University of Sussex

A Galaxy Cluster Finding Algorithm for Large-Scale Photometric Surveys

Leon Baruah

Submitted for the degree of Doctor of Philosophy

University of Sussex

January 2015

Declaration

I hereby declare that this thesis has not been and will not be submitted in whole or in part to another University for the award of any other degree.

Signature:

Leon Baruah

UNIVERSITY OF SUSSEX

Leon Baruah, Doctor of Philosophy

A Galaxy Cluster Finding Algorithm for Large-Scale Photometric Surveys

Summary

As the largest gravitationally bound objects in the Universe, galaxy clusters can be used to probe a variety of topics in astrophysics and cosmology. This thesis describes the development of an algorithm to find galaxy clusters using non-parameteric methods applied to catalogs of galaxies generated from multi-colour CCD observations. It is motivated by the emergence of increasingly large, photometric galaxy surveys and the measurement of key cosmological parameters through the evolution of the cluster mass function.

The algorithm presented herein is a reconstruction of the successful, spectroscopic cluster finding algorithm, C4 (Miller et al., 2005), and adapting it to large photometric surveys with the goal of applying it to data from the Dark Energy Survey (DES). AperC4 uses statistical techniques to identify collections of galaxies that are unusually clustered in a multi-dimensional space. To characterize the new algorithm, it is tested with simulations produced by the DES Collaboration and I evaluate its application to photometric datasets. In doing so, I show how AperC4 functions as a cosmology independent cluster finder and formulate metrics for a "successful" cluster finder.

Finally, I produce a galaxy catalog appropriate for statistical analysis. C4 is applied to the SDSS galaxy catalog and the resulting cluster catalog is presented with some initial analyses.

"There is a single light of science, and to brighten it anywhere is to brighten it everywhere."

— Isaac Asimov

Acknowledgements

I am grateful to Dr Kathy Romer, Prof. Chris Miller, and Dr Jon Loveday, who took on various supervisory roles during the course of this Ph.D. I cannot undervalue the thanks I owe my main supervisor, Kathy Romer, for taking me on as a student, for her thorough proof-reading of this thesis, and lending her support when I really needed it. I am indebted to the help of Chris Miller, who provided me with the original C4 algorithm that formed the basis of this thesis. I would like to thank him for his constructive advice, for hosting me at Michigan, for the time he has spent helping me flesh out my ideas and putting me on the right track when I was in need of direction. His generous donation of compute time on the FLUX HPC cluster made this thesis possible.

I am also grateful to the Astronomy department at Sussex for their advice over the years: I heard it well even though some of it may have taken time to sink in! Further thanks go to Andrew Liddle whose willingness to sit and listen to me and kind advice will not be forgotten.

Thank you to my examiners, Jon Loveday and Alfonso Aragón-Salamanca, for their considered feedback and hugely thoughtful input to the full development of this thesis.

Numerical computations were done on the Sciama High Performance Compute (HPC) cluster (supported by the ICG, SEPNet and the University of Portsmouth), and the FLUX HPC cluster at the University of Michigan (UMich). I would like to acknowledge financial support from 2007 to 2011 through a Science and Technology Facilities Council studentship. I was also able to attend various DES collaboration meetings thanks to funding from STFC, Ohio State University, and the DES collaboration itself. I attended an NVO Summer School part funded by NVO and received funding for my visit to Michigan.

This thesis would also not have been possible if it weren't for key contributions from around the world. I'd like to thank the many faces of the DES collaboration, who have been both welcoming and inspiring in many ways during my studentship. I must thank the DES simulations team, particularly Michael Busha and Risa Wechsler for driving the production and high quality of the DES Catalog Simulations. I thank Brian Gerke for sharing his membership matching code, and allowing me to collaborate in its development. I am

grateful for Dongryeol Lee's technical assistance and advice on the kth Nearest Neighbour finder. I must thank Gary Burton for technical assistance and his prompt, helpful, and friendly support, on using the Sciama computing cluster at Portsmouth. I also thank Peter Thomas for donating IDL licenses to the Sciama cluster, which in turn permitted me to use it.

I'd like to thank the many people who I've lived with during my Ph.D., and tolerated my sometimes-possibly-coherent rambles, here (in no particular order): Lauren Coleman, Matt Ryan, Sophie Osborne, Isaac & Yuko Roseboom, Cath Berry, Alex Seabrook, Tabitha Rohrer, Ness Kabas, Bernard Mills, and Monica Ross. They have made my time in Brighton all the richer in many ways. I'm particularly thankful to Pete Offord, Antony Lewis, and Ruth Pearson, for giving me a place to stay during my hauntingly regular spates of homelessness.

I must thank my fellow starting DPhil students, whose respective journeys have been inspirational; particularly Anna Jordanous and Sandra Deshors, whose continued friendship and support have been a lifeline.

Likewise, I would like like to thank my many friends and coworkers at the Physics and Astronomy dept, those that have since left, as well as those that since come and gone, for their help (and hindrance!) over the years, and making Sussex an enjoyable place to be. Special shout outs go to (in no particular order): David Parkinson, Leonidas Christodoulou, Duncan Farrah, Will Demeri-Watson, Donough Regan, Gemma Anderson, Darren Baskill, Nicola Mehrtens, Matt Thomson, Ippocratis Saltas, and Gwen Lefeuvre; all of whom have given me respite from sanity and insanity alike, and have broadened my horizons.

Finally, I thank my family, Richie, Manju, and Dwijen, for their continued love and support and for always just being there. Their belief in me from the time that I started this journey up until the present day has been unwavering; I don't know where I would be without it. This is for you.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction to Cosmology

1.1.1 The Cosmic Expansion

In the earlier half of the 20^{th} century, Edwin Hubble noted that several spiral nebulae appeared to be in recession; moving away from the Milky Way. Using observations of Cepheid variables (a kind of star) as *standard candles*, he was able to calculate the distance to these nebulae and, by relating them to their recessional velocities, discovered one of the most fundamental observations of cosmology: that of the expanding universe. The expansion as observed by Hubble is commonly expressed as,

$$v = Hd\,, \tag{1.1}$$

where v is the recessional velocity of the galaxy, d is the distance, and H is the *Hubble parameter*. The relation states that the more distant an observable object is in the universe, the faster it is receding from us.

A photon emitted by galaxy traversing the expanding universe will itself expand with the universe, and its wavelength, λ, will increase. When this photon is observed some time later, its wavelength will have increased by a factor

$$z = \frac{\lambda_{obs} - \lambda_{em}}{\lambda_{em}}\,, \tag{1.2}$$

where λ_{em} and λ_{obs} describe the photon's emitted and observed wavelengths, respectively, and z is called the *redshift*. Relating the wavelength shift to a difference in velocity between the source and the observer via the Doppler law,

$$z = \frac{\lambda_{obs} - \lambda_{em}}{\lambda_{em}} = \frac{v}{c}\,, \tag{1.3}$$

where c is the speed of light. Thus, the redshift, z, is a measure of the observed recessional velocity, which is in fact a measure of the expansion of the universe in the time taken between the emission and observation of a photon. The wavelength of a photon at the time of emission, λ_{em}, is proportional to the scale factor of the universe at that time, a_{em}. It follows that

$$1 + z = \frac{\lambda_{obs}}{\lambda_{em}} = \frac{a_{obs}}{a_{em}}\,, \tag{1.4}$$

where a_{obs} is the scale factor today and is conventionally defined as $a_{obs} = 1$.

The formal description of the Standard Model of Cosmology rests on two postulates:

The Cosmological Principle , which states that the universe appears homogeneous and isotropic on sufficiently large scales. From an observer's point of view, the Cosmological Principle implies that that observations of the state of the universe will appear to be consistent from any viewpoint and that we (or any observer in the universe) do not occupy a privileged position in space.

Einstein's theory of General Relativity , which folds together space and time into a single entity called *space-time*, where the presence of matter induces curvature in the fabric of space-time. This curvature of space-time is General Relativity's description of gravity.

On this basis, the expanding universe can be described with the Friedmann equation,

$$H^2 = \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{k}{a^2}\,, \tag{1.5}$$

where G is the gravitational constant, ρ is the energy density, and k describes the geometry of the universe. A useful quantity to define is the *critical density*, ρ_c, the density required that the universe's geometry is flat, i.e., $k = 0$, and the related ratio, Ω, also known as the density parameter.

$$\rho_c = \frac{3H^2}{8\pi G}, \tag{1.6}$$

$$\Omega = \frac{\rho}{\rho_c}. \tag{1.7}$$

As measured (to excellent precision) by WMAP (Spergel et al., 2003), the universe evolves in a matter-dominated expansion such that $\rho \propto a^{-3}$ in a universe that has flat or very close to flat curvature $k = 0$ and the energy density of the universe is very close to $\Omega = 1$.

In 1998, observations of high redshift supernovae Riess et al. (1998) provided the first strong evidence that the expansion of the universe was accelerating, i.e., $\ddot{a} > 0$, and is often quantified by the *deceleration parameter*, q (current value q_0):

$$q = -\frac{a\ddot{a}}{\dot{a}^2}, \tag{1.8}$$

which is negative for acceleration. In the Friedmann equation, this acceleration is parameterized by the *cosmological constant*, Λ, and appears as,

$$H^2 = \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{k}{a^2} + \frac{\Lambda}{3}. \tag{1.9}$$

Defining the density parameters of the curvature and cosmological constant by Ω_k and Ω_Λ, respectively, as

$$\Omega_k = -\frac{k}{a^2H^2} \text{ and } \Omega_\Lambda = \frac{\Lambda}{3\mathrm{H}^2}\,, \tag{1.10}$$

the equation becomes

$$\Omega_m + \Omega_k + \Omega_\Lambda = 1 \tag{1.11}$$

$$\text{or} \tag{1.12}$$

$$\Omega_m + \Omega_\Lambda = 1 \tag{1.13}$$

since curvature has been found to be consistent with $k = 0$ (Hinshaw et al., 2013), and in which Ω_m describes the energy density of all matter.

A second key equation is the fluid equation, which can be used to describe Λ as a fluid, with energy density ρ_Λ and presssure p_Λ

$$\dot{\rho_\Lambda} + 3\frac{\dot{a}}{a}\left(\rho_\Lambda + \frac{p_\Lambda}{c^2}\right) = 0\ ,\ \text{where } \rho_\Lambda = \frac{\Lambda}{8\pi G}. \tag{1.14}$$

If Λ is constant, then ρ_Λ is also, so its derivative goes to zero, and the equation holds only if the pressure $p_\Lambda = -\rho_\Lambda c^2$. This gives an indication of the exotic nature of the Λ fluid, where increasing the pressure *reduces* the density. If one imagines holding a ball of such a fluid, the more one squeezed the ball, the less resistance it would produce in response to being squeezed, i.e. the easier it is to squeeze.

It is a possibility that the observed accelerated expansion of the universe may be a transient state of the expansion, which can be described by reparameterizing the relationship between p_Λ and ρ_Λ as

$$p_\Lambda = w\rho_\Lambda c^2\,, \tag{1.15}$$

which is described as the equation of state for *dark energy*. If Λ is constant then $w = -1$, but an accelerating expansion can be brought about where $w < -1/3$. Recent observational

campaigns have been focussing on measuring the accelerating expansion, and the value of w.

1.1.2 Dark Energy

Whilst the nature of dark energy remains elusive, its existence as either the cosmological constant or time-varying *quintessence* presents us with some consequences for cosmology. Dark energy contributes about 70 percent of the energy density of the universe, and matter (dark and baryonic) about 30 percent. That the two are comparable is a remarkable coincidence. In terms of Ω, these are parameterized as $\Omega_m \simeq 0.3$ and $\Omega_\Lambda \simeq 0.7$. The CMB angular power spectrum shows a near-Euclidean space-time ($k = 0$, thus $\Omega_k = 0$ in Equation (1.11)), implying that $\Omega_m + \Omega_\Lambda = 1$. A dark energy component explains how a $\Omega_m \simeq 0.3$ universe is possible in such a geometry.

A cosmological constant, Λ, could be interpreted as a *vacuum energy* of free space; however, quantum theory predicts an energy roughly 10^{120} greater than observed. Observations at increasing redshift would help determine whether $\dot{w} \neq 0$, and decide whether the dark energy exists in the form of the cosmological constant or a time-varying quintessence.

1.2 Introduction to Galaxy Clusters

1.2.1 History

Building on earlier work on remote galaxies, Shapley (1933) assembled the first extragalactic catalog of 25 groups of galaxies from studies of the general distribution of galaxies across the sky. The galaxies identified in this survey were not uniformly distributed and were observed to cluster independently of the distribution of stars, confirming that this clustering was extragalactic in origin. Groups of galaxies in this catalog were populated by up to 472 galaxies to a plate magnitude of ∼18.0, although this membership count, or $N_{\rm gals}$, was made fairly arbitrarily, as the paper notes:

> "Some of these systems are spheroidal in form, centrally concentrated and of rather definite boundary. The majority of the groups, however, *are of irregular form, with indefinite boundaries, and small in membership.*"

During the following decades, several compilations of groups and clusters of galaxies were made, most notably by large surveys performed by Abell (1958) and Zwicky et al. (1961). It was the preparation of the Abell (1958) cluster catalog, with its exacting selection criteria for what galaxy membership qualifies as cluster content, that allowed stat-

istical comparison to be made between, and of, clusters. The Abell catalog characterized clusters as objects with *intrinsic* properties, rather than otherwise unrelated galaxies that appear as concentrations as viewed from Earth. From a starting sample of 2,712 clusters from the Palomar Observatory Survey, 1,682 were selected that met specific criteria for inclusion. These "Abell clusters" were defined as objects with the following properties:

1. They contained more than 50 galaxies within some fixed radial aperture. This radius was deemed to be arbitrary, so long as the same physical distance was used for each cluster. In Abell (1958), the radius used was approximately 1.5 Mpc, although the quoted measurement was half this size, due to the use of a Hubble constant twice that found in more recent extragalactic studies.

2. The magnitude of the faintest galaxy was no more than 2 magnitudes fainter than the 3rd brightest galaxy in that aperture.

3. They were distant enough that they did not extend over several photographic plates. The photographic plate was the standard astronomical observation medium before the CCD used in the modern era was developed and deployed on telescopes. Each plate covered a $6\overset{\circ}{.}6$ square area, so this discounted the Virgo cluster, but did include the Coma Cluster (Bower et al., 1992).

4. They appeared at high galactic latitudes. As galaxies at the time were identified by visual inspection of the photographic plates, this criterion was placed to escape the dense stellar fields in the galactic disk that would otherwise obscure areas of cosmological interest.

With these first characterizations of clusters by their observed content (i.e., their galaxies), statistical measurements and physical insight could be made concerning the contents and distribution of these clusters of galaxies. Indeed, Abell was able to indicate that the clusters were distributed in a way such that they themselves were clustered. Measurements of the number and brightness of the constituent galaxies, their velocity dispersions and their morphologies were built up into a picture of a typical cluster of galaxies, or *galaxy cluster*.

The launch of the Einstein X-ray Observatory (HEAO-2) in 1978, the first imaging X-ray telescope, introduced a new dimension to the contemporary observations of clusters. Previous X-ray observations of the galaxies in clusters had been performed by mounting X-ray detectors on rockets and balloons (Bahcall, 1977), but most of the detections were

limited in their energy and angular resolution. The X-ray satellite UHURU gave the first indication of clusters being powerful X-ray sources but similarly suffered limited angular resolution, such that the chance of there being a coincidental, galactic X-ray source could not be ruled out. Forman et al. (1979) undertook observations of galaxies in the nearby Virgo cluster in X-ray bands using the Einstein X-ray Observatory, probing energies up to $\lesssim$ 4keV. Their observations pointed out that, in addition to the previously seen low energy X-rays associated with clusters (assumed to be from the galaxies themselves), there was a very luminous, very energetic component aligned with the centre of the cluster. This energetic component was quickly allied to the cluster itself in the form of hot intergalactic gas occupying the deep gravitational potential formed by the cluster (as proposed by Felten et al., 1966), trapped by the well's gravitational forces and heated through shocks and adiabatic compression as it falls into the well to temperatures of millions of degrees. The X-ray observation of clusters has since become a mature component of cluster astronomy, and this hot gas component is now understood as the *Intracluster Medium*, or ICM, which helps inform a complete picture of the baryonic matter component in clusters of galaxies.

1.2.2 Galaxy Clusters as Astrophysical Probes

After the epoch of reionization, when the universe became transparent, matter fell into the gravitational potentials formed by the underlying *dark matter* distribution (see below), simultaneously increasing the matter content of those potentials and the depth of those potentials. Over time, the potentials themselves fall into each other and merge, creating ever larger potentials with increasing amounts of matter; this scenario is commonly known as *hierarchical structure formation*. These potentials assemble into giant 3-dimensional structures, forming a network of galaxies distributed in the form of filaments, walls, and clusters, interspersed by voids from which matter has largely been displaced. This interconnected series of structures are collectively referred to as *Large Scale Structure* (LSS).

Clusters of galaxies typically contain many tens or hundreds of galaxies that are gravitationally bound together. Their constituent galaxy populations are typically dominated by passively evolving, E/S0 ridgeline (*early type Elliptical/Lenticular*) galaxies, with the centre of the cluster commonly occupied by a very massive, bright elliptical type galaxy. Clusters are interesting because they collapse independently of large scale structure; the accumulation of matter in these overdensities affects particles such that their movement is dominated by the cluster's gravitational potential rather than the Hubble Flow.

In 1933, Fritz Zwicky measured the relative velocities of individual galaxies in the

Coma Cluster by examining their redshifts. By looking at the distribution of redshifts, or *velocity dispersion*, in the Coma Cluster, he measured the potential of the cluster by applying the Virial Theorem,

$$\langle \Phi \rangle = -2\langle T \rangle \,, \tag{1.16}$$

which states that the average potential energy, Φ, of an isolated gravitational bound system is equal to minus twice the averaged kinetic energy, T. He found that the gravitational potential described by the movement of the galaxies in the cluster required ~ 100 times more mass than observed in the stars and galaxies. This missing matter problem led to the initial proposal for the existence of *dark matter*, a non-baryonic, gravitationally attractive mass component of the energy density of the universe.

The physics of dark matter is one of the most significant observational and theoretical challenges in modern day astrophysics. However, our inability to probe its composition with telescopes that focus on the detection of objects across the electromagnetic spectrum hasn't hindered our ability to infer its existence. In addition to the velocity dispersion measurements in clusters, velocity dispersions of stars in elliptical galaxies, spiral galaxy rotation curves, and the power spectrum of the Cosmic Microwave Background (CMB) provide evidence for the existence of dark matter.

Figure 1.1 shows the iconic galaxy cluster 1E 0657-56, the "bullet cluster", which provides more evidence of the existence of dark matter (Markevitch et al., 2004; Clowe et al., 2004). The optical component of the image shows two concentrations of galaxies that have recently passed through each other, whilst the X-ray component appears to trail each concentration of galaxies, with the smaller gas component showing a very prominent shock front, identifying it as the "bullet". This is interpreted as the galaxy components of the clusters (treating the bullet cluster as two "bullet" and "target" clusters) passing through each other, whilst the gas components interact, slowing them down. Measurements of this interaction gives insight to extreme plasma physics. Using *lensing*, the total mass in the cluster can be measured by identifying correlated radial deformation of background galaxy images, which is caused by the bending of light by the cluster's gravitational potential. The bulk of the mass is traced by this lensing and identified by the blue component of the image. The optical mass component can be evaluated from the galaxy luminosities, whilst the X-ray gas mass can also be measured by its temperature. Studies of this cluster show that the gas mass exceeds that of the galaxies, whilst the detected lensing signal is correlated with the galaxy distributions, directly pointing to a further mass component

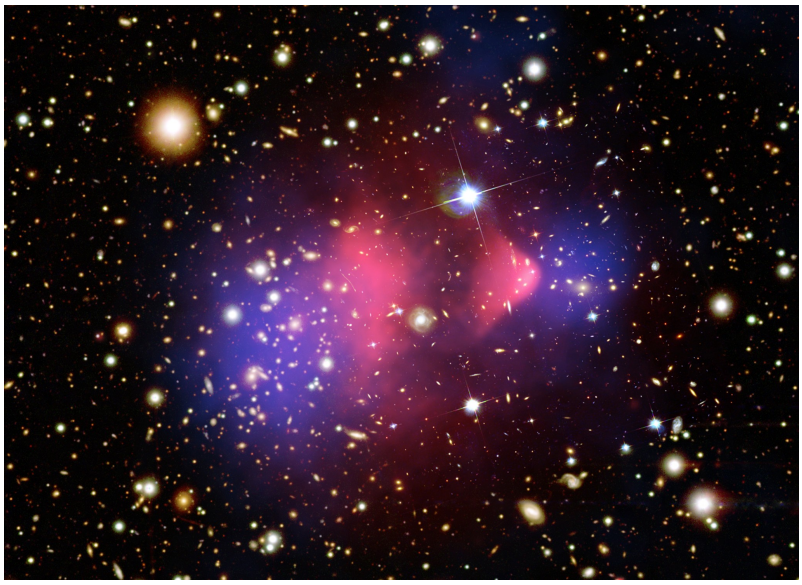


Figure 1.1: This is a composite image of galaxy cluster 1E 0657-56, the "Bullet Cluster," which is a system of two galaxy clusters during a merging event that are observed in the plane of the sky. As well as the optical distribution of galaxies taken by the Hubble Space Telscope, the X-ray emission from the hot ICM gas is highlighted in red, and a lensing map of the background galaxy populations indicates the bulk of the mass in blue. Having recently passed through each other, the lensing distortion shows the bulk of the mass has remained with the galaxies, whilst the gas has been impeded. [Figure credit: NASA APOD (Markevitch et al., 2004; Clowe et al., 2004)]

allied to the galaxies, i.e., the dark matter.

Being able to determine the strength of the potential in which a galaxy cluster sits allows us to investigate the astrophysics within the cluster. The evolution of galaxies within these massive potentials is an active area of ongoing research as a cluster potential provides a near consistent environment for a large sample of galaxies. The identification of galaxies by morphology and comparing them to less dense (*field*) environments (Hutchings et al., 2002; Boselli and Gavazzi, 2006), or to similar environments in the past (Harrison et al., 2012; Stott et al., 2012), builds up a picture of the assembly of different types of galaxies and their evolution against the evolving mass distribution in the universe.

Observations in the X-ray give observers the opportunity to probe the gas content of galaxy clusters, giving insight into the composition of the galaxies in the cluster and of the feedback processes governing the reinjection of energy and material into the ICM. The ICM acts as a fossil record of galaxy interactions, from measurements of the global distribution of metals (Maughan et al., 2008) to observations of massive bubbles formed in the gas (Brüggen and Kaiser, 2002). At high redshift, the picture of a passively evolving core group of galaxies begins to break down, as spectroscopic followup of clusters (found by their X-ray emission) reveal some of the cluster galaxies in $z \gtrsim 1.5$ clusters are undergoing

a period of star formation, or *starburst* (Fassbender et al., 2011; Hilton et al., 2010).

The diverse opportunities for probing galaxy clusters make them interesting astrophysical laboratories. Observing how different environmental conditions affect the assembly of stellar and galactic mass affords insights into the formation of structure against the background of the evolving universe. Similarly, acknowledging the relationship between astrophysical observations and properties of the clusters themselves, we can use clusters to infer some of the fundamental properties of the universe.

1.3 Cosmological Constraints from Galaxy Clusters

Galaxy clusters are the largest, gravitationally bound structures in the observed universe, signposting the peaks of the large-scale matter distribution and allowing us to use them as a measure of matter and density distribution. Considering the universe as a one-dimensional density field, galaxy clusters represent the peaks of this density field. They represent an interesting demographic in the universe, with their assembly being sensitive to the size and distribution of the most massive matter density perturbations left from the epoch of recombination. Galaxy cluster cosmology is concerned with understanding the matter distribution of the universe on large scales.

Galaxy clusters are massive enough that their composition should be representative of the whole universe, displaying ratios of stars to gas to dark matter that are consistent with any large volume of the universe and being broadly insensitive to their location in the sky or their redshift. Through a combination of X-ray and optical measurements (as demonstrated by the bullet cluster; Figure 1.1, §1.2.2), the composition of galaxy clusters is found to be of order 3% stars, 12% gas, and 85% dark matter (Voit, 2005), which is consistent with proportions in the universe measured by WMAP (Komatsu et al., 2011). Knowledge of the distribution of clusters allows us to extrapolate to the universe as a whole, informing us about Ω_m. Indeed, cluster measurements predicted a non-unity Ω_m (White et al., 1993) before WMAP ushered in the age of precision cosmology (Spergel et al., 2003).

Knowing the baryon fraction of a galaxy cluster (more commonly referred to as the *gas fraction*, since it is the chief baryon component), and holding the hypothesis that the relative composition of galaxy clusters is constant throughout the universe, Allen et al. (2008) and Vikhlinin et al. (2009) measured the compositions and distances to clusters with X-ray observations alone and were able to apply constraints on the nature of the dark energy.

Larger samples of galaxy clusters enable cosmology by simply counting the numbers of clusters of a given mass throughout the universe. Given the early universe consisted of Gaussian-distributed density perturbations, the first structures to form in the universe would inhabit the deepest gravitational potentials formed by these perturbations, and so the evolved size of these potentials is sensitive to their initial amplitude and number density. Relating this to observations of clusters in the present day, number densities of clusters are used to determine the density contrast of the universe on scales of $8\,h^{-1}\,\mathrm{Mpc}$, giving the cosmological variable σ_8. By measuring σ_8 at differing redshifts, one can build a picture of the evolution of the density contrast of matter, which is itself sensitive to values of Ω_m. σ_8 is typically found to be in the region $0.8 \lesssim \sigma_8 \lesssim 0.9$ (Eke et al., 1996).

How galaxies are distributed in galaxy clusters, or with respect to the matter-density distribution of the universe at large, has implications for cosmology (Lima and Hu, 2004; Rozo et al., 2007) and is an area of active research (Fassbender et al., 2011; Budzynski et al., 2012; Harrison et al., 2012; Stott et al., 2012).

1.3.1 Cosmological Measurements with Cluster Counts

Observations of clusters in different wavebands (§1.4), have led to a number of independent methods to constrain cosmology. Methods such as sampling the gas fraction of clusters rely on gas measurements (through X-ray and SZ observations) of the most massive clusters to determine Ω_m. As intracluster gas is not sampled in optical data, alternative measures of cosmology are drawn from examining the distribution of clusters in large-scale surveys.

Cosmology derived from cluster counts (the distribution of clusters in the universe) is itself motivated by a model of galaxy formation formulated by Press and Schechter (1974).

A spherical region of radius R will contain a mass $M = {}^{4\pi}/_{3}\rho_0 R^3$, where ρ_0 is the background density. The premise of the Press-Schecter approach uses the assumption that the initial density field is Gaussian, so the probability of being in a region of the universe of density, δ_M, is

$$P(\delta_M)d\delta_M = \frac{1}{\sqrt{2\pi}\sigma_M}\exp\left(-\frac{\delta_M^2}{2\sigma_M^2}\right)d\delta_M\,, \tag{1.17}$$

where δ_M is the density of a volume that has been smoothed on a scale R that encloses a mass M, and σ_M indicates the RMS of the smoothed density field.

Press-Schecter then state a spherical volume within this Gaussian density field (smoothed on some scale, R) that contains a mass M will undergo gravitational collapse if its density

exceeds some critical threshold, δ_c, such that,

$$P_{collapse}(M) = \int_{\delta_c}^{\infty} P(\delta_M) d\delta_M \,. \tag{1.18}$$

This has the notable drawback that half of the universe is underdense and does not undergo gravitational collapse. By observation of structure in the universe, this cannot be true. Press and Schechter remedied this by including the multiplicative factor of 2, seen in the equation below, to account for matter in underdense regions accreting to "neighbouring lumps in overdense regions". This was consequently explained by various groups (Peacock and Heavens, 1990; Bower, 1991; Bond et al., 1991) as a consequence of points that are located in underdense regions being in regions above δ_c when scales larger than R are considered, and should therefore be included in the fraction of collapsed objects that are more massive than $M(R)$.

To calculate the number of collapsed objects of mass M, $P_{collapse}(M+dM)$ is subtracted from the $P_{collapse}(M)$, where dM is a small mass interval, and we multiply through by the average number density, ρ/M, per unit volume.

$$\begin{aligned} n(M)dM =& 2\frac{\rho_0}{M}[P_{coll.}(M) - P_{coll.}(M+dM)] & (1.19)\\ =& -2\frac{\rho_0}{M}\frac{dP_{coll.}}{dM}dM & (1.20)\\ =& -2\frac{\rho_0}{M}\frac{dP_{coll.}}{d\sigma_M}\frac{d\sigma_M}{dM}dM \,. & (1.21) \end{aligned}$$

Evaluating the derivative $^{dP_{coll.}}/_{d\sigma_M}$, one finds,

$$n(M)dM = -\sqrt{\frac{2}{\pi}}\frac{\rho_0}{M}\frac{d\sigma_M}{dM}\frac{\delta_c}{\sigma_M^2}\exp\left(-\frac{\delta_c^2}{2\sigma_M^2}\right)dM \,. \tag{1.22}$$

Cosmology is derived through the number counts, the estimations of M, and σ_M, which are redshift dependent.

1.4 Observations of Galaxy Clusters

Cluster cosmology is effectively the study of the massive gravitational potentials in the universe that contain matter. The baryons that occupy the galaxy cluster, chiefly in the form of stars in galaxies and intracluster gas, form only a fraction of the total mass of a cluster. It is these baryons that provide the primary observations of galaxy clusters. From the early days of cluster measurement, the number of galaxies, or *richness*, of a rich cluster was seen to scale with velocity dispersion, and was taken as a quantity that scaled with mass. The mass of a cluster can be 'measured' by a number of features besides velocity

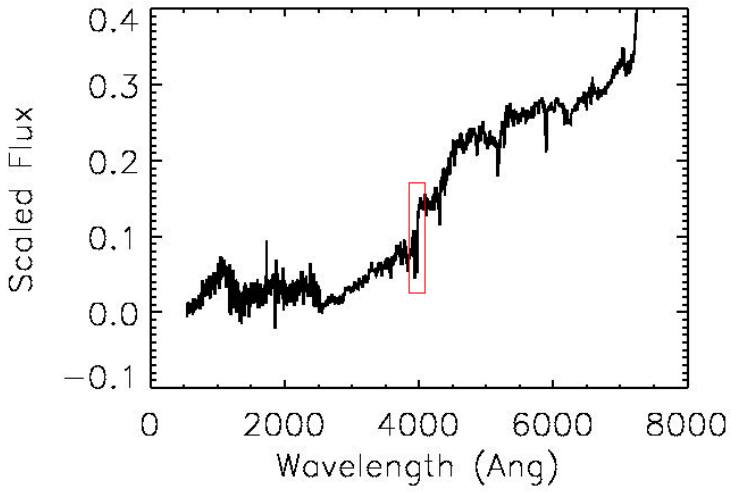


Figure 1.2: The above figure is an example of a Luminous Red Galaxy (LRG) spectrum, constructed by co-adding 784 rest-frame LRG spectra obtained from the SDSS DR4 spectroscopic galaxy catalog (Adelman-McCarthy et al., 2006) using CasJobs*. The vertical axis describes flux whilst the horizontal axis describes wavelength. The red box identifies the 4000Å break.

dispersion and richness. These features are interpreted as measures of mass through the use of *scaling relations*, which quantify the observable in terms of the total mass of the cluster, inclusive of baryonic and dark matter.

1.4.1 Optical

From their initial discovery as extragalactic structures (§1.2.1), clusters of galaxies have long been known to consist of high densities of galaxies that have distinct characteristics that distinguish them from the field. The central cluster populations are usually dominated by early type galaxies, primarily S0 and elliptical types, which show no evidence of ongoing or recent star formation and are thus said to be passively evolving. By contrast, field populations of galaxies contain actively star-forming galaxies, mostly spiral and irregular types (*late types*), in addition to early type galaxies.

Bright, passively evolving, elliptical galaxies are commonly known as *Luminous Red Galaxies*, or *LRGs*, and are common amongst the early type galaxies that occupy a cluster. LRGs have a strong, specific spectral feature that enables them to be identified using photometric data: *the 4000Å break*, as shown in Figure 1.2.

The 4000Å break provides a strong indication of the shape of a galaxy's *spectral energy distribution* (SED) when it is decomposed into broad band magnitudes as observed by a multiband photometric survey (e.g., SDSS, §2.1.1). This is achieved through the

*http://skyserver.sdss3.org/casjobs/

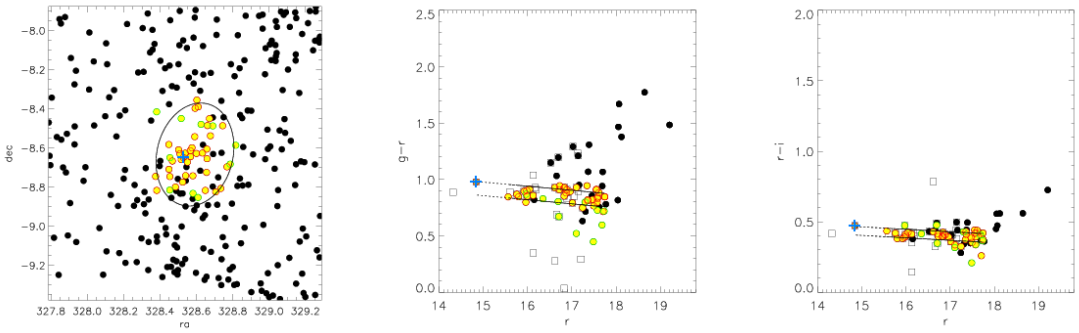


Figure 1.3: Cluster at $z \simeq 0.1$ from SDSS DR2 spectroscopic data (Miller et al., 2005). The plots are the projected sky distribution (*left*), the $(g-r)$ versus r-band colour-magnitude diagram (*centre*) and the $(r-i)$ versus r-band colour-magnitude diagram. The Blue cross is the Brightest Cluster Galaxy. Black filled circles are every spectroscopic galaxy projected onto the sky whilst yellow filled circles are C4 galaxies within 1 Mpc of the cluster centre. The galaxies identified as part of the cluster form a distribution known as the red sequence which is indicated by the 1σ errors around a straight line fit to the galaxies, represented by the dotted lines. [Figure credit: Chris Miller *priv. comm.*]

measurement of the LRGs broad band *colour*, which measures the difference between the magnitude of one photometric band and the band immediately red-ward of it. When the magnitudes that comprise the colour straddle the 4000Å break, then as the break moves to longer wavelength due to cosmological redshift, colour very accurately tracks the redshift of the LRG.

The *Brightest Cluster Galaxy*, or *BCG*, by definition, is the brightest galaxy in a cluster environment. BCGs have characteristic properties; Figure 1.3 shows an example cluster (found by C4, §1.5.6) with the BCG identified in blue. Sitting at the bright end of the red sequence, they represent the brightest passively evolving galaxies and are typically located at or near the centre of a cluster. They are the most massive stellar structures assembled in the universe and are of astrophysical and cosmological interest in and of themselves (Bower et al., 1992; Bernardi et al., 2007; Budzynski et al., 2012) since their formation histories have implications for galaxy mass assembly in general. Using the Millennium Simulations (Springel et al., 2005), De Lucia et al. (2007) find that BCGs assembled most of their mass late in the universe's lifetime, around $z \simeq 0.5$, through the dry merger (a merging of galaxies without gas such that no star formation is triggered) of smaller, passively evolving ellipticals falling towards the cores of galaxy clusters. By contrast, Collins et al. (2009) observe populations of BCGs in galaxy clusters at $z > 1$ which, compared to their $z \simeq 0$ counterparts, appear to have already assembled most of their stellar mass.

Cluster mass has traditionally been calculated by velocity dispersion of the galaxies, as per Zwicky et al. (§1.2.2). In more recent times, other probes using the galaxy component

of the clusters, such as caustics (Geller et al., 1999), the red sequence richness (Rykoff et al., 2012), or scaling to the BCG alone (Stott et al., 2012) have been used to estimate cluster mass directly, or by proxy to other observables.

Also within the optical regime, lensing has been used to estimate line-of-sight measures of cluster mass. Weak Lensing works by statistically measuring the distortion of the detected shapes of background populations of galaxies by the mass of the cluster potential. This probe is unique in that none of the astrophysics of the cluster are probed, and thus it delivers a measurement of the total cluster potential; the sum of the baryonic and dark components (Kneib and Natarajan, 2012).

1.4.2 X-ray

As introduced in sections 1.2.2 and 1.3, X-ray observatories are used to measure properties of the overdensity of hot gas that occupies the centre of the cluster potential. The hot gas is optically thin and, as evidenced by its observation in X-ray wavelengths, emits X-ray radiation at temperatures of $\sim$30 to 100 million degrees. The X-ray emission itself is produced by thermal Brehmsstrahlung of the electrons in the cluster gas. Assuming the gas is in hydrostatic equilibrium, the X-ray emission can then be used to quantify gas, the most direct measurement being the X-ray temperature. Spectroscopic studies of the gas component reveal the presence of iron and other metals distributed in the ICM (Peterson et al., 2003). Numerous scaling relations exist to relate X-ray properties such as temperature (as mentioned above), luminosity, or gas mass, to the total mass of the cluster.

X-ray surveys have been highly successful at finding clusters of galaxies (Maughan et al., 2008; Vikhlinin et al., 2009; Mehrtens et al., 2011) and measuring their various gas properties and relating them directly to mass or to observations made at other wavelengths.

1.4.3 Microwave

Observations of clusters at microwave wavelengths depend upon the *Sunyaev-Zel'dovich Effect*. Photons from the CMB that were emitted at the time of recombination can be used as probes of the cluster gas. Through a process of *inverse Compton scattering*, a fraction of the passing CMB photons scatter off electrons in the hot intracluster gas and *gain* energy. This fractional energy gain is seen as a distortion in the CMB black body spectrum as a function of frequency, where lower frequency CMB photons are displaced to higher frequencies. The result is a net "shift" in frequency of the black body spectrum,

and it is the magnitude of this shift that can be used to measure the temperature of the gas, again assuming hydrostatic equilibrium of the gas.

An often cited advantage of the Sunyaev-Zel'dovich effect (SZ effect), is its insensitivity to redshift compared with other probes. However, to produce a significant distortion in the CMB, detection of clusters in this way is limited to the most massive clusters. Nevertheless, SZ-based searches for clusters of galaxies have recently begun delivering samples of clusters (Menanteau et al., 2010; Reichardt et al., 2012; Planck Collaboration et al., 2013a) which will provide another independent probe of cluster mass.

1.5 Optical Cluster Finding

With the advent of large format CCDs and digitised data, new cluster finding techniques were developed. In the following subsections, I give some background to cluster finding, before outlining some cluster finders developed for use with digitised data.

1.5.1 History of optical cluster finding

The key goal of cluster finding, in optical wavebands or otherwise, is to describe the gravitational potentials in which they reside. The search for these potentials with optical data has a long history (§1.2.1). Quasi 3-dimensional methods (Huchra and Geller, 1982) relied on redshift to describe local clusters and groups of galaxies, recovering typically of order ∼100s to ∼1000s of clusters over the whole sky. Limitations of photographic plates (Couch et al., 1991) permitted 2-dimensional cluster finding at higher redshifts by taking positions of faint galaxy populations at the expense of photometry. This led to a great increase in the risk of spurious association via projection effects, where groups of galaxies that appear close in 2 dimensions are not physically associated in 3 dimensions. To minimize these projection effects, a range of new cluster finding methods have been developed.

Matched Filter (Postman et al., 1996) represents one of the first generations of cluster finders to use measurements of the data itself to avoid spurious associations. It enhances the contrast between cluster galaxy distributions with respect to foreground and background galaxy distributions by extracting the maximum likelihood cluster galaxies, determined by a modelled Gaussian background distribution, and convolving it with a function of luminosity-weighted magnitude, and angular-weighted radial separation, within a moving $1.25\,h^{-1}$ Mpc box. Adaptive Matched Filter (Kepner et al., 1999) follows on from Postman et al., fitting a model of cluster and field populations to galaxies, but adapting

to errors in the observed redshift, and upon finding clusters, filtering further to produce more precise estimates of cluster richness and redshift.

1.5.2 Red Sequence

The red sequence cluster finding technique was developed by Gladders and Yee (2000) to detect clusters in two-band optical/near-IR imaging data. The eponymous "red sequence" describes the observation that the cores of previously observed rich clusters chiefly consist of a population of passively evolving, elliptical galaxies.

By employing colour cuts to eliminate foreground contamination and hone in on concentrations of LRGs (within $1.33\,h^{-1}\,\mathrm{Mpc}$), Gladders and Yee (2000) delivered a method that is both simple and effective at locating the cores of typical rich clusters. As well as locating the cluster in the sky, by right ascension and declination (R.A. and dec, respectively), the Cluster Red Sequence finder also locates clusters in redshift without the use of spectroscopic information, ushering an increase in the immediate availability of cluster distribution measurements throughout the surveyed universe. As such, locating the red sequence became the de facto measurement of cluster redshifts, and led to the creation of several other red sequence based cluster finders. The exploitation of the cluster red sequence as a finding tool was a cornerstone in modern day cluster finding techniques.

1.5.3 MaxBCG

MaxBCG (Koester et al., 2007) represents another cornerstone in cluster finding. It uses the Brightest Cluster Galaxies, or BCGs, to locate potential cluster locations, then searches some radial area around it for galaxies that form a red sequence, or E/S0 ridgeline, and then uses that red sequence to derive a redshift for that cluster. It has produced one of the largest galaxy cluster catalogs to date, containing 13,823 galaxy clusters from the photometric Sloan Digital Sky Survey (SDSS, §2.1.1), covering 7,500 square degrees of sky between redshifts 0.1 and 0.3.

To identify the BCGs, it first takes a full galaxy catalog and removes the least likely BCG candidates by evaluating galaxy colours $g - r$, $r - i$ and an i-band magnitude. The algorithm then checks galaxies up to $1\,h^{-1}\,\mathrm{Mpc}$ around each BCG and evaluates the likelihood that it is indeed the most likely BCG, and if it is the most likely cluster centre. For positive identifications, MaxBCG then returns the BCG and its associated (red sequence) cluster membership, which form the MaxBCG cluster catalog.

1.5.4 Voronoi Tesselation

Voronoi tesselation, as applied to cluster finding, is a 2-dimensional process that identifies overdensities of galaxies by characterizing their clustering strength with the aid of a Voronoi diagram. A Voronoi diagram takes a Euclidian, 2-dimensional distribution of points and draws a line half-way between each pair of points. These bisecting lines, or boundaries, all join such that any point along a boundary is equidistant between the nearest two points. As such, the area enclosed by a set of boundaries will contain a single point, and is called a Voronoi cell. Voronoi cluster finders plot galaxies in a 2-dimensional Voronoi grid (nominally, in R.A. and dec.) then by locating groups of some minimum number of Voronoi cells, which individually have areas smaller than some threshold, they define clusters.

As mentioned above, projecting galaxies onto a 2-dimensional plane this way often leads to projection issues, or line-of-sight blending, where galaxies that appear close in 2-dimensions are not physically connected (i.e., they are separated in redshift by distances far greater than they appear in projection). To combat this issue, Voronoi techniques employ some additional condition to divide the data in some dimension analogous to redshift prior to the location of Voronoi clusters.

Kim et al. (2002) used a colour-magnitude filter to divide the galaxy catalog into regions of $(g - r)$ colour versus r-band magnitude, which is precisely the division of the galaxy samples into redshifts using the aforementioned red sequence (§1.5.2). However, this may not have avoided line-of-sight issues, as galaxy colours and magnitudes may have scattered into the red sequence at non-cluster sites.

Soares-Santos et al. (2010) used the photometric redshift as the dimension in which the input galaxy catalog is first divided. The Voronoi algorithm was run in redshift shells, and the neighbouring redshift slices are then recombined. Where Voronoi groups shared a minimal common area (in R.A./dec space), they are identified as the same cluster. This approach employed an earlier version of a Dark Energy Survey (DES, chapter 4.1) galaxy catalog simulation, or CatSim (§4.3), to determine its efficacy. Applied to real data, this robust approach delegates line-of-sight projection issues to the photometric redshift algorithm, which may introduce projection issues of its own.

More recently, a cluster finding algorithm called ORCA (Overdense Red-sequence Cluster Algorithm, Murphy et al., 2011) has been used to find clusters in optical data. ORCA used the red sequence to target galaxies in *griz* colour space at all redshifts, using a linear colour-magnitude fit to 126 members of Abell cluster 2631 (Abell et al., 1989) as

a model. ORCA further subdivided the galaxies into redshift bins by using spectroscopy or redshifts fitted with photometry to doubly avoid projection problems. Its application to SDSS Stripe 82 has been published by Geach et al. (2011), finding 4,098 clusters.

1.5.5 Cut and Enhance

The *Cut and Enhance* method (CE, Goto et al., 2002) focussed on detecting clusters of galaxies in multi-colour photometric surveys, such as the Early Data Release (EDR, Stoughton et al., 2002) of the Sloan Digital Sky survey (SDSS). CE attempted to minimize bias by minimizing the number of assumptions on cluster properties. The Cut and Enhance method works by employing colour cuts and a density enhancement algorithm to upweight pairs of galaxies that are close in both angular separation and colour within a $\sim 1.5\,h^{-1}$ Mpc radius. The colour cuts employed were numerous (30 single colour cuts and 4 colour-colour cuts). The justification for this approach was that if a method assumes a luminosity function or radial profile then the resulting sample will be biased towards that detection model. CE took advantage of the tight colour-magnitude relation of galaxies in clusters, e.g., the E/S0 ridgeline. The CE catalog contains cluster samples out to z$\sim$0.4.

Goto et al. (2002) investigated clusters unique to the detection algorithm (compared to maxBCG and Voronoi methods) and record an example of a blue spiral cluster and a cluster of faint ellipticals without a BCG. The Cut and Enhance (CE) cluster catalog was compared to Matched Filter (MF), Voronoi Tesselation (Kim et al., 2002, VTT) and maxBCG catalogues using a simple $6'$ cone search. They find that CE and maxBCG return more clusters, but this is primarily down to differences in thresholds, where maxBCG and CE contain lower richness systems. CE and MF compare well, but MF misses several low redshift systems which are visually found to be compact. The majority of clusters in the cross sample have spherical morphologies. maxBCG and CE match well, with their differences being highlighted by the fact that CE can probe bluer, star-forming galaxies where they may dominate a cluster, whilst maxBCG finds fainter higher redshift objects whose members fall below the CE magnitude cut. CE contains a high fraction of VTT clusters, but both fail at higher redshifts due to cluster galaxies falling below their respective magnitude cuts (i.e. are fainter than the algorithms allow them to detect).

1.5.6 C4

The premise of the C4 algorithm (Miller et al., 2005) is that optical clusters and groups of galaxies are dominated at their core by a single, co-evolving population of galaxies that

possess similar spectral energy distributions. It operated in optical wavebands, identifying clusters as overdensities in seven dimensional space, minimizing projection effects of prior optical cluster finding algorithms.

In C4, each galaxy's clustering measurement was evaluated within a $1\,h^{-1}$ Mpc R.A./Dec aperture at the galaxy's redshift. Within a single aperture and redshift bin, galaxy colour-clustering, or C4-clustering, was evaluated within 4-dimensional colour space. Once C4-clustering was measured for all galaxies, each galaxy was tested against the null hypothesis that the galaxy can be drawn from the field, where the null hypothesis is built by measuring the 4-dimensional colour space at random locations across the survey. C4 rejected the most field-like galaxies, leaving a sample of highly C4-clustered galaxies, which were then assembled into clusters.

C4 obtained a new sample of 748 clusters of galaxies identified in the spectroscopic sample of the Second Data Release (DR2, Abazajian et al., 2004) of the SDSS. The C4 cluster finding algorithm, its application to SDSS DR2, and a summary of some followup work with the C4 cluster catalog are described in detail in chapter 2.

1.6 Evaluating Cluster Finders

Verification of cluster catalog quality has a fairly straightforward history. Early cluster catalogs, e.g., Abell (1958), included a superset of cluster identifications, upon which various selection criteria were laid down and examined for known nearby clusters such as Coma, Leo, or Virgo. This is an early example of estimating the completeness of a cluster finder, where completeness is taken to mean the fraction of clusters (that exist) that have been captured by a given cluster finder/selection process.

Huchra and Geller (1982), with an alternate selection criteria including velocity dispersion in conjunction with celestial proximity to identify clusters, characterised groups that appeared irregular (based on their choice of parameters in their selection criteria). This is an early example of estimating the purity of a cluster finder, where purity is taken to mean the fraction of clusters (captured by a cluster finder) that can be considered real physical associations of galaxies. Note that Huchra and Geller clarify that some of these irregular associations may be down to a fraction of the galaxies in a given group being incorrectly included as part of that group.

In recent years there has been a trend to attempt to redefine these earlier definitions of completeness and purity in terms favourable to any given cluster finder. Koester et al. (2007); Murphy et al. (2011) refer to purity as the fraction of galaxies included in their

clusters that are interlopers (using simulations to characterise this fraction), and use catalog completeness as the quality measure for their cluster finders. Szabo et al. (2010) mention a purity measurement but never give one, describing instead measures of catalog completeness. Koester et al. (2007); Hao et al. (2010) use the clusters they have found to produce synthetic clusters, and place them randomly across the sky and redetect them with their respective algorithms to estimate completeness. This approach may be prone to overestimating completeness, as the completeness estimate is being trained on clusters the algorithm has already found. Rykoff et al. (2013) take a similar tack to Koester et al. and Hao et al., but use a richness estimator (Rykoff et al., 2012) to generate synthetic clusters at different redshifts, which probes completeness further but may still be limited by the characteristics of the clusters that are initially identified.

Completeness and purity form the key qualitative measures for establishing cluster finder quality. With the advent of computer simulations, the matched filter (Postman et al., 1996, §1.5.1) technique was trained with synthetic data to establish the accuracy of this cluster finding technique, with respect to cluster detections made at various depths, choosing parameters that provide a compromise between minimizing spurious detections and maximizing completeness at depths close to the survey magnitude limit. This, high purity and high completeness regime is desirable for all cluster finders.

In chapter 5, I expand upon these definitions of completeness and purity, and develop and present a framework for evaluating cluster finding that optimises both using the F-measure, or F1 score (§5.1.3). I note that in this thesis, false associations of galaxies to a cluster are not classed as cluster impurity, but classed as poor membership assignment; i.e., purity refers to the quality of the cluster catalog, rather than the quality of cluster membership in that catalog. Similarly, completeness is estimated through use of simulations, where the number of underlying halos is known and thus the recovered fraction of these halos represents the cluster catalog completeness.

1.7 Outline of Thesis

In this introductory chapter, I have reviewed a sample of galaxy cluster finders, the practicalities involved in their use, and concepts used in the measurement of success of a cluster finder. New optical and near-infrared galaxy surveys are in the process of being switched on, increasing the availability of galaxy data. The galaxy catalogs produced from these surveys will form more representative samples of further reaches of the universe such that statistical measurement of these populations can constrain how the universe evolved to

its current state. Clusters of galaxies found in these surveys can be verified by a number of complementary probes (§1.4), and will provide greater constraints on cosmological parameters.

This thesis introduces the AperC4 cluster finder in chapter 3, which is developed from the C4 cluster finder (§1.5.6). I review C4 in detail in chapter 2 to lend context to AperC4's adaptation to photometric data. In chapter 4, I introduce an upcoming galaxy survey called the Dark Energy Survey (DES), and the synthetic data they produced to test astronomical tools prior to the publication of real data. In chapter 5, I look at evaluating cluster catalogs and develop a tool for this purpose. In chapter 6, the AperC4 algorithm is applied to a simulation of the SDSS galaxy catalog to help tune the algorithm and assess its effectiveness. In chapter 7, I assess the characteristics of the real SDSS DR8 galaxy catalog and use it to find clusters with AperC4, presenting a preliminary cluster catalog. In chapter 8, I review the thesis and suggest directions for the development of further work.

Chapter 2

The Original C4 Algorithm

In this chapter, I will be reviewing the C4 algorithm presented by Miller et al. (2005, henceforth referred to as M05) and introducing the concepts therein that will become relevant to the work I present in further chapters. In section 2.1, I will outline the Sloan Digital Sky Survey (SDSS) and go on to describe the second data release (DR2) of the SDSS spectroscopic galaxy catalog. In section 2.2, I will describe the application of the C4 algorithm by M05 (from hereon, C4 products associated with Miller et al. will be denoted $C4_{M05}$, whilst general C4 concepts will remain unsubscripted), performed on the SDSS DR2 spectroscopic galaxy sample. I will also introduce the key C4 concepts of the false discovery rate and the k-NN distance. In section 2.3, I discuss the findings from M05 and some of the science that followed. In section 2.4, I describe some of the shortcomings of M05.

I note that this chapter is primarily a review of the M05 paper, however I have included my own figures to explain certain concepts.

2.1 SDSS DR2

2.1.1 The Sloan Digital Sky Survey (SDSS)

The Sloan Digital Sky Survey (SDSS York et al., 2000) is a large area photometric and spectroscopic survey covering the Northern Galactic Cap, employing a dedicated 2.5m f/5 modified Richey-Chrétien telescope with a 3° field-of-view situated at Apache Point Observatory, New Mexico (the SDSS telescope). The SDSS collaboration publicly releases raw and reduced data at regular intervals in the form of Data Releases (DR) and has completed three main campaigns, SDSS-I, SDSS-II, and SDSS-III, which consist of an Early Data Release (EDR) and Data Releases 1-10 (DR1 to DR10). SDSS is currently under-

taking its fourth campaign; SDSS-IV, which continues the survey from DR11 onward. Up until DR8, each Data Release contained the multi-colour imaging data, and photometric and spectroscopic catalogs of the increasing area of sky covered by SDSS (DR8 presented the last SDSS photometric data release). Prior to DR9*, the photometric portion of the survey was performed on moonless (dark), photometric nights with good seeing, whilst spectroscopy was performed when these conditions could not be fulfilled.

The survey is taken under a *survey coordinates* system which is a direct rotation of equatorial coordinates such that η, the survey latitude, and λ, the survey longitude, correspond to R.A./dec as in Table 2.1. Contrary to conventional latitude and longitude, η is constant along great circles, and runs between $-90 < \eta < 90$; whilst λ, which runs from $-180 < \lambda < 180$ around the sphere.

R.A.	dec	η (eta)	λ (lambda)
0.0	90.0	57.0	5.0
275.0	0.0	0.0	90.0
185.0	32.5	0.0	0.0

Table 2.1: Survey coordinates η and λ (survey latitude and longitude, respectively) are a pure rotation of R.A. and dec coordinates.

The SDSS Photometric Survey

The imaging and photometric data is gathered by a 120-megapixel CCD camera that images a wide field ($1.5\,\square^\circ$; $\square^\circ$ will be used to signify a square degree from hereon) by scanning the sky at the sidereal rate (drift scanning). The CCDs are arranged in rows, leading with a row of astrometric CCDs followed by 6 columns of CCDs in each of the five wavebands in the photometric array (one row per waveband) and a trailing row of astrometric CCDs. The 6 columns of photometric CCDs are themselves separated, so any area of sky covered by SDSS is scanned twice, offset by the width between CCD columns, to cover the full area. The SDSS survey area is subdivided into stripes, which are the sum of two strips of the SDSS camera. Each strip consists of 6 separate scans, covered by all the wavebands in the photometric array, which are individually referred to as camera columns or *camcols*.

The SDSS photometry (Fukugita et al., 1996; Gunn et al., 1998) covers optical wavelengths

*From DR9 onward, SDSS became devoted to spectroscopic campaigns.

through five filters: u-band, centred on an effective wavelength of 3549Å and a full width half-maximum (FWHM) of 560Å; g-band, a wider blue-green band centred on 4774Å with a FWHM of 1377Å; r-band, a red band centred at 6231Å, with a FWHM 1371Å (comparable to the g-band FWHM); i-band, a far-red band centred on 7615Å with a FWHM of 1510Å; and z-band, a near-infrared band centered on 9132Å with a FWHM of 940Å. These filters are coated on the red (long wavelength) and blue (short wavelength) ends of their spectral range, resulting in a fairly sharp cut in observing efficiency at the FWHM. In the case of the z-band, which has no red-end filter, the response is defined by the CCD: Near-IR sensitivity cuts out in the z-band due to water vapour and other atmospheric absorption effects at 9300-9700Å. The full filter transmission curves against wavelength are shown in Figure 2.1 (taken from Bartelmann and White, 2002).

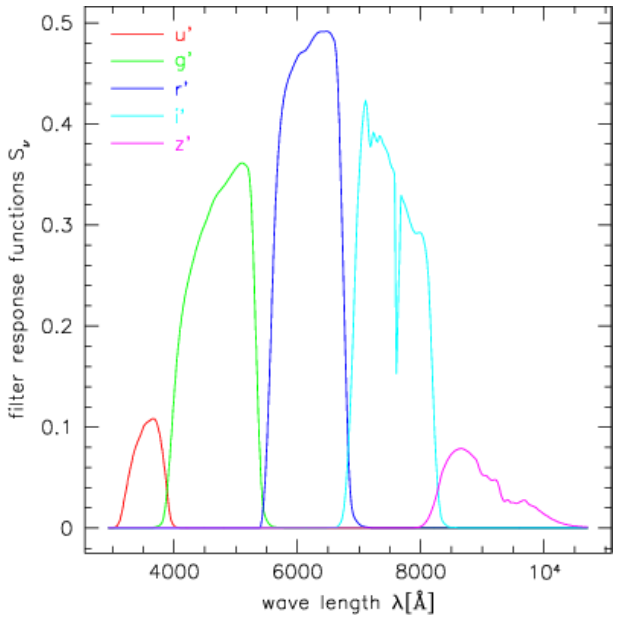


Figure 2.1: The SDSS photometric filter response curves. The horizontal axis shows wavelength and the vertical axis describes the filter transmission efficiency. (Figure credit: Bartelmann and White, 2002).

The image analysis software extracts sources, point-like and extended, from the imaging data and produces a catalog with various observable and metadata measurements (Lupton et al., 1999; Stoughton et al., 2002). The 95% detection limits for repeatability of source detection (of stars) in each band are given as $u < 22.0$, $g < 22.2$, $r < 22.2$, $i < 21.3$ and $z < 20.5$ for an air mass of 1.4, which is the median of the survey. Each pixel corresponds to a scale of $0''.4$, which allows oversampling of the median $0''.8$ seeing at Apache Point, delivering images with a FWHM of around 1".

Problems that have been reported from the SDSS data include a natural red-leak at 7100Å in the u-band due to the vacuum in the camera causing the interference film on the filter to shift redward (Smith et al., 2002). The leak varies from column to column, displays

time-variability and is unfocussed due to the redder wavelengths refracting differently to wavelengths of the bluer range of u-band, making correction difficult. Abazajian et al. (2004) also reports a slight downward bias in the determination of the sky level in the photometry, which, combined with the u-band leak, lead to of order 10% errors in the u-band Petrosian fluxes of large galaxies.

Measuring flux for galaxy photometry is notably harder than for stars, due to differing radial surface brightness profiles and no defined edges, so SDSS employs several photometric systems:

SDSS asinh magnitudes (luptitudes): The main SDSS survey employs a photometric scale, where magnitudes are output as inverse hyperbolic sine (or "asinh") magnitudes, also known as *luptitudes*, as detailed by Lupton et al. (1999). They are designed to behave as standard log-scale magnitudes at high signal-to-noise, but gracefully reduce to zero at low signal-to-noise, where the recorded flux is close to the detection limit of the survey.

Petrosian magnitudes (`petroMag`): To minimize biases in galaxy flux measurement, SDSS adopts a modified form of the Petrosian system (Petrosian, 1976). Galaxy flux is integrated within a circular aperture whose radius is defined by the galaxy's azimuthally averaged surface brightness profile. The aperture radius is defined in the r-band to ensure consistent flux measurement across the SDSS bands.

PSF magnitudes (`psfMag`): The PSF (Point Spread Function) magnitude measures the total flux of an object by fitting a PSF model to it, which should perfectly model the magnitude of point-like sources of stars or distant (unresolved) quasars but works less effectively for extended sources like galaxies. The PSF magnitude becomes useful in star-galaxy separation as one expects significant deviations between a PSF magnitude and a magnitude that more accurately measures galaxy flux.

Model magnitudes (`modelMag`): Model magnitudes for objects are output by the SDSS pipeline, which attempts to model the PSF magnitude for point sources and the Petrosian magnitude for extended sources. These model magnitudes are 2-D profiles fitted to the object images in each band; one an exponential profile and the other a deVaucouleurs profile which both deliver magnitude estimates (and associated variables) of the object. A further `cmodelMag` magnitude delivers a linear combination of the best fit exponential and deVaucouleurs fits (in each waveband) that best fits the image.

The SDSS Spectroscopic Survey

The SDSS spectroscopy is used to validate photometric measurements and highlights the differences found between photometric, spectroscopic, and synthetic, magnitude and colour data. SDSS spectroscopy is performed on objects detected in the imaging portion of the survey that are classified, as point source or extended, and measured by the SDSS image analysis software. Different classes of objects (e.g. stars, galaxies, QSOs) are selected from the imaging for spectroscopic followup by two spectrographs, each of which has a blue and red channel, with wavelength ranges of 3800-6150Å and 5800-9200Å respectively, separated by a dichroic filter. Each spectrograph is equipped with 320 fibers, totaling a maximum of 640 spectra per telescope pointing, and each fiber has a $3''$ diameter*. Spectra are obtained through plates (tiles) that cover the entire 3° field of view, where holes are drilled for each fiber. Because of the thickness of the cladding that house the optical fibers, fiber holes on any given plate are separated by at least 55", and target galaxies that are separated by less than this distance are said to "collide". However, these fiber collisions are minimized by tiling plates (overlapping adjacent plates) such that the sampling of targets is greater than 92% for all targets and greater than 99% for targets that do not collide (Blanton et al., 2003a). The reduced spectra are given in vacuum wavelengths in the heliocentric frame, and several techniques are used to measure redshift, characterize lines and better describe sources (initially identified and characterized in the photometric SDSS). In the following section, I describe the spectroscopic catalogs produced by the SDSS.

2.1.2 Selection of Spectroscopic DR2 Galaxy Catalog

Abazajian et al. (2004) details the DR2 portion of the spectroscopic survey as covering $2627\,\square^\circ$ (Figure 2.2) with 367,360 spectra, of which, 260,490 are classified as galaxies. SDSS spectroscopic galaxy observations are limited to $r \leq 19.5$ by the fixed exposure time. The galaxy target selection consists of two components. The first is described in Strauss et al. (2002) and consists of galaxies with r-band Petrosian magnitudes $r < 17.77$ and Petrosian half-light surface brightness $\mu_{50} \lesssim 24.5$ mag arcsec^{-2} (the `MAIN` galaxy sample). This first selection acts as a further star-galaxy separator (on top of the star-galaxy separation performed by the SDSS photometric pipeline), and is effective at removing nearly all stellar contamination whilst maintaining a high completeness for the spectroscopic galaxy

*note that these specifications are historical; as of DR8, the spectrographs are equipped with 1,000 fibers, each with a diameter of $2''$.

