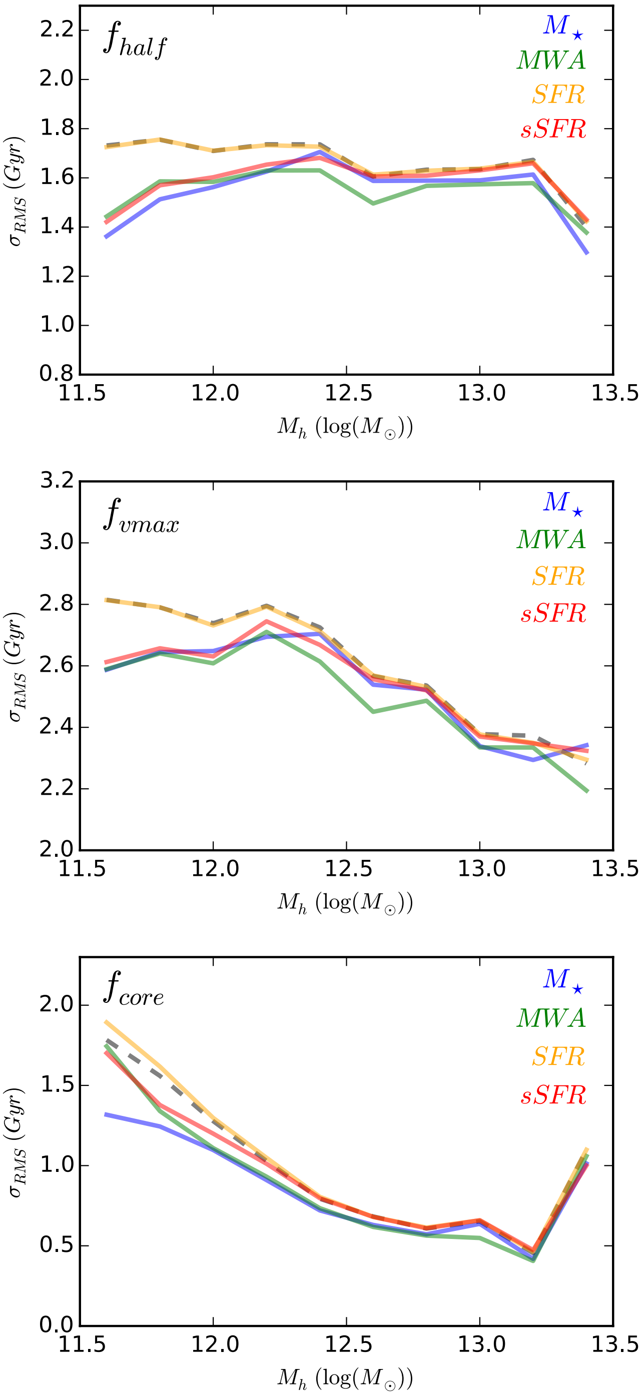

Figure 4.4: Shows the RMS error as a function of halo mass when trying to compute halo formation time from single parameters: stellar mass (blue), mass-weighted age (green), star formation rate (yellow), and specific star formation rate (red). The prediction from halo mass alone is shown as the grey dashed line. This is shown for three different definitions of formation time: f_{half} (top), f_{vmax} (middle), and f_{core} (bottom).

do this, we use a random forest (see section 4.2.2). Equivalently to before, we split data into bins of halo mass of size $\Delta log(M_{halo}) = 0.2$, and further split these samples into "training" and "test" samples, of size 80% and 20% respectively. Instead of fitting a polynomial, we train a random forest to infer our three definitions of formation time in each mass bin, using several combinations of parameters. We try three different combinations of paramteters: stellar mass, MWA and sSFR, all of these plus s/h ratio, and all these plus the full 13 bin SFH. We do not add SFR since in our preliminary tests it is bad predictor, and furthermore, the information for SFR is fully contained within the stellar mass and sSFR. For the same reason, we do not use halo mass (since we have s/h ratio and stellar mass).

Again, we compute RMS errors for all bins of halo mass such that results can be compared with figure 4.4. We run several random forests with different numbers of trees and minimum leaf sizes, finding that a minimum leaf size of ∼20 performs best, and finding no change in results with increasing number of trees. Results are therefore presented from a forest with 200 trees and minimum leaf size of 20. We present the computed RMS errors in prediction in figure 4.5.

It is firstly apparent that when combining multiple parameters, formation time can be constrained significantly better compared to considering single parameters (as in figure 4.4). Looking at both f_{half} and f_{core}, RMS error in our random forest predictions is roughly halved at lower masses compared with predictions from halo mass alone. RMS error is only slightly smaller at higher halo masses, again suggesting that formation time has only a small effect on galaxy properties here. f_{vmax} remains very difficult to constrain across all halo masses, although some improvement is shown when considering multiple parameters.

Looking at f_{half} and f_{vmax}, simply considering $M_{\star}$, MWA, and $sSFR$ works very well. Adding s/h ratio improves results only slightly, likely because we have already binned by halo mass, so evolution in formation time with halo mass is small. Adding the full SFH further improves results, implying that each parameter provides some different information. Now considering f_{core}, formation time is well inferred by $M_{\star}$, MWA, and $sSFR$. Adding s/h ratio and SFH improves upon this negligibly, implying that all the information required is contained in the first three parameters.

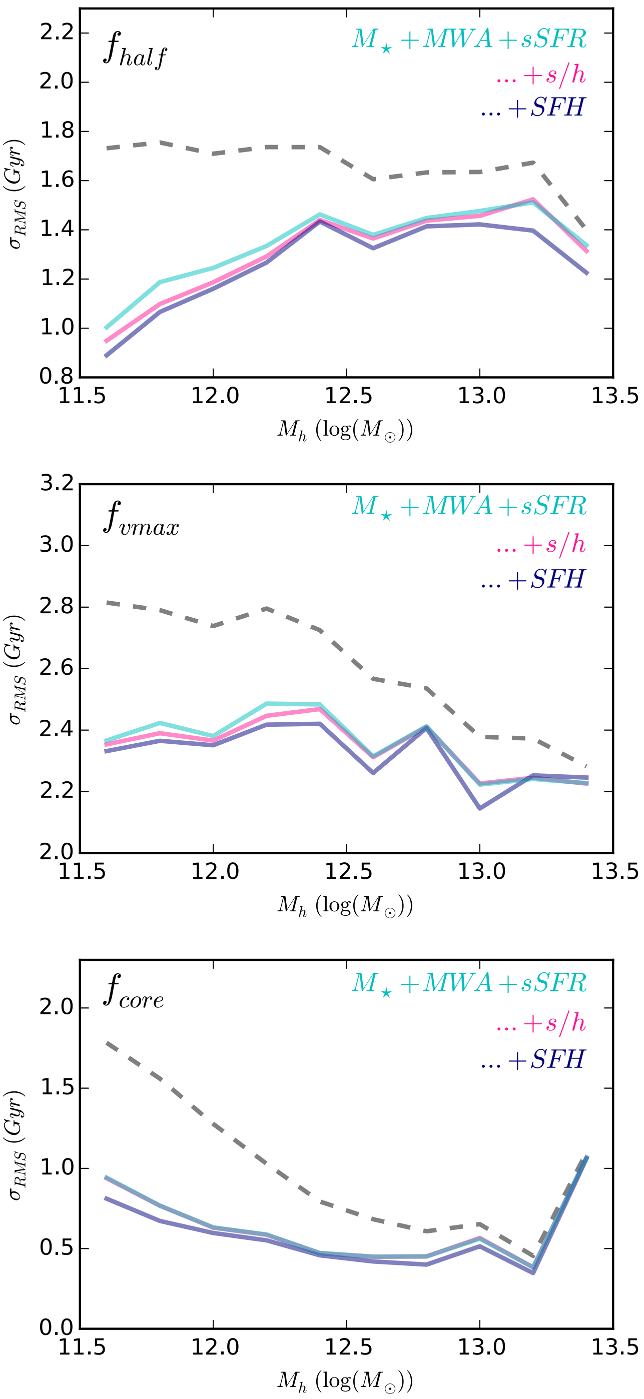

Figure 4.5: Shows the RMS error as a function of halo mass when trying to compute halo formation time from combinations of parameters using random forests. Results using stellar mass, mass-weighted age and specific star formation rate is shown in cyan, all of these parameters plus stellar-halo mass ratio (magenta), and all these four parameters plus the 13 bin star formation history (dark blue). The prediction from halo mass alone is shown as the grey dashed line. This is again shown for three different definitions of formation time: : f_{half} (top), f_{vmax} (middle), and f_{core} (bottom).

4.4.4 With observational errors

In sections 4.4.2 and 4.4.3 we showed that we can improve the prediction of halo formation time over using individual parameters by applying machine learning to many parameters. As this data is created from mock catalogues, these parameters have zero measurement error. In the real world, however, these parameters will be computed from spectra or photometry (with some signal-noise ratio), so the recovered values of parameters will have associated errors. These will depend on the signal-to-noise and wavelength range of observations, and also the choice of fitting method.

To compute mock errors for our data, we make use of the data described in Tojeiro et al. (2017), which contains the true and recovered parameters (stellar mass, SFR, MWA and SFH) obtained by running the VESPA code (Tojeiro et al., 2007) on synthetic spectra of 999 mock galaxies (taken from the same catalogue of LGalaxies used in this study). These spectra are designed to match the signal-to-noise and wavelength range of the GAMA survey (Driver et al., 2011). We obtain the error in recovery of each of these parameters (e.g. for stellar mass: $\Delta M_\star = M_{\star,recovered} - M_{\star,true}$) for all mock galaxies, and then compute the covariance between errors of all parameters, including all bins of SFH. This shows us how correlated the errors in recovery are between all parameters, since if one parameter is overestimated, another parameter may be more likely underestimated or overestimated also. The computed covariance is used to draw random errors from a multivariate normal distribution. These errors are applied to our mock data described in section 4.3 to simulate adding GAMA-like observational errors on our recovered parameters.

