GALAXY ANGULAR CLUSTERING EVOLUTION IN THE SDSS CO-ADD IMAGING DATA

by

Jeremy Brewer

B.Sc. in Physics, Rhodes College, 2000

M.Sc. in Physics, University of Pittsburgh, 2002

Submitted to the Graduate Faculty of

the Department of Physics & Astronomy in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2008

UNIVERSITY OF PITTSBURGH

DEPARTMENT OF PHYSICS & ASTRONOMY

This dissertation was presented

by

Jeremy Brewer

It was defended on

August 14th 2008

and approved by

Andrew Connolly, University of Pittsburgh & University of Washington

Arthur Kosowsky, University of Pittsburgh

David Turnshek, University of Pittsburgh

Vladimir Savinov, University of Pittsburgh

Jeff Schneider, Carnegie Mellon University

Dissertation Director: Andrew Connolly, University of Pittsburgh & University of Washington

GALAXY ANGULAR CLUSTERING EVOLUTION IN THE SDSS CO-ADD IMAGING DATA

Jeremy Brewer, PhD

University of Pittsburgh, 2008

We study the evolution of the angular clustering of galaxies as a function of redshift, luminosity, and type. We utilize redshift estimates computed from broadband photometry, so we require precise flux measurements. For this reason, we chose the SDSS co-added imaging data set from stripe 82 and obtained 1% error photometry with a custom image processing pipeline. We measured the angular clustering of galaxies $w(\theta)$ and inverted it to obtain the real space correlation function $\xi(r)$, which we fit as a power law with parameters r_0 and γ. Finally, we use our measured $\xi(r)$ fits to constrain galaxy formation models and find that luminous galaxies are found in higher mass dark matter halos, in agreement with theory and previous results in the field.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

If I could do one near perfect thing, I'd be happy.

– Stuart Murdoch

Kid, I've flown from one side of this galaxy to the other, and I've seen a lot of strange stuff, but I've never seen anything to make me believe that there's one all-powerful Force controlling everything. 'Cause no mystical energy field controls my destiny. It's all a lot of simple tricks and nonsense.

– Han Solo on Dark Energy

In the beginning...
Oh, long before that,
When Light was deciding who should be in and who should be out of Spectrum,
Yellow was in trouble, even then.
Seems that Green – you know how Green can be – didn't want Yellow in.
Some silly primal envy I suppose, but for whatever cause, the effect was bad on Yellow
And caused Yellow to weep yellow tears for several eternals (before there were years)
Until Blue
Heard
What was up
Between Green and Yellow
And took Green aside for a serious talk in which Blue pointed out
That if Yellow and Blue were to get together –
Not that they would, but *if* they did, a gentle threat –
They could make their own Green.
"Oh" said Green with some understanding.
Naturally, by a sudden change of hue, Green saw the light and Yellow got in.
Worked out fine –
Yellow got lemons,
And Green
Got limes.

– "Yellow" by Ken Nordine, from *Colors*

I would first like to thank my adviser, Andrew Connolly, for agreeing to take me on as a student at a particularly difficult time. His mentoring ~~allowed~~ enabled me to develop the programming skills that proved so useful in my job search. Not many advisers in physics would have given me the independence necessary for branching out into computer science, and for that I am grateful.

I would like to thank my wife, Bambi, and son, Oliver, for their love and support, especially through the final, hectic months of this thesis.

Sam Schmidt contributed greatly to this work through both discussion and endless tweaking of the various photometric redshift code inputs. His input has significantly improved the quality of this thesis.

I thank Ryan Scranton for developing the angular correlation function and masking code used in this thesis, both of which would have been difficult to write from scratch. He also provided a large amount of discussion and guidance, particularly towards the end of this thesis.

In the early days of my research, Simon Krughoff was always around to provide helpful suggestions and tests for my calibration pipeline. I am grateful for his useful feedback and advice.

Finally, I want to thank Andrew Zentner (and his wife) for somehow managing to provide me with halo model code shortly after his wife gave birth. His halo model review pre-print also proved very helpful in deciphering the many components of the model.

1.0 INTRODUCTION

In the late 1920s, Edwin Hubble demonstrated that there are galaxies outside of our Milky Way receding at velocities proportional to their distance from us. This result, implying an expanding universe composed of many galaxies, planted the seeds for our current understanding of how the universe originated, commonly referred to as the "Big Bang" theory. The Big Bang theory rests on three key observations: the aforementioned Hubble diagram demonstrating expansion, light element abundances consistent with Big Bang nucleosynthesis theory, and the primordial blackbody radiation known as the cosmic microwave background (CMB). From these observations, we theorize that the universe was once much denser and hotter and that the structure we see today (*i.e.* galaxies) arose from a much smoother, more homogeneous universe. Understanding how large scale structure grew from this environment is one of the primary goals of cosmology today.

1.1 THEORETICAL DEVELOPMENTS

Recently, observations have motivated the need for new physics to drive the expansion of the Big Bang: dark matter, dark energy, and inflation. Dark matter was originally proposed in 1933 by Fritz Zwicky to explain discrepancies in the rotation curves of spiral galaxies, but it is now known that dark matter is also needed to explain the large scale structure in the universe today. Additionally, we believe that dark matter is non-baryonic (not made of protons and neutrons) and interacts only gravitationally. Though its exact nature is unknown, physicists are currently searching for theoretical particles (*e.g.* gravitinos, axions) postulated to be this missing dark matter.

Even more mysterious is dark energy, the unknown yet dominant form of energy proposed to explain the observed expanding and accelerating universe. Though no theory currently exists to explain dark energy, it has several interesting properties. First, its energy density remains relatively constant with the expansion of the universe. Second, it has negative pressure, making it an exotic substance with no known origin. Measuring the properties of dark energy using weak gravitational lensing is one of the primary science goals of multiple future surveys including the Dark Energy Survey (DES), the Sloan Digital Sky Survey (SDSS) III, and the Large Scale Synoptic Survey (LSST).

Inflation is the current theory for explaining why large scale structure is so isotropic on scales larger than the comoving horizon (*i.e.* outside of the distance light could have travelled). Because photons in the CMB have nearly the same temperature (to one part in 10^5) even at very large scales, they must have been in equilibrium and thus in causal contact. Inflation solves the horizon problem by proposing that the universe expanded exponentially fast when it was 10^{-35} seconds old; unfortunately this solution requires matter or energy with negative pressure, similar to dark energy. It is possible that dark energy arose from inflation, though it is also possible that the form of dark energy driving the expansion of the universe today is distinct from that which drove inflation.

Another important development in cosmology over the last decade is the emergence of galaxy formation models which reproduce the observed statistical clustering of galaxies. Galaxy evolution can be decomposed into three contributions: luminosity evolution due to changes in the galaxy's internal stellar population, number evolution due to galaxy mergers, and spatial evolution due to large scale structure evolution. All of these components are poorly understood, and worse, they are intertwined – merging galaxies, for instance, experience bursts of star formation and hence become brighter after the merger. The current models for explaining galaxy formation, collectively known as the *halo model*, seek to sidestep these complications by first modeling the dark matter and later sprinkling galaxies throughout the dark matter halos, spherical structures in which all of the dark matter is postulated to reside. This approach seeks only to reproduce the statistical properties of galaxies rather than evolution of individual galaxies. The halo model uses this approach because dark matter only interacts gravitationally, making it is easy to simulate its clustering. Because

mass is easily modeled, there is a desire to associate the properties of galaxies only with the mass of their host halo so that galaxies are independent of their environment. Though still primitive, these models offer a promising approach to studying galaxy formation and evolution.

1.2 OBSERVATIONAL DEVELOPMENTS

Observationally, cosmology has amassed an enormous amount of high quality data within the last decade, particularly from the Sloan Digital Sky Survey (SDSS). Consider, for example, that in 1985 the state-of-the-art in galaxy surveys measured the positions and redshifts of 1100 galaxies. Today, SDSS has measured spectra of 1 million galaxies and broadband photometry of 217 million objects. This wealth of data enables more sophisticated studies of galaxy properties. First, important statistical properties such as galaxy number counts, clustering, and luminosity distributions (termed the *luminosity function*) are better constrained due to reduced Poisson noise. Second, galaxies can be separated into subpopulations by type (*e.g.* spiral or elliptical), luminosity (which is known to correlate with type), and redshift, enabling the study of how galaxy properties vary with these properties. In other words, we are now able to ask questions such as "Do all types of galaxies cluster in the same way?" and "How do galaxy properties evolve with time?". Future surveys such as LSST will amass even more data and probe even longer time scales for galaxy evolution, further constraining galaxy formation models.

Together, the increase in observational data and improvements in galaxy formation models present an excellent opportunity to study the evolution of galaxy clustering. The simplest statistic for measuring galaxy angular clustering is the two point angular correlation function $w(\theta)$ which measures how much more or less likely than random a pair of galaxies will be found at a given separation θ on the sky. In practice, $w(\theta)$ is measured by counting pairs of galaxies between a data set and a randomly generated data set; see equation 5.3 for details. The first wide field galaxy survey designed to study galaxy clustering was the Lick Observatory Sky Atlas. Shane and Wirtanen (1967) counted the distribution of galaxies brighter

than apparent magnitude 19 (by hand!), and Totsuji and Kihara (1969) first measured the clustering length r_0 using their results. In the 1970s, Groth and Peebles (1977) re-calculated the 2 and 3 point correlation function with corrections for plate-to-plate limiting magnitude and counting errors. Surveys since this time (Maddox et al., 1990; Collins et al., 1992; Connolly et al., 2002, *e.g.*) have refined the clustering analysis by probing larger areas of the sky and greater depths, with the current state of the art for local galaxy clustering measurements coming from SDSS data. Connolly et al. (2002) measured $w(\theta)$ using only positional information from the SDSS Early Data Release (EDR), and Budavári et al. (2003) used SDSS data with photometric redshifts (discussed below) to investigate how clustering varies with luminosity and type. Zehavi et al. (2005) measured the real space correlation function $\xi(r)$ from the SDSS spectroscopic data and used it to constrain the halo model of galaxy formation. For non-local galaxies, other groups have measured the clustering of high redshift galaxies (Coil et al., 2008) and very high redshift quasars (Shen et al., 2007) using DEEP2 and SDSS data respectively. Additionally, Zheng et al. (2007) have obtained preliminary results for relating the halo properties of galaxies in DEEP2 to those in SDSS, illustrating that work in this area is currently ongoing.

Galaxy clustering measurements have consistently found that the angular correlation function $w(\theta)$ is well described on small angular scales by a power law: $w(\theta) = A\theta^{1-\gamma}$. As shown in Figure 5.9, the amplitude A decreases with apparent magnitude with γ remaining approximately constant. Thus, fainter galaxies are less strongly clustered than bright galaxies. This result is consistent with Limber's well known scaling relation given in Equation 5.33; in essence, the number of spurious galaxies along the line of sight increases with survey depth and smears out the clustering strength. This is illustrated graphically in Figure 1.1 where one can easily see that the increase in galaxy numbers with depth reduces the clustering signal.

In order to relate galaxy clustering measured from angular positions on the sky to the true 3-D structure of galaxies, one must invert the angular correlation function using the distances to each galaxy. The most straightforward way of determining distance to a galaxy is by running the light of a galaxy through a spectrograph so that the various wavelengths of light are separated. Features of known rest wavelength can then be identified; the shift

in these features determines the redshift and hence distance to the galaxy (see Equations 4.1 and 5.9). Spectroscopy is very time consuming because the galaxy's light is spread over a larger area of the detector, requiring longer integration times, so a significant additional time investment is needed if one wishes to compute the true spatial distribution. This problem can be overcome by using a faster redshift estimation technique which utilizes only broadband photometry that can be obtained roughly 100x more quickly than spectra. The techniques for estimating redshift in this way are termed *photometric redshifts* or photozs for short. Because photometric redshifts use less information to estimate the redshift, they are inherently less accurate, but this can theoretically be overcome (in a statistical sense) with the larger number of galaxies. Because photometric redshifts rely on photometric measurements (*i.e.* magnitudes), it is essential to have well calibrated magnitude measurements.

The primary science goal of this thesis is to study the evolution of galaxy clustering with luminosity, type, and redshift using photometric redshifts. Because photometric redshifts require high quality photometry, we chose to measure galaxy clustering in the co-added imaging data set stripe 82 from SDSS. This data set consists of stacks of SDSS images co-added together from multiple passes over the same area of sky near the equator. The multiple images should improve the photometry of the co-added images over those of the SDSS main sample and probe higher redshifts than Budavári et al. (2003); we also hope to constrain the halo model by inverting $w(\theta)$ to obtain the real space correlation function $\xi(r)$. As the main SDSS pipeline cannot currently be used to process the co-added sample, much of this thesis is devoted to describing a custom pipeline we developed.

This thesis is broken into 6 additional chapters. In chapter 2, we cover the details of our image processing. In chapter 3, we present the method used to classify stars and galaxies. In chapter 4, we detail the process used to obtain photometric redshifts for the galaxies in our sample. In chapter 5, we describe the computation of the angular correlation function, and finally in chapter 6 we use our results to constrain parameters in the halo model.

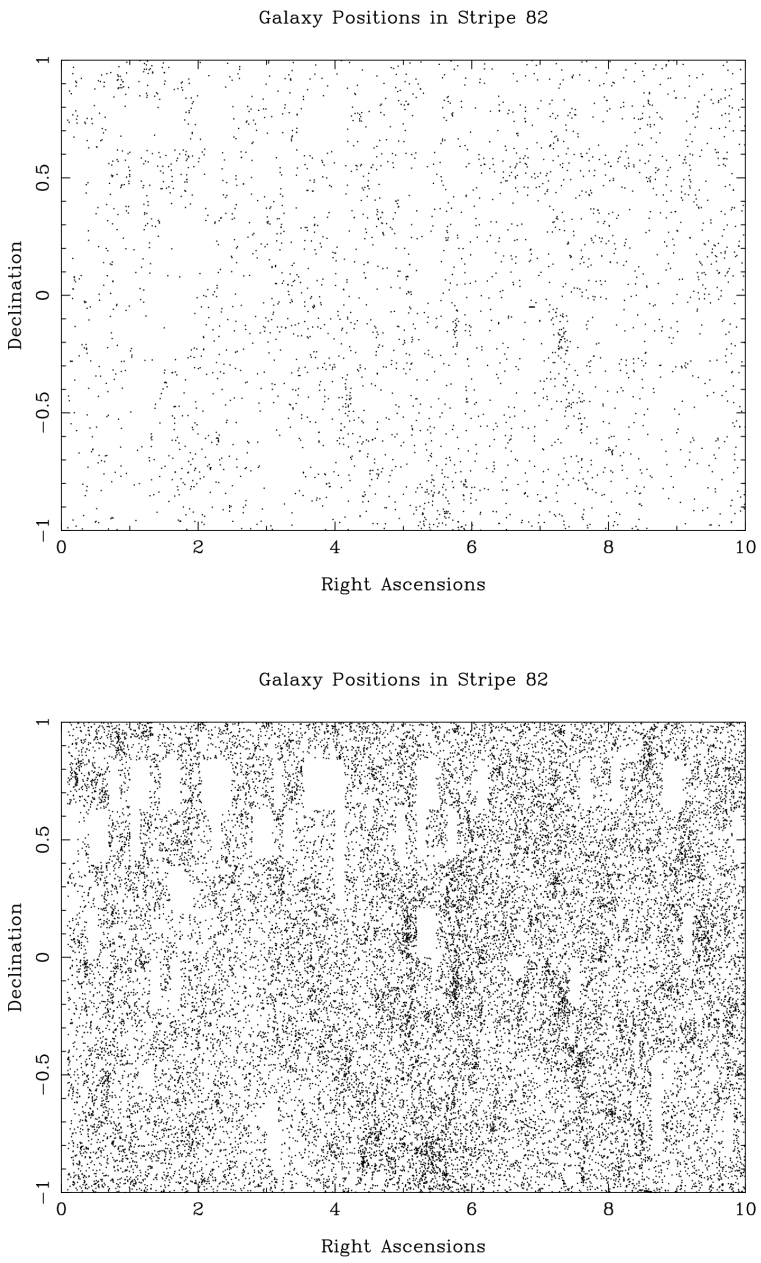


Figure 1.1 Visual representation of clustering evolution with apparent magnitude. The top figure shows galaxies with magnitudes $r \leq 18$, and the bottom shows galaxies with $r \leq 20$. As the number of spurious galaxy encounters along the line of sight increases, the clustering strength is "smeared out". Areas without galaxies are present due to masking.

2.0 IMAGE CALIBRATION

In order to analyze astronomical objects in imaging data, the images must first be processed to detect sources and extract their properties (positions, fluxes, etc.). The software that detects and measures these properties is termed an *imaging pipeline*. Developing an image pipeline that produces accurate calibration in astronomy is non-trivial as it involves precise measurements of both position and flux. Consider, for example, that the SDSS pixel scale is 0.396 arcsec/pixel (Gunn et al., 1998), so 1 pixel can capture a dime held at a distance of 2.9 miles. The photometric properties need to be similarly precise – an object with an apparent magnitude of 22 has a flux roughly equivalent to that of a 100W light bulb placed on the surface of the moon.

The SDSS survey team has developed a robust automated pipeline for obtaining precise optical imaging data (Lupton et al., 2001). For this thesis, data from SDSS DR5, the 5^{th} data release which includes data taken through June 2005, were used to assist in calibration. DR5 contains five bands of photometric data for 217 million objects spread over 8000 deg^2 and roughly 1 million spectra over 5700 deg^2. For the photometric data, there are 5 pass bands: u (3551 Å), g (4686 Å), r (6165 Å), i (7481 Å), and z (8931 Å). The photometric filters are discussed in detail in Fukugita et al. (1996). The r band is 95% complete to an AB magnitude of 22.2 with 2% RMS error and a median point spread function width of 1.4 arcsec (Adelman-McCarthy et al., 2007). For an overview of the SDSS DR5 catalog, see Adelman-McCarthy et al. (2007), for details on photmetric calibration pipeline see Lupton et al. (2001), and for astrometric calibration see Pier et al. (2003).

In addition to the primary survey, the SDSS imaging camera has obtained repeat observations of a stripe along the Celestial Equator spanning $22^h20^m < \alpha < 3^h20^m$, $-1.25^\circ < \delta < 1.25^\circ$ in J2000 coordinates, also known as stripe 82. The southern equatorial stripe

is observed in the fall months when the southern Galactic cap is visible in the northern hemisphere. This area of the sky was repeatedly imaged to enable studies of variable objects such as supernovae and to enable co-added imaging for probing fainter magnitudes. Though SDSS covers a wide area, it is a shallow survey only probing to a median redshift of roughly 0.1. With the co-added imaging data set, we can probe to fainter redshifts of roughly 0.4 at $r = 22$ in a volume limited survey and study evolution of galaxy clustering over a much longer period of time. It is this feature of the co-added imagery that this thesis aims to utilize – co-addition of imaging data to probe fainter sources and therefore the evolution of galaxies.

The SDSS imaging camera is mounted on a drift scan telescope, meaning that the telescope remains stationary while the sky passes overhead. As such, the SDSS imaging pipeline was engineered to deal with a constant stream of data rather than a set of random pointings. Work is currently underway to modify the imaging pipeline (Adelman-McCarthy et al., 2007) to process the co-added imagery, but results are still currently unavailable as of this writing. Instead, we chose to develop our own calibration pipeline which leveraged the SDSS main survey pipeline.

To develop our custom pipeline, we took the novel approach of applying web services to astronomical image calibration. This approach enabled us to utilize the entire DR5 catalog without storing any data on local disk and still achieve processing times of 2-5 sec/field with photometric RMS r band magnitude errors of 0.069 (see Figure 2.3). The details of this custom pipeline and all relevant calibration checks comprise the remainder of this chapter.

2.1 OVERVIEW OF WEB SERVICES

The term *web services* encompasses several protocols for making remote procedure calls (RPC), *i.e.* calling a function on another machine. For this thesis, two kinds of web services were utilized: XML-RPC and SOAP (formerly Simple Object Access Protocol). Both of these involve sending synchronous request messages in an XML format over plain text via the standard web protocol HTTP. The difference between these two services is analogous

to the difference between a dynamic language (*e.g.* Python) and statically typed language (*e.g.* C++) – XML-RPC is much simpler, easier to write, and more flexible, but SOAP potentially offers type safety on the client side which could detect errors at compile time rather than run time. In the simpler XML-RPC, the client can send any message it wants (including malformed ones) to the server, which decides whether the message is valid and either processes the request or returns an error. In SOAP, the server provides an XML description of its services named the Web Services Description Language (WSDL) which client authors compile into code in their preferred language called stubs. The stubs handle the conversion of types and functions/methods in the native language into XML messages that the server understands, so theoretically it should be impossible to send a malformed message to the server. Additionally, this approach enables code editors to implement automatic code completion for the methods the server supports since they have been serialized into native code.

In practice, though, SOAP's advantages over XML-RPC disappear due to degeneracies in how the WSDL can be specified which result in message transmission difficulties. In particular, SOAP clients that are written in a different language or toolkit than the server have trouble formatting messages in the precise form expected by the server. Often, these differences are quite trivial (*e.g.* the addition of a namespace for a few tags), but they result in technically malformed messages from the standpoint of the server. Even worse, upgrading the SOAP library used to generate the stubs can subtly change how the output message is formatted and completely break an application. This is a huge problem because SOAP libraries are still quite young and under active development. To work around these issues, we resorted to constructing SOAP messages using simple string formatting rather than by generating stubs. This approach turned out to be both easier to develop and faster performance-wise: for a typical query to Open SkyQuery, the "by hand" method was at least 3x faster.

One crucial element to achieving good performance using web services and large amounts of data is an efficient method for sending binary data; in particular, the server should ideally not pass large chunks of binary data through an XML parser. For the XML-RPC server we developed, we used a simple custom HTTP POST path */data* that indicates to the

server that the accompanying message is binary data; this approach utilizes the fact that XML-RPC servers are just web servers that parse XML messages. SOAP offers a similar potential solution called *attachments*, where a placeholder message is inserted into the actual XML which points to the accompanying binary part. Unfortunately, at the time our code was under development, attachments were poorly supported, and often the XML parsers would examine the attachments instead of ignoring them as they should. Another potential SOAP solution is to send a URL to the server which then downloads the data it needs. The drawback of this approach is that the client must be able to run some sort of file server.

With an efficient way to send large amounts of binary data, it is possible to achieve excellent performance using web services: for the photometric calibration pipeline, the typical processing time per 12 MB image was 4-5 seconds ($\approx$ 10 GB/hour). Furthermore, web services can be used to implement a simple approach to parallelism for tasks that do not require high volume message passing – simply write a "master server" that farms out incoming messages to a network of normal servers. This approach can be further improved with the use of asynchronous messaging.

The remainder of this section is devoted to outlining the web services used for this thesis.

2.1.1 Open SkyQuery

Open SkyQuery (Budavári et al., 2004) is a distributed database system that provides access to multiple astronomical surveys using a SQL-like syntax called ADQL. In addition to the usual SQL functions, ADQL provides a region operation that returns all sources within a given circle on the sky and a cross match operation that returns matches between 2 data sets. The cross match operation is made more useful by the fact that users can upload their own data to temporary tables to compare against other surveys. Each individual survey catalog is stored on a SkyNode, a server running a SQL database and implementing the ADQL query language as a web service. Additionally, the data stored on SkyNodes are indexed using a hierarchical triangular mesh (HTM), a hashing algorithm that significantly improves spatial searching performance for spherically distributed data.

All of the reference data used in calibration were retrieved using queries to Open Sky-

Query. The region command (as opposed to simple α, δ limits) is the fastest way to perform spatial data queries because it makes use of the HTM indexing.

For cross matching, we found Open Sky Query's algorithm to be lacking. Most significantly, it does not return a unique cross match between the two lists of points; that is, a point in the 1^{st} list may be matched to multiple points in the 2^{nd} list and vice versa. Additionally, the cross matching maximum radius is specified as a χ^2 threshold instead of a physically meaningful distance. For these reasons, we performed cross matches by first employing a region query to Open SkyQuery then locally running an $\mathcal{O}(n \log n)$ algorithm[1] that ensures the match is unique with respect to both lists. This approach has the added benefit that a temporary table does not need to be uploaded to Open SkyQuery, resulting in a performance gain.

In addition, the queries from each co-add image overlap spatially (both because the images themselves overlap and because the queries are spherical). This enables the remote SDSS database to utilize caching to further improve query time. As an example, an initial run of a representative query to select the first 10 objects within a radius of 5 arcmin took $\approx$3 seconds, but subsequent runs require only 0.5 seconds.

2.1.2 WESIX

Web Enabled Source Identification with X-Matching (WESIX) (Krughoff and Connolly, 2008) is a web service front end we developed to SExtractor (Bertin and Arnouts, 1996), a source extraction program widely used in astrophysics. There are a large number of configuration options controlling how SExtractor detects and measures sources, and WESIX supports nearly all of them. The programmable nature of WESIX enabled us to write source extraction programs that make multiple passes on the input image and adjust parameters between each pass easily. In fact, one way of viewing WESIX is merely as a scripting framework for generating SExtractor configuration files that is general enough to enable running SExtractor remotely. All source extraction on images was performed using WESIX.

[1] It is possible to cross match in $\mathcal{O}(n)$ time using hash tables.

2.2 PHOTOMETRIC PIPELINE

In this section, we outline the specifics of the photometric pipeline used to extract sources and their properties from the co-add imaging data.

2.2.1 Magnitude Calibration

Magnitude calibration is broken into three steps for each input image: 1.) extract sources from the input image and measure their positions on the sky using WESIX, 2.) retrieve reference sources (in this case from SDSS DR5) in the same area of the sky using Open SkyQuery, and 3.) compare the image and reference sources to determine the proper photometric calibration. Source extraction is performed in two passes to optimize signal-to-noise. In the first pass, only extremely bright ($\geq 20\sigma$ above background) image sources are used. These sources are cross matched against the reference sources to determine star-galaxy classification, then the average full-width-half-max (FWHM) of stars is computed to estimate the point spread function (PSF). On the second source extraction pass, a Gaussian convolution filter with approximately the same FWHM is applied to improve signal-to-noise; to ensure a close match in convolution FWHM, we generated Gaussian convolution filters with FWHM ranging from 2 to 4 pixels in steps of 0.1 pixels. In addition, the SExtractor detection threshold is set to 5σ per total number of effective pixels of the convolution filter to optimize signal-to-noise for point sources. The image sources obtained on the second pass are then cross matched against bright reference sources ($16 < r < 19$), after which an iterative sigma clipping fit is applied to determine the magnitude calibration. The fit used is a least squares fit with perpendicular offsets because there is no true independent variable, and it is performed only over stars in the valid magnitude range. The cuts in magnitude space are made roughly perpendicular to the fit line to avoid Malmquist bias. Finally, the calibration fit (including the slope) is applied to the image magnitudes and the final catalog is output for the image. An example calibration is shown in Figure 2.1.

It is important to note that each image is calibrated independently, including images of the same area of sky in different pass bands. We performed several checks (discussed later

in this chapter) to ensure that the magnitude calibrations did not vary significantly with position on the sky.

This magnitude calibration procedure has a failure rate of ∼0.1% for the co-added data set, where success is defined as having a fit with slope between 0.9 and 1.1. With the inclusion of "empty images" which include no data (whether because this area of the sky was unobserved or some co-add pipeline bug is unknown to me), the failure rate rises to ∼1.3%. Unfortunately, the failure rate of the matched aperture catalog (discussed later in this chapter) is ∼4% due to the additional requirement that all 5 bands for a given field must be calibrated successfully as a group.

2.2.2 Catalog Collation

The above calibration procedure is applied to every image in the co-added stripe 82. To construct a catalog of co-add sources, it is necessary to collect the catalogs output for each image into a single catalog. There are 2 difficulties in this procedure. First, the images overlap both in right ascension and declination, so care should be taken not to double count sources. The second difficulty is that a decision must be made as to what a "source" is – is it every object in every band measured, or is there a particular pass band that an object must present in to be considered a source?

To solve the overlap problem, we determined the non-overlapping boundaries between adjacent images in both right ascension and declination by taking $\frac{1}{2}$ the distance between the image centers as the boundary. The cut in declination is simpler because camcols follow lines of constant declination for stripe 82, so the limits can easily be pre-computed. For fields within a given camcol, we used the halfway point between image centers of the two closest successfully processed images as the non-overlapping boundary. This was done to allow a few more sources in the areas normally excluded in the overlap region. A similar approach could have been used for declination as well, but we preferred a simpler declination cut so that camcols could be processed independently.

With respect to how a source is defined, we take the r band as our primary detection band, meaning that sources are objects only if they are detected in r. If an object is not

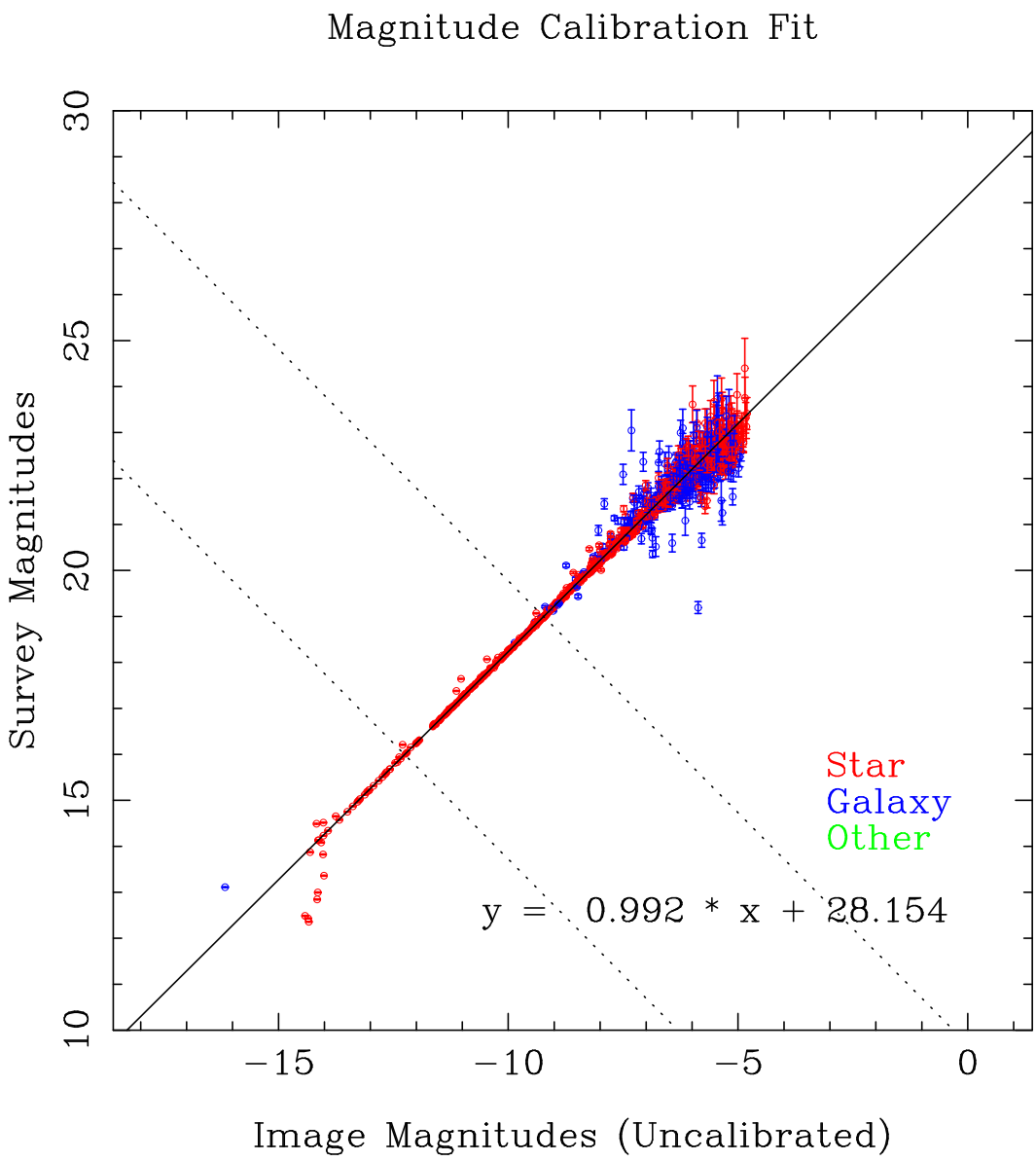


Figure 2.1 Magnitude Calibration Example. Points are color coded by object type with filled circles for points used for the fit and hollow circles for rejected points.

detected in another pass band, the measured properties for that band are set to a bad flag value (-999). Our cross matching code is used to identify object matches between pass bands. An additional result of this decision is that the right ascension and declination are taken from the values in the r band, so it is the r band coordinates that are used to determine the boundaries in the paragraph above.

With these two problems solved, it is a straightforward matter to collate all of the individual catalogs. This process consists of reading the log files to determine which r band images were successfully processed, computing the non-overlapping boundaries, reading each set of *ugriz* catalogs for the successfully processed r band images, matching objects between bands and filling in missing objects with bad flags, and outputting all of the measured properties to a new catalog file. This process was done for each camcol independently, then the results were uploaded to a SQL database to enable easier object selection.

As one final note, we created a unique ID for each object by joining the values for run, rerun (always 1), camcol, and field together. This yields a unique 64-bit integer key called the objID. Having a unique objID makes updating the SQL database much easier as no spatial querying is needed; for this reason, the objID was used as the primary key in the SQL database.

2.3 CALIBRATION TESTS

In this section we present a set of tests performed to verify the quality of the calibration was consistent across the entire stripe.

The most obvious test to do is to compare the magnitudes of the final calibrated co-add catalog against the SDSS DR5 sample used to calibrate it. Figure 2.2 demonstrates this comparison for stars for all bands over the entire stripe, compromising 1.8 million objects. Figure 2.3 shows Δmag for each of the bands over the same region. The fits were performed using an iterative σ clipping algorithm with both slope and intercept as free parameters. For the r band, the RMS scatter was $\sigma_r = 0.069$ or $r_{\rm err} \approx 0.4\%$. The highest error measured was in the u band with $\sigma_u = 0.178$ and $u_{\rm err} \approx 1\%$. These results demonstrate that we met our

target goal of 1% photometry error which lies within the scatter of the single epoch SDSS photometry.

Figure 2.4 shows a comparison of the co-add colors to SDSS DR5 colors over the entire magnitude range. While the 1 to 1 trend is visible in the plots, there are additional lines of degeneracy indicating some bias for our measured colors. However, it is important to note that there are 700,000 to 1.8 million points on these plots, so the density of outliers is over-emphasized visually. The percentage of points plotted lying within the 3σ line is 93.25% for $u-g$, 92.22% for $g-r$, 93.11% for $r-i$, and 91.47% for $i-z$, so the total number of outliers is less than 10% for every plot.

Additionally, because the calibration is performed on an image by image basis, it is useful to test whether the calibration varies appreciably within a single camcol and between neighboring camcols. Figure 2.5 shows how the calibration varies within a single camcol, and Figure 2.6 shows how it varies between camcols. For both of these plots, the difference in magnitudes shown is only for sources with $16 < r < 19$, *i.e.* those for whom the calibration should be best. The RMS scatter in ra for the r band is $\sigma_r = 0.011$, $r_{\rm err} = 0.06\%$, and in the z band $\sigma_z = 0.031$, $z_{\rm err} = 0.18\%$; the RMS scatter in dec for r and z is identical. Though the average error is insignificant, the u band shows a zero point variation in dec of roughly 0.015. While this trend in the u zero point is unsettling, the magnitude is nonetheless small enough to disregard.

Finally, Figure 2.7 shows the number counts of objects in each band. The r band is complete to 22.98, an improvement over the single epoch r limit of 22.2.

2.4 MATCHED APERTURE CATALOG

In addition to the catalog described above, we developed a version of the calibration pipeline that measured the magnitudes using matched apertures between all of the photometric band passes. In other words, objects were first detected in r, then the same apertures were placed in each of the other band passes. This is done so that extended objects have more consistent flux measurements, which should result in more accurate color measurements (flux ratios).

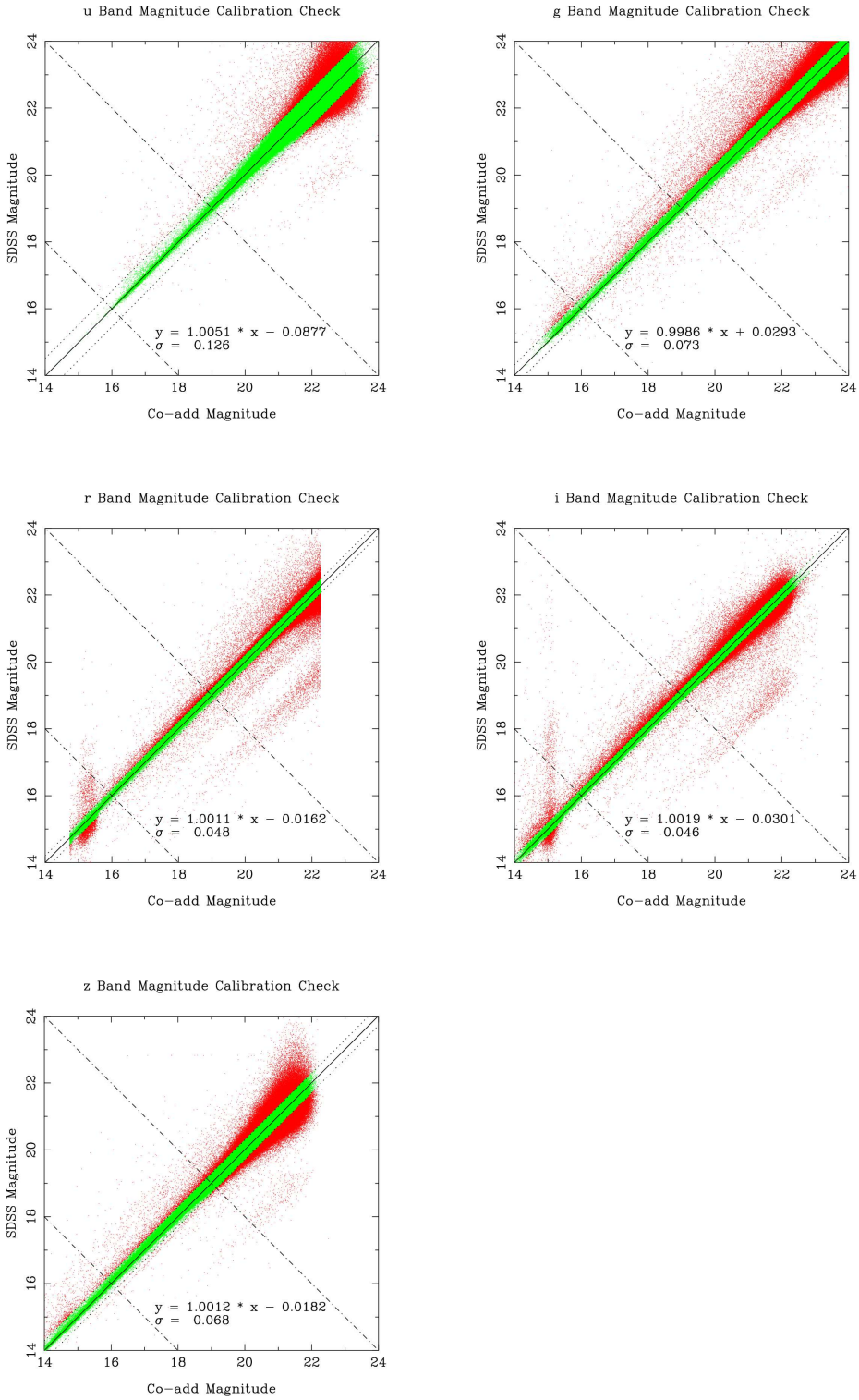


Figure 2.2 Comparison of co-added imaging MAG_AUTO and SDSS DR5 model magnitudes for all of stripe 82. All objects plotted are stars found in both the co-added catalog and SDSS. The fit was performed over all magnitudes using iterative 3σ clipping. Dotted lines show the 3σ cut and dash-dot lines show the range of magnitude space over which the calibration was initially determined.