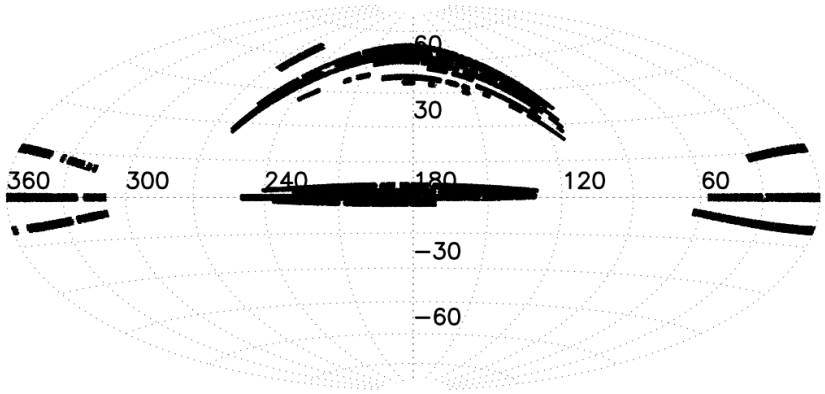


Figure 2.2: Aitoff projection of the sky, where the 2627 $\square^\circ$ coverage of the SDSS DR2 spectroscopic campaign is described by the shaded region. The horizontal axis shows R.A. and the vertical axis shows declination [Figure credit: SDSS DR2 website*].

sample. Strauss et al. (2002) find the main cause of incompleteness to be blending with saturated stars.

The second sample is a volume limited luminous red galaxy (LRG) sample (Eisenstein et al., 2001), that selects red galaxies in the SDSS DR2 region. Modifications to the Eisenstein et al. (2001) selection cuts, originally used on EDR, are made by Abazajian et al. (2004) on DR2. These changes were motivated primarily to select a uniform sample of LRGs over the redshift range $0.2 \lesssim z \lesssim 0.7$ and to compensate for errors/problems discovered in the photometry. The LRG sample is further subdivided into two selections, Cut 1 and Cut 2, since the 4000Å break used to trace LRGs transitions from the g-band to the r-band at a redshifts greater than $z \gtrsim 0.38$. Some further, high-redshift quasar specific cuts (Richards et al., 2002) are made to the sample, chiefly in the i-band. Details on Cut 1 and Cut 2 are as follows.

Cut I

I note here that Cut 1 isn't sufficient for LRG selection below redshift of $z < 0.15$ as it is too permissive, allowing lower luminosity sources to enter the main LRG sample at these low redshifts. However, this caveat simply concerns selection of LRGs from the `MAIN` sample, and does not supplant their use in target selection. To trace LRGs, $c_{||}$, defined in Equation 2.1, is used as an effective redshift indicator:

$$c_{||} = 0.7(g - r) + 1.2[(r - i) - 0.177] \ . \tag{2.1}$$

Then to select galaxies as a function of redshift, $c_{||}$ is evaluated against $r_{\rm petro}$, the

*http://www.sdss.org/DR2/coverage/

r-band Petrosian magnitude:

$$r_{\rm petro} < 13.116 + c_{||}/0.3, \tag{2.2}$$

which sets the luminosity threshold as a function of redshift. To obtain a reasonable signal-to-noise, a further apparent magnitude constraint is applied:

$$r_{\rm petro} < 19.2. \tag{2.3}$$

The cut given by $c_{\perp}$, in Equation 2.4, restricts the selection of galaxies along the locus in the $g - r$ versus $r - i$ colour-colour diagram (Equation 2.1, above):

$$|c_{\perp}| < 0.2 \tag{2.4}$$

Targets with very low surface brightness are cut by

$$\mu_{r,\rm petro} < 24.2\,{\rm mag\,arcsec}^{-2}, \tag{2.5}$$

where

$$\mu_{r,\rm petro} = r_{\rm petro} + 2.5\log_{10}(2\pi R_{50}^2), \tag{2.6}$$

as low surface brightness objects are often the product of errors in the SDSS reduction pipeline (Strauss et al., 2002).

To separate stars from galaxies, the PSF and model magnitudes are separated by

$$r_{\rm PSF} - r_{\rm model} > 0.24, \tag{2.7}$$

since for point-like sources (i.e., stars), the model magnitude approaches a delta function, and thus becomes equivalent to the PSF magnitude.

The only other imposed cuts set the general colour space containing the LRG colour locus:

$$g - r < 2.5, \tag{2.8}$$

$$r - i < 1.5. \tag{2.9}$$

Cut 2

To select galaxies above $z \gtrsim 0.38$, another set of colour and magnitude cuts are made, collectively called 'Cut 2'. So that a reasonable signal-to-noise is obtained, the Petrosian r-band magnitude is limited to

$$r_{\rm petro} < 19.5, \tag{2.10}$$

which corresponds to a slightly fainter limit than in Cut 1 (Equation 2.3).

To restrict the selection of galaxies along the locus in the $g-r$ versus $r-i$ colour-colour diagram, nominally to separate regions of low and high redshift, Equation 2.11, and isolate LRGs from the bulk of the late-type stellar locus, Equation 2.12, the following colour cuts are applied:

$$|c_\perp| < 0.449 - (g-r)/6\,, \tag{2.11}$$

$$g - r > 1.296 + 0.25(r-i)\,. \tag{2.12}$$

Targets with very low surface brightness are cut by

$$\mu_{r,\mathrm{petro}} < 24.2\,\mathrm{mag\,arcsec}^{-2}, \tag{2.13}$$

as with Cut 1 in Equation 2.5.

For star-galaxy separation, a tougher constraint is required than Cut 1 (Equation 2.7), so the tolerance between the PSF and model magnitudes is increased to

$$r_{\mathrm{PSF}} - r_{\mathrm{model}} > 0.4. \tag{2.14}$$

And again, the general colour space containing the LRG locus is set by:

$$g - r < 2.5, \tag{2.15}$$

$$r - i < 1.5. \tag{2.16}$$

I note here that deeper LRG observations that are selected by Cut 2 data may affect clustering measurements, and hence, cluster finding at higher redshifts where the galaxy population is biased towards these populations. But due to the ease of identification of LRGs in colour space, creating a catalog of LRGs for spectroscopic followup based on colour selection is easier than doing so for the general galaxy population.

2.2 The Original C4 Cluster Finding Algorithm

In this section, I describe in detail the original $\mathrm{C4_{M05}}$ algorithm from M05, introduced in section 1.5.6. Section 2.2.2 reviews the $\mathrm{C4_{M05}}$ method. In section 2.2.3, I describe the subsequent statistical extraction of cluster-like galaxies with the False Discovery Rate. Section 2.2.4 introduces k-NN clustering and its usage in the $\mathrm{C4_{M05}}$ algorithm to assign $\mathrm{C4_{M05}}$ galaxies to clusters (§2.2.5).

2.2.1 Introductory Notes

The main product of M05 is a highly complete and highly pure cluster sample at low redshifts using the $C4_{M05}$ algorithm. It was one of the first catalogs of clusters and groups constructed directly from spectroscopic data of the Sloan Digital Sky Survey (SDSS; York et al., 2000), specifically, spectroscopic galaxies from the Second SDSS Data Release (DR2; Abazajian et al., 2004). $C4_{M05}$ is limited to low redshift systems ($z < 0.17$) due to its dependence on spectroscopic redshift to locate cluster members. The underlying premise of $C4_{M05}$ is that optical clusters and groups of galaxies are dominated by a single, co-evolving population of galaxies, e.g. the "E/S0 ridgeline" or "red sequence", that have similar spectral energy distributions. A variety of other modern optical cluster finders (§1.5), also use this premise to identify clusters.

$C4_{M05}$ can be distinguished from other optical cluster finders by its utilisation of all galaxy colour information without explicitly modelling assumptions on the occupation of galaxies along a red sequence (or E/S0 ridgeline; as per Gladders and Yee, 2000; Koester et al., 2007; Hao et al., 2010; Gilbank et al., 2011; Murphy et al., 2011; Rykoff et al., 2013). Since the red sequence is already known to be a compact formation in redshift-colour space, clusters where the red sequence dominates are captured by the $C4_{M05}$ algorithm. The one (or more) co-evolving group(s) of galaxies which cluster in colour space do not necessarily have to be red sequence galaxies. Cluster finders that do not rely on red-sequence modelling tend search for enhancements in 3-dimensions (nominally R.A., dec, and z; as per Soares-Santos et al., 2008; Wen et al., 2009; Farrens et al., 2011), where their reliance on spectroscopic/photometric redshifts can be sensitive to spectroscopic failures (Cunha et al., 2012), photometric bias (Banerji et al., 2008) or catastrophic photometric errors (Padmanabhan et al., 2005).

$C4_{M05}$ captures cluster red sequences, given the tight colour dispersion of the red sequence against magnitude, without the need for modelling (as per Hao et al., 2010) or tuning to a cluster observation (as per Murphy et al., 2011). In common with matched filter and adaptive matched filter methods (§1.5.1), $C4_{M05}$ identifies galaxies that are more likely to belong to cluster distributions. But in place of identifying cluster galaxies assuming a modelled distribution of cluster/field galaxies in position, redshift, and luminosity function space (as per Szabo et al., 2010, an other matched filter methods), M05 use the data itself to calibrate the C4-clustering signal for a given galaxy in $C4_{M05}$.

$C4_{M05}$ incorporates 4-dimensional colour space overdensities with 3-dimensional position overdensities (including redshift). By employing the 4-dimensions of colour space

($u-g$, $g-r$, $r-i$ and $i-z$) from the SDSS photometry combined with the redshift measurements from the SDSS spectroscopic campaign, $C4_{M05}$ mitigates the problem of misidentifying groups of galaxies that appear close in projection that are not physically associated in real clusters (§1.5.1). By minimizing projection effects, $C4_{M05}$ produced a highly pure catalog of clusters, allowing further research into cluster astrophysics and studies into the galaxy populations that occupy them. $C4_{M05}$ was amongst the first cluster finders to use mock galaxy catalogs (Wechsler, 2004), representative of the real universe, to evaluate its selection function. The implementation of mock catalogs in the evaluation of cluster finders has since become common practice (Li and Yee, 2008; Hao et al., 2010; Soares-Santos et al., 2010; Murphy et al., 2011). The new C4 algorithm presented in this thesis will also be evaluated with simulated data in chapter 6, using a mock SDSS catalog simulation (§4.3).

The $C4_{M05}$ cluster finding algorithm obtained a new sample of 748 clusters of galaxies (recovering 92% of Abell clusters from a qualitative selection, see §2.3.4) identified using the spectroscopic SDSS DR2 galaxy sample. The catalog covers $\sim$2600 deg^2 of sky and ranges in redshift from $0.02 \leq z \leq 0.17$. The catalog provides the following information for each cluster:

- sky location;
- mean redshift;
- galaxy membership;
- summed r-band optical luminosity (L_r);
- and velocity dispersion.

Correlations between these parameters and the dark matter halo properties (§2.3) were explored by M05, using mock SDSS catalogs constructed from the ΛCDM Hubble Volume Sky survey output.

2.2.2 C4-clustering

Processing the DR2 spectroscopic galaxy catalog is a computationally heavy task and so this sample is divided into a multidimensional grid and then evaluated for *C4-clustering* where each galaxy's volume-colour density probability is evaluated to deliver a *p-value*.

Specific SDSS DR2 Galaxy Selections used in Miller et al., 2005

Further to the target selection of SDSS spectroscopic galaxies (§2.1.2), additional constraints are applied to the galaxy sample before $C4_{M05}$ is run. Objects with SDSS warning flags (output by the SDSS spectroscopic pipeline) set for: low-confidence redshift; no currently available spectrum; and no red or blue end are omitted. The `zConf` variable (an SDSS pipeline output that estimates a 'confidence' value, from 0 to 1, that the redshift is reliable) is also constrained such that `zConf`> 0.7. This results in 249725 unique galaxies that are processed by the $C4_{M05}$ algorithm.

Clustering in 3D Space

For efficient processing, a three dimensional grid is created to contain the input galaxy sample in equatorial (R.A., dec) and redshift bins. The extent of the grid in radial and redshift dimensions is given to the algorithm as an input, and the grid is built to accommodate the number of galaxies that are contained by the most populated grid point. The equatorial grid containing DR2 is divided into segments $\sim 1\,\square^\circ$ in size (units of $1^\circ \times 1^\circ$ in R.A. and dec, respectively), and the redshift dimension is divided into comoving coordinate bins of $50\,h^{-1}\,\mathrm{Mpc}$. Each of these redshift bins represents a redshift element within which the colour clustering is measured. An equatorial $1.0\,h^{-1}\,\mathrm{Mpc}$ kernel aperture is defined for $C4_{M05}$ around a galaxy providing R.A./dec constraints within each $1\,\square^\circ$ box. This aperture is calculated with

$$\begin{matrix} C4_{M05}\text{ kernel radius} \\ \text{in arcminutes} \end{matrix} = 1.0\,h^{-1}\,\mathrm{Mpc} \times \frac{H_0}{c} \times \frac{(1+z)^{1.5}}{(1+z)^{0.5}-1} \times \frac{180}{\pi} \times 60, \qquad (2.17)$$

where the $1.0\,h^{-1}\,\mathrm{Mpc}$ term is defined as the $C4_{M05}$ kernel radius, H_0 is $100\,h\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ and the final two terms convert the result from radians to arcminutes.

Should this aperture extend past the edge of the equatorial grid element, the algorithm simply selects galaxies in the next element within the radial aperture. The galaxies within the radial aperture in the target galaxy's redshift bin then represent the volume-limited sample around each galaxy in the shape of a narrow cylinder.

I note here that the $C4_{M05}$ kernel radius in arcminutes is equivalent to twice the transverse comoving separation in an $\Omega_m = 1$ cosmology. The M05 cluster sample is limited to redshifts between $0.02 < z < 0.17$, where differences in cosmology affect the ratio of kernel radius to comoving separation little, e.g., an assumption of Planck cosmology ($H_0 = 67.4$, $\Omega_M = 0.314$. $\Omega_\Lambda = 0.686$, Planck Collaboration et al. (2013b)) produces a

discrepancy of 8% at the upper redshift limit of $z = 0.17$ from this 2:1 ratio. If examining this ratio with Planck cosmology at higher redshifts, the ratio increases such that at $z \gtrsim 2$, the kernel radius is approximately three times larger than the transverse comoving distance. This is a notable concern for cluster finding at advanced redshifts with $C4_{M05}$ that I address in Chapter 3.

Clustering in Colour Space

A 4-dimensional colour box is defined for each galaxy from the systematic and statistical error components of that galaxy as seen in Equation 2.18. The statistical error, $\sigma_{xy}(\text{stat})$, refers to the observed error for the two magnitudes (x, y) summed in quadrature. The systematic error, $\sigma_{xy}(\text{sys})$, refers to the intrinsic scatter in the colour-magnitude relation (CMR) of cluster galaxies (Visvanathan and Sandage, 1977; Bower et al., 1992). A multiplier, γ, is associated with $\sigma_{xy}(\text{sys})$, which allows tuning of the size of the colour box. It is important to note that the colour box is so tuned in order to capture galaxies but is not an attempt to model the CMR.

$$\sigma_{xy} = \sqrt{\gamma\sigma_{xy}^2(\text{sys}) + \sigma_{xy}^2(\text{stat})} \tag{2.18}$$

where

$$\sigma_{xy}^2(\text{stat}) = \sqrt{\sigma_x^2 + \sigma_y^2} \tag{2.19}$$

Motivated by Goto et al. (2002, §1.5.5), the systematic components for the $u-g$, $g-r$, $r-i$, and $i-z$ SDSS colours were set as $\sigma_{ug}(\text{sys}) = 0.15$, $\sigma_{gr}(\text{sys}) = 0.12$, $\sigma_{ri}(\text{sys}) = 0.1$, and $\sigma_{iz}(\text{sys}) = 0.1$, respectively. The values are representative of the widths of the CMR, decreasing with increasing wavelength. Each colour error is evaluated on condition that the magnitudes, x and y, fall within the specified survey limits and on condition the errors on the magnitudes, σ_x and σ_y, are not greater than 10 magnitudes, such that $C4_{M05}$ excludes catastrophic failures. The limit on the magnitude uncertainties exists to exclude these failures, but in practice these do not occur in the DR2 spectroscopic sample. If a galaxy does not contain any magnitudes above the survey limits, or all the magnitude uncertainties are greater than 10 magnitudes, then it is excluded from $C4_{M05}$ processing.

Then, looping through each galaxy, in each spatial grid coordinate, a box in colour space (see Figure 2.3) is created in each colour out to σ_{xy}. For neighbours of the galaxy,

i.e., the galaxies contained within the spatial aperture, their error and calculated colour is used to define a gaussian probability in colour space, where the probability of the neighbour lying within the colour box is evaluated. Every galaxy within the volume aperture is counted as a candidate neighbour. Where neighbour candidates are not inside the colour box, but there still exists some finite probability within $6\sigma_{xy,neighbour}$ of the neighbour being in the colour aperture, that probability is also evaluated. Where there is incomplete colour information, it is not evaluated, thus if a neighbouring galaxy is too faint to be detected in bluer bands, but is spatially local and clusters in redder bands, then its C4-clustering is not biased through modelling of the bluer bands. The sum of the probabilities of these neighbours being in the target galaxy's colour aperture is then treated as the target galaxy's colour density within the volume aperture, or $N^{\text{target}}_{\text{neighbour}}$ in volume V, where

$$N^{\text{target}}_{\text{neighbour}} = \sum_{n \in V} \int_{-\sigma_{xy}}^{\sigma_{xy}} P(\sigma_{xy,\text{neighbour}})\ d\sigma_{xy}\ , \tag{2.20}$$

where $(n \in V)$ are all the neighbours (galaxies), n, in the target galaxy's volume aperture, V.

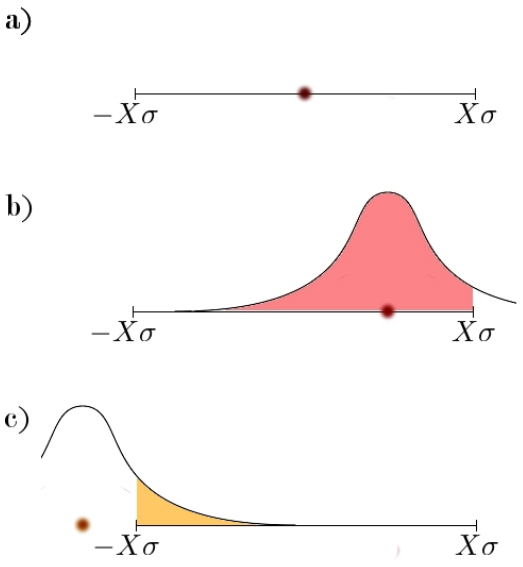


Figure 2.3: a) Using the uncertainty from Equation 2.18, a box around the galaxy's colour space is defined out to $X\sigma_{xy}$, where X is equal to unity in Miller et al. (2005). b) The probability of the neighbour lying in the galaxy's colour box (pink area) is given by integrating the neighbour's colour probability within the box limits. c) Where a neighbour is not inside the colour box, it may still contribute towards the galaxy's colour clustering (yellow area).

The same colour box and volume aperture is then moved randomly across the survey footprint, centring on other galaxies in the input galaxy catalog (but in the same redshift

bin). The colour density around these random points is evaluated 30 times to build up a distribution of colour densities for each target galaxy's colour-volume aperture, which acts as a model of the field. Figure 2.4 shows an example of a cluster galaxy and a field galaxy, plotted in colour-colour space.

This colour-volume aperture density model forms a Poisson distribution, which is then approximated to a Gaussian as per Anscombe (1948), against which the likelihood of the target galaxy occurring in this distribution is calculated out to $\pm 6\sigma$. Explicitly, this calculation is

$$p\text{-value} \;=\; 1 \;-\; 2 \times \left(\sqrt{N^{\text{target}}_{\text{neighbour}} + \frac{3}{8}} \;-\; \sqrt{N^{\text{aperture}}_{\text{neighbour}} + \frac{3}{8}} \right) \;, \qquad (2.21)$$

where the square root components represent the approximations to Gaussian for the Poisson neighbour counts of the target galaxy and aperture model median, respectively. The p-value in Equation 2.21 is the probability of the target galaxy occurring in the field given this aperture. Once p-values are evaluated for each galaxy in the catalog, the algorithm then eliminates galaxies that are most likely to exist in the field by employing a cut in probability enforced by a False Discovery Rate, or FDR. Section 2.2.3 introduces the FDR concept and its application to C4_{M05}.

2.2.3 False Discovery Rate (FDR)

The False Discovery Rate, as formulated by Benjamini and Hochberg (1995), is a statistical method that can be used to evaluate the complicated uncertainties under multiple hypothesis testing.

To explain, let me formulate a null hypothesis, H_0, to be tested against an alternative hypothesis, H_1, based on some test statistic, S. Under hypothesis testing, a rejection region is set, and H_0 is accepted where S falls outside that region, and H_0 is rejected where it occurs inside, and hence the alternative hypothesis, H_1, is accepted in this region.

When treating data in this manner, there are two sources of error. A false discovery, or "type I error", occurs for data where the null hypothesis is true but has been rejected, whilst a false non-rejection, or "type II error", occurs where the alternative hypothesis is true, but has been (falsely) characterized by the null hypothesis H_0. Figure 2.5 demonstrates these different kinds of error that may occur when classifying data in this manner. Often, type II errors are referenced to as "power", which is 1 minus the probability of a type II error, i.e., the probability of rejecting the null hypothesis given that it is false, and it is desirable to maximize power. However, increasing power also comes at the expense

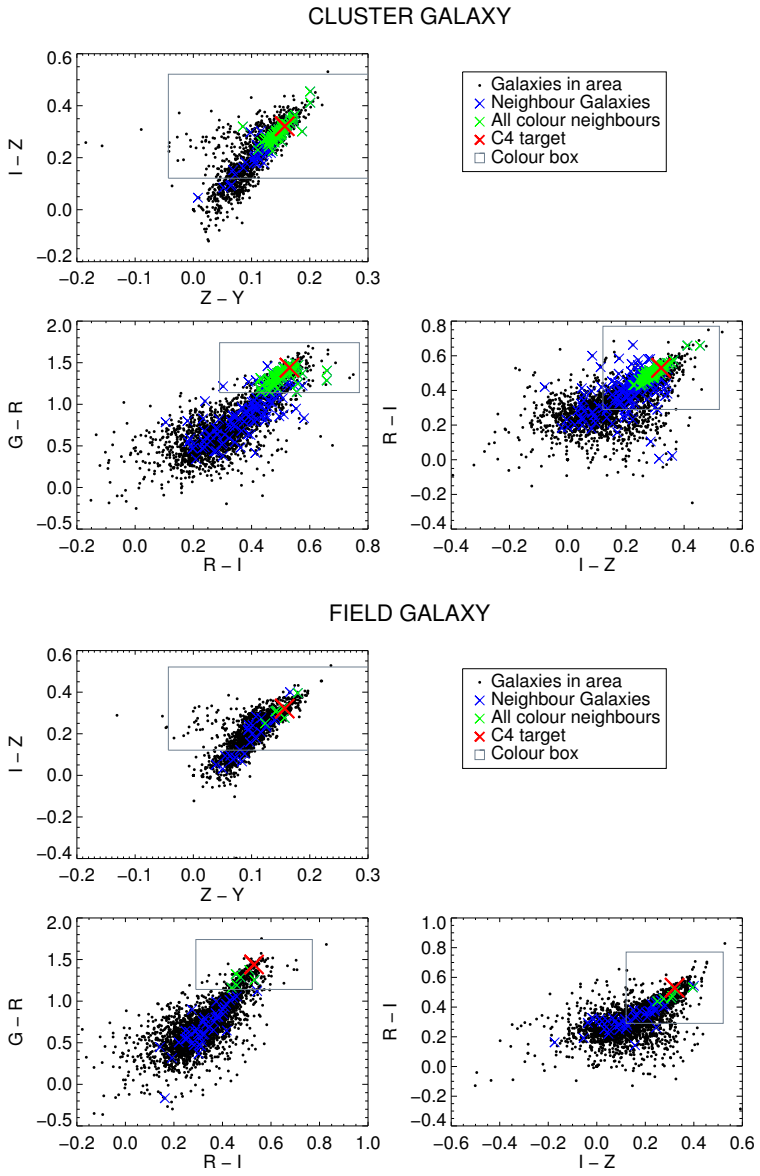


Figure 2.4: The upper three plots describe a cluster galaxy in 4 dimensions of colour-colour space, whilst the lower three describe a field galaxy. The target galaxy in each case is indicated by a red cross, and the surrounding grey box. The blue crosses represent all the galaxies within the volume aperture whilst the green crosses denote galaxies within the volume aperture that also fall inside all four colour apertures. The cluster galaxy 4-colour aperture shows a significantly higher number of volume-colour neighbours than those of the field galaxy. The galaxies and apertures in this figure are generated from $C4_{M05}$ applied to a mock galaxy catalog, which uses $g-r$, $r-i$, $i-z$, and $z-Y$ colours.

of increasing the fraction of type I errors, as seen in Figure 2.6.

The decision to reject a test statistic is driven by whether it is above some critical threshold. The probabilistic analogue to this is to use the p-value, which is the probability of observing a test statistic that is at least as extreme as the observation made, assuming that the null-hypothesis is true. For example, an observation with a p-value of 0.05 is equivalent to saying that the likelihood of this observation agreeing with the null hypothesis

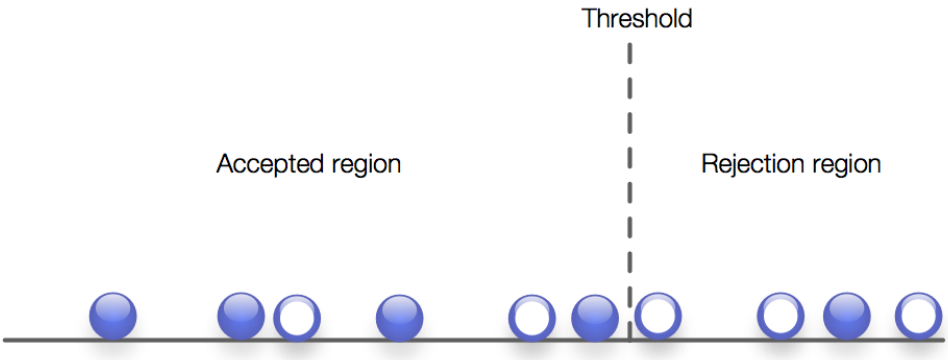


Figure 2.5: In this diagram, the circles correspond to values of some given test statistic. Filled and open circles represent data where the null hypothesis is true or false, respectively, for that test. The test rejects the null hypothesis where the test statistic places data to the right of the threshold. The filled circles to the right of the threshold are false discoveries or false rejections (type I errors). The open circles to the left of the threshold are false non-rejections (type II errors).

is 5%.

The relevance of multiple hypothesis testing to $C4_{M05}$ can be summed up in the sense that we can observe a colour-magnitude relation between galaxies in a cluster where none exists in the field. However, one cannot be certain that a colour-magnitude relation (CMR) exists solely due to the physical properties of galaxies within an individual cluster (having been identified by the p-values generated in section 2.2.2) or whether a more global effect dominates the CMR for clusters, as exemplified by red sequence cluster finding techniques (Gladders and Yee, 2000; Koester et al., 2007; Hao et al., 2010). Hence, the observed colour-magnitude relation of clusters is treated as a property of a galaxy within their associated $C4_{M05}$ colour-volume aperture. The null hypothesis tested, for each galaxy, is that no C4-clustering (and hence no colour-magnitude relation) is observed for that given galaxy at its redshift (§2.2.2).

With the False Discovery Rate (FDR) an acceptable proportion of false rejections, or error rate, is defined which determines a threshold p-value that defines such a rejection region. The FDR works algorithmically as follows:

1. An ascending array of length n with values i/n is created, where n is the number of galaxies and i is an integer from $1 \leq i \leq n$.

2. The p-values, p, for all n galaxies are sorted from low to high, $p_1, p_2, p_3, \cdots, p_n$, such that $p_i < p_{i+1}$.

3. The i/n array is multiplied by the FDR threshold value to give a new array, f.

4. The element t is determined by the minimum value of i where $p_i > f_i$.

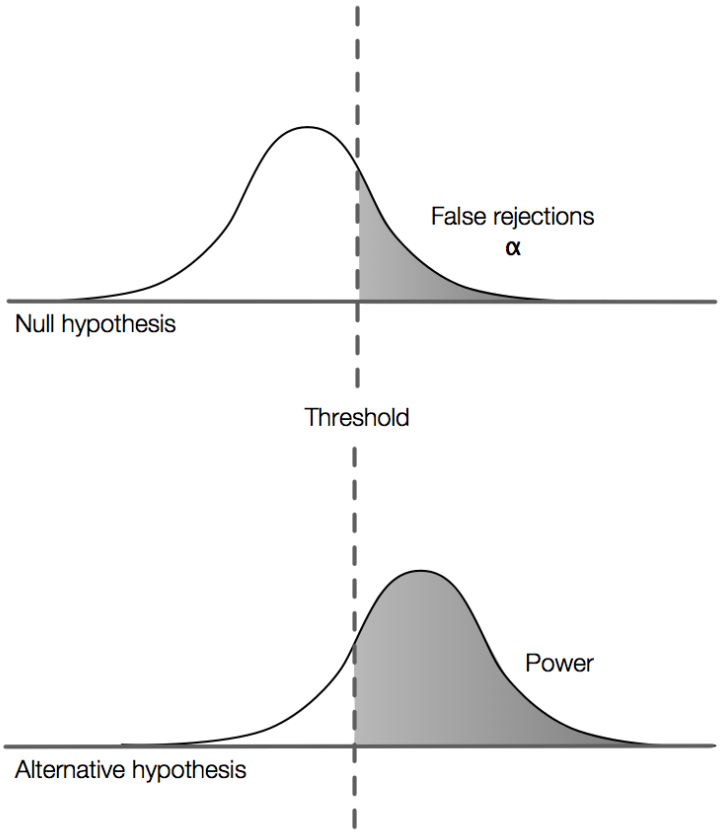


Figure 2.6: This schematic illustrates how changing the threshold affects both the probability of a type I error and the power. The first curve (above) describes the distribution of the null hypothesis, H_0, according to the test statistic. The second curve (below) describes the distribution of the test statistic for an alternative hypothesis, H_1. The dashed vertical line describes the threshold for accepting (left of threshold) or rejecting (right of threshold) the null hypothesis. The shaded areas to the right of the threshold, under both curves, describe the probability of rejecting the null hypothesis The shaded area under H_0 represents the probability of rejecting the null hypothesis when it is true, resulting in a type I error, α. Under H_1, the shaded region represents the probability of rejecting the null hypothesis when it is false, i.e., the power of H_1.

5. The p-value of the FDR threshold is given by this value p_t.

6. The galaxies described by the p-values $p_1, p_2, \cdots, p_{t-1}$ are kept as cluster galaxies, and galaxies with $p \geq p_t$ are rejected.

This algorithm delivers a sample of rejected null hypotheses with a known error rate, i.e., the FDR. Since the discriminator is the colour-clustering of a galaxy in a 3 dimensional volume, we are left with a sample of galaxies that are the most clustered (in a $C4_{M05}$ defined colour-volume aperture), and contain a known maximum proportion of galaxies that may belong to the field population (i.e., are drawn from the null hypothesis).

2.2.4 *k*-NN Distance

A k-NN algorithm, or k-nearest neighbours algorithm, classifies objects based on the kth nearest object within some n-dimensional space, where those n-dimensions relate to properties of that object. This is the simplest of all instance-based, machine learning algorithms, where an object is simply classified by the majority of its neighbours 1,2,...k. For example, if $k = 1$, then the object is assigned the class of its neighbour.

At this stage in the algorithm, $C4_{M05}$ (cluster) galaxies have been identified but not placed into clusters. To place the galaxies into clusters, $k-$NN distances are calculated for each cluster galaxy. The redshift and equatorial coordinates (converted to radians as per equations 2.22 and 2.23) are employed to place the cluster galaxies (as determined by FDR) into a 3-dimensional comoving coordinate system. When placing these galaxies into a (comoving) cartesian coordinate system, the distance metric (Equation 2.24) assumes $H_0 = 100\,h\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ and a flat geometry for the universe and interprets the redshift as being cosmological in origin (i.e., peculiar motions are unaccounted for). In spherical coordinates, the galaxy positions become:

$$\theta_{gal} = R.A. \times \frac{\pi}{180}, \tag{2.22}$$

$$\phi_{gal} = (90 - dec.) \times \frac{\pi}{180}, \text{ and} \tag{2.23}$$

$$R_{gal} = \frac{c}{H_0} \times \left(2.0 - \frac{2.0}{\sqrt{1.0+z}}\right) \tag{2.24}$$

where z in Equation 2.24 is the redshift of the galaxy. Equation 2.24 approximates the comoving distance in a flat cosmology of $\Omega_m = 1$ and $\Omega_\Lambda = 0$. These polar coordinates then recombine to give the cartesian coordinates (x_{gal}, y_{gal} and z_{gal}) for the galaxies in the survey as:

$$x_{gal} = R_{gal} \times \sin(\theta_{gal})\cos(\phi_{gal}), \tag{2.25}$$

$$y_{gal} = R_{gal} \times \sin(\theta_{gal})\sin(\phi_{gal}), \text{ and} \tag{2.26}$$

$$z_{gal} = R_{gal} \times \cos(\theta_{gal}). \tag{2.27}$$

These coordinates are then given to a k-NN code, that determines the metric distances to the 6th nearest neighbour* of each galaxy. Distances in this k-NN code are calculated through the use of *KD trees*, which partition points in a multidimensional space (for $C4_{M05}$ these are the three comoving dimensions described above).

*the 6th nearest neighbour is an arbitrary choice in M05

A KD tree that contains N points allows an algorithm to find a point's nearest neighbour to be found in a time proportional to $\mathcal{O}(\log N)$ or all points' nearest neighbours in $\mathcal{O}(N \log N)$ time. A KD tree works by placing all points within a box in the dimensions defined. Then this box is partitioned into two new boxes along the axis of one dimension, leaving an equal number of points in each partition. Each box is then partitioned in the remaining dimension(s) and the process repeats, cycling through the dimensions, until the final box (or node) contains either one or two points, (i.e., the nearest neighbours are together or in individual boxes divided in the last dimension) and the distance between them is calculated. Calculating the distances of each point to their respective kth nearest neighbours is further optimised by taking advantage of the KD tree structure to examine neighbouring nodes/branches of nodes, reducing the time of a $\mathcal{O}(k \log N)$ operation to one that scales as $\mathcal{O}(\log k \log N)$ for each point. These distances are then employed in assigning the galaxies to individual clusters.

2.2.5 Assigning Galaxies to Clusters

The 6th nearest neighbour to each $\mathrm{C4_{M05}}$ galaxy is used to establish a pseudo-density measure, importing the hypothesis that a galaxy in an overdense region will be closer to its 6th nearest neighbour than a galaxy in a less dense region. The list of $\mathrm{C4_{M05}}$ galaxies is ordered by the 6th nearest neighbour distance (smallest to largest i.e., most to least dense) and the galaxy with the smallest nearest neighbour distance (most dense) is set as the cluster centre. Then, neighbouring galaxies are assigned to the cluster centre, defining a spherically over-dense region. Selecting galaxies radially from the cluster centre, a number density is evaluated for the cluster at each galaxy's radius until the number density is below $1.2\,h^{-3}\,\mathrm{galaxies/Mpc^3}$, at which point the cluster radius is set at the outermost galaxy above that density threshold. If the threshold density is not reached before the cluster has grown to $5\,h^{-1}\,\mathrm{Mpc}$ in radius, then assignment of galaxies to the cluster centre is halted at this radius. If any of these cluster centres finds fewer than three galaxies as being the entire membership of a cluster above the density threshold, then they, and hence the "cluster" they form, are excluded. The remaining galaxies are then re-ordered by their 6th nearest neighbour and the process iterates until there are less than two remaining galaxies.

The penultimate step of applying an N_{gals} cut to the cluster catalog is taken and all candidate clusters with fewer than 8 members within $1\,h^{-1}\,\mathrm{Mpc}$ are rejected to allow a reliable velocity dispersion measurement. Finally, a "significance" cut demands that, in a

fixed volume around the cluster defined by its members, at least 10% of all galaxies in any given cluster's volume have been identified as $C4_{M05}$ cluster members. These remaining clusters form the $C4_{M05}$ cluster catalog.

2.3 Results of the Miller et al. Application of C4 to SDSS DR2

As mentioned previously, the $C4_{M05}$ cluster catalog contains 748 clusters of galaxies from $\sim$2600 $\square^\circ$ of SDSS DR2 spectroscopic data. These clusters range from a minimum of 10 members to over 200, and cover a redshift interval of $0.03 \leq z \leq 0.17$. Using simulated, or *mock*, galaxy catalogs developed by Wechsler (2004), the $C4_{M05}$ catalog is estimated to be 95% pure (where *purity** is the fraction of $C4_{M05}$ clusters discovered that match to a simulation halo), and $\sim$90% complete (where *completeness** is the fraction of simulated halos recovered by matching to the $C4_{M05}$ clusters found in the simulation) above $M_{200} \gtrsim 2 \times 10^{14} M_\odot$, where M_{200} is the mass within a radius that is greater than 200 times the critical density, and within $0.03 \leq z \leq 0.12$.

2.3.1 Purity and Completeness

Purity and completeness were evaluated using the mock catalogs, relating halo mass to cluster observables. These mock catalogs were designed such that the simulated galaxies match the distribution of galaxy colours, luminosities and their environmental dependence in the real SDSS data. By testing $C4_{M05}$ with these mock catalogs, simulated clusters found by $C4_{M05}$ could be directly related to the properties of the clusters' halos in the simulation, avoiding ambiguity in the definition of mass between the theoretical (simulated halo catalog) and observable (mock $C4_{M05}$ cluster catalog) measurements of each mock cluster. M05 found that the mock $C4_{M05}$ catalog was more than 99% pure for luminosities greater than $3 \times 10^{11} L_\odot$ (in r-band) and more than 90% complete for systems with a halo mass of $M_{200} \gtrsim 2 \times 10^{14} M_\odot$. Whilst completeness fell to around $\sim$50% at $M_{200} \simeq 1 \times 10^{14} M_\odot$, purity remained high, dropping to 90% for lower luminosity clusters.

By examining radial variations of the velocity dispersion for each cluster, M05 found that significant contamination of velocity dispersions arose when measuring clusters embedded in large scale structure. A structure contamination flag (SCF) with indices SCF= $[0, 2]$ was created to identify clusters with least (0) to most (2) contamination.

*§1.6

Employing a scaling relation to velocity dispersions of clusters, with SCF= 0 (minimum structure), delivered the smallest scatter against the halo masses. However, r-band luminosity was also found to be a good proxy for estimating halo-mass, with a scatter of 20% in the simulations, which was marginally larger than the best case (SCF= 0) scenario for velocity dispersions, but was not significantly affected by intervening large scale structure. The apparent magnitudes of the cluster members were first K-corrected (Blanton et al., 2003b) and extinction corrected (Schlegel et al., 1998) before being converted to luminosities, as per Fukugita et al. (1996).

2.3.2 Calibrating C4 Parameters

The simulated galaxy catalogs were also used to help calibrate parameters utilized by the $C4_{M05}$ algorithm. M05 did not test all parameters, limiting investigations to the impact of $C4_{M05}$ kernel radius (§2.2.2), the size of the redshift bins, the size of the colour box (§2.2.2), the FDR threshold, and the amount of colour information utilized on the purity and completeness of the mock $C4_{M05}$ catalog.

$C4_{M05}$ was run with kernel radii set at 0.5, 1.0, 2.0, and 6.0 h^{-1} Mpc. The kernel radius of $0.5\,h^{-1}$ Mpc appeared too small, causing completeness to fall 20% below that of the other parameters at halo masses $M_{\rm halo} \lesssim 5 \times 10^{14} M_{\odot}$, but did show an extremely high purity. This suggests that whilst $C4_{M05}$ was not finding as many clusters with the smallest kernel radius, the clusters it was finding were systems that existed in the simulation. Larger kernel radii made little difference to the purity and completeness of the mock $C4_{M05}$ catalog, maximising at $1\,h^{-1}$ Mpc and varying very little ($\lesssim 4\%$) at larger radii at the probed halo masses, $M_{\rm halo} \gtrsim 5 \times 10^{13} M_{\odot}$.

The size of the redshift bins was varied as comoving lengths of 25, 50, 100 and 200 h^{-1} Mpc. The $C4_{M05}$ mock cluster catalog was found to be independent of the line of sight aperture. This was attributed to the dearth of multiple clusters lying along the line of sight containing galaxies with similar global colours and that by not using k-corrections on the SDSS colours, the redshift dimension has already been accounted for.

To test the colour box, γ was allowed to vary as 1, 2, 4, and 6. Much like the smallest radial aperture, the smallest colour box, i.e., where $\gamma = 1$, produced a highly incomplete but highly pure catalog. As the colour box size is increased, the completeness becomes greater at the cost of lower purity. Between increasing $\gamma = 4$ to $\gamma = 6$, the completeness fraction changes very little but there is a significant drop in purity. This implies that at values of γ above $\gamma = 4$, few galaxies that are added to the $C4_{M05}$ mock galaxy catalog

that actually belong to the mock halos. The purity of the $\gamma = 4$ sample is seen to be fairly high ($> 90\%$), and so $\gamma = 4$ was used for the $C4_{M05}$ run on DR2.

The FDR threshold was varied as 5%, 10%, 15%, 20% and 50%. The least conservative threshold of 50% produces the highest completeness, at the expense of the purity. As the FDR threshold becomes more stringent, completeness falls whilst purity rises. However, this variation in completeness and purity is of order $\sim 10\%$, and so M05 chose an FDR threshold of 10%, which maximises purity, with the view that users of the $C4_{M05}$ catalog can have greater confidence that the cluster systems contained within the $C4_{M05}$-SDSS cluster catalog are real clusters.

To examine the effect of the amount of colour information available, the purity and completeness of the $C4_{M05}$ algorithm was calculated separately for each colour component and pairs of colours, then compared with the $C4_{M05}$ catalog that utilises all four colours. They found that using all four colours results in a catalog with much higher purity than using two colours, which in turn delivered catalogs with a higher purity than the single colour catalogs. However, completeness was found to show an inverse trend, with the single colour catalogs being more complete than the two colour catalogs, which in turn were more complete than the four colour catalog. This was found to be because the catalogs created with less colour information contained approximately twice the number of clusters as the four colour catalog, and so matched to the halo catalog because there were more opportunities to match. By scattering the cluster coordinates and matching to the halo catalog again (with the same $10'$ search radius), the single and double colour catalogs were seen to match more readily than the four colour catalog, which contains the fewest clusters.

2.3.3 Fiber Collisions

Fiber collisions may result in the loss of the Brightest Cluster Galaxies (BCGs) which account for a large proportion of the cluster luminosity (Bernardi et al., 2007). To test this, $C4_{M05}$ was tested with the mock data before and after applying fiber collisions, and $C4_{M05}$ was run on the real data (after fiber collisions) and SDSS photometric cluster galaxies missed by the tiling algorithm (that are close to the cluster red sequence, and otherwise fulfil the galaxy selection; section 2.1.2) were added in projection (since well-constrained redshift data was not available for the photometric sample). In the mock data, fiber collisions were replicated by identifying galaxy pairs with separations of less than 55" and retaining the brighter partner, and randomly selecting the remaining partners such

that 70% of all colliding galaxies were retained in the sample. Applied to the mock data, it was found that the fiber collisions result in a small (∼5%) drop in completeness (fewer systems found) and a small increase in purity (higher fraction of systems found match to simulated halos), and so fiber collisions are ruled to play a minimal role in cluster finding. However, the r-band luminosities are systematically underestimated, with bright massive clusters being more affected than smaller, dim clusters. They find that when BCGs were missed in the mock catalogs, clusters are dimmer by ∼24%, whilst clusters that were affected by fiber collision, but retain their BCGs, are dimmer by only ∼6%.

2.3.4 Comparison to Other Cluster Catalogs

Further comparisons were made to the existing Abell (§1.2.1) and RASS-SDSS (Popesso et al., 2004) cluster catalogs which have different selection techniques. The Abell cluster catalog was amassed by eyeballing the optical photographic plates from the Palomar All Sky Survey. The RASS-SDSS catalog identifies the optical counterparts to clusters discovered in the X-ray ROSAT All-Sky Survey. M05 find ∼90% completeness compared with these other cluster catalogs, taking into account different factors affecting each of the catalogs.

The survey area contains 346 Abell clusters within $10'$ of a $C4_{M05}$ cluster, with 123 Abell clusters uniquely matched to $C4_{M05}$ clusters, and five Abell clusters matching to two different $C4_{M05}$ clusters. To allow fair comparison between the Abell and $C4_{M05}$ clusters, constraints were placed on the survey volume covered, limiting the depth to $0.03 \leq z \leq 0.12$, where M05 assume the Abell cluster catalog to be highly complete, and find $C4_{M05}$ to be 90% complete above $2 \times 10^{14} M_{\odot}$. $C4_{M05}$ finds 71% of the 104 Abell clusters selected this way. Examination of the 30 missing Abell clusters revealed six were too deep for the $C4_{M05}$ algorithm, eight have too few spectra, and, by visual inspection, six are found not to be genuine clusters. Discounting these 20, and including another three Abell clusters which were blended into other nearby clusters, the recovery rate for the $C4_{M05}$ algorithm was found to be 92%.

Using the same $0.03 \leq z \leq 0.12$ volume constraints on the Popesso et al. (2004) cluster sample, $C4_{M05}$ found 39 of 43 RASS-SDSS clusters in the DR2 area. Allowing for a slightly larger matching radius at lower redshifts (since comoving distance decreases with decreasing redshift at fixed angular size), cluster deblending and edge effects, the $C4_{M05}$-RASS recovery rate was improved to 98%.

2.3.5 Follow-on Work

A number of publications have used the $C4_{M05}$ catalog to investigate various aspects of cluster astrophysics and cosmology. $C4_{M05}$ has also been examined in comparison to other cluster/group finders (e.g. group finders such as those in Berlind et al. (2006) and Yang et al. (2007)). A body of work also exists looking at the various properties of $C4_{M05}$ clusters and extending the mass-cluster observable work of M05.

Milosavljević et al. (2006) measures the luminosity gap (the difference in photometric magnitude between the two most luminous galaxies from a given system of galaxies) in the M05 catalog to quantify the dynamical age of the $C4_{M05}$ clusters, identifying fossil groups and then comparing them to theoretical predictions. Poggianti et al. (2006) uses the M05 catalog as a second SDSS cluster sample to measure the $O_{II}-\sigma$ relation in clusters and then infer its evolution with a high-z dataset from the ESO Distant Cluster Survey (EDisCS, White et al., 2005). Similarly, De Lucia et al. (2007) examines the EDisCS sample complemented with the M05 cluster catalog to examine the build-up of the red sequence in galaxy clusters since $z\sim0.8$.

Bernardi et al. (2006) uses the M05 catalog, to evaluate the environmental effect of early-type galaxy evolution, using it as a further indicator of a group's environment (as well as using the density of the galaxy groups themselves; groups found near clusters are classed as being in a dense environment). Aguerri et al. (2007) looks at the morphological distribution, velocity dispersion profiles and fraction of blue galaxies in clusters in SDSS-DR4, using cluster catalogs obtained prior to the SDSS, deriving and comparing their results with $C4_{M05}$ and RASS-SDSS. Using the Galaxy Zoo (Lintott et al., 2008) to classify galaxy morphologies, Bamford et al. (2009) looks at the dependence of morphology and colour on environment, using the M05 cluster sample to identify groups and group environments.

von der Linden et al. (2007) improves upon the clusters found by $C4_{M05}$ cluster finder, which is limited to the galaxies selected by the spectroscopic targeting algorithm (§2.1.2) and recoup the Brightest Cluster Galaxy (BCG) memberships that may have been missing from the M05 clusters due to fiber collisions (§2.3.3) and/or colour/magnitude contraints. With the BCGs identified in the M05 cluster catalog, Bernardi et al. (2007) looks at their properties within their respective clusters and discuss the implications for cluster formation histories. At other wavelengths, Best et al. (2007) discusses the connection and physical mechanisms for coincidences between (radio-loud active galactic nuclei) AGN and the $C4_{M05}$ cluster BCGs. Schawinski et al. (2007) uses GALEX near-UV (NUV)

photometry of a sample of early-type galaxies selected from $C4_{M05}$ clusters, particularly BCGs, to study the UV colour-magnitude relation.

2.4 Shortcomings of C4

In this section, I will briefly overview some of the shortcomings of the $C4_{M05}$ algorithm that may limit its success (§2.3) in the future.

2.4.1 Spectroscopic Redshift Limitations

$C4_{M05}$'s reliance on spectroscopic redshifts to group galaxies into clusters reduces its efficacy in higher redshift regimes (M05 is limited to $z < 0.17$), where data is limited due to galaxy magnitudes being too faint for spectroscopic confirmation by SDSS. $C4_{M05}$ leverages statistical power by being able to utilise data over large volumes of sky, but spectroscopic surveys are typically either wide and shallow, or narrow and deep.

In addition to SDSS spectroscopy, there are only a handful of current generation wide-field spectroscopic surveys: BOSS ($\sim 10,000\,\square^\circ$, Eisenstein et al., 2011), WiggleZ ($\sim 1,000\,\square^\circ$, Drinkwater et al., 2010), and GAMA ($\sim 250\,\square^\circ$, Driver and team, 2008), whilst most spectroscopic observation is focussed on following up astrophysical objects of interest or narrow field-of-view surveys e.g., COSMOS ($\sim 2\,\square^\circ$, Scoville et al., 2006). Spectroscopic surveys also take a significant amount of time to cover the volume of universe imaged by photometric surveys to a reasonable completeness, leaving C4 with limited future prospects for cluster finding.

2.4.2 Association of Galaxies to Clusters

When the $C4_{M05}$ algorithm forms C4 clusters from the C4 galaxy catalog (§2.2.5), it assumes the galaxies are (approximately) spherically distributed around a cluster centre with a densely populated centre. Whilst the premise of a concentrated population of galaxies is the very definition of a galaxy cluster, there are issues with treating the spectroscopic redshift as a cosmological distance. Galaxies are considered as being in a cluster when their motion is dominated by the gravitational potential of the cluster; moreso than the Hubble flow (the apparent motion of galaxies due to the expansion of the universe).

This dissociation from the Hubble flow may well introduce unaccounted line-of-sight fragmentation and/or blending in subsequent C4 catalogs due to the assignment of cluster centres being made in comoving space, where the redshift dimension is simply converted to a comoving distance (Equation 2.24), and the assignment of galaxies to that cluster

centre, which also uses the assumption that the source of the redshift is the Hubble flow. Assuming hydrostatic equilibrium, galaxies in a rich galaxy cluster move with random peculiar velocities (median line-of-sight velocities) of typically $\sim 750\,\mathrm{km\,s^{-1}}$ (Bahcall, 2000). Using the assumed cosmology of $\mathrm{H}_0 = 100\,h\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, means galaxies whose peculiar velocities are aligned with their recessional velocities appear a median $7.5\,h^{-1}\,\mathrm{Mpc}$ from the cluster centre, i.e., *outside* the maximum radius that galaxies are assigned to a cluster centre.

2.4.3 Cosmology

$\mathrm{C4_{M05}}$ uses an $\Omega_\mathrm{m} = 1$ cosmology to convert between redshift and comoving space (§2.2.2), which is a reasonable approximation at the limited redshifts of the $\mathrm{C4_{M05}}$ catalog. To convert a redshift into a comoving distance, one must account for the underlying cosmology, and expansion history, of the universe between the observed galaxy and the observer, and improper assumptions on a chosen cosmology may impact the efficacy of C4 at higher redshifts.

2.4.4 Ease of Use

The $\mathrm{C4_{M05}}$ algorithm is specifically tailored to the SDSS DR2 dataset. For example, $\mathrm{C4_{M05}}$ needs information telling it how many galaxies are in the input catalog and the limits of a spatial grid in which to compute the C4-clustering (§2.2.2). To adapt $\mathrm{C4_{M05}}$ to different data sources, changes can be made to $\mathrm{C4_{M05}}$ by going into the source code and changing the hard coded values. Repeated runs of $\mathrm{C4_{M05}}$ also require hard coding of the filenames throughout the algorithm if unique $\mathrm{C4_{M05}}$ parameter combinations are to be recorded for later comparison. This may need to be done several times since C4 parameters will need to be repeatedly run on a dataset and evaluated before an optimal parameter set is determined for the production of the final $\mathrm{C4_{M05}}$ cluster catalog from that input.

When $\mathrm{C4_{M05}}$ ingests the DR2 data, it takes in all of the DR2 galaxy information it is given. In the future, when datasets become far larger than the spectroscopic DR2, running C4 will require increased computing resources such that full data ingestion may not be possible with only a single CPU.

The $\mathrm{C4_{M05}}$ algorithm also has a noticeable shortcoming in that with the production of a cluster catalog, the galaxy membership only exists as a number count, meaning reanalysis is necessary to determine the actual members of the "found" C4 cluster for further

population/substructure studies (§2.3.5). Whilst this my not be challenging with shallow datasets, increasing the depth of the data given to C4 may result in an increased number of systems found along a single line-of-sight, leading to confusion when re-associating those C4 galaxies to their C4 defined cluster.

An esoteric programming criticism can be made of $C4_{M05}$ in that variables are defined multiple times throughout the algorithm, such that if a parameter value is changed, it must be changed in multiple places. If the parameter is not changed in some of those places, then the C4 catalog produced will be inconsistent with the user's expectations.

2.5 Summary

This chapter reviews the methodology and subsequent science produced as a result of the M05 $C4_{M05}$ catalog. This is a highly complete and pure cluster sample, as determined by comparison to existing cluster catalogs and tests with simulated galaxy/halo catalogs (§2.3).

In section 2.1, I have described the photometric and spectroscopic SDSS campaigns and discussed the selection of spectroscopic galaxies from the photometric survey that form the spectroscopic DR2 galaxy catalog.

In section 2.2, I described the original $C4_{M05}$ cluster finding algorithm, detailing its main components, and in section 2.3 I explored the science results from $C4_{M05}$ and some of the work that followed. In section 2.4, I spell out some of the shortcomings of $C4_{M05}$ with regards to its application to future data.

As well as the main $C4_{M05}$ measurement of colour-clustering (§2.2.2), the use of key concepts such as the FDR to identify the cluster galaxies and the use of k-NN to approximate cluster centres (§2.2.3 and §2.2.4, respectively) give the output $C4_{M05}$ catalogs a competitive level of statistical certainty. It is these concepts that define the $C4_{M05}$ algorithm and are common to the AperC4 algorithm, introduced in this thesis, that will be explored in subsequent chapters.

Chapter 3

AperC4: a New Cluster Finding Algorithm

In this chapter, I introduce the new AperC4 algorithm presented in this thesis and compare it to the original C4 algorithm (see chapter 2). I refer to the M05 C4 algorithm and products as $C4_{M05}$ and refer to the new C4 algorithm as AperC4 to differentiate them, whilst general C4 concepts are not assigned a subscript. AperC4 is named for its use of C4 concepts but sampling over multiple angular apertures rather than depending on redshift information. In section 3.1, I discuss the motivations for building AperC4 and consider how the changes from $C4_{M05}$ to AperC4 will affect cluster finding. I present the outline of this new algorithm, the AperC4 workflow, and the free parameters associated with the algorithm, in section 3.2. In section 3.3, I discuss how AperC4 aperture choices are made, then run through some thought experiments on the purpose of using multiple apertures, outlining their efficacy in different test cases. I then describe the advantages of using $p(z)$ over photo-z in breaking membership degeneracies between AperC4 catalogs.

3.1 Motivations for Modifications

3.1.1 Why Modify C4?

As discussed in §1.5, and at length in chapter 2, $C4_{M05}$ is an advanced cluster finding algorithm that lent itself to a range of scientific applications. Where $C4_{M05}$'s strength lies is in the purity of the M05 cluster catalog, and its high completeness measured against cluster catalogs from other wavelengths, which made it ideal for studies of the astrophysical nature of galaxy clusters (see §2.3.5). However, $C4_{M05}$'s shortcomings (§2.4) mean that it has limited future prospects.

The premise for creating a new C4 algorithm is to extend $C4_{M05}$ into the photometric regime, which is only limited to the magnitude limit of the imaging data supplied by any given survey telescope, rather than the shallower magnitude limit of followup spectroscopic observations. Additional motivation for this transition comes from the present generation of photometric surveys, e.g., DES (§4.1), Pan-STARRS (Kaiser et al., 2002), KiDS (de Jong et al., 2013), and VIKING (Sutherland, 2012), which are/will be delivering large volumes of extragalactic data for a variety of astronomical and cosmological applications (Seo and Eisenstein, 2003; Munshi et al., 2008; Venemans et al., 2012; Kirk et al., 2013). Proposed future surveys, such as LSST (LSST Collaboration, 2012), and Euclid (Laureijs et al., 2011), will provide even greater volumes of data to deliver better constraints and characterisation on the matter-energy content of the late-time universe. Detection of clusters in such data demand that any new cluster finding algorithm can computationally handle large volumes of data in a consistent way.

3.1.2 Redshift Considerations

$C4_{M05}$ is engineered towards spectroscopic data sets (specifically SDSS DR2), incurring limitations based on the depth of spectroscopic data (§2.4.1), and depends on assumptions on the deviation of redshifts of those cluster galaxies (§2.4.2).

APERC4 makes a large departure from $C4_{M05}$ by considering neither spectroscopic nor photometric redshifts when establishing the most cluster-like galaxies. Redshifts are employed only *after* candidate clusters have been defined by APERC4(§3.2.3).

Redshift information is used in the form of galaxy $p(z)$s (probabilistic redshift information derived from photometry; see §6.1) rather than spectroscopy. APERC4 uses $p(z)$s to optimize galaxy membership of the clusters identified, deprojecting clusters that may lie along the line-of-sight. Using the recorded APERC4 membership information and invoking external redshift information, galaxies that have been assigned to more than one cluster are reassigned to their (single) optimal cluster. This reassignment should allow discrimination of cluster substructure from smaller groups/clusters (§3.3.3).

A number of methods exist to give observed galaxies redshifts via photometric information (Csabai et al., 2003; Collister and Lahav, 2004; Wang et al., 2007; Cunha et al., 2009; Gerdes et al., 2009; Arnouts and Ilbert, 2011). APERC4 is set up to use any redshift mechanism available. To do so, it treats redshifts as a probability distribution, $p(z)$ across the entire observed redshift range of the photo-z/$p(z)$/spectro-z method. This is in tune with the more recent photometric redshift algorithms of Cunha et al. (2009) and Gerdes

et al. (2009), where the galaxies themselves are assigned a $p(z)$. The $p(z)$ given by these methods is typically based on the respective algorithms calibrating to a subset of galaxies with known spectroscopic redshifts, from some defined set of sources, and comparing them to their observed properties in the photometric survey, nominally from the SDSS (I discuss details of the Cunha et al., 2009, algorithm further in section 6.1).

3.1.3 On Apertures and Cosmology

The utilisation of predetermined angular apertures is a defining feature of AperC4. The use of apertures is motivated by the use of $p(z)$ redshift information and its translation to physical distances. Cluster finders (§1.5) will typically look for concentrations of galaxies within a $1\,h^{-1}\,\mathrm{Mpc}$ to $1.5\,h^{-1}\,\mathrm{Mpc}$ radius kernel at the proposed redshift of the cluster, using this as a physical cut off for cluster membership. When employing $p(z)$ information, there is no plausible single angular search radius that can be constructed from a comoving separation. Simultaneously, to translate a physical/comoving separation to an angular separation, one must also assume some form of cosmology (and as mentioned in §2.4.3, $\mathrm{C4_{M05}}$ already requires an update to its cosmological assumptions).

So to address the difficulty of associating a physical radius to a galaxy with a probabilistic redshift distribution, and the assumption of a known background cosmology, AperC4 functions by utilising multiple, explicit angular radii. Given a cosmology, an angular separation on the sky will relate to a physical transverse separation at a given redshift. Given a different cosmology, the same redshift and the same angular separation, all that changes is the size of that physical separation. By sampling enough differing angular radii, features on different physical scales at all redshifts may be probed. In this sense, AperC4 is cosmologically independent, avoiding complications in choice of input cosmology by not invoking one! AperC4 then relies on the 4-dimensional colour-clustering of cluster galaxies within these angular radii to identify clusters.

I save discussion of the finer points of choosing a set of angular apertures, and their effect on cluster finding, for section 3.3.

3.1.4 Computational Modifications

To address the limitations of $\mathrm{C4_{M05}}$'s ease of use (§2.4.4), the running of the C4 algorithm is modified in AperC4 such that:

- the algorithm is able to recognise parameter value combinations already evaluated, avoiding recalculation where unnecessary;

- if recalculation of a specific set of parameters is necessary, then it is possible to rerun the algorithm and save the new catalog results without necessarily losing/overwriting the old set of results;
- parameters have "legal" ranges, i.e., parameter values are limited to a finite range or set of options, where appropriate;
- APERC4 does not need to know the dimensions of the input galaxy catalog ahead of runtime (§3.2.3), i.e.,
  - the redshift range of the input galaxy catalog can be unknown;
  - the area of the survey (R.A./dec limits) covered by the input catalog can also be unknown.
  - the number of input galaxies does not need to be calculated beforehand.
- APERC4 outputs a member catalog that specifies which cluster a C4 galaxy has been assigned to (§2.4.2).
- output APERC4 catalogs are given identifiable names with user defined prefixes;
- output APERC4 catalogs and intermediary files can be placed in a specified directory.

From a programming perspective, the APERC4 algorithm is programmed such that:

- user-defined APERC4 parameter values are initialised at one location within the algorithm (when the algorithm is first called), except for mathematical/physical constants (e.g., π, e, c, etc.) which remain hardcoded;
- parameters have consistent names across the APERC4 software modules for ease of use by other programmers (§3.2.10);
- rather than being fixed to $1\,\square^\circ$ (as in $C4_{M05}$, §2.2.2), the scale of the APERC4 galaxy grid is allowed to vary, which permits optimisation of the algorithm to computing resources (§3.2.3);
- to deal with catalogs that cannot be handled by a computing system's memory, the APERC4 algorithm breaks the survey area into overlapping tiles.

These computing resource modifications include additional features that anticipate the usage of other, larger datasets by the APERC4 algorithm. This new dynamism of the APERC4 algorithm precipitated its translation from the `C` programming language to `C++`. These modifications shorten APERC4's effective running time when running the

algorithm multiple times, allowing a faster turnaround time for parameter exploration or sanity checks.

The original $C4_{M05}$ algorithm employs some external code that no longer compiles under current operating systems. In section 3.2.5, I replace the kth nearest neighbour code used by M05 with the newer code, MLPACK (Curtin et al., 2011). To facilitate this change, the AperC4 algorithm is written in a modular format, which allows components such as this $k-$NN module to be replaced without impacting the rest of the algorithm. This modularisation gives AperC4 a level of future-proofing over the original $C4_{M05}$ algorithm.

