MNRAS **000**, 1–6 (2017) Preprint 27 July 2017 Compiled using MNRAS LATEX style file v3.0

SDSS Stripe 82 : finding quasars in the forced photmetry haystack

Krzysztof Suberlak,[1]⋆ Željko Ivezić, [1] Yusra AlSayyad[1]
[1]*Department of Astronomy, University of Washington, Seattle, WA, United States*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT
We provide variability and color-selected quasar sample up to g < 23.5 . We reproduce many results from previous research, and find that using the LSST Stack reprocesse SDSS S82 data allows reaching fainter objects, and extending the reach of possible quasar candidate classification.

1 INTRODUCTION

This report aims to outline the process of analyzing the forced photometry reprocessed SDSS Stripe 82 data with the aim of improving the quasar selection using combined color and variability cuts.

Quasars are some of the brightest sources of radiation in the Universe. Most galaxies have gone through a phase of rapid accretion onto the central supermassive black hole, which resulted in emission of radiation from the hot accretion disk. As a result of physical processes in the disk, the radiation exhibits stochastic variability that can be mathematically described by a damped random walk, or a process with a certain covariance, and characteristic decay timescale. The amplitude of variability and characteristic timescale can be linked to physical properties of the disk, and are therefore of high interest. Beyond that, quasars are relevant as astrophysical probes, since the quasar luminosity function is related to the evolution of galactic initial mass function. (McGreer+2013)

Traditionally quasars can be found by color cuts because at least the nearby ones occupy a specific region in the color space (Sesar+2007, Ivezic+2003, Bovy+2011). However, it has long been recognized (Fan+1999) that as the redshift of the quasar increases, it's color changes because we probe different regions of the intrinsic spectral energy distribution. Around redshift 2 quasars cross through the stellar locus in u-g vs g-r color space (Yang+2016, Richards+2015 , Jiang+2014), so that without an additional selection criterion they are indistinguishable from stars (especially M dwarfs at redsfhits 5-6, Yang+2016). At higher redshifts quasars again occupy a region that can be confused with RR Lyrae in the color space. Variability has been successfully employed for quasar selection for a number of years (Palanque-Delabrouille 2011, 2013, 2016, Schmidt+2010, VandenBerk+2004, MacLeod+2011, MacLeod+2013, Peters+2015). This possible because stars in the main stellar locus are not variable, and RR Lyrae or Cepheids have a very specific variability pattern, very distinct from DRW. Eclipsing Binaries exhibit only very occasional deep dips in magnitude, which would be well distinguished from quasars also based on variability alone. Combining variability and color information can therefore provide a way to select quasars with minimally small contamination.

Stripe 82 is a very special SDSS field, observed multiple times, initially as part of the supernova survey. Each object in this equatorial stripe has between 60-180 epochs, and this allows study of variability. Coupled with supreme, well-calibrated SDSS photometry, S82 is a favorable testbed for selection and classification studies.

The S82 data was reprocessed in the Summer of 2013 as a result of the LSST Data Challenge, that was designed to test the capability of then in-development LSST Stack[1]. The S82 forced photometry dataset was also used as the testbed of database ingest into the Prototype Data Access Center[2].

The reprocessing included preparing i-band coadds, source detection on the coadds, and forced photometry on these locations in all epochal u,g,r,i,z data. The reprocessing effort was shared between the NCSA and IN2P3 to test the portability of algorithms. A five degree overlap was processed in both Data Processing Centers (DPC).

2 S82 REPROCESSING

We use data from all SDSS runs up to an including run 7202 (Data Release 7), including all 6 SDSS camera columns. Stripe 82 survey covered an equatorial strip of the sky, defined by declination limits of $\pm 1.27\,\mathrm{deg}$, extending from R.A. $\approx 20^h(320\,\mathrm{deg})$ to R.A. $\approx 4^h(55\,\mathrm{deg})$ (**??**). Observations conducted prior to September 2005 (part of SDSS I-II) had a more sparse sampling than SDSS-III, and the SDSS Supernova Survey, which ran between September 1st - November 30th each year between 2005-2007.

The SDSS Stripe 82 DR7 data was processed in two

[1] see https://confluence.lsstcorp.org/display/DM/Properties+of+the+2013+SDSS+Stripe+82+reprocessing
[2] DMTN-029 : Loading SDSS Stripe 82 data into PDAC https://dmtn-029.lsst.io

data centers : NCSA (National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, IL) and IN2P3 (Institut national de physique nucléaire et de physique des particules in Paris, France). NCSA processed data from $-40\,\mathrm{deg}\,(+320\,\mathrm{deg}) < RA < +10\,\mathrm{deg}$ and IN2P3 with $+5\,\mathrm{deg} < RA < +55\,\mathrm{deg}$. There is a 5 deg overlap, used to confirm that the data processing pipeline in both data centers yields identical data products.

All epochs (individual images) were background-subtracted, and then scaled from the Digital Unit counts to fluxes by comparing standard objects against the (**?**) catalog (similar to Jiang+2014). We downloaded the data from the NCSA storage at `https://lsst-web.ncsa.illinois.edu/~yusra/S13Agg/rawDataFPSplit/`.

The available data includes the DeepSource tables, which constitute of i-band coadd detection information, and the DeepForcedPhot tables, which contain forced photometry seeded from i-band detections.

The DeepSource tables and DeepForcedPhot tables can be joined on deepSourceId, which is called objectId in DeepForcedPhot.

Yusra AlSayyad and Ian McGreer calculated summary aggregate metrics on the S82 S13 data, using the code available at `https://github.com/imcgreer/QLFz4`. This information was used in McGreer+2013, and AlSayyad 2016 PhD Thesis.

In the following sections we describe the construction of DeepSource and DeepForcedSource tables, and what data products can be extracted from each catalog.

2.1 DeepSource tables: coadd source detection

Source extraction often needs to address the problem of blended sources. In such case the process of deblending assigns a single ParentSourceId to a blended clump. For an object which is a parent (eg. a galaxy), ParentSourceId is null. Solitary sources which are not blended in clumps are their own parents.