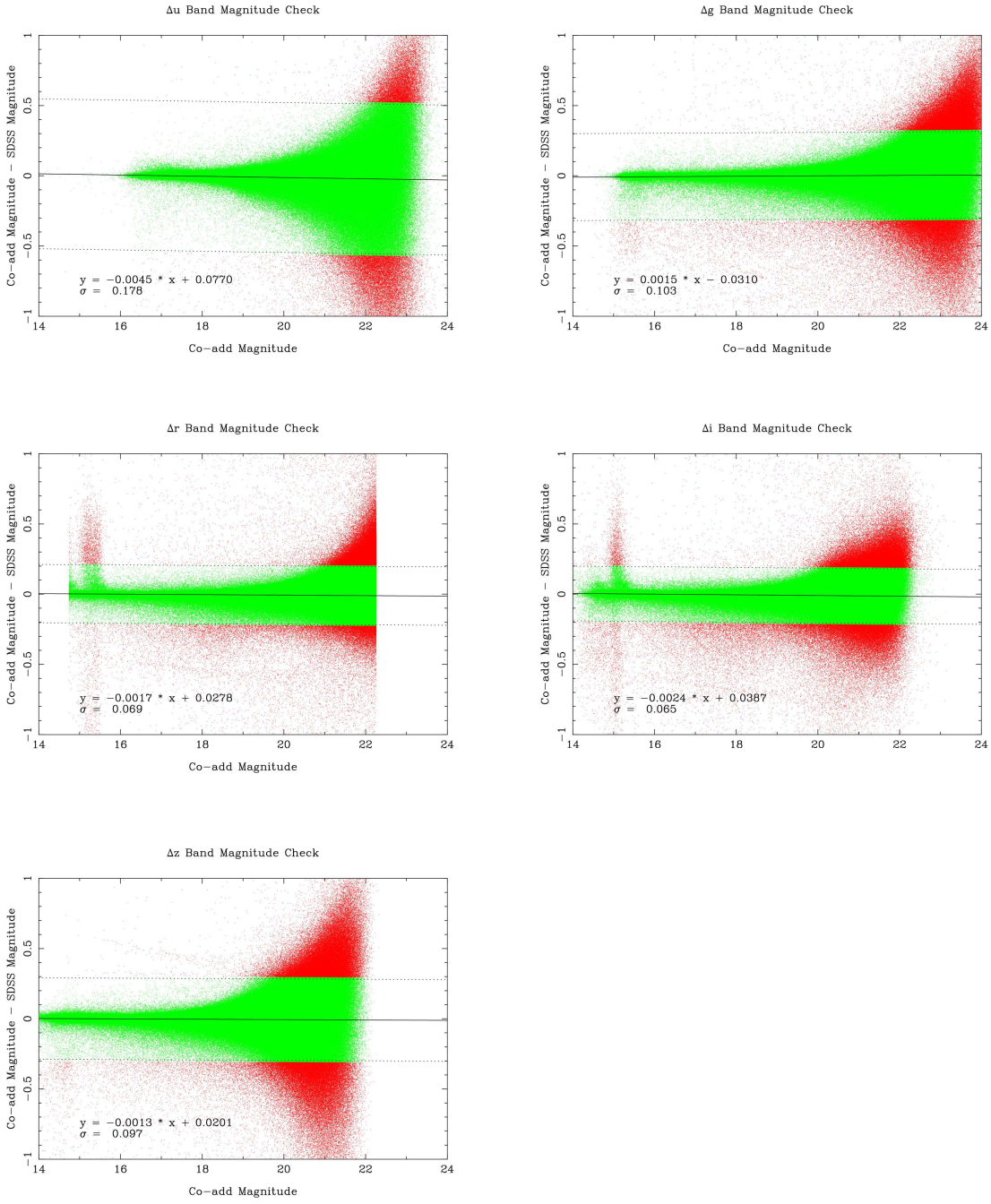


Figure 2.3 Comparison of co-added imaging MAG_AUTO and SDSS DR5 model magnitudes for all of stripe 82. All objects plotted are stars found in both the co-added catalog and SDSS. The fit was performed over all magnitudes using iterative 3σ clipping. Dotted lines show the 3σ cut.

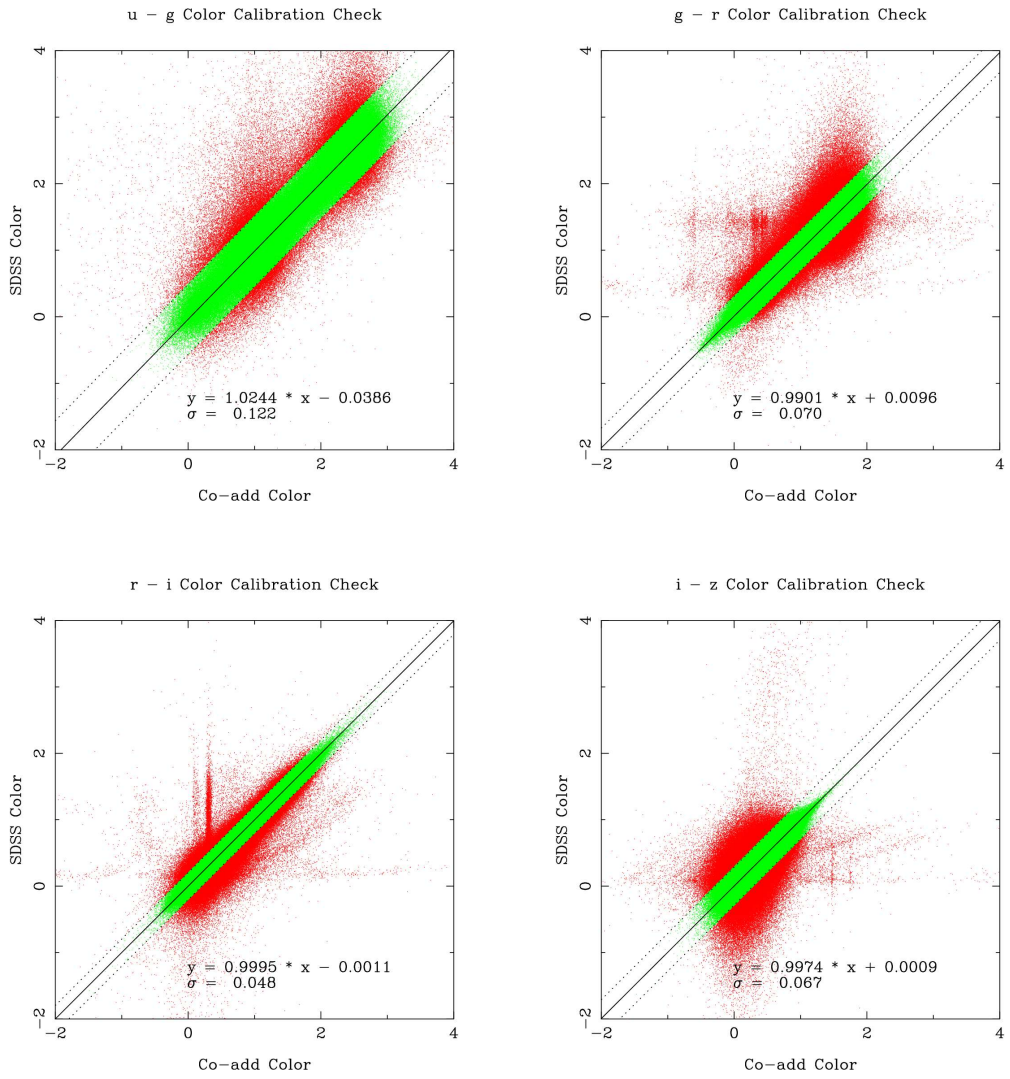


Figure 2.4 Comparison of co-added MAG_AUTO colors and SDSS DR5 model colors for all of stripe 82. All objects plotted are stars found in both the co-added catalog and SDSS. The fit was performed over the entire sample using iterative 3σ clipping. Dotted lines show the 3σ cut.

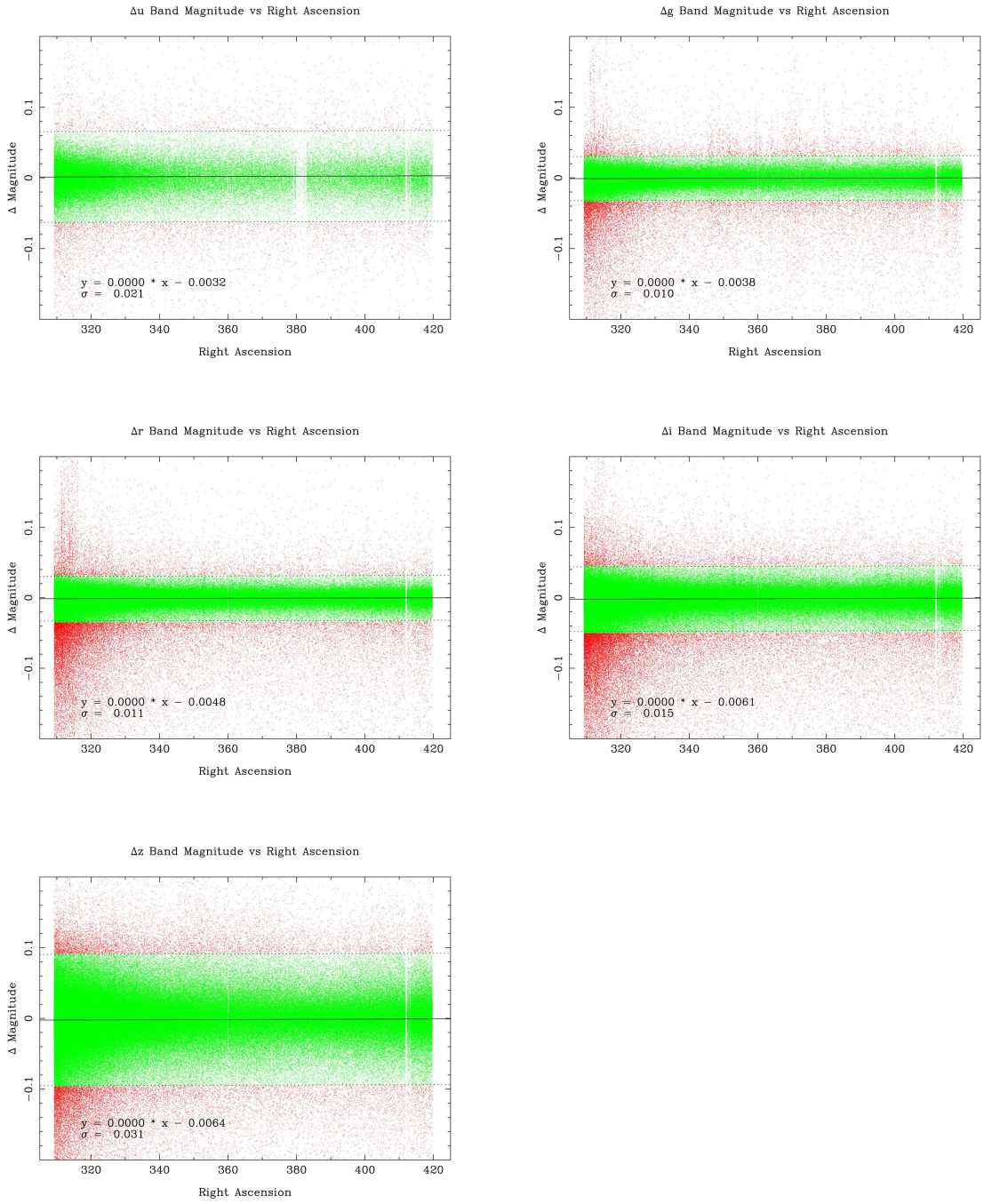


Figure 2.5 Comparison of co-added imaging MAG_AUTO and SDSS DR5 model magnitudes for all of in stripe 82 as a function of right ascension. All objects plotted are stars found in both the co-added catalog and SDSS. The fit was performed only over bright magnitudes ($16 < r < 19$) using iterative 3σ clipping in order to demonstrate that the magnitude zeropoint does not appreciably vary within a camcol. Dotted lines show the 3σ cut.

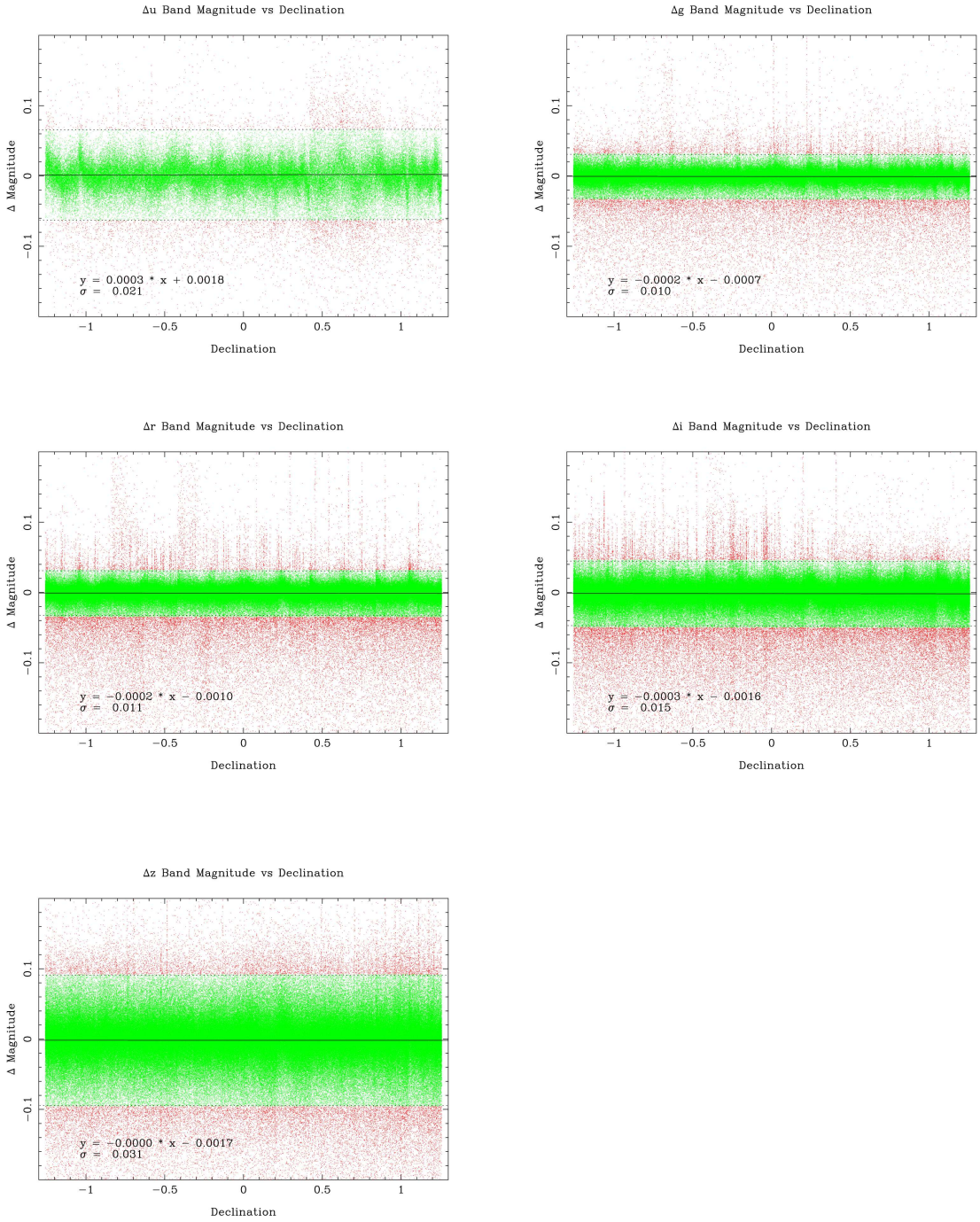


Figure 2.6 Comparison of co-added imaging MAG_AUTO and SDSS DR5 model magnitudes for all of stripe 82 as a function of declination. All objects plotted are stars found in both the co-added catalog and SDSS. The fit was performed only over bright magnitudes ($16 < r < 19$) using iterative 3σ clipping in order to demonstrate that the magnitude zeropoint does not appreciably vary between camcols. Dotted lines show the 3σ cut.

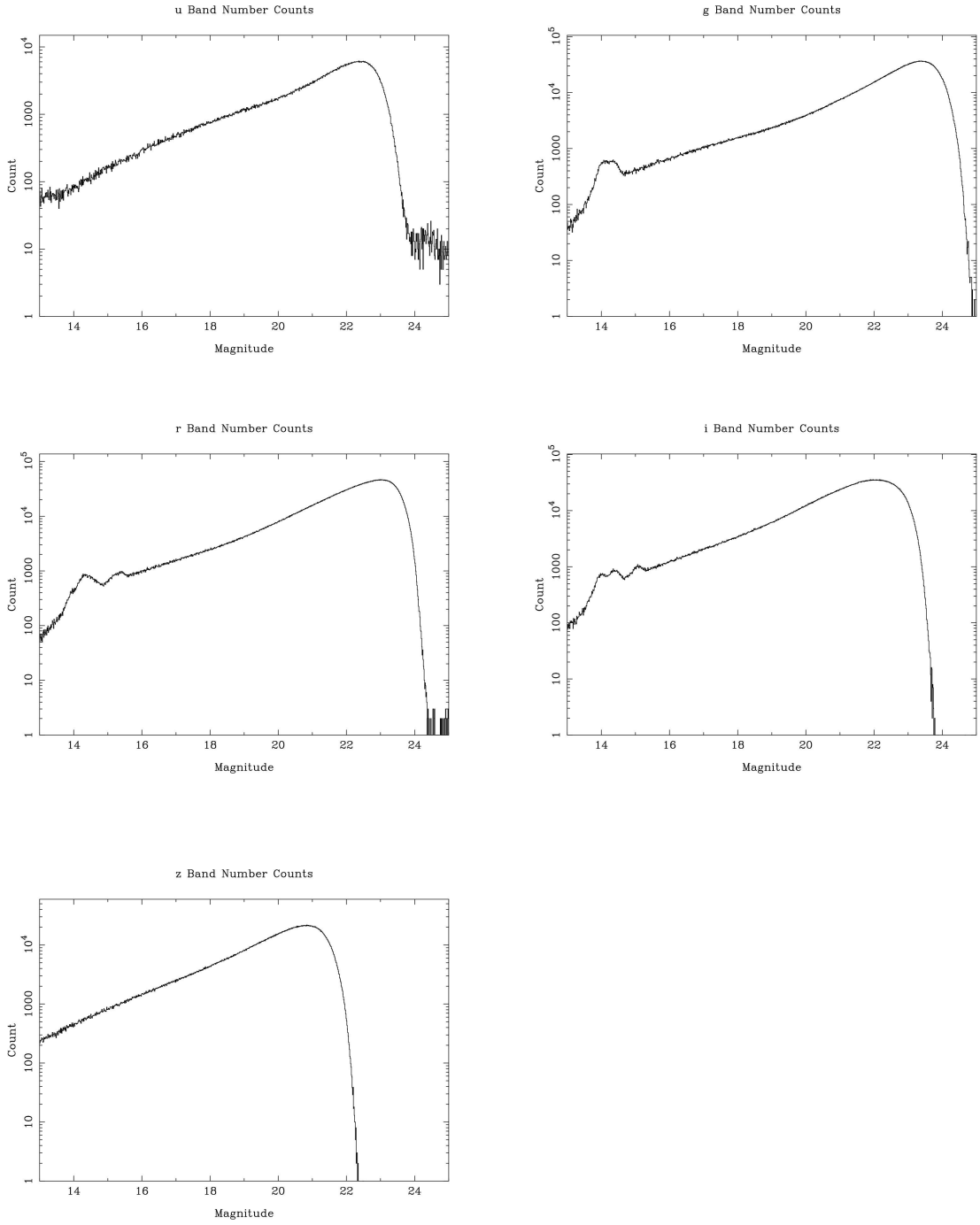


Figure 2.7 Number counts of co-added imaging MAG_AUTO for all of stripe 82. These plots show both stars and galaxies.

Because photometric redshifts are measured from the colors of galaxies (which are extended), the use of matched aperture magnitudes should increase the accuracy of our photometric redshifts.

3.0 STAR/GALAXY CLASSIFICATION

We are interested in the clustering properties of galaxies which are correlated with the large scale structure of the universe. Stars, on the other hand, are correlated with the plane of the Milky Way, which unfortunately runs through the middle of stripe 82. It is therefore necessary that we develop a robust method for classifying objects in our catalog as stars or galaxies to minimize stellar contamination of the galaxy clustering signal.

So, given an object, how does one determine its type? There are two pieces of information that are typically used for star/galaxy separation. First, galaxies are extended objects, and stars are point sources, assuming they are not sufficiently bright to cause diffraction spikes or CCD blooming. Second, the intrinsic spectra of galaxies and stars are different; hence their colors can be used for classification if a full spectrum is not available.

To classify based on object size, one can use magnitudes at two different apertures and compute their difference, which is termed *concentration*. By construction, the concentration of stars should be smaller than galaxies; moreover, there should be two visible populations in a concentration histogram (Scranton et al., 2002). The drawback to using concentration or object size is that faint galaxies are also very compact, so at some limiting magnitude it will become impossible to classify objects.

For color based classification, one compares the colors of the object to a multidimensional color manifold for some training set of stars and galaxies. The simplest form of this approach is to define some color cut (a piece-wise hyper-plane in color space) to classify the sample (Coil et al., 2004, *e.g.*). The drawback to using color cuts is that they introduce complicated biases into object selection – very red stars, for example, could easily be misclassified as high redshift galaxies. Additionally, because colors compare magnitudes in different pass bands, the limiting magnitude at which colors become significantly affected by noise is determined

by the band with the largest error (u or z for SDSS); that is, one cannot compare an object's magnitude in the r band to the u band if it is only detected in r.

For the reasons outlined above, we chose to use concentration distributions to classify objects in the co-added imaging data set. This approach is straightforward and we feel that it introduces less type bias into the galaxy selection. As a final note, one could try to combine these two approaches, but more sophisticated classification is beyond the scope of this thesis.

The remainder of this chapter gives an overview of the classification algorithm and concludes with suggested improvements to our algorithm.

3.1 FITTING THE CONCENTRATION DISTRIBUTION

Because the ratio of galaxies to stars increases with apparent magnitude, the distribution of concentrations is a strong function of apparent magnitude. For this reason, it is necessary to break up the sample to be classified into apparent magnitude bins and compute the concentration distribution in each bin; we used the SDSS r band (which has the best photometry) with bins of size 0.5. We compute the concentration by comparing an aperture magnitude at 3σ of the PSF[1] for the image to SExtractor MAG_AUTO, the optimal flux measurement for galaxies and stars convolved with Gaussian seeing. MAG_AUTO fits an elliptical aperture to the object's light distribution and then applies an algorithm similar to the "first moment" algorithm of Kron (1980). The concentration distribution in each bin can then be used to classify objects in that bin, which given all of the bins, yields a classifier as a function of magnitude.

For this thesis we attempted to use two classification schemes, one parametric and one non-parametric. These are discussed below.

[1]The point spread function (PSF) for each image is estimated as part of the calibration process.

3.1.1 Parametric Classification

The form of concentration distribution suggests that it could be modeled by a sum of two Gaussians. The advantage of this parameterization is obvious – one can associate one Gaussian with stars and the other with galaxies and assign a probability that an object is a star or galaxy based on the likelihood ratio of these two distributions. A probabilistic classification would enable the object selection to be optimized for the science question at hand, *e.g.* select only objects with probability $> 95\%$ of being a galaxy. Additionally, at faint magnitudes, when the concentration distribution becomes broad with only one peak, it would still be possible to fit the distribution and hence classify objects.

As a first approach, we applied the Expectation-Maxmization (EM) algorithm (Hastie et al., 2001) to Gaussian mixture models to fit the concentration distribution. EM is an iterative maximum likelihood technique that determines an initial estimate of the likelihood from the starting parameters (expectation step) and then varies the parameters to maximize the expected likelihood (maximization step). This process is then repeated with the new parameters input for the next expectation step. For a simple 2 Gaussian model, each iteration amounts to fuzzy classification – each object is assigned a weight between 0 and 1 describing how likely it belongs to each population, then the average and variance for each population are computed using those weights. The new parameters are then used to update the weights on the next iteration.

Figure 3.1 shows the results for an idealized concentration distribution. Unfortunately, as Figure 3.2 demonstrates, the real concentration distributions are not equivalent to a sum of two Gaussians, though they are close. The main reason the fit fails is that the broader Gaussian (for galaxies) cannot fit the distribution at small concentrations where the distribution abruptly goes to 0; this happens because below some limiting aperture, there simply is no flux. To compensate, the right Gaussian is shifted to higher concentration, which makes the overall fit worse in the most important area – the region between the two Gaussians.

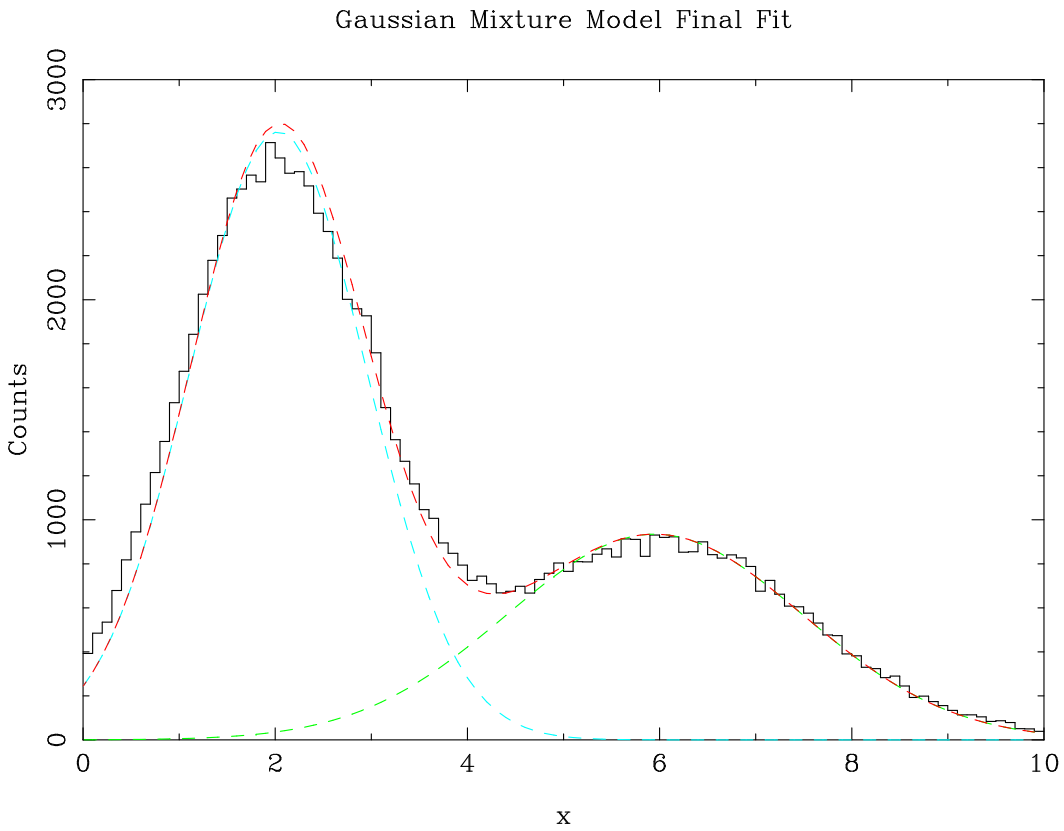


Figure 3.1 EM Gaussian mixture model fit to an idealized distribution of 2 Gaussians. The red line is the sum of both Gaussians, and the blue and green lines represent the contributions from each individual Gaussian.

3.1.2 Non-parametric Classification

The algorithm for our non-parametric classifier is simpler than the EM approach but more robust to the actual distributions: find the location of the valley between the two peaks in the concentration distribution. We used the *mean shift* algorithm (Carreira-Perpiñán, 2007; Fukunaga and Hostetler, 1975) to determine the location of the valley. The mean shift algorithm is a simple adaptive gradient ascent method applied to *kernel density estimation* (Hastie et al., 2001), a method for approximating distributions using a smoothing kernel. We used a Gaussian kernel so that the density estimate at point x given a distribution of data $x_i \in (x_1, x_2, \ldots x_N)$ and smoothing parameter h is

$$p(x) \approx \frac{1}{N\sqrt{2\pi h^2}} \sum_{i=1}^{N} \exp \frac{-(x - x_i)^2}{2h^2} \tag{3.1}$$

One nice feature of the kernel density estimate is that there is no dependence on bin size as in a histogram; the trade-off is that h is a parameter that needs to be tuned for the distribution. Typically there is some range of h that works well for a distribution. We used a value of $h = 0.02$, determined from "by eye" comparisons to histograms.

Given an estimate for $p(x)$, we can locate the extrema by taking $\frac{\partial p(x)}{\partial x}$ and setting it to 0. Doing so and solving for x gives the definition of the mean shift

$$x_{new} = \frac{\sum_{i=1}^{N} x_i \exp \frac{-(x-x_i)^2}{2h^2}}{p(x)} \equiv \text{meanshift} + x \tag{3.2}$$

I have written x_{new} on the left to signify that in order to compute this value, one must input some starting value of x. In fact, this is precisely how mean shift is used: input some starting value of x, compute the mean shift, set $x_{new} = x + \text{meanshift}$, and iterate until converged to within some tolerance. By choosing a series of initial starting values, one can locate all of the maxima of a distribution. Finally, note that the presence of $p(x)$ in the denominator is what makes mean shift an adaptive algorithm – it automatically moves away from areas with small probability densities.

In order to locate the valley between the two peaks, we first use mean shift to find both peaks by starting from one small concentration value and one large one[2]. Next we use

[2]The location of the peaks need not be very accurate as they are only used to determine which of the

bisection to pick concentration values (with the peaks as the initial bracketed region) and determine which of the two peaks mean shift converges to from the mid point value, which is used to update the bracket of the boundary between the peaks. This process is repeated and the "root"[3] bracket updated until its width is within a small tolerance. By definition, the midpoint of the bracket is approximately the location of the valley we are seeking.

Figure 3.3 shows the results of our algorithm for a bright magnitude bin, Figure 3.4 shows a bin near the middle of our magnitude range, and Figure 3.5 shows a faint bin. Finally, 3.6 shows a magnitude bin which is too faint to be classified using our algorithm. The minimum r band magnitude we can classify to is 22.261314; however, the number of galaxies is greater than stars at this point, so one can assume all sources fainter than this are galaxies with only a small amount of contamination. Similarly, sources that are brighter than $r = 14.738327$ cannot be classified, and we also flag objects that have abnormally high or low concentrations.

The concentration cut for each bin is recorded at the average magnitude value for that bin, then we interpolate to find the concentration cut to use for an arbitrary r band magnitude as shown in Figure 3.7. The brighter magnitude bins ($r < 18$) have only one prominent peak for stars, so the concentration cut used is constant with magnitude. To estimate the effect of errors on our concentration cuts, we also classified objects using $\pm 5\%$ variations in the cuts. As shown in Figure 3.8, this has a negligible effect on the resulting number counts.

3.2 EXTENDING STAR/GALAXY CLASSIFICATION

There are a number of ways in which the star/galaxy classification might be improved. The obvious next step is to improve the parametric classifier so that a probability can be assigned and fainter objects classified. One approach to this would be to use kernel density estimation to obtain $p(x)$ then use a non-linear fitting technique to fit two Gaussians to $p(x)$, ignoring the fit below some concentration threshold to avoid shifting galaxy population to fit the low

two populations a given point will migrate towards

[3] In this case there is no true root, as the starting value must converge to one of the two peaks.

concentration tail[4]. Additionally, it might be possible to fit only the region between the peaks; the other regions do not need to be particularly accurate given that they are assured to be stars or galaxies.

Additionally, it should be possible to incorporate prior information using Bayes' theorem into the classification. One candidate for a prior is an object's distance from the plane of the Milky Way because objects within the plane are more likely to be stars. This information can be quantified by using a dust reddening map and converting the reddening value into a probability. A second potentially useful prior could be constructed using the colors of the sample. Given the wealth of main sample SDSS galaxy data, it seems feasible that the color manifolds of stars and galaxies could be sampled well enough to use as a training set, and hence a useful prior.

[4]The parameterized $p_{\rm fit}(x)$ should be forced to 0 for values below this concentration

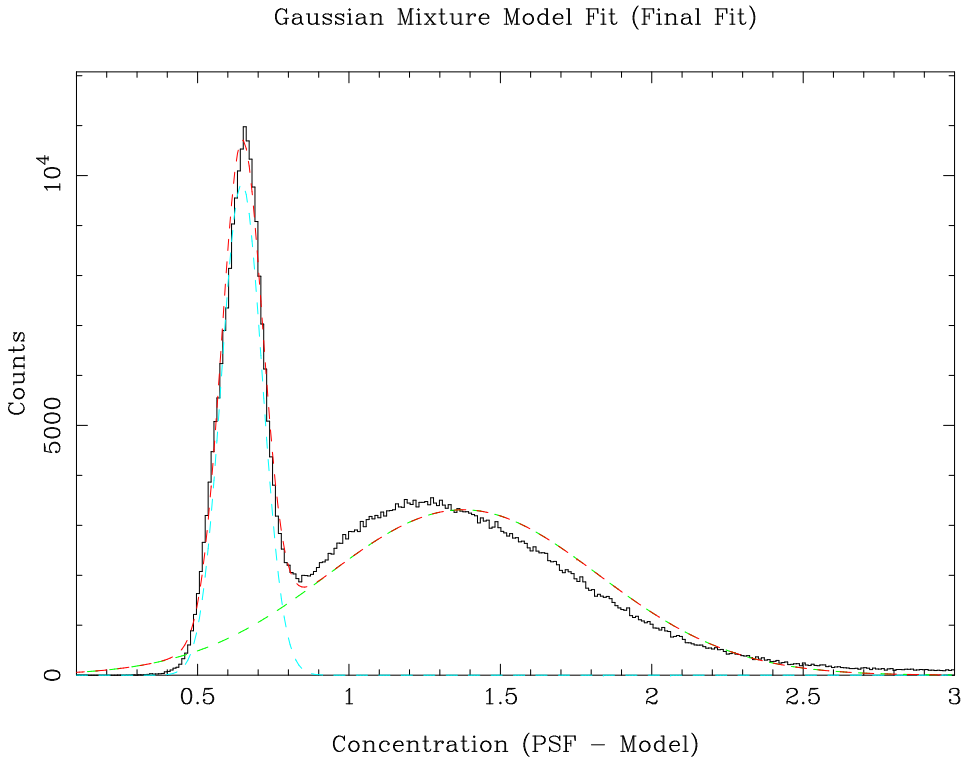


Figure 3.2 EM Gaussian mixture model fit to an actual concentration distribution. The red line is the sum of both Gaussians, and the blue and green lines represent the contributions from each individual Gaussian.

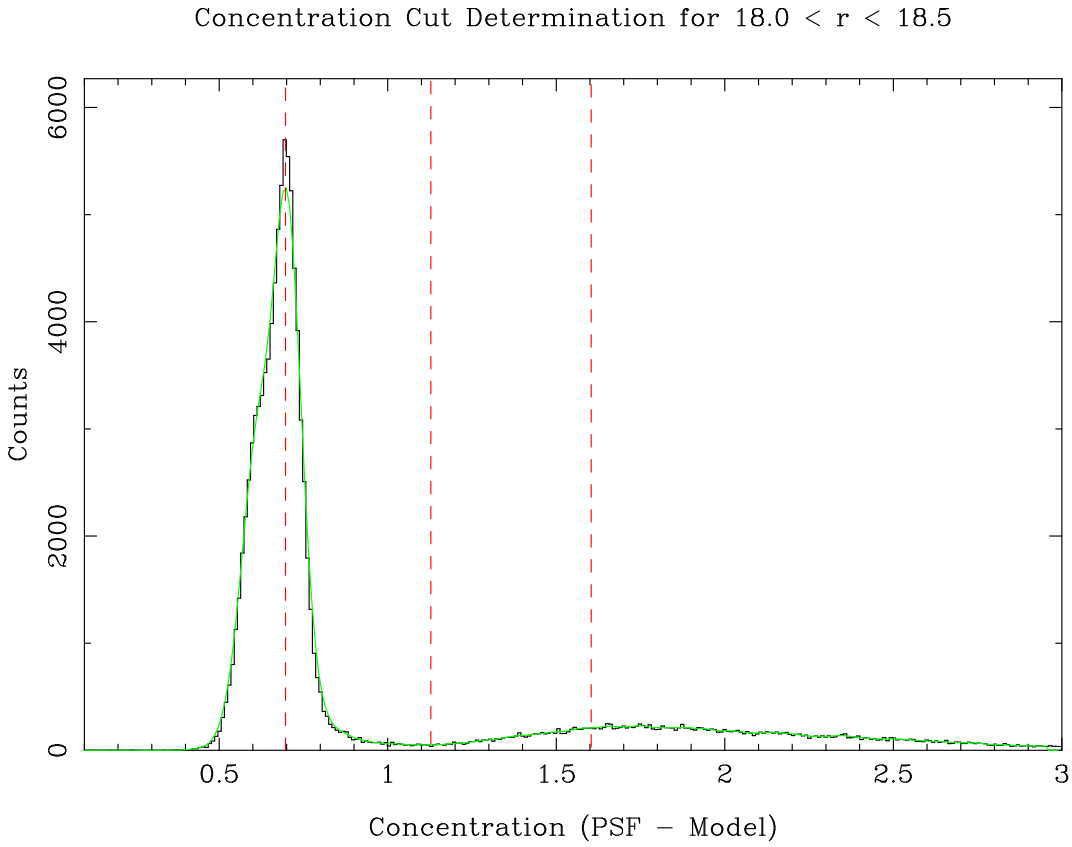


Figure 3.3 Concentration cut example for bright sources. The black line shows a histogram of the distribution and the green line shows the kernel density estimate (which is used to find the peaks). The dashed red lines show the location of the peaks and valley.

