The Extremely Luminous Quasar Survey in the SDSS Footprint. I. Infrared-based Candidate Selection

Jan-Torge Schindler1, Xiaohui Fan1, Ian D. McGreer1, Qian Yang2,3, Jin Wu2,3, Linhua Jiang2, and Richard Green1

1 Steward Observatory, University of Arizona, 933 North Cherry Avenue, Tucson, AZ 85721, USA
2 Kavli Institute for Astronomy and Astrophysics, Peking University, Beijing 100871, China
3 Department of Astronomy, School of Physics, Peking University, Beijing 100871, China

Received 2017 September 14; revised 2017 October 25; accepted 2017 November 6; published 2017 December 6

Abstract

Studies of the most luminous quasars at high redshift directly probe the evolution of the most massive black holes in the early universe and their connection to massive galaxy formation. However, extremely luminous quasars at high redshift are very rare objects. Only wide-area surveys have a chance to constrain their population. The Sloan Digital Sky Survey (SDSS) has so far provided the most widely adopted measurements of the quasar luminosity function at $z > 3$. However, a careful re-examination of the SDSS quasar sample revealed that the SDSS quasar selection is in fact missing a significant fraction of $z \gtrsim 3$ quasars at the brightest end. We identified the purely optical-color selection of SDSS, where quasars at these redshifts are strongly contaminated by late-type dwarfs, and the spectroscopic incompleteness of the SDSS footprint as the main reasons. Therefore, we designed the Extremely Luminous Quasar Survey (ELQS), based on a novel near-infrared JK/W2 color cut using Wide-field Infrared Survey Explorer mission (WISE) AllWISE and 2MASS all-sky photometry, to yield high completeness for very bright ($m_i < 18.0$) quasars in the redshift range of $3.0 \leq z \leq 5.0$. It effectively uses random forest machine-learning algorithms on SDSS and WISE photometry for quasar–star classification and photometric redshift estimation. The ELQS will spectroscopically follow-up ~230 new quasar candidates in an area of ~12,000 deg$^2$ in the SDSS footprint to obtain a well-defined and complete quasar sample for an accurate measurement of the bright-end quasar luminosity function (QLF) at $3.0 \leq z \leq 5.0$. In this paper, we present the quasar selection algorithm and the quasar candidate catalog.

Key words: galaxies: nuclei – quasars: general

1. Introduction

Quasars are the most luminous nontransient light sources in the universe. Powered by the accretion onto supermassive black holes (SMBHs) in the centers of galaxies, they provide important probes for the formation and evolution of structure in the universe up to the highest redshifts. As strong light beacons, their emission traverses the intergalactic medium on their way to us and allows us to study its properties and evolution. Furthermore, quasars produce large quantities of ionizing photons that drive the He reionization of the universe (e.g., Haiman & Loeb 1998; Madau et al. 1999; Miralda-Escudé et al. 2000). The discovery of luminous quasars 0.8 Gyr after the Big Bang (Mortlock et al. 2011) places strong constraints on the formation and growth mechanisms of SMBHs.

A fundamental probe of the growth and evolution of SMBHs over cosmic time is the quasar luminosity function (QLF). It is a measure of the spatial number density of quasars as a function of magnitude (or luminosity) and redshift. From $z = 0$ on, the number density of quasars increases (Schmidt 1968) up to the peak epoch of quasar activity at redshifts around $z = 2 – 3$. At redshifts beyond $z \approx 3$, the quasar number density is found to decline strongly (e.g., Schmidt et al. 1995; Fan et al. 2001; Richards et al. 2006; Ross et al. 2013). The QLF is best fit with a broken power law (Boyle et al. 1988, 2000; Pei 1995) with a steep power-law slope at high luminosities and a flatter power-law slope toward lower luminosities. The slopes, the break point, and the overall normalization are known to evolve with redshift and may provide insight into the physical mechanism of BH growth across cosmic time and the structure formation of the universe. For instance, the ratio of bright quasars to faint quasars decreases from redshift $z \approx 4$ to $z \approx 1$, an indication that the brighter quasars, associated with the more massive black holes, finish their evolution first (Ueda et al. 2003; Marconi et al. 2004; Labita et al. 2009).

Investigations into the evolution of the QLF have previously found the bright-end slope to be flattening (Koo & Kron 1988; Schmidt et al. 1995; Fan et al. 2001; Richards et al. 2006). However, more recent measurements have now established that it remains steep up to redshifts of $z \sim 6$ (Jiang et al. 2008; Croom et al. 2009; Willott et al. 2010; McGreer et al. 2013; Yang et al. 2016).

The difficulty in measuring the bright-end QLF is partly due to the rapid decrease in spatial density of quasars toward high luminosities and the overall decline of their number density toward higher redshifts. Therefore, only wide-area surveys allow for reliable measurements of the bright-end slope.

In addition, purely optical color selections are biased against certain redshift ranges (Richards et al. 2006; Ross et al. 2013). While this bias was well accounted for in previous calculations of the QLF, the low number of quasars at the brightest end does not allow for secure statistics.

Including its latest phase, the Sloan Digital Sky Survey (SDSS; York et al. 2000) has covered 14,555 deg$^2$ in the northern hemisphere with five-band (ugriz) optical imaging and extensive spectroscopic follow-up. The first phase of the survey allowed for the discovery of the first $z \gtrsim 5$ quasar (Fan et al. 1999). The quasar surveys of the first and second phases of SDSS led to the DR7 quasar catalog (DR7Q; Schneider et al. 2010), which contains $\geq 105,000$ quasars.
The SDSS-III Baryon Oscillation Spectroscopic Survey (BOSS; Eisenstein et al. 2011; Dawson et al. 2013) and the SDSS-IV extended Baryon Oscillation Spectroscopic Survey (eBOSS; Dawson et al. 2016) have carried out dedicated spectroscopic follow-up of galaxies and quasars up to $z \approx 3$. The DR14 quasar catalog (DR14Q; I. Paris et al. 2017, in preparation; see also Páris et al. 2017 for DR12Q), the largest quasar sample to date, now includes more than 500,000 known quasars.

However, we have discovered that the SDSS quasar surveys have missed many bright (SDSS $m_1 < 18.5$) higher-redshift ($3.0 < z < 5.0$) quasars. The optically based color selection of quasar candidates in SDSS, BOSS, and eBOSS has relatively low completeness in redshift regions, where the stellar locus overlaps with quasars in optical color space (Richards et al. 2006; Ross et al. 2013). Furthermore, $z > 3$ quasars are not explicitly targeted in BOSS and eBOSS, and the spectroscopic follow-up of SDSS-I/II is not fully complete, especially in the fall sky (R.A. $> 270$ deg and R.A. $< 90$ deg) of the SDSS footprint. As a result, a substantial fraction of bright $3.0 < z < 5.0$ quasars were missed in previous estimations of the QLF.

The power of near- to mid-infrared (mid-IR) photometry to comprehensively select quasars that are otherwise indistinguishable from stars in optical bands was exploited with the advent of large infrared surveys. New quasar selection methods for comprehensively select quasars that are otherwise indistinguishable from stars in optical bands was exploited with the advent of large infrared surveys. New quasar selection methods were developed using the infrared $K$-band excess in the UKIRT (UK Infrared Telescope) Infrared Deep Sky Survey (UKIDSS; Warren et al. 2000; Hewett et al. 2006; Maddox et al. 2008) to efficiently separate quasars and stars at lower ($z < 3$; Chiu et al. 2007) and higher ($z > 6$; Hewett et al. 2006) redshifts. A range of efforts combined optical and near-infrared (near-IR) photometry. Barkhouse & Hall (2001) used the Two Micron All Sky Survey (2MASS; Skrutskie et al. 2006) and optical photometry from the Veron-Cetty & Veron catalog (Veron-Cetty & Veron 2000), whereas Wu & Jia (2010) and Wu et al. (2011) combined SDSS and UKIDSS photometry. In the mid-IR, WISE (Wright et al. 2010) provided deep photometry to further increase the efficiency of $z < 3.2$ quasar selections (Wu et al. 2012).

We designed the Extremely Luminous Quasar Survey (ELQS) to re-examine the SDSS footprint. This paper (Paper I) motivates the survey and outlines our candidate selection. A follow-up paper (Paper II) will contain the spectroscopic results of the ELQS spring sample along with a first estimate of the bright-end QLF. At completion of the survey, we will summarize all spectroscopic discoveries, calculate the bright-end QLF over the entire surveyed footprint ($\sim 12,000$ deg$^2$), and discuss the resulting implications (Paper III).