Sources with parents brighter than 17 mag in the coadded i-band are considered unreliable, since they were not handled well by the deblender . ParentSourceId is null for objects that are their own parents (or in other words, are not blended).

Sources were detected and measured in i-band coadds. Locations for the sources with coadd i-band psfMag < 23.5 were used as a seed for forced photometry across epochs and bands. The data was simultaneously processed at NCSA ($-40 < \mathrm{RA} < 10$) and IN2P3 ($5 < \mathrm{RA} < 55$) - there is a 2.5 x 5 square degrees overlap for validation. Imaging at each Data Release Production (DRP) was processed independently. Fig. 1 shows the location on the sky of the forced photometry seeds (coadd iPsfMag < 23.5, parent iPsfMag > 17 mag).

Imposing a cut of 23.5 mag in coadded i-band , the main DeepSource NCSA -processed catalogs used contain 5474350 primary sources , and 1957486 non-primary sources (deblender parents and secondary detections). IN2P3-processed portion of the S82 includes 4998901 primary sources, and 1882303 non-primary sources.

Both NCSA and IN2P3-processed regions of S82 were divided into smaller sections called patches[3] , as illustrated on Fig. 2.

The extendedness was defined by a cutoff in iPsfMag - iModelMag. Wherever iPsfMag - iModelMag > 0.085645 , extendedness $= 1$, and 0 otherwise. This is illustrated on Fig. 3, and Fig. 4.

2.2 DeepForcedSource : forced photometry tables

For convenience and data storage, about 20 patches per filter are included per filter-patch file (eg. 'g00_22.csv' including forced photometry for all objects within patches J 0 to 22, in g filter). In particular, the following patch files were processed in NCSA: 00_21, 22_43, 44_65, 66_87, 88_109, 110_131, 132_153, 154_175, 176_181, 365_387, 388_409, and the following in IN2P3 : 155_176, 176_197,197_218, 218_239, 239_260, 260_281, 281_302, 302_323, 323_344, 344_365, 365_386.

The forced photometry tables, organized by filter-patch, contain the `id`, `objectId`, `exposure_id`, `mjd`, `psfFlux`, `psfFluxErr`, sorted by `objectId`, where Flux is measured in calibrated $[ergs/cm^2/sec/Hz]$.

For each light curve we calculate features. One main choice is that the features (described in Sec. 3) can be calculated either based on all epochs, or on seasonally-binned epochs, or on seasonal averages.

Before calculating features, we convert flux to Janskys to avoid floating point error on very small numbers. After removing all rows that are missing either flux, flux error, or time stamp, we calculate the signal-to-noise ratio for each epoch :

$$S/N = \frac{psfFlux}{psfFluxErr} \tag{1}$$

Wherever $S/N < 2$ (including negative values of forced photometry flux), by representing each flux measurement as a Gaussian likelihood $\mathcal{N}(psfFlux, psfFluxErr)$, we apply a Bayesian prior (including information about expected counts as a function of flux, and removing negative tail), and based on such truncated likelihood , we calculate mean, median, 2σ level and the RMS (called faintMean, faintMedian, faintTwoSigma, faintRMS, as described in the Faint Pipeline Report `https://github.com/suberlak/Faint_pipeline_report`). For each flux measurement where $S/N < 2$ we replace psfFlux with faintMean, and psfFluxErr with faintRMS.

This is the basic pre-processing applied to all light curves. What follows is either calculating full lightcurve-based features, or seasonally averaged features. We first describe the possible features in general, and then discuss the calculation of features for both full light curve and seasonally averaged light curve cases.

3 TIME SERIES FEATURES

For any time series (a collection of points with time stamps) it is possible to calculate a variety of features. An excellent overview is included in Sokolovsky+2016, who makes an

[3] patch boundaries can be found `https://lsst-web.ncsa.illinois.edu/lsstdata/dr-w2013/coaddBounds.txt`

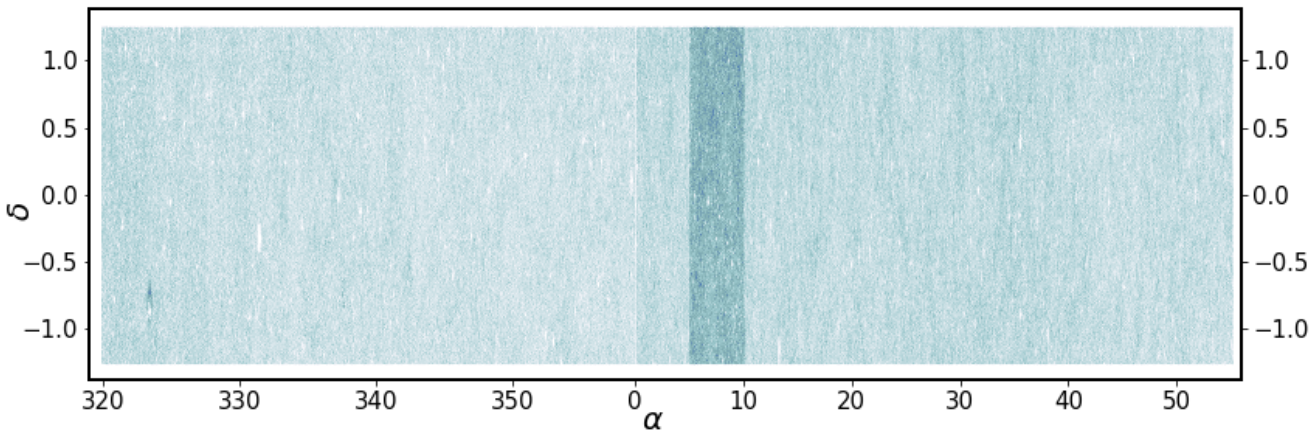

Figure 1. Combined NCSA (left) and IN2P3 (right) i-band forced photometry seeds. Apart from requiring these primary sources to have the coadd magnitude iPsfMag < 23.5, and have parents fainter than 17 mag, there are no other cuts applied to the 9491361 objects shown. In particular, for NCSA of 5474350 sources we remove 585920 with parents brighter than 17 mag, and for IN2P3 of 4998901 sources we remove 395970. This leaves 4888430 objects in NCSA, and 4602931 in IN2P3, so that in total we have 9491361 objects. One can clearly see the 2.5x5 degrees overlap region (which has a higher source density). To aid plotting so many objects we subsampled 25% of all objects : randomness ensures that all spatial features are preserved. There are some regions where there are no objects satisfying our criteria, seen here as blank spots. In particular, region around ra=323.36258, dec=-0.82325 is an M2 globular cluster, and the sheer stellar density prevented from using any sources there as seeds.

