Title
Searching for Quasars and Beyond

Author
Kirkpatrick, Jessica Ann

Publication Date
2012

Peer reviewed|Thesis/dissertation

Searching for Quasars and Beyond

by

Jessica Ann Kirkpatrick

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Physics

in the

Graduate Division
of the
University of California, Berkeley

Committee in charge:
Doctor David Schlegel, Co-chair
Professor Saul Perlmutter, Co-chair
Professor Adrian Lee
Professor Joshua Bloom

Fall 2012

Searching for Quasars and Beyond

Abstract

Searching for Quasars and Beyond

by

Jessica Ann Kirkpatrick

Doctor of Philosophy in Physics

University of California, Berkeley

Doctor David Schlegel, Co-chair
Professor Saul Perlmutter, Co-chair

The SDSS-III Baryon Oscillation Spectroscopic Survey (BOSS), a five-year spectroscopic survey of 10,000 deg^2, achieved first light in late 2009. One of the key goals of BOSS is to measure the signature of baryon acoustic oscillations (BAO) in the distribution of Lyα absorption from the spectra of a sample of $\sim$150,000 $z > 2.2$ quasars in conjunction with measuring the redshifts of 1.6 million luminous red galaxies with high completeness to $i \approx 19.9$ at $z \approx 0.7$. One of the biggest challenges in achieving this goal is an efficient target selection algorithm for quasars in the redshift range $2.2 < z < 3.5$, where their colors tend to overlap those of the far more numerous stars. During the first year of the BOSS survey, quasar target selection methods were developed and tested to meet the requirement of delivering at least 15 quasars deg^{-2} in this redshift range, with a goal of 20, out of 40 targets deg^{-2} allocated to the quasar survey. To achieve these surface densities, the magnitude limit of the quasar targets was set at $g \leq 22.0$ or $r \leq 21.85$.

In this thesis I present a new method for quasar target selection using photometric fluxes and a Bayesian probabilistic approach. For our purposes I target quasars using Sloan Digital Sky Survey (SDSS) photometry to a magnitude limit of $g = 22$. The efficiency and completeness of this technique is measured using the Baryon Oscillation Spectroscopic Survey (BOSS) data, taken in 2010. This "likelihood" technique was used for the uniformly selected (CORE) sample of targets in BOSS year one spectroscopy to be realized in the 9th SDSS data release. When targeting at a density of 40 objects deg^{-2} (the BOSS quasar targeting density) the efficiency of this technique in recovering $z > 2.2$ quasars is 40%. The completeness compared to all quasars identified in BOSS data is 65%.

An extension of the "likelihood" technique is also described. This *SDSS-XDQSO* technique

builds models of the distributions of stars and quasars in flux space down to the flux limit by applying the extreme-deconvolution method to estimate the underlying density. I convolve this density with the flux uncertainties when evaluating the probability that an object is a quasar. This approach results in a targeting algorithm that is more principled, more efficient, and faster than other similar methods.

With BOSS's new catalog of quasar and galaxy data, exciting new science can be done. Whether luminous quasars reside in dark matter halos of the same mass and accrete at different rates, or live in halos of different masses and accretion is near the Eddington limit, is still an open question. Here, I present measurements of the luminosity-dependence of quasar clustering, using QSO data from the Sloan Digital Sky Survey (SDSS) Data Release 7, 2dF-SDSS LRG and QSO Survey (2SLAQ), and SDSS-III: Baryon Oscillation Spectroscopic Survey (BOSS). In my quasar sample I have 3100 spectroscopically confirmed quasars with a redshift range of ($0.5 < z < 1.0$), luminosity range of ($-27 < M < -21$), down to $i-$band 22.14. In my galaxy sample I have 5.23 million photometric galaxies brighter than $i-$band = 23.50, selected from the CFHT (Canada-France-Hawaii Telescope) Survey of Stripe-82 (CS82). The cross-correlation is well described by a power law with slope 1.77 ± 0.1 and $r_0 = 5.05 \pm 0.14\ h^{-1}$ Mpc, which is consistent with previous findings. I determine a large-scale quasar bias, $b_{QSO} = 1.46 \pm 0.18$, at redshift $z = 0.7$. When I divide the quasar sample into low/high luminosity samples I find luminosity depended quasar clustering at a $4.56\,\sigma$ significance level.

To my parents, for making me believe
that I can accomplish anything.

Contents

List of Figures

List of Tables

Acknowledgments

There were many people who supported me over the last seven years, without which I never could have finished this thesis. I am so grateful for my amazing support network.

In my academic life, first and foremost I would like to thank Alexia Schulz and Nic Ross. Your patience, guidance, support, and insight were invaluable. I am a doctor because of you and I will be forever grateful for your mentorship and friendship. I am also very thankful for the opportunity to collaborate on projects with some truly incredible scientists: Martin White, Adam Myers, Joe Hennawi, Alexie Leauthaud, Nikhil Padmanabhan, and Erin Sheldon – I learned so much from you all. Berkeley professors Adrian Lee, Ori Ganor, Marjorie Shapiro, and JD Jackson encouraged me to "keep at it" even in the face of failure. Fellow classmates Ben Kain, Dan Brooks, Hannah Fakhouri, Michelle Yong, and Roger Mong – you all helped me so much with my classes and prelims, I never would have passed without you. Donna Sakima and Anne Takizawa shielded me from a lot of Berkeley's bureaucracy. You both should be paid much, much more. My academic siblings Eric Huff, Matt George, Victoria Martin, Vaishali Bhardwaj, Lauren Anderson, and Demitri Muna made research much more enjoyable. Finally, I would like to thank my co-advisors David Schlegel and Saul Perlmutter. I wish we had spent more time together, there is still so much I could learn from you. You are both a rare combination of extremely brilliant and incredibly likable. Thank you for your support.

There are many scientists who mentored me throughout my life. I would like to especially thank Andrew Elby for teaching me to love physics, and Dan Snowden-Ifft for teaching me to love research. George Schmiedeshoff, Alec Schramm, Mark Garcia, Betina Pavri, and Bruce Downing also played major roles in my scientific development – thank you for encouraging me to pursue physics.

It takes a village to raise a PhD. I never would have survived graduate school without my amazing emotional support network. Angie Little, Yolanda Hagar Slichter, Sybil Lockhart, and Tim Ruckle – you are so important to me. Thank you for your unconditional love and support. I am so blessed to have you in my life. Mary Ann Leff and my lunchtime friends at the YWCA – our meetings were essential to me. Thank you for helping me do something I never could have done by myself.

The greatest gift of graduate school are the amazing people I met. Kevin Young, Michelle Young, Ansley Scott, Mike Weber, Dan Slichter, Jennie Guzman, Josh Meyers, Lauren Tompkins, Kyle Barbary, Laura Garcia, Henry Garcia, Laura Kogler, Julie Oppermann, and Eric Jabart – thank you for putting up with my inappropriate jokes and constant complaining. You are the best friends a person could have. I also had some really wonderful housemates during graduate school. Ben Hooberman, Jacob Becklund, Ben Gellar, Peter Battaglino, Kristin Maravilla, Elise Post, Andrea Lee, and Frankie Li – thank you for making my home life an oasis of relaxation, laughter, and joy. I would like to thank my high school friends Annalise Vislocky, Mariko Blakemore, and Devorah Kasarjian Bajkowski for continuing to love me even though I am a big nerd. Sorry about all the Fisher matrices talk.

Finally, I want to thank my incredibly loving and supportive family. Mom – you are my inspiration. Your graceful example taught me that brains and femininity are not mutually exclusive. Having a mother with a PhD was essential in creating a tangible belief that I too could become a doctor. Dad – I got my math brain from you. You always helped me see the lessons and skills I was learning in any situation. Thank you for simultaneously reminding me that I can do anything and that I am nothing but trouble. Kathryn – you are my number one fan and supporter. This never would have been possible without you cheering for me on the sidelines. You are my best friend and it was such a gift to see you so many times during my research work. I'm so grateful to have been born into this wonderful family. Toni and Cleo – your support never wavered. Writing my thesis was much more enjoyable with you curled up on my lap. Adam – We did it! Meeting you made it all worth it.

Preface

Three of the six chapters in this thesis represent already published material. The text here closely follows the published version. As these were all multiple author papers, I list the full published references below. The co-authors of these papers have given me permission to reproduce these results in this thesis.

Chapters (2) & (4):
Title: The SDSS-III Baryon Oscillation Spectroscopic Survey: Quasar Target Selection for Data Release Nine (Ross et al. 2012).
Authors: Nicholas P. Ross, Adam D. Myers, Erin S. Sheldon, Christophe Yche, Michael A. Strauss, Jo Bovy, Jessica A. Kirkpatrick, Gordon T. Richards, Eric Aubourg, Michael R. Blanton, W. N. Brandt, William C. Carithers, Rupert A.C. Croft, Robert da Silva, Kyle Dawson, Daniel J. Eisenstein, Joseph F. Hennawi, Shirley Ho, David W. Hogg, Khee-Gan Lee, Britt Lundgren, Richard G. McMahon, Jordi Miralda-Escude, Nathalie Palanque-Delabrouille, Isabelle Paris, Patrick Petitjean, Matthew M. Pieri, James Rich, Natalie A. Roe, David Schiminovich, David J. Schlegel, Donald P. Schneider, Anže Slosar, Nao Suzuki, Jeremy L. Tinker, David H. Weinberg, Anya Weyant, Martin White, & Michael Wood-Vasey.
Journal: ApJS Volume 199, Issue 1, article id. 3 (2012).
Link: arXiv:1105.0606v1 [astro-ph.CO]

Chapter (3):
Title: A Simple Likelihood Method for Quasar Target Selection (Kirkpatrick et al. 2011)
Authors: Jessica A. Kirkpatrick, David J. Schlegel, Nicholas P. Ross, Adam D. Myers, Joseph F. Hennawi, Erin S. Sheldon, Donald P. Schneider, and Benjamin A. Weaver.
Journal: The Astrophysical Journal, Volume 743, Issue 2, article id. 125 (2011).
Link: arXiv:1104.4995v2 [astro-ph.CO]

Title: Think Outside the Color Box: Probabilistic Target Selection and the SDSS-XDQSO Quasar Targeting Catalog (Bovy et al. 2011b)
Authors: Jo Bovy, Joseph F. Hennawi, David W. Hogg, Adam D. Myers, Jessica A. Kirkpatrick, David J. Schlegel, Nicholas P. Ross, Erin S. Sheldon, Ian D. McGreer, Donald P. Schneider, and Benjamin A. Weaver
Journal: The Astrophysical Journal, Volume 729, Issue 2, article id. 141 (2011).

Link: arXiv:1011.6392v2 [astro-ph.CO]

Although the work described in Chapter (5) is currently unpublished, there are numerous contributions that made it possible. Martin White, David Schlegel, Nicholas Ross, and Alexie Leauthaud all advised me in this work and contributed to the analysis. On the CFHT Team, Hendrik Hildebrandt, Jean-Paul Kneib, Martin Makler, and Ludovic Van Waerbeke all helped with the reduction of and mask for the CFHT data. Ian McGreer provided help with the black-hole mass calculations. Adam Myers and Yue Shen both helped with the clustering analysis.

Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III web site is http://www.sdss3.org/.

SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, University of Cambridge, University of Florida, the French Participation Group, the German Participation Group, the Instituto de Astrosica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington, and Yale University.

The observations reported here were obtained in part at the MMT Observatory, a facility operated jointly by the Smithsonian Institution and the University of Arizona. Some MMT telescope time was granted by NOAO (program 2008B-0282), through the Telescope System Instrumentation Program (TSIP). TSIP is funded by NSF.

The 2SLAQ Redshift Survey was made possible through the dedicated efforts of the staff at the Anglo-Australian Observatory, both in creating the 2dF instrument and supporting it on the telescope.

This work was supported by the Director, Office of Science, Office of High Energy Physics, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231

Chapter 1

Introduction

"The astronomical discovery that ordinary matter comprises only 4% of the mass-energy density of the universe is the most dramatic in cosmology in the past several decades, and it is clear that new physics will be needed to explain the non-baryonic dark matter and dark energy. At the same time, data from particle physics suggests a corresponding need for physics beyond the Standard Model. Discovering and understanding the fundamental constituents and interactions of the universe is the common subject of particle physics and cosmology. In recent years, the frontier questions in both fields have become increasingly intertwined. The emerging common themes that astrophysics and particle physics are addressing have crystallized a new physics-astronomy community." — LSST Science Book

1.1 Quasars as Astrophysical Tools

Quasi-stellar radio sources (quasars/QSOs) are very energetic and distant active galactic nuclei. Quasars have been vital astrophysical tools since their discovery over forty years ago (Matthews & Sandage 1963; Schmidt 1963). Individually, quasars provide examples of energetic and extreme physics, e.g., the broad absorption line phenomenon, first characterized by Bahcall & Goldsmith (1971); Weymann et al. (1981, 1991), and more recently by Gibson et al. (2009); Allen et al. (2011). Quasars are also tracers of structure at large scales (Croom et al. 2005; Myers et al. 2006, 2007a,b; Shen et al. 2007, 2009; Ross et al. 2009) and small scales (Hennawi et al. 2006a; Myers et al. 2007b, 2008), providing constraints on their host dark matter halos and the quasar phenomenon (Shankar 2009; Shen 2009). Indeed, due to their intrinsic luminosity ($\sim$ few 10^{46} erg s^{-1}) they can be seen to high ($z \sim 6$) redshifts (Fan et al. 2006b; Willott et al. 2010), enabling constraints to be placed on the epoch of

reionization (EoR; Fan et al. 2006a). Furthermore, there is now strong evidence that there is a link between luminous AGN activity and the formation and evolution of massive galaxies (see Cattaneo et al. 2009, for a review).

Quasars have proved to be cosmological "backlights" and the absorption features blueward of the Lyman-α (1216 Å) emission line have been seen in traditionally bright, quasar spectra (Rauch 1998, and references therein). The absorption between the Lyman-α and the Lyman-limit (912 Å), caused by neutral hydrogen, is regularly referred to as the Lyman-α forest (LyαF; Lynds 1971). The region between Lyman-α and Lyman-β is typically used for LyαF science because it is not confused by Lyman-β absorption lines. Meiksin (2009) has a contemporary review, with the importance of the LyαF in cosmology discussed by Croft et al. (1998) and White (2003), and measured, via the ("1-D") Power Spectrum, by Croft et al. (2002) and McDonald et al. (2006).

The Lyα-forest begins to be redshifted into in the near-UV/blue bands at $z \sim 2$. The atmosphere extincts $> 50\%$ of the light below 3500 Å, and CCD detectors are also less efficient at those shorter wavelengths. If LyαF lines of sight are observed at high spatial density, then they can be used to perform precise measurements of the expansion rate and distance scale (and thus constrain cosmological world models, e.g. White 2003; McDonald 2003; McDonald & Eisenstein 2007; McQuinn & White 2011).

1.2 Quasar Target Selection

Previous quasar surveys, such as the Sloan Digital Sky Survey (SDSS; Schneider et al. 2010) and the Ango-Australian Telescope (AAT) Two-Degree Field (2dF) QSO Redshift Survey (2QZ; Croom et al. 2004), have historically performed quasar target selection by searching for relatively bright quasars ($i < 19.1$, $z < 3$ objects for SDSS). However previous methods, either the traditional "UVX" – selecting star-like objects with unusually blue broadband colors, Sandage (1965) or "color-boxes" Richards et al. (2006); Croom et al. (2009), § 3.3.1, begin to fail at fainter magnitudes because they do not fully account for the large errors at lower fluxes. This motivated our development of a selection technique which better handles the flux errors as one approaches the flux limit.

Furthermore, at ($z > 2$) redshifts, broad-band optical color selections fail, since the colors of these "mid-z" quasars are similar to those of stars (in particular early A and F stars, Fan 1999, Richards et al. 2002) and the quasars "pass over the stellar locus". Simultaneously, quasars become much fainter e.g., an $M_g = -23$ quasar at $z = 2$, has g-band ~ 21.7, and thus the photometric errors are consequently larger.

With the final data release of the SDSS (Abazajian et al. 2009), an incredibly powerful database of over 100,000 spectroscopically confirmed quasars (Schneider et al. 2010) and

1,000,000 photometric potential quasar targets (Richards et al. 2009a) is now publicly available. In particular, the advent of a mega-quasar catalog is impressive, and the Richards et al. (2009a) photometric sample consists of the largest sample of photometrically classified $z > 2.4$ quasars in the literature. This catalog has also been the dataset of several cosmology studies including: studies of cosmic magnification bias (Scranton et al. 2005a); the investigation into the clustering of quasars on large (Myers et al. 2006, 2007a) and small (Hennawi et al. 2006a; Myers et al. 2007b, 2008) scales; the Integrated Sachs-Wolfe (ISW) effect (Giannantonio et al. 2006; Giannantonio et al. 2008); binary quasars (Hennawi et al. 2006b, 2010; Prochaska & Hennawi 2009; Shen et al. 2010); and the near infrared photometric properties of quasars (Peth et al. 2011).

1.3 Searching for Quasars

The SDSS-III: Baryon Oscillation Spectroscopic Survey (BOSS; Eisenstein et al. 2011; Dawson et al. 2012) is specifically targeting $z > 2.2$ QSOs in order to observe 150,000 LyαF lines of sight. The key aim of BOSS is to measure the absolute cosmic distance scale and expansion rate with percent-level precision at three distinct cosmological epochs: redshifts $z = 0.3, 0.6$ using luminous red galaxies (LRGs) and $z \sim 2.5$ using the LyαF, via the baryon acoustic oscillation (BAO) technique. BOSS has dedicated 40 fibers per deg^2 to QSO target selection for measuring the BAO signal.

The BOSS LyαF/Quasar Survey targets objects thought to be $z > 2.2$ quasars to perform a LyαF BAO measurement. Since the foreground LyαF is independent of the background quasar, there is freedom to use multiple selection methods without biasing the BAO results. The methods used for BOSS targeting include the "Kernel Density Estimator" (KDE; Richards et al. 2004), an "extreme-deconvolution" method (XDQSO; Bovy et al. 2011b), and a Neural Network method (NN; Yèche et al. 2010). The BOSS QSO target selection (Ross et al. 2012) combines all these different methods with different photometric catalogs (SDSS, UKIDSS, GALEX and quasars found using their flux time-variability information (Palanque-Delabrouille et al. 2011).

My thesis has involved extensive work on the BOSS quasar target selection team. I developed a new target selection method (see Chapter 3) which was the primary target selection method for BOSS during commissioning and year-one.

1.4 And Beyond

There is tremendous interest today in understanding the formation of black holes and QSOs to inform our understanding of galaxy formation. Central black hole masses (m_{bh}) correlate with galaxy masses and therefore (indirectly) with dark matter halo masses. However, using quasars, we can only roughly estimate the m_{bh} because the quasar luminosities (which we measure directly) also depend on accretion rate. This creates a very interesting problem to solve if we want to trace the build-up of m_{bh} in the universe over cosmic time; understand the role that massive black holes play in quenching star formation in galaxies; or use quasars as precision tracers of large scale structure. There are competing theories for quasar accretion rates: Some people (e.g., Ferrarese & Merritt 2000) say that the gas feeding should scale with galaxy bulge mass and therefore m_{bh}. This predicts that quasar luminosity should correlate with halo mass and therefore that quasar clustering should depend strongly on quasar luminosity. Others (e.g., Hopkins et al. 2006) say that gas feeding is highly variable, in which case quasar clustering should not depend strongly on quasar luminosity.

In recent years, quasar catalogs have grown big enough in terms of the absolute number, redshift, and luminosity range to accurately understand their clustering properties (Porciani et al. 2004; Croom et al. 2005; Porciani & Norberg 2006; Hennawi et al. 2006a; Myers et al. 2007a,b; Shen et al. 2007; da Ângela et al. 2008) and distinguish between the above scenarios. However, because quasars are extremely rare ($\bar{n} \sim 10^{-6}\, h^3\, \mathrm{Mpc}^{-3}$ at $z \sim 0.5$), their clustering signal is dominated by shot-noise from Poisson fluctuations in the counts of objects.

With BOSS's new catalog of quasar and galaxy data, exciting new science can be done. For the first time there is enough quasar data to perform statistically significant measurements of luminosity dependent quasar clustering. As part of my thesis work, I have detected (for the first time ever) a luminosity dependent clustering signal (see Chapter 5).

1.5 Layout of Thesis

This thesis is has the following structure. In Chapter (2) I discuss an overview of quasar target selection and the methods used for BOSS Quasar Target Selection. In Chapter (3) I discuss in detail the two primary (CORE) BOSS quasar target selection algorithms: Likelihood and it's extension XDQSO. In Chapter (4) I discuss the results of BOSS quasar target selection. In Chaper (5) is discuss a new measurement I performed with the BOSS quasars which discovered luminosity dependent quasar clustering. In Chapter (6) I discuss the major findings and conclusions of this thesis.

Chapter 2

Quasar Target Selection

The BOSS quasar survey will pioneer a novel cosmological probe by mapping the large scale distribution of neutral hydrogen by absorption in the Lyα forest. Measurements of BAO in the three-dimensional correlation function in this neutral hydrogen will provide the first direct measurements of angular diameter distance at redshifts $z > 2$. Constructing a map of the distant Universe using QSOs as sight-lines is currently limited by the ability to observe a large enough density of quasars. According to McQuinn & White (2011) in order to *perfectly* map the BAO feature on $100h^{-1}$ Mpc scales, a surface density of $\sim$35 quasars deg^{-2} in the redshift range of $2 < z < 3$ is required.

The BOSS team estimates a minimum of 15 quasars deg^{-2} at redshifts $2.2 < z < 3.5$ and $g < 22.0$ are required to make this measurement. The precision of the BAO measurement shows an approximately linear improvement with the surface density of quasars, and the BOSS team's goal is to target as many quasars as possible. Fortunately, because the quasars are nearly uncorrelated with the intervening density field, Lyα measurements are insensitive to the exact details of quasar target selection and do not require the same uniform sample that is essential to the galaxy BAO measurement.

The BOSS LyαF/Quasar Survey will target objects thought to be $z > 2.2$ quasars to perform a LyαF BAO measurement. Since the foreground LyαF is independent of the intrinsic properties of the background quasar, there is freedom to use multiple selection methods without biasing the BAO results. The methods used for BOSS targeting include the "Kernel Density Estimator" (KDE; Richards et al. 2004), the "Likelihood" method (Kirkpatrick et al. 2011, and Chapter 3) an "Extreme-Deconvolution" method (XDQSO; Bovy et al. 2011b, and Chapter 3), and a Neural Network method (NN; Yèche et al. 2010). The BOSS QSO target selection used for the first year of observations (Ross et al. 2012) combines all these different methods, with different photometric catalogs such as the five-band optical photometry of SDSS (York et al. 2000), the infrared photometry of UKIDSS (Lawrence et al. 2007), the

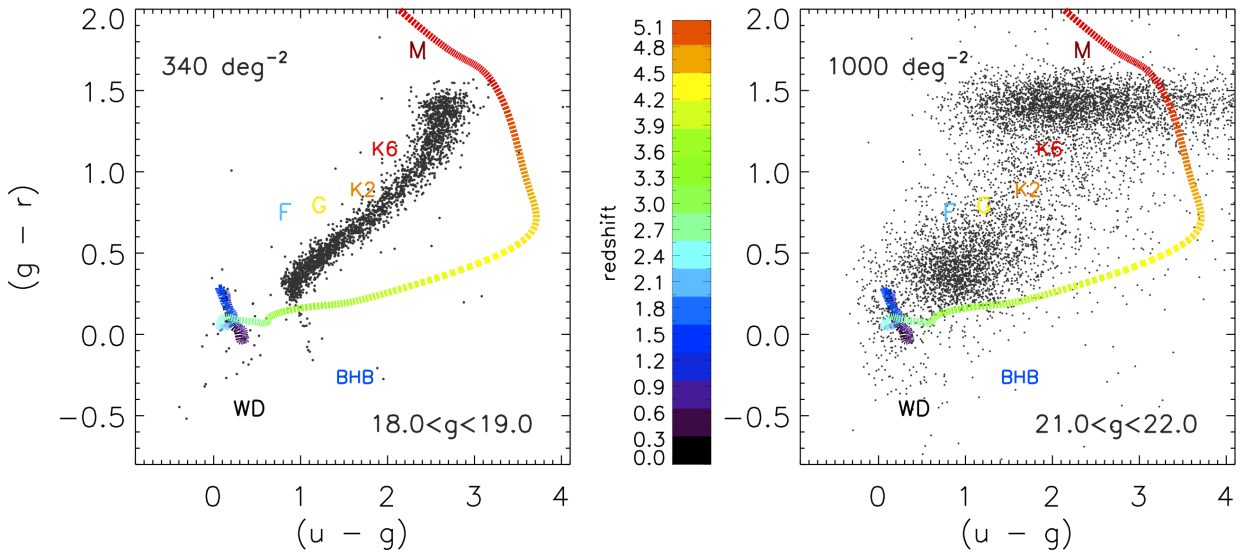


Figure 2.1: Color-color diagrams of point sources drawn from 7 deg^2 (the BOSS spectrograph field of view) in the SDSS photometric database. *(Left)* 2,400 objects with $18.0 < g < 19.0$, and *(Right)* 7,000 objects with $21.0 < g < 22.0$. Most of the objects shown are stars; low-redshift ($z < 2.2$) quasars lie preferentially in the region $u - g < 0.6, g - r > 0$ where very few stars are found. At $z > 2.2$, quasars become systematically redder in $u - g$ as the Lyα forest moves into the u-band and Lyα emission moves into g. At $z \sim$2.7, quasars have colors similar to those of blue horizontal branch (BHB) stars. The larger photometric errors at faint magnitudes broaden the stellar locus considerably (especially in the u-band for the reddest stars, which gives rise to the spread at $g - r \sim 1.5$), illustrating the challenges involved in selecting faint objects by their colors. Tracks for the quasar locus, as presented in Bovy et al. (2011a) are also shown, with the corresponding redshift given by the color-bar legend. Approximate surface densities are quoted, and stellar classifications are given as a guide.

ultraviolet photometry of GALEX (Martin et al. 2005) and quasars found using their flux time-variability information (Palanque-Delabrouille et al. 2011).

The BOSS survey requirements are for spectroscopy of 15 or more $z > 2.2$ quasars deg^{-2} (150,000 quasars over the BOSS footprint of 10,000 deg^2) (Eisenstein et al. 2011). Combining calculations from McDonald & Eisenstein (2007) and McQuinn & White (2011) with the luminosity function given by Jiang et al. (2006), it was found that targeting to a magnitude of $g < 22$ *with perfect completeness* will provide a surface density of $\sim$35, $z > 2.2$ quasars deg^{-2}. This magnitude limit is approaching the detection limit of SDSS photometry (Abazajian et al. 2004), meaning that photometric errors will significantly broaden the stellar locus (Figures 2.1, 2.3 and 2.4) and star-galaxy separation will be a factor. Contamination at

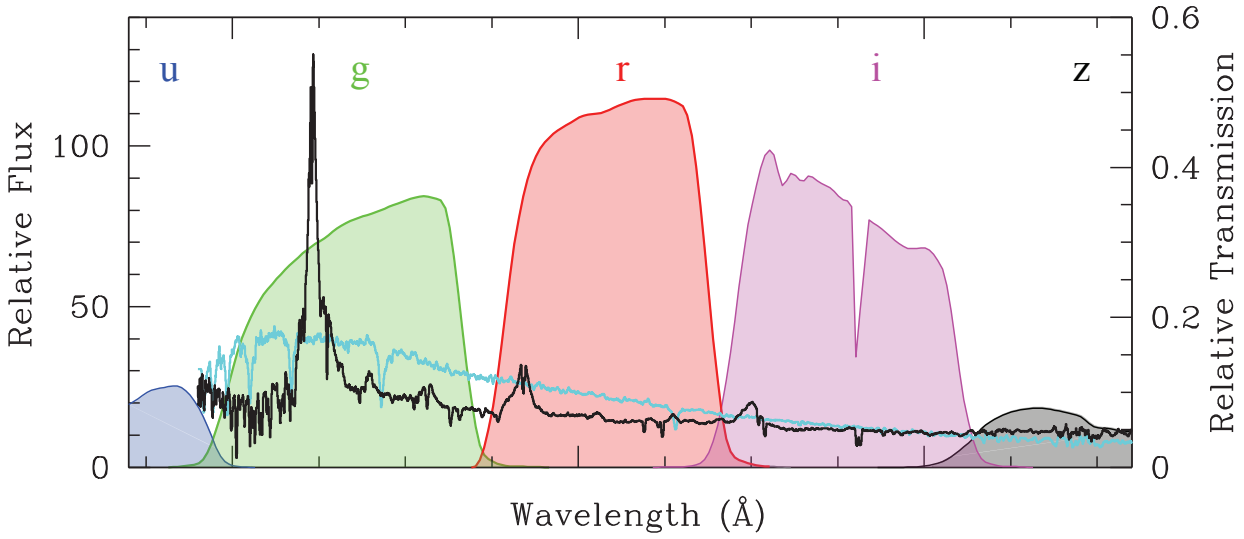


Figure 2.2: This figure shows the spectroscopy of a quasar (black) and star (blue) with similar photometry (*ugriz* histograms). This exemplifies why quasars are hard to distinguish from stars using only photometry. Both these objects have very similar photometry, but look vastly different when you compare their spectra.

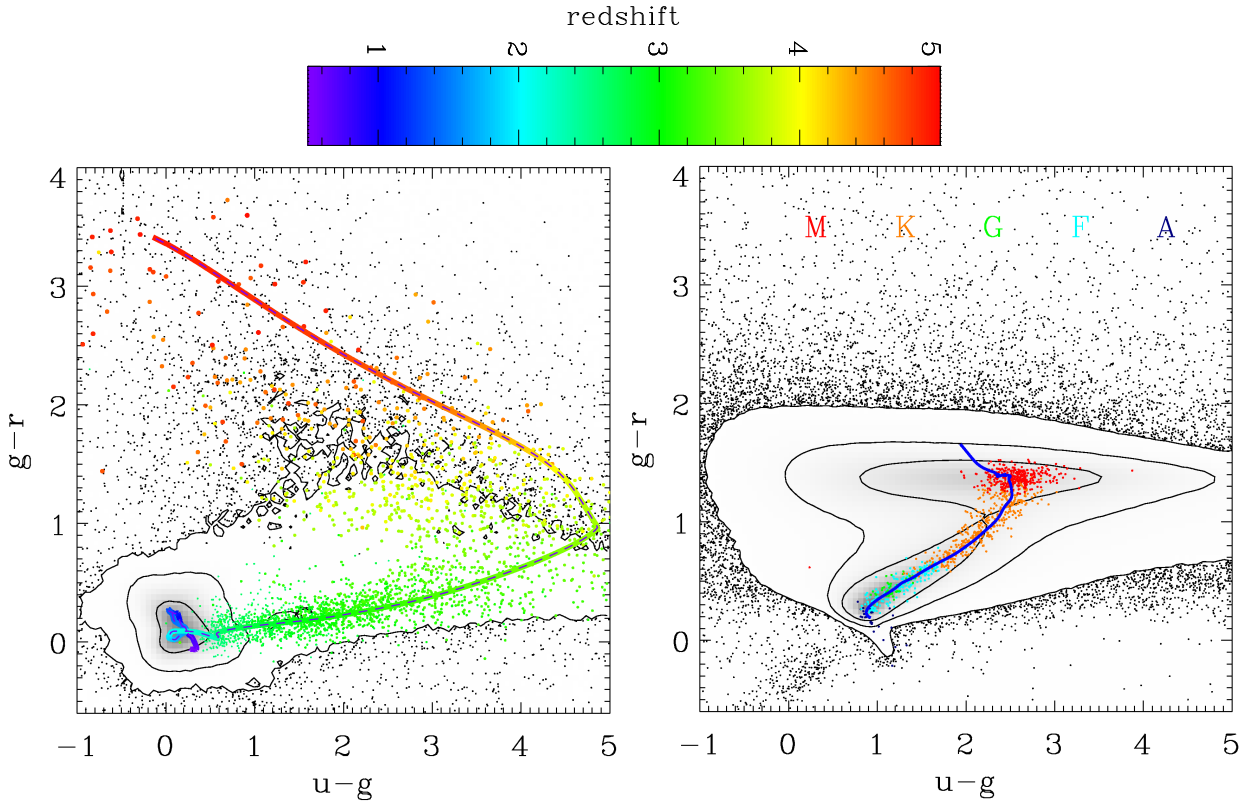


Figure 2.3: (*Left*) Color-color (*ug* vs. *gr*) distribution of SDSS quasars, de-reddened with $i < 21$ mag. The grayscale is linear in the density and the contours contain 68, 95, and 99 percent of the distribution. A sparse sampling of objects falling outside the outermost contour is shown as individual black points. A twenty-percent random sampling of objects with $z \geq 2.5$ in the *SDSS* DR7 quasar catalog (Schneider et al. 2010) is plotted as redshift color-coded points according to the color-bar above (lower redshift quasars are omitted for clarity). Higher redshift objects are plotted as larger points. A fit to the quasar locus from Hennawi et al. (2010) is shown by the dashed black line, similarly color-coded to indicate redshift. (*Right*) Color-color (*ug* vs. *gr*) distribution of SDSS stars, de-reddened with $i < 21$ mag. A fit to the stellar locus using spectroscopically confirmed stars from Hennawi et al. (2010) is shown in blue. Some representative classes of stars (M,K,C,F,A) along the stellar locus from *SDSS* are shown as colored. From Bovy et al. (2011b).

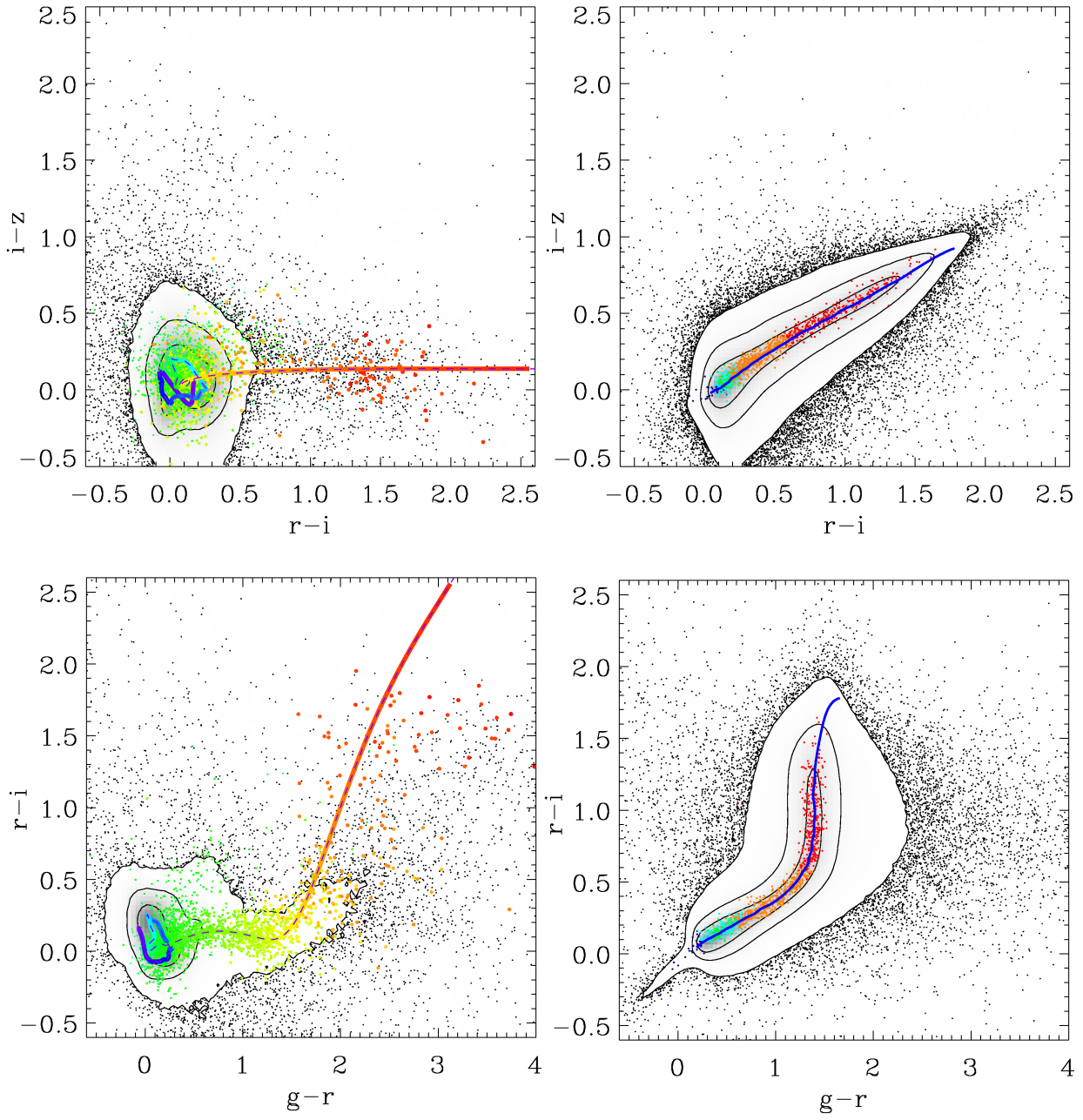


Figure 2.4: Same as Figure (2.3) but (ri vs. iz) and (gr vs. ri) color space. These figures show the width of both the quasar (*Left*) and stellar (*Right*) loci. From Bovy et al. (2011b).

both the bright and the faint end of the BOSS target range is mainly from metal-poor halo A and F stars, faint lower redshift ($z \sim 0.8$) quasars, and compact galaxies.

To put these requirements into perspective, the final quasar catalog from the original SDSS-I/II quasar survey (Schneider et al. 2010) contained 17,582, $z > 2.2$ objects over 9380 deg^2, while the 2dF-SDSS LRG And QSO (2SLAQ) survey (Croom et al. 2009), which observed to $g < 21.85$ and concentrated on UV-excess objects, contained 1,110 such quasars selected over 192 deg^2. The original 2dF QSO redshift survey (2QZ; Croom et al. (2004)) focused on the redshift range $z < 2.1$. Thus BOSS is attempting to observe almost an order of magnitude more $z > 2.2$ quasars than those known from previous surveys.

This challenge required a new approach to quasar target selection. The first year of the BOSS survey ("Year One"; 2009 September through July 2010) was devoted in part to refining the algorithms for selecting these objects. The resulting sample of quasars at $z > 2.2$ is comparable in size to the SDSS high-redshift quasar sample, and of course reaches much fainter magnitudes with much higher surface density. Thus the new sample itself represents the best test of our selection algorithms, and the target selection team modified those algorithms multiple times through the year. Year One included roughly three months of commissioning of the upgraded BOSS spectrographs and instrument control software as well as a steady ramp-up to full efficiency operations, so it includes well under 20% of the anticipated final sample for the five-year BOSS survey.

Background quasars have no causal influence on structure in the LyαF at the BAO scale.[1] Hence the sample of quasars used for LyαF cosmological studies may be quite heterogeneous, with the only consequence that the window function of the survey will depend on the distribution of the quasars for which exist spectra. Since the precision of the BAO measurement improves rapidly with the surface density of quasars (at fixed spectroscopic signal-to-noise ratio (S/N)), the BOSS team has implemented a target selection scheme in BOSS that can maximize the number of quasars found at $z > 2.2$ in any area of the sky, taking advantage of any available information (e.g., auxiliary data). In Year One, the target selection team explored a variety of methods, settling on our final target selection algorithms late in the year.

At the same time, in order to use the quasars themselves for statistical studies (such as luminosity functions or clustering analyses), BOSS must also produce a uniformly selected sample over the full footprint and based only upon SDSS imaging, which is referred to hereafter as CORE (§ 2.2.1). However, the target selection team changed the definition of the CORE sample several times over Year One, as we tested various algorithms. Therefore, our fully uniform quasar sample will not include data from this first year of the survey. However, statistical studies (luminosity functions, clustering, and so forth) can utilize all

[1]There may however be some measurement bias at the $0.1-1\%$ level for the flux power spectrum, optical depth and the flux probability distribution, due to gravitational lensing effects, (see e.g., Loverde et al. 2010).

five years of BOSS data by including moderate incompleteness corrections for Year One selection relative to the final CORE algorithm (see §5.4). The evolution of our algorithms is described in detail in Ross et al. (2012).

This chapter is organized as follows. § 2.1 describes the SDSS photometry on which the target selection algorithms are most heavily based. § 2.2 summarizes the four main methods used for selecting quasars (Richards et al. 2009a; Yèche et al. 2010; Kirkpatrick et al. 2011; Bovy et al. 2011b). These four papers suggest different, but complementary, methods, and the BOSS quasar target selection team has used a union of these techniques in different combinations through the survey.

A cosmological model throughout this chapter with $\Omega_b = 0.046$, $\Omega_m = 0.228$, $\Omega_\Lambda = 0.725$ is used (Komatsu et al. 2011). All optical magnitudes are quoted in, and based upon, the SDSS approximation to the AB zero-point system (Oke & Gunn 1983; Adelman-McCarthy et al. 2006), while all near-infrared (NIR) magnitudes are based on the Vega system. Throughout this thesis, "magnitude" refers to SDSS Point Spread Function (PSF) magnitudes (Stoughton et al. 2002).

2.1 SDSS Photometry

BOSS uses the same imaging data as that of the original SDSS-I/II survey, with an extension in the South Galactic Cap (SGC). These data were gathered using a dedicated 2.5 m wide-field telescope (Gunn et al. 2006) to collect light for a camera with 30 2k×2k CCDs (Gunn et al. 1998) over five broad bands - *ugriz* (Fukugita et al. 1996); this camera has imaged 14,555 unique deg^2 of the sky, including 7,500 deg^2 in the North Galactic Cap (NGC) and 3,100 deg^2 in the SGC (Aihara et al. 2011). The imaging data were taken on dark photometric nights of good seeing (Hogg et al. 2001), and objects were detected and their properties were measured (Lupton et al. 2001; Stoughton et al. 2002) and calibrated photometrically (Smith et al. 2002; Ivezić et al. 2004; Tucker et al. 2006; Padmanabhan et al. 2008), and astrometrically (Pier et al. 2003).

Padmanabhan et al. (2008) present an algorithm which uses overlaps between SDSS imaging scans to photometrically calibrate the SDSS imaging data. BOSS target selection uses data calibrated using this algorithm from the SDSS Data Release Eight (DR8) database (Sec. 3.3; Aihara et al. 2011). The 2.5°-wide stripe along the celestial equator in the Southern Galactic Cap, commonly referred to as "Stripe-82" was imaged multiple times, with up to 80 epochs at each point along the stripe spanning a 10-year baseline (Abazajian et al. 2009). In Chapter 3, I will discuss how the commissioning phase of BOSS used co-added catalogs in SDSS Stripe 82, generated by averaging the photometric measurements from ∼20 individual repeat scans. Also, Roughly 50% of the SDSS footprint has been imaged more

than once (Aihara et al. 2011); combining the photometric measurements in these overlap regions further reduces the flux errors.

Using the imaging data, BOSS quasar target candidates are selected for spectroscopic observation based on their PSF fluxes and colors in SDSS bands. Fluxes that are used for quasar target selection are corrected for Galactic dust extinction according to the maps of Schlegel et al. (1998). All objects classified as point-like that are brighter than $g = 22$ or $r = 21.85$ are passed to the various quasar target selection algorithms. The joint magnitude limit was imposed due to concerns of the LyαF moving into the g-band at $z \approx 2.3$ resulting in suppressed flux in g-band at redshifts greater than this. In practice, almost all the targets satisfy both these conditions. Throughout this Chapter, magnitudes use the asinh scale at low flux levels, as described by Lupton et al. (1999) although most cuts are imposed using the linear fluxes rather than these magnitudes.

2.2 Methods for BOSS Quasar Target Selection

2.2.1 Philosophy of CORE and BONUS

The BOSS team divided the quasar targets into two samples CORE and BONUS, each with approximately half the target fiber allocation.

The CORE sample is uniformly selected over the BOSS footprint and based only upon SDSS imaging. This is used for statistical studies of quasars, such as measurements of the luminosity function and the clustering of quasars. While these goals do not drive the survey's technical requirements to measure the BAO signal, BOSS will also provide an unprecedented spectroscopic dataset for studies of quasars themselves. Thus, design choices that are roughly neutral with regard to cost and impact on the cosmology goals are guided by these additional science considerations.

