Varlability seLection of Astrophyslcal sources iN PTF (VILLAIN)
I. Structure function fits to 71 million objects
Bruun, S. H.; Agnello, A.; Hjorth, J.

Published in:
Astronomy and Astrophysics

DOI:
10.1051/0004-6361/202244945

Publication date:
2023

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
https://doi.org/10.1051/0004-6361/202244945
VarlabilLity seLectiOn of AstrophysIcal sources iN PTF (VILLAIN)

I. Structure function fits to 71 million objects

S. H. Bruun1, A. Agnello1,2, and J. Hjorth1

1 DARK, Niels Bohr Institute, University of Copenhagen, Jagtvej 128, 2200 Copenhagen N, Denmark
e-mail: sofie.bruun@nbi.ku.dk; adriano.agnello@nbi.ku.dk; jens@nbi.ku.dk
2 STFC Hartree Centre, Sci-Tech Daresbury, Keckwick Lane, Daresbury, Warrington WA4 4AD, UK

Received 9 September 2022 / Accepted 11 April 2023

ABSTRACT

Context. Light-curve variability is well-suited to characterising objects in surveys with high cadence and a long baseline. This is especially relevant in view of the large datasets to be produced by the Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST).

Aims. We aim to determine variability parameters for objects in the Palomar Transient Factory (PTF) and explore differences between quasars (QSOs), stars, and galaxies. We relate variability and colour information in preparation for future surveys.

Methods. We fit joint likelihoods to structure functions (SFs) of 71 million PTF light curves with a Markov chain Monte Carlo method. For each object, we assume a power-law SF and extract two parameters: the amplitude on timescales of one year, A, and a power-law index, γ. With these parameters and colours in the optical (Pan-STARRS1) and mid-infrared (WISE), we identify regions of parameter space dominated by different types of spectroscopically confirmed objects from SDSS: Candidate QSOs, stars, and galaxies are selected to show their parameter distributions.

Results. QSOs show high-amplitude variations in the R band, and the highest γ values. Galaxies have a broader range of amplitudes and their variability shows relatively little dependency on timescale. With variability and colours, we achieve a photometric selection purity of 99.3% for QSOs. Even though hard cuts in monochromatic variability alone are not as effective as seven-band magnitude cuts, variability is useful in characterising object subclasses. Through variability, we also find QSOs that were erroneously classified as stars in the SDSS. We discuss perspectives and computational solutions in view of the upcoming LSST.

Key words. techniques: photometric – quasars: general – methods: statistical – methods: data analysis – surveys

1. Introduction

Large, wide-field surveys allow us to identify rare objects, study parameter distributions of different object classes, and select sources for further examination. Spectroscopy can produce high-quality classifications, but often only photometry is available. With the prospect of deep, 10-yr light curves from the Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST; Ivezić et al. 2019), it is important to explore photometric selection based on light-curve variability as a function of timescale for a large dataset.

This method is particularly suitable for analysing certain types of objects with distinctive variability parameters; for example for distinguishing quasars (QSO) from stars (van den Bergh et al. 1973; Ulrich et al. 1997; Schmidt et al. 2010; Myers et al. 2015). Variability can also be used to advance our understanding of the physical nature of the objects, as the mechanisms behind variability operate on different timescales (Schmidt et al. 2012); such as QSO accretion-disk instabilities (Rees 1984) or changes in obscuration (Hopkins et al. 2012).

With identification of distant QSOs, we can study structure formation in the early Universe and the cosmological parameters determining its expansion, for example through measurements of baryon acoustic oscillations (BAO) in the Lyα forest (Turner 1991; Song et al. 2016; du Mas des Bourboux et al. 2020; Alam et al. 2021; Secrest et al. 2021). Myers et al. (2015) selected a fraction of the SDSS eBOSS QSO targets for BAO based on variability in light curves from the Palomar Transient Factory (PTF; Law et al. 2009; Rau et al. 2009). QSOs are also useful for redshift-drift tests of cosmic acceleration (Sandage 1962; Loeb 1998; Kim et al. 2015; Alves et al. 2019). Lensed QSOs are particularly interesting because they can be used for time-delay cosmography, which works better with highly variable sources (Refsdal 1964; Treu & Marshall 2016). Lensed QSOs can be discovered as extended objects with high variability (Kochanek et al. 2006).

Faint photometric standards are also of special interest. Large future observatories such as the Vera C. Rubin Observatory, the Extremely Large Telescope (Tamai et al. 2016), the Thirty Meter Telescope (Skidmore et al. 2015), and the Giant Magellan Telescope (Johns et al. 2012) can be used to observe objects of greater magnitude than those of typical standard stars. These latter have magnitudes of 11.5 < V < 16 in the Landolt (1992) sample, which has, along with the Stetson database of secondary standards (Stetson et al. 2019; Stetson 2000), since been expanded and curated as described in Pancino et al. (2022). The combined sample mainly includes sources with 13 < V < 21. However, the distribution on the sky is not uniform, and being able to distinguish variable from non-variable sources is useful for calibration even at lower magnitudes, such as the 20.6 R-band limit of the PTF.

There are several ways of characterising variability, and these can be even more useful than colours for selection of AGN (De Cicco et al. 2021). Baldassare et al. (2020) demonstrated the
potential of PTF variability for selection of AGN in low-mass galaxies by fitting 50,000 light curves with a damped random walk (DRW) model. Ward et al. (2022) found variable AGN in the Zwicky Transient Facility (ZTF; Bellm et al. 2019) using a combination of variability measures.

In this paper, we apply structure functions (SFs) to explore the variability of objects. This is a well-known technique first applied in astronomy by Simonetti et al. (1985), which is computationally efficient for large sets of data with gaps (Moreno et al. 2019). SF models describe the difference in magnitude $\Delta m$ across a time interval $\Delta t$. As in Hook et al. (1994) for example, we specifically consider a power-law model, which is often applied for AGN and QSOs. Both single power law models and DRW models pose challenges; even 20 yr baselines cannot constrain DRW models completely (Suberlak et al. 2021; Stone et al. 2022).

Sánchez et al. (2017) found the Bayesian SF defined by Schmidt et al. (2010) to be the best and most stable SF for noisy light curves with irregular sampling. Schmidt et al. (2010) applied power-law SFs to light curves from The Sloan Digital Sky Survey (SDSS) and defined a variability parameter region for selection of QSO candidates with criteria chosen by eye. This gave relatively complete and pure candidate sets for SDSS S82 light curves, but with poorer performance for light curves from Pan-STARRS1 (PS1; Chambers et al. 2016). Schmidt et al. (2010) used sets of known QSOs, F/G stars, and RR Lyrae, with stars outnumbering QSOs by a factor of 30.

In this paper, we aim to analyse the results and performance of a similar method but applied to the entire PTF survey. We show where variability is most effective in breaking degeneracies and which forms of variability are common among different types of objects. Specifically, matching with spectroscopic SDSS classifications of QSOs, stars, and galaxies, which we assume as ground truth, we can evaluate new variability selection criteria and compare with those of Schmidt et al. (2010). By including as many objects from the PTF survey as possible, the properties of the total light-curve sample are more representative of PTF sources than if we only included specific stellar subtypes, for example.