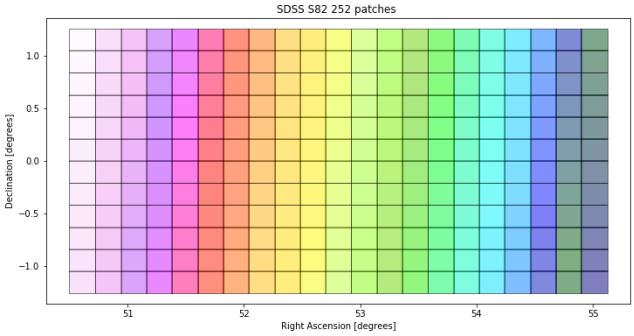

Figure 2. An illustration of 252 patches out of 4920 patches. Patches are denoted by (I,J) coordinate , where I runs in δ (from South to North), and J along α, from higher to lower right ascension values.

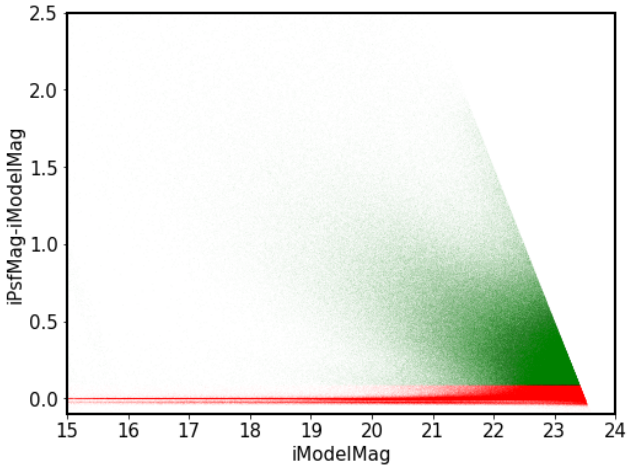

Figure 3. The NCSA data for objects with iPsfMag < 23.5, as above. There is 5474350 such objecs, and we subsampled every fourth object at random for plotting (1370001 objects are shown). We did not remove objects with parents < 17 mag. This scatter plot demonstrates the simple method of defininig extendedness : as a threshold in iPsfMag-iModelMag = 0.085645. All sources above that line are considered extended (green), and those above : compact (red). Extended sources have extendedness 1, otherwise it is 0. This plot also illustrates an important point that the uncertainty in classifying a given coadd source as 'extended' is magnitude-dependent, and the fainter the object, the less definite is this distinction. This is because distant, faint galaxies become more compact, and thus the green objects approach the red objects in the faint limit.

important distinction between scatter-based features (that he calls indices), and correlation-based ones. In our treatment we almost exclusively calculate scatter-based indices (χ^2, weighted standard deviation, weighted mean, weighted median, interquartile range). We encourage the reader to also consult existing packages meant for feature engineering (FATS: Feature Analysis for Time Series, Nun+2015 https://github.com/isadoranun/FATS; UPSILoN : AUtomated Classification for Periodic Variable Stars using MachIne LearNing Kim&Bailer-Jones 2016 https://github.com/dwkim78/upsilon; TSFRESH : Time Series Feature extraction based on scalable hypothesis tests, Christ+2016 https://github.com/blue-yonder/tsfresh).

Given a collection of fluxes with errors, and observation times : $\{y_i, e_i, t_i\}$, where i can run over all epochs of a given object (eg. 0 to 180 for an S82 object), or over the observations within a given section of a light curve (eg. a season, with approximately 40 points per season), or over the seasonally-binned light curve (with usually 4 seasons per light curve). Thus y can mean the flux value measured at a