3.2 Introduction to the New Algorithm

In this section, I shall describe the new cluster finding algorithm, AperC4. In subsection 3.2.1, I give a brief overview of the AperC4 process. In the subsections that follow, §3.2.2 to §3.2.9 , I describe the various processes of the AperC4 pipeline in finer detail, describing deviations from the $C4_{M05}$ processes. In subsection 3.2.10, I summarise the AperC4 input parameters, giving their default values. I will discuss the rationale for these processes in section 3.3.

3.2.1 AperC4 Algorithm Outline

For clarity, I describe the survey area as processed for a single aperture as an "*aperture-slice*". In step 1 (below), the survey area is split into overlapping tiles, hence, each aperture-slice is also divided into these galaxy tiles. I refer to a single galaxy tile that has been processed by a single aperture as an "*aperture-tile*". The AperC4 algorithm flow can be summarised in the following eight steps.

Step 1: Split galaxy survey into tiles (§3.2.2). AperC4 ingests the galaxy catalog and splits it into overlapping tiles.

Step 2: Calculate p-values (§3.2.3). p-values are calculated for each galaxy, as per §2.2.2 using every aperture in every tile.

Step 3: Identify C4 galaxies (§3.2.4). Calculate and apply the FDR (§2.2.3) on each aperture slice to produce "AperC4 galaxies", i.e., cluster-like galaxies as identified at some aperture scale.

Step 4: Find $k-$NN distance (§3.2.5). Find $k-$NN distance in 2-D to identify the galaxies in the densest areas for each aperture slice.

Step 5: Form clusters (§3.2.6). AperC4 associates galaxies to the densest galaxy regions identified in Step 4 above, down to a lower threshold surface density of galaxies (i.e. some number of galaxies per $\square^\circ$in the survey area). These clusters form the initial cluster catalogs and member catalogs.

Step 6: Redshift assignment (§3.2.7). The clusters in each aperture slice are assigned a weighted redshift distribution based on the redshifts of their constituent members.

Step 7: Aperture consolidation and membership optimization (§3.2.8). Clusters in aperture slices are combined where their redshift and memberships agree, else shared members between clusters with contrasting redshifts are assigned to their optimal cluster. This step produces a single cluster catalog with uniquely defined memberships.

Step 8: Apply $N_{\mathbf{gals}}$ threshold (§3.2.9). AperC4 removes clusters with fewer than some arbitrary number of galaxies. The remaining clusters and their members form the final AperC4 cluster and member catalog.

3.2.2 Step 1. Survey treatment: Tiling

AperC4 ingests the galaxy catalog and splits it into overlapping tiles. The only condition on these tiles is that they are large enough to be representative of the galaxy content of the universe. The size of these tiles is determined by the `SPLITSIZE`* and `OVERLAP` parameters which describe the minimum tile side length in degrees in both R.A./dec and the minimum overlap length in degrees. A typical tile will cover (`SPLITSIZE` $\times$ `OVERLAP` $\times \cos(\text{dec})$) $\square^\circ$, where dec is the mean declination of the tile, excepting survey edges which can be as small as (`SPLITSIZE`^{2} $\times \cos(\text{dec})$) $\square^\circ$. The size of the minimum overlap should be guided by the twice the angular extent of the largest association of galaxies (i.e., galaxy cluster) one may expect to find.

During this step, apparent magnitude limits can be placed on the galaxy catalog for each band of the survey with the `SURVEY_LIMITS` parameter. In this thesis, I apply AperC4 to SDSS style multi-band data, so the number of bands is limited to the five bands of the SDSS (§2.1.1). An optional `MAGLIM` parameter can be employed to increase/decrease all the band limits supplied by `SURVEY_LIMITS`.

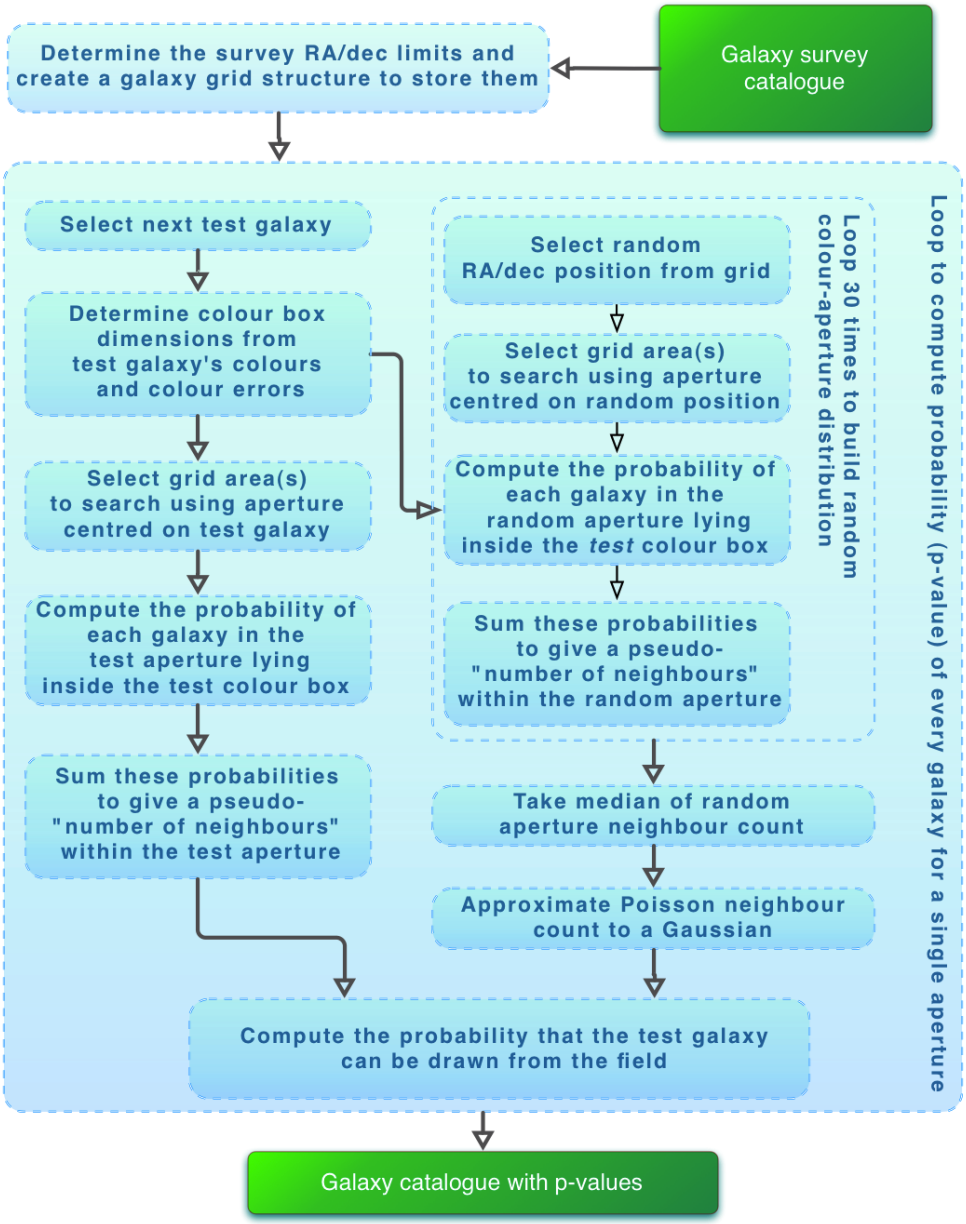


Figure 3.1: The APERC4 p-value calculation for a single aperture-tile is summarised in this flowchart. The blocks in green represent the galaxy catalog ingested (*top* green block) and the galaxy catalog output with the p-values (*bottom* green block). The blue blocks represent the processes used in every p-value calculation and are discussed in greater detail in section 3.2.3.

3.2.3 Step 2. AperC4-clustering: p-value Measurement

Adaptive Grid Creation

Starting with the creation of the grid, APERC4 is adaptive to the spatial dimensions of the input galaxy catalog, calculating the size of the grid needed to hold the data, creating that grid (with the C++ MULTIARRAY (Garcia and Lumsdaine, 2005) library from the

*In this thesis, `PARAMETERS` are written in this typeface.

BOOST project*), and will itself calculate the amount of data (number of galaxies) it needs to place in the grid.

As well as assigning the dimensional limits of the R.A. and dec. grid, I incorporate a new parameter, `RADECSCALE`, that allows the grid to be subdivided further into fractions of a degree. This affects the computational balance between CPU time and memory usage. `RADECSCALE` can be understood as the number of grid elements contained in a degree or as the square root of the number of grid elements per square degree (e.g. setting `RADECSCALE` = 6 results in a grid that holds 36 elements per square degree, each element being $10' \times 10'$). In the case of low memory systems, the `RADECSCALE` parameter can be set to allow one grid element per degree, which requires fewer grid elements to hold the galaxy data in memory, but when evaluating APERC4-clustering all galaxies sharing the same grid element are examined for their APERC4-clustering properties, and thus requires more CPU time. Where a system has access to more memory, `RADECSCALE` can be set to allow more grid elements per square degree, reducing the number of neighbouring galaxies evaluated for APERC4-clustering for each galaxy, thus reducing CPU time employed. This trade-off, to be made between processing time and memory usage, does not affect the APERC4 clustering itself.

Evaluating AperC4-clustering

Once the galaxy survey is split into tiles (§3.2.2), each tile is processed multiple times using the different aperture scales, or radii, which are used to construct angular apertures around each galaxy in a tile. This gives multiple APERC4-clustering p-values per galaxy (Step 2 in §3.2.1). The calculated p-values using the different apertures are kept as separate tiles themselves, such that a single input galaxy tile will produce multiple aperture tiles. For clarity, I will describe the process for a single aperture size in a single galaxy tile (aperture-tile). Each aperture radius is input into the algorithm from an array called `APERTURESIZE`. No redshift limit is involved in this calculation.

For each galaxy (target galaxy), APERC4 creates a colour-area aperture. When APERC4 creates a radial aperture for the galaxy, it searches all grid elements covered by the angular aperture, where the angular aperture is given explicitly to the algorithm.

Colour clustering is as described in section 2.2.2. However, a new multiplier is applied to the colour error, σ_{xy}, which in turn affects γ. M05 noted that the median of σ_{xy}(stat) = 0.02 for the DR2 data changes very little over the magnitude range, partly due to the fact

*http://www.boost.org/

that these are the brightest galaxies in the DR2 data, and were thus satisfied with σ_{xy}(sys) being the dominant term of Equation 2.18. By introducing the free parameter, $X_{\sigma_{xy}}$, as a multiplier to σ_{xy}, the relationship between the colour box size and Equation 2.18 becomes:

$$\begin{matrix}\text{Colour} \\ \text{box size}\end{matrix} \;=\; X_{\sigma_{xy}}\sigma_{xy} \;=\; X_{\sigma_{xy}}\sqrt{\gamma\sigma_{xy}^2(\text{sys}) + \sigma_{xy}^2(\text{stat})} \tag{3.1}$$

.

This scaling factor translates to the X seen in Figure 2.3. This formulation for the colour box keeps the γ parameter dependent on the overall scaling for the colour box. The systematic colour component was primarily instantiated from measurements of magnitude variations in 158 standard stars across the SDSS survey area (Smith et al., 2002), as measured with a 20-inch telescope with a non-SDSS filter set observed at Apache Point and interpolating them to the SDSS telescope system (the SDSS telescope could not directly observe the stars as they would quickly saturate the SDSS telescope CCDs). An improved photometric system was developed by Padmanabhan et al. (2008), entitled *übercalibration*, that significantly reduces this systematic colour component across the whole survey area. Additionally, as the survey area is treated in overlapping tiles by ApERC4 (§3.2.2), this systematic colour component is further reduced with respect to each individual galaxy. Thus, the $X_{\sigma_{xy}}$ term is kept outside the quadratic combination of systematic and statistical terms so that its meaning is clear for catalogs created where $\gamma = 0$. In the code itself, the $X_{\sigma_{xy}}$ term is indicated by the `SIGMACOLOUR` parameter, whilst γ is named `GAMMAFACTOR`.

The evaluation of $N_{neighbour}$ changes from that in Equation 2.20 to:

$$N_{\text{neighbour}}^{\text{target}} = \sum_{n\in A}\int_{-X\sigma_{xy}}^{X\sigma_{xy}} P(\sigma_{xy,neighbour})\; d\sigma_{xy}\;, \tag{3.2}$$

where the limits of the colour space integral are now increased/decreased by the scaling factor, X, and $(n \in A)$ specifies all the neighbouring galaxies, n, within an aperture, A, defined by some angular radius around the target galaxy.

ApERC4 evaluates the magnitudes of the target galaxy and each potential neighbour galaxy against the given magnitude limit of the survey, before evaluating the colour clustering. This is done for both magnitude components, of each colour, for both target galaxy and potential neighbour galaxies. This acts to maximize the amount of colour information available to each galaxy, by not throwing out galaxies that drop out of one or more SDSS bands (but not more than three out of the five) whilst respecting the limitation of detection accuracy of the survey instrument(s) in each respective waveband. Identification and selection of galaxies (on which ApERC4 is run) must be done prior to submitting the

galaxy data to APERC4. No weighting is applied to the colours in the absence of full colour information, and the colour-aperture probabilities (p-values; equations 2.21 and §2.2.3) are simply drawn from the combined probabilities of the employed colours.

For a single aperture-tile, this step is presented in Figure 3.1 as a flow chart. This process is repeated for all aperture-tiles in an aperture-slice, and for all aperture-slices (i.e., the full survey for all apertures).

3.2.4 Step 3. Identify C4 galaxies

Once the p-values for the galaxies are output, a FDR (§2.2.3) is employed to extract samples of galaxies (one set of p-values per aperture slice) that are the most colour-clustered in their respective apertures (more accurately, the least likely to have been drawn from the colour-clustering distribution generated by using a respective galaxy's colour box to sample the sky). The p-value distribution (Figure 3.2) can also be used to assess the effectiveness of the colour-aperture test. A strong test would result in a fraction of galaxies at a p-value of $p = 0.0$ and the remainder at $p = 1.0$, i.e., the test clearly discriminates between two types of galaxy, those that sparsely or densely populate a colour-aperture, resulting in a binary distribution. A weak test would result in a significant fraction of galaxies in the catalog having a p-value of $p \sim 0.5$, i.e., the test cannot discriminate field/cluster galaxies by how densely the colour-volume apertures are populated.

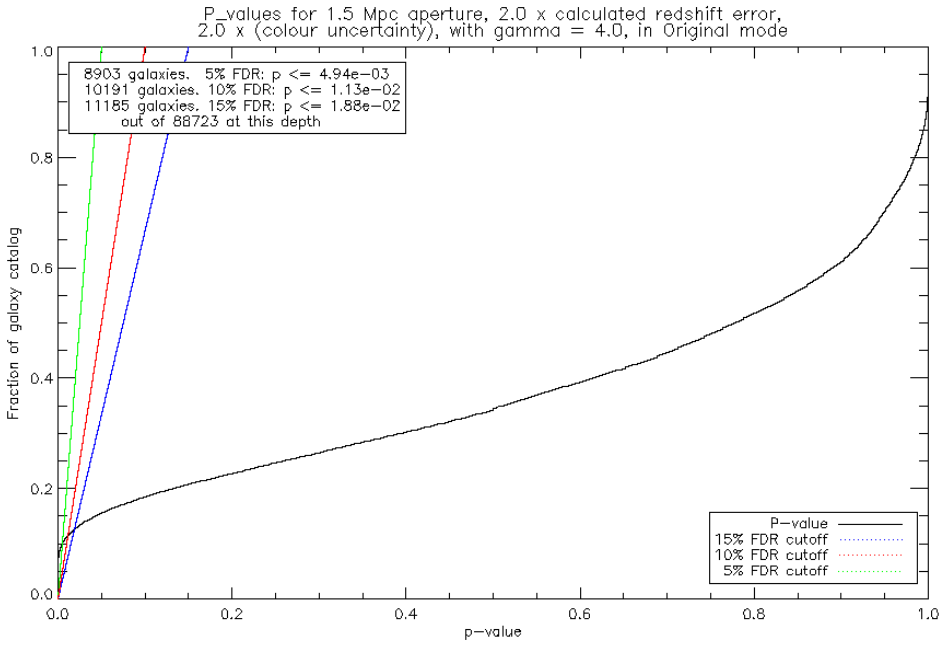


Figure 3.2: In this diagram, the p-values of the SDSS DR2 galaxies are plotted as a fraction of the catalog. The cuts representing a false discovery rate of 5%, 10% and 15% are highlighted in green, red and blue respectively

The aperture-tiles are combined into a single aperture-slice, eliminating duplicate galaxies in the overlap regions, and an appropriate FDR threshold is calculated for and applied to each aperture-slice, independently. The FDR is simply named `FDR` in the algorithm.

3.2.5 Step 4. Determining Cluster Centres with kth Nearest Neighbour

Once the FDR is applied to the aperture-slice galaxy catalogs, the galaxies' kth nearest neighbour distances are evaluated. The value of k is given to the algorithm as the `K_IN_KNN` parameter. The algorithm used is from part of the MLPACK* distribution (developed by Curtin et al. 2011). As the redshift dimension is unconstrained, the catalog is treated as a 2-dimensional map (utilising equations 2.22 & 2.23) and the kth nearest neighbour distance is calculated for each galaxy in the celestial plane.

At this stage the aperture-tiles are resized to allow parallelisation of the cluster forming processes (see Figure 3.3). This can be done because the survey tiles no longer need to be considered as a statistically representative portion of the survey area; the important thing is that the largest clusters can be adequately characterised by a set of galaxies captured in a single tile. These aperture-tiles still overlap by an amount set by the `OVERLAP` parameter set earlier (§3.2.2), but now use 3× `OVERLAP` to define the length of a tile axis. This ensures the time taken to process the aperture-tiles scales approximately linearly with increased survey area, helping reduce the effective run time of the algorithm (one of the stated aims of AperC4; see §3.1.4).

Note that the aperture slices are all still separate.

3.2.6 Step 5. Forming Clusters

As described in section 2.2.5, the kth nearest neighbour distances are used to sort the AperC4 galaxies (from smallest $k-$NN distance to the largest). Treating the distance as an inverse proxy for density, the first of these AperC4 galaxies (with the smallest nearest neighbour distance, and thus in the densest environment) is assigned as the cluster centre. Then, galaxies are added to the AperC4 cluster radially, in the plane of the sky from the centre, until the galaxies within a threshold radius defined by `APERTURESIZE` are exhausted. The galaxies assigned to the AperC4 cluster(s) are removed from the candidate list and the smallest kth nearest neighbour distance of the remaining unassigned galaxies is nominated as the next cluster centre, and the process repeats until all galaxies

*http://mloss.org/software/view/152/

are assigned or any remaining galaxies are isolated within an aperture radius. Isolated pairs of galaxies are also rejected at this stage.

The galaxies that pass the FDR threshold are significantly C4-clustered ($p \simeq 0.0$, §3.2.4) with neighbouring galaxies within an aperture radius (`APERTURESIZE`) in their respective aperture-slice. Real clusters can have multiple component colour behaviour (De Lucia et al., 2007), so no attempt is made to discriminate galaxies by colour within the aperture radius. Nearby galaxies (that cluster in colours dissimilar to the defined cluster centre) may be included, nominally due to their own AperC4-clustering significance.

The redshift dimension is unbound in AperC4 meaning the uncertainty on the redshifts of these clusters will be larger than those on the spectroscopic redshifts utilized in $C4_{M05}$. Indeed, by not bounding the redshift dimension, AperC4 permits increased blending of clusters along the line-of-sight. However, this method completely mitigates the problem of co-added recessional and peculiar velocities of spectroscopic methods and has no sensitivity to large errors in photometric redshift.

The AperC4 process up to this stage is given in a flowchart in Figure 3.3. To accelerate the process, the survey area is redivided into smaller overlapping tiles (as explained in §3.2.5). The R.A./dec dimensions of these subdivided tiles are defined by 3×`OVERLAP`, and overlap by 1×`OVERLAP`.

A further modification to the original algorithm means AperC4 retains galaxy membership for the clusters. AperC4 assigns IDs to the aperture-slice clusters and correspondingly linking these IDs to AperC4 galaxies in the aperture-slice galaxy catalogs. The result of this step is N_{aperture} aperture-slice cluster catalogs, where N_{aperture} is the number of apertures given to the algorithm, i.e., one cluster catalog per aperture-slice. Each of these cluster catalogs has a galaxy membership but they may contain galaxies duplicated between them, i.e., slices may contain galaxies whose AperC4-clustering strength persists at multiple apertures. These duplicate galaxies may lie in an equivalent cluster in other aperture-slices, or the aperture-slice clusters may be fragmented/blended with respect to one another. This allows AperC4 to combine the aperture-slices into a single cluster catalog, which I will discuss in the sections following.

3.2.7 Step 6. Assignment of Redshift to Clusters

The clusters in each aperture-slice are assigned a $p(z)$ by combining the product-sum of the $p(z)$ distributions of their respective members. Hence the initial $p(z)$ for an aperture-slice

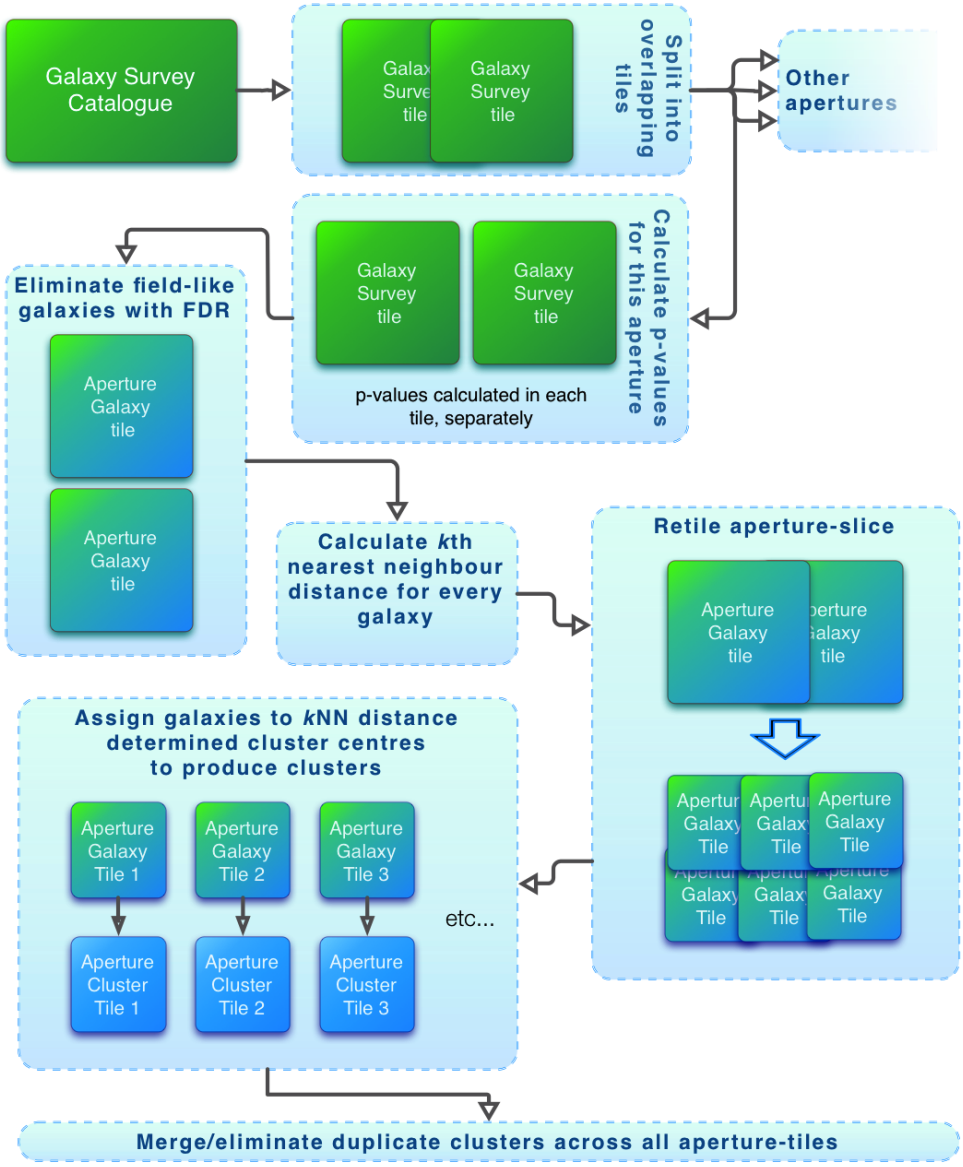


Figure 3.3: The APERC4 process for a single aperture is given in this flowchart. The dashed-line, pale blue boxes describe a process whilst the solid-line green, blue-green, and blue boxes indicate the data is the same as input, has been treated, or is new, respectively.

cluster with n members can be given by:

$$\text{Cluster } p(z) \ = \prod_{i=1}^{n} p(z_i). \tag{3.3}$$

where $i = 1, 2, \ldots, n$ simply indicates the galaxy is a member of the cluster; the order in which the product-sum is evaluated plays no role.

It can be noted that gaussian $z \pm \sigma_z$ style galaxy redshift information can also be interpolated onto an observed redshift range, by breaking such gaussians into discrete bins.

Each cluster $p(z)$ is then normalised to unity to give a representative redshift probability distribution function (PDF) for each cluster in each aperture-slice. Now that the

aperture-slice clusters have been given $p(z)$ information, the following section will show how APERC4 combines the tiles to give a single cluster catalog.

3.2.8 Step 7. Accounting for Fragmented Clusters

Each aperture-tile covers the same portion of survey for each aperture-slice. Considering a single tile area (i.e., the N_{aperture} tiles that cover a common area), the aperture-slice clusters within that area are combined in the following way.

The aperture-slice clusters are first given new, unique IDs such that the members do not get erroneously assigned to clusters defined in alternate aperture-slices, and the memberships of the galaxies in each aperture-slice are adjusted to retain this information. The tiles (of common area) are searched for duplicate galaxies, i.e., where they are found in more than one aperture-slice. Where a duplicate galaxy is found, APERC4 calculates the product of the $p(z)$ of the duplicate galaxy and the $p(z)$s of each aperture-slice cluster (§3.2.7) that the galaxy appears. The galaxy is assigned to the cluster where

$$p(z_{\text{clst|aptr}}) \times p(z_{\text{galaxy}}) \tag{3.4}$$

is maximised, where $p(z_{\text{clst|aptr}})$ is the cluster $p(z)$ calculated in Equation 3.3 for a given aperture. The galaxy is excised from all other aperture-slice clusters. The product-sum $p(z)$ for the clusters where the galaxy was removed is then recalculated and the process repeats for all duplicate galaxies found. This ensures that the memberships of the aperture-slice clusters are now unique amongst each other, i.e., a galaxy will belong to only one aperture-slice cluster in one aperture-slice.

Once the galaxies across all aperture-slices are no longer duplicated, the $p(z)$s of the aperture-slice clusters that shared members prior to the membership optimisation (above) are tested to see if they are at the same redshift (and therefore describe the same cluster, but with possibly differing memberships). The cross-$p(z)$ threshold parameter, `PZTHRESHOLD`, defines the minimum cross-probability of a cluster found with aperture A being the same as a cluster found with aperture B. If threshold is exceeded, i.e., $p(z)_A \times p(z)_B > p(z)_{\text{THRESHOLD}}$, then cluster A is merged with cluster B and their memberships are combined.

Once the aperture-slices contain no duplicate galaxies between them, and their clusters are unique in membership, the aperture-slices are combined into the APERC4 cluster catalog tile. This is repeated for every aperture-tile, such that there is now one cluster catalog for each tile area. The overlapping tile areas are similarly consolidated, and duplicate

members in the tiled clusters are treated the same way as above to form a single AperC4 cluster catalog.

This step serves the purpose of separating out clusters blended along the line-of-sight, whilst similarly merging cluster substructures that may appear separate in one aperture-slice but not another.

3.2.9 Step 8. Richness Cut

An $N_{\rm gals}$ cut-off is applied, acting as a richness cut removing AperC4 clusters with a membership of fewer than `NGALLIM` members, where `NGALLIM` is a parameter to test over.

Clusters with low memberships may often be chance superpositions rather than true 3D overdensities, due to photometric redshift uncertainties, and so this parameter serves to improve the purity of the final cluster sample. However, clusters with memberships of $N_{\rm gals} \leq$ `NGALLIM` may also include real systems, so completeness is negatively impacted by this limit.

In AperC4 this acts to eliminate galaxies that have become isolated through the (de)fragmentation step (§3.2.8). As galaxies get removed from their non-optimal clusters, the cluster $p(z)$ approaches the $p(z)$ distributions of its remaining members. In the limit of one member, one finds $p(z_{\rm cluster}) \mapsto p(z_{\rm galaxy})$. By definition, one cannot really define a cluster of galaxies with a membership of one. Similarly, when the number of galaxies in a cluster is small, the sum-product $p(z)$ is more sensitive to uncertainties in the redshift of any of its constituent members.

3.2.10 AperC4 Algorithm Input Parameters

Both AperC4 and $\mathrm{C4_{M05}}$ algorithms require a number of key parameters to run. As seen in this section, AperC4 utilizes free parameters from the $\mathrm{C4_{M05}}$ algorithm, frees or removes previously fixed parameters from the same, and implements new parameters. The key algorithmic parameters are described in Table 3.1, where parameters marked with a circle outline, ○, represent parameters common between $\mathrm{C4_{M05}}$ and AperC4; those marked with a filled circle, ●, are introduced in the AperC4 algorithm, and those with a half filled circle, ◐, have changed between the $\mathrm{C4_{M05}}$ and AperC4 algorithms.

Table 3.1: Descriptions of the set of parameters introduced, or modified from the old C4 algorithm, in the new APERC4 algorithm that affect the output APERC4 cluster catalog.

Parameter name (and origin)	**Variable type**	**Parameter values or range**	**Description**
`MAGLIM` ●	Float	$0.0 - 9.9$ (*1d.p.*)	`MAGLIM` is the *relative* magnitude cut made in all bands, allowing all the magnitude limits to be adjusted at once. The parameter modifies all magnitude limits that are set by hand (see `SURVEY_LIMITS` below). Usage: `MAGLIM = 0.0`
`SURVEY_LIMITS` ●	Array of 5 floats	Unbound	`SURVEY_LIMITS` are the photometric limits that define the 5 bands of the survey, i.e., $[u,g,r,i,z]$. The limits are all modified by `MAGLIM`. Usage: `SURVEY_LIMITS = [22.0,22.2,22.2,21.3,20.5]`
`SPLITSIZE` ●	Float	$0.0 - 9.9$ (*1d.p.*)	`SPLITSIZE` is the length/breadth in RA/Dec of an aperture-tile used in sections 3.2.2 to 3.2.5. This is the minimum size in degrees for RA/Dec for the side of a tile. Usage: `SPLITSIZE = 10.0`
`OVERLAP` ●	Float	$0.0 - 9.9$ (*1d.p.*)	`OVERLAP` is the extra length/breadth in RA/Dec of an aperture-tile used in sections 3.2.2 to 3.2.5. This is the amount of linear overlap, in degrees, each tile will have in a given dimension. *N.B.* `(OVERLAP + SPLITSIZE)` *represents the maximum size in degrees for RA/Dec for the side of a tile.* It is then used in sections 3.2.6 to 3.2.8 to define a new aperture-tile scale. Usage: `OVERLAP = 2.0`
`APERTURESIZE` ●	Array of Floats	$0.1 - 999.9$ (*1d.p.*)	`APERTURESIZE` is the set of radii of the APERC4-clustering apertures, in arcminutes. No limit is placed on the number of apertures that can be evaluated. Usage: `APERTURESIZE = [9.0, 5.0, 3.0, 2.0]`

Continued on next page

Table 3.1 – *Continued from previous page*

Parameter name (and origin)	**Variable type**	**Parameter values or range**	**Description**
`SIGMACOLOUR` ●	Float	$0.0 - 9.9$ (*1d.p.*)	`SIGMACOLOUR`, or $X_{\sigma_{xy}}$, is the number of σ_{xy} deviations (where xy is a colour described by wavebands x and y; Equation 3.1) in colour space that AperC4 uses to define the colour dimensions of the colour-aperture of the p-value calculation. Usage: `SIGMACOLOUR = 2.0`
`GAMMAFACTOR` ◐	Float	$0.0 - 9.9$ (*1d.p.*)	`GAMMAFACTOR`, or γ, describes a multiplicative factor on the systematic uncertainty in the colours of cluster galaxies, as in Equation 3.1. Usage: `GAMMAFACTOR = 0.0`
`FDR` ○	Float	$0.1 - 99.9$ (*1d.p.*; up to *3s.f.*)	`FDR` is the False Discovery Rate (§2.2.3); the maximum acceptable percentage of non-colour-clustered objects (i.e., field galaxies). Usage: `FDR = 10.0`
`K_IN_KNN` ○	Integer	<(no. of C4 galaxies)	`K_IN_KNN` defines k, where k is the neighbour out to which the $k-$NN procedure measures the distance. Usage: `K_IN_KNN = 6`
`XPZTHRESHOLD` ●	Float	$0.0 - 1.0$	`XPZTHRESHOLD` defines the minimum integrated cross-probability of two clusters occurring at the same redshift. Usage: `PZTHRESHOLD = 0.3`
`NGALLIM` ◐	Integer	$0 - 999$ (*3s.f.*)	`NGALLIM` is the $N_{\rm gals}$ cut-off: the minimum number of galaxies an AperC4 cluster must contain in order to keep it in the catalog. Usage: `NGALLIM = 8`

In addition to these parameters are a set of variables and flags, used for ApERC4 data handling, that *do not* affect the content of the output ApERC4 catalogs. These help ApERC4 index the runs by the parameters in Table 3.1, allow the user to observe the status of the ApERC4 software as it runs, utilise the directory structure of the computing system being operated to access different ApERC4 components, and store the output ApERC4 data for future analysis and/or repeated use. The parameters that help systematize and augment the running of the ApERC4 algorithm, but do not affect the output ApERC4 catalog, are given in Table 3.2.

Table 3.2: Descriptions of the set of parameters introduced to the new ApERC4 algorithm that affect the running of the algorithm and organisation of the data output(s).

Parameter name (and origin)	**Variable type**	**Parameter values**	**Description & example call**
`VERSION` ●	String	Unbound	`VERSION` is a basic description of the data being used, and is mainly used as a tag for ApERC4 to quickly find previous runs with a given parameter combination on this data. Usage: `VERSION = 'DR2'`
`RADECSCALE` ●	Float	$0.0 - 9.9$ (*1d.p.*)	`RADECSCALE` represents the number of grid divisions within a degree. Larger numbers result in faster running at the expense of greater memory use and vice versa for smaller numbers. This parameter is useful for system optimization (§3.1.4 & §3.2.3). Usage: `RADECSCALE = 6.0`
`PLOTOUT` ●	Integer (flag)	0 1	`PLOTOUT` is a flag that plots figures on-screen as ApERC4 is run: 0 suppresses plots; 1 shows plots. Usage: `PLOTOUT = 0`
`READOUT` ●	Integer (flag)	0 1	`READOUT` is a flag that makes the code more or less verbose, i.e., print more or less text during a ApERC4 run: 0 suppresses screen output; 1 shows more output. Usage: `READOUT = 0`
`INPUTDIR` ●	String	Unbound	`INPUTDIR` is the directory structure from which ApERC4 reads galaxy information. It also creates subdirectories to store catalog tiles. Usage: `INPUTDIR = '/path/to/galaxy_files/'`

Continued on next page

Table 3.2 – *Continued from previous page*

Name (and origin)	**Type**	**Values**	**Description & example call**
METADATADIR ●	String	Unbound	METADATADIR is the directory to which APERC4 saves metadata. Usage: `METADATADIR = '/path/to/metadata/'`
OUTPUTDIR ●	String	Unbound	OUTPUTDIR is the directory to which all reduced data is written. This is also the destination directory of the final APERC4 catalogs. Usage: `OUTPUTDIR = '/path/to/outputs/'`
C4PPDIR ●	String	Unbound	C4PPDIR is the directory that holds the APERC4-clustering (§3.2.3) C++ files of APERC4. The C++ program itself should be stored as <C4PPDIR>/Release/FillGrid. Usage: `C4PPDIR = '/path/to/program/'`
PREFIX ●	String	Unbound	PREFIX is simply a string that prefixes all output filenames to make runs distinctive. Usage: `PREFIX = 'First_DR8_Run'`
POSTFIX ●	String	Unbound	POSTFIX appends to filenames before the galaxies are assigned to cluster centres (§3.2.6) to make runs distinctive. Usage: `POSTFIX = 'June_2011'`
WRITESAVEFILE ●	Integer (flag)	0 1	WRITESAVEFILE states whether the information of a given run is to be saved to the metadata file: 0 doesn't save metadata; 1 saves metadata. This switch is included to prevent parallel runs failing due to different runs being unable to access and write to a file simultaneously. Usage: `WRITESAVEFILE = 0`

3.3 Discussion

3.3.1 p-value Calculation with Defined Apertures

$C4_{M05}$ calculates the p-values of galaxies belonging to the field in colour space (as described in §2.2.2) by setting volume apertures made by a $50\,h^{-1}$ Mpc cylinder length and kernel radius corresponding to $1\,h^{-1}$ Mpc from Equation 2.17. Both these dimensions are sensitive to assumed cosmologies, and Equation 2.17 is only consistent for low redshifts (such as $0.03 < z < 0.17$ used in M05). The key to both $C4_{M05}$ and APERC4 methods is that an

enhancement of galaxies in colour-space are indicative of a common stellar population, and hence, common evolutionary stage. APERC4 leverages the fact that redshift information is already embedded in the observed-frame SED, and so uses the observed-frame colours without constraining redshift.

Rather than invoke a cosmology, APERC4 is furnished with a set of apertures, whose radius is described by an angular separation. Because APERC4 is not given any redshift information, and hence never transforms galaxy positions into a 3-dimensional co-moving space, it is cosmology independent. However, there does need to be some decision making as to what apertures to give to APERC4. The aperture choices themselves can be *informed* by cosmology and typical cluster scales. Table 3.3 shows a single aperture can be thought of as probing different physical scales at different radii. The aperture selection in Table 3.3 can be seen to correlate to 1 Mpc scales at redshifts between $0.1 < z < 1$ when assuming a WMAP-9 cosmology ($H_0 = 70.0$, $\Omega_m = 0.279$, and $\Omega_\Lambda = 0.721$, Hinshaw et al., 2013).

Using other cosmologies affects the aperture radii in Table 3.3 such that the extremes of the redshift range may not get probed at specifically 1 Mpc, or simply transforms the alignment of the 1 Mpc elements but still at scales suited to galaxy clusters. For example, the aperture scales for $z = 0.05$ do not specifically probe 1 Mpc, but smaller radii that are still consistent with cluster scales (i.e., 0.53 Mpc) are implicitly included. In this way, despite using a cosmology to establish explicit apertures, the apertures themselves can be related to cluster radii for various redshifts in any number of cosmologies. By treating the algorithm as multiple trials of different apertures, multiple physical scales in the universe are tested. Note that the transverse physical distance of the aperture radius simply correlates to a different redshift if the underlying cosmology departs from that measured by WMAP, i.e., once the p-values for an aperture are calculated they can be reused for different cosmologies.

In the following subsection (§3.3.2), I will set out some thought experiments to reason how the aperture range is meant to capture low and high redshift clusters.

3.3.2 Aperture Gedanken

The cartoons in this section broadly describe the range of situations in which the APERC4 algorithm described in section 3.2 will encounter. I will first discuss the effect of the aperture size on a cluster, relative to its radial extent, then demonstrate the method's application to examples of clusters of different masses and redshifts.

The figures below show the effect of different aperture sizes placed on a galaxy cluster

Aperture Radius (arcmin)	9.0	6.3	5.0	3.7	3.0	2.7	2.5	2.2	2.0
z	Mpc scale (where $H_0 = 70.0$, $\Omega_M = 0.279$, $\Omega_\Lambda = 0.721$)								
0.05	0.53	0.37	0.29	0.22	0.18	0.16	0.15	0.13	0.12
0.10	1.00	0.70	0.55	0.41	0.33	0.30	0.28	0.24	0.22
0.15	1.42	0.99	0.79	0.58	0.47	0.42	0.39	0.35	0.31
0.20	1.79	1.25	0.99	0.73	0.60	0.54	0.50	0.44	0.40
0.30	2.42	1.69	1.34	0.99	0.81	0.72	0.67	0.59	0.54
0.40	2.92	2.04	1.62	1.20	0.97	0.88	0.81	0.71	0.65
0.50	3.32	2.32	1.84	1.37	1.11	1.00	0.92	0.81	0.74
0.60	3.64	2.55	2.02	1.50	1.21	1.09	1.01	0.89	0.81
0.70	3.90	2.73	2.16	1.60	1.30	1.17	1.08	0.95	0.87
0.80	4.10	2.87	2.28	1.68	1.37	1.23	1.14	1.00	0.91
0.90	4.26	2.98	2.37	1.75	1.42	1.28	1.18	1.04	0.95
1.00	4.38	3.07	2.43	1.80	1.46	1.31	1.22	1.07	0.97

Table 3.3: Assuming a WMAP-9 cosmology, this Table gives a matrix of Mpc scales that correspond to the aperture radii (in arcminutes; top row) at various redshifts (left column) assuming the cosmology given in the second row of the table. Highlighted in red are scales within 5% of 1 Mpc.

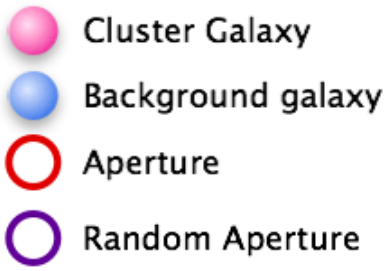


Figure 3.4: Key for gedanken figures

using the symbols described in the key (Figure 3.4). A background galaxy is one that is both in the projected plane, and in the APERC4 colour-aperture, but is not bound to

the cluster. They are pictured here as discrete entities for illustrative purposes only. In AperC4 proper, and assuming the colours are somewhat unique to cluster environments and redshift, multiple background galaxies would contribute small fractional probabilities towards entering the same 4-D colour space as the cluster galaxies. With significantly more galaxies in the RA-dec aperture, without colour space enhancement clusters may be difficult to properly identify, i.e., line-of-sight contamination.

When using an aperture that is too large for the cluster the significance of the cluster is more likely to get washed out by background contribution to the neighbour count in colour-aperture space, as shown in Figure 3.5

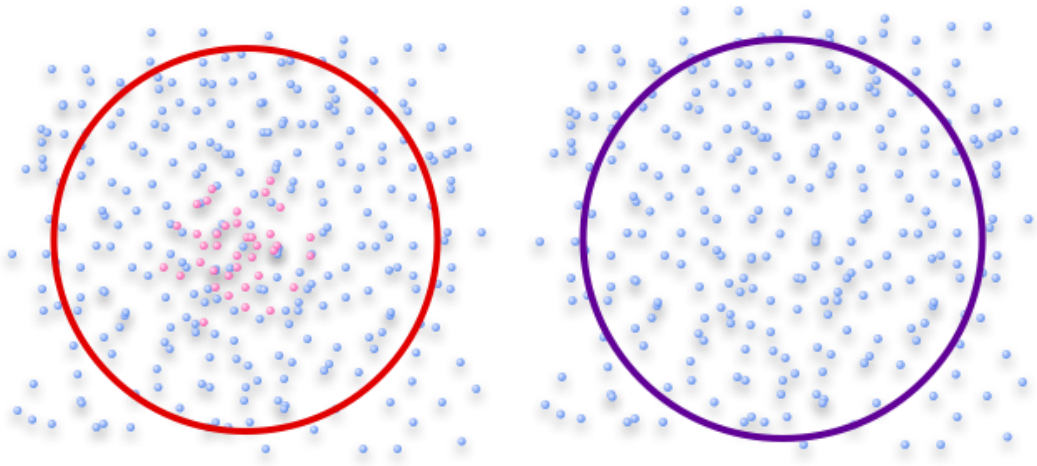


Figure 3.5: This cartoon shows an aperture too large for a cluster. The Figure shows an aperture centred on a central cluster galaxy (*left*) and on a random field galaxy (*right*). We see that the aperture is dominated by background noise, so the p-value will indicate the cluster galaxy is more field-like than it is in reality.

Figure 3.6 shows that when the aperture approaches the size of the cluster, then the contribution of the cluster galaxies to the neighbour count becomes stronger compared to the random/model aperture.

For yet smaller apertures, Figure 3.7 illustrates there should still be an enhancement in the signal from cluster centres, but clearly this falls as the aperture shrinks to below the scale of the cluster concentration.

For apertures well below the scale of the cluster, it can be seen that galaxies at the centre of the cluster may still display an enhancement against the field (Figure 3.8). However, outside the central regions of the cluster, the signal (detected cluster galaxies) is confused with noise (detected background galaxies) and so the true 'size' of the cluster will be underestimated.

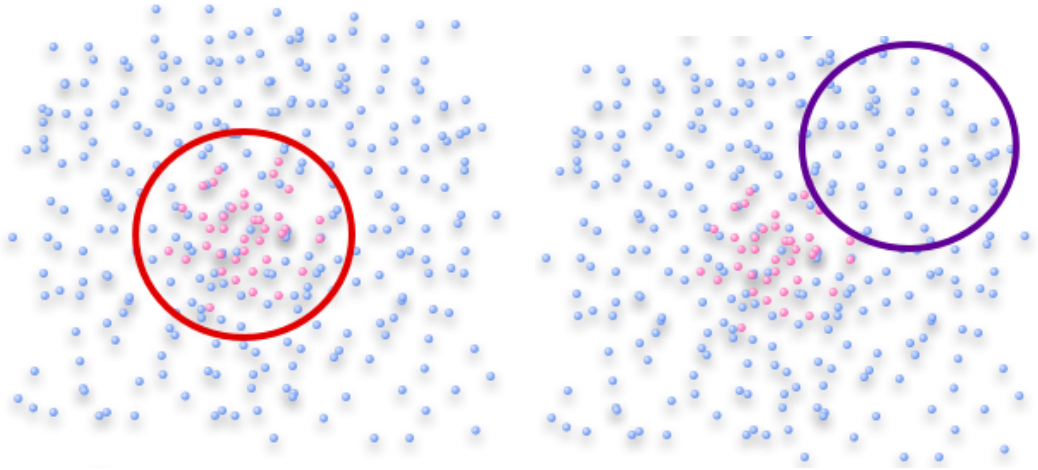


Figure 3.6: This cartoon describes an aperture approaching size of the cluster. The Figure shows an aperture centred on a cluster galaxy (*left*) and on a random field galaxy (*right*). Here the enhancement of cluster galaxies is greater than the field, hence the galaxy that the aperture is centred on is less likely to be drawn from the field population.

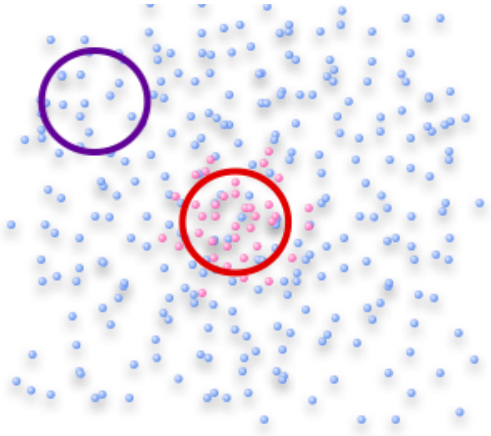


Figure 3.7: This cartoon shows an aperture that is slightly smaller than the cluster extent. The Figure shows an aperture centred on a cluster galaxy (red aperture) and on a random field galaxy (purple aperture). The enhancement of the central cluster galaxies can now clearly be seen within an aperture centred on the cluster.

Large Mass/Radius - Low Redshift Cluster

Large radius, low redshift clusters are dependent on the maximum aperture size applied. The choice of largest aperture should consider the largest low redshift cluster we wish to consider (e.g. Figure 3.7), however, since the aperture is larger so is the background signal, meaning the colour-clustering signal may be diluted. AperC4 may still detect cluster members close to the core as per Figure 3.8, but the ability to identify members on the periphery will be reduced accordingly.

Furthermore, if an aperture centres on a high redshift supercluster, this may also cause an enhancement to the signal, meaning identification of cluster members at low z encounter

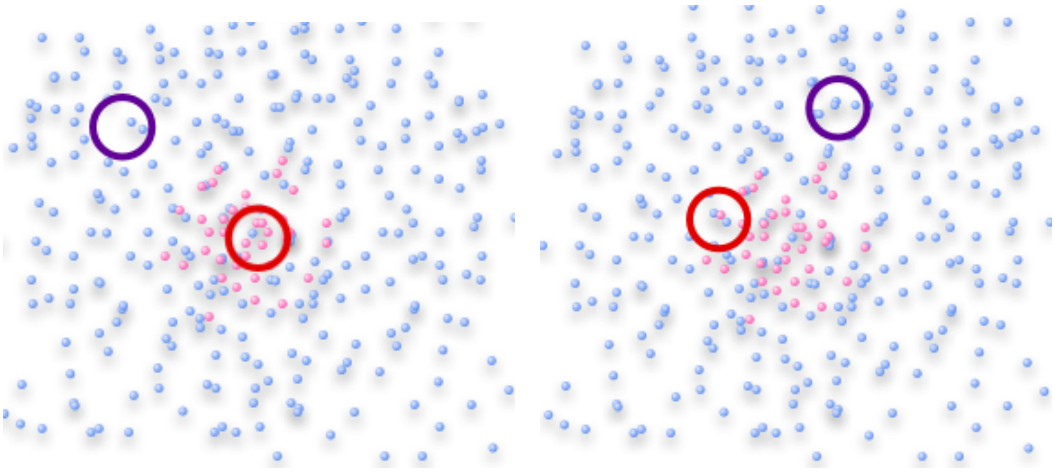


Figure 3.8: Aperture approaching size of the cluster. Figure show an aperture centred on a cluster (red aperture) and on a random field purple aperture). Where the aperture is placed on the cluster centre (*left*) the cluster galaxy enhancement is somewhat greater than the field whilst cluster galaxies on the cluster periphery (*right*) are indistinguishable from the field.

more issues than small aperture high z cluster member detections (e.g. reconsidering the random apertures in Figure 3.8 as centering on genuine clusters at higher redshift). This may be a non-issue as superclusters at high-z are a rarity (Einasto et al., 2014).

Small Mass/Radius - Low Redshift Cluster

At low redshift, small clusters can be found easily since the apertures that are explored should include smaller apertures to find high redshift clusters. When employing an aperture that is too large, the detection is equivalent to washing out the signal, as per Figure 3.5. Since the high redshift apertures will explore smaller aperture sizes, then the full range of scales are explored for these clusters, and so the detection is maximised between Figures 3.6 and 3.7. For apertures significantly smaller than the low redshift cluster then, as per Figure 3.8, the detection is not made.

Finding small clusters at low redshift in the ApERC4 framework depends very much on the aperture used being the right angular size, based on the clusters' true extent and redshift. ApERC4 will tend towards washing out the signal (Figure 3.5) when large apertures are applied as the cluster is of intrinsically small radius. ApERC4 recovers these clusters by testing over multiple aperture scales, and so signal-to-noise (colour-clustered cluster galaxies to colour-clustered background galaxies) maximises at the scale most appropriate to the cluster. If that signal-to-noise is deemed significant in an aperture (by FDR), then it will persist into the ApERC4 galaxy catalog.

Large Mass/Radius - High Redshift Cluster

High mass/radius high-z clusters are detected in the same as a low redshift small radius clusters described above.

Small Mass/Radius - High Redshift Cluster.

The small radius, high redshift clusters are dependent on the *minimum* aperture size applied. The choice here is dependent on the smallest aperture we decide to apply. High z clusters with small radial extents will encounter problems with the aperture being too large (Figure 3.5) most of the time, and so will only be detected as the area of the aperture approaches the size of the cluster (Figure 3.6). A concern regarding the all-z approach here is that the low-z groups (which can also be colour clustered) also count as enhancements against the background. So in these regimes we pick up both high-z galaxies in clusters and low-z galaxies in groups (Note that since low-z group and high-z cluster galaxies will have different colour space occupation due to different SEDs, different redshifts, etc, *both* will be included as AperC4 galaxies).

3.3.3 $p(z)$ Blending and Fragmentation

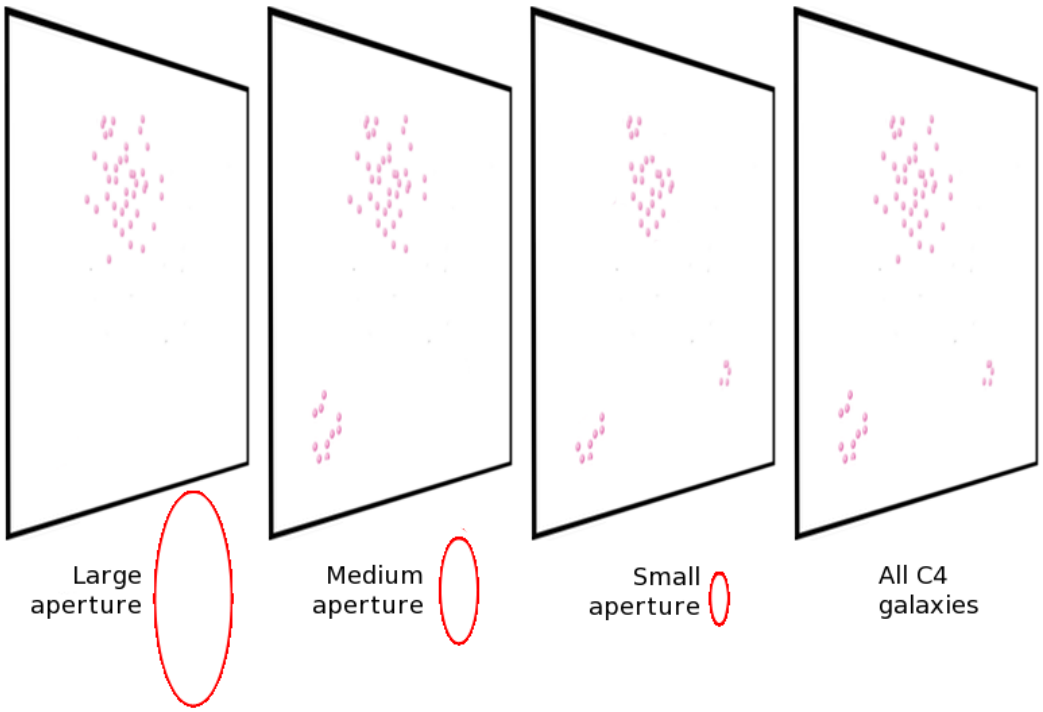


Figure 3.9: This figure shows an example of the AperC4 catalogs produced for three apertures, here pictured as large, medium and small angular radii (*far left, centre-left, and centre-right*, respectively). The final AperC4 cluster catalog (*far right*) should be a combination of these separate aperture-slice catalogs, where their galaxy members are not duplicated between clusters.

Using the cartoon pictured in Figure 3.9, one can see the large and medium aperture sizes identify the large concentration of galaxies at the top of the tile, whilst the small aperture identifies two apparent concentrations. By testing the $p(z_{\text{galaxy}})$ of the cluster members in the small aperture catalog against those of the medium/large apertures, one can see whether the concentrations are separated in redshift or are part of the same cluster. Examining the cluster in the lower left of the tile, it is too small for the largest aperture to find effectively, whilst the smallest aperture only recovers a fraction of the members. The smallest aperture identifies small clusters that neither the medium or large apertures can find, as can be seen by the identification of the cluster on the bottom right of the tile in Figure 3.9.

By giving all the clusters in each aperture-slice a preliminary $p(z)$ from the product-sum of the member $p(z)$ (Equation 3.3), one can test whether duplicate galaxies are best suited to a cluster by evaluating the product of the galaxy $p(z)$ and each aperture-slice cluster $p(z)$ to which the galaxy belongs (Equation 3.4).

Figure 3.10 assigns some imagined redshifts for the galaxies identified in Figure 3.9. If one takes the large group at the top of the figure, which consists of both high and low redshift members (red and green points, respectively), one can imagine its coalescence into a single cluster using the large aperture would give a $p(z)$ distribution strongly dominated by the $z = 0.1$ component, with some small enhancement at $z = 0.7$. But the same group is split into two pieces in the smallest aperture, with each component ($z = 0.1$ and $z = 0.7$) given a separate $p(z)$ that is more indicative of each cluster's redshift. When comparing the products of each high redshift galaxy $p(z)$ with the aperture-slice cluster $p(z)$s, the shared members in the high redshift component (red) will look more like the separate group found with the small aperture, and so the two radially blended clusters are given separate identifications.

As a counter-example, the gold galaxies in Figure 3.10 at $z = 0.3$ are all part of the same cluster, and the medium-sized aperture has found all the members, whilst the smallest aperture has only identified a single concentrated piece. As both cluster $p(z)$s will indicate the same redshift, the galaxies will be assigned according to slight deviations in their $p(z)$ distribution. Having "exchanged" members, AperC4 tests the $p(z)$s of the clusters in the different apertures against each other, and if they are found to agree above some probabilistic threshold (`PZTHRESHOLD`), then they are merged.

By utilising the differently identified candidate clusters (from the various aperture-slices) as initial models for the distribution of the galaxies they identify, and then applying

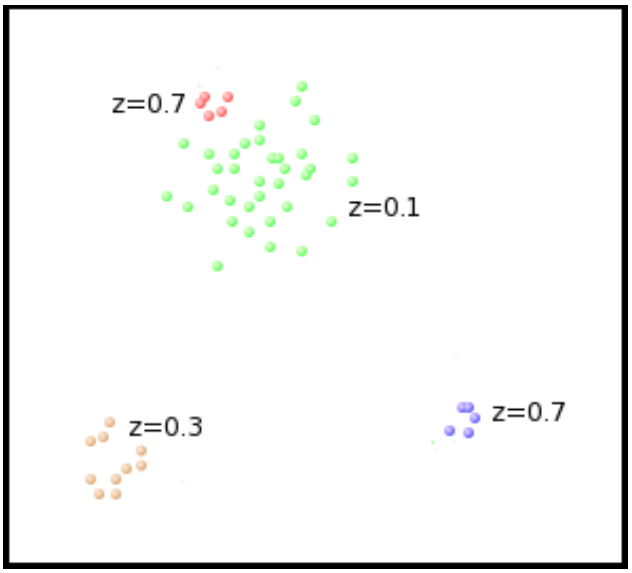


Figure 3.10: Figure shows an example of the combined APERC4 catalog produced by uniquely assigning the galaxies found in multiple apertures (seen in Figure 3.9) to their correct cluster. Galaxies can be assigned to their cluster by employing $p(z)$ information (§3.2.7 and §3.2.8).

$p(z)$ information, line-of-sight blending and radial fragmentation issues can be curbed.

3.3.4 Comparison of AperC4 and C4$_{\mathbf{M05}}$ Algorithms

As discussed in section 3.2, APERC4 introduces a large number of new parameters for cluster finding, seen in Table 3.1, as well as for data-handling and organisation, seen in Table 3.2. Additionally, the functionality of C4$_{\mathrm{M05}}$ has changed in APERC4 as a result of allowing some of these new parameters to be varied, and in the methodology of APERC4 itself.

Grid Treatment

APERC4's adaptive grid creation contrasts with the C4$_{\mathrm{M05}}$ algorithm where the grid size is hardcoded into the algorithm such that the program needs to be edited and recompiled for different datasets to account for the maximum number of galaxies out of all the grid bins, and the minimum and maximum R.A., dec., and redshift. This made it difficult to process alternative/future data sets or subsamples without having to reprogram specific lines of code and recompile the program each time.

In writing the APERC4 code, I incorporated the C++ BOOST library, allowing flexibility in assigning both the R.A. and dec. limits of the grid, which in turn then allow the dynamic creation of a grid structure to process any incoming galaxy catalog efficiently and without requiring any prior knowledge of the survey's limits. By being adaptable,

APERC4 uses memory resources more efficiently than $C4_{M05}$ when creating the grid. In removing the redshift dimension, APERC4 is no longer dependent on an input cosmology to determine colour-clustering strength, and hence grid creation, and assignment of galaxies to the grid, are accelerated. Additionally, the `RADECSCALE` parameter is very useful in accelerating the algorithm in computing environments where memory (RAM) is more available. The grid is also no longer fixed into subdivisions of $1\,\square^\circ$, and allows the extent of the aperture to cover more than two grid spaces, as opposed to the $C4_{M05}$ which simply sought out its neighbouring $1\,\square^\circ$ grid element.

Some more subtle effects can occur as a result of survey tiling in APERC4, such as when `SPLITSIZE` and `OVERLAP` become too small. If both `SPLITSIZE` and `OVERLAP` are too small, then the 2-dimensional (R.A. and dec.) area covered by a survey tile will not contain a representative sample of galaxies to be able to fairly evaluate APERC4-clustering at a given location, e.g., if a tile were the angular size of a cluster, then tiles without clusters would give some fraction of galaxies a low likelihood of belonging to the field (i.e., a low p-value) and tiles that do contain clusters would give a fraction of genuine cluster galaxies a high likelihood of belonging to the field. If `OVERLAP` is too small, then the areas where tiles overlap would suffer from edge effects, and the galaxies along these tile edges appear to have a high likelihood of belonging to the field, as a part of their volume aperture becomes empty. $C4_{M05}$ avoids this by ingesting the whole survey at once. By tiling, APERC4 allows the p-value calculation to be parallelised, also reducing the memory resources required by any single CPU when processing the survey area.

p-value Calculation

The front end processes that lead to the determination of the galaxies' p-values, i.e., APERC4-clustering, are no longer single estimations of the 4-dimensional colour space enhancements of galaxies in a 3-dimensional space (C4-clustering) but multiple estimations of the same enhancements in 2-dimensional spaces at multiple scales.

For each galaxy (target galaxy), APERC4 creates a colour-area aperture, as opposed to the colour-volume aperture created in $C4_{M05}$. One other difference between the $C4_{M05}$ and APERC4 algorithms' treatments of neighbour search radii (radius of the $C4_{M05}$-/APERC4-clustering volume/area aperture; §3.3.1) is how they deal with apertures that extend beyond the grid point containing the galaxy whose neighbours are being determined. In $C4_{M05}$, the algorithm is allowed to extend to one neighbouring grid location in R.A. or dec. (but not both) during its search. In APERC4, the number of neighbouring grid locations

available to search is limited only to the grid locations that at least partially lie inside the aperture used to evaluate APERC4-clustering. I argue here, that in allowing grid points to be accessed in this way allows APERC4 to treat each galaxy correctly, whilst galaxies near the corner of a grid point in $C4_{M05}$ may have a suppressed C4-clustering measurement due to their inability to search a diagonally neighbouring grid location.

The effective redshift extent of the 'volume' aperture in APERC4 is the full redshift extent of the survey, whilst it is bins of $50\,h\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ in $C4_{M05}$. M05 investigated the effect of redshift bin size and found the measured C4-clustering to be independent of this length. But by having an explicit angular aperture size, and no redshift limitation, no calculation of comoving separation is involved in APERC4, i.e., the explicit aperture radius replaces the $C4_{M05}$ kernel radius calculated in Equation 2.17, and thus no cosmology is invoked. However, now all the galaxies live on an RA-dec plane, signal-to-noise in APERC4 is far below that of $C4_{M05}$. Also, since apertures corresponding to low redshift are large, this results in a lot of computation for galaxies whose real location is at high redshift. Indeed, all galaxies will have p-values calculated for each aperture size.

