MNRAS **000**, 1–5 (2015) Preprint 5 April 2016 Compiled using MNRAS LATEX style file v3.0

SDSS Stripe 82 : quasar variability from forced photometry

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

1 INTRODUCTION

2 DATA ANALYSIS

We use data from all SDSS runs up to an including run 7202 (Data Release 7), including all 6 SDSS camera columns.

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

The data were coadded, and all objects detected in the i-band coadds were assigned a deepSourceId (== objectId). For star-galaxy separation, 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. This amounts to 40 milion i-band detections down to 3σ. 8 million of those are brighter than 23^{rd} mag. Thus the total number of photometric measurements is : (40 million i-band detections) x (80 epochs) x (5 filters) = 16 billion measurements, including (8 million i-band detections i < 23) x (80 epochs) x (5 filters) = 3 billion measurements brighter than 23^{rd} mag.

Forced photometry was performed in u,g,r,i,z on the individual epoch images (NOT difference imaging), in locations specified by i-band coadds. (DIFFERENCE imaging is when photometry is done on coadd - individual-epoch-image) . Therefore in some cases the flux reported for a given aperture is negative, because after background subtraction noise oscillates around 0, and when it is scaled up, it can have negative values. [stored in rawDataFPSplit] The background in the optical bands is bright, and if we assume that the measured number of bacground counts oscillates around the value B_0, then the measured background count B is distributed as a Gaussian of width σ_B : $B - B_0$ $\mathcal{N}(0, \sigma_B)$. The noise is Poissonian, i.e. depends on the number of counts, and since for optical mesaurements the number of counts is large, $\sigma_B = \sqrt{B}$. On a 4kx4k CCD, with 16 Mpix, 5σ (corresponding to 1 false detection in a million), we would expect about 16 false detections. Now considering the distribution of the probability (likelihood) of flux measurement $L(F|data)$, for bright sources it is a very narrow Gaussian centered on the measured F_S, width σ_F on the level of $1-2\% \approx 0.01-0.02$ mag. However, for faint sources the probability, centered around the F_S, is much wider, so that there is a nonzero probability of negative flux measurement. A Bayesian way to address this issue is to impose the prior $p(F)$, since we understand that physically flux cannot be negative, so that the posterior probability $p(F|data) \propto L(F|data)p(F)$. A simple flat prior, being 0 for $F < 0$ and 1 otherwise, would not affect the measured F_S for bright sources, but for faint sources it would move the distribution (posterior) to be above zero flux. This would be the upper limit on the flux of that source. Therefore we decided to apply the Bayesian prior in case where $\langle F_L \rangle < k\sigma$, with $k = 2$ (for a Gaussian likelihood this corresponds to 2% probability of $F_L < 0$).

2.1 Faint sources

To test our 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 5σ level as the robust 25-th percentile (or median) of the ensemble F_{true} distribution : $\sigma_F = (1/5)F_{25\%}$ (in reality, σ increases for fainter observations, but this is a good approximation). Thus 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 $F^i_{obs} < 2\sigma F$, we consider $p(F)$ - a Gaussian likelihood associated with i-th measurement: $p_i(F) = \mathcal{N}(\mu = F^i_{obs}, \sigma = \sigma_F)$. Each measurement F_{obs} is a mean of this likelihood: $F_{obs} = \langle p_i(F) \rangle$. We call it $p(F)$ for short :

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

After imposing the Bayesian uniform prior $p(F)$ becomes a truncated Gaussian (without the negative part), centered on F^i_{obs}, with a width σ_F. Thus for the truncated $p(F)$ the mean ceases to be F^i_{obs}, but it can be defined as

$$F_{mean} = \int_0^\infty F\,p(F)dF \quad (2)$$

We also define the median as

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

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 2σ level as two areas $B = 0.05A$, or :

$$\int_{F_{2\sigma}}^{\infty} p(F)dF = 0.05 * \int_{0}^{\infty} p(F)dF \quad (4)$$