Another approach to photometric selections is based on colours. We examine colour–colour and colour–magnitude diagrams of a spectroscopically confirmed sample by matching magnitudes in mid-infrared (IR) from the Wide-field Infrared Survey Explorer (WISE; Wright et al. 2010) and in the optical from PS1. This will allow us to examine simple selection criteria based on these.

We describe the datasets (and the data-cleaning process) used in this work in Sect. 2. We introduce the methodology in Sect. 3. In Sect. 3.1, we define models of the PTF light-curve SFs, and we fit them in Sect. 3.2. The metrics used for evaluation of the selection of objects are defined in Sect. 3.3. The results of the SF fitting are plotted and described in Sect. 4. Based on the variability parameters and matched colour information, we choose photometric selection criteria and examine the properties of selected objects in Sect. 5. We discuss these properties in Sect. 6 and summarise our findings in Sect. 7.

2. Data

2.1. PTF

The Palomar Transient Factory was a project with the Palomar 48 Schmidt telescope at Palomar Observatory in California. This includes imaging in DR1, DR2, and DR3 from 1 March 2009 to 28 January 2015 covering 20,000 deg$^2$ (Law et al. 2009; Rau et al. 2009). The PTF light-curve database is based on a subset of these images with data from 598,975,024 objects in $R$ and $g$.

As most data points are in $R$, these are chosen for the analysis of this paper (as mag$_{\text{autocorr}}$ and magerr$_{\text{auto}}$). The $R$ band is adapted from the 658 nm Mould-$R$ filter described in Cuillandre et al. (2000), has a limiting magnitude of 20.6, and photometry is in the AB magnitude system (Law et al. 2009). From this database, we also consider the timestamps, PTF object ID, RA, Dec and median R, all queried in 648 patches of $10\times10$ via the IRSA Catalog Search Tool. We split 310 large patch files into 6340 files, keeping data points with the same object ID together. File by file, we group data points by ID to assemble light curves. These are cleaned and fitted in Sects. 2.5 and 3.1 in parallel using 14 computing nodes for six months.

2.2. SDSS

The Sloan Digital Sky Survey DR17 (Blanton et al. 2017; Abdurro’uf et al. 2022) includes 5789 200 spectra (Smee et al. 2013; Drory et al. 2015; Wilson et al. 2019) from the Apache Point Observatory in New Mexico, USA (Gunn et al. 2006). Based on DR17 from February 2021 (with data both in PhotoObj and SpecObj) 866,338 QSOs, 962,162 stars, and 2,790,253 galaxies are spectroscopically confirmed. Their coordinates, redshifts, and spectroscopic classes are queried via CasJobs on SciServer (Taghizadeh-Popp et al. 2020). We cross-match by creating a k-d tree (Manewongvatana & Mount 1999) of the PTF coordinates. We then query the tree to find the nearest SDSS object within 2 arcsec, if it exists. By inspecting SDSS images, we find that this radius best avoids spurious matches.

2.3. WISE

The Wide-field Infrared Survey Explorer is a space telescope observing in the mid-IR (Wright et al. 2010). It had a cryogenic phase from December 14, 2009 to February 2011 and a post-cryogenic phase has been ongoing since 2013 (NEOWISE). With the latest update from February 2021, the combined AllWISE program (Cutri et al. 2021) data release II/328 contains 748 million objects.

We query WISE objects with the CDS X-Match Service (Boch et al. 2012) within five arcseconds of the mean PTF object ID together. File by file, we group data points by ID to assemble light curves. These are cleaned and fitted in Sects. 2.5 and 3.1 in parallel using 14 computing nodes for six months.

1 This analysis expands the first queries and parameter exploration presented by Bruun (2020), MSc thesis (unpublished).

A111, page 2 of 16

---

2 As the Intermediate Palomar Transient Factory for DR3.

3 www.ptf.caltech.edu/iptf

4 PTF team, DOI: 10.26131/IRSA156, url: https://irsa.ipac.caltech.edu/Missions/ptf.html, IPAC.

5 Using the Image List tool at skyserver.sdss.org/dl17/VisualTools/list
2.4. PS1

The Panoramic Survey Telescope and Rapid Response System includes two telescopes in Hawaii: Pan-STARRS1 and Pan-STARRS2. We use the $g$ (481 nm), $r$ (617 nm), and $z$ (866 nm) bands (Tonry et al. 2012) from the Pan-STARRS1 survey DR1 with 1.92 billion objects (‘vizier:II/349/ps1’; Ochsenbein et al. 2000; Chambers et al. 2016). This covers the sky at J2000.0 declination $>30^\circ$, including the PTF footprint. The PS1 magnitudes are in the AB system. As in WISE, in PS1 we query within five arcseconds of PTF objects and select one set of magnitudes, magnitude errors, and coordinates per source.

2.5. Data cleaning

Data cleaning is needed for reliable fitting results. For PTF, extreme magnitudes outside the range of $12 < R < 22$ are removed. If a light curve contains too few data points, fit parameters are difficult to constrain, and if the time span is too short, we cannot detect variability over long timescales. Therefore, we require that each light curve must contain at least 20 data points and span at least 1 yr. At $80^\circ < RA < 90^\circ$ and $0^\circ < Dec < +10^\circ$, light curves are so well sampled that in order to speed up the analysis in Sect. 3.2 we select a maximum of 1500 random data points for some sources.

Magnitude outliers affect variability fits, and so we want to remove them while preserving the real variability. A moving weighted median (MWM) is computed for each data point $i$ in each light curve. This is based on the magnitudes $R_{\text{close,i}}$ of the seven closest neighbouring data points in time within a window of five days (2.5 days to each side). Data points more than three standard deviations away from the MWM are removed, taking into account both the error of the data point and of the weighted median. A data point with magnitude $R$ and error $\sigma_R$ is removed if

$$\frac{|R - \text{MWM}|}{\sqrt{\sigma_R^2 + \text{MAD}^2}} > 3,$$

where MAD is the median absolute deviation. We compute the MAD of every part of the MWM as the weighted median of the absolute distances of the data points used in the computation of the MWM. That is, the MAD of a specific value of MWM is based on the up to seven data points within the time window of data point $i$:

$$\text{MAD} = \text{weighted median} (|R_{\text{close,i}} - \text{MWM}|).$$

If fewer than seven data points are close enough to $i$ in time, they are still used, and the data point $i$ itself has a relatively higher weight compared to those few data points, leading to a higher probability of acceptance. This is desirable, as we have less information to base rejection on.

After data cleaning, the remaining light curves are distributed on the sky according to Fig. A.1, mostly covering the northern hemisphere and avoiding low Galactic latitudes. Table 1 contains sample statistics including objects matched in SDSS, WISE, and PS1. 70920904 objects are analysed, spanning 365–2147 days and containing 20–5579 data points.