We first describe the photometric data used for our candidates selection (Section 2). Thereafter, we carefully analyze why previous SDSS surveys missed bright higher-redshift quasars (Section 3). We further develop a solution to the incomplete selection via near-IR/infrared photometry and discuss the rejection of extended objects in Section 4. We continue to describe in detail how we employ machine-learning algorithms to classify quasar candidates and obtain redshift estimates in Section 5. At last, we present the construction of the ELQS quasar candidate catalog (Section 6) and then summarize the results of this paper in Section 7.

All magnitudes are displayed in the AB system (Oke & Gunn 1983) and corrected for galactic extinction (Schlafly & Finkbeiner 2011), unless otherwise noted. We adopt the standard flat $\Lambda$CDM cosmology with $H_0 = 70$ km s$^{-1}$ Mpc$^{-1}$, $\Omega_m = 0.3$, and $\Omega_k = 0.7$, generally consistent with recent measurements (Planck Collaboration et al. 2016).

2. Photometry

For our quasar candidate selection, we use a combination of near-IR 2MASS and mid-IR WISE photometry, complemented with optical photometry from SDSS.

2.1. The Sloan Digital Sky Survey (SDSS)

For all SDSS sources, we use the point-spread function asinh magnitudes (Lupton et al. 1999) in the five optical band passes ($ugriz$; Fukugita et al. 1996). Throughout this paper, we only use AB magnitudes. We therefore converted the SDSS $u$-band and $z$-band magnitudes with $m_{AB} = u' - 0.04$ mag and $z_{AB} = \prime + 0.02$ mag. All magnitudes are corrected for galactic extinction using the extinction values from the Casjobs Data Release 13 (DR13) PhotoObjAll or PhotoPrimary tables (Schlafly & Finkbeiner 2011). The imaging data of the SDSS survey were completed in 2009 and cover a unique area of 14,555 deg$^2$. The magnitude limits (95% completeness for point sources) in the five optical bands are 21.6, 22.2, 22.2, 21.3, and 20.7 for $u$, $g$, $r$, $i$, and $z$, respectively.

2.2. The Wide-field Infrared Survey Explorer (WISE)

WISE mapped the entire sky at 3.4, 4.6, 12, and 22 μm ($W_1$, $W_2$, $W_3$, and $W_4$). The recent AllWISE data release combines the data from the cryogenic and post-cryogenic (Mainzer et al. 2011) phases of the mission. The AllWISE source catalog achieved 95% photometric completeness for all sources with limiting magnitudes brighter than 19.8 and 19.0 (Vega: 17.1, 15.7) in $W_1$ and $W_2$. Saturation affects sources brighter than 8, 7 in the $W_1$ and $W_2$ bands. We restrict ourselves to photometry of the $W_1$ (3.4 μm) and $W_2$ (4.6 μm) infrared bands and convert them to AB magnitudes using $W_{1AB} = W_1 + 2.699$ and $W_{2AB} = W_2 + 3.339$. The WISE photometry is then extinction-corrected using the extinction coefficients $A_{W1}, A_{W2} = 0.189, 0.146$ with the extinction values from the SDSS photometry.

2.3. The Two Micron All Sky Survey (2MASS)

2MASS has mapped the entire sky in the near-IR bands $J$ (1.25 μm), $H$ (1.65 μm), and $K_s$ (2.17 μm). We use the 2MASS point source catalog (PSC), which was preparsed to the WISE AllWISE source catalog. The 2MASS PSC is generally complete at a level of 10σ photometric sensitivity for all sources brighter than 16.7, 16.4, and 16.1 (Vega: 15.8, 15.0, 14.3) in the $J$, $H$, and $K_s$ bands, respectively. Yet, the 2MASS PSC includes all sources with at least a signal-to-noise ratio of $S/N \geq 7$ in one band or $S/N \geq 5$ detections in all three bands. Furthermore, due to the confusion of sources close to the galactic plane, the photometric sensitivity is a strong function of galactic latitude. Based on the online documentation, we estimate the limiting magnitudes of the 2MASS PSC for higher latitudes to be $J = 17.7$, $H = 17.5$, and $K_s = 17.1$. We

---

3. http://irsa.ipac.caltech.edu/cgi-bin/Gator/nph-scan?submit=Select&projshort=WISE
4. Figure 7 on https://www.ipac.caltech.edu/2mass/releases/allsky/doc/sec2_2.html.
Table 1
Number Counts of Quasars Missed in SDSS to the Total Number of Known Quasars in the SDSS-matched MQC

| i-band Magnitude Bins | 14.0 ≤ $m_i$ < 16.0 | 16.0 ≤ $m_i$ < 17.0 | 17.0 ≤ $m_i$ < 17.5 | 17.5 ≤ $m_i$ < 18.0 | 18.0 ≤ $m_i$ < 18.5 | 18.5 ≤ $m_i$ < 19.0 | 19.0 ≤ $m_i$ |
|-----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------|
| Fall sky and 2.5 ≤ $z$ < 5 |
| Missed QSOs | 2 | 4 | 10 | 17 | 34 | 40 | 50 |
| Known QSOs | 3 | 12 | 26 | 123 | 438 | 1080 | 2180 |
| Fraction | 0.67 | 0.44 | 0.31 | 0.15 | 0.07 | 0.03 | 0.02 |
| Spring sky and 2.5 ≤ $z$ < 5 |
| Missed QSOs | 3 | 3 | 13 | 14 | 28 | 43 | 58 |
| Known QSOs | 7 | 33 | 126 | 404 | 1215 | 3014 | 6152 |
| Fraction | 0.43 | 0.09 | 0.10 | 0.03 | 0.02 | 0.01 | 0.01 |

Figure 1. In this figure, we examine the fraction of quasars known in the MQC but missed by SDSS at redshifts of 2.5 ≤ $z$ < 5. Quasars that were identified by SDSS are termed recovered. Left: number distributions of missed and recovered quasars as a function of $i$-band magnitude bins in the SDSS fall footprint. Middle: fractions of missed quasars to the total number of quasars in the MQC as a function of $i$-band magnitude bins, divided into a fall sky and spring sky sample. Right: number distributions of missed and recovered quasars as a function of $i$-band magnitude bins in the SDSS spring footprint. This figure illustrates that the majority of missed quasars are bright ($m_i < 17.5$) and located in the fall sky of the SDSS footprint.

3. Bright High-z Quasars Missed By SDSS

In order to investigate why and to what extent SDSS missed bright high-redshift quasars, we match a compilation of known quasars in the literature, the Million Quasar catalog (hereafter MQC, Version 5.2; Flesch 2015) with SDSS DR14 photometry and the SDSS DR7 and DR14 quasar catalogs. The MQC includes spectroscopically confirmed quasars as well as high-probability quasar candidates from various selection methods. For this exercise, we will exclude all candidates.

We restrict ourselves to brighter quasars with $m_i < 19.5$ and at 2.5 ≤ $z$ < 5. The sample is further divided into a spring (90 deg < R.A. < 270 deg) and fall (R.A. > 270 deg and R.A. < 90 deg) portion of the SDSS footprint. All quasars that have SDSS photometry, but are not included in the DR7 and DR14 quasar catalogs, will be termed missed quasars. We investigate the fraction of missed quasars to all quasars found in MQ in Table 1 and Figure 1. Both clearly illustrate that the fraction of missed quasars is larger in the fall sky than in the spring sky. For 2.5 ≤ $z$ < 5 quasars with $i$-band magnitudes $m_i < 17.0, 17.5, 18.5$, the fraction of missed to known quasars is $f ≈ 0.40, 0.39, 0.11$ in the fall sample and $f ≈ 0.15, 0.11, 0.03$ in the spring sample. These numbers demonstrate the poorer spectroscopic completeness of the SDSS survey in the fall sky. Furthermore, the fractions increase as we restrict ourselves to brighter $i$-band magnitudes, emphasizing that especially bright quasars were missed.

The majority (>90%) of the SDSS and BOSS quasars at 2.5 ≤ $z$ < 5 were observed during the BOSS campaign. In the SDSS and BOSS five-band optical color space, quasars at $z = 2.5–3.5$ overlap significantly with the stellar locus. The survey completeness of SDSS (Richards et al. 2006, their Figure 6) and BOSS (Ross et al. 2013, their Figure 6) show the low completeness in those regions. Furthermore, as the candidates get brighter, the contamination of the stellar population increases, as demonstrated by a steep decrease in completeness toward brighter magnitudes in Ross et al. (2013, their Figure 6).

...
To illustrate the confusion between quasars at $z = 2.5–3.5$ and stars, we show the ugriz color-space diagrams in Figure 2. The stellar locus is shown with the black contours, while we display the missed quasars with SDSS, 2MASS PSC, and AllWISE photometry as filled circles color-coded by redshift. It should be noted that this is a subset of all missed quasars, because some of them are not detected by 2MASS and WISE. In panels (a)–(c), the majority of the missed quasars overlap significantly with the stellar locus. The SDSS inclusion regions for higher-redshift quasars (Richards et al. 2002) also miss some of the $z > 3.5$ quasars that scatter into the stellar locus.