The introduction of `SIGMACOLOUR`, $X_{\sigma_{xy}}$, and its relationship to the `GAMMAFACTOR` parameter, γ, in section 3.2.3 has been changed such that the statistical component of the colour uncertainty is not effectively ignored when rescaling the colour aperture, contrary to the method of M05 who only scale the systematic component. The modification of the colour box from Equation 2.18 to 3.1 removes the implicit assumption of M05 that the statistical error will always be dominated by the systematic component, but still retains the facility to tune the colour box size with γ. The `GAMMAFACTOR` component, γ, is retained in the algorithm, to allow for adjustments to the colour box based on the dispersion of galaxy magnitudes/colours across the survey area. `GAMMAFACTOR` still influences the size of the colour box jointly with `SIGMACOLOUR`, $X_{\sigma_{xy}}$. In $C4_{M05}$, this permitted the systematic component of Equation 3.1 to dominate the statistical component when the magnitude uncertainties exceeded the typical size of the CMR. However, since APERC4 decomposes the survey area into local tilings, and the introduction of übercalibrated magnitudes, which greatly reduce survey-wide photometric biases, one expects the systematic term to be greatly diminished.

The new APERC4 parameters are explicitly stated so that multiple tests can be made to evaluate their impact on cluster finding. Clearly parameters like `MAGLIM` and `SURVEY_LIMITS` affect the number of galaxies seen and the strength of the APERC4-clustering in euclidian space (assuming faint galaxies are distributed amongst bright galax-

ies) and colour space (assuming the uncertainty of a galaxy magnitude increases as the galaxy magnitude becomes fainter). Parameters like `APERTURESIZE` (§3.2.3) directly limit the number of galaxies being evaluated in the colour-aperture, and so if they are too large, the difference between APERC4-clustered and field objects may diminish due to the presence of APERC4 cluster galaxies somewhere in the aperture, and if they are too small, the magnitude of the difference may become too small to distinguish between cluster and field galaxy neighbour counts (§3.3.2).

kth Nearest Negihbour

A different $k-$NN program is used (to the $C4_{M05}$ $k-$NN method, section 2.2.4), to assess the $k-$NN clustering of the APERC4 galaxies (§3.2.5). The change in $k-$NN program is due to the particular software having become inoperable under newer operating systems.

Forming Clusters

`APERTURESIZE` is also reused in the assignment of galaxies to clusters. The $C4_{M05}$ number density threshold (2.2.5) is no longer involved in the assignment of galaxies to clusters to *a*) facilitate assignment of galaxies to clusters when employing photometric redshifts at a later stage, since galaxies in the cluster may be scattered out of the cluster due to the increased uncertainty in redshift if including photo-z prior to assignment, and *b*) account for additional peculiar velocity components that affect cluster galaxy redshifts along the line-of-sight. The reason for the exclusion of the density threshold is because the extended redshift interval over which APERC4 performs (compared to $C4_{M05}$) introduces competing observational effects. At higher redshifts, fewer faint cluster members are likely to be detected in a given survey, and so the 2-D membership count of a cluster at high redshift will be lower than a contemporary cluster at low redshift. Conversely, the angular separation between galaxies in in a cluster corresponds to their physical separation and cosmology of the universe between the observer and the redshift of the cluster (as can be seen in Table 3.3). If one compares a hypothetical high redshift (e.g., $z \sim 1$) cluster to an identical cluster at low redshift (e.g., $z \sim 0.1$), and all galaxies in both cases are observed, then the surface density of the high redshift cluster would appear larger due to the smaller angle subtended. But (keeping with this hypothetical comparison), it is unlikely that as many galaxies from the high redshift cluster would be seen as the low redshift case, due to the decrease in observed flux from the cluster members falling below the detection limit of an observational campaign, thus the surface density would decrease for higher redshift

clusters.

As no redshift information is included before determining cluster centres (§3.2.5), in the interest of simplicity whilst acknowledging that there will indeed be some confusion in cluster centering of low redshift objects, the $k-$NN distances are treated in 2-D (compared to 3-D in $C4_{M05}$). Using R.A. and dec, lower redshift clusters are less radially concentrated than identical clusters at high redshift (Table 3.3) shows the relationship of radial distance with physical separation at various redshifts for an assumed cosmology) and as such their nearest neighbour distances are larger, which may bias assignment of these galaxies to higher redshift clusters along or near the line of sight versus a lower redshift cluster source. Limiting the $k-$NN to angular distances may also result in misplacement of cluster centres at low redshifts, which are radially biased by high redshift clusters and large scale structure near, or along, the line-of-sight. Similarly, APERC4 employs a 2-dimensional assignment algorithm to place galaxies into clusters as opposed to the spherical assignment method of $C4_{M05}$.

Alternative formulations for cluster centre identification were tested. The redshift dimension may be estimated by photometric redshift, which has greater uncertainty than spectroscopic redshifts. Increasing the dimensionality to include colours was also trialled, but without some meaningful metric to convey power between angular scales and magnitude scales it is difficult to assess the degree to which neighbour "distances" decomposed radially or in colour space.

Richness Thresholds

Parameters like `NGALLIM` have more complicated, but very definite, effects on output APERC4 catalogs, where raising these values generally reduces the total number of clusters found by the algorithm, increases the fraction of real clusters, and incurs some blending or fragmentation in the output cluster sample. This parameter has also changed slightly from its $C4_{M05}$ definition of being the total number of galaxies (both C4 galaxies and "background" galaxies) within a volume defined by $C4_{M05}$ galaxies. The APERC4 N_{gals} defines the number of APERC4 galaxies in a cluster. The APERC4 N_{gals} cut-off has changed in character in anticipation of photometric data.

The choice of cutting out clusters at populations lower than 8 members is applied in M05 to eliminate chance superpositions whilst maximising completeness/purity. The APERC4 definition differs slightly from the original C4 implementation, where the number of galaxies (not necessarily C4 galaxies) within $1\,h^{-1}$ Mpc is treated as the N_{gals} threshold,

and $> 10\%$ of all the galaxies in the cluster volume must be C4 galaxies. This would not work as effectively in a photometric context, since the aperture may contain a reduced number of cluster galaxies compared to its spectroscopic counterpart, and an increased number of non-cluster galaxies, both of which are more likely to have scattered in/out of the colour-aperture due to the effectively unlimited redshift extent of the initial AperC4 cluster definition.

Moreover, uncertainties in source classification and optical detection limit become increasingly problematic in photometric data, causing cluster volumes to become populated with non-cluster objects. For the same reasons, calculating the 'significance' of AperC4 clusters, as defined by the fraction of C4 galaxies to all galaxies within the cluster volume (as per $C4_{M05}$; §2.2.5), is deemed unhelpful when photometric uncertainty is introduced. Additionally, with the absence of fiber collisions limiting the source density of the survey area in photometric data, the galaxy number count will be far greater than in the DR2 spectroscopic data. The validity of a significance cut at such low thresholds becomes highly questionable, so AperC4 rejects this threshold. Instead, the $N_{\rm gals}$ cut-off is redefined as the minimum total AperC4 galaxy population of a cluster and is employed as the quality threshold for clusters in the AperC4 catalog.

In summary, the interaction between parameters and their effect on the output AperC4 catalogs can be used to reveal nuances in the data and the AperC4 algorithm, itself. Understanding the individual components of the AperC4 algorithm can help one interpret the effect(s) on an output cluster catalog observed by changing parameter values.

3.4 Summary

In summary, AperC4 represents a reformulation and update of the original $C4_{M05}$ code and algorithm. As opposed to the majority of modern day cluster finders (including $C4_{M05}$) it does not invoke a cosmology. AperC4 introduces a large degree of flexibility and adaptability to the cluster finding process and affords a greater insight into, and control of, the selection function of the AperC4 algorithm.

In this chapter, I have introduced a new algorithm, AperC4, derived from the M05 algorithm. In section 3.1, I discussed the motivations for AperC4 and its key features.

In section 3.2, I described the new algorithm, summarising the associated parameters in section 3.2.10. In section 3.3, I presented some examples to explain how AperC4 works. I also discussed how the AperC4 apertures and other parameters affect cluster finding,

and contrast their function in AperC4 to their use in $C4_{M05}$.

Having introduced the AperC4 algorithm, I discuss how to verify an output cluster catalog in chapter 5 and deliver an optimal cluster catalog. In chapter 6, I use DES galaxy catalog simulations, which I will introduce in chapter 4, to test the cluster finding ability of AperC4 with simulated photometric data and further investigate the sensitivity of parameters to cluster finding. In chapter 7, I process the SDSS DR8 galaxy catalog and produce an AperC4-SDSS DR8 cluster catalog.

Chapter 4

Introduction to DES and CatSim

In this chapter, I introduce the Dark Energy Survey, highlighting the cosmological probes that will use DES data to describe the universe in section 4.2. In preparation for the DES, a suite of simulations was assembled to test cosmological tools to optimise them for survey data. These simulations are introduced in section 4.3.

I note that this chapter is a summary of the DES and the DES collaboration aims and simulated products as they stand at the time of writing.

4.1 The Dark Energy Survey (DES)

4.1.1 Overview

The Dark Energy Survey is an optical to near-infrared survey proposed to cover 5000 $\square^\circ$ of the South Galactic Cap to ∼24th magnitude in $grizY$ bands (Tucker et al., 2007). This area and depth covers a larger cosmological volume than the SDSS, reaching redshifts of around $z \sim 1.3$. This huge increase on the previous generation of optical surveys will give photometric and position measurements of ∼300 million galaxies and is predicted to detect $\gtrsim$100,000 galaxy clusters. The large depth and area of the survey will allow weak lensing measurements to probe the matter distribution of the universe. The huge data output (∼2 petabytes) has demanded a concerted data management effort to collect and redistribute data collected from the DES Telescope to the DES Collaboration (Mohr et al., 2012). The DES Survey will employ the 4 metre Victor M. Blanco Telescope located in Cerro Tololo, Chile, and has been given 525 nights over 5 years to complete its campaign goals.

In exchange for this amount of Telescope time, the DES Collaboration built a new CCD instrument, DECam (§4.1.2), to replace the existing wide-field MOSAIC CCD camera that

was already in place. In addition to the construction of a new instrument and photometric filter set, the DES Collaboration invested a significant amount of time and resources into upgrading the telescope itself. A non-exhaustive list of these modifications includes:

- upgrading the Telescope Control System, nominally to improve aging systems, but necessary to increase the speed of repositioning of the telescope so that it can be pointed 2° from any starting position within $\sim$ 17 seconds;the readout time of the camera. This ensures observations are not delayed by the slewing of the telescope.

- upgrading the readout electronics attached to the CCD cameras from the $\sim$ 100 second electronics previously in place to sub-20 seconds.

- implementing environmental control in the telescope dome, so the observing quality through the night is less affected by temperature fluctuations in the dome inducing relative offsets between the detector and the telescope components.

- cooling the CCD chips themselves, providing a stable temperature background so that the CCDs perform efficiently.

- a new wide field corrector, to improve upon the previous corrector which was deemed to be non-optimally configured (Doel et al., 2008).

DECam achieved first light in September 2012. There followed several months of Science Verification (both of the camera and of DES operations), through to February 2013. The DES proper started in September 2013 and completed its first season of operations in February 2014. The second (of five) DES season started in August 2014 and runs through to the end of January 2015. At the time of writing, only DES science verification (SV) data is available to the collaboration. The SV data products were not finalised until well into 2014 and so arrived too late for consideration in this thesis.

4.1.2 DECam

To achieve these aims, the DES collaboration commissioned, built and mounted a 519 megapixel optical/near-infrared CCD camera with wide field corrector, named DECam (Flaugher et al., 2010), mounted onto the prime focus cage of the 4 metre Victor M. Blanco Telescope (Blanco, hereafter). The individual CCD pixels have an angular scale of $0\farcs26$ per pixel, equivalent to the previous MOSAIC CCDs, but covering a $2.2\,\square^\circ$ Field of View (FoV), a huge increase over the previous $0.36\,\square^\circ$ FoV.

DECam features guide and focus sensors on the focal plane (Diehl et al., 2008) allowing greater sensitivity to the precise telescope orientation and direction necessary for accurate astrometry. This is essential for weak lensing measurements, as varying image distortions throughout an observation may affect the weak lensing signal in ways that would be difficult to assess, weakening the output cosmological constraints.

Cryogenic cooling of the CCDs to 180K ensures low noise CCD readout, which, in observational terms, will mean real objects detected in the DECam images can be measured to greater significance (high signal-to-noise), and fainter objects can be observed with reduced uncertainty in their photometry.

Upon first light, the DECam achieved produced high quality images with a seeing of 1.2 arcseconds within hours of being turned on. A week on, calibration of the multiple new pieces of equipment had produced images with a seeing of 0.8 arcseconds, significantly improving upon previous typical seeing at Blanco (1.39 and 1.23 arcsec in the r and z-bands, respectively, Mehrtens et al., 2011).

4.2 Cosmology with DES

The intended goal of the Dark Energy Survey is to constrain the late time acceleration of the expansion of the universe. Four complementary optical probes that approach the analysis of optical distribution of matter are proposed for this purpose: weak lensing, galaxy clustering, type Ia supernovae, and clusters of galaxies. In the following sub-sections, I briefly overview the way they address the dark energy that drives the accelerating cosmic expansion.

4.2.1 Weak Lensing

Weak lensing concerns the distortion of the appearance of distant galaxies by the gravitational effect of matter, close to the line-of-sight, bending the path of the background light. This gravitational lensing of matter on fields of distant galaxies is called *cosmic shear*, and is the key measurement of weak lensing. As touched upon in section 1.4, one of the defining features of weak lensing is that it probes both baryonic and non-baryonic matter components, and is robust to both the physical and dynamic state of the matter.

The statistical properties of cosmic shear measured as a function of photometric redshift will be sensitive to expansion rate at that redshift, informing the growth rate and distance-redshift relation. Cosmic shear is a relatively small signal, measured at the 1-2% level (Bacon et al., 2000), and is sensitive to the intrinsic alignment of galaxies in cluster

size potentials (Smargon et al., 2011), but the accompaniment of photometric redshifts to the weak lensing signal vastly reduces the negative impact of the intrinsic alignments (Heymans and Heavens, 2003). With the large area and observation of a huge number (300 million) of galaxies with photometric redshifts, DES weak lensing will reduce uncertainties in weak lensing cosmological parameter constraints from a factor 2 to around $\sim 30\%$ (Huterer et al., 2006).

4.2.2 Galaxy Clustering

Galaxy clustering employs galaxies as a tracer for the distribution of matter in the large-scale structure of the universe. Of particular interest is the feature in galaxy clustering known as the Baryon Acoustic Oscillation (BAO) peak, in the distribution of the galaxies. Briefly, before the universe became transparent, sound waves initiated by instabilities in overdensities of the primordial universe traversed through the unified photon-baryon fluid outward in spherical shells. After some time, the expansion of the universe cools to a point where the photons and baryons decouple and the photons free stream away (forming the observed CMB), whilst the baryons (no longer supported by the pressure of the photon-baryon fluid) undergo gravitational collapse in a shell at a fixed radius from the central overdensity that initiated the sound wave. Given the time for these sound waves to occur, harmonics in the photon-baryon fluid would have persisted up until decoupling.

Galaxy clustering studies (Eisenstein et al., 2005; Percival et al., 2007) show that there is a statistically preferred separation between galaxies at scales of $\sim 100\,h^{-1}\,\mathrm{Mpc}$, that signify the frozen out sound waves, or acoustic oscillations, of the baryons at decoupling. Since this harmonic scale is characteristic of the photon-baryon fluid, observed in the CMB at decoupling, measurements of the acoustic peak in the local galaxy distribution can act as a cosmological yardstick for measuring the acceleration and expansion of the Universe. Combining the DES data with VISTA (Visible and Infrared Survey Telescope for Astronomy; also located in Chile) observations in the JHK bands, will provide galaxy power spectra out to $z \simeq 2$ with uncertainties of order 20% (Banerji et al., 2008).

4.2.3 Type Ia Supernovae

Measurements of type Ia supernovae (SNIa) hinge upon their reliability as standard candles. SNIa are thought to be carbon-oxygen (CO) white dwarf stars that accrete matter from a companion star until they reach the *Chandrasekhar mass limit* of $\sim 1.4\,\mathrm{M}_\odot$. At this point, the electron degeneracy pressure supporting the white dwarf is exceeded and

the white dwarf undergoes collapse, igniting carbon fusion in a runaway thermonuclear explosion that results in its annihilation. This model of an accreting white dwarf means that the mass of the SNIa progenitor will nearly always be close to the Chandrasekhar limit, and so the resulting supernova will appear the same for any such CO white dwarf and companion star system. This physical picture motivates their use as "*standard candles*."

Observations of SNIa have determined a relationship between the shape of the light curve and its peak luminosity, enabling their brightness to be used as a relative distance measurement, determining the deceleration parameter, q_0, and hence show the evolution of the Universe's expansion. The DES Supernova Survey expects to measure ∼1900 SN light curves (and followup spectroscopy of the host galaxies, though at another telescope facility) in 40 $\square^\circ$. For comparison, the work of Riess et al. (1998), which made the first observation of an accelerating universe with Type Ia supernovae that was awarded the 2011 Nobel Prize, employed a total of 50 supernovae. Simulations from Bernstein et al. (2009) indicate DES will deliver dark energy constraints a factor of 4.6 times better than contemporary surveys.

4.2.4 Clusters of Galaxies

As covered in section 1.3, DES galaxy clusters will measure cosmology through finding their abundances at various redshifts. Changes in the abundance and rate of formation of clusters depends on nature of dark energy (Frieman et al., 2008). 1000 $\square^\circ$ of the survey area coincide with the South Pole Telescope (SPT) Sunyaev-Zel'dovich Survey (Reichardt et al., 2012) region, which will improve upon independent cosmological measurements of galaxy clusters by a factor of two by allowing cross-calibration of mass measurements of clusters (§1.4) discovered in the DES (Cunha et al., 2009).

One of the great strengths of cluster cosmology is the multiple ways that cluster mass can be probed. For example, weak lensing of clusters, *cluster shear*, produces a stronger, more readily detectable, lensing signal (as demonstrated by the bullet cluster, Figure 1.1; §1.2.2). The complementarity of cluster measurements with other probes should allow observations of DES clusters help build concordance between X-ray, weak lensing, and SZ measurements (Rozo et al., 2012).

4.3 DES Catalog Simulations

4.3.1 Introduction to CatSim

The DES simulations working group have produced a series of image and catalog simulations (Lin et al., 2010) to permit the development of data reduction pipelines (Mohr et al., 2012) and science analysis codes (Soares-Santos et al., 2010; Cunha et al., 2012) in advance of DES first light. These simulations have been useful in making analytical forecasts of cosmological parameter constraints (Tinker et al., 2012) and quantifying necessary followup observation campaigns that will help improve upon these parameter constraints (Wu et al., 2010) in a DES-like survey context.

In chapter 6, I use the SDSS catalog simulations, CatSim, to evaluate AperC4 cluster finding. The SDSS catalog simulations were produced by constructing galaxy catalogs that resembled real data to presently observable limits. The simulated observations were extended to the DES specifications, but the SDSS version will be the sole focus in this thesis.

4.3.2 Simulation Construction

Generically, a galaxy catalog is constructed around an existing N-body simulation, e.g., Hubble Volume Simulation (Evrard et al., 2002), Millennium Simulation (Springel et al., 2005), MICE Simulation (Fosalba et al., 2008), etc. These N-body simulations work by populating a Gpc-scale volume with 'dark matter particles' of some unit mass, or mass resolution; the scale and mass resolution of these simulation "boxes" being limited to the computing resources used to produce them (e.g., TeraGrid, Catlett, 2005). This volume is assigned some cosmological parameters and the dark matter particles are distributed with some initial positions and velocities according to input initial conditions at high redshift. The volume is allowed to evolve under gravity. Snapshots of the simulation are taken at various stages of the simulations evolution, which translate to snapshots at decreasing redshifts. These snapshots are interpolated into a 'light cone', which replicates the distribution of dark matter at all the simulated redshifts, from the initial redshift down to $z = 0$ along a conical volume with the apex at $z = 0$, such that if viewed from the apex to the base, the cone represents a dark matter distribution equivalent to that "observed" from a fixed point in the universe, i.e., a telescope. By grouping together dark matter particles that are gravitationally bound, the simulation becomes discretized into dark matter '*halos*'.

CatSim has been developed with Carmen, one of the four types of box from the Large Suite of Dark Matter Simulations* (or LasDamas, McBride et al., 2011). LasDamas used a single cosmological model for all its simulation products, setting:

- $\Omega_m = 0.25$;
- $\Omega_\Lambda = 0.75$;
- $\Omega_b = 0.04$;
- $H_0/100 = (h =) 0.7$;
- $\sigma_8 = 0.8$;

where Ω_b is the baryonic component of the matter density of the universe, Ω_m. Carmen consists of a $1000^3\, h^{-1}\,\mathrm{Mpc}$ box containing 1120^3 particles with a particle mass of $4.938 \times 10^{10}\,\mathrm{M}_\odot\, h^{-1}$. The initial conditions are set with CMBfast (Seljak and Zaldarriaga, 1996) computing the power spectrum of the density fluctuations, an initial density field with positions and velocities determined by second-order Lagrangian perturbation theory (2LPT, Crocce et al., 2006), and an initial redshift of $z = 49$. The dark matter simulation is gravitationally evolved with the Gadget-II code (Springel et al., 2005) using a collisionless model for the dark matter particles. Dark matter halos are identified with a parallel friends-of-friends (FOF) code, Ntropy-fofsv, developed with the Ntropy framework (Gardner et al., 2007).

With this simulated distribution of dark matter halos (and particles) in the Carmen lightcone, *Halo Occupation Distribution* (HOD) parameters are applied to the FOF halos. HOD parameters determine how galaxies populate dark matter halos by defining a conditional probability P(N|M) that a virial mass M (i.e., a halo) contains N galaxies, and prescribing relative spacial and velocity distributions of those galaxies, thus connecting the observable galaxy with the underlying matter distribution (Berlind and Weinberg, 2002, and references therein).

In the DES Catalog Simulations, CatSim, galaxies are assigned to the LasDamas dark matter distribution with the ADDGALS (Adding Density Determined GAlaxies to Lightcone Simulations) algorithm (Busha and Wechsler, 2008), which is outlined as follows. A 'central' galaxy is always placed at the dark matter particle closest to the centre-of-mass of each halo and given the mean velocity of the halo. Galaxies are then assigned to the halo by randomly selecting dark matter particles from the halo and using their position and velocity information. The dark matter particles considered for

*http://lss.phy.vanderbilt.edu/lasdamas/

galaxy assignment are limited to those with a mass greater than $\sim 10^{12}\,\mathrm{M}_{\odot}\,h^{-1}$ (the mass resolution of the output galaxy catalog) as the simulation halos are smoothed on a lagrangian scale of $1.8\times10^{13}\,\mathrm{M}_{\odot}\,h^{-1}$. The galaxies themselves are distributed according to an HOD model that resembles the SDSS DR7 galaxy distribution at $z = 0.1$, characterizing this $z = 0.1$ distribution with the galaxy luminosity function and luminosity-dependent correlation function taken from multiple volume-limited SDSS galaxy samples.

The SDSS samples also provide a list of r-band galaxy magnitudes, which are assigned to the galaxy particles in the simulation with an input *Brightest Cluster Galaxy* (*BCG*) to halo relation, and a probabilistic relation between absolute r-band galaxy magnitudes and dark matter density, $\mathrm{P}(\delta\mathrm{M}|r_{\mathrm{mag}})$. The BCG r-band magnitude is applied separately, as the algorithm does not otherwise identify the BCG as being at or near the centre of the halo. Fainter galaxies are added to the simulation to randomly selected dark matter particles and assigned r-band magnitudes (N.B. Busha and Wechsler do not explicitly mention how faint these fainter galaxies are, and so this thesis does not describe how the luminosity-dependent correlation function is extended to fainter magnitudes). Once the simulated r-band galaxy catalog is constructed, a training set of *spectral energy distributions (SED*s) from the SDSS catalog was used to map colours onto the simulated galaxies so that they match the SDSS colour-environment relation (Hogg et al., 2004). The $z = 0.1$ SEDs assigned to the galaxies are then passively evolved (k-corrected, as per Blanton et al., 2003b) to the redshift of the galaxies themselves and the simulated band filters are applied to return the galaxies' apparent magnitudes and colours. Noise is then applied to the magnitudes to simulate the seeing effects and pixel noise that would be experienced in a real survey.

4.3.3 CatSim Products

From the construction of a CatSim (§4.3.2), two main catalogs are produced which are then used to give feedback on both the simulations and the tools trained on/with them. The first is the halo catalog and the second is the mock galaxy catalog. The halo catalog contains the list of identified dark matter halos that were assigned a galaxy by ADDGALS, down to a mass of $\mathrm{M}_{\mathrm{halo,min}} \geq 5\times10^{12}\,\mathrm{M}_{\odot}\,h^{-1}$. This 'truth' halo catalog describes the mass of the halo by M_{200}, which is the mass of the halo within R_{200}, which describes the radius from the halo centre at which the matter density drops below 200 times the critical density. The halo catalog also contains the halo's position in terms of R.A., dec and redshift as well as the simulation's spacial and velocity coordinates. Summed luminosity

information from all of the galaxies, the brightest 20, and the BCG alone are also recorded. Each of the halos are assigned an ID, which is then used to link galaxy information to the simulated dark matter distribution.

Several generations, or *versions*, of the CatSim catalogs have been released since November 2007, with > 20 major or minor revisions to the algorithm or output data being documented by each subsequent version. This thesis employs version 1.0 of the Aardvark release of the CatSim catalog, which was released to the DES collaboration in April 2013. This release consists of a $10313\,\square^\circ$ simulated patch of sky, to full DES depth, out to a redshift of out to $z = 1.35$, populated with ADDGALS galaxies. These galaxies populate the catalog down to $[grizy]_{\rm mag} = [26.0, 25.5, 24.8, 24.3, 22.5]$ based on Jim Annis' adjustment of the 5σ detection threshold of the "$5000\,\square^\circ$ Baseline Survey" (originally calculated for four bands in Tucker et al., 2007).

The mock galaxy catalog contains a lot of information on the simulated galaxies, the key components for AperC4 being the 'observed' galaxies' R.A. and declination, photometric redshift (or $p(z)$), and the magnitudes in each band with associated uncertainties. Different redshifts delivered by various photo-z algorithms are supplied, with associated redshift uncertainties. In this thesis, I use zCarlos $p(z)$ (§6.1) when running AperC4 on CatSim as this was the only "observable" redshift information (beyond the simulated input redshift itself) supplied at the time of the catalog's April 2013 release. Each galaxy is given its own unique ID number and the ID of the halo (from the halo catalog; HALOID) to which it is associated. By connecting the halo catalog and the galaxy catalog in this way, it is possible to see how well AperC4 identifies halos and halo galaxies when they are placed into clusters (e.g., this permits "membership matching" between the observed and simulated clusters and is used to evaluate blending/fragmentation of observed clusters with respect to their simulated counterparts; §5.2.1 and §5.1.2). This thesis does not explore other aspects of the galaxies contained in the supplementary galaxy catalog, such as simulation coordinate information, simulated lensing information, ellipticity, angular size, simulated observable values without the addition of noise, etc.

Simulated SDSS DR8 catalogs were also derived from the Aardvark v1.0 simulations, consisting of additional SDSS-like magnitudes derived from the simulated galaxy catalog. In addition to the simulated redshifts, the CatSim-DR8 galaxies have been given zCarlos redshifts (§6.1) which are trained on a subset of simulated galaxies (that are treated as spectroscopic observations; §6.1.2). No other observable redshift is derived from the CatSim-DR8 sample so this thesis limits itself to cluster finding with the zCarlos $p(z)$'s.

Due to a bug that mislabels the identification numbers of galaxies between the $p(z)$ files and the source galaxy catalogs (that the authors were unaware of until I informed them), this analysis is limited to $\sim 213\,\square^\circ$ of the full simulated survey area.

4.4 Summary

In this chapter, I have introduced the key elements of DES (§4.1) and given an outline of the cosmological probes intended to be used on DES data that will constrain the nature of the cosmic expansion (§4.2). I also describe the mock galaxy catalogs (CatSim) that have been developed by the DES simulations working group (§4.3). These catalogs have been made by the same team that developed simulations for the M05 paper, but covering a wider redshift range and forming a complete representation of the galaxy population to be observed, no longer limited by a finite number of spectral fibers per imaging area, or fiber collisions (§2.3.3).

The DES CatSim mock galaxy catalog is a very useful resource for training and evaluating observational tools against a known input cosmology. In chapter 6, I apply the new AperC4 algorithm introduced in chapter 3 to the Aardvark v1.0 iteration of the CatSim mock catalog. I evaluate AperC4's ability to recover the underlying halo population through comparison to the CatSim halo catalog, using statistical tools I introduce in the following chapter.

Chapter 5

Evaluating Cluster Finding

In this chapter, I discuss how APERC4 will be evaluated through matching cluster catalogs derived from simulated data to the simulation input.

In section 5.1, I introduce the measures used to evaluate cluster catalogs. In section 5.2, I introduce the membership matching evaluation method as previously used by the DES Cluster Working Group and my alterations to that algorithm. In section 5.4, I test my version of the matching algorithm and finally, I summarise my findings in section 5.5.

5.1 Evaluation Measures

As previewed in section 1.6, completeness and purity are employed as the de facto measures for cluster catalog quality. Rozo et al. (2007) demonstrate how these quality measures are degenerate with cluster cosmology (§1.3.1), so knowledge of these measures is extremely useful in qualifying the reliability of cluster catalogs for measurement of cosmological parameters.

5.1.1 Completeness and Purity

In this thesis, I employ *completeness* as one indicator of cluster catalog quality (as has been used historically by cluster finders such as Abell (1958), where it describes the fraction of known clusters recovered; §1.6). To demonstrate this quality, I consider a new cluster catalog, X, as produced by some halo/cluster finding algorithm run on some simulated galaxy catalog, and measure it against the underlying halo catalog, A, of that simulated catalog, where the halos/clusters that populate the simulation are known. For clarity, I will refer to the galaxies grouped together by the cluster finder X as clusters, whilst the simulated (true) galaxy groupings in A will be referred to as halos.

Equation (5.1) shows that completeness measures the fraction of halos in catalog A that have been identified in cluster catalog X. If the number of A halos found in (or matched to) the X catalog is equal to the total number of A halos, then the completeness fraction is 100%. As the number of matched A halos falls, so does the completeness measurement. If no clusters in catalog X are found in halo catalog A, then the completeness is 0%, irrespective of the number of clusters in catalog X.

$$X\text{completeness} = \frac{N(A \text{ clusters matched to } X \text{ clusters})}{N(A\text{clusters})} \tag{5.1}$$

However, completeness alone cannot evaluate the effectiveness of cluster finding by algorithm X. If catalog X were made to arbitrarily contain a uniform distribution of clusters such that at least one X cluster could be matched with one A halo (at the expense of introducing a large number of X clusters that don't match to any A, considered real, halo) then, in terms of completeness, catalog X would appear to represent catalog A.

To better determine cluster catalog quality, I use *purity* as complement to the completeness measure (akin to minimising spurious detections by Postman et al. (1996); §1.6). Equation (5.2) shows that the purity measures the fraction of clusters in catalog X that exist in catalog A. As the number of matched X clusters falls, so does catalog purity. Similar to completeness, if catalog X doesn't find any halos from catalog A, then the purity of X will be 0%, regardless of the number of clusters in catalog X.

$$X\text{purity} = \frac{N(X \text{ clusters matched to } A \text{ clusters})}{N(X\text{clusters})} \tag{5.2}$$

The only way to maintain high purity is for X to find clusters that have been defined in A. This could lead to overtuning to A, if there exist halos in A that are produced due to some non-physical feature of the simulation. Consider a situation where catalog A describes a large proportion of realistic clusters (from the total cluster population of the observed cosmological volume) and an additional large number of erroneous associations. Based on this situation, cluster finder X would only achieve a high completeness if it also identified the same idiosyncratic associations as clusters. Considering purity, if all of the clusters in X match to some fraction of the halos in A, then the catalog will display 100% purity against some fractional completeness. If the example of the uniform X cluster catalog were taken again, the purity would be seen to suffer as the increased number of clusters in catalog X that do not match A halos would cause purity to fall.

If a fraction of the clusters in X match halos in A, and the techniques employed by X create some differences to the underlying catalog A, then the measured purity and

completeness should characterize the ability of X to describe A. Given that clusters are real objects, then the matching of objects in X to objects that do not exist in A (e.g., line-of-sight associations, or fragmented A halos) should be a rarer occurrence than matching to real objects in A, and so there will be some threshold description of clusters to which X and A agree, before their characterizations of clusters diverge/devolve.

Note that A may also come in the form of some archival cluster catalog, but I do not consider that here, as such archival catalogs (a) may not perfectly capture the cluster population of the universe, and (b) would simply tune X to A, where the quality of A, and thus X, will remain unknown.

5.1.2 Unique Matching

Supplementing purity and completeness information by setting a uniqueness constraint can help describe the type of differences occurring between cluster and halo catalogs. Without a uniqueness constraint, one can achieve high completeness and purity by producing a catalog X that matches a subset of A multiple times. For example, if two catalogs, X and A, contain 100 clusters/halos each, then to achieve 100% completeness and 100% purity, X clusters can either match to each A halo one-to-one, or match to a single A halo 100 times. The uniqueness constraint serves to penalise the latter case, demanding that once the A halo has been matched by one X cluster, it cannot be matched any further, thereby reducing completeness and purity to more representative fractions.

Deviations between the non-unique and unique completeness and purity fractions can also indicate a degree of blending/fragmentation of clusters between one catalog and the other. For example, if catalog X's completeness does not change with the addition of the uniqueness threshold, but the purity does, then that indicates that multiple X clusters are matching to some number of A halos, i.e. clusters in catalog X are fragmented with respect to halos in catalog A. If in a situation where X purity remains constant and completeness falls, that indicates clusters in catalog X match to multiple A halos, i.e., X clusters are blended with respect to halos in catalog A.

5.1.3 F-measure

Given there are two qualitative measures of a cluster catalog, completeness and purity, I introduce a composite measure called the F-measure, or F1 score (Murphy, 2012). The F-measure has been used as a qualitative measure for various computational search processes such as text retrieval. In the context of scoring, cluster catalog quality is the harmonic

mean of the completeness and purity:

$$\text{F1 score} = \frac{2 \cdot \text{Purity} \cdot \text{Completeness}}{\text{Purity} + \text{Completeness}} \tag{5.3}$$

An advantage of the harmonic mean over the arithmetic mean, $^{(\text{Purity} + \text{Completeness})}/_{2}$, is that it indicates success where both purity and completeness are non-marginal, but will fall off if one of the measures tends towards zero. For example, if one produces a catalog where purity is 50% and completeness is 50%, both arithmetic and harmonic means go to 0.5. But if we consider a cluster survey that returns one true cluster in a cosmological volume that contains many ($N_{\text{cluster}} \gg 1$) clusters, purity will tend towards one whilst completeness tends towards zero. The arithmetic mean will still tend towards 0.5, whilst the harmonic mean will fall toward zero.

The F-measure allows cluster catalogs to be evaluated by a single statistic. In this analysis, completeness and purity are measured in bins of mass and redshift, and the F-measure is calculated for each of these bins. To assess a cluster catalog, the F1 scores for a range of mass/redshift bins are summed to give a single, global F1 score. As such, an 'optimal' catalog is obtained by choosing the catalog where the global F1 score is maximised.

5.2 Method of Evaluation

5.2.1 Membership Matching

The challenge for ApERC4, or any other optical cluster finder, is to identify clusters down to a determinable mass threshold so that they are useful for cosmology, as discussed in section 1.3. The new ApERC4 algorithm (chapter 3) now produces an accompanying galaxy catalog, which relates survey galaxies to the ApERC4 clusters identified, i.e., gives the explicit galaxy membership of each galaxy cluster.

In the absence of C4_{M05} galaxy memberships, M05 defined matches as being where a given C4_{M05} cluster centre is close to a simulation halo centre (*proximity matching*). The matching between C4_{M05} clusters and extant simulations made the assumption that differences between the centring of the clusters and the halos (as with catalogs X and A in §5.1.1, respectively) did not influence the matching. Since I did not have access to the galaxy memberships of the C4_{M05} clusters in the simulations it was trained on, this assumption could not be tested. Furthermore, in the event of cluster blending or fragmentation, the closest proximity centres between the C4_{M05} and simulated catalogs may not necessarily represent the best match between them.

As the CatSim (ADDGALS) galaxies are the observable tracer of the simulation halos, and are tagged as belonging to these halos by the halo identification numbers, then using these galaxies for cluster matching removes the element of ambiguity between AperC4 clusters and CatSim halos being matched by proximity of their centres. Membership matching also prevents confusion in matching by cluster and halo centres, where cluster finders have misinterpreted galaxy properties and placed clusters at dissimilar redshifts from their associated halo. In this thesis, matching clusters-to-halos by their shared galaxy populations, and vice versa, is called *membership matching*.

Membership matching (compared to proximity matching) allows a more fine-grained definition of what is meant by unique and non-unique matching (§5.1.2), and quantifies how 'good' a cluster-halo match is by proportion of shared membership. A 'match' in membership matching is counted where at least one galaxy from a halo is found to be in a cluster, or vice versa. When matching non-uniquely, the number of matches from halo-to-cluster and from cluster-to-halo shouldn't be different, as there are no limits to the matching. When matching uniquely (§5.1.2), the cluster/halo that contains the most members of the halo/cluster that it is being matched to is counted as the sole "best" match.

5.2.2 Rank Matching

Membership matching allows one to qualitatively, and directly, assess the quality of a match between any individual cluster-halo pair. However, this does not help decide the suitability of a given cluster catalog for cosmology unless the matching for a whole catalog is assessed. Since we want to (eventually) be able to derive cosmology utilising these clusters then it would be useful to examine how well the clusters are recovered as a function of both redshift and mass. Rather than trying to determine masses for the clusters (in addition to locating them), the matching algorithm utilises rank matching, to allow qualitative (as opposed to quantitative) assessment of a cluster catalog compared to some underlying model (the halo catalog).

Accurate mass determination of clusters is a desirable goal for cosmology, but to accurately estimate a cluster mass is difficult, even comparing between observational probes (Noh and Cohn, 2012; Rozo et al., 2012), and is sensitive to the mass model employed in the construction of CatSim halos (or indeed halos of any simulation being matched to). Using ranking mechanisms for matching allows the determination of a non-parametric relationship between the (ranked) mass in CatSim and the ranking mechanism of AperC4,

i.e., showing that such a relationship truly exists. The matching by rank is evaluated with the Spearman rank-order correlation coefficient, ρ or r_s (Press et al., 2007), which is the linear correlation coefficient of the ranks, i.e., if the AperC4 cluster ranks are denoted by X_i and the matched CatSim halo mass ranks are denoted by A_i, such that the cluster ranked X_i matches to the halo ranked A_i, then the rank-order correlation coefficient is given by,

$$\rho = \frac{\sum_i (X_i - \bar{X})(A_i - \bar{A})}{\sqrt{\sum_i (X_i - \bar{X})^2}\sqrt{\sum_i (A_i - \bar{A})^2}} \ , \tag{5.4}$$

where $\bar{X}$ and $\bar{A}$ are the mean rank for clusters and halos, respectively. The Spearman rank-order correlation coefficient ranges in value from $-1 \leq \rho \leq 1$, where 1 represents a monotonic linear correlation, -1 represents a monotonic anti-correlation, and 0 means no correlation. A monotonic correlation means that any increase in AperC4 rank means an increase in CatSim rank, whilst the monotonic anti-correlation means any increase in AperC4 rank results in a decrease in CatSim rank.

To do the ranked membership matching, I modified a code originally written by Brian Gerke for the Cluster Working Group (CWG) of the DES collaboration, which matches clusters-to-halos and vice verse using cluster/halo memberships and a fiducial ranking for mass (§5.3 below). I will first review Gerke's matching algorithm and then discuss my modifications.

5.2.3 Gerke Matching Algorithm

Gerke's algorithm works as follows:

1. AperC4 (or some other cluster finder) produces a list of clusters with R.A., dec, redshift and rank properties, and a list of associated cluster members.

2. Each halo mass, M_{200} (or M_{true} in Gerke's algorithm), is assigned to the clusters in order of rank, i.e., the mass of the nth most massive halo is assigned as the mass, $M_{\rm obs}$, of the nth ranked cluster. This simplification means a cluster does not have to physically be associated (by proximity or membership) to the halo it is ranked equal to (nor, indeed, associated to any halo).

3. Divide the lists of clusters/halos into redshift bins.

4. For each redshift bin, perform membership matching between clusters and halos in four different ways:

(a) In order of descending mass/rank order, for each halo, find the cluster with the largest proportion of halo members (galaxies identified as belonging to a single halo), regardless of whether the cluster has already been matched. Where a cluster has already been matched, it can be matched again by subsequent halos up to a total of five times (five times is an arbitrary threshold set by Gerke). This is non-unique, halo-to-cluster matching.

(b) In rank order, for each halo, find the cluster with the largest proportion of halo members and, once the cluster has been matched, remove it from the list of matching candidates. This is unique (§5.1.2), halo-to-cluster matching.

(c) In rank order, but now for each cluster, find the halo with the largest proportion of cluster members, non-uniquely (as clusters are matched to halos in step 4a), allowing up to five clusters to match to any given halo. This is non-unique, cluster-to-halo matching.

(d) In rank order, for each cluster, find the halo with the largest proportion of cluster members, uniquely (as clusters are matched to halos in step 4b). This is unique, cluster-to-halo matching.

5. Having matched clusters to halos and vice versa, find the non-unique and unique two-way matches between clusters and halos.

6. Dividing the matches into bins of mass according to the rank-mass relation of the halos, compute matching statistics for unique and non-unique matches as a function of halo/cluster mass and redshift.

Having assigned the clusters masses from the halo catalog, and matched halos to clusters (and vice versa) by membership, it is useful to evaluate the scatter of cluster mass against halo mass. Examining the matches by mass, from most massive to least, the mass at which the cluster mass, $M_{\rm obs}$, begins to deviate significantly from halo mass, $M_{\rm true}$, indicates the robustness of the cluster ranking by mass. By dividing the clusters/halos into redshift bins in step 3, one can also gauge how well a cluster catalog matches the halo catalog and the effectiveness of the ranking mechanism as a function of redshift (see §5.2.5; for the rest of this subsection, and for purposes of illustration, I will momentarily neglect redshift dependence and only consider a single redshift bin).

However, because it relies on the mass-rank relation of the halos, this algorithm makes the assumption that both halo and cluster catalogs contain the same mass halos/clusters in each redshift bin, e.g., the masses of the top 100 halos in a redshift bin are identical to

the masses of the top 100 clusters in the same redshift bin. That is to say, the halos and clusters are simply being binned by rank.

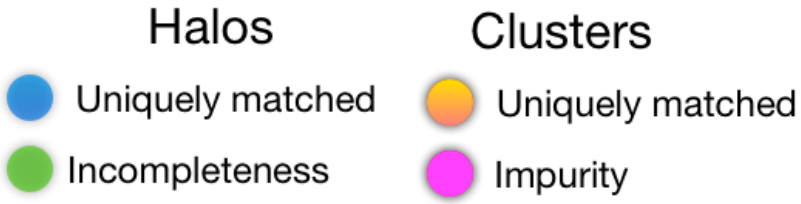


Figure 5.1: This figure is the symbol key for the ranking example cartoons in Figures 5.2 to 5.4. The dots in the left column represent halos, coloured blue where they are matched to a cluster, and green where they are not. The dots in the right column represent clusters, coloured orange where they are matched to a halo, and pink where they are not.

Considering a catalog with low completeness but high purity in Figure 5.2, since the mass-rank bins are divided by the number of halo objects within a mass range, the cluster mass-rank bin will count clusters which match to halos in lower halo mass-rank bins. The outcome is that catalogs with low completeness but high purity will appear to be highly complete in the highest mass-rank bins, with the fall in completeness only affecting the lower mass bins.

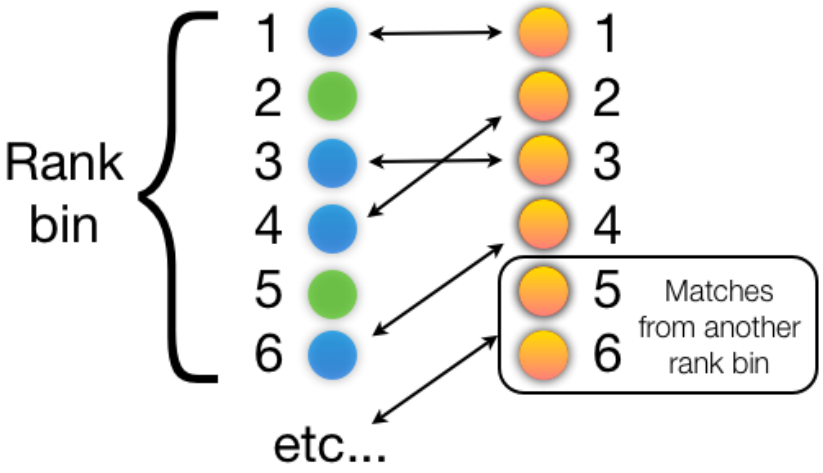


Figure 5.2: This cartoon is a representation of the Gerke algorithm's treatment of matching a highly pure but incomplete cluster catalog to the halo catalog (key is given in Figure 5.1). The numbers to the left and right of the dots represent the rank of the halos and clusters, respectively. The arrows between the dots are representative of their unique two-way matching. By relying on a 1:1 halo rank to cluster rank relation, the cluster mass-rank bin will simply be populated by lower mass clusters that match to a lower halo mass-rank bin. Thus the cluster mass-rank bin appears complete, despite there being fewer clusters of the same mass range as the halo mass-rank bin.

If considering a catalog that is highly complete but impure (i.e., contains spurious detections) in a given mass range, as in Figure 5.3, the purity measured in that mass range is limited to the number of halos that occupy that mass-rank bin. Therefore, impurity in a mass-rank bin will cascade into lower mass-rank bins such that catalogs will appear to be more pure in higher mass-rank bins than they actually are. The assumed equivalency between the masses of clusters and halos (in a mass-rank bin) means that a cluster catalog may look pure in higher high mass-rank bins, since all the halos in the same mass-rank bin have been matched to.

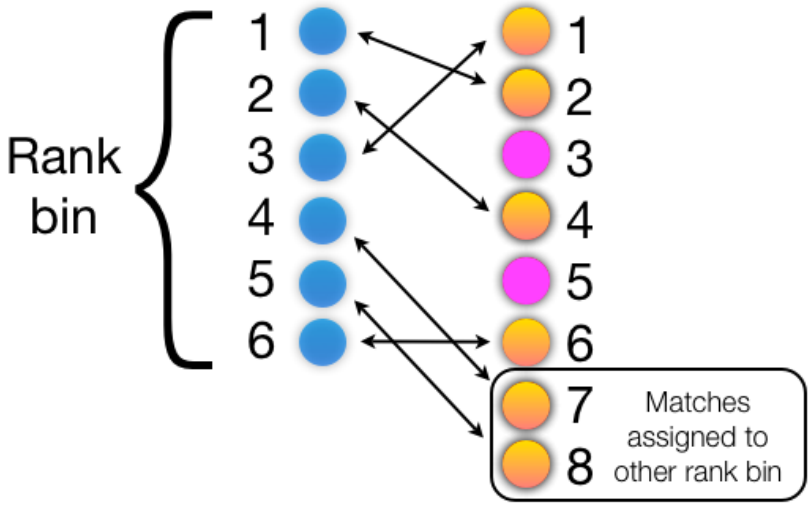


Figure 5.3: This cartoon is a representation of the Gerke algorithm's treatment of matching an impure but highly complete cluster catalog to the halo catalog (key is given in Figure 5.1). The numbers to the left and right of the dots represent the rank of the halos and clusters, respectively. The arrows between the dots are representative of their unique two-way matching. In this case, the cluster mass-rank bin will be populated by well described clusters and impurities (non-real clusters), and so the catalog appears complete as all the halos in the bin have been matched to. The Gerke matching algorithm will claim that the cluster catalog is the appropriate purity in the highest mass-rank bin, but due to the matched clusters that rank below the bin limit (the rank bin above being occupied by impurities), the next mass-rank bin will have increased purity. As such, purity can be enhanced by cluster matches from higher mass-rank bins cascading into lower mass-rank bins.

Further to the cartoons above, one can see that the matched clusters within a mass-rank bin may not even be physically related to the halos that occupy the equivalent halo mass-rank bin. The dependence of measured completeness and purity on the rank of occupation of mass-rank bins motivates the modified matching code in section 5.2.4.

In summary, the Gerke algorithm assumes that a cluster finder cannot find more halos than are actually present in any mass/redshift bin. This is true, by definition, but if a cluster finder "overfinds" it means it is identifying false clusters (i.e., it is associating

galaxies along the line of sight, or describing a fragment of a halo), which should be characterised by decreased purity of its redshift/mass bin. With the Gerke algorithm, catalog completeness and purity is impacted through binning by rank and assuming cluster catalog rankings scale 1:1 with halo catalog mass-ranks.

5.2.4 Modified Matching Algorithm

To address some of the shortcomings of the Gerke matching algorithm, I modified the matching algorithm such that it operates as follows:

1. AperC4 (or some other cluster finder) produces a list of clusters with R.A., dec, redshift and rank properties, and a list of associated cluster members.

2. Divide the lists of clusters/halos into redshift bins.

3. Before assigning halo masses to the clusters, for each redshift bin, perform cluster-halo membership matching in the four ways used in the Gerke algorithm (steps 4a through 4d in §5.2.3, above).

4. Having matched clusters to halos and vice versa, find the unique (§5.1.2), two-way matches between clusters and halos.

5. For each unique two-way match, assign the mass of the halo to the cluster.

6. Looping through each redshift bin again, examine the clusters in rank order. Any clusters which are not unique two-way matches are assigned a mass by interpolating between the masses of the unique two-way matched clusters that are ranked immediately above and below these clusters.

7. Divide the halos and clusters in each redshift bin into bins of mass according to the mass-bin limits and compute matching statistics as a function of mass and redshift bins.

In contrast to Gerke's method (where cluster's are assigned a mass by the halo mass-rank relation; §5.2.3), a cluster is assigned mass of the halo it is uniquely, two-way matched to (i.e., the least ambiguous matches between the cluster and halo catalogs), and rank is only used relative to those clusters in the cluster catalog that already have a mass assigned to them (i.e., they are matched well). Figure 5.4 demonstrates this modification to the matching algorithm.

Remaining clusters have their masses interpolated from the masses of unique, two-way matched clusters ranked above and below them. The cluster masses are now trained to halo mass (M200). As a result, cluster catalog purity and completeness are now evaluated in the appropriate mass/redshift bins, properly reporting impurities and incompleteness across all mass bins.

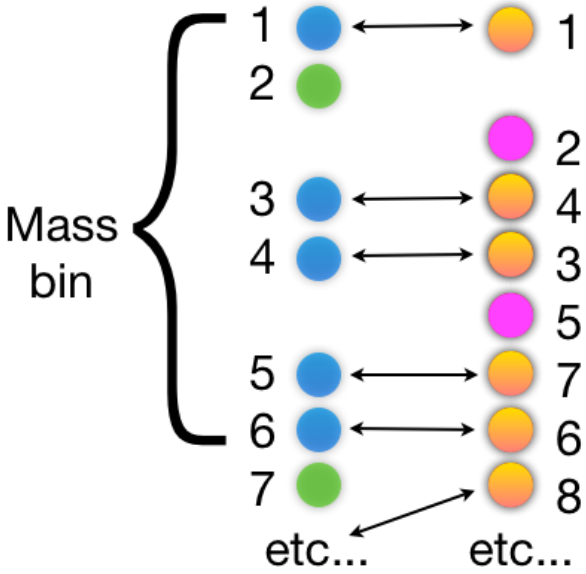


Figure 5.4: This cartoon is a representation of my modified algorithm's treatment of matching an impure but highly complete cluster catalog to the halo catalog (key is given in Figure 5.1). The numbers to the left and right of the dots represent the rank of the halos and clusters, respectively. The arrows between the dots are representative of their unique two-way matching. By assigning the halo mass of a two way match to its cluster counterpart (superseding the cluster rank), the cluster mass-rank bin will be populate the same mass range as the halo mass-rank bin. Thus, the bin is representative of the cluster/halo mass. To illustrate this, both clusters and halos in this figure are ordered by descending mass, hence the cluster ranks appear out of order.

The number of clusters is now independent of the number of halos in any given mass/redshift bin. The Spearman ranking coefficient (§5.2.2) characterises how well clusters are ranked using the unique two-way matches, which is the same measure used by Gerke. The ranking coefficient characterises the catalog only using two-way unique matches. As such, an objectively highly complete and pure cluster catalog will be shown to be poorly/well ranked from the ranking coefficient without affecting its measured completeness/purity.

5.2.5 Redshift Binning

The cluster-halo matching that occurs through the matching algorithms is performed in bins of redshift. The matching algorithms described above take in a range of redshift bin

limits (step 3 in §5.2.3 and step 2 in §5.2.4), into which the cluster and halo catalogs are distributed. This is so the mass ranking of any given cluster/halo is evaluated against other clusters/halos where the completeness of the galaxy sample is consistent. Thus, nearby clusters with more visible members are not compared with clusters at high redshift, where galaxy membership may appear poorer due to a greater proportion of high redshift galaxies falling below the magnitude limit of a (simulated or real) survey.

Redshift binning also limits the effect of any underlying evolution in the cluster mass function (Eke et al., 1996; Böhringer et al., 2014). It can also highlight any redshift dependence of the mass proxy calculation (although this goes beyond the scope of this thesis; alternative cluster finders would be necessary to distinguish such an effect from cluster finding performance).

Redshift bins should be appropriately sized, such that the bin volumes contain enough clusters in the redshift interval to characterise the mass function (at that redshift) well, whilst ensuring that differences in galaxy completeness within a redshift bin are small. I note that membership matching (§5.2.1) means that clusters may be matched to halos, and vice versa, where they do not occupy the same redshift bin. As such, the measured success of finding a given cluster is not dependent on a cluster finder assigning the correct redshift to that cluster.

Hence, redshift binning helps to diagnose a cluster finder's selection function, providing a breakdown of cluster catalog completeness and purity such that the effects of survey completeness, evolution of the cluster mass function, and/or redshift dependence of the mass proxy, can be accounted for.

5.2.6 Matching Algorithm Output

The matching algorithm outputs fours sets of summary plots describing:

- the *non-uniquely* matched clusters and halos plotted by mass, in $\log(M/\,\mathrm{M}_\odot)$ units, for each redshift bin (e.g., Figure 5.5);
- the *uniquely* matched clusters and halos plotted by mass for each redshift bin (e.g., Figure 5.6);
- the completeness, purity and F1 scores of the APERC4 clusters, by any possible match or best match, as a function of $\log M_{\mathrm{obs}}$ for each redshift bin, and the associated scatter (e.g., Figure 5.7); and
- the deviation of the APERC4 cluster centres from the halo centres in radial (R.A./dec) and redshift directions (e.g., Figure 5.8).

In the following subsections, I describe the information contained in example Figures 5.5 through 5.8, and how this reflects on the quality of the example APERC4 cluster catalog (for which, in this example, the input parameters will remain unspecified). To render the information, I use redshift bins specified by the minimum/maximum limits $z_{lim} = [0.0, 0.2, 0.3, 0.4, 0.6]$ (as guided by §5.2.5). For Figure 5.7 the clusters are further subdivided into mass bins of size $\delta \log {}^{M_{\rm obs}}/_{\rm M_\odot}$ (or $\delta \log {}^{M_{\rm true}}/_{\rm M_\odot}$) $= 0.25$.

Non-unique Mass Scatter Plots

Figure 5.5 gives an example of a mass scatter plot from the matching between an APERC4 catalog and the halo catalog. The four plots show the matching and mass scatter for the four redshift ranges specified by the five redshift limits mentioned above. As explained in the caption, the blue dots represent one-way matching of halos to clusters (halo-cluster matches). Where halos do not match to any cluster (i.e., no halo members are found in any cluster), they are placed next to the vertical axes (where $\log M_{\rm obs} < 13$ is treated as not matched). Similarly, the red dots represent one-way matching of clusters to halos (cluster-halo matches), and where there is no match, the cluster is placed along the horizontal axes in line with its assigned mass, $M_{\rm obs}$. Smaller points represent matches other than the best match for each halo/cluster; up to five non-unique matches are made for each cluster/halo (the arbitrary five matches from 4a and 4c in §5.2.3).

The black points represent the two-way matching between clusters and halos and, by definition, don't include missed matches. The black points mostly occur along the $M_{\rm obs} = M_{\rm true}$ relation, as these clusters have their mass calibrated from the best uniquely two-way matched halo. Those that do not lie on this relation indicate non-unique two-way matches: clusters/halos that match to each other, and are mutually amongst their top five matches, but do not represent the best unique match, i.e., their observed mass has been interpolated by rank. The $M_{\rm obs} \neq M_{\rm true}$ black points indicate where multiple halos have been blended into a single cluster or multiple clusters are fragments of a single halo.

Unique Mass Scatter Plots

The coloured points and arrangement of plot items in the example unique mass scatter plot in Figure 5.6 is the same as those in the non-unique mass scatter plots above (Figure 5.5). The black points all occur along the $M_{\rm obs} = M_{\rm true}$ relation, as is the design of this algorithm. Figure 5.6 also includes rank information about the unique two-way matches in these plots. The purple lines represent $\log(M) \pm 0.4$ which is used to define the 1σ

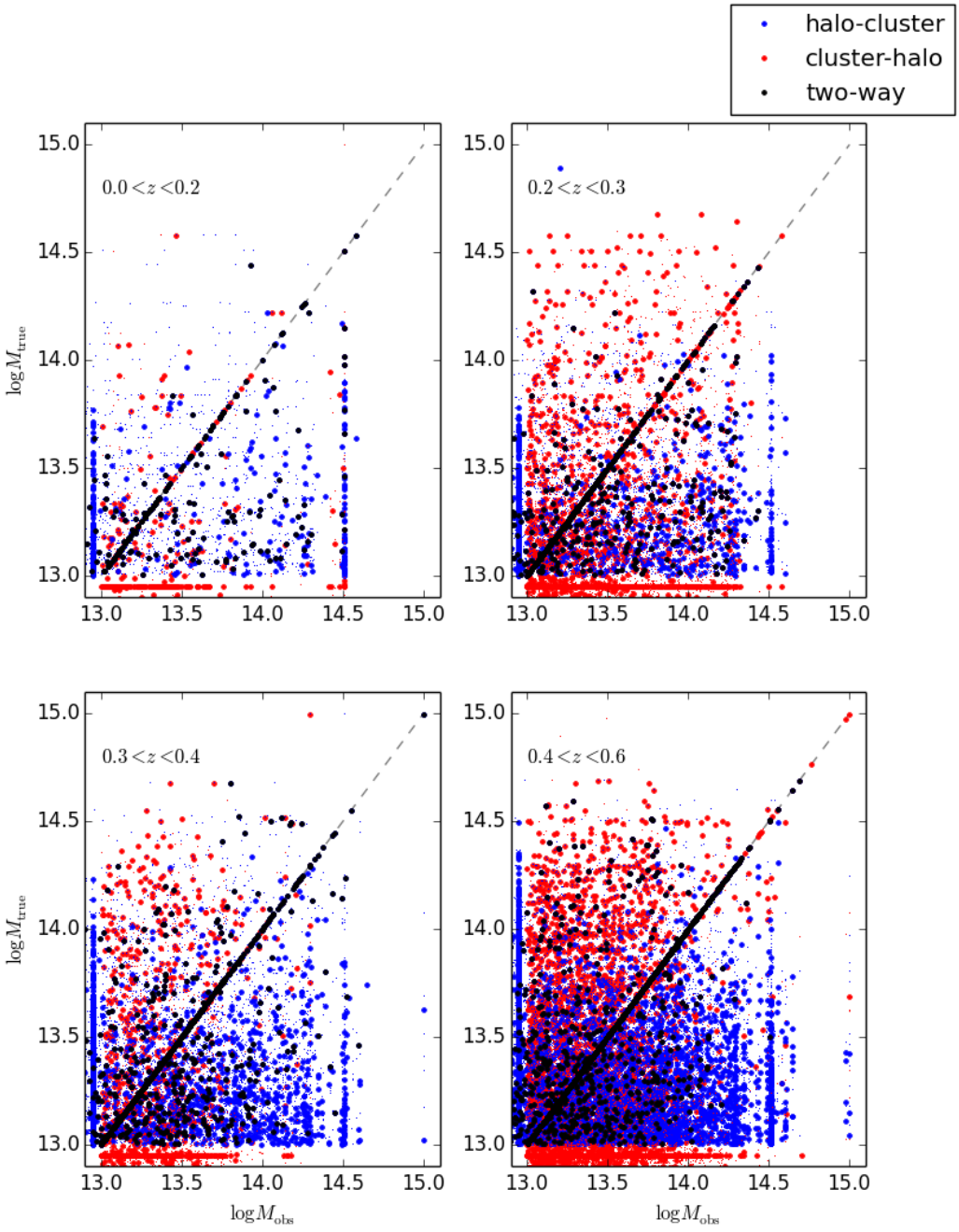


Figure 5.5: Example of a set of mass scatter plots produced by non-unique matching. The blue dots represent halos that match to clusters but not the other way around; the red dots represent clusters that match to halos but not the other way around; and the black dots identify halos/clusters that are two-way matches. The smaller points represent non-unique matches where the cluster/halo has already been matched. The dashed line shows where $M_{\rm obs} = M_{\rm true}$.

