

# SDSS Stripe 82 : quasar variability from forced photometry

Krzysztof Suberlak,<sup>1\*</sup> Željko Ivezić,<sup>1</sup> Yusra AlSayyad,<sup>1</sup>

<sup>1</sup>*Department of Astronomy, University of Washington, Seattle, WA, United States*

Accepted XXX. Received YYY; in original form ZZZ

## ABSTRACT

### 1 INTRODUCTION

Many objects in the universe, from stellar to extragalactic scales, vary on timescales less than a few hundred years. Lightcurves carry a wealth of information allowing one to infer various physical properties of a planet or a galaxy. If a lightcurve is poorly sampled, the inferred characteristics are less certain. Yet, since all astronomical observations suffer from a detection threshold, a very faint variable object in some epochs may be undetectable. Forced photometry rescues the information from very faint epochs by performing a measurement in all epochs in a location from the co-added images. An inherent challenge to such set of measurements is an interpretation of noise-dominated flux. To circumvent this problem many studies apply a magnitude cutoff few magnitudes above the detection limit, which reduces the amount of available data. Indeed, in order to fully utilize information present in time-domain surveys, such as Large Scale Synoptic Telescope, Palomar Transient Factory, or Sloan Digital Sky Survey, and properly characterize faint variable objects, we need to properly handle the faint flux measurements. A new methodology would allow an unbiased study of such faint variable objects, including quasars, RR Lyrae, Cepheids, and a wealth of other variable sources. With the advent of precision time-domain astronomy surveys it is crucial to apply the best possible faint forced photometry algorithms and thus make full use of the data.

## 2 METHODS

### 2.1 Data Overview

#### 2.1.1 Stripe 82

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  $\pm 1.27^\circ$ , extending from R.A.  $\approx 20^h(320^\circ)$  to R.A.  $\approx 4^h(55^\circ)$  (Sesar+2010). 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 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  $-40^\circ < RA < +10^\circ$  and IN2P3 with  $+5^\circ < RA < +55^\circ$ . There is a  $5^\circ$  overlap, used to confirm that the data processing pipeline in both data centers yields identical data products. The entire strip was split into smaller patches

All epochs (individual images) were background-subtracted, and then scaled from the Digital Unit counts to fluxes by comparing standard objects against the Ivezić+2007 catalog (similar to Jiang+2014).

#### 2.1.2 Source Detection

Sources were detected in the i-band coadds. Each detection in the coadded images was assigned a `deepSourceId` (elsewhere called `objectId`). Considering a dense region with clumped stars and/or galaxies, the entire clump was considered as one parent source (with single `ParentSourceId`). 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. The result of this procedure were 40 million i-band detections down to  $3\sigma$ . 8 million of those are brighter than  $23^{rd}$  mag. Part of Stripe82 processed in NCSA yielded 20978391 detections (`iCoaddA11.csv`). The part that does not overlap with IN2P3 has 16520093 sources (`iCoaddPhotometryA11.csv`), of which 16514187 are brighter than  $30^{mag}$  (5906 less) (`DeepSourceNCSA_i_1t300.csv`).

#### 2.1.3 Forced Photometry

On positions specified by the detection data AlSayyad+2015 performed forced photometry in all SDSS photometric bands, on the individual epoch images. It is different from image differences technique, where the photometry is done on a difference between a coadd and an individual epoch image. The total number of photometric measurements (combining NCSA and IN2P3) was  $(40 \text{ million i-band detections}) \times (80 \text{ epochs}) \times (5 \text{ filters}) = 16 \text{ billion measurements}$ , including  $(8 \text{ million i-band detections } i < 23) \times (80 \text{ epochs}) \times (5 \text{ filters}) = 3.2 \text{ billion measurements brighter than } 23^{rd} \text{ mag}$ . For each patch the raw lightcurves contain the `id`, `objectId`, `exposure_id`, `mjd`, `psfFlux`, `psfFluxErr`, sorted by `objectId`, measuring flux in  $[ergs/cm^2/sec/Hz]$  (`rawDataFPSplit/bandPatchStart_PatchEnd.csv`).

## 2.2 Analysis

### 2.2.1 Faint Sources

Each measurement of flux is affected by the background noise. The bright optical background can have two contributions to faint source detection. First, the oscillation of background around the mean may lead to spurious detections. We can understand it by modelling the distribution of background counts as a Gaussian centered around the mean value  $B_0$ :  $B - B_0 \sim \mathcal{N}(0, \sigma_B)$ . The noise itself is Poissonian, and for a large number of photons hitting the detector the width of the distribution  $\sigma_B$  is proportional to the square-root of counts:  $\sqrt{B}$ . Thus on a 4kx4k CCD, similar to those used by SDSS, we would expect about 1 false detection in a million at  $5\sigma$  level - 16 per CCD. Second contribution of background variation is the unphysically low flux at source location. Since the background oscillates around a mean value, an individual epoch may have a lower-than-average background value which after mean background subtraction yields a negative value of flux at source location. Thus especially for variable sources, the location where it was detected in coadds may have a negative flux value because if the source is below detection threshold in an individual epoch, we are measuring the background noise oscillation. The background noise can significantly affect the measurement if it is as strong as the flux of a source in some epochs.