Figure 2. We show the quasars with full SDSS, 2MASS PSC, and AllWISE photometry that SDSS missed as filled circles, colored by redshift in the SDSS optical color space (AB magnitudes) in panels (a)–(c). The numbers reference the missed quasars described in Section 3.1. A significant fraction of the missed quasars are always overlapping with the stellar locus (black contours). We further display the SDSS high-redshift inclusion boxes (ugr in (a), $z \geq 3.0$; gri in (b), $z \geq 3.6$; riz in (c), $z \geq 4.5$). Panel (d) shows the population of missed quasars in the near-IR $J-K_s$ and $K_s-W_2$ color space of 2MASS and WISE. The quasars are well-separated from the stellar locus. We display our new JKW2 color cut as the thick black line.

Fiber collisions during observations and quality criteria on the photometry may have contributed to the number of missed quasars. In addition, quasar lenses can have extended morphologies and will have been likely rejected. In some cases, objects with extremely bright apparent magnitudes ($m_i \lesssim 15.5$) were excluded for spectroscopic follow-up to avoid scattered light from nearby fibers. However, the main reasons SDSS and BOSS missed bright high-redshift quasars are the ugriz optical-based candidate selection and the spectroscopic incompleteness of the fall sky footprint.
3.1. Notes on the Brightest Missed High-redshift Quasars in the SDSS Spring Sky Footprint

We selected all missed quasars with \( m_i < 17.0 \) and \( 2.5 < z < 5 \) that fall into the well-surveyed spring sky of the SDSS footprint to understand why they were missed. We analyze whether the non-fatal and fatal image quality flags and the high-redshift inclusion regions of the original selection (Richards et al. 2002) play a role in them being overlooked. All of the quasars below with \( z \geq 2.8 \) are selected with our selection method and will be included in the ELQS spring sample (Paper II).

3.1.1. Q1208+1011

This object is a known quasar lens (Bahcall et al. 1992; Magain et al. 1992, \( m_i = 16.77 \)) with a redshift of \( z = 3.80 \). One of the non-fatal quality flags is raised, and it is not included in any of the high-redshift inclusion regions. Finally, this quasar is located too close to the stellar locus in the optical-color space (see Figure 2) to be selected by previous SDSS campaigns.

3.1.2. APM 08279+5255

This object is a well-known, lensed broad absorption line quasar with multiple images at \( z = 3.91 \) (Ihata et al. 1999). It has an SDSS \( i \)-band magnitude of \( m_i = 14.84 \). While it does not show any fatal or non-fatal flags, it is also not selected in the high-redshift inclusion regions. The colors of this quasar (see Figure 2) place it well away from the stellar locus in the \( u-g-r \) color space, but it is too bright for the SDSS spectroscopic target list to avoid scattered light from nearby fibers.

3.1.3. SDSS J1622+0702A

This object is a bright \( (m_i = 16.87) \) binary quasar at \( z = 3.26 \) (Hennawi et al. 2010). The fatal and non-fatal flags are not raised, and it actually falls into the \( u-g-r \) inclusion region. This quasar has the correct target flag in SDSS, but it was not observed, because of fiber collisions with a nearby galaxy.

3.1.4. B1422+231

This lensed quasar has a total of four components and is measured to be at \( z = 3.62 \) (Patnaik et al. 1992). It is very bright with \( m_i = 15.31 \). While its image quality flags are not raised and it actually falls into the \( u-g-r \) inclusion region, it is marked as extended (\( \text{type} = 3 \)) by SDSS. However, it would not have been rejected by our selection for its morphology (see Section 4.3).

3.1.5. CSO 167

CSO 167 is a \( z = 2.56 \) quasar (Sanduleak & Pesch 1984; Everett & Wagner 1995) with \( m_i = 16.50 \). None of the fatal or non-fatal image quality flags are raised. It is not expected to and also does not fall into one of the high-redshift inclusion regions. This quasar is too close to the stellar locus (see Figure 2) to be selected in previous SDSS campaigns.

3.1.6. CSO 1061

Similar to CSO 167, CSO 1061 (Sanduleak & Pesch 1989) is at a lower redshift (\( z = 2.67; m_i = 16.33 \)) and therefore is not expected to be selected by any of the inclusion regions. None of the fatal or non-fatal image quality flags are raised. This object was also not observed because it is too close to the stellar locus.

4. Bright Quasar Selection Based on Near-infrared Photometry

Our new survey for extremely luminous quasars, ELQS, is designed to bypass the limitations of a purely optical quasar candidate selection by harnessing the information of near-IR and infrared photometry. For this purpose, we examined the distribution of stars and quasars in the color space of the 2MASS and WISE all-sky surveys. We discovered that the combination of \( J-K \) and \( K_{\text{r}}-W2 \) colors offers a clear separation of stars and quasars and designed a \( JKW2 \) color cut in this color space.

This color cut is very efficient in rejecting stars and in concert with a measure to eliminate galaxies, the quasar fraction should be fairly pure. Nevertheless, the fraction of bright stars, which make the cut, to bright quasars is non-negligible, and the color cut does not discriminate among quasars of different redshifts. In fact, the majority of quasar candidates will be at \( z < 2.5 \) and therefore also reduce the efficiency of our selection. Hence, it is necessary to estimate photometric redshifts and to classify candidates at a later stage (Section 5).

For an early overview, we illustrate the general steps of our quasar candidate selection in Figure 3. They will be described in detail in Section 6.

4.1. The JKW2 Color Cut

While we have shown the overlap between the missed quasars and the stellar locus in optical color space in Figure 2 panels (a)–(c), we also show the 2MASS and WISE \( J-K \), \( K_{\text{r}}-W2 \) color space in panel (d). Here, the stellar locus clearly separates from the distribution of missed quasars. After examining the distribution of known quasars to known stars in this color–color space, we qualitatively determined the color cut that separates these distributions best. In Vega magnitudes, the color cut reads

\[
K_{\text{S Vega}} - W_2\text{Vega} \geq 1.8 - 0.848 \cdot (J_{\text{Vega}} - K_{\text{S Vega}}),
\]

while in AB magnitudes it changes to

\[
K_s - W_2 \geq -0.501 - 0.848 \cdot (J - K_s).
\]

The color cut separates quasars from stars because of their difference in the \( K-W2 \) flux ratio (or \( K-W2 \) color). This flux ratio is fundamentally different in stars (up to spectral type T) and quasars (up to redshifts of \( z \approx 5 \)). For those stars and quasars, the stellar flux declines more strongly than the quasar flux between the \( K \)-band and the \( W2 \)-band. Hence, quasars will have a redder \( K-W2 \) flux ratio (or color). Emission lines, like \( 
H\alpha \) in the \( K \)-band at \( z \approx 2.1-2.50 \) or \( H\beta \) in the \( K \)-band at \( z \approx 3.2-3.8 \), do affect the \( K-W2 \) flux ratio, but their influence is negligible. As a result, the \( JKW2 \) color cut cannot discriminate between quasars at different redshifts.

Six known quasars lie beneath the \( JKW2 \) color cut within the stellar locus in Figure 2, panel (d). No spectra were available for any of these objects in the literature. A closer examination of their identifications and photometry concluded that either the photometry or the identification reference seems to be unreliable.
We test the color cut in a $\sim 70 \text{deg}^2$ region (R.A. = 120 – 130 deg, decl. = 40 – 50 deg), shown in Figure 4. For this purpose, we selected all sources in SDSS DR13 (PhotoObjAll) brighter than SDSS $m_i = 18.5$, matched the sources with the AllWISE source catalog, including matched 2MASS PSC photometry, and retrieved spectral identification from the SDSS DR13 (SpecObj) catalog, where possible.

The full test field includes a total of 1,327,439 sources detected with full SDSS photometry, of which 1,155,203 sources have full 2MASS and WISE $W1$ and $W2$ photometry.

For our purposes, we estimate the efficiency of the color cut using the knowledge about the spectroscopically identified quasars (at all redshifts), stars, and galaxies, and the total number of sources with 2MASS and WISE $W1$ and $W2$ colors. The results are summarized in Table 2.

The color cut clearly separates stars and quasars. It reduces the total number of sources in the test region to 25,470 out of 1,155,203 (2.0%). The stellar contamination is greatly reduced. Only 0.37% (8/2138) of spectroscopically identified stars in SDSS DR13 make the color cut. Of all 526 quasars in this region, 301 have 2MASS and WISE $W1$ and $W2$ colors, and of these, 298 are included in the JKW2 color cut.

Considering only the quasars with full near-IR photometry, the color cut has an estimated completeness of $\sim 99\%$.