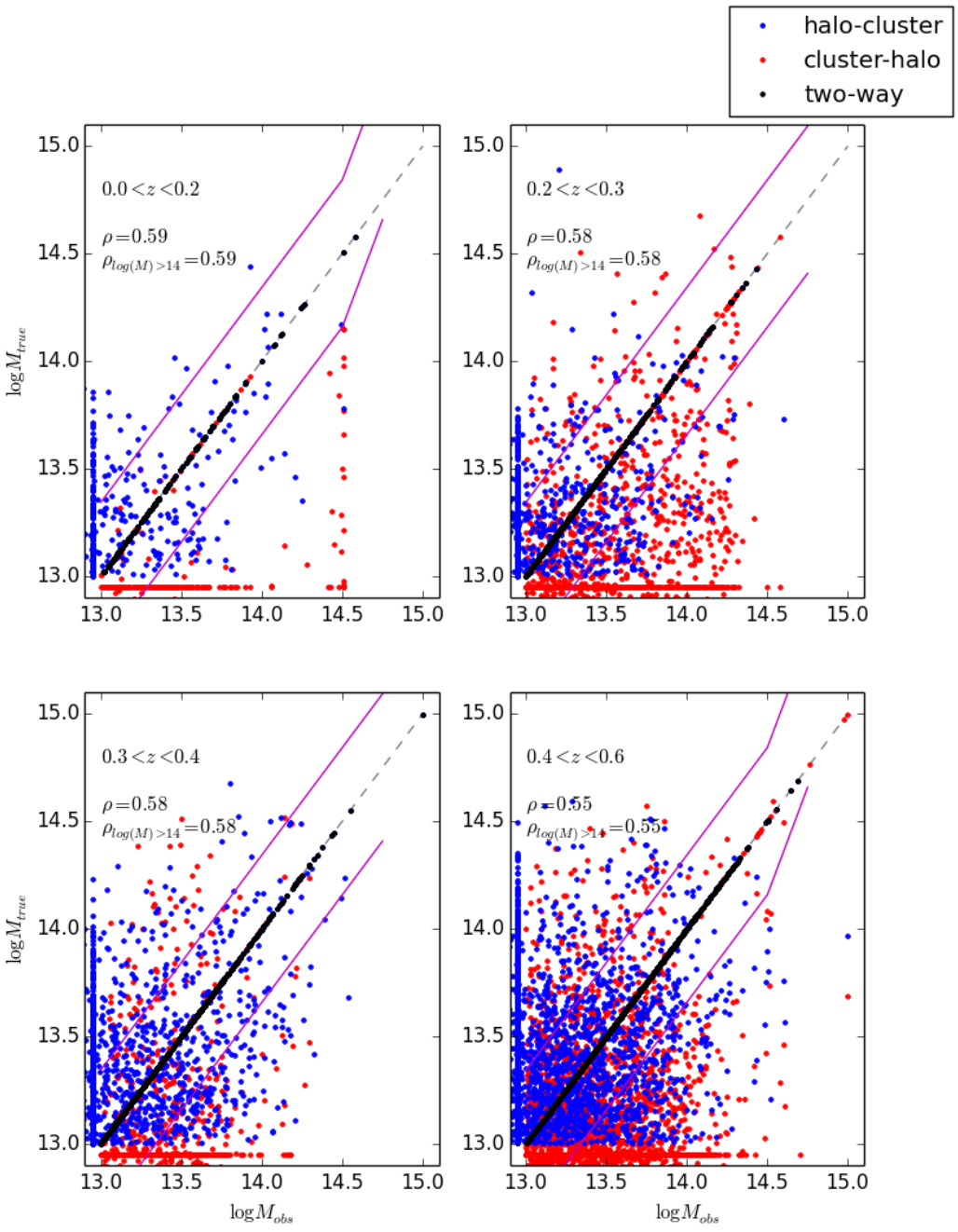


Figure 5.6: Example of a set of mass scatter plots produced by *unique* matching. The blue dots represent halos that best match to clusters but not the other way around; the red dots represent the inverse case; and the black dots identify halos/clusters that best match to each other. The dashed line shows where $M_{\rm obs} = M_{\rm true}$. The purple lines represent the 1σ scatter around the $M_{\rm obs} = M_{\rm true}$ relation.

scatter on the median $M_{\rm obs}$ in bins of $M_{\rm true}$. The value for ρ is the Spearman rank-order correlation coefficient (§5.2.2), which is limited to the range $-1 \leq \rho \leq 1$. That ρ is greater than zero indicates that there is a positive correlation between the halo mass and AperC4 rank.

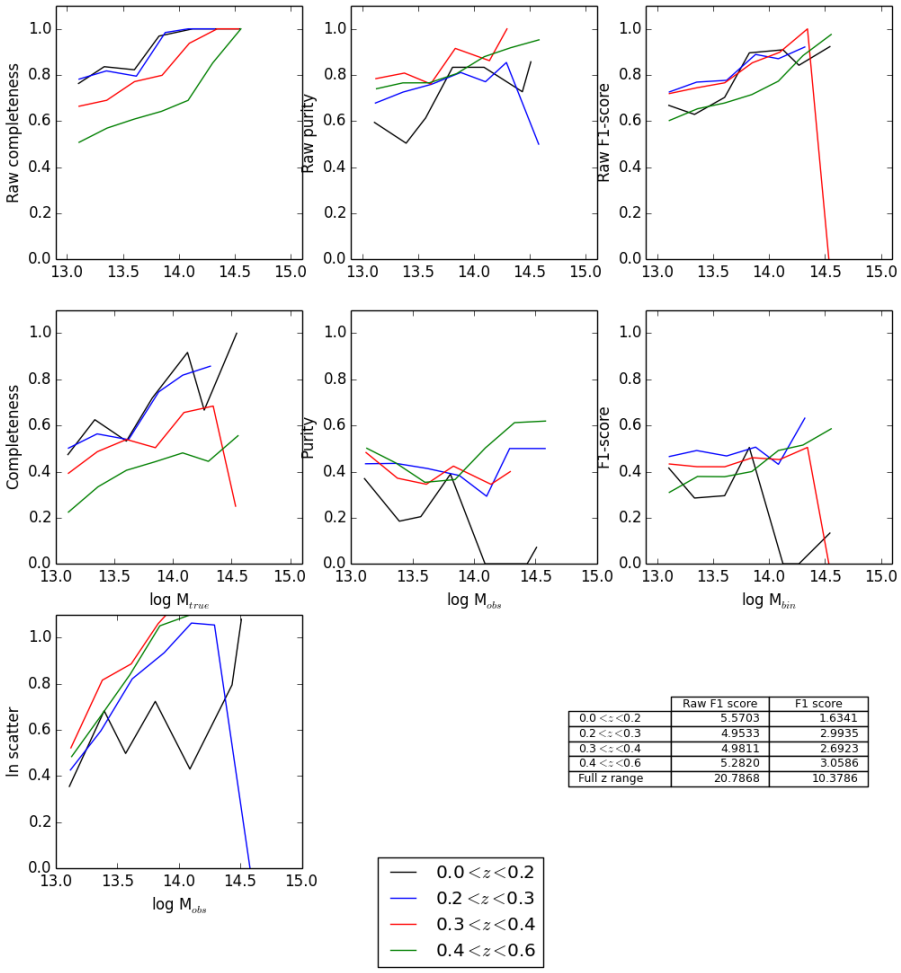
Comparing the distribution of the non-unique matches to the unique matches, there is a noticeable change in the distribution of halo-cluster matches (red points). Some halos match uniquely to nothing, and so end up along the vertical axes in the plots of their respective redshift ranges. The decreased scatter in the unique mass-scatter plots implies that, with the uniqueness constraint, the halos do describe clusters that are ranked well, but their memberships tend to be dominated by those of higher ranked clusters. A similar effect occurs with the cluster-halo matches, however, very few additional cluster-halo matches end up matching to nothing, implying that, in this catalog, clusters blend multiple halos more than fragmenting them.

Purity and Completeness Plots

Figure 5.7 is an example of the set of completeness and purity plots delivered by the matching algorithm, with the redshift bins identified by the legend at the bottom. In order to compute purity and completeness, a power law is fit to the σ-clipped standard deviation of $M_{\rm obs}$ versus $M_{\rm true}$ in bins of $M_{\rm true}$, and rejecting those matches outside a 3σ band about $M_{\rm obs} = M_{\rm true}$, as seen by the purple lines around the grey dashed line in the plots of Figure 5.6. Purity and completeness are given by the fraction of clusters within 3σ of the log scatter.

The "raw" completeness and purity of the AperC4-CatSim matching, as given by the upper left and middle plots, is effectively representative of the non-unique matching in Figure 5.5, showing the fraction of clusters or halos that match to any halo or cluster, respectively, regardless of the quality of the match. The fall in completeness with increasing redshift and/or decreasing mass shows that the cluster finder is less successful at finding low mass and/or high redshift clusters. The general decline in purity from high mass to low mass shows that the cluster finder falsely identifies clusters at lower masses more often than at high mass. The upper-rightmost plot shows the F1 score (the harmonic mean of completeness and purity) for each mass and redshift bin.

The middle row of plots in Figure 5.7 show the effect of applying the uniqueness constraint on the matching. For this catalog, completeness has fallen from its high raw completeness, indicating that a fraction of clusters in all mass-redshift bins describe more



	Raw F1 score	F1 score
0.0 <z<0.2	5.5703	1.6341
0.2 <z<0.3	4.9533	2.9935
0.3 <z<0.4	4.9811	2.6923
0.4 <z<0.6	5.2820	3.0586
Full z range	20.7868	10.3786

Figure 5.7: Example of a set of purity and completeness plots. The upper and middle plots show the log mass of the matches (horizontal axes) against, from left to right, the fractional completeness, purity, and F1 score (vertical axes) of each mass bin, where mass is binned by $log(\text{Mass}) = 0.25$ bins. The lowest left plot shows the log scatter in the $M_{\text{obs}} = M_{\text{true}}$ relation in bins of M_{true}. The lowest bottom right table gives the total F1 score in bins of redshift and for all redshifts for non-uniquely (raw) and uniquely matched catalogs. The colours of the lines represent the matching redshift ranges given in the bottom centre table

than one halo (blending). When looking at purity, there has been a fall from the raw purity, indicating that a fraction of halos in all mass-redshift bins describe more than one cluster.

The lower left hand plot shows the log scatter in bins of $M_{\rm obs}$. The increase in scatter for the higher redshift bins indicates that higher mass clusters are more poorly ranked than lower mass clusters. However, it should be noted that higher mass halos/clusters are rarer than lower mass objects, and so misestimated rankings result in a greater difference in mass for highly ranked objects.

As an aside, where the halo catalog and cluster catalog are identical, the scatter goes to zero, as expected. However, this will also be true where the catalogs are identical but their rankings are not, since the clusters will get their masses from their halo match, irrespective of rank. In this rare situation, one will be able to diagnose the disordered ranking using the Spearman ranking coefficient (see §5.4.5).

In summary, the raw plots show the proportions of AperC4-CatSim clusters that relate to real halo objects. The addition of the uniqueness constraint show how AperC4's identification of the CatSim halos (by membership) is additionally impacted by fragmentation or blending. The scatter indicates how well cluster masses have been assigned based on the two-way halo matches. Finally, the F1 score gives an objective measure of cluster catalog's quality.

Centring Plots

Figure 5.8 presents the offsets between the cluster centre and the halo centre in redshift and radial dimensions. The left hand plots show the distribution of these offsets, whilst the right hand plots show the individual offsets as scatter against redshift. For this example catalog, the deviation in redshift from matching clusters to halos is unbiased, with an approximately gaussian error described by full-width half maximum (FWHM) of $\Delta z \simeq \pm 0.02$. The same redshift deviation is seen to apply to halo-cluster and two-way matching.

The radial centering offset indicates that clusters are most likely to deviate from a halo centre in R.A./dec by a median $0\farcm3$ offset. Only a handful of cluster centres deviate from the halo centres by more than 4 arcminutes, where 4 arcminutes corresponds to $\lesssim 2.1\,$Mpc (using the simulation cosmology) at the upper limit of the simulation's redshift range.

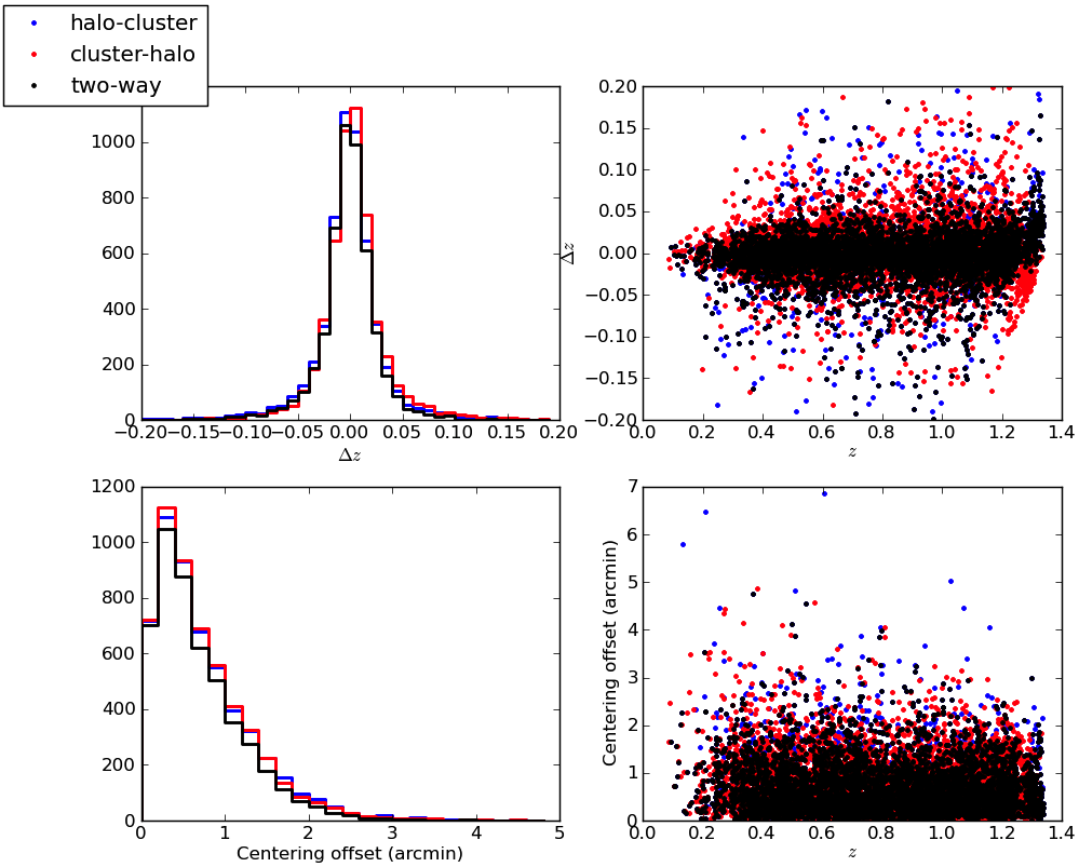


Figure 5.8: Example of a set of centering plots. The upper plots show the deviation between matched CatSim halo and cluster centres in the redshift dimension with (*left*) a histogram of the difference in z and (*right*) each halo-cluster offset against halo redshift as a scatter plot. The lower plots show the radial centring offset between matched CatSim halo and cluster centres in arcminutes with (*left*) a histogram of the halo-cluster radial offsets and (*right*) a scatterplot of the radial separation against redshift of each matched halo-cluster. The blue, red and black points/lines represent the halo to cluster, cluster to halo and two-way matched offsets.

5.3 Ranking the AperC4 Clusters by Mass Proxy

The AperC4 cluster catalog delivers clusters with unique memberships and $p(z)$ś. But to evaluate the output catalogs, one must invoke a mass proxy to deliver a rank.

An additional step to the algorithm flow (as outlined in §3.2.1) is applied to estimate the relative mass distribution of the clusters in the form of a rank. It does not affect the final cluster or member catalogs, and is therefore not part of the cluster finding algorithm. This is to allow estimation of the completeness and purity of the returned clusters in bins of redshift and mass.

Rank clusters: Clusters are ranked from most massive to least massive with the following mass proxy ranking mechanisms:

(a) by $N_{\rm gals}$

(a) by Abell richness.

These establish a relative mass proxy between clusters in a given catalog such that higher ranked clusters are estimated to be more massive than lower ranked clusters. The performance of these proxies is given by the Spearman rank-order correlation coefficient (§5.2.2) for the unique two-way matched clusters (§5.2.4), and will affect the completeness and purity in the relevant mass-redshift bin (§5.2.6) where clusters are not uniquely two-way matched to a simulation halo.

At this point, it is appropriate to explain why I have not used a red sequence richness estimator (e.g. Rozo et al., 2008). Red sequence cluster finders would be better suited to ranking/estimating the mass of these clusters. By including non-red sequence galaxies in a cluster, red-sequence richness measurement increases the scatter in richness vs X-ray luminosity comparisons (Rykoff et al., 2012). As AperC4 makes no distinction between red-sequence and other galaxy colours when designating galaxies to clusters, a red-sequence richness estimate from these clusters will include this extra mass-ranking scatter.

5.3.1 $N_{\rm gals}$ Ranking

The $N_{\rm gals}$ ranking simply ranks AperC4 cluster masses by the number of galaxy members, $N_{\rm gals}$, with the most populated (highest $N_{\rm gals}$) cluster being ranked as the most massive, and the least populated as the least massive.

Whilst this will be indicative of mass for survey volumes where the galaxy census is close to complete, lower luminosity galaxies at high redshifts fall below the detection limit

of the survey, thus the memberships of cluster systems will appear to decrease as a function of redshift for a given mass.

$N_{\rm gals}$ also may suffer from projection effects. But since the galaxies which constitute the field have mostly been eliminated through use of FDR (§3.2.4), these projection effects are assumed to be minimal. That said, there may be some misattribution of galaxies to cluster potentials as a result of poorer $p(z)$ estimations, allowing galaxies to favour an incorrect cluster redshift. This in turn would increase scatter in the mass-ranking relation.

Furthermore, the relationship between cluster $N_{\rm gals}$ and mass has some intrinsic scatter. Galaxies represent a small fraction of the total baryon content of a cluster (around 90% of the baryon component being the hot X-ray gas at the centre of the cluster potential Giodini et al., 2009), which in turn represents a small fraction of the total matter content of the cluster (the cluster potential being dominated by dark matter Zwicky, 1933).

5.3.2 Abell Ranking

The Abell richness ranking in this thesis is derived from Abell's specification of galaxy clusters (§1.2.1) applied to the $N_{\rm gals}$ richness measurement (§5.3.1). Abell's first characterisation of clusters through their constituent galaxies attempts to cut foreground/background galaxies by limiting the range of magnitudes between the third brightest cluster member to two magnitudes fainter, when measuring the cluster properties. Originally, Abell also specified further selection criteria such as a minimum $N_{\rm gals} > 50$, that cluster selection is limited to high galactic magnitudes, and that the > 50 galaxies above the faint magnitude limit are within a 1.5 Mpc aperture.

The criterion applied here is that the magnitude of the faintest galaxy counted in the richness measure is no more than 2 magnitudes fainter than the 3rd brightest galaxy in the cluster. The motivation for this criterion is to compensate for the fact that fewer galaxies will be seen at high redshifts, due to their increased distance causing their observed magnitudes to fall below the detection limit of a given astronomical survey. The other Abell criteria are not included, as SDSS nominally probes galaxies at latitudes above/below the galactic disk, and quality of clusters by $N_{\rm gals}$ is probed as a separate parameter.

Note that Abell ranking is not tested in this chapter, as its impact on the results will be consistent with the $N_{\rm gals}$ ranking test in section 5.4.6. It is tested on data in chapters 6 and 7.

5.4 Testing the Matching Code

To test the matching code, a series of catalogs derived from the CatSim halo/galaxy catalog (§4.3) were created to test how impurity and incompleteness are evaluated by the matching algorithm. The derived catalogs each consist of two parts: the 'cluster catalog', which consists of the cluster/halo rank, the R.A., dec, and z; and the 'member catalog', which consists of the galaxy ID and the rank of the cluster to which it is associated. Hence, the derived catalogs are consistent with AperC4 catalog content. Referring to section 5.1.1, the derived catalogs are each treated as catalog X, whilst the full CatSim catalog to which it is matched, is treated as catalog A.

5.4.1 Pure and Complete Cluster Catalog

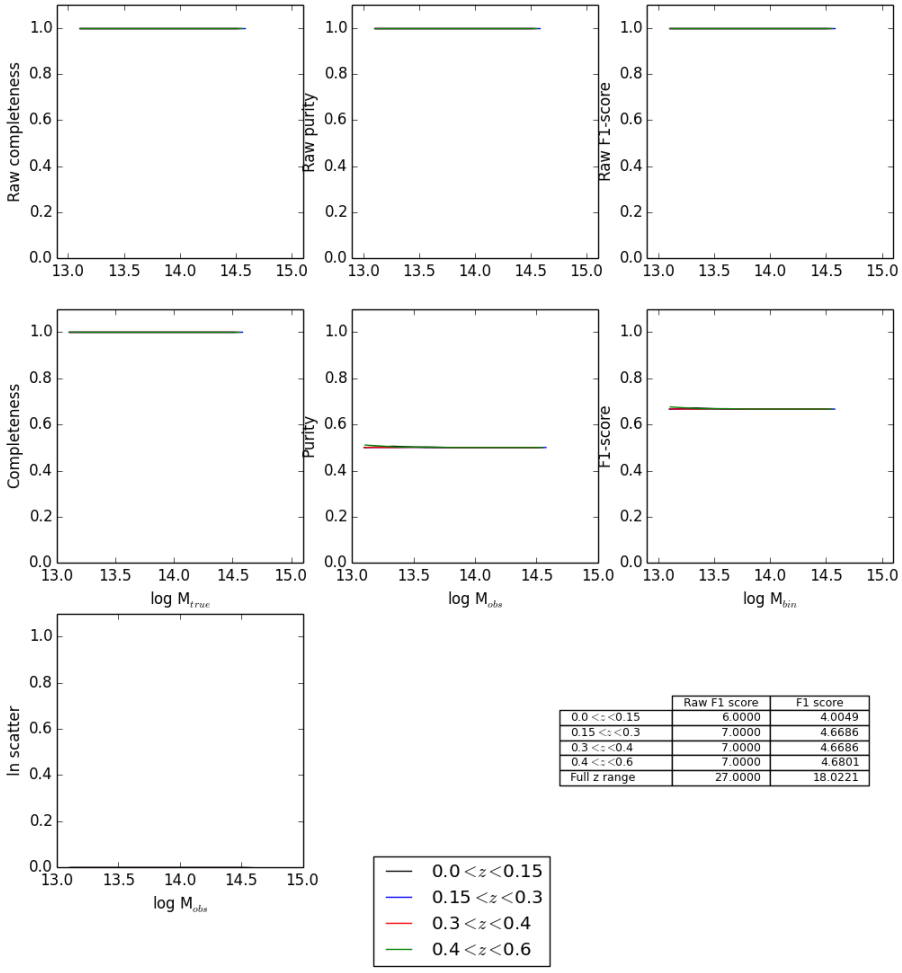
A null test was performed where the halo/galaxy catalog is matched to itself. The matching code returned 100% completeness and 100% purity across all mass-redshift bins. The global F1 score was given as F1 = 27, which represents the maximum possible score of any cluster catalog derived from this halo catalog, using these mass and redshift binning schemas (the F1 score of 27 implies there are 27 mass-redshift bins).

5.4.2 Impure and Complete Cluster Catalog

An impure but complete halo catalog was produced by dividing each cluster with more than one observable member into two parts with an approximately 3:1 redistribution of the constituent members. Figure 5.9 shows that whilst the 'Raw purity' is 100%, as one would expect since all clusters were derived from the true halos, the uniqueness constraint pushes this down to 50%, as only one of each derived cluster pair is permitted to be the best two-way match, indicative of the fragmentation in the derived cluster catalog. If raw purity were also 50%, that would imply that half of the clusters in the catalog do not relate to a halo (i.e., to a real cluster) at all. The global F1 score falls by roughly a third, from 27.00 to 18.02, as one would expect through Equation (5.3) in 27 mass-redshift bins.

5.4.3 Incomplete and Pure Cluster Catalog

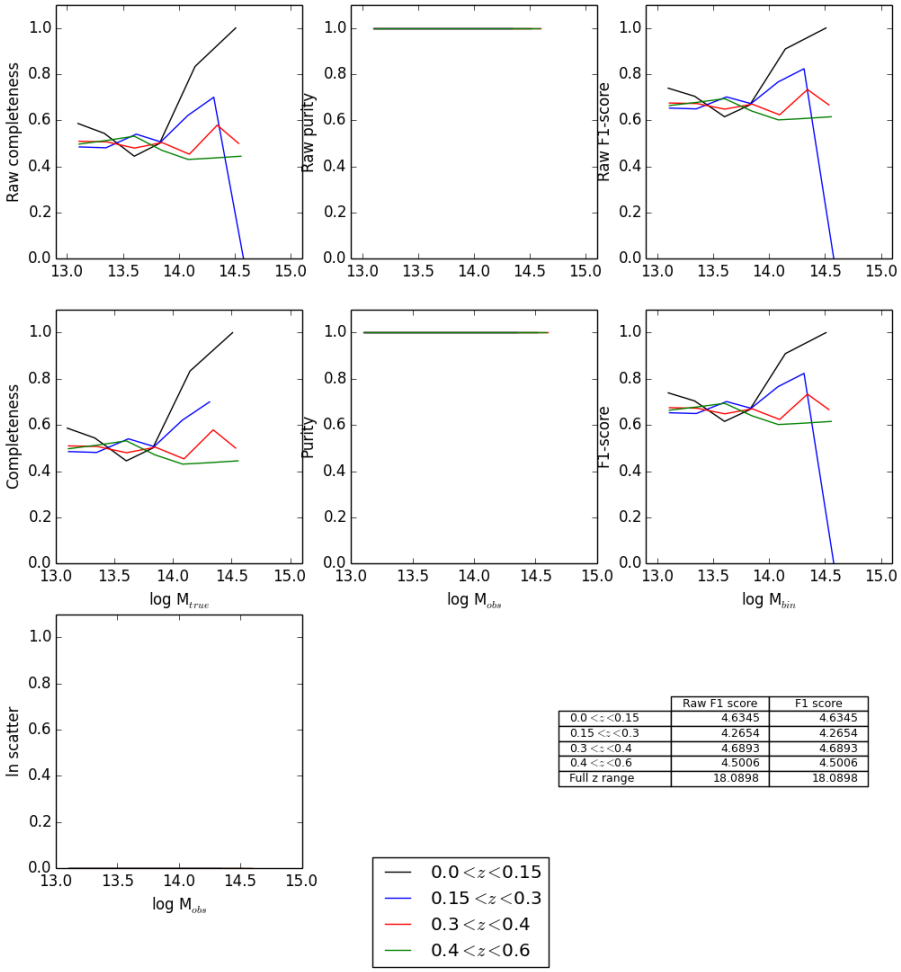
An incomplete cluster catalog was produced from the input CatSim catalog by ranking the clusters (halos) by M_{200} and, iterating through rank, removing every other cluster and its corresponding members from the catalog. This was then matched to the full catalog. Figure 5.10 shows that unlike the matching results for the impure catalog (§5.4.2 above), the incompleteness impacts the raw completeness measure and varies with both mass and



	Raw F1 score	F1 score
0.0 <z<0.15	6.0000	4.0049
0.15 <z<0.3	7.0000	4.6686
0.3 <z<0.4	7.0000	4.6686
0.4 <z<0.6	7.0000	4.6801
Full z range	27.0000	18.0221

Figure 5.9: Completeness and purity plots for a CatSim catalog generated at 50% purity. Subfigures are described as in Figure 5.7, §5.2.6. Observe that the raw purity does not fall to 50% because all the clusters can be matched to the true cluster population. Since the impure catalog is constructed by dividing each cluster into two parts, the uniqueness constraint applied to the second row of plots only allows for one out of two clusters to be matched to each halo, resulting in 50% purity.

redshift. The drop in raw completeness shows that these systems are completely missed by the constructed incomplete catalog. The variation of completeness by mass and redshift is because different proportions of clusters fall into each mass-redshift bin, the greatest variance occurring in the higher mass bins, where the fewest clusters can be recovered. If raw completeness were much higher than 50%, that would indicate the cluster catalog has blended the underlying halo (true cluster) population. The global F1 score is reduced by a third, from 27.00 to 18.09, consistent with Equation (5.3), and the impure catalog F1 score (§5.4.2).



	Raw F1 score	F1 score
0.0 <z<0.15	4.6345	4.6345
0.15 <z<0.3	4.2654	4.2654
0.3 <z<0.4	4.6893	4.6893
0.4 <z<0.6	4.5006	4.5006
Full z range	18.0898	18.0898

Figure 5.10: Completeness and purity plots for a CatSim catalog generated at 50% completeness. Subfigures are described as in Figure 5.7, §5.2.6. The completeness is around 50% in all mass-redshift bins, but varies due to the redshift bins not containing equal proportions of clusters of a similar mass. The incomplete catalog is constructed without reference to redshift bins and so this variance is expected.

5.4.4 Incomplete and Impure Cluster Catalog

An incomplete and impure cluster catalog was created by combining the techniques to produce the impure and incomplete catalogs described above (§5.4.2 and §5.4.3, respectively). The global F1 score is F1 = 13.42, which approximates the score of 13.5 expected through use of Equation (5.3) in 27 mass-redshift bins.

5.4.5 Ranking by $N_{\mathbf{gals}}$

To ensure the effect (or lack of effect) of ranking on the matching, the null test cluster catalog from section 5.4.1 was ranked using cluster $N_{\rm gals}$ as the ranking mechanism (§5.3.1). The completeness and purity were maintained at 100%, whilst the Spearman ranking coefficient (§5.2.2) delivers positive values below unity, as seen in Figure 5.11. The positive values for ρ imply that there is a general trend for $N_{\rm gals}$ rank to decrease with decreasing mass, but there is scatter in the mass-rank relation such that the $N_{\rm gals}$ rank does not monotonically decline with mass. The smaller values for $\rho_{log(M)>14}$ imply that $N_{\rm gals}$ is no better at ranking mass for higher mass cluster populations.

5.4.6 $N_{\mathbf{gals}}$ Limited Catalogs

The final step of the APERC4 algorithm is to cut on $N_{\rm gals}$, nominally to cut out spurious associations of galaxies. Excluding clusters with $N_{\rm gals} < 8$ from the null test cluster catalog (from §5.4.1) only affects completeness at masses below $log(M) < 13.75$ in the highest redshift bin ($0.4 < z < 0.6$) where completeness falls to 70% by $log(M) = 13.0$. In Figure 5.12, the $N_{\rm gals}$ threshold is raised to exclude clusters with $N_{\rm gals} < 32$. The catalog with $N_{\rm gals} \geq 32$ shows $> 90\%$ completeness above $log(M) > 14.0$ for redshifts below $z < 0.4$, with the highest redshift bin declining to $\sim 40\%$ at the same mass threshold. The F1 score reflects this by dropping from 26.7 in the $N_{\rm gals} \geq 8$ catalog to 20.3 in the $N_{\rm gals} \geq 32$ catalog.

This is the effect of an observer based condition on the output catalog. Whilst the previous tests have been concerned with the accuracy of catalog quality reporting, this test shows that by limiting clusters to those with more than $N_{\rm gals} \geq 32$, lower mass clusters are missing in significant proportions, with fewer clusters found at high redshifts.

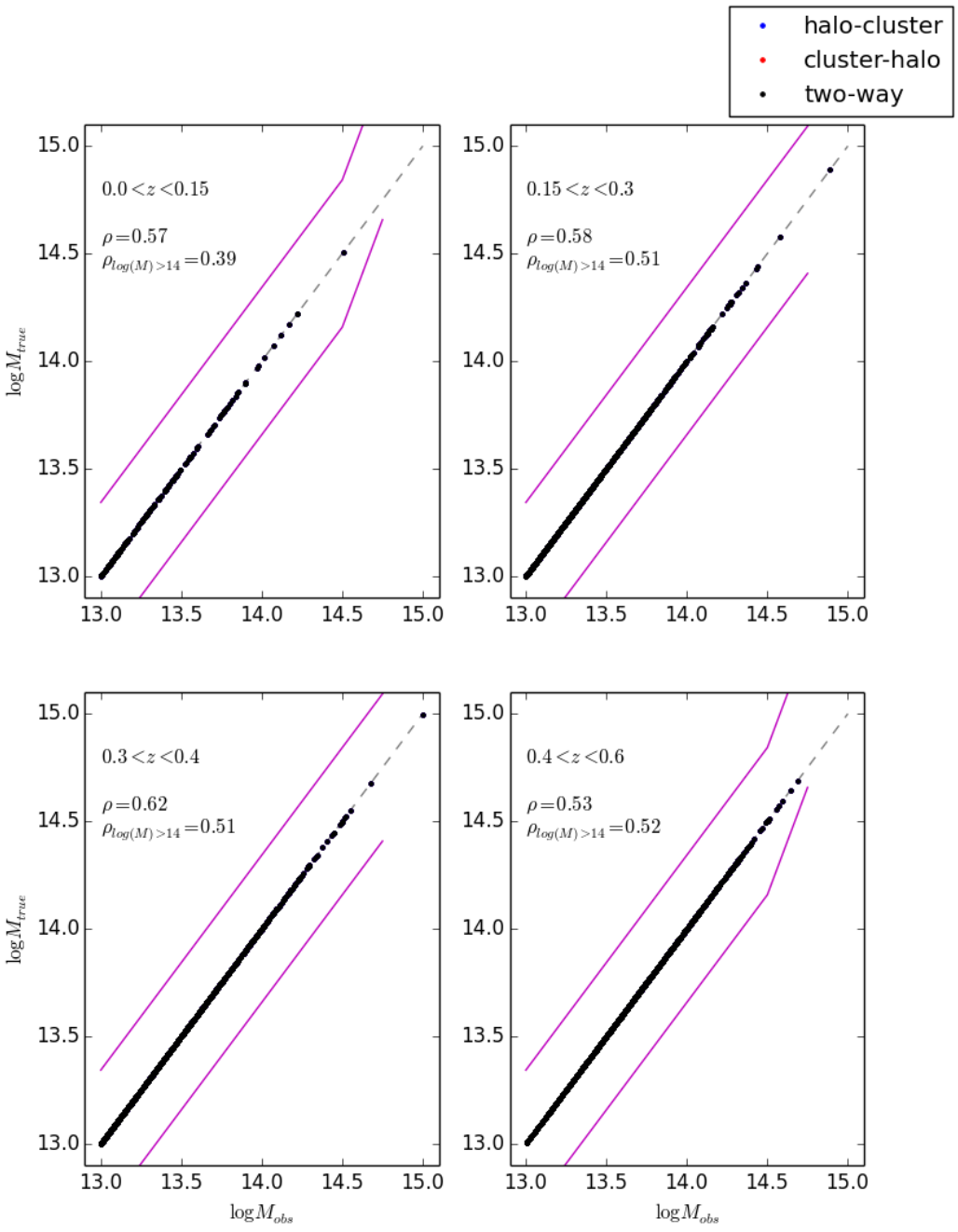
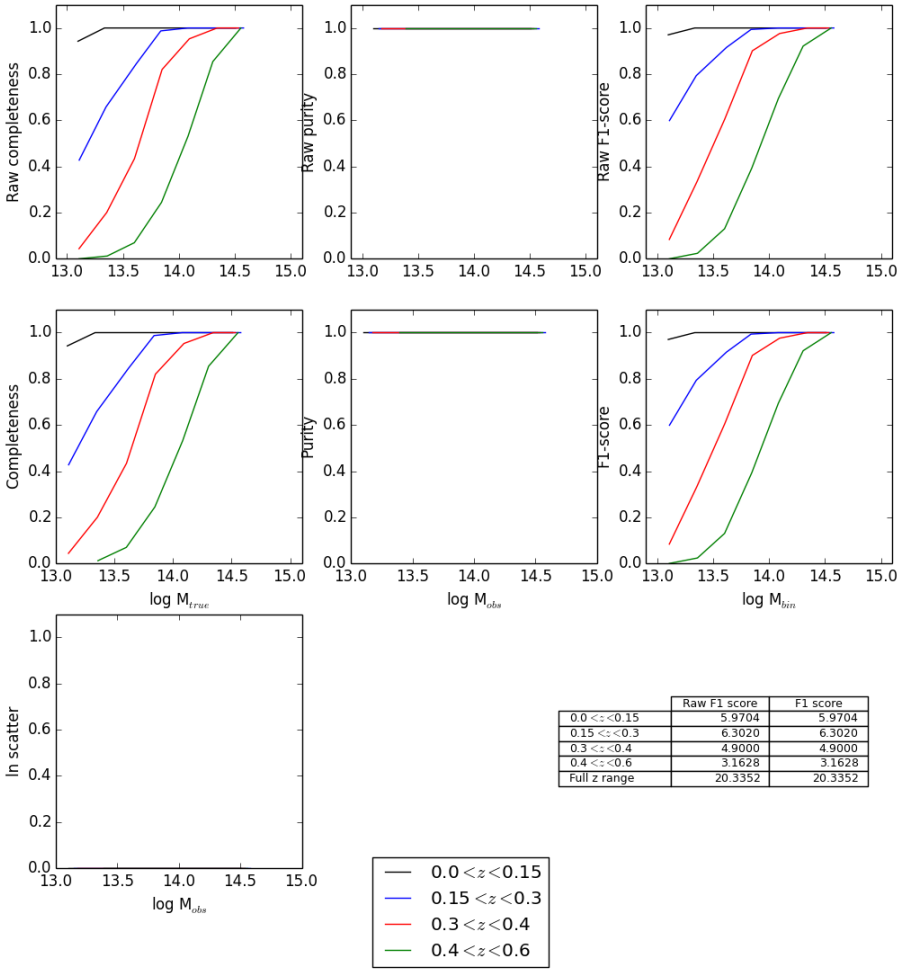


Figure 5.11: Mass scatter plots for a CatSim catalog ranked by $N_{\rm gals}$. Subfigures are described as in Figure 5.6, §5.2.6. The matched clusters make up the black dots along the $M_{obs} = M_{true}$ relation as they are calibrated to their uniquely two-way matched halo catalog counterpart. The values of ρ and $\rho_{log(M)>14}$ show that $N_{\rm gals}$ ranks mass in a consistent way across all redshift bins.



	Raw F1 score	F1 score
0.0 <z<0.15	5.9704	5.9704
0.15 <z<0.3	6.3020	6.3020
0.3 <z<0.4	4.9000	4.9000
0.4 <z<0.6	3.1628	3.1628
Full z range	20.3352	20.3352

Figure 5.12: Completeness and purity plots for a CatSim catalog limited to clusters where $N_{\rm gals} \geq 32$. Subfigures are described as in Figure 5.7, §5.2.6. The completeness of all redshift bins drops as a function of both mass and redshift, with completeness falling significantly in the highest redshift bin.

5.5 Summary

In this chapter, I introduced the framework I will be using to evaluate the cluster catalog produced with APERC4 run on simulation data. In section 5.1, I re-introduced the concepts of completeness and purity as measurements of cluster catalog quality, introduced the F-measure/F1 score as a single quality parameter for a given cluster catalog, and suggested that a uniqueness constraint will supplement the measure of a catalog's quality.

In section 5.2, I introduce the membership matching evaluation method as previously used by the DES Cluster Working Group and my alterations to that algorithm. I describe the link between halo membership and halo mass, membership matching between cluster finder and simulation, and how that is used to then evaluate cluster finding at different masses and redshifts. I briefly overview two ranking mechanisms in section 5.3, which allow cluster catalogs to be assessed in a context relevant to cluster cosmology.

In section 5.4, I test my version of the matching algorithm with some catalogs generated from the input halo catalog, with known imperfections and evaluate the the matching algorithms ability to evaluate cluster catalogs of different flavours. Of note, the F-score is shown to behave as predicted and is an extremely useful single point statistic that can be used for optimisation of any cluster finder.

Since clusters are tracers of the matter distribution of the universe, the completeness and purity measures of a cluster catalog have historically been made relative to some other catalog rather than some absolute definition. More recently, n-body simulations have matured to the point where observational tools can be trained on them. Being in a paradigm where every cluster finder works to 90% purity and completeness is down to people's wildly different definitions of purity and completeness (§1.6). The F-measure analysis in this chapter has proved to be a robust tool in assessing cluster catalog quality when using simulated information. With this in mind, and with the analysis tools presented in this chapter, I will be employing DES CatSim (Chapter 4), which has been produced with SDSS magnitude information, to obtain an optimal parameter set to use on SDSS DR8 (Chapter 6).

Chapter 6

Characterisation of AperC4 with SDSS Catalog Simulations

To develop AperC4, it is necessary to work with a well understood dataset that simultaneously tests cluster finding and verifies modifications to the algorithm. The Dark Energy Survey (DES) Catalog Simulation, or CatSim (§4.3), provides a representation of the universe that will be captured by the DES observational campaign, linking the cosmological distribution of dark matter halos to the observable galaxies that occupy them. Using the same simulations, SDSS magnitudes have been applied to the simulated galaxies, making it representative of SDSS DR8. In this chapter, using this simulated dataset, I apply AperC4, introduced in chapter 3, with the evaluation framework of chapter 5 to deliver an optimal parameter set for use with the SDSS DR8 proper (chapter 7).

In section 6.1, I review the zCarlos $p(z)$ method and overview its incorporation into the CatSim pipeline. In section 6.2, I apply AperC4 to CatSim, describing the input parameters tested in AperC4 cluster finding. In section 6.4, I demonstrate how the F-measure is used to identify an optimal parameter set and the impact of $p(z)$ quality on the output cluster catalog, before summarising in section 6.5.

6.1 CatSim Implementation of zCarlos $p(z)$

6.1.1 Introduction to zCarlos $p(z)$ Method

This thesis employs probabilistic redshift, or $p(z)$, information in the process of AperC4 cluster finding. Where a photometric source has been identified as a galaxy, $p(z)$ presents a series of redshift intervals (or bins) where the the likelihood of the galaxy belonging to each interval is evaluated such that the sum of these probabilities is $\Sigma p(z) = 1$, for

all galaxies. Specifically, this thesis focusses on the $p(z)$ method developed from a series of papers: Oyaizu et al. (2008), Lima et al. (2008), and Cunha et al. (2009), which are referred to as zCarlos* $p(z)$s. This is a training set based photo-z estimator (Csabai et al., 2003; Collister and Lahav, 2004), i.e., it uses a subset of photometric galaxies with spectroscopy to derive a relationship between spectroscopic redshift and their photometric observables such as their magnitudes or colours. For information on alternative methods, a recent comparison of several photometric redshift estimators (e.g., training-set, template, model-based) is presented in a paper by Sánchez et al. (2014).

The motivation for the zCarlos $p(z)$ construction method comes from Oyaizu et al. (2008), who show that biases in photometric redshift (photo-z) error estimates can be reduced by employing spectroscopic training data. They also show that catastrophic photo-z failures, i.e., where the photometric redshift is estimated to be significantly above or below the true redshift ($|z_{\rm photo} - z_{\rm true}| \gg \sigma_z$, Bernstein and Huterer, 2010; Padmanabhan et al., 2005), can be characterised by their large estimated photo-z errors. Oyaizu et al. (2008) split their spectroscopic sample into bins of magnitude in the five SDSS/DES bands, each containing around 100 galaxies, and establish a nearest-neighbour error estimate (i.e., characterizing the photo-z error, $|z_{spec} - z_{phot}|$, for photometric galaxies by their proximity in magnitude-space to the spectroscopic subsample) computed against a neural network photo-z code. In doing so, they show that their training set based error estimator is superior to template/model based analogs when trying to characterise the error distribution, and is robust even when the spectroscopic subset is not fully representative of the full photometric sample.

Building on this work, Lima et al. (2008) introduce a method to estimate the underlying $N(z)$ (galaxy count as a function of redshift) of photometric galaxy catalogs. Each galaxy is assigned a weighting in a set of redshift bins based on a nearest neighbour technique in multidimensional colour-magnitude space occupied by spectroscopic samples. This weighting method is somewhat analogous to how p-values are employed to measure the colour-clustering of galaxies in $C4_{M05}$ and AperC4 (§2.2.2 and §3.2.3, respectively). They stipulate a condition that the spectroscopic sample(s) must cover the same range of photometric observations as the galaxies being assigned $p(z)$s, i.e. the spectroscopic sample must occupy the same colour-magnitude space as the photometric sample. However, they demonstrate that the redshift distribution of the spectroscopic training sample does *not* have to be the same as the redshift distribution of the photometric sample being

*The zCarlos name is an informal reference employed by the DES simulations working group for Carlos Cunha's implementation of the method described in this section.

trained, thus offering advantages against training-set photo-z estimation methods that are sensitive to such biases. They comment that single-number photo-z estimates, where the $p(z)$ is narrowly peaked, offer precise (reduced scatter) and accurate (low bias) measurements. However, broad, skewed or multiply peaked $p(z)$s lead to poorer, and possibly catastrophic, single-number photo-z estimates. By constructing a $p(z)$ for each galaxy, they are able to avoid systematic errors in individual photo-z measurements that would collectively bias the $N(z)$ distribution.

Tests of the zCarlos $p(z)$ construction are outlined in Lima et al.'s sister paper, Cunha et al. (2009), where they deliver a catalog of $p(z)$s for $\sim$ 78 million SDSS DR7 galaxies. The $p(z)$ calculation algorithm was tested with SDSS DR6 galaxy data and on CatSim simulation data, albeit a much earlier version than the one used in this thesis (§4.3.3). With the observed data, they find that the quality of the $p(z)$ estimates degenerates for $r > 21.5$ recommending a cut in brightness of at least $r < 21.8$. They determine that this is due to the *faint* photometric sources being trained to solely spectroscopic DEEP/DEEP2 data (brighter spectroscopic sources are taken from a variety of surveys, including the spectroscopic SDSS campaign), leading to selection effects biasing the output $p(z)$s for individual galaxies.

6.1.2 zCarlos Applied to CatSim

Cunha et al. (2009) mentions that their simulated photometric sample, excluded 43% of the galaxies as they weren't well represented in their training set (i.e., lie outside the photometric space of the spectroscopic sample). For comparison, Cunha et al. say 98% of the real SDSS DR6 photometric galaxies with $r < 22$ are well represented by the spectroscopic sample. I note here that this discrepancy may affect results if AperC4 identifies galaxies from this excluded 43% as being cluster-like galaxies. At the time of writing, it is unknown if this problem exists in CatSim Aardvark v1.0.

The DES CatSim simulation used in this thesis is Aardvark v1.0 (§4.3.3), for which the zCarlos $p(z)$s have been constructed. The DES Simulations Working Group treated 149,370 simulation galaxies brighter than $r < 21.8$ as the spectroscopic training sample, to calibrate the zCarlos $p(z)$ code. $\sim$1.61 million galaxies were given $p(z)$s in $\sim 213\,\square^\circ$ of the Aardvark v1.0 CatSim catalog, using *ugriz* magnitudes to simulate SDSS DR8 information. It wasn't possible to test on a larger area due to the zCarlos information supplied by the DES Simulations Group having inconsistent IDs with the galaxies in CatSim in the larger catalog simulation.

Figure 6.1 shows that the total $p(z)$ information recovers the underlying $N(z)$ with much smaller scatter than treating the maximum likelihood of each galaxy $p(z)$ as the measured, single-point redshift. Indeed, utilising the full $p(z)$ reduces characteristic deficiencies of single photometric redshifts such as the typical underestimation of redshift for galaxies at high z, and the overestimation of redshift for galaxies at low z.

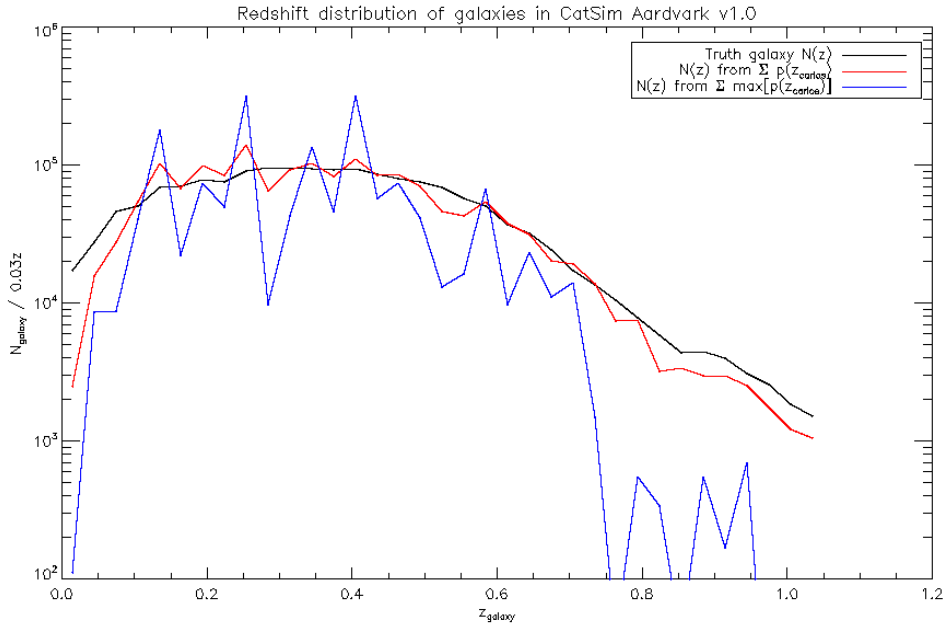


Figure 6.1: This plot shows $N(z)$ using the input redshifts of the simulated CatSim DR8 galaxies (black line), the $N(z)$ produced by totalling the simulation galaxy zCarlos $p(z)$s (red line), and the $N(z)$ produced by counting the peak of the each zCarlos $p(z)$ as the galaxy redshift (blue line). Using the full $p(z)$ information results in an $N(z)$ that recovers the underlying galaxy redshift distribution to a greater accuracy than assigning the maximum likelihood redshift to each galaxy.

6.1.3 zCarlos $p(z)$ of CatSim Halos

In section 3.2.7, I introduced the product-sum of the galaxy memberships as the basis for giving the apeture-slice clusters an initial $p(z)$. As a demonstration of this procedure, Figure 6.2 shows two halos at low and intermediate redshift ($z = 0.099$ and $z = 0.421$, respectively) where the constituent galaxy $p(z)$s have been combined to indicate the redshift of the cluster using the product-sum approach. By inspection, the individual $p(z)$s of each galaxy in a cluster are not necessarily consistent with one another, but once consolidated, the redshift of the underlying halo becomes apparent. For comparison, the median of the individual $p(z)$s is also plotted. For the lower redshift halo, this peaks in the same location as the product-sum $p(z)$, whilst the example at higher redshift doesn't peak as strongly and displays multiple peaks.

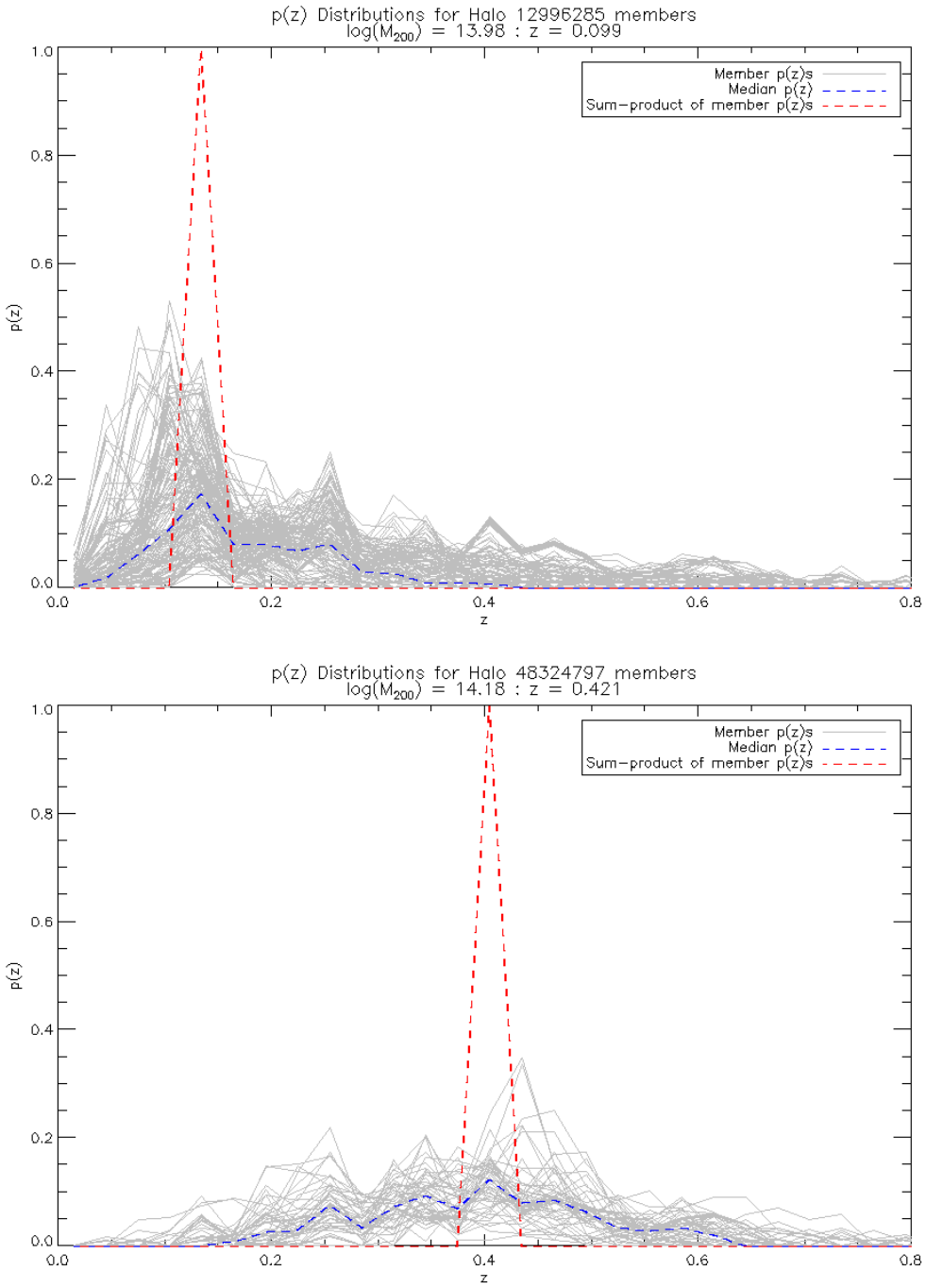


Figure 6.2: These plots show the $p(z)$ distributions for two CatSim halos. The lines represent the member galaxies (grey), the median of the member galaxy $p(z)$s (blue), and the product-sum of the member galaxy $p(z)$s (red). The upper panel shows a halo at $z = 0.099$, whilst the lower panel shows a halo at $z = 0.421$, of mass $\log(M/\mathrm{M}_\odot) \sim 14$. In both cases, the redshift of the halo is strongly indicated by the peak sum-product redshift bin.

The product-sum is, however, limited to the redshift resolution of the $p(z)$ binning schema in the CatSim employed, $z_{\mathrm{bin}} = 0.03$, hence the uncertainty on the product-sum $p(z)$ has a lower limit of $\pm\Delta z = 0.015$ around the median redshift of each bin.

I note here that the $p(z)$s produced for CatSim are adjusted before they are combined

in the product-sum. As can be seen in Figure 6.2, some of the galaxy $p(z)$ values go to zero, indicating with certainty the galaxy does not exist in a given redshift bin. But as 6.1.1 explains, this is the result of assigning zero weight to a galaxy's probability in a given redshift bin, which should be interpreted as there being no evidence for the galaxy lying at this redshift, rather than no possibility of the galaxy lying at this redshift. Indeed, such a result would mean that the product-sum could be biased, eliminating the probability of multiple galaxies being in a redshift bin if any individual $p(z)$ goes to zero in that bin. For the purpose of disallowing such events, the zero $p(z)$ bins are assigned the minimum non-zero $p(z)$ value in the whole $p(z)$ catalog, then each $p(z)$ distribution is renormalised to unity. This has the effect of retaining the weighting schema of Cunha et al. (2009) between the previously non-zero $p(z)$ bins, but at the expense of expressing that the minimum non-zero $p(z)$ value does not contain any useful redshift information relative to other non-zero $p(z)$ bins.

6.2 AperC4 Applied to CatSim

In chapter 3, I detailed the improvements made to the $C4_{M05}$ algorithm, highlighting the updated mechanisms. In this subsection, I will apply AperC4 to CatSim (§4.3.3) with the parameters in Table 6.1. Then, using the analysis tools described in chapter 5, the optimal AperC4 parameters are chosen in section 6.3.

These tests are performed on a $\sim 213\,\square^\circ$ segment of CatSim Aardvark v1.0. This is partially to facilitate processing of the data multiple times over a range of parameters within a short turnaround time, but also due to IDs between the $p(z)$ files and the main catalog files not correlating across the whole survey area.

With reference to the complete parameter list and descriptions in Table 3.1, the parameters test with AperC4 are given in Table 6.1. In this table, the aperture range refers to the set of aperture-slices that are consolidated during membership optimisation (§3.2.8). Three sets of apertures are chosen with *a*) the full set of apertures given in Table 3.3; *b*) a subset of those apertures, chosen such that the corresponding physical scale does not exceed 1.5 Mpc at higher redshifts ($0.5 < z < 1$, assuming WMAP-9 cosmology); and *c*) a single aperture of $2.5'$, to act as a null test of the aperture-slice consolidation.

The cluster cross-$p(z)$ threshold values are chosen such that they represent scenarios where identified clusters that shared members, and have similar $p(z)$ distributions, are merged often to rarely (a cross-$p(z)$ threshold of 0.00 being the most often, and a cross-$p(z)$ threshold of 0.9 being least often) or not at all ($p(z)$ threshold = 1.1). Note that a

cross-$p(z)$ threshold does not include the threshold value, since at a cross-$p(z)$ threshold of 0.00, that would simply mean the rejection of all $p(z)$ information and merge all structures with shared members.

Parameter	**Values**		
Aperture range	9.0′, 6.3′, 5.0′, 3.7′, 3.0′, 2.7′, 2.5′, 2.2′, 2.0′	3.0′, 2.7′, 2.5′, 2.2′, 2.0′	2.5′
FDR	0.2	1.0	
	4.0	10.0	
kth Nearest Neighbour	6	12	24
Cluster cross-$p(z)$ threshold	0.00	0.01	
	0.30	0.60	
	0.90	1.10	
N_{gals} cut-off	8	16	24
	32	40	48
	64		

Table 6.1: Set of parameters (see Table 3.1 for details on their specific properties), and their values, tested on $\sim 213\,\square^\circ$ of the Aardvark v1.0 CatSim galaxy catalog with the AperC4 algorithm. Each parameter (*left column*) takes one value (*right column*), with the exception of the *Aperture range* parameter, where multiple aperture sizes are processed in a single AperC4 run (§3.2).

The remaining variables from Table 3.1 are set as follows:

`MAGLIM` is kept at 0.0, such that the magnitude limits come from the `SURVEY_LIMITS` parameter, alone.

`SURVEY_LIMITS` are taken from analysis of the Sheldon et al. (2012) SDSS galaxy sample (§7.1.3) and applied to CatSim (§4.3.3). The apparent magnitude limits are set to

$$[ugriz]_{\text{mag}} = [22.21, 21.19, 20.65, 19.94, 19.44]$$

.

`SPLITSIZE` **and** `OVERLAP` are set to 10° and 2°, respectively, ensuring tile sizes are at minimum $(100 \times \cos(\text{dec}))\,\square^\circ$ and at most $(144 \times \cos(\text{dec}))\,\square^\circ$. The declination of

the CatSim sample is around $-30°$, so these tile sizes approximate to $70\,\square°$ and $102\,\square°$, respectively.

`SIGMACOLOUR` is set to $X_{xy} = 2.0$, informed by tests on $C4_{M05}$ when given photometric redshift information.

`GAMMAFACTOR` is set to $\gamma = 0$, using the reasoning that the dispersion in colour across the survey is mitigated by dividing the survey into tiles (as discussed in §3.3.4).

6.3 Optimal Parameter Selection with $p(z)$ Information

The optimal combination of parameters (parameter set) is determined by maximisation of the F-measure, introduced in section 5.1.3. For the F1 score maximisation, I allow the cluster catalog to match only to halos above $\log(M/\,\mathrm{M}_\odot) > 13.5$, such that spurious detections of halos below that mass threshold are not made, and so tests the cluster ranking mechanism, i.e., if low mass halos are found, then the corresponding clusters should be ranked low. The F1 score is calculated in bins of $\Delta \log(M/\,\mathrm{M}_\odot) = 0.25$. Furthermore, I limit the F1 score to bins above $\log(M/\,\mathrm{M}_\odot) \geq 13.75$ (or $M \gtrsim 6e13\,\mathrm{M}_\odot$) as these are the halos of interest for calibrating cosmology (Wu et al., 2010).

To evaluate the matching, the clusters are divided into redshift bins by the location of the peak of their respective $p(z)$s. This may introduce additional error, as this simplifies the $p(z)$ distribution in a manner that is opposed to the intention of avoiding redshift bias in the galaxy $p(z)$ distributions (§6.1). However, the nature of membership matching (§5.2.1) still allows clusters and halos to match with each other even if their redshifts differ by more than one bin. Differences in redshift will be characterised in the centering figure (e.g., Figure 5.8).

In the following subsections, I will look at optimisation of AperC4 using zCarlos $p(z)$ information and verifying the optimal parameter sets by running the analysis with the simulation redshift information (which I refer to as *ideal*, or *idealised*, redshift information).

6.3.1 Optimisation using zCarlos

Employing the zCarlos $p(z)$ information, Figure 6.3 shows the maximum F1 scores for any value combination of two parameters, keeping all other parameter values free.

The optimal parameter set used in chapter 7 is described by the left-most column of Table 6.2. The next four near optimal parameter sets and accompanying F1 scores using the zCarlos redshifts are given to add context to the discussion in section 6.3.3. Given

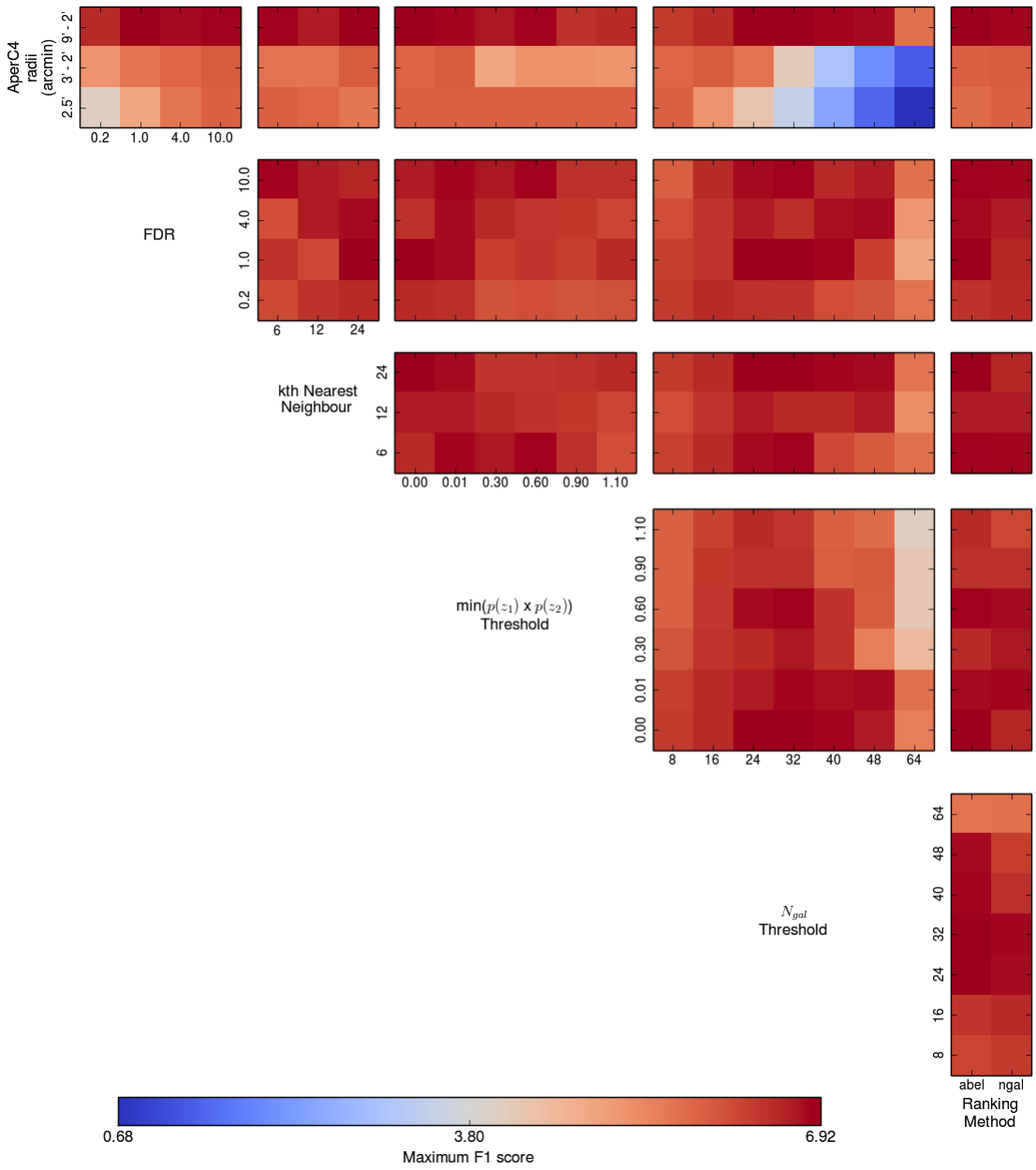


Figure 6.3: This 6-D plot shows the maximum F1 score from 3024 cluster catalogs, found for any combination of two parameters from Table 6.1, when employing zCarlos $p(z)$ redshift information. The colour bar on the bottom indicates the range of maximum F1 scores when fixing two parameters ranges between 0.68 (dark blue) and 6.92 (deep red).

that the matching code will deliver F1 scores that deviate from the expected F1 score by as much as 0.09 during testing (e.g. F1 scores of 18.02 and 18.09 for expected F1 scores of 18.00 for impure and incomplete catalogs respectively, §5.4), the optimal catalog parameter set should be amongst the top three parameter sets. I will discuss the optimal parameter set in section 6.3.3, after examining the optimal parameter sets found using idealised $p(z)$ information.

