Variability classification of PTF lightcurves

Master thesis

Written by *Sofie Helene Bruun*

November 2, 2020

Supervised by

Jens Hjorth and Adriano Agnello

University of Copenhagen

UNIVERSITY OF COPENHAGEN

FACULTY:	SCIENCE
INSTITUTE:	DARK, Niels Bohr Institute
AUTHOR(S):	Sofie Helene Bruun
EMAIL:	ndl212@alumni.ku.dk
TITLE AND SUBTITLE:	Variability classification of
PTF lightcurves -	
SUPERVISOR(S):	Jens Hjorth and Adriano Agnello
HANDED IN:	2.11.2020
DEFENDED:	November 2020

NAME ______________________________

SIGNATURE __________________________

DATE ______________________________

Abstract

Determining the nature of astrophysical sources can be done in many ways, and one of them is by the study of variations in luminosity. This thesis analyses how variability of lightcurves can be used for classification of sources. This could be useful in e.g. studying quasars or cataloguing standard stars for improved photometric calibrations. Four samples of lightcurves from the Palomar Transient Facility (PTF) are used: All sources in 3 patches covering 10×10 deg^2 each, quasars spectroscopically confirmed by SDSS, quasar candidates selected from WISE colours and known lensed quasars. Their magnitude changes over different time scales are fitted to a power law with an affine-invariant Markov chain Monte Carlo ensemble sampler. The fit parameters then describe their variabilities and can be used for classification. A suggested cut in variability for quasar selection in SDSS lightcurves is tested on the PTF lightcurves, and while it does select more sources in the quasar samples than in a random sample, it still only classifies 14 % of the spectroscopically confirmed quasars correctly – but 51 % out of those consistent to two σ with variabilities following a power law with non-zero exponent and amplitude. So while variability classification in PTF does have potential, optimisation of methods for data selection and utilisation of the variability information is needed.

Contents

Chapter 1

Introduction

1.1 Motivation

If the luminosity evolution of an object is known, is it possible to infer what kind of object it is, based on how it varies over time? That is the question this project seeks to answer. Or, more specifically, if we know the lightcurves of several objects, is the way their lightcurves vary useful in separating and classifying different types of astronomical objects?

Some objects may vary over time scales of days, some over months or years. Some do not seem to vary at all. So how can these cases be described? There will of course always be some variation in measured luminosity due to uncertainties. But after taking that into account, there is sometimes still some excess variation left. This excess can be fitted to estimate intrinsic variations of the source. For each time scale in the lightcurve, a so called structure function can estimate the excess variations, as introduced to astronomy by Simonetti et al [22]. With this, one can use the time separation of epochs to estimate the expected change in luminosity.

Objects with different variability will have different fit parameters when a structure function is fitted to the data, and therefore they can be recognised by their different positions in fit parameter space. But if their intrinsic variabilities depend more on the individual objects than the overall type, it will be difficult to classify objects this way. Large uncertainties can also make it difficult to get precise estimates of the intrinsic variability, as any variability besides that from uncertainties will then not be clearly necessary to explain the total variations. This project will study the extent to which the classifications are possible based on lightcurves from the Palomar Transient Facility (PTF).

This can be done by analysing a representative selection of lightcurves of the survey (or simply all of them). But one can also use a list of coordinates to known objects of a specific type and ask: What characterises the variabilities of these objects, and are they different from those of a random sample? Or from those of another type of object? If the parameters describing the variabilities are different, the method of variability classification of lightcurves can help distinguish this type of object from others. Not necessarily perfectly, but if the parameter distributions are clearly different, it means the variabilities contain information about what type of object the lightcurve belongs to, and that we can extract this information to classify objects – possibly in combination with other parameters.

The study of sources representative of all objects in a survey can also reveal how the objects behave in general. This exploratory research could for example split known categories into different variability subtypes. And if this is done using machine learning, the method would find them in a way that rely little on human biases.

There are multiple interesting perspectives in classifying objects this way. After determining which properties are associated with which kinds of objects, the method can automatically be applied to new objects. When we know which properties to look for, new sources can be classified by fitting a structure function and identify its position in the fit parameter space. And with machine learning, the variability of the object can be compared with every type of object the algorithm knows and it will output a probability that the new source belongs to each class.

Secondly, if an object has well-determined variability properties, this can be used to study the physics of the specific object. If the estimator is correct and two objects have different intrinsic variations in luminosity, there must be a physical reason for this.

Knowledge of the variability properties of sources can also be useful, not only for classifying single objects to understand them, but also for using variability classification for studying populations and extending object catalogues. An example is faint stars with low variability. If the luminosity of a star is known and we know it to be stable, it can be used as a standard star for calibrating measurements when measuring the luminosity of a nearby object. But if the star is too bright, it could saturate the equipment for a deep observation. Therefore, it is important to create a sample of faint standard stars.

Another example is quasars. The name is derived from "quasi stellar radio source", since these sources often appear as point sources on the sky, like most stars, but are actually extremely luminous objects much further away. If the centre of a galaxy is bright due to mass accretion around a supermassive black hole it is called an active galactic nucleus (AGN), and the brightest of these are called quasars. An accretion disk forms around the black hole, and as some of the matter falls in, $\sim$10 % of the rest mass of the in falling matter is converted to energy. This is what makes them so luminous [7]. These extremely bright AGN, are visible even at high redshifts (up to z $\sim$7.5 when the universe was only 690 Myr old [3]) and are therefore useful probes of the early universe. For example for studying galaxy clustering based on the positions of quasars[23].

Since gravity can bend light and act as a lens, if the light from a quasar passes something massive on its way to us, this can significantly impact how it looks when it arrives. When this happens, the quasar is called a lensed quasar. The detected point source can be amplified (or distorted and magnified for extended objects) and it is even possible to see multiple images of the same object. This means some lensed sources can be recognised in imaging by not being fitted well by the point spread function of the survey as a single point source – the point spread function describes how sources would normally have their shapes blurred in the images [12]. If the light from a quasar takes multiple paths around a lensing galaxy, it will produce multiple images with different travel times. This difference in travel time and travel distance depends on cosmology, and therefore lensed quasars can be used for measuring the shape and components of our universe [24]. But it has also been suggested to search for lensed quasars as variable, extended sources. Therefore, it is relevant to classify the variability of their lightcurves [15].

Larger databases of different objects will of course also give us better statistics of the properties those objects – including their variability.