We also produce errors for our halo masses. To do this, we measure the scatter in halo mass at fixed stellar mass from data in Tojeiro et al. (2017) to be roughly 0.5. From Tinker et al. (2017), the intrinsic scatter in this relationship is 0.18, so to add observational scatter to this relationship, we draw errors in halo mass from a normal distribution of size $\sigma_{log(M_h)} = \sqrt{0.5^2 - 0.18^2} = 0.47$ and allocate these to all halo masses in the mock catalogue.

We now compute halo formation times using the same method as in section 4.4.3 but for data with mock observational errors added. We do this using both single parameters, and by running random forests on multiple parameters such that we can compare with the previous results. Note that again we run several random forests with different numbers of trees and minimum leaf sizes, finding that using a larger minimum leaf size of ∼200 performs best, compared with ∼20 in the noiseless case. We also find this time that results are worsened for very small numbers of trees ($\lesssim 10$), so again show results for 200 trees. We also investigate augmenting the training data by copying the sample multiple times, adding different random noise to galaxy observables in copy (such that there is more data to train on), but find no significant difference in results. We show the resulting predictions for single parameters in figure 4.6, and random forests in figure 4.7.

Looking at figure 4.6, it is clear that when observational errors are added, it becomes much more difficult to predict the halo formation times f_{half} and f_{vmax} using individual parameters. This is particularly true at low masses, where we originally saw that most parameters gave us more information over just using the halo mass. The added noise on these parameters appears to make this much more difficult. Looking at f_{core}, this is generally much worse predicted, however even with this noise, predictions using individual parameters still appear to do much better than considering halo mass alone, particularly when using stellar mass. This is likely because stellar mass is highly correlated with f_{core} as shown in figure 4.3, and also because the error in stellar mass computed by VESPA is relatively small when compared with other parameters.

The improvement when using many parameters to infer formation time, shown in figure 4.7, is worsened with this added noise. This is likely because parameters are now significantly less correlated with formation time due to the added observational errors. This is further demonstrated by the fact that the random forests prefer a larger minimum leaf size, indicating that extracting more general trends between parameters performs better (i.e. smaller leaf sizes are over-fitting the training sample). Machine learning still offers better predictions than using individual parameters to predict f_{half} and f_{vmax}, particularly at lower masses, however this is not as significant as in the noiseless case. Predictions of f_{core}, however, are drastically improved when using multiple parameters. For example, at the lowest masses, σ_{RMS} goes from ~ 1.8 when using just stellar mass, to ~ 1.2 when using data from all parameters.

Machine learning appears to present improvement over using single parameters for GAMA-like data, however, although f_{core} is very well predicted, the improvement appears to be small for f_{half} and f_{vmax}. Surveys with higher quality spectra or more accurately recovered parameters would therefore benefit more from using machine learning techniques. Improvement also appears to be more significant at lower halo masses, where observed parameters are more cor-

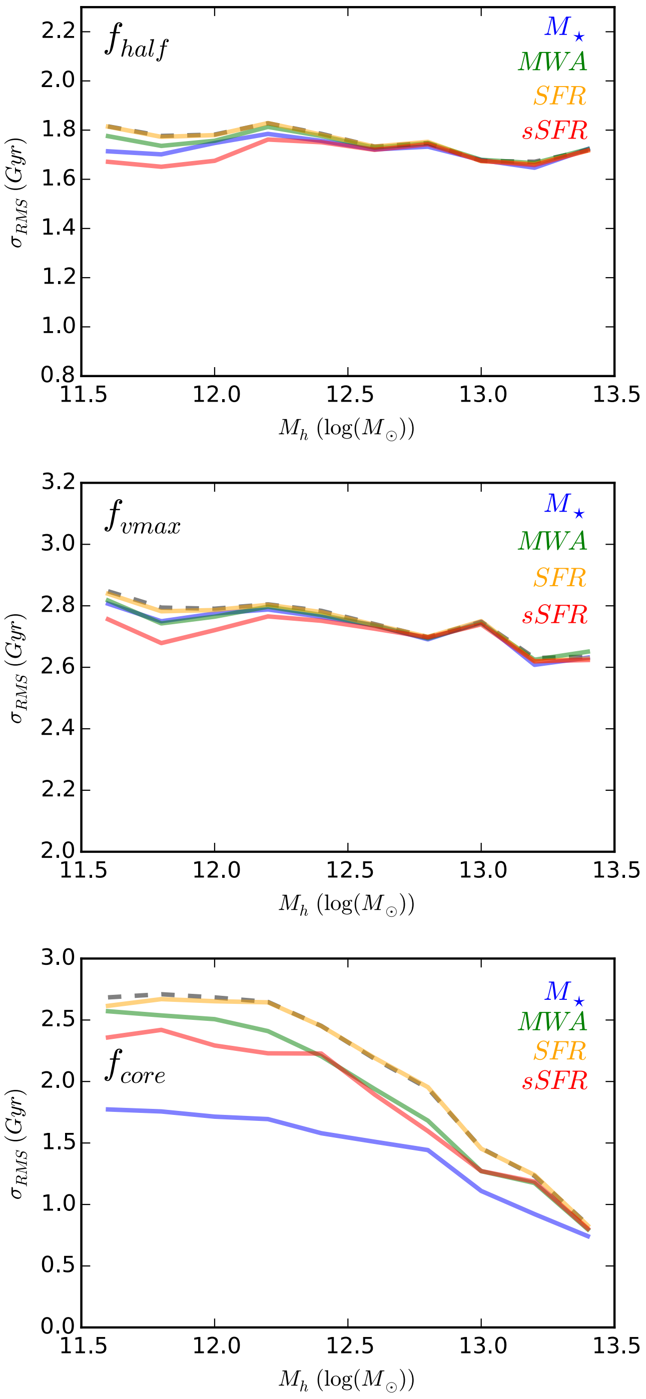

Figure 4.6: Same as figure 4.4, but with errors added to all observable parameters (including halo mass). Size of errors are computed by running VESPA on LGalaxies spectra and measuring offsets from true values. Halo mass errors are computed from the scatter in the measured stellar-halo mass ratio (after removing intrinsic scatter).

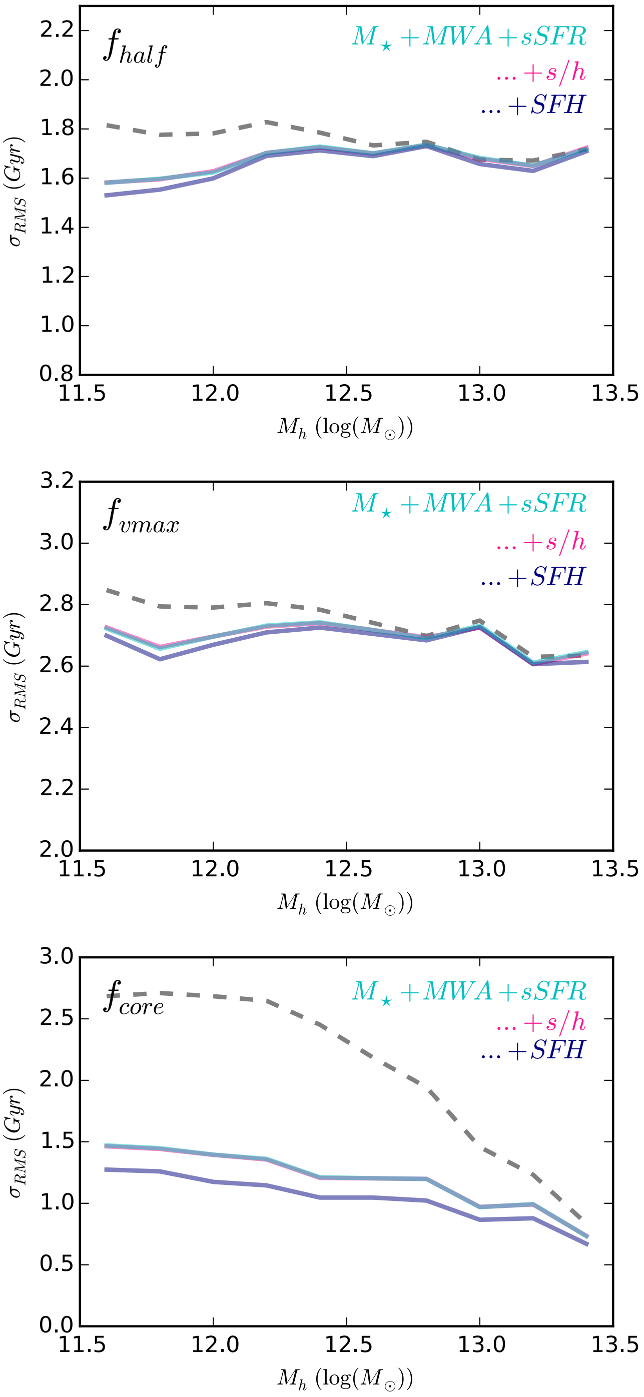

Figure 4.7: Same as figure 4.5, but with errors added to all observable parameters as in figure 4.6.

related with halo mass, so surveys targeting lower halo mass galaxies (i.e. fainter magnitude limits) may also benefit from this.