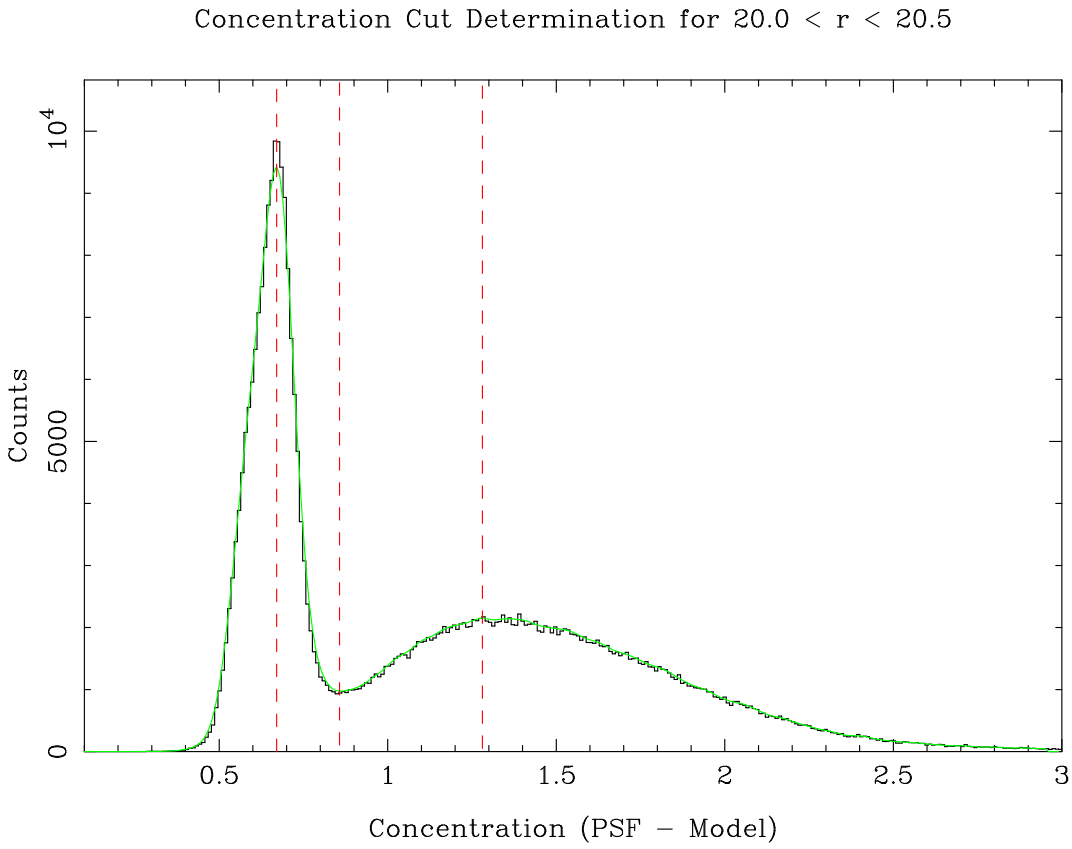


Figure 3.4 Concentration cut example for sources near the middle of our magnitude range. The black line shows a histogram of the distribution and the green line shows the kernel density estimate (which is used to find the peaks). The dashed red lines show the location of the peaks and valley.

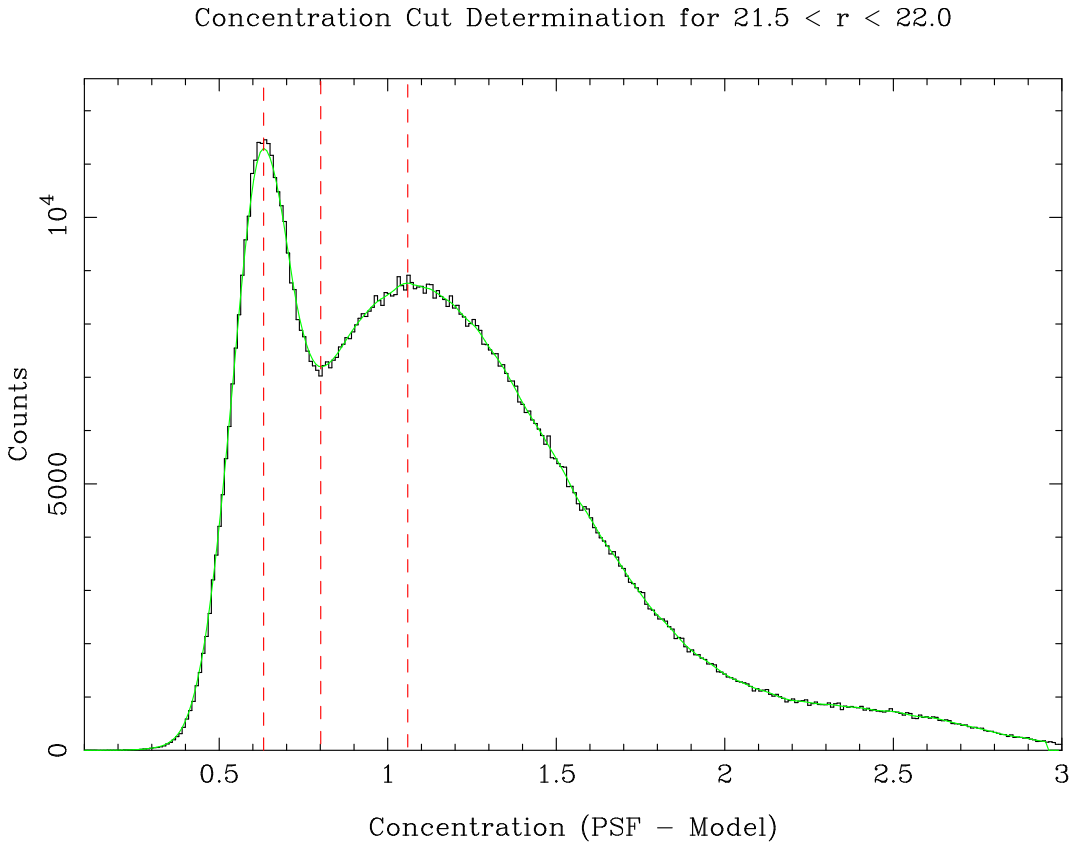


Figure 3.5 Concentration cut example for faint sources. The black line shows a histogram of the distribution and the green line shows the kernel density estimate (which is used to find the peaks). The dashed red lines show the location of the peaks and valley.

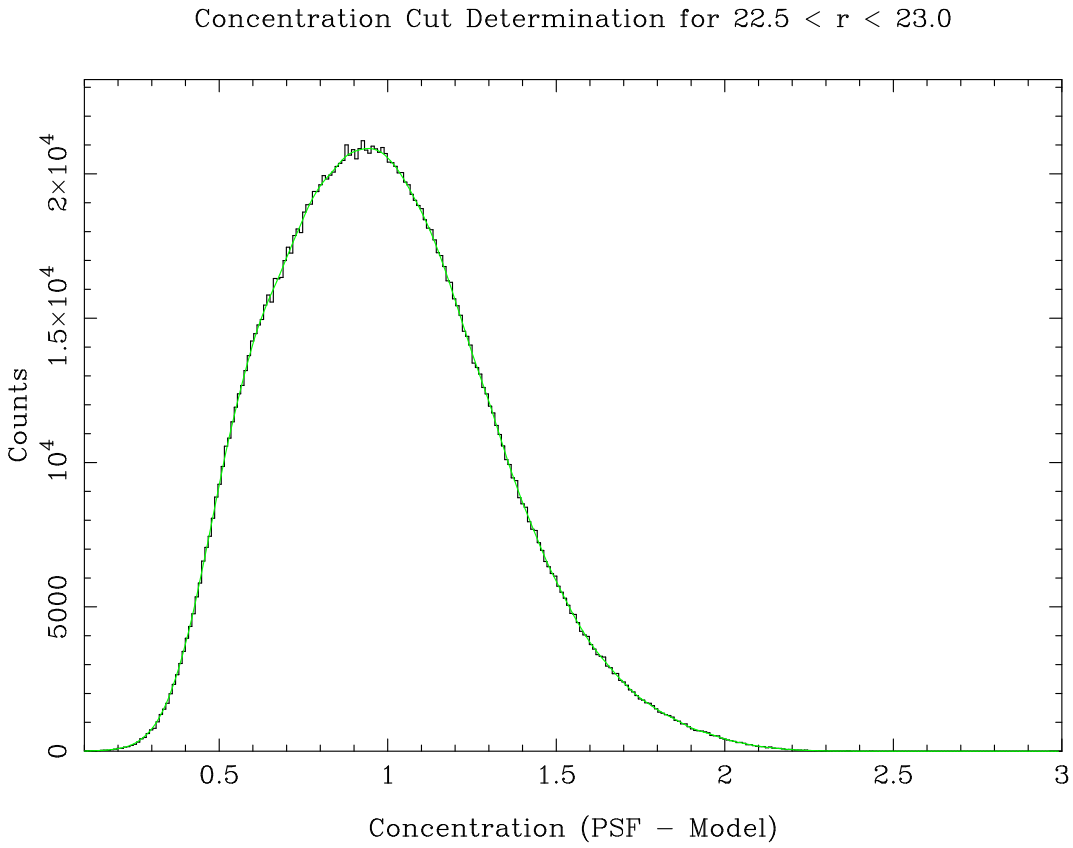


Figure 3.6 Concentration distribution for sources which are too faint to be classified

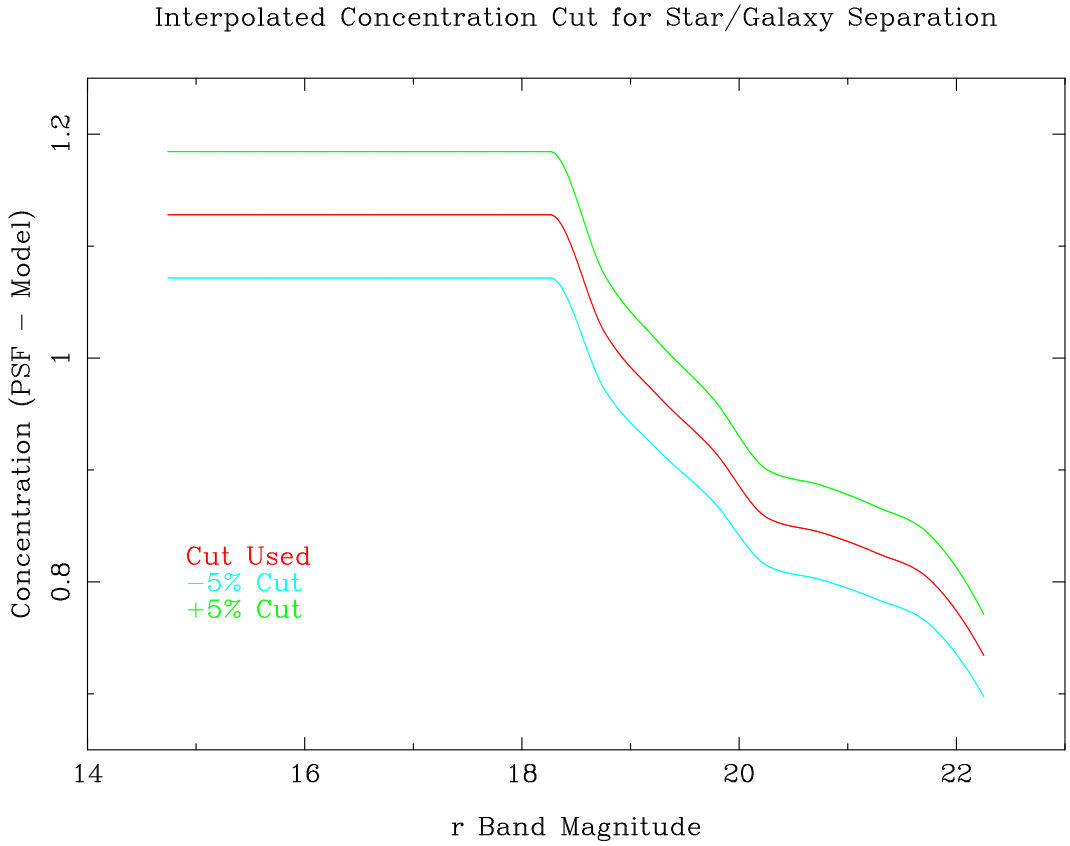


Figure 3.7 Interpolated concentration cut as a function of r band magnitude. The red line shows the cut used for classification, and the blue and green lines show $\pm 5\%$ values used to check number count variation with shifts in the cut used.

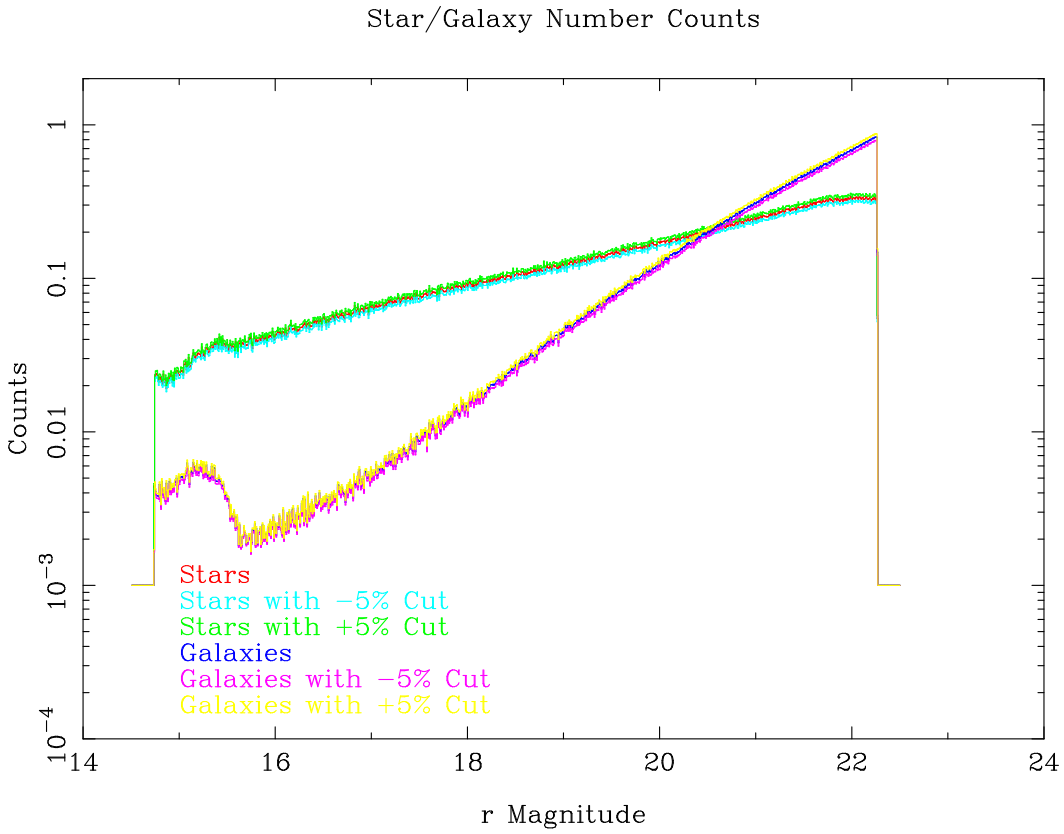


Figure 3.8 Star and galaxy number counts as a function of magnitude. There are 3 lines for both stars and galaxies – the actual number counts and the counts from using ±5% different concentration cuts.

4.0 PHOTOMETRIC REDSHIFTS

The expansion of the universe shifts the wavelengths of light emitted by stars and galaxies in an manner analogous to the Doppler effect. Because these objects are receding from us, they appear redder than if they were at rest. The shift from the emitted frame wavelength λ_e to observed frame wavelength λ_o is related to the redshift z through

$$\lambda_o = (1 + z)\lambda_e \tag{4.1}$$

We can exploit the fact that the redshift is due to the object's recession velocity to infer distance and time (Hogg, 1999); for this reason, redshift is the natural cosmological measure of time. Obtaining redshifts for our galaxy sample is thus essential to studying the evolution of galaxy clustering and to inferring the underlying 3-D distribution of galaxies.

Redshifts can be determined by running the light from an object through a spectrograph which separates its different wavelengths. Features of known rest wavelength (*e.g.* emission lines) can then be used to estimate the redshift with high accuracy. Additionally, the type and subtype (*e.g.* B0 star) of the object can be determined with high accuracy, ensuring that only galaxies are included in the clustering measurement. Unfortunately, taking a spectrum requires much longer exposure times than broadband photometry because the light is spread out over a larger physical area on the detector. As a concrete example, in SDSS there are approximately 100x more objects in the photometric sample than the spectroscopic sample.

The shorter observing times of broadband photometry make it attractive for large scale surveys because the larger number of objects allows for higher source density and subsampling of populations (*e.g.* how do blue galaxies cluster compared to red ones). Obviously, some method of estimating redshift using broadband photometric filters is highly desirable

as it would allow for larger redshift surveys. Collectively the techniques for estimating redshift from broadband photometry are termed *photometric redshifts* (Baum, 1962; Connolly et al., 1995; Bolzonella et al., 2000; Benítez, 2000; Csabai et al., 2003), as opposed to the traditional *spectroscopic redshifts*. Photometric redshifts, or "photozs" for short, are most simply explained as a low resolution (5 points for SDSS) spectrum that can only identify broad spectral features. In particular, there are two significant breaks in galaxy spectra at 912 Å (the Lyman break, due to neutral hydrogen absorption) and 4000 Å (the H-K or 4000 Å break, due to absorption by doubly ionized calcium and the Balmer series) that provide significant information even in low resolution. Of course, because photozs are estimated with less information, they are inherently less accurate, but the increased number counts can theoretically be used to reduce the scatter and allow for statistical measurements with high precision. In addition, many applications may not require high redshift accuracy as long as there is no bias in the redshift estimate. For galaxy clustering, only a few redshift bins are required to measure the evolution, so the measurement is only affected if the contamination due to incorrectly estimated redshifts within those bins is significant.

Photometric redshift techniques fall into two broad categories: empirical and template based. The empirical techniques tend to follow standard data mining approaches – given some training set of data with known redshifts and colors (*i.e.* a multi-dimensional color manifold), compare the colors of each object to be classified to those in the training set to determine redshift. For example, the weighted average of redshifts of the k nearest neighbors (with k as an input parameter) in color space can be used as an estimator. Errors can be estimated by jack-knifing the training set because randomly removing points in color space measures how well sampled the color manifold is. As with all training set algorithms, the redshift estimation will only be good if the training set accurately represents the color space and redshift distribution of the unclassified data set. One consequence of this is that empirical techniques can only be used for shallow photometric surveys because the spectroscopic sample used for training will have longer integration times and hence a brighter limiting magnitude.

Template based estimation seeks to overcome the shallow limitation by building the training set out of a small set of template spectra that are generated from theoretical galaxy models or empirical averages of multiple objects of similar type. Typically, the template set

includes elliptical, spiral, irregular, and star burst galaxies as well as hybrid types interpolated between each discrete spectrum. Each template spectrum can then be redshifted over an arbitrary range and have its colors measured by convolving the spectrum with the filter curves of the desired survey. Doing this for each template over some redshift range produces a training set which can then be used as in the empirical methods. Usually, though, most template based codes simply compute a χ^2 measurement of color distance for the entire template-redshift grid and take the minimum value as the redshift (*i.e.* use only 1 nearest neighbor). The error (or rather, the redshift probability density) can be estimated by turning the χ^2 into a probability for either the entire 2-D type-redshift grid or the grid marginalized over type.

It is important to note that the template based methods are not without problems. First, there is an implicit assumption that galaxy types do not evolve with time; the degree to which this assumption is violated is unknown. Second, we are presuming that we can accurately describe any galaxy in the universe as one of a handful of templates; alternatively stated, we can only estimate redshifts for objects which are well described by our template set. Third, at high redshift the color tracks of each template begin to cycle and create degeneracies, though *a priori* knowledge such as magnitudes can be used to partially correct for this. This problem is made worse for larger redshift grids. Additional degeneracies arise when distinguishing spectral features pass through a gap between two filters. All of these problems underscore that improvements to photometric redshift techniques are needed and research in this area is still currently ongoing.

The rest of this chapter outlines in detail the procedure used to obtain photometric redshifts for the co-added imaging set.

4.1 BAYESIAN PHOTOMETRIC REDSHIFTS

To estimate redshifts we use BPZ (Benítez, 2000), a template based photometric redshift code that implements an apparent magnitude Bayesian prior. A magnitude prior improves photoz quality by resolving color-type degeneracies through *a priori* knowledge. This makes

intuitive sense because we know that brighter galaxies are more likely to be nearby and hence at low redshift. Additionally, BPZ improves redshift error estimation by outputting the type marginalized $p(z)$ for each object.

The magnitudes/colors that we input into BPZ were those generated from the matched aperture catalog so that objects' fluxes were measured with the same aperture in all bands. Objects that were not detected in a given band (excluding the r band) had their magnitudes flagged so that BPZ would treat them as non-detections and use the 1 σ detection limit[1] for an upper bound flux threshold as additional information when determining the redshift.

We customized BPZ to improve photoz quality and error estimation by modifying its default behavior in three ways: we developed a prior more suitable for SDSS data, we parameterized the marginalized $p(z)$ for each object, and we created an alternate template set. We discuss each modification below.

4.1.1 Estimating the Prior

The probability that a given galaxy with colors C and apparent magnitude m has redshift z is given by applying Bayes' Rule to the set of templates T (Benítez, 2000):

$$p(z|C,m) = \sum_T p(z,T|C,m) \propto \sum_T p(z,T|m)p(C|z,T) \tag{4.2}$$

Here it is assumed that C and m are independent. The $p(C|z,T)$ term is simply the standard likelihood computed by comparing the object's colors to that of the templates. The first term is the apparent magnitude prior which can be further decomposed

$$p(z,T|m) = p(T|m)p(z|T,m) \tag{4.3}$$

These two terms are parameterized for a given training set following the method of Benítez (2000):

$$p(T|m) = \begin{cases} f_t e^{-k_t(m-m_0)} & \text{early and spiral} \\ 1 - p(T=\text{early}|m) - p(T=\text{spiral}|m) & \text{irregular} \end{cases} \tag{4.4}$$

[1]The 1 σ detection limit was approximated using the completeness limits for the SDSS main sample and assuming the relative offsets from the r band remained constant.

$$p(z|T,m) \propto z^{\alpha_t} \exp\left[-\left(\frac{z}{z_{0t}+k_{mt}(m-m_0)}\right)^{\alpha_t}\right] \tag{4.5}$$

There are 11 free parameters to be determined: $\{\alpha_t, z_{0t}, k_{mt}, k_t\}$ where t denotes the type used in the prior. The prior type t is distinct from the template type T because the prior is parameterized for 3 basic types: early/elliptical, spiral, and irregular. To apply the prior, each template type must be associated with one of the 3 prior types. Additionally, the fractions f_t at m_0 (the magnitude above which to apply the prior) must be determined, though they can be measured directly from the sample.

The parameters are estimated using Maximum Likelihood Estimation (MLE) by maximizing the log likelihood function with a simplex method (Galassi et al., 2006, *e.g.*):

$$\log \mathcal{L} = \prod_i p(T_i|m_i)p(z_i|T_i,m_i) \tag{4.6}$$

Here i labels objects in the data set used to estimate the prior. The normalization of $p(z|T,m)$ is the most computationally expensive part of this calculation. A useful optimization is to compute the normalization on a grid in m (for all 3 types) and interpolate rather than compute the normalization for each individual object.

We simplify the parameter estimation by solving for the set of k_t and f_t independently by maximizing the simpler log likelihood function

$$\log \mathcal{L}_2 = \prod_i p(T_i|m_i) \tag{4.7}$$

Application of the algorithm outlined above to a training set consisting of objects with known redshifts, magnitudes, and type (which can be determined using a simple color comparison with the templates at the known redshift) is straightforward.

For our prior, we used a mixed prior for SDSS r band computed from 2 data sets, the SDSS spectroscopic sample (Adelman-McCarthy et al., 2007) and the VIMOS VLT Deep Survey (VVDS) (Le Fèvre et al., 2003, 2004). We used the SDSS prior with $m_0 = 16$ for $r < 20$ and the VVDS prior with $m_0 = 20$ for fainter magnitudes. The full list of prior parameters for SDSS is given in Table 4.1 and the VVDS prior in Table 4.2. Incorporating the

prior reduced a small number of catastrophic outliers and improved the slope and intercept in Figure 4.1 by $\approx 5\%$.

4.1.2 Estimating Redshift Probability

Most photoz codes do not provide realistic error estimates of the redshifts they output. Furthermore, the idea that an object's photometric redshift is a single value is incorrect; it should be regarded as a probabilistic estimate and hence a spread of values with associated likelihood. Because template based photoz codes compute a χ^2 value on a grid of redshift and type, the most accurate estimate of $p(z)$ would be this 2-D grid[2]. For our BPZ parameters, though, this would require a total of 1500×6 points[3] for each galaxy. Marginalizing this 2-D grid over type yields a factor of 6 improvement in size, but it still represents a substantial amount of data to store for each galaxy. A better solution is to parameterize $p(z)$. Schmidt (2007) showed that the type marginalized $p(z)$ is well approximated by a double Gaussian with 5 free parameters $\{\alpha,\sigma_1,\sigma_2,\mu_1,\mu_2\}$:

$$p(z) \simeq \frac{\alpha}{\sqrt{2\pi\sigma_1^2}} \exp\left[\frac{(z-\mu_1)^2}{2\sigma_1^2}\right] + \frac{(1-\alpha)}{\sqrt{2\pi\sigma_2^2}} \exp\left[\frac{(z-\mu_2)^2}{2\sigma_2^2}\right] \tag{4.8}$$

Our experience confirms the accuracy of this fit for a wide variety of shapes of $p(z)$: we were able to find excellent fits for over 99.9% of our galaxies using an implementation of the Levenberg-Marquardt algorithm (Galassi et al., 2006). Most of the failures are caused by extremely bimodal distributions with $z = 0$ and $z = 1.5$ degeneracies, implying that they are either objects at high redshift or objects with unusual colors (*e.g.* incorrectly classified stars).

Given a compact and analytic expression for $p(z)$, it is possible to represent photometric redshifts in a more statistically correct way. For instance, it is possible to compute $\frac{\mathrm{d}n}{\mathrm{d}z}$ by summing $p(z)$ for each galaxy rather than taking a histogram of reported photoz values (*i.e.* the peak values of $p(z)$). It is also possible to more intelligently bin objects in redshift by

[2]Of course, if the object is not well described by the template set, the type-redshift grid will not accurately describe $p(z)$.

[3]We compute redshifts on a grid from 0.01 to 1.5 in steps of 0.001.

integrating $p(z)$ over the bin and either making a confidence cut or assigning a weight based on the probability the object actually lies within the bin.

4.1.3 Template Selection and Tweaking

The default BPZ template set contains the 4 empirical templates of Coleman et al. (1980) and 2 theoretical star burst spectra (Kinney et al., 1996). For this thesis, we obtained elliptical and spiral templates which have been optimized for our data set (S. Schmidt, private communication). These "tweaked" templates were derived from the original templates using the method of Csabai et al. (2003). Briefly, this process involves adjusting the template spectra to more closely follow the observed tracks of the data sample through color space; in this way, the templates are trained using the data they will later classify.

Of all of the work we did to improve redshift quality, template tweaking had the most significant effect.

4.2 PHOTOMETRIC REDSHIFT RESULTS

Evaluating the quality of photometric redshifts is difficult because the most obvious metric, comparing to a sample of spectroscopic redshifts, will only cover a small fraction of the data. Additionally, the comparison will typically consist of only the brightest sources which have the best photometry and hence the smallest errors. To work around the latter issue, we combined spectroscopic samples from multiple surveys: the SDSS spectroscopic sample, the Canadian Network for Observational Cosmology Field Galaxy Redshift Survey (CNOC2) (Yee et al., 1998; Lin et al., 1999), and DEEP2 (Davis et al., 2003). Figure 4.1 shows the results of an iterative sigma clipping fit to photoz vs specz.

Another useful comparison is to plot the color tracks of both the templates and the data points and compare how well the templates span the color manifold of the data. Unfortunately color space is 4-D, so we must plot in color slices, making interpretation of overlapping color tracks more difficult (do they really overlap, or is it a projection effect?). As is evi-

dent in figures 4.2, 4.3, and 4.4 the ellipticals clearly have a distinct color track that is well modeled by the templates.[4] As a result, one would expect that elliptical galaxies have more accurate redshifts.

As a final test, we computed $\frac{\mathrm{d}n}{\mathrm{d}z}$ from the photoz distribution in 3 different ways. First, we computed a histogram of the photoz values and then interpolated to obtain a smooth function. Second, we summed the $p(z)$ distributions for each object. Finally, we used the parameterization of Baugh and Efstathiou (1993) with the median photometric redshift z_m:

$$\frac{\mathrm{d}n}{\mathrm{d}z} = \frac{3z^2}{2(z_m/1.412)^3} \exp\left[-\left(\frac{1.412z}{z_m}\right)^{\frac{3}{2}}\right] \tag{4.9}$$

The distribution of redshifts was computed using an apparent magnitude cut of $16 \leq r \leq 21$ and an absolute magnitude cut of $-22 \leq M_r \leq -18$. Bolzonella et al. (2000) apply a similar absolute magnitude cut to implement a crude luminosity function prior (it eliminates obviously suspect galaxies); for our data, we noticed a significant improvement in the shape of $\frac{\mathrm{d}n}{\mathrm{d}z}$ when applying this cut. As seen in Figure 4.5, the observed $\frac{\mathrm{d}n}{\mathrm{d}z}$ as computed from summed $p(z)$ approximately matches how a typical distribution should look, as parameterized by Equation 4.9 with a median redshift of 0.233. For comparison, we also show the distribution without absolute magnitude cuts (median $z_m = 0.254$) in Figure 4.6.

Our significantly improved $\frac{\mathrm{d}n}{\mathrm{d}z}$ demonstrates that we can remove objects with poorly estimated $p(z)$ distributions simply by making cuts in absolute magnitude. Alternatively stated, our most accurate redshift predictions are for objects with intrinsic luminosity near what we expect for our sample. The advantage of this is obvious – we now have a simple way of finding objects that are poorly described by the photoz template set. Unfortunately, there was insufficient time to further investigate the physical mechanisms by which selecting objects in luminosity removes photoz outliers. One possible explanation is that because absolute magnitude is very sensitive to object type[5], this cut becomes a sanity check on the estimated object type. Another possibility is that objects which are intrinsically bright or faint will tend to be more degenerate in color space as they will appear brighter or fainter

[4] This is the reason why luminous red galaxies (LRGs) are often studied at high redshift – they have more reliable photometric redshifts.

[5] At $z = 1$, the K correction can be off by ≈ 4 magnitudes if the object type is incorrect. See the next chapter for more information.

than the typical galaxy at their redshift.

Spectral Type	α_t	z_{0t}	k_{mt}	f_t	k_t
Early/Elliptical	2.24	0.062	0.052	0.585	0.043
Spiral	2.00	0.048	0.038	$0.250 + 0.145$	-0.045
Irregular	1.38	0.026	0.021	...	...

Table 4.1 Prior parameters used for $r < 20$ derived from SDSS spectroscopic sample with $m_0 = 16$. We estimated the fractions of the Sbc and Scd spiral templates separately because they were so abundant in the SDSS sample.

Spectral Type	α_t	z_{0t}	k_{mt}	f_t	k_t
Early/Elliptical	1.957	0.3214	0.1963	0.25	0.557
Spiral	1.598	0.2911	0.1667	0.54	0.100
Irregular	0.9638	0.1700	0.1290	...	...

Table 4.2 Prior parameters used for $r > 20$ derived from VVDS spectroscopic sample ($m_0 = 20$).

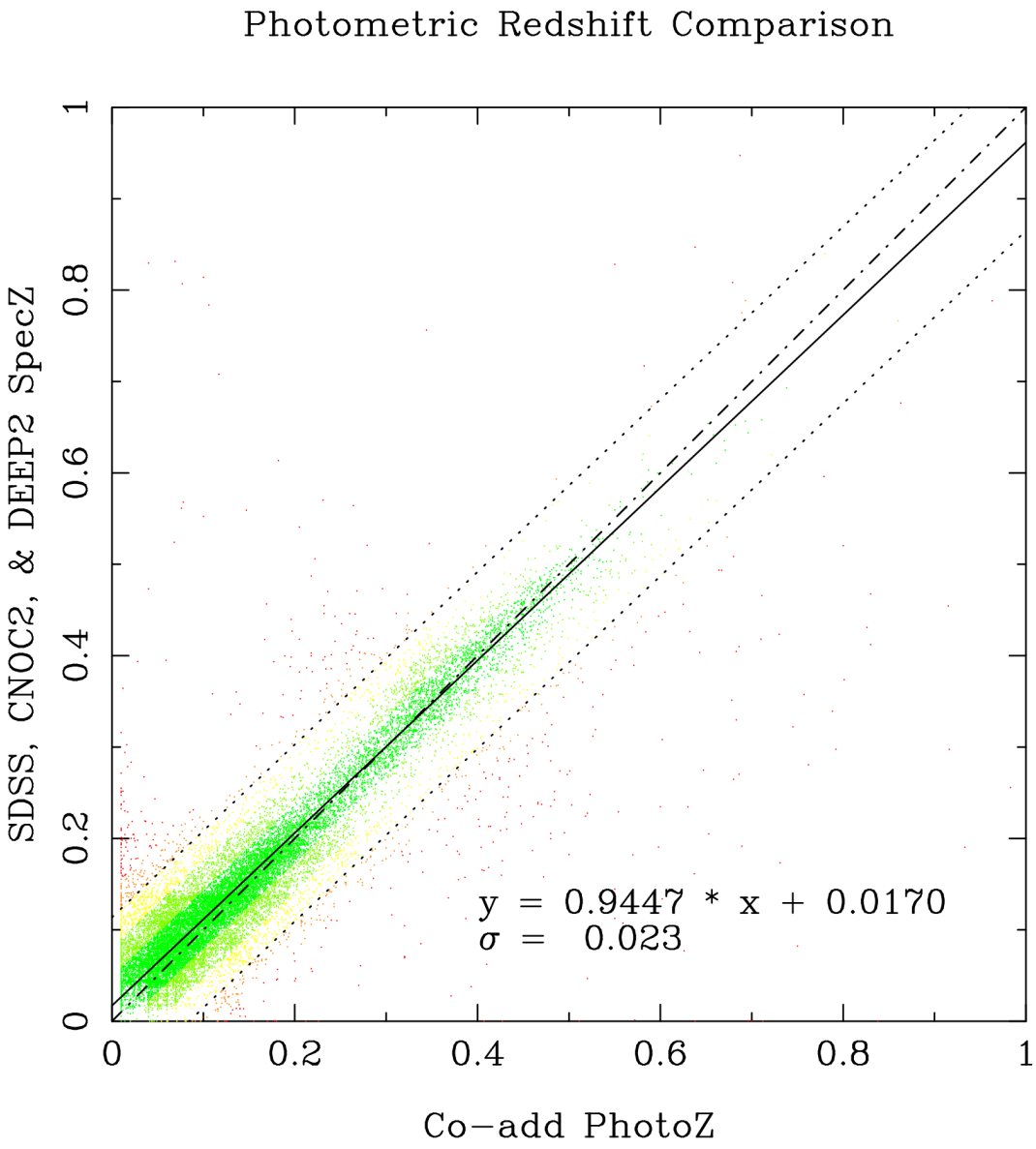


Figure 4.1 Comparison of photometric redshifts to spectroscopic redshifts from SDSS, CNOC2, and DEEP2. The fit was an iterative 3σ clipping algorithm with 2 DOF and a maximum of 5 iterations. The solid line shows the best fit and the dash-dot line shows the 1 to 1 line. The dotted lines show the 3σ cut. Color indicates the outlier status: bright green = $< 1\sigma$, dark green = 1-2σ, yellow = 2-3σ, orange = 3-4σ, and red = $\geq 4\sigma$.

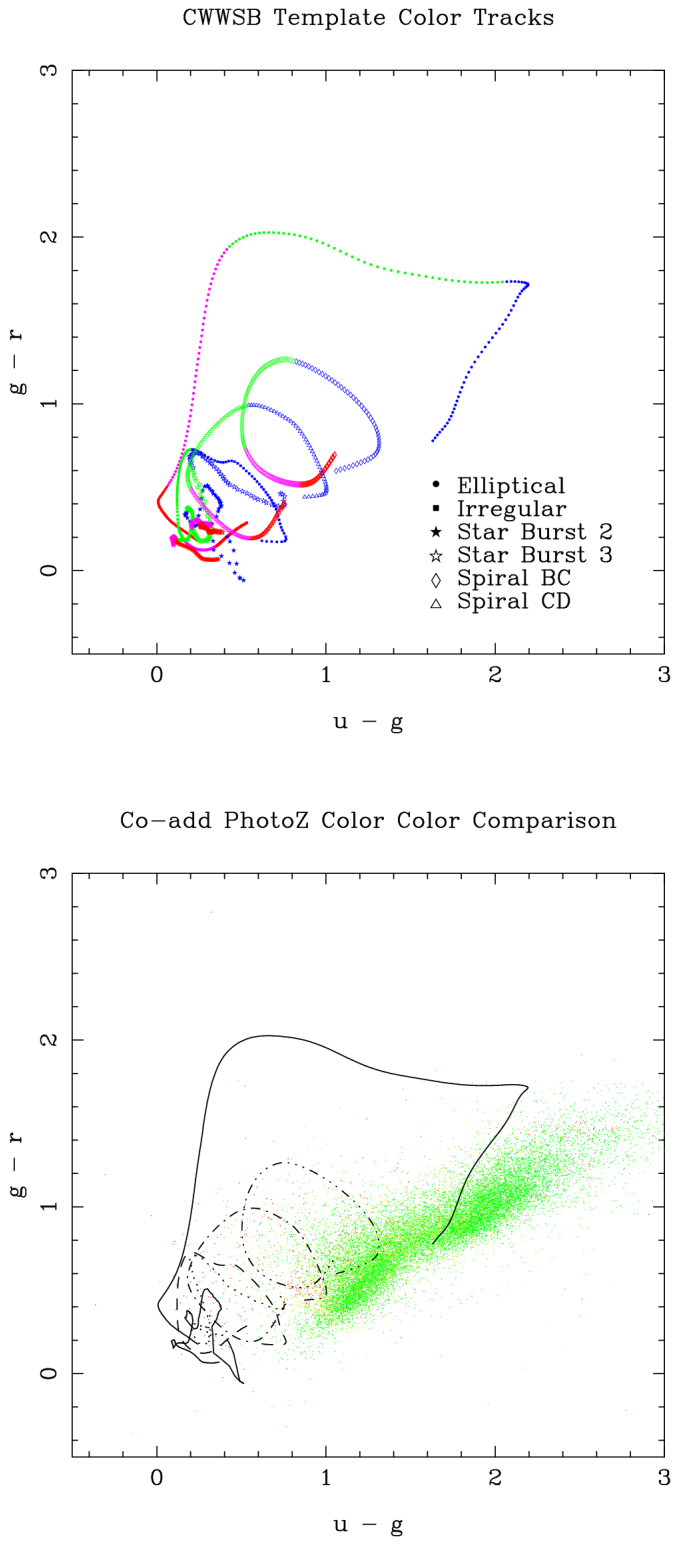
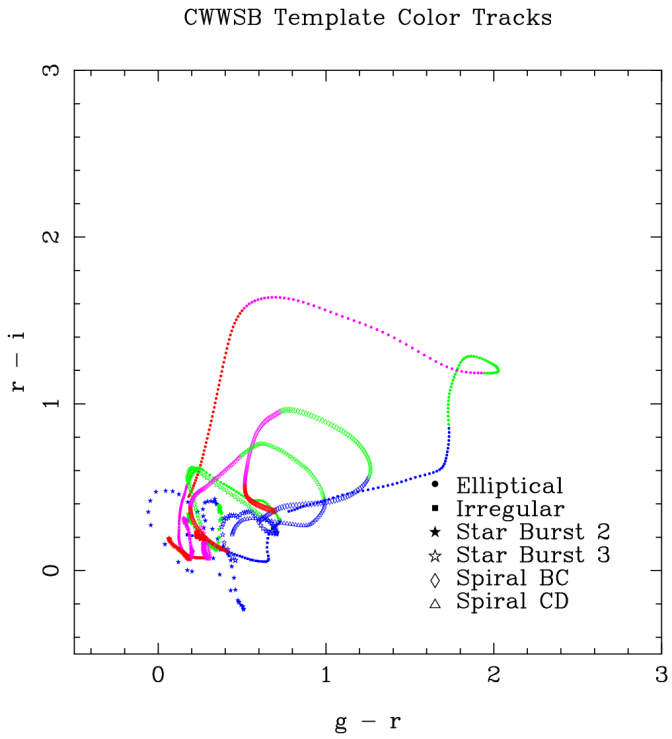


Figure 4.2 Color tracks of templates and SDSS co-add data for $u - g$ and $g - r$. The top panel shows the evolution of the templates from $z = 0$ to 2 and the bottom panel shows co-add data overlaid on the template tracks. In the top panel, point shape indicates template type and color indicates redshift, with blue $= 0 \leq z < 0.5$, green $= 0.5 \leq z < 1.0$, magenta $= 1.0 \leq z < 1.5$, and red $= 1.5 \leq z < 2.0$. In the bottom panel, line type indicates template type and color indicates the outlier status from Figure 4.1.