Parameter	**Value**				
F1 score	6.920	6.913	6.846	6.819	6.806
Aperture range	$9' - 2'$	$9' - 2'$	$9' - 2'$	$9' - 2'$	$9' - 2'$
FDR	1.0	1.0	10.0	1.0	10.0
kth Nearest Neighbour	24	24	6	24	6
Cluster cross-$p(z)$ threshold	0.00	0.00	0.60	0.00	0.01
N_{gals} cut-off	32	24	32	40	32
Ranking method	Abell	Abell	Abell	Abell	N_{gals}

Table 6.2: The top five optimal AperC4 parameter set trained on $\sim 213\,\square^\circ$ of the Aardvark v1.0 CatSim galaxy catalog using zCarlos $p(z)$ redshift information. The leftmost column constitutes the optimal parameter set.

6.3.2 Idealised $p(z)$ Information

To consider the impact of zCarlos $p(z)$ on the output cluster catalog quality, the analyses of output catalogs was repeated using the input simulation redshifts of the galaxies converted into an idealised $p(z)$. Ideal $p(z)$ information was generated with the same structure as the zCarlos $p(z)$ files. This represents a 'best case scenario' model for aperture defragmentation and AperC4 cluster redshift estimation with a zCarlos-like $p(z)$.

All the bins in each ideal $p(z)$ are set to $p(z_{\text{bin}}) = 0.01$, except the bin of the galaxy's redshift, which is set to 1 and then the ideal $p(z)$ is renormalised to 1. AperC4 was then run on the full range of parameters.

The top eight optimal parameter sets (within 0.09 of the maximum F1 score) and accompanying F1 scores using ideal redshift information are given in Table 6.3.

Using the results of the idealised $p(z)$ information allows us to consider the impact of redshift quality on the output AperC4 catalog (§6.3.3).

6.3.3 Optimal Parameter Set

Aperture range from $9'$ to $2'$ is most favoured by the F1 score (first row of Figures 6.3 and 6.4). This can be seen across all tested parameter combinations with this aperture range. The next most favoured is the $3'$ to $2'$range, with the single $2.5'$aperture being least favoured. AperC4 clearly performs better when it has more information on the colour clustering of galaxies, i.e., by including galaxies that appear colour

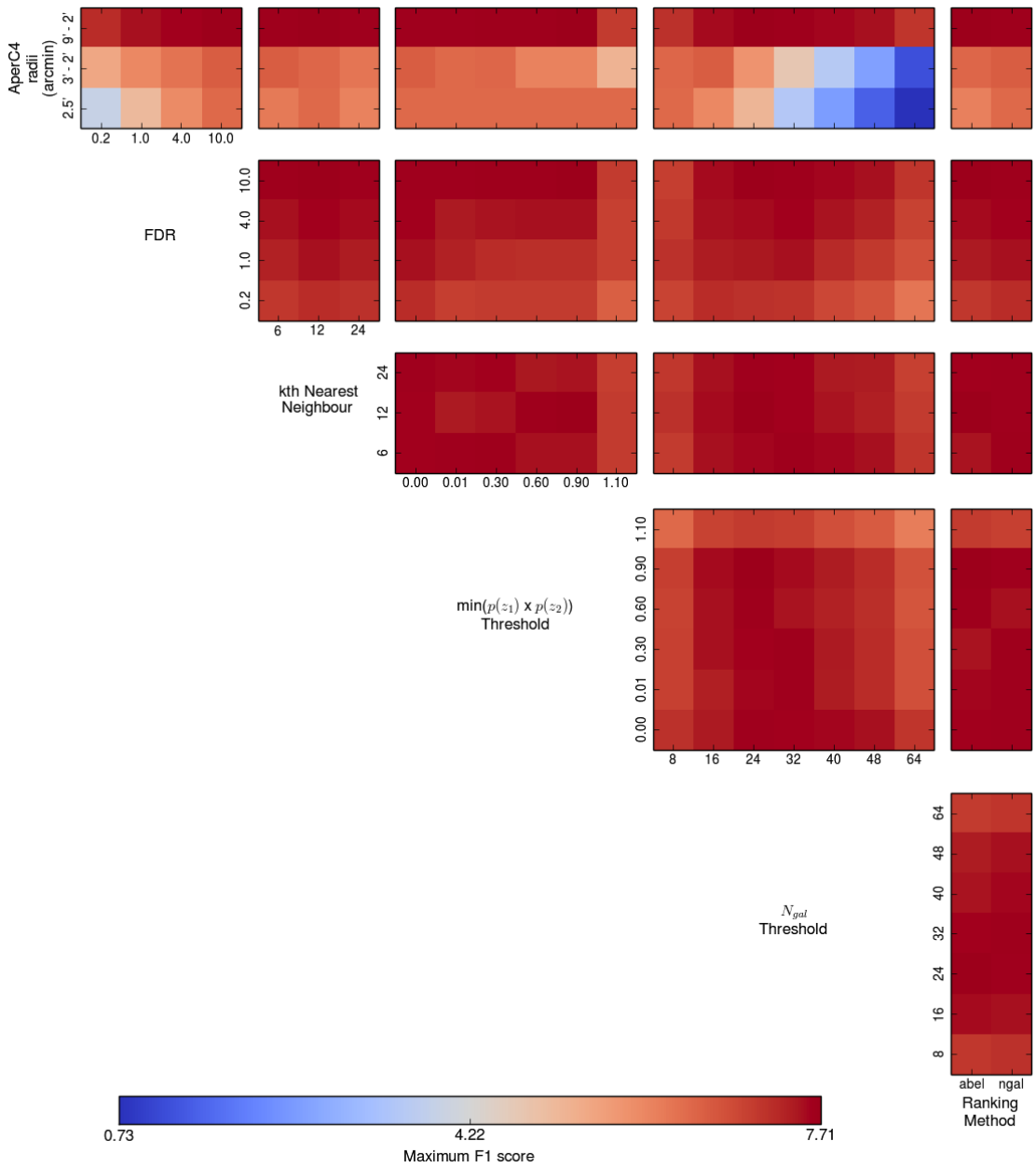


Figure 6.4: This 6-D plot shows the maximum F1 score from 3024 cluster catalogs, found for parameter combinations from Table 6.1 using a ideal $p(z)$ derived from the simulation redshift. The colour bar on the bottom indicates the range of maximum F1 scores when fixing two parameters ranges between 0.73 (dark blue) and 7.71 (deep red). [Note there is a difference in the range of these F1 scores to Figure 6.3]

clustered in larger aperture sizes. Note that the physical scale represented by these larger apertures can be far larger than 1 Mpc at advanced redshifts, as can be seen in Table 3.3 (§3.3.1).

FDR of 1% is most favoured by the F1 score (second row/first column of Figures 6.3 and 6.4). The FDR acts to reduce the number of field galaxies in the AperC4 aperture-slices, but at the expense of reducing the number of cluster galaxies (i.e.,

Parameter	**Value**							
F1 score	7.715	7.672	7.664	7.648	7.645	7.638	7.630	7.628
Aperture range	$9' - 2'$	$9' - 2'$	$9' - 2'$	$9' - 2'$	$9' - 2'$	$9' - 2'$	$9' - 2'$	$9' - 2'$
FDR	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
kth Nearest Neighbour	12	6	12	6	12	24	6	24
Cluster cross-$p(z)$ threshold	0.90	0.30	0.60	0.01	0.90	0.00	0.00	0.00
N_{gals} cut-off	24	32	24	32	24	24	32	32
Ranking method	Abell	N_{gals}	Abell	N_{gals}	N_{gals}	N_{gals}	N_{gals}	Abell

Table 6.3: The top eight optimal APERC4 parameter sets trained on $\sim 213\,\square^{\circ}$ of the Aardvark v1.0 CATSIM galaxy catalog using ideal $p(z)$ information generated from simulation redshifts.

decrease the number of false rejections at the expense of decreased power; see Figure 2.6). In Figure 6.3, we see that the range of FDR thresholds only weakly affects output catalog quality with no apparent trend between $1\% < \text{FDR} < 10\%$, and $FDR = 0.2\%$ being mildly disfavoured against the other FDR values.

If we compare the zCARLOS runs to the ideal $p(z)$ runs in section 6.3.2, we find that there is a slight trend towards favouring a higher FDR, i.e. allowing a larger proportion of non-cluster galaxies allows the capture of more true cluster galaxies. The impact of having less redshift information means that even with a larger population of true cluster galaxies, the increased presence of non-cluster galaxies causes the cluster $p(z)$s to be less well defined, leading to spurious associations being defined as clusters.

kth Nearest Neighbour of $k = 24$ is the optimal value favoured by the F1 score (third row/second column of Figures 6.3 and 6.4). There is no clear trend favouring one value of k over another, suggesting this parameter does not significantly affect APERC4 cluster finding.

Cluster cross-$p(z)$ threshold of cross-$p(z) = 0.00$ is the optimal value favoured by the F1 score (fourth row/third column of Figures 6.3 and 6.4). Like FDR, this varies very gently across the parameter combinations, displaying a slight preference towards low thresholds when using the zCARLOS $p(z)$. Other than where merging is disallowed, the optimal cross-$p(z)$ threshold for the catalog produced with ideal $p(z)$ shows very

little preference for low or high cross-$p(z)$thresholds, indicating that the clusters which do straddle bins, have some shared non-zero probability outside of the peak $p(z)$ bin, which allows them to merge. Where merging is disallowed (cross-$p(z)$ = 1.1), the catalog is clearly non-optimal when using the ideal galaxy redshifts, which is mostly due to clusters straddling the $p(z)$ bin edges remaining fragmented. This effect is less dramatic when using zCarlos $p(z)$s since (by design) they tend to be less constrained towards a single redshift.

N_{gals} cut-off of 32 galaxies per cluster is the optimal value favoured by the F1 score (fifth row/fourth column of Figures 6.3 and 6.4). The lower N_{gals} cut-off F1 scores imply that completeness increases less than purity decreases when lowering the N_{gals} threshold. Above an $N_{\mathrm{gals}} = 32$ threshold, the F1 scores are driven by fewer real and false clusters where purity increases less than completeness decreases.

Ranking method the Abell ranking is favoured by the F1 score across all parameters (fifth column of Figures 6.3 and 6.4), and is therefore the optimal ranking. Interestingly, the Abell ranking favours higher N_{gals} thresholds, which is part of its historical definition for galaxy clusters (§1.2.1), but the degree to which this trend is observed is minimal.

6.4 Discussion

6.4.1 AperC4 Catalog Statistics with zCarlos Redshifts

Having determined the optimal parameter set in section 6.3.3, the optimal catalog will be examined using the matching framework (presented in chapter 5) in this section. In order to bin by redshift, the $p(z)$s of the AperC4 clusters have been reduced to their peak value, which may incur added scatter (§6.1.2) when looking at the redshift separations but does not affect matching as this is done by membership.

Approximately two-thirds of the matched clusters are two-way in non-unique matching. Figure 6.5 shows the non-uniquely matched clusters and halos are fragmented with respect to one another. AperC4 cluster finding below $z < 0.1$ is poor, which is to be expected where the aperture extents probe $\lesssim 1\,h^{-1}\,\mathrm{Mpc}$ scales (assuming WMAP9 cosmology; see Table 3.3) at the upper limit of the lowest redshift bin. The presence of blue points along the vertical axis indicates AperC4 is not finding these halos. Below $z < 0.4$ very few halos $\log(M/\,\mathrm{M}_{\odot}) \gtrsim 14$ are missed by AperC4 completely, whilst above $z > 0.4$, significant numbers of halos are also missed up to $\log(M/\,\mathrm{M}_{\odot}) \lesssim 14.75$. The red points along the

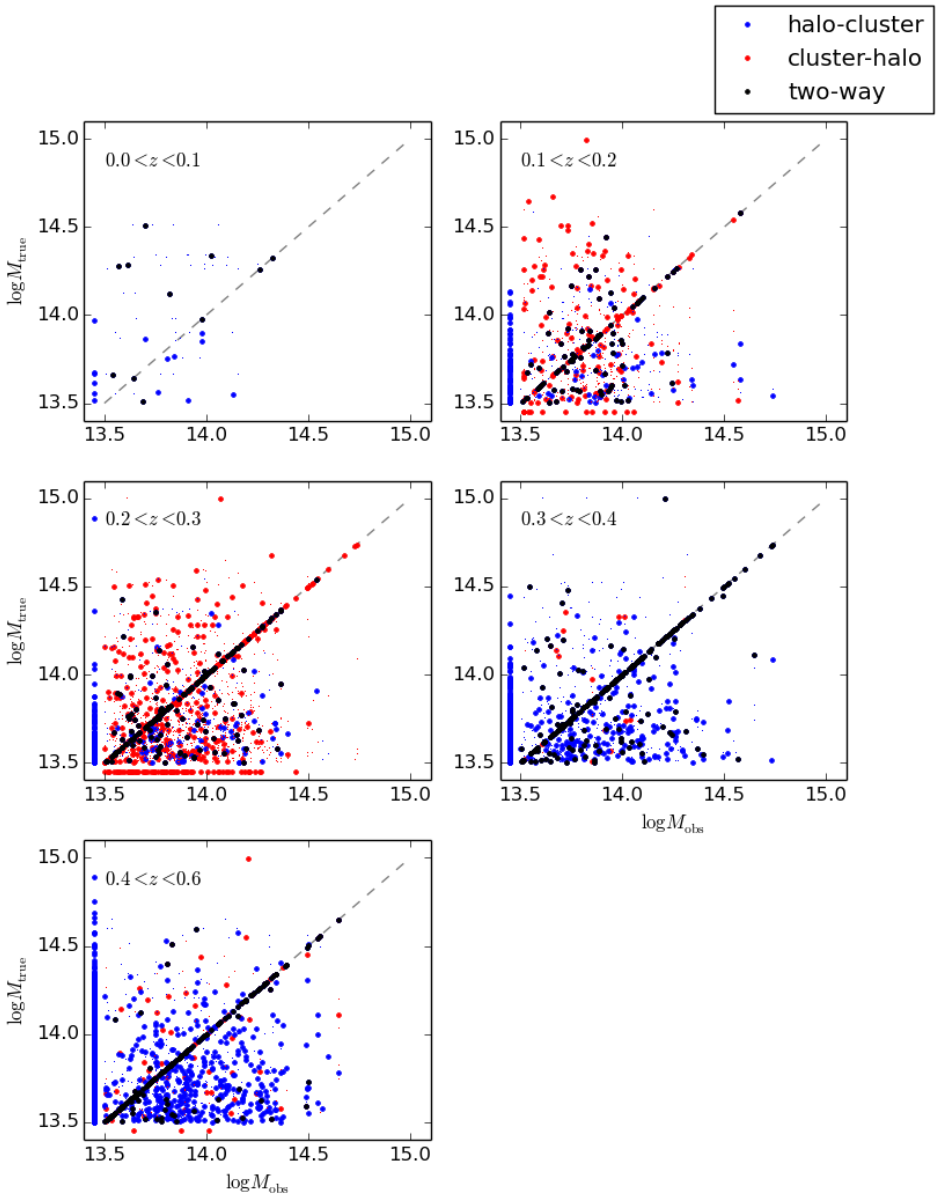


Figure 6.5: This plot shows the non-unique scatter of matched ApERC4-CatSim cluster-halos as per Figure 5.5 (§5.2.6) from the optimal ApERC4 catalog. At redshifts $0.3 < z < 0.4$ and $z > 0.4$, the number of halo-to-cluster matches (blue points) outnumbers the number of cluster-to-halo matches, indicating the cluster catalog tends to blend halos in these redshift bins. In the $0.1 < z < 0.2$ and $0.2 < z < 0.3$ bins, the situation appears reversed and multiple clusters match to single halos, indicating fragmentation in this bin.

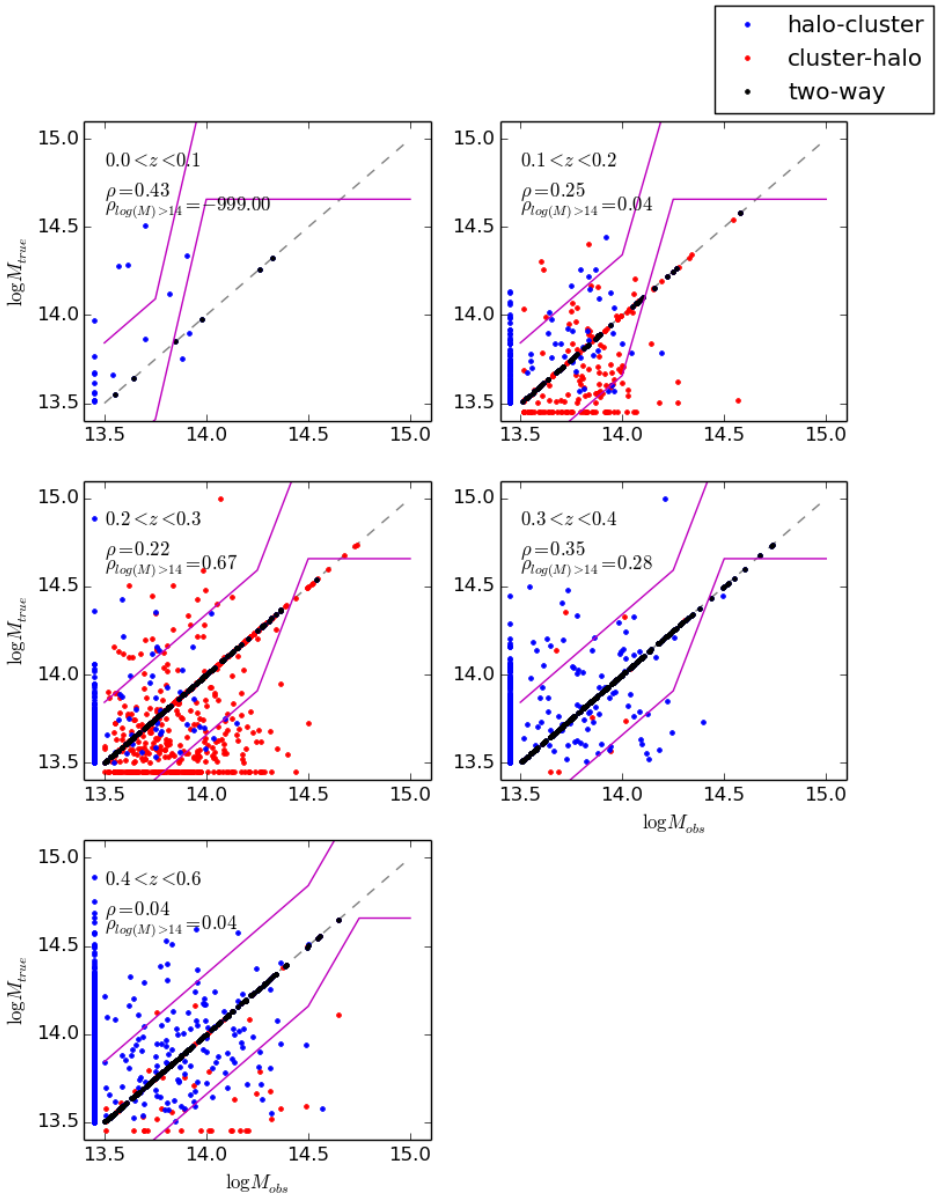


Figure 6.6: This plot shows the unique scatter of matched ApERC4-CatSim cluster-halos as per Figure 5.6 (§5.2.6) from the optimal ApERC4 catalog. ApERC4 can be seen to produce clusters ranked below $\log(M/\,\mathrm{M}_\odot) \lesssim 14$ that do not uniquely match to any halo or redshifts below $z < 0.4$ indicating fragmentation occurs in these regimes. Above $z > 0.4$ the fragmentation occurs up to $\log(M/\,\mathrm{M}_\odot) \lesssim 14.5$. The misfit scatter lines (purple) above $\log(M/\,\mathrm{M}_\odot) \gtrsim 14.5$ are because the bins aren't occupied well enough by non-two-way matches in this mass range to get a proper handle on scatter limits (the two-way matches in these bins are calibrated to the halo mass, hence there is no scatter by definition).

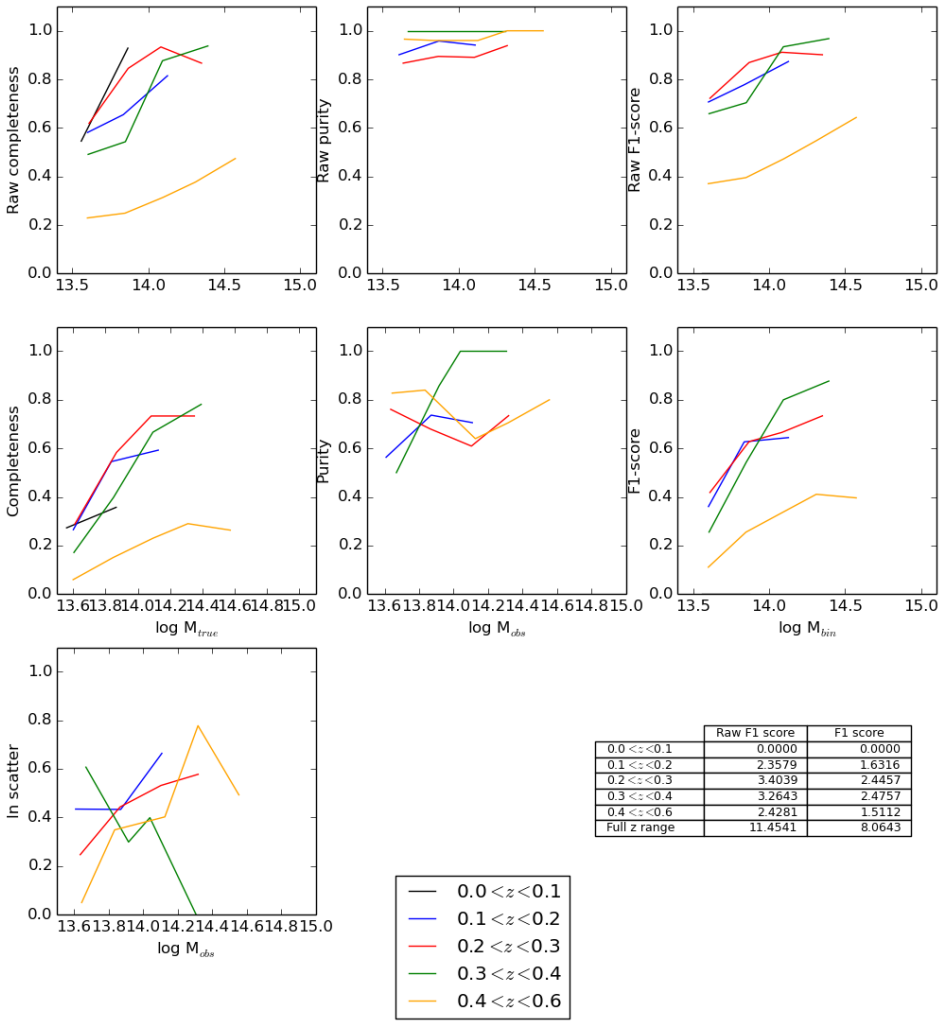
horizontal axis indicate APERC4 clusters that do not match any halo in this sample, and are thus false associations. Such false associations only appear to occur in significant proportions in the $0.2 < z < 0.3$ bin. The apparent abundance of cluster-halo matches in lower redshift bins, and halo-cluster matches in the higher redshift bins, indicates line-of-sight fragmentation, where the off-relation (i.e. $M_{\rm obs} \neq M_{\rm true}$) low z clusters describe a blend of low-z and high-z halos. This distribution of matching type by redshift also indicates that cluster redshifts tend towards redshift ranges that cover these bins. The apparent paucity of the plots indicates either a general dearth of clusters found (such as in the $0.0 < z < 0.1$ bin), or indicates that the majority of clusters are well described by (two-way matched to) their halo counterparts (such as in the $0.3 < z < 0.4$ bin).

The addition of the uniqueness constraint reduces the fraction of two-way matches to half the cluster sample. Figure 6.6 indicates that most of the poorly (non-unique) matched clusters occur where the clusters are ranked below $\log(M/\,{\rm M}_\odot) \lesssim 14$, with the exception of the highest redshift bin. At increasing redshift, the redshift bins include an increasing fraction of halos at the edge of detectability, i.e., they contain very few galaxies that are simulated within the magnitude limit of the SDSS, decreasing APERC4's ability to detect them as clusters.

Looking at the Spearman ranking coefficients (indicated by ρ in the figures; §5.2.2), the Abell ranking shows a positive correlation with mass at redshifts below $z < 0.4$, indicating that the ranking very broadly differentiates (positive but non-unity) between high mass and low mass clusters in these redshift bins. This correlation is close to zero in the highest redshift bin, indicating that Abell ranking is ineffective in this regime. The scatter fit (shown by the purple lines) is poor for high mass clusters/halos due to their lack of numbers, indicating that a larger sample is required to assess the scatter properly.

In Figure 6.7, raw completeness is behaving as one would expect, referring to earlier tests on matching with an $N_{\rm gals}$ limited catalog (§5.4.6), with low mass/high redshift objects being less readily detected than high mass/low redshift objects. Halos in the $0.4 < z < 0.6$ are found less effectively than the lower redshift bins. Raw purity is $\gtrsim 90\%$ in all redshift bins, which implies that nearly all of the clusters in the catalog match to at least one halo (i.e., contain at least one halo member) above $\log(M/\,{\rm M}_\odot) \gtrsim 13.5$. The F1-score on the right is the harmonic mean of these distributions.

The addition of the uniqueness constraint on matching shows a degree of fragmentation and blending in the cluster catalog. Completeness drops by around 20% as a fraction of clusters contain some blended halos. Relating to raw completeness, this blending appears



	Raw F1 score	F1 score
0.0<z<0.1	0.0000	0.0000
0.1<z<0.2	2.3579	1.6316
0.2<z<0.3	3.4039	2.4457
0.3<z<0.4	3.2643	2.4757
0.4<z<0.6	2.4281	1.5112
Full z range	11.4541	8.0643

Figure 6.7: This plot shows the purity, completeness, and F1 measures of matched AperC4-CatSim cluster-halos as per Figure 5.7 (§5.2.6) from the optimal AperC4 catalog. The redshift ranges are given in the key, and the mass bins are $\delta \log(M) = 0.25$ in size.

to be independent of redshift, whilst at lower masses, unique completeness drops to around 20%, indicating that the lower ranked/mass clusters are more likely to be blended. Purity drops to around 70% in most redshift bins with the inclusion of the uniqueness constraint, indicating that halos are getting fragmented across the survey, irrespective of mass or redshift. Considering all the purity information, the high raw purity indicates that whilst nearly all of the clusters relate to some halo object, some clusters identify some halos multiple times.

Scatter is small since clusters are designated the mass of the halo they are uniquely two-way matched to. The log-mass scatter falling to zero may be indicative of sampling problems, as the higher mass bins will contain fewer clusters/halos than lower mass bins, meaning that individual clusters with larger than average deviation from their true mass will have a more sizable effect on the measured scatter in the higher mass bins than at lower masses.

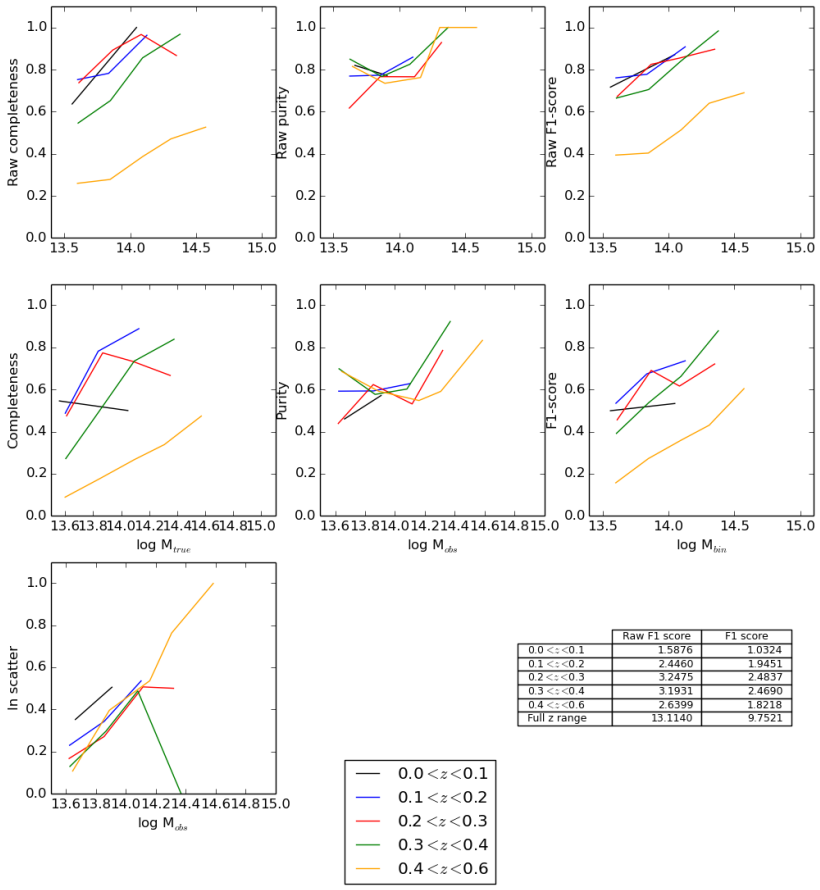
Figure 6.9 shows that defining the galaxy with the minimum k-NN distance as the cluster centre will tend to be offset from the true centre of the simulated halo. The majority of these points come from low mass systems, as seen from the distribution of points in the matching scatter plots above, and so may be indicative of fragmentation of the halo populations at the halo edges. Similarly, the displacement of cluster-halo and halo-cluster matches in redshift space indicates that high redshift halos are being matched to low redshift clusters, and/or cluster redshifts are biased low. The banding exhibited in the z versus Δz plot indicates that the combined $p(z)$s tend to peak in these locations.

6.4.2 AperC4 Catalog Statistics with Ideal Redshifts

The distribution of maximum F1 score by parameter combination in Figure 6.4 appears to be consistent with the optimal parameter set from the zCarlos $p(z)$s in Figure 6.3.

Idealized $p(z)$ Purity and Completeness Behaviour

In Figure 6.8, the distribution of purity and completeness using the ideal/simulated $p(z)$ catalog generated with the optimal parameter set (Table 6.2 in §6.3) shows some differences to the optimal catalog matching (generated with zCarlos). The optimum catalog with the ideal $p(z)$ looks worse, whilst the catalog behaves more consistently across all redshift/mass bins. The raw completeness has increased slightly, reaching near 100% above $log(M) > 14.0$. This is because the optimal ideal redshift catalog has a relaxed FDR of 10%, which includes a greater number of cluster-like galaxies at the expense of an increased



	Raw F1 score	F1 score
0.0 <z< 0.1	1.5876	1.0324
0.1 <z< 0.2	2.4460	1.9451
0.2 <z< 0.3	3.2475	2.4837
0.3 <z< 0.4	3.1931	2.4690
0.4 <z< 0.6	2.6399	1.8218
Full z range	13.1140	9.7521

Figure 6.8: This plot shows the purity, completeness, and F1 measures of matched AperC4-CatSim cluster-halos as per Figure 5.7 (§5.2.6) from the optimal AperC4 catalog but with *ideal* $p(z)$s, generated from the simulation redshift, substituted for the zCarlos $p(z)$. Figure 6.7 is the equivalent plot using zCarlos $p(z)$.

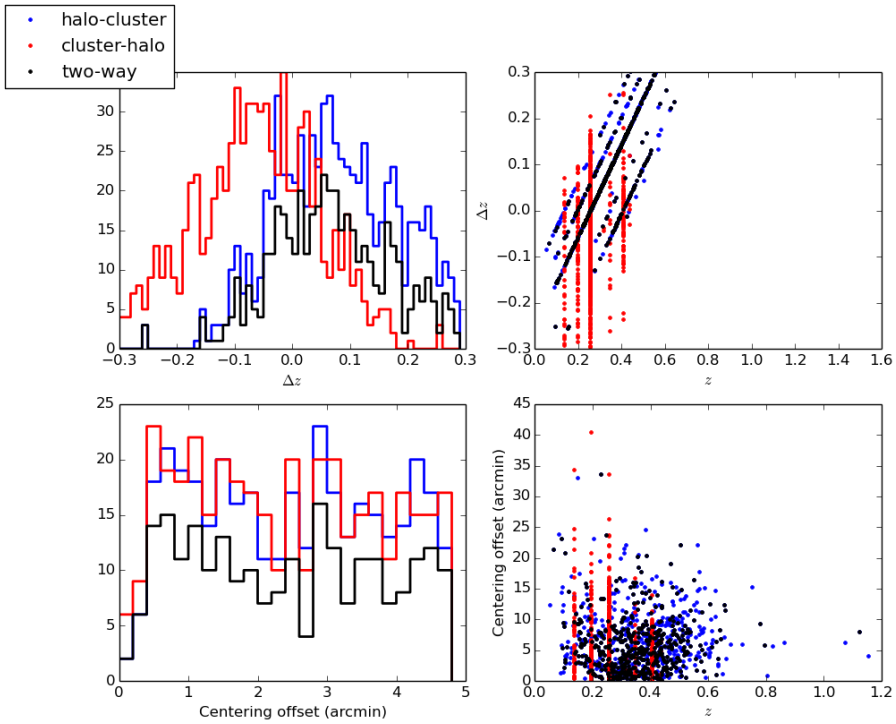


Figure 6.9: This plot shows the centering of matched AperC4-CatSim cluster-halos as per Figure 5.8 (§5.2.6) from the optimal AperC4 catalog. The upper left plot shows that the unique two-way matches and the halo-cluster matched halos are typically at higher redshift than the clusters they match to. Conversely, the cluster-halo matched clusters tend to be slightly below the correct halo redshift. The diagonal lines forming the two way matches in the upper right is a consequence of the quantised binning of redshift by the $p(z)$. It is important to note that this deviation is against the peak of the AperC4 cluster $p(z)$s and hence measuring a delta in this way is an over-simplification of the $p(z)$ (for the purpose of visualising this difference). The lower two plots show the centering offset of the cluster from its matched halo (and/or vice versa) in arcminutes. The lower plots show the matches are not well centred. A significant high redshift population of halos that are not two-way matched can be seen in the lower right plot, which also shows that AperC4 is able to find real clusters (unique two-way matched to simulation halo) above redhshifts $z > 0.6$.

fraction of contaminating field-like galaxies (§2.2.3).

As the galaxies have generally been moved into the cluster with the correct redshift binning, contaminating galaxies will either be isolated or put into small groups in various redshift bins, and thus excluded by the $N_{\rm gals}$ cut-off. Groups of contaminating galaxies that are not excluded by the $N_{\rm gals}$ cut-off join real clusters or form false associations, which *will* negatively affect catalog purity. Raw purity falls to $\sim 80\%$ below $log(M) \lesssim 14.0$, in the ideal $p(z)$ matching plots, indicating $\sim 20\%$ of the clusters in this mass range are false associations.

Comparing to the AperC4-zCarlos clusters, those clusters that were populated by galaxies from multiple halos have been divided into clusters at the proper halo redshift in the AperC4-ideal $p(z)$ catalog, and false associations of contaminant galaxies that occur randomly. As we allowed a 10% FDR, the maximum expected contamination is 10% in terms of galaxies, and since the Abell ranking is derived from $N_{\rm gals}$, the resultant decrease in purity (with mass proxy) is expected for decreasing $N_{\rm gals}$/mass as the false associations should not be expected to occur with greater $N_{\rm gals}$ than the simulated halos. This expectation draws from the fact that the contaminant galaxies in these false associations really do belong in the same redshift bin. Thus, any false association that is constrained within an aperture and redshift interval, and is shown to be colour clustered, is a questionable result of either the AperC4 cluster finding or the CatSim halo definition.

Idealized $p(z)$ Centering Behaviour

Figure 6.10 shows that fixing the redshift of the galaxies, and thus the derived redshift of the clusters, has the expected result of decreasing the redshift scatter of the clusters. The centering of clusters in the celestial plane also improves upon the zCarlos generated centres, despite no centring adjustment to the clusters in either ideal or zCarlos AperC4 catalog. The matches in this distribution are mostly unique, and so many of the cluster-halo matches in the optimal AperC4-zCarlos catalog no longer uniquely match to any halo, and are thus excluded from the centering evaluation.

6.4.3 Comparison between zCarlos and Ideal Redshift Cluster Catalogs

The F1-scores favour the ideal case cluster catalogs, peaking at 7.72 compared to 6.92 for the zCarlos catalogs. This indicates that whilst AperC4 is able to identify cluster galaxies well (based on the raw purities of the zCarlos and ideal catalogs; §6.4.1 and §6.4.2, respectively), a decrease in the quality of redshift information has a measurable

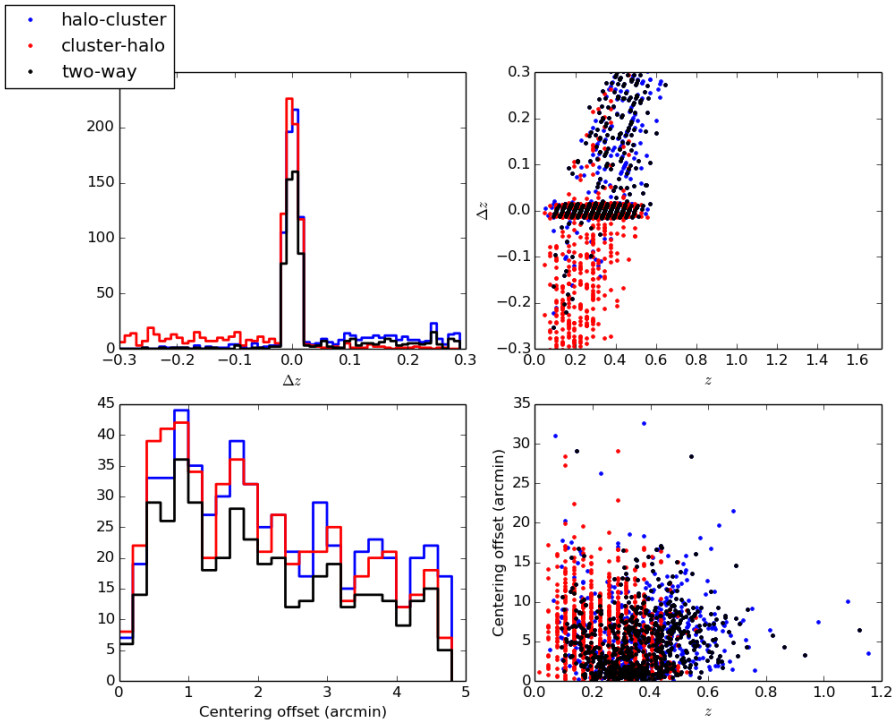


Figure 6.10: This plot shows the centering of matched AperC4-CatSim cluster-halos as per Figure 5.8 (§5.2.6) from the optimal AperC4 catalog but with a ideal $p(z)$, generated from the simulation redshift, substituted for the zCarlos $p(z)$. Figure 6.9 is the equivalent plot using zCarlos $p(z)$.

negative effect on the output AperC4 clusters. In terms of the F1 score itself, both zCarlos and ideal AperC4 catalogs optimise to approximately the same parameter set (§6.3 and §6.3.2). The exception in the trends is FDR, where having the true redshift information (i.e., ideal information) allows one to include more cluster and contaminant galaxy populations (§6.4.2), resulting in more clusters found at a rate consistent with the purity of zCarlos.

Examination of the completeness/purity distributions with the uniqueness constraint shows that the completeness of the ideal $p(z)$ cluster catalog is close to that of the zCarlos cluster catalog, implying that when AperC4 does find real clusters, it describes them fairly well. Examining the unique purity curves shows that an increasing number of false clusters are identified with decreasing mass, with approximately one false association for every three real halos. Comparison of the ideal $p(z)$ unique purity curves with the enhanced zCarlos $p(z)$ unique purity curves implies that the clusters produced with zCarlos carry a larger proportion of contaminant galaxies.

The decreased scatter in the redshifts of cluster centres when using the ideal $p(z)$ (Figure 6.10) compared to zCarlos $p(z)$ (Figure 6.9), is because the true redshift of the member galaxies is known, so the AperC4 cluster produced from these galaxies lies at

the same redshift. The apparent improvement in celestial centering is because the cluster galaxies are moving into the "correct" cluster during the defragmentation step (§3.2.8). Contaminant galaxies tend to move into smaller groups (as shown in Figure 6.8) that are either excluded by the $N_{\rm gals}$ cut, or simply do not uniquely match to any halo. Finally, because the redshifts are known, the halos above $z > 0.6$ are well identified, whereas the AperC4-zCarlos catalog fails to identify these clusters uniquely (Figure 6.9), revealing the limitations of the zCarlos $p(z)$ in this regime.

The limited area (213 $\square^\circ$) in which this analysis is performed means that it is prone to shot noise due to the limited number of halos in the higher mass bins across all redshifts. The area limitation being due to inconsistencies between galaxy identifications in the CatSim and zCarlos catalogs (§6.1.2). However, the optimal parameter set of the ideal $p(z)$ AperC4 catalog broadly agrees with that of the zCarlos-AperC4catalog, suggesting that the optimal parameter set constructed using zCarlos is close to, if not exactly the same as, the optimal parameter set for AperC4 on an SDSS DR8-like galaxy survey.

6.5 Summary

In section 6.1, I introduced the zCarlos $p(z)$ and its application to the DES CatSim (specifically, the SDSS DR8 realisation of CatSim) showing how $p(z)$s can be used to reconstruct $N(z)$ and discussing their treatment in AperC4 cluster finding. In section 6.2, I applied AperC4 to the CatSim realisation of DR8 over 213 $\square^\circ$over multiple parameter combinations, then utilised the F-measure to select an optimal parameter combination in section 6.3.

Then, in section 6.4, I qualitatively examined the AperC4-CatSim cluster catalog produced with the optimal parameter combination utilising the matching algorithm described in chapter 5. I also explored AperC4 catalog realisations over the same parameter combinations where AperC4 is given ideal $p(z)$s, generated from the simulation redshifts. This allowed me to evaluate the impact of $p(z)$ quality on the output cluster catalog, which in turn allowed better characterisation of the AperC4 catalog's successes and weaknesses. The use of simulated information is arguably the best way to to assess cluster catalog quality. Since there is no way to objectively know what's in the sky, data-based completeness/purity measures are self-defeating as there is only one sky, and previous cluster catalogs may well have undocumented features that make them less than reliable.

From this chapter, there now exists an optimal parameter set for AperC4 on simulated

SDSS DR8 data. In the following chapter, I will apply AperC4 to the observed SDSS DR8 with this well characterised, optimal parameter set to obtain a cluster catalog.

Chapter 7

The SDSS DR8-AperC4 catalog

Having determined an optimal set of AperC4 parameters, by using F-measure for an area that simulates the SDSS galaxy survey (described in chapter 6), AperC4 is applied to real SDSS data to find galaxy clusters.

In section 7.1, I overview the SDSS DR8 galaxy catalog and the Sheldon et al. (2012) galaxy selection used to construct an accompanying $p(z)$ catalog. I investigate this galaxy selection and introduce further cuts before applying AperC4 to the data in section 7.2. In section 7.3, I present the final AperC4-SDSS DR8 cluster catalog, overviewing its structure and employ GAMA data (Driver et al., 2010) to a subsample of clusters to establish the success and failure modes of AperC4. In section 7.4, I discuss the findings of section 7.3 and how they can be used to feed back to the AperC4 algorithm, before summarising in section 7.5

7.1 SDSS DR8

7.1.1 Introduction to DR8

SDSS DR8 represents the 8th Data Release of the SDSS (SDSS-III collaboration et al., 2011), and marks the first data release from the SDSS III campaign. DR8 totals 14,555 $\square^\circ$ of photometric coverage and 9,274 $\square^\circ$ of spectroscopic coverage (see §2.1.1). The DR8 catalog employs an alternative photometric calibration algorithm to the one employed by SDSS DR2 (§2.1.1). It simultaneously solves for calibration parameters and relative stellar fluxes by using overlapping SDSS observations: the *übercalibration* method (Padmanabhan et al., 2008). Übercalibration works by decoupling the problem of "absolute" calibrations (the translation of a measurement to a physical flux) from "relative" calibrations (internally consistent measurements) across the entire SDSS. The method employs

priors, in addition to standard stars, to reduce variation between disconnected regions. The advantage to APERC4 cluster finding that these übercalibrated magnitudes give, is the reduction of systematic error (Equations 3.1 [§3.2.3] and 2.19 [§2.2.2]) between two galaxies (of similar colours at the same redshift in different regions of the survey) down to a 1-2% level. This allows a greater degree of reliability in the p-value estimation, which evaluates a given galaxy's relative colour density (as explained in section 2.2.3) and motivates the exclusion of the systematic component of the colour-aperture (by setting $\gamma = 0$).

During the SDSS I & II campaigns, three stripes in the south galactic cap were observed: one along the celestial equator and two others north and south of the equator. The equatorial stripe (Stripe 82, Annis et al. 2011) was observed repeatedly with the intention of finding variable objects and, when co-added, to reach a magnitude limit roughly 2 magnitudes deeper than the main SDSS sample (York et al., 2000). The Stripe 82 imaging and co-added catalog was released alongside the Seventh Data Release (DR7, Abazajian and Survey 2008) of the SDSS, marking the completion of the SDSS II phase of the survey. DR8 included 2,395 $\square^\circ$ of additional imaging around the Southern Galactic Cap to give 3,173 $\square^\circ$ contiguous coverage in this region.

The Southern Galactic Cap will also be covered by the DES (Flaugher, 2005), as well as being an area of interest for other surveys covering a range of wavelengths (e.g. radio (VLA, Hodge et al. 2011), SZ (ACT, Reese et al. 2011), infrared (UKIDSS, Warren et al. 2007; VHS, Banerji et al. 2008), optical (BOSS, Eisenstein et al. 2011; CFHTLS, Brimioulle et al. 2008) and X-ray (Stripe 82 X, Urry 2010)). In covering this area, galaxy photometry can be trained, employing spectroscopic data for training redshifts and calibrating absolute photometry between surveys. The multi-wavelength coverage of this region also offers the opportunity of deeper insight into the cosmological, astrophysical and statistical properties of the extragalactic information contained inside the shared survey area.

A DR8 cluster catalog is highly desirable as a means to investigate cluster populations, at greater depth, with the well characterized SDSS hardware/software calibrations and pipelines. Cross-wavelength information on these clusters can then be used as a means to evaluate mass proxies / astrophysical properties against one another. I note that cluster catalogs for DR8 already exist in the form of redMaPPer (a red sequence method, Rykoff et al., 2013), and a Friends-of-Friends algorithm (Wen and Han, 2013). APERC4 is complementary to these catalogs by not being a red sequence detection method or relying on photo-zs for cluster galaxy identification.

7.1.2 Sheldon et al. Galaxy Selection

The $p(z)$ information utilised in this analysis is sourced from Sheldon et al. (2012) (henceforth referred to as S12), which are the application of the zCarlos $p(z)$method (§6.1) to SDSS DR8. Therein, users of the DR8 $p(z)$ catalog are advised to "choose a subset of the data that suits their needs" and that the catalog "should be a superset of objects that can be further trimmed." In this subsection, I will first describe the Sheldon et al. (2012) superset galaxy selection, before introducing my own selection subset.

Star-galaxy separation is achieved by determining the difference between the $r-$band `modelMag` and `psfMag` (described in §2.1.1) such that

$$r_{\rm PSF} - r_{\rm model} > 0.135, \tag{7.1}$$

which is slightly relaxed from the DR2 spectroscopic galaxy target selection in Equation 2.7 (§2.1.2). S12 mention that at their magnitude limit of $r = 21.8$, that the stellar contamination is relatively large. They determine stellar contamination of a few percent at $r = 21$, increasing to 10% at $r = 22$.

All objects identified in SDSS are each assigned 59 different flags during source extraction and identification, which can be used to reject objects identified as galaxies that may be unreliable and/or poorly measured. The S12 identified "galaxy" sources are run through a preliminary cleaning, using the standard "clean photometry" selection*, which removes sources that are: duplicated in overlapped fields (not flagged as `PRIMARY`); detected below a 5σ significance in the original imaging frame, the $\times 2$, or the $\times 4$ binned images (`BINNED`{1,2,4}); where a half-light radius could not be measured (`NOPROFILE`); that are marked as a blend but not deblended, usually due to proximity to the frame edge or because the object is too large (`BLENDED && NODEBLEND`); that are too close to the edge of a frame to be satisfactorily centroided (`PEAKCENTER`); where the pixel was *not* examined for an object (`NOTCHECKED`); where objects have been measured twice due to their high flux and thus duplicated (the rejected duplicate is flagged as `BRIGHT`); where the object is too bright for the detector (`SATURATED`); and where a source has been deblended into more than 25 objects (`DEBLEND_TOO_MANY_PEAKS`).

Note that, by definition, `PRIMARY` selected sources include a combination of the flags `!BRIGHT && (!BLENDED || NODEBLEND || nchild == 0)`, such that the source is neither duplicated nor can be deblended further.

S12 excise objects where the extinction corrected (Schlegel et al., 1998) model flux is not well determined in at least one photometric band, and require objects to be detected

*http://cas.sdss.org/dr4/en/help/docs/realquery.asp#flags

in at least one band in $[ugriz] < [21, 22, 22, 20.5, 20.1]$, demanding a 5σ detection in both the r and i bands (through the `BINNED`{1,2,4} flag). The `cModelMag` $r$band is restricted to a range within $15 \leq r_{\rm cModel} \leq 21.8$, and limiting the extinction corrected `modelMag` to the range $[15.0, 29.0]$ to ensure reasonable galaxy colours.

The source catalog is celestially limited to the BOSS footprint (see Figure 7.7). Objects near stars in the TYCHO2 catalog (Høg et al., 2000) were also excised using a magnitude dependent masking radius. Finally, objects found in images where the SDSS u amplifier was not working* were removed.

The S12 catalog contains 58,533,603 objects.

7.1.3 Galaxy Selection for AperC4

Stoughton et al. (2002) detail shortcomings of the SDSS source extraction/identification pipeline (known as `frames`), mentioning difficulties such as: cosmic-ray rejection; electronics cross-talk producing faint ghosts in z-band; the lack of diffraction spike subtraction/interpolation; faint ghosts from the physical edges of the filters (and urges suspicion of the reality of sources fainter than $r > 24$ mag arcsec^2); deblending of low surface brightness features from large galaxies; and ghosting in the u- filter introducing some systematic error into the flat fielding of the images. They give a disclaimer that the problems they identify are by no means exhaustive, and whilst the `frames` framework had been regularly updated as the SDSS timeline progressed, some issues may still impact the DR8 catalog data.

Further to the S12 galaxy selection (§7.1.2) above, further cuts were investigated. A random sample of 5 million S12 DR8 galaxies were then used to follow-up and evaluate prospective cuts.

Flags

The flags applied to S12 are very thorough, with the `PRIMARY` and `NOTCHECKED` flags being the most effectual selections in reducing the input DR8 galaxy catalog.

The `MAYBE_CR`, and `MAYBE_EGHOST` flags were investigated. This `MAYBE_CR` flag signals an object may be a cosmic ray, describing objects detected in a single-filter that have not been interpolated over, i.e., the object's measurements are detected in the data and are uncorrected (as opposed to detections that *are* flagged as containing cosmic rays and interpolated over). `MAYBE_EGHOST` indicates that the object may be an electronics ghost of

*http://www.sdss.org/dr7.1/start/aboutdr7.1.html#imcaveat

a bright star in the given band. Objects flagged with these keywords that did fall into the S12 selection make up slightly less than 1% of the S12 catalog. Eyeball inspection of ~ 100 flagged objects at various magnitudes was undertaken, revealing no source that could be attributed to cosmic rays or to the electronics replication of a nearby bright star, and thus the flag may be attributed to the quality of the detection in a non-optimal band. As the S12 selection demands detection in both r and i bands, the likelihood of these two flags suggesting either genuine electronics ghosts or cosmic rays is low. As such, no additional flags are used to constrain the S12 galaxy catalog.

Surface Brightness

Based on the suspicions of Stoughton et al. (2002), sources with a r-band half-light surface brightness, μ_r, fainter than

$$\mu_r > 24.5 \text{mag arcsec}^2 \qquad \text{for} \quad r_{\text{petro}} \leq 19, \text{ and} \tag{7.2}$$

$$\mu_r > 5.5 + r_{\text{petro}} \qquad \text{for} \quad r_{\text{petro}} \geq 19, \tag{7.3}$$

were inspected for improper identifications (see purple points in Figure 7.1). However, the removal of the images where the u amplifier was malfunctioning (already undertaken by S12) appears to have reduced the presence of ghosting in the S12 catalog in this surface brightness regime. Remaining galaxies appear to have either a very red SED, or a fairly blue SED, implying that the low r-band surface brightness of these objects is a consequence of the object having a spectrum that peaks in redder, or bluer, bands, respectively. As such, no faint surface brightness limits are applied to the S12 sample.

The green points in Figure 7.1 describe a noticeable band of galaxies that extends from an r-band half-light surface brightness of $\mu_r \gtrsim 16$ to $\mu_r \lesssim 20.5$ mag arcsec^{-2} and an r-band magnitude of $r_{petro} < 20.5$ mag, which originates from a small number of bright stars satisfying the S12 star-galaxy separation criteria (Equation 7.1), resulting in their misclassification (Strauss et al., 2002). The stellar track is cut until it reaches the main bulk of the galaxy population in Figure 7.1, which may result in a fraction of real galaxies being cut along with stars in the stellar track. With reference to Figure 7.2, examination of these 'galaxies' reveals that the bright end is primarily populated with very close binary stars, which are complicated to deblend successfully. The faint end shows objects where the galaxy-star appearance of the source isn't readily distinguishable by eye, and so the fainter end of this cut describes some distribution of both stars and galaxies.

This stellar locus cut is implemented and excises 45,717 sources from the 58,533,603 (0.07%) objects in the S12 catalog.

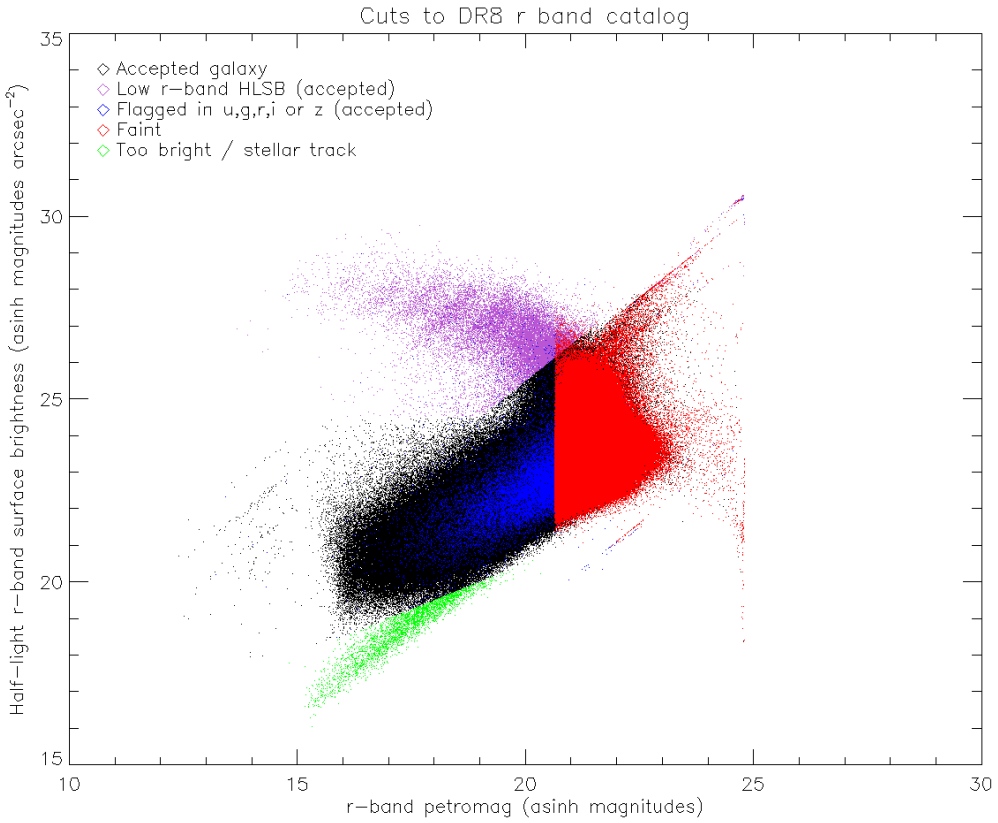


Figure 7.1: Petrosian magnitude vs half-light surface brightness of DR8 galaxies selected, highlighting the 5 million investigated galaxies. The blue points describes galaxies selected by the `MAYBE_CR` and `MAYBE_EGHOST` flags; the purple points describe galaxies investigated at low surface brightness; the green area contains 'galaxies' excluded through surface brightness cuts; red points describe galaxies that are below all magnitude limits (§7.1.4); and the remaining galaxies, represented by black points, form the sample on which APERC4 is run. The galaxies identified by green and red points are excluded from the Sheldon et al. galaxy sample, whilst the other selections remain.

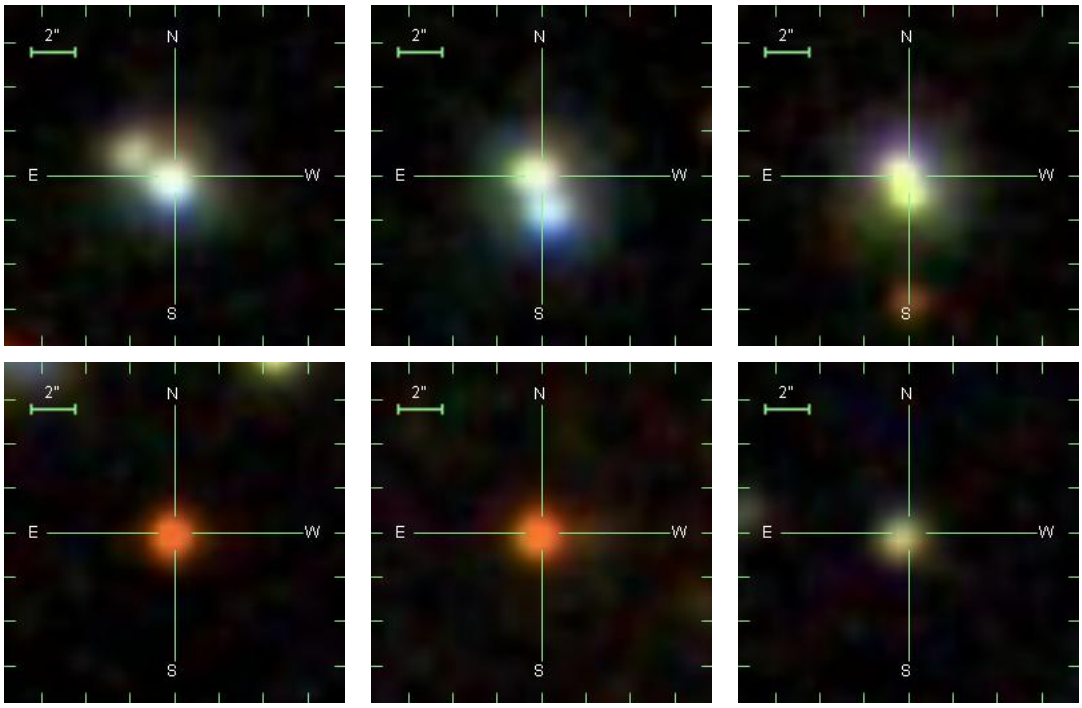


Figure 7.2: These cutouts demonstrate 'galaxies' identified by surface brightness cuts (green area in Figure 7.1). The top row shows three example galaxies identified at magnitudes brighter than $r < 16.5$, whilst the bottom row shows three 'galaxies' in the same surface brightness cut at magnitudes fainter than $r > 19.0$. Each image is $15''$ on each side. The bright examples show close binaries, where the SDSS deblender has been unsuccessful in splitting the sources into their constituent parts. The faint examples are somewhat less easy to distinguish as either star or galaxy by eye.

7.1.4 Magnitude Limits

The S12 magnitude limits used for selection are $[ugriz] < [21, 22, 22, 20.5, 20.1]$, which were employed to ensure a reasonable detection in each band. These are within the Stoughton et al. (2002) 95% completeness limits for point-like (PSF-like, e.g., stars) objects in the SDSS ($[ugriz] < [22, 22.2, 22.2, 21.3, 20.5]$). However, for cluster finding, variability in completeness across the survey area may lead to biases in the locations of where clusters are found. The selection of `PRIMARY` objects ensures no duplicate objects exist in areas of the sky that have been observed repeatedly. In areas where survey runs overlap, repeat observations increase the opportunity for objects to be detected near the magnitude limit, and thus these regions may appear to contain more galaxies than others.

Completeness limits suitable for cluster finding are determined empirically. The S12 galaxies are taken and counted in magnitude bins of 0.01 dex for each band, independently, and examined for turnover from power law. The power law is fitted to these histograms between the magnitude bins where there are at least 1,000 galaxies in a given bin, and the bins where the maximum number of galaxies are found (see Figure 7.3). The peak magnitude bins ($[ugriz] = [22.65, 22.21, 21.71, 20.95, 20.35]$; vertical red dashed lines in Figure 7.3) resemble the S12 magnitude limits, with the exception of u-band, where the peak occurs 0.65 magnitudes deeper. The magnitude limit for a given band is then determined by the highest magnitude bin that remains above the fitted power law.