Using our definition, the mean is :

$$F_{mean} = \int_0^{\infty} \frac{F}{\sqrt{2\pi\sigma_F^2}} \exp\left(-\frac{(F-F_{obs})^2}{2\sigma_F^2}\right) = \frac{\sigma_F}{\sqrt{2\pi}} \exp\left(-\frac{F_{obs}^2}{2\sigma_F^2}\right) + F_{obs}\,\mathrm{sf}\left(\frac{-F_{obs}}{\sigma_F}\right) \quad (5)$$

where we used $z = (F - F_{obs})/\sigma_F$, and noticed that

$$\int_{z_0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \mathrm{CCDF}(z_0) = 1 - \mathrm{CDF}(z_0) = \mathrm{sf}(z_0) \quad (6)$$

is a complementary cumulative distribution function (CCDF), also known as the survival function (sf).

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

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-x_{obs})^2}{2}\right) \quad (7)$$

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

$$p(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z)^2}{2}\right) \quad (8)$$

In z-space, the median from

$$\int_0^{x_{med}} p(x)dx = \int_{x_{med}}^{\infty} p(x)dx \quad (9)$$

becomes

$$\int_{x_0}^{z_{med}} p(z)dz = \int_{z_{med}}^{\infty} p(z)dz \quad (10)$$

with $x_0 = -x_{obs}$

This expression for z_{med} can be evaluated : rhs is a survival function (sf) = 1 - cumulative density function (cdf) :

$$\int_{z_{med}}^{\infty} p(z)dz = \mathrm{sf}(z_{med}) \quad (11)$$

and the lhs, assuming $z_{med} > x_0$, is :

$$\int_{x_0}^{z_{med}} p(z)dz = \int_{-\infty}^{z_{med}} p(z)dz - \int_{-\infty}^{x_0} p(z)dz = \mathrm{cdf}(z_{med}) - \mathrm{cdf}(x_0) \quad (12)$$

Rearranging, and using the percent point function (ppf) - the inverse of the cdf, we find:

$$z_{med} = \mathrm{ppf}\left(\frac{1+\mathrm{cdf}(x_0)}{2}\right) \quad (13)$$

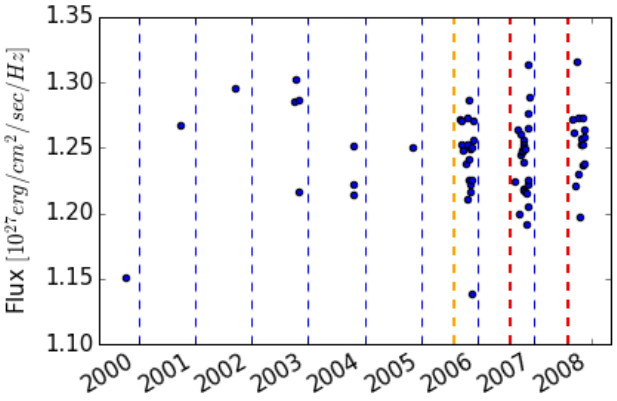

Figure 1. A plot showing an example lightcurve for an object id 217720894888346422. Jan 1st of each year (blue), August 1st of 2005 (orange) and August 1st of each subsequent year (red) is indicated by vertical dashed lines. Observations prior to August 1st of 2005 have sparser cadence, whereas those 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.

and transforming back to F space:

$$F_{med} = F_{obs} + \sigma_F\ \mathrm{ppf}\left(\frac{1+\mathrm{cdf}(x_0)}{2}\right) \quad (14)$$

with x_0 and x_{obs} as above.