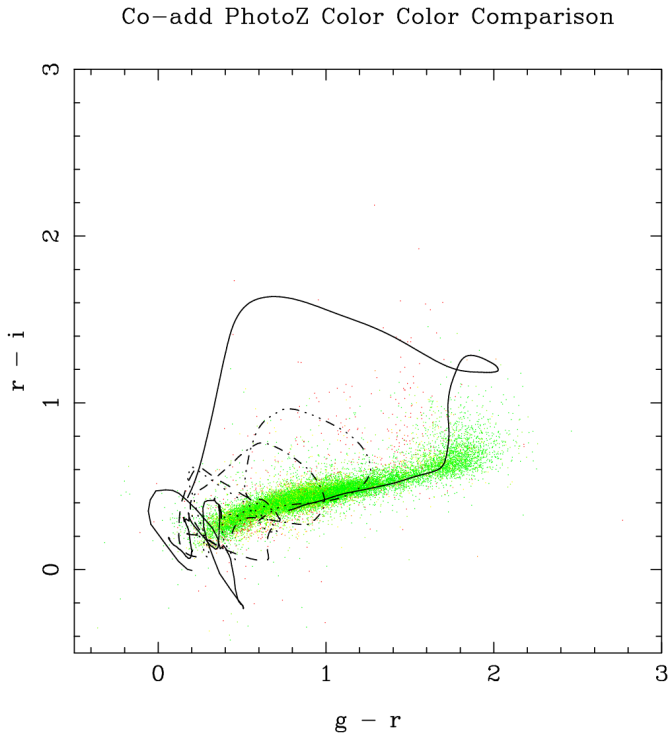


Figure 4.3 Color tracks of templates and SDSS co-add data for $g - r$ and $r - i$. The top panel shows the evolution of the templates from $z = 0$ to 2 and the bottom panel shows co-add data overlaid on the template tracks. In the top panel, point shape indicates template type and color indicates redshift, with blue $= 0 \leq z < 0.5$, green $= 0.5 \leq z < 1.0$, magenta $= 1.0 \leq z < 1.5$, and red $= 1.5 \leq z < 2.0$. In the bottom panel, line type indicates template type and color indicates the outlier status from Figure 4.1.

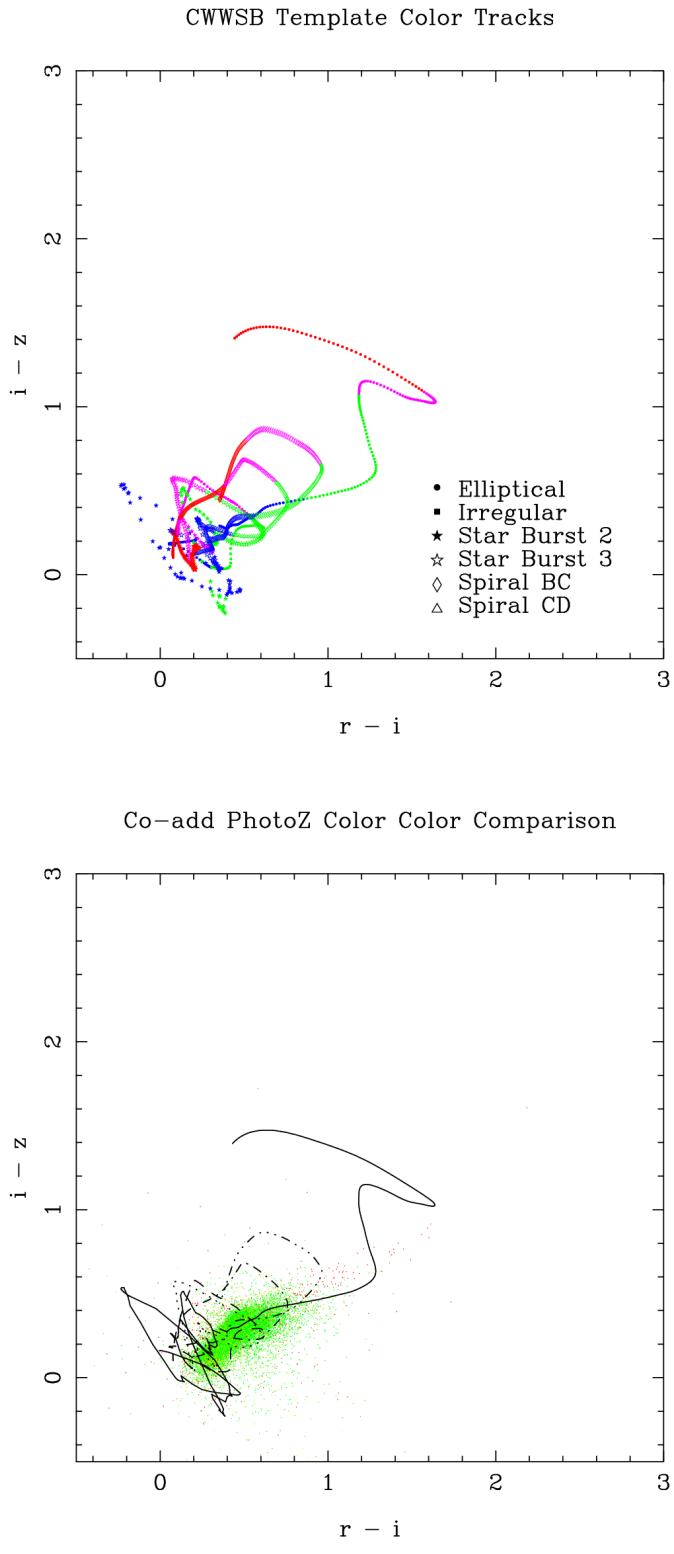


Figure 4.4 Color tracks of templates and SDSS co-add data for $r - i$ and $i - z$. The top panel shows the evolution of the templates from $z = 0$ to 2 and the bottom panel shows co-add data overlaid on the template tracks. In the top panel, point shape indicates template type and color indicates redshift, with blue $= 0 \leq z < 0.5$, green $= 0.5 \leq z < 1.0$, magenta $= 1.0 \leq z < 1.5$, and red $= 1.5 \leq z < 2.0$. In the bottom panel, line type indicates template type and color indicates the outlier status from Figure 4.1.

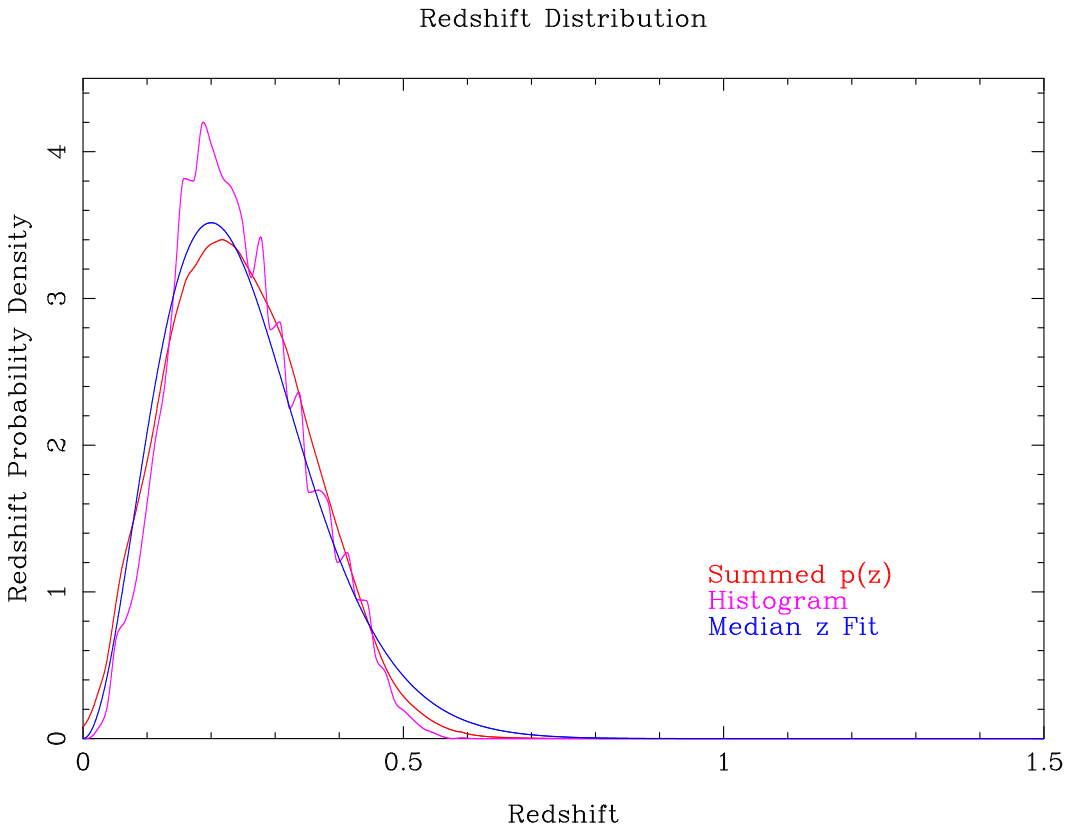


Figure 4.5 Comparison of 3 ways of estimating $\frac{dn}{dz}$ for the co-add data set. The sample was selected with an apparent magnitude cut of $16 \leq r \leq 21$ and an absolute magnitude cut of $-22 \leq M_r \leq -18$. The median redshift was 0.233.

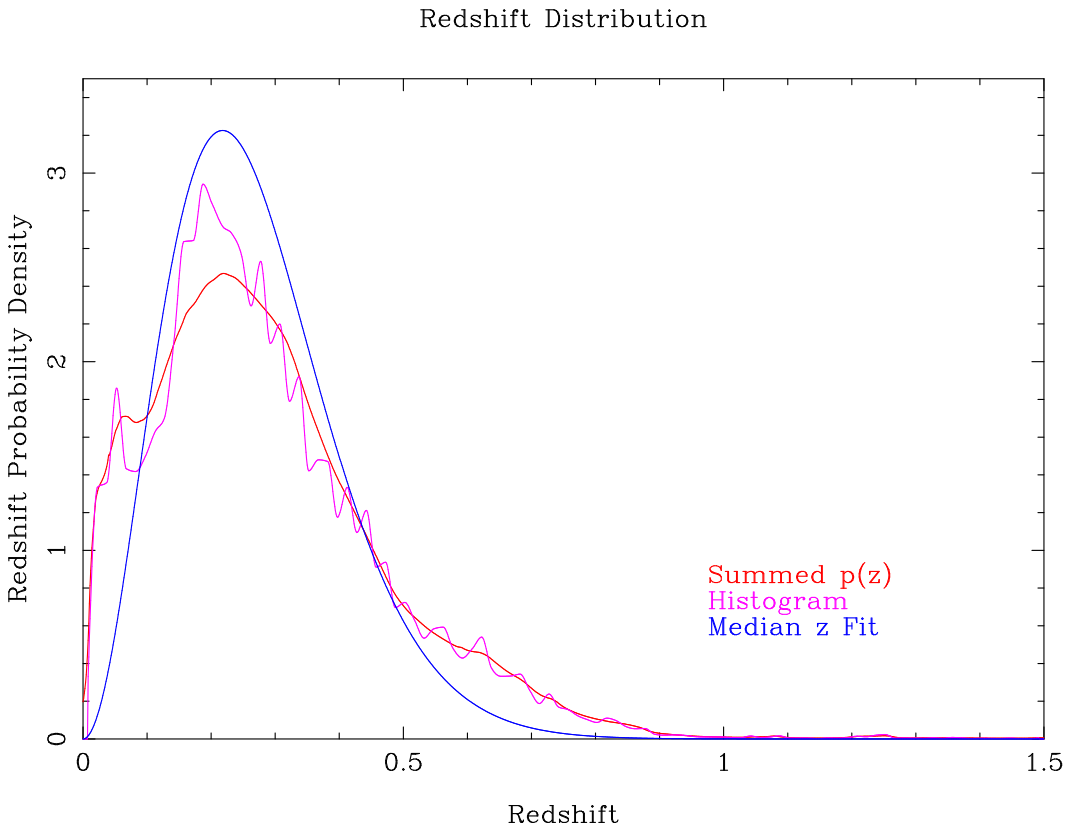


Figure 4.6 Comparison of 3 ways of estimating $\frac{dn}{dz}$ for the co-add data set. The sample was selected with an apparent magnitude cut of $16 \leq r \leq 21$. The median redshift was 0.254.

5.0 THE ANGULAR CORRELATION FUNCTION

The standard galaxy clustering measurement is the two point angular correlation function, $w(\theta)$. The angular correlation function is defined as how much more or less likely than random a pair of galaxies will be found with a separation on the sky of θ within a solid angle $\mathrm{d}\Omega$:

$$\mathrm{d}P = \bar{n}(1 + w(\theta))\mathrm{d}\Omega \tag{5.1}$$

Here $\bar{n}$ is the mean number of galaxies per solid angle. $w(\theta)$ is therefore an excess probability, measuring the probability that an event will occur more or less frequently than a random event. For a Gaussian random field, $w(\theta)$ and its Fourier transform pair the power spectrum fully specify the distribution of galaxies. Even if the case of non-Gaussianity, the angular correlation function remains a useful statistic for quantifying galaxy clustering.

The two point angular correlation function has been measured since the very first large scale galaxy surveys which probed brighter magnitudes (Groth and Peebles, 1977; Maddox et al., 1990; Collins et al., 1992), and it has consistently been found to be well described by a power law on small scales with a steeper decrease beginning at $\sim 1^\circ$. For reasons which will become apparent later, the power law form of $w(\theta)$ is typically written as (Martínez and Saar, 2002)

$$w(\theta) = A\theta^{1-\gamma} \tag{5.2}$$

The slope and intercept in log-log space are then $1-\gamma$ and $\log A$ respectively with $\gamma \approx 1.7$ More recent measurements that probe fainter magnitudes (Connolly et al., 2002; Willmer et al., 2006) revealed that the power law slope remains roughly constant but the intercept

decreases with fainter magnitudes. Similar results are observed when galaxies are binned in luminosity or redshift.

In practice, $w(\theta)$ is measured by counting pairs of galaxies in an annulus of radius θ and width $\pm\delta\theta$ for multiple values of θ. Because $w(\theta)$ is an excess probability, the pair counting must be done for both the data set and a set of random points uniformly distributed over the sphere in the same area of the sky[1]; a statistical estimate of $w(\theta)$ can then be computed from the pair counts. The well known Landy-Szalay estimator (Landy and Szalay, 1993) is a minimum variance estimator and hence requires the fewest number of random points to estimate $w(\theta)$:

$$w(\theta) = \frac{DD - 2DR + RR}{RR} \tag{5.3}$$

Here DD represents the number of pairs between the data set compared to itself, DR pairs between the data and random sets, and RR pairs between random and itself. The size of the random set is typically 6-10x that of the data set so that Poisson errors in the random data set do not effect the $w(\theta)$ measurement; for this thesis we generated 10x more randoms than data.

The remainder of this chapter is devoted to the many details associated with the measurement of the angular correlation function and its interpretation in the framework of a power law correlation function.

5.1 COMPUTING THE CORRELATION FUNCTION AND ITS ERROR

To compute $w(\theta)$ we use the code of R. Scranton (private communication). Scranton's code uses a hierarchical 2-D grid data structure based on SDSSPix (Tegmark et al.) to efficiently locate points on the sphere[2]. On small scales, pairs are counted exactly by examining neighboring grid cells. On large scales, pairs are counted using an approximate scheme

[1]The randoms are effectively a Monte Carlo integration over the annulus – the integral is too difficult to perform analytically

[2]The grid projection breaks down near the poles, but fortunately our data lie near the equator.

which utilizes the hierarchical nature of the grid cells to group together many cells and hence reduce the number of cells to examine.

The approximate counting algorithm is necessary to achieve good performance on large scales because a huge number of tiny (compared to the annulus radius) cells must be examined to perform an exact pair count. This is because spatial data structures are efficient at finding all points near a given location, but they become inefficient when a significant fraction of points must be examined (*i.e.* for a large search radius). The transition scale between these two pair counting methods is a configurable flag typically set to 0.09°.

The same hierarchical grid data structure is used to describe the area of the survey in a set of grid cells (also called pixels) termed a *map* or *mask*, depending on whether the area is to be added or subtracted respectively. Scranton provides a C++ API called STOMP (Scranton et al., 2008) for efficiently performing a variety of map related operations such as area computation, addition/subtraction, and intersection.

Measuring $w(\theta)$ requires minimizing systematic contaminants of the galaxy clustering signal as possible. Additionally, it is important that the random data points are generated to fill the same geometry on the sky as the survey. The map facilities provided by STOMP can help accomplish both of these goals. To compute $w(\theta)$ we we combined several maps using STOMP: a map describing the basic area of stripe 82, a mask to remove bright stars (which create many spurious objects along diffraction spikes), a mask to remove areas from fields which we failed to calibrate, and a mask to remove highly reddened ($E(B-V) > 0.2$) areas of the stripe due to the galactic plane. This final mask was necessary because our star-galaxy classification appears to be correlated with the galactic plane. Obtaining reliable star-galaxy classification near the galactic plane is difficult due to the increased star number counts and dust (which tends obscure local stars so they appear to be faint galaxies), so this is not surprising. The final area we used for the calculation was 175.95 deg^2.

The errors for $w(\theta)$ are computed by creating a set of jack-knife samples (*i.e.* by generating multiple random data sets) and observing the change in $w(\theta)$. The simplest form of this approach is simply to leave out 1 random chunk of the random data and observe the change in $w(\theta)$. We used 48 jack-knife samples, which is 2x the number of bins in θ. A covariance matrix containing information about how the errors in separate angular bins are correlated

is also determined from the jack-knife errors. We use this covariance matrix when computing fits to $w(\theta)$ to account for cross correlations between angular bins. For more details on the calculation of $w(\theta)$ and its errors, see Scranton et al. (2002).

5.2 LIMBER'S EQUATION

In the limit of small separations, $w(\theta)$ can be related to the real space two point correlation function $\xi(r)$ through an expression known as Limber's equation (Peebles, 1980, 1993; Martínez and Saar, 2002). In this section, we provide a brief derivation of all of the relevant equations for relating $w(\theta)$ to $\xi(r)$.

To compute the angular correlation function, one must integrate over the comoving lines of sight for 2 objects at 3-D positions $\vec{r_1}, \vec{r_2}$ separated by an angle θ on the sky:

$$w(\theta) = \int\int \xi(\vec{r_1}, \vec{r_2}) r_1^2 r_2^2 \phi(r_1)\phi(r_2)\, \mathrm{d}r_1\, \mathrm{d}r_2 \tag{5.4}$$

Here $r_1 = |\vec{r_1}|, r_2 = |\vec{r_2}|$, $\phi(r)$ is the radial selection function, and we have assumed that there are no curvature effects. The radial selection function incorporates the limitations of a survey by quantifying the probability of observing a galaxy at radial distance r. The selection function is normalized so that $\int_0^\infty \phi(r) r^2\, \mathrm{d}r = 1$.

Next we change variables to $u = r_1 - r_2$ and $r = \frac{1}{2}(r_1 + r_2)$ and note that the largest contributions to the integral come from when u is small so that $r_1 \approx r_2$. If we further assume that $\theta \ll 1$ so that $\cos\theta \approx 1 - \frac{\theta^2}{2}$, the distance d between the two galaxies becomes

$$d = \sqrt{|\vec{r_1} - \vec{r_2}|^2} = \sqrt{r_1^2 + r_2^2 - 2 r_1 r_2 \cos\theta} \approx \sqrt{u^2 + r^2\theta^2} \tag{5.5}$$

Finally, we change variables to u and r in the integral (the determinant of the Jacobian is 1 so that $\mathrm{d}r_1\, \mathrm{d}r_2 = \mathrm{d}r\, \mathrm{d}u$) to obtain Limber's equation:

$$w(\theta) = \int_0^\infty r^4 \phi^2(r)\, \mathrm{d}r \int_0^\infty \xi\left(\sqrt{u^2 + r^2\theta^2}\right)\, \mathrm{d}u \tag{5.6}$$

In practice the actual integration limits are finite as they are determined by the non-zero range of the selection function. We can relate the radial selection function to the observable redshift distribution $\frac{\mathrm{d}n}{\mathrm{d}z}$ by noting that the number of observed galaxies in a small radial shell is equivalent to the number of galaxies observed in a small width of redshift:

$$\phi(r)r^2\,\mathrm{d}r = \frac{\mathrm{d}n}{\mathrm{d}z}\,\mathrm{d}z \tag{5.7}$$

Here, we have again assumed a flat cosmology. Additionally, the redshift distribution must be normalized so that $\int_0^\infty \frac{\mathrm{d}n}{\mathrm{d}z}\,\mathrm{d}z = 1$. This equation is equivalent to the assumption that the true number density of galaxies is constant with comoving volume (*i.e.* a homogeneous universe) – to see this, simply divide both sides of Equation 5.7 by $r^2\,\mathrm{d}r = \frac{\mathrm{d}V}{\mathrm{d}\Omega}$. Substitution into Equation 5.6 yields a more practical expression of Limber's equation

$$w(\theta) = \int_0^\infty \left(\frac{\mathrm{d}n}{\mathrm{d}z}\right)^2 \left(\frac{\mathrm{d}r}{\mathrm{d}z}\right)^{-1} \mathrm{d}z \int_0^\infty \xi\left(\sqrt{u^2 + r^2(z)\theta^2}\right)\,\mathrm{d}u \tag{5.8}$$

The term $\frac{\mathrm{d}r}{\mathrm{d}z}$ is given by the standard cosmography measures (Hogg, 1999) with r as the comoving line of sight distance:

$$\frac{\mathrm{d}r}{\mathrm{d}z} = \frac{D_H}{E(z)} \tag{5.9}$$

$$E(z) = \sqrt{\Omega_M(1+z)^3 + \Omega_k(1+z)^2 + \Omega_\Lambda} \tag{5.10}$$

$$H(z) = H_0 E(z) \tag{5.11}$$

$$D_H = \frac{c}{H_0} \tag{5.12}$$

In order to compute Limber's equation, one must assume a set of cosmological parameters. For this thesis, we used the latest WMAP cosmology parameters (Hinshaw et al., 2008): $\Omega_M = 0.28$, $\Omega_\Lambda = 0.72$, $\Omega_k = 0$, $H_0 = 71$ km/s/Mpc. Integration of Equation 5.8 is straightforward once $\frac{\mathrm{d}n}{\mathrm{d}z}$ has been measured.

Finally, we can gain additional insight into the physical scales of galaxy clustering by assuming that the real space correlation function $\xi(r)$ is a power law:

$$\xi(r) = \left(\frac{r}{r_0}\right)^{-\gamma} \tag{5.13}$$

Under this assumption it can be shown (Peebles, 1980) by substitution of Equation 5.13 into 5.8 that $w(\theta)$ is also a power law

$$w(\theta) = r_0^{\gamma} H_{\gamma} \theta^{1-\gamma} \int_0^{\infty} r^{1-\gamma} \left(\frac{\mathrm{d}n}{\mathrm{d}z}\right)^2 \left(\frac{\mathrm{d}r}{\mathrm{d}z}\right)^{-1} \mathrm{d}z \tag{5.14}$$

$$H_{\gamma} = \frac{\Gamma(\frac{1}{2})\Gamma(\frac{\gamma-1}{2})}{\Gamma(\frac{\gamma}{2})} \tag{5.15}$$

This is equivalent to the previous power law expression for $w(\theta)$ in Equation 5.2 with

$$A = r_0^{\gamma} H_{\gamma} \int_0^{\infty} r^{1-\gamma} \left(\frac{\mathrm{d}n}{\mathrm{d}z}\right)^2 \left(\frac{\mathrm{d}r}{\mathrm{d}z}\right)^{-1} \mathrm{d}z \tag{5.16}$$

Note that $A \propto r_0^{\gamma}$. If we measure $w(\theta)$ and fit a line to it in log-log space, we can use the measured slope and intercept to derive r_0. We can do this by integrating Equation 5.16 with $r_0 = 1$ and γ as measured from our $w(\theta)$; the value of A is then independent of r_0. Taking the ratio of $A(r_0 = 1)$ with the A derived from the $w(\theta)$ fit yields the desired value of r_0:

$$r_0 = \left[\frac{A}{A(r_0 = 1)}\right]^{\frac{1}{\gamma}} \tag{5.17}$$

Thus, by using Limber's equation we can relate the correlation scale length r_0 to the clustering amplitude of $w(\theta)$.

5.3 SAMPLE SELECTION

For this thesis we created two samples from the co-added imaging data. The first sample is an apparent magnitude limited survey consisting of all galaxies with magnitudes $16 \leq r \leq 21$. This sample is easy to define but difficult to interpret physically because all galaxies are grouped together simply by how bright they are on the sky; hence galaxies within such a

sample do not share any intrinsic properties. In particular, this selection leads to inherent bias at the faint end of the sample because there is a tendency to select galaxies that are intrinsically more luminous. The clustering in an apparent magnitude sample is therefore affected by the limitations of the survey (*i.e.* the fact that we cannot resolve every galaxy in a given patch of sky). For this reason, we only used the magnitude limited sample as a check for systematic errors in our calibration.

The second sample we created is a volume limited sample which removes the incompleteness in our galaxy selection. A volume limited sample is defined so that the number of galaxies per redshift per comoving volume element is constant. The limits for the volume limited sample can be determined by plotting redshift vs absolute magnitude and finding a rectangular region that lies entirely within the resulting curve. As Figure 5.1 shows, we define our volume limited sample with cuts of $0.1 \leq z \leq 0.3$ and $-23 \leq M_r \leq -20$. We initially hoped to probe higher redshifts, but we would require an apparent magnitude cut of $r = 22$ to reach $z = 0.4$; photometric redshift errors forced us to use the shallower $r = 21$ cut, reducing both our number counts and upper redshift limit.

Because the photometric redshifts have an associated error, the absolute magnitudes computed using photozs are also inherently noisy. Thus, selecting a sample using a hard cut in both photometric redshift and absolute magnitude does not yield a true volume limited sample. As such, we must model the contamination in both redshift and absolute magnitude in order to compute the true redshift distribution, a topic covered in detail in the following section.

5.4 ESTIMATING THE REDSHIFT DISTRIBUTION

Limber's equation requires an estimate of $\frac{\mathrm{d}n}{\mathrm{d}z}$, but the inherent scatter in photometric redshifts makes estimating the true redshift distribution difficult. We previously mentioned three methods for estimating $\frac{\mathrm{d}n}{\mathrm{d}z}$: computing a histogram of the photoz values, summing $p(z)$ for each galaxy, and computing the median redshift and using Equation 4.9. In this section we outline another method following Budavári et al. (2003) for computing the redshift

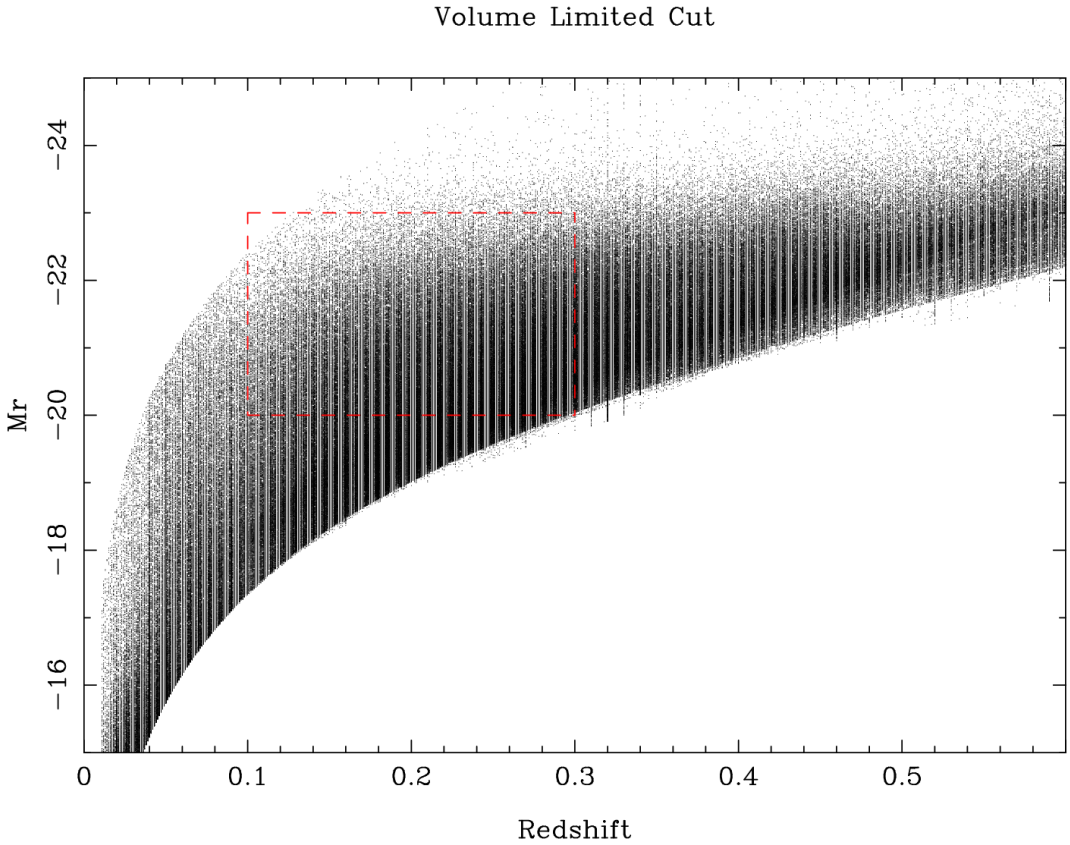


Figure 5.1 Selection used for the volume limited sample. The dashed red line indicates the cuts we used to define the sample. The white stripes in the sample indicate that BPZ does not produce a continuous distribution in z for our templates.

distribution which utilizes the luminosity function and estimations of contamination in the redshift and luminosity cuts.

To relate $\frac{\mathrm{d}n}{\mathrm{d}z}$ to the luminosity function, the distribution of absolute magnitudes, we must first define a few basics. Absolute magnitude is the magnitude an object would have if it were located 10 parsecs away. For a flat universe, the absolute magnitude M is

$$M = m - 5\log\left[\frac{r(z)(1+z)}{10\,\mathrm{pc}}\right] - K(z,T) \tag{5.18}$$

where $r(z)$ is the comoving line of sight distance and K is the K-correction which depends on redshift and type T. The K-correction adjusts magnitudes to how they would be measured in the object's rest frame (Hogg et al., 2002):

$$K = -2.5\log\left[\left(\frac{1}{1+z}\right)\left(\frac{\int \lambda F(\frac{\lambda}{1+z})R(\lambda)\,\mathrm{d}\lambda}{\int \lambda F(\lambda)R(\lambda)\,\mathrm{d}\lambda}\right)\right] \tag{5.19}$$

Here $R(\lambda)$ is the filter transmission curve used to measure the apparent magnitude, which is roughly a Gaussian shape, and $F(\lambda)$ is the flux in units of energy/length/sec^2/wavelength. To compute the K-correction, one needs an estimate of the object's flux as a function of wavelength λ. Because we estimated the object's type when computing the photometric redshift, we can simply use the spectrum for that template when computing K-corrections. For interpolated types, we normalized the spectra so that $\int F(\lambda)\lambda\,\mathrm{d}\lambda = 1$ and then took linear combinations of neighboring types. For example, for the type 1.66, we used $\frac{1}{3}$ of type 1 and $\frac{2}{3}$ of type 2 after normalizing both.

The distribution of absolute magnitudes is known as the luminosity function, $\phi(M)$. The luminosity function is similar to $\frac{\mathrm{d}n}{\mathrm{d}z}$ in that it is often used to determine the selection function; hence it is an important fundamental quantity that has been measured for many surveys. The luminosity function is often parameterized using a Schechter function (Lin et al., 1999, *e.g.*):

$$\phi(M) = 0.4\ln(10)\phi_* 10^{0.4Pz}\left[10^{0.4(M_*(z)-M)}\right]^{\alpha+1}\exp\left[-10^{0.4(M_*(z)-M)}\right] \tag{5.20}$$

$$M_*(z) = M_*(0) - Qz \tag{5.21}$$

The luminosity function is specified with the set of parameters $\{\phi_*,\ \alpha,\ M_*(0),\ Q,\ P\}$, with the last two parameters estimating density and luminosity evolution with redshift respectively. The luminosity function for SDSS has been measured using spectroscopic data (with a median redshift of 0.1) by Blanton et al. (2003); we use their fit for the r band luminosity function: $\phi_* = 1.49\,10^2 h^3$, $\alpha = -1.05$, $M_*(z = 0.1) = -20.44 + 5\log h$, $Q = 1.62$, $P = 0.18$. Here h is related to Hubble's constant through $h = \frac{H_0}{100}$[3]; we use $h = 0.71$. Note that these values are fits for $z = 0.1$, so we use $z' = z - 0.1$ in 5.20.

Given an expression for the luminosity function and an estimate of the K-corrections, we can relate the redshift distribution to the luminosity function (Dodelson et al., 2002, *e.g.*):

$$\frac{\mathrm{d}n}{\mathrm{d}z} \propto \frac{r^2(z)\frac{\mathrm{d}r}{\mathrm{d}z}}{(1+z)^3} \int_{M_{\min}(z)}^{M_{\max}(z)} \phi(M)\,\mathrm{d}M \tag{5.22}$$

Again, $r(z)$ is the comoving line of sight distance and $\frac{\mathrm{d}r}{\mathrm{d}z}$ is given by Equation 5.9. Note that the integration limits are a function of redshift. In an apparent magnitude limited survey, the limits are given by a range in apparent magnitude, and as we probe different redshifts, the limits in absolute magnitude change according to Equation 5.18. That is, the range of the luminosity function that contributes to $\frac{\mathrm{d}n}{\mathrm{d}z}$ varies with redshift.

Finally, we note that the integration limits depend on the type of galaxy used for the K-correction in Equation 5.18. For this reason, we compute three integrals in Equation 5.22 for early, spiral, and irregular galaxies with BPZ spectral types 1.0, 2.33[4], and 4.0 respectively. We then weight each integral according to the expected fractions we determined for the photoz prior in Equation 4.4 using the average apparent magnitude (*i.e.* $\frac{1}{2}(m_{\min} + m_{\max})$). For a large range in apparent magnitude, this is not a good approximation, but we can break up the $\frac{\mathrm{d}n}{\mathrm{d}z}$ calculation into small bins in apparent magnitude and average them according to the number of objects in each bin.

For a volume limited survey, the limits in Equation 5.22 are fixed. Similarly, the non-zero range of $\frac{\mathrm{d}n}{\mathrm{d}z}$ is fixed. However, because we are using photometric redshifts, these cuts will have contamination from objects with redshifts outside the desired range due to errors in the redshift estimation. Because the photozs are estimated from apparent magnitudes, we expect

[3]This odd choice of parameterization is due to historical uncertainty in H_0.
[4]We used an interpolated type here because there are 2 spiral templates.

there to be a strong correlation between the contamination in the volume limited survey and apparent magnitude. For this reason, we will estimate the contamination as a function of apparent magnitude by taking small (width = 0.2) bins in apparent magnitude and estimating the contamination in both z and M_r in each bin. We then use these contamination estimates to compute $\frac{\mathrm{d}n}{\mathrm{d}z}$ using the method outlined below.

We first approximate the average photometric redshift error in a particular magnitude bin. The error of an individual object is not Gaussian, but the average of the errors in the bin should be roughly Gaussian because of the central limit theorem. The probability of obtaining photometric redshift z_p given the true redshift z is then

$$p(z_p|z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(z_p - z)^2}{2\sigma^2}\right] \tag{5.23}$$

We actually want to compute the reverse, $p(z|z_p)$, which we can obtain using Bayes' Rule:

$$p(z|z_p) = \frac{p(z)p(z_p|z)}{p(z_p)} \tag{5.24}$$

We then want to select a set of redshifts using a window function $W(z)$ that is a simple step function:

$$W(z) = \begin{cases} 1 & 0.1 \leq z \leq 0.3 \\ 0 & \text{otherwise} \end{cases} \tag{5.25}$$

The probability that a galaxy is selected by this window function (in this particular bin) is

$$\begin{aligned} p(z|W) &\propto \int p(z_p)W(z_p)p(z|z_p)\,\mathrm{d}z_p \\ &\propto p(z)\int W(z_p)p(z_p|z)\,\mathrm{d}z_p \\ &\propto p(z)W_{\mathrm{eff}}(z) \end{aligned} \tag{5.26}$$

where $W_{\mathrm{eff}}(z)$, the effective window function, is the window function convolved with the uncertainty in photometric redshift. We estimate the true redshift distribution $p(z)$ using

the luminosity function and Equation 5.22. However, we will see that there is an effective luminosity function that must be used rather than Equation 5.20. Finally, we note that $p(z|W)$ is not normalized because of the window function, so it must be normalized in each individual magnitude bin.

The errors in M follow a similar derivation. We estimate the average error in absolute magnitudes computed with photozs M_{z_p} given the true absolute magnitude M as a Gaussian (central limit theorem):

$$p(M_{z_p}|M) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(M_{z_p} - M)^2}{2\sigma^2}\right] \tag{5.27}$$

Then, using an analogous window function $W(M)$ that is 1 for $-23 \leq M \leq -20$ and 0 otherwise, we compute the probability of selecting an object with absolute magnitude M:

$$\begin{aligned} p(M|W) &\propto \int p(M_{z_p}) W(M_{z_p}) p(M|M_{z_p})\, \mathrm{d}M_{z_p} \\ &\propto p(M) \int W(M_{z_p}) p(M_{z_p}|M)\, \mathrm{d}M_{z_p} \\ &\propto p(M) W_{\mathrm{eff}}(M) \end{aligned} \tag{5.28}$$

$$p(M) = \frac{\phi(M)}{\int \phi(M)\, \mathrm{d}M} \tag{5.29}$$

Thus, we see that $p(M|W)$ defines an effective luminosity function $\phi_{\mathrm{eff}}(M) = \phi(M) W_{\mathrm{eff}}(M)$ for this apparent magnitude bin. It is this $\phi_{\mathrm{eff}}(M)$ that we integrate to obtain $p(z)$ in Equation 5.27. As a final note on the luminosity function, to evaluate $\phi(M)$ we need a redshift value, so we use the average redshift for this magnitude bin.

Now it only remains to estimate the average errors for z_p and M_{z_p}. Ideally, we could simply compare to spectroscopic data and average the true errors. Unfortunately, the fainter magnitude bins do not have enough points to allow for a good average error estimate when comparing to SDSS spectroscopic data, so we instead estimate the errors using the $p(z_p)$ fits

in Equation 4.8. We compared our estimates to those from the true errors in the bright bins to verify our estimates. The redshift errors were computed as

$$< z >= \int zp(z)\,\mathrm{d}z \tag{5.30}$$

$$< z^2 >= \int z^2 p(z)\,\mathrm{d}z \tag{5.31}$$

$$\sigma_z =< z^2 > - < z >^2 \tag{5.32}$$

Comparison to the true errors revealed that we underestimate the average photoz error by roughly a factor of 2, so we used $2\sigma_z$ as our redshift error estimate (see Figure 5.2). To compute the errors in M_r, we generated random points according to the Gaussian distribution we fit to $p(z)$[5] and then computed σ_{M_r} of this distribution. While the z error estimate was underestimated, our absolute magnitude error estimation was approximately correct. Both error estimates displayed the expected behavior – monotonic increase of both σ values with apparent magnitude. The average error in both z and M_r are plotted for 6 representative bins in figures 5.2 and 5.3 respectively.