Since the one (imaging) dataset that is available over the entire BOSS footprint is the SDSS single-epoch photometry (including the new coverage in the SGC, Aihara et al. (2011)), we define quasar CORE targets as a sample of 20 targets deg^{-2}, which are selected *only from this single-epoch imaging data, using a uniform algorithm.* As described below, the efficiency of the CORE sample is near our goal of 50% (i.e. $\sim$10 out of 20 CORE targets deg^{-2} are $z > 2.2$ quasars). The CORE sample is designed to have a well understood, uniform, and reproducible selection function.

In contrast, the "BONUS" sample is selected using as many methods and additional data as deemed necessary to achieve our desired quasar density. The BONUS sample has a target density of 20 deg^{-2}. The number of BONUS targets added in each region of sky is adjusted

to assure that the total density of targets, CORE + BONUS, is uniform across the sky. However, the number of BONUS targets was extended up to 60 targets deg^{-2} (and then 40 targets deg^{-2}), during the BOSS Commissioning and early science phases, for a total (CORE+BONUS) of 80 (and then 60) targets deg^{-2}. The efficiency of BONUS selection is generally lower than that of CORE, despite the use of multiple algorithms and auxiliary data, simply because the relatively "easy" targets have already been picked by CORE and are therefore are not included in BONUS.

Prior to BOSS, there was no existant survey that successfully targeted $z > 2.2$ quasars to the depth and surface density and with the efficiency needed. The first year of BOSS spectroscopy was therefore largely a commissioning year for quasar target selection, during which the team gathered the quasar sample needed to test the various algorithms. In particular, it was only at the end of the year that we settled on the final CORE and BONUS algorithms. Thus, the nominal CORE-selected objects from the first year are *not* a uniformly selected sample. Sec. 5.4 describes the completeness of the final CORE sample in Year One spectroscopy.

Through this first year, the team worked on and refined a variety of algorithms for BOSS target selection, as it was not clear from the outset that any single method could meet our scientific goals. These methods include:

- The Non-parametric Bayesian Classification and Kernel Density Estimator (KDE; Richards et al. 2004, 2009a), which measures the densities of quasars and stars in color-color space from training sets. Richards et al. (2009a) showed that this was able to identify quasars at $2.2 < z < 3.5$ from SDSS photometry with an efficiency of $46.4 \pm 5.8\%$, down to a magnitude limit of $i = 21.3$, approximately $\sim$0.5 magnitudes brighter than the BOSS limit.

- A likelihood approach (Kirkpatrick et al. 2011, and Chapter 3), which determines the likelihood that each object is a quasar, given its photometry and models for the stellar and quasar loci. This method was used as CORE in Year One of data taking.

- A Neural Network (NN) approach from Yèche et al. (2010), which takes as input the SDSS photometry and errors.

- A variant of the likelihood approach, which accounts for the observational errors more properly when determining the stellar locus, called "Extreme Deconvolution" (*XDQSO* Bovy et al. 2011a). Bovy et al. (2011b) present full details on how the *XDQSO* method can be used to describe a probabilistic quasar target selection technique, called "*XDQSO*", that uses density estimation in flux space to assign quasar probabilities to all SDSS point sources. *XDQSO* was not used in Year One target selection, but it did become the CORE method in Year Two.

Below are descriptions of each of these methods, and I introduce a variant of the NN, the

"Combined Neural Network" (a.k.a. the NN-Combinator), which incorporates information from all the methods and produces the BONUS sample. There are several ancillary methods of selection, including objects associated with FIRST radio sources (§ 2.2.8) and repeat observations of previously known $z > 2.2$ quasars (§ 2.2.9) which are outlined below.

2.2.2 Kernel Density Estimation and χ^2 Cuts

Kernel Density Estimation assigns each pixel in multi-dimensional color-color space an expected density of stars and quasars, allowing a probability that any given object is a quasar to be estimated (Gray & Moore 2003; Gray & Riegel 2006; Riegel et al. 2008). Richards et al. (2004, 2009a) have applied it to the SDSS imaging data to produce photometric quasar catalogs with $\approx 10^6$ quasars. The principles of the KDE are as follows. A sample of objects of known classification (stars and quasars) serves as a training set, from which the smoothed distributions of quasar and star probability as a function of color are constructed. This allows one to compute the probability that any object of interest from the test set is a star, "KDE-star-density", or quasar, "KDE-quasar-density" (e.g. Figure 8 in Richards et al. 2009a). Based on these probabilities, the "KDE probability" is defined as:

$$\mathrm{KDE}_{Prob} = \frac{\text{KDE-quasar-density}}{\text{KDE-quasar-density} + \text{KDE-star-density}} \tag{2.1}$$

which can be used to decide whether a given object should be targeted as a quasar. As described in Section 3.5 of Richards et al. (2009a), for my purposes, the quasar density is defined as just for those objects with $2.2 < z < 3.5$; all other quasars are put into the "star" category.

Richards et al. (2004, 2009a) actually define two KDEs, split at $g = 21$, with separate color loci (different "trainings") for the bright and faint estimations. This approach crudely accounts for the very different photometric errors of the two sets, given that the KDE method, as implemented, does not take errors explicitly into account.

Roughly 45% of objects in the KDE catalog of Richards et al. (2009a) in the "mid-z" range (i.e. the redshift range of interest to BOSS) are not stars (Table 4, Richards et al. 2009a), based on an analysis of the classification efficiency using clustering (e.g., Myers et al. 2006). In the absence of significant contamination by galaxies at the faint end of the KDE catalog, the KDE algorithm is thus about 45% efficient at the Richards et al. (2009a) target density of 18.6 mid-z quasars deg^{-2}.

BOSS needs a higher efficiency, so the target selection team applied an additional cut beyond that of the Richards et al. papers to improve the efficiency of the KDE method. This cut is based on the χ^2_{star} statistic introduced by Hennawi et al. (2010), which quantifies how far a

given object is from the stellar locus:

$$\chi^2_{star} = \sum_{m=ugriz} \frac{[f^m_{\rm data} - A f^m_{\rm model}]^2}{[\sigma^m_{\rm data}]^2 + A^2[\sigma^m_{\rm model}]^2} \tag{2.2}$$

where f is the flux in each of the five SDSS bands ($m = ugriz$) for the data and for the model, $\sigma^m_{\rm data}$ is the flux error in each band, $\sigma^m_{\rm model}$ is the model uncertainty in each band, and A is a normalization. Following Hennawi et al. (2010), the stellar locus is defined by a set of ≈14,000 stars with accurate photometry from SDSS spectroscopic plates, on which all point sources were targeted above a flux limit of $i < 19.1$ regardless of color (Adelman-McCarthy et al. 2006). The minimum distance to the stellar locus, χ^2_{star}, can the be computed by minimizing the value $\chi^2(A, g-i)$, where A is the normalization constant relating the data to a model, $f^m_{\rm data} = A f^m_{\rm model}$, and $g-i$ is the color chosen as a proxy for stellar temperature. The distribution of the minimum distance to the stellar locus, i.e. range of χ^2_{star}, is shown in Figure 3 of Hennawi et al. (2010). The crucial strength that the χ^2_{star} cut adds to our KDE selection is the rejection of objects that have colors consistent with those of quasars, but have flux errors that make them consistent with the stellar locus as well.

The key parameters for the KDE method are the minimum thresholds for selection in both KDE_{prob} and $\chi^2_{\rm star}$. Early in Year One, CORE objects were selected solely by the KDE algorithm; at that time, a limit $\chi^2_{\rm star} \geq 7$ was applied. Later, when KDE was no longer the CORE algorithm, this criterion was relaxed to $\chi^2_{\rm star} \geq 3$.

2.2.3 Likelihood Method

Full details of the Likelihood method, including an in-depth analysis of its performance, are presented in Chapter 3 and Kirkpatrick et al. (2011). It is also summarized briefly here. Like KDE, the Likelihood method starts with a sample of known quasars, and a sample of "Everything Else" (EE in what follows), i.e., stars and galaxies, with *ugriz* photometry and errors. One defines likelihoods that a given object with fluxes f^m and errors σ^m ($m = ugriz$) is drawn from the quasar or EE catalog by summing a χ^2-like statistic over the full training set:

$$\mathcal{L}_{\rm quasar} = \sum_i \prod_m \sqrt{\frac{1}{2\,\pi(\sigma^m_i)^2}} \exp\left(-\frac{[f^m - {\rm quasar}^m_i]^2}{2\,(\sigma^m)^2}\right) \tag{2.3}$$

$$\mathcal{L}_{EE} = \sum_i \prod_m \sqrt{\frac{1}{2\,\pi(\sigma^m_i)^2}} \exp\left(-\frac{[f^m - EE^m_i]^2}{2\,(\sigma^m)^2}\right). \tag{2.4}$$

The sums are over all objects i in the training set. By restricting the sum to those training-set quasars in a specific redshift range, one can define an equivalent likelihood that the object in question is in this redshift range; in Year One, this was done by summing over those quasars

with $z > 2.2$. Given these likelihoods, one defines a probability that the object is a quasar to be targeted (compare with Equation 2.1):

$$\mathcal{P} = \frac{\mathcal{L}_{\rm quasar}(z > 2.2)/A_{\rm quasar}}{\mathcal{L}_{EE}/A_{EE} + \mathcal{L}_{\rm quasar}({\rm all}\ z)/A_{\rm quasar}} \tag{2.5}$$

where the As normalize for the possibly different effective solid angles of the quasar and EE training sets. In the denominator, the likelihood sum is over quasars at all redshifts, not just those at $z > 2.2$.

Like the KDE method above, this method makes use of the varying densities of objects in color space, and includes a χ^2 selection. Note that it correctly utilizes the flux errors in determining whether a given object belongs to the quasar or EE class. Potential quasar targets can be ranked by their probability $\mathcal{P}$. The Likelihood threshold is defined as ($\mathcal{P} \geq 0.234$); all objects are targeted as quasars for $\mathcal{P}$ above this value. The Likelihood method was chosen as the CORE algorithm near the end of Year One.

2.2.4 Artificial Neural Network

An Artificial Neural Network (NN) is used at two stages of the selection process. Full details of this algorithm may be found in Yèche et al. (2010). As in the previous methods, the team defines training sets of known quasars, and objects that are not quasars.

For the first stage, the NN is used with 10 inputs for each object (the SDSS g-band magnitude, the five SDSS magnitude errors and the four SDSS colors). The training set for non-quasars is a set of ∼30,000 SDSS point sources from SDSS DR7 (Abazajian et al. 2009), selected over the magnitude range $18.0 < g < 22.0$ and with Galactic latitude $b \approx 45°$ to average the effects of Galactic extinction. The training set for quasars consisted of spectroscopically confirmed quasars from the 2QZ (Croom et al. 2004), 2SLAQ (Croom et al. 2009), and the SDSS (Schneider et al. 2010) quasar catalogs.

The NN developed for target selection has four layers of "neurons" (see Figure 3 of Yèche et al. 2010). The fourth layer only has one neuron, providing a single output parameter, y_{NN}. The quantity y_{NN} quantifies the probability that an input object is a quasar, although since y_{NN} can be greater than 1, it is not a probability in the formal sense. A photometric redshift estimate, $z_{p\rm NN}$, is also generated (see Section 5 of Yèche et al. 2010), with a cut placed on this photometric redshift estimate, $z_{p\rm NN} > 2.1$.

2.2.5 Extreme Deconvolution

More details of the Extreme deconvolution (*XDQSO* Bovy et al. 2011a) method are presented in Chapter 3. *XDQSO* describes the underlying distribution function of a series of points in parameter space (e.g., quasars in color space), by modeling that distribution as a sum of Gaussians convolved with measurement errors. Bovy et al. (2011b) apply extreme deconvolution to the problem of quasar target selection, using flux data from the SDSS DR8. The so-called *XDQSO* method is conceptually similar to the Likelihood method, but explicitly models the non-uniform errors of the training set from which the quasars and stellar/EE loci are derived. Indeed, the Likelihood method effectively double-counts the errors of the training set, since the *observed* distribution of fluxes from which the Likelihood training set is built is the true *underlying* distribution convolved with the uncertainty distribution. *XDQSO* avoids this double-counting by deconvolving the underlying distribution of the training set.

XDQSO constructs a model of the distribution of the fluxes of stars and quasars in different redshift ranges based on training samples of known stars and quasars. *XDQSO* then builds a model of the relative-flux distribution as a mixture of 20 Gaussian components and fits this model to the training data, taking the heteroscedastic nature of the SDSS flux uncertainties fully into account. The *XDQSO* model for the relative-flux distribution is fit in narrow bins in i-band magnitude and combined with an apparent-magnitude dependent prior based on star counts in Stripe-82 and the Hopkins et al. (2007a) quasar luminosity function. The probability for an object to be a mid-redshift quasar ($2.2 < z < 3.5$) is given by the ratio between the number density of mid-redshift quasars and that of stars plus all quasars at the object's fluxes (in the spirit of Equation 2.5). The probability that a given object is a mid-z quasar is then:

$$P(\mathrm{QSO_{midz}}|\{f^m\}) \propto P(\{f^m/f^i\}|\mathrm{QSO_{midz}})\, P(f^i|\mathrm{QSO_{midz}})\, P(\mathrm{QSO_{midz}})\,, \qquad (2.6)$$

where m indexes the fluxes and f^i is the SDSS i-band flux. The first factor on the right is given by the *XDQSO* model for the relative-flux (i.e., color) distribution of quasars, while the second and third factors are obtained from the quasar luminosity function. The underlying relative-flux distribution is convolved with the object's flux uncertainties before evaluation. The expressions for stars and high/low redshift quasars are similar. Probabilities are normalized assuming that these classes exhaust the possibilities ($P(\mathrm{QSO_{midz}}) + P(\mathrm{QSO_{hiloz}}) + P(\mathrm{star}) = 1$). Objects are ranked on their mid-redshift quasar probability for targeting.

Since *XDQSO* target selection properly takes the flux uncertainties into account both in the training and the evaluation stage, it can be trained and evaluated on data of low signal-to-noise ratio. It can also incorporate data from surveys other than SDSS in a straightforward way, as described for near-infrared and ultraviolet surveys below. The performance of *XDQSO* using Stripe-82 data is given in Bovy et al. (2011b). The catalog of SDSS objects selected by *XDQSO* is available through the SDSS-III DR8 Science Archive Server.[2] The

[2] http://data.sdss3.org/sas/dr8/groups/boss/photoObj/xdqso/xdcore/

XDQSO method was not used during Year One, but *XDQSO* was set *a*nd fixedas the CORE method for Year Two and the remainder of the BOSS.

2.2.6 The UKIRT Infrared Deep Sky Survey

Lawrence et al. (2007) presents an overview of the United Kingdom Infrared Telescope (UKIRT) Infrared Deep Sky Survey (UKIDSS). The UKIDSS is a collection of five surveys of different coverage and depth using the Wide-Field Camera (WFCAM; Casali et al. (2007)) on UKIRT. WFCAM has an instantaneous field of view of 0.21 deg^2, and the various surveys employ up to five filters, *ZYJHK*, covering the wavelength range 0.83-2.37μm. The photometric system and calibration are described in Hewett et al. (2006) and Hodgkin et al. (2009), respectively. The pipeline processing is described in Irwin et al. (2011, *in prep.*) and the WFCAM Science Archive (WSA) by Hambly et al. (2008). The astrometry is accurate to 0.1″.

The UKIDSS Large Area Survey (ULAS) aims to map ∼4,000 deg^2 of the Northern Sky, which, when combined with the SDSS, produces an atlas covering almost an octave in wavelength. The target point-source depths of the survey are $Y = 20.3, J = 19.5, H = 18.6, K = 18.2$ (Vega); the ULAS does not image in the WFCAM Z-band. Unlike the SDSS, the ULAS multiband photometry is not taken simultaneously (e.g. Sec. 5.2 of Dye et al. 2006; Lawrence et al. 2007, Sec. 4.2), so the four bands have different coverage maps, with the H and K bands obtained together, and Y and J obtained separately. For example, the ULAS "DR8Plus"[3] coverage is 2,670 deg^2, 2,685 deg^2, 2,795 deg^2 and 2,810 deg^2, in Y, J, H and K respectively.

The UKIDSS NIR photometry was used to improve target selection in two complementary techniques. The first is to classify quasars by their "K-excess" ("KX"; e.g., Warren et al. 2000; Croom et al. 2001; Sharp et al. 2002; Chiu et al. 2007; Maddox et al. 2008; Smail et al. 2008; Wu & Jia 2010; Peth et al. 2011). The power-law quasar SED has an excess in the K-band over a blackbody stellar SED, allowing quasars to be identified (and stars rejected) that would be normally excluded from an optical color-only quasar selection algorithm - even for dust reddened quasars. Peth et al. (2011) investigated the KX method and provided an SDSS-UKIDSS matched quasar catalog. However, the very low yield (from admittedly a small target sample) caused the team to drop this method.

The second method of inclusion of NIR photometry is to improve quasar classification, and of particular importance for BOSS, photometric redshift estimation, in the *XDQSO* method. Including the NIR flux information removes many of the optically-based redshift degeneracies known for quasars (see (Bovy et al. 2011a)). Models were trained for SDSS-only fluxes and various combinations of SDSS+UKIDSS data. z ∼2.5 quasars have $(i - K)$ ∼2.1

[3] http://surveys.roe.ac.uk/wsa/dr8_las.html

(e.g., Peth et al. 2011); thus given the BOSS quasar survey magnitude limit of $i \sim 21.8$, the ULAS catalog is too shallow to guarantee 5σ detections of all sources. We therefore measure aperture magnitudes in the UKIDSS images at the positions of SDSS object counterparts; even low-significance detections can be used by *XDQSO* . Bovy et al. (2011a) gives technical details. The SDSS (optical) only model is used by *XDQSO* to generate targets for CORE, where the upper limit of the mid-z bin is $z = 3.5$. For BONUS, the SDSS+UKIDSS model is used to generate targets as an input to the NN-Combinator with an upper limit of the mid-z bin extended to $z = 4.0$. This was implemented in BONUS from the middle of Year Two onwards, with significant gains in the yield of $z > 2.2$ quasars.

2.2.7 GALEX: The Far and Near UV

The Far (1350 - 1750 Å) and Near (1750 - 2750 Å) ultraviolet (FUV and NUV respectively) photometry from the GALEX Small Explorer mission (Martin et al. 2005) also provide information that could help discriminate between hot stars and $z \sim 0.8$ quasars, both of which should have considerably more flux in the UV than a $z > 2$ quasar because of Lyα absorption along the line of sight in the latter.

The *XDQSO* technique is trained on SDSS, UKIDSS and GALEX input data. Thus the team can now perform 11-dimensional quasar target selection using the FUV/NUV,$u, g, r, i, z, Y, J, H, K$ bands. The relevant GALEX surveys are relatively shallow, e.g. $m_{\rm UV} \approx 20.5$ AB, so most potential BOSS quasar targets are not detected at high significance. Despite this, our tests (detailed in Ross et al. (2012)) confirmed that GALEX measurements—even at low significance—do help with target selection.

2.2.8 Radio Selection

As in the SDSS-I/II quasar survey, objects that are detected in the FIRST radio survey (Becker et al. 1995) are also incorporated in target selection. Radio stars are rare, thus most radio sources with faint, unresolved optical counterparts are quasars. Optical stellar objects with $g \leq 22.00$ or $r \leq 21.85$ which have FIRST counterparts within $1''$ are considered as potential quasar targets, irrespective of the radio morphology.

In the early BOSS commissioning data, the team simply selected all such radio matches. This approach targeted a substantial number of quasars with $z < 2.2$, and thus an additional color cut as placed, $(u - g) > 0.4$, to exclude UV–excess sources at lower redshift. Bluer FIRST sources are not rejected outright, but are required to pass one of the regular optical color selections to be selected. Ross et al. (2012) describes this in more detail.

2.2.9 Previously Known Objects

The density of $z > 2.2$ quasars known before BOSS started was $\sim$2 objects deg^{-2}. Given the superior throughput of the BOSS spectrographs over those of SDSS-I/II, the team decided to re-observe these objects for improved Lyα forest clustering signal. Moreover, this allows vital checks of survey quality and uniformity, and the data can be used to study the spectroscopic variability of quasars. We thus target previously known spectroscopically confirmed $z > 2.15$ quasars from the literature. Such objects are included as targets if they match a point source in the target imaging to within 1.5″, or if they match a point source in the target imaging to within 2″ and match the magnitude of that object to within 0.5.

The catalogs of previously known quasars used include the SDSS DR7 quasar catalog (Schneider et al. 2010), the 2SLAQ quasar catalog (Croom et al. 2009), the 2QZ survey (Croom et al. 2004), the AAT-UKIDSS-SDSS (AUS) survey (Croom et al., *in prep*), and the MMT-BOSS pilot survey.

To compare and check our moderate resolution spectra of generally fainter quasars to those taken by 10 m class telescopes using high-resolution spectrographs (e.g. KECK-HIRES and VLT-UVES), the team also mined the data archives (the NED[4], the Keck Observatory Archive[5] and the ESO Science Archive Facility[6]) and added those quasars with $z > 2.15$ that were not included from the above catalogs.

The full sample of known quasars contains $\sim 18,000$ $z > 2.15$ objects. Those objects in the BOSS footprint are given the highest targeting priority in tiling (Blanton et al. 2003).

The team also vetos previously known low ($z < 2.15$) redshift quasars identified from the SDSS-I/II, 2QZ, 2SLAQ and MMT surveys, never assigning them spectroscopic fibers. The team is confident that we are not inadvertently rejecting any real $z > 2.2$ quasars, since the vast majority of these objects were visually inspected and identified in the SDSS, 2QZ and MMT surveys (Schneider et al. 2010; Croom et al. 2005).

2.2.10 Combinations of Methods

Combining results from several of the methods described above in target selection requires a method to merge the (overlapping) ranked lists from these methods into a single ranked catalog. The challenge is shown in Figure (2.5), which shows the surface density of the union of those objects selected by the KDE, Likelihood, and NN methods with no further refinement, to yield an average target density of $\sim$ 60 targets deg^{-2}. The tidal stream of

[4]http://nedwww.ipac.caltech.edu/
[5]http://www2.keck.hawaii.edu/koa/public/koa.php
[6]http://archive.eso.org/

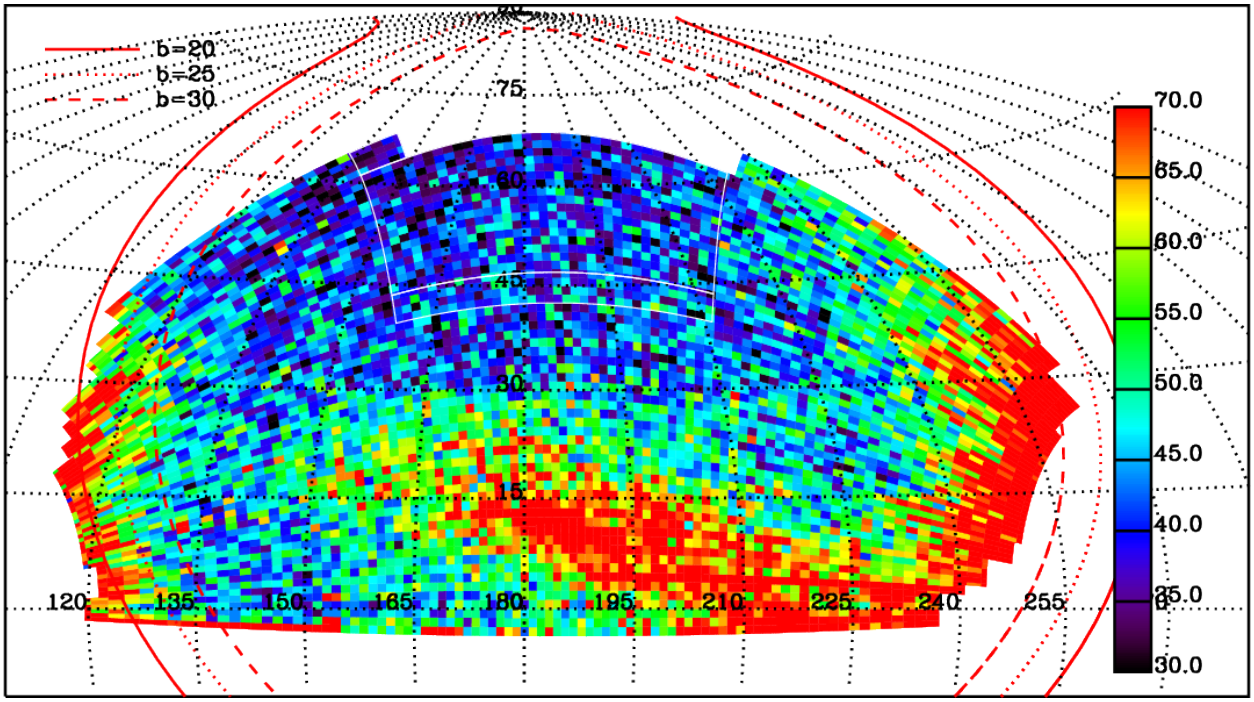


Figure 2.5: The BOSS quasar target surface density in Equatorial coordinates in the NGC, from a run of the BOSS QTS with a selection made by combining the three Year One methods, KDE, Likelihood and NN, in such a way that the average target density over the full given NGC area was ~ 60 quasar targets deg^{-2}. The color indicates the local number density of targets per square degree. The tidal stream of the Sagittarius dwarf spheroidal galaxy is prominent in the region $180° < \alpha < 240°$, and $0° < \delta < +15°$. The white lines show the "Blind Test Area", described in Ross et al. (2012).

the Sagittarius dwarf spheroidal galaxy (Ibata et al. 1995; Belokurov et al. 2006) is quite striking in this figure, spanning $180° < \alpha < 240°$ and $0° < \delta < +15°$. The target density in Figure (2.5) varies from 35 to 70 deg^{-2}. A description of the tuning and ranking of these various methods can be found in Ross et al. (2012).

NN-Combinator

The team found that the outputs of the three methods could be used as inputs into a neural net to improve the yield of $z > 2.2$ quasars. This approach is called the NN-Combinator. This approach can easily be expanded to allow input from additional selection techniques.

The NN-Combinator used the data from Stripe-82 obtained by BOSS as an input training set. The NN-Combinator was the selection method for BONUS from for most of Year One in the survey, drawing on the inputs of KDE, Likelihood, and NN.

In Year Two, with the advent of the *XDQSO* method, the results of this method were added to the NN Combinator. In particular, near the end of Year Two, a version of *XDQSO* that included data from UKIDSS (§ 2.2.6) which selected targets to $z = 4$ was used; the version of *XDQSO* used for CORE used SDSS single-epoch photometry only and did not incorporate UKIDSS data.

2.2.11 Rationale and Summary

As the above makes clear, BOSS quasar selection has been through a complex series of changes during its first two years. Here we recall the reasons for this complexity and summarize the main points of this history.

BOSS quasar target selection is complex because

- for the survey's defining science goal, measurement of BAO in the Lyα forest, the primary requirement is a high surface density of quasars in the relevant redshift range, not simplicity or homogeneity of selection,
- selection of quasars in the desired redshift range from single-epoch SDSS imaging is difficult because of proximity to the stellar locus and substantial photometric errors near the magnitude limit for BOSS selection,
- pre-BOSS quasar samples provided inadequate training sets in the desired magnitude and redshift range, so the quasars the team discovered in this first year allowed us to refine our algorithms as the year proceeded.

Roughly speaking, the effective survey volume for measurement of Lyα forest clustering shows an approximately linear improvement with the surface density of quasars, so even modest gains in efficiency have a significant science impact.

As discussed in §2.2.1, the goal of CORE selection *is* to provide a homogeneously selected sample suitable for quasar science. Ideally, the team would have frozen the CORE algorithm at the very beginning of BOSS, but the higher imperative of maximizing efficiency has led us to alter CORE as our algorithms improved. The team started by using KDE+χ^2 as the CORE algorithm but switched to Likelihood based on its greater flexibility and simplicity. Finally, we switched from Likelihood to *XDQSO* based on its better performance (at a level of $\sim$one additional high-z quasar deg^{-2}). The team intends to maintain a fixed CORE algorithm for Years 2$-$5 of the survey, and for many purposes we anticipate that completeness corrections will allow use of Year One data in statistical studies of the quasar population (see §5.4).

Beyond CORE, whatever combinations of data and methods can maximize our targeting efficiency are used, including known quasars, FIRST candidates, and the BONUS sample. Because the methods described in §§2.2.2-2.2.5 have complementary strengths, the team draws on all of them in creating the BONUS sample. We have tried different methods of forming a combined BONUS list during the first year, and the teamf settled on the NN-combinator (§2.2.10) as our primary tool for doing so. The individual methods feeding into the NN-combinator use co-added SDSS photometry where it is available in overlap regions, in contrast to CORE, which relies on single-epoch photometry to ensure uniformity. Auxiliary data such as UKIDSS and GALEX photometry are fed into the *XDQSO* selection, which in turn is fed into the NN-combinator.

Chapter 3

Likelihood and *XDQSO* Target Selection

The BOSS LyαF/Quasar Survey will target objects thought to be $z > 2.2$ quasars to perform a LyαF BAO measurement. Since the foreground LyαF is independent of the intrinsic properties of the background quasar, there is freedom to use multiple selection methods without biasing the BAO results. The methods used for BOSS targeting include the "Kernel Density Estimator" (KDE; Richards et al. 2004), an "Extreme-Deconvolution" method (XDQSO; Bovy et al. 2011b), and a Neural Network method (NN; Yèche et al. 2010). The BOSS QSO target selection used for the first year of observations (Ross et al. 2012) combines all these different methods, including the Likelihood method described in this chapter, with different photometric catalogs such as SDSS (York et al. 2000), UKIDSS (Lawrence et al. 2007), GALEX (Martin et al. 2005) and quasars found using their flux time-variability information (Palanque-Delabrouille et al. 2011).

In this chapter, I describe a new method for quasar target selection. My method models data in 5-filter flux space, then calculates likelihood estimates that a given object is a $z > 2.2$ quasar. Because a given survey has a finite number of spectroscopic fibers (observing time allocation) to dedicate towards quasar targeting, this method attempts to prioritize selection by calculating a probability that a potential target is a quasar based on these likelihood calculations. Targets are ranked by likelihood probability. This method differs from KDE in that it incorporates the photometric errors for each object into the likelihood calculations; also KDE only imposes a single magnitude prior and color-distribution, whereas I model the QSO density as a function of magnitude and evolution of color distribution.

The layout of this chapter is as follows. § 3.1 describes the method used to calculate the likelihoods, and training catalogs that are generated and used for Likelihood target selection. In § 3.2 I give an overview of the BOSS Data and the performance of the Likelihood method

using this data. In § 3.3 I discuss testing and optimization of the method, as well as future work and possible improvements. I use the terms "quasar" and "QSO" interchangeably to refer to quasi-stellar, type-I broad line objects. All Right Ascension (RA) and Declinations (Dec) discussed are J2000.

My Likelihood method was used for the uniformly selected sample (which I refer to hereafter as "CORE") in the first year of the BOSS QSO target selection (Ross et al. 2012) I released my calculated likelihood probabilities as a data product in the SDSS Data Release 9.

3.1 Likelihood Method and Catalog Generation

3.1.1 Likelihood Method

Recent work has approached target selection within a Bayesian statistical framework over more traditional color-box approaches (Richards et al. 2004; Bovy et al. 2011b; Mortlock et al. 2012). Spectroscopic target selection can be viewed as a classification problem. Given a set of photometric target objects (O) with attributes ($\mathbf{a}$) and a discrete set of astronomical object classes, one would like to assign a target to a particular class. For my purposes I am simply interested in the question: "Is the object a quasar?" Thus I have two classes: quasar (QSO) and non-quasar (i.e., all other observable objects: stars + galaxies + anything else) hereby referred to as Everything Else (EE).

The probability that an object O is a quasar (in class QSO) given a vector of object attributes $\mathbf{a}$, is provided by Bayes' theorem (Sivia & Skilling 2006):

$$\mathcal{P}(O \in QSO \mid \mathbf{a}) = \frac{\mathcal{P}(\mathbf{a} \mid O \in QSO)\, \mathcal{P}(O \in QSO)}{\mathcal{P}(\mathbf{a})} \tag{3.1}$$

where $\mathcal{P}(\mathbf{a} \mid O \in QSO)$ is the conditional probability that given attributes $\mathbf{a}$, object O is a quasar; $\mathcal{P}(O \in QSO)$ is the prior probability that O is a quasar (prior in the sense that it does not take into account any information about the object attributes); $\mathcal{P}(\mathbf{a})$ is the marginal probability of an object with attributes $\mathbf{a}$ occurring at all, and acts as a normalizing constant. In my case:

$$\mathcal{P}(\mathbf{a}) = \mathcal{P}(\mathbf{a} \mid O \in QSO)\, \mathcal{P}(O \in QSO) + \mathcal{P}(\mathbf{a} \mid O \in EE)\, \mathcal{P}(O \in EE) \tag{3.2}$$

because $QSO \cup EE$ contain all possible classifications (or outcomes) for object O.

I used the term "likelihood" to denote the conditional probabilities $\mathcal{P}(\mathbf{a} \mid O \in QSO)$ and $\mathcal{P}(\mathbf{a} \mid O \in EE)$ in Equations (3.1) and (3.2). In the case where the attributes of a target object are measured with a significant amount of uncertainty, one can imagine $\mathbf{a}$ is a noisy

measurement of an underlying true attribute vector $\mathbf{a}'$. I then calculated the likelihood by marginalizing over all possible of values of $\mathbf{a}'$: [1]

$$\begin{aligned}\mathcal{P}(\,\mathbf{a} \mid O \in QSO) &= \int \mathcal{P}(\,\mathbf{a}, \mathbf{a}' \mid O \in QSO)\; d\mathbf{a}' \\ &= \int \mathcal{P}(\,\mathbf{a} \mid \mathbf{a}', O \in QSO)\; \mathcal{P}(\,\mathbf{a}' \mid O \in QSO)\; d\mathbf{a}' \\ &= \int \mathcal{P}(\,\mathbf{a} \mid \mathbf{a}')\; \mathcal{P}(\,\mathbf{a}' \mid O \in QSO)\; d\mathbf{a}' \;.\end{aligned} \tag{3.3}$$

For my purposes, $\mathcal{P}(\,\mathbf{a}' \mid O \in QSO)$ is just the empirical distribution observed in a discrete set of high signal-to-noise objects which are already classified to be either quasars or non-quasars. The attributes are the photometric fluxes (f) in the five SDSS color filters ($\mathbf{f} = \{u, g, r, i, z\}$) and are independent of each other. Because the empirical distribution places a δ-function at each training example in Equation (3.3), the integral becomes a sum over all objects (O') with attributes $\mathbf{a}'$ in the training sets:

$$\int \mathcal{P}(\,\mathbf{a} \mid \mathbf{a}')\; \mathcal{P}(\,\mathbf{a}' \mid O \in QSO)\; d\mathbf{a}' = \sum_{O'} \mathcal{P}(\,\mathbf{a} \mid \mathbf{a}') \;. \tag{3.4}$$

Like other recent publications (Bovy et al. 2011b; Mortlock et al. 2012), I use a Gaussian distribution, $\mathcal{P}(\,\mathbf{a} \mid \mathbf{a}')$, for the uncertainties of the attributes. Thus a Gaussian distribution is used for the errors (σ_f) in the object fluxes, and fluxes f and $f^{O'}$ for one of the target object attributes ($\mathbf{a}$) and training object attributes ($\mathbf{a}'$) respectively. The likelihood ($\mathcal{L}$) for a single flux f then becomes:

$$\mathcal{L} = \mathcal{P}(\,f \mid O \in QSO \text{ or } EE) \simeq \sum_{O'} \sqrt{\frac{1}{2\pi\sigma_f^2}} \exp\left[-\frac{[f - f^{O'}]^2}{2\sigma_f^2}\right] . \tag{3.5}$$

When I consider all five SDSS fluxes, there is a multiplicative sum over these attributes and Equation (3.5) becomes:

$$\mathcal{L} = \mathcal{P}(\{u, g, r, i, z\} \mid O \in QSO \text{ or } EE) \simeq \sum_{O'} \prod_{f=u,g,r,i,z} \sqrt{\frac{1}{2\pi\sigma_f^2}} \exp\left[-\frac{[f - f^{O'}]^2}{2\sigma_f^2}\right] \tag{3.6}$$

Note that the above equations become equalities when the training catalogs completely represent the object flux-space. For my target object fluxes (f), I used SDSS photometric PSF fluxes from the SkyServer (http://www.sdss3.org/dr8/) under the standard SDSS data releases.

[1] In the derivation of Equation (3.3), I assume the noisy observation $\mathbf{a}$ is independent of the classification of O given $\mathbf{a}'$, therefore: $\mathcal{P}(\,\mathbf{a} \mid \mathbf{a}', O \in QSO) = \mathcal{P}(\,\mathbf{a} \mid \mathbf{a}')$.

All fluxes and their errors are corrected for Galactic extinction in the SDSS filters using the prescription in Schlegel et al. (1998). Because the sum is done in flux space rather than color space, object errors are independent. Also, my method preserves the luminosity function information, whereas the absolute flux information is lost when using colors. My training catalogs use stacked (co-added) fluxes (see Sections 3.1.3 and 3.1.4), whereas the targets are single epoch fluxes (see Section 3.2.2). Therefore in the above equations, the errors in the catalog fluxes ($f^{O'}$) are ignored because the signal-to-noise ratio of the catalog fluxes are much greater than the signal-to-noise ratio of my potential target fluxes (f).

The QSO likelihood can separated into redshift bins (Δz) so that I can tune Equation (3.1) to a desired target redshift range. This simply requires having redshifts for the objects (O') in the QSO training catalog and subdividing this data into redshift bins. I used a width $\Delta z = 0.1$ (e.g. $0.5 \rightarrow 0.6$, $0.6 \rightarrow 0.7$ ... $4.9 \rightarrow 5.0$). This results in the following final equations for the QSO and EE likelihoods:

$$\mathcal{L}_{QSO}(\Delta z) = \mathcal{P}(\,\mathbf{f} \mid O \in QSO(\Delta z)) \simeq \sum_{O' \in QSO(\Delta z)} \prod_{f=\mathbf{f}} \sqrt{\frac{1}{2\pi\sigma_f^2}} \exp\left[-\frac{(f - f^{O'})^2}{2\sigma_f^2}\right] \tag{3.7}$$

$$\mathcal{L}_{EE} = \mathcal{P}(\,\mathbf{f} \mid O \in EE) \simeq \sum_{O' \in EE} \prod_{f=\mathbf{f}} \sqrt{\frac{1}{2\pi\sigma_f^2}} \exp\left[-\frac{(f - f^{O'})^2}{2\sigma_f^2}\right]. \tag{3.8}$$

The Gaussian normalizations add a multiplicative constant to each likelihood ($\mathcal{L}$), which is the same for both Equation (3.7) and Equation (3.8) for a given target and cancel when calculating the probabilities in Equation (3.1).

The prior probabilities in Equation (3.1) and Equation (3.2) are the relative surface densities of quasars and everything else on the sky, and thus normalize Equation (3.7) and Equation (3.8). I do this by defining the prior probabilities to be the inverse of the effective sky area (A) of the QSO and EE catalogs:

$$\mathcal{P}(O \in QSO) = \frac{1}{A_{QSO}}\,, \qquad \mathcal{P}(O \in EE) = \frac{1}{A_{EE}}\,. \tag{3.9}$$

By inserting Equation (3.8), Equation (3.7), and Equation (3.9) into Equation (3.1) I can get probability that a single potential target object (O) is a quasar (QSO) in a target redshift range ($\Delta z_{\rm target}$):

$$\mathcal{P}(O \in QSO(\Delta z_{\rm target}) \mid \mathbf{f}\,) \simeq \frac{\displaystyle\sum_{\Delta z_{\rm target}} \left(\frac{\mathcal{L}_{QSO}(\Delta z)}{A_{QSO}}\right)}{\displaystyle\frac{\mathcal{L}_{EE}}{A_{EE}} + \sum_{\Delta z_{\rm all}} \left(\frac{\mathcal{L}_{QSO}(\Delta z)}{A_{QSO}}\right)}\,. \tag{3.10}$$

In the numerator, $\mathcal{L}_{QSO}(\Delta z)$ is summed over the desired quasar target redshift range ($\Delta z_{\rm target}$), whereas the denominator contains all objects in both catalogs summed over the entire redshift range ($\Delta z_{\rm all}$). This probability is exact in the limit of perfect training catalogs (infinite objects and zero errors). A probability is calculated for every potential target using the full *QSO* and *EE* catalogs.

3.1.2 Imaging Data

BOSS uses the same imaging data as that of the original SDSS-I/II survey (York et al. 2000), with an extension in the South Galactic Cap (SGC). These data were gathered using a dedicated 2.5 m wide-field telescope (Gunn et al. 2006) to collect light for a camera with 30 2k×2k CCDs (Gunn et al. 1998) over five broad bands - *ugriz* (Fukugita et al. 1996); this camera has imaged 14,555 deg^2 of the sky, including 7,500 deg^2 in the North Galactic Cap (NGC) and 3,100 deg^2 in the SGC (Aihara et al. 2011). The imaging data were taken on dark photometric nights of good seeing (Hogg et al. 2001), and objects were detected and their properties were measured (Lupton et al. 2001; Stoughton et al. 2002) and calibrated photometrically (Smith et al. 2002; Ivezić et al. 2004; Tucker et al. 2006; Padmanabhan et al. 2008), and astrometrically (Pier et al. 2003).

Padmanabhan et al. (2008) presents an algorithm which uses overlaps between SDSS imaging scans to photometrically calibrate the SDSS imaging data. BOSS target selection uses these "ubercalibrated" data from the SDSS Data Release Eight (DR8) database (Aihara et al. 2011). The 2.5° stripe along the celestial equator in the Southern Galactic Cap, commonly referred to as "Stripe-82" was imaged multiple times, for up to 80 epochs spanning a 10-year baseline (Abazajian et al. 2009). A coaddition of these data (Adelman-McCarthy et al. 2008) goes roughly two magnitudes fainter than the single-scan images which make up the bulk of the SDSS imaging data.

3.1.3 QSO Catalog

Because there are relatively few previously observed quasars in the desired BOSS redshift range ($z > 2.2$) with sufficiently small flux errors to precisely describe the quasar color locus, for my purposes the QSO Catalog is generated by a Monte Carlo technique (Hennawi et al. 2010) to provide a less biased and more complete sample than is available from the SDSS quasar catalog. The Monte Carlo simulation uses a model of the quasar luminosity function based on the studies by Jiang et al. (2006) to compute the density of quasars as a function of redshift and $i-$magnitude. The Jiang et al. (2006) luminosity function is used because it extends fainter than the luminosity function of Richards et al. (2006) and thus better matches the high redshift quasars in the BOSS redshift regime. SDSS Data Release

5 spectroscopically confirmed quasars (DR5QSO; Schneider et al. 2007) are the photometric inputs to the Monte Carlo. The simulation generates 9.94 million unique (i-magnitude, redshift) pairs down to $i = 22.5$ (0.5−mag fainter than the BOSS magnitude limit) with a distribution given by the luminosity function. Each simulated quasar (O') is then matched to the SDSS quasar (DR5QSO) with the nearest redshift to O'. The SDSS photometry of the DR5QSO quasar is rescaled such that its $i-$magnitude matches that of the simulated quasar (O'). I assume that quasar colors are not a function of magnitude in the redshift range of interest, and thus can be extrapolated in this manner to deeper fluxes. Thus this technique preserves the relative fluxes while providing a more complete coverage of the flux space than only using known SDSS quasars. Objects with redshifts in the range desired for BOSS targeting ($z > 2.2$) are included in the numerator of Equation (3.10). The location in ugr color-color space of the $z > 2.2$ objects in the QSO Catalog is shown by the blue contours in Figure (3.1).