Some have described the variability dependence on time scale with a power law. Schmidt et al. [20] applied this model on lightcurves from the Sloan Digital Sky Survey, to make rough variability criteria for quasar selection. This project will apply their model to a different survey, The Palomar Transient Facility, and test if lightcurves behave differently. In the future, the Legacy Survey of Space and Time (LSST) at the Vera C. Rubin Observatory will obtain lightcurves with $\sim$ 200 epochs over 10 years and thereby huge potential for variability analysis. Therefore it is relevant to check how the method performs on different surveys, especially when it can be used for automatic classification of large data sets as LSST is expected to include 3.2×10^{10} observations [13].

1.2 Surveys

To analyse variability over different timescales, one must use lightcurves from a survey with multiple epochs taken over a long time span. For example, Schmidt et al. [20] used Sloan Digital Sky Survey (SDSS) lightcurves from the The Sloan Foundation 2.5m Telescope at Apache Point Observatory in New Mexico. These lightcurves have ~ 60 epochs taken over $\sim$5 years over 320 deg^2 of the sky (the whole sky is $\sim$41253 deg^2, so they covered 7.8%). This is in the bands: u (3543 Å, ultra-violet), g (4770 Å, green), r (6231 Å, red), i (7625 Å, near infrared) and z (9134 Å, infrared). The limiting magnitudes of filters defines the limits of how faint sources can be detected – in the same order as the filters, they are 22.0, 22.2, 22.2, 21.3 and 20.5 [1]. The higher the magnitude, the fainter the object.

The SDSS database contains spectroscopic classifications including classifications of quasars. The positions of these are used in this project for querying PTF lightcurves at known quasar coordinates. These positions are taken from the BESTDR8 database over 9274 deg^2 (22 % of the sky) [34].

The Palomar Transient Facility is another wide-field survey with time domain imaging. The data is taken with the Palomar 48 Schmidt telescope at the Palomar Observatory in California. The Palomar Transient Facility Data Release Three (DR3) contains lightcurves from 1,010 nights taken over ~ 6 years and covers most of the sky – see Fig. 1.1. The data is in the g and R bands (6258 Å, red) with magnitudes in the AB magnitude system [17] and most data being in the R band. Therefore, the R band lightcurves are selected for variability analysis in this study. Note that the coverage map is of all epochal images in R – not the subset included in lightcurves. The lightcurve database includes data from $\sim 65\%$ of nights with R or g band data. The PTF lightcurve database contains 598,975,024 objects [33]. The limiting magnitudes are $R \sim 20.6$ and $g \sim 21.3$, so the survey is not as deep as SDSS [17]. A successor to the PTF survey exists, the Zwicky Transient Facility (ZTF), but this has only been running since 2017, so the lightcurves are shorter[4].

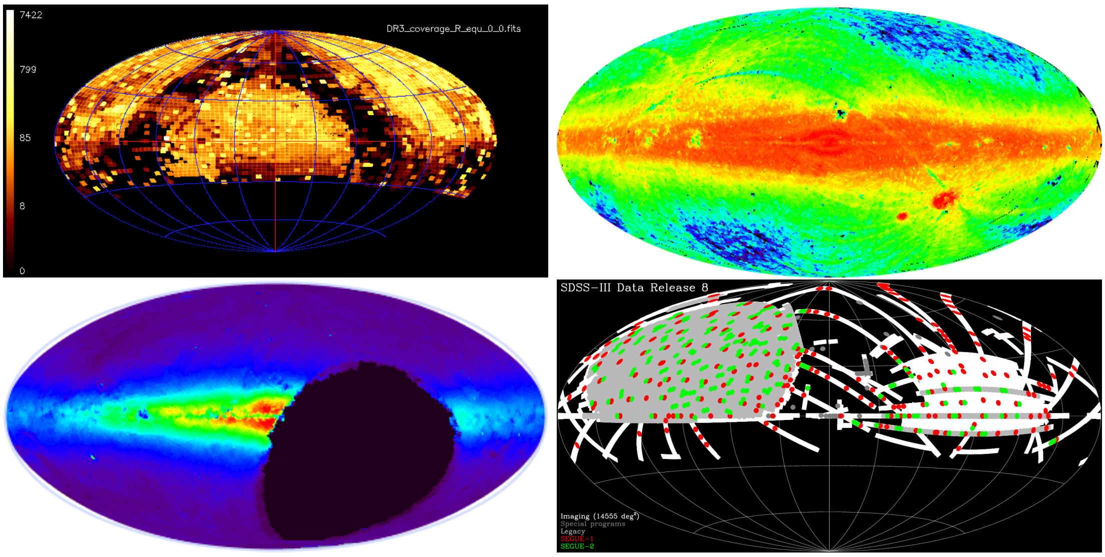

Figure 1.1: Coverage maps of the four surveys used in this project. Upper left: R band PTF data from DR3, DR2 and DR1 with 3.47 million images [33]; upper right: WISE with 747 million objects [29]; lower left: PS1 with 1.919 billion objects [30]; and lower right: the spectroscopic coverage (grey, red and green – white is imaging) of SDSS DR8 with 1.843 million spectra [34].

.

In addition to PTF lightcurves, colours from other surveys can aid the classifications and help explore the nature of the sources. For this, magnitudes from The Wide-field Infrared Survey Explorer (WISE) and Panoramic Survey Telescope and Rapid Response System (Pan-STARRS) are chosen. They cover a wide range of wavelengths useful for comparing colours with those of known objects. For example, quasars will usually appear bluer than stars in the bands of these surveys [2].

WISE is a space telescope with the bands $W1$, $W2$, $W3$ and $W4$ with 3.4, 4.6, 12 and 22μm respectively, covering the mid-infrared. Data from the AllWISE program data release II/328 contains 747 mio. objects spread over the whole sky, and the magnitudes are in the Vega magnitude system [26].

The Pan-STARRS data releases DR1 and DR2 are combined in the survey Pan-STARRS1 (PS1). This data is taken by the Pan-STARRS1 telescope in Hawaii, covering the sky at J2000.0 declination> -30 deg. The filters are g (481 nm), r (617 nm), i (752 nm), z (866 nm) and y (962 nm) and magnitudes are given in the AB magnitude system [25].

So together, WISE and PS1 cover most of the sky in the optical to mid infrared.

This project will perform queries to these PTF, WISE and PS1 over 4 searches and compare the resulting distributions of objects and variabilities.

- Queries for all sources in some square "box" patches on the sky
- Queries at coordinates to objects that are known to be blue in WISE and are therefore quasar candidates
- Queries at the coordinates of known lensed quasars
- And queries at the coordinates of quasars that have been spectroscopically selected in SDSS.