We recalculate the flux for all 'faint' ( $< 2\sigma$ ) sources by truncating the flux likelihood function. Each forced photometry 'observed' flux in a single epoch can be thought of as a mean of the 'intrinsic' flux likelihood. Such Gaussian likelihood is centered on the 'observed' flux  $F$ , with a width corresponding to the measurement error  $\sigma_F$ :  $L(F) \sim \mathcal{N}(F, \sigma_F)$ . For example, the bright sources can be thought of as having a very narrow Gaussian likelihood, with a width on the level of  $1 - 2\% \approx 0.01 - 0.02$  mag. On the other hand, the faint sources, with larger uncertainties, have very wide Gaussian tails. The tails of their likelihood can extend below zero, and thus the forced-photometry measurement is underestimating the flux. In reality no physical source of flux can have a negative flux. This negative portion of the likelihood stems from the background fluctuation, or small signal to noise ratio. We address the issue of unphysical likelihood and underestimated flux by recalculating flux for all sources with less than ' $2\sigma$ ' detection. This means that we recalculate flux for sources where there is a 2% or higher probability of negative flux.

In our treatment we are explicitly using a prior understanding of the flux behavior of any astrophysical object. Without any further knowledge about the nature of the source, flat prior is the least informative Bayesian prior. Any additional information about the nature of object, and thus expected variability pattern, could affect the choice of prior to be more specific. For instance, consider a sinusoidal flux variability. If the flux of an object over many epochs is expected to vary in a sinusoidal fashion, i.e.  $F(t) = F_{min} + \sin(t)$ , the probability of a given flux measurement is a cosine, ranging from  $F_{min}$  to  $F_{max}$ . With that prior, without any measurement taken, the flux of the object is most likely  $(F_{min} + F_{max})/2$ , i.e. at the peak of the cosine likelihood function. However, as soon as one measures  $(F_i, e_i)$  from that source, the probability distribution of a flux at that epoch becomes a convolution of cosine prior information

with the Gaussian curve of width  $e_i$ , centered on  $F_i$  (assuming Gaussian errors). However, without any a-priori information about the variability pattern of the considered object, the least informative Bayesian prior we can impose is a flat one:  $p(F) = 0$  for  $F < 0$ , and 1 elsewhere. Thus the posterior probability is

$$p(F|data) \propto L(F|data)p(F) \quad (1)$$

To test the method we generate fiducial lightcurves (DRW / sinusoidal), with a uniform sampling ( $N = 100 \div 1000$ ). Based on the generated flux ( $F_{true}$ ) we define the  $5\sigma$  level as the robust 25-th percentile (or median) of the ensemble  $F_{true}$  distribution:  $\sigma_F = (1/5)F_{25\%}$  (in reality,  $\sigma$  increases for fainter observations, but this is a good approximation). We define  $F_{obs} = F_{true} + F_{noise}$ , where the Gaussian noise  $F_{noise} = \sigma_F \mathcal{N}(0, 1)$  was added to each point. For a weak signal, defined as  $F_{obs}^i < 2\sigma_F$ , we consider  $p(F)$  - a Gaussian likelihood associated with  $i$ -th measurement:  $p_i(F) = \mathcal{N}(\mu = F_{obs}^i, \sigma = \sigma_F)$ . Each measurement  $F_{obs}$  is a mean of this likelihood:  $F_{obs} = \langle p_i(F) \rangle$ . We call it  $p(F)$  for short:

For each epoch, based on the raw forced photometry measurement, we calculate new descriptors of faint fluxes. We define a faint measurement by  $F_i < 2\sigma_F$ , i.e. where the flux is less than twice the flux error. We assume that the flux is the mean of the Gaussian likelihood  $p_i(F) = \mathcal{N}(\mu = F_i, \sigma = \sigma_F)$ :

$$p(F) = \frac{1}{\sqrt{2\pi\sigma_F^2}} \exp\left(-\frac{(F - \mu)^2}{2\sigma_F^2}\right) \quad (2)$$

so that  $F_i = \langle p(F) \rangle$ . For faint measurements we truncate the negative part of  $p(F)$ , and recalculate the mean, median, rms, and  $2\sigma$  level. Thus the mean is

$$F_{mean} = \frac{\int_0^\infty F p(F) dF}{\int_0^\infty p(F) dF} \quad (3)$$

where we normalized the truncated Gaussian likelihood.

We define the median as

$$\int_0^{F_{median}} p(F) dF = \int_{F_{median}}^\infty p(F) dF \quad (4)$$

The rms level is

$$F_{rms}^2 = \frac{\int_0^\infty (F - F_{mean})^2 p(F) dF}{\int_0^\infty p(F) dF} \quad (5)$$

Finally, since for a Gaussian distribution the area contained between  $\mu \pm \sigma$  is 95.5% of the total area under the curve, for the truncated Gaussian we define the  $2\sigma$  level as:

$$\int_{F_{2\sigma}}^\infty p(F) dF = 0.05 * \int_0^\infty p(F) dF \quad (6)$$

Both for median and for the  $2\sigma$  level the normalization cancels out. For details, see Appendix ...