3.1.4 Everything Else Catalog

The Everything Else (EE) Catalog is generated using stacked SDSS "Stripe-82" imaging, allowing the construction of a large point source catalog with variability information and smaller errors than possible using single-epoch SDSS imaging. Stripe-82 is the 2.5° wide region on the celestial equator between RA= −45° and RA= 45° where SDSS repeatedly scanned. Non-photometric data were ignored, and the photometric images were processed with a version of the SDSS photometric reduction pipeline similar to that in data release eight (Aihara et al. 2011). The photometric depth is $r \sim 22.5$ magnitude (5σ) for point sources, with high completeness and accurate star-galaxy separation to $r \sim 22$ magnitude. These data were combined at the catalog level to produce co-added PSF photometry. Typically 20 observations were included for each object, resulting in a co-added catalog with typical errors of 6.1%, 2.4%, 3.0%, 7.1% and 27% at 22nd magnitude in the u, g, r, i, z filters, respectively.

The EE Catalog is further trimmed to a clean sample of non-variable point sources for inclusion in the likelihood calculations of Equation (3.10). The 23.9% of objects that are blended with neighboring objects are rejected, thus reducing the effective footprint of this catalog from 225 deg^2 to 171.2 deg^2. Objects with high variability are explicitly excluded from the catalog under the presumption that these are dominated by quasars (Schmidt et al. 2010), and I explicitly add quasars into the numerator and denominator of Equation (3.10) such that the computed probability remains in the range [0,1]. These variable objects are identified as those with a reduced χ^2 of the fit to a constant r-band flux exceeding 1.4. This reduces the catalog to an effective area of 150 deg^2. The result is a catalog with 1,042,262 photometric fluxes that represent all non-quasar types of objects. I determined the contamination of $z > 2.2$ quasars in this set is less than 0.5% by comparing this catalog with those for which I have spectra. In Figure (3.1) the red contours show the urg color-color space of the objects in the EE catalog.

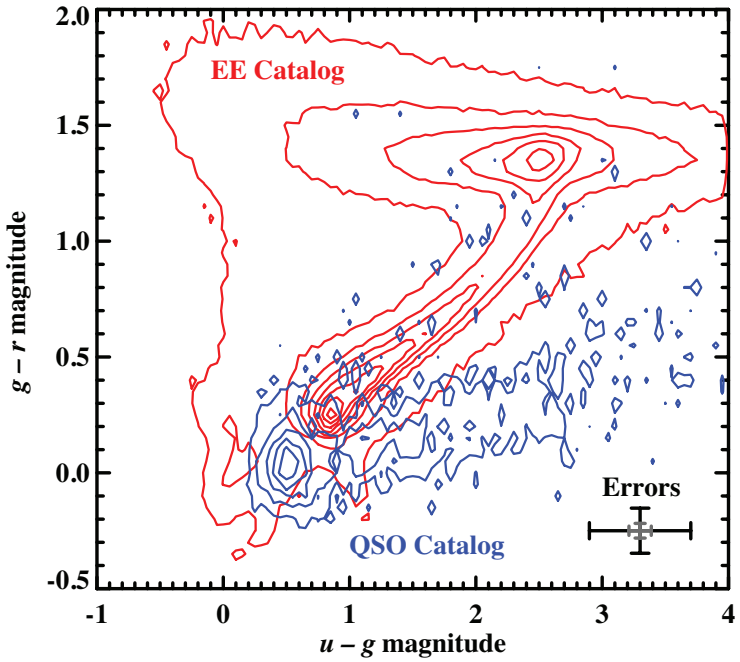


Figure 3.1: Contour plot of the $u-g$ and $g-r$ colors of the Everything Else (*red*) and QSO (*blue*) Catalogs. The region of overlap, where target selection becomes difficult, is at $u-g \approx 1$ and $g-r \approx 0$. The error bars are the SDSS single-epoch $g-r$ and $u-g$ magnitude errors at g=22 (*black*) and g=20 (*grey*).

3.2 BOSS Data & Likelihood Performance

3.2.1 BOSS Stripe-82 Data

In September of 2009, BOSS started taking spectroscopic data. During the first year of data taking, several target selection methods were employed. In addition to likelihood method, three other selection techniques were deployed: the KDE method developed to classify quasars by separating them from stars in color space (Richards et al. 2004), an "extreme-deconvolution" method (XD Bovy et al. (2011b), Section 3.4) and a new approach based on artificial neural networks (Yèche et al. 2010). Previously spectroscopically confirmed quasars, as well as objects with high variability (Palanque-Delabrouille et al. 2011) over consecutive Stripe-82 runs were also targeted during this time.

Stripe-82 target selection used co-added catalog data from SDSS as the potential target fluxes. Because the co-added photometry has a higher signal-to-noise ratio than any single-epoch data run and the target fiber density in this region was higher than the rest of the survey, BOSS QSO completeness is highest in this region. Once observed, all of the quasar targets were automatically classified and then visually examined.

Based on the objects selected in Stripe-82, I found that the performances of the four methods were not identical as a function of the magnitude and redshift of the objects (Ross et al. 2012). This behavior is likely due to the different strategies adopted in the training of the methods.

3.2.2 Likelihood Performance

Although I targeted a number of tiles for spectroscopy during the first year of data taking, observational success was varied. Due to a combination of poor observing conditions and equipment glitches, spectroscopic completeness (the fraction of total spectroscopic observations in a tiling region which yielded a high confidence spectroscopic identification upon visual inspection) was a strong function of the region in which a target was tiled. In this thesis, I only test my method using observations in Stripe-82 regions with a spectroscopic completeness of $> 90\%$. In Figure (3.2) I show the tiles used for testing.

To test the performance of my likelihood method, I calculated probabilities using Equation (3.10) on single-epoch data in regions of Stripe-82 with high spectroscopic completeness and compared that target list with the BOSS "truth table" (which includes targets from all targeting methods, quasars targeted using variability, and all previously known quasars). This is a fair test because targeting in this region was conducted using co-added photometry and thus I am not testing the likelihood method on a region that was targeted with the same

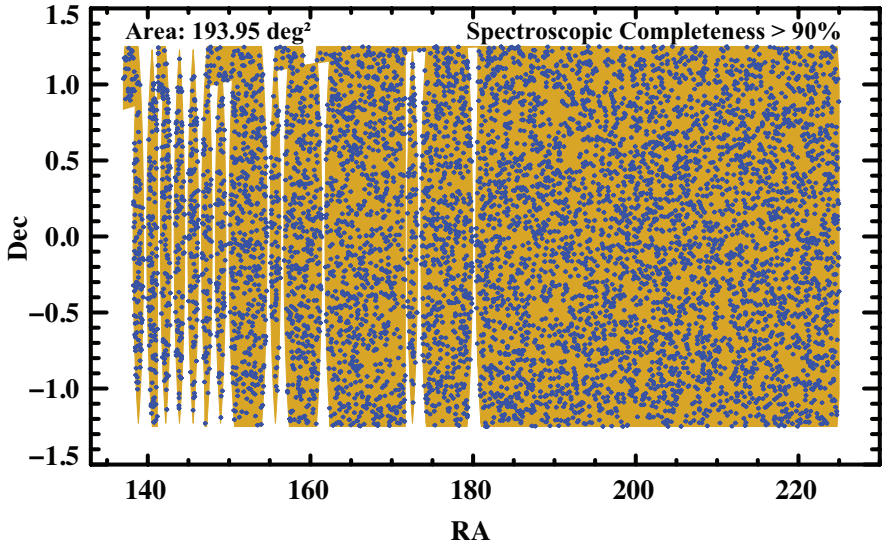


Figure 3.2: Right Ascension (RA) versus Declination (Dec) of BOSS QSO data used for the likelihood method testing and luminosity function testing. Testing was done in the Stripe-82 calibration band with regions of high ($> 90\%$) spectroscopic completeness. The blue points are spectroscopically confirmed quasars and the yellow regions are the sky tiles that were observed. Note that the vertical and horizontal scales are not the same.

photometry.

Likelihood probabilities were calculated for 592,847 objects; of those, the top 7,757 likelihoods were selected ($\mathcal{P} > 0.245$) for a target density of 40 objects per deg^2. Figure (3.3) shows the distribution of the probabilities for the recovered[2] and false-negative (missed) QSOs as well as for the false-positive stars (wrongly) targeted by the method. I found an efficiency ($\mathcal{E}$ = Recovered QSOs / Total Targets) of 40% and completeness ($\mathcal{C}$ = likelihood recovered QSOs / total BOSS recovered QSOs) of 65%. Figure (3.5) shows ugr color-color plots of BOSS quasars recovered (magenta) and missed (cyan) by the likelihood method as well as false positive contamination stars (red) that were targeted by the method.

There is of course the inevitable trade-off between $\mathcal{E}$ and $\mathcal{C}$. The more fibers given to quasar targets, the more QSOs are found (greater $\mathcal{C}$), but the accuracy of targeting a quasar decreases (lower $\mathcal{E}$). This is shown in Table (3.1) where the rate of new targeted QSOs is shown to steadily decrease as a function of targets deg^{-2}.

By comparing the target objects to the catalog contours in Figure (3.5), it is clear that the likelihood method fails mostly in the region of overlap between the two catalogs. Figure (3.4) shows the redshift distributions of the targeted and missed quasars and the limitation of SDSS DR5 catalog at $z > 2$. Table (3.1) shows the detailed testing results.

3.3 Testing and Improvements

3.3.1 Likelihood versus Color-Box

In order to see how my likelihood method performed against the traditional "color-box" selection, I compared the number of $z > 2.2$ quasars the likelihood method was able to recover versus a simple color-box selection, using the BOSS data on Stripe-82 (the same data set as in Section 3.2.2). Please note that my color-box, described below, is a relatively simple selection in only $(u-g)$ vs. $(g-r)$ magnitudes color-space, designed to adequately sample the location where z 2.7 QSOs reside. This color-box is not the same as the "inclusion region" from Richards et al. (2002)(Section 3.5.2, Fig. 7) but the $(g-r) < 0.43(u-g)$ cut is inspired by their high-redshift color-selection.

The color cuts I used for my tests are:

$$\begin{gathered}[((g-r) < -0.13 \cdot (u-g) - A) \text{ and } ((u-g) > 0.3) \text{ and } (r < 22.0)] \\ \text{or} \\ [((g-r) < 0.43 \cdot (u-g) + B) \text{ and } (0.3 < (u-g) < 2.0) \text{ and } (r < 22.0)]\end{gathered} \tag{3.11}$$

[2] I define recovered/missed QSOs to be quasars in the desired BOSS redshift range ($z > 2.2$).

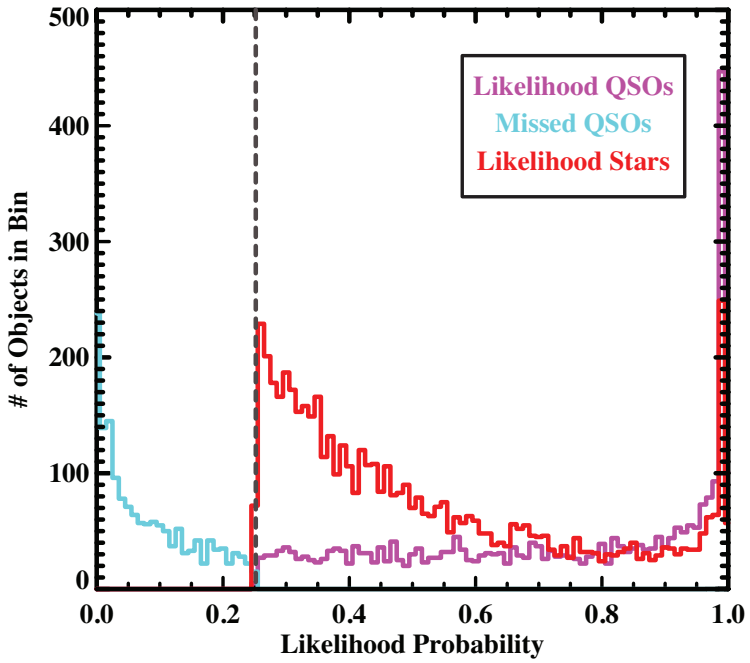


Figure 3.3: The probability ($\mathcal{P}$) distributions of the likelihood method recovered QSOs (magenta, 4617 total), false-negative QSOs that were missed by the likelihood method (cyan, 1566 total), and false-positive stars that were incorrectly targeted by likelihood method (red, 5743 total). The vertical gray dashed line shows the likelihood $\mathcal{P}$ threshold used for targeting ($\mathcal{P} > 0.245$). The spike around $\mathcal{P} = 0$ in the cyan curve are quasars that fall in the midst of the stellar locus and therefore are found by the method to have a very low probability of being QSOs. Most of these quasars are targeted because they are previously spectroscopically confirmed QSOs or by their flux variability. The likelihood distribution of the probabilities for the untargeted stars (true-negative) are not included in the plot, but constitute an additional 742,662 objects.

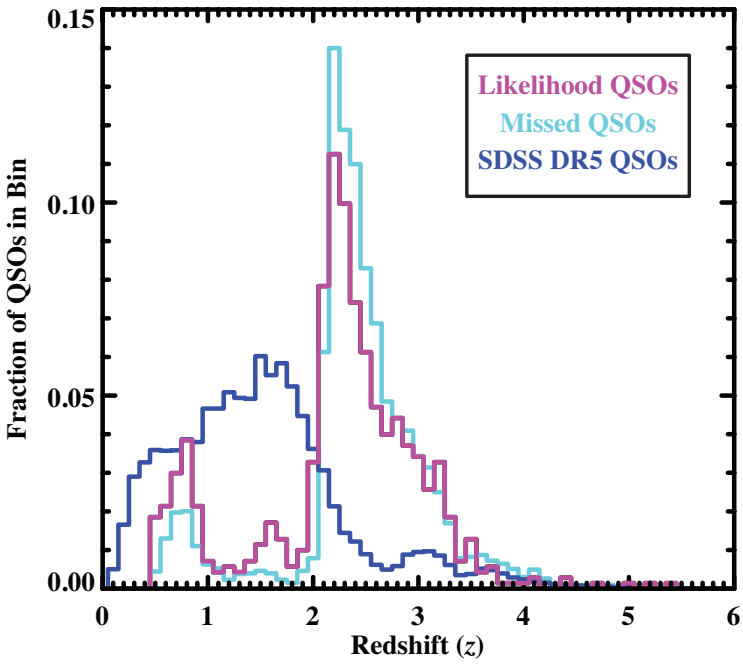


Figure 3.4: The redshift distributions of the likelihood method recovered QSOs (magenta) and false-negative QSOs that were not targeted (missed) by the likelihood method (cyan), compared with SDSS DR5 QSOs (blue).

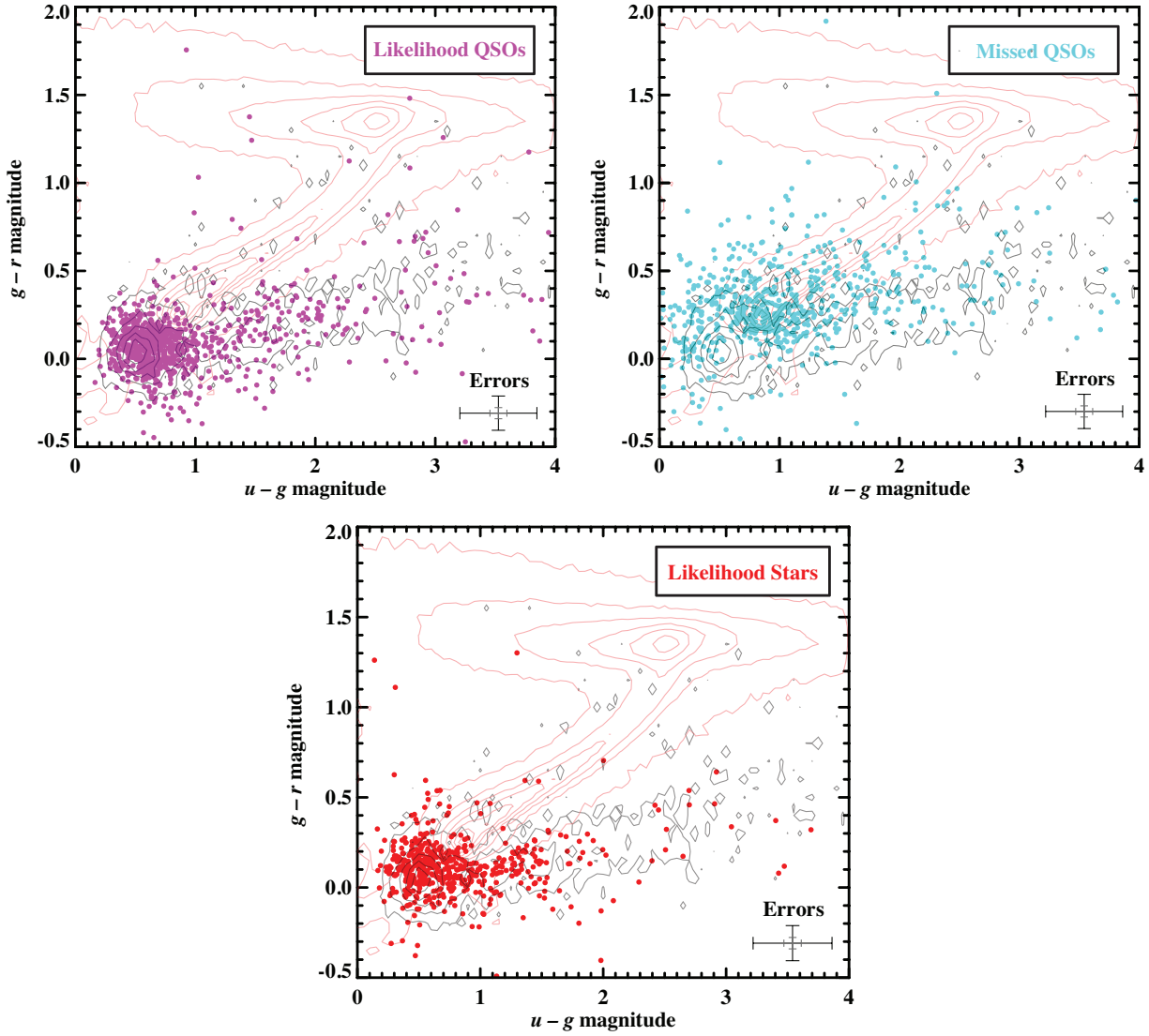


Figure 3.5: Color-color diagrams of BOSS QSOs recovered by the likelihood method (*magenta*), false-negative QSOs that were not targeted (missed) by likelihood method (*cyan*), and false-positive stars that were wrongly targeted by likelihood method (*red*). These plots show recovered/missed ($z > 2.2$) QSOs. It is clear when comparing these plots with Figure (3.1) that the problematic region for likelihood targeting is where the two catalogs overlap near $u - g = 1$, $g - r = 0.25$. For context the *QSO* Catalog and *EE* Catalog contours plot from Figure (3.1) are included in the above plots. The error bars are the SDSS single-epoch $g-r$ and $u-g$ magnitude errors at $g = 22$ (*black*) and $g = 20$ (*grey*). The targeting decisions were computed in flux space rather than the color space shown in the figures.

Table 3.1. Likelihood Stripe-82 Results, Jiang et al. (2006) Luminosity Function

Targets per deg^2	Likelihood $\mathcal{P}$ Threshold	Total Targets	QSOs Recovered	QSOs Missed	$\mathcal{C}$ (%)	$\mathcal{E}$ (%)
5	0.974	969	669	3811	15	69
10	0.833	1938	1276	3227	28	66
20	0.535	3878	2166	2401	47	56
40	0.245	7757	3087	1657	65	40
60	0.136	11636	3595	1331	73	31
80	0.088	15515	3965	1108	78	26
100	0.063	19394	4219	980	81	22
140	0.037	27152	4618	806	85	17

Note. — The $\mathcal{E}$ and $\mathcal{C}$ as a function of dedicated target fibers (targets deg^{-2}). These values are for $z > 2.2$ recovered/missed QSOs. There is a trade-off: the more fibers given to targets, the more QSOs are found (greater $\mathcal{C}$), but the accuracy of finding a quasar decreases (lower $\mathcal{E}$). The values for threshold, $\mathcal{C}$ and $\mathcal{E}$ will of course depend on Galactic latitude (Ross et al. 2012). BOSS year-one data targeted using the likelihood method at 20 targets deg^{-2} for the CORE sample.

where offsets A and B in Equation (3.11) are defined as:

$$A = (0.01 \cdot k) - 0.32, \qquad B = -(0.01 \cdot k) - 0.28 \tag{3.12}$$

and k is an integer varying from 0 to 29, $k = [0, 1, 2, \cdots, 27, 28, 29]$.

In Figure (3.6) I show the above color cuts applied to the Stripe-82 potential quasar targets. The red points are targets that are not selected by the color-box. The points that are other colors at the bottom center of the figure are targets that were selected by the cuts in Equation (3.11) with the different offsets from Equation (3.12). The color-box selection cuts with the most targets ($k = 0$) are closest to the stellar locus (shown as grey contours). The cuts then get more restrictive going down towards the x-axis, where less and less targets are selected.

Figure (3.7) show the results of $z > 2.2$ quasars recovered by the above color cuts compared to those recovered by the likelihood method. This is consistent with the results in Table (8) of Ross et al. (2012) which shows 6.45 mid-z quasars are recovered from 20 targets deg^{-2} using their (slightly different) color-box selection. Likelihood out performs the color-box selection method by recovering over twice as many BOSS quasars at 20 targets deg^{-2}.

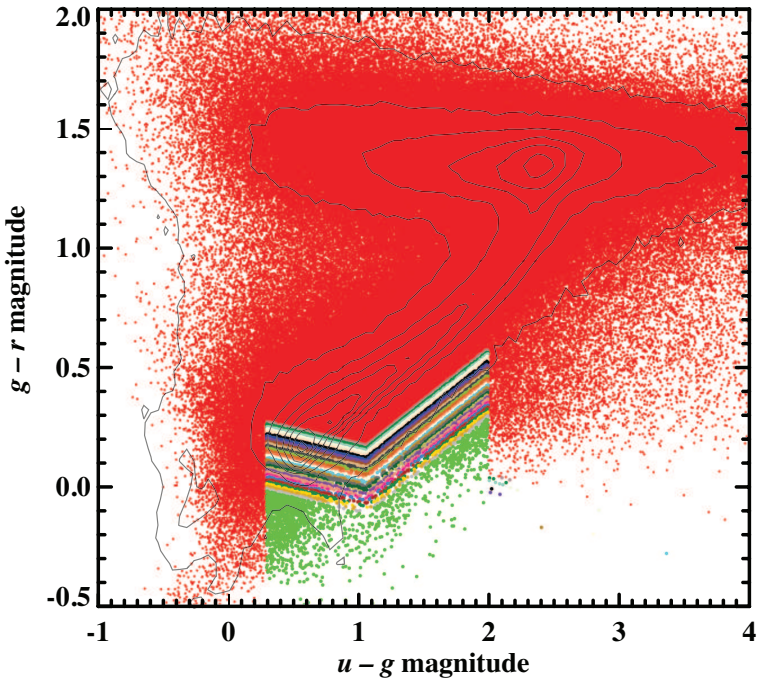


Figure 3.6: A color-color diagram showing the cuts applied for color-box selection. The red points are potential targets that were not selected by the color-box method. The other colored regions are targets selected by the cuts in Equation (3.11) with the different offsets from Equation (3.12). The color-box selection cuts with the most targets is at the top (closest to the stellar locus, shown as grey contours), and the cuts get more restrictive moving towards the x-axis.

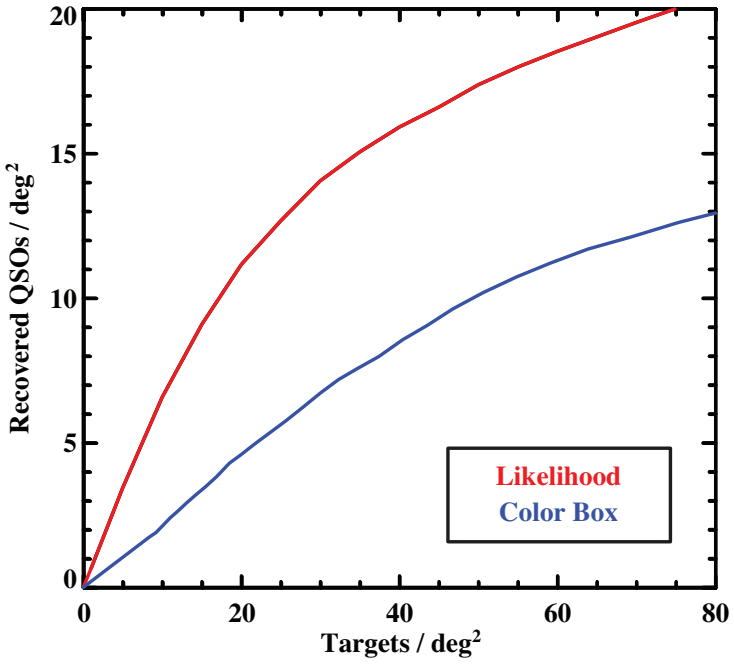


Figure 3.7: The number of BOSS $z > 2.2$ quasars recovered as a function of targets deg^{-2} for the likelihood method (red) and a traditional color-box technique (blue). Likelihood out performs the color-box selection method by recovering over twice as many BOSS quasars at 20 targets deg^{-2}

3.3.2 Luminosity Function Testing

I tested the performance of three different quasar luminosity functions (LFs) as inputs to the *QSO* Catalog. The LFs enter into the generation of the *QSO* Catalog by determining the density of quasars as a function of redshift and $i-$band magnitude. All the other details of the Monte Carlo remain the same as described in Section (3.1.3). The *EE* Catalog is not dependent on these LFs so this catalog stays the same for these tests. The three functions tested are from Jiang et al. (2006), Richards et al. (2006), and Hopkins et al. (2007a). The inputs and results from this testing is shown in Figure (3.8).

The quasar redshift distributions for these three luminosity functions are shown in the left panel of Figure (3.8). The performance of the method did not change significantly for the three different LF priors. Figure (3.8, *Center*) shows the number of quasars successfully recovered as a function of the number of dedicated QSO target fibers per deg^2. Notice the shape of this function, the rate of newly recovered quasars drops off significantly beyond 40 targets deg^{-2}. The performance of all three LFs is essentially identical up to 20 targets deg^{-2}. The redshift distributions of the recovered quasars are slightly different as shown in Figure (3.8, *Right*), so using different LF could be used to help tune the targeting redshifts. This method applied to BOSS targeting is not sensitive to uncertainties in the luminosity function.

Ultimately it was decided that Jiang et al. (2006) was the best luminosity function for my purposes because it was more efficient at recovering $z > 2.2$ QSOs than Richards et al. (2006) and Hopkins et al. (2007a). More detailed values for the performance of these different luminosity functions can be seen in Table (3.2).

While I am not currently making a proper comparison of the redshift distributions of BOSS quasars and the redshift distributions of these luminosity functions, a future improvement would be to use a luminosity function generated from the redshift distribution of BOSS quasars, properly adjusted for the targeting selection function imprinted upon it, as the input to the Monte Carlo to see if this approach improves target selection. Another promising modification would be to add the photometry from BOSS quasars to the inputs to the Monte Carlo simulation.

3.3.3 Weighted Likelihoods

I also tested adjusting Equation (3.7) to incorporate a weighting factor to optimize (in redshift-magnitude space) selection of objects with a high dark energy figure of merit (Albrecht et al. 2006). This weighting is done by simply adding a factor ($w_{O'}$) inside the product

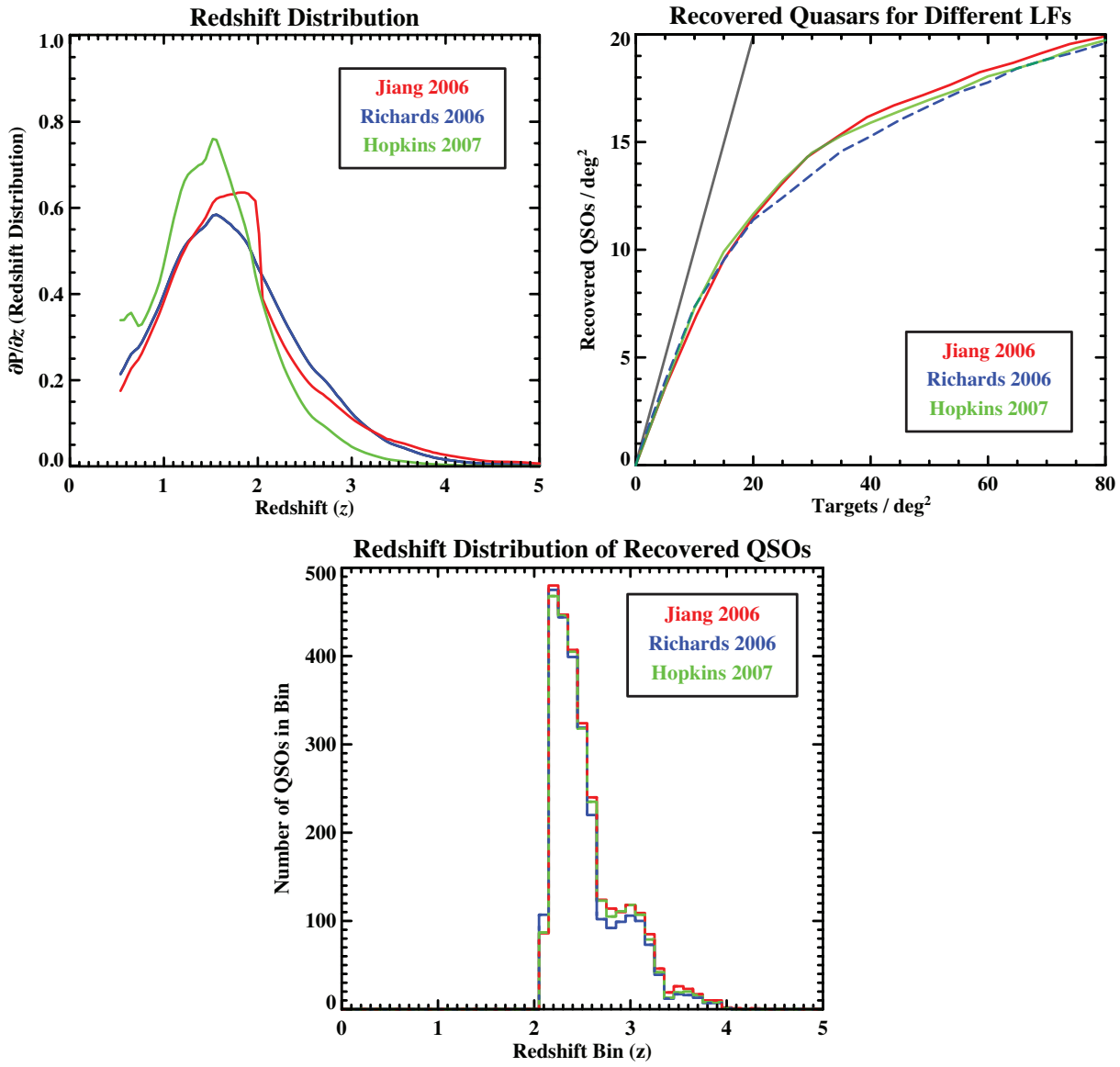


Figure 3.8: (*Top Left*) - Redshift distribution for Jiang et al. (2006) (red), Richards et al. (2006) (blue) and Hopkins et al. (2007a) (green) luminosity functions. (*Top Right*) - The number of BOSS $z > 2.2$ quasars recovered as a function of targets deg^{-2} for the three luminosity functions (i.e. different priors). The performance of all three LFs is almost identical for target densities up to ~ 20 targets deg^{-2}, at which point the Richards model starts to perform sightly worse. The gray line shows 100% efficiency, and emphasizes that very high efficiency is achieved if a small number of targets are selected. (*Bottom*) - The redshift distributions of the recovered $z > 2.2$ quasars for the three luminosity functions.

Table 3.2. Likelihood Luminosity Function Testing

Luminosity Function	Targets per deg^2	Likelihood $\mathcal{P}$ Threshold	Total Targets	QSOs Recovered	QSOs Missed	$\mathcal{C}$ (%)	$\mathcal{E}$ (%)
Jiang et al.	20	0.535	3878	2166	2401	47	56
	40	0.245	7757	3087	1657	65	40
Richards et al.	20	0.237	3978	2089	2466	46	54
	40	0.079	7757	2939	1808	62	37
Hopkins et al.	20	0.383	3878	2154	2401	47	56
	40	0.158	7737	3046	1676	64	39

Note. — The above table shows the completeness ($\mathcal{C}$) and efficiency ($\mathcal{E}$) as a function of dedicated target fibers (targets deg^{-2}) for three different luminosity functions. The three LFs tested are from Jiang et al. (2006), Richards et al. (2006), and Hopkins et al. (2007a). These values are for $z > 2.2$ recovered/missed QSOs.

based on the value of the QSO catalog quasar flux ($f^{O'}$) and redshift (z):

$$\mathcal{L}_{QSO}(\Delta z) = \sum_{O' \in QSO(\Delta z)} \prod_{f=\mathbf{f}} \frac{w_{O'}}{\sqrt{2\pi\sigma_f^2}} \exp\left[-\frac{(f - f^{O'})^2}{2\sigma_f^2}\right] \tag{3.13}$$

I tested adjusting the likelihood method in this manner with weights ($w_{O'}$) calculated by Pat McDonald (private communication, see Figure 3.9, and Table 3.3). Here the weight is determined by the contribution of the quasar's LyαF to the BAO signal; a higher weight yields a higher signal-to-noise BAO measurement.

The weight is a functional derivative of the overall BAO distance error squared with respect to the luminosity function. It can therefore be integrated over any achieved luminosity function (or summed over a set of quasars) to produce an estimate proportional to the BAO distance error squared that one would expect to achieve from that data set. There are two relevant factors affecting the value of a quasar: the noise level in the spectrum, and the density of quasars at a given redshift. The low redshift cutoff comes primarily from the degradation in the signal-to-noise ratio at the blue end of the spectrograph, while the high-z tail-off comes from the diminishing density of quasars with which to perform a cross-correlation for LyαF calculations.

While using these weights did recover QSOs with a higher on average BAO signal, as expected fewer total quasars were recovered using this scheme, see Figure (3.10). Ultimately it was decided to optimize the number of recovered quasars rather than the BAO signal for BOSS target selection. Therefore this weighting scheme was not used in the final likelihood

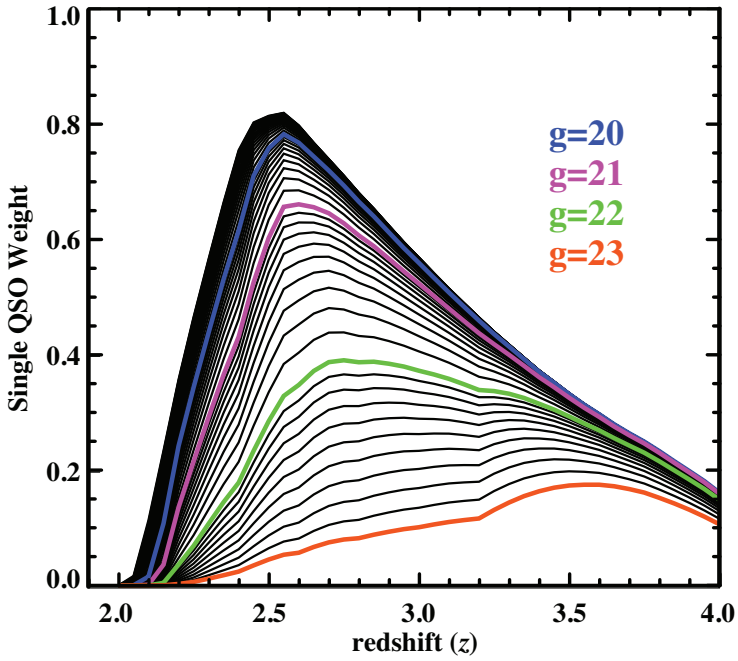


Figure 3.9: Weight representing the effectiveness as a dark energy BAO probe of a QSO as a function of magnitude and redshift. The lines are at 0.1 magnitude intervals. Brighter quasars have a higher weight, and so do quasars centered around a $z \sim 2.5$. For a detailed table of the numbers in this plot, see Table (3.3).

Table 3.3. Likelihood Weights

r-mag	$z =$ **2.0**	**2.25**	**2.5**	**2.75**	**3.0**	**3.25**	**3.5**	**3.75**	**4.0**	**4.25**	**4.5**
17.5	0.00	0.487	0.822	0.713	0.570	0.441	0.334	0.250	0.162	0.162	0.162
18.1	0.00	0.476	0.818	0.712	0.569	0.441	0.334	0.250	0.161	0.161	0.161
18.7	0.00	0.464	0.814	0.710	0.568	0.440	0.333	0.250	0.161	0.161	0.161
19.3	0.00	0.446	0.807	0.708	0.567	0.439	0.333	0.250	0.161	0.161	0.161
19.9	0.00	0.411	0.794	0.704	0.564	0.438	0.332	0.249	0.161	0.161	0.161
20.5	0.00	0.350	0.758	0.692	0.557	0.434	0.330	0.247	0.159	0.159	0.159
21.1	0.00	0.288	0.698	0.665	0.541	0.424	0.323	0.242	0.155	0.155	0.155
21.7	0.00	0.198	0.565	0.597	0.499	0.401	0.307	0.227	0.144	0.144	0.144
22.3	0.00	0.113	0.381	0.446	0.388	0.331	0.269	0.196	0.118	0.118	0.118
22.9	0.00	0.041	0.187	0.254	0.247	0.231	0.173	0.109	0.061	0.061	0.061
23.5	0.00	0.008	0.045	0.054	0.049	0.046	0.026	0.015	0.004	0.004	0.004

Note. — Shown is a subsample of the values for the weights in Equation (3.13). For a full length, downloadable, table of weights, please see the electronic version of the Likelihood paper Kirkpatrick et al. (2011).

targeting algorithm. However, depending on the goals of the user, a weighting scheme could be useful for future targeting purposes.

3.4 The *XDQSO* Likelihood Extension

The likelihood method inspired a similar targeting approach, extreme-deconvolution quasar targeting (*XDQSO*; Bovy et al. 2011b). The training sets used in *XDQSO* are almost identical to the *QSO* and *EE* Catalogs used in the likelihood method. However, instead of representing these catalogs as a discrete set of thousands (or millions) of objects, *XDQSO* uses an extreme-deconvolution (XD) fit to these catalogs, such that they are represented by a small set of Gaussian distributions. Below is a detailed description of the *XDQSO* density model work done with Jo Bovy.

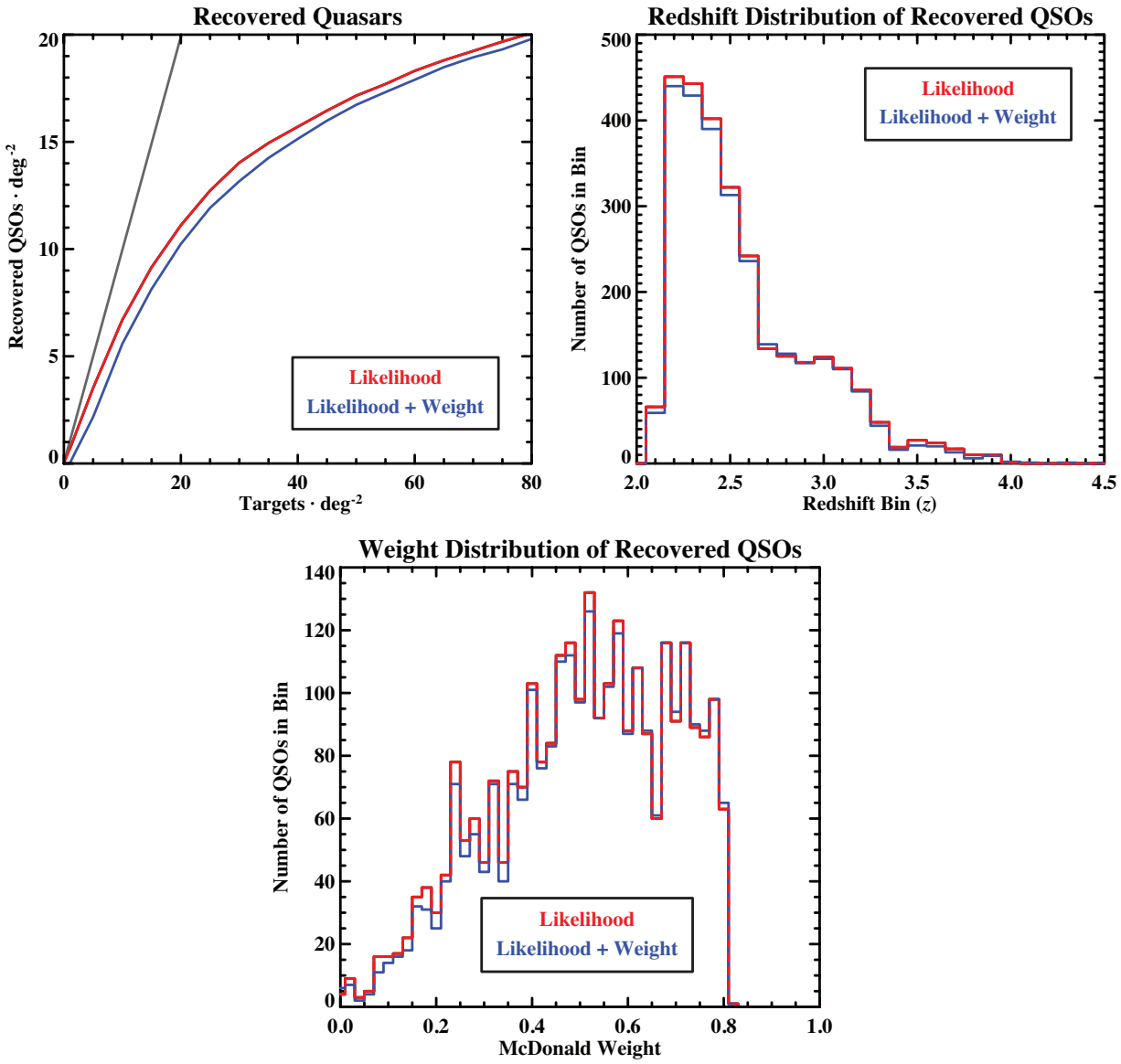


Figure 3.10: (*Top Left*) - The number of BOSS $z > 2.2$ quasars recovered as a function of targets deg^{-2} for the likelihood method with and without weights. The gray line shows 100% efficiency, and emphasizes that very high efficiency is achieved if a small number of targets are selected. (*Top Right*) - The redshift distributions of the recovered $z > 2.2$ quasars. (*Bottom*) - The weight distributions of the recovered $z > 2.2$ quasars. Notice that using the likelihood + weights recovers QSOs with a higher BAO value.

3.4.1 Extreme-Deconvolution Density Model

To estimate the density of stars and quasars in flux space we use XD[3] (Bovy et al. 2011a). At the faint end ($i \gtrsim 20$ mag) of the magnitude range of interest here the flux uncertainties in the training set are substantial, even though they are much smaller than the single-epoch uncertainties. Deconvolution therefore avoids adding in uncertainties twice at the density-evaluation phase. While XD can handle missing data as well, we do not need this feature here. However, this capability of XD is crucial when we want to extend the current framework to include near-infrared (NIR), ultraviolet (UV), or variability information, since these data will not be available for every object in the training set (see Section 3.5).

XD models the underlying, deconvolved, distribution as a sum of K Gaussian distributions, where K is a free parameter that needs to be set using an external objective. It assumes that the flux uncertainties are known, as is the case for point-spread function (PSF) fluxes for point sources in SDSS (Ivezić et al. 2003; Scranton et al. 2005b; Ivezić et al. 2007). XD consists of a fast and robust algorithm to estimate the best-fit parameters of the Gaussian mixture. For example, we were able to fit the color distribution of the full stellar catalog of 701,215 objects in only a few hours using 30 four-dimensional Gaussians. It is robust in the sense that even a poor initialization quickly leads to an acceptable solution. We are interested not so much in the true underlying distribution function as in finding a good fit to the observed density (after convolving the model with the data uncertainties) without overfitting, so it is not absolutely necessary to find the exact best fit in the complicated likelihood surface. Therefore, we use the simplest version of XD that does not use the heuristic search extension or priors on the parameters (Bovy et al. 2011a).