1.3 Structure function fitting

Each lightcurve contains points with times t, magnitudes m and magnitude errors σ. To analyse variability over different timescales, all points are compared to all other points. For a pair of points i and j, the differences in time and magnitude are $\Delta t_{ij} = t_j - t_i$ and $\Delta m_{ij} = m_j - m_i$, respectively. It is these changes in magnitude depending on time difference that can be modelled by a structure function. They can behave differently depending on the nature of the lightcurve, but let us assume it is useful to describe them with a power law as proposed by Schmidt et al. [20]. We then define the structure function as the squared variability:

$$SF(\Delta t) = A^2 \left(\frac{|\Delta t|}{t_0} \right)^{2\gamma}, \tag{1.1}$$

where t_0 is a specific timescale and Δt is some time difference (any value, not a specific measured Δt_{ij}). The magnitude difference is modelled with 2 parameters – the amplitude A of variations on the timescale t_0, and the power law index γ, where both are ≥ 0. Negative amplitudes would be nonphysical, and it is expected that the shortest time scales would have the smallest magnitude differences.

Table 1.1 gives the units of some parameters for which they are not always stated in the following. If a logarithm is taken, the parameter is of course divided by its units.

Parameter	**Units**
A	mag
gamma	none
q	mag/day
t	days

Table 1.1: Units of some parameters.

1.3.1 Interpretation of structure functions

For large variations, A would be large – it simply scales up the magnitude differences modelled by the rest of the function. γ describes how fast the structure function increases with Δt. On short timescales, the magnitudes have not had time to change much. On larger scales, for $\gamma < 1$, the magnitude differences would increase until Δt spans enough time that all timescales with different variabilities are covered, and then the curve flattens. If $\gamma > 1$, it means the variations keep increasing the magnitude variations, and the full scale of Δm is not covered. That is, the largest intrinsic magnitude differences have likely not been observed, because the lightcurves are too short. $\gamma \sim 0$ can occur if the lightcurve behaves as white noise [20] – per definition that would mean all frequencies are equally represented and so there is no overall change in Δm over Δt. If the lightcurves are long enough to cover all scales of Δm that would mean they behave as white noise on the longest timescales. It can require many years of observations to reach these and thereby describe the full variability. For models that describe AGN variability with a damped random walk, $\gamma \sim 0.5$ is expected, but it can take $\sim$ 10 years to estimate[16]. If the lightcurve signal is dominated by another type of noise, this would also impact the structure function - how would depend on which type.

1.3.2 Estimating the structure function

If the power law structure function of Eq. 1.1 is applied to a lightcurve, we can model the intrinsic variability for each Δt_{ij} as

$$V_{mod,ij} = A\left(\frac{|\Delta t_{ij}|}{t_0}\right)^{\gamma}. \tag{1.2}$$

Including the uncertainties we have the effective variability

$$V_{eff,ij}^2 = V_{mod,ij}^2 + \sigma_i^2 + \sigma_j^2. \tag{1.3}$$

Since some lightcurves will follow a general trend, and a constant rise or fall in magnitude should not count in the variability, it can be useful to add a parameter q for the overall, linear slope of the lightcurve. This can be subtracted from the magnitude differences by computing $\Delta m_{ij} - q\Delta t_{ij}$.

To find the likelihood of a set of values for A, γ and q, all pairs of Δm_{ij} and Δt_{ij} in the lightcurve must be evaluated. Assuming a given Δt_{ij} gives a gaussian distribution of Δm_{ij}, the likelihood contribution of each pair is

$$L_{ij} = \frac{1}{\sqrt{2\pi V_{eff,ij}^2}} \exp -\frac{\left(\Delta m_{ij} - q\Delta t_{ij}\right)^2}{2V_{eff,ij}^2}. \tag{1.4}$$

And the total likelihood is then

$$L = \prod_{i,j}\left(L_{ij}\right). \tag{1.5}$$

For optimisation purposes, it is easier to use log likelihood and this would be

$$LLH = Lp = \sum_{i,j}\left(\ln L_{ij}\right) + \ln(p) \tag{1.6}$$

if we include p as the prior.

1.3.3 Prior

For small errors, one could use

$$\ln(p) = \ln\frac{1}{1+\gamma^2} + \ln\frac{1}{A} + \ln\frac{1}{1+q^2}, \tag{1.7}$$

as an uninformative prior, but this does not hold when the errors $\sigma_{ij} = \sqrt{\sigma_i^2 + \sigma_j^2}$ are comparable to Δm_{ij}. For example, $\frac{1}{A}$ could reward an algorithm for choosing lower and lower values of A when the data cannot constrain it well. Here, we will instead adopt a uniform prior for simplicity – corresponding to not adding a prior.

1.3.4 MCMC

To find A, γ and q, and the distributions of their probable values for each lightcurve, these are used as free parameters in a fit using an MCMC (Markov Chain Monte Carlo) method [18, 10] with several walkers. The MCMC uses the data (lightcurves) and priors to estimate posterior distributions of each parameter. An initial value of A, γ and q is chosen for each walker that will then jump randomly in parameter space by taking random steps away from its current position. For each suggested set of fit parameters, some specific distribution of data would be expected. So the likelihood of getting each data point (a set of Δm_{ij} and Δt_{ij}) is computed and combined into a total likelihood of seeing this light curve if it is described well by the suggested fit parameters. The likelihood is described in Section 1.3.2. The probability of accepting the new fit parameter values as the next step in the walker is then determined by the difference in log likelihood and a random threshold – so the new value will sometimes be more unlikely than the previous step. After the "burn in" phase, where the walkers are still trying to reach an equilibrium, the walkers sample fit parameters values with a distribution that is representative of their probability. This is useful for estimating not only median values but also errors even when these are asymmetrical.

The MCMC used in this project is defined by the Python emcee package [8]. This is an affine-invariant ensemble sampler meaning it is invariant to affine transformations (transformations where parallel lines stay parallel) and is insensitive to parameter covariances. The walkers interact, so the steps taken by each of them are conditional on the positions of other walkers [21].

A and γ can never be negative, and so, the values will be extremely low for some objects. A might be estimated as 10^{-6} but we cannot measure 10^{-6} of a magnitude, so it consistent with 0 - unless combined with a high γ making it noticeable over long time scales.