Table 1. Sample sizes.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full PTF light curve database</td>
<td>598 975 024</td>
</tr>
<tr>
<td>Cleaned PTF R-band sample</td>
<td>70 920 904</td>
</tr>
<tr>
<td>Matched</td>
<td></td>
</tr>
<tr>
<td>SDSS spectra</td>
<td>1 748 047</td>
</tr>
<tr>
<td>WISE</td>
<td>57 007 069</td>
</tr>
<tr>
<td>PS1</td>
<td>70 891 378</td>
</tr>
<tr>
<td>WISE and PS1</td>
<td>56 991 591</td>
</tr>
<tr>
<td>All surveys</td>
<td>1 613 916</td>
</tr>
</tbody>
</table>

Notes. Sample sizes before and after selecting and cleaning PTF light curves and matching with sources in SDSS, WISE, and PS1.

3. Methodology

We describe the variability of each object by defining power-law models of their structure functions. The models are fit to structure functions of individual objects to extract variability descriptors as fit parameters. These are used for exploring relations in the data and photometric selection of object classes in Sects. 4–5; but first, we define metrics for evaluation of selection quality in Sect. 3.3.

3.1. Structure functions

For each object, we analyse variability by comparing every pair of data points in its light curve. For each pair of data points $i$ and $j$, we have a difference in magnitude $\Delta m_{ij}$ and in time $\Delta t_{ij}$. By comparing all pairs, we find the timescale dependence of differences in magnitude with a SF (Simonetti et al. 1985), and model this variability.

The total effective variability $V_{\text{eff}}$ of each object consists of intrinsic variability, which we describe with a structure function SF, and noise, $\sigma_m$. Assuming the model describes the data well, $V_{\text{eff,ij}}$ is close to the observed $\Delta m_{ij}$. For SF, we choose a power law following the notation of Schmidt et al. (2010):

$$V_{\text{eff,ij}}^2 = SF_{ij}^2 + \sigma_{m,ij}^2 + \sigma_{m,ij}^2 = (\Delta m_{ij})^2,$$

$$SF_{ij} = A \left(\frac{|\Delta t_{ij}|}{t_0}\right)^\gamma, \quad A \geq 0.$$

Equation (4) has two free parameters, $A$ and $\gamma$; the former quantifies the amplitude of the variations, in units of magnitude, on the timescale $t_0$, and the power-law index $\gamma$ describes how the amplitudes depend on timescale. We choose $t_0 = 1$ yr (observed frame). One could correct for the factor of $1 + z$ to find rest frame $\Delta t_{ij}$, but redshift information is limited, and SDSS redshifts are not independent of the spectroscopic labels we use for comparison.

We expect most objects to have $0 < \gamma < 1$. Here, $\gamma > 0$ shows that variability increases with timescale, and for $\gamma > 1$ this is accelerating, which we would mostly expect to see for short light curves with highly uncertain $\gamma$. This could be objects with an overall positive or negative trend. $\gamma < 0$ means most variability is found to be on short timescales, for example in transient tails, but we do not expect to include these objects due to the minimum observed time span of 1 yr. Only $\gamma$ consistently shows correlation with physical black-hole parameters in the literature, namely an anti-correlation with the bolometric luminosity, $L_{\text{bol}}$, and the Eddington ratio, $\lambda_E$ (De Cicco et al. 2022).
3.2. Fitting

Structure function models are fitted to data using the emcee Python package (Foreman-Mackey et al. 2013) for affine-invariant sampling with Markov chain Monte Carlo (MCMC; Goodman & Weare 2010). As the walkers of the MCMC jump to different $A$ and $\gamma$ values, they evaluate the likelihood of observing the light curve given the variability parameters. This likelihood is

$$L_{ij} = \frac{1}{\sqrt{2\pi \sigma_{\Delta m_{ij}}^2}} \exp \left( -\frac{(\Delta m_{ij})^2}{2\sigma_{\Delta m_{ij}}^2} \right)$$

(5)

for each pair of data points, assuming a Gaussian distribution of $\Delta m_{ij}$. For the total log posterior (LP) of the light curve, we use the same prior $\ln(p)$ as in Schmidt et al. (2010):

$$\ln(p) = \ln \left( \frac{1}{1 + \gamma^2} \right) + \ln \left( \frac{1}{A} \right),$$

(6)

$$LP = \sum_{i,j} \ln(L_{ij}) + \ln(p).$$

(7)

The MCMC runs with eight walkers for 500 steps where the first 200 are discarded as burn-in. We find these values to balance accuracy and speed. From each fit, we retain the median values of $A$ and $\gamma$ and their 16th and 84th percentiles as 1σ uncertainties.

3.3. Evaluation metrics

To study parameter distributions of different object types, we photometrically select QSOs, stars, and galaxies. The quality of selection criteria is evaluated on purity and completeness, which are estimated using the spectroscopically confirmed subset. We compute these with the number of true positives (TPs), false negatives (FNs), and false positives (FPs):

$$\text{Completeness} = \frac{TP}{TP + FN}, \quad \text{Purity} = \frac{TP}{TP + FP}.$$  

(8)

4. Variability and colour distributions

We present plots and statistics for two datasets: all fitted objects and those with spectroscopic classifications. In Sect. 5, we also assemble photometric selections. The datasets allow us to compare distributions of three classes: QSOs, stars, and galaxies. The plots include variability parameters ($A$ and $\gamma$) and colour diagrams of $W2$ versus $W1 - W2$ and $g - r$ versus $z - W1$. We know stars, galaxies, and QSOs to have good separation in optical versus IR colours (Maddox & Hewett 2006; Lang et al. 2016). $W1 - W2$ was also used by Wright et al. (2010), and Assef et al. (2013) combined $W2$ and $W1 - W2$ for selecting AGN.

We query and fit the data in batches as described in Sect. 2.1. For 6340 files with approximately 1 million light-curve data points per file, the batch processing time is about 1 h on a computing node with 32 cores. The maximum time is 11 days for one file.

4.1. Full Palomar Transient Factory sample

In Fig. 1, we show parameter distributions for the full PTF sample. The top panel is a 2D histogram of $A$ and $\gamma$ values, illustrating the variability of the 70920904 fitted light curves. We see two large clusters, at $\log A \sim -0.8$ and $\log A \sim -5$ (base 10).
Variations of just $10^{-5}$ magnitudes over timescales of 1 yr indicate that the values of the latter cluster are spurious.

The middle and bottom panels of Fig. 1 display a colour–colour and a colour–magnitude diagram. As these are based on parameters in WISE and PS1, they show distributions for the 80% (56 991 591) of the PTF objects with matches in both of these surveys (see Table 1). Matches in the two surveys are found for 93% of QSOs, 75% of stars, and 98% of galaxies.