Once we have estimated the contamination, we can then compute a normalized $p(z|W)$ for every magnitude bin and weight each bin by the number of objects it contains. The weighted sum then gives our estimate of $\frac{\mathrm{d}n}{\mathrm{d}z}$. Because of the exponential nature of galaxy number counts, the contamination will always be dominated by the faint end which contains more galaxies than the previous bins and also has the largest average error.

Figure 5.4 compares our new method of estimating $\frac{\mathrm{d}n}{\mathrm{d}z}$ to our previous methods of summing individual $p(z)$ distributions and fitting the median redshift. As in Figure 4.5, adding an absolute magnitude cut improves the agreement between the various estimations. For the volume limited cut, though, the large discrepancy in redshift distributions justifies our use of the $\frac{\mathrm{d}n}{\mathrm{d}z}$ estimate outlined above.

Finally, we show $\frac{\mathrm{d}n}{\mathrm{d}z}$ for bins in apparent magnitude (Figure 5.5), absolute magnitude (Figure 5.6), type (Figure 5.7), and redshift (Figure 5.8) as computed using the method

[5]This was only done for speed, and we could have simply generated according to $p(z)$.

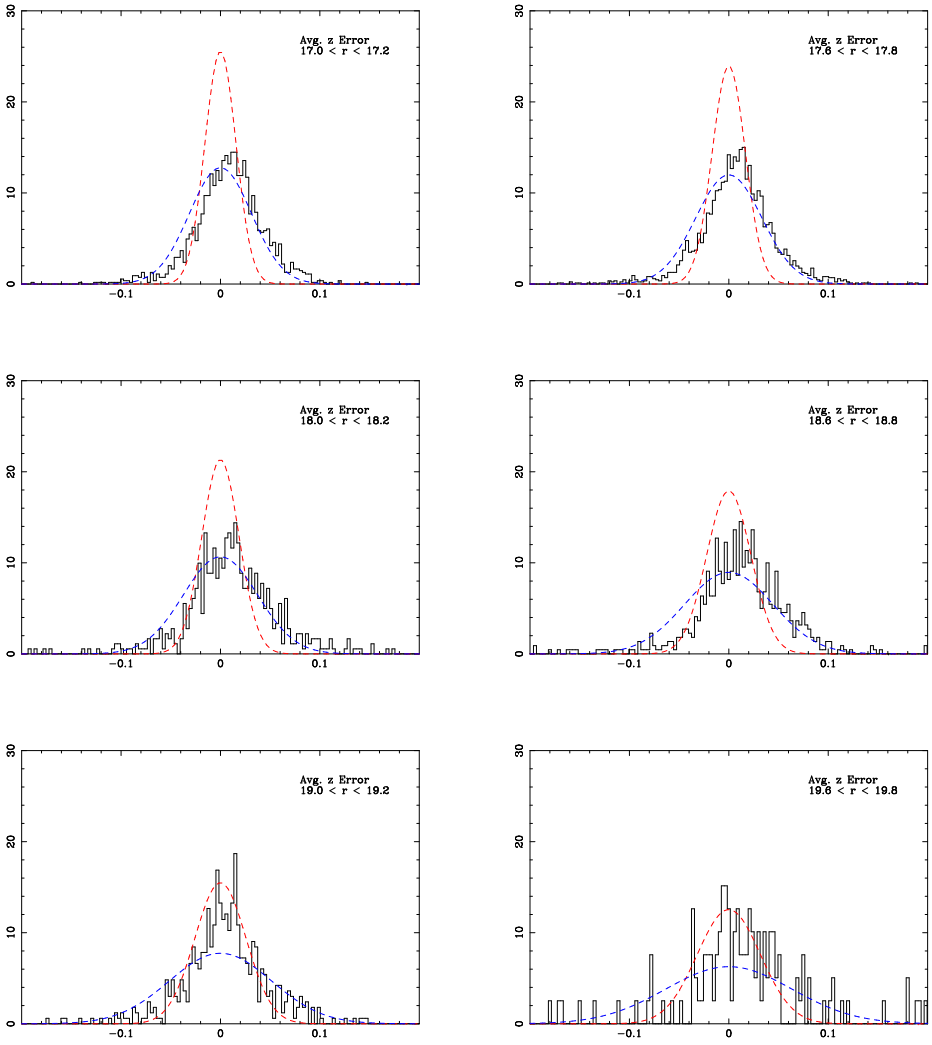


Figure 5.2 Average redshift error as a function of r band apparent magnitude. The histogram is taken from comparison with spectroscopic redshifts from SDSS, CNOC2, and DEEP2. The red line is our Gaussian estimate from photometric redshifts, and the blue line is the same but with twice the estimated σ for comparison.

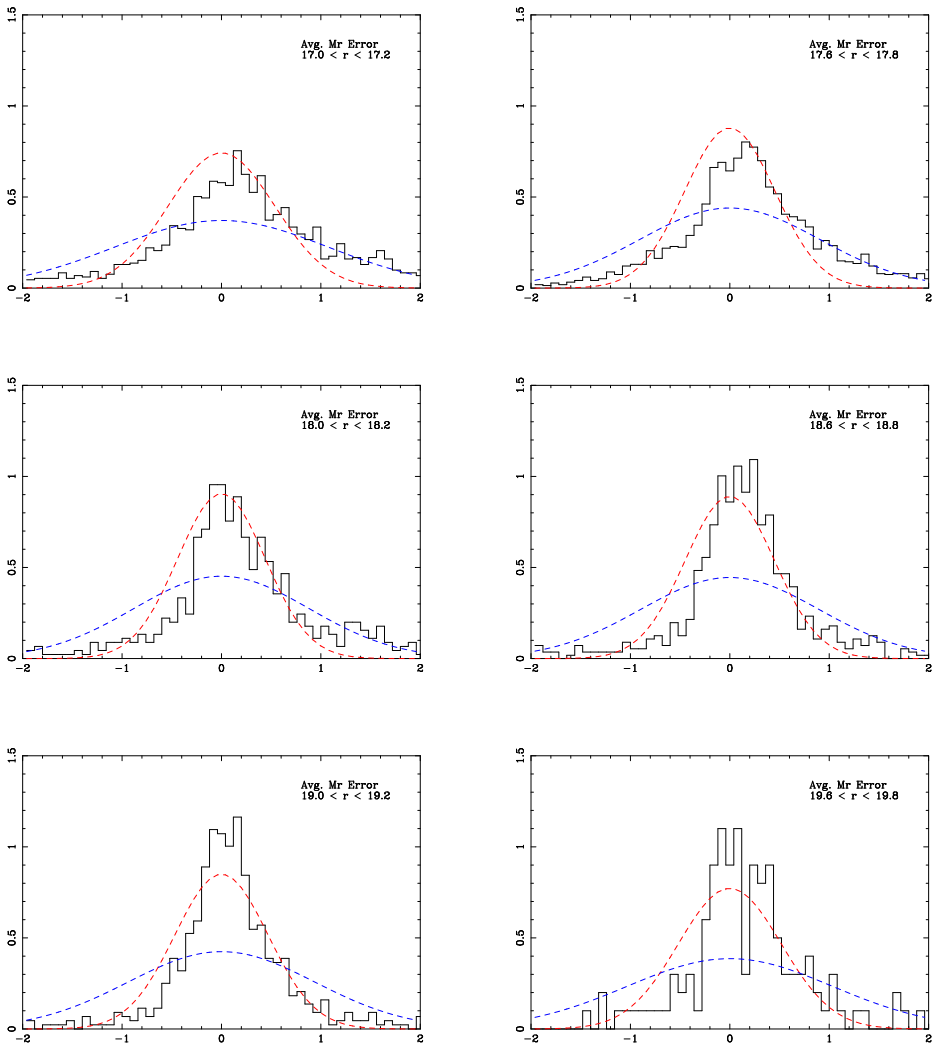


Figure 5.3 Average r band absolute magnitude error as a function of r band apparent magnitude. The histogram is taken from comparison with spectroscopic redshifts from SDSS, CNOC2, and DEEP2. The red line is our Gaussian estimate from photometric redshifts, and the blue line is the same but with twice the estimated σ for comparison.

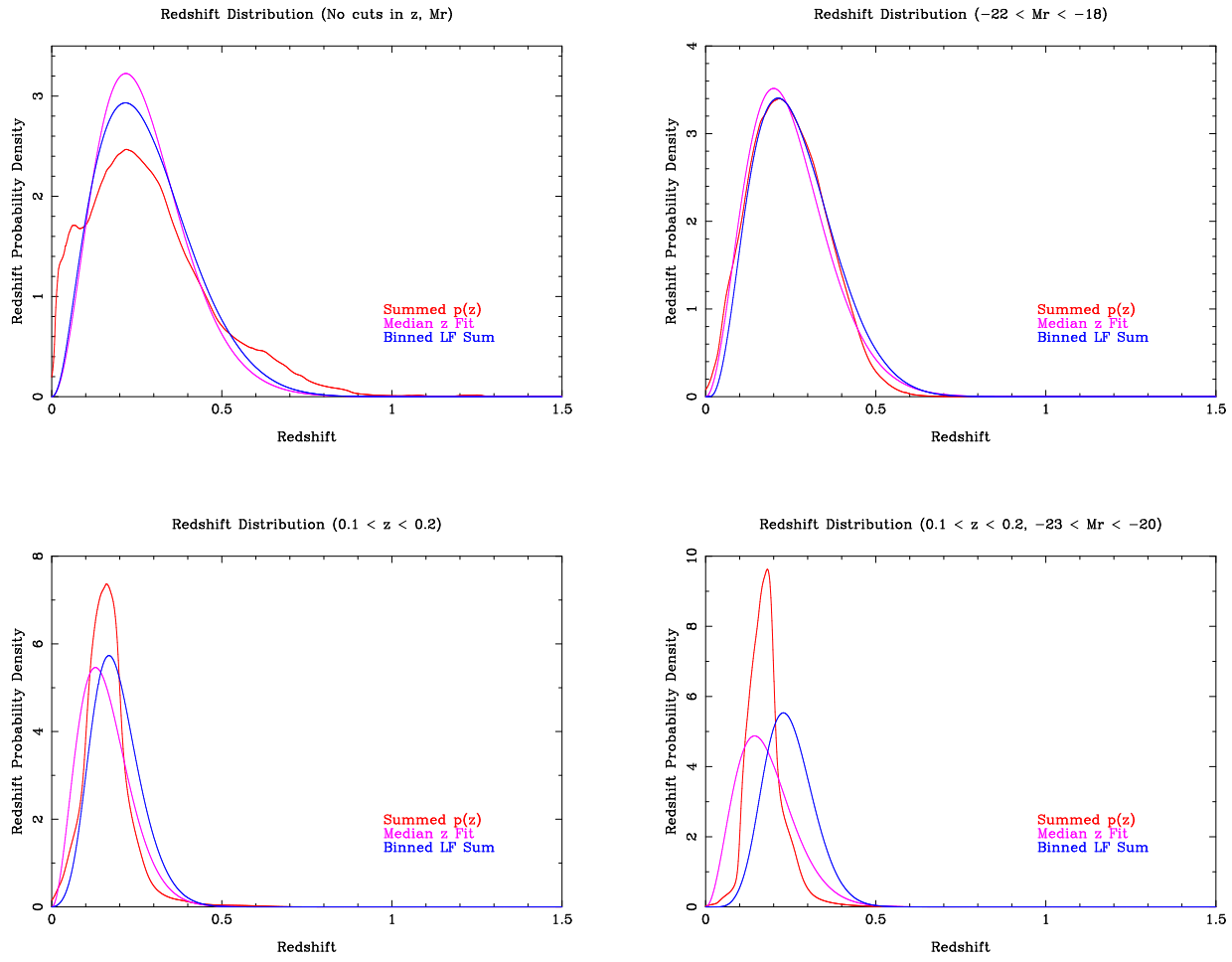


Figure 5.4 Comparison of different $\frac{\mathrm{d}n}{\mathrm{d}z}$ estimations. The methods are very similar for the cuts in the top 2 panels. The methods are the most consistent for the case with an absolute magnitude cut (upper right), in agreement with the results in Figure 4.5. The discrepancy in the volume limited cut (bottom right) is partially because of the z cut (see the lower left panel) and partially because of the different M_r cut.

outlined in this section. For all calculations requiring $\frac{dn}{dz}$, we used the technique described in this section.

5.5 ANGULAR CORRELATION RESULTS

The evolution of $w(\theta)$ with r band apparent magnitude is shown in Figure 5.9 with the corresponding fit parameters in Table 5.1. The intercept decreases from -1.7 to -2.5 and the slope decreases from -0.72 in the two brightest bins to -0.8 in the two faintest bins. Thus, the evolution is primarily in A with γ remaining approximately constant. These results are consistent with the scaling relation (Peebles, 1973) that results from Limber's equation. Given two samples that only differ in depth D, the selection function $\phi(r, D)$ in Equation 5.6 must depend only on the ratio $\frac{r}{D}$. It can be shown in this case that (Fall, 1979)

$$w(\theta) = \left(\frac{1}{D}\right) F(\theta D) \tag{5.33}$$

where F is a function which is determined by $\xi(r)$ but depends on θ only through θD. At a fixed physical scale θD, the clustering strength decreases inversely with depth D because the number of uncorrelated galaxies along the line of sight is proportional to D. Hence, projection effects appear to "smear out" the clustering signal due to increased numbers of galaxies along the line of sight.

Figure 5.10 compares our $w(\theta)$ measurement against that of Connolly et al. (2002) which was measured from the Early Data Release (EDR) of SDSS. The EDR had magnitude zero point calibration issues which result in fit intercepts inconsistent with our measurement, though the slopes are consistent (≈ -0.7) in the brightest two bins. In the two fainter bins, our measured slopes steepen to ≈ -0.8 whereas the EDR slopes remain near -0.7. The change in slope is clearly related to increased deviation from a pure power law – the fit becomes a strong function of the limits in θ used for these bins. The deviation from a power law is most likely a result of unknown systematics in our magnitude calibration and/or star-galaxy classification.

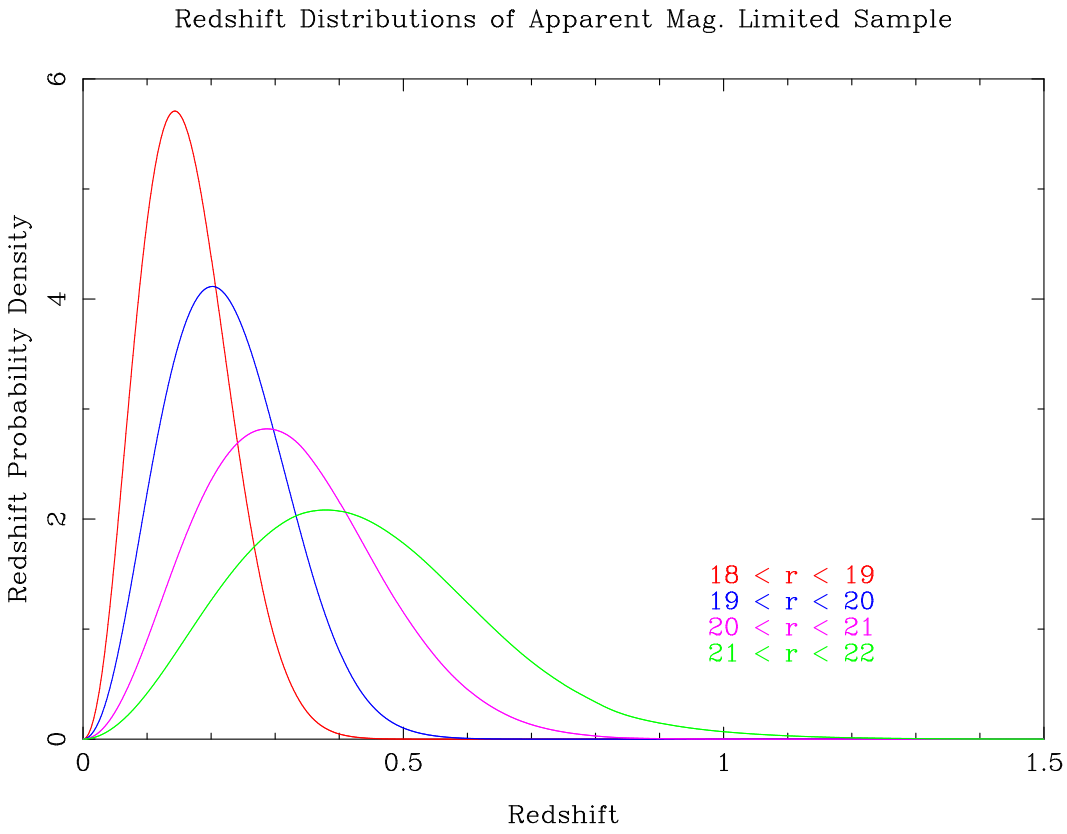


Figure 5.5 Redshift distributions for each bin in the apparent magnitude limited sample. Here $\frac{dn}{dz}$ is computed by summing the effective luminosity function in apparent magnitude bins.

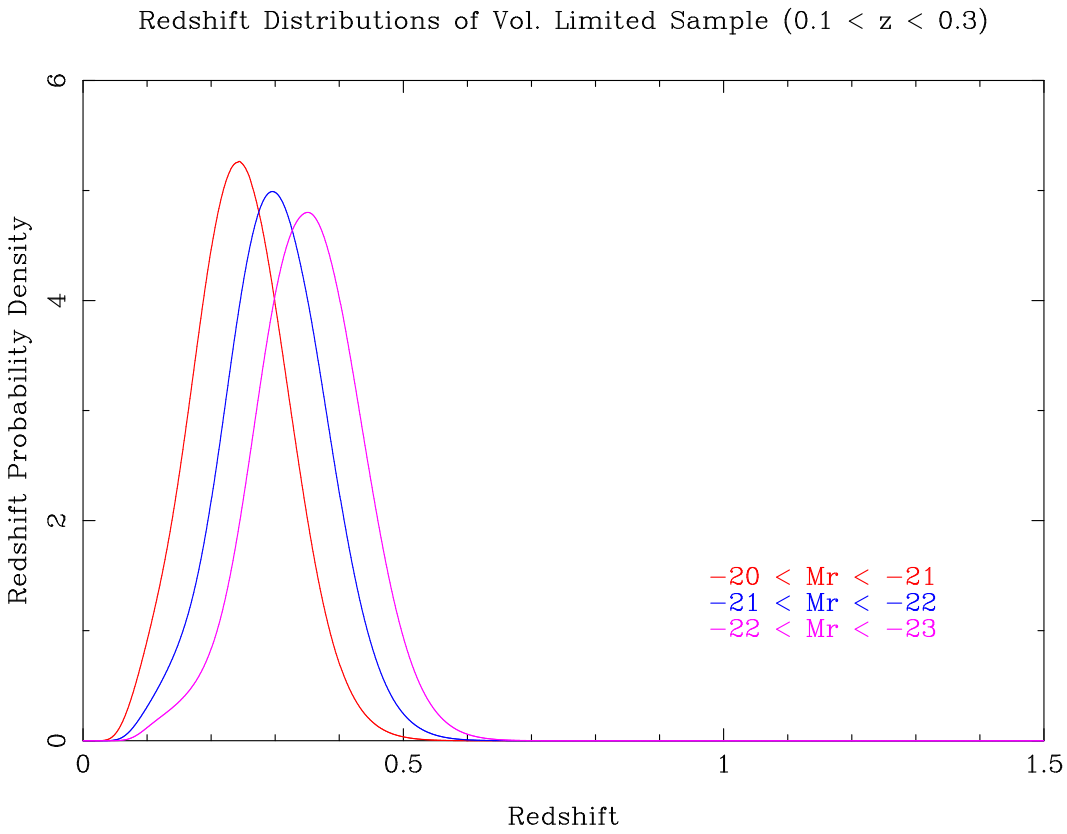


Figure 5.6 Redshift distributions for 3 absolute magnitude bins in the volume limited sample. Here $\frac{\mathrm{d}n}{\mathrm{d}z}$ is computed by summing the effective luminosity function in apparent magnitude bins.

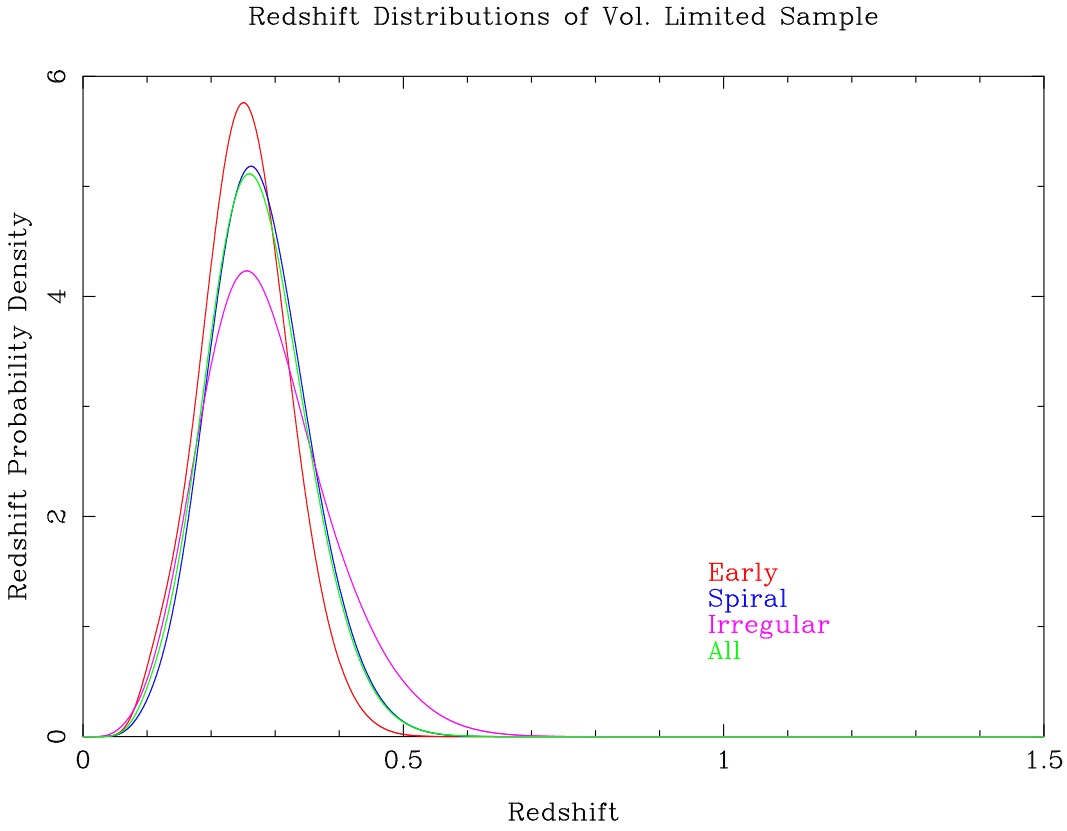


Figure 5.7 Redshift distributions for the 3 prior types. Here $\frac{\mathrm{d}n}{\mathrm{d}z}$ is computed by summing the effective luminosity function in apparent magnitude bins. The $\frac{\mathrm{d}n}{\mathrm{d}z}$ of all galaxy types is shown for comparison; spirals clearly dominate the redshift distribution.

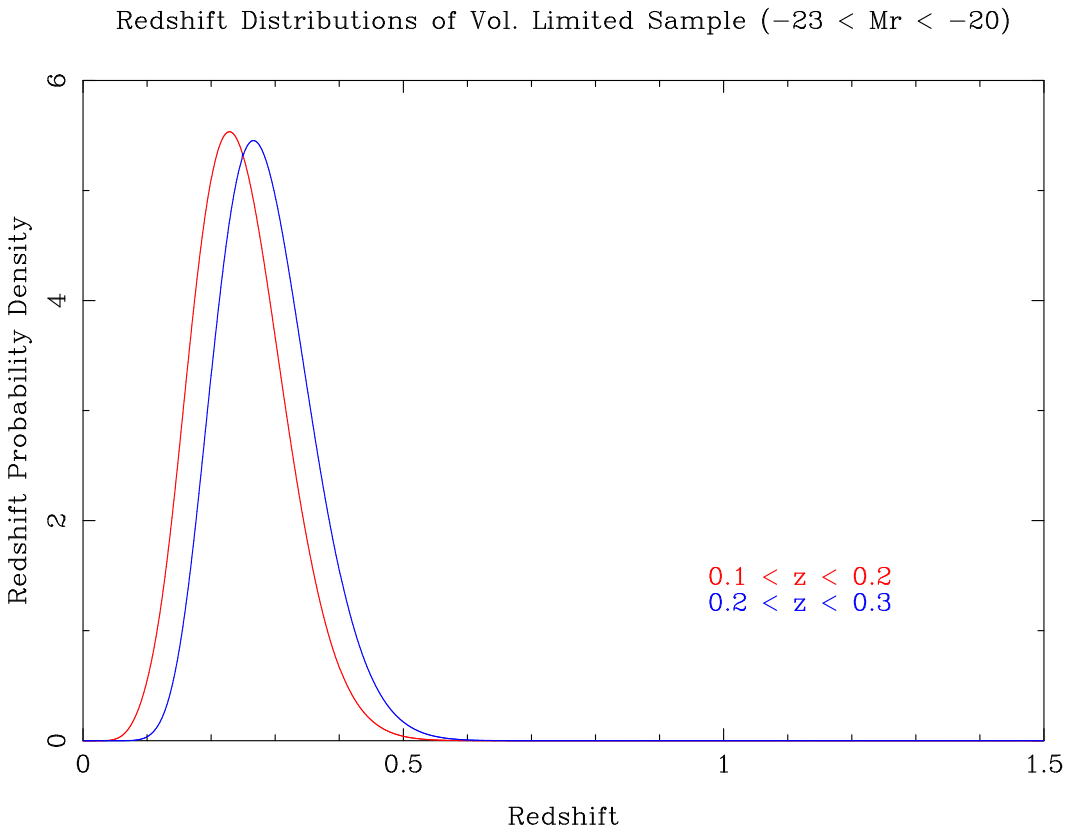


Figure 5.8 Redshift distributions for 2 redshift bins in the volume limited sample. Here $\frac{dn}{dz}$ is computed by summing the effective luminosity function in apparent magnitude bins.

We performed a second measurement of evolution in apparent magnitude bins, this time using the i band, in Figure 5.11 with fit parameters given in Table 5.2. This measurement shows a similar trend in A to that of the r band with the intercept decreasing with apparent magnitude from -1.86 to -2.41; however, the slope actually *increases* from -0.78 to -0.61, displaying even more power law deviation than the r band. We made this measurement for direct comparison to the results of DEEP2 (Coil et al., 2004), which is presented in Figure 5.12. The differences in this comparison stem from the fact that the DEEP2 galaxy population is very different from that in the SDSS sample. The DEEP2 survey team selected their objects with color cuts designed to predominantly select high redshift (> 0.7) galaxies; thus, the galaxies in the DEEP2 are high redshift galaxies with a particular range in color space rather than the set of all observable galaxies. Hence, we expect the clustering measurements to differ. In particular, we expect higher redshift galaxies to be less strongly clustered than our sample, which is exactly what is observed in Figure 5.12.

Figure 5.13 shows the evolution of $w(\theta)$ with luminosity in the volume limited sample with fit parameters in Table 5.3. Here the intercept decreases from -1.74 to -1.77 and the slope increases from -0.92 to -0.68. Here we see that selecting in absolute magnitude changes the slope of the sample appreciably; this is the expected result, as we expect type to be correlated with absolute magnitude. Despite this, though, our result still agrees well with the results of Budavári et al. (2003) who found that there is marginal evolution in γ with luminosity; this is shown in Figure 5.14. This comparison demonstrates the best agreement of any comparison we made, and it should – both samples are taken from SDSS data and use nearly identical volume limited cuts. The agreement is a nice result, though, as it shows that our imaging pipeline and photometric redshifts are consistent with those of the main SDSS sample. The potential advantage of our sample, increased photometric accuracy, seems to be outweighed by the larger number counts of Budavári et al. (2003) who use roughly a factor of 8 more galaxies.

We present the evolution of $w(\theta)$ with type in Figure 5.15 with fit parameters in Table 5.4. The fits here demonstrate a decrease in intercept from -1.60 to -2.06 and an increase in slope of -0.85 to -0.77, consistent with Budavári et al. (2003) who found that both γ and A change with type. We used the broader definition of type from the BPZ prior consisting

of only early/elliptical ($t < 1.5$), spiral ($1.5 < t < 3.5$), and irregular ($t > 3.5$) galaxies. Unfortunately even with this broad grouping, we did not have enough irregular galaxies to estimate $w(\theta)$ reliably, so we omitted them from the figure and table. As expected, the early galaxies (which have higher quality photometric redshifts) display a smooth correlation function. We again compare to the results of Budavári et al. (2003) in Figure 5.16 and find rough agreement. This comparison should be taken qualitatively, though, as we use different luminosity cuts and different photometric redshift templates. For the comparison, we used the $t < 0.02$ values for ellipticals and the $0.3 < t < 0.65$ values for spirals from Budavári et al. (2003).

Budavári et al. (2003) and Zehavi et al. (2002) found that γ does not vary appreciably with luminosity even though it does with type, but we found evidence that γ does vary with luminosity. To explain their discrepancy, Budavári et al. (2003) proposed a simple bimodal model for the galaxy population consisting of red and blue galaxies. In this model, luminosity cuts brighter than $M_{r*} \approx -20$ do not select objects by type due to similar luminosity functions for red and blue galaxies (Baldry et al., 2004). Hence, the shape of the correlation function should be roughly the same, as is shown in the measurement. An alternative explanation, and one which seems more probable, is that photometric redshifts have errors in M_r which introduce a wider variety of galaxy types into a particular absolute magnitude bin than should be present. This "mixing" of galaxy types obscures the type selection one would normally expect with absolute magnitude cuts; the $w(\theta)$ measurement is then averaged over a larger spread of types and the type evolution is lost.

Figure 5.17 shows the evolution of $w(\theta)$ with redshift within our volume limited sample with fit parameters in Table 5.5. Once again, the evolution primarily occurs in A with the intercept decreasing from -1.53 to -1.78 and the slope changing from -0.796 to -0.804. Because the shape of $w(\theta)$ depends only upon A and does not change with redshift, the bimodal galaxy population model suggests that the relative mix of red and blue galaxies is roughly constant over our redshift interval $0.1 \leq z \leq 0.3$. Thus, we would expect the evolution of galaxies over this interval to be minimal. This seems unlikely with such a large redshift interval, giving further doubt to the bimodal population model.

We also show the evolution of the clustering length r_0 with apparent magnitude (Figure

5.18), absolute magnitude (Figure 5.19), galaxy type (Figure 5.21), and redshift (Figure 5.22). The errors in r_0 presented are derived solely from the associated linear fit errors[6]. In general, our estimates for r_0 are systematically higher than previous surveys (Hudon and Lilly, 1996, *e.g.*) which previously found $3.8 < r_0 < 5.4$ in units Mpc h^{-1}. For our volume limited absolute magnitude evolution, we were able to directly compare to the results of Budavári et al. (2003) (Figure 5.19, red line). Here we are again systematically higher; the inconsistency is greater than the expected errors, implying an additional source of error. This is almost certainly due to differences in the redshift distribution used to estimate r_0, which is highly sensitive to the width of $\frac{dn}{dz}$. Unfortunately it is impossible to obtain a reliable estimate of the error in r_0 due to $\frac{dn}{dz}$ because the error in contamination for the faintest bins cannot be determined due to insufficient spectroscopic information. Because of the nature of galaxy number counts, the faintest bins undoubtedly contribute the most to the variation in the redshift distribution. To quantify the necessary changes in $\frac{dn}{dz}$ for the luminosity binned samples, we varied the average photometric redshift error in each apparent magnitude bin by a constant factor; adjusting the contamination in each bin effectively narrows or widens the distribution. We find that this simple model can successfully account for the discrepancy in two of the three luminosity bins using a factor of 0.5 as shown by the blue line in Figure 5.19. This implies our average photoz error estimation is a factor of 4 too large since we originally used a factor of 2. It is enlightening to see what change this causes in the redshift distribution, so we show this in Figure 5.20; the solid line gives our original $\frac{dn}{dz}$ estimate and the dashed line our inferred $\frac{dn}{dz}$ chosen to match the values of r_0. From this plot it is evident that the faintest bin $-20 < M_r < -21$ is significantly less peaked, implying a larger r_0 value consistent with the final point in Figure 5.19. Comparing this estimated $\frac{dn}{dz}$ to our other techniques for estimating the redshift distribution shown in Figure 5.4 reveals that the summed $p(z)$ may be a more accurate estimation of the redshift distribution than the contamination approach, though clearly a more substantial analysis is needed. All of this demonstrates a fundamental difficulty in de-projecting $w(\theta)$ to obtain $\xi(r)$ using photometric redshifts – there is a potentially significant source of error from the shape of $\frac{dn}{dz}$ which cannot

[6]The fit errors dominate over the errors from the luminosity function parameters; see (Budavári et al., 2003) for typical error values from the luminosity function.

be estimated without spectroscopic redshifts.

For spectroscopic redshift samples, there appears to be disagreement in the literature as to the effect uncertainty in $\frac{\mathrm{d}n}{\mathrm{d}z}$ has on the measurement of r_0. Hudon and Lilly (1996) found that the errors in r_0 due to $\frac{\mathrm{d}n}{\mathrm{d}z}$ were roughly 3%, much smaller than the other sources of error. On the other hand, Coil et al. (2004) note with some concern the variety of methods used to estimate the redshift distribution for spectroscopic samples – for example, some use $\frac{\mathrm{d}n}{\mathrm{d}z} \propto z^2 \exp\left(-\frac{z}{z_0}\right)$, some use $\frac{\mathrm{d}n}{\mathrm{d}z} \propto z^2 \exp\left[\left(-\frac{z}{z_0}\right)^2\right]$, and Hudon and Lilly (1996) used the "$\frac{1}{V_{\max}}$ formalism". Thus it seems possible that the errors due to the redshift distribution have not been fully addressed even for spectroscopic redshift samples. For photometric redshift samples, though, the uncertainty in the redshift distribution is clearly a problem which deserves more exploration. A recent paper by Newman (2008) proposes a novel technique for measuring $\frac{\mathrm{d}n}{\mathrm{d}z}$ which we were unable to test due to time constraints. Finally, we note that it would be worthwhile to develop an unbiased estimator for the true median redshift given a photometric redshift distribution (perhaps through comparison to Newman's technique) so that Equation 4.9 could be easily applied to photometric redshift distributions.

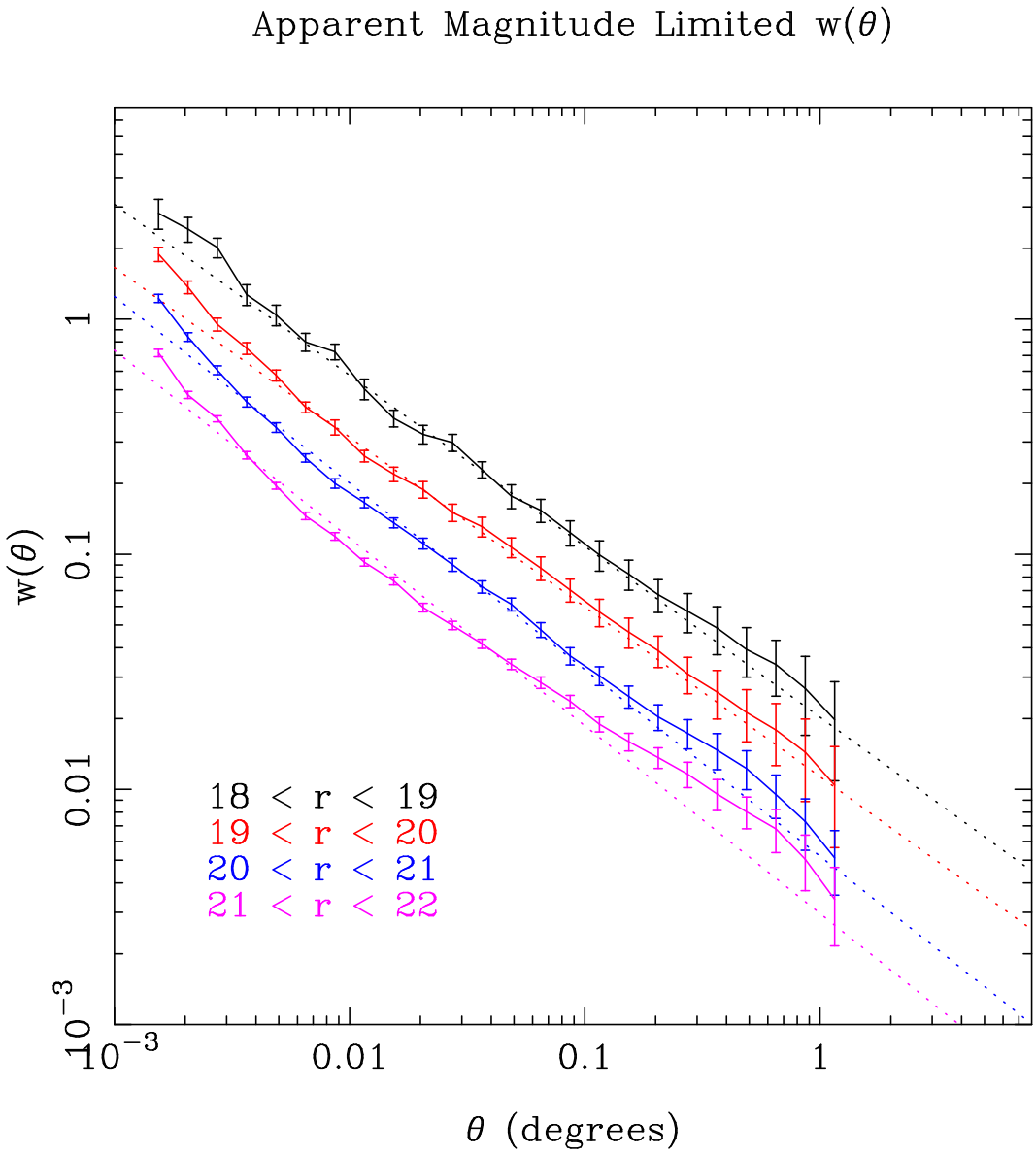


Figure 5.9 Evolution of $w(\theta)$ with r band apparent magnitude. The sample used was apparent magnitude limited. The fit was performed using all points with $\theta < 0.1°$. The fit parameters are given in Table 5.1.