1.3.5 Empirical structure function

As seen in Eq. 1.4, when the MCMC optimises the gaussian, it is creating a gaussian function for each i, j with mean $q\Delta t_{ij}$ and width $V_{eff,ij}$. Then it checks how well the measured Δm_{ij} matches this gaussian. q adjusts the mean of the gaussian, whereas A and γ adjust the width. So the variations in the lightcurve that will spread out the values of Δm_{ij} are modelled by $V_{eff,ij}$, if we include both intrinsic variations and variability caused by measurement errors.

A $V_{eff,ij}$ that is too small, corresponds to a sharp peak and gives a low L_{ij} for the real values of Δm_{ij} that would vary enough to be far from the peak. A large $V_{eff,ij}$ would spread gaussian giving a lower L_{ij} than it needed to be for the Δm_{ij}. In a good fit, the fit parameters ensure that

$$V_{eff,ij} \approx \Delta m_{ij} - q\Delta t_{ij}. \tag{1.8}$$

So to evaluate how well the fit matches the data, it can be useful to plot these together. To have an expression only depending on the fit, let us isolate $V_{mod,ij}$ using Eq. 1.3:

$$V_{mod,ij}^2 + \sigma_i^2 + \sigma_j^2 \approx (\Delta m_{ij} - q\Delta t_{ij})^2 \tag{1.9}$$

$$V_{mod,ij}^2 \approx (\Delta m_{ij} - q\Delta t_{ij})^2 - \sigma_i^2 - \sigma_j^2 \tag{1.10}$$

$$A^2 \left(\frac{|\Delta t_{ij}|}{t_0} \right)^{2\gamma} \approx (\Delta m_{ij} - q\Delta t_{ij})^2 - \sigma_i^2 - \sigma_j^2. \tag{1.11}$$

It is not possible to make the two sides completely independent in this case, since q appears on the right hand side and all parameters are fitted together.

If the expression to the left corresponds to using the structure function1.1 on a specific time difference, we could use the expression to the right to create an empirical estimate of what the structure function would be – that is, create an empirical structure function. This is done by averaging over bins of data with similar $|\Delta t_{ij}|$:

$$SF_{emp}(\Delta t) = \frac{1}{N}\sum_{i,j}\left(\Delta m_{ij} - q\Delta t_{ij}\right)^2 - \sigma_i^2 - \sigma_j^2, |\Delta t_{ij}| \sim \Delta t \text{ and } i < j, \tag{1.12}$$

where N is the length of the sum. One could use $i \neq j$ instead, but then N would double, leading some to use a notation with $2N$ instead[9]. The uncertainty in each bin is

$$\sigma_{SF,emp}(\Delta t) = \frac{1}{\sqrt{N}} RMS\left(\left(\Delta m_{ij} - q\Delta t_{ij}\right)^2 - \sigma_i^2 - \sigma_j^2\right). \tag{1.13}$$

Chapter 2

Method

2.1 HPC

The High Performance Computing Centre at the University of Copenhagen (HPC/UCPH) [31] contains the DARK cluster with 70 nodes. Each node can use up to 64 GM memory, has 32 cores and jobs can take a maximum of 14 days. These jobs are submitted in jobscripts via the SLURM system for managing Linux clusters [36].

2.2 Querying

2.2.1 Box search

The PTF lightcurves database is currently not set up for all sky surveys, but it is possible to query everything by dividing the sky into smaller patches. This could be done with astroquery queries of the dataset on the NASA/IPAC Infrared Science Archive (IRSA) [27] from inside python, but it has a maximum box query area of $60'' \times 60''$. It is faster to make "all sky" queries with strict limits on J2000.0 right ascension (RA) and J2000.0 declination (Dec) as Structured Query Language (SQL) queries to IRSA through a terminal or via the web interface Gator [32].

A script is created for doing this with 10 by 10 degree boxes covering the whole sky with RA 0 to 360 degrees and Dec -90 to 90 degrees. Of course, this means the patch sizes vary, since the observed spherical projection of objects is converted to a rectangular map. For a query from RA 0 to 10 and Dec 20 to 30, it would look like this:

```
curl -v -o /storage/dark/shbruun/output_all/ptf_11_0p0_10p0_20p0_30p0.csv
"https://irsa.ipac.caltech.edu/TAP/async?QUERY=SELECT+
LC.ra,+LC.dec,+LC.obsmjd,+LC.mag_autocorr,+LC.magerr_auto,+LC.oid
+FROM+ptf_lightcurves+AS+LC+WHERE+
LC.ra>=0.0+AND+LC.ra<10.0+AND+LC.dec>=20.0+AND+LC.dec<30.0+AND+
LC.fid>1+AND+LC.nobs>=10+AND+LC.oid>0+AND+
LC.mag_autocorr>-10+AND+LC.mag_autocorr<30
+ORDER+BY+LC.oid&FORMAT=CSV&PHASE=RUN"
```

This command sends an asynchronous query with some constraints, selects some columns to save and saves the data in CSV format to a file called "ptf_11_0p0_10p0_20p0_30p0.csv" (query file no. 11) in the HPC folder /storage/dark/shbruun/output_all. fid stands for Filter ID and here it selects the R filter. The number of observations in each lightcurve is $nobs$, and there must be at least 10 – but we will have to check this again later in the analysis, as some points could be discarded as outliers. OID stands for Object ID and must be > 0 here to avoid including cases where points were not identified as belonging to a specific object. Ideally, mag_autocorr would already have constraints of > 11 and < 22 here, but the queries were completed in an early stage of the project, so they are imposed later.

In total, 648 SQL queries are sent to 21 different nodes. Max 32 per node, as they have 32 cores each. Each asynchronous command returns a location containing the results if the query is complete. However, the IRSA website cannot handle too many queries at a time – therefore, not all of them return data when they are sent over a short time span. And some patches simply do not contain any PTF lightcurve data fulfilling the criteria.

So another script is created to check the phase of each query and download the results if they are ready. Then, the result files are sorted by size in case the analysis will be to slow for the larger files.

2.2.2 Coordinate search

If coordinates to interesting objects are already known, it can be faster to find their PTF lightcurves by sending individual queries.

For each set of coordinates, PTF is queried $5''$ around it with astroquery.irsa [27] for the position, magnitude, magnitude error, time and OID associated with each data point.

Lensed quasars

The lensed quasars coordinates are from the CASTLES [19] and SQLS [12] databases of known lensed quasars.

WISE colour-selected quasars

The WISE quasar candidates are selected with an all-sky query in Gator with the constraints

```
((w1mpro-w2mpro)>(0.7 + sqrt(power(w1sigmpro,2)+power(w2sigmpro,2)))) and dec
>-30 and w1mpro >10 and w1mpro <17 and w2mpro >9 and w2mpro <15.6
```

meaning the colours are blue and we are looking at positions where *PTF* is likely to have data (a few PTF points are registered at lower declinations).