The XD method works by iteratively increasing the likelihood of the underlying, K Gaussian model given the data. It is an extension of the Expectation Maximization (EM) algorithm for fitting mixtures of Gaussians in the absence of noise (A. P. Dempster & Rubin 1977) to the case where noise is significant or there are missing data. The algorithm basically iterates through an expectation (E) and a maximization (M) step. During the expectation step the data and the current estimate of the underlying density are used to calculate the expected value of (a) indicator variables that for each data point indicate which Gaussian it was drawn from and (b) the true, noiseless value of each data point and its second moment. In the maximization step, these expected values are used to optimize the amplitude, mean, and variance of each of the K Gaussians. The E and M steps are iterated until the likelihood ceases to change substantially. The algorithm has the property that after each EM iteration the likelihood of the model is increased.

[3]Code available at `http://code.google.com/p/extreme-deconvolution/` .

3.4.2 Construction of the Model

The full model consists of fitting the flux density in a number of bins in i-band magnitude for the various classes of objects. We describe the model in a single bin first for a single example class. Throughout we will use the EE class as the example.

We opted for a binning approach because the true five-dimensional distribution of fluxes has a dominant power-law shape corresponding to the number counts as a function of apparent magnitude. However, most of the information for discriminating between quasars and stars is not in this power-law behavior, but in the behavior of colors or relative fluxes. While the latter can be represented well by mixtures of Gaussian distributions (Equation 3.14), the power-law behavior cannot without using large numbers of Gaussians. For this reason we chose to take out the power-law degree of freedom, i.e., the overall behavior of the density as a function of apparent magnitude. Neither the color distributions of quasars or that of stars is a strong function of apparent magnitude, such that the binning described below does not introduce strong assumptions about the behavior of the color distributions. Any weak magnitude dependence of the color distributions is captured in our model since we use narrow bins in i-band magnitude.

In a single bin in i-band magnitude, we separate the absolute flux from the flux relative to i in the likelihood in Equation (3.3) as follows:

$$p(\{f_j\}|O \in \text{`star'}) = p(\{f_j/f_i\}|f_i, O \in \text{`star'})\, p(f_i|O \in \text{`star'})\,, \tag{3.14}$$

where we now specify that the attributes $\{a_j\}$ are the $ugriz$ fluxes $\{f_j\}$, $\{f_j/f_i\}$ are the fluxes relative to i, and f_i is the i-band flux. We model these two factors separately.

We model the first factor using XD in narrow bins in i-band magnitude described in detail below. We use relative fluxes rather than colors—logarithms of relative fluxes—since the observational uncertainties are closer to Gaussian for relative fluxes than they are for colors. Except for the absence of a logarithm, relative fluxes are similar to colors and have the same number of degrees of freedom, viz., four. The fact that fluxes must be larger than zero while the Gaussian mixture model does not contain any such constraints, which is one reason to model the logarithm of the fluxes rather than the fluxes themselves, does not matter greatly here because most of the objects in the training set are at least five-sigma detections. However, for $z > 3$ quasars the u-band has zero flux and the flux measurement can be negative, in which case magnitudes are badly behaved; relative fluxes remain well-behaved in this case. To evaluate the XD probabilities during training, we always convolve the underlying model with the objects' uncertainties. We assume that the relative-flux uncertainties are Gaussian—which is a good assumption because the i-band magnitude is always measured at a reasonable signal–to–noise ratio—such that the convolution of the Gaussian mixture with the uncertainties results in a Gaussian mixture, with an object's uncertainty variance added to the model variance for each of the components.

Table 3.4. Total Number Counts in $17.75 \leq i < 22.45$.

	$z < 2.2$ Quasar	$2.2 \leq z \leq 3.5$ Quasar	$z > 3.5$ Quasar	Star
Number counts ($\deg^{-2}$)	140.72	50.70	6.13	5209.38

We model the four-dimensional relative fluxes $\{f_j/f_i\}$, which each come with an individual, non-diagonal, four by four uncertainty covariance, using 20 Gaussian components. The number 20 was decided upon as follows. For a few bins we performed XD fits with 5, 10, 15, 20, 25, and 30 components. While fits with less than 20 components overly smoothed the observed distribution, models with more than 20 components used the extra components to fit extremely low significance features in the observed distribution. Because of the scale of the full model (see below) no bin-by-bin objective method for setting the number of components was pursued, although we did verify that all of the resulting fits were reasonable.

We model the second factor, $p(f_i|i \in \text{'star'})$, by first combining it with the number factor $P(i \in \text{'star'})$. This combined factor becomes the number density as a function of apparent magnitude. This quantity will always be expressed in units of $\deg^{-2}$. For the EE class we model the number density directly using the number counts of the training data, by spline interpolating the histogram of i-band magnitude number counts per square degree. For the 'quasar' class we use a model for the quasar luminosity function (Hopkins et al. 2007b) to calculate the number density of quasars as a function of apparent i-band magnitude; we multiply these theoretical number densities with a simple model for the SDSS incompleteness of point sources

$$I(i) = \left(1 + \exp\left(\frac{i - 21.9}{0.2}\right)\right)^{-1}, \tag{3.15}$$

designed to reproduce the incompleteness as given in Abazajian et al. (2003). The $p(f_i|i \in$ class) factors for the various target classes are shown in Figure (3.11). The total number densities for the various classes are given in Table 3.4.

The full model consists of 47 bins of width 0.2 mag between $i = 17.7$ and $i = 22.5$, spaced 0.1 mag apart such that adjacent bins overlap. We further divide the 'quasar' class into three subclasses corresponding to low-redshift ($z < 2.2$), medium-redshift ($2.2 \leq z \leq 3.5$; the *BOSS* quasar redshift range), and high-redshift ($z > 3.5$) quasars. The XD fits for all but the brightest bin for a given class are initialized using the best-fit parameters for the previous bin. There are typically $\approx 20,000$ objects in each bin for the stellar training data; for the quasars there are 85,998 low-redshift, 14,060 medium-redshift, and 3,519 high-redshift quasars in each bin.

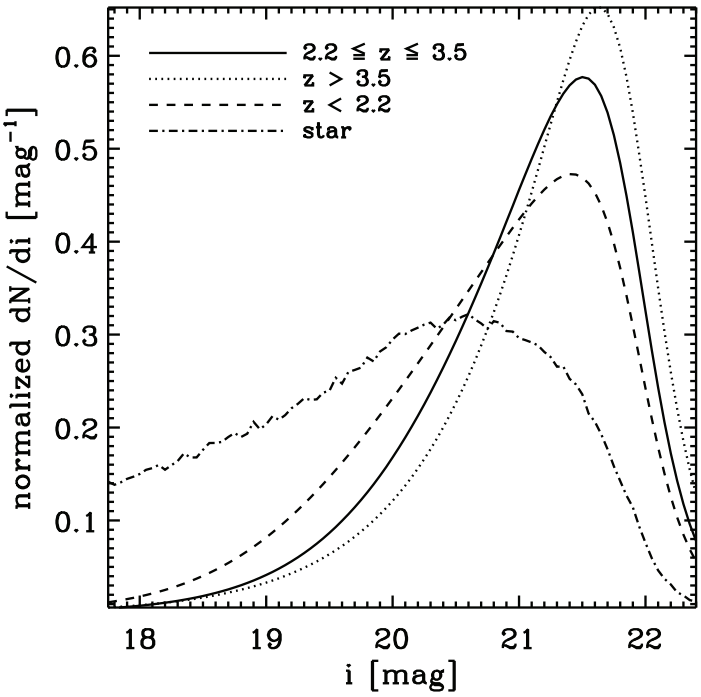


Figure 3.11: Number counts $p(f_i|i \in \text{class})$ for the various target classes. These have the expected property that higher redshift quasars are fainter since they are more distant.

In each of 4×47 bins we fit 20 four-dimensional Gaussians, yielding a total of $4 \times 47 \times (20 \times 15 - 1) = 56,212$ parameters.

3.4.3 Comparison Between *XDQSO* and Likelihood

The Likelihood method is very similar to the *XDQSO* however instead of using a small series of Gaussians to represent the data, it instead uses a delta function—corresponding to a model of the underlying density consisting of delta functions centered at the location of each object in the respective training set. These delta functions are convolved with the flux-uncertainties at evaluation such that a smooth density is obtained nevertheless. The Likelihood technique uses the same EE and QSO training data as the *XDQSO* catalog. Rather than simulating relative fluxes only, a full quasar catalog with five-dimensional *ugriz* fluxes in the relevant magnitude range is simulated.

The Likelihood technique uses the flux uncertainties to smooth the discrete underlying delta-function distribution of its training sets. However, since it does not use an optimized bandwidth, there is the danger that the density might be undersmoothed in certain regions. At the faint end, where the training set has a significant contribution from the uncertainties, the Likelihood method also effectively convolves with the uncertainties twice. This follows because the training set is a sample from the observed distribution of fluxes, rather than the true underlying distribution; the former is the latter convolved with the uncertainty distribution. Finally, the calculation of the quasar and star probabilities is extremely slow compared to *XDQSO*. In our comparison below we use cached versions of the Likelihood catalog created by the *BOSS* target selection team.

In Figure (3.12) we first compare the *XDQSO* quasar probabilities in the *BOSS* mid-redshift range with the probabilities calculated using the Likelihood method for sources in Stripe-82. We select targets at 20 targets $\deg^{-2}$ and show those targets selected by both techniques or by only one of the techniques. While many of the targets cluster around the one-to-one line, there is a distinct population of targets that receive high Likelihood probabilities, yet low *XDQSO* probabilities. A similar population of high *XDQSO*-only targets is absent, indicating that the Likelihood method indeed has problems with undersmoothing.

In Figure (3.13) we compare the *XDQSO* catalog with the Likelihood technique at different target densities. The two methods perform similarly, with a slightly better performance for the *XDQSO* catalog over the whole range.

A further comparison between the *XDQSO* and the Likelihood methods for medium-redshift quasar selection was performed during the Fall of 2010 using *BOSS* observations of an ≈ 200 $\deg^2$ region just north of Stripe-82. Both methods were given similar target densities to allow for a direct comparison. As of 2010 November 22, *XDQSO* was given 4,593 targets,

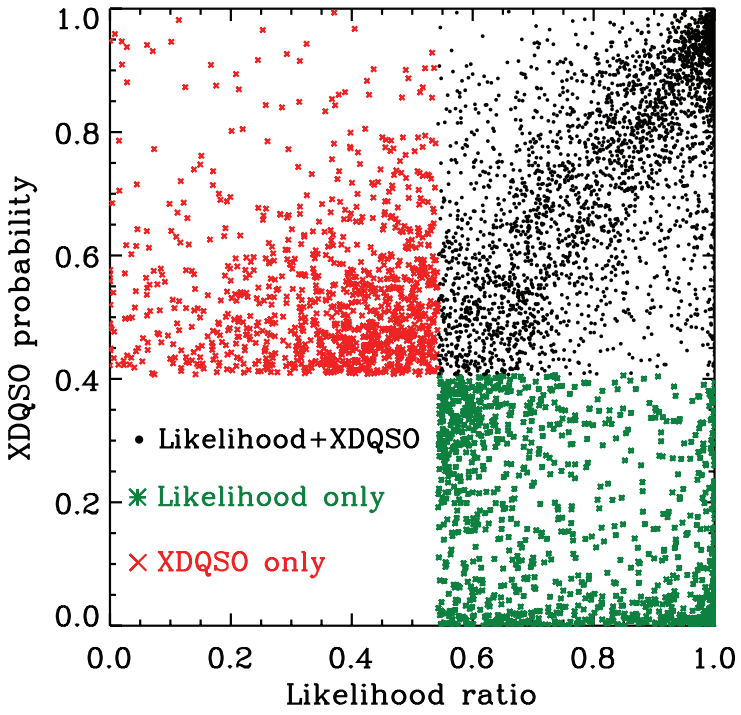


Figure 3.12: Comparison of the mid-redshift ($2.2 \leq z \leq 3.5$) quasar probability for the *XDQSO* and Likelihood methods at 20 targets deg^{-2} for sources in *SDSS* Stripe-82. Targets selected by both methods are on the upper right, Likelihood-only targets are on the lower right, and targets exclusive to *XDQSO* are on the upper left.

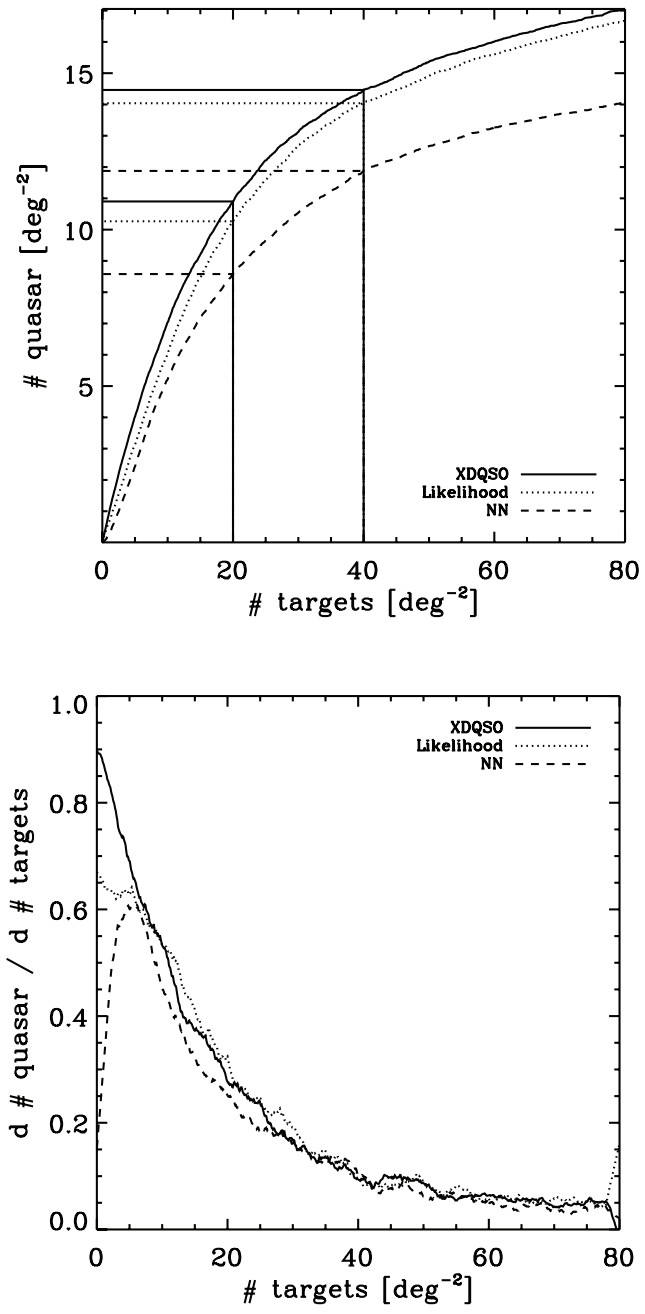


Figure 3.13: Number of confirmed $2.2 \leq z \leq 3.5$ quasars as a function of the target density for different target selection methods used in the *BOSS* (*XDQSO*, Bovy et al. (2011b); Likelihood, Kirkpatrick et al. (2011); NN, Yèche et al. (2010)). Input target densities relevant to the *BOSS* target selection are highlighted. This uses *BOSS* observations of sources in *SDSS* Stripe-82.

while Likelihood received 4,853 targets. Of the 4,593 *XDQSO* targets 2,194 were classified as $2.2 \leq z \leq 3.5$ quasars, while of the 4,853 Likelihood targets only 2,056 turned out to be medium-redshift quasars. From this test and that in Figure (3.13) we conclude that the *XDQSO* technique's performance is about 10 percent better than that of the Likelihood method for the selection of medium-redshift quasars.

3.5 Further Extensions

One of the main advantages of the general target selection technique and in particular of the specific Likelihood and *XDQSO* implementation described in this chapter is that it can easily be extended in a variety of ways. These extensions can be changes to the model—such as different number count priors, the addition to the model of other data such as NIR or UV observations—or the combination of the flux-based selection described here with target selection based on variability information. All of these extensions are described briefly here.

Most of these extensions involve only some of the factors in Equations (3.3) and (3.14). Since we provide all of these factors separately in the method catalogs, extensions that do not change all of the factors can use some of the information in the catalog. For example, extensions that only change the number count priors, e.g., $p(f_i \mid i \in \mathrm{EE})$ or $P(i \in \mathrm{EE})$, will not need to re-calculate the relative-flux likelihoods—the most expensive of the factors in Equation (3.14)–but can instead re-use the catalog values.

3.5.1 Additional NIR or UV data

Quasar selection from broad-band fluxes can be improved by the addition of NIR or UV fluxes to the optical fluxes used to create the Likelihood or *SDSS-XDQSO* catalogs (e.g., Warren et al. 2000; Maddox et al. 2008; Richards et al. 2009b; Jimenez et al. 2009; Worseck & Prochaska 2011). For example, the *Galaxy Evolution Explorer* (*GALEX*; Martin et al. (2005)) has completed a near full-sky survey in the ultraviolet (UV) and the *UKIRT Infrared Deep Sky Survey* (*UKIDSS*; Lawrence et al. (2007)) is observing a large part of the SDSS footprint in the NIR. However, this situation is complicated by the fact that these surveys are generally shallower than the optical fluxes available from *SDSS*, such that most of the objects in the SDSS catalog are not detected at high significance in these surveys.

Since these surveys have point-spread functions that are worse than that of the *SDSS*, low signal-to-noise measurements of the NIR and UV fluxes of many of the objects in the SDSS catalog can be obtained by forced photometry of the *GALEX* or *UKIDSS* images at the SDSS positions, which can be regarded as truth because of the difference in resolution. Because there are gaps in these surveys, it will still be the case that some objects in both

the training set and the evaluation set will not have measured NIR or UV fluxes.

To use these low signal-to-noise or non-existent fluxes, it is necessary to employ a classifier than can handle the data uncertainties correctly and that can handle missing data, both for training and for evaluation of the quasar probabilities. The *XDQSO* method described in here is the only technique to date that can do this task naturally and it is therefore the only method available that—in a straightforward way—can use all of the available information for an object to classify it as a star or quasar.

3.5.2 Variability

With the opening of the time-domain in the near-future with surveys such as *Pan-STARRS* Kaiser et al. (2002); Morgan et al. (2008), *LSST* Ivezić et al. (2008); LSST Science Collaboration (2009), *Skymapper* Keller et al. (2007), and *WFIRST*, the selection of quasars based on their variability has recently received some attention (Kozlowski et al. 2010; Schmidt et al. 2010; Butler & Bloom 2011; MacLeod et al. 2011). Some of these techniques currently amount to drawing the equivalent of "color-boxes" in variability space to select quasars (Schmidt et al. 2010; MacLeod et al. 2011), while others perform more sophisticated model selection (Butler & Bloom 2011). However, it is clear that the variability technique could be brought under the umbrella of probabilistic target selection by doing density estimation in the space of variability attributes (such as parameters of the structure function) in a similar manner as we did in flux space in this chapter.

The variability data can be used to form a variability-likelihood similar to the relative-flux likelihood used in Equation (3.14). If we assume that the variability of a quasar is independent of its (relative) flux—not necessarily a good assumption—we can combine the relative-flux and variability likelihoods by simply multiplying the quantities. Alternatively, we can perform density estimation in the combined space of relative fluxes and variability parameters, and use the combined likelihood instead of the relative-flux likelihood in Equation (3.14)—this will capture any (relative) flux dependence of the variability of quasars. Combining variability and flux information is our best hope to create extremely efficient quasars surveys in the future that are free from the biases associated with color or variability selection alone.

3.5.3 Other Extensions

The data used as inputs to Equation (3.14) can be further improved upon using existing data and the *XDQSO* target selection technique. For example, we used a star count model that is not a function of Galactic coordinates, but we know that the number density of stars

is a strong function of Galactic latitude and Galactic longitude. Both the total stellar number counts in Table 3.4 and the stellar number counts as a function of apparent magnitude in Figure (3.11) could be re-calculated using models for the stars counts in different directions. However, this will not lead to significant changes in the calculated quasar and star probabilities, as the distribution on the celestial sphere of photometrically selected quasars and stars already follows the expected celestial distributions of quasars and stars quite well (Figure 3.14).

We also used relative-flux distributions for the stars that do not depend on the celestial location of an object. However, the colors of stars do change with proximity to the plane of the Galaxy. Therefore, we could have used a model that reconstructs the relative flux distribution of stars as a function of the position on the sky. Such a model is hard to produce from the current data, if we do not want to rely on theoretical models for this dependence, since we have no way *a priori* to separate quasars from stars all over the sky. What we can do is model the relative-flux distribution of all point sources as a function of position on the sky, from the single-epoch SDSS fluxes available on the SDSS footprint. In order to use the noisy single-epoch fluxes properly, it is again necessary to use a technique such as *XDQSO* that uses the uncertainties correctly. Using the model for the quasar fluxes that we have been using in this chapter we can calculate quasar probabilities by using the quasar model in the numerator of Equation (3.2) and the model for all point sources in the denominator. However, it is then possible for the probability to exceed one, since the probability is not explicitly normalized.

We could also use a training set consisting of point sources over the entire SDSS footprint rather than the 150 deg^2 area of Stripe-82 to increase our sampling of rare stellar-locus outliers. As mentioned above, red stellar-locus outliers outnumber high-redshift quasars and are a significant contaminant for high-redshift quasar selection. Our current training set does not contain enough of these red stellar outliers to model their relative flux distribution. By extending our stellar training sample to the full $\approx 10^4$ deg^2 SDSS footprint we would have about 100 times more stellar outliers and we could model their color distribution. This would significantly improve high-redshift quasar target selection.

3.6 Summary and Conclusions

In this chapter I :

- Developed a new "Likelihood" method for quasar target selection using photometric fluxes and a Bayesian probabilistic approach;
- Demonstrated that this leads to the recovery of 15.9 ($z > 2.2$) quasars deg^{-2} from the

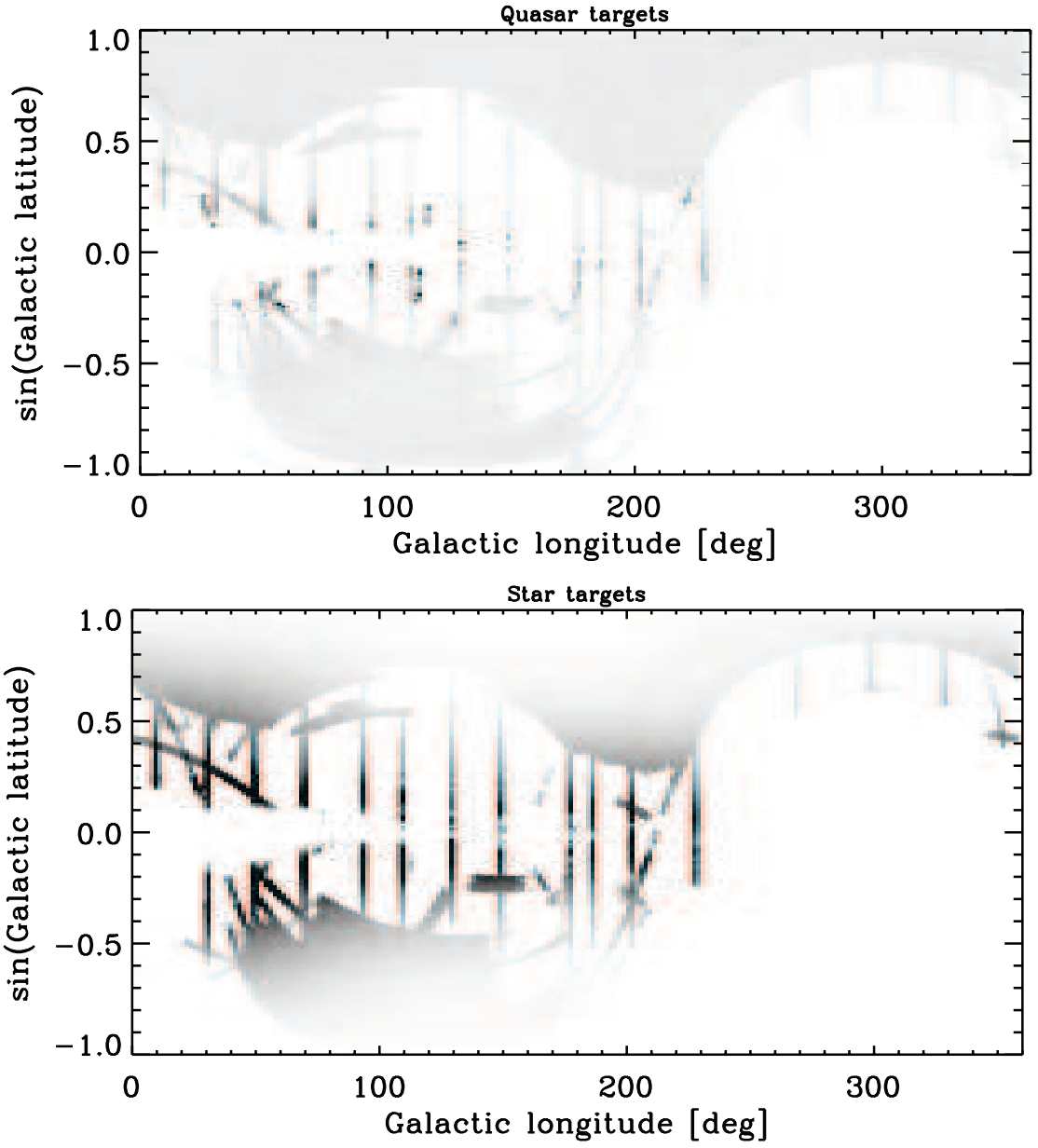


Figure 3.14: Sky distribution of quasar ($P(\text{quasar}) \geq 0.5$) and star ($P(\text{star}) \geq 0.95$) targets. The contrast for the star targets is saturated near the Galactic plane.

SDSS Stripe-82 dataset when targeting at 40 targets deg^{-2}, with a completeness of 65% and efficiency of 40%;

- Showed that the likelihood method recovers twice as many quasars as traditional "color-box" selection;
- Tested for the effects of changing the input QSO catalog, using different luminosity functions and adding a weighting scheme to the likelihood calculations.
- Discussed in detail an extension of the "Likelihood" method: *XDQSO* .
- Compared these two methods and their respective benefits and drawbacks.
- Described future extensions to these methods.

After a commissioning period in September-November 2009, the QSO targeting fibers were divided into a uniformly selected CORE sample and a non-uniformly selected BONUS sample (Ross et al. 2012). The likelihood method, using the Jiang luminosity function, was used for targeting the CORE sample (20 targets deg^{-2}) for the first year of BOSS data taking. The rest of the fibers (BONUS sample) were targeted by a combination of the output of the likelihood, KDE and NN methods using a neural network. This approach allows the team to combine both different methods and different photometric catalogs (SDSS, UKIDSS, GALEX) in the BONUS selection.

After the first year of data taking, the CORE sample targeting switched to using the *XDQSO* method (Bovy et al. 2011b) and likelihood was then used in the BONUS sample as well as one of the inputs to the NN. This switch was made because *XDQSO* performed slightly better at recovering high-redshift quasars and the priority of the target selection team was to maximize number of selected quasars. I released the probabilities from Equation (3.10) in the project data releases of SDSS data.

The Likelihood and *XDQSO* target selection technique can be extended to include low signal-to-noise data in NIR and UV filters and to include other information such as quasar variability. It is the low signal-to-noise ratio regime at the faint edge of surveys that often contains the most interesting objects. *XDQSO* is the only target selection technique currently available that can calculate robust quasar probabilities taking the data-uncertainties fully into account. Since the most successful photometric quasar catalogs are based on calculating good photometric quasar probabilities (Richards et al. 2009a), these techniques or similar will be essential to create the best and largest photometric quasar catalogs in upcoming surveys such as *Pan-STARRS* and *LSST*.

Chapter 4

BOSS QSO Target Selection Results

4.1 Results

This chapter presents the results of quasar spectroscopy carried out during the Year One of BOSS data taking — from MJD 55176 (2009 December 11) through MJD 55383 (2010 July 06). These results are described in greater detail in Ross et al. (2012). The distribution of BOSS Year One quasars on the celestial sphere is shown in Figure (4.1).

4.1.1 Global Properties and Efficiencies

There were 54,909 spectra of objects targeted as quasars, of which 52,238 were unique objects. These were observed over over a footprint of 878 deg^2, giving a mean surface density of 63.8 targets deg^{-2}. Of the 54,909 (52,238 unique) spectra, 35,305 (33,556) had high-quality redshifts. Of the 33,556 unique objects with high-quality redshifts, 11,149 are stars, while 13,580 are QSOs with $z > 2.20$. The remaining 8,827 objects are mostly quasars at $z \sim 0.84$ and $\sim$1.6, and low-z compact galaxies; see Figure (4.2). Of the 13,580 high redshift objects, 2,317 were previously spectroscopic confirmed quasars; thus the first year of BOSS observations resulted in the spectroscopic confirmation of 11,263 new $z > 2.2$ quasars.

Figure (4.2) shows the redshift distribution of BOSS quasars from the first year, and compares it with that from the SDSS DR7 quasar sample (Schneider et al. 2010) and the 2SLAQ survey (Croom et al. 2009). This plot is very similar, but not identical, to that shown in the SDSS-III overview paper of Eisenstein et al. (2011). Of course, the DR7 sample is selected over the full SDSS-II imaging area, approximately 9,380 deg^2, while the BOSS Year One data come from observations of 880 deg^2. Already BOSS has slightly more quasars in the $z = 2.2 - 2.8$ range, while at higher redshifts the DR7 sample remains larger.

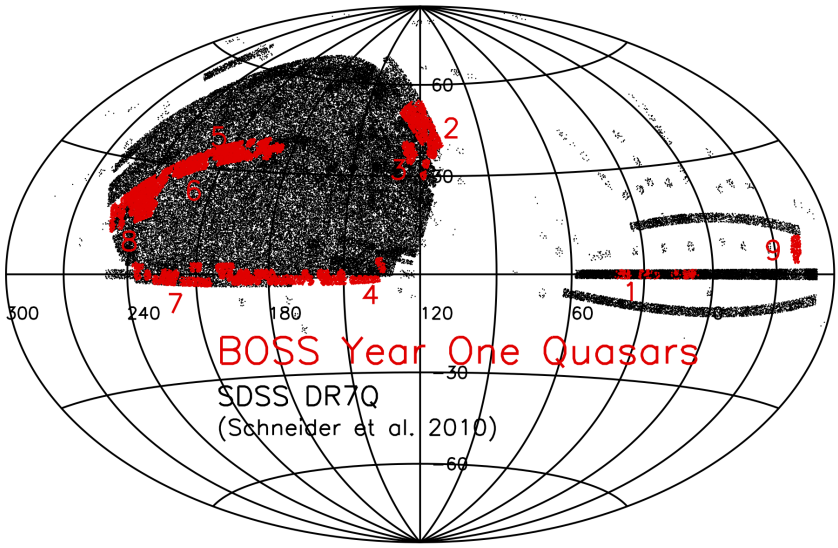


Figure 4.1: Sky distribution of the 14,287 quasars in the BOSS Year One quasar survey (J2000 equatorial coordinates), in red. The nine regions ("chunks") observed by BOSS are labeled accordingly, and the dotted lines are drawn at Galactic latitudes $b = \pm 25°$. The spectroscopically confirmed SDSS-I/II DR7 quasar catalog (Schneider et al. 2010) is shown for comparison in black.

Degeneracies in the color-redshift relation of quasars lead to the selection of low-z quasars in BOSS. The quasars at $z \sim 0.8$ have Mg II λ2800 Å at the same wavelength as Lyα at redshift $z \sim 3.1$, giving these objects similar broad-band colors, while the large number of objects at $z \sim 1.6$ is due to the confusion between λ1549 C IV and Lyα at $z \approx 2.3$. We shall come back to this feature when comparing the performance of the NN, KDE, and Likelihood methods in § 4.1.3. The tail of objects at $z \gtrsim 3.5$ includes a significant contribution from re-observations of previously known quasars.

Figures (4.3) and (4.4) present our key results, the efficiency of the current target selection algorithms. For these tests, we have constructed a control sample of targets on Stripe-82, where our spectroscopy is more complete than anywhere else on the sky, albeit still not perfect. Here we include data from Year Two, where Stripe-82 was retargeted using a variability selection for quasars (Palanque-Delabrouille et al. 2011). Stripe-82 also has high completeness because quasars are selected from co-added photometry, with much smaller photometric errors.

For Figure (4.3), we select the quasar targets in our normal way from *single-epoch* data, with the first 20 targets deg^{-2} selected by the *XDQSO* CORE algorithm. Targets are ranked in order of probability, and the plot shows the number of $z > 2.2$ quasars deg^{-2} vs. the

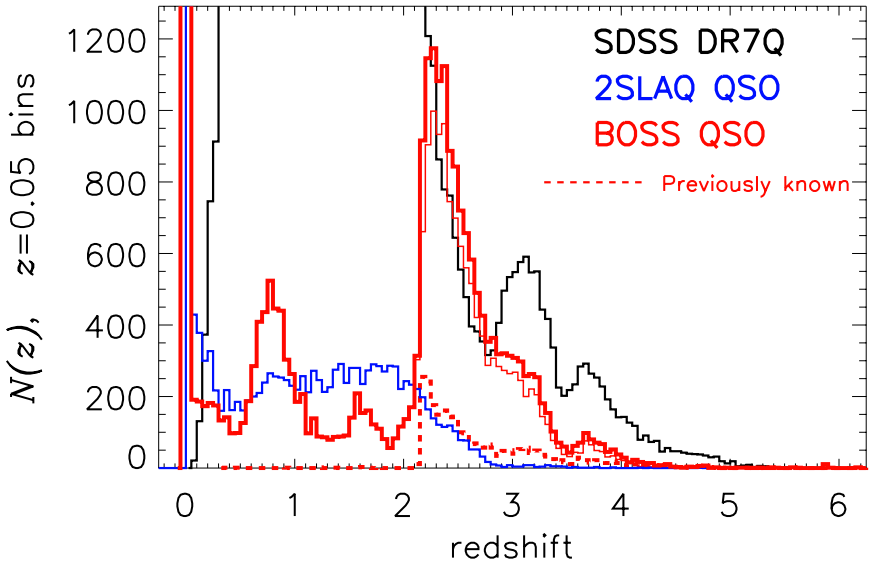


Figure 4.2: The redshift histogram of BOSS Year One quasars (solid red thick histogram). The dashed red line represents those objects known prior to BOSS observations, while the distribution of newly confirmed quasars is given by the thin red line. For comparison the SDSS DR7 quasars from Schneider et al. (2010) (selected over a much larger sky area) are shown by the black histogram, while the 2SLAQ quasar data (Croom et al. 2009), are in blue.

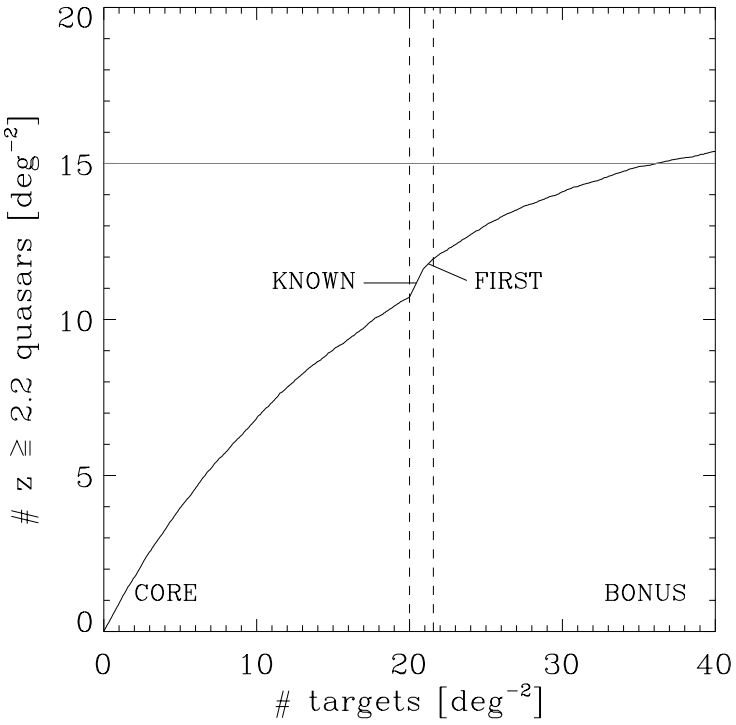


Figure 4.3: Cumulative number of quasars with $z > 2.2$ as a function of the rank of the target for the Stripe-82 control sample with single-epoch photometry. At 20 fibers deg^{-2}, the *XDQSO* CORE algorithm selects 10.7 quasars deg^{-2}, while previously known and FIRST sources add an average of 1.5 quasars deg^{-2}. At 40 fibers deg^{-2}, the total surface density of $z > 2.2$ quasars selected by our current algorithms from single-epoch SDSS photometry is 15.4 deg^{-2}. Note that these numbers represent an average over a wide range of Galactic latitudes, and therefore stellar contaminations.

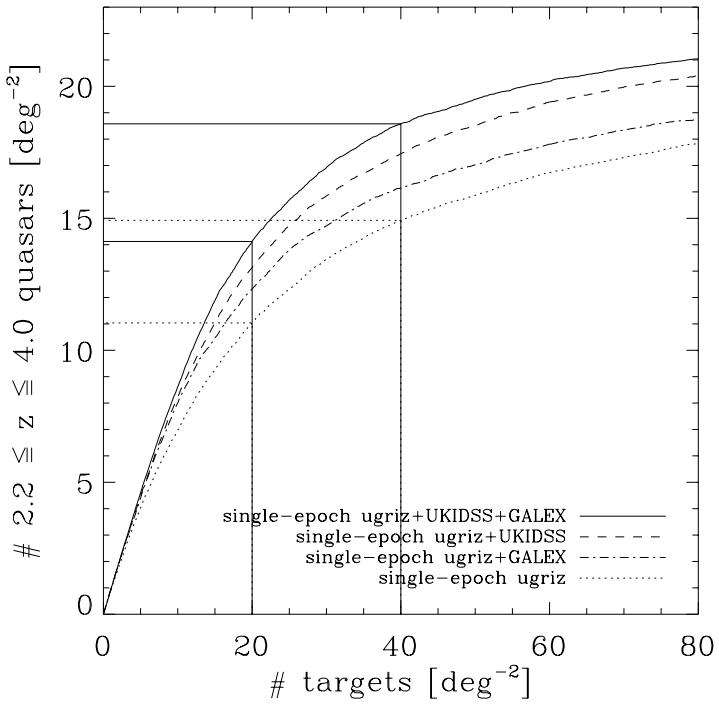


Figure 4.4: Similar to Figure (4.3), but showing the impact of adding GALEX photometry, UKIDSS photometry, or both to SDSS single-epoch photometry. This Figure is based on Stripe-82 data and *XDQSO* selection for all targets.

number of targets deg^{-2}, with the slope of the curve indicating the efficiency of selection. The CORE algorithm selects 10.7 $z > 2.2$ quasars deg^{-2} from its 20 targets. We then show the average contribution of KNOWN and FIRST quasars, totaling 1.6 high-z quasars deg^{-2}. This increment assumes a surface density of 0.9 known high-z quasars deg^{-2} (and 0.7 deg^{-2} from FIRST), which is consistent with our Year One data but lower than the surface density of known pre-BOSS high-z quasars on Stripe-82, which is unusually well studied. Finally, we add the BONUS targets from the NN-combinator, again in rank order. At 40 targets deg^{-2}, we are just above the minimum BOSS goal, with a mean density of 15.4 $z > 2.2$ quasars deg^{-2}. Stripe-82 samples a wide range of Galactic latitude and thus stellar density; we therefore anticipate that this test should be representative of selection efficiency averaged over the full BOSS survey region. We also found from early observations, that adding additional fibers beyond the nominal 40 deg^{-2}, led to only very minimal gains in yield.

Figure (4.4) shows the impact of adding UKIDSS and GALEX data to single-epoch SDSS photometry. For this test we use the *XDQSO* algorithm alone, since this is where these auxiliary data sets currently enter our selection procedures, and we extend the efficiency curves up to 80 targets deg^{-2}. At 40 targets deg^{-2}, the efficiency for *XDQSO* with single-epoch SDSS imaging alone is 15.0 $z > 2.2$ quasars deg^{-2}. Adding GALEX data improves the efficiency to 16.2 deg^{-2}, adding UKIDSS improves it to 17.3 deg^{-2}, and adding both improves it to 18.6 deg^{-2}. Thus, both of these data sets can significantly enhance the efficiency of BOSS quasar target selection in regions where they are available. Stripe-82 has medium-deep ("MIS") GALEX data, and the improvement with shallower ("AIS") coverage will be smaller, but our tests indicate that GALEX addition will still improve the selection.

Figure (4.5) shows the redshift distribution of all known quasars on Stripe-82 as a function of redshift, as well as those selected by the single-epoch SDSS algorithms illustrated in Figure (4.3) above. The ratio of the two measures the completeness of BOSS single-epoch quasar selection relative to known quasars in this well studied region, ranging from 40% to 70% over our critical redshift range $2.2 < z < 3.5$. Of course, this remains a lower limit to the true completeness at the BOSS magnitude limit, though in the $2.2 < z < 3.5$ redshift range we anticipate that the BOSS Stripe-82 sample selected from co-added photometry and variability has high completeness (Palanque-Delabrouille et al. 2011).

Figure (4.6) shows examples of BOSS spectra of quasar targets from the Year One data. From top to bottom: a $z > 5$ quasar found by the Likelihood method (and not selected by any other method); a newly discovered $z = 2.6$ quasar at a typical S/N; a $z = 3.5$ quasar selected only by the KX method; a re-observed BAL quasar showing spectroscopic variability over 3377 days in the observed frame; a star at our typical S/N; and a $z = 1.5$ quasar with our typical S/N.

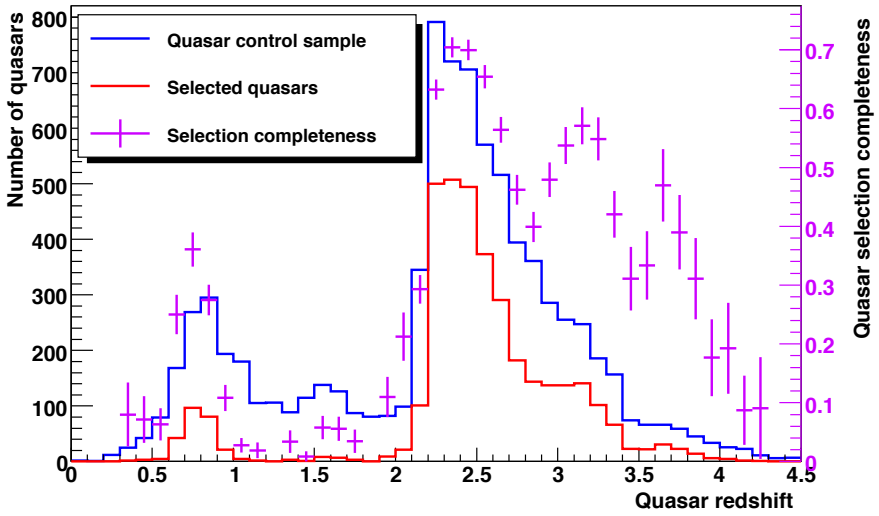


Figure 4.5: Completeness of BOSS single-epoch target selection vs. redshift, on Stripe-82. The blue histogram shows the redshift distribution of all spectroscopically confirmed quasars on Stripe-82. The red histogram is for those quasars that pass the BOSS single-epoch target selection for a threshold tuned to produce 40 targets deg^{-2}. Purple points with Poisson error bars show the ratio of the two, i.e., the selection completeness (right-hand scale).

4.1.2 Magnitude, Color and the $L - z$ Plane

Figure (4.7) shows the distribution of quasar targets from the BOSS first-year data which are spectroscopically confirmed as either stars or $z > 2.2$ quasars, in the *(u-g)* vs. g color-magnitude plane. The distribution of stars at the bright end, $g < 18$, and the lack of bright $z > 2.2$ quasars, led us to impose the bright $i = 17.8$ limit. Objects fainter than $g = 22$ are brighter than our r band limit of 21.85 mag.