The quasar sample in the $70 \text{deg}^2$ test region is rather small. Therefore, we test our color cut with all quasars from the DR7 and DR12 quasar catalogs that have 2MASS photometry in the $J$ and $K_s$ bands and WISE $W2$ photometry. This sample includes 5945 quasars, out of which 5927 are included in the color cut, a fraction of more than 99\%.

Galaxies, however, straddle the JKW2 color cut and only half of all spectroscopically identified galaxies (2706/5517) are excluded. Therefore, we believe the majority of the remaining 25,470 sources above the color cut to be galaxies.

In total, the JKW2 color cut manages to eliminate the majority of stellar sources while retaining a highly complete sample of bright quasars uniform in redshift and SDSS $i$-band magnitude. However, in order to efficiently use the color cut, we need, first, to reject galaxy contaminants and second, to estimate the photometric quasar redshifts.

4.2. Photometric Completeness of 2MASS and WISE

The 2MASS and WISE surveys do not reach the depth of optical surveys, like SDSS. Therefore, fainter sources might only show spurious detections or no detections at all. Even for a bright quasar survey, this is a concern regarding the photometric completeness. We analyze the fraction of sources in the MQC without detections in the 2MASS and WISE bands necessary for the JKW2 color cut. More than 99.5\% of $m_i < 18.0$ quasars at $z > 2.5$ in the SDSS-matched MQC are detected with $S/N > 5$ in the $W1$ and $W2$ bands. 2MASS is shallower but still allows for about $80\%$ of $m_i < 18.0$ quasars at $z > 2.5$ in the SDSS-matched MQC to be detected in $J$ and $K_s$. The photometric completeness of all surveys will be taken account when we calculate the overall survey completeness in Paper II.
4.3. Galaxy Rejection Using the Petrosian Radius

The SDSS survey uses the type parameter to differentiate between different source types. It allows point sources (type = 6) and sources that are classified as extended (type = 3) to be distinguished from one another. However, we do not generally want to exclude quasar lenses in our selection, and thus do not use this parameter. Instead, we turn to a more quantitative measure, the Petrosian radius.

At redshifts above $z = 2.5$, even lensed quasars will be reasonably compact (petroRad$_i < 2''$) if they are not strongly distorted, but might still be categorized as extended objects (SDSS flag type = 3; e.g., the lensed quasar B1422+231).

In Figure 5, we show that a Petrosian radius of $2''$ in the SDSS i-band strongly reduces the number of galaxies in the total test region while mainly reducing the number of quasars between $0 \leq z \leq 1.0$, where its host is likely resolved in the SDSS photometry. In the targeted redshift range of $2.5 \leq z \leq 4.0$, this cut on the Petrosian radius in the SDSS i-band does not reject any quasars.

We summarize the effects of the cuts in the Petrosian radius on the 70 deg$^2$ test region (Figure 4) in Table 3. The limit on the Petrosian radius reduces the number of spectroscopically identified quasars from 301 to 207, out of which 205 (99%) are within the JKW2 color cut. As expected, the Petrosian radius limit does not significantly alter the number of stars within the color cut. However, it reduces the total number of spectroscopically identified galaxies that make the JKW2 color cut from 2706 to 15, and the total number of sources from 25,470 to 1296. The majority of the rejected sources will be galaxies with a contribution from low-redshift ($z \leq 1.0$) quasars.

5. Photometric Redshift Estimation and Quasar–Star Classification

The JKW2 color cut is very efficient in rejecting stellar sources, and the additional limitation on the Petrosian radius excludes the majority of galaxies. However, we currently have no measure to exclude low-redshift quasars, since the JKW2 color cut does select quasars independent of their redshift. About $\sim 95\%$ of the DR7 and DR12 quasars that make the JKW2 color cut are at lower redshifts. Although this distribution is biased due to the selection of SDSS and BOSS quasar candidates, it still shows us that our sample will suffer from a large quantity of lower-redshift quasars. In addition, bright quasars ($m_i \leq 18.5$) would make up less than half of the sources selected by these two criteria, the other half still being stellar contaminants.

Hence, we decided to use the SDSS, 2MASS, and WISE photometry to further estimate their photometric redshifts and also to classify quasar candidates. Our methods of choice here are two supervised machine-learning algorithms, support vector machines (SVMs; Vapnik 1995; Burges 1998; Vapnik 1998), and random forests (Breiman 2001).

While the amount of galaxies included in the JKW2 color cut might seem non-negligible, the visual inspection of photometric cutouts efficiently reduces them further, and our spectroscopic observations show that they are insignificant contaminants. Consequently, we decided not to include them in the classification process.

We will first explain the construction of the training sets that both algorithms will rely on, then introduce the methods themselves and finally discuss the results.

5.1. Training Sets

5.1.1. The Empirical Quasar Catalog

In order to devise training sets for the classification and photometric redshift estimation, we combined the SDSS quasar catalogs with 2MASS All-Sky and WISE AllWISE photometry.

We select all quasars from the SDSS DR7 and DR12 quasar catalogs (DR7Q and DR12Q, respectively) and matched them with SDSS DR13 (PhotoObjAll) sources in a 2$''$ radius to obtain additional information about the sources. The resulting...
catalog is then matched to WISE AllWISE and 2MASS All-Sky photometry in a 1° radius to make sure that the matches are accurate. We exclude magnitude outliers (2 < i and i > 29; i = apparent i-band magnitude) from the catalog. The magnitudes are then corrected for galactic extinction (m_i = extinction-corrected apparent i-band magnitude). Fluxes and flux errors (where possible) are then calculated from the extinction-corrected magnitudes. To ensure that the magnitude error distributions are free of outliers, objects with positive values in the SDSS flags INTERP and DEBLEND_AT_EDGE are excluded.

The resulting quasar catalog is termed the “empirical” quasar catalog and has a total of 215,087 objects with full SDSS photometry. An overview of the different sample sizes drawn from this catalog is given in Table 4.

The column “photometry” refers to the photometric information included in this training set. Not all sources in the full empirical quasar catalog have information in the 2MASS or WISE filter bands because the 2MASS and WISE surveys do not always reach the same depth as the SDSS photometry. If we mandate values in the WISE W1 and W2 bands or the 2MASS bands and their corresponding errors, the size of the training set decreases.

One additional question that naturally arises is: Can we find a bright population of quasars that does not exist in the training set? The distribution of the brightest quasars does not have inherently different spectral slopes. As a consequence, their flux ratios (or colors) do not differ substantially from less brighter ones. Since the main features used in the machine-learning methods are the flux ratios of adjacent photometric bands, bright quasars should be comprehensively selected without problems.

5.1.2. The Empirical Star Catalog

To construct a catalog of stars, we restricted our sample to only spectroscopically classified stars in the SDSS footprint. Some of those spectroscopic validations were a result of quasar selection and as such, our star catalog may be biased toward stellar classes that can be confused with quasars. However, this works to our advantage as common quasar contaminants will be overrepresented in the sample.

Using the Casjobs6 interface, our query automatically added SDSS photometry to the 2MASS PSC and the WISE AllWISE catalogs where objects were prematched. To ensure good-quality photometry in the SDSS bands, we used the following quality flags: WARNING = 0, INTERP = 0, BINNED1 = 0, not (EDGE, NOPROFILE, NOTCHECKED, PSF_FLUX_INTERP, SATURATED), BAD_COUNTS_ERROR = 0, DEBLEND_NOPEAK = 0, INTERP_CENTER = 0, COSMIC_RAY = 0.

Table 3

| Criterion                  | petroRad_i < 3.0 arcsec | petroRad_i < 2.0 arcsec | petroRad_i < 1.5 arcsec |
|----------------------------|-------------------------|-------------------------|-------------------------|
| Full sample                | 1,155,203               | 965,887                 | 841,925                 |
| Galaxies                   | 5517                    | 578                     | 39                      |
| Quasars                    | 301                     | 257                     | 207                     |
| Stars                      | 2138                    | 2112                    | 2081                    |
| No spectral id.            | 1,147,247               | 962,940                 | 839,598                 |

The Astrophysical Journal, 851:13 (16pp), 2017 December 10

Schindler et al.