SDSS spectroscopically selected quasars

10,000 SDSS quasars classified from spectroscopy are selected with the following SQL query submitted to the SDSS database via the SDSS website [35]:

```
SELECT TOP 10000
   p.ra,p.dec
FROM PhotoObj AS p
   JOIN SpecObj AS s ON s.bestobjid = p.objid
WHERE
   s.class='QSO'
```

2.2.3 PTF lightcurves

After the PTF lightcurve points have been gathered, it is time to clean the data, group points together by OID and find related points in WISE and Pan-STARRS .

First, PTF points with magnitudes < 11 or > 22 are removed. Then points with identical OID's are identified, and if at least 10 points belong to a lightcurve it is kept and the mean coordinates to it are computed. The median PTF magnitude is computed and the most extreme outlier is removed if it is more than 3 magnitudes away from the median. Then the code checks if the lightcurves still have $>$10 points.

Finding the closest lightcurve (Coordinate search)

In the case of a search at coordinates for specific objects, the lightcurve closest (and within $5''$) to each set of coordinates is identified.

This process works by creating a KD tree [5] with the coordinates to all the lightcurves. The KD tree organises the points in a tree structure of k dimensions – 2 in this case, since the coordinates contain 2 values per point. This allows for fast searches for all lightcurves within a certain distance to each query coordinate – $5''$ in this case. The squared euclidian distances are then computed for these lightcurves and the nearest object identified.

All unique lightcurves associated with a set of query coordinates are then kept.

2.2.4 Additional surveys

For all lightcurves, WISE and PS1 are queried $5''$ around their mean coordinates. To avoid sending too many queries, the lightcurve positions are used in a version where they are rounded to nearest $0.1''$ and then duplicates are removed.

The queries are performed with astropy.Xmatch [28] on the Vizier datasets. In both cases, the saved information for every point is coordinates, OID's, magnitudes and magnitude errors. For WISE ('vizier:II/328/allwise'), the magnitudes are $W1$ and $W2$, and for PS1 ('vizier:II/349/ps1'), they are g, r, i, z and y.

For every point, it is checked that none of the columns returned NaN – otherwise the point is removed.

The points are then grouped by OID and the closest source to each PTF lightcurve is found – but this time, the sources do not have to be unique. That is, some WISE or PS1 sources can be associated with multiple PTF lightcurves, as this project is primarily about analysing those. The associated magnitudes from WISE and PS1 are still the most likely ones to belong to the same object based on the distance. The lightcurves without counterparts in both WISE and PS1 are removed. When nearest WISE or PS1 object contains multiple points, only minimum magnitude is saved as this is brightest point – but for PS1 the magnitudes must be $>$10 due to outliers.

In the end, the saved data consists of:

- PTF:
  - OID's
  - Lightcurves with:
    - Magnitudes (R)
    - Magnitude errors
    - Times
    - Positions (RA, Dec)
- WISE:
  - Magnitudes ($W1$, $W2$)

Magnitude errors

- PS1:

  Magnitudes (g, r, i, z, y)

  Magnitude errors

2.3 Structure function fitting

For the structure function fitting, the code starts by loading the results of the querying process for the relevant search process. Then, it sets the hyperparameters for the MCMC and fits the sources one by one. For every object, it extracts the times, magnitudes and magnitude errors for the lightcurve, fits them with the MCMC from the emcee python package described in Section 1.3.4, and uses the resulting fit parameter distributions to estimate the median values, their errors and the errors on the structure function.

The MCMC uses 16 walkers and 500 steps to explore the 3 dimensional parameter space of A, γ and q. The 150 first steps are discarded, giving the burn in 150 steps to finish. This seems to be enough based on plots of the chains made by each walker, such as the example in Fig. 3.10. If the burn in did not finish completely, the bias will be small after basing the parameter estimates on the last 350 steps.

The log likelihood function used is described in Section 1.3.2. If a walker proposes a step with negative A or γ, the log likelihood is simply set to $-\infty$, so that such a step can never be accepted.

The initial values $A_{ini,i}, \gamma_{ini,i}$ and $q_{ini,i}$ for each walker i are chosen randomly to make sure they explore different parts of the parameter space – at least to begin with.

Shorter lightcurves will likely need to explore a wider space, since they have fewer points to constain the parameters – and the RMS used in A_{ini}. So in these cases, the spread of initial guesses includes a dependency of the number of points in the lightcurve, N. Specifically, they are set to

$$s\left(\begin{bmatrix} A \\ \gamma \\ q \end{bmatrix}\right) = \begin{bmatrix} 0.02 \\ 0.5 \\ 0.0001 \end{bmatrix} \frac{\sqrt{20 \times 19}}{\sqrt{N(N-1)}}. \tag{2.1}$$

This s is chosen because the values in the array on the right hand side worked well in early experiments with a lightcurve of length $N = 20$. One could make further tests by evaluating e.g. the average uncertainties of the fit parameters depending on the initial values.

The inital values of all walkers are chosen in the following way:

- A is initiated using the RMS of the magnitude differences from the mean within 5 σ of the median. That is

  $$RMS = \sqrt{\frac{1}{N}\sum_{i}^{N} |m_i - \langle m \rangle|^2}, |m_i - \text{median}(m)| < 5\sigma_i, \tag{2.2}$$

  unless < 3 points are close to the median - then RMS uses all points in the lightcurve.

  The code now choose a uniformly random value in the interval $A_{ini} \in [\min(0, RMS - s(A)), RMS + s(A)]$

- $\gamma_{ini} \in [\min(0, 0.15 - s(\gamma)), 0.15 + s(\gamma)]$

- $q_{ini} \in [0 - s(q), 0 + s(q)]$

A describes the amplitude of the variations and so RMS from the mean should indicate the right scale of this parameter – notice that the σ_i also contribute to RMS, however.

After the MCMC has sampled values of the fit parameters, it saves the single best value and its errors for each parameter. The median values are used as the best fit values. Errors are computed by sorting all the parameter values and selecting the values one standard deviation away from the median to each side. The differences between these values and the median are then used as the asymmetrical errors on the fit parameter. 68 % of the generated values are within these errors of the median.