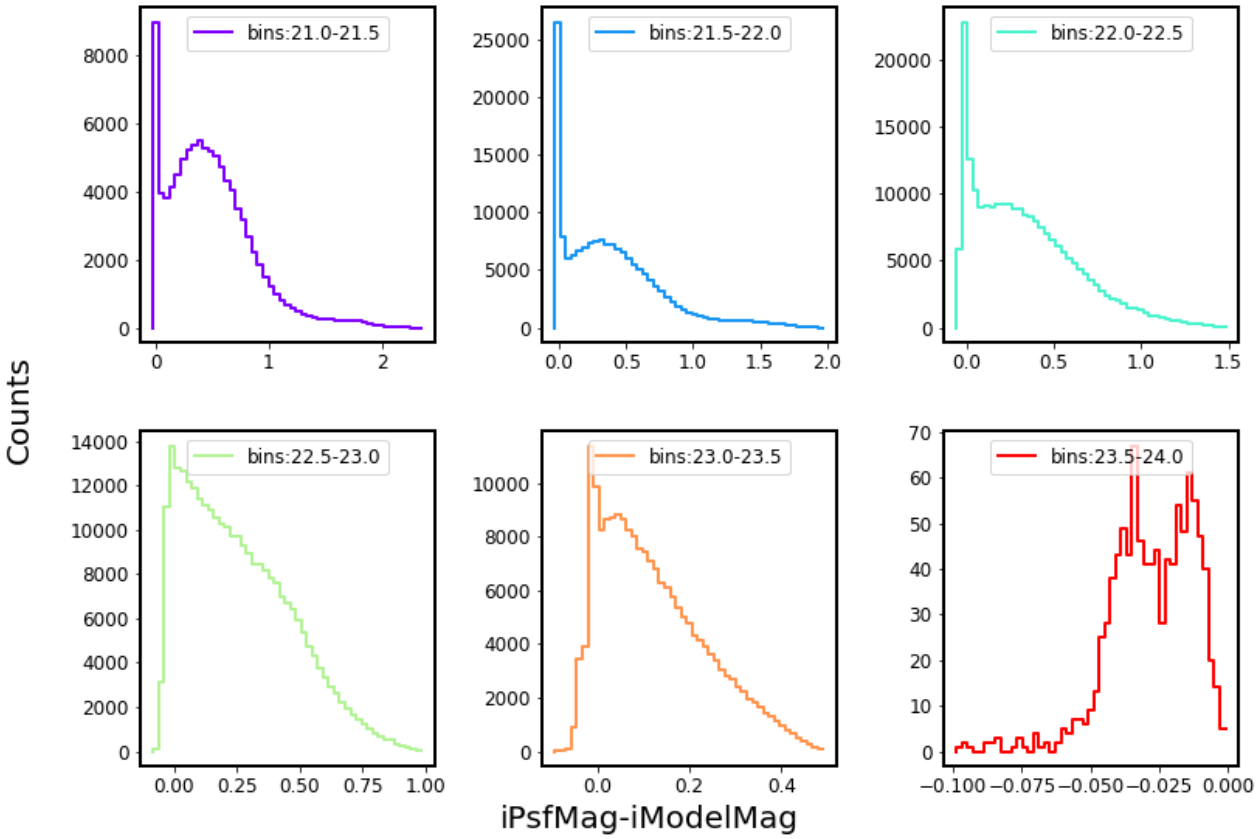

Figure 4. Histograms of vertical bins through Fig. 3. With the fainter bins in iModelMag, the two distributions in iPsfMag - iModelMag become blended. This includes only NCSA data, with the same selection and subsampling as Fig. 3 (showing 1142716 sources in total). See Palanque-Delabrouille+2016 Fig.2 for more on how iPsfMag-iModelMag can be used to select quasars. Specifically, nearby quasars may still have the host galaxy visible, which would in some cases not pass this criterion and be assigned extendedness 0.

given epoch with forced photometry, but in the following definitions it can itself be an aggregate quantity, eg. a weighted mean of fluxes within a given season. For clarity we separately define features based on raw measurements (fluxes), or aggregate quantities (eg. mean seasonally-binned fluxes).

First, consider raw forced photometry measurements. An example light curve from S82 is shown on Fig. XXX (a light curve with overplotted mean ,median, and vertical double-arrows representing the distribution spread....). We summarize the temporal information (from observation times t_i) as : minimum(t), maximum(t), mean(t).

Then we summarize the value of flux by error-weighted mean and median ($\mu_w(y,e)$, $med_w(y,e)$), together with uncertainties on those quantities (error on the mean and related error on the median). Both are useful to estimate the value of flux where most points can be found, but median is less sensitive to outliers.

We finally consider the noisiness of the data: the vertical dispersion of the flux values . This can be summarized by the robust interquartile-based deviation σ_G, weighted standard deviation about the weighed mean σ_w, χ^2_{DOF}, robust χ^2_R, and mean signal-to-noise ratio $\mu(y/e)$. The difference between robust χ^2 and the χ^2 per degree of freedom is that median-based robust χ^2 is less sensitive against any outliers, by the virtue of definition of the median. We calculate χ^2 as follows (note, that as mentioned before, y_i could be a set of 120 points over entire light curve, or perhaps just a sub-section of those 40 points within a single season - Fig. 5) :

$$\chi^2_{DOF} = \frac{1}{N-1}\sum \frac{(y_i - \mu(y))^2}{e_i^2} \tag{2}$$

$$\chi^2_R = 0.7413(q75(Z) - q25(Z)) \tag{3}$$

where

$$Z = \frac{y_i - median(y_i)}{e_i} \tag{4}$$

For weighted quantitites we introduce for each point a probability p_i (often called weight). In our case we choose $p_i = 1/e_i^2$, which makes sense because we would expect that a given flux measurement is less reliable the larger the error (it could be a spurious measurement, bad pixel, etc - we are not sigma-clipping time series).

That way, the weighted mean is :

$$\mu_w(y) = \frac{\sum p_i y_i}{p_i} \tag{5}$$

And the weighted standard deviation :

$$\sigma_w(y_i) = \sqrt{\frac{\sum p_i(y_i - \mu_w(y))^2}{\sum p_i}} \quad (6)$$

as shown in the Variability Report https://github.com/suberlak/Variability_report, with all probabilities equal , the weighted mean becomes the mean :

$$\mu(y) = \frac{\sum y_i}{N} \quad (7)$$

and the standard deviation :

$$\sigma(y_i) = \sqrt{\frac{\sum (y_i - \mu(y))^2}{N}} \quad (8)$$

The robust interquartile σ is :

$$\sigma_G = 0.7413(q75(y) - q25(y)) \quad (9)$$

and the median $med(y)$ is the fiftieth percentile :

$$med(y) = q50(y) \quad (10)$$

The error on the weighted mean is :

$$\sigma(\mu_w(y,p)) = 1/\sqrt{\sum p_i} \quad (11)$$

with the weights p_i as above .

The error on the median is related to the error on the weighted mean:

$$\sigma(med(y)) = \sqrt{\pi/2}\sigma(\mu_w(y,p)) \quad (12)$$

Another measure of noisiness is more involved - it includes modelling the light curve as a distribution of points with associated errors: $\{y_i, e_i\}$ (time information is lost). The so-called intrinsic variability σ_0 answers the question: to what extent are the errors consistent with a combination of Gaussian distributions $\mathcal{N}(y_i, e_i)$? Our method of calculating the intrinsic variability σ_0 follows Ivezic+2014 chapter 5, and is further described in the Variability Report https://github.com/suberlak/Variability_report.