Figure (4.8) shows the SDSS $(u - g)$, $(g - r)$, $(r - i)$, and $(i - z)$ colors as a function of redshift for the BOSS Year One data. Also shown are the mean color in redshift bins (thin solid line), and the model of Bovy et al. (2011a) (thick colored line). This model is systematically bluer than the data at low redshift; BOSS target selection systematically excludes UV-excess quasars, and thus those low-redshift quasars that happen to enter the sample are redder than the average quasar. The trends with redshift are due to various emission lines moving in and out of the SDSS broadband filters, and the onset of the Lyα forest and Lyman-limit systems Fan (1999); Richards et al. (2002); Hennawi et al. (2010); Bovy et al. (2011b); Peth et al. (2011); Prochaska & Hennawi (2009); Worseck & Prochaska (2011). McGreer et al. (2012, *in prep*) will present a detailed analysis of this diagram, and its implications for our completeness.

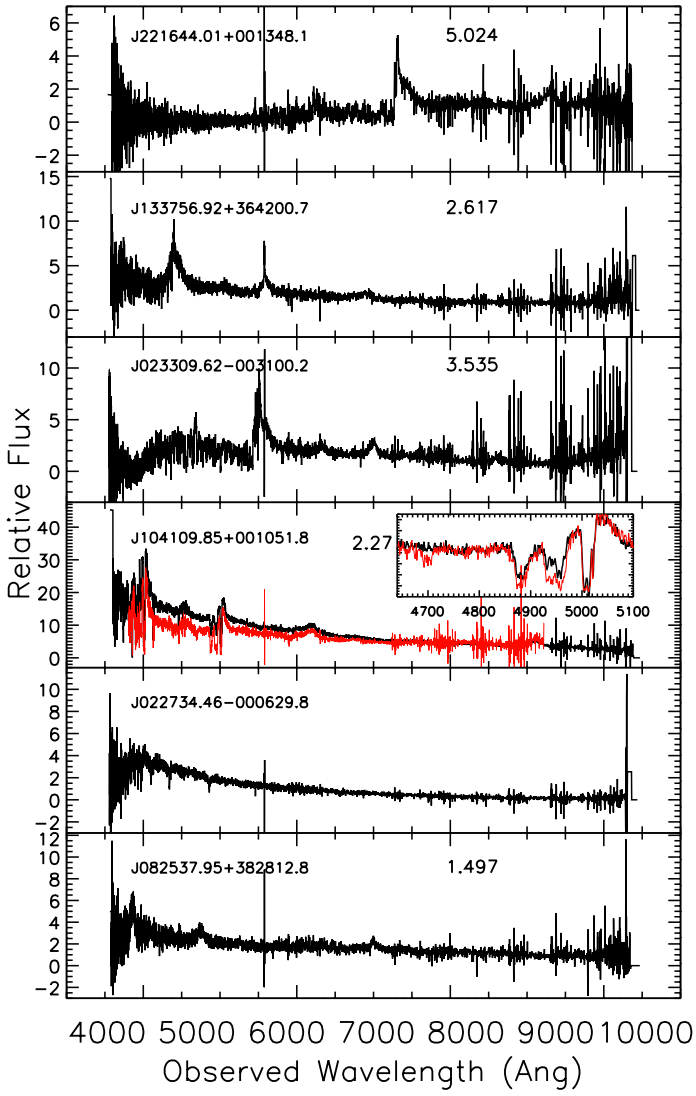


Figure 4.6: Examples of spectra of BOSS quasar targets. The SDSS object name and pipeline redshift are given in each panel (except for the star). From top to bottom: a $z > 5$ quasar found by the Likelihood method; a newly discovered $z = 2.6$ quasar at the typical S/N; a $z = 3.5$ quasar selected only by the KX method (§ 2.2.6); a re-observed BAL quasar showing spectroscopic variability (black line is the BOSS spectrum; red is from SDSS, a spectrum taken 3377 days earlier); a star with our typical S/N and a $z = 1.5$ quasar with our typical S/N. The feature at 5577Å in all spectra is a residual from a sky line.

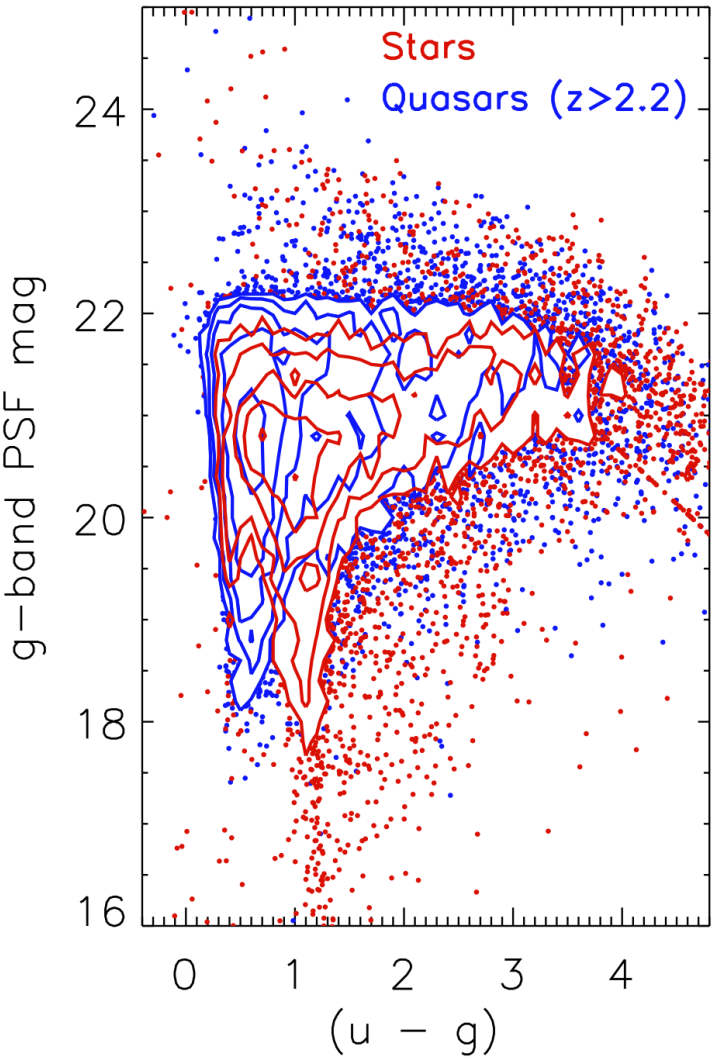


Figure 4.7: Color-magnitude diagram ($(u-g)$ vs. g) for objects spectroscopically classified as stars (red contours and points) and $z > 2.2$ quasars (blue contours and points). Only objects with good spectra are shown. The quasars are systematically bluer; there are very few quasars with $g < 18$.

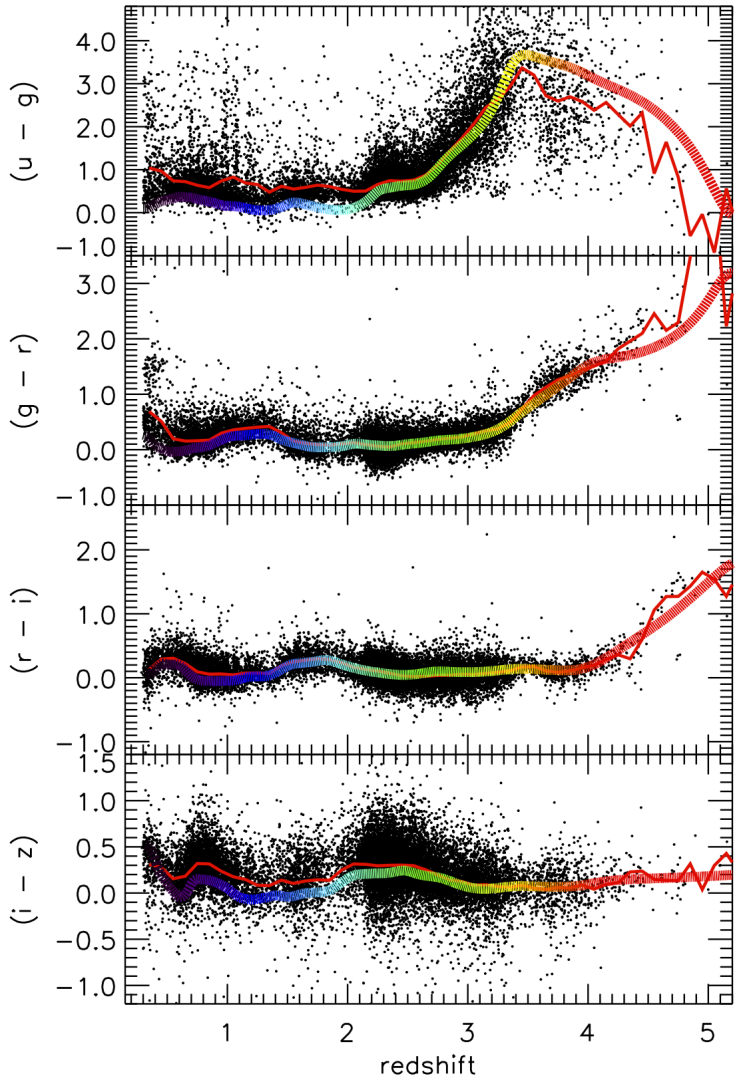


Figure 4.8: SDSS colors vs. redshift for quasars in the BOSS Year One data. The thin solid line is the mean color in bins of redshift, while the thick colorful line is from the model of Bovy et al. (2011a). The model is systematically bluer than the data at low redshift because BOSS systematically excludes UV-excess sources.

Table 4.1. The number of targets Year One Data

Selection Method	# Quasar targets	# with good spectra	and with $z > 2.20$	or are stars
Totals	52,238	33,556	13,580	11,149
KDE	34,503 (4794)	20,993 (2693)	9,050 (229)	7,607 (1,856)
NN	16,747 (975)	13,267 (710)	7,743 (135)	3,604 (504)
Likelihood	29,150 (2325)	21,975 (1647)	11,244 (447)	4,483 (724)

Note. — The number of unique quasar targets from the first year of BOSS spectroscopy, broken down by the three key selection methods. Numbers in parentheses indicate the number of objects selected by the indicated method *only*. Because these methods were applied non-uniformly, this table is provided as an *informational guide*, and not as a direct comparison between methods (see text for further explanation).

Figure (4.9) shows the SDSS color-color diagrams for the first year BOSS quasars, for all quasars with good redshifts above $z = 2.2$. This figure illustrates the redshift dependence of quasar colors as the Lyα emission line moves from the g band to the r-band at $z \approx 3.5$. Quasars with $2.2 < z < 3.5$ lie in the range $-0.3 < (g-r) < 0.6$, while objects with $z > 3.5$ generally have $(g-r) > 0.8$.

Figure (4.10) shows the distribution of objects in the redshift-luminosity ("$L-z$") plane for three recent large quasar surveys: SDSS (black points), 2SLAQ (cyan) and BOSS (red). There are $\approx 105,000$ objects in the SDSS DR7 catalog, and $\approx 9,000$ $g \leq 21.85$ low-redshift quasars from the 2SLAQ Survey (Croom et al. 2009). We calculate the absolute i-band magnitudes, M_i, using the observed i-band PSF magnitudes and the k-corrections given in Table 4 of Richards et al. (2006). The three surveys together cover the $L-z$ plane well, with a dynamic range in luminosity of ≈ 4 magnitudes at any given redshift up to $z \sim 3.5$. This coverage will be vital for calculating the evolution of the faint end of the quasar luminosity function, and placing strong constraints on the luminosity dependence of quasar clustering.

4.1.3 Comparison of Algorithms

The original motivation for the implementation of multiple target selection algorithms was the lack of evidence prior to BOSS observations that a single method could select $z > 2.2$ quasars down to $g \approx 22$ with our required efficiency. With the Year One data now in hand, we can compare the effectiveness of our different methods. However, due to the continually changing nature of the BOSS QTS over this year, where different methods were used as CORE and BONUS, these comparison will be generally qualitative in nature. The

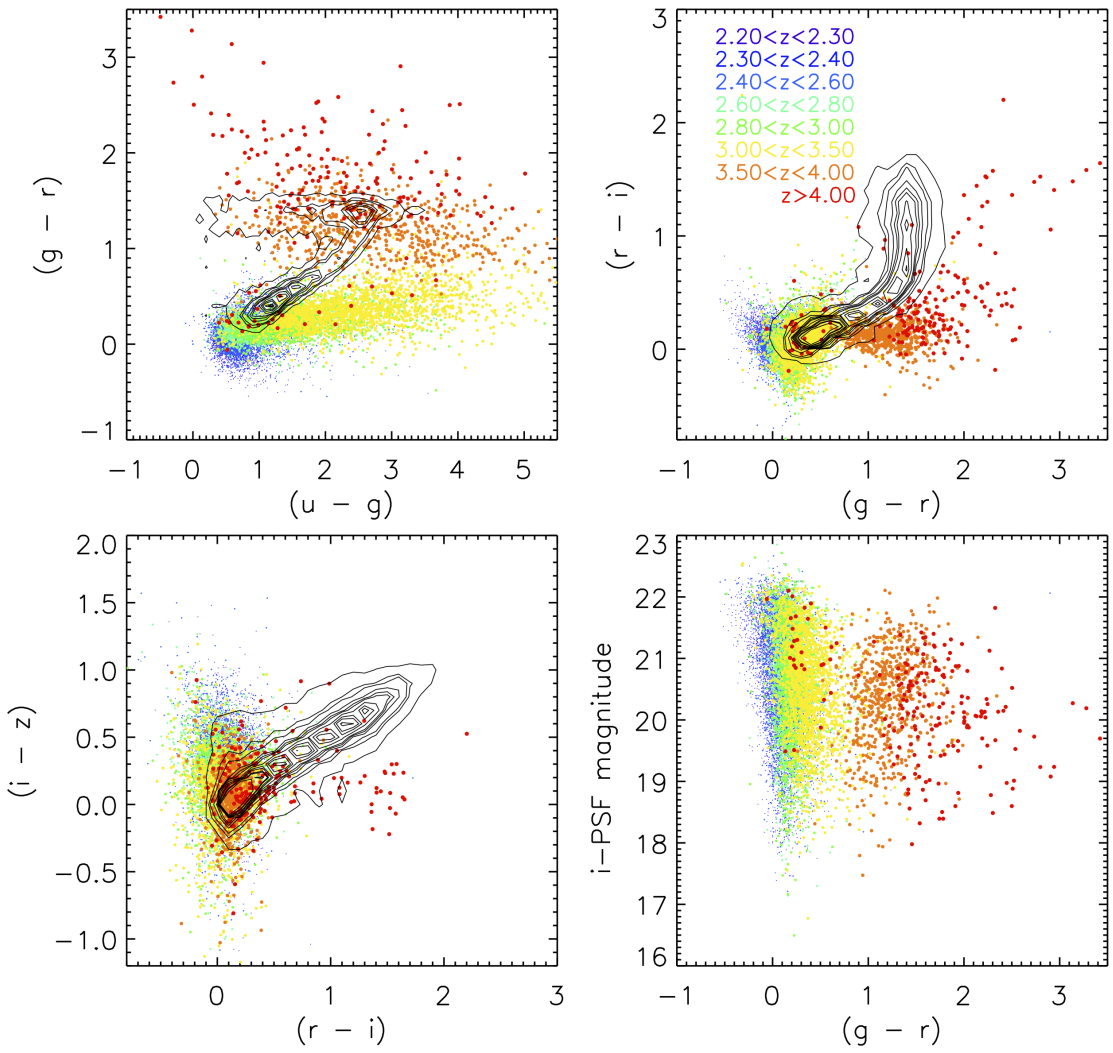


Figure 4.9: Color-color diagrams for the First Year data for all spectroscopically confirmed quasars with good redshifts above $z = 2.2$. The stellar locus is shown as contours. (*Top Left*), *ugr*; (*Top Right*), *gri*, (*Bottom Left*), *riz*. The horizontal swath of both stars and quasars at $g - r \sim 1.5$ in the $u - g, g - r$ color-color diagram is caused by the large u-band photometric errors in the reddest objects. The colors of points encode their redshifts; the sizes of the points vary for clarity. The lower right panel shows the i magnitude as a function of the $g - r$ color.

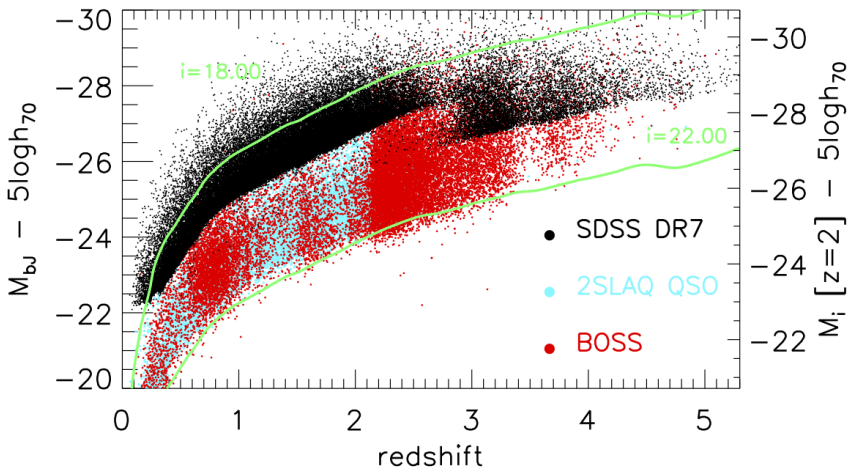


Figure 4.10: The $L - z$ plane for three recent quasar surveys: SDSS-I/II, (black points), 2SLAQ (cyan) and BOSS (red). The luminosity assumes $H_0 = 70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$. There are $\approx 105,000$ objects in the SDSS DR7 catalog and $\approx 9,000$ $g \leq 21.85$ low-redshift quasars from the 2SLAQ Survey (Croom et al. 2009). The three surveys together give a dynamic range in luminosity of ≈ 4 magnitudes at any given redshift up to $z \sim 3.5$. The luminosity corresponding to magnitude limits of $i = 22$ on the faint end and $i = 18$ on the bright end are shown. The coverage here can be compared to Figure (5) in Croton (2009).

interested reader is referred to the discussions in Chapter (3) and Bovy et al. (2011b) for further comparisons.

As an aid for our discussions, in Table 4.1 we list the number of targets from this first year, broken down by the three key selection methods. Again, given the non-uniform selection over this year, this table is provided as an *informational guide* only; it should not be used as a direct comparison between methods.

The redshift distributions for objects with reliable redshifts selected by our three main methods (NN, KDE, and Likelihood) are given in Figure (4.11). Again, because of the non-uniform manner in which these methods were applied during Year One, this plot should not be interpreted as a quantitative comparison between the methods. There is substantial overlap between the methods; many objects are selected by more than one technique. The three histograms have similar shapes over the range $2.2 < z < 3.5$. While NN avoids being confused by $z \sim 1.5$ objects, and KDE avoids objects at $z > 3.5$, all three methods select a substantial number of objects at $z \sim 0.8$.

Figures (4.12), (4.13) and (4.14) show the color-color and the color-magnitude distributions of $z > 2.2$ quasars selected by the Likelihood, NN and KDE methods, respectively. The figures show in orange and black the ratio of numbers of objects selected by each method to

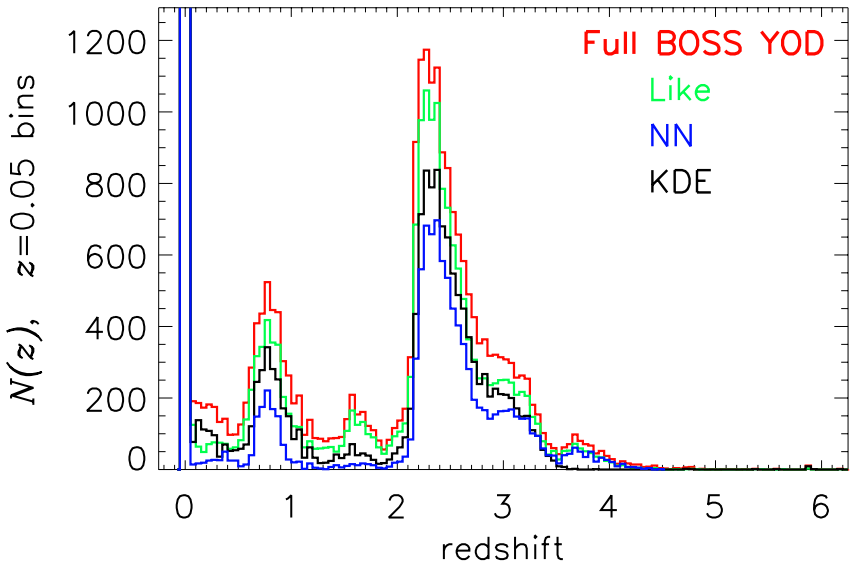


Figure 4.11: The BOSS quasar redshift distribution for objects with reliable redshifts, selected by our three main methods from Year One. The green, blue and black histograms give the redshift distributions for the Likelihood, NN and KDE methods, respectively. The red histogram is the full sample from Figure (4.2). These methods were not applied uniformly through Year One, so this plot is shown for *qualitative and informative* purposes only, and should not be used as a direct comparison between the methods. The KDE, NN and Likelihood algorithms are *not* mutually exclusive, with many objects selected by more than one method.

the total number of Year One quasars, at each point in color space. This ratio is normalized to the global ratio of targets from Column 4 of Table 4.1; thus a point in color space with a value $> 100\%$ is one where the method in question outperforms the total selection on average. The difference between the three methods is clear in the $(u-g)$ vs. $(g-r)$ diagrams. The contours for the Likelihood method are fairly flat away from the stellar locus. NN performs well at $(u-g) \sim 0.6$, $(g-r) \sim 0$ and in those regions of color-color space corresponding to higher-redshift quasars, but does more poorly elsewhere. KDE selects objects only over a very narrow range in $(g-r)$. From the $(g-r)$ vs. i-band color-magnitude diagram (bottom right panels of the figures), we see that the Likelihood method was more efficient at selecting fainter, $i \gtrsim 21.0$ quasars, while the NN tends to select the brighter $i \lesssim 20.0$ objects at all $(g-r)$ colors.

These trends can be understood given the methodology of these algorithms. The Likelihood method down-weights objects close to the stellar locus as the denominator of Equation (2.5) gets large, which is why Likelihood selects few objects there. Otherwise, the Likelihood method traces the overall BOSS Year One sample in color-color and color-magnitude space. The Likelihood method did not place any cuts on photometric redshift, and hence samples the high redshift distribution of the BOSS data well, especially at $(g-r) \gtrsim 1$ (corresponding to redshift $z > 3.5$). See § 3.2.2 for full details of the Likelihood performance.

At the crux of an artificial neural network is the sample of objects used to train it (see Yèche et al. (2010) and references therein, and Section 2.2.4). The training set for the NN we have used was based on the SDSS quasar catalog and the 2SLAQ surveys, and did *not* use data from the MMT pilot survey or the AUS survey. Thus, this training set was geared towards brighter quasars ($i < 20.2$), giving rise to the tendency for NN to select the brighter quasars.

The KDE training set included only $2.2 < z < 3.5$ quasars, and thus the redshift histogram drops to zero at $z = 3.5$ (Figure 4.11). This is related to the fact that KDE quasars inhabit a much narrower range of the $(g-r)$ *vs.* $(r-i)$ color-color plane than the other two methods. In summary, Figures (4.12)-4.14 reflect the relative strengths and trainings of these methods; ultimately, the three methods complemented each other well.

4.1.4 The Blind Test Area

After spectroscopy from the first few chunks had been analyzed, it became clear that the survey would have to decide on a single method for the CORE, and that we would have to restrict ourselves to the nominal target density of 40 targets deg^{-2}. Thus, we designed a test to decide which combinations of methods gave the best yields for the CORE and BONUS selections.

The "Blind Test Area" is a region of sky of ~ 1000 deg^2 in the NGC at high declination

Table 4.2. Surface Density of BOSS Quasars by TS Method

Method	Threshold @20/deg^2	Threshold @40/deg^2	$N_{\rm quasar}$ @20/deg^2	$N_{\rm quasar}$ @40/deg^2	Score @20/deg^2	Score @40/deg^2
KDE	0.904	0.599	9.45	11.35	4.79	5.71
Likelihood	0.543	0.234	8.70	12.23	4.39	5.89
Weighted Like	0.262	0.108	8.89	12.33	4.58	5.98
NN	0.852	0.563	7.62	10.84	4.00	5.51
NN Comb.	0.853	0.573	9.37	12.81	4.69	6.26
Color Box	n/a	n/a	6.45		3.41	

Note. — The surface density of spectroscopically confirmed $2.2 < z < 3.5$ quasars from early (Chunk 1, 2 and 3) BOSS spectroscopic data that would be recovered by various methods, and the thresholds of the key parameters required to yield a surface density of 20 or 40 deg^{-2} in the blind survey region (§ 4.1.4). The Weighted Likelihood incorporated a weighting function which optimizes the S/N of the Lyα forest clustering signal. The redshift and flux distribution of the resulting quasar sample determines this signal, as quantified by the score in the last two columns.

($\delta > +40°$) and high Galactic latitudes, shown by the thin white line in Figure (2.5). This area is used for tuning the threshold of each method to a particular target density. The resulting thresholds were then applied to existing data to determine the selection efficiency.

Table 4.2 summarizes these tests. This table gives the surface density of $2.2 < z < 3.5$ quasars from early (Chunk 1, 2 and 3) BOSS spectroscopic data that would be recovered by various methods at various thresholds of their key parameters when they are tuned to yield a surface density of 20 or 40 deg^{-2} in the blind survey region. The effectiveness of each quasar spectrum for Lyα forest studies depends on its redshift (and thus the spectral coverage of the forest) and its brightness (and thus the S/N of the spectrum). This "value" is quantified by a score of each quasar, motivated by the checks performed in McDonald & Eisenstein (2007); summing this over the expected quasars per square degree gives the numbers in Table 4.2. These scores do not include contributions from quasars outside the redshift range $2.2 < z < 3.5$. "Weighted Likelihood" was an adaption of the Likelihood method to maximize this score, as discussed in detail in § 3.3.3.

We also tried selecting quasars using a simple color region isolating the region where $z \sim 2.7$ quasars are found, akin to the mid-z box used by Richards et al. (2002), but this did not deliver an efficiency close to our requirements.

Although Table 4.2 shows that the KDE method returns the most $z > 2.2$ quasars (9.45 deg^{-2}) at the CORE target density of 20 deg^{-2}, after much deliberation, we selected the Likelihood method as CORE for the latter stages of Year One, since it is a simpler algorithm

to understand and explain, it has a more uniform spatial selection, and is easier to reproduce. Further tests showed that using the Neural Network in its "Combinator" mode for BONUS would yield the highest number of high-z quasars overall. The difference when weighting by the Lyα forest score was too small to motivate us to include it; see the discussion in McQuinn & White (2011).

However, tests of the Year One data with the *XDQSO* method (Bovy et al. 2011b) showed it selected about 1 $z > 2.2$ quasar deg^{-2} more than Likelihood. Thus in Chunks 12 and 13 the union of Likelihood and *XDQSO* was treated as CORE, allowing us to test them directly against one another (Bovy et al. 2011b). In Chunks 12 and 13, 2426 out of 4710 *XDQSO* targets had good spectra and $2.2 < z < 3.5$, for an efficiency of 52%, while Likelihood obtained 2296 quasars from 5086 targets, for a 45% efficiency. This result is our motivation for declaring *XDQSO* to be CORE for the rest of the BOSS quasar survey.

4.2 The Completeness of CORE in Year One

Studies of clustering in the Lyα forest are not biased by the distribution of background quasars used to illuminate Lyα forest absorption. Thus the Year One BOSS quasar sample can be used for these studies. Indeed, Slosar et al. (2011) have performed a first clustering analysis of Lyα forest flux from the BOSS Year One data.

However, given the changes in QTS throughout the first year, the quasar sample described in this chapter is far from sufficiently uniform to be used directly for studies of the statistics of the quasars themselves, such as measurements of their luminosity function or clustering. The goals of the CORE sample is to have such a uniformly-selected sample of quasars, but as the definition of CORE changed several times during commissioning, CORE objects in the first year do not represent a statistical sample.

The project settled on the *XDQSO* algorithm (§ 3.4; Bovy et al. (2011b)) for the CORE method at the end of Year One, and will use it for the rest of the survey. It is therefore useful to apply this algorithm to the photometry used in the Year One spectroscopy, and determine the completeness of the Year One targeted chunks. Table 4.3 and Figure (4.16) give the results of this test. Given the placement and overlap of the spectroscopic plates, each chunk can be uniquely divided into sectors covered by a unique combination of plates. The completeness of the targeting: i.e., the fraction of the *XDQSO* CORE sources that were actually targeted in Year One, is measured for each sector separately. Encouragingly, these targeting completeness values are generally 80% or higher, which indicates that statistical analyses of the final CORE sample should be able to incorporate Year One data by introducing moderate weighting factors. The lower targeting completeness (65%) on Chunk 11 highlights a subtle point: the completeness for CORE-selected *quasars* should be higher

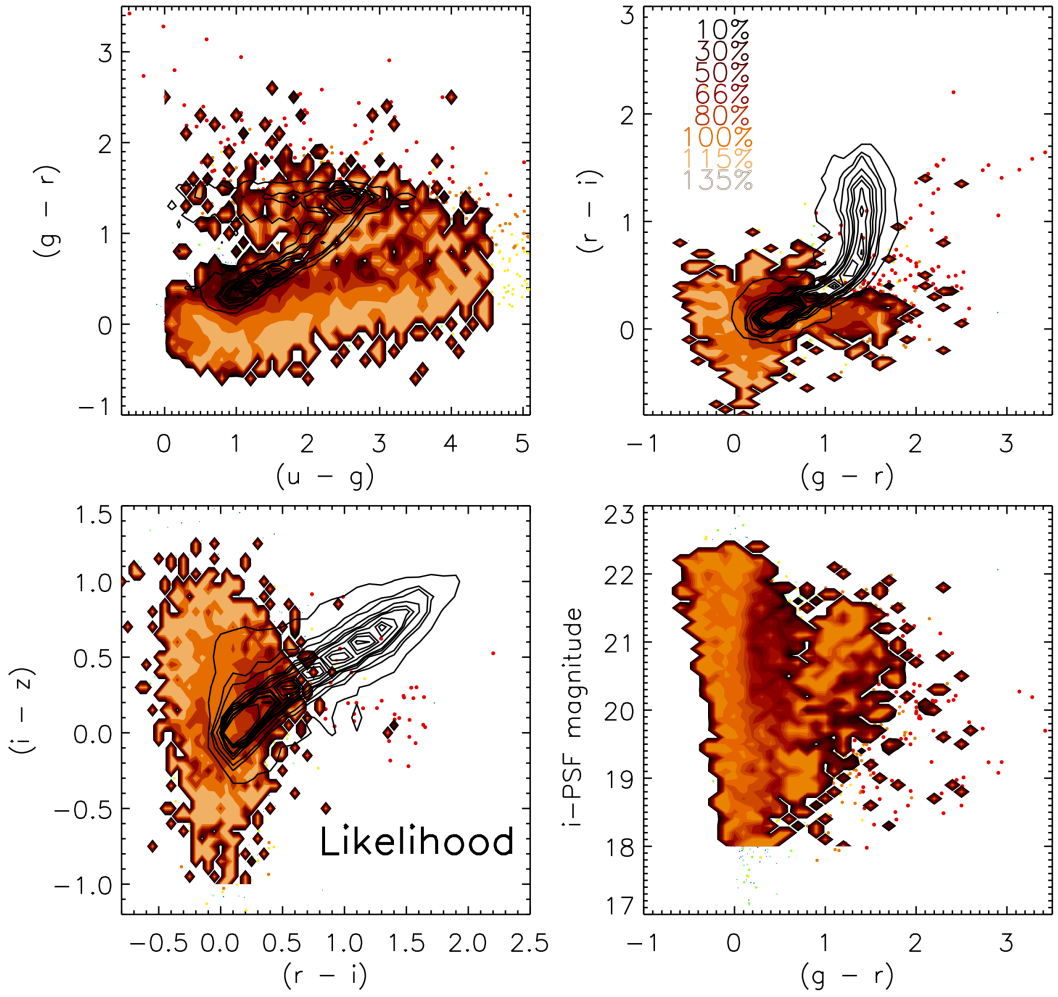


Figure 4.12: Distributions in color-color and color-magnitude space for $z > 2.2$ quasars selected by the Likelihood method in Year One. The black contours give the location of the stellar locus, while the orange contours give the ratio, at each point of color space, of $z > 2.2$ quasars selected by Likelihood to all Year One BOSS quasars, normalized to the global ratio of the two. Quasar numbers were smoothed with a tophat of width 0.10 mag in $u - g$ and $g - r$, and 0.05 mag in $r - i$ and $i - z$, before taking ratios.

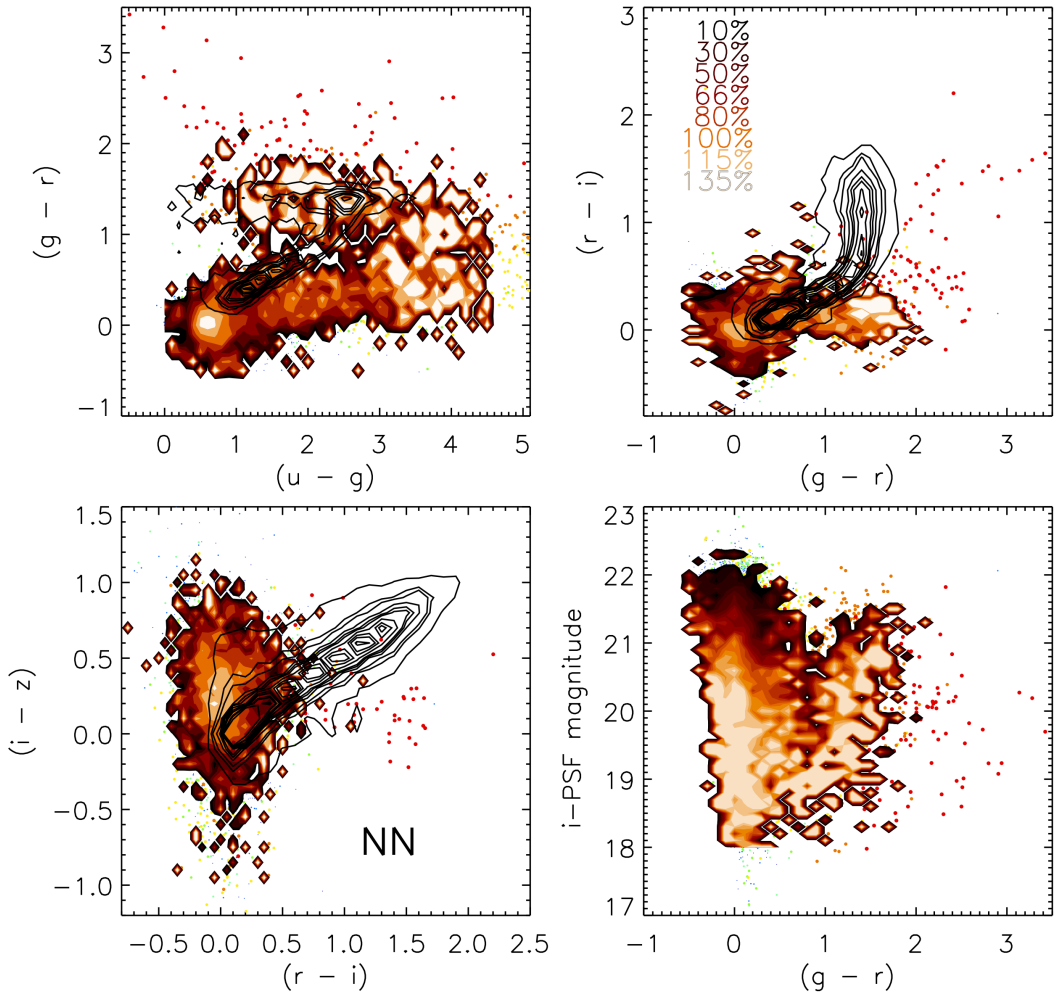


Figure 4.13: As in Figure (4.12), for the NN method.

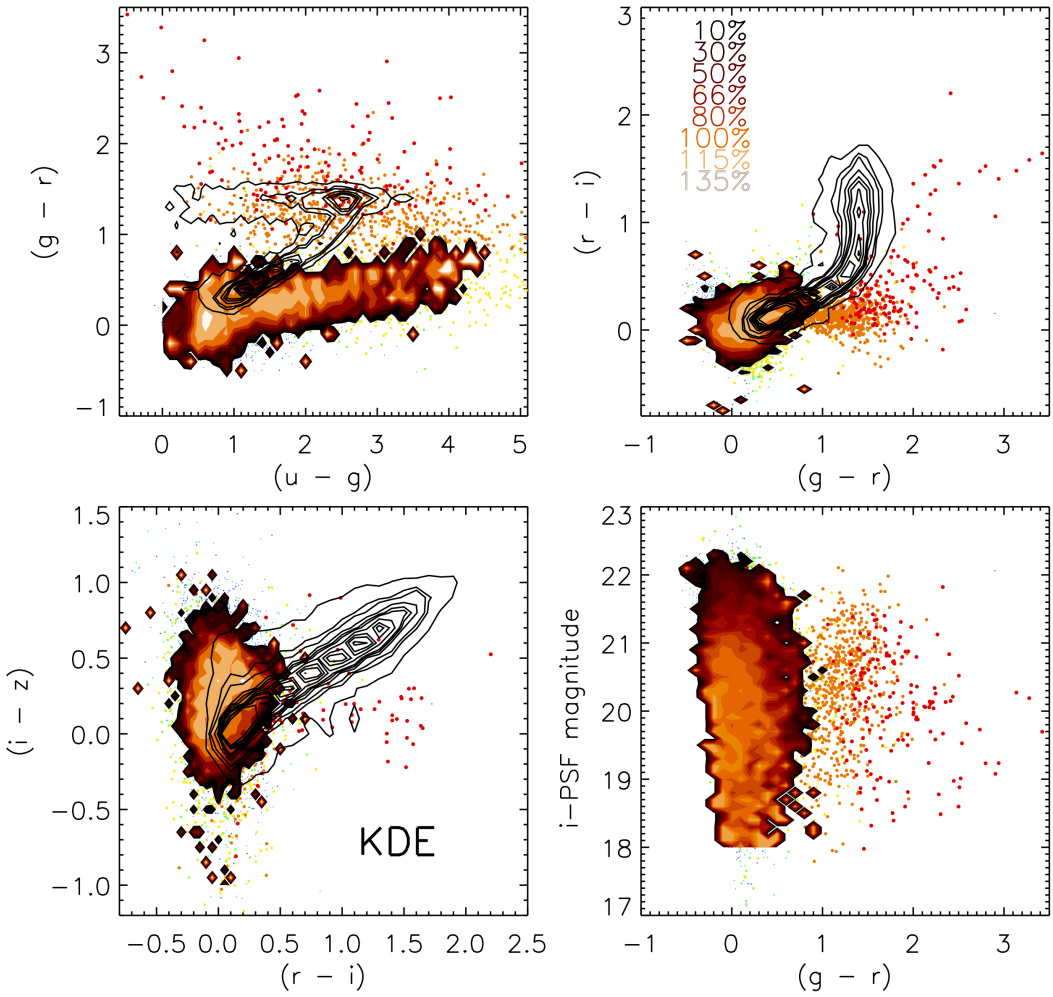


Figure 4.14: As in Figure (4.12), for the KDE method.

Table 4.3. BOSS Quasar Targeting Completeness

Chunk	Area (deg^2) $C \geq 0.75$	Effective Area (deg^2) $C \geq 0.75$	Mean C
11	70.6	58.2	0.654
2	130.1	120.4	0.905
3	85.9	79.4	0.830
4	246.1	230.4	0.861
5	243.0	232.0	0.952
6	182.6	171.2	0.933
7	205.0	185.8	0.836
8	75.5	65.7	0.814
9	84.1	71.6	0.822
10	71.7	60.7	0.813

Note. — C of objects that would have been targeted by the *a posteriori XDQSO* CORE algorithm, which were actually targeted, for each Year One chunk. Chunk 11 has greater area coverage than Chunk 1, thus we list it instead. The second column gives the solid angle (in deg^2) of the region of each chunk in which the completeness is greater than 0.75, the third column lists the same value but for *effective area* (i.e area $\times$ completeness) and the fourth column tabulates the mean completeness over the chunk. See also Figure (4.16).

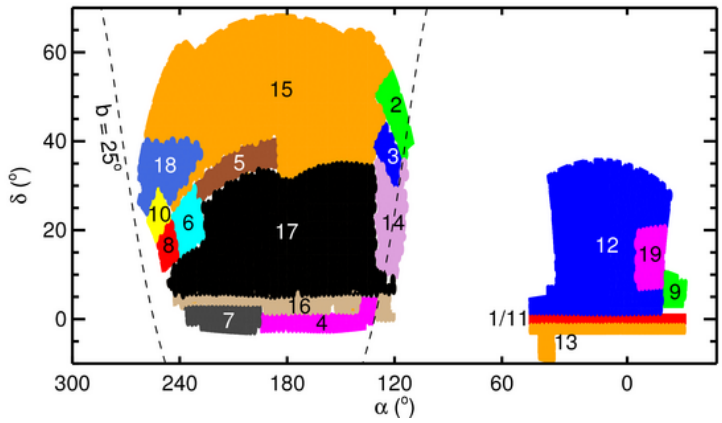


Figure 4.15: The targeting footprint for the SDSS-III:BOSS Lyα forest/Quasar Survey. The various chunks are indicated by different colors. Chunks 16, 17 and 18 lie within the footprint of Chunk 15. The full targeting footprint is 10,200 deg^2, with a total of $\approx$430,000 tiled targets. 180,000 of these targets had spectra by the end of Year Two observations. The global Year One quasar target density is 60.4 targets deg^{-2}, and the mean target density over all chunks shown is 47.9 targets deg^{-2}. The dashed line is at Galactic latitude $b = 25°$.

than the completeness for CORE targets as a whole, because the true quasars are the most likely to also be selected by one of our other algorithms. In the case of Chunk 11, the deeper Stripe-82 photometry eliminates many noisy stellar contaminants in the single-epoch *XDQSO* target list, but it probably selects nearly all of the true quasars selected by CORE.

For calculations of the quasar luminosity function, one must also account for the incompleteness of the *XDQSO* CORE sample relative to the full population of quasars. This can be quantified, for example, using the extensive targeting on Stripe-82 (Palanque-Delabrouille et al. 2011). Similarly, to determine completeness as a function of position on the sky for quasar clustering work it is necessary to determine the fraction of quasars hiding among the unclassifiable spectra. Ongoing visual inspections of these spectra will address this question to some extent.