Table 4

| Data Set      | Photometry | Constraints | Size     |
|---------------|------------|-------------|----------|
| Emp. QSOs     | SDSS       | m_i < 18.5  | 215,087  |
| Emp. QSOs     | SDSS+W1W2  | m_i < 18.5  | 12,408   |
| Emp. QSOs     | SDSS+W2    | m_i < 18.5  | 153,890  |
| Emp. QSOs     | SDSS+2MASS+ W1W2 | m_i < 18.5 | 12,388   |
| Emp. QSOs     | SDSS+2MASS+ W2 | m_i < 18.5 | 4815     |
| Emp. QSOs     | SDSS+2MASS+ W1W2 | m_i < 18.5 | 4021     |
| Emp. QSOs     | SDSS+2MASS+ W2 | m_i < 18.5 | 4015     |
| DR13 Stars    | SDSS       | m_i < 18.5  | 387,854  |
| DR13 Stars    | SDSS+W1W2  | m_i < 18.5  | 219,375  |
| DR13 Stars    | SDSS+W2    | m_i < 18.5  | 245,326  |
| DR13 Stars    | SDSS+2MASS+ W1W2 | m_i < 18.5 | 197,798  |
| DR13 Stars    | SDSS+2MASS+ W2 | m_i < 18.5 | 174,218  |
| DR13 Stars    | SDSS+2MASS+ W1W2 | m_i < 18.5 | 159,211  |
| DR13 Stars    | SDSS+2MASS+ W2 | m_i < 18.5 | 209      |

Table 5

| Spectral Class | # of Objects |
|----------------|--------------|
| O              | 547          |
| OB             | 533          |
| B              | 4638         |
| A              | 47,714       |
| F              | 158,233      |
| G              | 29,953       |
| K              | 60,233       |
| M              | 64,691       |
| L              | 1329         |
| T              | 172          |
| WD             | 15,782       |
| CV             | 3363         |
| Carbon         | 321          |

The resulting catalog contains a total of 387,854 objects that have full SDSS photometric information. A detailed list of the size of the different subsamples, which have more photometric information, is given in Table 4.

For the purpose of quasar–star classification, we summarized the number of stellar subclasses of SDSS DR13 (SpecObj) into the following stellar classes: O, OB, B, A, F, G, K, M, L, T, WD, CV, and Carbon. The number of objects per class is shown in Table 5.

5.2. Introduction to Random Forest Methods

The random forest method (Breiman 2001) uses an ensemble of decision (or regression) trees to vote for the most popular…

6 http://skyserver.sdss.org/casjobs/
class (value) regarding a classification (regression) problem. The method is a supervised machine-learning method. Therefore, it requires a training set to learn from. Furthermore, random forests are non-parametric, inherently allow for multiple (>2) classes, and avoid the problem of overfitting.

Since random forests rely on decision (or regression) trees, we will shortly introduce their operation.

The features of the training set, which are fluxes and flux ratios for our purposes, generate the multidimensional input space for the classification or regression. A decision tree divides this input space into cuboid regions along the feature axis. Each of these regions contains one or more data points that determine the target class or target value for that cuboid.

This structure is built to correspond to a binary tree, where at each node a decision on one input feature is made on how to split the tree into two branches. This is called recursive binary partitioning. The decision is reached using a criterion that encourages the formation of regions, where the majority of the data points belong to only one class.

To predict the target class or target value for a new data point, the class or value of the cuboid region that the data point falls into is chosen.

Decision trees are invariant under scaling of the feature values and to the inclusion of irrelevant features. They also produce reviewable models of the data that are easy to visualize and inherently work with multiple (>2) classes.

In the random forest method, each decision tree is fit on a bootstrap subsample of equal size to the original input sample. During the construction of the decision tree, a random subset of all available features is used to determine the best splits for all nodes. In our case the number of random features considered is taken to be the square root of the total number of features. A full decision tree is grown without pruning the tree during the construction. The forest contains a large number of these trees (>100), each of which gives a classification (target value) for every object. The class with the most votes is then chosen to be the final class of the object. For random forest regression, the mean of all target values is calculated to be the final regression value.

While one decision tree is prone to overfitting the training sets, random forests overcome this problem through the averaging process of many randomized decision trees.

Breiman et al. (2003) first proposed this machine-learning technique as an astronomical classification tool to find quasars. In astronomy, the method has since been applied for the classification of variable stars (Dubath et al. 2011; Richards et al. 2011) and variable quasars (Pichara et al. 2012), photometric redshift estimation (mainly aimed at galaxies; Carliles et al. 2010; Carrasco Kind & Brunner 2013), and quasar classification (Carrasco et al. 2015).

We are using the implementation of the random forest classifier and regressor provided by the scikit-learn (Pedregosa et al. 2011) Python library with many of the default parameters.

For the construction of the binary tree, we use the Gini impurity to determine the best split at each step. While the original random forest method (Breiman 2001) lets each classifier vote for the final class, this implementation averages the probabilistic predictions of each classifier to find the final probabilities for each class.

We adopt the hyperparameters that control the size of the decision trees (min_samples_split, max_depth) as well as the size of the forest (n_estimators) to find the best classification/regression model. The best-fit values for these hyperparameters are found using a limited grid search.

5.3. Introduction to SVMs

SVMs (Vapnik 1995, 1998; Burges 1998) offer a sparse kernel method for classification, regression, and novelty detection. Similar to random forests, SVMs belong to supervised machine-learning methods and therefore rely on a well-constructed training set. Fundamentally a two-class classifier, the extension of SVMs to multiclass (>2) classification is problematic.

In the case of classification, the algorithm calculates a decision boundary (hyperplane) to divide the multidimensional feature space into two regions according to the two classes. The decision boundary is constructed to maximize the smallest distance between itself and any of the data samples. In the end, only a small subset of original data points is necessary to define the decision boundary. These data points are the support vectors. One strength of the SVM lies in this reduction of information to only a few data samples that define how to split the feature space. As a result, the SVM is very fast in making predictions.

SVMs as kernel methods use an algorithm that allows for the use of a kernel function to implicitly transform the original feature space into a higher-dimensional feature space. In such an algorithm, the input vector enters only in the form of its scalar product and is then substituted by the kernel. Therefore, the coordinates in the higher-dimensional feature space do not have to be calculated explicitly, only their inner product. The kernel allows for complex decision boundaries that are nonlinear in the original feature space and allow for more complicated distributions of the two classes.

However, many data sets do not have fully separable classes even if nonlinear kernels are used. As a solution, the algorithm allows for training data to lie on the “wrong” side of the decision boundary. These data points are then misclassified but ignored by the algorithm. They are called slack variables and allow for a better generalization of the classification. In older formulations of the SVM algorithm, the amount of slack variables are controlled with the parameter C.

Similar to other regression problems, SVM regression also seeks to minimize a regularized error function. This error function incorporates the previously introduced slack variables as well as an ε-insensitive error function (Vapnik 1995). Therefore, data points within ε of the regression model have no associated error, and the number of data points outside of this region is controlled using the slack variable parameter C.

SVMs have been widely used in astronomy for galaxy (Wadadekar 2005; Wang et al. 2008) and quasar (Han et al. 2016) photometric redshift estimation and source classification (Gao et al. 2008; Huertas-Company et al. 2008; Kim et al. 2012; Peng et al. 2012; Kurcz et al. 2016).

We use the scikit-learn (Pedregosa et al. 2011) implementation of the support vector regression for a comparison of the photometric redshift estimation with the random forest method above. We use the radial basis function kernel (kernel = "rbf"), which adds the hyperparameter gamma. In order to estimate the optimal hyperparameters C, epsilon, and gamma, we carry out limited grid searches.
Furthermore, it is important to note that the features of the input and test vectors need to be normalized because SVMs are not scale invariant.

5.4. Photometric Redshift Estimation

With the JK2W color cut rejecting the majority of stars and the limit on the Petrosian radius filtering out unwanted galaxies, lower-redshift quasars become major contaminants for our survey targeted at \( z \geq 2.8 \) quasars. Hence, it is critical to estimate the quasar redshift and use this photometric redshift as a criterion for our candidate selection. Instead of relying on optical-color cuts (e.g., Richards et al. 2002), we aim to utilize the full photometric information given with SVM regression (SVR) and random forest (RF) regression.

The training sets for both algorithms are drawn from the empirical quasar catalog, which is based purely on the SDSS DR7 and DR12 quasar catalogs, as described above.

We use three different training sets built from the empirical quasar data to test the effects of different feature sets and the size of the training sets on the regression outcome. The first (SDSS; Table 6, rows 1 and 5) includes all sources with full SDSS photometry, the second (SDSS+1W2; Table 6, rows 3 and 7) includes \( W1 \) and \( W2 \) information in addition to full SDSS photometry, whereas the last (SDSS+1W2, \( m_i < 18.5 \); Table 6, rows 4 and 8) uses only sources with SDSS \( i \)-band magnitude brighter than \( m_i > 18.5 \) of the SDSS+1W2 subset. In general, it is advisable to always include the largest amount of information possible. However, because we want to evaluate the benefit of including the \( W1 \) and \( W2 \) bands in addition to the SDSS photometry, we limit ourselves to the SDSS photometry for a test case on the SDSS+1W2 subset (Table 6, rows 2 and 6).