### 2.2.2 Variability

In what follows, for faint measurements we choose to replace  $F$  with  $F_{mean}$ . For all objects we calculate statistics based on entire lightcurves. We denote  $(F_{obs}, \sigma_{obs})$  as  $(x_i, \sigma_i)$ :

- mean weighted by  $w_i = (\sigma_i)^{-2}$

$$\bar{x} = \frac{\sum_{i=1}^N w_i x_i}{\sum_{i=1}^N w_i} \quad (7)$$

- weighted mean error

$$\sigma_{\bar{x}} = \left( \sum w_i \right)^{-1/2} \quad (8)$$

- weighted standard deviation

$$\sigma_{st.dev.w.} = \left( \frac{\sum w_i (x_i - \bar{x})^2}{\frac{N-1}{N} \sum w_i} \right)^{1/2} \quad (9)$$

- error on standard deviation

$$\sigma_s = \frac{1}{\sqrt{2}} \sqrt{\frac{N}{N-1}} \sigma_{\bar{x}} \quad (10)$$

- robust interquartile width

$$\sigma_G = 0.7414 * (75\% - 25\%) \quad (11)$$

- median as the 50-th percentile, median error

$$\sigma_{median} = \sqrt{\pi/2} \sigma_{\bar{x}} \quad (12)$$

Variations in object brightness have two main origins: an error-induced noise, and an intrinsic variability. A lightcurve consists of a set of  $N$  measurements of brightness  $x_i$  with errors  $e_i$ . In this analysis we assume that  $x_i$  are drawn from a Gaussian distribution  $\mathcal{N}(\mu, \sigma)$ , and that errors  $e_i$  are homoscedastic. We describe this distribution with two parameters: mean  $\mu$ , and width  $\sigma$ . To increase efficiency, we employ a two-step approach after Ivezić+2014. First, we find approximate values of  $\mu_0$  and  $\sigma_0$ , and then we evaluate the full logarithm of the posterior pdf in the vicinity of the approximate solution. With a Bayesian approach, we find  $\mu_{full}, \sigma_{full}$  by maximizing the posterior probability distribution function (pdf) of  $\mu, \sigma$  given  $x_i$  and  $e_i$ :  $p(\mu|x_i, \sigma_i)$  (see Fig. 2.2.2, and Appendix B for the detailed calculation).

For each lightcurve, we also calculate mean-based  $\chi_{DOF}^2$  and median-based  $\chi_R^2$  (the latter is more robust against any outliers in the distribution):

$$\chi_{dof}^2 = \frac{1}{N-1} \sum \left( \frac{x_i - \langle x_i \rangle}{e_i} \right)^2 \quad (13)$$

and

$$\chi_R^2 = 0.7414(Z_{75\%} - Z_{25\%}) \quad (14)$$

with  $Z = (x_i - \text{median}(x_i))/e_i$ .

On Fig. 2.2.2 we plot the stages of calculating  $\mu$  and  $\sigma$ . Left and middle panels compare the two methods of calculating variability parameters. The initial approximation (dashed) is based on bootstrapped resampling of  $(x_i, e_i)$  points from the lightcurve. By randomly resampling the lightcurve  $M$  times, instead of a single sample with  $N \approx 10-70$  points we have  $M$  samples. The histogram of  $M = 1000$  values for  $\mu, \sigma$  from resampling is plotted with dashed lines. We use the approximate values to provide bounds for the 200x70 grid of  $\mu, \sigma$ , used to evaluate the full posterior likelihood density function (right panel). This ensures that, despite using a coarse grid to improve computational speed, we still resolve the peak of the underlying distribution.

All variability parameters describe in a certain way the lightcurve variability.  $\sigma_{full}$  corresponds to the spread of the flux distribution. For a non-variable source,  $\chi^2$  would be centered about 1, with a width  $\sigma$  of  $\sqrt{2/N}$ , where  $N$  is the number of points per lightcurve. Thus we would expect that for any distribution 50% of sources would have  $\chi^2 > 1$ . Therefore a "3 $\sigma$ " variability detection would require  $\chi^2 > \chi_{limit}^2$ , with  $\chi_{limit}^2 = 1 + 3\sqrt{2/N}$ . We call a source a 'robust variability candidate' if  $\sigma_{full} > 0$  and  $(\chi_{DOF}^2 > \chi_{limit}^2 \text{ or } \chi_R^2 > \chi_{limit}^2)$ , i.e. we require the full  $\sigma$  to be larger than 0, and either robust or DOF  $\chi^2$  to be larger than the limiting  $\chi^2$ .

Initially, we evaluate variability parameters  $\mu_{full}, \sigma_{full}, \chi_{dof}^2$ , and  $\chi_R^2$  based on all points of the lightcurve. Only for variable sources, as defined above, we calculate the variability parameters for the seasonally-binned portions of the lightcurve. The  $\chi_{DOF}^2$  vs.  $\chi_R^2$  plotted on Fig. ... shows that these distributions are similar for seasons as well as full lightcurves.

### 2.2.3 Colors

Since the reported fluxes are not extinction-corrected, we use a table of  $E(B-V)$  in a direction of a given source to correct for the galactic extinction. We use the formula  $x_{corr} = x_{obs} + A_x * E(B-V)$ , where  $x$  is u,g,r,i,z, and  $A_x$  is 5.155, 3.793, 2.751, 2.086, 1.479 for each filter respectively [Schlegel 98,  $A_v$  are for  $RV = 3.1$ , also suggested by Eddie Schlafly]

Colors  $x-y$  for an object with observations over many epochs are defined as the difference in magnitudes  $m_x - m_y$ . To find  $m_x$ , we need to define the average brightness of an object in a given filter. With a special treatment of faint sources, substituting  $(F_{obs}, \sigma_F)$  for each faint observation by  $(\langle F_{exp} \rangle, rms)$ , we analyse updated lightcurves, addressing sparse sampling (see Fig. 2.2.3).