As the peak bin occurs after the distribution departs from this power law, the fitted power law includes bins below 100% completeness, thus magnitude limits may be slightly fainter than a 100% completeness level. Iteration for the top magnitude bin for the power law was considered, but r and i distributions deviate from power law very slightly as to make the faint end limit very bright, so this method was rejected. A possible reason for this may be that the accelerating expansion of the universe means that the volume considered is not truly Euclidean, thus the number density of sources in bins of magnitude will also deviate. A further caveat is that the completeness limits will be strongly surface-brightness dependent for galaxies (Blanton et al., 2005).

From identifying where the fitted power law turns over, the magnitude limits are determined as,

$$[ugriz]_{\rm mag} = [22.21, 21.19, 20.65, 19.94, 19.44]$$

Again, the u-band appears deeper than the S12 limit, whilst the remaining bands are between 0.56 and 1.45 magnitudes shallower. These form the magnitude limits applied the the galaxies input into the AperC4 algorithm. Galaxies are excluded if no single band

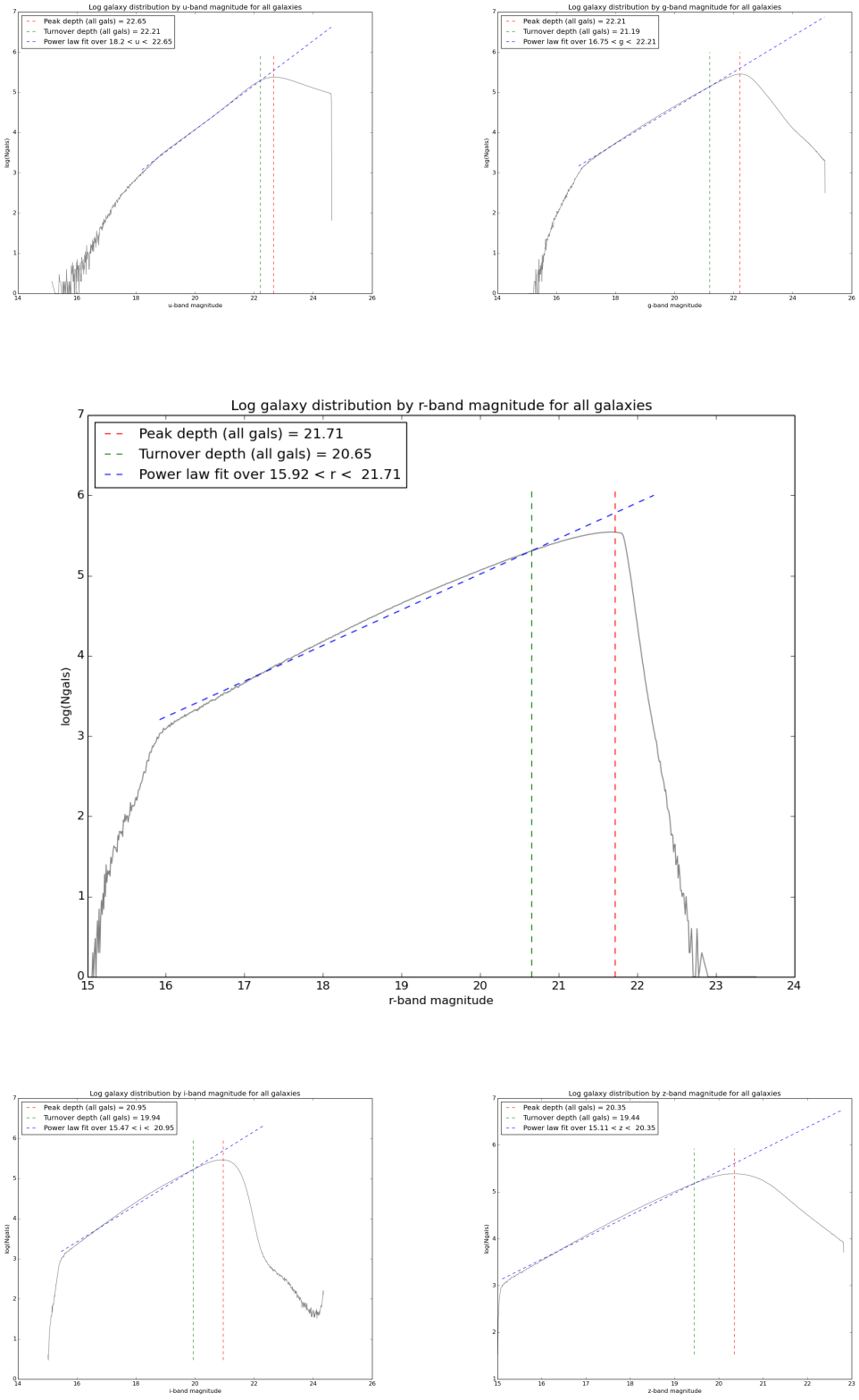


Figure 7.3: Magnitude limits and limits determined from S12 galaxy catalog. Each plot corresponds to the *u, g, r, i* and *z* SDSS bands from left-to-right, top-to-bottom, respectively. The red dashed line marks the location of the peak magnitude bin, the blue dashed line denotes the best fit power law, and the green dashed line marks the magnitude bin at which turnover from power law is indicated.

passes these thresholds (these are represented by the red points in Figure 7.1).

7.2 AperC4-SDSS DR8 Cluster Catalog Assembly

In this section, I will describe some of the verification processes that are applied to the AperC4 products that are created during the running of the algorithm (outlined in §3.2).

7.2.1 Survey Tiling

When AperC4 ingests the galaxy catalog (step 1; §3.2.2), it creates tiles of equal side length in R.A. & declination. During the simulation runs (chapter 6), the data are presented in tiles bounded by constant R.A. and dec. The S12 catalog is bounded to the BOSS (Eisenstein et al., 2011, Baryon Oscillation Spectroscopic Survey) footprint (see upper plot of Figure 7.4*), which requires the algorithm to perform the tiling on regions that are not bound by constant R.A. or dec. Furthermore, the BOSS survey footprint covers two contiguous regions that cover the Southern and Northern Galactic Cap, which are connected by a limited number of runs where the intermediary galaxies have been excised.

The AperC4 algorithm divides the galaxies into tiles of equal side (`SPLITSIZE` + `OVERLAP`) = (10 + 2) in R.A. and dec. (see lower plot of Figure 7.4), starting from the lowest R.A. of the input galaxy catalog. In this case, the lowest R.A. is 0°, and the algorithm can be seen to overlap where the celestial coordinates wrap from 360° to 0°. Where there are no galaxies, AperC4 does not attempt to record a tile. With this tiling, I can demonstrate that the area of the tiles is not equal for tiles at different declination, i.e., the tiles at dec = 60° are approximately half the width (and hence, half the area) of those at dec= 0°. At the survey edges, AperC4 merges tiles that have fewer than 100,000 galaxies (typical tile number densities in the CatSim catalog for equivalent sized tiles is around $\sim 200,000$ galaxies) with neighbouring AperC4 tiles, until all tiles contain at least this number of galaxies. This minimum number of galaxies ensures that tiles that are too slim at declinations towards the poles are also suitably populated.

7.2.2 p-value Determination and FDR

After calculating p-values (step 2, §3.2.3), the AperC4 galaxies are determined by applying the FDR threshold (step 3, §3.2.3) such that the remaining galaxies are the most cluster-like. Figure 7.5 shows the effect of the FDR values tested on simulation galaxies in section 6.2, on the DR8 dataset for the 3.7′ aperture size. The galaxies coloured purple and

*Footprint survey file from http://data.sdss3.org/sas/dr9/boss/lss/boss_survey.fits

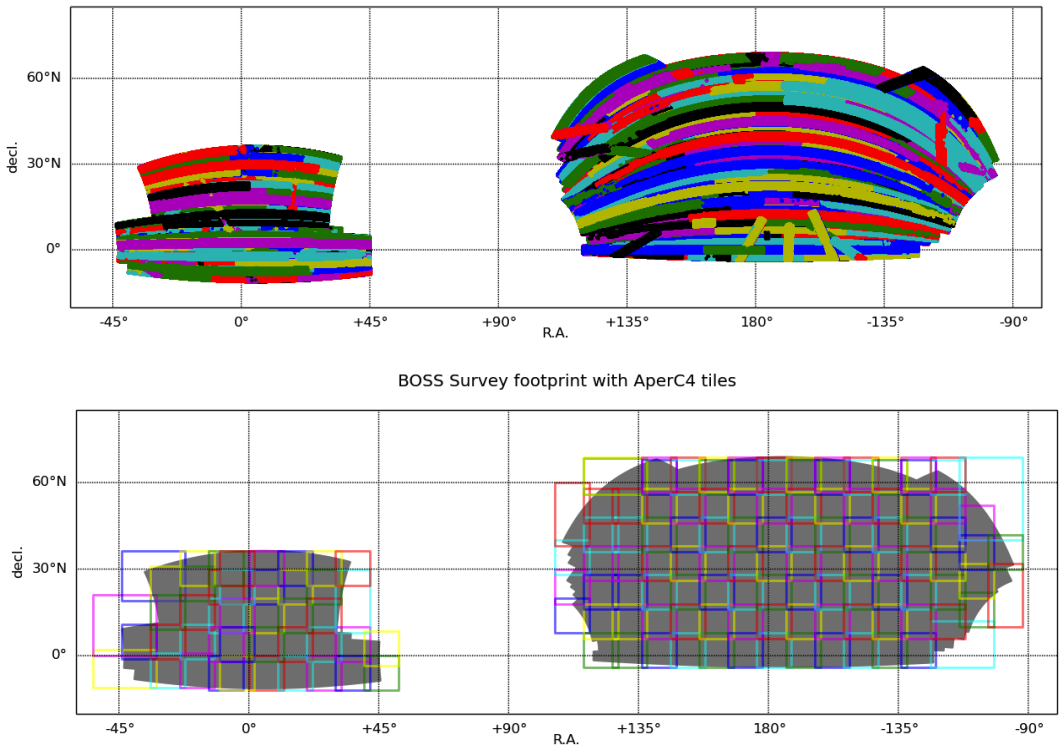


Figure 7.4: This figure shows the distribution of SDSS DR8 runs (*top*) and the AperC4 tiling (*bottom*) of the S12 galaxies in the BOSS footprint. The S12 galaxies are sourced from 559 DR8 runs, and each run is randomly coloured in the upper plot. The AperC4 tiles are outlined in random colours in the lower plot, where tile dimensions have been determined by the `SPLITSIZE` and `OVERLAP` parameters.

deep blue (FDR of 0.4% and 1%, respectively) make it past the optimal FDR= 1% selection (§6.3.3), whilst those coloured light blue and green (FDR of 4% and 10%, respectively) indicate galaxies that are close to, but do not pass, this criteria. In general, the light blue and green galaxies are found around groups of deep blue/purple galaxies, demonstrating that the p-values can be used to identify cluster galaxies fairly well. However, we can also notice some groups of chiefly green galaxies, e.g., at $(\mathrm{R.A.}, \mathrm{dec}) = (181.5°, 32.3°)$, which are not included post-FDR cut since the inclusion of such groups would likely increase the number of false projections included in the final catalog at a greater rate than the inclusion of real clusters, thus negatively impacting purity (§6.3).

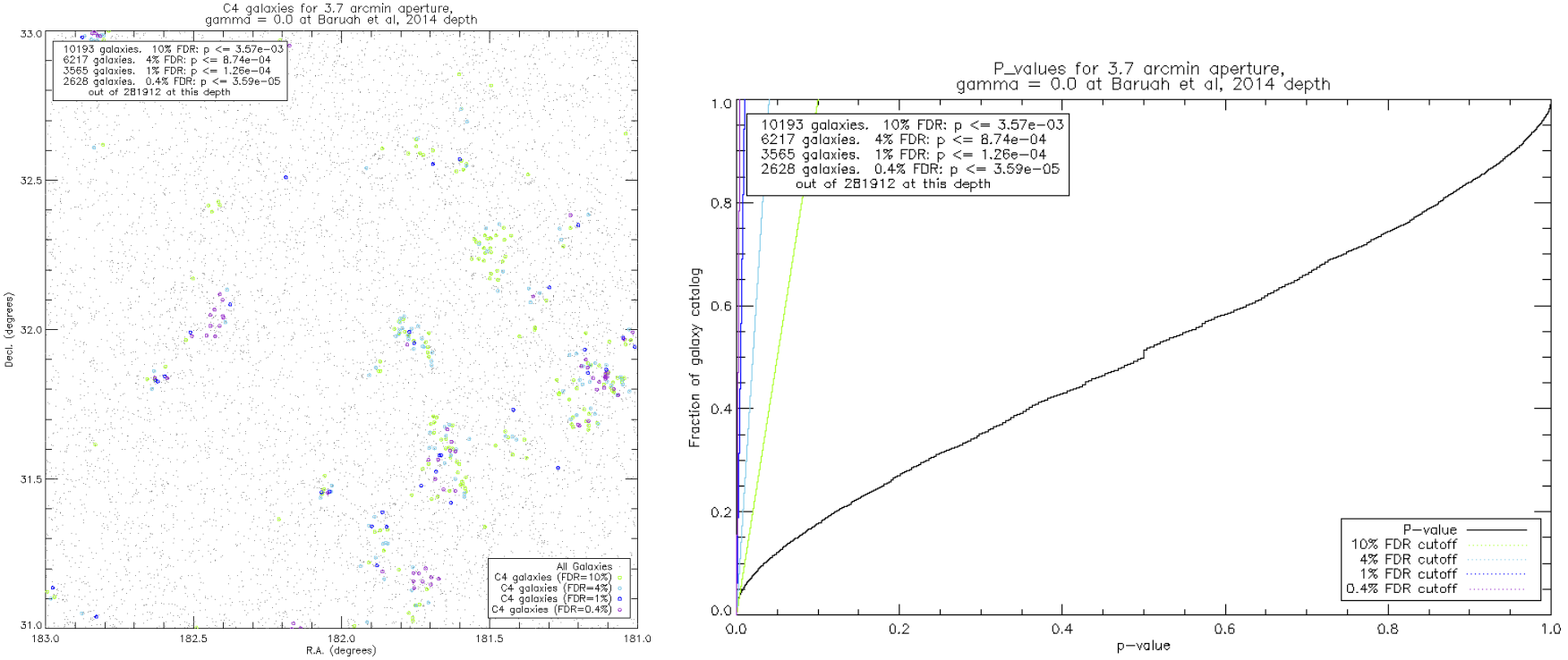


Figure 7.5: This figure shows the distribution of DR8 galaxies found with the 3.7′ aperture for a ∼4□° area (*left*) and a corresponding p-value distribution for those DR8 galaxies and other galaxies within their tile (§7.1.3 & §7.2.1) processed by AperC4 (*right*). Background DR8 galaxies are shown as black points. The galaxies shown that pass FDR are 10%, 4%, 1% and 0.4% (colour coded purple, blue, light blue, and green, respectively) which select the coloured galaxies in the left plot (note that the galaxies selected at smaller FDR values are subsets of the larger FDR values). These galaxies are selected by their p-value in the right plot, where the p-value distribution to the left of the intercept between the p-value distribution (black line) and the correspondingly coloured FDR threshold.

The p-value distribution in Figure 7.5 shows that whilst there are galaxies that can be differentiated between field and cluster-like, i.e., galaxies with $p = 1$ or $p = 0$, most of the catalog lies between the two extremes. Contrasting with the spectroscopic DR2 p-value distribution (constructed with a comoving aperture in $C4_{M05}$) in Figure 3.2, it can be seen that the ability to distinguish between cluster-like and field-like galaxies has decreased by not constraining the redshift dimension and by having a single static aperture. In the following subsection, I will demonstrate how APERC4 leverages this reduced statistical power by employing multiple apertures.

7.2.3 k-NN Centre Determination and Forming Aperture-Slice Clusters

Using the optimal $k-$NN distance, $k = 24$, (i.e., distance to the 24^{th} neighbour, §6.3.3), to determine the cluster centres (step 4, §3.2.5) then associating galaxies to those centres (step 5, §3.2.6), APERC4 produces nine aperture-slice cluster catalogs (there being nine apertures in the optimal parameter set). Recall that no redshift information has been used up to this point, and that the galaxies have been associated to the centres based on their proximity and the initial $k-$NN density of the central galaxy. Groupings of 3 galaxies or fewer are excised from the aperture-slice catalogs.

Figure 7.6 shows the distribution of cluster centres and their associated galaxies from three aperture-slices (at $2.7'$, $3.7'$, and $6.3'$). At $2.7'$, the APERC4 galaxies identified are easily grouped into distinct, low population groupings (aperture-slice cluster). As the aperture size increases to $3.7\ '$, more aperture-slice clusters are found, with slightly larger radii stemming from the fact that they probe larger physical associations than $2.7'$ at some common distance, or conversely probe the same size physical associations at $2.7'$, but at closer distances. At $3.7'$, structures can start to appear fragmented, such as the concentration at $(\text{R.A.}, \text{dec}) = (182.4^\circ, 32.0^\circ)$, which the algorithm determines as two concentrations since the centres are more than $3.7'$apart. As the aperture size is increased further, so the number of galaxies that qualify as APERC4-clustered increases, as can be seen with the $6.3'$aperture.

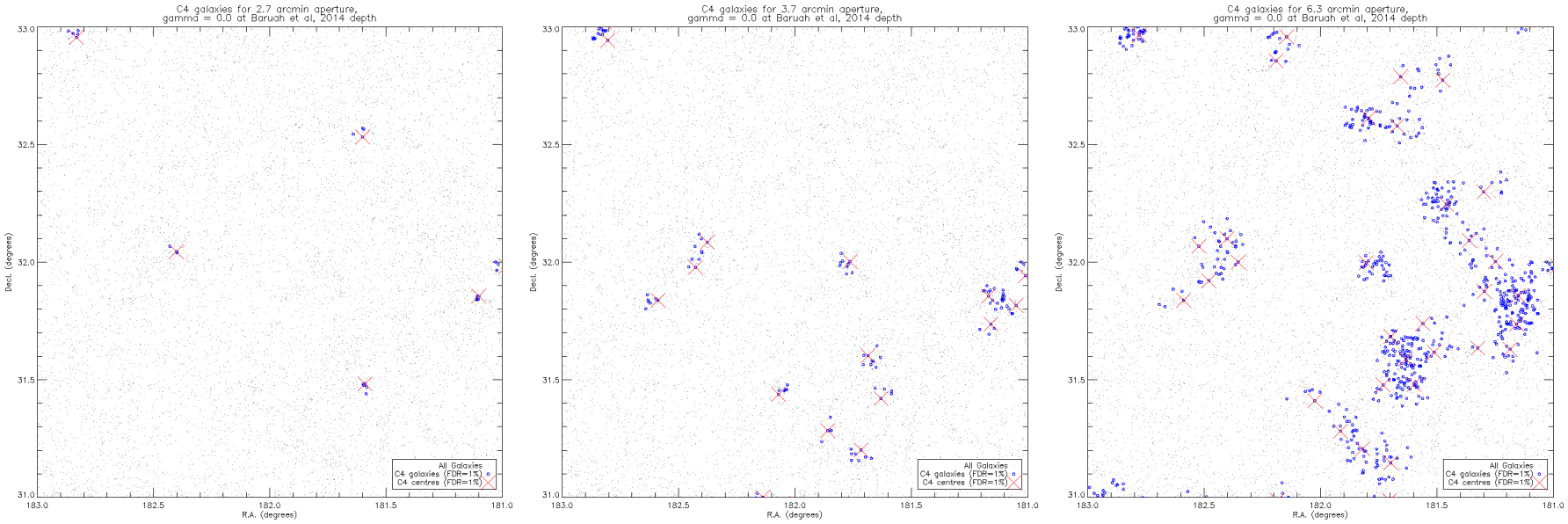


Figure 7.6: This figure shows the distributions of DR8 galaxies identified as AperC4 galaxies with the 2.7′ (*left*), 3.7′ (*middle*), and 6.3′ (*bottom*) apertures for the same area of sky as Figure 7.5. Background DR8 galaxies are shown as black points, the AperC4 galaxies (that passed the FDR= 1% threshold) that have been assigned to cluster centres are shown in blue, whilst the aperture-slice cluster centres are denoted by the large red crosses.

The aperture-slice clusters found with the larger apertures appear fragmented in the celestial plane, with multiple centres applied to apparent 2-D groupings. If clusters from other aperture-slices combine galaxies from neighbouring groups, then it's possible for these groupings to merge, provided their $p(z)$s indicate the galaxies exist at the same redshift. Hence, the final cluster radii (the radii of merged aperture-slice clusters) will not be limited to the maximum aperture radius employed.

In Figure 7.6, a concentration of galaxies can be seen at approximately (R.A., dec) = $(181.6°, 32.6°)$ in the $2.7'$ and $6.3'$aperture-slices, but not the $3.7'$aperture-slice. This concentration did exist in the $3.7'$catalog but was not significant enough to pass the FDR threshold, and can be seen as a small concentration of green galaxies (FDR= 10%) in Figure 7.5. Whether this grouping will form substructure from the the grouping(s) in the $6.3'$aperture or describe a separate cluster along the line of sight will be determined by the addition of $p(z)$ information.

7.2.4 Combining the Aperture-Slice Clusters with $p(z)$ Information and Producing the Final Cluster Catalog

The aperture-slice clusters are given $p(z)$s (step 6, §3.2.7) by combining the S12 $p(z)$ information of their constituent galaxies. The clusters are then merged/fragmented such that each galaxy only belongs to one cluster (step 7; §3.2.8). As found through training on CatSim, clusters that share some finite probability of lying at the same redshift are merged (i.e., cluster cross-$p(z) > 0.0$; §6.3.3). In section 7.3.2, I will use spectroscopic GAMA data (Driver and team, 2008) to examine this merging/defragmentation process.

The aperture-slice tiles have been combined into tiles with clusters with unique galaxy memberships. These tiles are then themselves merged to be sure that duplication of clusters/galaxies does not occur between tile overlaps, and then clusters with fewer than $N_{\rm gals} < 32$ (determined to be the optimum threshold that identifies the best balance between true clusters and false associations; §6.3.3) are excluded from the fully merged catalog (step 8; §3.2.9). The remaining clusters and associated galaxies form the final AperC4 cluster catalog.

7.3 The AperC4-SDSS DR8 Cluster Catalog

7.3.1 AperC4-SDSS DR8 Catalog Summary

The AperC4-SDSS DR8 Cluster Catalog contains 83,880 clusters containing 1,898,189 galaxies between them. The cluster catalog itself comes in three parts: the cluster catalog (Table 7.1), the member catalog (Table 7.2) and the cluster $p(z)$ catalog (Table 7.3).

Table 7.1: Schema of the final AperC4 cluster catalog.

Name	Format	Description
ID	Long64	AperC4 cluster identification number
RA	Double	Right Ascension in decimal degrees (J2000)
DEC	Double	Declination in decimal degrees (J2000)
RADIUS	Double	Angular separation of central & outermost galaxies (arcminutes)
NGALS	Long	Number of galaxies associated to this cluster, i.e., $N_{\rm gals}$.
KNN_CENTRE_DISTANCE	Double	kth nearest neighbour distance (radians)
NGALRANK	Long	$N_{\rm gals}$ rank (§5.3.1)
ABELRICHNESS	Long	Number of galaxies brighter than $(m_3 - 2)$ in r-band (§5.3.2)
ABELRANK	Long	Rank based on ABELRICHNESS

Table 7.2: Schema of the final AperC4 member catalog.

Name	Format	Description
RA	Double	Right Ascension in decimal degrees (J2000)
DEC	Double	Declination in decimal degrees (J2000)
NEIGHBOUR_COUNT	Double	Colour clustering measurement of galaxy (Equation 3.2).
MEDIAN_NEIGHBOUR_COUNT	Double	Median model colour clustering measurement
N_RANDOM_NEIGHBOUR_COUNTS	Integer	Number of times the aperture was randomly placed.
P_VALUE	Double	p-value (Equation 2.21).
UG	Double	$u - g$ colour, calculated with SDSS MODELMAG magnitudes.
UG_ERR	Double	Uncertainty on $u - g$ (Equation 3.1).
GR	Double	$g - r$ colour, calculated with SDSS MODELMAG magnitudes.
GR_ERR	Double	Uncertainty on $g - r$ (Equation 3.1).
RI	Double	$r - i$ colour, calculated with SDSS MODELMAG magnitudes.
RI_ERR	Double	Uncertainty on $r - i$ (Equation 3.1).
IZ	Double	$i - z$ colour, calculated with SDSS MODELMAG magnitudes.
IZ_ERR	Double	Uncertainty on $i - z$ (Equation 3.1).
MAG_R	Double	SDSS r-band MODELMAG magnitude
MAG_I	Double	SDSS i-band MODELMAG magnitude
APERTURE	Double	Aperture where $p(z_{\rm cluster}) \times p(z_{\rm galaxy})$ is optimal (§3.2.8).
ID	Long64	SDSS DR8 CAS object identifier
KNN_DISTANCE	Double	kth nearest neighbour distance in aperture-slice catalog (radians)
CLUSTERID	Long64	AperC4 cluster identification number
RADIUS	Double	Angular distance to the cluster centre (arcminutes)
N_INSIDE	Long	Number of galaxies in the cluster that are within RADIUS.

Member catalog schema continued on next page

Table 7.2 – *Member catalog schema continued from previous page*

Name	Format	Description
DENSITY	Double	Surface density of APERC4 galaxies, $^{\texttt{N\_INSIDE}}/_{\pi(\texttt{RADIUS})^2}$.
CLUSTERNGALS	Long	Cluster $N_{\rm gals}$
PZFILE	String	Name of S12 file containing galaxy's $p(z)$ information.
NGALRANK	Long	Cluster $N_{\rm gals}$ rank (§5.3.1)
ABELRANK	Long	Cluster rank based on ABELRICHNESS (§5.3.2)

Table 7.3: Schema of the final APERC4 cluster $p(z)$ catalog.

Name	Format	Description
ID	Long64	APERC4 cluster identification number
WEIGHT	Double	Minimum $p(z)$ probability of constituent members.
NGALS	Long	Cluster $N_{\rm gals}$.
P_OF_Z	Double[35]	Cluster $p(z)$, where bins are defined by S12, and calculated by the product sum of the member galaxies (§3.3)

The full set of APERC4 catalogs can be downloaded from:

`http://astronomy.sussex.ac.uk/~lb92/AperC4/catalog.html`

Figure 7.7 shows the distribution of the 83,880 APERC4-DR8 clusters, evaluated with S12 $p(z)$, plotted on the BOSS footprint. There is an apparent structure to the distribution of clusters in the catalog, suggesting APERC4 is selecting clusters at some preferred location.

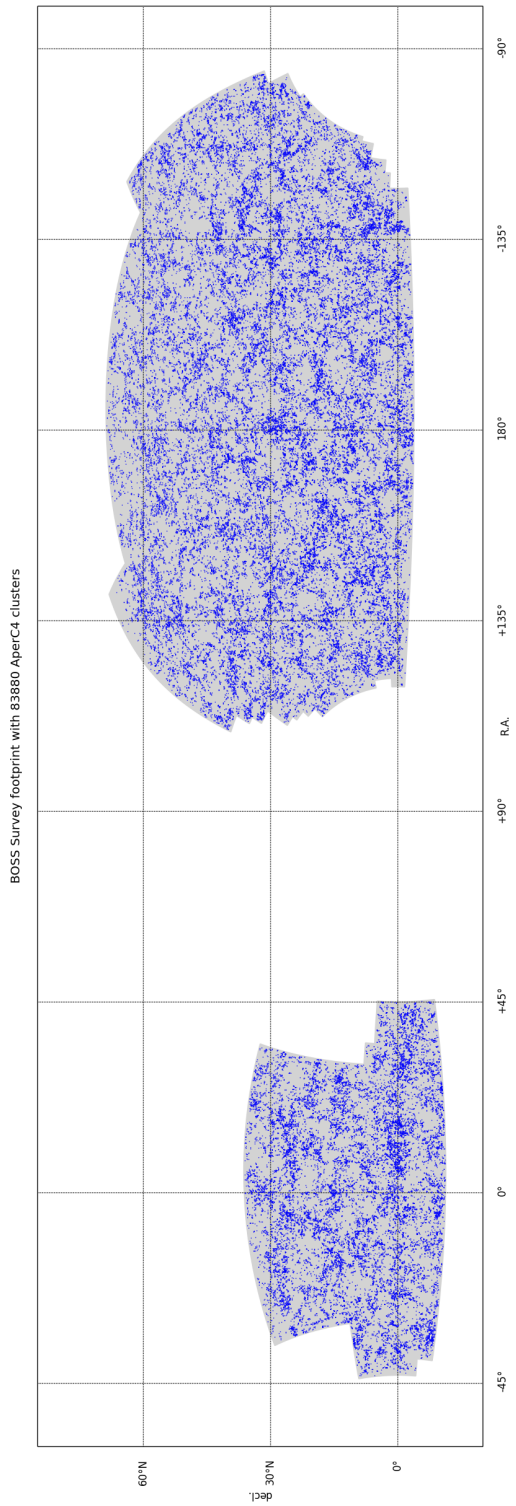


Figure 7.7: Distribution of AperC4-DR8 clusters (blue points) across the BOSS survey footprint (shaded region).

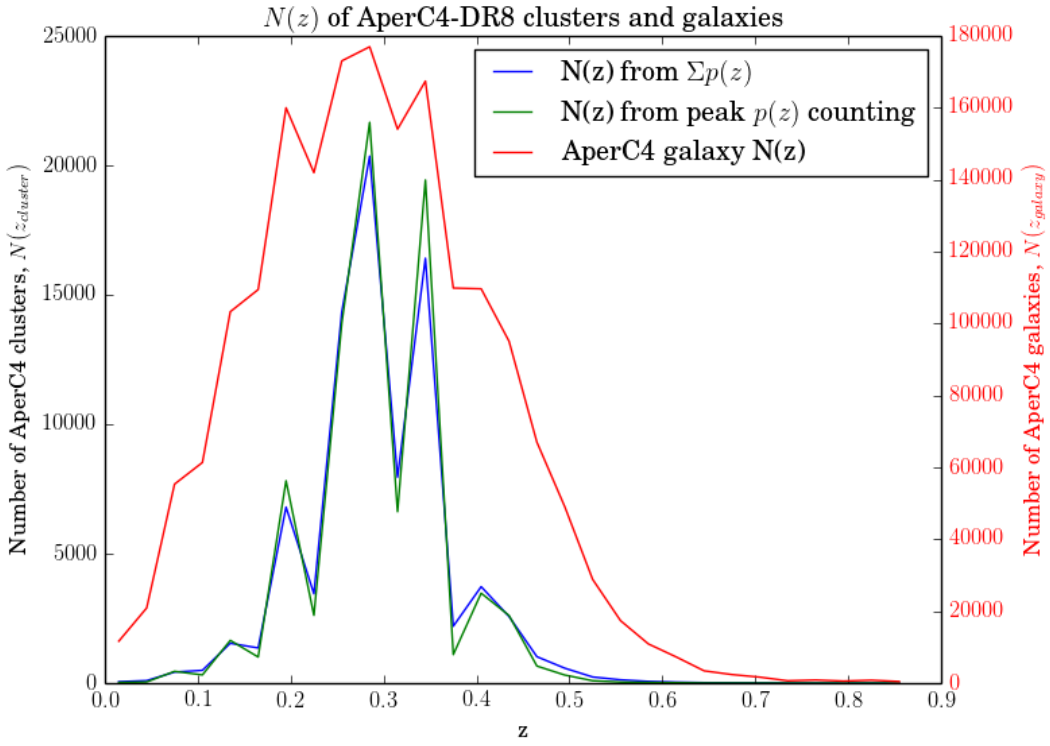


Figure 7.8: $N(z)$ of AperC4-DR8 clusters (left vertical axis) through summation of the cluster $p(z)$s (blue line), counting the locations of the $p(z)$ peaks (green line). Also shown is the $N(z)$ of the constituent AperC4 galaxies (red line; right vertical axis).

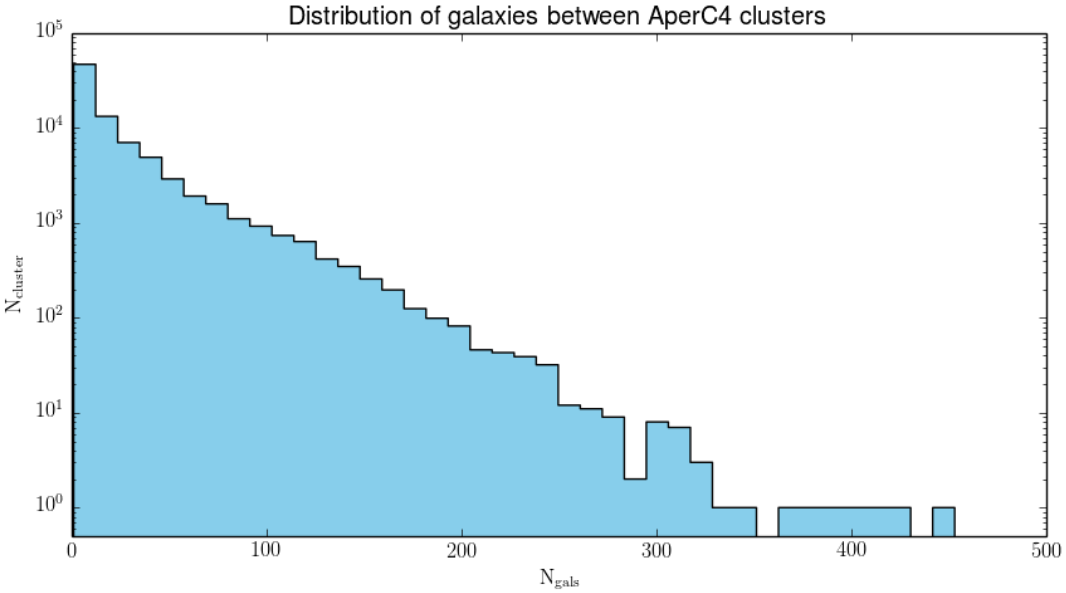


Figure 7.9: $N_{\rm gals}$ of AperC4-DR8 clusters.As one would expect, the higher $N_{\rm gals}$ clusters are rarer than those with lower $N_{\rm gals}$. The richest cluster identified by AperC4 contains ∼450 galaxies.

Figure 7.8 shows the $N(z)$ of the AperC4 clusters and of their constituent galaxies. The clusters appear to be preferentially found in the $0.24 < z < 0.27$, $0.27 < z < 0.30$ and

$0.33 < z < 0.36$ bins, where those particular bins contain more than double the number of clusters in any other bin. By comparison, the $N(z)$ of the AperC4 galaxies (per redshift bin) appears to increase until $z = 0.2$, then remains approximately constant until $z = 0.36$, at which point the number of galaxies falls (nominally because fainter galaxies have been excluded by the magnitude limits of the catalog). The clusters are populated as one would expect (Figure 7.9), with richer clusters being rarer than clusters with lower $N_{\rm gals}$. I will discuss the nature of this $N(z)$ distribution and projected structure in section 7.4.1.

7.3.2 GAMA Spectroscopy in AperC4-SDSS DR8 Clusters

To lend some context to the cluster redshifts, I employ spectroscopic redshift measurements of SDSS galaxies taken from the GAMA (Galaxy And Mass Assembly) survey.

The GAMA Survey

The GAMA survey (Driver et al., 2010) is a multi-wavelength photometric and spectroscopic survey, designed to expand upon the Millennium Galaxy Catalog (Liske et al., 2003), 2dFGRS (2 degree Field Galaxy Redshift Survey; Lewis et al., 2002), and SDSS spectroscopic campaign to measure galaxy properties and environments on scales from 1 kpc to 1 Mpc. GAMA covers 3 regions covering a total of $\sim 180\,\square^\circ$ to a magnitude of $r < 19.8$. This thesis employs the second data release from GAMA (DR2; Liske et al, *in prep.*).

Matching AperC4 to GAMA

The galaxies for which GAMA provides spectroscopy are targeted from the SDSS DR6 photometric galaxy catalog, which allows unambiguous matching to the SDSS DR8 galaxies employed in this thesis. As the SDSS DR6 object identification numbers (`objID`s) differ from the `objID`s of SDSS DR8, it isn't possible to match galaxies by `objID`. However, since the object identification system (excluding the numbering schema) between SDSS DR6 and DR8 is largely unchanged, GAMA galaxies are matched to AperC4-DR8 galaxies by proximity, with an upper limit of $2''$.

Of 221,373 galaxies with spectroscopic redshifts in the GAMA catalog, 9,743 are found in the AperC4 member catalog, distributed among 872 AperC4 clusters. Figure 7.10 summarises the redshift distribution of GAMA galaxies amongst AperC4 clusters, showing lower redshift clusters (as indicated by the peak of the cluster $p(z)$) will tend to contain galaxies with higher spectroscopic redshifts ($z_{\rm GAMA}$) whilst higher redshift clusters will tend

to contain galaxies from lower redshifts. The mean GAMA redshifts appear to indicate that the peak of the cluster $p(z)$s are fairly accurately positioned, whilst the large error bars indicate low precision. This is indicatory of a problem with assignment of galaxies to clusters (e.g., blending along the line of sight), the location of the peak $p(z)$ of the cluster (through its product-sum calculation [§3.2.7], or the S12 $p(z)$s used to calculate them), and/or the overall AperC4 selection of galaxies. I note that the peak of the cluster $p(z)$ may not be the best description of the full $p(z)$ (§6.1.1).

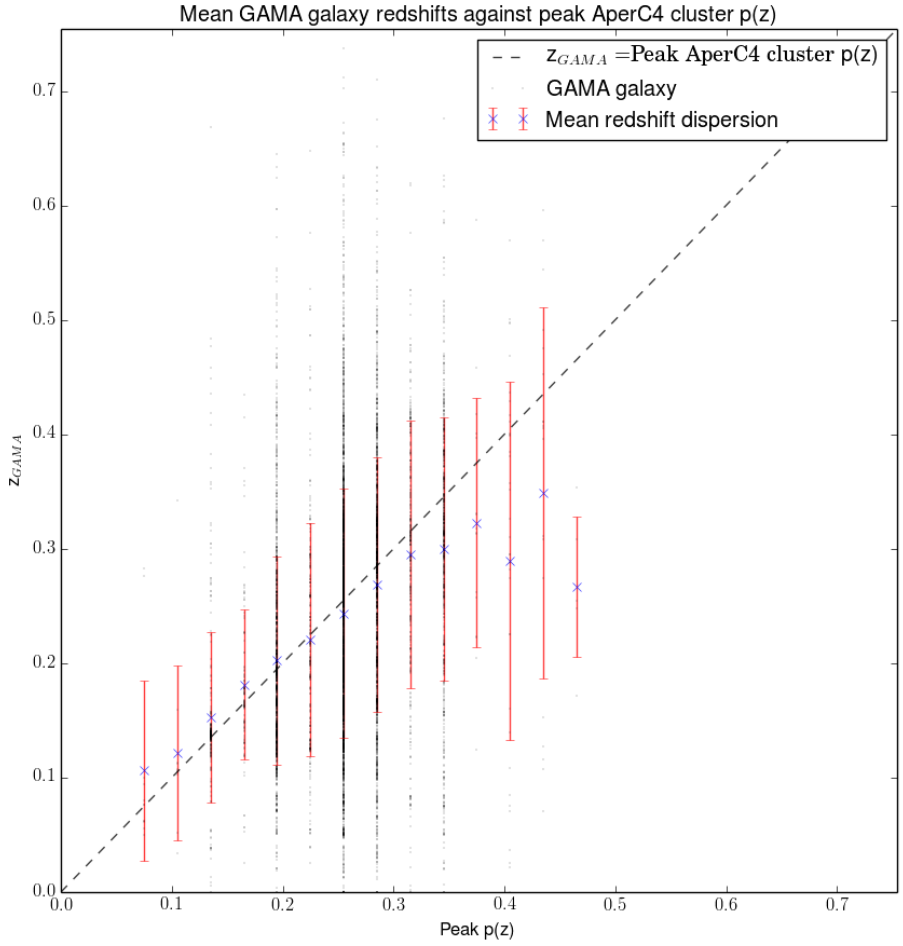


Figure 7.10: Mean GAMA galaxy redshifts against peak AperC4 cluster $p(z)$. The GAMA galaxies are grouped into bins where their corresponding AperC4 cluster is most likely located (using AperC4 $p(z)$). The black dashed line shows where the GAMA galaxy redshift, $z_{\rm GAMA}$, is equal to the location of the AperC4 cluster $p(z)$ peak. The blue crosses represent the mean $z_{\rm GAMA}$ of the galaxies that belong to clusters in that bin, and the red bars correspond to the 1σ error on the mean.

In Figure 7.11, the GAMA galaxies are distributed amongst the 872 AperC4 clusters they are found in, and the mean GAMA cluster redshift ($\overline{z}_{\rm GAMA}$) is calculated by taking the

mean redshift of the GAMA members in a given AperC4 cluster. The distribution seems to show the peak $p(z)$ of the clusters are near to the mean z of their constituent GAMA galaxies, allowing for a slight overestimation of peak $p(z)$ redshift at low $\overline{z}_{\rm GAMA}$ ($z \lesssim 0.1$), and an underestimation of peak $p(z)$ at high $\overline{z}_{\rm GAMA}$ ($z \gtrsim 0.4$). The median standard deviation of the GAMA galaxy redshifts around the mean redshift of their respective clusters is around $\delta z \approx 0.1$, where 9% have $\delta z \leq 0.01$ and 1% have $\delta z \geq 0.2$.

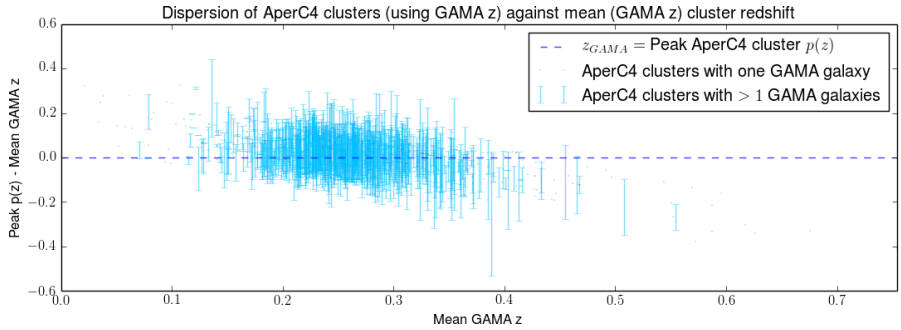


Figure 7.11: Using GAMA redshifts, this plot shows the dispersion of AperC4 clusters ($\sigma(z_{\rm GAMA})$) against mean spectroscopic cluster redshift ($\overline{z}_{\rm GAMA}$). The dotted line (dark blue) represents where $\overline{z}_{\rm GAMA}$ is the same as the midpoint of the $p(z)$ bin at which the cluster $p(z)$ peaks. The error bars (light blue) each indicate the redshift dispersion of a cluster as calculated by its spectroscopic members. Points (grey) indicate clusters where only one member has a GAMA redshift. Where error bars/points are vertically offset from the dashed line indicates the difference between the cluster $\overline{z}_{\rm GAMA}$ and peak $p(z)$.

7.3.3 Example AperC4 Clusters with GAMA Redshifts

With spectroscopic redshift information, and the knowledge that the catalog generated is approximately $\sim 70\%$ pure (with respect to the underlying distribution of matter, §6.4.1), it is possible to examine some of the success/failure modes of the AperC4 membership assignment and $p(z)$ construction. I now present some example clusters from the AperC4 cluster catalog in Figures 7.12 to 7.16. Figure 7.12 details the contents of the subfigures in the figures that follow.

Successful AperC4 Identifications

Figures 7.12 and 7.13 demonstrate two clusters identified by AperC4 in the $0.30 < z < 0.33$ and $0.15 < z < 0.18$ redshift bins, respectively. The redshift distributions of the members with spectroscopic redshifts in GAMA for these clusters mostly lie in the same bin. However, there are occurrences in both cases of a fraction of members with spectroscopy being line-of-sight contamination.

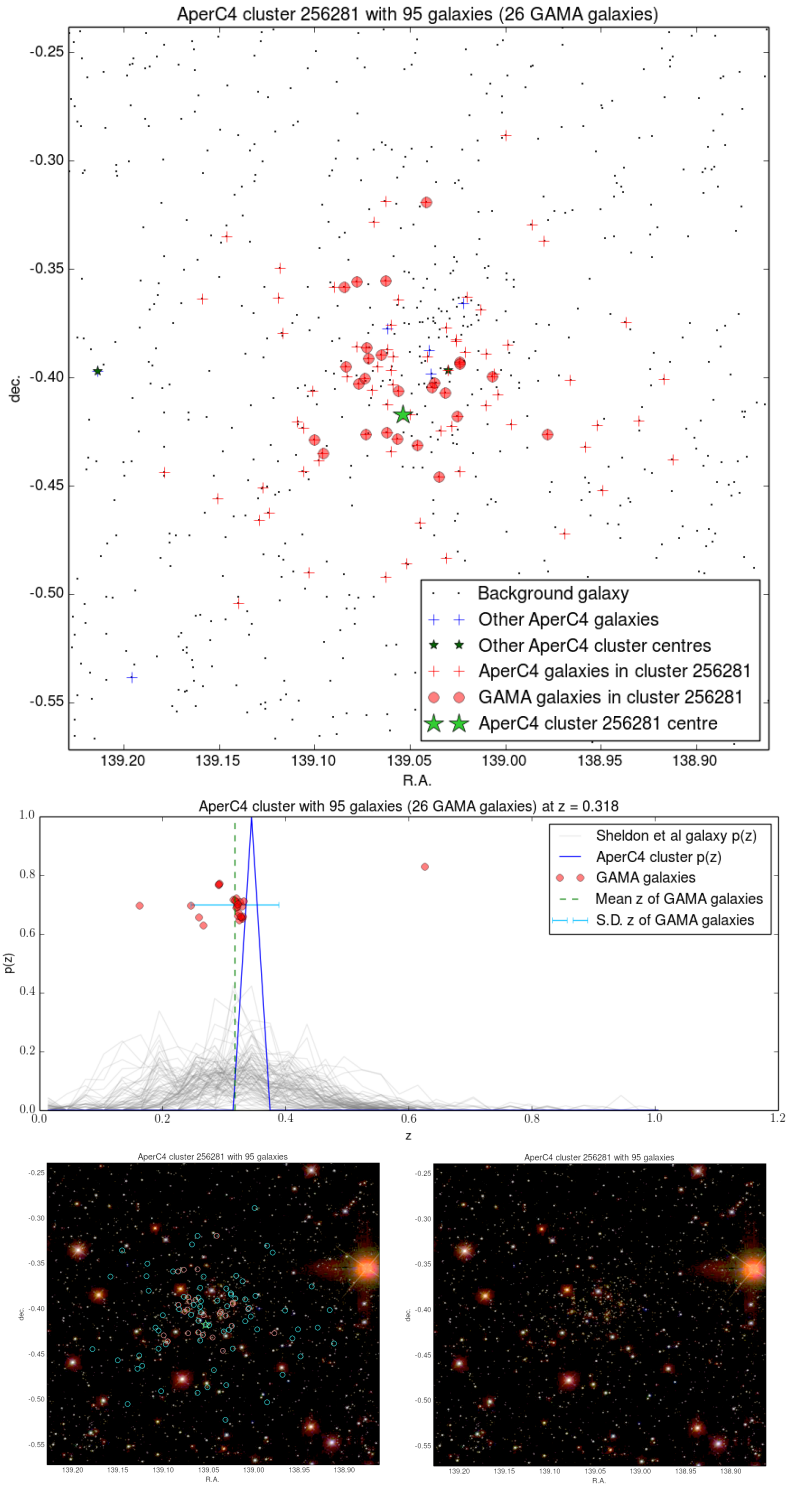


Figure 7.12: This figure shows AperC4 cluster ID 256281 and its 95 members. *Top:* Projection of the sky around the cluster centre (green star) and members with (red circles) and without (red plusses) GAMA redshifts. Also shown are other nearby AperC4 cluster centres (black stars), their members (blue plusses) and background galaxies (black points). *Middle:* The cluster's $p(z)$ (blue line) is compared to the constituent S12 galaxy $p(z)$s (grey lines). Also shown are GAMA redshifts of the cluster members (red dots), where vertical displacement around $p(z) = 0.7$ is proportional to displacement in dec. The mean redshift of those GAMA members (green dotted line) and 1σ standard deviation (horizontal cyan bar) are also shown. *Bottom:* SDSS DR8 SkyServer images of the cluster region [*bottom left*] unmarked, and [*bottom right*] with the cluster centre (green star), and members with GAMA redshifts (red ring) and without (cyan ring), indicated. The mean GAMA redshift of $z = 0.318$ coincides with the peak $p(z)$ of the cluster, indicating a good identification, although $\sim$⅓ GAMA galaxies are associated line-of-sight.

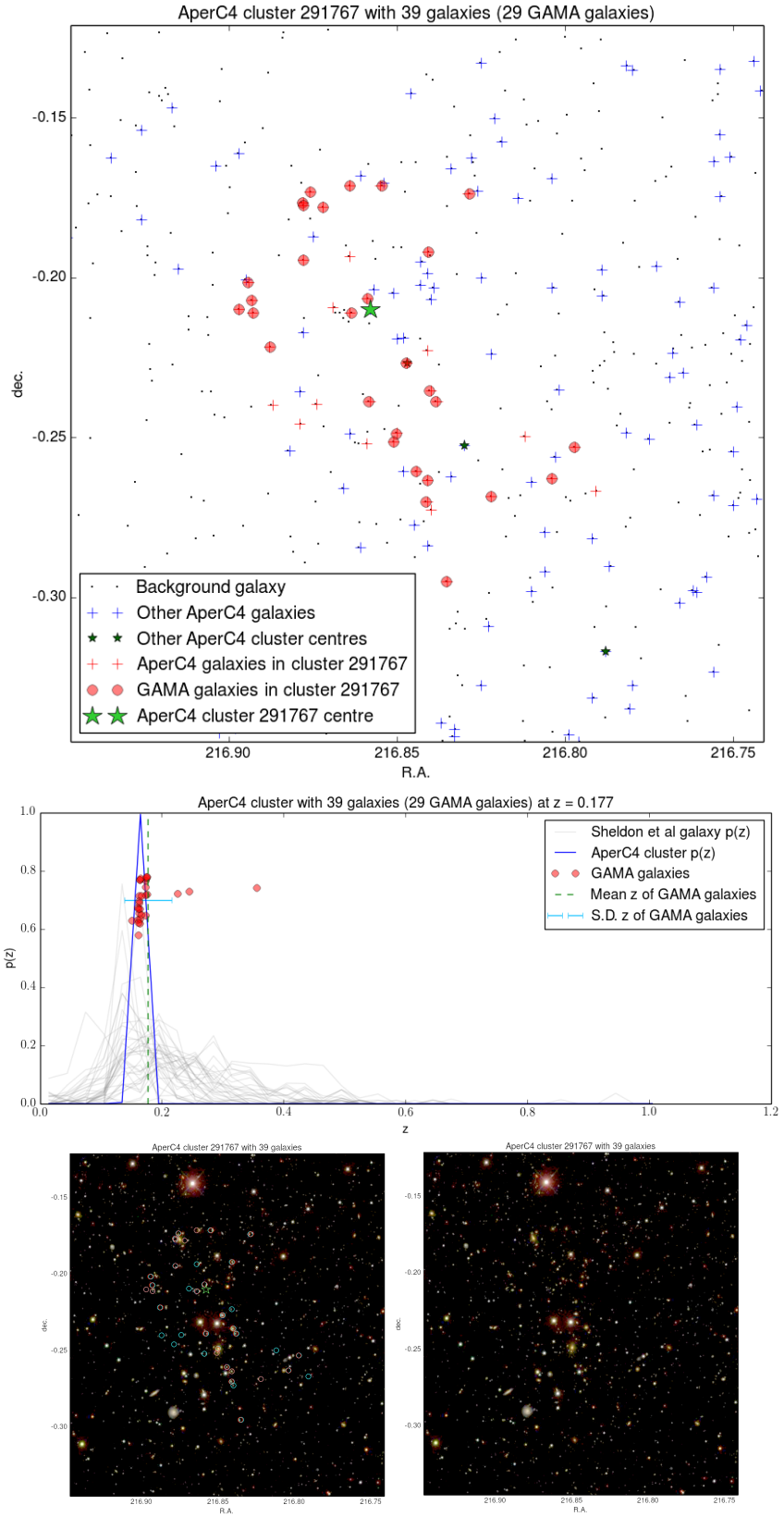


Figure 7.13: This figure shows AperC4 cluster ID 291767 and its 39 members. A description of the subplots is given in Figure 7.12. The mean GAMA redshift of $z = 0.177$ coincides with the peak $p(z)$ of the cluster, indicating a good identification where only 3 of 29 GAMA members are at a different redshift from the peak $p(z)$ of the cluster.

Common Redshift Bin Blend

Figures 7.14 and 7.15 demonstrate two clusters identified by AperC4 in the $0.21 < z < 0.24$ and $0.18 < z < 0.21$ redshift bins, respectively. Inspection of the spectroscopic members in these clusters reveals there are two redshift concentrations of galaxies occupying the $0.21 < z < 0.24$ bin in both clusters. Due to the binning of the $p(z)$s, AperC4 is unable to successfully defragment these two clusters. Furthermore, the fact that the two spectroscopic redshift concentrations are in the same $p(z)$ bin, implies that AperC4 has been unable to blend (or deblend) these two clusters with the $p(z)$ information of the galaxies themselves.

The projections also reveal several other AperC4 cluster centres and galaxies in close proximity to the clusters plotted, implying further fragmentation has occurred with these systems. Again, using the members with GAMA spectroscopic redshifts, there are occurrences in both cases of a members being line-of-sight contamination.

Line of Sight Blend

Figure 7.16 shows another unsuccessfully deblended cluster. This cluster demonstrates that whilst many AperC4 galaxies appear to belong in clusters (albeit, the galaxies with measured spectroscopic redshifts), AperC4 is not able to assemble them into groups well with S12 $p(z)$ information. Despite the presence of several other AperC4 cluster centres local to this group (in projection), the deblending process has not been able to assign these galaxies into groups of common redshift. The peak $p(z)$ of the cluster itself doesn't appear to align with any of the apparent concentrations of members in redshift space (at $z \approx 0.08$, $z \approx 0.26$, and $z \approx 0.42$), instead falling into the $0.30 < z < 0.33$ bin. This is one consequence of the cross-$p(z)$ threshold being set to 0.0 (§6.3.3), which permits aperture-slice clusters to merge if they share any members.

7.4 Discussion

7.4.1 AperC4-SDSS DR8 Structure

In section 7.3.1, the AperC4 galaxy cluster distribution appears to show some structure when projected across the survey area (Figure 7.7), and in redshift space (Figure 7.8). Putting these results together, AperC4 is preferentially forming clusters in the $0.24 < z < 0.30$ range, which is causing the projection to be dominated by structure in this redshift interval. The galaxies selected by AperC4 chiefly reside in the $0.16 < z < 0.42$

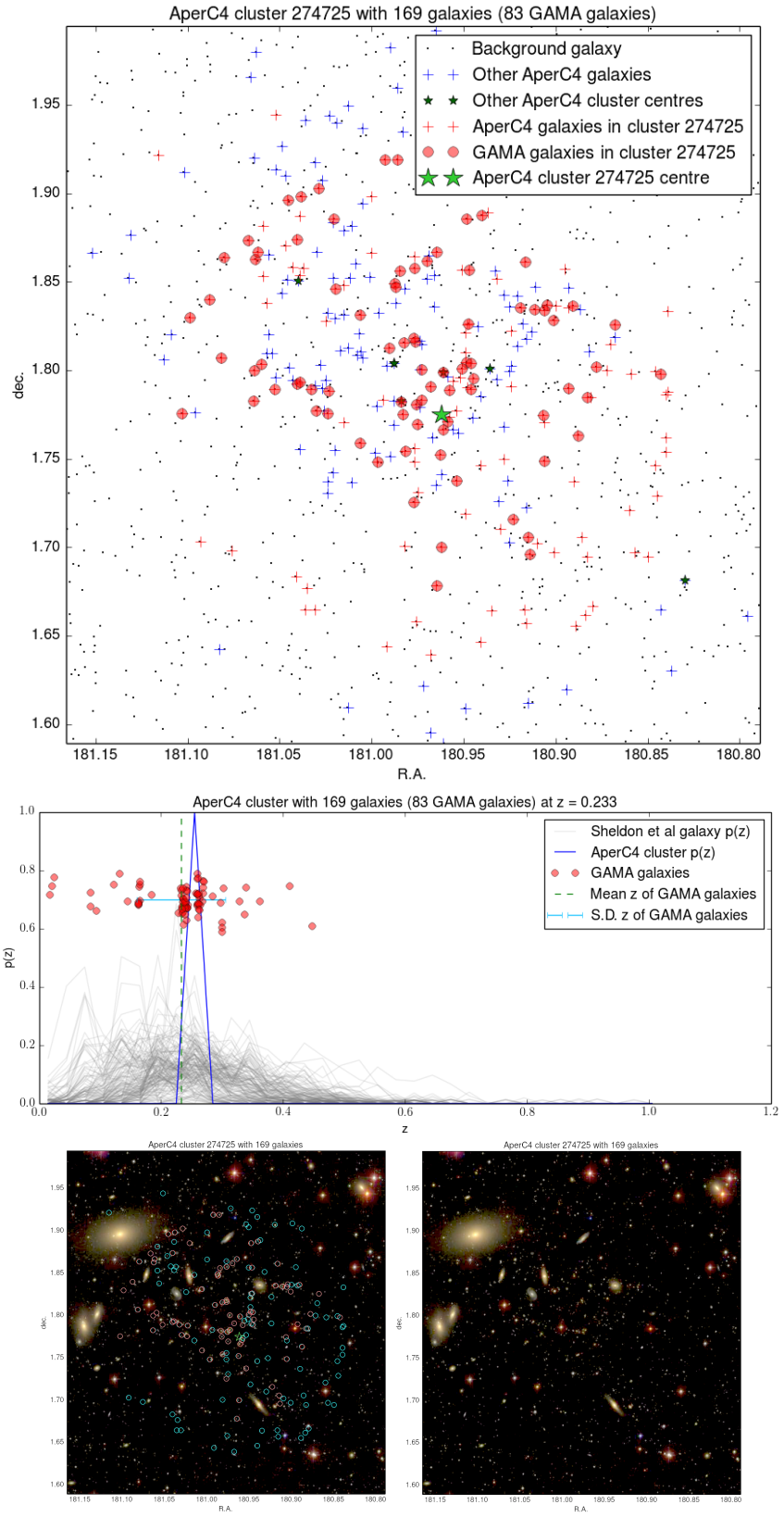


Figure 7.14: This figure shows AperC4 cluster ID 274725 and its 169 members. A description of the subplots is given in Figure 7.12. This cluster centre is identified close to many other cluster centres, including cluster 274974 (Figure 7.15). The spectroscopic redshifts from GAMA indicate that there are *two* clusters within the same $p(z)$ bin as the cluster $p(z)$ peak. The galaxy $p(z)$ information appears to be too coarse to distinguish these clusters.

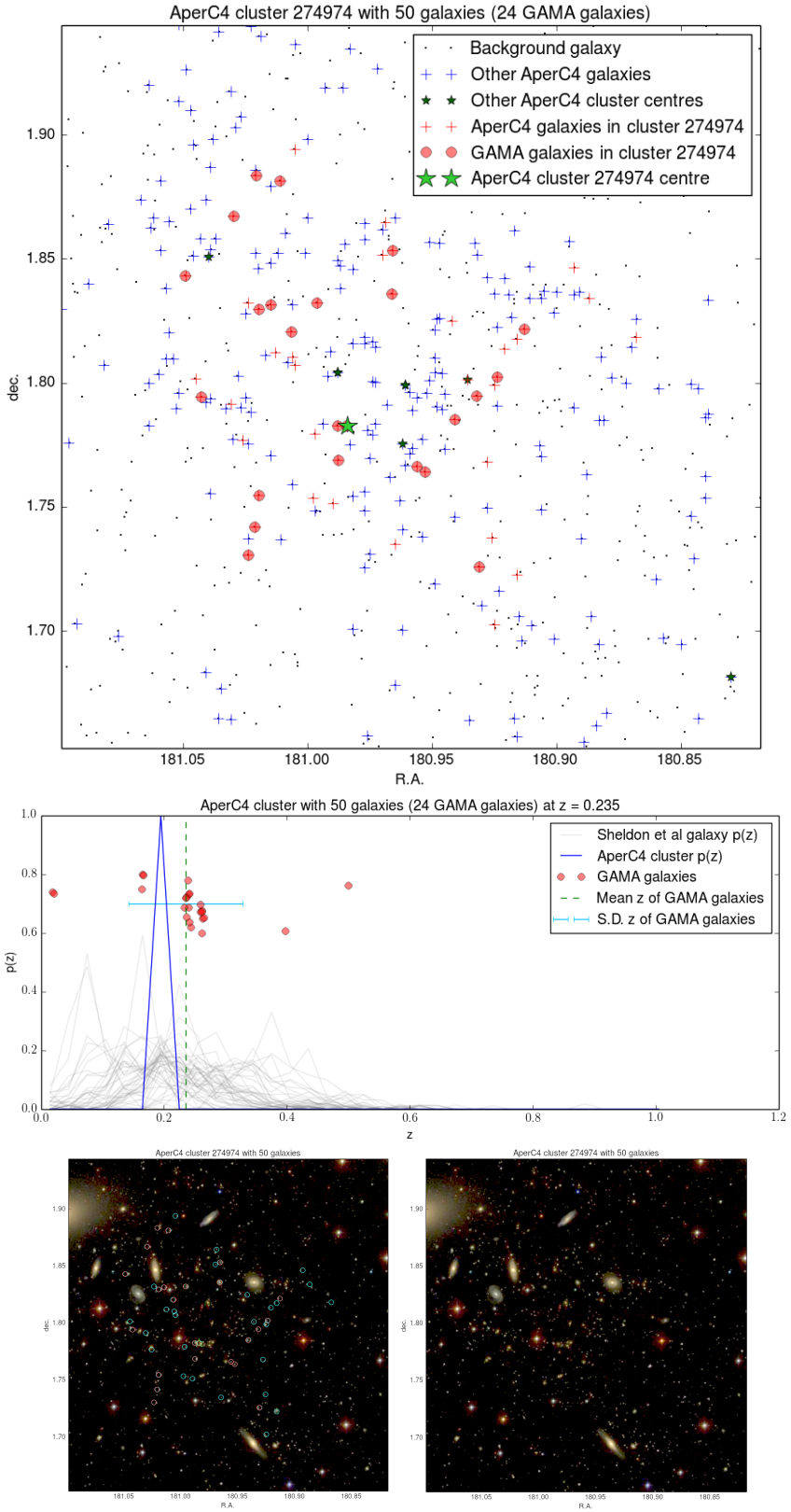


Figure 7.15: This figure shows AperC4 cluster ID 274974 and its 50 members. A description of the subplots is given in Figure 7.12. This cluster centre is identified close to many other cluster centres, including cluster 274725 (Figure 7.14). The spectroscopic redshifts from GAMA indicate that the two apparent clusters do not occur at the peak cluster $p(z)$. This may be an artefact of shot noise in the S12 $p(z)$ catalog or the general quality of the member $p(z)$s being less accurate than those constituting cluster 274725.