4.2. Spectroscopically confirmed sample

Figure 2 contains similar plots to Fig. 1, but only for data with spectroscopic SDSS classes. In the variability parameter plot, this is 2.5% of the fitted sources (1 748 047), and the colour diagrams show the 2.276% (1 613 916) with matches in SDSS, WISE, and PS1. The colours denote different classes: red for QSOs, green for stars, and blue for galaxies. These are scaled based on relative density, with full saturation for areas of parameter space with one class and maximum relative density, white for areas without data and black or grey for areas with high relative densities of multiple classes. We note that this means a space that is more green than blue can still have more galaxies than stars. The full dataset includes more galaxies than stars and QSOs combined (see Sect. 2.2).

In the top panel of $A$–$\gamma$ space, we notice the two clusters of Fig. 1. The one at $\log(A) \sim -0.8$ has high relative densities of all classes, but they are spread out along different axes in the $A$–$\gamma$ plane; more galaxies have $\log(A) < -1$ and more QSOs have $\gamma > 0.1$. Stars are mostly in the same areas as galaxies, but with more spread in $\gamma$ and higher relative density at $\log(A) < -3$. In the middle panel, the stellar locus is immediately recognisable, as are the main distributions of QSOs and galaxies (Lang et al. 2016; Ansari et al. 2021). The latter two include an overlap most apparent at $g - r \sim 0.4$ and $z - W1 > 0.3$. The bottom panel shows that the three classes have different distributions in $W1 - W2$ versus $W2$ as well, but with a large overlap between stars and galaxies at $W2 > 15$.

5. Photometric selection

We next explore how variability relates to broad-band colours through photometric selections using colours, variability, or both. Simple photometric selections of QSOs, stars, and galaxies allow us to inspect differences between parameter distributions of photometrically and spectroscopically selected objects. Based on the distributions of Fig. 2, we define regions of parameter space dominated by QSOs, stars, or galaxies. Schmidt et al. (2010) used a similar approach in $A$ and $\gamma$. For reliable selections, we focus on purity and accept a low completeness. We avoid areas of high degeneracy between classes and achieve separate selection criteria for variability parameters and for colours. All criteria are listed in Table 2. While these are simple, linear criteria in parameter space that can be used to study relations between colour and variability, more advanced techniques exist and have been deployed for example by Ansari et al. (2021) in colour space. A study of this kind is presented in a companion paper (Bruun et al. 2023).

We also examine the properties of the QSO selection criteria in Schmidt et al. (2010) for sources with large fit parameters compared to their uncertainties:

$$\frac{A}{\sigma_A} > 2, \quad \frac{\gamma}{\sigma_\gamma} > 2.$$  \hspace{1cm} (9)

Fig. 2. All spectroscopically confirmed stars (green), QSOs (red) and galaxies (blue) from SDSS plotted in variability fit parameters (top) and colours: $g - r$ vs. $z - W1$ (middle) and $W2$ vs. $W1 - W2$ (bottom). Marker colours show the object class and blend to grey or black when multiple classes occupy the same parameter region. A heat map of all spectroscopically confirmed objects is found in Fig. B.2 for comparison with Fig. 1.
Table 2. Selection criteria.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>QSO candidates</th>
<th>Star candidates</th>
<th>Galaxy candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma$</td>
<td>$&gt; 0.13$</td>
<td>...</td>
<td>$&lt; 0.1$</td>
</tr>
<tr>
<td>$\log_{10}(A)$</td>
<td>$&gt; 0.11$</td>
<td>$&gt; 0.01$</td>
<td>$&lt; 0.1 \land &gt; 0.01$</td>
</tr>
<tr>
<td>$g - r$</td>
<td>$&lt; 0.2$</td>
<td>...</td>
<td>$&gt; 1$</td>
</tr>
<tr>
<td>$z - W1$</td>
<td>$&gt; 2.8$</td>
<td>$&lt; 2 \lor 1 + 1.25(g - r)$</td>
<td>$&gt; 2.3 + 1.25(g - r)$</td>
</tr>
<tr>
<td>$W1 - W2$</td>
<td>$&gt; 0.75$</td>
<td>$&lt; -0.05$</td>
<td>$&gt; 0.25 \land &lt; 0.35$</td>
</tr>
<tr>
<td>$W2$</td>
<td>...</td>
<td>$&lt; 15$</td>
<td>$&lt; 15$</td>
</tr>
</tbody>
</table>

Notes. Criteria for selection of candidate QSOs, stars, and galaxies. These are based on the SDSS distributions of Fig. 2 and the parameter regions are plotted in Fig. B.2.

5.1. Selection properties

The selection criteria of Table 2 are based on purity and completeness of the labelled objects in Fig. 2 and are then applied to the larger, unlabelled datasets of Fig. 1. The criteria are illustrated in Fig. B.2 with those of Schmidt et al. (2010) for comparison.

Table 3 lists sample statistics of candidates selected within these parameter space regions of variability, colour, or both. The Schmidt criteria are included for comparison. To estimate completeness, colour-based selections are given as percentages of the total set of objects with colour matches in WISE and PS1. Using both colour and variability criteria, the selections have low completeness and are very pure. However, selection bias in the spectroscopic sample affects purity estimates (see Sect. 6.3). If we disregard the bias, a naïve purity estimate for colour-selected QSO candidates is 98.7%, assuming 15.5% of all PTF light curves with colour matches are QSOs. We obtain a purity of 99.3% for colour- and variability-selected QSOs, and of 59.3% with just variability criteria (assuming 15.3% QSOs in the latter case because WISE and PS1 matches are not required). Most candidates selected with colour and variability are registered with the same label as main type in SIMBAD; Fig. D.1 shows statistics for each class for comparison.

Gaia DR3 (Gaia Collaboration 2016, 2023) sources within one arcsecond of PTF sources have lower proper motions for QSO candidates. This is especially clear when colour information is included in QSO selection, as shown in Fig. E.1. Objects with stellar colours and variability have the highest proper motions, as expected. Assuming a maximum of one Gaia match per source, only 3.7% of SDSS galaxies have a match in Gaia with a measured proper motion, and the fraction decreases to 0.5% for galaxy candidates selected by colour and variability. Further details are given in Appendix E.

If instead we use the Schmidt criteria and require the sources to have $A$ and $\gamma$ significantly different from zero (Eq. (9)), we find the most common SDSS class to be QSOs. However, these criteria lead to more contamination from stars and galaxies than the QSO criteria of Table 2.

5.1.1. Colour selection

Figure 3 shows distributions of colour-based selections from the large dataset of fitted sources in Fig. 1. The overall pattern is the same as in Fig. 2, but the stars are more spread out. More of them have very low $A$ or high $\gamma$, and relatively few are found at $\log(A) \sim -0.8$ and $\gamma \sim 0$. According to Table 3, the colour-selected candidates have similarly high purities of 98.6–99.7% for all classes, but with varying completeness, of namely 6.82% for galaxies, 7.92% for stars, and 38.3% for QSOs.

5.1.2. Variability selection

We plot the variability-selected sources in Fig. 4. Relying solely on $A$ and $\gamma$ is challenging according to the SDSS labels. For example, while the variability criteria for stars catch 53.4% of stars and 19.8% of galaxies, they select more galaxies due to different population sizes. Naturally, criteria based on overlapping classes in $A-\gamma$ space lead to overlapping selections in colour space.