In z space , the 2σ areas A and B are :
A = sf(x_0) and B = sf(z_B), so to find z_B we use the inverse survival function isf : z_B = isf(0.05A). Thus transforming back to F-space we have:

$$F_{2\sigma} = F_{obs} + \sigma_F\,(\mathrm{isf}(0.05\,\mathrm{sf}(x_0)) \quad (15)$$

2.2 Photometric colors

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},σ_F) for each faint observation by ($< F_{exp} >$,rms), we analyse updated lightcurves, addressing sparse sampling. Observations conducted prior to September 2005 were part of SDSS I-II, which had more sparse sampling than SDSS-III. September 1 - November 30 of 2005-2007 marks the SDSS Supernova Survey (see Fig. 2.2).

16

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

$$< F >= \frac{\sum w_i F_i}{\sum w_i} \qquad \sigma_{<F>} = \left(\sum w_i\right)^{-1/2} \quad (16)$$

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, σ_i) point prior to March 2006, and a single point per every subsequent year, where (F_i, σ_i) is (*mean*, *meanErr*) or (*median*, *medianErr*). We look for evidence of intrinsic variability by calculating χ^2 for mean,

$$\chi^2_{dof} = \frac{1}{N-1} \sum \left(\frac{F_i - < F >}{\sigma_i} \right)^2 \quad (17)$$

and the robust version based on the median: $\chi^2_R = 0.7414(Z_{75\%} - Z_{25\%})$ with $Z = (F_i - median)/\sigma_i$.

We further analyze the lightcurves by calculating σ_0 (the approximate value), and σ_{full}, following AstroML Fig.5.8. σ_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 only have 10 measurements, we have $x_{i,boot}$ which has 1000 random samples. Median is the 50-th percentile of any sample. Following [Ivezic+2014], chapter 5, we use a sample median to estimate $\mu_0 = median(x_{i,boot})$, 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 $e_{50} = median(e_{i,boot})$, we estimate σ_0 as :

$$\sigma_0 = (variance_{approx})^{1/2} = (\zeta^2 \cdot \sigma_G^2 - e_{50}^2)^{1/2} \quad (18)$$

where

$$\zeta = \frac{median(\widetilde{\sigma_i})}{mean(\widetilde{\sigma_i})} \quad (19)$$

and

$$\widetilde{\sigma_i} = (\widetilde{variance})^{1/2} = (\sigma_G^2 + e_i^2 - e_{50}^2)^{1/2} \quad (20)$$

For the marginalized σ_{full}, we calculate logarithm of the posterior probability distribution for the grid of μ and σ values as:

$$\log L = -0.5 \sum \left(\ln(\sigma^2 + e_i^2) + \frac{(x_i - \mu)^2}{(\sigma^2 + e_i^2)} \right) \quad (21)$$

We shift the maximum value of logL by subtracting the maximum value of logL, thus calculating the likelihood :

$$L = e^{\log L - max(\log L)} \quad (22)$$

We then marginalize over μ or σ :

$$p(\sigma) = \sum_{\mu} (L_{\sigma,\mu}) \qquad p(\mu) = \sum_{\sigma} (L_{\sigma,\mu}) \quad (23)$$

and normalize the probability :

$$p_{norm}(\sigma) = \frac{p(\sigma)}{\int p(\sigma) \mathrm{d}\sigma} \qquad p_{norm}(\mu) = \frac{p(\mu)}{\int p(\mu) \mathrm{d}\mu} \quad (24)$$

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

2.3 Extinction Correction

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, Av are for RV = 3.1, also suggested by Eddie Schlafly]

The SDSS Stripe 82 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 with $-40\,(+320) < RA < +10$ and IN2P3 with $+5 < RA < +55$ degrees. In the NCSA data there are 20978391 sources (`iCoaddAll.csv`). Removing those that overlap with IN2P3, we are left with 16520093 sources in the coadd detection data (`iCoaddPhotometryAll.csv`), and 16514187 sources (5906 less) in `DeepSourceNCSA_i_lt300.csv`. There are 12373162 sources with median photometry, matched with E(B-V) data (`medianPhotometry.csv`) , and of these, 5892054 brighter than 23 mag, with calculated median flux and colors (`ugrizMetrics.csv`). Before individual bands are aggregated into one, we have individual bands treated separately, with metrics calculated for each band, eg. `i_metrics.csv`. In this file we have both annual aggregate metrics and full lightcurve aggregate metrics, 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 τ, amplitude, expected for a DRW for a QSO, we calculate the likelihood of a given lightcurve belonging to a QSO.