4.5 Application to GAMA data

Since we have demonstrated that we can infer halo formation times more accurately using multiple galaxy observables at once, we now apply this method to real data. We do this with the view of investigating how halo formation time changes with environment for halos of fixed mass. To do this we use a sample of central galaxies from the GAMA survey described in section 4.3.2, leaving us with VESPA estimates of $M_\star$, MWA, SFR, sSFR, and SFH for 11,700 galaxies.

As in section 4.4.4, we define a sample of simulated galaxies, and apply VESPA-like errors and offsets to recovered parameters, as well as halo mass errors. We train random forests as in section 4.4.4 for different combinations of parameters, however, now train on the full simulated galaxy sample (rather than just the training sample). After training the forests, we feed in the GAMA data to compute formation times for all galaxies in the sample using three different combinations of parameters:

- $M_\star$ and M_{halo}
- $M_\star$, M_{halo} and sSFR
- $M_\star$, M_{halo} and MWA

We do not show results training on the full SFH since random forests are sensitive to differences between data sample and training sample, and it is unclear how comparable SFHs are between our simulations and the real universe (since we must interpolate VESPA SFHs to match the simulation SFH bins, and also, when computing mock-errors, measure the covariance between 16 bins using only 999 galaxies). We note, however, that trends do not change significantly when including SFH. Furthermore, we only compute formation times for f_{half}, since this both captures the hierarchical nature of halo formation (unlike f_{core}), and as shown in section 4.4.4, it is more easily inferred than f_{vmax}.

Since we are interested in testing if at fixed mass, formation time varies with environment, we make use of the two measures of environment described in section 4.3.2. Firstly, we consider the geometric environment definitions of Eardley et al. (2015), which measure the dimensionalily of collapse (i.e.: 0D: voids, 1D: sheets, 2D: filaments, 3D: knots). We split the GAMA sample into 5 equal size bins of halo mass between $11.5 < log(M_{halo}) < 14$, and com-

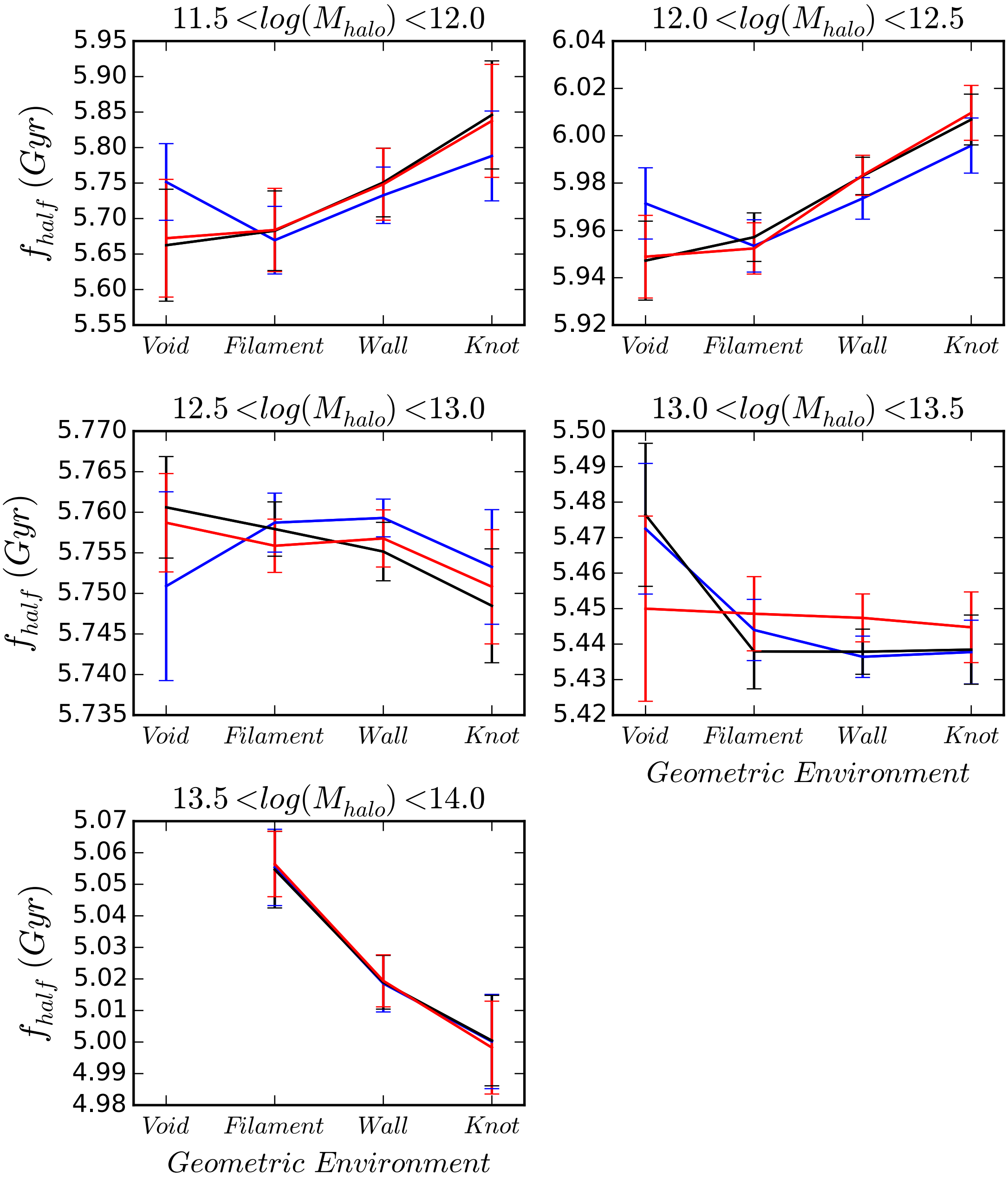

Figure 4.8: The computed median formation time for GAMA data as a function of geometric environment, in a number of halo mass bins using random forests. Formation times are computed using three different sets of parameters: $M_\star$ and M_{halo} (black), $M_\star$, M_{halo} and sSFR (blue), and $M_\star$, M_{halo} and MWA (red). The errorbars represent the standard error.

pute the median formation time in each environment. Following Tojeiro et al. (2017), since formation time is sensitive to differences halo mass (also shown in our figure 4.1), and environment is correlated with halo mass, in each halo mass bin, we weight our galaxies. Weights are chosen such that the weighted halo mass distribution for each environment is the same as that of the full sample in that halo mass bin. This allows us to ensure any trends of formation time with environment are not due to differences in halo mass.

We show the resulting formation times as a function of environment in figure 4.8 training on three different sets of parameters. It is firstly visible that, as seen in simulations in figure 4.1, halos in higher mass bins are generally younger than those at lower masses. This trend is true in figure 4.8 for all bins except the $11.5 < log(M_{halo}) < 12$ bin, which is younger than the next most massive bin. This could be due to the GAMA magnitude limit affecting lower halo masses more, meaning only halos hosting brighter galaxies are selected in this bin. To explain this trend, this would then require brighter galaxies to be preferentially present in younger halos, which, if halo mergers are correlated with the quenching of star formation, would make sense, since galaxies which quench later should have younger (hence brighter) stellar populations.

Now considering environment, in the two highest halo mass bins, halos in dense environments (e.g. knots) appear to form later than in voids. In the intermediate mass bin ($12.5 < log(M_{halo}) < 13$), there appears to be little difference in formation time with environment, however in the lowest two mass bins, the trend appears to reverse, and low mass halos in dense environments are older. Some small differences are seen depending on which set of parameters are chosen to train the model, however the global trends are in good agreement. These results are in general agreement with simulations (Gao et al., 2005; Wechsler et al., 2006; Hahn et al., 2009), which find that at high masses, older halos are less biased, and lower masses older haloes are more biased.

We now consider our second measure of environment, computed using DisPerSE (see section 4.3.2). All halos are allocated a value of d_{node}, describing the distance to nearest node. We define bins 5 bins in d_{node} for the GAMA sample such that each bin contains an equal number of galaxies when considering the full sample. We again split the GAMA sample in to 5 bins of halo mass between $11.5 < log(M_{halo}) < 14$, and for each bin, compute the median formation time for each of our five bins of environment. The median formation times, again