Thus for a given object we average all sparser observations prior before SDSS-III, and calculate annual averages for all subsequent years. We calculate weighted mean and the rms as

$$\langle F \rangle = \frac{\sum w_i F_i}{\sum w_i} \quad \sigma_{\langle F \rangle} = \left( \sum w_i \right)^{-1/2} \quad (15)$$

with weights as  $w_i = 1/\sigma_i^2$ . We also calculate the robust median and the median error:  $\sqrt{\pi/2} \sigma_F$  [robust  $\sigma_G = 0.7414 * (75\% - 25\%)$ , based on the interquartile range]. Then lightcurve for a given object is reduced to one  $(F_i, \sigma_i)$  point prior to March 2006, and a single point per every subsequent year, where  $(F_i, \sigma_i)$  is  $(\text{mean}, \text{meanErr})$  or  $(\text{median}, \text{medianErr})$ .

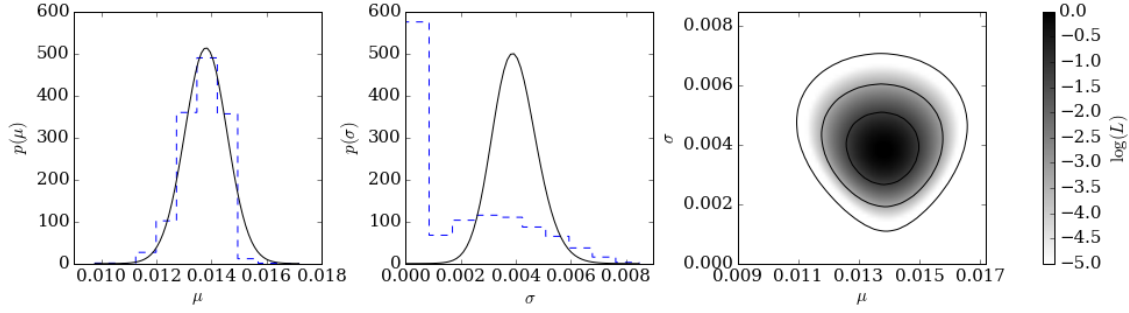
The resulting average flux is converted to magnitude, and the color is  $c = m_x - m_y$ , with combined errors of band lightcurves added in quadrature

## 3 RESULTS

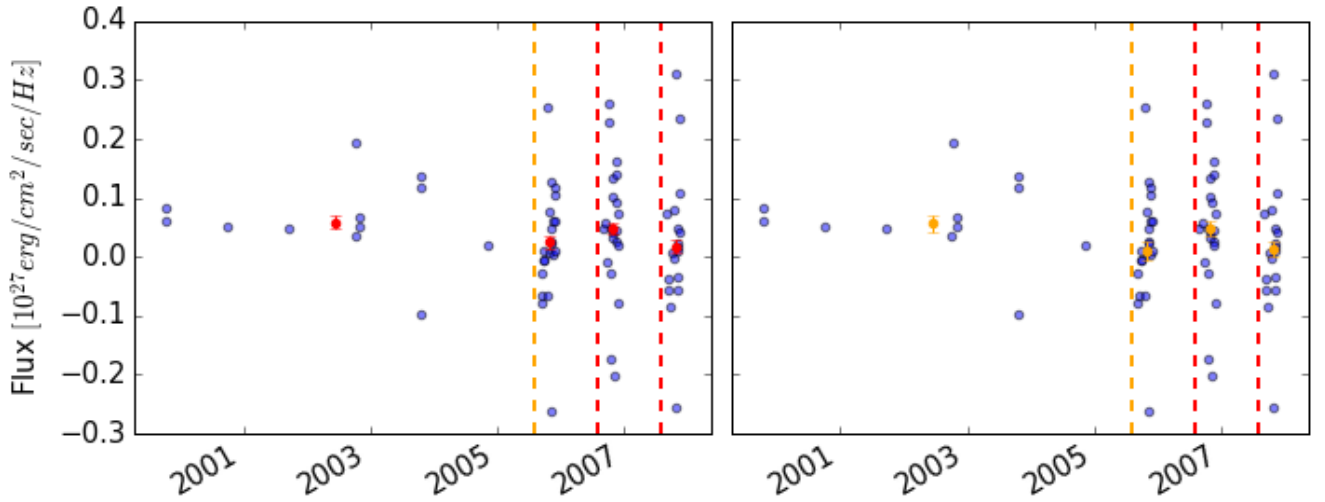
## 4 CONCLUSIONS

## 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



**Figure 1.** Two-step approach to finding  $\mu$  and  $\sigma$  via  $\mu_0$  and  $\sigma_0$  for an object 217720894888346446. In this calculation we use raw psf flux, before employing the faint source treatment outlined in Section 2.2.1. On the left and middle panels, solid lines trace marginalized posterior pdfs for  $\mu$  and  $\sigma$ , while dashed lines depict histogram distributions of 10,000 bootstrap resamples for  $\mu_0$  and  $\sigma_0$ . The right panel shows the logarithm of the posterior probability density function for  $\mu$  and  $\sigma$ .



**Figure 2.** A plot showing an outcome of seasonal averaging for an object id 217720894888346425. The left panel (red dots) shows (mean, meanErr), and the right panel (orange) shows (median, medianErr), instead of seasonal points (blue). Vertical dashed lines as on Fig. 2.2.3

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.