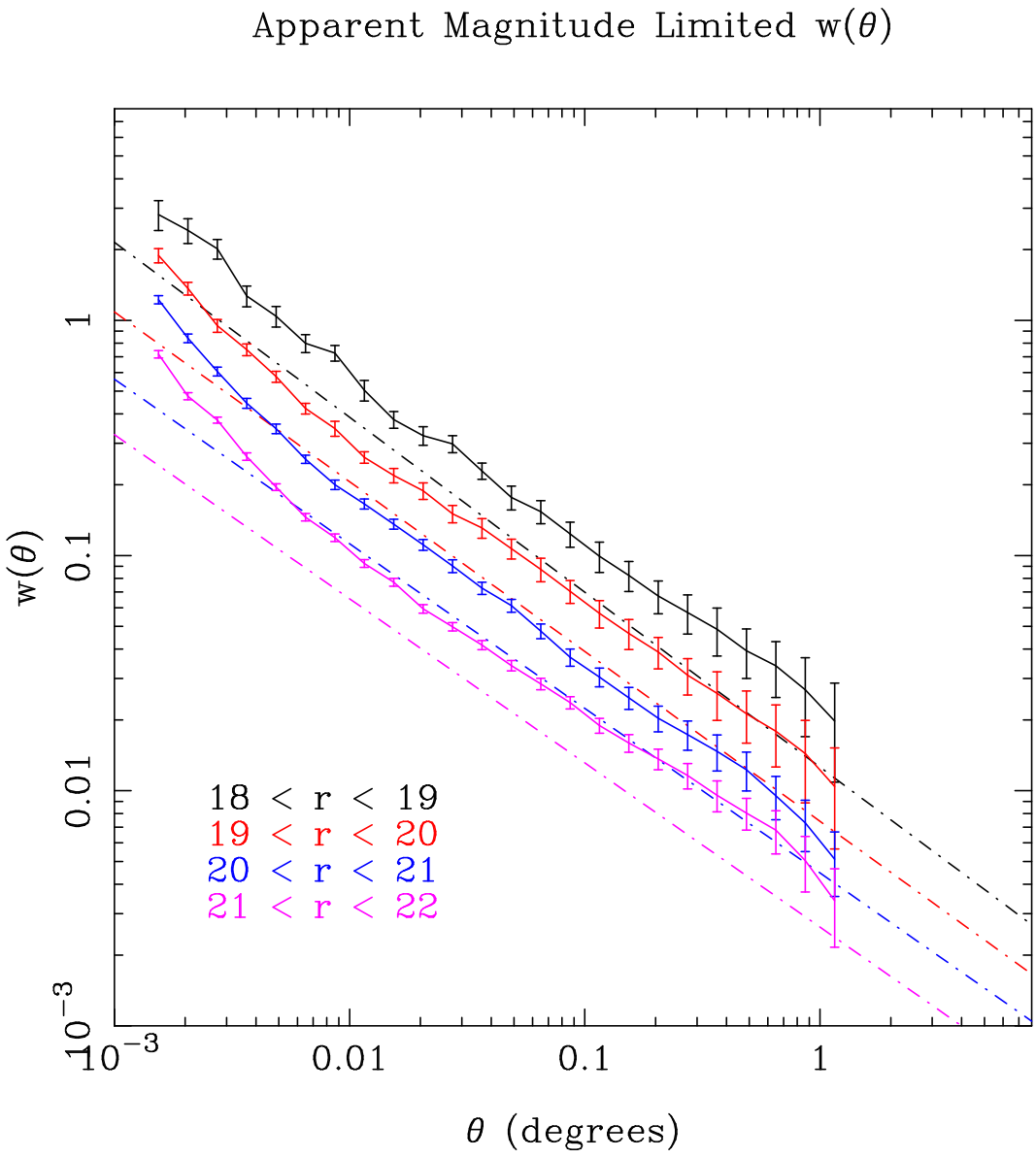


Figure 5.10 Comparison of $w(\theta)$ evolution in apparent magnitude bins to the results of Connolly et al. (2002) (the dash-dot lines). The discrepancies are due to calibration issues in magnitude zero points for the SDSS EDR (Early Data Release). See also Figure 5.9 and Table 5.1.

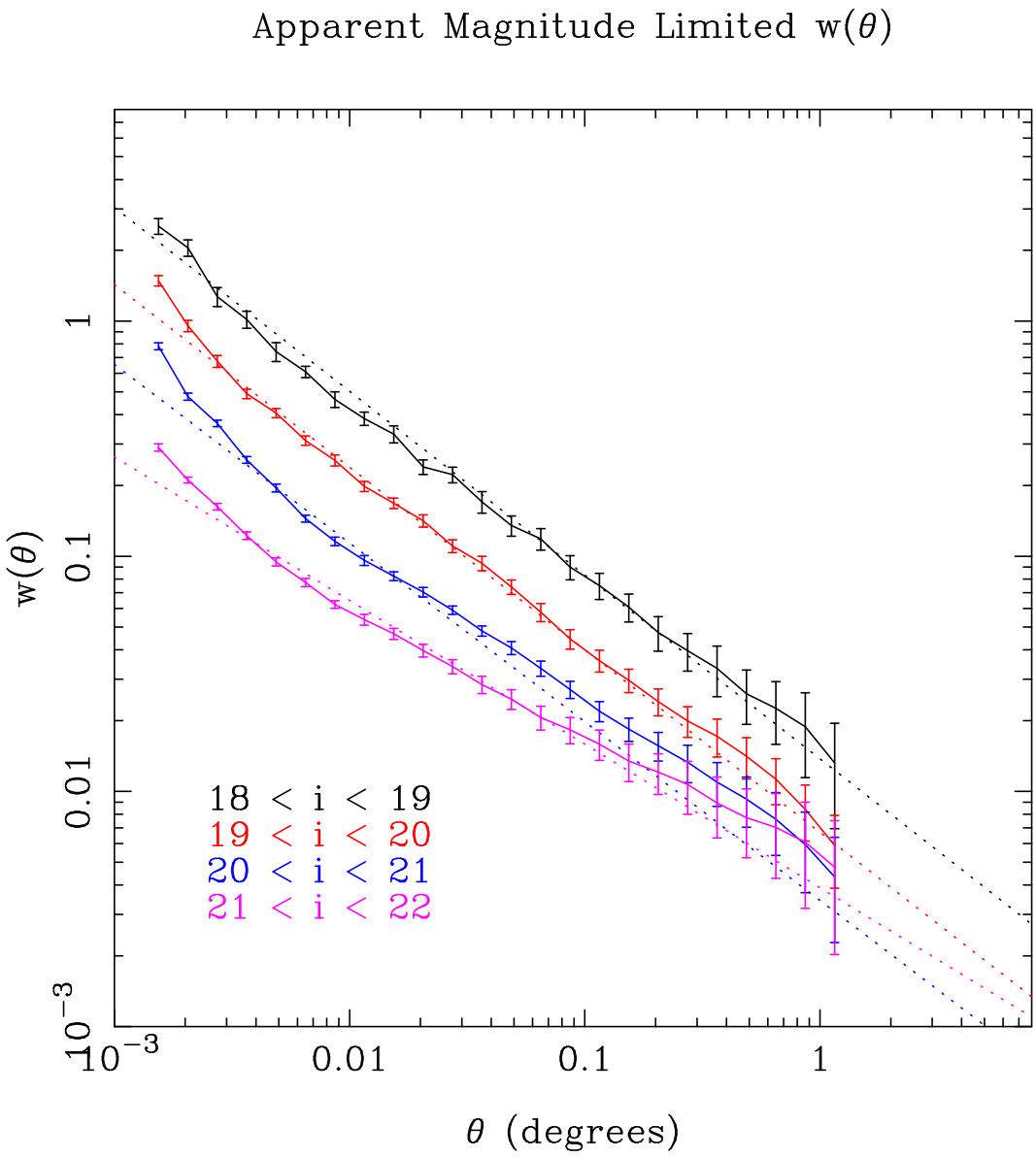


Figure 5.11 Evolution of $w(\theta)$ with i band apparent magnitude. The sample used was apparent magnitude limited. The fit was performed using all points with $\theta < 0.1°$. The fit parameters are given in Table 5.2. These fits were performed for comparison to the DEEP2 results. See Figure 5.12.

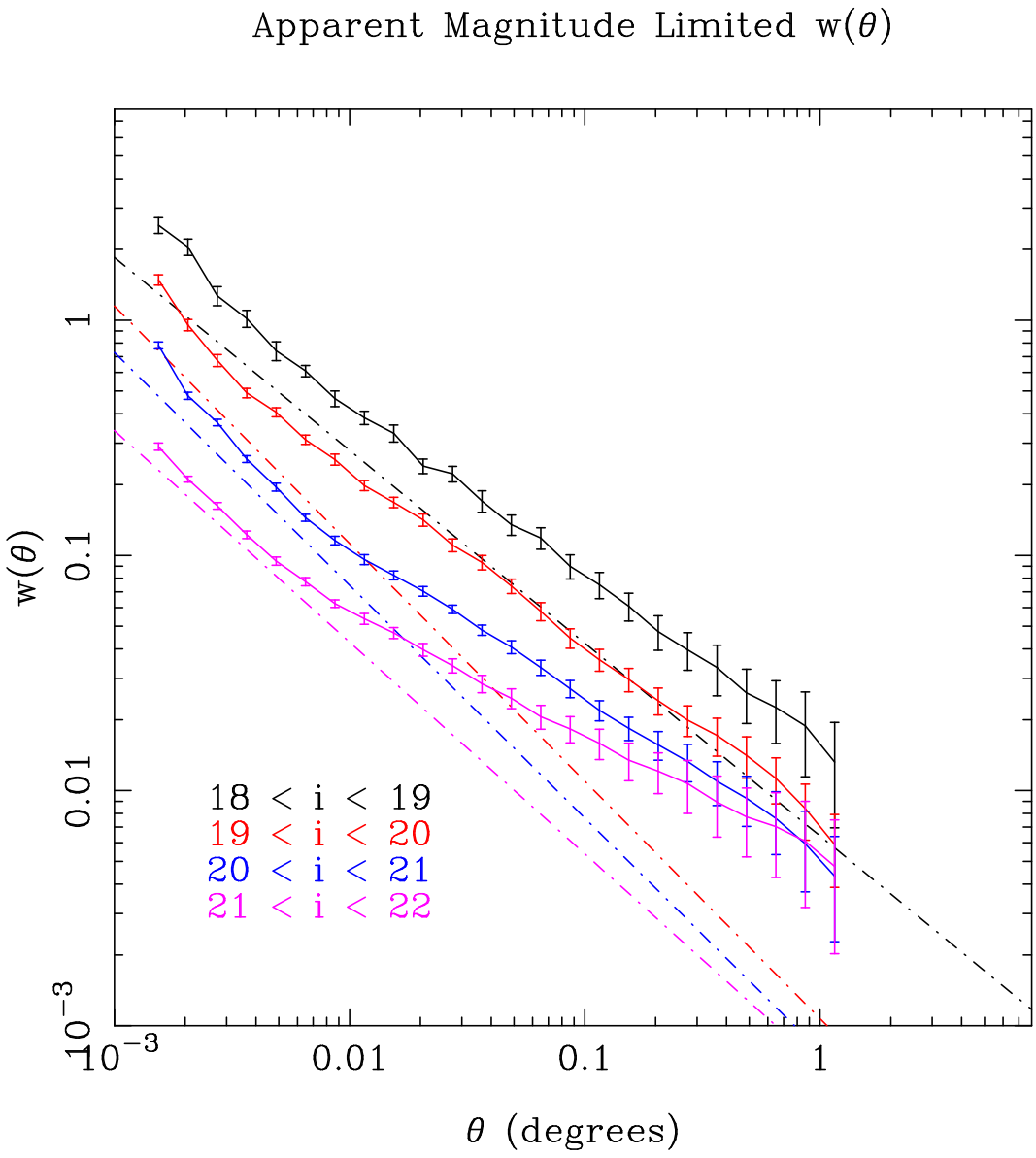


Figure 5.12 Comparison of $w(\theta)$ evolution in i band apparent magnitude to the results of Coil et al. (2004) (the dash-dot lines). The discrepancies are due to differences in sample selection. For SDSS, the selection was i band apparent magnitude limited. For DEEP2, objects were selected using a color cut to predominantly select high redshift (> 0.7) objects before an apparent magnitude cut was applied. See also Figure 5.11 and Table 5.2.

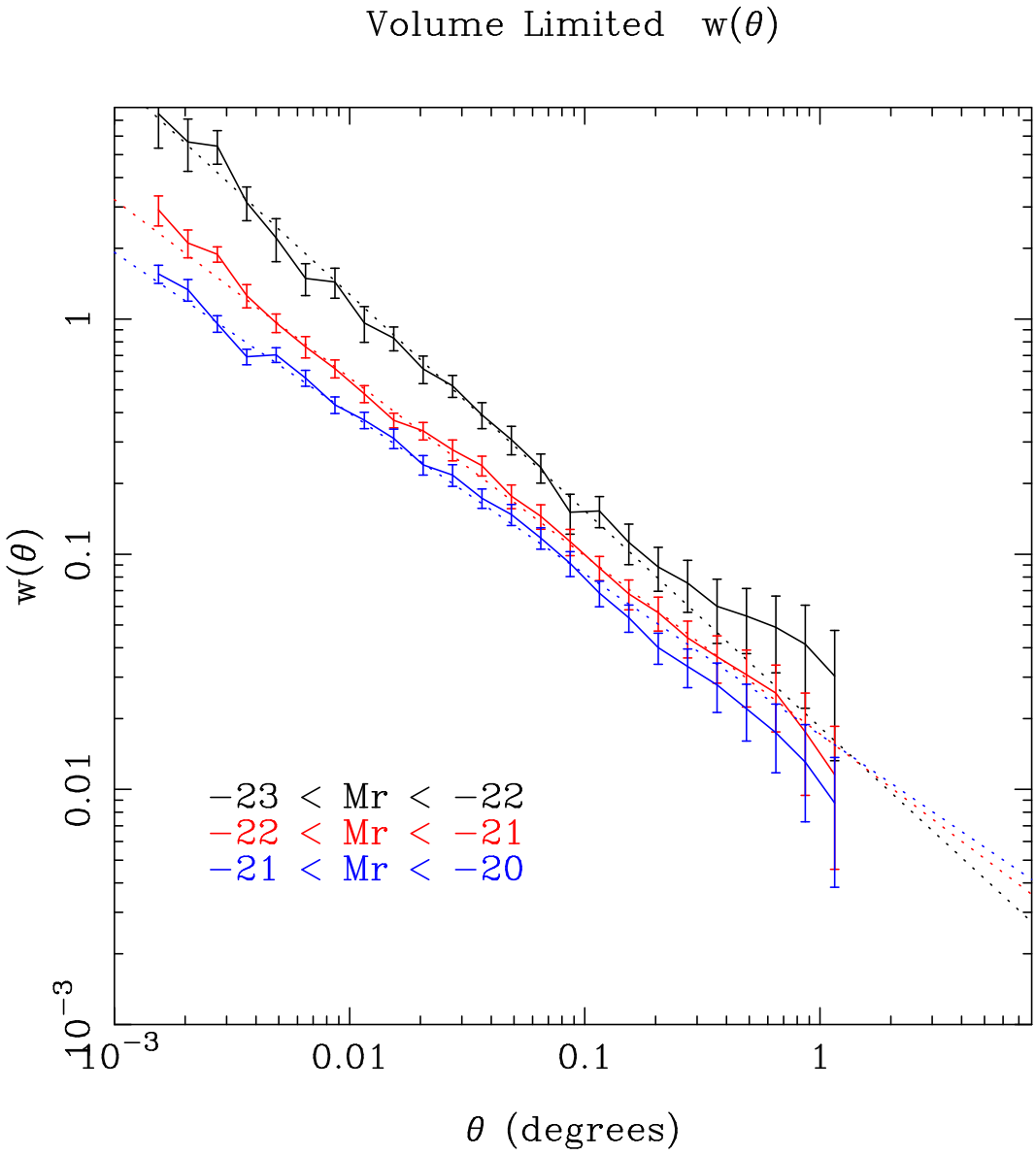


Figure 5.13 Evolution of $w(\theta)$ with absolute magnitude. The sample used was volume limited. The fit was performed using all points with $\theta < 0.1°$. The fit parameters are given in Table 5.3.

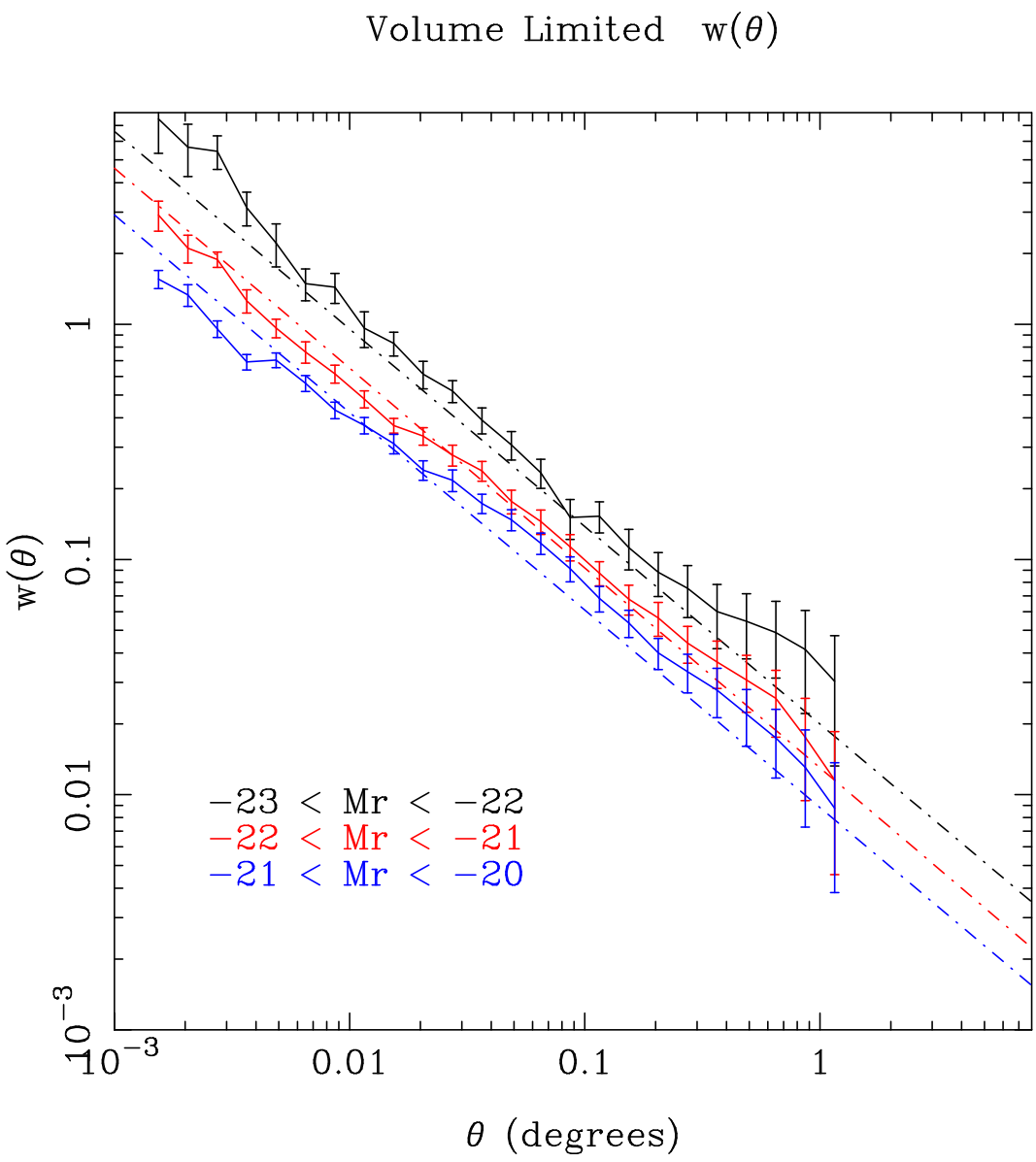


Figure 5.14 Comparison of $w(\theta)$ evolution in absolute magnitude bins to the results of Budavári et al. (2003) (the dash-dot lines). The fits agree reasonably well, though it is clear that the larger sample size (roughly a factor of 8) smooths out $w(\theta)$ considerably. See also Figure 5.13 and Table 5.3.

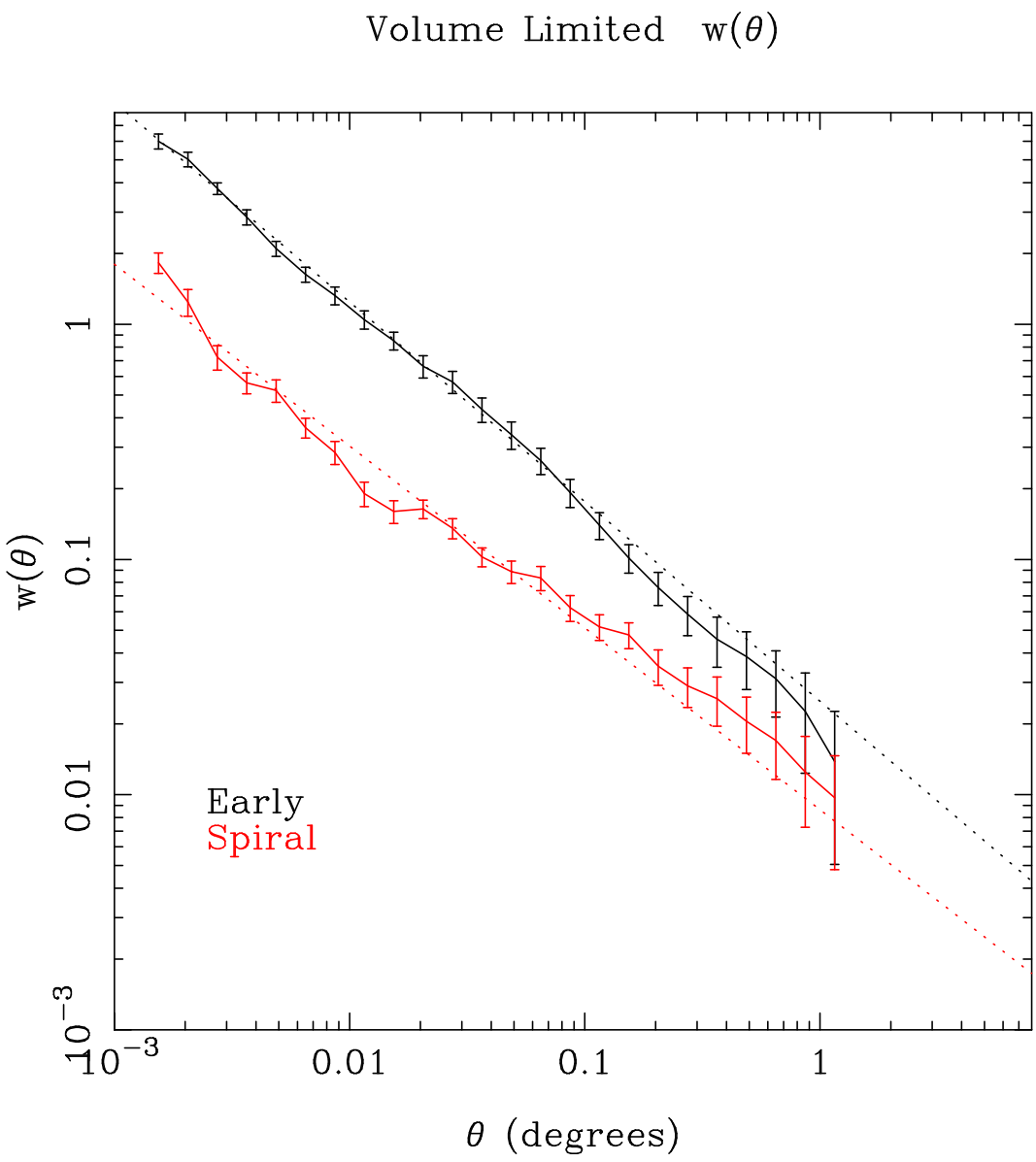


Figure 5.15 Evolution of $w(\theta)$ with galaxy type. The sample used was volume limited. The fit was performed using all points with $\theta < 0.1°$. The fit parameters are given in Table 5.4. Irregulars are not included due to low number counts (≈ 1000) which yield incorrect covariance matrices in our $w(\theta)$ code.

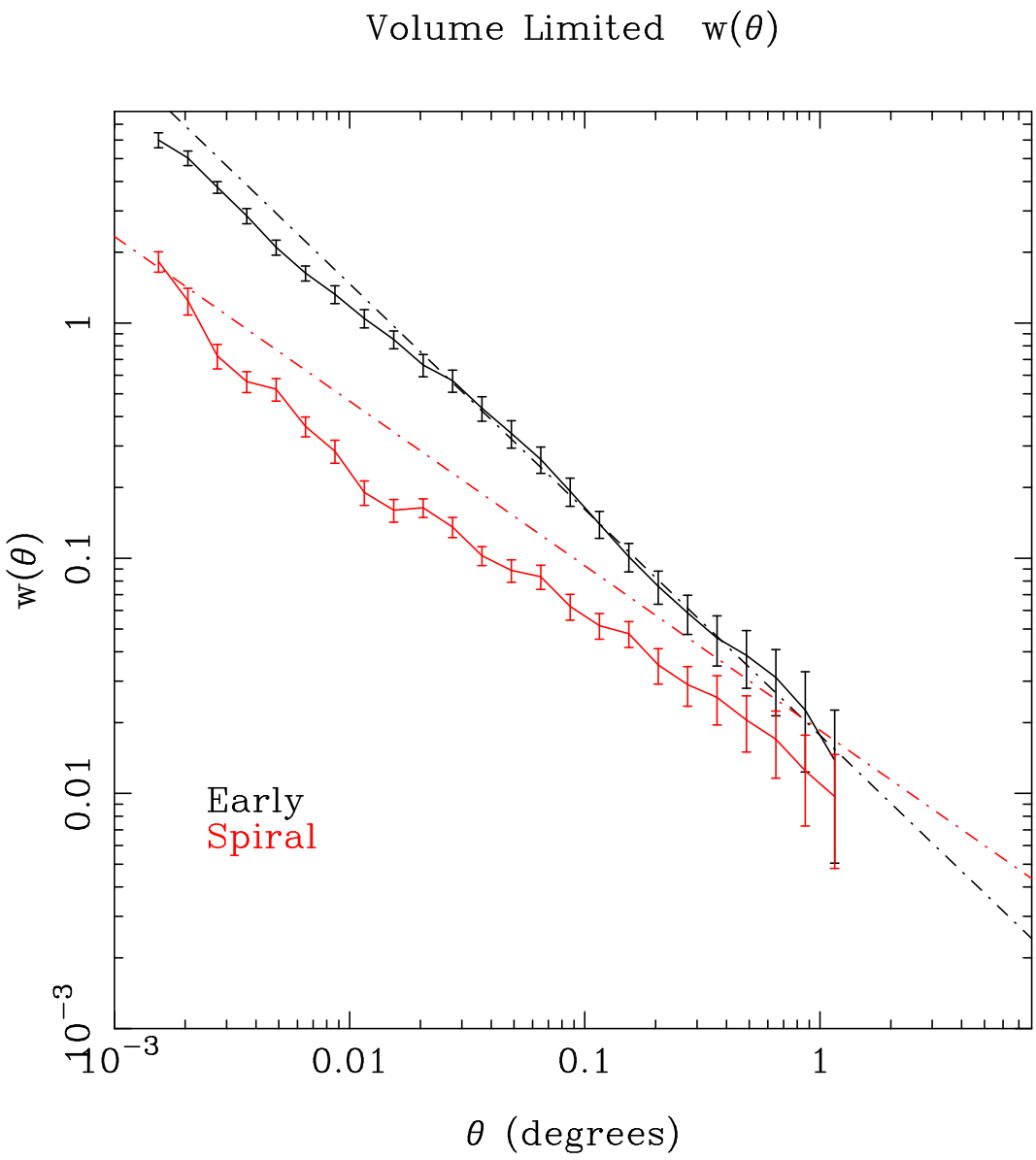


Figure 5.16 Comparison of $w(\theta)$ evolution with type to the results of Budavári et al. (2003) (the dash-dot lines). Some discrepancy is due to different luminosity bins, and some of it to different spectral templates, though our results are in qualitative agreement. See also Figure 5.15 and Table 5.4.

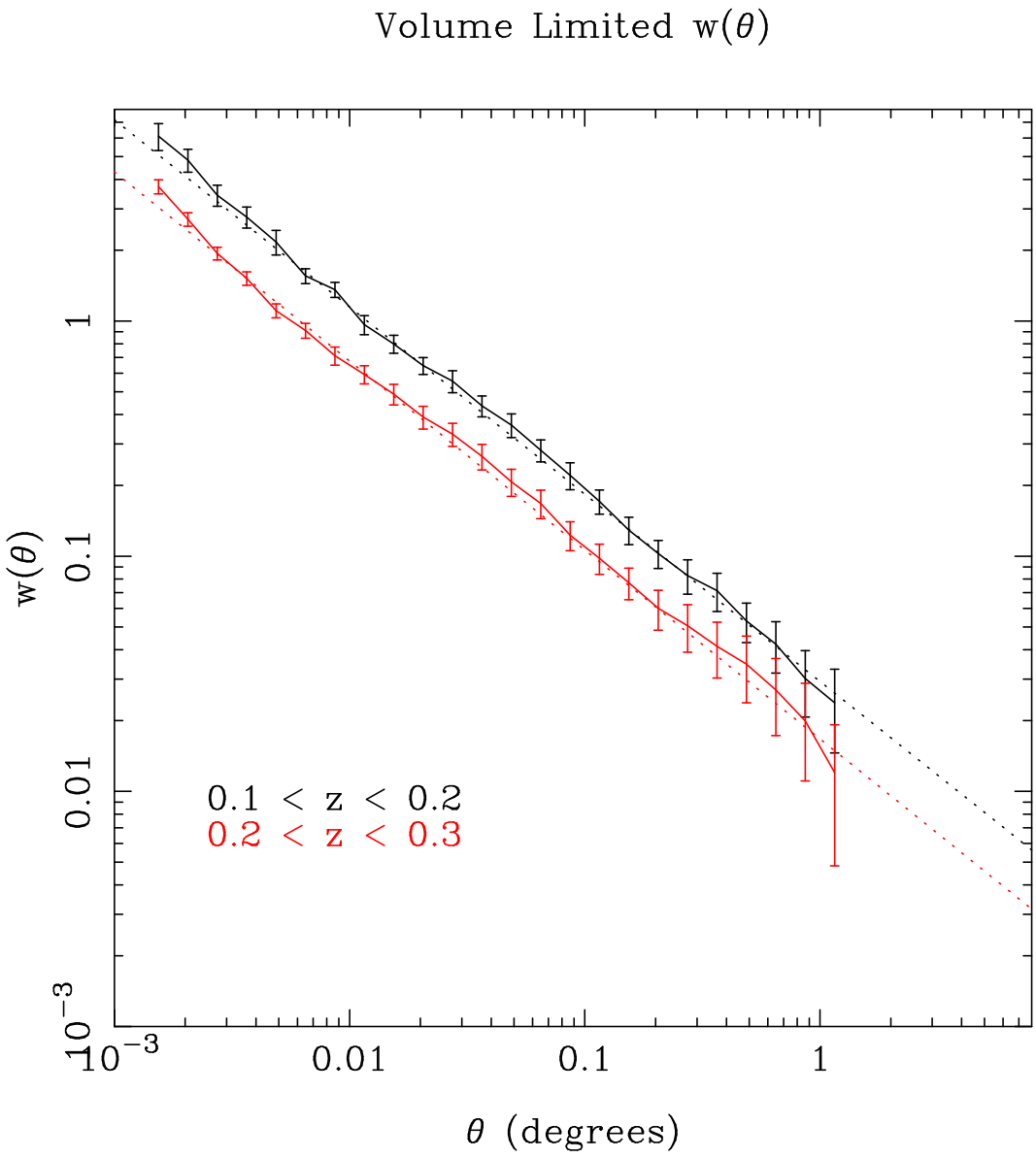


Figure 5.17 Evolution of $w(\theta)$ with redshift. The sample used was volume limited. The fit was performed using all points with $\theta < 0.1°$. The fit parameters are given in Table 5.5.

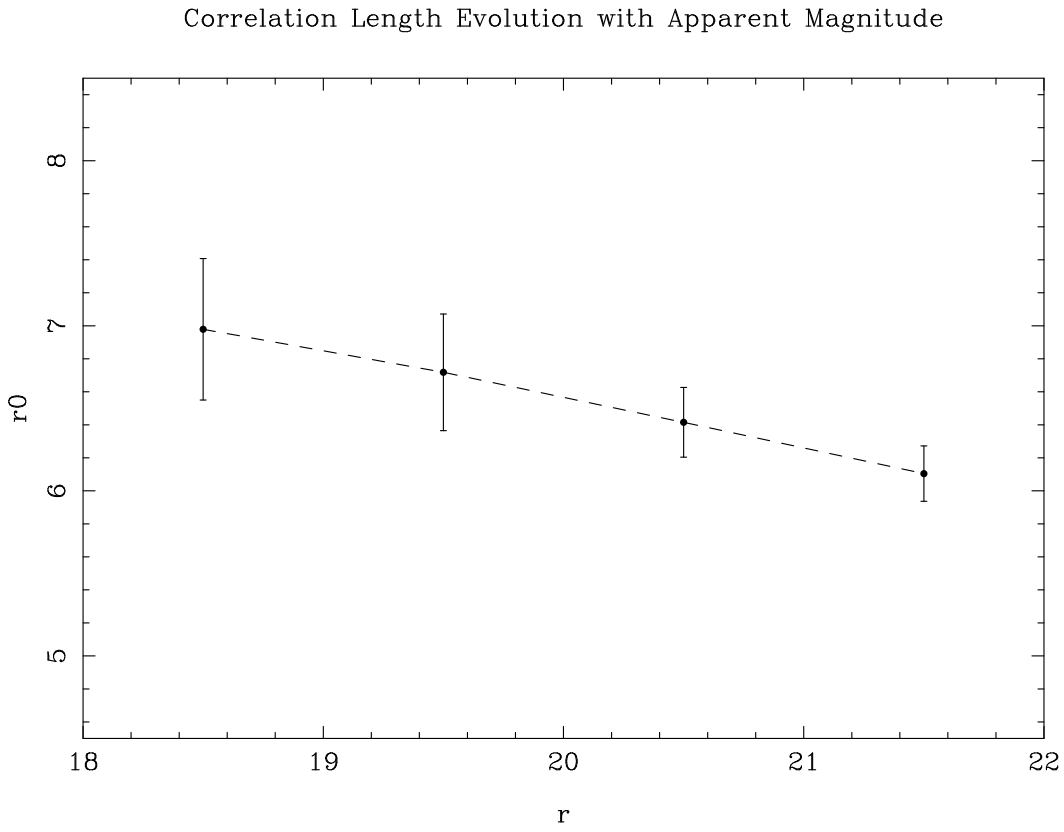


Figure 5.18 Evolution of r_0 with apparent magnitude. The sample used was apparent magnitude limited. The fits used to derive r_0 are from Figure 5.9 and Table 5.1.

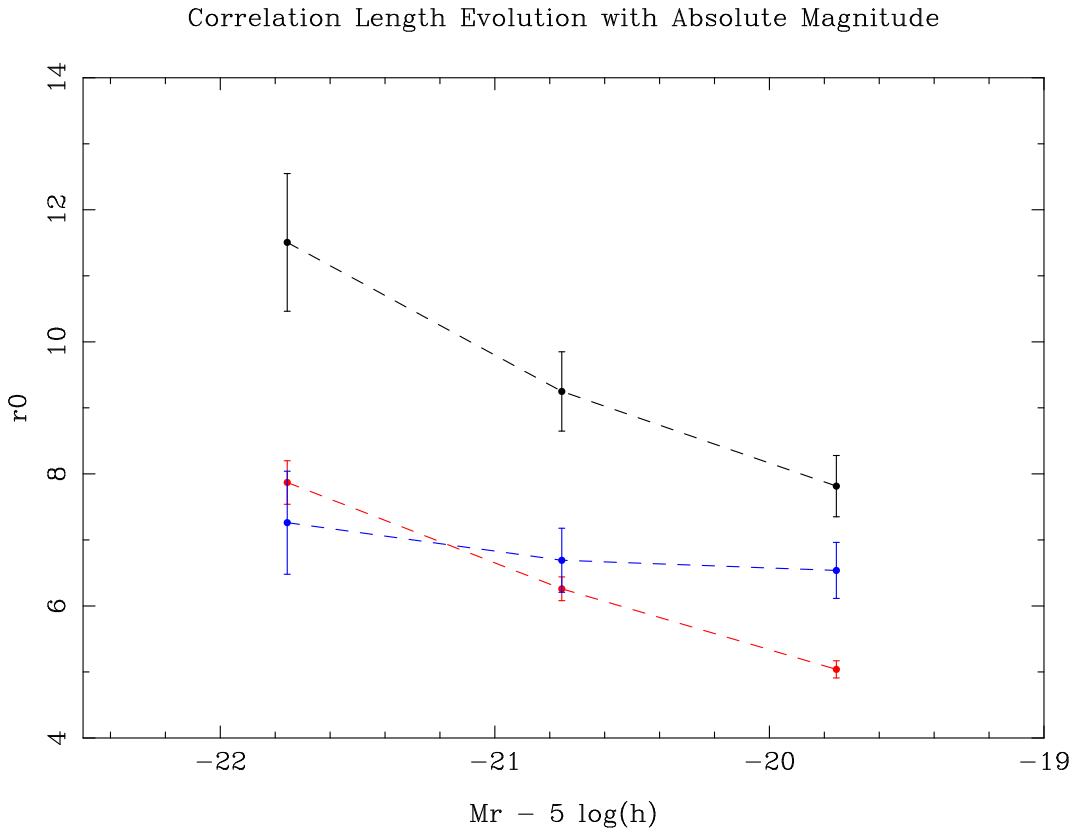


Figure 5.19 Evolution of r_0 with absolute magnitude. The black line shows our results, the red line the results of Budavári et al. (2003), and the blue line our results with a more narrow $\frac{dn}{dz}$ (see Figure 5.20). The sample used was volume limited. The fits used to derive r_0 are from Figure 5.13 and Table 5.3.

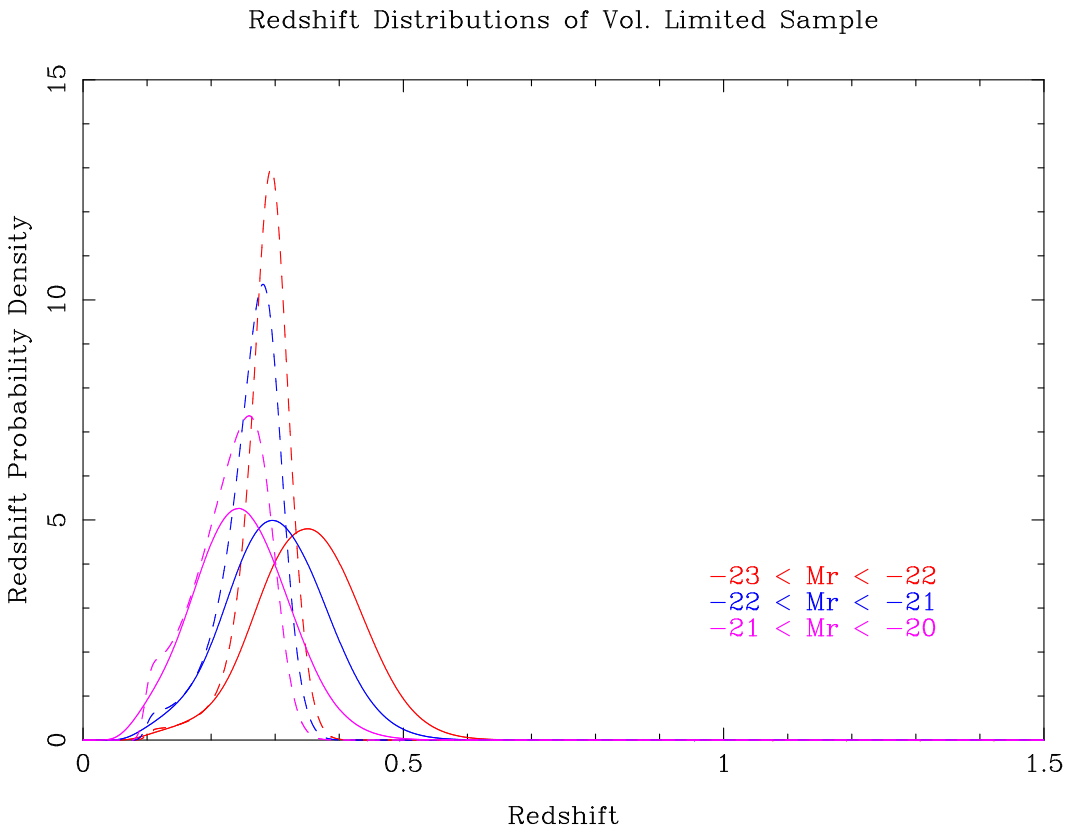


Figure 5.20 Comparison of redshift distributions for the volume limited sample as a function of contamination in photometric redshift. The solid line shows our derived $\frac{\mathrm{d}n}{\mathrm{d}z}$ from Figure 5.6 and the dashed line that of a factor of 4 smaller photometric redshift contamination in each apparent magnitude bin. The dashed curves are the derived redshift distributions which approximately match the results of Budavári et al. (2003). See also figure 5.19.