We convert the aggregate flux quantities (in Janskys) to magnitudes :

$$\text{psfMean} = -2.5 * \log_{10}(\text{psfFluxMean}) + 8.90 \quad (13)$$

$$\text{psfMeanErr} = \frac{2}{\ln(10)} \text{psfFluxMeanErr/psfFluxMean} \quad (14)$$

and similarly for median-based quantities.

Aggregate information is stored in patch-filter system following the input. We gather information about the same objects across all SDSS filters by merging the patch-filter aggregates. The majority of objects have extinction information , and thus their color is corrected using extinction maps for RV = 3.1, from Schlegel+98 :

$$m_{corr} = m + A(m) * E(B-V) \quad (15)$$

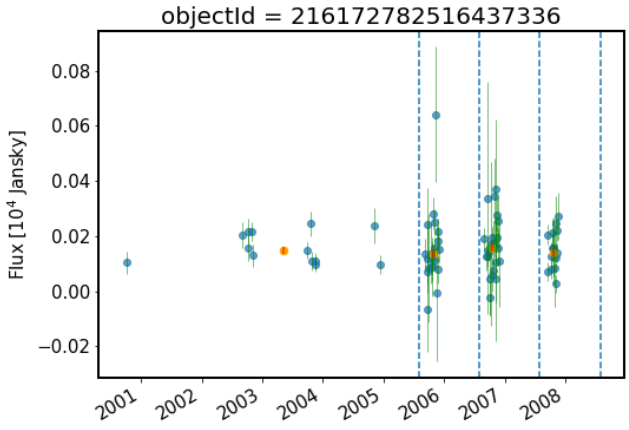

Figure 5. A plot showing a forced photometry light curve for an example object from NCSA-processed filter-patch g00_22. We mark August 1st of each year aftrer 2005 with vertical dashed line : this marks the beginning of each season. Observations have uneven cadence: this is because more observations were scheduled since the commencement of the SDSS-III Supernova Survey in September 1st 2005. For this reason we decided to consider all epochs prior to August 1st 2005 as a single season. All consecutive seasons span a period of one calendar year. We overplot the weighted mean $\mu_w(y,e)$, and the error on the weighted mean $\sigma(\mu_w(y,p))$. Each season is thus represented by the aggregate flux-related quantity with an associated error (eg. mean, median), the measure of scatter (χ^2, σ_G, etc.), and the mean time of observations within each season.

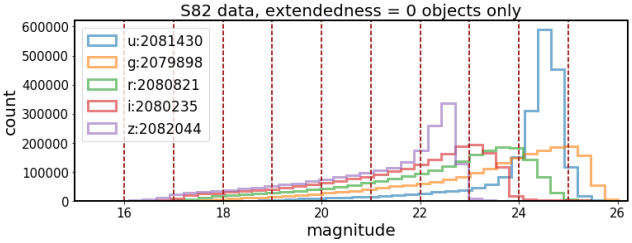

Figure 6. A histogram of all compact sources in S82.

where $m \in \{u,g,r,i,z\}$ is the magnitude of an object in a given filter, and correspondingly $A_m \in \{5.155, 3.793, 2.751, 2.086, 1.479\}$.

At this stage we also add to each object variability metrics table the positional ra,dec data from DeepSource tables . We only consider for the following analysis objects that do not have deblender parents brighter than i=17 mag.

4 SEARCHING FOR QSO

Using the light curve based average properties we combine NCSA and IN2P3 light curve aggregate information. Given the caveat of removing objects with bright parents, and those that did not have E(B-V) information, we start with 7135337 objects. Fig. 6 shows the summary of distribution of compact sources (extendedness =0). We also plot the histogram of magnitude vs error on Fig. 7.

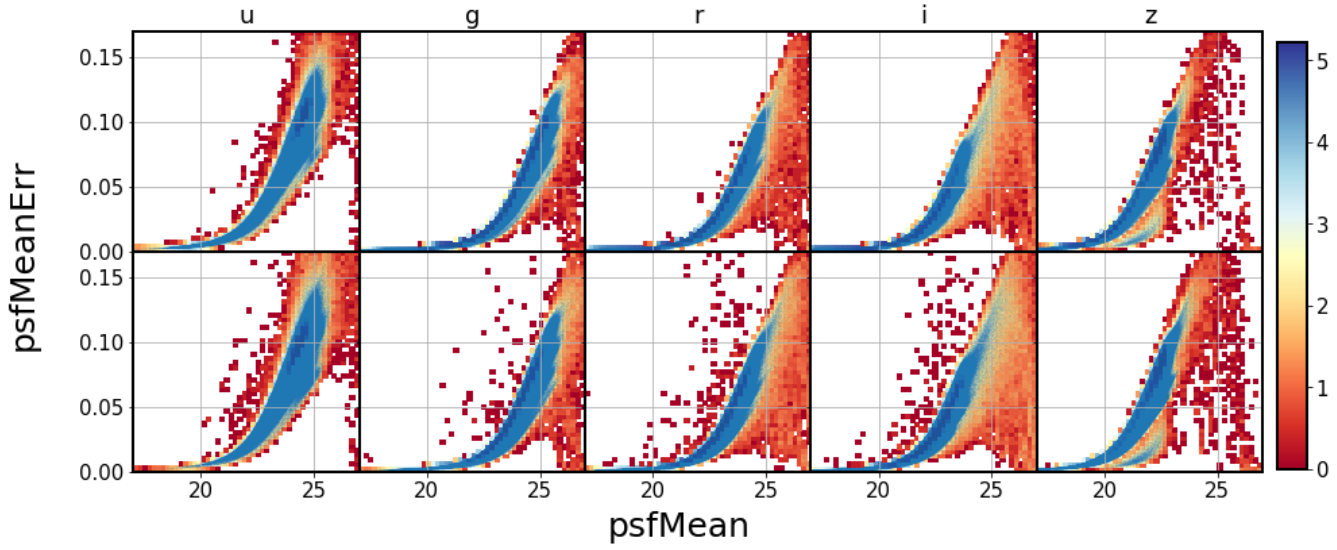

Figure 7. Mean-based magnitudes vs errors for S82 objects with extendedness=0 (top), and extendedness=1 (bottom). From left to right, data in u,g,r,i,z filters. The color scale corresponds to $\log_{10}$(counts). The fainter the object, the higher the mean error (psfMeanErr is based on psfMeanFluxErr, which is the error on the weighted mean of fluxes).