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 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

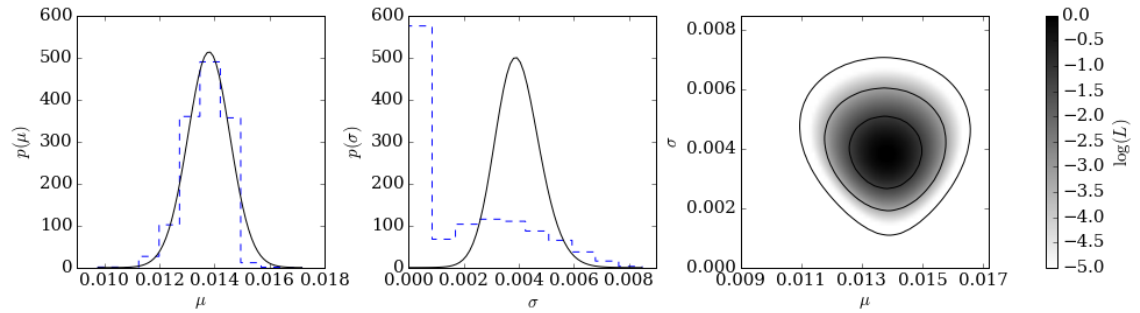

Figure 2. A plot equivalent to AstroML Fig 5.7 and 5.8 for lightcurve measurement points (x_i, e_i) for an object 217720894888346446 (without any averaging, raw forced photometry flux measurements). The solid lines show marginalized posterior pdfs for μ (left) and σ (middle). The dashed histograms show the distributions of approximate estimates for μ and σ, for 10,000 bootstrap resamples of the same data set. The right panel shows the logarithm of the posterior probability density function for μ and σ.