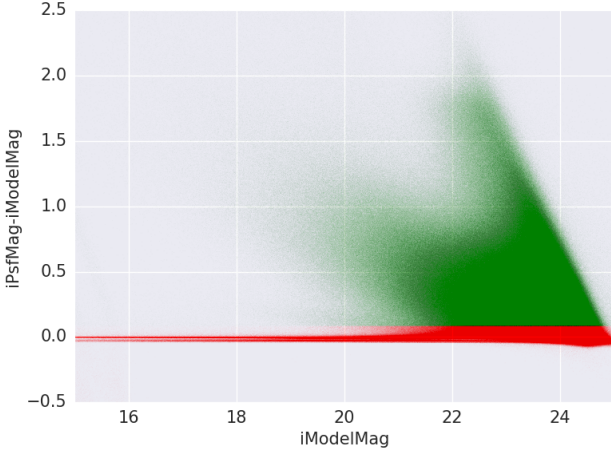
## APPENDIX A: TREATMENT OF FAINT SOURCES

In our calculations we used the `scipy` implementation of the following often used integrals of Gaussian distributions :

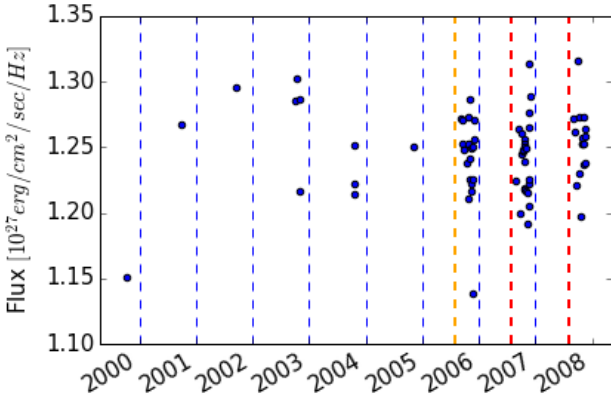
- cumulative density function, that is an area under the Gaussian distribution from  $-\infty$  to  $x_0$  :

$$\text{cdf}(x_0) = \int_{-\infty}^{x_0} \mathcal{N}(\mu, \sigma) dx = \int_{-\infty}^{x_0} \frac{\exp(-(x-\mu)^2/2\sigma^2)}{\sqrt{2\pi\sigma^2}} dx \quad (\text{A1})$$

- point percent function, that is an inverse of the cumulative density function: if  $A = \text{cdf}(x_0)$ , then  $x_0 = \text{ppf}(A)$
- survival function (also known as the complementary cumulative distribution function), that is an area under a Gaussian distribution from  $x_0$  to  $\infty$



**Figure 3.** A plot showing NCSA sources detected in coadds, removing the outliers beyond the edges of the plot. The coloring corresponds to the `extendedness` parameter calculated in the pipeline based on the `iPsFMag-iModelMag` : red being 0 (compact), and green being 1 (extended). As `iModelMag` increases, the separation becomes less certain, as more distant galaxies are more compact.



**Figure 4.** A plot showing an example lightcurve for an object id 217720894888346422. Jan 1st of each year (blue), August 1st of 2005 (orange) and August 1st of each subsequent year (red) is indicated by vertical dashed lines. Observations prior to August 1st of 2005 have sparser cadence, whereas those after that date have more frequent observations. This is due to the SDSS-III Supernova Survey which begun Sept 1st 2005. All points to the left of August 1st 2005 (orange line) are averaged together. Points to the right of August 1st 2005 are seasonally averaged.

$$\text{sf}(x_0) = \int_{x_0}^{\infty} \mathcal{N}(\mu, \sigma) dx = \int_{-\infty}^{\infty} \mathcal{N}(\mu, \sigma) dx - \int_{-\infty}^{x_0} \mathcal{N}(\mu, \sigma) dx = 1 - \text{cdf}(x_0) \quad (\text{A2})$$

In our faint flux treatment we assume that each flux measurement has an associated Gaussian likelihood, and that the width and mean of the likelihood correspond to the measured flux and the measurement error respectively.

For a source where signal-to-noise  $< 2$  (in our case, ratio

of flux to error), we remove the negative portion of the likelihood, since there is no physical likelihood that a flux would be negative. Thus for mean, we integrate from 0 instead of  $-\infty$  :

$$F_{\text{mean}} = \frac{\int_0^{\infty} F p(F) dF}{\int_0^{\infty} p(F) dF} = I_0 / I_1 \quad (\text{A3})$$

where we need to normalize by the integral over the positive part of the Gaussian likelihood.

We evaluate

$$I_0 = \int_0^{\infty} \frac{F}{\sqrt{2\pi}\sigma_F^2} \exp\left(-\frac{(F-F_{\text{obs}})^2}{2\sigma_F^2}\right) dF = \frac{\sigma_F}{\sqrt{2\pi}} \exp\left(-\frac{F_{\text{obs}}^2}{2\sigma_F^2}\right) + F_{\text{obs}} \text{sf}\left(\frac{-F_{\text{obs}}}{\sigma_F}\right) \quad (\text{A4})$$

and

$$I_1 = \int_0^{\infty} p(F) dF = \int_0^{\infty} \frac{\exp\left(-\frac{(F-F_{\text{obs}})^2}{2\sigma_F^2}\right)}{\sqrt{2\pi}\sigma_F^2} dF = \text{sf}(-F_{\text{obs}}/\sigma_F) \quad (\text{A5})$$

so that

$$x_{\text{mean}} = \frac{\exp(-x_{\text{obs}}^2/2)}{\text{sf}(-x_{\text{obs}})\sqrt{2\pi}} + x_{\text{obs}} \quad (\text{A6})$$

where we scaled  $F_{\text{obs}}$  by  $\sigma_F$  (i.e.  $F_{\text{mean}} = x_{\text{mean}} \cdot \sigma_F$ ).