We do not build a training set based on full SDSS, 2MASS, and \( W1 \) and \( W2 \) photometry because the number of quasars with full 2MASS photometry is too small (~1000) to allow for sufficient training in the large feature space. In addition, a cut on the \( i \)-band magnitude at \( m_i < 18.5 \) strongly reduces the number of higher-redshift quasars in the training set. As a consequence, the feature space is not sufficiently populated at higher redshifts, and the method is biased against those quasars.

For training sets with only SDSS features, we use the four adjacent flux ratios \((u/g, g/r, r/i, i/z)\) from the five photometric bands and the SDSS \( i \)-band magnitude. When we include WISE photometry (SDSS+1W2), we expand the flux ratios accordingly \((+z/W1, W1/W2)\) and add the \( W1 \) magnitude to the feature set.

For each subset, we use a grid search on the hyperparameters of the random forest and SVM regression to determine the best regression model. We calculate the regression results for all subsets using both machine-learning methods to compare them to each other.

The grid for the random forest regression has the following hyperparameters: \( n_estimators = [50, 100, 200, 300], \) \( min_samples_split = [2, 3, 4], \) and \( max_depth = [15, 20, 25] \) (36 combinations).

For the SVM regression, we use a grid of \( C = [10, 1.0, 0.1], \) \( gamma = [0.01, 0.1, 1.0], \) \( epsilon = [0.1, 0.2, 0.3] \) (27 combinations).

We use five-fold cross-validation on the full training set and always train on 80% of the full training set, while the remaining 20% are used to test the regression. Before we continue to discuss the results of the calculations, we will first introduce the typical regression metrics.

5.4.1. Regression Metrics

To evaluate the success of the photometric redshift estimation methods, we use the standard \( R^2 \) regression score as well as the residual between the photometric and spectroscopic redshifts, \( \Delta z = \hat{z}_{\text{spec}} - \hat{z}_{\text{spec}}. \)

The \( R^2 \) score, also called the coefficient of determination, gives a measure of the goodness of fit of a model. Let us assume that the true redshifts are denoted by \( z_i \) and the predicted redshift values are denoted by \( \hat{z}_i \). The \( R^2 \) score is then calculated from the total sum of squares \( SS_{\text{tot}} \) and the residual sum of squares \( SS_{\text{res}} \), where \( \hat{z} \) is the mean redshift, according to

\[
SS_{\text{tot}} = \sum_i (z_i - \bar{z})^2, \quad SS_{\text{res}} = \sum_i (z_i - \hat{z}_i)^2, \quad R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}.}
\]

In the best-case scenario, the residual sum of squares will go to 0 and the \( R^2 \) score will reach 1. If the model always predicts the mean value of the training data \( \bar{z} \), then it results in an \( R^2 \) score of 0. For bad models, it is possible that the \( R^2 \) reaches negative values. We use \( R^2 \) as a measure to compare models with one another to find the model that best represents the data. Therefore, we focus on the relative values of the different models and do not interpret the absolute one.

In the literature, the goodness of the photometric redshift estimation is often measured by the fraction of test quasars \( N \) with absolute redshift residuals \( |\Delta z| = |z_i - \hat{z}_i| \) smaller than a given residual threshold \( \epsilon \) (Bovy et al. 2012; Peters et al. 2015;
Richards et al. 2015),

$$\delta_e = \frac{N(|z_i - \hat{z}_i| < e)}{N_{\text{tot}}}. \quad (5)$$

Typical values chosen are $e = 0.1, 0.2, 0.3$. However, in many cases the redshift normalized residuals are used instead,

$$\delta_e = \frac{N(|z_i - \hat{z}_i| < e \cdot (1 + z_i))}{N_{\text{tot}}}. \quad (6)$$

5.4.2. Results

The results of the regression calculations using RF regression and SVR are shown in Table 6.

The top half of the table details the results from the RF method, whereas the bottom half shows the corresponding results for the SVR. Both algorithms perform similarly well for all four training and feature sets used. If the WISE W1 and W2 photometry is included, the RF method performs slightly better.

A comparison between the qualitative results of the two methods is given in Figure 6. The left panel shows the results of the second and sixth rows of Table 6 while the right panel shows the third and seventh rows of the same table. The orange and blue curves are the histograms of the redshift residuals $\Delta z$ for the SVR and RF regression, respectively. While the RF method shows a tighter distribution of the histogram around $\Delta z = 0$, the different colored curves show the same general qualitative behavior in both panels.

The main difference between both methods lies in the computing time. For the subset referenced with the * (Table 6), the computing time for RF regression amounted to 281 s, whereas the same subset calculated with SVR took 4682 s for similarly good results. This is an inherent drawback of the SVR method as the computing timescales with $O(N^3)$, where the number of objects in the training set is $N$. For comparison, the random forest method computed timescales with the number of trees in the Forest $T$ and the depth of each tree $D$ as $O(T \cdot D)$. Hence, for large training sets, the random forest regression should always be preferred if both regression methods perform equally well.

If one compares the $R^2$ score of the second and third rows of Table 6, it becomes obvious that the inclusion of more photometric features clearly improves the regression results. While the $R^2$ score summarized the quantitative effect in one number, we show the qualitative differences in Figure 6. From the left to the right panels, the distribution of test quasars tightens significantly around $\Delta z = 0$ and the second bump around $\Delta z \approx -0.8$ disappears.

The price is the reduction of the training sample from 172,069 quasars to 123,112. However, a comparison between the first and second rows of Table 6 shows that a larger training set improves the regression results if the same number of features are considered. This suggests that an addition of training objects still leads to a better characterization of the target values in the feature space.

The best regression results are achieved on the SDSS+W1/W2 training set with $m_i < 18.5$. Because we ignore the distribution of fainter quasars here, the photometric errors on this training set are smaller. As a result, the regression achieves better results. However, we have to advise caution here. The $m_i < 18.5$ cut severely reduces the number of training and test objects. Because the redshift distribution of the empirical training set is dependent on the $i$-band magnitude of the quasars, it biases the training set against higher-redshift sources.

We choose the RF method with the SDSS+W1/W2 training set and all available features for our photometric redshift estimation (marked with a * in Table 6).

Figure 7 shows the density map of the random forest regression test results using this training set as a function of the measured spectroscopic redshift. The color of the map corresponds to the number of test objects per rectangular bin. The photometric redshift of the majority of test objects is well estimated. However, a distribution of outliers still persists at spectroscopic redshifts of $z \approx 0.8, 1.6, \text{ and } 2.0-2.3$.

5.4.3. Comparison to Other Methods

The difficulty in comparing different methods for redshift estimation to another is that the same training set should be used in all cases to create the model. Otherwise, model
comparisons, even using the same regression metrics, are not meaningful.

In a recent study on the photometric selection of quasars (Yang et al. 2017, ApJ in press), the authors incorporate asymmetries into a model of the relative flux distributions of quasars. As part of their work, they carry out a comparison (see their Table 1; Yang et al. 2017, ApJ in press) between their new method (Skew-t) and a range of frequently used algorithms in the literature (XDQSOz, Bovy et al. 2012; KDE, Silverman 1986; CZR, Weinstein et al. 2004) based on the same photometric test and training sample. Their method achieves slightly better results than the others on the same training set based only on SDSS photometry.

While their training/test sample includes a range of high-redshift quasars from the literature, the majority of the sample is also based on the DR7 and DR12 catalog quasars. Therefore, we compare our results (Table 6) with theirs in Table 7. The RF and SVR are outperformed by the Skew-t method on the SDSS feature/training set (see $\delta_{02}$ = 0.1); meanwhile, the RF method performs equally well, once the WISE W1 and W2 bands are included. Hence, random forests can keep up with other modern methods of photometric redshift estimation.

5.5. Quasar–Star Classification

To classify our quasar candidates as stars or quasars, we only use the random forests, because the SVM algorithm is not suited for simultaneous classification with more than two classes.

The training set for the classification problem is built from the empirical quasar and star catalogs described above. Both catalogs are added to form the full empirical training set on which the random forest classification for different stellar spectral classes and quasar redshift classes will be trained.

However, we limit ourselves to stars with spectral classes A, F, G, K, and M since these classes have enough objects for sufficient training and the main contaminants of quasars with redshifts $z = 2$–5 are K and M stars. Most L and T dwarfs are too faint for our ELQS and conflict only with even higher-redshift quasars in color space. We also excluded O and B stars because they are very rare, and their blue colors are hardly confused with $z > 2$ quasars.