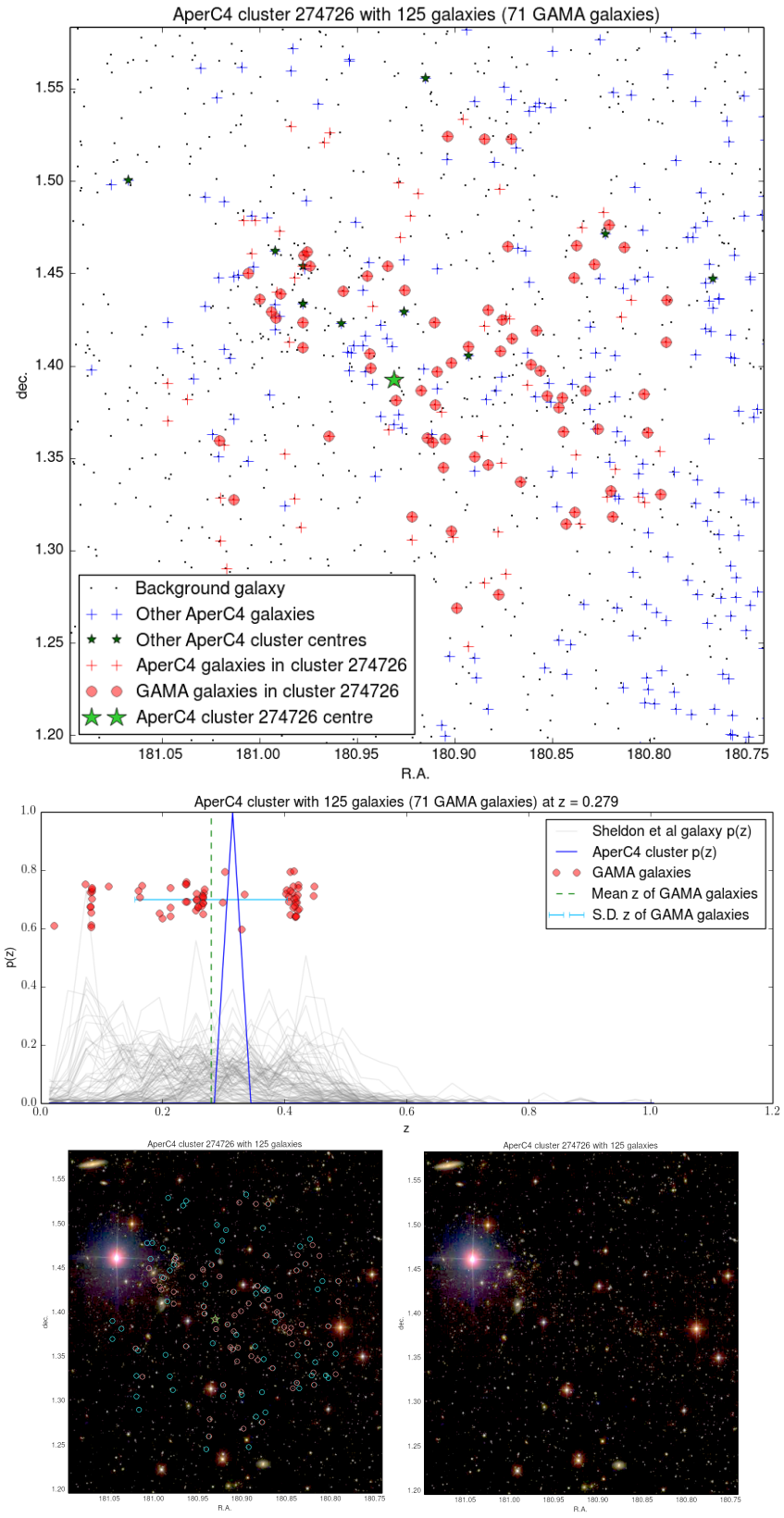


Figure 7.16: This figure shows AperC4 cluster ID 274726 and its 125 members. A description of the subplots is given in Figure 7.12. The spectroscopic members of this cluster appear across a large redshift range ($\delta z \approx 0.5$), with three apparent concentrations at $z \approx 0.08$, $z \approx 0.26$, and $z \approx 0.42$.

redshift range (taking the galaxy $N(z)$ in Figure 7.8, to be representative of the true redshift distribution of galaxies), so this projected structure is likely sourced from the way that AperC4 is combining galaxies into clusters.

Looking at the S12 galaxy $p(z)$s in the example clusters at the end of section 7.3.2, we see that the individual galaxy $p(z)$s extend across several redshift bins, sometimes with clear peaks (e.g. Figure 7.13) but mostly with no clear single redshift (Figures 7.14 to 7.16). In clusters where the majority of the galaxies do not have a singly preferred $p(z)$ bin, it is possible for galaxies to be drawn from multiple $p(z)$ bins so long as they have a finite probability of sharing a single bin. This is allowed by the optimal cluster cross-$p(z)$ threshold being set at zero (§6.3.3), allowing any aperture-slice clusters to merge so long as they have a shared membership. This is favoured by the F1 score as increasing the threshold adds more impurities (false associations) to the cluster catalog than real clusters. Only a handful of their respective members indicating a probability greater than $p(z) > 0.4$ of lying in any redshift bin. This noisy $p(z)$ information may also be partly responsible for unsuccessful deblends in the cluster finding algorithm.

The generous merging allowed by this cross-$p(z)$ threshold is also responsible for the large angular diameter ($\gtrsim 10'$) of clusters at $z > 0.1$ (e.g., Figures 7.14 to 7.16). I note here that the optimal catalog parameters when given ideal redshift information in a simulation (§6.4.2) showed that there is call for a more stringent cross-$p(z)$ threshold when redshifts better approximate the truth. However, the degree to which this stringency is applied was not well characterised, nominally due to the inherent bias of the distribution of clusters with redshift.

The optimal cluster cross-$p(z)$ threshold being set at zero also incurs a second penalty for the richest clusters, which is the fidelity of numbers under the programming language IDL. Once probabilities drop below $\sim 1 \times 10^{-340}$, IDL suffers a floating underflow, such that numbers below this limit are set to zero. As the minimum probability in any galaxy $p(z)$ is set to $\sim 10^{-3}$, and cluster $N_{\rm gals}$ can exceed 112 members (see Figure 7.9), it is possible for such a cluster $p(z)$ bin to go to zero whilst not accounting for the majority of cluster members. This may cause the richest clusters to be placed in the non-optimal redshift bin.

Line-of-sight structures appear to be where the galaxies at redshifts not in the $0.24 < z < 0.30$ and $0.33 < z < 0.36$ ranges (henceforth known at the peak AperC4 $p(z)$ bins) are being assigned to clusters, resulting in the cluster $N(z)$ seen in Figure 7.8. A thorough investigation of the catalog reveals that there are very few clusters seen at redshifts other

than the peak bins unless they are in celestial (RA/dec) proximity to clusters in the peak bins (Figure 7.16). This implies that likely cluster candidate galaxies are only being selected where they occur in celestial proximity to clusters in the peak AperC4 redshift bins. This in turn implies that the galaxy p-values tend towards favouring galaxies in these regions as being colour-clustered, whilst disfavouring galaxies clustered at other redshifts at other celestial locations.

The effect of galaxies in clusters not in the peak AperC4 $p(z)$ bins not appearing in the catalog is attributed to the fact that AperC4 does not employ redshift information when assigning p-values to the input galaxy catalog (§3.2.3). By *not* limiting the redshift extent of the aperture, cluster galaxies located at the redshift peak of the SDSS survey appear more clustered than the rest of the survey, as it will appear to have a higher density of galaxies than the rest of the survey (Equation 3.2). Furthermore, since the colour box created for a galaxy *not* in a peak AperC4 $p(z)$ bin is less likely to land on a similar coloured galaxy (§3.2.3), and since galaxies cluster (to some extent) in colour space, the model measurements for these galaxies will be artificially low, meaning their p-values will suggest they are non-fieldlike. This means that the galaxies not in the peak AperC4 $p(z)$ bins are not necessarily in clusters, but are drawn randomly, and thus groups/clusters found outside of the denser regions of Figure 7.7 are likely line-of-sight associations rather than true clusters.

7.4.2 Potential Improvements to AperC4

AperC4 functions, but it is not fit to identify clusters over a range of redshifts. By only considering AperC4-clustering on angular scales, AperC4 is biased to selecting galaxies in regions where the projected surface density of galaxies is maximised, which occur around cluster regions where the redshift density of galaxies is the most complete with respect to survey depth. At the same time, the current handling of $p(z)$ information may be volatile for rich clusters ($N_{\text{gals}} \gtrsim 100$), which may lead to the construction of inadequate $p(z)$s for those clusters. I suggest the following modifications be applied to AperC4 to alleviate these behaviours.

Limit Colour-Space Relocation of Model during p-value Measurement

One of the motivating factors in building AperC4 was to decouple galaxy redshifts from cluster galaxy selection (§3.1.2), only using redshift information after clusters had been identified (Steps 6 through 8, §3.2.7 to 3.2.9, respectively). This modification is made in

attempt to retain this feature.

During the p-value measurement process of the ApERC4 algorithm (Step 2, §3.2.3), the model count is constructed by moving the aperture to random galaxy locations across the survey footprint (as with $C4_{M05}$, §2.2.2). However, in $C4_{M05}$, the model count is limited to galaxies within the same $50\,h^{-1}$ Mpc bin, allowing C4-clustering to distinguish between clustered and non-clustered regions at various redshift intervals. By only applying angular limits to the colour box, ApERC4 aimed to maximise colour-clustering within a circular aperture (ApERC4-clustering) for a range of radii, but only succeeded in doing so where galaxies are aligned with clusters in the densest projected redshift slices.

Where the colour-aperture model fails, is the appreciation that when considering a target galaxy's colour, it is likely to be situated near galaxies of similar colour (as inferred by hierarchical structure formation, §1.2.2). Thus moving the colour box established by the target galaxy to the location of another galaxy, of a different colour, will up-weight the target colour-aperture's ApERC4-clustering density against the model. The reason this wasn't considered a problem, is the target galaxy is not counted in the p-value calculation, and the surrounding galaxies were assumed to occupy colour space randomly.

My proposed solution would be to ensure the model colour-aperture is centred on other galaxies *of similar colour*. Then the model should provide a consistent measure against which the target galaxy's ApERC4-clustering can be evaluated. This added constraint on the construction of the model can be justified by the fact that the aperture-colour space of the target galaxy already contains at least one galaxy of that colour: the target galaxy itself.

Effectively, this rephrases the question posed by the ApERC4 aperture-colour measurement from "How does the aperture-colour richness local to the target galaxy compare with any other non-local galaxy?" to "Given that ApERC4 has measured an aperture-colour richness around a galaxy of specified colour, how does that aperture-colour richness compare to a non-local galaxy of *the same* colour?", where non-locality is in the celestial sphere. Tolerance on the *sameness* of the non-local galaxy colour will need investigation. Whilst this may be sensitive to line-of-sight contamination as galaxy colours do not indicate a redshift by themselves, it encompasses the fact that the colour-space occupied by galaxies in galaxy clusters is not unique to galaxy clusters.

Calculate $p(z)$ in Log Space

Converting the product-sum given by Equation 3.3 to a log space calculation should be a relatively straightforward conversion. In doing so the product-sum of the probabilities will become the sum of the log-probabilities. This should allow calculation of values below the lower IDL limit of $\sim 1 \times 10^{-340}$ (in log-space this is merely -340) to be made consistently before exponentiating the log-probability to a normalised $p(z)$.

7.5 Summary

In section 7.1, I introduced SDSS DR8, highlighting the *übercalibration* update made to SDSS. This was followed by the S12 selection of galaxies in section 7.1.2. After investigating the S12 galaxies, examining various SDSS flags and surface brightness properties, I apply magnitude limits and surface brightness cuts in section 7.1.3 to produce the galaxy catalog to be run through the AperC4 algorithm.

In section 7.2, I ran through the steps of AperC4, observing how it identifies cluster galaxies until production of the final AperC4-DR8 cluster, $p(z)$, and member catalogs. In section 7.3, I described the output catalog formats and summarised the distribution of the AperC4 clusters. In section 7.3.2, I employed spectroscopic GAMA data to qualitatively assess the success and failure modes of AperC4, focussing on the assignment of galaxies to clusters and the redshift dispersion of GAMA galaxies in those clusters.

In section 7.4, I discussed the unexpected structure in the AperC4-DR8 cluster catalog (§7.4.1), using the investigation of AperC4 clusters with GAMA data in the previous section to determine shortcomings/biases in the AperC4 algorithm. I concluded with section 7.4.2 by discussing possible improvements to the AperC4 algorithm to redress these shortcomings.

Chapter 8

Thesis Summary and Further Work

8.1 Summary of Thesis

In this thesis, I presented the AperC4 cluster finder as a spiritual successor to the $C4_{M05}$ cluster finder (Miller et al., 2005). Also presented, is an optimisation framework using F1-scores, which helped tune the parameters of the AperC4 algorithm with the SDSS CatSim. Once an optimal parameter set was determined, AperC4 was applied to SDSS DR8 and the accompanying galaxy $p(z)$ catalog from Sheldon et al. (2012), finding 83,880 galaxy clusters.

In chapter 2, I reviewed at the $C4_{M05}$ algorithm in detail, describing key features such as the p-value colour-clustering measurement, the False Discovery Rate (FDR), and the kth Nearest Neighbour centering approximation. After summarising the successes of $C4_{M05}$, I discussed its shortcomings, highlighting its limitations with regard to use of spectroscopic galaxies, its inflexibility, and its use of non-concordance cosmology.

I presented the new AperC4 algorithm in chapter 3, providing motivations for its inception using the lessons learned from $C4_{M05}$. I explained the algorithm workflow and its associated parameters and provided theoretical scenarios for AperC4's desired behaviour for finding clusters. Of note, AperC4 does not use any cosmological information to find clusters, allowing output cluster catalogs to be used as cosmological measures without the need to determine its sensitivity to a given cosmology if some other cosmology is input. AperC4 is designed to use $p(z)$ information, which means it is able to use photometric or spectroscopic redshift information to determine clusters. AperC4 is also useful as an astrophysical probe of galaxy clusters, since galaxies have not been rejected on the basis

that they aren't in a red sequence.

In chapter 4, I introduced the Dark Energy Survey (DES) and the DES Catalog Simulations, CatSim. I employed a version of CatSim configured to simulate SDSS to test AperC4. To optimise the parameter set used by AperC4 on CatSim, I formulated the evaluation framework presented in chapter 5 which builds upon work by Brian Gerke for the DES Cluster Working Group. The key to the matching algorithm is its use of membership matching to relate galaxies that belong to halos (from the simulation) with the galaxies that are identified as being in clusters (as identified by a cluster finder). My major contribution was recognising that the selection functions of the cluster finders were previously misreported due in situations where the number of clusters in a catalog differs from the number of halos it is trying to locate. To remedy this, I calibrate the clusters (found by the cluster finder) to the masses of the halos (in the simulation) where they are least ambiguously matched (two-way unique matched), and interpolate the masses of the more ambiguous clusters on the basis of rank, where rank is a proxy to cluster mass. My other contribution was the inclusion of the F1-score, or F-measure, which combines a single cluster catalog's completeness and purity information, across the redshift and mass ranges of the clusters recovered from the simulation, into a single number. This F1-score is employed to objectively optimise the AperC4 parameter set such that it maximises completeness and purity simultaneously.

Following the introduction of AperC4, CatSim, and the cluster catalog evaluation framework, chapter 6 presents the applications of AperC4 to CatSim with zCarlos. 6,048 catalogs were produced and evaluated using unique combinations of AperC4 parameters, ranking methods, and galaxy redshift sources. The outcome of this chapter was an optimal set of AperC4 parameters for use on the SDSS (and SDSS-like) galaxy catalog(s). I examined the selection function of the optimum AperC4-CatSim-zCarlos catalog. To characterise the impact of probabilistic redshift information, I also examined the selection function of an idealised AperC4-CatSim catalog, where the galaxy redshifts were informed by the simulation input, and compared it with that of the AperC4-CatSim-zCarlos catalog.

With the optimal AperC4-SDSS parameter set in hand, I applied AperC4 to the SDSS DR8 galaxy catalog in chapter 7. After introducing SDSS DR8 and the S12 $p(z)$ catalog, I applied my own cuts to the S12 selection, in surface brightness and magnitude limits, to produce a galaxy catalog suitable for cluster finding. The key product of this chapter, and indeed of this thesis, is the AperC4-SDSS DR8 cluster catalog. After finding

some structure in the AperC4-SDSS DR8 cluster distribution, I used GAMA data to interrogate a subsample of the cluster catalog, and individual clusters in that subsample, to establish the success and failure modes of AperC4. After determining likely causes of the apparent structure and failure modes, I made suggestions for further work.

8.2 Further Work

In this section, I will discuss avenues for improvement of the AperC4 algorithm, before discussing uses for the evaluation framework and AperC4-SDSS DR8 catalog.

8.2.1 Algorithm Development

AperC4 has undergone much development from its $C4_{M05}$ origins. For example, Dietrich et al. (2014) investigated orientation bias of cluster finding on CatSim halos, utilising a photometric version of C4 that still employed the esoteric cosmological assumptions of $C4_{M05}$ (§2.4.3). Despite this shortcoming, the orientation bias measured from clusters output by the photometric C4 were found to be highly competitive with the other cluster finding techniques examined.

As discussed in section 7.4, AperC4 experiences various selection effects that may not suit it to cluster finding over a broad range of redshifts. The key improvements cited in section 7.4.2 were limiting the placement of the model colour-box to galaxies within the same colour space, and calculating the cluster $p(z)$ in log space to avoid computational arithmetic issues. Beyond these, there are a number of avenues that could be pursued to further enhance the AperC4 algorithm. I discuss three of these avenues below.

Cluster Centering

Beyond the initial seeding of the clusters with $k-$NN during the formation of the aperture-slice clusters, no work is performed by AperC4 to properly centre the clusters once the aperture-slices are merged. Centering could be avoided by including the memberships of each cluster, which allowed unambiguous matching to halo galaxies when using simulated data (chapter 6). However, centering information would be useful so that the optical clusters produced by AperC4 can be matched more readily to observed cluster catalogs that do not have membership information (e.g., X-ray clusters, Mehrtens et al. 2011; or SZ detections Reese et al. 2011; Planck Collaboration et al. 2013a) and hence cannot be unambiguously matched.

Using Alternative Photometric Redshift Data

Since most of work in cluster identification done before looking at redshift information, it is possible for AperC4 to be run with different redshift algorithms. Even if the catalog quality is non-optimal, it can still be compared to other redshift algorithms in the same way the AperC4-CatSim-zCarlos catalogs were compared to AperC4-CatSim catalogs that employed the input simulation redshifts (§6.3).

Increasing the Number of Apertures used by AperC4

Since including larger apertures led to a better F1-score with the basic investigation carried out in chapter 6, it may be worth investigating whether supplying AperC4 with more aperture radii increases the quality of the output catalog.

In anticipation of running AperC4 on deeper photometric datasets, such as those from DES (chapter 4, Tucker et al., 2007) and future data such as from LSST (LSST Collaboration, 2012), and Euclid (Laureijs et al., 2011), it would be worthwhile investigating the inclusion of smaller aperture sizes (i.e., extending the redshift range of Table 3.3).

8.2.2 Evaluation Framework

In terms of evaluating AperC4, running the cluster finder on larger simulated areas would give a better handle on high mass F1-scores. Additionally, it would be useful to run AperC4 on different simulation datasets, such that the F1-score can be used to characterise the cluster finder without being limited to sample variance, or indeed, cosmic variance. As AperC4 has been shown to be adaptable to the data it is applied to (running on CatSim with zCarlos $p(z)$s and SDSS DR8 with S12 $p(z)$s), this work is only impeded by available compute time.

In a larger sense, whilst it is useful to use the evaluation framework to optimise AperC4 or any other cluster finder on its own, it would also be useful to compare cluster finders to each other. This would allow an end user to select an appropriate cluster finder catalog (from a suite of such catalogs) through the knowledge of its strengths and weaknesses at various redshifts and mass ranges in comparison to other cluster finders. The evaluation framework as presented in this thesis has been distributed to the DES Brazil group*, who have incorporated it into their science portal.

* `http://des-brazil.linea.gov.br/`

8.2.3 Uses for the AperC4-SDSS DR8 catalog

Following a similar vein to $C4_{M05}$(§2.3.5), the AperC4-SDSS DR8 catalog can be used for a variety of cluster population studies such as cluster luminosity gaps (Milosavljević et al., 2006) or the evolution of the red sequence (De Lucia et al., 2007). Given that AperC4 does not restrict itself to red sequence cluster finding, it is well suited to studying the dependence of morphology and colour on environment (Bamford et al., 2009) and cluster blue fractions (Aguerri et al., 2007).

In terms of cosmology, and pending algorithm corrections, it may be possible to look at the abundance of AperC4-SDSS DR8 clusters to measure cosmology directly from the cluster catalog (§1.3.1; Lima and Hu, 2004), performing a similar analysis to Rozo et al. (2007) on MaxBCG clusters.

Bibliography

Kevork Abazajian, Jennifer K Adelman-McCarthy, Marcel A Agüeros et al. (2004). The Second Data Release of the Sloan Digital Sky Survey. *The Astronomical Journal*, 128:502. Cited on 19, 25, 26, 27, 30

K Abazajian and for the Sloan Digital Sky Survey (2008). The Seventh Data Release of the Sloan Digital Sky Survey. *arXiv*, astro-ph. Cited on 145

George O Abell (1958). The Distribution of Rich Clusters of Galaxies. *Astrophysical Journal Supplement*, 3:211. Cited on 4, 5, 19, 92, 112

George O Abell, Harold G Corwin, and Ronald P Olowin (1989). A catalog of rich clusters of galaxies. *Astrophysical Journal Supplement Series (ISSN 0067-0049)*, 70:1. Cited on 17

J. K. Adelman-McCarthy, M. A. Agüeros, S. S. Allam et al. (2006). The Fourth Data Release of the Sloan Digital Sky Survey. ApJS, 162:38–48, `arXiv:astro-ph/0507711`. Cited on 12

J A L Aguerri, R Sánchez-Janssen, and C Muñoz-Tuñón (2007). A study of catalogued nearby galaxy clusters in the SDSS-DR4. I. Cluster global properties. *Astronomy and Astrophysics*, 471:17. Cited on 45, 180

S. W. Allen, D. A. Rapetti, R. W. Schmidt et al. (2008). Improved constraints on dark energy from Chandra X-ray observations of the largest relaxed galaxy clusters. *Monthly Notices of the Royal Astronomical Society*, 383:879–896, `arXiv:0706.0033`. Cited on 9

James Annis, Marcelle Soares-Santos, Michael A Strauss et al. (2011). The SDSS Coadd: 275 deg^2 of Deep SDSS Imaging on Stripe 82. *arXiv.org*, astro-ph.CO. Cited on 145

F J Anscombe (1948). The transformation of Poisson, binomial and negative-binomial data. *Biometrika*, 35:246–254. Cited on 35

S. Arnouts and O. Ilbert (2011). LePHARE: Photometric Analysis for Redshift Estimate. Astrophysics Source Code Library. Cited on 50

D. J. Bacon, A. R. Refregier, and R. S. Ellis (2000). Detection of weak gravitational lensing by large-scale structure. *Monthly Notices of the Royal Astronomical Society*, 318:625–640, `arXiv:astro-ph/0003008`. Cited on 84

N. A. Bahcall (1977). X-ray clusters of galaxies - Correlations with optical morphology and galaxy density. *The Astrophysical Journal, Letters*, 217:L77–L82. Cited on 5

N. A. Bahcall (2000). Clusters and Groups of Galaxies. In A. N. Cox, editor, *Allen's Astrophysical Quantities*, page 613. Springer-Verlag. Cited on 47

Steven P Bamford, Robert C Nichol, Ivan K Baldry et al. (2009). Galaxy Zoo: the dependence of morphology and colour on environment. *Monthly Notices of the Royal Astronomical Society*, 393(4):1324–1352. Cited on 45, 180

Manda Banerji, Filipe B Abdalla, Ofer Lahav, and Huan Lin (2008). Photometric redshifts for the Dark Energy Survey and VISTA and implications for large-scale structure. *Monthly Notices of the Royal Astronomical Society*, 386:1219. Cited on 30, 85, 145

M Bartelmann and S D M White (2002). Cluster detection from surface-brightness fluctuations in SDSS data. *Astronomy and Astrophysics*, 388(2):732–740. Cited on 24

Y Benjamini and Y Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300. Cited on 35

Andreas A Berlind, Joshua Frieman, David H Weinberg et al. (2006). Percolation Galaxy Groups and Clusters in the SDSS Redshift Survey: Identification, Catalogs, and the Multiplicity Function. *The Astrophysical Journal Supplement Series*, 167:1. Cited on 45

A. A. Berlind and D. H. Weinberg (2002). The Halo Occupation Distribution: Toward an Empirical Determination of the Relation between Galaxies and Mass. *The Astrophysical Journal*, 575:587–616, `arXiv:astro-ph/0109001`. Cited on 88

Mariangela Bernardi, Joseph B Hyde, Ravi K Sheth, Chris J Miller, and Robert C Nichol (2007). The Luminosities, Sizes, and Velocity Dispersions of Brightest Cluster Galaxies: Implications for Formation History. *The Astronomical Journal*, 133:1741. Cited on 13, 43, 45

Mariangela Bernardi, Robert C Nichol, Ravi K Sheth, C J Miller, and J Brinkmann (2006). Evolution and Environment of Early-Type Galaxies. *The Astronomical Journal*, 131:1288. Cited on 45

G. Bernstein and D. Huterer (2010). Catastrophic photometric redshift errors: weak-lensing survey requirements. MNRAS, 401:1399–1408, `arXiv:0902.2782 [astro-ph.CO]`. Cited on 121

J. P. Bernstein, R. Kessler, S. Kuhlmann, and H. Spinka (2009). Dark Energy Survey Supernovae: Simulations and Survey Strategy. *ArXiv e-prints*, `arXiv:0906.2955 [astro-ph.CO]`. Cited on 86

P N Best, A von der Linden, G Kauffmann, T M Heckman, and C R Kaiser (2007). On the prevalence of radio-loud active galactic nuclei in brightest cluster galaxies: implications for AGN heating of cooling flows. *Monthly Notices of the Royal Astronomical Society*, 379:894. Cited on 45

Michael R Blanton, J Brinkmann, István Csabai et al. (2003a). Estimating Fixed-Frame Galaxy Magnitudes in the Sloan Digital Sky Survey. *The Astronomical Journal*, 125:2348. Cited on 26

Michael R Blanton, Huan Lin, Robert H Lupton et al. (2003b). An Efficient Targeting Strategy for Multiobject Spectrograph Surveys: the Sloan Digital Sky Survey "Tiling" Algorithm. *The Astronomical Journal*, 125:2276. Cited on 42, 89

M. R. Blanton, R. H. Lupton, D. J. Schlegel et al. (2005). The Properties and Luminosity Function of Extremely Low Luminosity Galaxies. *The Astrophysical Journal*, 631:208–230, `arXiv:astro-ph/0410164`. Cited on 151

H. Böhringer, G. Chon, and C. A. Collins (2014). The extended ROSAT-ESO Flux Limited X-ray Galaxy Cluster Survey (REFLEX II). IV. X-ray luminosity function and first constraints on cosmological parameters. A&A, 570:A31, `arXiv:1403.2927 [astro-ph.CO]`. Cited on 103

J R Bond, S Cole, G Efstathiou, and N Kaiser (1991). Excursion set mass functions for hierarchical Gaussian fluctuations. *Astrophysical Journal*, 379:440–460. Cited on 11

A. Boselli and G. Gavazzi (2006). Environmental Effects on Late-Type Galaxies in Nearby Clusters. *The Publications of the Astronomical Society of the Pacific*, 118:517–559, `arXiv:astro-ph/0601108`. Cited on 8

Richard G Bower (1991). The evolution of groups of galaxies in the Press-Schechter formalism. *Monthly Notices of the Royal Astronomical Society*, 248:332–352. Cited on 11

R G Bower, J R Lucey, and R S Ellis (1992). Precision Photometry of Early Type Galaxies in the Coma and Virgo Clusters - a Test of the Universality of the Colour / Magnitude Relation - Part Two - Analysis. *R.A.S. MONTHLY NOTICES V.254*, 254:601. Cited on 5, 13, 33

Fabrice Brimioulle, Michael Lerchster, Stella Seitz, Ralf Bender, and Jan Snigula (2008). Photometric redshifts for the CFHTLS-Wide. *arXiv*, astro-ph. Cited on 145

M. Brüggen and C. R. Kaiser (2002). Hot bubbles from active galactic nuclei as a heat source in cooling-flow clusters. *Nature*, 418:301–303, `arXiv:astro-ph/0207354`. Cited on 8

J M Budzynski, S Koposov, I G McCarthy, S L McGee, and V Belokurov (2012). The radial distribution of galaxies in groups and clusters. *arXiv.org*, astro-ph.CO. Cited on 10, 13

M. Busha and R. H. Wechsler (2008). Making Mock Galaxy Catalogs with ADDGALS. In *The Rencontres de Moriond (Cosmology Session)*. As used by the DES Collaboration. Cited on 88, 89

Charles E. Catlett (2005). Teragrid: a foundation for us cyberinfrastructure. In *Proceedings of the 2005 IFIP international conference on Network and Parallel Computing*, NPC'05, pages 1–1, Berlin, Heidelberg. Springer-Verlag. Cited on 87

D. Clowe, A. Gonzalez, and M. Markevitch (2004). Weak-Lensing Mass Reconstruction of the Interacting Cluster 1E 0657-558: Direct Evidence for the Existence of Dark Matter. *The Astrophysical Journal*, 604:596–603, `arXiv:astro-ph/0312273`. Cited on 7, 8

Chris A Collins, John P Stott, Matt Hilton et al. (2009). Early assembly of the most massive galaxies. *Nature*, 458:603. Cited on 13

A. A. Collister and O. Lahav (2004). ANNz: Estimating Photometric Redshifts Using Artificial Neural Networks. PASP, 116:345–351, `astro-ph/0311058`. Cited on 50, 121

Warrick J Couch, Richard S Ellis, Iain MacLaren, and David F Malin (1991). A uniformly selected catalogue of distant galaxy clusters. *Royal Astronomical Society*, 249:606. Cited on 15

M. Crocce, S. Pueblas, and R. Scoccimarro (2006). Transients from initial conditions in cosmological simulations. *Monthly Notices of the Royal Astronomical Society*, 373:369–381, arXiv:astro-ph/0606505. Cited on 88

István Csabai, Tamás Budavári, Andrew J Connolly et al. (2003). The Application of Photometric Redshifts to the SDSS Early Data Release. *The Astronomical Journal*, 125:580. Cited on 50, 121

C. Cunha, D. Huterer, and J. A. Frieman (2009). Constraining dark energy with clusters: Complementarity with other probes. *Phys Rev D*, 80(6):063532, arXiv:0904.1589 [astro-ph.CO]. Cited on 86

C. E. Cunha, D. Huterer, H. Lin, M. T. Busha, and R. H. Wechsler (2012). Spectroscopic failures in photometric redshift calibration: cosmological biases and survey requirements. *ArXiv e-prints*, arXiv:1207.3347 [astro-ph.CO]. Cited on 30, 87

Carlos E Cunha, Marcos Lima, Hiroaki Oyaizu, Joshua Frieman, and Huan Lin (2009). Estimating the redshift distribution of photometric galaxy samples - II. Applications and tests of a new method. *Monthly Notices of the Royal Astronomical Society*, 396(4):2379–2398. Cited on 50, 51, 121, 122, 125

Ryan Curtin, James Cline, Neil Slagle, Matthew Amidon, and Alexander Gray (2011). MLPACK: A Scalable C++ Machine Learning Library. In *BigLearning: Algorithms, Systems, and Tools for Learning at Scale*. Cited on 53, 59

J. T. A. de Jong, G. A. Verdoes Kleijn, K. H. Kuijken, and E. A. Valentijn (2013). The Kilo-Degree Survey. *Experimental Astronomy*, 35:25–44, arXiv:1206.1254 [astro-ph.CO]. Cited on 50

Gabriella De Lucia, Bianca M Poggianti, Alfonso Aragón-Salamanca et al. (2007). The build-up of the colour-magnitude relation in galaxy clusters since z 0.8. *Monthly Notices of the Royal Astronomical Society*, 374:809. Cited on 13, 45, 60, 180

H. T. Diehl, R. Angstadt, J. Campa et al. (2008). Characterization of DECam focal plane detectors. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 7021 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*. Cited on 84

J. P. Dietrich, Y. Zhang, J. Song et al. (2014). Orientation bias of optically selected

galaxy clusters and its impact on stacked weak-lensing analyses. MNRAS, 443:1713–1722, arXiv:1405.2923. Cited on 178

P. Doel, T. Abbott, M. Antonik et al. (2008). Design and status of the optical corrector for the DES survey instrument. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 7014 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*. Cited on 83

Michael J Drinkwater, Russell J Jurek, Chris Blake et al. (2010). The WiggleZ Dark Energy Survey: survey design and first data release. *Monthly Notices of the Royal Astronomical Society*, 401(3):1429–1452. Cited on 46

S P Driver, D T Hill, L S Kelvin et al. (2010). Galaxy and Mass Assembly (GAMA): survey diagnostics and core data release. *arXiv.org*, 1009.0614v1. Cited on 144, 163

Simon P Driver and the GAMA team (2008). Galaxy And Mass Assembly (GAMA). *arXiv.org*, 0807.0376v1. Cited on 46, 158

Maret Einasto, Erik Tago, Heidi Lietzen et al. (2014). Tracing a high redshift cosmic web with quasar systems. *Astronomy & Astrophysics*, 568:A46. Cited on 72

D J Eisenstein, J Annis, J E Gunn et al. (2001). Spectroscopic Target Selection for the Sloan Digital Sky Survey: The Luminous Red Galaxy Sample. *arXiv*, astro-ph. Cited on 27

Daniel J Eisenstein, David H Weinberg, Eric Agol et al. (2011). SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way Galaxy, and Extra-Solar Planetary Systems. *arXiv*, astro-ph.IM. Cited on 46, 145, 153

D. J. Eisenstein, I. Zehavi, D. W. Hogg et al. (2005). Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies. *The Astrophysical Journal*, 633:560–574, arXiv:astro-ph/0501171. Cited on 85

Vincent R Eke, Shaun Cole, and Carlos S Frenk (1996). Using the Evolution of Clusters to Constrain Omega. *Monthly Notices of the Royal Astronomical Society*, 282:263–280, astro-ph/9601088v1. Cited on 10, 103

A. E. Evrard, T. J. MacFarland, H. M. P. Couchman et al. (2002). Galaxy Clusters in Hubble Volume Simulations: Cosmological Constraints from Sky Survey Populations. *The Astrophysical Journal*, 573:7–36, arXiv:0110246. Cited on 87

S Farrens, F B Abdalla, E S Cypriano, C Sabiu, and C Blake (2011). Friends-of-friends groups and clusters in the 2SLAQ catalogue. *Monthly Notices of the Royal Astronomical Society*, 417:1402. Cited on 30

R. Fassbender, A. Nastasi, H. Böhringer et al. (2011). The X-ray luminous galaxy cluster XMMU J1007.4+1237 at z = 1.56. The dawn of starburst activity in cluster cores. *Astronomy and Astrophysics*, 527:L10, `arXiv:1101.3313 [astro-ph.CO]`. Cited on 9, 10

J. E. Felten, R. J. Gould, W. A. Stein, and N. J. Woolf (1966). X-Rays from the Coma Cluster of Galaxies. *The Astrophysical Journal*, 146:955–958. Cited on 6

Brenna Flaugher (2005). The Dark Energy Survey. *International Journal of Modern Physics A*, 20:3121. Cited on 145

B. L. Flaugher, T. M. C. Abbott, J. Annis et al. (2010). Status of the dark energy survey camera (DECam) project. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 7735 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*. Cited on 83

W. Forman, J. Schwarz, C. Jones, W. Liller, and A. C. Fabian (1979). X-ray observations of galaxies in the Virgo cluster. *The Astrophysical Journal*, 234:L27–L31. Cited on 6

P. Fosalba, E. Gaztañaga, F. J. Castander, and M. Manera (2008). The onion universe: all sky lightcone simulations in spherical shells. *Monthly Notices of the Royal Astronomical Society*, 391:435–446, `arXiv:0711.1540`. Cited on 87

Joshua Frieman, Michael Turner, and Dragan Huterer (2008). Dark Energy and the Accelerating Universe. *arXiv*, astro-ph. Cited on 86

M Fukugita, T Ichikawa, J E Gunn et al. (1996). The Sloan Digital Sky Survey Photometric System. *Astronomical Journal v.111*, 111:1748. Cited on 23, 42

Ronald Garcia and Andrew Lumsdaine (2005). MultiArray: a C++ library for generic programming with arrays. *Software: Practice and Experience*, 35(2):159–188. Cited on 55

J. P. Gardner, A. Connolly, and C. McBride (2007). Enabling Rapid Development of Parallel Tree Search Applications. *ArXiv e-prints*, `arXiv:0709.1967`. Cited on 88

James E Geach, David N A Murphy, and Richard G Bower (2011). 4098 galaxy clusters to z 0.6 in the Sloan Digital Sky Survey equatorial Stripe 82. *arXiv*, astro-ph.CO. Cited on 18

M. J. Geller, A. Diaferio, and M. J. Kurtz (1999). The Mass Profile of the Coma Galaxy Cluster. *The Astrophysical Journal, Letters*, 517:L23–L26, `arXiv:astro-ph/9903305`. Cited on 14

David W Gerdes, Adam J Sypniewski, Timothy A McKay et al. (2009). ArborZ: Photometric Redshifts Using Boosted Decision Trees. *arXiv*, astro-ph.CO. Cited on 50

David G Gilbank, M D Gladders, H K C Yee, and B C Hsieh (2011). The Red-Sequence Cluster Survey-2 (RCS-2): Survey Details and Photometric Catalog Construction. *The Astronomical Journal*, 141(3):94. Cited on 30

S Giodini, D Pierini, A Finoguenov et al. (2009). Stellar and total baryon mass fractions in groups and clusters since redshift 1. *arXiv.org*, `0904.0448v2`. Cited on 112

Michael D Gladders and H K C Yee (2000). A New Method For Galaxy Cluster Detection. I. The Algorithm. *The Astronomical Journal*, 120:2148. Cited on 16, 30, 37

Tomotsugu Goto, Maki Sekiguchi, Robert C Nichol et al. (2002). The Cut-and-Enhance Method: Selecting Clusters of Galaxies from the Sloan Digital Sky Survey Commissioning Data. *The Astronomical Journal*, 123:1807. Cited on 18, 33

J Gunn, M Carr, C Rockosi et al. (1998). The Sloan Digital Sky Survey Photometric Camera. *The Astronomical Journal*, 116:3040. Cited on 23

Jiangang Hao, Timothy A McKay, Benjamin P Koester et al. (2010). A GMBCG Galaxy Cluster Catalog of 55,437 Rich Clusters from SDSS DR7. *arXiv*, astro-ph.CO. Cited on 20, 30, 31, 37

Craig D Harrison, Christopher J Miller, Joseph W Richards et al. (2012). The XMM Cluster Survey: The Stellar Mass Assembly of Fossil Galaxies. *arXiv.org*, astro-ph.CO. Cited on 8, 10

C. Heymans and A. Heavens (2003). Weak gravitational lensing: reducing the contamination by intrinsic alignments. *Monthly Notices of the Royal Astronomical Society*, 339:711–720, `arXiv:astro-ph/0208220`. Cited on 85

M. Hilton, E. Lloyd-Davies, S. A. Stanford et al. (2010). The XMM Cluster Survey: Active Galactic Nuclei and Starburst Galaxies in XMMXCS J2215.9-1738 at z = 1.46. *The Astrophysical Journal*, 718:133–147, `arXiv:1005.4692 [astro-ph.CO]`. Cited on 9

G. Hinshaw, D. Larson, E. Komatsu et al. (2013). Nine-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Parameter Results. ApJS, 208:19, `arXiv:1212.5226 [astro-ph.CO]`. Cited on 3, 68

J A Hodge, R H Becker, R L White, G T Richards, and G R Zeimann (2011). High-resolution VLA Imaging of SDSS Stripe 82 at 1.4 GHz. *arXiv*, astro-ph.CO. Cited on 145

E. Høg, C. Fabricius, V. V. Makarov et al. (2000). The Tycho-2 catalogue of the 2.5 million brightest stars. A&A, 355:L27–L30. Cited on 147

D. W. Hogg, M. R. Blanton, J. Brinchmann et al. (2004). The Dependence on Environment of the Color-Magnitude Relation of Galaxies. *The Astrophysical Journal Supplement Series*, 601:L29–L32, `arXiv:astro-ph/0307336`. Cited on 89

J P Huchra and M J Geller (1982). Groups of galaxies. I - Nearby groups. *Astrophysical Journal*, 257:423. Cited on 15, 19

J. B. Hutchings, A. Saintonge, D. Schade, and D. Frenette (2002). Galaxy Morphology in the Rich Cluster Abell 2390. *Astronomical Journal*, 123:1826–1837, `arXiv:astro-ph/0201042`. Cited on 8

D. Huterer, M. Takada, G. Bernstein, and B. Jain (2006). Systematic errors in future weak-lensing surveys: requirements and prospects for self-calibration. *Monthly Notices of the Royal Astronomical Society*, 366:101–114, `arXiv:astro-ph/0506030`. Cited on 85

N. Kaiser, H. Aussel, B. E. Burke et al. (2002). Pan-STARRS: A Large Synoptic Survey Telescope Array. In J. A. Tyson and S. Wolff, editors, *Survey and Other Telescope Technologies and Discoveries*, volume 4836 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 154–164. Cited on 50

Jeremy Kepner, Xiaohui Fan, Neta Bahcall et al. (1999). An Automated Cluster Finder: The Adaptive Matched Filter. *The Astrophysical Journal*, 517:78. Cited on 15

Rita Seung Jung Kim, Jeremy V Kepner, Marc Postman et al. (2002). Detecting Clusters of Galaxies in the Sloan Digital Sky Survey. I. Monte Carlo Comparison of Cluster Detection Algorithms. *The Astronomical Journal*, 123:20. Cited on 17, 18

Donnacha Kirk, Ofer Lahav, Sarah Bridle et al. (2013). Optimising Spectroscopic and Photometric Galaxy Surveys: Same-sky Benefits for Dark Energy and Modified Gravity. *eprint arXiv:1307.8062*. Cited on 50

Jean-Paul Kneib and Priyamvada Natarajan (2012). Cluster Lenses. *arXiv.org*, astro-ph.CO. Cited on 14

B P Koester, T A McKay, J Annis et al. (2007). A MaxBCG Catalog of 13,823 Galaxy Clusters from the Sloan Digital Sky Survey. *The Astrophysical Journal*, 660:239. Cited on 16, 19, 20, 30, 37

E. Komatsu, K. M. Smith, J. Dunkley et al. (2011). Seven-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Interpretation. *The Astrophysical Journal, Supplement*, 192:18, `arXiv:1001.4538 [astro-ph.CO]`. Cited on 9

R Laureijs, J Amiaux, S Arduini et al. (2011). Euclid Definition Study Report. *eprint arXiv:1110.3193*. Cited on 50, 179

I J Lewis, R D Cannon, K Taylor et al. (2002). The Anglo-Australian Observatory 2dF facility. *Monthly Notices of the Royal Astronomical Society*, 333(2):279–298. Cited on 163

I H Li and H K C Yee (2008). Finding Galaxy Groups in Photometric-Redshift Space: The Probability Friends-of-Friends Algorithm. *The Astronomical Journal*, 135:809. Cited on 31

Marcos Lima, Carlos E Cunha, Hiroaki Oyaizu et al. (2008). Estimating the redshift distribution of photometric galaxy samples. *Monthly Notices of the Royal Astronomical Society*, 390(1):118–130. Cited on 121, 122

Marcos Lima and Wayne Hu (2004). Self-calibration of cluster dark energy studies: Counts in cells. *Physical Review D*, 70(4):043504. Cited on 10, 180

H. Lin, N. Kuropatkin, R. Wechsler et al. (2010). Dark Energy Survey Simulations. In *American Astronomical Society Meeting Abstracts #215*, volume 42 of *Bulletin of the American Astronomical Society*, page 470.07. Cited on 87

Chris J Lintott, Kevin Schawinski, Anze Slosar et al. (2008). Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey . *Monthly Notices of the Royal Astronomical Society*, 389(3):1179–1189. Cited on 45

J Liske, D J Lemon, S P Driver, N J G Cross, and W J Couch (2003). The Millennium Galaxy Catalogue: 16 . *Monthly Notice of the Royal Astronomical Society*, 344(1):307–324. Cited on 163

The LSST Collaboration (2012). Large Synoptic Survey Telescope: Dark Energy Science Collaboration. *eprint arXiv:1211.0310*. Cited on 50, 179

Robert H Lupton, James E Gunn, and Alexander S Szalay (1999). A Modified Magnitude System that Produces Well-Behaved Magnitudes, Colors, and Errors Even for Low Signal-to-Noise Ratio Measurements. *The Astronomical Journal*, 118:1406. Cited on 24, 25

M. Markevitch, A. H. Gonzalez, D. Clowe et al. (2004). Direct Constraints on the Dark Matter Self-Interaction Cross Section from the Merging Galaxy Cluster 1E 0657-56. *The Astrophysical Journal*, 606:819–824, `arXiv:astro-ph/0309303`. Cited on 7, 8

B. J. Maughan, C. Jones, W. Forman, and L. Van Speybroeck (2008). Images, Structural Properties, and Metal Abundances of Galaxy Clusters Observed with Chandra ACIS-I at 0.1 ¡ z ¡1.3. *The Astrophysical Journal, Supplement*, 174:117–135, `arXiv:astro-ph/0703156`. Cited on 8, 14

C. McBride, A. A. Berlind, R. Scoccimarro et al. (2011). LasDamas: Accurate Determination of the Abundance of Galaxy Clusters. In *American Astronomical Society Meeting Abstracts #217*, volume 43 of *Bulletin of the American Astronomical Society*, page 249.07. Cited on 88

Nicola Mehrtens, A Kathy Romer, E J Lloyd-Davies et al. (2011). The XMM Cluster Survey: Optical analysis methodology and the first data release. *eprint arXiv:1106.3056*, astro-ph.CO:3056. Cited on 14, 84, 178

Felipe Menanteau, Jorge González, Jean-Baptiste Juin et al. (2010). The Atacama Cosmology Telescope: Physical Properties and Purity of a Galaxy Cluster Sample Selected via the Sunyaev-Zel'dovich Effect. *The Astrophysical Journal*, 723:1523. Cited on 15

Christopher J Miller, Robert C Nichol, Daniel Reichart et al. (2005). The C4 Clustering Algorithm: Clusters of Galaxies in the Sloan Digital Sky Survey. *The Astronomical*

Journal, 130:968. Cited on 13, 18, 22, 29, 30, 31, 32, 34, 39, 41, 42, 43, 44, 45, 46, 48, 49, 53, 56, 67, 77, 79, 80, 91, 95, 176

Miloš Milosavljević, Christopher J Miller, Steven R Furlanetto, and Asantha Cooray (2006). Cluster Merger Variance and the Luminosity Gap Statistic. *The Astrophysical Journal*, 637:L9. Cited on 45, 180

J. J. Mohr, R. Armstrong, E. Bertin et al. (2012). The Dark Energy Survey Data Processing and Calibration System. *ArXiv e-prints*, `arXiv:1207.3189 [astro-ph.IM]`. Cited on 82, 87

D Munshi, P Valageas, L vanwaerbeke, and A Heavens (2008). Cosmology with weak lensing surveys. *Physics Reports*, 462(3):67–121. Cited on 50

D N A Murphy, J E Geach, and R G Bower (2011). ORCA: The Overdense Red-sequence Cluster Algorithm. *arXiv*, astro-ph.CO. Cited on 17, 19, 30, 31

Kevin P Murphy (2012). *Machine Learning.* A Probabilistic Perspective. MIT Press. Cited on 94

Yookyung Noh and J D Cohn (2012). Disentangling correlated scatter in cluster mass measurements. *Monthly Notices of the Royal Astronomical Society*, 426(3):1829–1844. Cited on 96

Hiroaki Oyaizu, Marcos Lima, Carlos E Cunha, Huan Lin, and Joshua Frieman (2008). Photometric Redshift Error Estimators. *The Astrophysical Journal*, 689(2):709–720. Cited on 121

Nikhil Padmanabhan, Tamas Budavari, David J Schlegel et al. (2005). Calibrating photometric redshifts of luminous red galaxies. *Monthly Notices of the Royal Astronomical Society*, 359(1):237–250. Cited on 30, 121

Nikhil Padmanabhan, David J Schlegel, Douglas P Finkbeiner et al. (2008). An Improved Photometric Calibration of the Sloan Digital Sky Survey Imaging Data. *The Astrophysical Journal*, 674:1217. Cited on 57, 144

J A Peacock and A F Heavens (1990). Alternatives to the Press-Schechter cosmological mass function. *Monthly Notices of the Royal Astronomical Society (ISSN 0035-8711)*, 243:133–143. Cited on 11

W. J. Percival, S. Cole, D. J. Eisenstein et al. (2007). Measuring the Baryon Acoustic Oscillation scale using the Sloan Digital Sky Survey and 2dF Galaxy Redshift Survey. *Monthly Notices of the Royal Astronomical Society*, 381:1053–1066, `arXiv:0705.3323`. Cited on 85

J. R. Peterson, S. M. Kahn, F. B. S. Paerels et al. (2003). High-Resolution X-Ray Spectroscopic Constraints on Cooling-Flow Models for Clusters of Galaxies. *The Astrophysical Journal*, 590:207–224, `arXiv:astro-ph/0210662`. Cited on 14

V Petrosian (1976). Surface brightness and evolution of galaxies. *The Astrophysical Journal*, 209:L1. Cited on 25

The Planck Collaboration, P A R Ade, N Aghanim et al. (2013a). Planck 2013 results. XX. Cosmology from Sunyaev-Zeldovich cluster counts. *arXiv.org*, 1303:5080. Cited on 15, 178

The Planck Collaboration, P. A. R. Ade, N. Aghanim et al. (2013b). Planck 2013 results. XVI. Cosmological parameters. *ArXiv e-prints*, `arXiv:1303.5076 [astro-ph.CO]`. Cited on 32

Bianca M Poggianti, Anja von der Linden, Gabriella De Lucia et al. (2006). The Evolution of the Star Formation Activity in Galaxies and Its Dependence on Environment. *The Astrophysical Journal*, 642:188. Cited on 45

P Popesso, H Böhringer, J Brinkmann, W Voges, and D G York (2004). RASS-SDSS Galaxy clusters survey. I. The catalog and the correlation of X-ray and optical properties. *Astronomy and Astrophysics*, 423:449. Cited on 44

Marc Postman, Lori M Lubin, James E Gunn et al. (1996). The Palomar Distant Clusters Survey. I. The Cluster Catalog. *Astronomical Journal v.111*, 111:615. Cited on 15, 20, 93

William H Press and Paul Schechter (1974). Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation. *The Astrophysical Journal*, 187:425. Cited on 10, 11

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery (2007). *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA, 3 edition. Cited on 97

Erik D Reese, Tony Mroczkowski, Felipe Menanteau et al. (2011). The Atacama Cosmology Telescope: High-Resolution Sunyaev-Zel'dovich Array Observations of ACT SZE-selected Clusters from the Equatorial Strip. *arXiv*, astro-ph.CO. Cited on 145, 178

C. L. Reichardt, B. Stalder, L. E. Bleem et al. (2012). Galaxy clusters discovered via the Sunyaev-Zel'dovich effect in the first 720 square degrees of the South Pole Telescope survey. *ArXiv e-prints*, `arXiv:1203.5775 [astro-ph.CO]`. Cited on 15, 86

Gordon T Richards, Xiaohui Fan, Heidi Jo Newberg et al. (2002). Spectroscopic Target Selection in the Sloan Digital Sky Survey: The Quasar Sample. *The Astronomical Journal*, 123:2945. Cited on 27

A. G. Riess, A. V. Filippenko, P. Challis et al. (1998). Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant. *Astronomical Journal*, 116:1009–1038, `arXiv:astro-ph/9805201`. Cited on 3, 86

E. Rozo, J. G. Bartlett, A. E. Evrard, and E. S. Rykoff (2012). Closing the Loop: A Self-Consistent Model of Optical, X-ray, and SZ Scaling Relations for Clusters of Galaxies. *ArXiv e-prints*, `arXiv:1204.6305 [astro-ph.CO]`. Cited on 86, 96

Eduardo Rozo, Eli S Rykoff, Benjamin P Koester et al. (2008). An Improved Cluster Richness Estimator. *eprint arXiv:0809.2797*. Cited on 111

Eduardo Rozo, Risa H Wechsler, Benjamin P Koester, August E Evrard, and Timothy A McKay (2007). Optically-Selected Cluster Catalogs as a Precision Cosmology Tool. *arXiv.org*, `astro-ph/0703574v1`. Cited on 10, 92, 180

E. S. Rykoff, B. P. Koester, E. Rozo et al. (2012). Robust Optical Richness Estimation with Reduced Scatter. *The Astrophysical Journal*, 746:178, `arXiv:1104.2089 [astro-ph.CO]`. Cited on 14, 20, 111

E. S. Rykoff, E. Rozo, M. T. Busha et al. (2013). redMaPPer I: Algorithm and SDSS DR8 Catalog. *ArXiv e-prints*, `arXiv:1303.3562 [astro-ph.CO]`. Cited on 20, 30, 145

C Sánchez, M Carrasco Kind, H Lin et al. (2014). Photometric redshift analysis in the Dark Energy Survey Science Verification data. *arXiv.org*. Cited on 121

K Schawinski, S Kaviraj, S Khochfar et al. (2007). The Effect of Environment on the Ultraviolet Color-Magnitude Relation of Early-Type Galaxies. *The Astrophysical Journal Supplement Series*, 173:512. Cited on 45

David J Schlegel, Douglas P Finkbeiner, and Marc Davis (1998). Maps of Dust Infrared Emission for Use in Estimation of Reddening and Cosmic Microwave Background Radiation Foregrounds. *Astrophysical Journal v.500*, 500:525. Cited on 42, 146

N Scoville, H Aussel, M Brusa et al. (2006). The Cosmic Evolution Survey (COSMOS) – Overview. *arXiv.org*, `astro-ph/0612305v1`. Cited on 46

The SDSS-III collaboration, Hiroaki Aihara, Carlos Allende Prieto et al. (2011). The Eighth Data Release of the Sloan Digital Sky Survey: First Data from SDSS-III. *arXiv.org*, `1101.1559v2`. Cited on 144

U. Seljak and M. Zaldarriaga (1996). A Line-of-Sight Integration Approach to Cosmic Microwave Background Anisotropies. *The Astrophysical Journal*, 469:437, `arXiv:astro-ph/9603033`. Cited on 88

Hee Jong Seo and Daniel J Eisenstein (2003). Probing Dark Energy with Baryonic Acoustic Oscillations from Future Large Galaxy Redshift Surveys. *The Astrophysical Journal*, 598(2):720–740. Cited on 50

H. Shapley (1933). Luminosity Distribution and Average Density of Matter in Twenty-five Groups of Galaxies. *Proceedings of the National Academy of Science*, 19:591–596. Cited on 4

Erin S Sheldon, Carlos E Cunha, Rachel Mandelbaum, J Brinkmann, and Benjamin A Weaver (2012). Photometric Redshift Probability Distributions for Galaxies in the SDSS DR8. *The Astrophysical Journal Supplement*, 201(2):32. Cited on 126, 144, 146, 147, 148, 151, 152, 153, 154, 158, 160, 164, 166, 168, 170, 172, 175, 176, 177, 179

A Smargon, R Mandelbaum, N Bahcall, and M Niederste-Ostholt (2011). Detection of intrinsic cluster alignments to 100 Mpc/h in the SDSS. *arXiv.org*, astro-ph.CO. Cited on 85

J Allyn Smith, Douglas L Tucker, Stephen Kent et al. (2002). The u'g'r'i'z' Standard-Star System. *The Astronomical Journal*, 123:2121. Cited on 24, 57

Marcelle Soares-Santos, Reinaldo R de Carvalho, James Annis et al. (2010). The Voronoi Tessellation cluster finder in 2+1 dimensions. *arXiv*, astro-ph.CO. Cited on 17, 31, 87

M Soares-Santos, R R de Carvalho, F La Barbera, P A A Lopes, and J Annis (2008). Cosmography with Galaxy Clusters. *arXiv*, astro-ph. Cited on 30

D. N. Spergel, L. Verde, H. V. Peiris et al. (2003). First-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Determination of Cosmological Parameters. *The Astrophysical Journal, Supplement*, 148:175–194, `arXiv:astro-ph/0302209`. Cited on 2, 9

Volker Springel, Simon D M White, Adrian Jenkins et al. (2005). Simulations of the formation, evolution and clustering of galaxies and quasars. *Nature*, 435(7042):629–636, `arXiv:0504097`. Cited on 13, 87, 88

John P Stott, Ryan C Hickox, Alastair C Edge et al. (2012). The XMM Cluster Survey: The interplay between the brightest cluster galaxy and the intra-cluster medium via AGN feedback. *arXiv.org*, astro-ph.CO. Cited on 8, 10, 14

Chris Stoughton, Robert H Lupton, Mariangela Bernardi et al. (2002). Sloan Digital Sky Survey: Early Data Release. *The Astronomical Journal*, 123:485. Cited on 18, 24, 147, 148, 151

Michael A Strauss, David H Weinberg, Robert H Lupton et al. (2002). Spectroscopic Target Selection in the Sloan Digital Sky Survey: The Main Galaxy Sample. *The Astronomical Journal*, 124:1810. Cited on 26, 27, 28, 148

W. Sutherland (2012). VIKING: the VISTA Kilo-degree INfrared Galaxy survey. In *Science from the Next Generation Imaging and Spectroscopic Surveys*. Cited on 50

Thad Szabo, Elena Pierpaoli, Feng Dong, Antonio Pipino, and James E Gunn (2010). An Optical Catalog of Galaxy Clusters Obtained from an Adaptive Matched Filter Finder Applied to SDSS DR6. *arXiv.org*, astro-ph.CO. Cited on 20, 30

J. L. Tinker, E. S. Sheldon, R. H. Wechsler et al. (2012). Cosmological Constraints from Galaxy Clustering and the Mass-to-number Ratio of Galaxy Clusters. *The Astrophysical Journal*, 745:16, `arXiv:1104.1635 [astro-ph.CO]`. Cited on 87

D. L. Tucker, J. T. Annis, H. Lin et al. (2007). The Photometric Calibration of the Dark Energy Survey. In C. Sterken, editor, *The Future of Photometric, Spectrophotometric and Polarimetric Standardization*, volume 364 of *Astronomical Society of the Pacific Conference Series*, page 187. Cited on 82, 90, 179

C Megan Urry (2010). Stripe 82 X. *XMM-Newton Proposal ID #06730002*, page 124. Cited on 145

Bram Venemans, Joe Findlay, Richard McMahon et al. (2012). Illuminating the Dark Ages: Very high redshift quasars in VIKING+KiDS. *Science from the Next Generation Imaging and Spectroscopic Surveys*, -1:36. Cited on 50

A. Vikhlinin, A. V. Kravtsov, R. A. Burenin et al. (2009). Chandra Cluster Cosmology Project III: Cosmological Parameter Constraints. *The Astrophysical Journal*, 692:1060–1074, `arXiv:0812.2720`. Cited on 9, 14

N Visvanathan and A Sandage (1977). The color-absolute magnitude relation for E and S0 galaxies. I - Calibration and tests for universality using Virgo and eight other nearby clusters. *Astrophysical Journal*, 216:214. Cited on 33

G. M. Voit (2005). Tracing cosmic evolution with clusters of galaxies. *Reviews of Modern Physics*, 77:207–258, `arXiv:astro-ph/0410173`. Cited on 9

Anja von der Linden, Philip N Best, Guinevere Kauffmann, and Simon D M White (2007). How special are brightest group and cluster galaxies? *Monthly Notices of the Royal Astronomical Society*, 379:867. Cited on 45

D Wang, Y X Zhang, C Liu, and Y H Zhao (2007). Kernel Regression For Determining Photometric Redshifts From Sloan Broadband Photometry. *arXiv*, astro-ph. Cited on 50

S J Warren, N J G Cross, S Dye et al. (2007). The UKIRT Infrared Deep Sky Survey Second Data Release. *eprint arXiv*, page 3037. Cited on 145

R H Wechsler (2004). Interpreting SDSS Cluster Masses and Abundances with Mock Catalogs. *Clusters of Galaxies: Probes of Cosmological Structure and Galaxy Evolution*. Cited on 31, 41

Z L Wen and J L Han (2013). Updated catalog of 132,684 galaxy clusters and evolution of brightest cluster galaxies. *arXiv.org*, `1301.0871v1`. Cited on 145

Z L Wen, J L Han, and F S Liu (2009). Galaxy Clusters Identified from the SDSS DR6 and their Properties. *The Astrophysical Journal Supplement Series*, 183(2):197–213. Cited on 30

S D M White, D I Clowe, L Simard et al. (2005). EDisCS – the ESO distant cluster survey. *Astronomy and Astrophysics*, 444(2):365–379. Cited on 45

Simon D M White, Julio F Navarro, August E Evrard, and Carlos S Frenk (1993). The baryon content of galaxy clusters: a challenge to cosmological orthodoxy. *Nature*, 366(6454):429–433. Cited on 9

H.-Y. Wu, E. Rozo, and R. H. Wechsler (2010). Annealing a Follow-up Program: Improvement of the Dark Energy Figure of Merit for Optical Galaxy Cluster Surveys. *The Astrophysical Journal*, 713:1207–1218, `arXiv:0907.2690 [astro-ph.CO]`. Cited on 87, 127

Xiaohu Yang, H J Mo, Frank C van den Bosch et al. (2007). Galaxy Groups in the SDSS DR4. I. The Catalog and Basic Properties. *The Astrophysical Journal*, 671:153. Cited on 45

Donald G York, J Adelman, John E Anderson et al. (2000). The Sloan Digital Sky Survey: Technical Summary. *The Astronomical Journal*, 120:1579. Cited on 22, 30, 145

F. Zwicky (1933). Die Rotverschiebung von extragalaktischen Nebeln. *Helvetica Physica Acta*, 6:110–127. Cited on 112

Fritz Zwicky, E Herzog, and P Wild (1961). Catalogue of galaxies and of clusters of galaxies, Vol. I. *Pasadena: California Institute of Technology (CIT)*. Cited on 4, 13