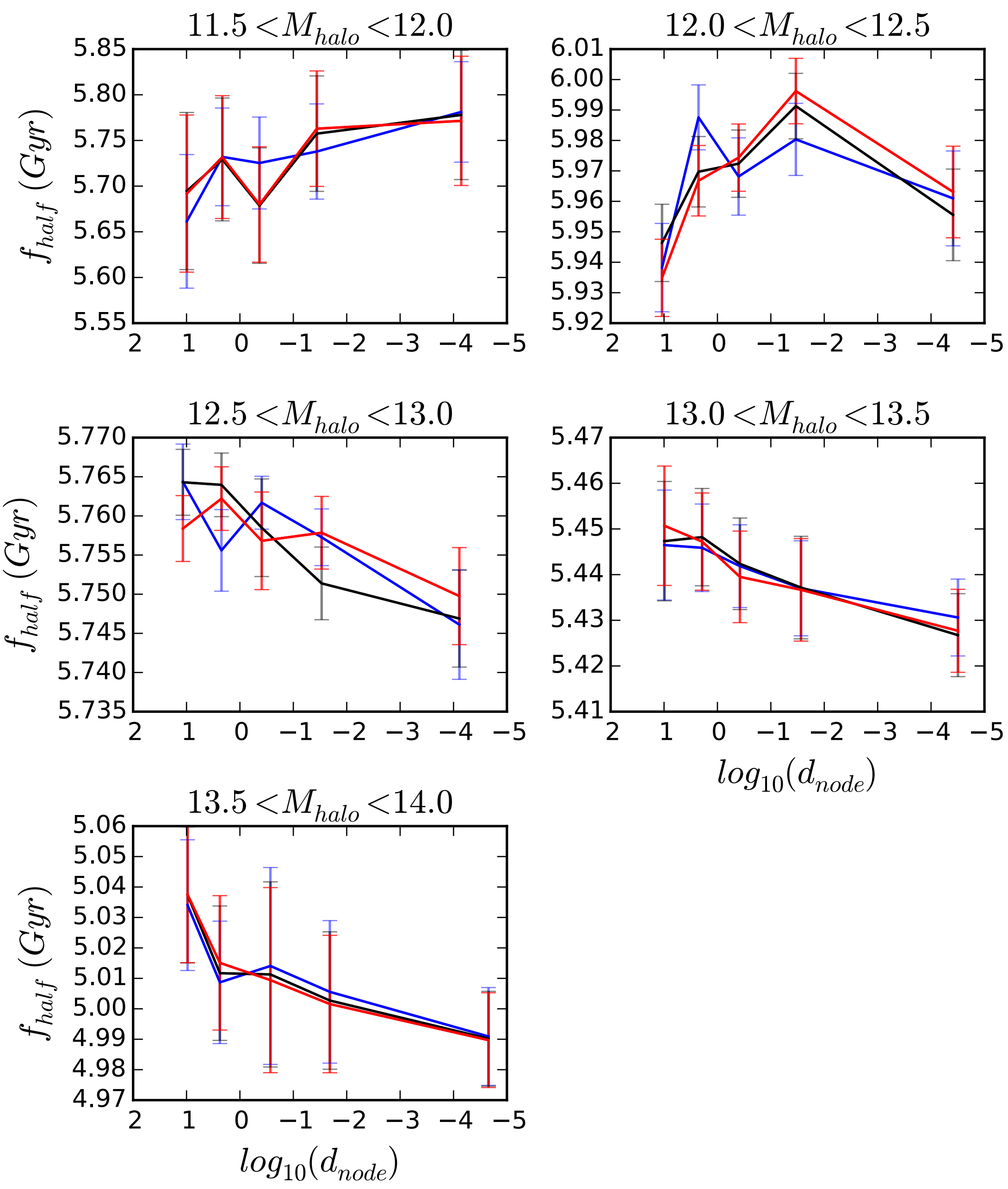

Figure 4.9: The computed median formation times of GAMA data in 5 bins of $log(d_{node})$, shown for 5 bins of halo mass. Formation times are computed using three different sets of parameters: just $M_\star$ and M_{halo} (black), $M_\star$, M_{halo} and sSFR (blue), and $M_\star$, M_{halo} and MWA (red). Note that the x-axis is flipped such that dense environments (i.e. knots) are on the right as in figure 4.8. The errorbars represent the standard error.

computed using three sets of parameters, are presented in figure 4.9.

The results show much of the same trends as with geometric environment, with the highest three mass bins showing that halos in dense environments (e.g. knots) form later. The lowest mass bin shows the opposite trend with halos in dense environments forming earlier, again agreeing broadly with figure 4.8. Interestingly, the trend appears to be in the process of flipping in the $12 < log(M_{halo}) < 12.5$ bin, showing that halos in both knots and voids are younger, with intermediate environments being older, although errorbars are fairly large. Considering also the trends in geometric environment, this implies a turning point at around $M_{halo} \sim 10^{12.5} M_{\odot}$.

These results agree well with Tojeiro et al. (2017), who also find that, using GAMA data and splitting galaxies by geometric environment, low-mass haloes residing in knots are older than those residing in voids, and conversely, that at high masses, haloes in knots are younger than those in voids.

4.6 Conclusions

We have investigated three different definitions of halo formation time, f_{half}, f_{vmax}, and f_{core}, in the LGalaxies semi-analytic model, showing that each of these definitions describe different aspects of the halo formation process. We investigate how these three formation times are correlated with a number of observable galaxy properties: stellar mass, MWA, SFR, sSFR, stellar-halo mass ratio, and SFH. All observables show some correlations with formation time, although f_{vmax} appears slightly less correlated.

We split the data into training and test samples to investigate how well formation times can be predicted from galaxy observables. When fitting polynomials to these relations and applying to a seperate sample, we find that all parameters offer improvement in formation time prediction over simply considering halo mass alone, except SFR, which performs particularly badly. Furthermore, f_{vmax} appears very difficult to estimate with typical RMS errors of ~ 2.5 Gyr, compared with ~ 1.6 Gyr and ~ 1 Gyr for f_{half} and f_{core} respectively.

We go on to investigate if we can improve on these predictions by applying random forests to multiple parameters at once, showing that considering $M_{\star}$, MWA and sSFR together roughly halves the RMS error in f_{half} and f_{core} at low halo masses, offering only small improvements at higher masses, which seems to suggest that formation time has a smaller effect on galaxy properties at higher masses. Adding stellar-halo mass ratio and SFH only slightly improves the prediction, suggesting that most of this information is already contained within the first three parameters. f_{vmax} remains very difficult to estimate, however random forests still perform better than considering parameters individually.

We then add GAMA-like errors to galaxy observables and halo mass, in order to see if this changes the accuracy of prediction, showing that individual parameters are then inferred much less accurately. This is particularly true for both f_{half} and f_{vmax}, however f_{core} is still fairly well predicted, and best predicted by stellar mass. When running random forests on these parameters, a lot of the improvement seen before is lost due to the errors in recovered GAMA observables. f_{half} is still better predicted with multiple parameters, however, and f_{core} is still very well predicted, with an RMS error of ~ 1 Gyr at all halo masses.

We go on to compute formation times for the real GAMA survey, training on LGalaxies data, and investigate how formation time changes with two different definitions of environment at fixed halo mass. We show that, firstly, as seen in simulations, halos in higher mass bins are generally younger than those at lower masses. Considering geometric environment, we show that high mass halos in dense environments are younger than in voids, and that this trend reverses at lower halo masses with halos in knots being older. Looking instead at environments computed using DisPerSE, we find the same general trend, although in the $12 < log(M_{halo}) < 12.5$ bin, halos in both knots and voids are younger, with intermediate environments being older. This, along with our results from geometric environment suggest a flipping point in the trend at around $M_{halo} \sim 10^{12.5} M_{\odot}$

These results are in general agreement with simulations (Gao et al., 2005; Wechsler et al., 2006; Hahn et al., 2009; Borzyszkowski et al., 2017), which find that at high masses, older halos are less biased, and lower masses older haloes are more biased. Hahn et al. (2009) and Borzyszkowski et al. (2017) show that at low masses, the growth of these haloes in the vicinity of high-mass haloes can be suppressed, corresponding to an earlier formation time. At high masses a dependence of bias with formation time is expected from considering Gaussian statistics of rare high-density peaks (Zentner, 2007; Dalal et al., 2008), and although we consider environment rather than clustering, bias is expected to increase monotonically with geometric environment (Fisher & Faltenbacher, 2016). These results furthermore agree with Tojeiro et al. (2017), who find the same trends at both high and low masses in GAMA data, splitting galaxies by geometric environment.

These results appear to show signs of halo assembly bias in GAMA data based on geometric environment. Newer surveys that have larger samples of galaxies, or better estimates of galaxy parameters (e.g. DESI (DESI Collaboration et al., 2016), EUCLID (Laureijs et al., 2011)) should be able to be perform similar studies with greater accurately. This will allow us to test assembly bias in the real universe in much more detail, and will allow us test if the magnitude of assembly bias seen in our current cosmological models is correct.

Appendices

4.A Formation time without halo mass

For some datasets, estimates of halo mass may be inaccurate or missing, or one might be interested in the formation history of the halo from the point of view of galaxy evolution. In this case, one might wish to compute formation time as a function of, for example, stellar mass instead. Here, we investigate how well formation time can be predicted using only parameters obtainable from spectra (stellar mass, MWA, SFR, and SFH), comparing this in different bins of stellar mass. We follow the same method as section 4.4.2, but instead split data in to bins of stellar mass of size $\Delta log(M_\star) = 0.2$ between $9.5 < log(M_\star) < 11.5$. We again split this data in to a "training" and "test" sample, of size 80% and 20% respectively and fit 2nd order polynomials to the training data, using the fit to predict formation times for the test data.

We present the resulting Root Mean Squared (RMS) error for each individual parameter in figure 4.A.1. It is clear that no galaxy observable gives significant improvement in formation time estimation over using stellar mass alone, however, small improvements are again seen when using SFR, MWA and sSFR at lower stellar masses. It is secondly visible that formation time estimates are not significantly different than estimates without halo mass 4.4, indicating that most of the halo mass information is contained within stellar mass and these other parameters.

We now compute formation times using the same method as in section 4.4.3, computing formation times using several parameters, by firstly training a random forest with these parameters. This time, however, we bin by stellar mass, and also do not input any parameters relating to halo mass in to our random forests. We show the resulting RMS error in prediction for two combinations of parameters in figure 4.A.2.