5.2. Ambiguous sources

Given the purity of the photometric selections in Table 3, the contaminating sources are expected to have a high rate of misclassification by SDSS. Checking the most confident photometric predictions with differing spectroscopic classifications, we find examples of spectra that appear to be more typical of the photometric candidate class. We inspect the spectra of the 32 photometric QSO candidates with stellar spectroscopic classifications, and judge $\sim 20\%$ to be QSOs and $\sim 50\%$ to be possible QSOs. Of these objects, 16 are in SIMBAD, and are registered as 8 QSOs, 4 BL Lacertae objects, 1 blazar, and just 3 stars. For example, SDSS J120429.34+495814.4, with the spectrum of Fig. B.3, has the characteristic broad lines of a QSO. This object is part of the Data Release 12 Quasar Catalog from SDSS (Paris et al. 2017), but is spectroscopically confirmed as a star in SDSS DR17. This misclassification analysis shows the power of the photometric selection criteria of this work, although we do not expect high rates of misclassification in the full SDSS spectroscopically confirmed sample.

Variable galaxies in the Table 2 QSO region of $A$ and $\gamma$ have higher redshifts, $W2$, and $z - W1$ than other galaxies (typically redshift $0.5 \text{ vs. } 0.1$; see Figs. C.1 and C.2). In SIMBAD, variable galaxies are more often registered as the brightest galaxy in a cluster than other galaxies are. The fraction increases from 9% to 15%. There is also a low but relatively much higher fraction of radio sources, which goes from 1.6% to 3.7%. SDSS QSOs have similar redshifts independently of variability, except more non-QSO-like variability at $z < 0.3$ (see Fig. C.1).

Spectroscopic QSOs variability-selected as stars or galaxies are more often found at $W1 - W2 < 0.75$ and $W2 < 14$ than other QSOs. However, this difference in WISE colours is smaller than for spectroscopic galaxies and stars in different photometric classes, as illustrated in Figs. C.2–C.4.

6. Discussion

Quasars, stars, and galaxies are distributed differently in $A$ and $\gamma$ and in colours, but with overlaps limiting the quality of simple selections. We see this in differences between the SDSS
Table 3. Selection statistics.

<table>
<thead>
<tr>
<th>Selection</th>
<th>QSOs</th>
<th>Stars</th>
<th>Galaxies</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>268 230</td>
<td>389 317</td>
<td>1 090 500</td>
<td>70 920 904</td>
</tr>
<tr>
<td>Relative frequency in full sample</td>
<td>0.3782%</td>
<td>0.5489%</td>
<td>1.538%</td>
<td>...</td>
</tr>
<tr>
<td>Relative frequency in spec sample</td>
<td>15.34%</td>
<td>22.27%</td>
<td>62.38%</td>
<td>...</td>
</tr>
<tr>
<td>Colour and variability QSO candidate</td>
<td>31 404</td>
<td>32</td>
<td>191</td>
<td>70 503</td>
</tr>
<tr>
<td>Completeness</td>
<td>12.53%</td>
<td>0.011%</td>
<td>0.018%</td>
<td>0.1237%</td>
</tr>
<tr>
<td>Purity</td>
<td>99.3%</td>
<td>0.10%</td>
<td>0.60%</td>
<td>...</td>
</tr>
<tr>
<td>Colour and variability star candidate</td>
<td>15</td>
<td>21 129</td>
<td>194</td>
<td>2 995 546</td>
</tr>
<tr>
<td>Completeness</td>
<td>0.006%</td>
<td>7.24%</td>
<td>0.018%</td>
<td>5.256%</td>
</tr>
<tr>
<td>Purity</td>
<td>0.07%</td>
<td>99%</td>
<td>0.91%</td>
<td>...</td>
</tr>
<tr>
<td>Colour and variability galaxy candidate</td>
<td>28</td>
<td>25</td>
<td>28 076</td>
<td>107 227</td>
</tr>
<tr>
<td>Completeness</td>
<td>0.011%</td>
<td>0.009%</td>
<td>2.62%</td>
<td>0.1881%</td>
</tr>
<tr>
<td>Purity</td>
<td>0.10%</td>
<td>0.09%</td>
<td>99.8%</td>
<td>...</td>
</tr>
<tr>
<td>Variability QSO candidate</td>
<td>75 690</td>
<td>12 704</td>
<td>39 284</td>
<td>3 390 617</td>
</tr>
<tr>
<td>Completeness</td>
<td>28.2%</td>
<td>3.26%</td>
<td>3.60%</td>
<td>4.781%</td>
</tr>
<tr>
<td>Purity</td>
<td>59.3%</td>
<td>9.95%</td>
<td>30.8%</td>
<td>...</td>
</tr>
<tr>
<td>Variability star candidate</td>
<td>14 005</td>
<td>207 701</td>
<td>216 312</td>
<td>22 854 151</td>
</tr>
<tr>
<td>Completeness</td>
<td>5.22%</td>
<td>53.4%</td>
<td>19.84%</td>
<td>32.225%</td>
</tr>
<tr>
<td>Purity</td>
<td>3.20%</td>
<td>47.4%</td>
<td>49.4%</td>
<td>...</td>
</tr>
<tr>
<td>Variability galaxy candidate</td>
<td>12 470</td>
<td>44 170</td>
<td>228 634</td>
<td>9 140 797</td>
</tr>
<tr>
<td>Completeness</td>
<td>4.65%</td>
<td>11.35%</td>
<td>20.97%</td>
<td>12.889%</td>
</tr>
<tr>
<td>Purity</td>
<td>4.37%</td>
<td>15.48%</td>
<td>80.1%</td>
<td>...</td>
</tr>
<tr>
<td>Colour QSO candidate</td>
<td>96 050</td>
<td>237</td>
<td>1 045</td>
<td>277 394</td>
</tr>
<tr>
<td>Completeness</td>
<td>38.3%</td>
<td>0.081%</td>
<td>0.098%</td>
<td>0.4867%</td>
</tr>
<tr>
<td>Purity</td>
<td>98.7%</td>
<td>0.24%</td>
<td>1.07%</td>
<td>...</td>
</tr>
<tr>
<td>Colour star candidate</td>
<td>17</td>
<td>23 111</td>
<td>319</td>
<td>3 495 123</td>
</tr>
<tr>
<td>Completeness</td>
<td>0.007%</td>
<td>7.92%</td>
<td>0.030%</td>
<td>6.133%</td>
</tr>
<tr>
<td>Purity</td>
<td>0.07%</td>
<td>98.6%</td>
<td>1.36%</td>
<td>...</td>
</tr>
<tr>
<td>Colour galaxy candidate</td>
<td>152</td>
<td>73</td>
<td>73 112</td>
<td>322 126</td>
</tr>
<tr>
<td>Completeness</td>
<td>0.061%</td>
<td>0.025%</td>
<td>6.82%</td>
<td>5.65%</td>
</tr>
<tr>
<td>Purity</td>
<td>0.21%</td>
<td>0.10%</td>
<td>99.7%</td>
<td>...</td>
</tr>
<tr>
<td>Schmidt region</td>
<td>85 012</td>
<td>24 576</td>
<td>79 279</td>
<td>5 484 750</td>
</tr>
<tr>
<td>Completeness</td>
<td>31.7%</td>
<td>6.31%</td>
<td>7.27%</td>
<td>7.734%</td>
</tr>
<tr>
<td>Purity</td>
<td>45.0%</td>
<td>13.01%</td>
<td>42.0%</td>
<td>...</td>
</tr>
</tbody>
</table>