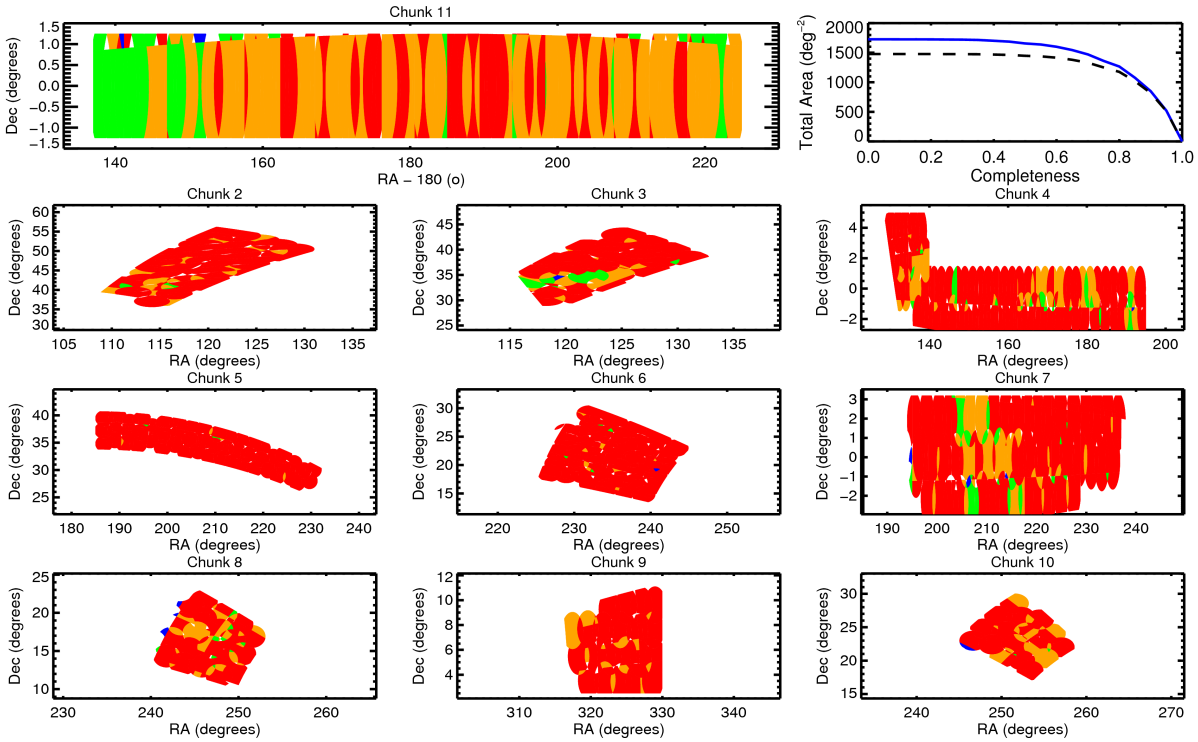


Figure 4.16: The fraction of the objects would be targeted using the final version of the *XDQSO* CORE quasar target selection, that were actually targeted in Year One. Each panel shows the area covered by a Chunk (2-11) from Year One. We use Chunk 11 on Stripe-82 in place of Chunk 1 (top-left panel) as Chunk 11 has superior areal coverage. Note that in some chunks, the scales on the RA and Dec axes are quite different. Color coding shows the spectroscopic completeness of the *a posteriori XDQSO* CORE sample for each area. Those areas in red have a targeting completeness above 0.75, orange have a completeness of 0.5–0.75, green have a completeness of 0.25–0.5 and the few areas in blue have a completeness below 0.25. The top right panel shows the cumulative area (blue solid line) and effective area (area × completeness; black dashed line) above a given level of targeting completeness for the *XDQSO* CORE sample.

Chapter 5

Quasar Luminosity Dependence

If one believes that quasars are the result of major galaxy mergers, than one can assume a relationship between the quasar's luminosity and the mass of the central black hole (m_{bh}) (Kauffmann & Haehnelt 2000; Wyithe & Loeb 2003; Volonteri et al. 2003). The fraction of gas accreted onto the black hole during each merger is chosen to match the observed relation between the velocity dispersion of the bulge and m_{bh} (Ferrarese & Merritt 2000). This results in a correlation between the quasar luminosity and the mass of the host dark-matter halo. Since the clustering properties of dark-matter halos strongly depend on their mass, the quasar clustering amplitude is thus expected to depend on luminosity.

However, it is only recently that samples of quasars have grown big enough (in terms of the number of objects) to study their clustering with some precision (Porciani et al. 2004; Croom et al. 2005; Porciani & Norberg 2006; Hennawi et al. 2006a; Myers et al. 2007a,b; Shen et al. 2007; da Ângela et al. 2008). One of the major problems with measuring the clustering of quasars is that they are extremely rare ($\bar{n} \sim 10^{-6}\,h^3\,\mathrm{Mpc}^{-3}$ at $z \sim 0.5$). Shot-noise from Poisson fluctuations in the counts of objects thus obscures their clustering signal. At low redshifts, this problem is exacerbated, requiring measurements in very broad redshift intervals.

To increase the signal to noise of the clustering signal, I use a cross-correlation technique similar to that discussed in Padmanabhan et al. (2009). I measure the clustering of approximately 3500, $0.5 < z < 1.0$ quasars from SDSS, 2SLAQ and BOSS quasar surveys with a sample of 5.23 million photometric galaxies brighter than $i-\mathrm{band} = 23.5$ from the CFHT (Canada-France-Hawaii Telescope) Survey of Stripe-82 (CS82). The galaxies have reliable photometric redshifts, trace the matter distribution in a way that is well understood, and have a much higher volume density ($\bar{n} \sim 10^{-1}\,h^3\,\mathrm{Mpc}^{-3}$) than the quasar sample. The cross-correlation can thus be well measured and inverted, using the known redshift distribution, to the underlying 3D clustering.

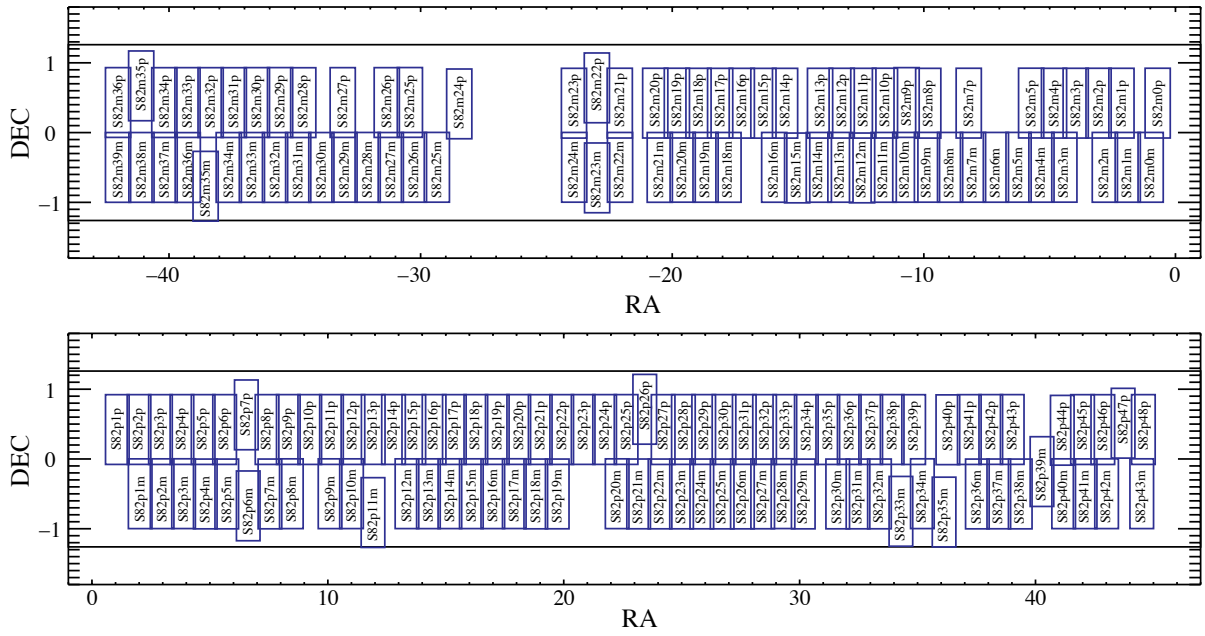


Figure 5.1: Layout of the CS82 fields (166 images).

I have measured the luminosity-dependence of quasar clustering which span ∼4 magnitudes in luminosity, at a given redshift. I find a strong ($4.56\,\sigma$) detection of luminosity dependent clustering in the cross-correlation. This is the first time that this signal has been detected. This constrains several parameters in the halo occupation distribution (HOD) model and their scaling with luminosity or black-hole mass. In order to interpret my observations, which lie largely in the non-linear regime, I use the halo catalog from an N-body simulation of the ΛCDM family.

The layout of this chapter is the following: In § 5.1, I discuss the various datasets I am using and the different breakdowns of the QSO population. In § 5.2, I discuss the statistical methods used in this chapter for calculating the clustering properties. In § 5.3, I present my results and various tests done on the stability of these results. In § 5.4, I compare my results to previous work. My conclusions are in § 5.5. Throughout this chapter I quote distances in comoving coordinates, magnitudes in the AB system, use log to indicate base-10 logarithms and ln to indicate natural logarithms.

5.1 Data

In this section, I give the relevant details for the data that I used to perform my cross-correlation measurements. The data is from various surveys, but it is all from observations made in the field known as "Stripe 82". Stripe-82 is a $\sim$250 deg^2 area stripe centered on the celestial equator in the Southern Galactic Cap, and was imaged multiple times during the SDSS-I/II (Abazajian et al. 2009). SDSS-III also performed spectroscopic observations on Stripe-82, which resulted in the discovery of over 9400 confirmed quasars (see Chapter 4.1). I describe the quasar spectroscopic sample below in § 5.1.2. First, I discuss my galaxy sample, which is derived from the "CS82" survey (Erben et al. 2013).

5.1.1 Galaxies: CFHT Stripe-82 Survey

The CFHT-Stripe-82, "CS82", is a completed imaging survey that has taken medium-deep, $i < 23.7$ data across 170 deg^2 of Stripe-82, using the MegaCam imager on the 4.2 meter Canada-France Hawaii Telescope on Mauna Kea (Hildlebrandt et al. 2012). The area of one image is 1 deg^2, and the total number of images is 166, with some minimal overlap between exposures for calibration. Bright stars along the stripe are avoided. Figure (5.1) shows the layout of the CS82 fields. The pixel size is 0.187″, and the imaging PSF varies from $\sim$ 0.4″ to $\sim$ 0.8″, with the median being around 0.6″. Images are 4 dithers with a total exposure time of 1640 seconds.

Catalog generation is performed by running Source Extractor (Bertin & Arnouts 1996). The basic catalogs, before any quality controls, contain about 11 million objects. Image masks are hand generated by the CS82 team, and cuts are made around bright stars, bad pixels, and regions which have imaging artifacts. After removing objects in masks, the catalog contains $\sim$6 million objects, and has a total area of 124 deg^2. The mean seeing is 0.6, shown in Figure (5.2) along with the magnitude and photometric redshift distributions of the galaxies.

5.1.2 QSO Datasets

My quasar dataset comes from three spectroscopic surveys; the Sloan Digital Sky Survey (SDSS; York et al. 2000), the 2dF-SDSS LRG and QSO Survey (2SLAQ; Croom et al. 2009), and the SDSS-III: Baryon Acoustic Oscillation Spectroscopic Survey (BOSS; Eisenstein et al. 2011). I give the necessary details of these three surveys below, and a summary of my QSO dataset is in Table (5.1). The footprint in RA-Dec for the QSOs is shown in Figure (5.3), the magnitude and photometric redshift distributions and the coverage in the luminosity-redshift, $L - z$, plane are shown in Figure (5.4).

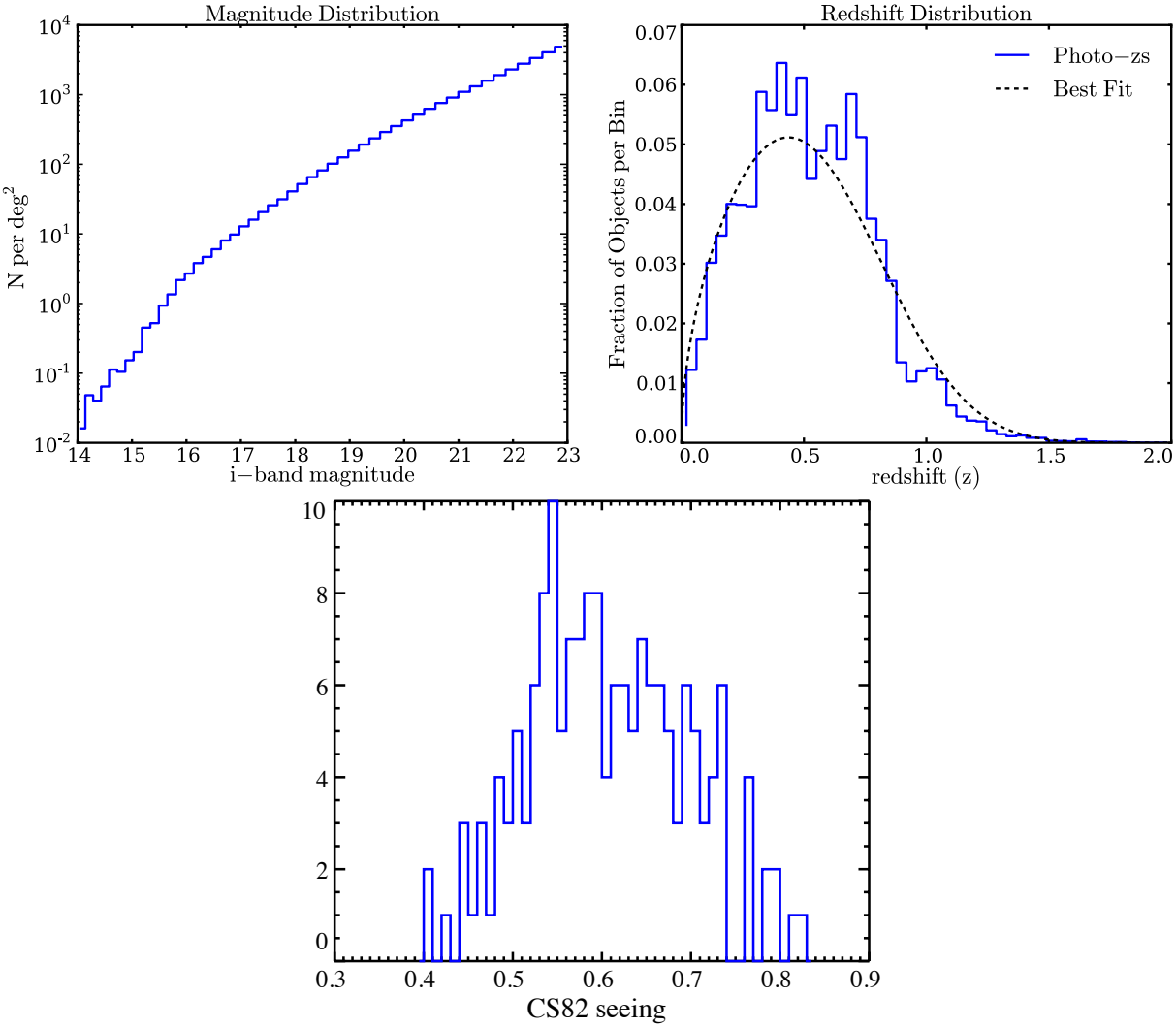


Figure 5.2: (*Top Left*) Histogram of the $i-$band magnitude distribution for CS82 data. (*Top Right*) The photometric redshift (photo-z) distribution (solid) for the CS82 data as well as the best fit curve to the distribution (dashed). (*Bottom*) The mean seeing for CFHT.

SDSS QSOs

The SDSS used a dedicated 2.5 meter wide-field telescope (Gunn et al. 2006) to collect light for 30 2k×2k CCDs (Gunn et al. 1998) over five broad bands - *ugriz* (Fukugita et al. 1996) - in order to image $\sim \pi$ steradians of the sky. The imaging data are taken on dark photometric nights of good seeing (Hogg et al. 2001), are calibrated photometrically (Smith et al. 2002; Ivezić et al. 2004; Tucker et al. 2006; Padmanabhan et al. 2008) and astrometrically (Pier et al. 2003), and then object parameters are measured (Lupton et al. 2001; Stoughton et al. 2002).

Using the imaging data, quasar target candidates are selected for spectroscopic observation based on their colors, magnitudes and detection in the FIRST radio survey (Becker et al. 1995), as described by Richards et al. (2002). Low-redshift, $z \lesssim 3$, quasar targets are selected based on their location in *ugri*-color space and the high-redshift, $z \gtrsim 3$, objects in *griz*-color space. Quasar candidates passing the *ugri*-color selection are selected to a flux limit of $i = 19.1$, but since high-redshift quasars are rare, objects lying in regions of color-space corresponding to quasars at $z > 3$ are targeted to $i = 20.2$. Furthermore, if an unresolved, $i \leq 19.1$ SDSS object is matched to within $2''$ of a source in the FIRST catalog, it is included in the quasar selection. A tiling algorithm then assigns these candidates to specific spectroscopic plates, in order to maximize target completeness (Blanton et al. 2003). No two fibers can be placed closer than $55''$, corresponding to $\sim 0.3\ \ h^{-1}$ Mpc at $\langle z \rangle = 0.76$, the mean redshift of my sample.

For my analysis, I use the SDSS Data Release Seven (DR7; Abazajian et al. 2009) and select quasars from the final version of the SDSS quasar catalog (DR7Q; Schneider et al. 2010). This catalog consists of spectroscopically identified quasars that have luminosities larger than $M_i = -22.0$ (measured in the rest frame) and at least one emission line with FWHM larger than 1000 km s^{-1}. Every object in the DR7Q had its spectrum manually inspected. There are 105,783 confirmed quasars over the 9,380 deg^2 spectroscopic DR7 print; the 19,137 DR7Q quasars with redshifts $0.5 < z < 1.0$ will be the SDSS sample I use in this investigation. Of these 19,137 low-redshift quasars, 1,725 are on Stripe-82.

2SLAQ QSOs

Croom et al. (2009) gives full details of the 2SLAQ quasar survey. 2SLAQ was a deep, $18 < g < 21.85$ (extinction corrected), sample aimed at probing in detail the faint end of the broad line active galactic nuclei luminosity distribution at $z < 2.6$. The candidate QSOs were selected from SDSS photometry and observed spectroscopically with the 2dF spectrograph on the Anglo-Australian Telescope. This sample covers an area of 191.9 deg^2 and contains spectra of 8764 QSOs, of which 7623 were discovered by 2SLAQ. Although the 2SLAQ targeted the two equatorial regions from the SDSS imaging data in both the North

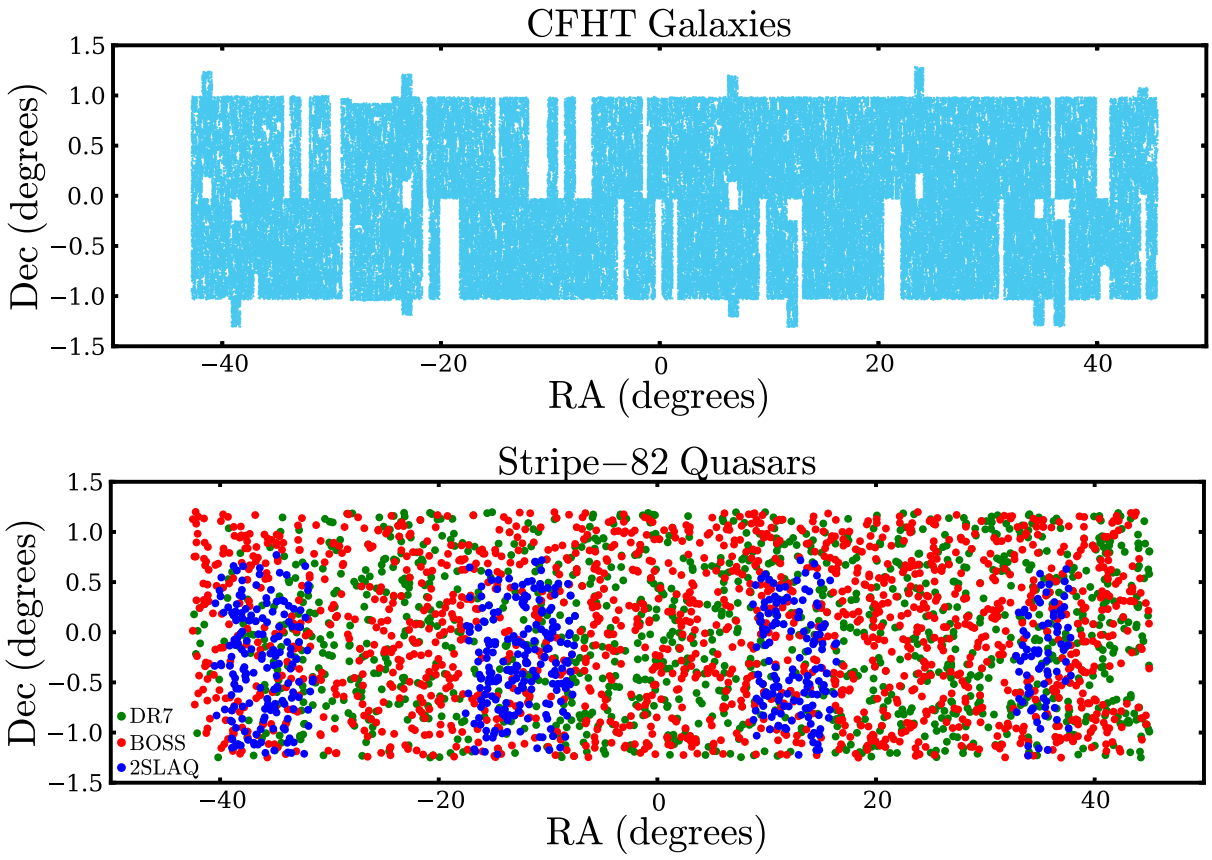


Figure 5.3: Top: Alternative view of Figure (5.1). RA versus Dec for the CS82 galaxy data. Bottom: RA versus Dec for the quasar data. The quasar dataset comes from three spectroscopic surveys: SDSS DR7 (green), 2SLAQ (blue), and BOSS (red).

Galactic Cap and the South Galactic Cap, due to various factors (e.g. telescope scheduling, weather), only ~1/3 of the 2SLAQ coverage, and hence catalog, was on Stripe-82. Imposing the redshift range $0.5 < z < 1.0$ results in 596 objects.

SDSS-III BOSS QSOs

The SDSS-III BOSS uses the same imaging data as that of the original SDSS-I/II survey, but BOSS specifically targets objects thought to be at high redshifts, $z > 2.2$ (for LyαF cosmology), and spectra were taken with the new BOSS spectrographs on the 2.5 meter Sloan telescope (Eisenstein et al. 2011).

Chapter (2) and Ross et al. (2012) have full details on the quasar target selection methods that were used for BOSS. In short, using the imaging data, BOSS quasar target candidates

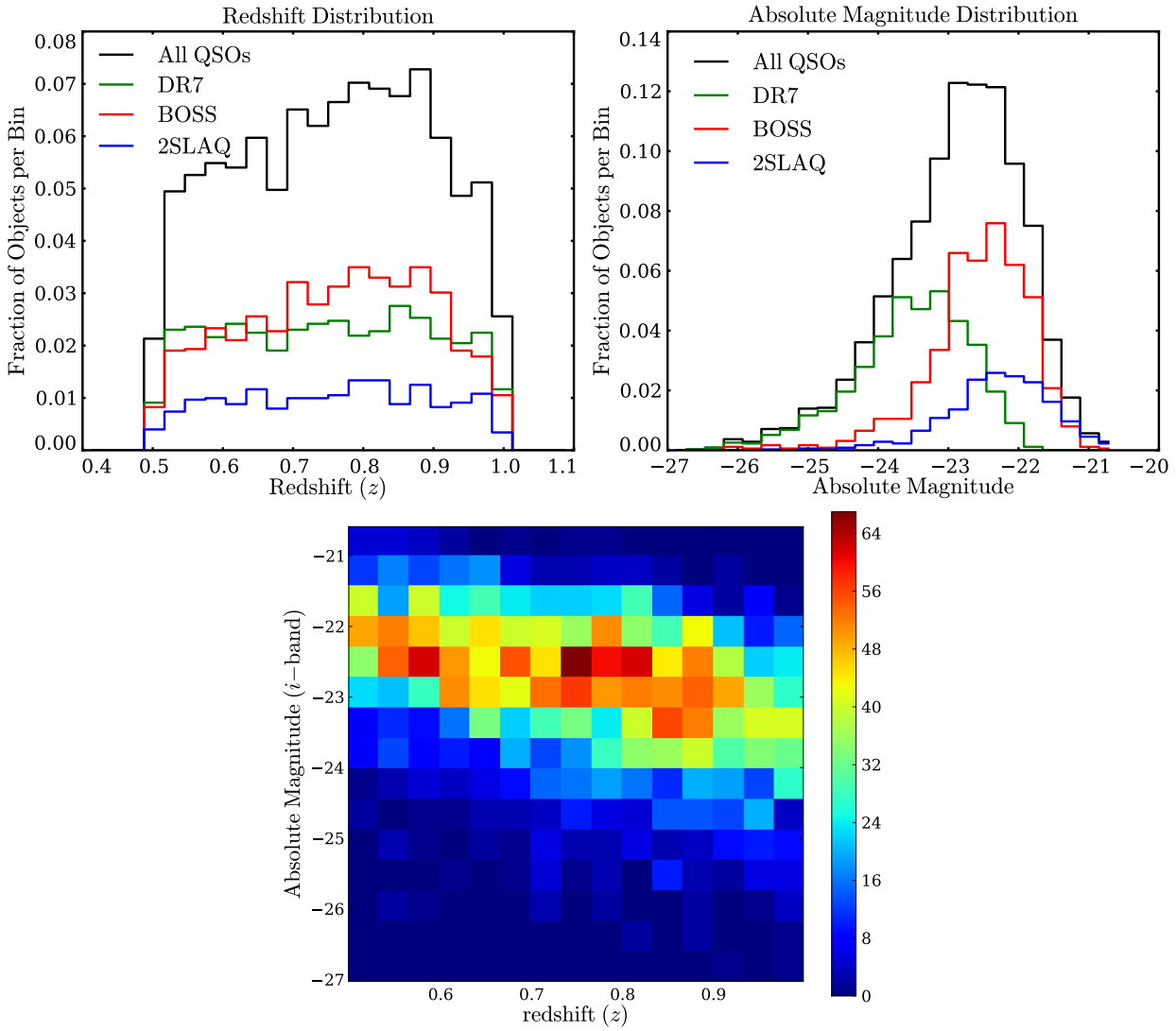


Figure 5.4: (*Top Left*) Redshift distribution of quasar data. The quasar dataset comes from three spectroscopic surveys: SDSS DR7, 2SLAQ, and BOSS. (*Top Right*) Absolute magnitude distribution of quasar data. (*Bottom*) Distribution in the $M - z$ plane.

Table 5.1. Properties of Quasar Sample

Survey	# QSOs	z range	$\bar{z}$	magnitude (M) range	$\overline{\mathrm{M}}$
SDSS DR7	1717	$0.50 < z < 1.0$	0.75	$-27.0 < M < -18.6$	-22.9
BOSS	622	$0.50 < z < 1.0$	0.79	$-25.5 < M < -19.6$	-22.5
2SLAQ	514	$0.50 < z < 1.0$	0.76	$-24.5 < M < -18.6$	-22.2

Note. — Properties of the quasar sample broken down by data source.

are selected for spectroscopic observation based on their fluxes and colors in SDSS bands. All objects classified as point like and having a magnitude limit of $g \leq 22$ or $r < 21.85$ (PSF magnitudes) are passed to the quasar target selection code.

Although the aim of BOSS is to target $z > 2.2$ quasars, degeneracies in the color-redshift relation of quasars lead to the unintentional selection of low-z quasars in BOSS. Thus, BOSS picks up a population of quasars at $z \sim 0.8$ which have Mg II λ2800 Å masquerading as Lyα emission at redshift $z \sim 3.1$, giving these objects similar broad-band colors. However, since I am employing a cross-correlation technique, I am at liberty to select, and use in my analysis, these heterogeneously selected low-z BOSS quasars. Details of the BOSS Quasar dataset are given in Table (5.1).

5.1.3 Quasar k-correction

When calculating Absolute Magnitudes, I assume a $(0.70, 0.30) = (\Omega_\Lambda, \Omega_{\mathrm{matter}})$ cosmology and I utilize the k-corrections from Richards et al. (2006), but normalizing to $z = 0.0$. Richards et al. (2006) generate their k-corrections in two parts, first by assuming a general power-law continuum (K_{cont}) and second, by modeling a set of quasar emission lines (K_{em}). Thus the contribution to the k-correction due to the continuum at a given redshift z is $K_{\mathrm{cont}} = -2.5(1+\alpha_\nu)\log(1+z)$, where α_ν is the power-law slope and is set to be -0.5 (e.g., Vanden Berk et al. 2001). Continuing to follow Richards et al. (2006) then gives:

$$M_i(z=0) = M_i(z=2) + 2.5(1+\alpha_\nu) + log(1+z) \tag{5.1}$$

for my quasar sample to $z = 0.0$.

5.2 Methods for Clustering Calculation

I measure the clustering in configuration space, rather than Fourier space. For rare objects, where shot-noise is an important or dominant piece of the error budget, configuration space

estimators have the advantage of more nearly independent errors. I compute the angular correlation function over a broad redshift range but replace the angular separation by the transverse separation R computed assuming that the galaxies are the same redshift as the quasar with which they are being correlated. I use the estimator:

$$w_\theta(R) = \frac{QG(R)}{QR(R)} - 1 \ , \tag{5.2}$$

where QG and QR are the quasar-galaxy and quasar-random pairs, and my notation makes explicit that I am binning in physical transverse separation, R, and not angle, θ. The random points reflect the imaging mask of the CS82 galaxies (§ 5.1.1) and I use 5× more random points than galaxies. Assuming that $w_p(R)$ does not evolve over the redshift bin,

$$w_\theta(R) = \langle f(\chi) \rangle w_p(R) \ , \tag{5.3}$$

where the average is done over the quasar redshift distribution and $f(\chi)$ is the (normalized) radial distribution of the galaxies. For a survey of solid angle Ω,

$$dN = \bar{n}(\chi) dV = \Omega \, \bar{n}(\chi) \, \chi^2 d\chi = \Omega \left(\frac{c\chi^2}{H(\chi)} \right) \bar{n} \ dz \tag{5.4}$$

or if I write $\mathcal{N} = N/\Omega$,

$$\frac{1}{N}\frac{dN}{dz} = \frac{1}{\mathcal{N}}\frac{d\mathcal{N}}{dz} = \frac{c\chi^2}{H}\bar{n} \Big/ \int \frac{c\chi^2}{H}\bar{n} \, dz = \frac{c\chi^2}{H} \ \frac{\bar{n}}{\mathcal{N}} \tag{5.5}$$

So that

$$f(\chi) = \frac{\bar{n} \ \chi^2}{\mathcal{N}} = \frac{1}{\mathcal{N}}\frac{d\mathcal{N}}{dz} \ \frac{dz}{d\chi} = \frac{H}{c} \ \frac{1}{\mathcal{N}}\frac{d\mathcal{N}}{dz} = \frac{H}{c} \ \frac{1}{N}\frac{dN}{dz} \tag{5.6}$$

Over a wide redshift interval I measure an "effective" cross-correlation, ξ_{12}, between two tracers which is

$$\xi_{12}(r, z_{\rm eff}) \simeq \langle \xi_{12}(r) \rangle = \frac{\int dz \ (dN/dz)_1 (dN/dz)_2 (H/c\chi^2) \ \xi_{12}(r,z)}{\int dz \ (dN/dz)_1 (dN/dz)_2 (H/c\chi^2)} \tag{5.7}$$

where $z_{\rm eff}$ is defined with a similar average to ξ. The single power of H/χ^2 arises because the correlation function is defined in terms of a number times a number density. I can now combine these equations to find

$$\langle \xi_{12}(r) \rangle = \frac{\int dz \ (dN/dz)_1 [f_2(\chi)/\chi^2] \ \xi_{12}(r,z)}{\int dz \ (dN/dz)_1 [f_2(\chi)/\chi^2]} = \int dz \ W_{12}(z) \xi_{12}(r,z) \tag{5.8}$$

Figure (5.5) shows $f(\chi)$ for the CS82 galaxies at several different magnitude cuts.

I estimate the errors on my measurements by bootstrap resampling (e.g. Efron & Gong 1983). Specifically, I subdivide the survey into approximately equal area regions $2^\circ \times 2^\circ$

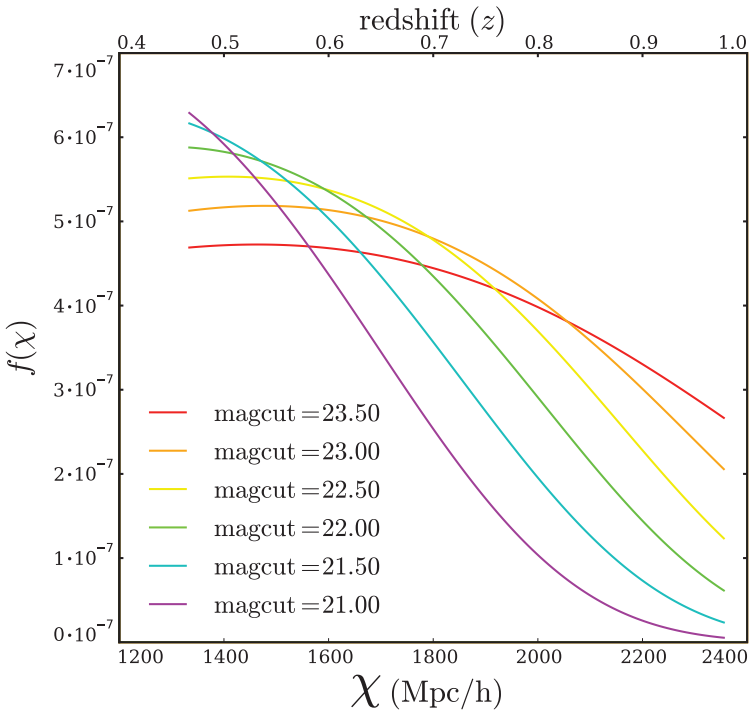


Figure 5.5: $f(\chi)$ for the CS82 galaxies at several different magnitude cuts.

Table 5.2. QSO-Galaxy Cross-Correlation Fit

R Range (Mpc/h)	$\langle f \rangle$	r_0	γ	W
[0.3,3]	$4.245 \cdot 10^{-4}$	5.05	1.77	66.02

Note. — Luminosity dependent quasar clustering using a cross-correlation technique between CS82 galaxies ($M < 23.5$) and SDSS, BOSS, and 2SLAQ quasars ($0.5 < z < 1.0$). The cross-correlation function was calculated in the R-range of [0.3-3.0] Mpc/h. A power-law fit to the data using Equation (5.9) found that $r_0 = 5.05$, and $\gamma = 1.77$. The collapsed single number from Equation (5.10) is $W = 66.02$.

spaced equally in right ascension. Both the average and error on any quantity is estimated by bootstrap resampling the pair counts based upon a random draw of regions (with replacement). Individual radial bins appear to be highly covariant, indicating that both sample variance and shot-noise are contributing significantly to the error budget. I find that individual bootstrap realizations for different radial bins and for different quasar samples are positively correlated. The diagonal component of the correlation matrix increases rapidly beyond a few Mpc (see Figure 5.6), because the geometry of the survey is narrow in the declination direction (the transverse dimension subtends approximately $50\,h^{-1}$Mpc at $z = 0.5$), and the bins become increasingly correlated as shot-noise becomes a smaller contribution than sample variance.

I find that the cross-correlation functions are well fit by power laws over the range where my constraints are tightest, as expected if quasars are hosted by massive halos[1]. If the 3D clustering signal is a power-law, $\xi(r) = (r_0/r)^\gamma$, then the projected clustering

$$\frac{w_p(R)}{R} = \frac{\sqrt{\pi}\,\Gamma[(\gamma-1)/2]}{\Gamma[\gamma/2]} \left(\frac{r_0}{R}\right)^\gamma . \tag{5.9}$$

Figure (5.7) shows the cross-correlation function $w_p(R)/R$ for the full quasar and galaxy samples. This figure also shows the best fit power-law.

For $\gamma = 2$ the prefactor is π, however my results are better fit by a slope $\gamma = 1.77$ for which the prefactor is approximately 3.79. (Note the non-trivial sensitivity of the amplitude of w_p near $R = r_0$ to the assumed slope.) Traditionally a power-law form is fit to the measured signal, and I give fits for r_0 holding $\gamma = 1.77$ fixed in Table (5.2). However in what follows I have chosen a different way of summarizing the data for two reasons. First, the interpretation in terms of r_0 is more complex for cross-correlations than for auto-correlations. Secondly

[1]The non-linear, scale-dependent bias of dark matter halos makes their correlation functions close to a power-law on Mpc scales.

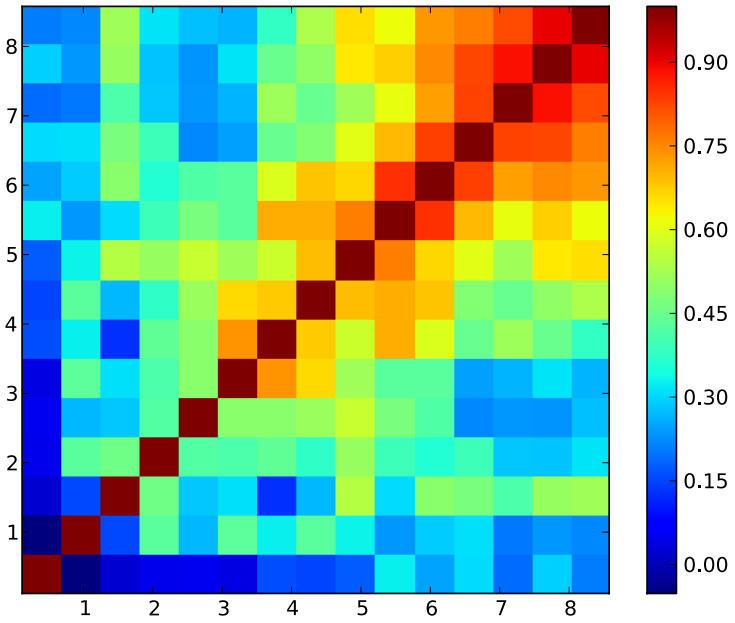


Figure 5.6: Covariance matrix for the correlation function.

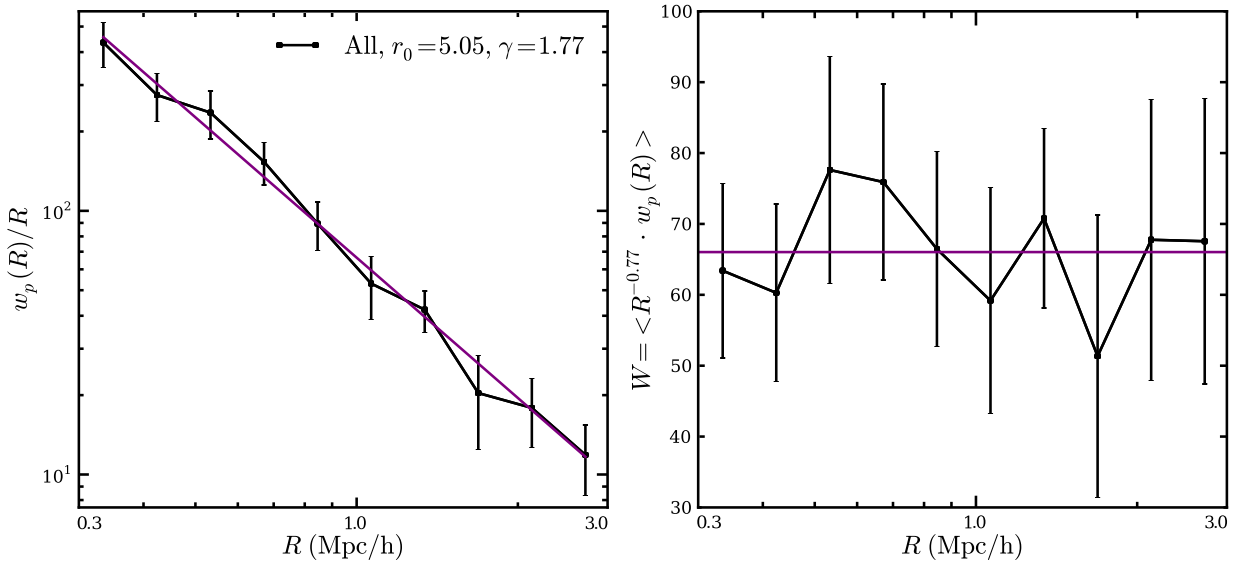


Figure 5.7: (*Left*) The cross-correlation function $w_p(R)/R$ for the full quasar and full galaxy samples. The purple line is the best-fit power-law using Equation (5.9) with values $\gamma = 1.77$ and $r_0 = 5.05$. (*Right*) The collapsed correlation function W shown in Equation (5.10) with $\mu = -0.77$.

a power-law fit requires us to estimate the covariance matrix and I have found that the covariance matrix returned by the bootstrap can be quite noisy (see Figure 5.6). Given how covariant the radial bins are, and how well the individual samples resemble a power-law, I choose to "collapse" the measurement down into a single number per quasar sample:

$$W \equiv \frac{1}{N_{\rm pnt}} \sum_i \left(\frac{R_i}{1\,h^{-1}{\rm Mpc}} \right)^{\mu} w_p(R_i) \quad . \tag{5.10}$$

where $N_{\rm pnt}$ is the number of bins in the sum. As shown in Figure (5.7), μ is chosen such that multiplying by R^{μ} flattens w_p so that I am doing an unweighted average of the amplitude at each scale. I can then determine the joint distribution of the set $\{W\}$ (one for each quasar sample) using the bootstrap technique. As the dimensionality of the data has been reduced the results are much more stable. While this method does not optimally use the data[2], it allows us to make statistically meaningful statements. The increasing variance and correlation of the large-R data points means that an unweighted average, such as this, becomes increasingly noisy as larger and larger scales are included. For this reason I restrict the sum to $0.3 < R < 3\,h^{-1}$Mpc, (see § 5.3.1 for more details).

5.3 Results

To measure the luminosity dependence of the QSO-Galaxy cross-correlation function, I split my quasar sample into the 1/3 brightest and 2/3 dimmest quasars in terms of absolute magnitude. The bright and dim samples have different redshift distributions. In order to separate effects in clustering due to the different redshift distributions and due to luminosity differences, I weight the correlation function such that the redshift distributions of the two sets are flat. This allows me to directly compare the two samples without contamination from the redshift differences (Figure 5.8). More discussion of the effects of this weighting are discussed in § 5.3.1 below. Table (5.3) gives the raw values for the weighted, cumulative, magnitude distributions for the bright and dim quasar samples also shown in Figure (5.8).

I calculated the cross-correlation functions as described in § 5.2. For $\langle f(\chi) \rangle$ I used the best fit function to the galaxy redshift distribution (as shown in Figure 5.2). More discussion of the effects of using different fits are discussed in § 5.3.1. When calculating the parameter W from Equation (5.10) I use the entire quasar data set (not just bright or dim) to fit the value for μ. Figure (5.9) shows $\omega_p(R)/R$ and W versus R for the bright and dim samples. The error bars are calculated from bootstrap resampling of the data. Figure (5.10) shows the distribution of the W for 10,000 bootstrap realizations of the data for the bright and dim samples. For 97% of the realizations $W_{dim} < W_{bright}$ which corresponds to a $1.9\,\sigma$ detection of luminosity dependent quasar clustering. Table (5.4) summarizes my findings. These numbers can be improved with various cuts on the data sets (see below).

[2]This average would be optimal for shot-noise limited measurements with $\xi \propto r^{-2}$.

Table 5.3. Magnitude Distribution of Bright and Dim QSO Samples

Absolute Magnitude	Dim Sample	Bright Sample	Absolute Magnitude	Dim Sample	Bright Sample	Absolute Magnitude	Dim Sample	Bright Sample
-25.99	0.00	0.01	-24.31	0.00	0.23	-22.63	0.43	1.00
-25.78	0.00	0.02	-24.10	0.00	0.32	-22.42	0.59	1.00
-25.57	0.00	0.03	-23.89	0.00	0.46	-22.21	0.72	1.00
-25.36	0.00	0.04	-23.68	0.00	0.62	-22.00	0.82	1.00
-25.15	0.00	0.06	-23.47	0.00	0.79	-21.79	0.91	1.00
-24.94	0.00	0.09	-23.26	0.01	1.00	-21.58	0.95	1.00
-24.73	0.00	0.12	-23.05	0.13	1.00	-21.37	0.98	1.00
-24.52	0.00	0.16	-22.84	0.29	1.00	-20.95	1.00	1.00

Note. — Shows the weighted, cumulative magnitude distribution for the bright and dim quasar samples. The weighting is done such that both samples have the same redshift distribution (see Figure 5.8).