The structure function SF from Eq. 1.1 will later be computed based on the best fit values of the fit parameters. But due to the asymmetrical errors of the fit parameters, the errors on SF for each lightcurve are not estimated with standard error propagation. Instead, all the fit parameter values (after burn in) from the MCMC samples are used. For each set of fit parameter values, SF is computed on a range of Δt values logarithmically spaced between the minimum and maximum Δt of that lightcurve. 10 Δt values are chosen per lightcurve to avoid saving too much σ_{SF} data. Then σ_{SF} is computed based on percentiles of the SF values by the same process used for estimating the errors on the fit parameters.

2.4 Classification

It would be interesting to classify with an unsupervised clustering algorithm, but to begin with, we can explore the properties of some rough classifications and compare with the variability quasar selection by Schmidt et al. [20].

The structure functions are computed with $t_0 = 1$ for simplicity since t is already measured in days, but Schmidt et al. used $t_0 = 1$ yr. So for comparison, the fit parameters are converted. We just need to change A to A_{yr} since

$$A \left(\frac{\Delta t}{t_0} \right)^{\gamma} = A \, (\Delta t)^{\gamma} \, t_0^{-\gamma} \tag{2.3}$$

corresponds to changing A by a factor $t_0^{-\gamma}$. Since t is given as a MJD (Mean Julian Date), Δt is in days and therefore the conversion is

$$A_{yr} = A t_{yr}^{-\gamma}. \tag{2.4}$$

Note that the errors in A_{yr} depend on γ too – so for large γ and σ_γ, A_{yr} is less reliable. For the exact errors on A_{yr}, it is easiest to rerun the analysis with t_0 = 1yr since the errors on A and γ are not gaussian (cf. the distributions in Fig. 3.9).

To analyse the results, 8 classes of objects are created. The point of them is not to create perfect classifications but to get an idea of how different types of objects behave. Two branches in the $A - \gamma$ plane are observed in the box search results, so their definitions are included here – and like for the Schmidt quasar candidates their selection criteria are defined by eye and illustrated in Section 3.1.1.

- **All**: All accepted sources.
- **Low** σ_m: $\frac{median(\sigma_m^2)}{median((\Delta m - q\frac{\Delta t}{t_0})^2)} < 1$, where t_0 =1 day. That is, the median squared magnitude error is lower than the median variation observed, so the structure function has some excess variation to fit.

- **Well defined**: $\frac{A}{\sigma_{A,-}} > 2$ and $\frac{\gamma}{\sigma_{\gamma,-}} > 2$, so the fit parameters are far from 0. This is based on $t_0 = 1$ day. If A is converted to A_{yr}, the errors change because γ is included in the conversion.
- **Quasar candidates**: "Well defined" and colour selected by $W1{-}W2 > (0.7{+}\sqrt{\sigma_{W1} + \sigma_{W2}})$.
- **Standard star candidates**: Colour selection by $W1 - W2 < (0.1 + \sqrt{\sigma_{W1} + \sigma_{W2}})$; A and q being consistent with 0: $\frac{A}{\sigma_{A,-}} < 3$, $\frac{q}{\sigma_{q,+-}} < 3$ using the error in the relevant direction depending on the sign of q; and A and q being near 0 $A{+}3\sigma_{A,+} < 0.01$, $q{+}3\sigma_{q,+} < 2.7{\times}10^{-5}$ mag/day, $q + 3\sigma_{q,+} < 2.7 \times 10^{-5}$ mag/day (a cut corresponding to a change of max 0.01 mag/yr, but it still has to be consistent with 0)
- **Schmidt quasar candidates**: "Well defined", $\gamma > 0.055$, $\gamma > 0.5 \log_{10} A_{yr} + 0.5$ and $\gamma > -2 \log_{10} A_{yr} - 2.25$.
- **Branch 1**: $\gamma < -1.47 \log_{10} A_{yr} - 0.05$ and $\log_{10} A_{yr} < -1.15$
- **Branch 2**: $\gamma > -1.47 \log_{10} A_{yr} - 0.05$, $\gamma < -1.47 \log_{10} A_{yr} + 0.21$, $\log_{10} A_{yr} > -1.95$ and $\log_{10} A_{yr} < -1.15$

Chapter 3

Results

3.1 Querying

3.1.1 Box search query results

The box searches with the fewest selected sources (that found something) are box 5 (0<RA<10 deg, -40<Dec<-30 deg), 6 (0<RA<10 deg, -30<Dec<-20 deg) and 71 (30<RA<40 deg,80<Dec<90 deg). These are selected for easier further processing and contain 16147 accepted sources combined. The positions along with the other completed PTF box queries are illustrated in Fig. 3.2. Fig. 3.1 shows a histogram of the number of points in each object in the three selected boxes – the vast majority contain 10-15 points. 59 objects (0.03 %) were removed for containing only 9 points after an outlier 3 mag from the median was removed, leaving the total of 16147 accepted objects.

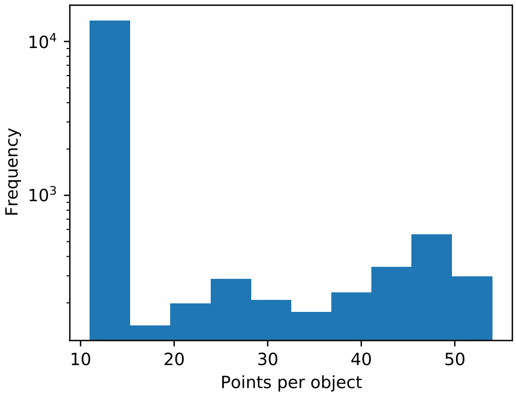

Figure 3.1: Histogram of the number of data points per accepted PTF lightcurve for the 3 selected boxes. Note that it is logarithmic in the y axis.

.

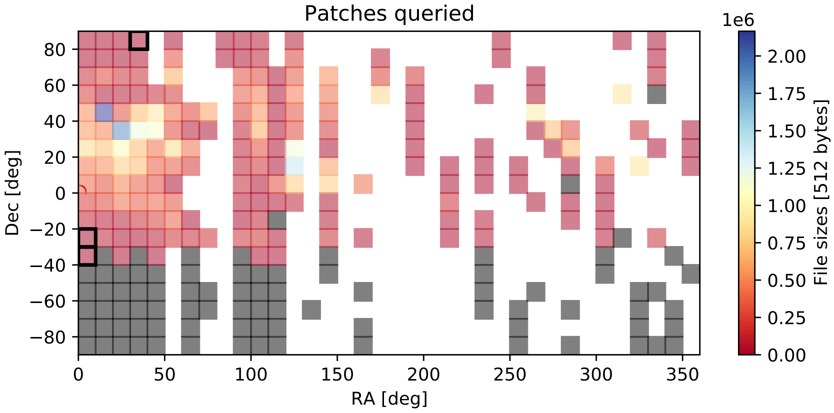

Figure 3.2: Map showing the file sizes of outputs from completed PTF queries. The red boxes with black outlines have been selected for further analysis. Black boxes represent completed queries that found no data.