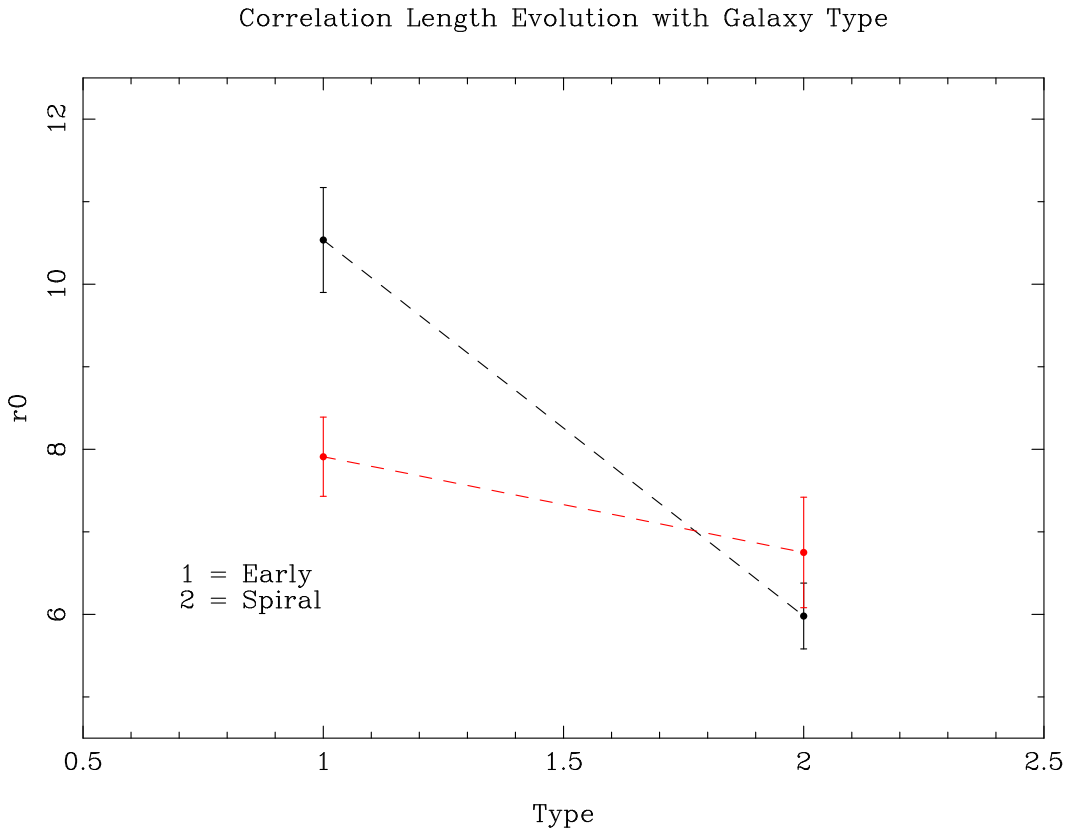


Figure 5.21 Evolution of r_0 with galaxy type. The black line shows our results and the red line the results of Budavári et al. (2003). The sample used was volume limited. The fits used to derive r_0 are from Figure 5.15 and Table 5.4.

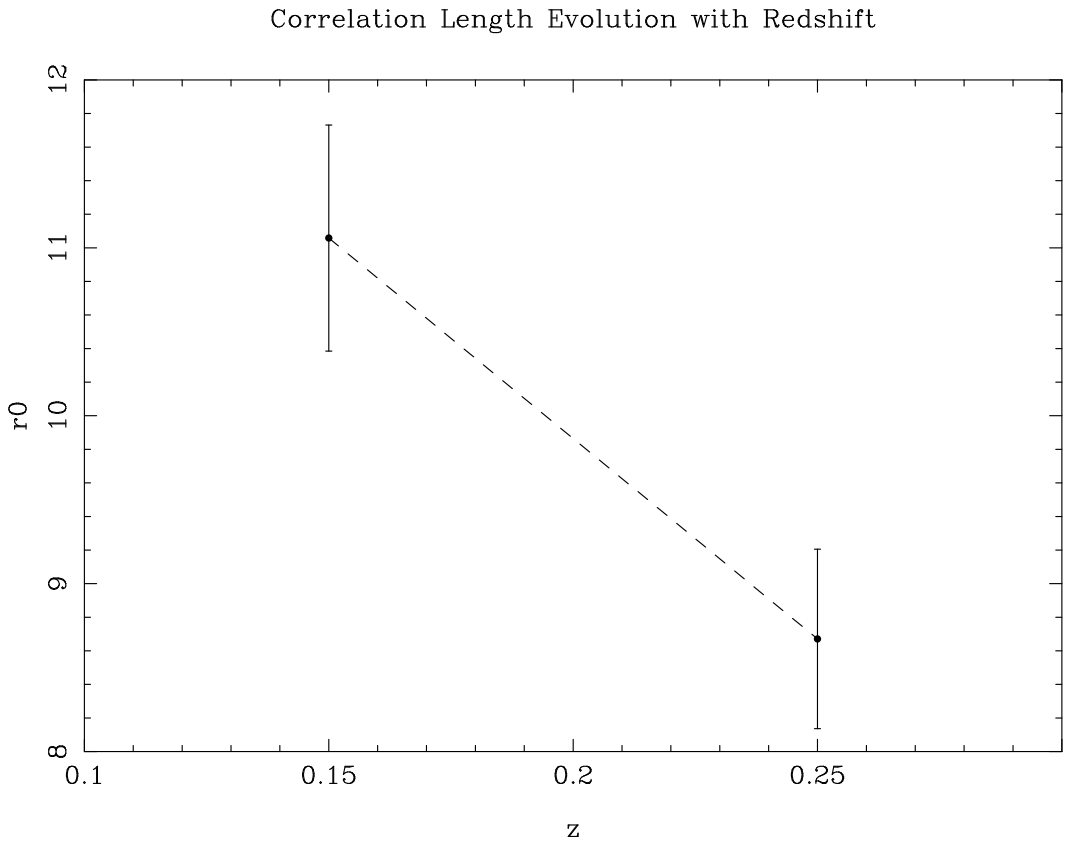


Figure 5.22 Evolution of r_0 with redshift. The sample used was volume limited. The fits used to derive r_0 are from Figure 5.17 and Table 5.5.

Magnitude	Slope	Intercept	γ	r_0 (Mpc h^{-1})	$\frac{\chi^2}{\mathrm{DOF}}$
$18 < r < 19$	-0.727 ± 0.023	-1.693 ± 0.045	1.727 ± 0.023	6.979 ± 0.437	15.41
$19 < r < 20$	-0.722 ± 0.017	-1.946 ± 0.037	1.722 ± 0.017	6.718 ± 0.354	16.55
$20 < r < 21$	-0.793 ± 0.011	-2.284 ± 0.025	1.793 ± 0.011	6.415 ± 0.214	17.79
$21 < r < 22$	-0.798 ± 0.009	-2.528 ± 0.019	1.798 ± 0.009	6.104 ± 0.162	27.43

Table 5.1 Fit parameters for $w(\theta)$ in apparent magnitude bins as shown in Figure 5.9.

Magnitude	Slope	Intercept	γ	r_0 (Mpc h^{-1})	$\frac{\chi^2}{\mathrm{DOF}}$
$18 < i < 19$	-0.781 ± 0.019	-1.861 ± 0.039	1.781 ± 0.019	5.777 ± 0.299	6.19
$19 < i < 20$	-0.776 ± 0.013	-2.174 ± 0.029	1.776 ± 0.013	5.201 ± 0.196	9.85
$20 < i < 21$	-0.760 ± 0.010	-2.464 ± 0.022	1.760 ± 0.010	4.917 ± 0.148	52.55
$21 < i < 22$	-0.611 ± 0.012	-2.410 ± 0.029	1.611 ± 0.012	5.672 ± 0.259	31.52

Table 5.2 Fit parameters for $w(\theta)$ in i band apparent magnitude bins as shown in Figure 5.11.

Magnitude	Slope	Intercept	γ	r_0 (Mpc h^{-1})	$\frac{\chi^2}{\mathrm{DOF}}$
$-23 < M_r < -22$	-0.920 ± 0.037	-1.736 ± 0.069	1.920 ± 0.037	11.506 ± 1.043	1.28
$-22 < M_r < -21$	-0.757 ± 0.023	-1.764 ± 0.047	1.757 ± 0.023	9.249 ± 0.601	4.24
$-21 < M_r < -20$	-0.684 ± 0.020	-1.768 ± 0.042	1.684 ± 0.020	7.815 ± 0.464	3.40

Table 5.3 Fit parameters for $w(\theta)$ in absolute magnitude bins as shown in Figure 5.13.

Type	Slope	Intercept	γ	r_0 (Mpc h^{-1})	$\frac{\chi^2}{\mathrm{DOF}}$
Early	-0.849 ± 0.020	-1.603 ± 0.045	1.849 ± 0.020	10.536 ± 0.636	6.62
Spiral	-0.773 ± 0.023	-2.064 ± 0.046	1.773 ± 0.023	5.980 ± 0.399	11.34

Table 5.4 Fit parameters for $w(\theta)$ in galaxy type bins as shown in Figure 5.15.

Redshift	Slope	Intercept	γ	r_0 (Mpc h^{-1})	$\frac{\chi^2}{\mathrm{DOF}}$
$0.1 < z < 0.2$	-0.796 ± 0.023	-1.533 ± 0.047	1.796 ± 0.023	11.058 ± 0.673	3.79
$0.2 < z < 0.3$	-0.804 ± 0.020	-1.777 ± 0.046	1.804 ± 0.020	8.671 ± 0.534	7.72

Table 5.5 Fit parameters for $w(\theta)$ in redshift bins as shown in Figure 5.17.

6.0 THE HALO MODEL

The early galaxy clustering theory of Neyman and Scott (1952) proposed to describe the galaxy distribution using galaxy clusters of various sizes. Given information about the distribution of sizes of these galaxy clusters, the arrangement of galaxies within these clusters, and how clusters are scattered across the sky (the "clustering of clusters"), one can describe the statistical properties of galaxy clustering. This theory faced a major drawback, though – none of the pieces of the model were understood due to insufficient data.

Presently, we know that the majority of the universe is made up of dark matter which was initially smoothly distributed. This enables the study of dark matter structure evolution through perturbation theory (Bernardeau et al., 2002) down to scales of a few megaparsecs; on smaller scales, the clustering becomes non-linear. Increases in computer efficiency have made possible numerical studies of the growth of non-linear dark matter structure as well. These studies have shown that dark matter evolves from its initially smooth conditions to a spiderweb-like distribution of knots and filaments; these knots are known as *halos*. Because simulations are limited by the number of particles that can be simulated in a reasonable amount of time, dark matter has been tackled from two different perspectives. First, high resolution, low volume simulations have mapped the structure of dark matter within and around a single halo (Navarro et al., 1997). Second, low resolution, high volume simulations have characterized the large scale distribution of dark matter halos throughout the universe (Jenkins et al., 2001). Together, these advances have enabled the treatment of dark matter using a similar approach to that of Neyman and Scott (1952) with great success; collectively, the models which describe dark matter using this approach are termed *the halo model* (Cooray and Sheth, 2002; Zentner, 2008). The halo model can be extended to describe the clustering of galaxies which are known to trace the dark matter distribution (White and Rees,

1978), making it a useful comparison for observations because we cannot directly observe dark matter.

The fundamental assumption of the halo model is that all dark matter resides within halos. The size of a halo (usually assumed to be spherical) is given by the *virial radius* – the radius needed to enclose some over-density of mass (typically ≈200). Thus, a halo is by definition a sphere of dark matter with a density 200 times that of the background. We can use the halo model to describe dark matter clustering by specifying 3 ingredients: 1.) the distribution of dark matter halos as a function of mass $\frac{\mathrm{d}n}{\mathrm{d}m}$, termed the *halo mass function* (Sheth and Tormen, 1999), 2.) the bias between halos and dark matter $b_{\mathrm{h}}(m)$ (Sheth and Tormen, 1999; Tinker et al., 2005), and 3.) the distribution of dark matter within halos $\lambda_m(\vec{x})$ (Navarro et al., 1997). Additionally, it is often assumed that halo clustering depends only upon halo mass, though it is possible to extend the formalism to include other halo properties. While it has been shown that halo clustering depends upon age (Gao et al., 2005) and concentration (Wechsler et al., 2006), these effects are small enough to justify neglecting them.

In this chapter we give a very brief overview of the halo model and its application to galaxy clustering and present results constraining the halo occupation distribution.

6.1 DARK MATTER CLUSTERING

Using the halo model, we can compute the 2 point correlation function of dark matter in real space. We can break this calculation into two terms: one dealing with correlations within a single halo and another dealing with correlations between different halos. These terms are known as the "one halo" and "two halo" terms, respectively. Clearly, the one halo term dominates at small scales (*i.e.* scales less than the typical virial radius), and the two halo term dominates at large scales. The correlation function is then

$$\xi(r) = \xi_{1\mathrm{h}}(r) + \xi_{2\mathrm{h}}(r) \tag{6.1}$$

We compute the one and two halo terms by counting pairs of infinitesimal masses (Zent-

ner, 2008):

$$\xi_{1\mathrm{h}}(r) = \frac{1}{\bar{\rho}^2} \int \mathrm{d}m\, m^2 \frac{\mathrm{d}n}{\mathrm{d}m} \int \mathrm{d}^3x\, \lambda_m(\vec{x})\lambda_m(\vec{x}+\vec{r}) \tag{6.2}$$

$$\xi_{2\mathrm{h}}(r) = \frac{1}{\bar{\rho}^2} \int \mathrm{d}m_1 \int \mathrm{d}m_2\, m_1 \frac{\mathrm{d}n}{\mathrm{d}m_1} m_2 \frac{\mathrm{d}n}{\mathrm{d}m_2} \int \mathrm{d}^3x \int \mathrm{d}^3y\, \lambda_{m_1}(\vec{x})\lambda_{m_2}(\vec{y})\xi_{\mathrm{hh}}(\vec{x}-\vec{y}+\vec{r}|m_1,m_2) \tag{6.3}$$

Here $\bar{\rho}$ is the mean mass density of the universe, $\xi_{\mathrm{hh}}(\vec{x}|m_1,m_2)$ is the cross correlation function for halos of masses m_1 and m_2, and $\lambda_m(\vec{x})$ is the density profile form of Navarro et al. (1997), commonly called the NFW profile:

$$\lambda_m(r) \propto \left[\frac{c(m)r}{r_{\mathrm{vir}}}\right]^{-1} \left[1+\frac{c(m)r}{r_{\mathrm{vir}}}\right]^{-2} \tag{6.4}$$

The concentration c is some weak function of halo mass specified by the model (Bullock et al., 2001, *e.g.*), and r_{vir} is the virial radius. Furthermore, it is usually assumed that

$$\xi_{\mathrm{hh}}(r|m_1,m_2) = b_{\mathrm{h}}(m_1)b_{\mathrm{h}}(m_2)\xi_{\mathrm{linear}}(r) \tag{6.5}$$

where $\xi_{\mathrm{linear}}(r)$ is the correlation function of dark matter as computed from linear perturbation theory.

Because this integral involves a convolution, it is usually computed in Fourier space where convolutions are simple multiplications. Several other approximations are needed to compute the integrals, the details of which are beyond the scope of this simple overview – see Zentner (2008) for details. We use the halo model code of A. Zentner (private communication) with the NFW profile (Navarro et al., 1997), concentrations from Bullock et al. (2001), and the halo mass function $\frac{\mathrm{d}n}{\mathrm{d}m}$ and bias $b_{\mathrm{h}}(m)$ from Sheth and Tormen (1999).

6.2 GALAXY CLUSTERING: THE HALO OCCUPATION DISTRIBUTION

We can also compute the correlation function of galaxies using the halo model if we assume that all galaxies reside within dark matter halos. This computation is further simplified if we

assume that the properties of galaxies within a halo depend only upon their host halo mass. To compute the galaxy correlation function, we need only specify the probability $p(N_{\rm gal}|m)$ that a halo of mass m host $N_{\rm gal}$ galaxies; this probability is known as the *halo occupation distribution* (HOD). As before, we denote the spatial distribution of galaxies within a halo of mass m as $\lambda(\vec{x}|m)$ and break the computation into correlations between galaxies within the same halo and galaxies in separate halos:

$$\xi_{\rm gg}(r) = \xi^{\rm 1h}_{\rm gg}(r) + \xi^{\rm 2h}_{\rm gg}(r) \tag{6.6}$$

Each term is given by counting pairs of galaxies (Zentner, 2008):

$$\xi^{\rm 1h}_{\rm gg}(r) = \frac{1}{\bar{n_{\rm gal}}^2} \int {\rm d}m \, \frac{{\rm d}n}{{\rm d}m} \langle N_{\rm gal}(N_{\rm gal}-1)\rangle_m \int {\rm d}^3x \, \lambda_m(\vec{x})\lambda_m(\vec{x}+\vec{r}) \tag{6.7}$$

$$\xi^{\rm 2h}_{\rm gg}(r) = \frac{1}{\bar{n_{\rm gal}}^2} \int {\rm d}m_1 \int {\rm d}m_2 \, \frac{{\rm d}n}{{\rm d}m_1} \langle N_{\rm gal}\rangle_{m_1} \frac{{\rm d}n}{{\rm d}m_2} \langle N_{\rm gal}\rangle_{m_2} \int {\rm d}^3x \int {\rm d}^3y \, \lambda_{m_1}(\vec{x})\lambda_{m_2}(\vec{y})\xi_{\rm hh}(\vec{x}-\vec{y}+\vec{r}|m_1,m_2) \tag{6.8}$$

Here $\bar{n_{\rm gal}}$ is the mean number density of galaxies. The cross correlation of halos $\xi_{\rm hh}$ is again simplified by assuming it can be expressed in terms of the halo bias $b_{\rm h}(m)$. The one halo term contains the mean number of galaxy pairs per halo:

$$\langle N_{\rm gal}(N_{\rm gal}-1)\rangle_m = \sum_{N_{\rm gal}=1}^{\infty} N_{\rm gal}(N_{\rm gal}-1)\, p(N_{\rm gal}|m) \tag{6.9}$$

The two halo term depends on the average number of galaxies within a halo:

$$\langle N_{\rm gal}\rangle_m = \sum_{N_{\rm gal}=1}^{\infty} N_{\rm gal}\, p(N_{\rm gal}|m) \tag{6.10}$$

These equations are fully determined once we have parameterized $p(N_{\rm gal}|m)$. Simulations indicate that it is useful to model the contributions to $p(N_{\rm gal}|m)$ as arising from two distinct populations of galaxies: central galaxies located at the center of halos and satellite galaxies scattered throughout the halo according to some distribution. It is known that smaller halos do not host galaxies, so the central galaxy component is often modeled as a step function such that halos with mass below some threshold $M_{\rm min}$ host 0 galaxies, and galaxies above

$M_{\rm min}$ contain 1 central galaxy. This distribution is known as a *nearest integer* distribution, and it has a second moment $\langle N(N-1)\rangle = 0$. The average contribution of satellite galaxies is parameterized as a power law:

$$\langle N_{\rm sat}\rangle = \left(\frac{m}{M_1}\right)^{\alpha} \tag{6.11}$$

Simulations have demonstrated that satellite galaxies follow a Poisson distribution within a halo of fixed mass m; this is convenient as the Poisson distribution is fully specified with the average $\langle N_{\rm sat}\rangle$.

The halo occupation distribution is thus fully specified with a given cosmology, redshift (the mass functions change with redshift), and set of HOD parameters $M_{\rm min}$, M_1, and α. Multiple studies (*e.g.* Zehavi et al., 2005) have found that $M_1 \approx 20 M_{\rm min}$, so the number of degrees of freedom can be reduced to 2; assuming this ratio effectively limits the parameter search to HOD models that are similar to a power law (Zentner, 2008). Hence, we can compare our $w(\theta)$ measurements to the HOD model and solve for α and $M_{\rm min}$ or M_1 instead of using a standard 2 parameter least squares fit.

6.3 COMPARISON TO HOD MODELS

To compare against our $w(\theta)$ measurements, we fixed $\xi(r)$ as a power law with γ and r_0 from our previous fits and then solved for α and M_1 with $M_{\rm min} = \frac{M_1}{20}$. We used a simple adaptive grid of size 10×10 to search for the optimal parameters: the cell containing the optimum values of α and M_1 from each iteration was expanded to a new 10×10 grid and the process repeated until both parameters were known to a relative tolerance of 10^{-4}. α was allowed to vary from 0.5 to 2.5 in evenly spaced steps, and $\log M_1$ from 11 to 16 (M_1 is in units $M_\odot h^{-1}$) in evenly spaced steps in log space. To estimate errors, we used 50 Monte Carlo iterations and varied γ and r_0 according to their measured errors; more iterations were computationally prohibitive. For the redshift, we used the average redshift for each bin.

Figure 6.1 shows the evolution of $\xi(r)$ with redshift for the volume limited sample. The dotted lines show the derived power law fits from Table 5.5, and the solid lines show the

optimal HOD fits given in Table 6.1. The dip seen in both HOD curves is due to the transition between the 1 halo and 2 halo terms. This general shape demonstrates a basic difficulty for our method of comparing to the halo model, namely that $\xi(r)$ is not well described by a power law in the intermediate regime. Nonetheless, the trend of decreasing amplitude with roughly constant slope as seen in the power law evolution is present in the HOD curves as well.

Figures 6.2 and 6.3 show the redshift evolution of α and M_1 respectively. From these, we see that α is nearly constant with redshift while M_1 decreases slightly. Thus, the redshift evolution in galaxy clustering appears to be due to an increase in the characteristic mass of halos that host galaxies. This scenario is consistent with the ΛCDM cosmological model where large structure grows bottom-up from mergers of smaller masses. Because we expect halos to grow in size over time, halos at later times can host more galaxies, so we expect clustering to increase at lower redshifts.

The evolution of $\xi(r)$ with luminosity is presented in Figure 6.4 using γ and r_0 parameters from Table 5.3. The $-23 < M_r < -22$ and $-21 < M_r < -20$ bins agree well with the power law, but the $-22 < M_r < -21$ bin is poorly described by the HOD. As seen in Figures 6.5 and 6.6, the HOD parameter space becomes is very degenerate near the solution for this particular bin. From these plots we see that α decreases for brighter objects and that M_1 shows the reverse trend, consistent with the results of Zehavi et al. (2005) and theoretical models. Semianalytic models predict that the most massive halos host the brightest galaxies, and it is also known that these massive halos are more strongly clustered. As a result we expect that brighter galaxies will be more strongly clustered (Cooray and Sheth, 2002). Our results are consistent with this explanation because we find that more luminous galaxies reside in more massive halos. Additionally, we find that power law slope of the distribution of satellite galaxies increases with luminosity, consistent with Zehavi et al. (2005) and numerical simulations (Gao et al., 2004, *e.g.*). That is, higher mass halos host a larger number of high luminosity galaxies.

For comparison, we also show the results of Zehavi et al. (2005) in Figures 6.5 and 6.6 as the red points and lines. No exact errors are listed in Zehavi et al. (2005) though they are of the order ± 0.05 in α and $\pm 0.5 \times 10^{13} h^{-1} M_\odot$ as reported in Zehavi et al. (2004). It is

important to regard this comparison as only qualitative because Zehavi et al. (2005) use a much shallower spectroscopic sample and a different binning for M_r; instead of considering a particular range of M_r, they take every galaxy brighter than a given threshold in M_r. Despite these differences, there are consistent trends between are two results. First, M_1 decreases with M_r as expected from theory. This means that intrinsically brighter galaxies are found within larger mass halos. Second, both feature a sudden changes in the trend for α for luminous galaxies. Zehavi et al. (2005) suggest that this sudden change is a result of using a step function for $M_{\rm min}$ in the high luminosity samples rather than a smooth roll-off. Further investigation of this idea would require modification of our underlying halo model, which is beyond the scope of this thesis.

Our constraints on the HOD models would be significantly improved with increased number counts and reduced systematics – both of which should smooth out the resulting $w(\theta)$ curve and reduce the error bars – and a more accurate representation of the redshift distribution. With a smoother $w(\theta)$ curve, it should be possible to detect deviations from a power law on intermediate scales. These deviations were first associated with the 1 halo to 2 halo transition by Zehavi et al. (2004). The 1 to 2 halo transition would provide a stronger constraint on the HOD models because more points would meaningfully contribute to the χ^2 (currently the small and large scale points dominate as we assume a power law). Additionally, the improved redshift distribution would enable more accurate determination of γ and r_0. Finally, with an accurate representation of $\frac{{\rm d}n}{{\rm d}z}$ we can compare the two point projected correlation function $w(r_p)$ (projected along the axis perpendicular to the line of sight) instead of $\xi(r)$ to reduce redspace distortions along the line of sight due to galaxy motion.

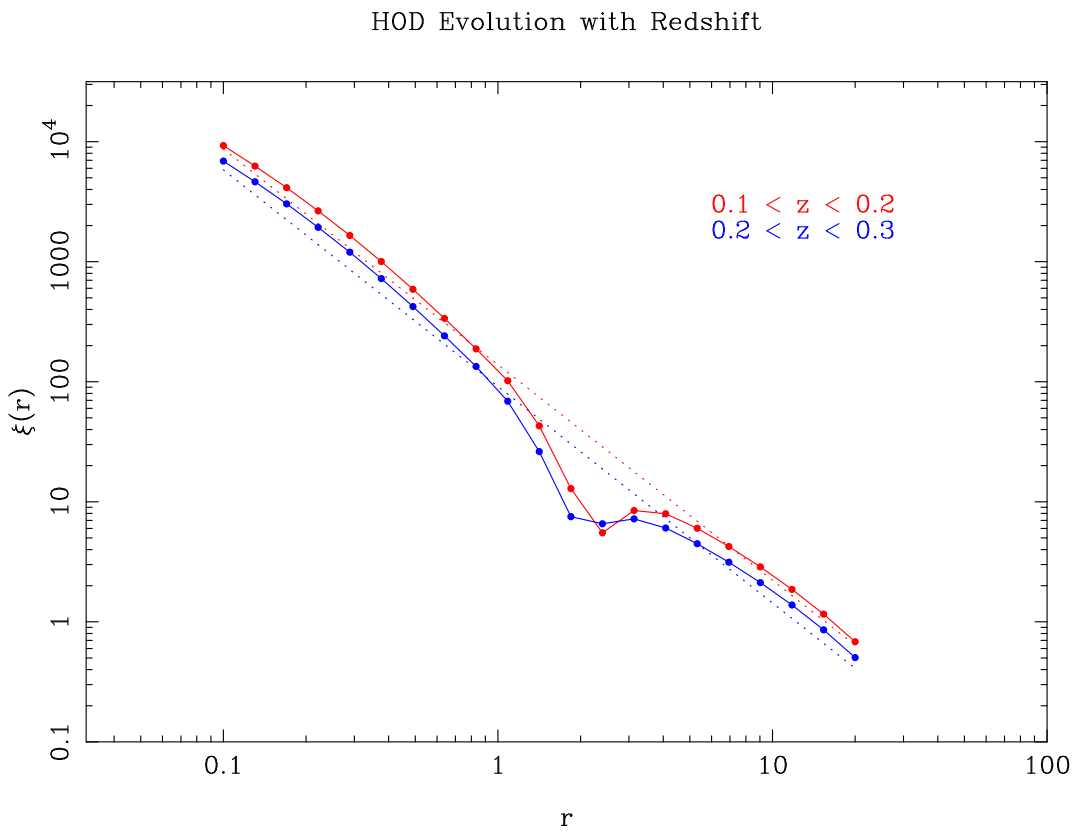


Figure 6.1 Evolution of best fit HOD $\xi(r)$ with redshift. The sample used was volume limited. The best fit parameters are listed in Table 6.1.

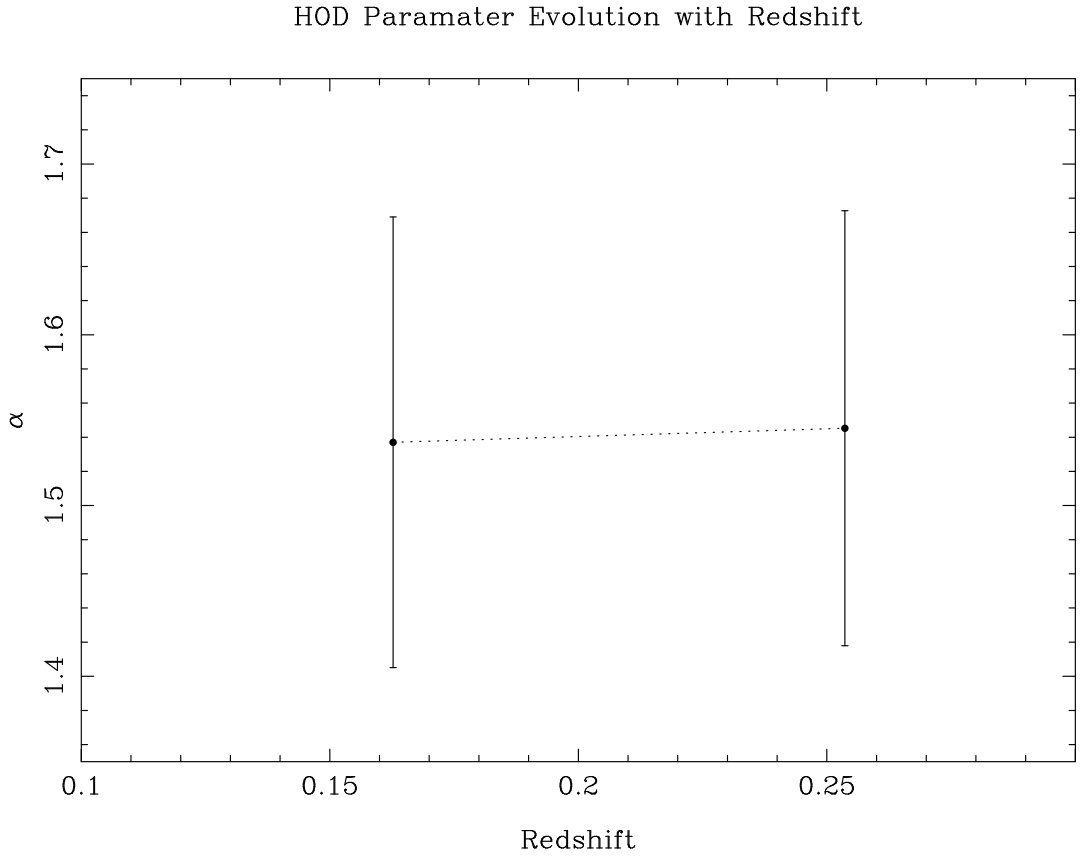


Figure 6.2 Evolution of HOD parameter α with redshift. The sample used was volume limited. The best fit parameters are listed in Table 6.1.

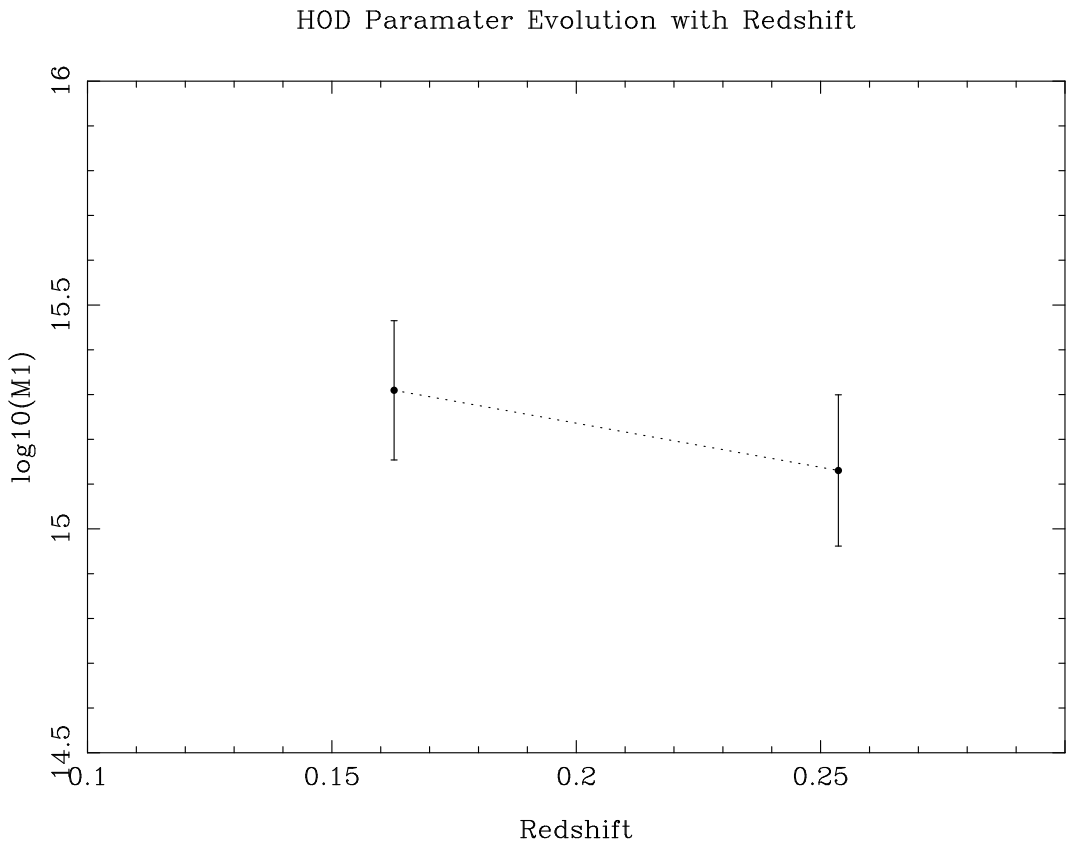


Figure 6.3 Evolution of HOD parameter M_1 with redshift. The sample used was volume limited. The best fit parameters are listed in Table 6.1.

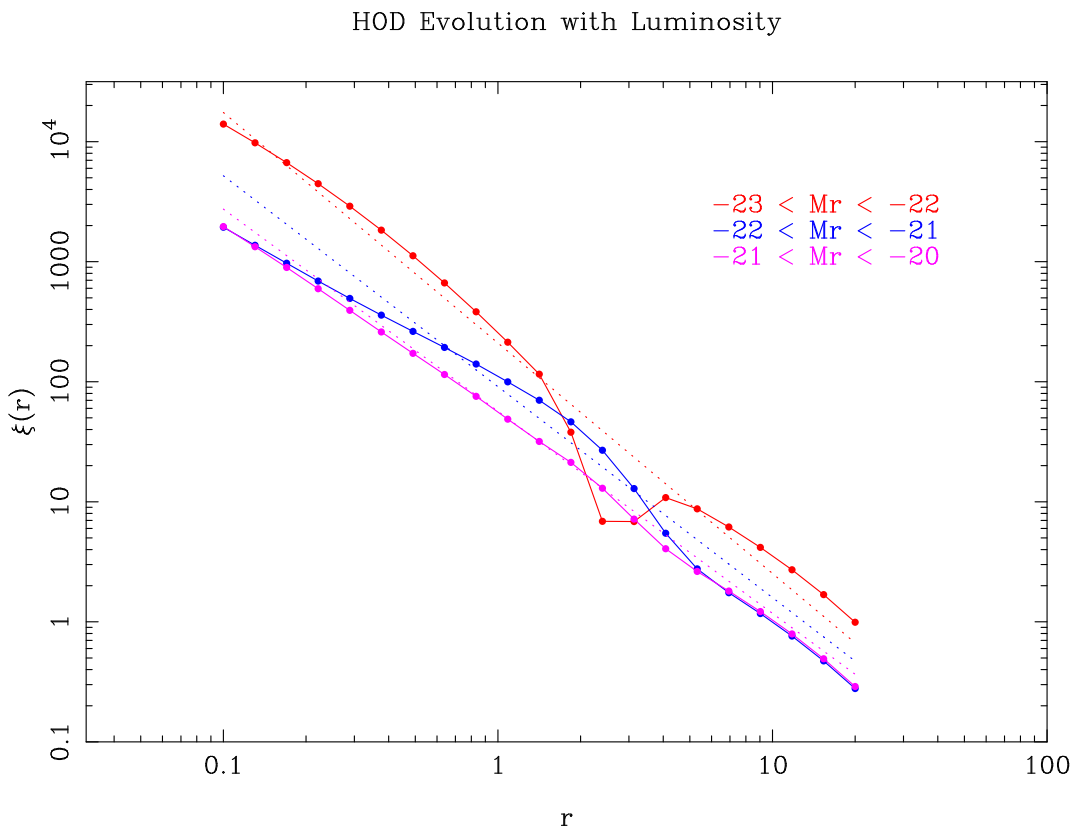


Figure 6.4 Evolution of best fit HOD $\xi(r)$ with absolute magnitude. The sample used was volume limited. The best fit parameters are listed in Table 6.2.

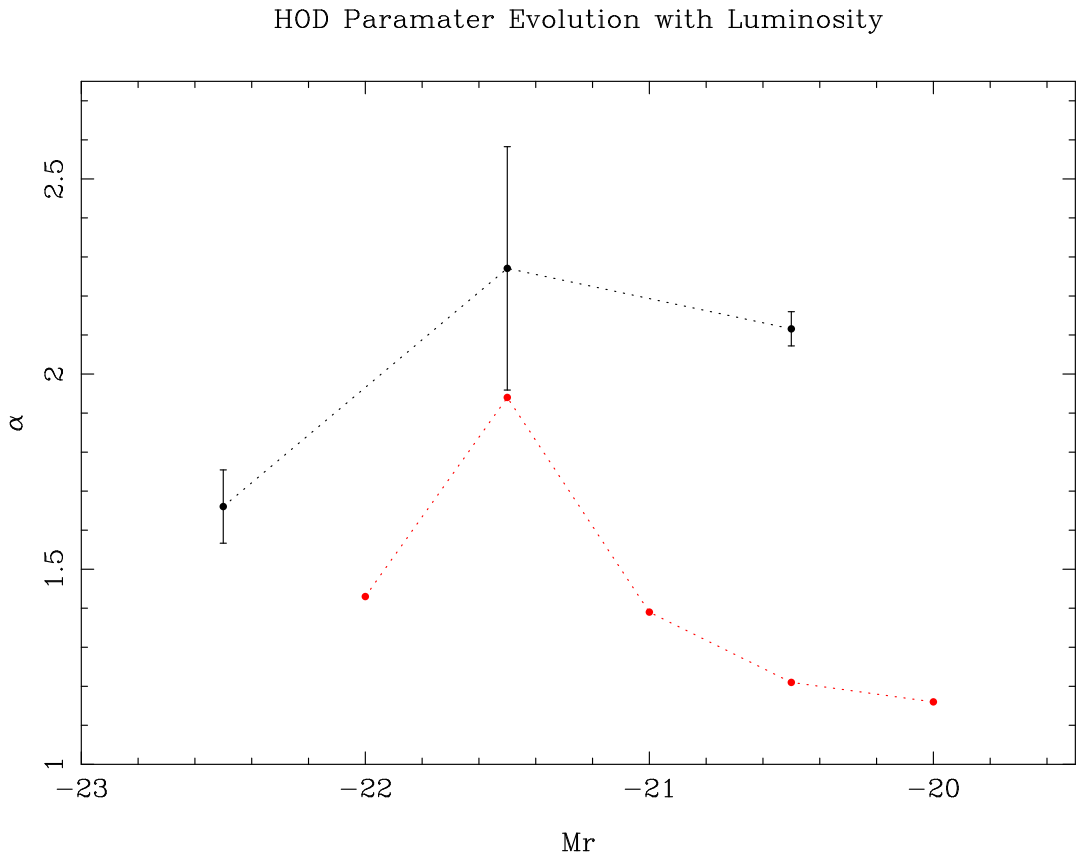


Figure 6.5 Evolution of HOD parameter α with absolute magnitude. The sample used was volume limited. The best fit parameters are listed in Table 6.2. The red line and points shows the results from Zehavi et al. (2005). Their sample is shallower in redshift and binned in M_r differently, so only qualitative comparisons should be made.