Since quasar colors change considerably over redshift, we divide quasars into four redshift classes similar to Richards et al. (2015). The classes are designed to split the quasars at redshifts, where dominating features change the $u$-band to $g$-band flux ratio. The redshift classes are “$\text{lowz}$” with $0 < z \leq 1.5$, “$\text{lowz}$” with $1.5 < z \leq 2.2$, “$\text{midz}$” with $2.2 < z \leq 3.5$, and “$\text{highz}$” with $3.5 < z$. At $z = 1.5$, the Ly$\alpha$ emission line is just blueward of the $u$-band and the CIV emission line is still in the $g$-band. The second break at $z = 2.2$ marks the point where the Ly$\alpha$ line is leaving the $u$-band and the last break at $z = 3.5$ is marked by a strong flux decrease in the $u$-band as the Ly$\alpha$ forest absorbs flux blueward of the Ly$\alpha$ line. We use these four quasar classes and five stellar classes to form the nine main labels for our classification problem. For evaluation purposes, we introduce the binary labels “STAR” and “QSO,” which encompass all stellar classes and all quasar classes, respectively. We will use them to calculate the completeness of the quasar selection against stars in general.

The RF algorithm is trained only on the photometric information. For classifications using only the SDSS color space, the features are the four adjacent flux ratios ($u/g$, $g/r$, $r/i$, $i/z$) from the five photometric bands and the SDSS $i$-band magnitude. When we include the 2MASS or WISE photometry, we expand the flux ratios accordingly and add the $J$-band magnitude or the $W1$ magnitude to the feature set.

We investigate the effects of two magnitude limits in the SDSS $i$-band, $m_i < 21.5$ and $m_i < 18.5$, and the use of different sets of photometric features (SDSS, SDSS+W1W2, and SDSS+2MASS+W1W2) on the results of the random forest classification. The six subsets of the full empirical training set have different sizes, since not every object always has the full photometric information required. If information in one photometric band or its associated error is missing, we reject the source from the training set.

For each of the six subsets, we find the best combination of the training hyperparameters by calculating the classification results on a grid of $n_{\text{estimators}} = [50, 100, 200, 300]$, $\text{min}_{\text{samples}}_{\text{split}} = [2, 3, 4]$, and $\text{max}_{\text{depth}} = [15, 20, 25]$ (36 combinations). We use five-fold cross-validation and always train on 80% of the full training set, while the remaining 20% are used to test the classification.

5.5.1. Classification Metrics

In order to measure the performance of the classification, we will introduce three standard classification metrics: the precision, the recall, and the $F_1$ score (Bishop 2006).
Training problems is the confusion matrix used. We show the results of the classification on the training set along with their respective constraints and features. The Astrophysical Journal, well as the grouped classes of all quasars (although they truly belong to other classes in the sample.

The harmonic mean of the precision and recall values is the traditional F-measure or balanced F-score. The \( F_1 \) score for each class individually is given by:

\[
P = \frac{tp}{tp + fp} \quad \text{and} \quad r = \frac{tp}{tp + fn}.
\]

The harmonic mean of the precision and recall values is:

\[
F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.
\]

For multiclass classification problems, one can define a precision, recall, and \( F_1 \) score for each class individually against all other classes.

Also, an average precision, recall, and \( F_1 \) score weighted by the number of occurrences in each true class can provide an idea of how well the classifier generally works for problems with multiple classes.

A helpful visualization for the results of the classification problems is the confusion matrix \( C \). Each entry \( C_{ii} \) is the number of objects known to be in class \( i \), but predicted in class \( j \). Therefore, the entries \( C_{ij} \) show the true positives \( tp \) for each class \( i \). All other values in row \( i \) show the number of false negatives \( fn \), objects predicted to belong to other classes, although they truly belong to class \( i \). All other values in column \( j \) are false positives \( fp \), the values predicted to belong to class \( j \), although they truly belong to other classes in the sample.

5.5.2. Results

We present the results of the random forest classifications in Table 8. It shows the six different subsets of the full empirical training set along with their respective constraints and features used. We show the results of the classification as the precision \( p \), recall \( r \), and \( F_1 \) measures for the “highz” quasar class as well as the grouped classes of all quasars (“QSO”) and stars (“STAR”).

As expected, it is evident from Table 8 that the inclusion of more photometric features always leads to better classification results at the expense of the training sample size.

The subsets with the same amount of features but different limits on the SDSS \( i \)-band show that the brighter samples \( (m_i < 18.5) \), the ones with the qualitatively better photometry, also show more accurate classifications. Again, the training sets for those three subsets are much smaller and will not be able to populate the entire feature space as well as a fainter limit of \( m_i < 21.5 \) would allow.

Since the number of higher-redshift quasars is a strong function of the \( i \)-band magnitude, any limitation of the training set in this regard will bias the classification of those objects. For example, the training set with full SDSS+2MASS+W1/W2 photometry and \( m_i < 18.5 \) \((m_i < 21.5)\) has only 38 (42) “highz” \((z > 3.5)\) training objects and 8 (13) test objects. This demonstrates that these classification results cannot be fully trusted because of the low number of objects that the precision, recall, and \( F_1 \) score is based upon.

Based on these insights, we adopt the SDSS+W1/W2 subset with \( m_i < 21.5 \) as the best training and feature set for the quasar classification in our ELQS quasar selection (see the * in Table 8). It achieves the highest completeness (recall) of “highz” quasars and “QSO” in general of all subsets in Table 8 with the hyperparameters set to \( n_{\text{estimators}} = 300, \text{min}_{\text{samples}}_{\text{split}} = 3 \) and \( \text{max}_{\text{depth}} = 25 \).

For this particular training and feature set, we show the full confusion matrix in Figure 8. The rows correspond to the true class of the object, while the columns show the predicted labels. The values on the diagonal are the correctly classified objects. Each entry in the matrix shows the total number of objects and the percentage of the objects in that entry with respect to the true class (the full row). The entries are color-coded based on this percentage, with a darker blue color corresponding to a higher percentage.

Although the stellar classification encounters difficulties for the A, F, and G stars, the later spectral types and the quasars are well classified with diagonal entries above 80%. Only a very small percentage of stars are classified as quasars (top-right corner), with the majority of stellar contaminants falling into the “midz” quasar class \((2.2 < z < 3.5)\). This is the redshift range in which the quasar distribution overlaps with the stellar locus the most in optical-color space.

Conversely, the bottom-left corner shows quasars classified as stellar sources. Again, the majority of these objects fall into the “midz” quasar class.

We can simplify this confusion matrix by grouping all stellar spectral classes and all quasar classes in the binary classes “QSO” and “STAR.” The binary classification results are shown in Table 9. Here, the top part summarizes the results of the confusion matrix in Figure 8. In the bottom part, we used the same training set but reduced the feature set to only include the SDSS photometry. As a result, the off-diagonal values are much higher, reflecting a larger fraction of stellar contaminants and a lower quasar completeness. This demonstrates how the information gained by including the WISE W1 and W2 bands enhances the performance of the classifications.

### Table 8

| Training/Test size | Constraints | Features | \( p/\tau/F_1 \) (highz) | \( p/\tau/F_1 \) (QSO) | \( p/\tau/F_1 \) (STAR) |
|-------------------|-------------|----------|--------------------------|--------------------------|--------------------------|
| 183,108/45,778    | SDSS fl.r., \( m_i < 18.5 \) | SDSS     | 0.92/0.71/0.80           | 0.88/0.96/0.92           | 1.00/0.99/1.00           |
| 167,076/41,770    | SDSS+W1/W2 fl.r., \( m_i < 18.5 \) | SDSS+W1/W2 | 0.97/0.86/0.91           | 0.91/1.00/0.95           | 1.00/0.99/1.00           |
| 129,934/32,485    | SDSS+2MASS+W1/W2 fl.r., \( m_i < 18.5 \) | SDSS+2MASS+W1/W2 | 0.88/0.88/0.88           | 0.93/0.99/0.96           | 1.00/1.00/1.00           |
| 442,529/110,634   | SDSS fl.r., \( m_i < 21.5 \) | SDSS     | 0.87/0.87/0.87           | 0.77/0.94/0.85           | 0.96/0.85/0.90           |
| 313,453/78,364    | SDSS+W1/W2 fl.r., \( m_i < 21.5 \) | SDSS+W1/W2 | 0.92/0.95/0.93           | 0.88/1.00/0.94           | 1.00/0.92/0.96           |
| 141,837/35,460    | SDSS+2MASS+W1/W2 fl.r., \( m_i < 21.5 \) | SDSS+2MASS+W1/W2 | 1.00/0.69/0.82           | 0.90/0.99/0.94           | 1.00/1.00/1.00           |
training set would be strongly biased against higher-redshift $(z \gtrsim 3)$ objects, which have less bright apparent magnitudes and are therefore less likely to be detected by 2MASS. Even though we are focusing on the brightest quasars, it is important for the redshift range above $z \approx 3$ to be well populated in the training set to achieve reliable results in the classification and regression.

As long as $J$- and $K$-band photometry are limited to the brightest objects in the training sets, the JKW2 color cut will play an important role in rejecting stellar sources.