.

3.1.2 WISE quasar candidate search query results

The WISE query returned 2081102 quasar candidates with Dec> 0 deg, and PTF searches for lightcurves at 10,000 of the positions, randomly selected. At the 10,000 coordinates queried, PTF finds 239939 points in total. 727 are removed for $m < 12$ and 54 for $m > 22$, however. The remaining points are found near 4543 of the original coordinates.

Assuming each OID corresponds to an object, 8246 different objects are found in PTF. 4223 of them have at least 10 points. Out of these, 2936 unique objects are identified as being the closest one to one of the coordinates.

3480 points in PS1 are close to these objects, and after searching for the closest matching object within $5''$ (not necessarily unique), all 2936 objects have data in PS1 and are therefore kept. All of the accepted PS1 objects contain 1 point. But 224 of these are removed for having PS1 magnitudes < 10.

So in the end, 10,000 coordinates led to 2712 accepted objects (at 27 % of the coordinates) with a total of 137533 PTF points (57% of the points found in PTF).

Fig. 3.3 contains maps of the original coordinates for the colour selected quasar candidates, the points found within $5''$ in PTF and the final selection of matching PTF sources. The number of sources is of course much lower, given that all accepted sources contain at least 10 points and almost half of the original points were not included in any accepted source. The number of points per object and per WISE quasar candidate coordinate are plotted in Fig. 3.4. Both decline steeply. One set of coordinates is within $5''$ of ~3500 PTF points, but they do not all belong to the same source (or many were not accepted for other reasons), as the maximum number of points in an accepted lightcurve is ~750.

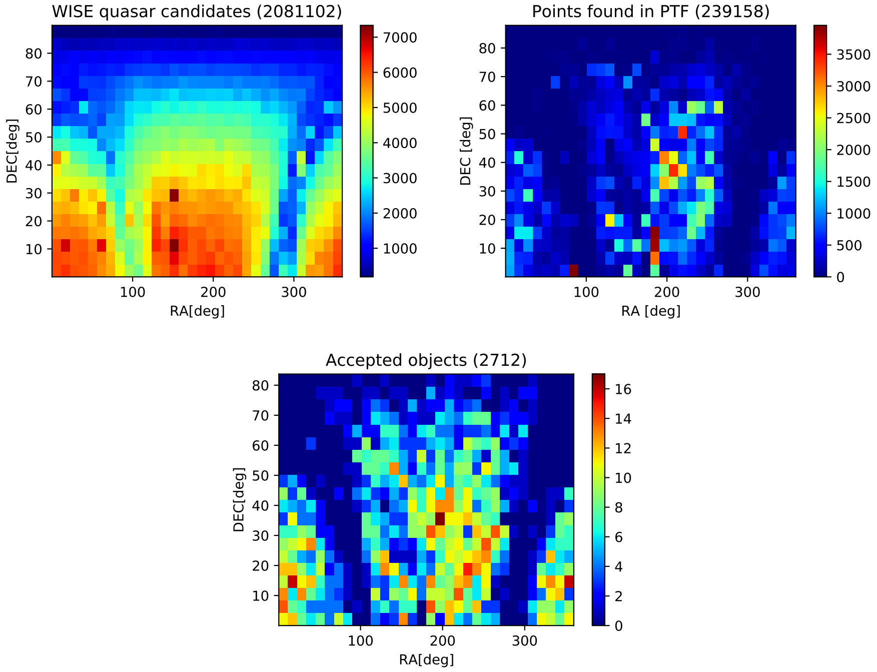

Figure 3.3: Left: Coordinates of quasars in WISE. Right: Matching points in PTF if ≥ 10 are found within a search cone of radius $5''$. Centre: The final selection of objects. Note the logarithmic y-axes.

.

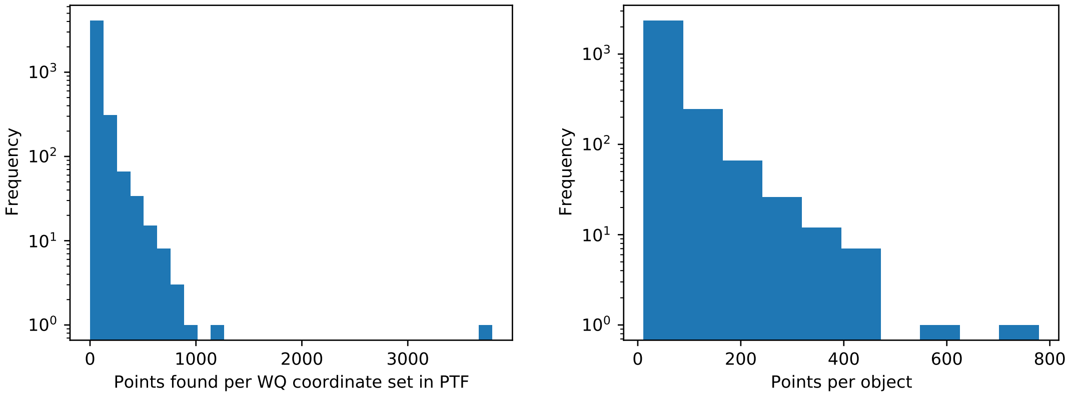

Figure 3.4: Left: Number of PTF points found per coordinate query for the WISE quasar candidates. Right: the number of points in each accepted object.

.

3.1.3 Lensed quasar search query results

PTF was searched at 337 coordinates of known lensed quasars and found data near all of them. Out of the 29047 PTF points, 85 were removed for m <12 and 6 for $m > 22$ – and 20563 were removed for being duplicates (meaning some of the lensed quasar coordinates are very close).

The PTF points had 314 different object ID's and 148 of those objects were removed for containing <10 points, so 161 are left. 158 of these were the closest unique object to a pair of lensed quasar coordinates. They were when matched by 135 WISE objects – and 150 PTF objects were matched to one of them. PS1 found 229 objects near these and all 150 objects were matched to one of them. 7 objects were then removed for at least one of PS1 magnitudes being <10. All the PS1 objects contained 1 point, but some WISE objects had multiple. The number of points found in WISE vs. PTF objects are plotted in Fig. 3.6.