Notes. Sample statistics of selections in variability and colour, including counts, completeness, and purity (see Eq. (8)). In row 1 (All), we compare spectroscopically classified objects to the full fitted source count and to the full set with spectroscopic classes to get relative frequencies in each set. In rows 5–7 and 11, for computing purity and completeness, selection counts are compared to the total spectroscopic counts. In rows 2–4 and 8–10, comparisons also require matches in WISE and PS1, because colours are used.

class distributions of Fig. 2 and the corresponding candidate class plots in Figs. 3 and 4 based on colour and variability, respectively. The overall patterns can still be recognised. Class information in the SF variability parameters is especially interesting for objects without spectroscopy and limited colour measurements. Variability alone does not give candidates that are as reliable as those from colour selection, but colour and variability combined can break degeneracies and select classes better than each of them separately.

6.1. $\alpha$-$\gamma$ clusters

In $\alpha$-$\gamma$ space, we see two large clusters. One of them is likely an artefact from objects with undetectable variability, leading the MCMC to suggest arbitrarily small $\alpha$ and a large range of $\gamma$ values. There is only one clear cluster at higher $\alpha$ ($\log(\alpha) \sim -0.8$), but the QSOs and galaxies are spread along different axes. This is most apparent in Fig. 2. The galaxies cover a broad range of $\alpha$, but their variability is rarely timescale dependent, at least not on most relevant scales. This is shown by the low $\gamma$ values. In contrast, QSOs are even more variable (high $\alpha$) with a clearer timescale dependence. Stars are spread out more evenly in $\alpha$ and $\gamma$, limiting selection purity and reflecting the diverse nature of stellar variability.

6.2. Comparison with SDSS light-curve analysis

The selection criteria of Table 2 are simple and focus on purity. For QSOs, they differ from those by Schmidt et al. (2010). Table 3 shows a purer QSO set with slightly lower completeness, than if we apply the Schmidt criteria. This may be due to slightly different fitting and outlier removal or differences in noise and
measurements between PTF and SDSS. The dataset presented in this paper is more representative of all observed object types, as it includes all sources from the PTF survey that pass through the data cleaning of Sect. 2.5, whereas Schmidt et al. (2010) introduced specific types of contaminants and in specific ratios.

6.3. Spectroscopic selection bias

With only 2.5% of sources having spectroscopic classifications, there is a need for other methods for identification. It also allows for significant selection bias in sources with SDSS spectra compared to sources with long PTF light curves. This affects purity estimates, and the choice and evaluation of selection regions.

There are differences in the distributions of sources with and without spectroscopic classes. This is especially clear if we compare Fig. 1 to Fig. B.2. For example, SDSS is missing spectra for objects at low W2, including a band of sources at low W1 – W2. At 10 < W2 < 12 and –0.5 < W2 – W1 < –0.25, we have a very mixed group, with sources typically marked as stars (21%) or binary star systems in SIMBAD.

SDSS also has a bias in favour of sources with z – W1 > 3. This is part of the reason why in Fig. 4, QSOs and galaxies are relatively infrequent at those values compared to Fig. 2. If more objects are included in star-dominated areas, they also include a larger fraction of the galaxies and QSOs. However, many of these are actually stars incorrectly selected according to variability. Variable stars late in the main sequence can be especially difficult to distinguish from QSOs and galaxies; the parameter region at 2.1 < z – W1 < 2.6 and 1.1 < g – r < 1.4 is dominated by stars for all three variability selections. Variability-selected star candidates in the area are at least 86% stars judging by the most common stellar classifications registered in SIMBAD (‘Star’, ‘low-mass*’, and ‘PM*’). Galaxy and QSO candidates are at least 79% and 72% stars, respectively, showing a small difference in the nature of these objects. For spectroscopic stars, those with QSO- or galaxy-like variability are more spread out in z – W1 and found at higher values of W1 – W2. This is illustrated in Fig. C.4. The variability does indicate a physical difference in these cases.

In W2 versus W1 – W2, the spectroscopic classes overlap at W2 > 15, and the variability selections overlap even more. Objects at –0.25 < W2 – W1 < 0 and 10 < W2 < 12 are mostly registered as stars in SIMBAD, and SDSS does not classify any of them as QSOs, although many have QSO-like variability.

SDSS-matched sources have relatively long time spans and more epochs per light curve, as shown in Fig. B.1. Hence, variability estimation of the full sample might be less accurate, spreading out sources in A-γ space. Therefore, removing the most sparsely sampled sources is important. To both include large datasets and be confident in the results, the balance of data cleaning will also be important in future surveys such as the LSST. Even for sources with >100 epochs over >5 yr, the spectroscopic classes still have an overlap at log A ~ –0.8 and γ ~ 0, but it is about 50% smaller in both log A and γ. Longer timescales may change the selected populations, for example by increasing the fraction of type II to type I AGN (De Cicco et al. 2015, 2019).

6.4. Photometric selection bias

We select pure sets of each class, but with low completeness and a bias for sources with specific parameters. With variability, we only select stars with low A; we know variable stars exist, but they are difficult to isolate, and so reduce completeness for stars and purity for galaxies and QSOs. Type I and Type II AGN differ in colour and SF, and so are not best selected by one simple set
of criteria either (De Cicco et al. 2022). The most densely populated variability region, at log $A \sim -0.8$ and $\gamma \sim 0$, is not covered by the criteria at all. The same goes for the dense colour diagram area at high W2. In Fig. 2, galaxies dominate a triangular area of $g - r$ versus $z - W1$ with two large clusters. The criteria only cover one of them. In SIMBAD, the cluster at high $g - r$ has more galaxies labelled as being part of a cluster, and especially as the brightest galaxy in a cluster. A more advanced selection method could solve these issues (Bruun et al. 2023).

The surveys are not completely representative of the sky, which is not observed uniformly (see Fig. A.1). We include fewer stars at low Galactic latitudes, where Galactic extinction has a greater affect on colours and there is a higher risk of mismatches with nearby sources. Stars have the fewest colour matches, indicating an under-representation in WISE or PS1. A change in ratios of object types and stellar subtypes affects purity. The reason is that the number of true or false positives depends on the selection of objects that are not equally difficult to distinguish from or as stars. For example, including more variable stars could lower the QSO selection purity, which is due to both the overlap with QSO variability and the increased prior probability of a classified object being a star. However, including more data would also mean there is more data to learn from. The balance of object-type frequencies is relevant if we apply Table 2 criteria to other datasets, and in evaluation of criteria of Schmidt et al. (2010) on PTF data. Completeness is computed independently for each class, but could also be affected by a change in the fractions of subtypes.