To find the median and the  $2\sigma$  level we transform from  $F$  space to  $x$  space, scaling by  $\sigma_F$ , so that  $x = F/\sigma_F$ , and thus the likelihood  $p(x) \sim \mathcal{N}(x_{\text{obs}}, 1)$  is :

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-x_{\text{obs}})^2}{2}\right) \quad (\text{A7})$$

We then transform from  $x$  to  $z$  space, with a translation by  $x_{\text{obs}} = F_{\text{obs}}/\sigma_F$  :  $z = x - x_{\text{obs}}$ , so that now  $p(z) \sim \mathcal{N}(0, 1)$ :

$$p(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \quad (\text{A8})$$

In  $z$ -space, the median from

$$\int_0^{x_{\text{med}}} p(x) dx = \int_{x_{\text{med}}}^{\infty} p(x) dx \quad (\text{A9})$$

becomes

$$\int_{x_0}^{z_{\text{med}}} p(z) dz = \int_{z_{\text{med}}}^{\infty} p(z) dz \quad (\text{A10})$$

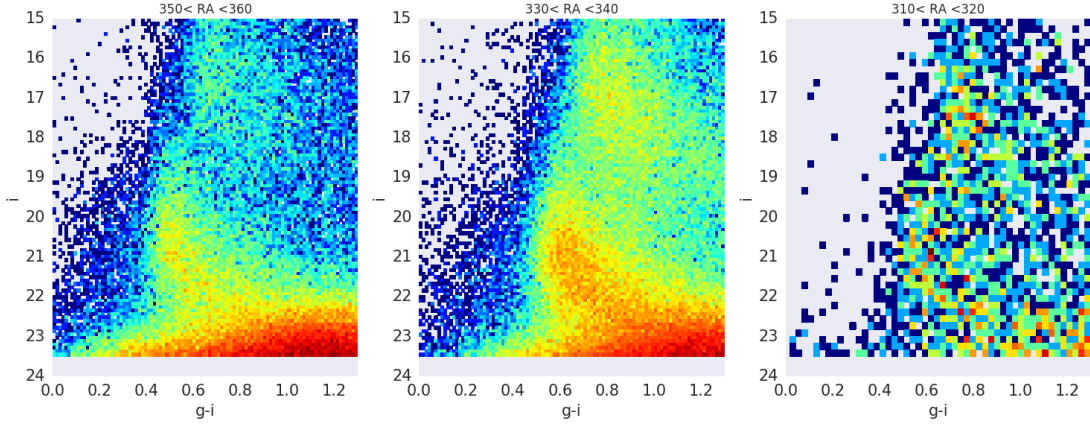
with  $x_0 = -x_{\text{obs}}$

We evaluate  $z_{\text{med}}$  analytically - the right hand side is the survival function :

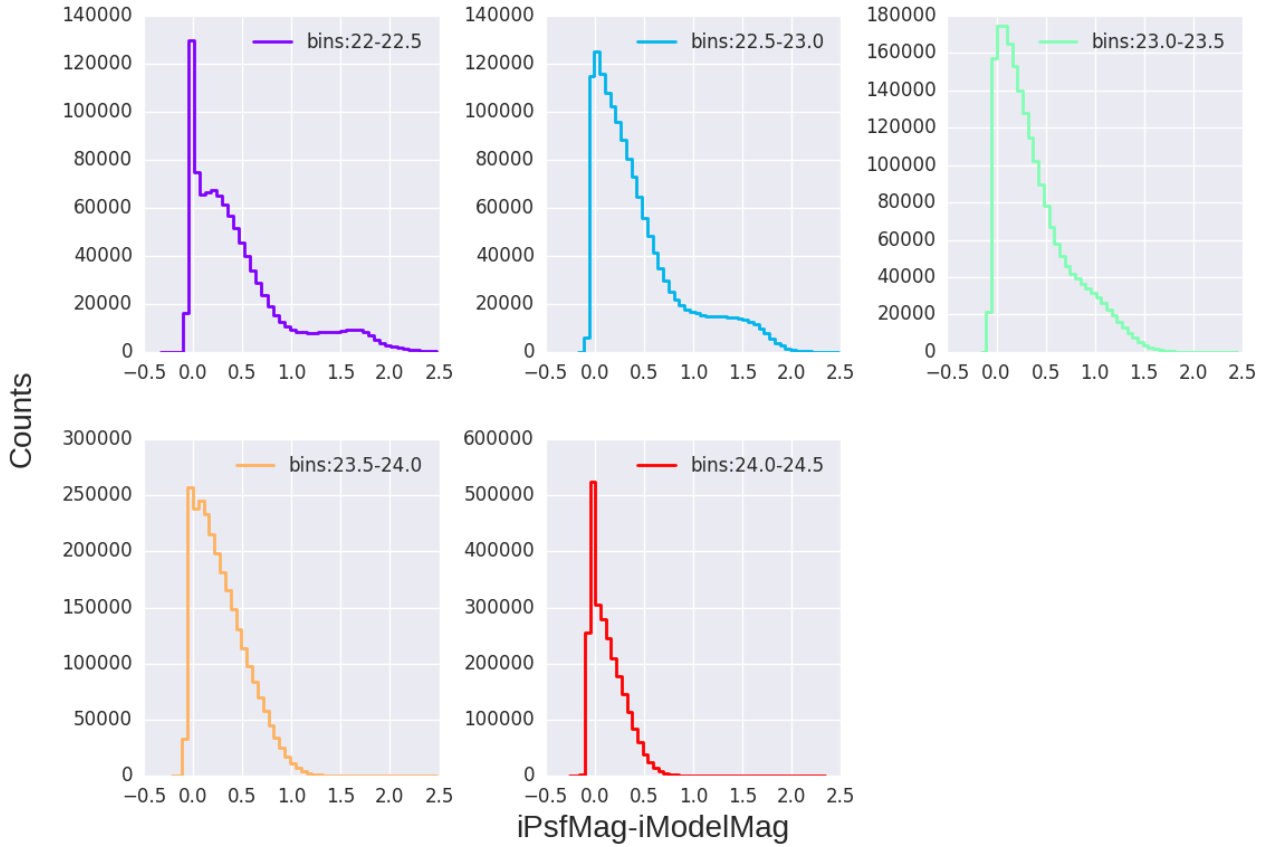
$$\int_{z_{\text{med}}}^{\infty} p(z) dz = \text{sf}(z_{\text{med}}) \quad (\text{A11})$$

and the left hand side, assuming that the median  $z_{\text{med}} > x_0$ , is :