We first select variable objects by selecting various regions in the $\chi^2_{DOF} - \chi^2_R$ space, as shown on Fig. 8

Given the χ^2-based variability selection, we show the impact of that in the color-color u-g vs g-r space by first reproducing Fig.3 of Sesar+2007 on Fig. 9.

Next we illustrate on Fig. 10 what is the impact of selecting a region in the u-g vs g-r space in the χ^2 space. The color-selected quasar sample uses the following conservative boundaries :

$$-0.5 < u - g < 0.5 \quad (16)$$

$$0 < g - r < 0.5 \quad (17)$$

We illustrate the competitiveness of this quasar sample by plotting differential counts (a raw product often used as a first look into dataset in Quasar Luminosity Function studies) on Figs. 11 and 12

ACKNOWLEDGEMENTS

Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS Web Site is http://www.sdss.org/.

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.

This paper has been typeset from a TEX/LATEX file prepared by the author.

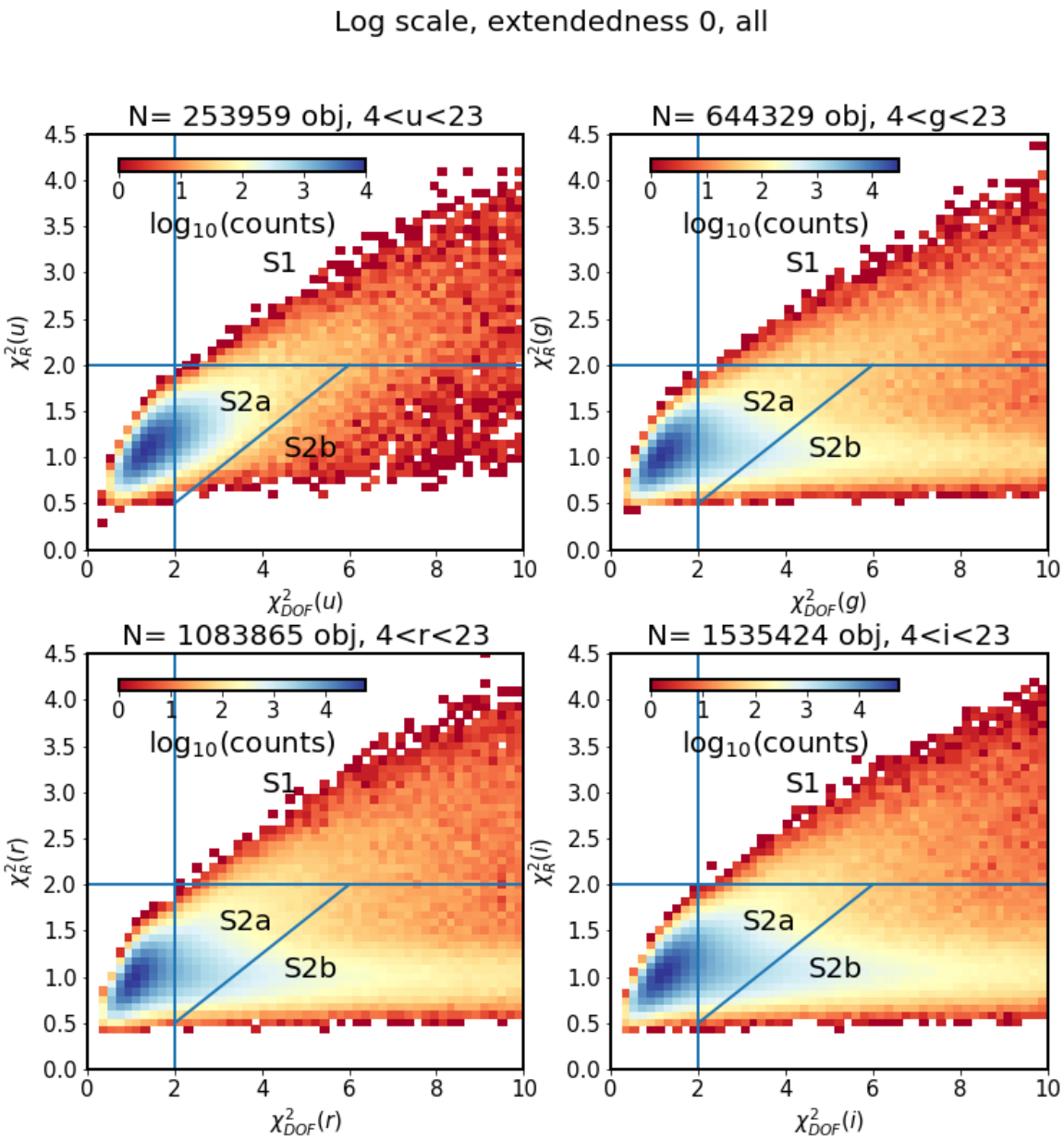

Figure 8. $\chi^2_{DOF} - \chi^2_R$ space calculated for data in u,g,r,i filters. Region S1 includes highly variable objects, S2a a transitional region, and S2b a region with possible variable candidates. We select variable candidates from S1 region, based on the g-band χ^2.