6.5. Perspectives

If we were to base selection criteria on distributions from SDSS-confirmed objects and evaluate on the same set, we would be overfitting. However, here the goal is only to estimate class distributions and relations between variability and colour. Machine learning could automatically classify all objects based on the SDSS-labelled subset. This would allow us to select more sources and examine probabilities of belonging to each class. It would also quantify how variability breaks degeneracies and improves selections based on colour and magnitude. This will be performed in Paper II of VILLAIN (Bruun et al. 2023), including a table of all variability parameters and classifications. Accurate photometric selections can identify new QSO candidates and prepare us for analysis of large surveys like the LSST. The optical variability of galaxies, including in the PTF R-band, is also known be useful for identifying AGN missed by other techniques (Baldassare et al. 2020). It would be interesting to study subtypes in terms of variability and redshift differences of type I and II AGN, as in De Cicco et al. (2022). Intermediate-redshift QSOs can have colours comparable to those of stars, and so variability could be more valuable for those (Yang et al. 2017). Another prospect is selecting non-variable stars for photometric calibration or for a homogeneous study of variability across stellar subtypes.

In the present study, we assume that the objects show simple power-law variability, but this is not necessarily a good model for all sources or on all timescales. One could fit for example exponential or DRW models instead and analyse the differences; although we expect the overall selections and challenges to be similar. Advanced models can capture more variability information but require more resources (Moreno et al. 2019). More parameters could be used for selection, such as proper motions or photometric redshifts.

7. Conclusion

We devised a procedure for the homogeneous analysis of 71 million PTF light curves. We fitted them with joint-likelihood SF models and studied regions in both variability and colour space. Structure function power-law variability is most useful outside the log($A) \sim -0.8$ and $\gamma \sim 0$ region. We select photometric sets of 99.3% spectroscopic purity for QSOs, 99% for stars, and 99.8% for galaxies. However, the spectroscopic classifications are incorrect for 20–50% of objects photometrically identified as QSOs but studied the Galactic Archival Telescope Data. This sample allows us to discover these rare cases and assemble a set of 31 404 QSO candidates according to colour and variability. With only variability, spectroscopic purity drops to 59.3% and with only colour, it is 98.7%.

Using SF joint likelihoods on the entire PTF survey, we show how variability might be used on future large datasets including the LSST. When new measurements are added to a light curve, complete reprocessing can be avoided, as likelihood information on the previous segment of the light curve is already stored in $A$ and $\gamma$. In a survey with a foreseen depth similar to that of the LSST, a colour-plus-variability method can provide a large sample of faint astrometric standards for the internal calibration of extremely large telescopes, which require objects beyond the depth of Gaia.

In each survey, and depending on computational resources, one must balance sample size and fitting accuracy via data cleaning and selection methods. The value of variability and colours depends on the survey and sources, but in general for PTF, cross-matching colours should be prioritised.

Considering simple cuts in both variability and colour, the completeness is at 12.5%, and so a machine learning method that balances purity and completeness has the potential to create larger QSO samples for studying cosmology, for example. This is examined in the companion VILLAIN paper (Bruun et al. 2023), where we also release a table of classifications and parameters for all fitted PTF sources. Such a large dataset would allow a complementary selection of rare objects, for example, such as lensed QSOs.

Acknowledgements. This work was supported by a Villum Investigator grant (project number 16599). S.H.B. was also supported by a grant from the Danish National Research Foundation (DNRF132). A.A. was also supported by a Villum Fonden grant (project number 12589). This work made 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. The Pan-STARRS1 Surveys were made possible through contributions by the Institute for Astronomy, the University of Hawaii, the Pan-STARRS Project Office, the Max-Planck Society and its participating institutes, the Max Planck Institute for Astronomy, Heidelberg and the Max Planck Institute for Extraterrestrial Physics, Garching, The Johns Hopkins University, Durham University, the University of Edinburgh, the Queen's University Belfast, the Harvard-Smithsonian Center for Astrophysics, the Las Cumbres Observatory Global Telescope Network Incorporated, the National Central University of Taiwan, the Space Telescope Science Institute, and the National Aeronautics and Space Administration under Grant No. NNX08AR22G issued through the Planetary Science Division of the NASA Science Mission Directorate, the National Science Foundation Grant No. AST-1238877, the University of Maryland, Eotvos Lorand University (ELTE), and the Los Alamos National Laboratory. The Pan-STARRS1 Surveys are archived at the Space Telescope Science Institute (STScI) and can be accessed through MAST, the Mikulski Archive for Space Telescopes. Additional support for the Pan-STARRS1 public science archive is provided by the Gordon and Betty Moore Foundation. The Pan-STARRS1 Surveys (Chambers et al. 2016), Pan-STARRS Data Processing System (Magnier et al. 2020a); Pan-STARRS Pixel Processing: Detrending, Warping, Stacking (Waters et al. 2020); Pan-STARRS Pixel Analysis: Source Detection and Characterization (Magnier et al. 2020b); Pan-STARRS Photometric and Astrometric Calibration (Magnier et al. 2020c); The Pan-STARRS1 Database and Data Products (Flewelling et al. 2020). Funding for the Sloan Digital Sky Survey IV has been
Appendix A: Sky maps

The sources with PTF light curves are mainly from the northern hemisphere. These are distributed according to the top diagram of Fig. A.1 after the data cleaning of Sect. 2.5. When we also match to SDSS spectroscopic classifications, the sources in the bottom diagram are left. The differences indicate a bias in photometry between the surveys, as mentioned in Sect. 6.4.

Appendix B: SDSS matched data

Of the fitted PTF sources, 2.5% are matched to spectroscopic classifications in SDSS. This subset has different parameter distributions from the full PTF sample. On average, the time spans are longer and the number of epochs is higher, as shown in Fig. B.1. The distributions of variability parameters $A$ and $\gamma$, and colours in the optical and mid-IR are shown in Fig. B.2 for SDSS matched data for comparison with the full sample in Fig. 1. The differences are discussed in Sect. 6.3. Figure B.2 also illustrates the selection criteria for stars, galaxies, and QSOs in Table 2 and the criteria of Schmidt et al. (2010):

\[
\begin{align*}
\gamma &> 0.055 \quad \text{(B.1)} \\
\gamma &> 0.5 \log_{10} A + 0.5 \quad \text{(B.2)} \\
\gamma &> -2 \log_{10} A - 2.25 \quad \text{(B.3)}
\end{align*}
\]

These criteria can point to potential misclassifications in SDSS, as discussed in Sect. 5.2, and an example of this is SDSS J120429.34+495814.4. This object is registered as a star in SDSS, but the selection criteria of this paper point to it being a QSO. The spectrum in Fig. B.3 has the broad emission lines characteristic of QSOs.