6. The ELQS Quasar Candidate Catalog

6.1. Area Coverage of the ELQS Survey

The ELQS survey includes all SDSS photometry with galactic latitudes $b < -20$ or $b > 30$. To estimate the effective area of the ELQS survey, we use the Hierarchical Equal Area isoLatitude Pixelization (HEALPix; Górski et al. 2005). The process of the calculation and the general parameters used are identical to the description in Jiang et al. (2016).

The effective area of the full ELQS survey is $11,838.5 \pm 20.1 \text{ deg}^2$, with a contribution of $7601.2 \pm 7.2 \text{ deg}^2$ from the spring (90 deg $< R.A. < 270$ deg) sky and a contribution of $4,237.3 \pm 12.9 \text{ deg}^2$ from the fall (R.A. $> 270$ deg and R.A. $< 90$ deg) sky.

6.2. Construction of the Candidate Catalog

With all tools at hand, we begin the construction of the ELQS quasar candidate catalog in the SDSS footprint. An overview is given in Figure 3. The general source selection is based on the WISE AllWISE catalog matched with photometry from the 2MASS PSC. Both surveys are all-sky and therefore include the galactic plane, where the source density of stars is extremely high and leads to confusion in the near- and infrared surveys. Therefore, we restrict the quasar candidates to larger galactic latitudes and only include sources with either $b < -20$ or $b > 30$. We also require the WISE $W1$ and $W2$ photometry to have $S/N > 5$ and positive $J$-band magnitudes for all objects. All sources that pass these criteria and obey the JKW2 color cut $K - W2 \gtrsim 1.8 - 0.848 \cdot (J - K)$ are selected in our WISE–2MASS–Allsky catalog. It comprises a total of 3,576,354 sources.

We then proceed to match all of these sources to the SDSS DR13 (PhotoPrimary) catalog in a 3\arcsec 96 aperture. We do not reject any objects based on their photometric flags. None of the fatal or non-fatal flags of Richards et al. (2002) are evaluated in our selection. We instead inspect the images of all quasar candidates ($\approx 400$ objects) in the very end to be as complete as possible. The matched SDSS–WISE–2MASS catalog has a total of 1,690,813 objects.

In the next step, we apply the criterion on the Petrosian radius $\text{petroRad} \leq 2.0$; see Section 4.3) to reject the majority of galaxies. We also require all sources to satisfy the $i$-band magnitude cut of $m_i < 18.5$.

For the remaining candidates, we calculate photometric redshifts and evaluate their quasar probabilities using RF regression and classification (see Section 5). This demands that all sources have quantified photometric errors for all SDSS and WISE $W1$ and $W2$ photometry. The classification calculates the most probable class of objects ($\text{rm_emp_mult_class_pred}$), the general quasar or star class ($\text{rf_emp_bin_class_pred}$), and the total probability of the object to belong to the quasar

5.6. On the Prospect of Using Only Random Forests for Quasar Selection

We demonstrated that random forests efficiently classify quasars and estimate their photometric redshifts using the SDSS optical bands in concert with WISE $W1$ and $W2$ photometry. In this context, we need to ask whether the JKW2 color cut is really necessary for our quasar candidate selection.

To evaluate these questions, we assume the RF classification contamination as shown in Table 9 and apply it to the spectroscopically unidentified sources in the 70 deg$^2$ test regions (Table 2). In this case, the number of stellar contaminants would at least rise from the $\sim 1000$ which make the color cut to $\sim 1700$. This illustrates how important the information of the 2MASS $J$- and $K$-bands is to reject stars.

Unfortunately, only a small fraction of all known quasars and stars in the training sets have well-detected 2MASS $J$ and $K$ photometry. The resulting training samples are too small to properly populate the multidimensional feature space to allow for sufficient training of the RF model. In addition, the quasar
class \((\text{rf\_emp\_gso\_prob})\). The regression delivers the best estimate for the photometric redshift \((\text{rf\_emp\_photoz})\).

With this information at hand, all objects that obey the photometric redshift cut of \(z_{\text{reg}} > 2.5\) and are generally classified as quasars \((\text{rf\_emp\_bin\_class\_pred} = \text{QSO})\) are identified as additional candidates. This results in a candidate catalog of 2253/1735\footnote{With \(S/N > 5\) in the WISE \(W1\) and \(W2\) bands and the quality requirements of the 2MASS PSC.} total/primary sources, out of which 920/876 are known quasars from the literature and 1333/859 are valid candidates for spectroscopic follow-up.

From this sample, we prioritize bright high-redshift objects with \(z_{\text{reg}} > 2.8\) and \(m_i < 18.0\), which will make up the ELQS spectroscopic survey. These criteria leave 742/594 total/primary candidates out of which 341/327 are known and 401/267 need to be followed up.

In the last stages, every candidate’s photometry will be inspected before observation to check for image defects or obviously extended sources. Of the total/primary sample, roughly 59%/69% are good candidates, 10%/8% are blended with other sources in WISE, 9%/6% are extended, and 22%/17% have bad image quality in at least one of the bands or are contaminated by bright sources close by.

The final ELQS quasar candidate catalog includes a total of 237 targets, out of which 184 are primary candidates.

### 7. Conclusion

In this paper, we showed that the SDSS and BOSS quasar surveys have systematically missed bright quasars at redshifts \(z \sim 3\) and above. This is mainly due to stellar contamination at redshifts, where quasars overlap in optical-color space with the stellar locus, and the surveys’ incomplete spectroscopic observations in the fall sky (R.A. > 270 deg and R.A. < 90 deg) of the SDSS footprint. We developed a more inclusive quasar selection algorithm that is based on a near-IR/infrared color criterion with high quasar completeness. Further inclusion of the SDSS optical photometry allows for galaxy rejection, classification of sources in stellar spectral types and quasar redshift classes, and photometric redshift estimation. The latter tasks are accomplished using the random forest machine-learning method on a training sample of SDSS DR13 spectroscopic stars and quasars from the SDSS DR7 and DR12 quasar catalogs.

Although the near-IR/infrared color criterion is very effective, the limiting magnitude of the 2MASS survey only allows using it at the brightest end of the quasar distribution. This limits the use of this criterion, as even our first estimates, taking the combined SDSS DR7 and DR12 quasar catalogs as a basis, show that we only reach 80% photometric completeness for \(m_i < 18.0\) quasars at \(z > 2.5\). Future infrared surveys (e.g., EUCLID) with deeper photometry will be able to exploit this color cut to create a fainter, highly complete quasar sample.

The total/primary high-priority quasar candidate catalog comprises 237/184 objects with \(z_{\text{reg}} \geq 2.8\) and \(m_i < 18.0\). Observations taken up to August 2017 have successfully identified a total of 67 quasars at \(z \geq 2.8\) in both total and primary candidate samples. We estimate the efficiency of our quasar selection on the spectroscopically completed ELQS spring sky sample. It includes 340 primary candidates of which 39 are newly identified quasars as part of ELQS, and 231 are known quasars in the literature. The remaining primary candidates were either low-redshift quasars (36) or identified not to be quasars (35) by our survey. This results in an efficiency of roughly \(\sim 79\%\). The efficiency predicted by the RF classification reaches 80%–90% for “midz” and “highz” quasars. The reason for our somewhat lower efficiency value is likely found in our loose quality criteria on the SDSS, WISE, and 2MASS photometry. We do not use any of the standard SDSS, 2MASS, or WISE quality flags and only rely on \(S/N > 5\) in the WISE \(W1\) and \(W2\) bands and the quality requirements of the 2MASS PSC.

In a forthcoming publication, we will present the spectroscopic observations of the completed spring sky footprint of SDSS, along with a discussion on the full completeness of the ELQS survey and a first estimation of the bright-end QLF. With the conclusion of the survey, a final publication will calculate the ELQS quasar luminosity over the full survey footprint and discuss the implications for the evolution of the brightest quasars and thus the most massive black holes.

J.T.S., X.F., and I.D.M. acknowledge support from the NSF grant AST 15-15115. Q.Y., J.W., and L.J. acknowledge support from the National Key R&D Program of China (2016YFA0400703) and from the National Science Foundation of China (grant 11533001).

This publication makes use of data products from the Two Micron All Sky Survey, which is a joint project of the University of Massachusetts and the Infrared Processing and Analysis Center/California Institute of Technology, funded by the National Aeronautics and Space Administration and the National Science Foundation. This publication makes use of data products from the Wide-field Infrared Survey Explorer, which is a joint project of the University of California, Los Angeles, and the Jet Propulsion Laboratory/California Institute of Technology, funded by the National Aeronautics and Space Administration.

Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS Web site is http://www.sdss.org.

SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, the Chilean Participation Group, the French Participation Group, Harvard-Smithsonian Center for Astrophysics, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU)/University of Tokyo, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatório Nacional/MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional