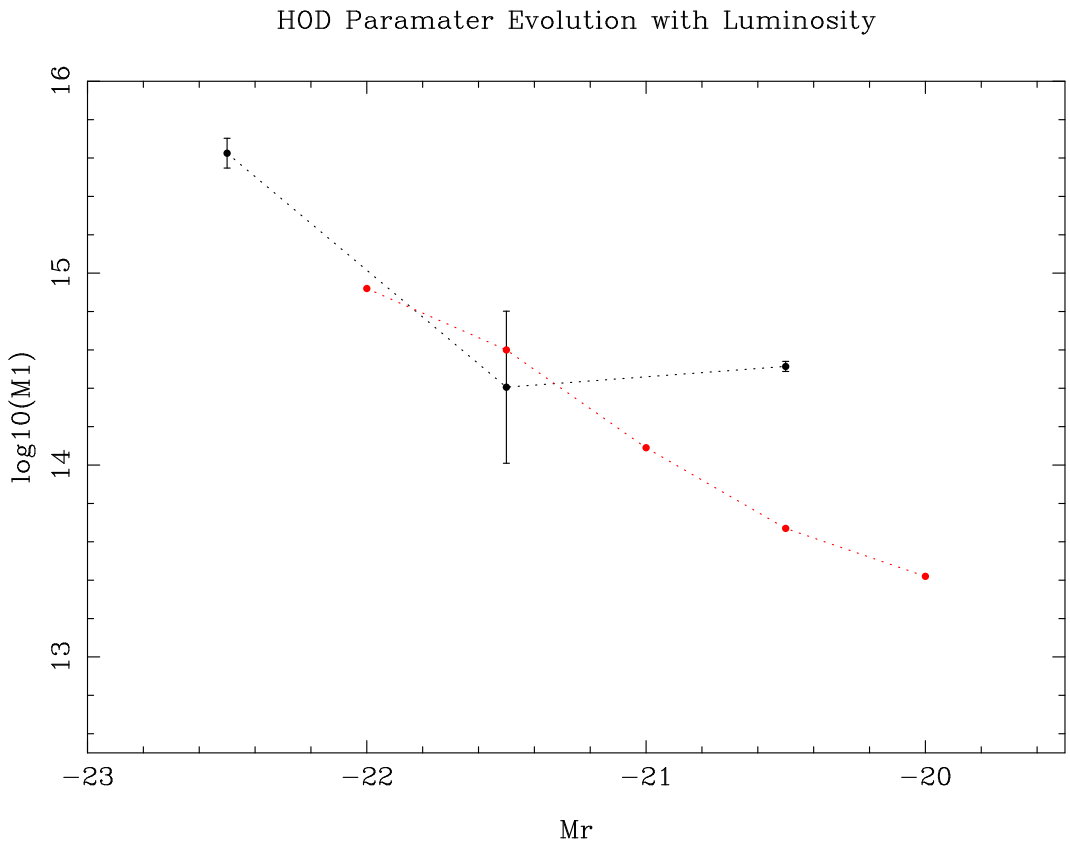


Figure 6.6 Evolution of HOD parameter M_1 with absolute magnitude. The sample used was volume limited. The best fit parameters are listed in Table 6.2. The red line and points shows the results from Zehavi et al. (2005). Their sample is shallower in redshift and binned in M_r differently, so only qualitative comparisons should be made.

Redshift	Avg. Redshift	α	$\log M_1$	$\log M_{\rm min}$
$0.1 < z < 0.2$	0.163	1.54 ± 0.132	15.31 ± 0.155	14.01
$0.2 < z < 0.3$	0.254	1.55 ± 0.127	15.13 ± 0.169	13.83

Table 6.1 Fit parameters for the HOD model in redshift bins as shown in Figure 6.1. Mass units are $M_\odot h^{-1}$.

Magnitude	Avg. Redshift	α	$\log M_1$	$\log M_{\rm min}$
$-23 < M_r < -22$	0.304	1.66 ± 0.094	15.63 ± 0.078	14.32
$-22 < M_r < -21$	0.291	2.27 ± 0.312	14.41 ± 0.396	13.10
$-21 < M_r < -20$	0.252	2.11 ± 0.044	14.51 ± 0.026	13.21

Table 6.2 Fit parameters for the HOD model in absolute magnitude bins as shown in Figure 6.4. Mass units are $M_\odot h^{-1}$.

7.0 CONCLUSIONS AND FUTURE WORK

Our goal was to measure the angular clustering evolution of galaxies in luminosity, type, and redshift using redshifts estimated from broadband photometry. Photometric redshifts require precise flux measurements, so we used the SDSS stripe 82 co-added imaging data set; this data set has co-added repeat scans of the same patch of the sky in order to obtain higher quality imaging. We developed a custom imaging pipeline to extract sources from this data with photometry errors less than 1% and no significant spatial zero point variation.

We classified objects as stars or galaxies by examining how point-like objects were as a function of magnitude. We estimated the size of objects using concentration, the difference of magnitudes at two different apertures. For a particular range of apparent magnitude, the distribution of concentrations has two visible populations, one for stars and one for galaxies; we applied the non-parametric mean shift algorithm to separate the two populations by locating the midpoint of the population peaks.

We computed photometric redshifts using the code of Benítez (2000) with three significant modifications. First, we computed a custom r band magnitude prior using SDSS and VVDS spectroscopic data. Second, we parameterized the type marginalized $p(z)$ so that we could describe the redshift of each galaxy as a probability density. Third, we used a "tweaked" template set that was optimized for our data. All of these improved the photometric redshift quality over the BPZ defaults.

We compare multiple methods for computing the true redshift distribution $\frac{\mathrm{d}n}{\mathrm{d}z}$ from a distribution of photometric redshifts. We used a naive histogram, a sum of $p(z)$ for each galaxy, the paramterization in Equation 4.9 with the median photometric redshift, and the method of Budavári et al. (2003) which convolves the $\frac{\mathrm{d}n}{\mathrm{d}z}$ from the luminosity function with an effective selection window. We demonstrate that these methods agree when we perform

a cut in M_r of roughly $M_* \pm 2$ as shown in Figure 5.4. This result demonstrates that we can accurately predict photometric redshifts and their errors for galaxies in a particular range of intrinsic luminosity.

We measured the angular correlation function $w(\theta)$ of galaxies. Our $w(\theta)$ result agrees with the closest available study (Budavári et al., 2003). We used Limber's equation to fit a power law to the real space correlation function $\xi(r)$ using our estimated $\frac{\mathrm{d}n}{\mathrm{d}z}$. Our results for r_0 are systematically higher than similar studies, though we qualitatively agree in the trend of r_0 increasing with luminosity. This discrepancy must be due to our estimate for $\frac{\mathrm{d}n}{\mathrm{d}z}$; in particular, the average photometric redshift error for our faintest bin has significant effect on the width of $\frac{\mathrm{d}n}{\mathrm{d}z}$ and thus r_0. We found that we can account for this discrepancy in 2 of our 3 luminosity bins by reducing our photometric redshift contamination by a factor of 4.

Finally, we related our $w(\theta)$ measurement to parameters from the halo occupation distribution to provide more physical insight into clustering evolution. We found that the characteristic mass of dark matter halos decreases with redshift and increases with luminosity while α only evolves with luminosity. The M_1 redshift evolution is consistent with the bottom-up formation of large scale structure in ΛCDM cosmology, and the M_1 luminosity evolution is consistent with semianalytic model results which find that the brightest galaxies are found in higher mass halos. However, our results do not improve upon those of Zehavi et al. (2005) due to the uncertainties in $\frac{\mathrm{d}n}{\mathrm{d}z}$ which prevented us from performing a direct comparison of $\xi(r)$ or the projected correlation function $w(r_p)$. Determining an accurate $\frac{\mathrm{d}n}{\mathrm{d}z}$ would significantly improve our constraints on the HOD.

One of the goals of this work was to obtain higher quality photometric redshifts using superior co-added imagery. While we did succeed in improving the quality of redshifts, the improvement was not enough to probe magnitudes fainter than $r = 21$, the same depth used by Budavári et al. (2003) with single epoch SDSS data. This severely limited both the number counts and the maximum redshift we could investigate, and as a result we were unable to improve on the results of Budavári et al. (2003). Further improving photometric redshifts is obviously a candidate for future work, though probing to $r = 22$ or fainter with just SDSS data will be difficult. However, new surveys such as UKIDSS will add infrared photometry for much of the footprint of SDSS, and adding additional magnitude measurements

to each galaxy can significantly improve photometric redshift results; photometric redshifts computed in this way are sometimes termed *super photozs*. Additionally, it may be possible to use empirical photometric redshifts, which have better error properties, for this purpose in the future, but a deeper spectroscopic sample is needed. Finally, the difficulty in properly determining $\frac{\mathrm{d}n}{\mathrm{d}z}$ from a photometric redshift distribution is an important outstanding problem that is just beginning to be addressed in the literature (Newman, 2008).

Star/galaxy classification is the second obvious area to improve. Devising an effective parametric classifier would allow for classification of fainter objects. Additionally, the probability that an object is a galaxy could be used as a weight in the $w(\theta)$ computation to further improve the measurement. Of particular importance to this data set is the performance of star/galaxy separation as a function of distance from the galactic plane, so this should be investigated thoroughly.

There appear to be additional systematics present in the apparent magnitude binned $w(\theta)$ plot in Figure 5.9, so further investigating the properties of the magnitude calibration would be worthwhile. In particular, studying how the convolution filter used affects source density in neighboring images should be enlightening, particularly for images near the galactic plane.

Finally, we note that future surveys such as LSST and improved photometric redshift techniques could significantly improve our constraints on galaxy formation models. LSST will offer more galaxies in more filters at greater depth, all of which would improve our results. The increased number counts would help smooth out $w(\theta)$ and reduce error bars, enabling us to probe the power law deviations at intermediate scales arising from the 1 to 2 halo transition. This would enable us to compare directly to the HOD models rather than using a power law approximation. The greater depth would improve our star/galaxy classification and reduce stellar contamination, further smoothing $w(\theta)$ and reducing errors in the measurements of γ and r_0. The addition of the y filter would provide one more point for our low resolution spectra for each galaxy and reduce photometric redshift degeneracies. Improved photoz codes would offer similar improvements, providing less contamination in our redshift bins and improved $\frac{\mathrm{d}n}{\mathrm{d}z}$ estimation. Thus, future developments in cosmology will provide even more insight into galaxy formation.

BIBLIOGRAPHY

J. K. Adelman-McCarthy, M. A. Agüeros, S. S. Allam, K. S. J. Anderson, S. F. Anderson, J. Annis, N. A. Bahcall, C. A. L. Bailer-Jones, I. K. Baldry, J. C. Barentine, T. C. Beers, V. Belokurov, A. Berlind, M. Bernardi, M. R. Blanton, J. J. Bochanski, W. N. Boroski, D. M. Bramich, H. J. Brewington, J. Brinchmann, J. Brinkmann, R. J. Brunner, T. Budavári, L. N. Carey, S. Carliles, M. A. Carr, F. J. Castander, A. J. Connolly, R. J. Cool, C. E. Cunha, I. Csabai, J. J. Dalcanton, M. Doi, D. J. Eisenstein, M. L. Evans, N. W. Evans, X. Fan, D. P. Finkbeiner, S. D. Friedman, J. A. Frieman, M. Fukugita, B. Gillespie, G. Gilmore, K. Glazebrook, J. Gray, E. K. Grebel, J. E. Gunn, E. de Haas, P. B. Hall, M. Harvanek, S. L. Hawley, J. Hayes, T. M. Heckman, J. S. Hendry, G. S. Hennessy, R. B. Hindsley, C. M. Hirata, C. J. Hogan, D. W. Hogg, J. A. Holtzman, S.-i. Ichikawa, T. Ichikawa, Ž. Ivezić, S. Jester, D. E. Johnston, A. M. Jorgensen, M. Jurić, G. Kauffmann, S. M. Kent, S. J. Kleinman, G. R. Knapp, A. Y. Kniazev, R. G. Kron, J. Krzesinski, N. Kuropatkin, D. Q. Lamb, H. Lampeitl, B. C. Lee, R. F. Leger, M. Lima, H. Lin, D. C. Long, J. Loveday, R. H. Lupton, R. Mandelbaum, B. Margon, D. Martínez-Delgado, T. Matsubara, P. M. McGehee, T. A. McKay, A. Meiksin, J. A. Munn, R. Nakajima, T. Nash, E. H. Neilsen, Jr., H. J. Newberg, R. C. Nichol, M. Nieto-Santisteban, A. Nitta, H. Oyaizu, S. Okamura, J. P. Ostriker, N. Padmanabhan, C. Park, J. J. Peoples, J. R. Pier, A. C. Pope, D. Pourbaix, T. R. Quinn, M. J. Raddick, P. Re Fiorentin, G. T. Richards, M. W. Richmond, H.-W. Rix, C. M. Rockosi, D. J. Schlegel, D. P. Schneider, R. Scranton, U. Seljak, E. Sheldon, K. Shimasaku, N. M. Silvestri, J. A. Smith, V. Smolčić, S. A. Snedden, A. Stebbins, C. Stoughton, M. A. Strauss, M. SubbaRao, Y. Suto, A. S. Szalay, I. Szapudi, P. Szkody, M. Tegmark, A. R. Thakar, C. A. Tremonti, D. L. Tucker, A. Uomoto, D. E. Vanden Berk, J. Vandenberg, S. Vidrih, M. S. Vogeley, W. Voges, N. P. Vogt, D. H. Weinberg, A. A. West, S. D. M. White, B. Wilhite, B. Yanny, D. R. Yocum, D. G. York, I. Zehavi, S. Zibetti, and D. B. Zucker. The Fifth Data Release of the Sloan Digital Sky Survey. ApJS, 172:634–644, October 2007. doi: 10.1086/518864.

I. K. Baldry, K. Glazebrook, J. Brinkmann, Ž. Ivezić, R. H. Lupton, R. C. Nichol, and A. S. Szalay. Quantifying the Bimodal Color-Magnitude Distribution of Galaxies. ApJ, 600: 681–694, January 2004. doi: 10.1086/380092.

C. M. Baugh and G. Efstathiou. The Three-Dimensional Power Spectrum Measured from the APM Galaxy Survey - Part One - Use of the Angular Correlation Function. MNRAS, 265:145–+, November 1993.

W. A. Baum. Photoelectric Magnitudes and Red-Shifts. In G. C. McVittie, editor, *Problems of Extra-Galactic Research*, volume 15 of *IAU Symposium*, pages 390–+, 1962.

N. Benítez. Bayesian Photometric Redshift Estimation. ApJ, 536:571–583, June 2000. doi: 10.1086/308947.

F. Bernardeau, S. Colombi, E. Gaztañaga, and R. Scoccimarro. Large-scale structure of the Universe and cosmological perturbation theory. Phys. Rep., 367:1–3, September 2002.

E. Bertin and S. Arnouts. SExtractor: Software for source extraction. A&AS, 117:393–404, June 1996.

M. R. Blanton, D. W. Hogg, N. A. Bahcall, J. Brinkmann, M. Britton, A. J. Connolly, I. Csabai, M. Fukugita, J. Loveday, A. Meiksin, J. A. Munn, R. C. Nichol, S. Okamura, T. Quinn, D. P. Schneider, K. Shimasaku, M. A. Strauss, M. Tegmark, M. S. Vogeley, and D. H. Weinberg. The Galaxy Luminosity Function and Luminosity Density at Redshift z = 0.1. ApJ, 592:819–838, August 2003. doi: 10.1086/375776.

M. Bolzonella, J.-M. Miralles, and R. Pelló. Photometric redshifts based on standard SED fitting procedures. A&A, 363:476–492, November 2000.

T. Budavári, A. J. Connolly, A. S. Szalay, I. Szapudi, I. Csabai, R. Scranton, N. A. Bahcall, J. Brinkmann, D. J. Eisenstein, J. A. Frieman, M. Fukugita, J. E. Gunn, D. Johnston, S. Kent, J. N. Loveday, R. H. Lupton, M. Tegmark, A. R. Thakar, B. Yanny, D. G. York, and I. Zehavi. Angular Clustering with Photometric Redshifts in the Sloan Digital Sky Survey: Bimodality in the Clustering Properties of Galaxies. ApJ, 595:59–70, September 2003. doi: 10.1086/377168.

T. Budavári, A. S. Szalay, J. Gray, W. O'Mullane, R. Williams, A. Thakar, T. Malik, N. Yasuda, and R. Mann. Open SkyQuery – VO Compliant Dynamic Federation of Astronomical Archives. In F. Ochsenbein, M. G. Allen, and D. Egret, editors, *Astronomical Data Analysis Software and Systems (ADASS) XIII*, volume 314 of *Astronomical Society of the Pacific Conference Series*, pages 177–+, July 2004.

J. S. Bullock, T. S. Kolatt, Y. Sigad, R. S. Somerville, A. V. Kravtsov, A. A. Klypin, J. R. Primack, and A. Dekel. Profiles of dark haloes: evolution, scatter and environment. MNRAS, 321:559–575, March 2001.

M. Carreira-Perpiñán. Gaussian mean shift is an EM algorithm. volume 29 of *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pages 767–776, May 2007.

A. L. Coil, J. A. Newman, N. Kaiser, M. Davis, C.-P. Ma, D. D. Kocevski, and D. C. Koo. Evolution and Color Dependence of the Galaxy Angular Correlation Function: 350,000 Galaxies in 5 Square Degrees. ApJ, 617:765–781, December 2004. doi: 10.1086/425676.

A. L. Coil, J. A. Newman, D. Croton, M. C. Cooper, M. Davis, S. M. Faber, B. F. Gerke, D. C. Koo, N. Padmanabhan, R. H. Wechsler, and B. J. Weiner. The DEEP2 Galaxy

Redshift Survey: Color and Luminosity Dependence of Galaxy Clustering at z˜1. ApJ, 672:153–176, January 2008. doi: 10.1086/523639.

G. D. Coleman, C.-C. Wu, and D. W. Weedman. Colors and magnitudes predicted for high redshift galaxies. ApJS, 43:393–416, July 1980. doi: 10.1086/190674.

C. A. Collins, R. C. Nichol, and S. L. Lumsden. The Edinburgh-Durham Southern Galaxy Catalogue. III - w(theta) from the full survey. MNRAS, 254:295–300, January 1992.

A. J. Connolly, I. Csabai, A. S. Szalay, D. C. Koo, R. G. Kron, and J. A. Munn. Slicing Through Multicolor Space: Galaxy Redshifts from Broadband Photometry. AJ, 110:2655–+, December 1995. doi: 10.1086/117720.

A. J. Connolly, R. Scranton, D. Johnston, S. Dodelson, D. J. Eisenstein, J. A. Frieman, J. E. Gunn, L. Hui, B. Jain, S. Kent, J. Loveday, R. C. Nichol, L. O'Connell, M. Postman, R. Scoccimarro, R. K. Sheth, A. Stebbins, M. A. Strauss, A. S. Szalay, I. Szapudi, M. Tegmark, M. S. Vogeley, I. Zehavi, J. Annis, N. Bahcall, J. Brinkmann, I. Csabai, M. Doi, M. Fukugita, G. S. Hennessy, R. Hindsley, T. Ichikawa, Ž. Ivezić, R. S. J. Kim, G. R. Knapp, P. Kunszt, D. Q. Lamb, B. C. Lee, R. H. Lupton, T. A. McKay, J. Munn, J. Peoples, J. Pier, C. Rockosi, D. Schlegel, C. Stoughton, D. L. Tucker, B. Yanny, and D. G. York. The Angular Correlation Function of Galaxies from Early Sloan Digital Sky Survey Data. ApJ, 579:42–47, November 2002. doi: 10.1086/342787.

A. Cooray and R. Sheth. Halo models of large scale structure. Phys. Rep., 372:1–129, December 2002.

I. Csabai, T. Budavári, A. J. Connolly, A. S. Szalay, Z. Győry, N. Benítez, J. Annis, J. Brinkmann, D. Eisenstein, M. Fukugita, J. Gunn, S. Kent, R. Lupton, R. C. Nichol, and C. Stoughton. The Application of Photometric Redshifts to the SDSS Early Data Release. AJ, 125:580–592, February 2003. doi: 10.1086/345883.

M. Davis, S. M. Faber, J. Newman, A. C. Phillips, R. S. Ellis, C. C. Steidel, C. Conselice, A. L. Coil, D. P. Finkbeiner, D. C. Koo, P. Guhathakurta, B. Weiner, R. Schiavon, C. Willmer, N. Kaiser, G. A. Luppino, G. Wirth, A. Connolly, P. Eisenhardt, M. Cooper, and B. Gerke. Science Objectives and Early Results of the DEEP2 Redshift Survey. In P. Guhathakurta, editor, *Discoveries and Research Prospects from 6- to 10-Meter-Class Telescopes II. Edited by Guhathakurta, Puragra. Proceedings of the SPIE, Volume 4834, pp. 161-172 (2003).*, volume 4834 of *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, pages 161–172, February 2003.

S. Dodelson, V. K. Narayanan, M. Tegmark, R. Scranton, T. Budavári, A. Connolly, I. Csabai, D. Eisenstein, J. A. Frieman, J. E. Gunn, L. Hui, B. Jain, D. Johnston, S. Kent, J. Loveday, R. C. Nichol, L. O'Connell, R. Scoccimarro, R. K. Sheth, A. Stebbins, M. A. Strauss, A. S. Szalay, I. Szapudi, M. S. Vogeley, I. Zehavi, J. Annis, N. A. Bahcall, J. Brinkman, M. Doi, M. Fukugita, G. Hennessy, Ž. Ivezić, G. R. Knapp, P. Kunszt, D. Q. Lamb, B. C. Lee, R. H. Lupton, J. A. Munn, J. Peoples, J. R. Pier, C. Rockosi, D. Schlegel, C. Stoughton, D. L. Tucker, B. Yanny, and D. G. York. The Three-dimensional Power

Spectrum from Angular Clustering of Galaxies in Early Sloan Digital Sky Survey Data. ApJ, 572:140–156, June 2002. doi: 10.1086/340225.

S. M. Fall. Galaxy correlations and cosmology. *Reviews of Modern Physics*, 51:21–43, January 1979.

M. Fukugita, T. Ichikawa, J. E. Gunn, M. Doi, K. Shimasaku, and D. P. Schneider. The Sloan Digital Sky Survey Photometric System. AJ, 111:1748–+, April 1996. doi: 10.1086/117915.

K. Fukunaga and L. D. Hostetler. The Estimation of a Gradient of a Density Function, with Applications in Pattern Recognition. volume 21 of *IEEE Trans. on Information Theory*, pages 32–40, January 1975.

M. Galassi, J. Davies, J. Theiler, B. Gough, G. Jungman, M. Booth, and F. Rossi. *GNU Scientific Library Reference Manual (2nd Ed.)*. Network Theory Ltd., 2006.

L. Gao, S. D. M. White, A. Jenkins, F. Stoehr, and V. Springel. The subhalo populations of ΛCDM dark haloes. MNRAS, 355:819–834, December 2004. doi: 10.1111/j.1365-2966.2004.08360.x.

L. Gao, V. Springel, and S. D. M. White. The age dependence of halo clustering. MNRAS, 363:L66–L70, October 2005. doi: 10.1111/j.1745-3933.2005.00084.x.

E. J. Groth and P. J. E. Peebles. Statistical analysis of catalogs of extragalactic objects. VII - Two- and three-point correlation functions for the high-resolution Shane-Wirtanen catalog of galaxies. ApJ, 217:385–405, October 1977.

J. E. Gunn, M. Carr, C. Rockosi, M. Sekiguchi, K. Berry, B. Elms, E. de Haas, Ž. Ivezić, G. Knapp, R. Lupton, G. Pauls, R. Simcoe, R. Hirsch, D. Sanford, S. Wang, D. York, F. Harris, J. Annis, L. Bartozek, W. Boroski, J. Bakken, M. Haldeman, S. Kent, S. Holm, D. Holmgren, D. Petravick, A. Prosapio, R. Rechenmacher, M. Doi, M. Fukugita, K. Shimasaku, N. Okada, C. Hull, W. Siegmund, E. Mannery, M. Blouke, D. Heidtman, D. Schneider, R. Lucinio, and J. Brinkman. The Sloan Digital Sky Survey Photometric Camera. AJ, 116:3040–3081, December 1998. doi: 10.1086/300645.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

G. Hinshaw, J. L. Weiland, R. S. Hill, N. Odegard, D. Larson, C. L. Bennett, J. Dunkley, B. Gold, M. R. Greason, N. Jarosik, E. Komatsu, M. R. Nolta, L. Page, D. N. Spergel, E. Wollack, M. Halpern, A. Kogut, M. Limon, S. S. Meyer, G. S. Tucker, and E. L. Wright. Five-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Data Processing, Sky Maps, and Basic Results. *ArXiv e-prints*, 803, March 2008.

D. W. Hogg. Distance measures in cosmology. *ArXiv Astrophysics e-prints*, May 1999.

D. W. Hogg, I. K. Baldry, M. R. Blanton, and D. J. Eisenstein. The K correction. *ArXiv Astrophysics e-prints*, October 2002.

J. D. Hudon and S. J. Lilly. The Clustering of Faint Galaxies and the Evolution of $\xi(r)$. ApJ, 469:519–528, October 1996.

A. Jenkins, C. S. Frenk, S. D. M. White, J. M. Colberg, S. Cole, A. E. Evrard, H. M. P. Couchman, and N. Yoshida. The mass function of dark matter haloes. MNRAS, 321: 372–384, February 2001.

A. L. Kinney, D. Calzetti, R. C. Bohlin, K. McQuade, T. Storchi-Bergmann, and H. R. Schmitt. Template Ultraviolet to Near-Infrared Spectra of Star-forming Galaxies and Their Application to K-Corrections. ApJ, 467:38–+, August 1996. doi: 10.1086/177583.

R. G. Kron. Photometry of a complete sample of faint galaxies. ApJS, 43:305–325, June 1980. doi: 10.1086/190669.

K. S. Krughoff and A. J. Connolly. WESIX. In M. J. Graham, M. J. Fitzpatrick, and T. A. McGlynn, editors, *The National Virtual Observatory: Tools and Techniques for Astronomical Research*, volume 382 of *ASP Conference Series*, pages 119–128. Astronomical Society of the Pacific, 2008.

S. D. Landy and A. S. Szalay. Bias and variance of angular correlation functions. ApJ, 412: 64–71, July 1993. doi: 10.1086/172900.

O. Le Fèvre, G. Vettolani, D. Maccagni, D. Mancini, A. Mazure, Y. Mellier, J. P. Picat, M. Arnaboldi, S. Bardelli, E. Bertin, G. Busarello, A. Cappi, S. Charlot, G. Chincarini, S. Colombi, M. Dantel-Fort, S. Foucaud, B. Garilli, L. Guzzo, A. Iovino, C. Marinoni, G. Mathez, H. McCracken, R. Pello, M. Radovich, V. Ripepi, P. Saracco, R. Scaramella, M. Scoreggio, L. Tresse, A. Zanichelli, G. Zamorani, and E. Zucca. Virmos-VLT deep survey (VVDS). In P. Guhathakurta, editor, *Discoveries and Research Prospects from 6- to 10-Meter-Class Telescopes II. Edited by Guhathakurta, Puragra. Proceedings of the SPIE, Volume 4834, pp. 173-182 (2003).*, volume 4834 of *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, pages 173–182, February 2003.

O. Le Fèvre, G. Vettolani, S. Paltani, L. Tresse, G. Zamorani, V. Le Brun, C. Moreau, D. Bottini, D. Maccagni, J. P. Picat, R. Scaramella, M. Scodeggio, A. Zanichelli, C. Adami, S. Arnouts, S. Bardelli, M. Bolzonella, A. Cappi, S. Charlot, T. Contini, S. Foucaud, P. Franzetti, B. Garilli, I. Gavignaud, L. Guzzo, O. Ilbert, A. Iovino, H. J. McCracken, D. Mancini, B. Marano, C. Marinoni, G. Mathez, A. Mazure, B. Meneux, R. Merighi, R. Pellò, A. Pollo, L. Pozzetti, M. Radovich, E. Zucca, M. Arnaboldi, M. Bondi, A. Bongiorno, G. Busarello, P. Ciliegi, L. Gregorini, Y. Mellier, P. Merluzzi, V. Ripepi, and D. Rizzo. The VIMOS VLT Deep Survey. Public release of 1599 redshifts to $I_{AB} \leq 24$ across the Chandra Deep Field South. A&A, 428:1043–1049, December 2004. doi: 10.1051/0004-6361:20048072.

H. Lin, H. K. C. Yee, R. G. Carlberg, S. L. Morris, M. Sawicki, D. R. Patton, G. Wirth, and C. W. Shepherd. The CNOC2 Field Galaxy Luminosity Function. I. A Description of Luminosity Function Evolution. ApJ, 518:533–561, June 1999. doi: 10.1086/307297.

R. Lupton, J. E. Gunn, Z. Ivezić, G. R. Knapp, and S. Kent. The SDSS Imaging Pipelines. In F. R. Harnden, Jr., F. A. Primini, and H. E. Payne, editors, *Astronomical Data Analysis Software and Systems X*, volume 238 of *Astronomical Society of the Pacific Conference Series*, pages 269–+, 2001.

S. J. Maddox, G. Efstathiou, and W. J. Sutherland. The APM Galaxy Survey - Part Two - Photometric Corrections. MNRAS, 246:433–+, October 1990.

V. J. Martínez and E. Saar. *Statistics of the Galaxy Distribution*. Chapman & Hall/CRC, 2002.

J. F. Navarro, C. S. Frenk, and S. D. M. White. A Universal Density Profile from Hierarchical Clustering. ApJ, 490:493–+, December 1997. doi: 10.1086/304888.

J. A. Newman. Calibrating Redshift Distributions Beyond Spectroscopic Limits with Cross-Correlations. *ArXiv e-prints*, 805, May 2008.

J. Neyman and E. L. Scott. A Theory of the Spatial Distribution of Galaxies. ApJ, 116: 144–+, July 1952.

P. J. E. Peebles. Statistical Analysis of Catalogs of Extragalactic Objects. I. Theory. ApJ, 185:413–440, October 1973.

P. J. E. Peebles. *Large Scale Structure of the Universe*. Princeton University Press, 1980.

P. J. E. Peebles. *Principles of Physical Cosmology*. Princeton University Press, 1993.

J. R. Pier, J. A. Munn, R. B. Hindsley, G. S. Hennessy, S. M. Kent, R. H. Lupton, and Ž. Ivezić. Astrometric Calibration of the Sloan Digital Sky Survey. AJ, 125:1559–1579, March 2003. doi: 10.1086/346138.

S. Schmidt. *Galaxy Evolution: The DRaGONS Survey and Luminosity Functions with Photometric Redshifts*. PhD thesis, University of Pittsburgh, Pittsburgh PA, United States, November 2007.

R. Scranton, D. Johnston, S. Dodelson, J. A. Frieman, A. Connolly, D. J. Eisenstein, J. E. Gunn, L. Hui, B. Jain, S. Kent, J. Loveday, V. Narayanan, R. C. Nichol, L. O'Connell, R. Scoccimarro, R. K. Sheth, A. Stebbins, M. A. Strauss, A. S. Szalay, I. Szapudi, M. Tegmark, M. Vogeley, I. Zehavi, J. Annis, N. A. Bahcall, J. Brinkman, I. Csabai, R. Hindsley, Z. Ivezic, R. S. J. Kim, G. R. Knapp, D. Q. Lamb, B. C. Lee, R. H. Lupton, T. McKay, J. Munn, J. Peoples, J. Pier, G. T. Richards, C. Rockosi, D. Schlegel, D. P. Schneider, C. Stoughton, D. L. Tucker, B. Yanny, and D. G. York. Analysis of Systematic Effects and Statistical Uncertainties in Angular Clustering of Galaxies from Early Sloan Digital Sky Survey Data. ApJ, 579:48–75, November 2002. doi: 10.1086/342786.

R. Scranton, K. S. Krughoff, and A. J. Connolly. STOMP Footprint Service. In M. J. Graham, M. J. Fitzpatrick, and T. A. McGlynn, editors, *The National Virtual Observatory:*

Tools and Techniques for Astronomical Research, volume 382 of *ASP Conference Series*, pages 85–98. Astronomical Society of the Pacific, 2008.

C. D. Shane and C. A. Wirtanen. The Distribution of Galaxies. *Lick Obs. Publications*, XXII Part I, 1967.

Y. Shen, M. A. Strauss, M. Oguri, J. F. Hennawi, X. Fan, G. T. Richards, P. B. Hall, J. E. Gunn, D. P. Schneider, A. S. Szalay, A. R. Thakar, D. E. Vanden Berk, S. F. Anderson, N. A. Bahcall, A. J. Connolly, and G. R. Knapp. Clustering of High-Redshift Quasars from the Sloan Digital Sky Survey. AJ, 133:2222–2241, May 2007. doi: 10.1086/513517.

R. K. Sheth and G. Tormen. Large-scale bias and the peak background split. MNRAS, 308: 119–126, September 1999.

M. Tegmark, X. Yongzhong, and R. Scranton. http://lahmu.phyast.pitt.edu/~scranton/SDSSPix/.

J. L. Tinker, D. H. Weinberg, Z. Zheng, and I. Zehavi. On the Mass-to-Light Ratio of Large-Scale Structure. ApJ, 631:41–58, September 2005. doi: 10.1086/432084.

H. Totsuji and T. Kihara. The Correlation Function for the Distribution of Galaxies. PASJ, 21:221–+, 1969.

R. H. Wechsler, A. R. Zentner, J. S. Bullock, A. V. Kravtsov, and B. Allgood. The Dependence of Halo Clustering on Halo Formation History, Concentration, and Occupation. ApJ, 652:71–84, November 2006. doi: 10.1086/507120.

S. D. M. White and M. J. Rees. Core condensation in heavy halos - A two-stage theory for galaxy formation and clustering. MNRAS, 183:341–358, May 1978.

C. N. A. Willmer, S. M. Faber, D. C. Koo, B. J. Weiner, J. A. Newman, A. L. Coil, A. J. Connolly, C. Conroy, M. C. Cooper, M. Davis, D. P. Finkbeiner, B. F. Gerke, P. Guhathakurta, J. Harker, N. Kaiser, S. Kassin, N. P. Konidaris, L. Lin, G. Luppino, D. S. Madgwick, K. G. Noeske, A. C. Phillips, and R. Yan. The Deep Evolutionary Exploratory Probe 2 Galaxy Redshift Survey: The Galaxy Luminosity Function to z˜1. ApJ, 647:853–873, August 2006. doi: 10.1086/505455.

H. K. C. Yee, M. J. Sawicki, R. G. Carlberg, H. Lin, S. L. Morris, D. R. Patton, G. D. Wirth, C. W. Shepherd, D. Ellingson, D. Schade, and R. Marzke. The CNOC2 Field Galaxy Redshift Survey. *Highlights of Astronomy*, 11:460–+, 1998.

I. Zehavi, M. R. Blanton, J. A. Frieman, D. H. Weinberg, H. J. Mo, M. A. Strauss, S. F. Anderson, J. Annis, N. A. Bahcall, M. Bernardi, J. W. Briggs, J. Brinkmann, S. Burles, L. Carey, F. J. Castander, A. J. Connolly, I. Csabai, J. J. Dalcanton, S. Dodelson, M. Doi, D. Eisenstein, M. L. Evans, D. P. Finkbeiner, S. Friedman, M. Fukugita, J. E. Gunn, G. S. Hennessy, R. B. Hindsley, Ž. Ivezić, S. Kent, G. R. Knapp, R. Kron, P. Kunszt, D. Q. Lamb, R. F. Leger, D. C. Long, J. Loveday, R. H. Lupton, T. McKay, A. Meiksin,

A. Merrelli, J. A. Munn, V. Narayanan, M. Newcomb, R. C. Nichol, R. Owen, J. Peoples, A. Pope, C. M. Rockosi, D. Schlegel, D. P. Schneider, R. Scoccimarro, R. K. Sheth, W. Siegmund, S. Smee, Y. Snir, A. Stebbins, C. Stoughton, M. SubbaRao, A. S. Szalay, I. Szapudi, M. Tegmark, D. L. Tucker, A. Uomoto, D. Vanden Berk, M. S. Vogeley, P. Waddell, B. Yanny, and D. G. York. Galaxy Clustering in Early Sloan Digital Sky Survey Redshift Data. ApJ, 571:172–190, May 2002. doi: 10.1086/339893.

I. Zehavi, D. H. Weinberg, Z. Zheng, A. A. Berlind, J. A. Frieman, R. Scoccimarro, R. K. Sheth, M. R. Blanton, M. Tegmark, H. J. Mo, N. A. Bahcall, J. Brinkmann, S. Burles, I. Csabai, M. Fukugita, J. E. Gunn, D. Q. Lamb, J. Loveday, R. H. Lupton, A. Meiksin, J. A. Munn, R. C. Nichol, D. Schlegel, D. P. Schneider, M. SubbaRao, A. S. Szalay, A. Uomoto, and D. G. York. On Departures from a Power Law in the Galaxy Correlation Function. ApJ, 608:16–24, June 2004. doi: 10.1086/386535.

I. Zehavi, Z. Zheng, D. H. Weinberg, J. A. Frieman, A. A. Berlind, M. R. Blanton, R. Scoccimarro, R. K. Sheth, M. A. Strauss, I. Kayo, Y. Suto, M. Fukugita, O. Nakamura, N. A. Bahcall, J. Brinkmann, J. E. Gunn, G. S. Hennessy, Ž. Ivezić, G. R. Knapp, J. Loveday, A. Meiksin, D. J. Schlegel, D. P. Schneider, I. Szapudi, M. Tegmark, M. S. Vogeley, and D. G. York. The Luminosity and Color Dependence of the Galaxy Correlation Function. ApJ, 630:1–27, September 2005. doi: 10.1086/431891.

A. Zentner. The Halo Model. In preparation, 2008.

Z. Zheng, A. L. Coil, and I. Zehavi. Galaxy Evolution from Halo Occupation Distribution Modeling of DEEP2 and SDSS Galaxy Clustering. ApJ, 667:760–779, October 2007. doi: 10.1086/521074.