Table 5.4. Cross-correlation Fit Details

QSO Division	R Range (Mpc/h)	$\langle f \rangle$	r_0	γ	W	Separation (%)	Result Strength
1/3 Bright 2/3 Dim	[0.3,3]	$4.24 \cdot 10^{-4}$ $4.26 \cdot 10^{-4}$	6.19 4.48	1.77	96.05 52.77	97	$1.9\,\sigma$

Note. — Luminosity dependent quasar clustering using a cross-correlation technique between CS82 galaxies ($M < 23.5$) and SDSS, BOSS, and 2SLAQ quasars ($0.5 < z < 1.0$). The quasars were divided into the 1/3 brightest and 2/3 dimmest, and a cross-correlation function was calculated between a range of [0.3-3.0] Mpc. A power-law fit to the data of the form $\xi(r) = (r_0/r)^\gamma$ found that $\gamma = 1.77$ and $r_0 = 6.19$ for the bright sample and $r_0 = 4.48$ for the dim sample.

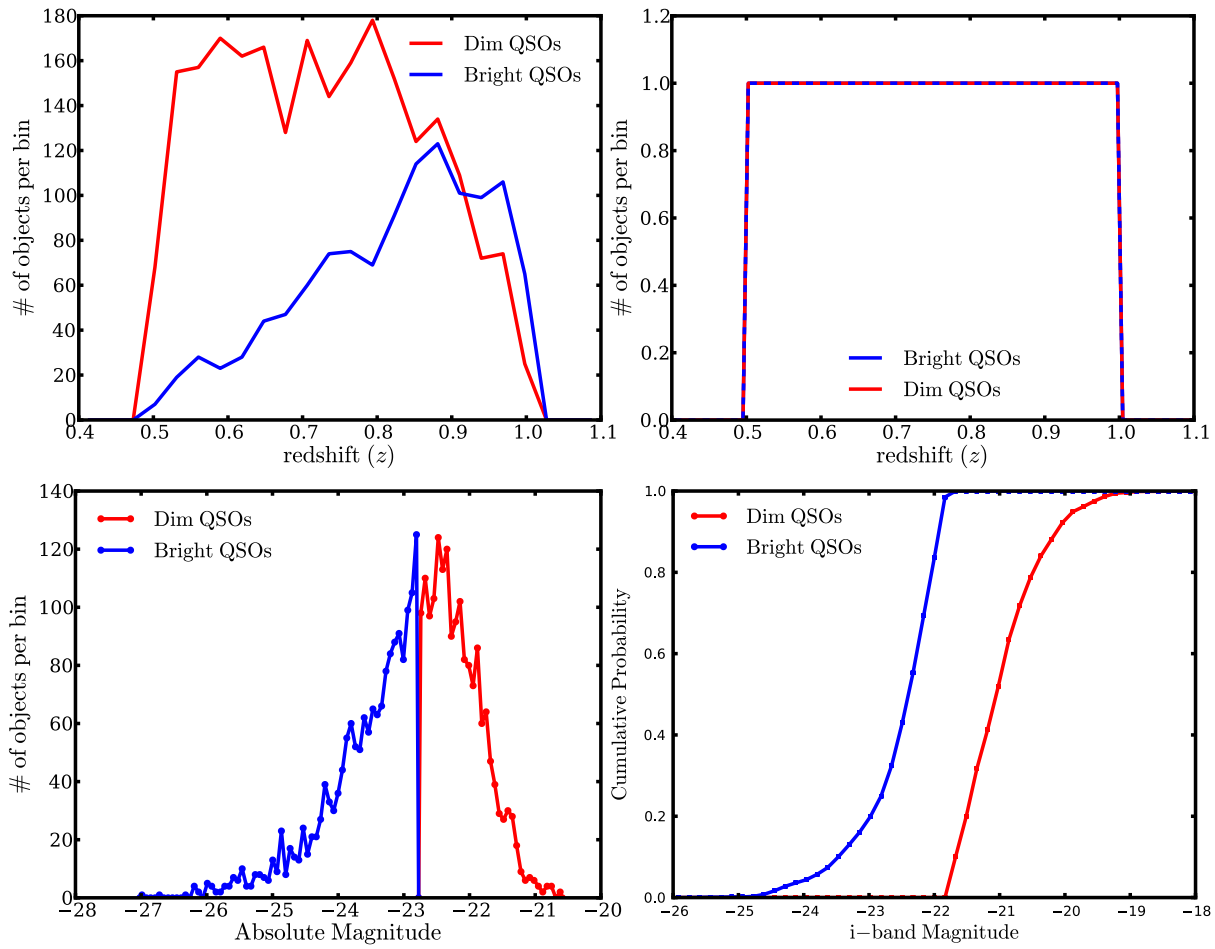


Figure 5.8: (*Top Left*) Redshift distribution of bright and dim quasar subsets. (*Top Right*) Redshift distribution of the bright and dim quasar subsets after weighting is applied. (*Bottom Left*) Absolute magnitude distribution of quasar data, split into the bright and dim subsets. (*Bottom Right*) Weighted, cumulative probability vs. magnitude. Table (5.3) shows the raw values from this figure.

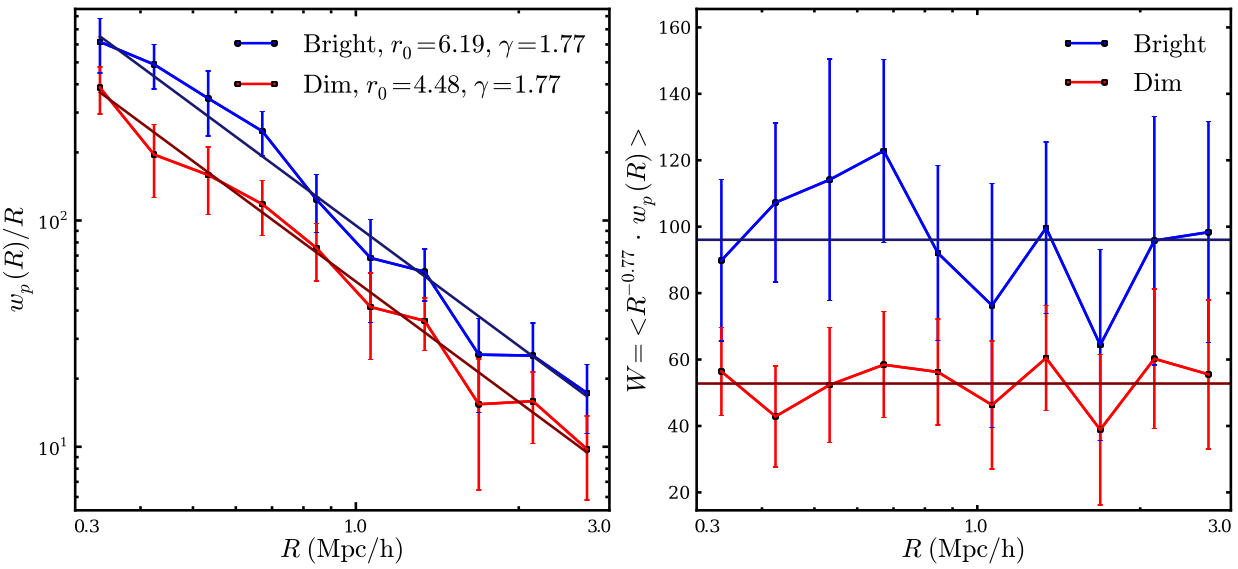


Figure 5.9: (*Left*) The galaxy-quasar cross-correlation function $w_p(R)/R$ versus R for the bright and dim samples. (*Right*) W versus R for the bright and dim samples. The error bars are calculated from bootstrap resampling of the data.

5.3.1 Testing

I did various tests to try and understand the stability of these results as well as the effect of changes to the various data sets. Here I present these tests. The tests performed involved different divisions to the quasar sample, different magnitude cuts on the galaxy sample, different correlation lengths, different methods of calculating $\langle f(\chi) \rangle$, different redshift ranges, and testing of the quasar weighting function.

Quasar Redshift Cuts

I first looked at the effect of restricting the quasar redshift range on cross-correlation function. The hope was that by applying a weight to the quasar data (Figure 5.8) I would separate out any difference in clustering strength of the quasar subsample due to redshift. However, I felt it was important to test this explicitly, by doing the calculation both using the weighting scheme and not, and also subdividing the quasar sets into different redshift bins and ranges.

First I looked at the effect of doing the calculation described in § 5.3 with and without the weighting. As one would expect when doing the calculation over a small redshift bin ($\Delta z = 0.1$), performing the weighting does not have a huge effect on the correlation functions of the luminosity dependence. This is because the redshift distributions are fairly flat over

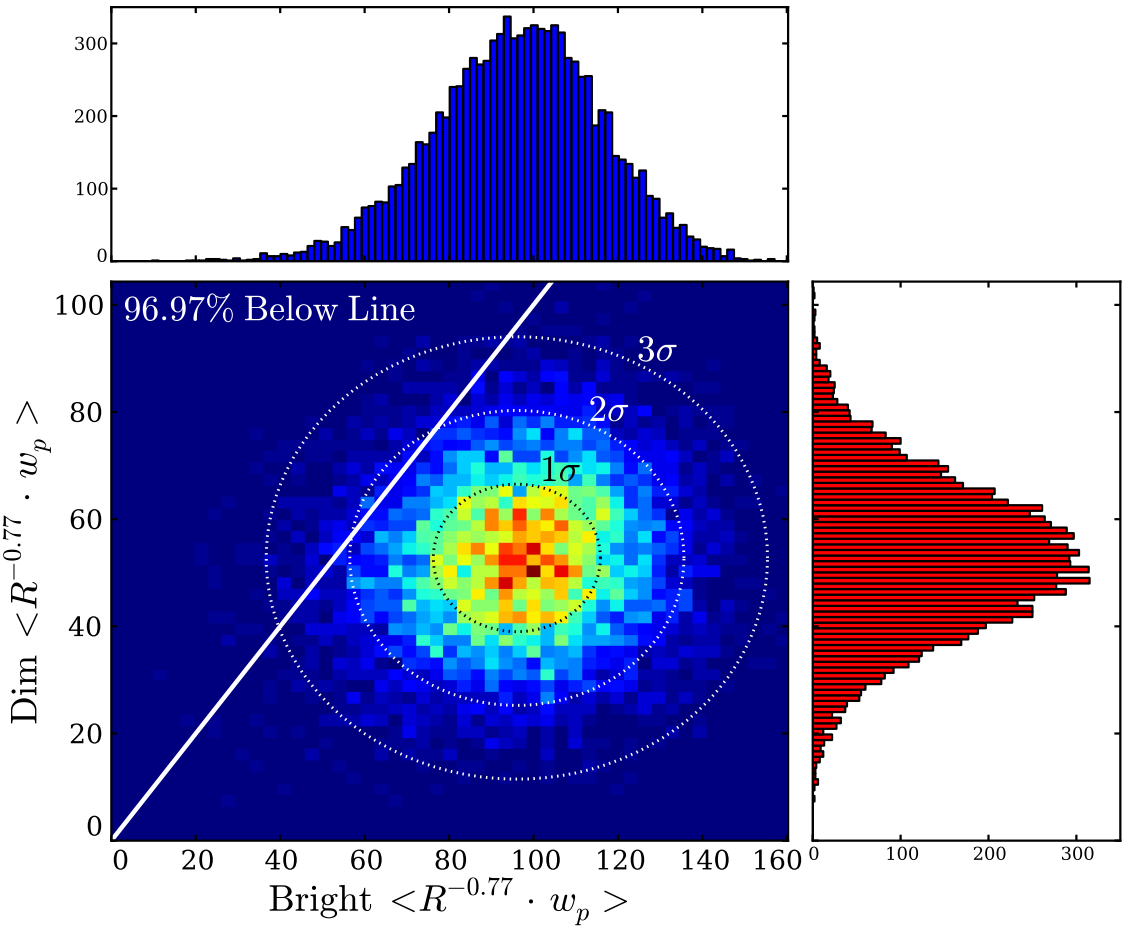


Figure 5.10: The distribution of clustering strength, as estimated from 10,000 bootstrap realizations, for the bright and dim quasar subsamples. The clustering strength is summarized as the average of $R^{-0.77} \cdot w_p(R)$ over the radial range $0.3 < R < 3\,h^{-1}$Mpc. 97% of the bright samples have a clustering strength greater than the dim samples. This results in a $1.9\,\sigma$ detection of luminosity dependent quasar clustering.

Table 5.5. Quasar Redshift Cuts Testing

z Range Weight	# in Sample	r_0	γ	W	Separation %	Result Strength
$0.5 < z < 1.0$	3,518					
Yes		6.19, 4.48	1.77	96.06, 52.77	96.97	$1.87\,\sigma$
No		6.34, 4.51	1.78	102.66, 54.05	99.59	$2.66\,\sigma$
$0.6 < z < 0.9$	2,078					
Yes		6.56, 3.90	1.90	120.06, 44.92	99.98	$3.56\,\sigma$
No		6.49, 4.13	1.86	112.58, 49.09	99.60	$2.73\,\sigma$
$0.6 < z < 0.7$	606					
Yes		5.17, 3.85	2.05	89.33, 46.80	87.73	$1.16\,\sigma$
No		5.10, 3.87	2.00	83.14, 45.94	86.34	$1.07\,\sigma$
$0.7 < z < 0.8$	708					
Yes		9.37, 4.43	1.72	182.91, 50.43	100.00	$4.44\,\sigma$
No		9.39, 4.84	1.69	182.08, 57.81	100.00	$4.30\,\sigma$
$0.8 < z < 0.9$	770					
Yes		3.18, 2.41	2.12	42.48, 27.66	65.56	$0.39\,\sigma$
No		3.47, 2.93	2.03	50.26, 36.03	64.63	$0.39\,\sigma$

Note. — Luminosity dependent quasar clustering with and without weighting and using different quasar redshift ranges. All the above tests were done with the same galaxy sample described in § 5.3 with a galaxy magnitude cut of < 23.5. The two values for r_0 and W are the fits for the bright (*Left*) and dim (*Right*) samples. The weighting has little effect for smaller redshift ranges where the redshift distribution of the bright and dim samples is fairly flat. It is interesting to note that the luminosity dependent signal improves with some of the smaller redshift ranges: $0.6 < z < 0.9$ and $0.7 < z < 0.8$. I decided to use the range $0.6 < z < 0.9$ moving forward, to maximize separation, but also maximize statistics.

these ranges. However, for larger ranges the weighting does seem to have an effect. The results of this testing can be found in Table (5.5).

Next I looked at the effect of doing the calculation using different quasar redshift ranges. I found that certain smaller redshift ranges had a stronger separation of the bright and dim clustering signals. The range $0.7 < x < 0.8$ has the strongest followed by $0.6 < z < 0.9$. Because the number of QSOs in the $0.7 < x < 0.8$ is significantly lower, I decided to use the range $0.6 < z < 0.9$ moving forward, to maximize separation, but also maximize statistics. The results of this testing can be found in Table (5.5).

Table 5.6. Galaxy Magnitude Cuts Testing

Galaxy Mag Cut	# in Sample	r_0	γ	W	Separation %	Result Strength
< 23.75	6,317,666	5.02	1.82	68.25	99.92	$3.24\,\sigma$
< 23.50	5,230,545	4.90	1.90	69.53	99.98	$3.56\,\sigma$
< 23.25	4,301,764	5.12	1.83	70.30	99.97	$3.55\,\sigma$
< 23.00	3,520,689	5.11	1.84	70.79	99.69	$2.88\,\sigma$
< 22.75	2,873,080	5.19	1.85	72.78	99.99	$4.14\,\sigma$
< 22.50	2,336,058	5.49	1.79	78.27	100.00	$4.15\,\sigma$
< 22.25	1,894,427	6.05	1.76	89.32	100.00	$4.56\,\sigma$
< 22.00	1,532,691	6.32	1.75	93.97	100.00	$3.89\,\sigma$
< 21.75	1,236,255	6.09	1.70	87.55	99.99	$4.05\,\sigma$
< 21.50	994,165	7.17	1.60	107.79	99.95	$3.51\,\sigma$
< 21.25	795,861	7.30	1.59	112.48	100.00	$4.10\,\sigma$

Note. — Luminosity dependent quasar clustering with varying galaxy magnitude cuts. The redshift range of the quasar sample is $0.6 < z < 0.9$. The values for $\langle f \rangle$, r_0, γ and W are for the full quasar sample cross-correlated with the full galaxy sample less than the magnitude cut. The separation and result strength are calculated the same way as in Table 5.4 and represent the luminosity dependent quasar clustering signal.

Galaxy Magnitude Cuts

I next looked at the effect of galaxy magnitude cuts on the cross-correlation function. By using brighter galaxies, my galaxy sample consists of higher-precision data, and thus the cross-correlation has smaller errors. However, only using brighter galaxies reduces the number of objects in my galaxy catalog and thus increases the noise in the cross-correlation. Thus there are competing factors at play, and the goal is to find a magnitude cut that maximizes the cross-correlation separation. I performed the same calculation described in § 5.3 but with different magnitude cuts on the galaxy sample. Table (5.6) summarizes the results of the effect of these galaxy magnitude cuts on the strength of the luminosity dependent clustering signal. Note that for this testing I am now using the quasar redshift range $0.6 < z < 0.9$.

Correlation Length Testing

As I already discussed in § 5.3, the correlation length (R) over which I calculate the cross-correlation affects the clustering separation between the dim and bright samples. This is because at small R there are not very many objects, and so the poisson errors are large. At

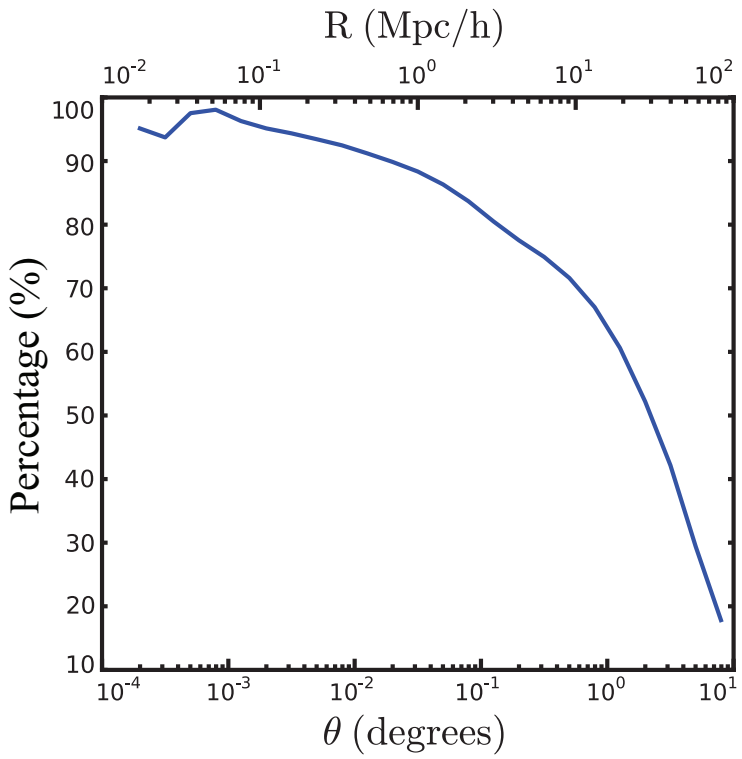


Figure 5.11: Geometric effects on the number of pairs in the correlation. This figure show the percentage of pairs in the cross-correlation calculation as a function of separation length. 100% means there is no effect of the geometry on the pair counts (this happens at very small separations). 50% means that 1/2 of the pairs are lost due to edge effects of the geometry. This is why the error bars get large with larger R in the cross-correlation function calculation.

large R, I start to lose signal because of the geometry of the stripe. Figure (5.11) shows the effects of the stripe's geometry on the number of correlation pairs as a function of R and θ.

I performed the cross-correlation at several different separation ranges, and this has a large effect on the results. Table (B.1) shows the calculation for several different correlation ranges, and how this affects the strength of the separation result. This table shows that the more "central" I choose the separation bins, the stronger the bright and dim samples are separated. Figure (5.11) demonstrates this as well.

Table 5.7. Correlation Length Testing

Correlation Range (Mpc)	r_0	γ	W	Separation %	Result Strength
$0.01 < R < 100$	5.64	1.67	84.58	98.03	$2.0\,\sigma$
$0.1 < R < 10.0$	5.18	1.79	71.02	99.79	$2.9\,\sigma$
$0.3 < R < 3.0$	4.90	1.90	69.53	99.98	$3.6\,\sigma$
$1.0 < R < 2.0$	4.16	2.49	82.65	100.00	$3.7\,\sigma$

Note. — Luminosity dependent quasar clustering with different correlation lengths. All the above tests were done with the same galaxy sample described in § 5.3 with a galaxy magnitude cut of < 23.5 and a quasar redshift range of $0.6 < z < 0.9$. Figure (5.11) shows the collapsed correlation functions and density plot for these four different correlation lengths.

Calculating $\langle f(\chi) \rangle$

In order to translate from $w_\theta(R)$ to $w_p(R)$ I need $\langle f(\chi) \rangle$, the radial distribution of galaxies (see Equation 5.3). The CS82 galaxy data is photometric, however the redshift distribution is well understood by the CFHT team. In order to test the photometric redshifts, I used several different fits to the distribution, from several different sources, to observe how these fits effect the cross-correlation calculation. Figure (5.12) shows the photometric redshifts (photo-z's) for the CS82 galaxies, as well as three fits to the data described below:

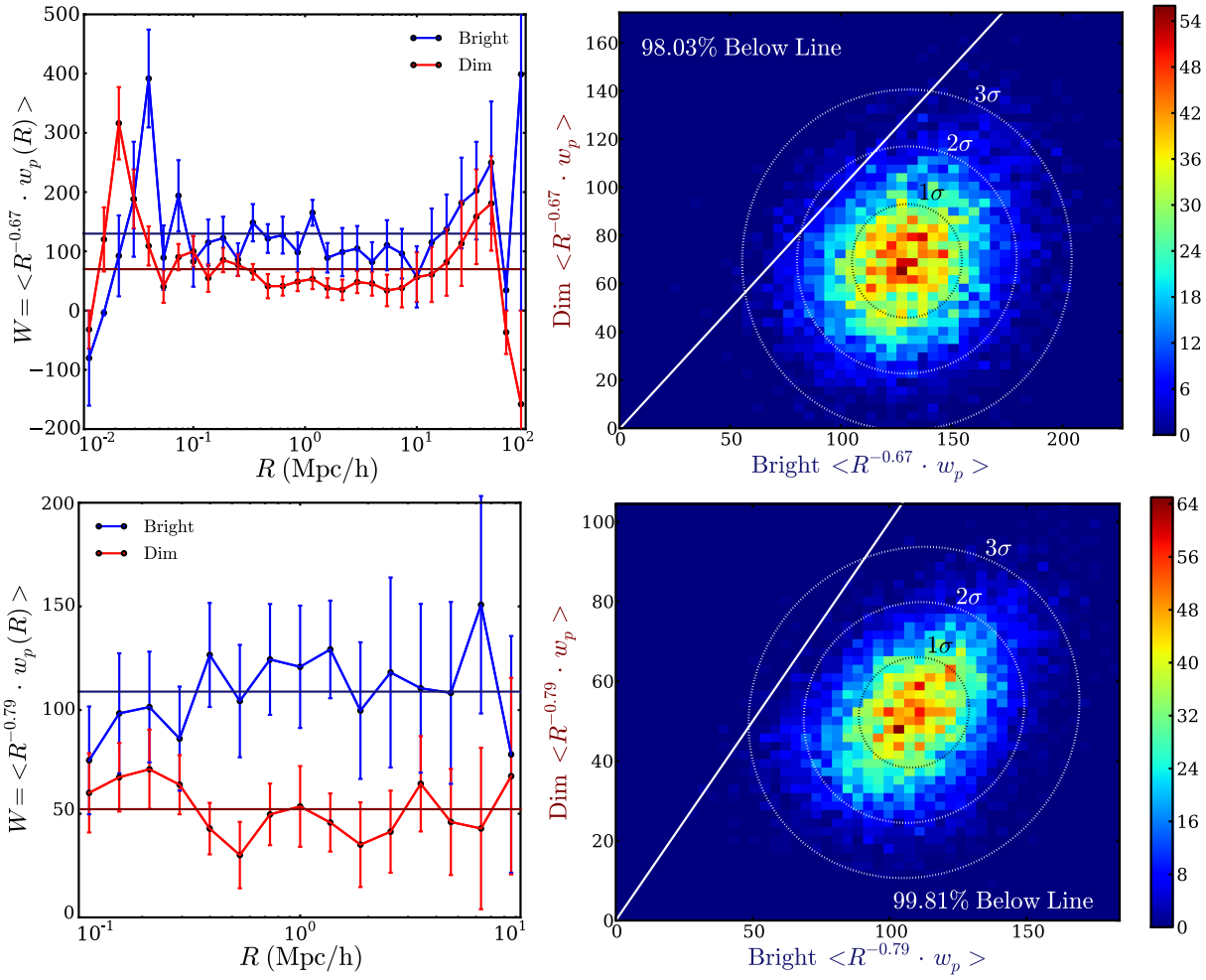
1. LSST Science Book Fit (LSST Science Collaboration 2009): Uses the LSST Science Book galaxy distribution function:

$$\begin{aligned} z_0 &= 0.0417 \cdot [\text{Mag Limit}] - 0.744 \\ \frac{dN}{dz} &= \frac{z^2 e^{-z/z_0}}{2{z_0}^3} \end{aligned} \tag{5.11}$$

2. SDSS Co-add Fit (Huff et al. 2011): Uses a smoothed fit to the redshift distribution calculated in (Huff et al. 2011) based on the SDSS Coadded Photometric Catalogs. The smoothing fit uses the following equation:

$$\frac{dN}{dz} = (z/a)^b \cdot e^{-z/c} \tag{5.12}$$

where a, b, and c are the best fit parameters. See Appendix (A) for the raw redshift distribution data from Huff.


Bright
Dim
$W = <R^{-0.67} \cdot w_p(R)>$
R (Mpc/h)
98.03% Below Line
3σ
2σ
1σ
Dim $<R^{-0.67} \cdot w_p>$
Bright $<R^{-0.67} \cdot w_p>$
Bright
Dim
$W = <R^{-0.79} \cdot w_p(R)>$
R (Mpc/h)
3σ
2σ
1σ
99.81% Below Line
Dim $<R^{-0.79} \cdot w_p>$
Bright $<R^{-0.79} \cdot w_p>$

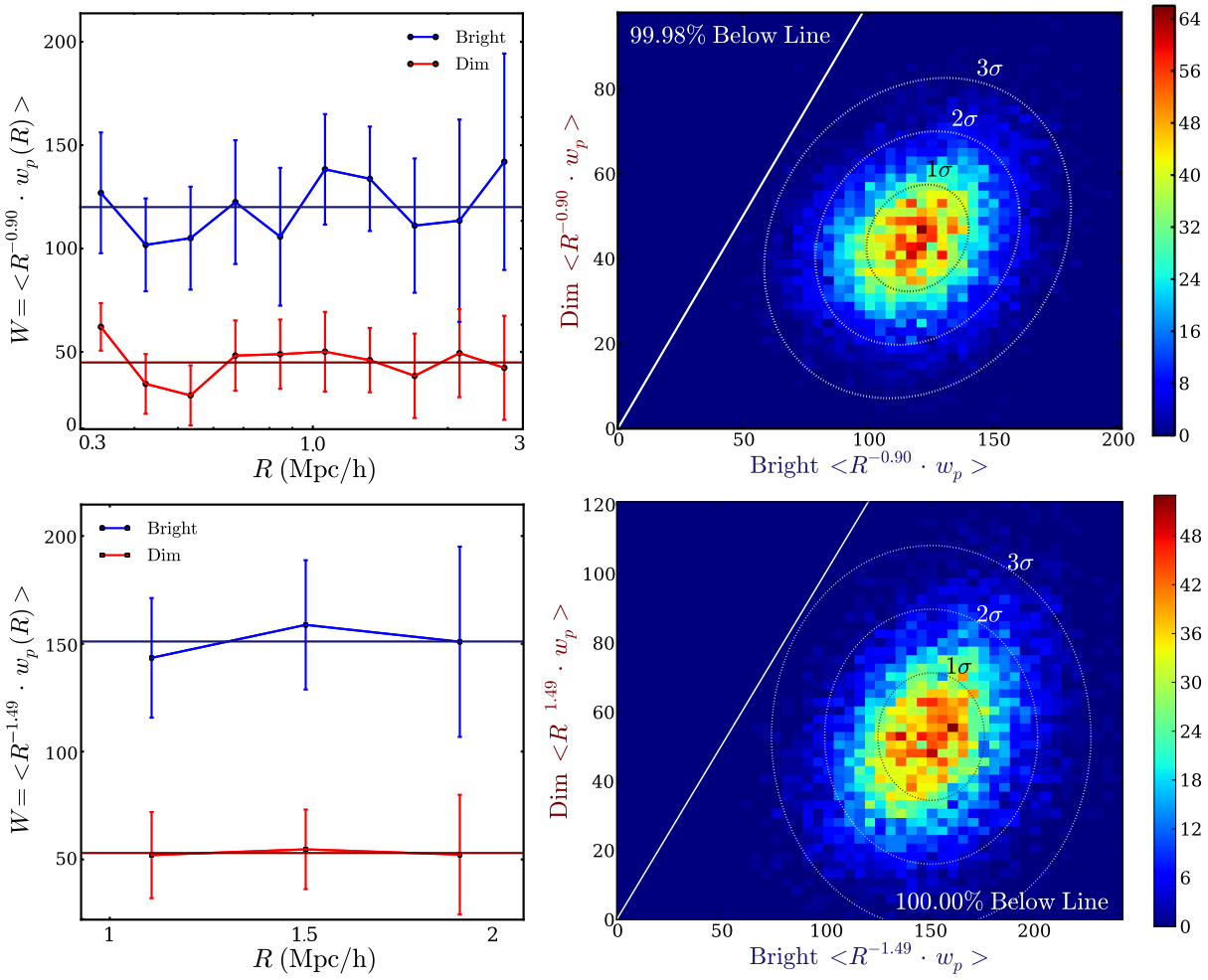


Figure 5.11: Luminosity dependent quasar clustering with different correlation lengths. All the above tests were done with the same galaxy sample described in § 5.3 with a galaxy magnitude cut of < 23.5 and a quasar redshift range of $0.6 < z < 0.9$. Table (B.1) shows the details of these runs.

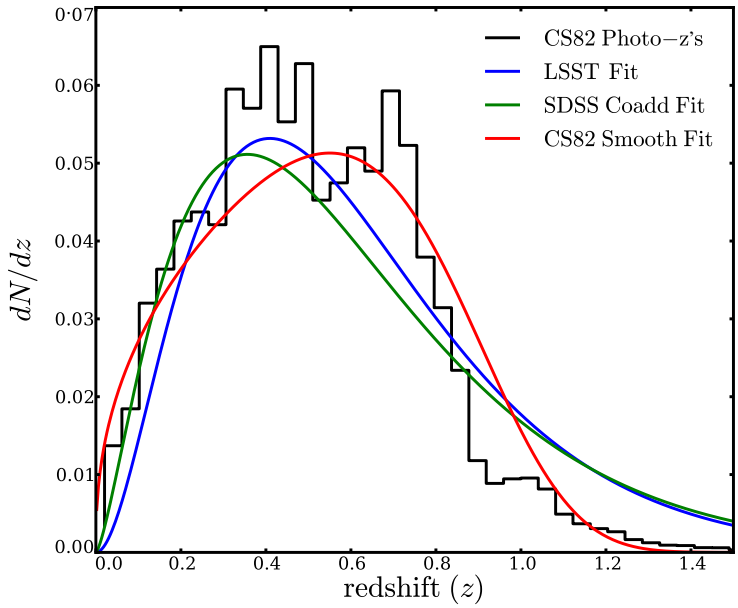


Figure 5.12: The photometric redshifts (photo-z's) for the CS82 galaxies with three fits to the data. The fits are described by Equations (5.11,5.12,5.13)

3. CS82 Smooth Fit (Erben et al. 2013): Uses a smoothed fit to the photometric redshift distributions calculated by the CS82 team. This fit uses the following equation:

$$\frac{dN}{dz} = (az^b + c)e^{-dz^e} \tag{5.13}$$

where a, b, c, d and e are the best fit parameters. See Appendix (B) for the raw redshift distribution data from the CS82 Team.

While the different redshift distribution fits give different values of r_0, they do not have a substantial effect on the clustering separation (see Table 5.8). I use the CS82 Smooth Fit for all other analysis/tests described in this chapter.

More Quasar Sample Divisions

Thus far I have discussed luminosity dependent quasar clustering, however I also tested to see if there was a clustering dependence on other divisions of the quasar sample. I also looked at the effect of dividing by black hole mass, and redshift on the clustering dependence. None of these other quasar sample divisions had as dramatic of an effect on the clustering signal as luminosity. Table (5.9) compares these other sample divisions with the luminosity/magnitude division. In order to estimate the black hole mass, I did the following. First I used models

Table 5.8. Redshift Distribution Fits Testing

Fit Method	r_0	γ	W	Separation %	Result Strength
LSST Science Book Fit	6.33	1.64	88.42	95.05	$1.66\,\sigma$
SDSS Coadd Fit	6.60	1.64	94.73	95.05	$1.66\,\sigma$
CS82 Smooth Fit	5.90	1.64	78.67	95.12	$1.67\,\sigma$

Note. — Luminosity dependent quasar clustering with different redshift distribution fits for the galaxy data. All the above test were done with the same galaxy sample described in § 5.3 with a galaxy magnitude cut of < 22.5 and a quasar redshift range of $0.5 < z < 1.0$. Figure (5.12) shows the different fit functions which are also described by Equations (5.11,5.12,5.13). While the different redshift distribution fits give different values of r_0, they do not have a substantial effect on the clustering separation. I use the CS82 Smooth Fit for all other analysis/tests described in this chapter.

from Ian McGreer (McGreer 2012) to get the monochromatic luminosity from the observed quasar flux magnitudes. The BOSS/SDSS spectroscopic pipeline provides the Hβ line widths. I can then calculate the black hole mass, following Vestergaard and Peterson (Vestergaard & Peterson 2006):

$$\log\left(\frac{M_{BH}}{M_\odot}\right) = \log\left[\left(\frac{\mathrm{FWHM(H}\beta)}{1000\ \mathrm{km\ s^{-1}}}\right)^2\left(\frac{\lambda L_\lambda(5100\text{Å})}{10^{44}\ \mathrm{egs\ s^{-1}}}\right)^{0.50}\right] + (6.91 \pm 0.02). \tag{5.14}$$

While separating by black hole mass does seem to have a clustering dependence, it is not as strong as that with luminosity. This is because using Hβ to calculate the line widths is a rough estimate, and introduces noise into the signal. I did not find a significant redshift dependence.

5.3.2 Luminosity Dependent Quasar Clustering

As discussed above changing the redshift range of the quasar sample, the magnitude cut of the galaxy sample, or the correlation length over which I am calculating the cross-correlation all effect the strength of the luminosity dependent clustering signal. The best signal of luminosity dependent quasar clustering occurs with the samples shown in Table (5.10) below.

Therefore, I find a $4.56\,\sigma$ detection of luminosity dependent quasar clustering. For this result, I am cross-correlating 2,078 quasars in the redshift range $0.60 < z < 0.90$ with 1,894,427 galaxies with a galaxy magnitude cut < 22.25. Figure (5.13) shows the correlation functions for this best-result run.

Table 5.9. Quasar Sample Divisions

Sample Division	r_0	γ	W	Separation %	Result Strength
Absolute Magnitude					
1/3 Brightest, 2/3 Dimmest	6.19, 4.48	1.77	96.06, 52.77	96.97	$1.87\,\sigma$
1/3 Brightest, 1/3 Dimmest	6.19, 4.89	1.77	96.09, 62.07	92.05	$1.40\,\sigma$
1/2 Brightest, 1/2 Dimmest	5.09, 4.48	1.77	68.00, 53.07	75.33	$0.68\,\sigma$
Black Hole Mass					
1/3 Largest, 2/3 Smallest	6.09, 4.45	1.77	91.73, 53.72	95.82	$1.68\,\sigma$
1/3 Largest, 1/3 Smallest	6.09, 4.96	1.77	91.73, 64.97	79.74	$0.86\,\sigma$
1/2 Largest, 1/2 Smallest	5.27, 4.68	1.77	71.56, 58.24	70.70	$0.56\,\sigma$
Redshift					
2/3 Furthest, 1/3 Closest	5.43, 4.66	1.77	76.07, 58.11	84.14	$1.00\,\sigma$
1/3 Furthest, 1/3 Closest	5.07, 4.63	1.77	69.59, 57.36	74.83	$0.67\,\sigma$
1/2 Furthest, 1/2 Closest	5.12, 5.02	1.77	69.75, 64.92	60.79	$0.28\,\sigma$

Note. — Testing of different quasar sample divisions. All the above tests were done with a correlation range R of $(0.3 < R < 3.0)$, a quasar redshift range of $(0.5 < z < 1.0)$, and a galaxy magnitude cut of < 23.5.

Table 5.10. Luminosity Dependent Quasar Clustering Best Results

Redshift Range	Galaxy Mag Cut	Correlation Length	r_0	γ	Separation %	Result Strength
$0.60 < z < 0.90$	< 22.25	$0.30 < R < 3.00$	8.83, 4.43	1.76	100.00	$4.56\,\sigma$
$0.70 < z < 0.80$	< 22.50	$0.30 < R < 3.00$	11.74, 5.07	1.59	100.00	$4.77\,\sigma$

Note. — The above runs have the best signal of luminosity dependent quasar clustering. The cross-correlation functions were calculated in the R-range of [0.3-3.0] Mpc/h. A power-law fit to the data using Equation (5.9) found that $r_0 = 8.83/4.43$ (for the bright/dim samples), and $\gamma = 1.76$. with a quasar redshift range of $0.60 < z < 0.90$ and a galaxy magnitude cut of < 22.25.

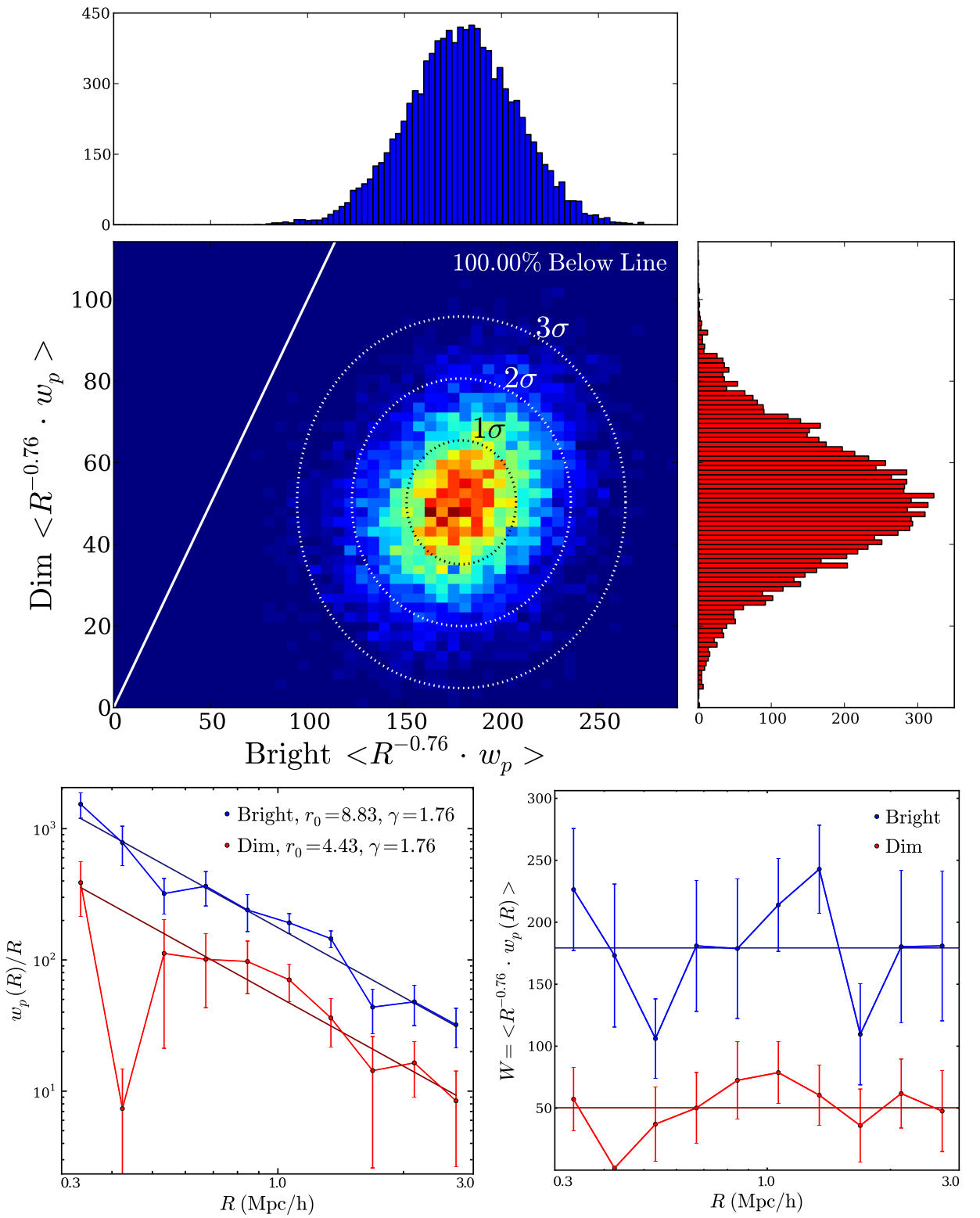


Figure 5.13: The strongest luminosity dependent quasar clustering cross-correlation signal. I cross-correlate 2,078 quasars in the redshift range $0.60 < z < 0.90$ with 1,894,427 galaxies with a galaxy magnitude cut < 22.25. I cross-correlate in the range $0.3 < R < 3.0$. This run finds a $4.56\,\sigma$ detection of luminosity dependent quasar clustering.

5.3.3 Calculating Bias

Quasars are biased tracers of the underlying dark matter. The linear scale-independent bias factor is normally defined in terms of the auto-correlation function (Shen et al. 2009):

$$b^2 = \xi/\xi_m \tag{5.15}$$

where b is the sample bias, ξ is the auto-correlation function of the sample and ξ_m is the matter correlation function. For a cross-correlation between two samples (1 and 2), I can modify Equation (5.15) as follows:

$$b_1 \cdot b_2 = \xi_{(1,2)}/\xi_m \tag{5.16}$$

where b_1 and b_2 are the biases of the two samples, $\xi_{(1,2)}$ is the cross-correlation function, and ξ_m is the matter correlation function. Recall from Equation (5.9) that $\xi \propto w_p/R$, therefore I can write the quasar bias in terms of my projected cross-correlation function, w_p:

$$b_1 \cdot b_2 = w_{p(1,2)}/w_{p,m} \quad \rightarrow \quad w_{p(1,2)} = b_1 \cdot b_2 \cdot w_{p,m} \tag{5.17}$$

$$\therefore w_{p(\mathrm{QSO,CS82})} = b_{\mathrm{QSO}} \cdot b_{\mathrm{CS82}} \cdot w_{p,m}. \tag{5.18}$$

I can estimate $w_{p,m}$ by measuring the correlation function of a dark matter simulation at the same redshift as my cross-correlation. However, Equation (5.18) still has two unknowns (b_{QSO} and b_{CS82}). Therefore in order to determine the quasar bias, I needed to perform the cross-correlation using a sample with a well-understood bias. I therefore perform a cross-correlation with the BOSS CMASS galaxy sample, where the bias is well measured to be $b_{\mathrm{CMASS}} = 2.0$. Thus I can find the quasar bias as follows:

$$w_{p(\mathrm{CMASS,\ CS82})} = b_{\mathrm{CMASS}} \cdot b_{\mathrm{CS82}} \cdot w_{p,m} \tag{5.19}$$

Inserting $b_{\mathrm{CMASS}} = 2.0$ into Equation (5.19) allows me to solve for b_{CS82}:

$$w_{p(\mathrm{CMASS,CS82})} = 2 \cdot b_{\mathrm{CS82}} \cdot w_{p,m} \quad \rightarrow \quad b_{\mathrm{CS82}} = {}^1\!/_2\, w_{p(\mathrm{CMASS,CS82})}/w_{p,m}\ . \tag{5.20}$$

Then I can solve for b_{QSO} by inserting Equation (5.20) into Equation (5.18):

$$w_{p(\mathrm{QSO,CS82})} = b_{\mathrm{QSO}} \cdot b_{\mathrm{CS82}} \cdot w_{p,m} \tag{5.21}$$

$$w_{p(\mathrm{QSO,CS82})} = {}^1\!/_2\, b_{\mathrm{QSO}} \cdot w_{p(\mathrm{CMASS,CS82})} \tag{5.22}$$

$$b_{\mathrm{QSO}} = \frac{2 \cdot w_{p(\mathrm{QSO,CS82})}}{w_{p(\mathrm{CMASS,CS82})}}\ . \tag{5.23}$$

Figure 5.14 shows plots for $w_{p(\mathrm{CMASS,CS82})}$, $w_{p(\mathrm{QSO,CS82})}$, and $w_{p,m}$ for samples centered around $z = 0.7$. As I mentioned above, $w_{p,m}$ is from a dark matter simulation. I determined the quasar bias as a function of redshift for several different magnitude cuts on the galaxy sample and show that this quasar bias is stable, see Table 5.11.