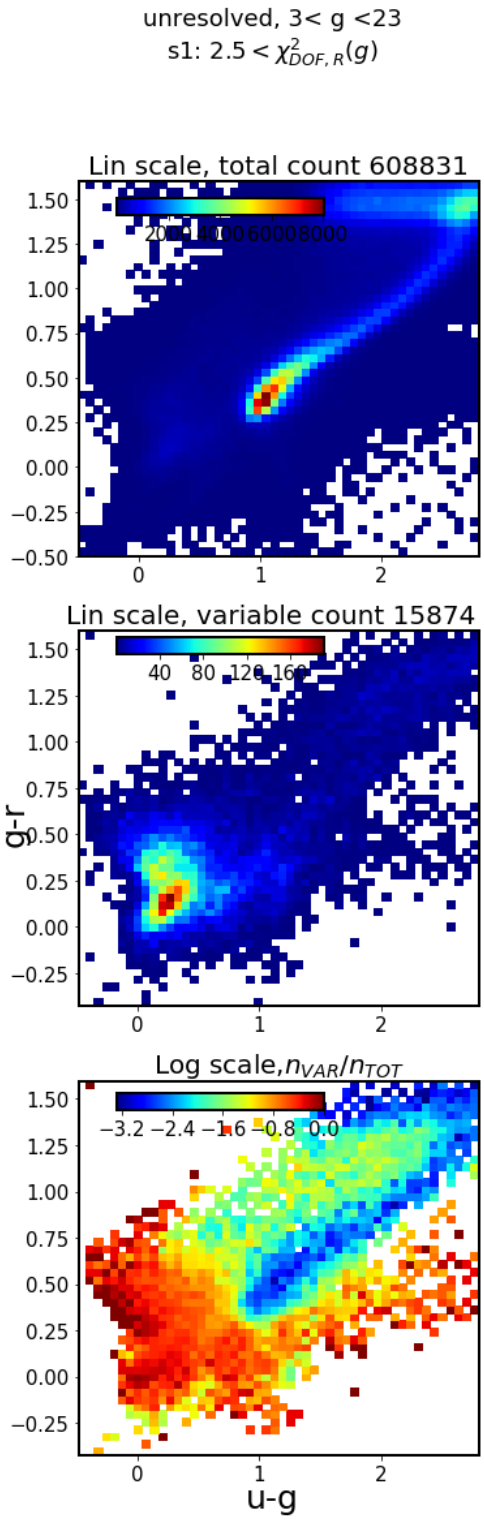

Figure 9. Reproduction of Fig.3 of Sesar+2007 with the new reprocessed S82 data. From top to bottom : total counts, variable counts, ratio of variable to total count.

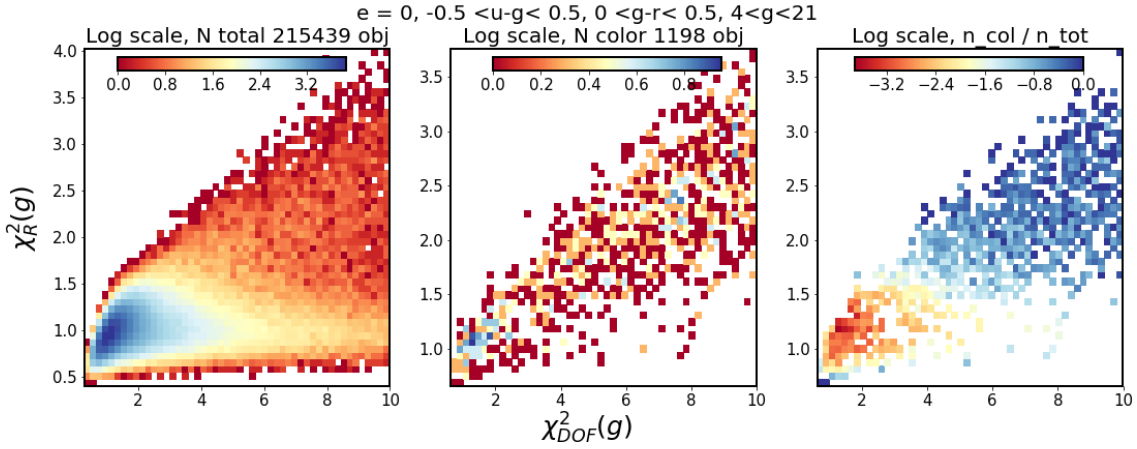

Figure 10. An impact of color selection in variability space. Left panel contains all objects with 4<g<21, and extendedness=0 . The middle panel depicts the count of those objects that fulfill color criteria $-0.5 < u-g < 0.5$ and $0 < g-r < 0.5$. The right panel shows counts of the ratio of number of color-selected objects to total number, i.e. ratio of middle to left panels.