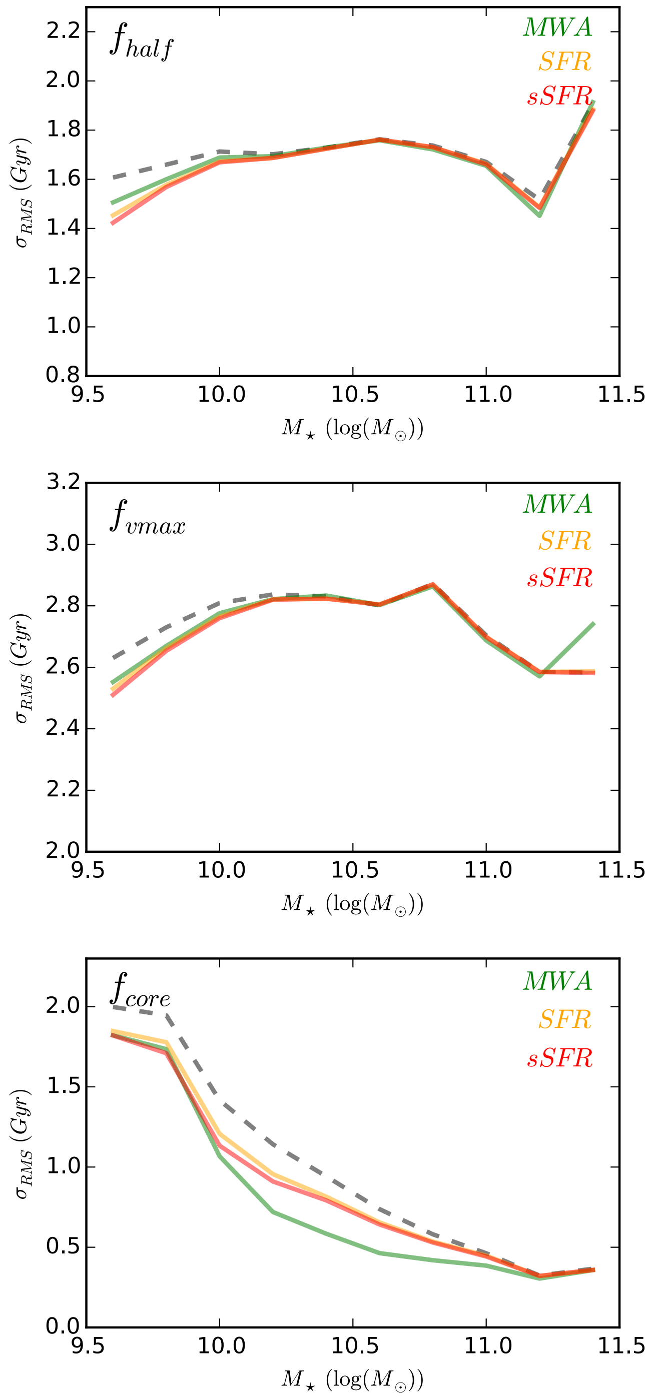

Figure 4.A.1: Same as figure 4.4, but binning by stellar mass instead. We show the RMS error as a function of stellar mass when trying to compute halo formation time from single parameters: mass-weighted age (green), star formation rate (yellow), and specific star formation rate (red). The prediction from just stellar mass is shown as the grey dashed line.

Trends are again comparable to the case where halo mass is included (figure 4.5), with RMS errors being significantly smaller than the single parameter case, and a greater improvement seen at lower stellar masses. This demonstrates that in the case where we have perfect measurements of stellar mass, MWA, sSFR and SFH, halo formation time can be accurately inferred, even without halo masses.

For consistency with section 4.4.4, we test how well formation times can be computed without halo masses for GAMA-like data. We add noise to all parameters using the same method as section 4.4.4. The single parameter predictions are shown in figure 4.A.3, and random forest predictions in 4.A.4. The individual parameters show virtually no improvement over using stellar mass alone, likely due to the large errors in MWA, sSFR and SFH. Previously, we halo mass and stellar mass, which both have small errors, relative to other parameters. Using random forests again shows some improvement, particularly at low stellar masses, and for f_{core}, although the improvement is not as significant as the case with halo mass. We note that for datasets with higher quality spectra, but no halo mass estimates, formation time estimates should be significantly better.

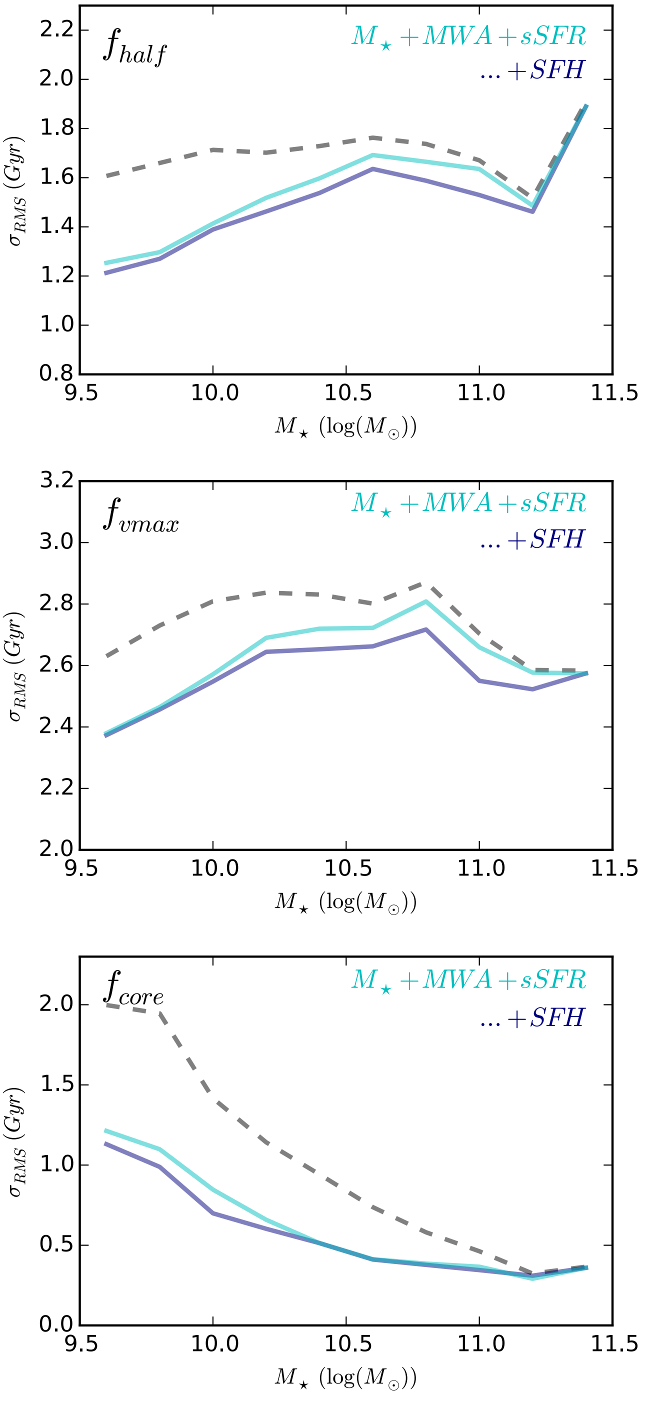

Figure 4.A.2: Same as figure 4.5, but binning by stellar mass instead. RMS errors for a random forest using stellar mass, mass-weighted age and specific star formation rate is shown in cyan, then these three parameters plus the full 13 bin star formation history (dark blue)

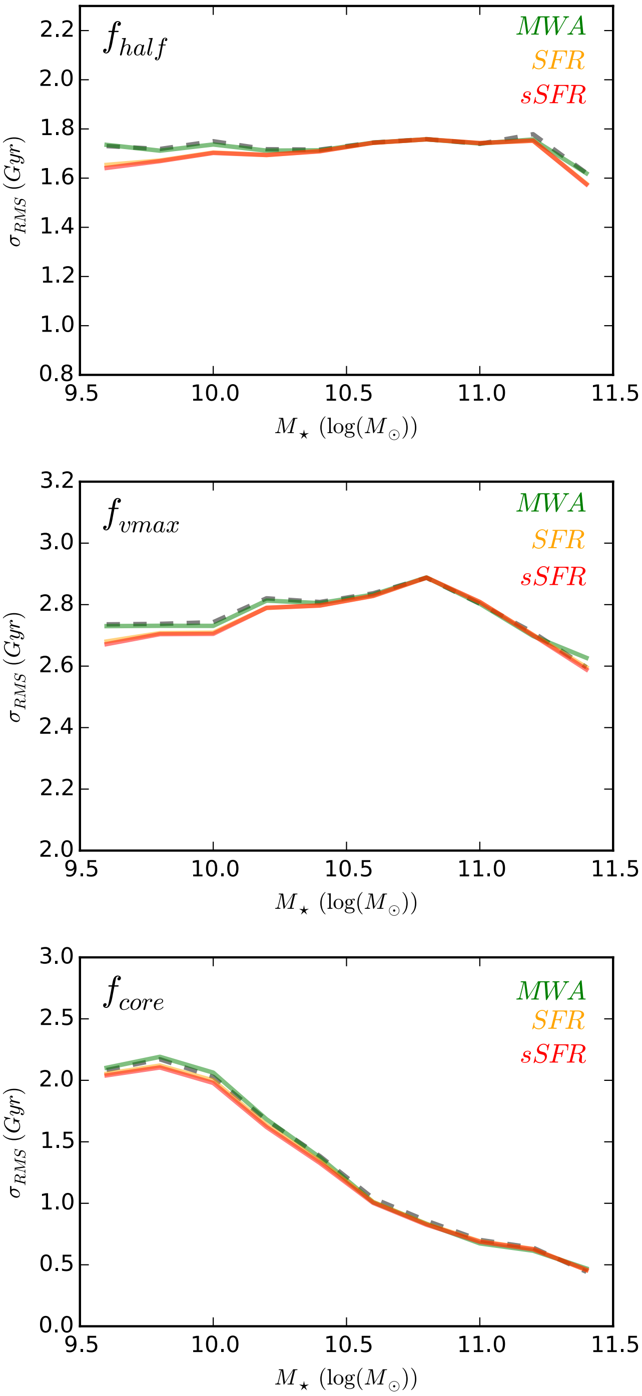

Figure 4.A.3: Same as figure 4.A.1, but with errors added to all observable parameters as in figure 4.6.