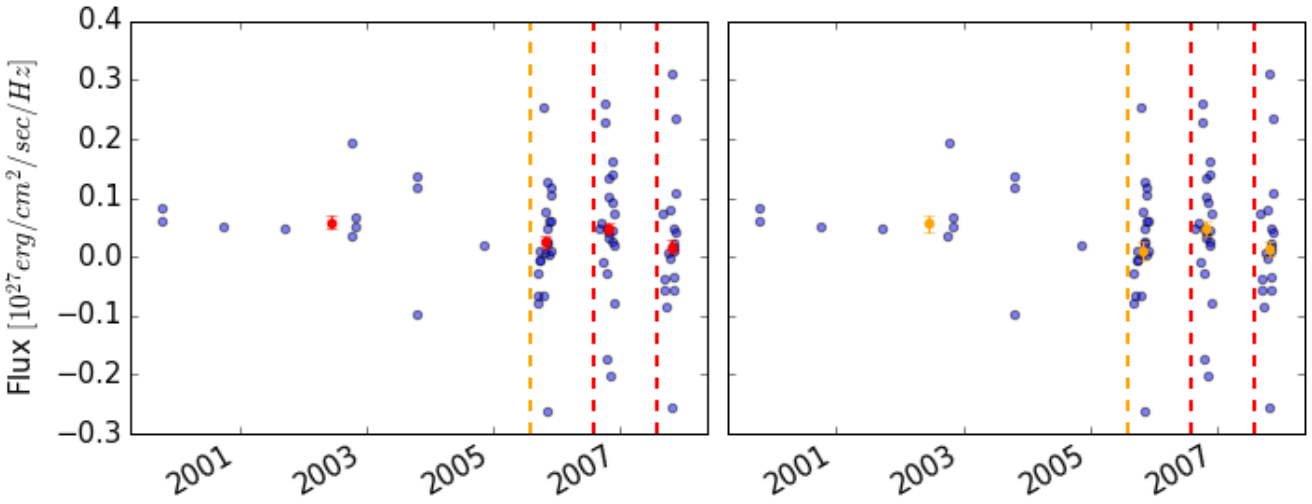

Figure 3. A plot showing an outcome of seasonal averaging for an object id 217720894888346425. The left panel (red dots) shows (mean, meanErr), and the right panel (orange) shows (median,medianErr), instead of seasonal points (blue). Vertical dashed lines as on Fig. 2.2

University, the United States Naval Observatory, and the University of Washington.

5 APPENDIX

5.1 How ugriz metrics were made

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).

5.2 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 $< 2\sigma$: $F_{2\sigma}$.

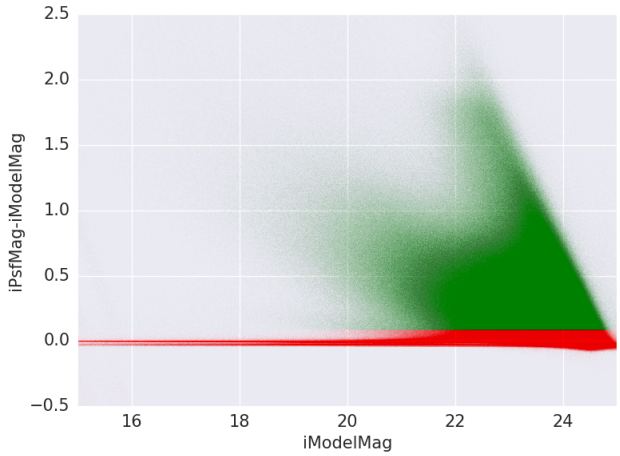

Figure 4. A plot showing NCSA sources detected in coadds, removing the outliers beyond the edges of the plot. The coloring corresponds to the `extendedness` parameter calculated in the pipeline based on the iPsfMag-iModelMag : red being 0 (compact), and green being 1 (extended). As iModelMag increases, the separation becomes less certain, as more distant galaxies are more compact.

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

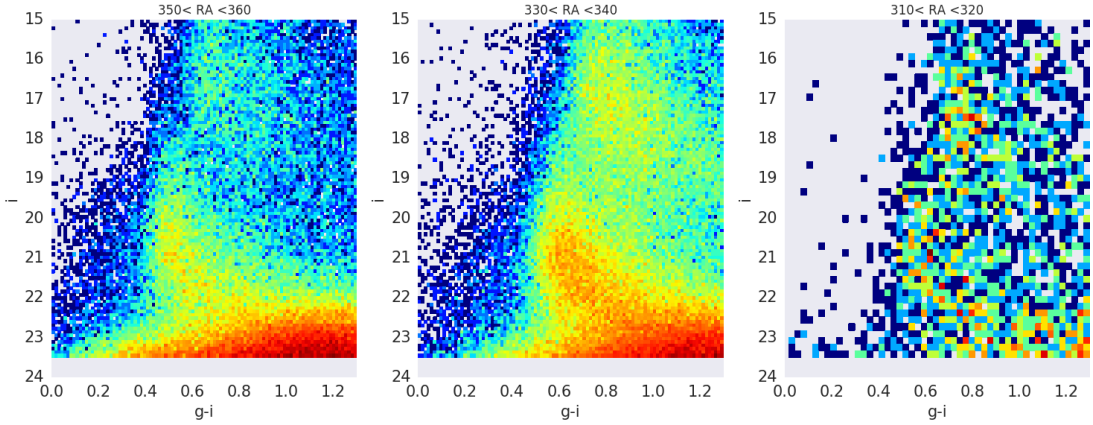

Figure 5. A color-magnitude plot , reproducing the results of Sesar+2010 , Fig.23 . We show here only NCSA-processed sources, which is why certain RA ranges are omitted or have less sources. We only select sources with `extendedness=0` parameter (stars). The scale is showing the $\log_{10}$ of count.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