**Figure 5.** A color-magnitude plot, reproducing the results of Sesar+2010, Fig.23. We show here only NCSA-processed sources, which is why certain RA ranges are omitted or have less sources. We only select sources with `extendedness=0` parameter (stars). The scale is showing the  $\log_{10}$  of count. All sources have their colors corrected for extinction. On first two panels the features of Sagittarius Stream are clearly visible.



**Figure 6.** The histograms show the count of sources in 5 magnitude bins, corresponding to the vertical cut through Fig. 2.2.3. It helps to verify how well can the extended and compact sources be separated based solely on the `iPsMag-iModelMag`

$$\int_{x_0}^{z_{med}} p(z)dz = \int_{-\infty}^{z_{med}} p(z)dz - \int_{-\infty}^{x_0} p(z)dz = \text{cdf}(z_{med}) - \text{cdf}(x_0) \quad (\text{A12})$$

Rearranging, and using the percent point function (ppf) we find:

$$z_{med} = \text{ppf}\left(\frac{1 + \text{cdf}(x_0)}{2}\right) \quad (\text{A13})$$

and transforming back to  $F$  space:

$$F_{med} = F_{obs} + \sigma_F \text{ppf}\left(\frac{1 + \text{cdf}(x_0)}{2}\right) \quad (\text{A14})$$

with  $x_0$  and  $x_{obs}$  as above. We also normalize this expression by  $\int_0^\infty p(F)dF$  and  $\sigma_F$ :

$$x_{med}^{norm} = x_{obs} \quad (\text{A15})$$

In  $z$  space, the  $2\sigma$  areas  $A$  and  $B$  are:

$A = \text{sf}(x_0)$  and  $B = \text{sf}(z_B)$ , so to find  $z_B$  we use the inverse survival function  $\text{isf} : z_B = \text{isf}(0.05A)$ . Thus transforming back to  $F$ -space we have:

$$F_{2\sigma} = F_{obs} + \sigma_F (\text{isf}(0.05 \text{sf}(x_0))) \quad (\text{A16})$$

We also find the root-mean-square:

$$F_{rms}^2 = \frac{\int_0^\infty (F - F_{mean})^2 p(F)dF}{\int_0^\infty p(F)dF} = I_0/I_1 \quad (\text{A17})$$

this can be evaluated by numerical integration, scaling by  $\sigma_F$ , so that  $x_{mean} = F_{mean}/\sigma_F$ ,  $x_{obs} = F_{obs}/\sigma_F$ :

$$F_{rms}^2 = \frac{\sigma_F^2 \int_0^\infty (x - x_{mean})^2 \exp(-(x - x_{obs})^2/2) dx}{\int_0^\infty \exp(-(x - x_{obs})^2/2) dx} \quad (\text{A18})$$

We derived analytical expression for rms:

$$x_{rms} = \left(\frac{I_0}{I_1 \sigma_F^2}\right)^{1/2} \quad (\text{A19})$$

where  $I_1$  is our normalization, as in calculation of  $F_{mean}$ , and as

$$\frac{I_0}{\sigma_F^2} = \frac{1}{2} \text{erf}\left(\frac{x_{obs}}{\sqrt{2}}\right) + \frac{1}{\sqrt{2\pi}} e^{(-x_{obs}^2/2)} (2\Delta x - x_{obs}) + (\Delta x)^2 \text{sf}(-x_{obs}) \quad (\text{A20})$$

with  $\Delta x = x_{obs} - x_{mean}$

## APPENDIX B: CHARACTERIZING VARIABILITY

We further characterize the variability of lightcurves by calculating  $\sigma_0$  (the approximate value), and  $\sigma_{full}$ , following Ivezić+2014, chapter 5.

$\sigma_0$  is found in the following way: if by  $(x_i, e_i)$  we denote the measurement and associated error, then the bootstrapped sampling of  $(x_i, e_i)$  is sampling each vector at a number of  $N$  random indices (eg.  $N=1000$ ). Thus instead of  $x_i$  which may include only  $N=10$  measurements, we have  $x_{i,boot}$  which has  $N=1000$  random samples. Median is the 50-th percentile of any sample. Following [Ivezić+2014], chapter 5, we use the sample median to estimate  $\mu_0 = \text{median}(x_{i,boot})$ , and an interquartile range width estimator to estimate the standard deviation:  $\sigma_G = 0.7413(X_{75\%} - X_{25\%})$  for  $X = x_{i,boot}$ . With the median error  $e_{50} = \text{median}(e_{i,boot})$ , we estimate  $\sigma_0$  as:

$$\sigma_0 = (\text{variance}_{approx})^{1/2} = (\zeta^2 \cdot \sigma_G^2 - e_{50}^2)^{1/2} \quad (\text{B1})$$

where

$$\zeta = \frac{\text{median}(\tilde{\sigma}_i)}{\text{mean}(\tilde{\sigma}_i)} \quad (\text{B2})$$

and

$$\tilde{\sigma}_i = (\widetilde{\text{variance}})^{1/2} = (\sigma_G^2 + e_i^2 - e_{50}^2)^{1/2} \quad (\text{B3})$$