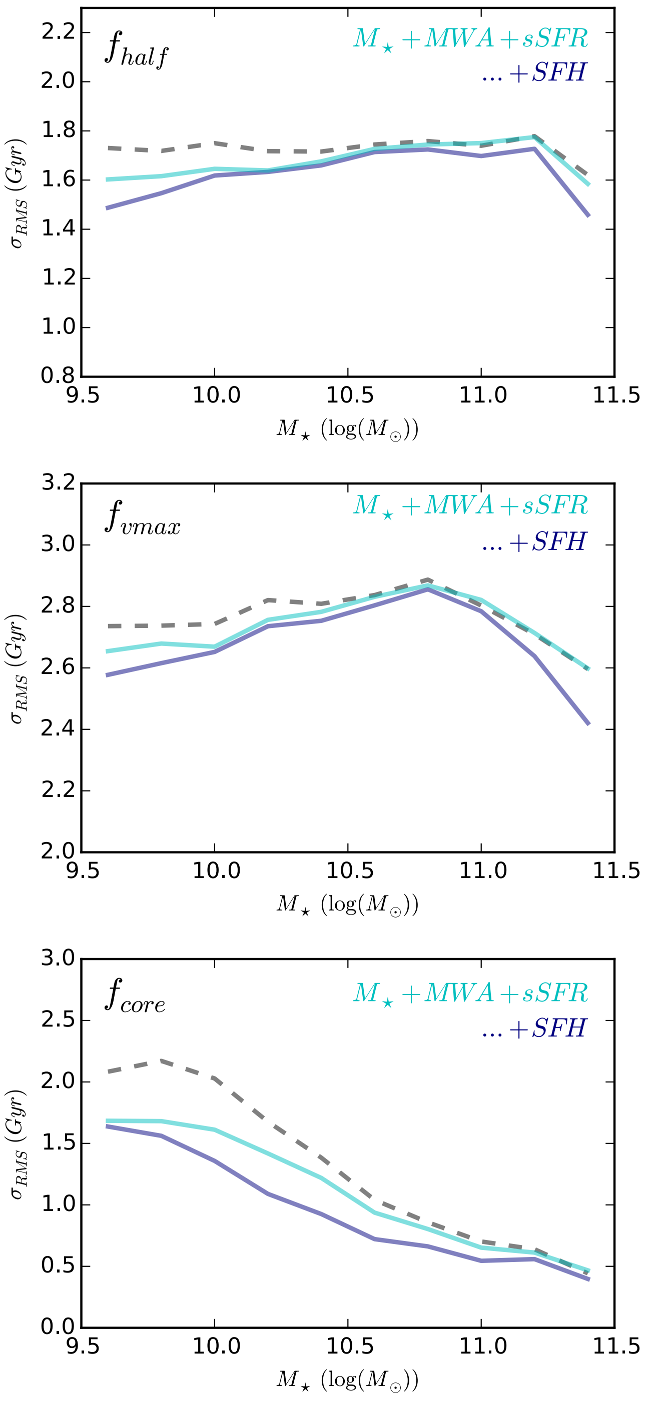

Figure 4.A.4: Same as figure 4.A.2, but with errors added to all observable parameters as in figure 4.6.

5

Conclusions

In this thesis, we have performed of galaxy evolution using a number of different observational probes. In chapter 2, we implement a method of clustering redshifts based on the method of Ménard et al. (2013), testing this on the BOSS survey, splitting galaxies by colour in to a blue and red sample. Recovering the redshift distribution of the blue sample by crosscorrelation with the red sample is very successful, even when assuming no evolution in bias for the unknown sample. We test the recovery of a sub-sample of the unknown sample, showing that noise in the recovered $\phi(z)$ increases significantly as the number of galaxies in the unknown sample decreases.

We have also investigated how the method performs using mock galaxy catalogues defined from semi-analytic models. We define a mock SDSS survey and mock BOSS survey with comparable magnitudes and photometric errors to the real surveys. We firstly investigate how the bias evolves in the mock SDSS, finding that clustering amplitude for magnitude limited samples increases at high redshifts, likely due to the fact that these galaxies are intrinsically luminous and hence strongly biased. Clustering amplitude also increases at low redshifts in these samples, implying that, although an increasing satellite fraction may have some effect on this, amplitude in the model is too strong here.

We apply our implementation of clustering redshifts to recover redshift distributions of mock SDSS galaxies in small bins of i, $g-r$, and $r-i$, using a bias correction defined from how clustering amplitude evolves in the model. We find that at low magnitudes, redshift distributions are recovered well, and the choice of bias correction has little effect. At fainter magnitudes, distributions are recovered well with a bias correction, however, with no correction applied, redshift distributions are biased towards low redshifts. The significance of the bias correction here is likely because redshift distributions are much wider at faint magnitudes and hence clustering is more affected by the bias evolution. This increased width is likely due to the fact that photometric error is higher at faint magnitudes, hence redshift is less well constrained in each bin. We note that this may not be as significant a problem for future photometric surveys, for example DES or DECaLS, with better photometry and smaller photometric error.

We finally investigate how the choice of clustering scale affects the recovered distributions, finding that crosscorrelations over larger scales are noisier, leading to more error in recovered $\phi(z)s$. We therefore suggest using the smallest scale possible when recovering redshift distributions of real data.

In chapter 3, we extend this work, showing that the clustering redshifts method can be applied successfully to real photometric data from the SDSS, recovering redshift distributions out to $z=0.8$, as a function of magnitude and colour. We show that these redshift distributions are broadly consistent with distributions from the GAMA survey over the same bins.

We show that stellar masses and luminosities can be computed by binning semi-analytic models in colour and redshift space, and applying these masses and luminosities to galaxies over the same bins of magnitude/colour and redshift. We test how well mass functions are recovered using this method, considering two different semi-analytic models (LGalaxies and SAGE). We show that both models produce similar mass estimates for low to intermediate mass galaxies ($M_\star < 10^{11.25} M_\odot$), however estimates differ at higher masses. After testing how well both models re-produce galaxies in the BOSS survey, we conclude that high mass galaxies in LGalaxies best represent the real universe, so opt to compute masses using this model. We also test how well luminosity functions can be recovered using the same method, finding that both models well recover the luminosity function of the mock data.

We apply this method to real data, recovering mass function and luminosity functions for a large ($\sim 7000\ deg^2$) sample of galaxies from the SDSS ($i < 21$), allowing us to understand their evolution with little sample variance. We find little evolution at high masses between $0.2 < z < 0.8$, suggesting that the most massive galaxies form most of their mass before this time, and do not evolve significantly in mass afterwards. The lack of evolution over these redshifts agrees well with other studies, for example, Pérez-González et al. (2008); Moustakas et al. (2013); Leauthaud et al. (2016); Guo et al. (2018). In our study, the effect of a bias correction on the recovered mass functions is generally comparable to, or smaller than, the error, however this may not be the case for future large-volume surveys. Our luminosity functions show some evolution with redshift, possibly due to passive evolution.

We also produce targeting completeness measurements for BOSS using these mass functions, suggesting that over the redshift range $0.2 < z < 0.7$, BOSS is around 80% complete at high masses ($M_\star > 10^{11.4} M_\odot$), and falling to almost zero below $M_\star < 10^{11} M_\odot$. In our highest redshift bin ($0.7 < z < 0.8$) BOSS is strongly affected by incompleteness, and is only about 30% complete at the highest masses $M_\star \gtrsim 10^{11.6} M_\odot$. We also demonstrate that when comparing mass functions or completeness estimates between methods, significant offsets can be present, which require correction.

Our completeness estimates are in good agreement with Guo et al. (2018), finding that BOSS is around 80% complete above $M_\star \gtrsim 10^{11.3} M_\odot$ between $0.2 < z < 0.6$, with completeness falling off significantly at higher redshifts. Similar completeness estimates are also found in Leauthaud et al. (2016), however at higher completeness is found at the highest masses when compared with our estimates or Guo et al. (2018).

Ongoing and future large-volume spectroscopic surveys, for example eBOSS, DESI and EUCLID (Laureijs et al., 2011), will produce large number of spectra out to higher redshifts. This will firstly allow for better clustering redshifts estimates due to having a larger, more densely populated reference sample, but also produce large spectroscopic galaxy samples, for which incompleteness must be understood. Combining these data with ongoing and future photometric surveys, for example, The Dark Energy Camera Legacy Survey (DECaLS) (Dey et al., 2018), and The Dark Energy Survey (DES) (DES Collaboration et al., 2017), will allow for redshift distributions to be computed out to higher redshifts, and in much smaller bins of colour, due to these new surveys reaching much deeper and having much smaller photometric error. This will allow us to not only to understand the completeness of these spectroscopic samples, but also compute stellar mass and luminosity functions over the largest volumes possible.

The methods used in these chapters, will therefore be important tools for the next generation of galaxy surveys in order to fully utilise these large databases, and to understand the galaxy populations present. Furthermore, producing mass functions, luminosity functions, and redshift distributions will provide tighter constrains for the next generation of simulations, allowing us to better test our models of cosmology and galaxy evolution.