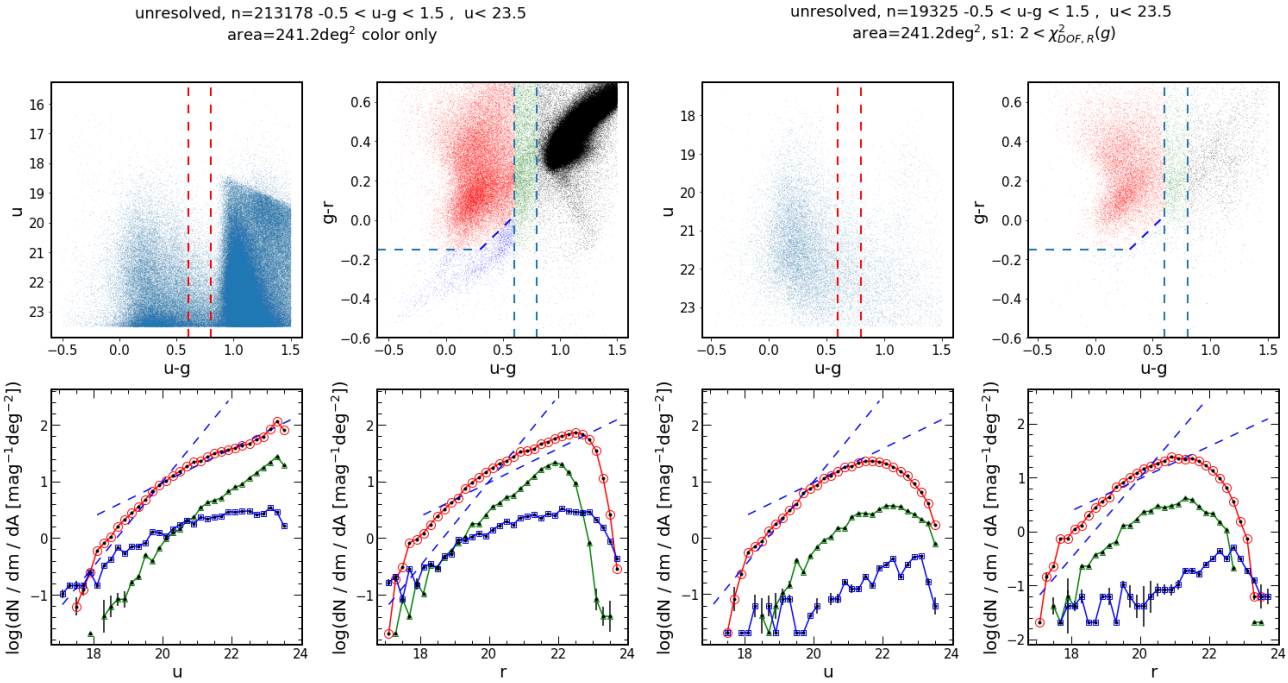

Figure 11. Two four-panel figures side by side for the ease of comparison. Both contain counts in the u vs u-g and u-g vs g-r, colored by regions marked on the panel, and differential counts in u and r (see Fig.1 of Ivezic+2004). The counts contain only objects as selected by u-g vs g-r color cuts, to colors on the differential counts panels corresponds to the colors on u-g vs g-r space. Red objects are mostly quasars, blue: hot stars (one can easily pick out two white dwarf branches, see Fig.23 and 24 in Ivezic+2007), green : transitional region, black : main sequence stars . Left set of four panels : only color-selected object . Right set : variability and color-selected objects. Note that including $\chi^2_{DOF,R}(g) > 2$ cut we vastly reduce contamination due to main sequence stars, and hot white dwarfs.

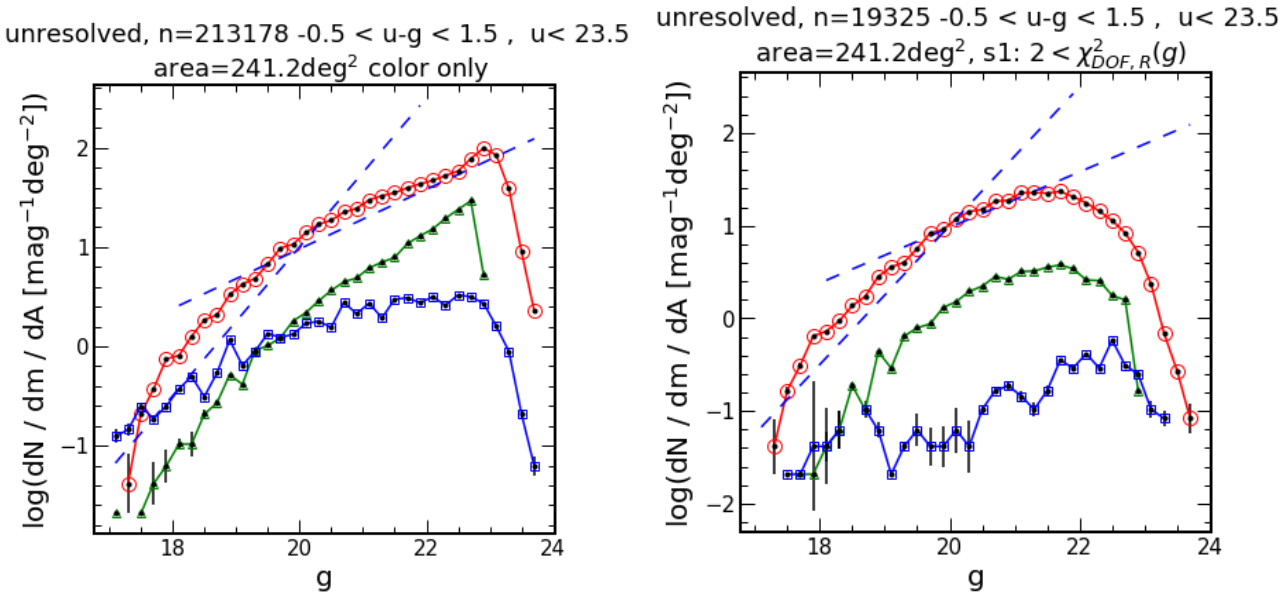

Figure 12. As Fig. 11, but showing counts in g band. Left : only color selection, right : color and variability selection.

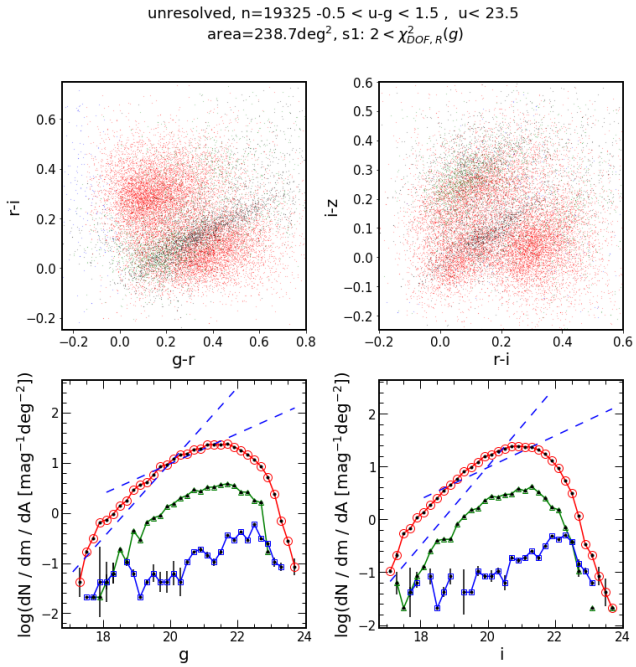

Figure 13. We also plot the counts in g-r vs r-i and r-i vs i-z, colored by u-g vs g-r regions as in Fig. 11

.