For the marginalized  $\sigma_{full}$ , we calculate logarithm of the posterior probability distribution for the grid of  $\mu$  and  $\sigma$  values as:

$$\log L = -0.5 \sum \left( \ln(\sigma^2 + e_i^2) + \frac{(x_i - \mu)^2}{(\sigma^2 + e_i^2)} \right) \quad (\text{B4})$$

We shift the maximum value of  $\log L$  by subtracting the maximum value of  $\log L$ , thus calculating the likelihood:

$$L = e^{\log L - \max(\log L)} \quad (\text{B5})$$

We then marginalize over  $\mu$  or  $\sigma$ :

$$p(\sigma) = \sum_{\mu} (L_{\sigma, \mu}) \quad p(\mu) = \sum_{\sigma} (L_{\sigma, \mu}) \quad (\text{B6})$$

and normalize the probability:

$$p_{norm}(\sigma) = \frac{p(\sigma)}{\int p(\sigma) d\sigma} \quad p_{norm}(\mu) = \frac{p(\mu)}{\int p(\mu) d\mu} \quad (\text{B7})$$

To characterize lightcurve variability we first calculate for the entire lightcurve of an object the approximate  $\mu_0$  and  $\sigma_0$  using bootstrapped resampling of the lightcurve. This yields the boundaries for the more exact calculation of the full 2D log-likelihood performed on a grid of  $\mu$  and  $\sigma$  values. Thus the more accurate  $\sigma_{full}$  and  $\mu_{full}$  are found as a maximum of the 2D log-likelihood distribution (see Fig. 2.2.2).

## APPENDIX C: MAKING OF UGRIZ METRICS

Colors can be calculated in two ways: using the median of forced photometry over all epochs (object detected in coadded i-band has photometry in all epochs:

`ugrizMetrics.csv`), or the median over single-epoch detections (only when an object was above the detection threshold for a single epoch : `medianPhotometry.csv`). The median over all detections will be biased (especially for faint sources) towards higher brightness. On the other hand, the median over all epochs will be more representative of the true brightness of an object in a given filter. If a median brightness is negative, we can use zero point magnitudes and in such cases median over all epochs will be an upper limit on brightness, but still less biased than median over all detections. Therefore we choose to use median over all epochs to calculate colors (see Fig. 3 for an example).

#### APPENDIX D: ZERO FLUX MAGNITUDES

If the median flux of an object over all epochs is negative (an outcome of forced photometry on fluctuating noise), we cannot define its magnitude in that filter. In such situation one can revert to using for each negative flux the zero point magnitude ( $m_1$ ) - the magnitude for a source with a flux of 1 count per sec, different for each exposure. The zero point magnitude for each exposure with negative flux is calculated from the Flux of 0 magnitude source,  $F_0$ , as  $m_1 = 2.5 \log_{10} F_0$ . For that object the new median magnitude in that filter will be the upper limit. We did not use this method, since a better way is to calculate the  $2 - \sigma$  flux limit for each flux measurement  $< 2\sigma : F_{2\sigma}$ .

#### APPENDIX E: LIGHTCURVE METRICS

For each object we calculate lightcurve-derived metrics. Denoting `psfFlux` and `psfFluxErr` as  $y$  and  $\sigma_y$ , we find the number of measurements per lightcurve ( $N$ ), the mean flux, the median flux (the 50th quartile), median flux error `e50`, mean flux error `e_mean`,  $\sigma_G$  (based on interquartile flux range  $0.7413(q75-q25)$ ),  $\chi^2$ :

$$\frac{1}{N-1} \sum \left( \frac{y - \text{mean}(y)}{\sigma_y} \right)^2 \quad (\text{E1})$$

mean weighted by the `WVar` - the inverse variance (`WeightedMean`), and the standard deviation weighted by inverse variance and corrected for intrinsic scatter (`WeightedStdCorr`):

$$\text{WVar} = \left( \sum \frac{1}{\sigma_y^2} \right)^{-1} \quad (\text{E2})$$

$$\text{WeightedMean} = \text{WVar} \sum \frac{y}{\sigma_y^2} \quad (\text{E3})$$

$$\text{WeightedStdCorr} = \left[ \frac{\text{WVar}}{N-1} \sum \frac{(y - \text{WeightedMean})^2}{\sigma_y^2} \right]^{1/2} \quad (\text{E4})$$

From these metrics, we can calculate the catalog photometry :

$$\text{median\_mag} = -2.5 \log_{10} \text{median} - 48.6 \quad (\text{E5})$$

There are 5892054 sources with catalog photometry brighter than 23 mag (`ugrizMetrics.csv`).

We can also calculate median photometry over all individual epochs detections, cross-matched by extinction tables [HOW ? ] There are 12373162 sources with median photometry, matched with E(B-V) data (`medianPhotometry.csv`).

For each band we calculate metrics describing the lightcurve behavior for a given band, including the Butler & Bloom classifier, which can for high S/N objects, where it has a good discriminating power. It's advantage over the full DRW analysis for each lightcurve is that by assuming a range of  $\tau$ , amplitude, expected for a DRW for a QSO, we calculate the likelihood of a given lightcurve belonging to a QSO (`i_metrics.csv`).

This paper has been typeset from a  $\text{\LaTeX}$  file prepared by the author.