In chapter 4, we investigate three different definitions of halo formation time, f_{half}, f_{vmax}, and f_{core}, in the LGalaxies semi-analytic model, showing that each of these definitions describe different aspects of the halo formation process. We investigate how these three formation times are correlated with a number of observable galaxy properties: stellar mass, MWA, SFR, sSFR, stellar-halo mass ratio, and SFH. All obserbles show some correlations with formation time, although f_{vmax} appears slightly less correlated.

We split the data in to training and test samples to investigate how well formation times can be predicted from galaxy observables. When fitting polynomials to these relations and applying to a seperate sample, we find that all parameters offer improvement in formation time prediction over simply considering halo mass alone, except SFR, which performs particularly badly. Furthermore, f_{vmax} appears very difficult to estimate with typical rms errors of ~ 2.5 Gyr, compared with ~ 1.6 Gyr and ~ 1 Gyr for f_{half} and f_{core} respectively.

We go on to investigate if we can improve on these predictions by applying random forests to multiple parameters at once, showing that considering $M_{\star}$, MWA and sSFR together roughly halves the rms error in f_{half} and f_{core} at low halo masses, offering only small improvements at higher masses, which seems to suggest that formation time has a smaller effect on galaxy properties at higher masses. Adding stellar-halo mass ratio and SFH only slightly improves the prediction, suggesting that most of this information is already contained within the first three parameters. f_{vmax} remains very difficult to estimate, however random forests still perform better than considering parameters individually.

We then add GAMA-like errors to galaxy observables and halo mass, in order to see if this changes the accuracy of prediction, showing that individual parameters are then inferred much less accurately. This is particularly true for both f_{half} and f_{vmax}, however f_{core} is still fairly well predicted, and best predicted by stellar mass. When running random forests on these parameters, a lot of the improvement seen before is lost due to the errors in recovered GAMA observables. f_{half} is still better predicted with multiple parameters, however, and f_{core} is still very well predicted, with an RMS error of ~ 1 Gyr at all halo masses.

We go on to compute formation times for the real GAMA survey, training on LGalaxies data, and investigate how formation time changes with two different definitions of environment at fixed halo mass. We show that, firstly, as seen in simulations, halos in higher mass bins are generally younger than those at lower masses. Considering geometric environment, we show that high mass halos in dense environments are younger than in voids, and that this trend reverses at lower halo masses with halos in knots being older. Looking instead at environments computed using DisPerSE, we find the same general trend, although in the $12 < log(M_{halo}) < 12.5$ bin, halos in both knots and voids are younger, with intermediate environments being older. This, along with our results from geometric environment suggest a flipping point in the trend at around $M_{halo} \sim 10^{12.5} M_{\star}$

These results appear to show signs of halo assembly bias in GAMA data based on geometric environment. This is in general agreement with simulations (Gao et al., 2005; Wechsler et al., 2006; Hahn et al., 2009; Borzyszkowski et al., 2017), which find that at high masses, older halos are less biased, and lower masses older haloes are more biased. Newer surveys (e.g. DESI (DESI Collaboration et al., 2016), EUCLID (Laureijs et al., 2011)) will have larger samples of galaxies, or better estimates of galaxy parameters, so should be able to be perform similar studies with even greater accurately. This will allow us to test assembly bias in much more detail, and make detailed comparisons between observations and simulations to help us better constrain cosmology and galaxy evolution. Furthermore, machine learning techniques will become increasingly useful, as more astronomical data is collected. Applying these techniques to more accurately recovered parameters, or even raw galaxy spectra, might allow us to constrain formation time with greater accuracy, or even measure full halo assembly history.

Bibliography

Abbas, U., & Sheth, R. K. 2006, MNRAS, 372, 1749A

Abdalla, F. B., et al. 2011, MNRAS, 417, 1891A

Abolfathi, B., et al. 2018, ApJS, 235, 42A

Agertz, O., et al. 2013, ApJ, 770, 25A

Aihara, H., et al. 2011, ApJS, 193, 29A

——. 2018, PASJ, 70S, 8A

Alam, S., et al. 2015, ApJS, 219, 12A

——. 2017, MNRAS, 470, 2617A

Anderson, L., et al. 2012, MNRAS, 427, 3435A

Armitage, T. J., et al. 2019, MNRAS, 484, 1526A

Artale, M. C., et al. 2018, MNRAS, 480, 3978A

Asquith, R., et al. 2018, MNRAS, 480, 1197A

Ata, M., et al. 2018, MNRAS, 473, 4773A

Aumer, M., et al. 2013, MNRAS, 434, 3142A

——. 2014, MNRAS, 441, 3679A

Baldry, I. K., et al. 2008, MNRAS, 388, 945B

——. 2012, MNRAS, 421, 621B

——. 2018, MNRAS, 474, 3875B

Ball, N. M., et al. 2008, ApJ, 683, 12B

Banerji, M., et al. 2010, MNRAS, 406, 342B

Bardeen, J. M., et al. 1986, ApJ, 304, 15B

Barnes, J., & Hut, P. 1986, Natur, 324, 446B

Bates, D. J., et al. 2019, MNRAS, 486, 3059B

Bautista, J. E., et al. 2017, A&A, 603A, 12B

Behroozi, P. S., et al. 2013, ApJ, 762, 109B

Bell, E. F., et al. 2003, ApJS, 149, 289B

——. 2004, ApJ, 608, 752B

Benítez, N. 2000, ApJ, 536, 571B

Bernyk, M., et al. 2016, ApJS, 223, 9B

Beutler, F., et al. 2014, MNRAS, 443, 1065B

Blanton, M. R., et al. 2017, AJ, 154, 28B

Bolzonella, M., et al. 2000, AA, 363, 476B

Borzyszkowski, M., et al. 2017, MNRAS, 469, 594B

Boylan-Kolchin, M., et al. 2009, MNRAS, 398, 1150B

Brammer, G. B., et al. 2011, ApJ, 739, 24B

Bruzual, G., & Charlot, S. 2003, MNRAS, 344, 1000B

Bundy, K., et al. 2015, ApJ, 798, 7B

Calderon, V. F., & Berlind, A. A. 2019, arXiv:1902.02680

Calzetti, D., et al. 2000, ApJ, 533, 682C

Capozzi, D., et al. 2017, arXiv:1303.4409

Carliles, S., et al. 2010, ApJ, 712, 511C

Chaves-Montero, J., et al. 2016, MNRAS, 460, 3100C

Chen, et al. 2012, MNRAS, 421, 314C

Christodoulou, L., et al. 2012, MNRAS, 425, 1527C

Coil, A. L. 2013, pss6, book, 387C

Coil, A. L., et al. 2011, ApJ, 741, 8

Cole, S., & Kaiser, N. 1989, MNRAS, 237, 1127C

Cole, S., et al. 2000, MNRAS, 319, 168C

——. 2001, MNRAS, 326, 255C

Colless, M., et al. 2003, arXiv:astro-ph/0306581

Collister, A. A., & Lahav, O. 2004, PASP, 116, 345C

Comparat, J., et al. 2017, preprint (arXiv:1711.06575)

Cora, S. A., et al. 2018, MNRAS, 479, 2C

Crain, R. A., et al. 2015, MNRAS, 450, 1937C

Cresswell, J. G., & Percival, W. J. 2009, MNRAS, 392, 682C

Croton, D. J., et al. 2006, MNRAS, 365, 11C

——. 2016, ApJS, 222, 22C

Csabai, I., et al. 2003, AJ, 125, 580C

Dalal, N., et al. 2008, PhRvD, 77, l3514D

Davis, M., & Geller, M. J. 1999, ApJ, 208, 13D

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

Davis, M., et al. 1985, ApJ, 292, 371D

——. 1988, ApJ, 333L, 9D

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

——. 2016, AJ, 151, 44D

De Lucia, G., & Blaizot, J. 2007, MNRAS, 375, 2D

De Lucia, G., et al. 2010, MNRAS, 406, 1533D

Del Popolo, A., & Le Delliou, M. 2017, Galax, 5, 17D

DES Collaboration, et al. 2017, preprint (arXiv:1708.01530)

DESI Collaboration, et al. 2016, preprint (arXiv:1611.00036)

Dey, A., et al. 2018, preprint (arXiv:1804.08657)

——. 2019, AJ, 157, 168D

Di Matteo, P., et al. 2007, A&A, 468, 61D

Dodelson, S. 2003, moco, book, D

Driver, S. P., et al. 2009, A&G, 50e, 12D

——. 2011, MNRAS, 413, 971D

Duffy, A. R., et al. 2010, MNRAS, 405, 2161D

Eardley, E., et al. 2015, MNRAS, 448, 3665E

Eisenstein, D. J., et al. 2001, AJ, 122, 2267E

——. 2011, AJ, 142, 72E

Faber, S. M., et al. 2007, ApJ, 665, 265F

Feldman, H. A., et al. 1994, ApJ, 426, 23F

Fisher, J. D., & Faltenbacher, A. 2016, arXiv:1603.06955

Freeman, K. C. 1970, ApJ, 160, 811F

Frenk, C. S., et al. 1983, ApJ, 271, 417F

Fry, J. N. 1996, ApJ, 461L, 65F

Gabor, J. M., et al. 2010, MNRAS, 407, 749G

Gao, L., & M., S. D. 2007, MNRAS, 377L, 5G

Gao, L., et al. 2005, MNRAS, 363L, 66G

Gatti, M., et al. 2018, MNRAS, 477, 1664G

Gerdes, D. W., et al. 2010a, ApJ, 715, 823G

——. 2010b, ApJ, 715, 823G

Gonzalez-Perez, V., et al. 2014, MNRAS, 439, 264G

Governato, F., et al. 2012, MNRAS, 422, 1231G

Grützbauch, R., et al. 2011, MNRAS, 411, 929G

Guedes, J., et al. 2011, ApJ, 742, 76G

Gunn, J. E., et al. 1998, AJ, 116, 3040G

——. 2006, AJ, 131, 2332G

Guo, H., et al. 2018, ApJ, 858, 30G

Guo, Q., et al. 2011, MNRAS, 413, 101G

Guzzo, L., et al. 2014, A&A, 43, 108G

Hahn, O., et al. 2009, MNRAS, 398, 1742H

Han, J., et al. 2015, MNRAS, 446, 1356H

Henriques, B. M. B., et al. 2013, MNRAS, 431, 3373H

——. 2015, MNRAS, 451, 2663H

Heymans, C., et al. 2013, MNRAS, 432, 2433H

Ho, T. K. 1995, in Proceedings of the Third International Conference on Document Analysis and Recognition - Volume 1, 278