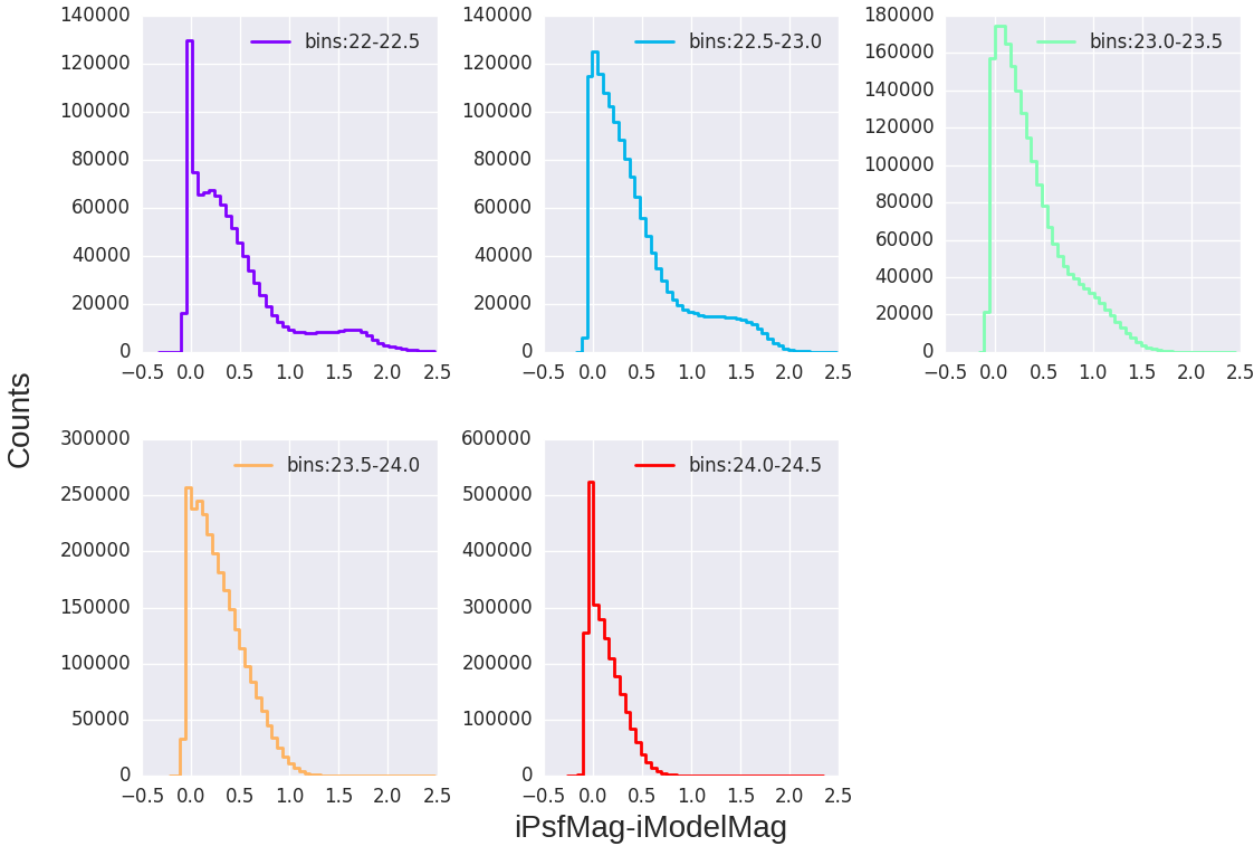

Figure 6. The histograms show the count of sources in 5 magnitude bins, corresponding to the vertical cut through Fig. 2.3. It helps to verify how well can the extended and compact sources be separated based solely on the iPsfMag-iModelMag