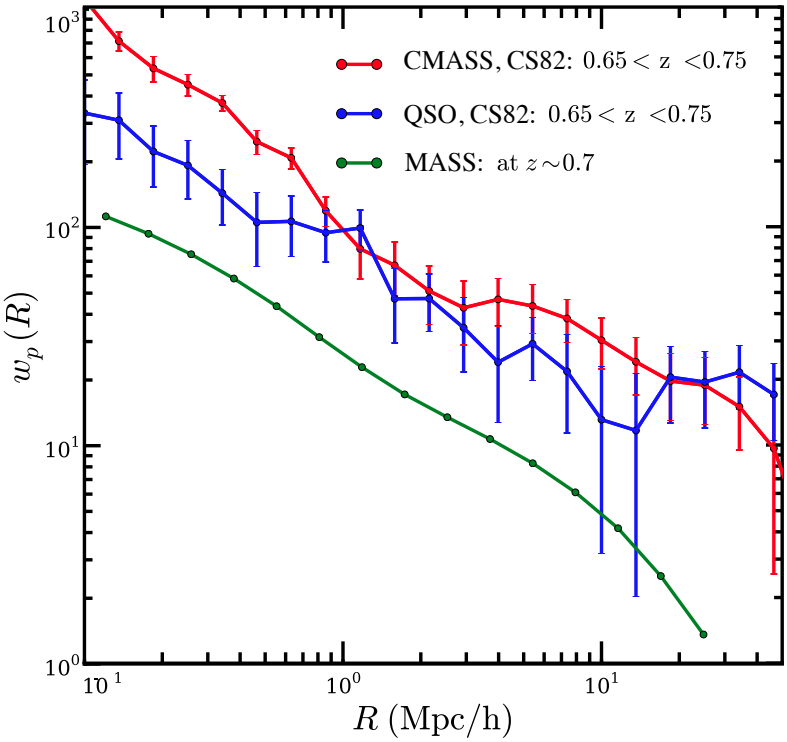


Figure 5.14: Correlation functions used to calculate the quasar bias. First a cross-correlation between the CMASS galaxies and the CS82 galaxies (red) is performed to determine the bias of the CS82 galaxies. Then a cross-correlation between the quasars and the CS82 galaxies (blue) is performed to determine the bias of the quasars. The bias of the mass is determined by a dark matter simulation (green). All of these cross-correlations are done in a restricted redshift range to reduce redshift effects.

Table 5.11. Quasar Bias Calculation, Fixed Redshift

redshift	Mag Lim	$w_{p,m}$	$w_{p(\mathrm{CMASS,CS82})}$	$w_{p(\mathrm{QSO,CS82})}$	b_{CS82}	b_{QSO}
$0.65 < z < 0.75$	23.50	28.88	85.45	62.2585	1.48±0.37	1.46±0.18
$0.65 < z < 0.75$	23.00	28.88	92.45	69.3466	1.60±0.41	1.50±0.21
$0.65 < z < 0.75$	22.50	28.88	94.70	77.5768	1.64±0.43	1.64±0.18
$0.65 < z < 0.75$	22.00	28.88	99.01	81.3316	1.71±0.47	1.64±0.17
$0.65 < z < 0.75$	21.50	28.88	114.58	87.4214	1.98±0.58	1.53±0.22
$0.65 < z < 0.75$	21.00	28.88	124.07	90.5306	2.15±0.89	1.46±0.49

Note. — Numbers for calculating the quasar bias for various galaxy magnitude cuts. The bias calculation is performed as outline in Equations (5.15) through (5.23). The quasar bias is stable to within $\sim 15\%$.

5.4 Discussion

5.4.1 Previous Results

There has been a steady run of results regarding quasar clustering over the last decade or so as large spectroscopic samples have become available. Several papers (Porciani et al. 2004; Croom et al. 2005; da Ângela et al. 2008; Ross et al. 2009; Shen et al. 2009) measure the quasar auto-correlation function, generally across $0.5 < z < 2.2$ with a mean redshift closer to $z \sim 1.4$.

Padmanabhan et al. (2009) measures a quasar-galaxy cross-correlation, as do Coil et al. (2007). Donoso et al. (2010) in the same redshift range, also do a cross-correlation, but this time with Radio-Loud AGN (RLAGN) and photometric SDSS LRGs. These authors find a range of biases depending on how they divide their sample.

Table (5.12) summarizes my results compared with previously published work[3]. My results strongly favor the general consensus that the bias of low-redshift quasars is ∼1 with the bias increasing slightly with increasing redshift (see Figure 5.15); this is also consistent with the models of Hopkins et al. (Hopkins et al. 2007a) as well as the previous extrapolations by Croom et al. (2005). There are two significant exceptions: First Myers et al. (2007a) find 1.93 ± 0.14 based on a photometrically selected sample of quasars. However, it is possible that contamination by a high-redshift population could boost the measured bias values. The more intriguing discrepancy is with (Mountrichas et al. 2009) who analyze a similar sample to mine , also in cross-correlation with LRGs, and find biases between 1.90 ± 0.16 and 1.45 ± 0.11 depending on the particular LRG and quasar sample they cross-correlate against. These results are also discrepant with da Ângela et al. (2008), Padmanabhan et al. (2009), and Krumpe et al. (2010, 2012) (with whom I am consistent) who analyze the same sample in auto-correlations. Furthermore, the scatter in the different subsamples analyzed by Mountrichas et al. (2009) significantly exceeds their quoted errors, suggesting either a systematic in their analysis or an underestimate of their errors. Using the observed scatter between the different subsamples as an estimate of the error yields a value consistent with my measurement.

I also compared my cross-correlation to a similar analysis done by Shen et al. (2012). Shen cross-correlated DR7 quasars with BOSS galaxies. I find the same quasar bias as Shen (see Figure 5.16) when I performed a cross-correlation using a quasar set with the same mean redshift as Shen's. Shen performed his cross-correlation with the CMASS galaxy sample, which has a bias of $b_{\rm CMASS} = 2.0$. I performed my cross-correlation with the CS82 galaxy sample, which has a bias of $b_{\rm CS82} = 1.5$. Therefore if these two data-sets have the same

[3]I do caution the reader that the errors for a number of these measurements are simply Poisson errors, and ignore correlations between different scales and are therefore likely underestimated.

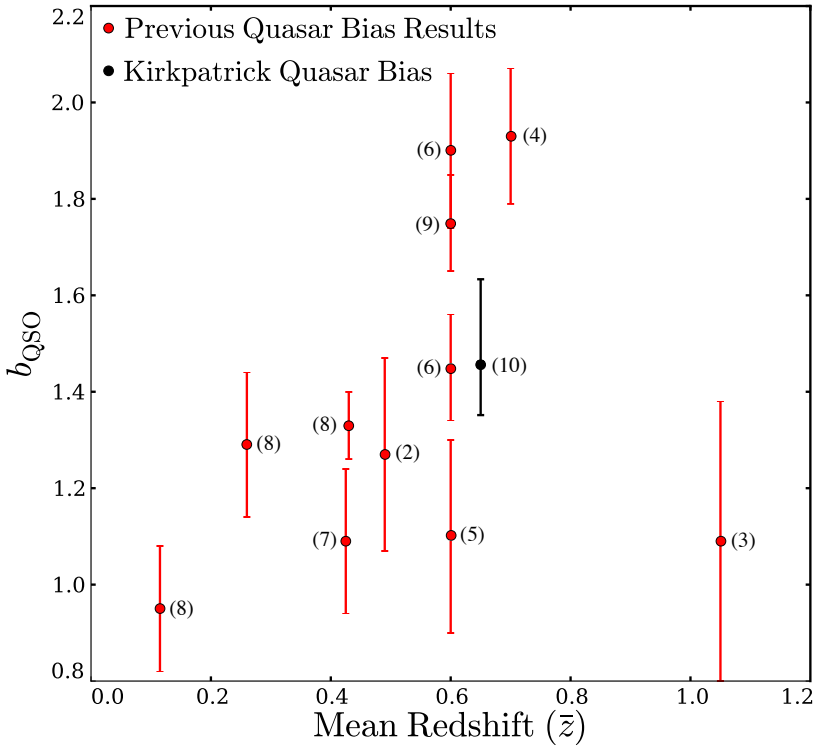


Figure 5.15: A summary of previous low redshift quasar clustering results, compared with results from this work, scaled to the cosmology assumed here. The numbers refer to the data described in Table (5.12). My results strongly favor the general consensus that the bias of low-redshift quasars is ∼1 with the bias increasing slightly with increasing redshift. There are three significant exceptions (Myers et al. 2007a; Mountrichas et al. 2009; Donoso et al. 2010). These results are discrepant with da Ângela et al. (2008), Padmanabhan et al. (2009), and Krumpe et al. (2010, 2012)– with whom I am consistent – who analyze the same sample in auto-correlations.

Table 5.12. Previous Quasar Clustering Results

z	$L_{\rm min}$	b_{QSO}	Reference
$z < 0.3$	$0.4L_\star$	0.97 ± 0.05	(1)
$0.3 < z < 0.68$	$0.4L_\star$	1.27 ± 0.20	(2)
$0.7 < z < 1.4$	$0.1L_\star$	1.09 ± 0.29	(3)
$0.4 < z < 1.0$	$0.1L_\star$	1.93 ± 0.14	(4)
$z \sim 0.6$	$0.4L_\star$	1.10 ± 0.20	(5)
$z \sim 0.6$	$0.4L_\star$	1.90 ± 0.16	(6)
$z \sim 0.6$	$2.5L_\star$	1.45 ± 0.11	(6)
$0.25 < z < 0.6$	$L_\star$	1.09 ± 0.15	(7)
$0.07 < z < 0.16$	$?L_\star$	0.95 ± 0.13	(8)
$0.16 < z < 0.36$	$?L_\star$	1.29 ± 0.15	(8)
$0.36 < z < 0.50$	$?L_\star$	1.33 ± 0.07	(8)
$0.40 < z < 0.80$	$?L_\star$	1.75 ± 0.10	(9)
$\mathbf{0.65 < z < 0.75}$	$?\mathbf{L}_\star$	$\mathbf{1.46 \pm 0.18}$	(10)

Note. — A summary of previous low redshift quasar clustering results, compared with results in this work, scaled to the cosmology assumed here. (1) Croom et al. (2004). (2) Croom et al. (2005). (3) Coil et al. (2007). (4) Myers et al. (2007a). (5) da Ângela et al. (2008) . (6) Mountrichas et al. (2009). (7) Padmanabhan et al. (2009). (8) Krumpe et al. (2010, 2012). (9) Donoso et al. (2010). (10) The work presented in this chapter. For the results from (Coil et al. 2007), I scale the relative bias presented there by the large scale bias $b = 1.22$ of all the galaxies (Coil et al. 2007).

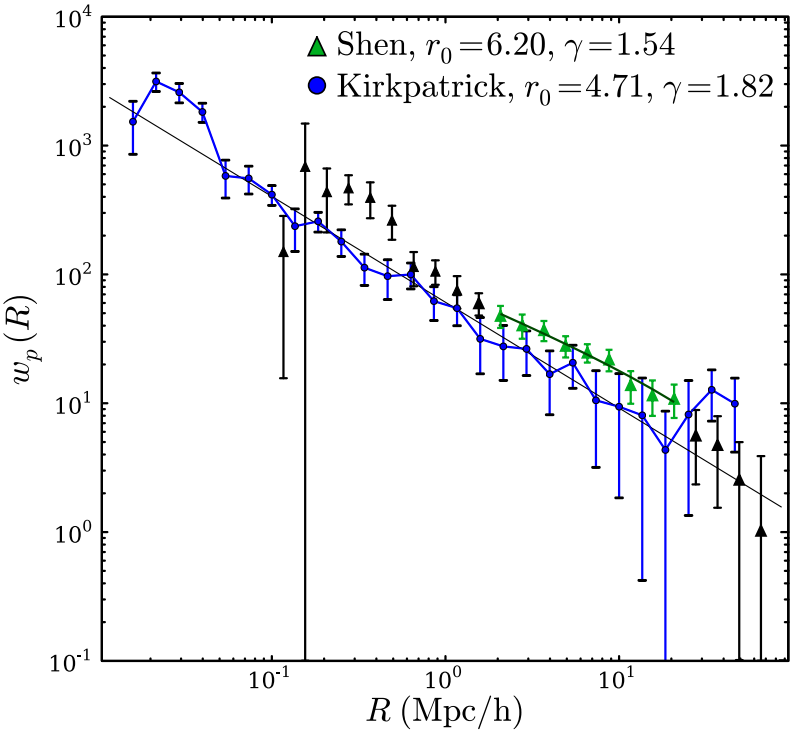


Figure 5.16: Comparison between my cross-correlation function (Kirkpatrick) to that done by Shen et al. (2012) which is a cross-correlation of DR7 quasars and BOSS galaxies. Shen's sample has a larger r_0 which is expected because the CMASS galaxy sample (that Shen uses in his cross-correlation) has a larger bias than the CS82 sample (that I use in the cross-correlation).

quasar bias, the r_0 fits to our two cross-correlation functions should have the same ratio as the two galaxy sample biases:

$$\frac{b_{\rm CS82}}{b_{\rm CMASS}} = \frac{1.5}{2.0} = 0.750, \qquad \frac{r_{0,\rm Kirkpatrick}}{r_{0,\rm Shen}} = \frac{4.7}{6.2} = 0.758 \tag{5.24}$$

Therefore my sample has the same quasar bias as Shen's (within errors).

However, Shen did not measure a luminosity dependent signal with his cross-correlation calculation (see Figure 5.17). This could perhaps be because Shen is using the 10% most luminous quasars and comparing their clustering with the bottom 90%, whereas my calculation compares the top 33% with the bottom 66%. Also my galaxy sample is 35× larger than Shen's galaxy sample and so my cross-correlation function has much smaller errors. When I do the same analysis as Shen on my quasar sample (10%/90%), I do not see as strong of a luminosity dependence ($1.56\,\sigma$ result). Figure (5.18) shows my reproduction of the Shen analysis side by side with Shen's findings.

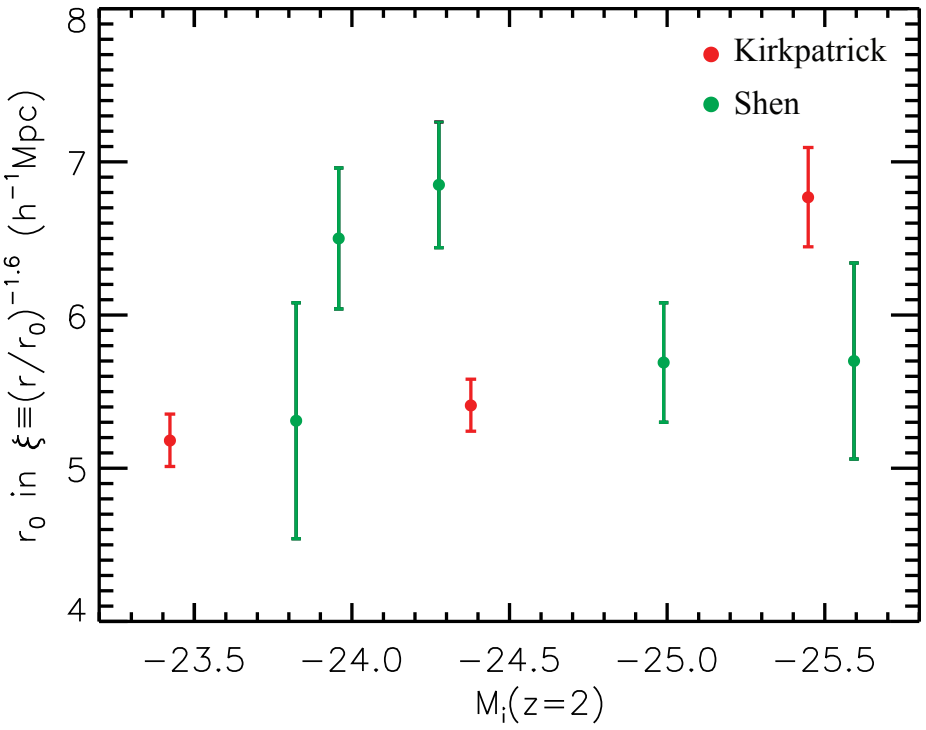


Figure 5.17: This Figure shows the values of r_0 in the fit given by Equation (5.9) as a function of absolute i-band magnitude (M_i) for various samples using my cross-correlation technique (red, Kirkpatrick) and that done by Yue Shen (green, Shen) which is a cross-correlation of DR7 quasars and BOSS galaxies. The red points seem to have an increased r_0 with increased magnitude, whereas Shen's do not.

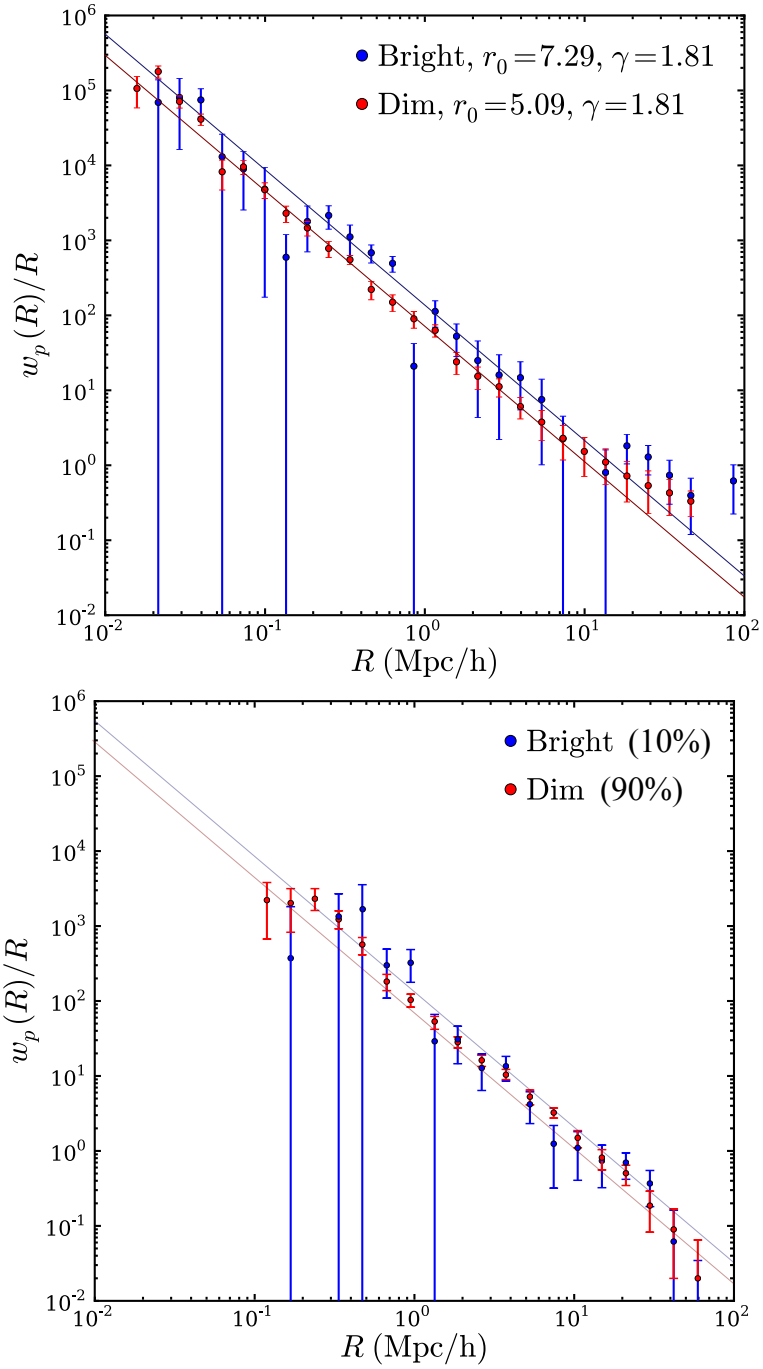


Figure 5.18: Reproduction of the Shen et al. (2012) result. Shen splits the sample by the 10% brightest quasars and the 90% dimmest quasars. Shen (*Bottom*) does the cross-correlation between DR7 quasars and BOSS galaxies. I show this same bright/dim division but with my quasar and galaxy sample (*Top*). I adjust the quasar redshift range such that it has the same mean redshift as Shen's. I see a $1.56\,\sigma$ detection of luminosity dependence with this division. To ease comparison I have put the "best fit" lines from my correlation (*Top*) functions on the Shen plot (*Bottom*)

5.5 Summary and Conclusions

I measured the cross-correlation of a sample of ∼5.2 photometric galaxies from the CFHT and a volume limited sample of ∼2,078, $0.6 < z < 0.9$ QSOs. The cross-correlation of QSOs and galaxies is well described on all measured scales by a power law of slope 1.77 ± 0.1 and a scale length of $5.05 \pm 0.05\, h^{-1}$Mpc, consistent with observed slopes and amplitudes for local galaxies.

The large scale bias is 1.46 ± 0.18 at redshift $z = 0.7$ and is consistent with most previous measurements and theoretical models, the exceptions being (Myers et al. 2007a) and Mountrichas et al. (2009); possible reasons for this discrepancy are discussed in §5.4.

I see evidence for variations of the clustering/bias with luminosity. When I divide the quasar sample into low/high luminosity samples I find luminosity depended quasar clustering at a $4.56\,\sigma$ significance level.

Chapter 6

Conclusions and Future Prospects

I refer the reader to Chapter 4, §5.5, §3.6 for more detailed conclusions on the various topics explored in this thesis. Below is I highlight my main results and some future plans:

6.1 Quasar Target Selection

This thesis describes the BOSS quasar target selection algorithms during the first two years of BOSS observations. BOSS aims to obtain spectra of a sample of ∼150,000 $z > 2.2$ quasars, in order to probe structure in the Lyα forest to provide a percent-level measurement of the expansion history of the Universe, by measuring baryon oscillations in the Lyα forest clustering. This first year was a commissioning period for quasar target selection, and the algorithms for identifying quasar candidates varied significantly over the year.

The target selection team's key results are:

- We have performed quasar target selection (QTS) over 10,200 deg^2 of the SDSS-III imaging footprint, producing a list of 484,000 targets. These objects are selected to be at redshift $z > 2.2$, motivated by the need to observe the Lyα forest in the BOSS wavelength coverage.
- After a year of testing and evolution of the BOSS QTS, we settled on the *XDQSO* method as our uniformly-selected subsample (CORE) and a neural network Combinator for the BONUS sample.
- Having the BONUS selection allows us to implement improvements throughout the survey, e.g., through auxiliary photometric data. This has already been achieved with

the inclusion of NIR $YJHK$ photometry from the UKIDSS and UV data from GALEX, increasing our $z > 2.2$ quasar yields by $\sim 2 - 3$ deg^{-2}.

- We obtained spectra of 54,909 objects selected by the quasar target selection algorithms over a footprint of 878 deg^2 during the first year of observations. The mean target density is 63.8 targets deg^{-2}.

- Of these 54,909 spectra, 33,556 were unique objects and had high quality spectra. 11,149 had redshifts $z < 0.02$, and 13,580 had redshifts of $z > 2.20$ (of which 11,263 were not previously known).

- Our mean $z > 2.2$ quasar surface density was 15.46 $z > 2.20$ quasar deg^{-2}, with a global efficiency of 26.0%.

- The $z > 2.2$ objects selected by the three main methods used during Year One are found in different regions in color-color and color-magnitude space, reflecting in part the fact that the methods were trained for different redshift ranges. The three methods complemented each other well, and together select 60-70% of *all* quasars in our magnitude range with $2.2 < z < 3.5$.

- Working with single-epoch SDSS data, our current target selection algorithms slightly exceed the BOSS technical goal of selecting 15 $z > 2.2$ quasars deg^{-2} from 40 targets deg^{-2} (Eisenstein et al. 2011). The tests on Stripe-82 indicate an efficiency of 15.4 quasars deg^{-2}, of which 11.2 deg^{-2} come from known quasars plus the CORE selection at 20 targets deg^{-2} (Figure 4.3). We anticipate that use of auxiliary imaging data, including GALEX, UKIDSS, and additional SDSS epochs in overlap regions, will boost our efficiency by $1 - 4$ quasars deg^{-2}, significantly increasing the statistical power of BOSS Lyα forest clustering measurements.

- All BOSS spectra from the first two years of observations, August 2009 through to July 2011, will be made publicly available in the next SDSS data release, DR9.

The team continues to investigate ways to improve quasar target selection. We have already described the incorporation of data from ultraviolet (GALEX) and near-IR (UKIDSS). Data from the Wide-field Infrared Survey Explorer (WISE; Wright et al. 2010) will provide photometry at mid-infrared wavelengths for our targets; it is deep enough to detect at least the brighter quasars in the BOSS sample. Variability as measured from repeat scans is an important method, independent of colors, to separate quasars from stars. Building on the SDSS Stripe-82 study by Sesar et al. (2007), recent investigations (Palanque-Delabrouille et al. 2011; Butler & Bloom 2011; MacLeod et al. 2011; Richards et al. 2011; Kozłowski et al. 2011; Sarajedini et al. 2011) have re-invigorated the field of AGN identification through variability selection.

In addition to Stripe-82, roughly 50% of the SDSS imaging footprint has been imaged more than once (Aihara et al. 2011), primarily in overlaps between adjacent stripes. However, most of this area is observed only a few times, over timescales of days, rather than the desired month or year baselines that lead to efficient AGN selection.

In this regard, the Palomar Transient Factory (PTF; Law et al. 2009)[1] could be a natural dataset to use for this purpose. The PTF is an automated, wide-field imaging survey aimed at the exploration of the optical transient sky. PTF uses the 1.2m Schmidt telescope at Palomar Observatory with a 8 deg^2 field-of-view to perform large area transient searches. An area of several hundred deg^2 can be imaged in one night, typically in the Mould R-band but also in the SDSS g-band. We are actively investigating the inclusion of PTF imaging data into BOSS QTS.

PTF could also potentially aid BOSS QTS by improving star/galaxy separation at the faint end. Potentially any of the PTF variability methods could work with other transient/variability based surveys as well, e.g. the Pan-STARRS survey (Kaiser et al. 2002).

6.2 Luminosity Dependent Quasars

With BOSS's new catalog of quasar and galaxy data, exciting new science can be done. In this thesis I also presented measurements of the luminosity-dependence of quasar clustering using QSO data from SDSS, 2SLAQ, and BOSS.

This was done by measuring the cross-correlation of a sample of ∼5.2 photometric galaxies from the CFHT and a volume limited sample of ∼2,078, $0.6 < z < 0.9$ QSOs. The cross-correlation of QSOs and galaxies is well described on all measured scales by a power law of slope 1.77 ± 0.1 and a scale length of $5.05 \pm 0.05\,h^{-1}$Mpc, consistent with observed slopes and amplitudes for local galaxies.

I determined a large-scale quasar bias, $b_{QSO} = 1.46 \pm 0.18$, at redshift $z = 0.7$. When I divide the quasar sample into low/high luminosity samples I find luminosity depended quasar clustering at a $4.56\,\sigma$ significance level.

In the future, I hope to interpret my observations, which lie largely in the non-linear regime, using a halo catalog from an N-body simulation. This will allow me to constrain the halo mass of the bright and dim quasar's host halos, which in turn can help us understand the lifetimes of these QSOs (Cole & Kaiser 1989; Haiman & Hui 2001; Martini & Weinberg 2001) and their duty cycles.

Please give me a PhD.

[1] http://www.astro.caltech.edu/ptf/

Bibliography

A. P. Dempster, N. M. L. & Rubin, D. B. 1977, Journal of the Royal Statistical Society. Series B, 39, 1

Abazajian, K. N. et al. 2003, AJ, 126, 2081

—. 2004, AJ, 128, 502

—. 2009, ApJS, 182, 543

Adelman-McCarthy, J. K. et al. 2006, ApJS, 162, 38

—. 2008, ApJS, 175, 297

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

Albrecht, A. et al. 2006, ArXiv Astrophysics e-prints

Allen, J. T., Hewett, P. C., Maddox, N., Richards, G. T., & Belokurov, V. 2011, MNRAS, 410, 860

Bahcall, J. N. & Goldsmith, S. 1971, ApJ, 170, 17

Becker, R. H., White, R. L., & Helfand, D. J. 1995, ApJ, 450, 559

Belokurov, V. et al. 2006, ApJ Lett., 642, L137

Bertin, E. & Arnouts, S. 1996, A&AS, 117, 393

Blanton, M. R. et al. 2003, AJ, 125, 2276

Bovy, J., Hogg, D. W., & Roweis, S. T. 2011a, Annals of Applied Statistics, 5, 1657

Bovy, J. et al. 2011b, ApJ, 729, 141

Butler, N. R. & Bloom, J. S. 2011, AJ, 141, 93

Casali, M. et al. 2007, Astron. & Astrophys., 467, 777

Cattaneo, A. et al. 2009, Nature, 460, 213

Chiu, K., Richards, G. T., Hewett, P. C., & Maddox, N. 2007, MNRAS, 375, 1180

Coil, A. L., Hennawi, J. F., Newman, J. A., Cooper, M. C., & Davis, M. 2007, ApJ, 654, 115

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

Croft, R. A. C., Weinberg, D. H., Katz, N., & Hernquist, L. 1998, ApJ, 495, 44

Croft, R. A. C. et al. 2002, ApJ, 581, 20

Croom, S. M., Warren, S. J., & Glazebrook, K. 2001, MNRAS, 328, 150

Croom, S. M. et al. 2004, MNRAS, 349, 1397

—. 2005, MNRAS, 356, 415

—. 2009, MNRAS, 392, 19

Croton, D. J. 2009, MNRAS, 394, 1109

da Ângela, J. et al. 2008, MNRAS, 383, 565

Dawson, K. S. et al. 2012, ArXiv e-prints

Donoso, E., Li, C., Kauffmann, G., Best, P. N., & Heckman, T. M. 2010, MNRAS, 407, 1078

Dye, S. et al. 2006, MNRAS, 372, 1227

Efron, B. & Gong, G. 1983, American Statistician, 37, 36

Eisenstein, D. J. et al. 2011, AJ, 142, 72

Erben, T. et al. 2013, *in prep.*

Fan, X. 1999, AJ, 117, 2528

Fan, X., Carilli, C. L., & Keating, B. 2006a, ARA&A, 44, 415

Fan, X. et al. 2006b, AJ, 132, 117

Ferrarese, L. & Merritt, D. 2000, ApJ Lett., 539, L9

Fukugita, M., Ichikawa, T., Gunn, J. E., Doi, M., Shimasaku, K., & Schneider, D. P. 1996, AJ, 111, 1748

Giannantonio, T. et al. 2006, Phys. Rev. D, 74, 063520

Giannantonio, T. et al. 2008, Phys. Rev. D, 77, 123520

Gibson, R. R. et al. 2009, ApJ, 692, 758

Gray, A. G. & Moore, A. W. 2003, in Proceedings of the Third SIAM International Conference on Data Mining (SIAM), 203

Gray, A. G. & Riegel, A. W. 2006, in COMPSTAT 2006 - Proceedings in Computational Statistics, 845

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

—. 2006, AJ, 131, 2332

Haiman, Z. & Hui, L. 2001, ApJ, 547, 27

Hambly, N. C. et al. 2008, MNRAS, 384, 637

Hennawi, J. F. et al. 2006a, AJ, 131, 1

—. 2006b, ApJ, 651, 61

—. 2010, ApJ, 719, 1672

Hewett, P. C., Warren, S. J., Leggett, S. K., & Hodgkin, S. T. 2006, MNRAS, 367, 454

Hodgkin, S. T., Irwin, M. J., Hewett, P. C., & Warren, S. J. 2009, MNRAS, 394, 675

Hogg, D. W., Finkbeiner, D. P., Schlegel, D. J., & Gunn, J. E. 2001, AJ, 122, 2129

Hopkins, P. F., Hernquist, L., Cox, T. J., Di Matteo, T., Robertson, B., & Springel, V. 2006, ApJS, 163, 1

Hopkins, P. F., Richards, G. T., & Hernquist, L. 2007a, ApJ, 654, 731

Hopkins, P. F. et al. 2007b, ApJ, 662, 110

Huff, E. M., Eifler, T., Hirata, C. M., Mandelbaum, R., Schlegel, D., & Seljak, U. 2011, ArXiv e-prints

Ibata, R. A., Gilmore, G., & Irwin, M. J. 1995, MNRAS, 277, 781

Ivezić, Ž. et al. 2003, Mem. Soc. Astron. Italiana, 74, 978

—. 2004, Astronomische Nachrichten, 325, 583

—. 2007, AJ, 134, 973

—. 2008, ArXiv e-prints

Jiang, L. et al. 2006, AJ, 131, 2788

Jimenez, R. et al. 2009, ApJS, 181, 439

Kaiser, N. et al. 2002, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 4836, 154–164

Kauffmann, G. & Haehnelt, M. 2000, MNRAS, 311, 576

Keller, S. C. et al. 2007, PASA, 24, 1

Kirkpatrick, J. A. et al. 2011, ApJ, 743, 125

Komatsu, E. et al. 2011, ApJS, 192, 18

Kozłowski, S., Kochanek, C. S., & Udalski, A. 2011, ApJS, 194, 22

Kozlowski, S. et al. 2010, ApJ, 708, 927

Krumpe, M., Miyaji, T., & Coil, A. L. 2010, ApJ, 713, 558

Krumpe, M., Miyaji, T., Coil, A. L., & Aceves, H. 2012, ApJ, 746, 1

Law, N. M. et al. 2009, PASP, 121, 1395

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

Loverde, M., Marnerides, S., Hui, L., Ménard, B., & Lidz, A. 2010, Phys. Rev. D, 82, 103507

LSST Science Collaboration. 2009, ArXiv e-prints

Lupton, R., Gunn, J. E., Ivezić, Z., Knapp, G. R., & Kent, S. 2001, in Astronomical Society of the Pacific Conference Series, Vol. 238, 269

Lupton, R. H., Gunn, J. E., & Szalay, A. S. 1999, AJ, 118, 1406

Lynds, R. 1971, ApJL, 164, L73+

MacLeod, C. L. et al. 2011, ApJ, 728, 26

Maddox, N., Hewett, P. C., Warren, S. J., & Croom, S. M. 2008, MNRAS, 386, 1605

Martin, D. C. et al. 2005, ApJ Lett., 619, L1

Martini, P. & Weinberg, D. H. 2001, ApJ, 547, 12

Matthews, T. A. & Sandage, A. R. 1963, ApJ, 138, 30

McDonald, P. 2003, ApJ, 585, 34

McDonald, P. & Eisenstein, D. J. 2007, Phys. Rev. D, 76, 063009

McDonald, P. et al. 2006, ApJS, 163, 80

McGreer, I. D. 2012, *in prep.*

McQuinn, M. & White, M. 2011, MNRAS, 415, 2257

Meiksin, A. A. 2009, Reviews of Modern Physics, 81, 1405

Morgan, J. S., Burgett, W., & Teran, J. U. 2008, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7012

Mortlock, D. J. et al. 2012, MNRAS, 419, 390

Mountrichas, G. et al. 2009, MNRAS, 394, 2050

Myers, A. D. et al. 2006, ApJ, 638, 622

—. 2007a, ApJ, 658, 85

—. 2007b, ApJ, 658, 99

—. 2008, ApJ, 678, 635

Oke, J. B. & Gunn, J. E. 1983, ApJ, 266, 713

Padmanabhan, N., White, M., Norberg, P., & Porciani, C. 2009, MNRAS, 397, 1862

Padmanabhan, N. et al. 2008, ApJ, 674, 1217

Palanque-Delabrouille, N. et al. 2011, Astron. & Astrophys., 530, A122

Peth, M. A., Ross, N. P., & Schneider, D. P. 2011, AJ, 141, 105

Pier, J. R. et al. 2003, AJ, 125, 1559

Porciani, C., Magliocchetti, M., & Norberg, P. 2004, MNRAS, 355, 1010

Porciani, C. & Norberg, P. 2006, MNRAS, 371, 1824

Prochaska, J. X. & Hennawi, J. F. 2009, ApJ, 690, 1558

Rauch, M. 1998, ARA&A, 36, 267

Richards, G. T. et al. 2002, AJ, 123, 2945

—. 2004, ApJS, 155, 257

—. 2006, AJ, 131, 2766

—. 2009a, ApJS, 180, 67

—. 2009b, AJ, 137, 3884

Richards, J. W. et al. 2011, ApJ, 733, 10

Riegel, R., Gray, A., & Richards, G. 2008, in Proceedings of the 2008 SIAM International Conference on Data Mining, 208

Ross, N. P. et al. 2009, ApJ, 697, 1634

—. 2012, ApJS, 199, 3

Sandage, A. 1965, ApJ, 141, 1560

Sarajedini, V. L. et al. 2011, ApJ, 731, 97

Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, ApJ, 500, 525

Schmidt, K. B., Marshall, P. J., Rix, H. f., Jester, S., Hennawi, J. F., & Dobler, G. 2010, ApJ, 714, 1194

Schmidt, M. 1963, Nat, 197, 1040

Schneider, D. P. et al. 2007, AJ, 134, 102

—. 2010, AJ, 139, 2360

Scranton, R. et al. 2005a, ApJ, 633, 589

—. 2005b, ArXiv Astrophysics e-prints

Sesar, B. et al. 2007, AJ, 134, 2236

Shankar, F. 2009, New A Rev., 53, 57

Sharp, R. G. et al. 2002, MNRAS, 337, 1153

Shen, Y. 2009, PhD thesis, Princeton University

Shen, Y. et al. 2007, AJ, 133, 2222

—. 2009, ApJ, 697, 1656

—. 2010, ApJ, 719, 1693

—. 2012, *in prep.*

Sivia, D. & Skilling, J. 2006, Data analysis: a Bayesian tutorial, Oxford science publications (Oxford University Press)

Slosar, A. et al. 2011, J. Cosmology Astropart. Phys., 9, 1

Smail, I. et al. 2008, MNRAS, 389, 407

Smith, J. A. et al. 2002, AJ, 123, 2121

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

Tucker, D. L. et al. 2006, Astronomische Nachrichten, 327, 821

Vanden Berk, D. E. et al. 2001, AJ, 122, 549

Vestergaard, M. & Peterson, B. M. 2006, ApJ, 641, 689

Volonteri, M., Haardt, F., & Madau, P. 2003, ApJ, 582, 559

Warren, S. J., Hewett, P. C., & Foltz, C. B. 2000, MNRAS, 312, 827

Weymann, R. J., Carswell, R. F., & Smith, M. G. 1981, ARA&A, 19, 41

Weymann, R. J., Morris, S. L., Foltz, C. B., & Hewett, P. C. 1991, ApJ, 373, 23

White, M. 2003, in The Davis Meeting On Cosmic Inflation

Willott, C. J. et al. 2010, AJ, 139, 906

Worseck, G. & Prochaska, J. X. 2011, ApJ, 728, 23

Wright, E. L. et al. 2010, AJ, 140, 1868

Wu, X. & Jia, Z. 2010, MNRAS, 406, 1583

Wyithe, J. S. B. & Loeb, A. 2003, ApJ, 595, 614

Yèche, C. et al. 2010, Astron. & Astrophys., 523, A14+

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

Appendix A

SDSS Coadded Redshift Distributions from Eric Huff

In Chapter 5 I discuss different redshift distribution fits used to calculate $\langle f(\chi) \rangle$ for the CS82 galaxy redshift distribution. One of the fits used is the SDSS Coadded redshift distributions from Eric Huff Huff et al. (2011). Below are some data tables for the raw data from Huff.

Table A.1. Huff's Coadd Redshift Distributions

Redshift (photo-z)	dN (Normalized)	Redshift (photo-z)	dN (Normalized)
Magnitude cut:	**< 21.5**	**Magnitude cut:**	**< 22.0**
0.0	0.0	0.0	0.0
0.0394737	0.0379943	0.0394737	0.0355229
0.118421	0.0802963	0.118421	0.0725315
0.197368	0.0924514	0.197368	0.0824158
0.276316	0.113122	0.276316	0.100390
0.355263	0.106076	0.355263	0.0971328
0.434211	0.131585	0.434211	0.125494
0.513158	0.110114	0.513158	0.104105
0.592105	0.0992236	0.592105	0.0995395
0.671053	0.0650341	0.671053	0.0694157
0.750000	0.0607802	0.750000	0.0704682
0.828947	0.0446999	0.828947	0.0567152
0.907895	0.0194636	0.907895	0.0302315
0.986842	0.0177599	0.986842	0.0254004
1.06579	0.00988198	1.06579	0.0136944
1.14474	0.00422623	1.14474	0.00595355
1.22368	0.00296176	1.22368	0.00426750
1.30263	0.00208903	1.30263	0.00247040
1.38158	0.00181843	1.38158	0.00320377
1.46053	9.20902E-06	1.46053	0.000149467

Note. — Eric Huffs redshift distributions used to calculate $\langle f(\chi) \rangle$ for the CS82 galaxy redshift distribution

Appendix B

CS82 Redshift Distributions from Hendrik Hildebrandt

In Chapter 5 I discuss different redshift distribution fits used to calculate $\langle f(\chi) \rangle$ for the CS82 galaxy redshift distribution. One of the fits used is the CS82 Redshift Distributions from Hendrik Hildebrandt. Below are some data tables for the raw data from Hildebrandt.

Table B.1. Hildebrandt's CS82 Redshift Distributions

Redshift (photo-z)	dN (Normalized)
Magnitude cut:	**< 23.5**
0.0	0.0
0.08	2.627167072792613647e-02
0.13	3.134399509597542155e-02
0.18	3.529441409043507227e-02
0.23	3.870911392771612253e-02
0.28	4.300697810023142859e-02
0.33	4.547589529849720147e-02
0.38	4.382703621477411493e-02
0.43	4.612589672159694842e-02
0.48	4.873126363788367332e-02
0.53	4.824104089894157982e-02
0.58	5.094846918254932605e-02
0.63	5.164955866881062313e-02
0.68	5.343783400910655929e-02
0.73	5.535582916538753923e-02
0.78	5.184191591272272609e-02
0.83	4.553532800235304129e-02
0.88	3.796597880106748063e-02
0.93	3.153478676737584557e-02
0.98	2.640665009325228810e-02
1.03	2.199739336197466949e-02
1.08	1.803833677855539233e-02
1.13	1.496303561541872940e-02
1.23	1.117727293201613813e-02
1.28	1.003514352373161174e-02
1.33	9.504021877169642127e-03
1.38	9.048931557387797017e-03
1.43	8.151680663174594790e-03
1.48	7.065283285840166100e-03
1.53	5.977949074098860491e-03
1.58	4.942460079061231602e-03
1.63	3.958382586506919293e-03
1.68	3.159450839359128488e-03
1.73	2.631441988049393159e-03
1.78	2.291512369212071396e-03
1.83	2.042756280698197063e-03
1.88	1.849573886020138740e-03
1.93	1.694132308656104684e-03

Note. — Hendrik Hildebrandt redshift distributions used to calculate $\langle f(\chi) \rangle$ for the CS82 galaxy redshift distribution