Ho, T. K. 2002, Pattern Analysis & Applications, 5, 102

Hocking, A., et al. 2018, MNRAS, 473, 1108H

Hockney, R. W., & Eastwood, J. W. 1988, csup, book, H

Holmberg, E. 1958, MeLuS, 136, 1H

Hopkins, P. F., et al. 2013, MNRAS, 430, 1901H

——. 2014, MNRAS, 445, 581H

Hubble, E. P. 1925, Obs, 48, 139H

——. 1926, ApJ, 64, 321H

——. 1929, PNAS, 15, 168H

——. 1936, The Realm of the Nebulae (Oxford University Press: Oxford), 79

Ilbert, O., et al. 2006, AA, 457, 841I

Jogee, S., et al. 2009, ApJ, 697, 1971J

Kaiser, N. 1984, ApJ, 284L, 9K

Kauffmann, G., et al. 1997, MNRAS, 286, 795K

Kimm, T., et al. 2009, MNRAS, 394, 1131K

Klypin, A., et al. 2016, MNRAS, 457, 4340K

Knebe, A., et al. 2013, MNRAS, 435, 1618K

——. 2015, MNRAS, 451, 4029K

——. 2018, MNRAS, 474, 5206K

Kraljic, K., et al. 2018, MNRAS, 474, 547K

Kroupa, P., et al. 2001, MNRAS, 322, 231K

Lacey, C., & Cole, S. 1994, MNRAS, 271, 676L

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

Laureijs, R., et al. 2011, preprint (arXiv:1110.3193)

Lawrence, A., et al. 2007, MNRAS, 379, 1599

Leauthaud, A., et al. 2016, MNRAS, 457, 4021L

Lehmann, B. V., et al. 2017, ApJ, 834, 37L

Li, Y., et al. 2008, MNRAS, 389, 1419L

Liddle, A. 2003, imcs, book, L

Mahabal, A., et al. 2008, AN, 329, 288M

Mandelbaum, R., et al. 2006, MNRAS, 368, 715M

Mann, R. G., et al. 1998, MNRAS, 293, 209M

Maraston, C., & Strömbäck, G. 2011, MNRAS, 218, 2785M

Maraston, C., et al. 2013, MNRAS, 435, 2764M

Marchesini, D., et al. 2009, ApJ, 701, 1765M

Matthee, J., et al. 2017, MNRAS, 465, 2381M

Matthews, D. J., & Newman, J. A. 2010, ApJ, 684, 88N

——. 2012, ApJ, 745, 180M

Ménard, B., et al. 2013, preprint (arXiv:1303.4722)

Mitchell, P. D., et al. 2013, MNRAS, 435, 87M

Mo, H. J., & White, S. D. M. 1996, MNRAS, 282, 347M

Montero-Dorta, A. D., et al. 2017, ApJ, 848L, 2M

Moustakas, J., et al. 2013, ApJ, 767, 50M

Muzzin, A., et al. 2013, ApJ, 777, 18M

Myers, A. D., et al. 2015, ApJS, 221, 27M

Nelson, D., et al. 2015, A&C, 13, 12N

Newman, J. A. 2008, ApJ, 684, 88N

Newman, J. A., et al. 2013, ApJ, 208, 1

Nikoloudakis, N., et al. 2013, MNRAS, 429, 2032N

Norberg, P., et al. 1988, MNRAS, 328, 64N

Ntampaka, M., et al. 2016, ApJ, 831, 135N

——. 2018, arXiv:1810.07703

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

Pacifici, C., et al. 2012, MNRAS, 421, 2002P

Pallero, D., et al. 2018, arXiv:1812.08802

Panter, B., et al. 2007, MNRAS, 378, 1550P

Pasquet, J., et al. 2019, A&A, 621A, 26P

Peebles, P. J. E., & Hauser, M. G. 1974, ApJS, 28, 19P

Peebles, P. J. E., et al. 1980, lssu, book, P

Pérez-González, P. G., et al. 2008, ApJ, 675, 234P

Perlmutter, S., et al. 1999, ApJ, 517, 565P

Phillipps, S., & Shanks, T. 1987, MNRAS, 227, 115P

Phillipps, S., et al. 1985, MNRAS, 212, 657P

Pillepich, A., et al. 2018, MNRAS, 473, 4077P

Planck Collaboration, et al. 2014, A&A, 571A, 16P

——. 2018, arXiv:1807.06209

Prakash, A., et al. 2016, ApJS, 224, 34P

Press, W. H., & Schechter, P. 1974, ApJ, 187, 425P

Qu, Y., et al. 2017, MNRAS, 464, 659Q

Rahman, M., et al. 2015, MNRAS, 447, 3500R

——. 2016a, MNRAS, 460, 163R

——. 2016b, MNRAS, 457, 3912R

Reddick, R. M., et al. 2013, ApJ, 771, 30R

Reid, B. A., et al. 2012, MNRAS, 426, 2719R

Riess, A. G., et al. 1998, AJ, 116, 1009R

Robotham, A. S. G., et al. 2011, MNRAS, 416, 2640R

Ross, A. J., et al. 2011, MNRAS, 417, 1350R

——. 2012, MNRAS, 424, 564R

Rubin, V. C., & Ford, W. Kent, J. 1970, ApJ, 159, 379R

Sánchez, C., et al. 2014, MNRAS, 445, 1482S

Sanders, D. B., et al. 1988, ApJ, 325, 74S

Schawinski, K., et al. 2014, MNRAS, 440, 889S

Schaye, J., et al. 2015, MNRAS, 446, 521S

Schmidt, S. J., et al. 2013, MNRAS, 431, 3307S

Schneider, M., et al. 2006, ApJ, 651, 14S

Scottez, V., et al. 2016, MNRAS, 462, 1683S

——. 2018, MNRAS, 474, 3921S

Seldner, M., & Peebles, P. J. E. 1976, ApJ, 227, 30S

Sheth, R. K., & Tormen, G. 1999, MNRAS, 308, 119S

Skibba, R., et al. 2006, MNRAS, 369, 68S

Smee, S. A., et al. 2013, AJ, 146, 32S

Somerville, R. S., & Davé, R. 2015, ARA&A, 53, 51S

Sousbie, T. 2011, MNRAS, 414, 350S

Sousbie, T., et al. 2011, MNRAS, 414, 384S

Springel, V., et al. 2005a, MNRAS, 364, 1105S

——. 2005b, MNRAS, 361, 776

Stinson, G. S., et al. 2013, MNRAS, 428, 129S

Stoughton, C., et al. 2002, AJ, 123, 485S

Strauss, M. A., et al. 2002, AJ, 124, 1810S

Suzuki, N., et al. 2012, ApJ, 746, 85S

Swanson, M. E. C., et al. 2008, MNRAS, 387, 1391S

Taylor, E. N., et al. 2011, MNRAS, 418, 1587T

Tegmark, M., & Peebles, P. J. E. 1998, ApJ, 500L, 79T

Tinker, J. L., et al. 2017, ApJ, 839, 121T

Tojeiro, R., et al. 2007, MNRAS, 381, 1252T

——. 2017, MNRAS, 470, 3720T

van Daalen, M. P., & White, Martin, . 2018, MNRAS, 476, 4649V

van Daalen, M. P., et al. 2016, MNRAS, 458, 934V

Vargas-Magaña, M., et al. 2013, AA, 554A, 131V

Wake, D. A., et al. 2008, MNRAS, 387, 1045W

Wang, L., et al. 2013, MNRAS, 431, 648W

Wechsler, R. H., & Tinker, J. L. 2018, ARA&A, 56, 435W

Wechsler, R. H., et al. 2006, ApJ, 652, 71W

White, M., et al. 2015, MNRAS, 447, 234W

White, S. D. M. 1978, MNRAS, 184, 185W

Wild, V., et al. 2009, MNRAS, 395, 144W

Yang, X., et al. 2006, ApJ, 638L, 55Y

York, D. G., et al. 2000, AJ, 120, 1579Y

Zarrouk, P., et al. 2018, MNRAS, 477, 1639Z

Zehavi, I., et al. 2005, ApJ, 630, 1Z

Zentner, A. R. 2007, IJMPD, 16, 763Z

Zentner, A. R., et al. 2014, MNRAS, 443, 3044Z

——. 2019, MNRAS, 485, 1196Z

Zolotov, A., et al. 2015, MNRAS, 450, 2327Z

Zu, Y., & Mandelbaum, R. 2016, MNRAS, 457, 4360Z

Zwicky, F. 1933, AcHPh, 6, 110Z