Fig. A.1. Sky distributions of fitted and accepted PTF light curves (top) and of those also matched to spectroscopic classifications in SDSS (bottom). The coordinates are equatorial, with RA increasing to the right. In grey areas, no data exist or remain after the data cleaning of Sect. 2.5.

Fig. B.1. Histograms of all PTF sources and sources matched to spectroscopic classifications in SDSS. The differences in time span and number of epochs show two ways in which the SDSS matched data are not representative of the entire PTF sample.
Fig. B.2. Heat maps of all data with spectroscopic classifications in SDSS. Comparing this with Fig. 1, we see SDSS has spectroscopic data focused on specific parts of the parameter space. Strict criteria for pure selection of stars (green), QSOs (red), and galaxies (blue) are overplotted. These are listed in Table 2. The fit-parameter panel also includes a black line showing the Schmidt et al. (2010) QSO selection.

Fig. B.3. Spectrum of SDSS J120429.34+495814.4 from the BOSS spectrograph. The source is classified as a star in SDSS-IV DR17, but it has the variability and colour parameters of a QSO candidate. The broad emission lines support the latter classification. This is one of 32 objects selected as a QSO candidate and whose SDSS spectra were classified as stars, of which 20% – 50% are QSOs based on visual inspection of the spectra. (SDSS-IV DR17, CC-BY license, skyserver.sdss.org/dr17/VisualTools/explore/summary?objId=1237658613058109587.)
Appendix C: Variability-selected objects

The photometric selections are performed based on either the variability criteria, the colour criteria, or both of Table 2. Some spectroscopically confirmed objects have conflicting photometric parameters, as discussed in 5.2. In Fig. C.1 the redshifts for spectroscopic galaxies are typically higher for variability-selected galaxy candidates than for star candidates. QSOs have similar variability except at $z < 0.3$ where more appear like stars or galaxies in $A$ and $y$.

Figures C.2–C.4 show colour diagrams for all combinations of spectroscopic and variability selected classes. This illustrates how the variability selection criteria are picking objects with different colour distributions, even when the objects are spectroscopically confirmed to belong to the same class. This is discussed further in Sects. 5.2 and 6.3.

Fig. C.1. Redshift distributions of galaxies (top) and QSOs (bottom) that appear as candidates of different classes based on variability criteria. The galaxies with QSO-like high variability are typically found at high redshifts of ~0.5, and those with star-like low variability have lower redshifts than other galaxies. SDSS QSOs are dominated by QSO-like variability except at $z < 0.3$. The bins of each diagram are normalised with respect to the total samples of spectroscopic SDSS galaxies and QSOs, respectively.

Fig. C.2. Galaxies with variability parameters typical of different types of sources. We see that the more variable galaxies (with QSO variability) have high $W2$ and $z - W1$. 

Fig. C.3. Colour diagram showing the distribution of galaxies with QSO variability (red points) compared to those with star variability (green points). The blue points represent galaxies with QSO and star variability. The bins of each diagram are normalised with respect to the total samples of spectroscopic SDSS galaxies and QSOs, respectively.

Fig. C.4. Colour diagram showing the distribution of QSOs with QSO variability (red points) compared to those with star variability (green points). The blue points represent QSOs with QSO and star variability. The bins of each diagram are normalised with respect to the total samples of spectroscopic SDSS galaxies and QSOs, respectively.
Fig. C.3. Quasars with different variability parameters are still mostly found in the same regions of the colour diagrams.

Fig. C.4. Stars are in the stellar locus of the upper panel, but with more spread for those with variability that is more typical of QSOs and galaxies. Galaxy-like variability is also found at higher $W_1 - W_2$ than for most stars. Based on inspection of SDSS imaging, the 73 SDSS stars with galaxy-like colours mostly resemble galaxies or a star–galaxy chance alignment. SDSS stars with $0.2 < W_1 - W_2 > 0.3$ and $W_2 < 13$ are also often not isolated. We do not see this for random subsets of all of the 12,470 SDSS stars with galaxy-like variability.
Appendix D: SIMBAD statistics

Candidate stars, QSOs, and galaxies are selected based on the criteria in Table 2. To understand the physical nature of the objects, we looked up the statistics of their ‘main type’ in SIMBAD. Most photometrically selected objects were registered with an expected label in SIMBAD, as shown in Fig. D.1.

Fig. D.1. SIMBAD ‘main type’ statistics for candidates by colour and variability. This shows the subtypes and contaminating sources for QSO (upper panel), star (middle), and galaxy (lower) candidates.

Appendix E: Gaia proper motions

We cross match with Gaia DR3 (Gaia Collaboration 2016, 2023) proper motions using the CDS X-Match Service (Boch et al. 2012). This is done for all sources within one arcsecond of the PTF coordinates. For those selected through variability, we limit the search to 1 million random sources of each class. In Fig. E.1 we show histograms of the sources with measured proper motions. As expected, these are generally lower for QSOs and higher for stars. The differences are smaller for variability-selected objects and they follow the distribution of spectroscopically confirmed stars almost perfectly, indicating that most of the objects with non-zero proper motions are stars. However, the objects with QSO-like variability have a higher spectroscopic purity (59.3%) in the SDSS. Sources selected by both colour and variability have distributions closer to those of the spectroscopically confirmed classes, especially for QSOs. We note that the sample sizes are also different, as not all sources have a match in Gaia and multiple Gaia sources can be detected per PTF source. For variability-selected objects, the Gaia sets are 60.6%, 92.0%, and 72.0% of the sizes of the QSO, star, and galaxy sets, respectively. For sources selected based on variability and colour, these numbers are 84.7%, 99.6%, and 0.5%, and for spectroscopically confirmed sources, they are 82.3%, 95.9%, and 3.7%. Gaia has proper motions for almost all stars and very few galaxies, as its design is optimised for detecting stars and the limit for proper-motion measurements is not as faint as the PTF or Pan-STARRS source-detection depth. As is shown by the comparison of SDSS and Gaia purity of the variability-selected QSOs, atypical proper motions may indicate misclassifications and mismatches between surveys, or interesting subclasses or flaring objects. The sum of the residuals between histograms of variability-selected QSOs and variability-selected stars (<0.01 for stars to calibrate the curves at high proper motions), when the total area under the curve is normalised to 1, is 0.29. This gives a lower bound on the QSO purity, in addition to the higher spectroscopic purity, as it does not take into account differences in how many objects are matched, chance alignments of stars, and so on. (Even SDSS QSOs usually have a proper motion measured nearby in Gaia.)
Fig. E.1. \textit{Gaia} proper motions for sources selected according to variability (top), colour and variability (middle), and spectroscopy (bottom). The histograms are normalised individually. All variability-selected sources approximately follow the distribution of SDSS stars, but when we also select by colour, the results are generally close to those of the spectroscopically confirmed objects. A high-purity and high-completeness cut at $\leq 2$ mas/yr, justified by the bottom panel, would yield $\approx 50\%$ purity in a sample of QSO candidates selected based on variability only.