143 objects with a total of 7101 PTF points were accepted after searching the 337 positions. So an object were accepted at 42 % of the search coordinates, and 24 % of the PTF points were accepted – or 84 % of the non-duplicate PTF points.

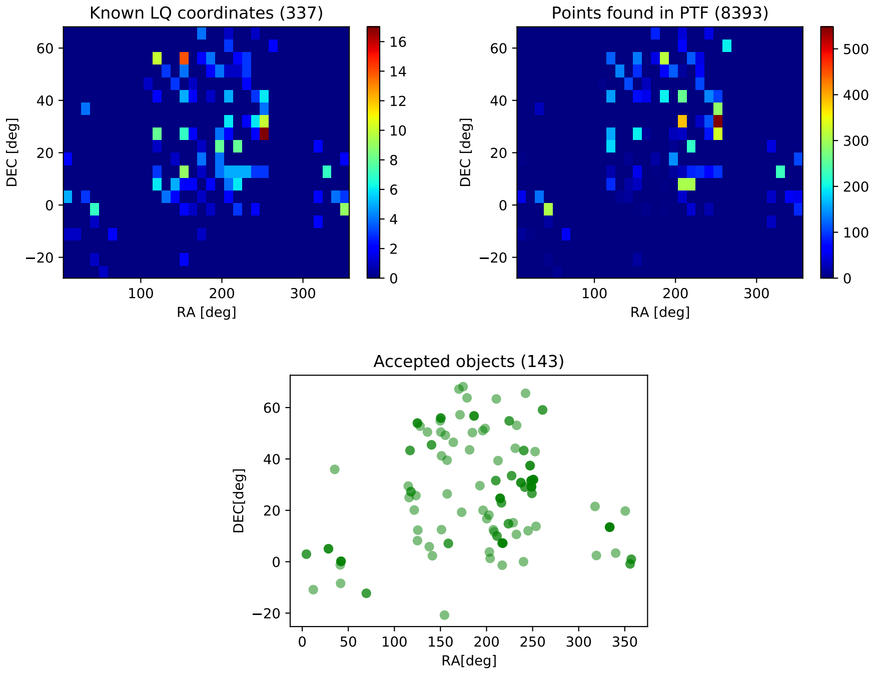

Figure 3.5: Left: Coordinates of quasars in WISE. Right: Matching points in PTF if ≥ 10 are found within a search cone of radius 5″. Centre: the final selection of objects.

.

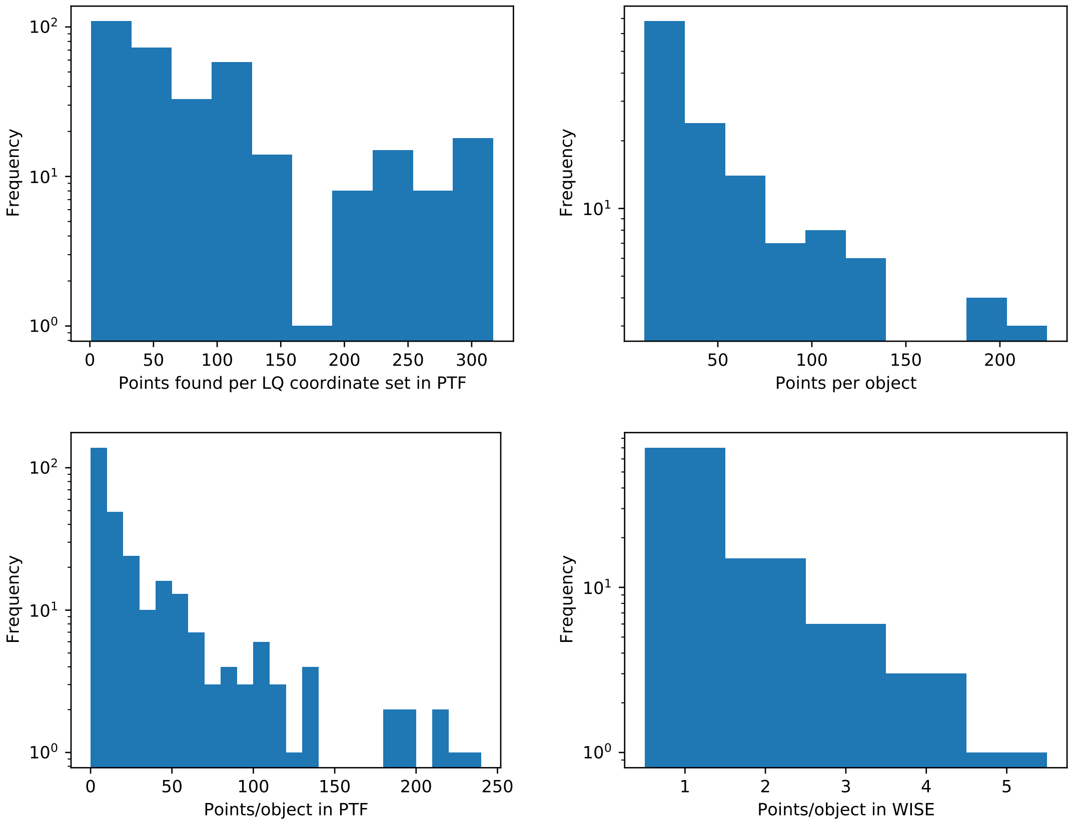

Figure 3.6: Upper left: Number of points found per coordinate query for the lensed quasars. Upper right: Number of PTF points in each accepted object. Lower left: Number of PTF points found per object if any. Lower right: The number of WISE points per WISE object associated with the PTF objects. The number of PS1 points per PS1 object are not plotted, as it was always 1.

.

3.1.4 SDSS quasar search query results

10000 SDSS coordinates to spectroscopically selected quasars were searched if Dec>0 deg – this was true for 8050 of the positions. PTF data was found at 7270 of those.

Out of the 585098 PTF points, 1688 were removed for having m<12 and 23 removed for m<22. None of them were duplicates. These points belonged to 13495 objects, and 9220 had >10 points.

6243 of these PTF objects were the closest unique one to a set of SDSS coordinates. 6050 WISE objects matched their coordinates, and 6034 of them were the closest match to a source. 6943 PS1 matched the sources left, and the PTF objects could all be matched to a close PS1 object. All WISE and PS1 objects contained 1 data point. But 97 objects were removed for having a PS1 magnitude <10.

So 5937 objects were accepted and matched to 8050 coordinate sets (74% of them). They contain 336874 PTF points in total, meaning 58 % of them were accepted.

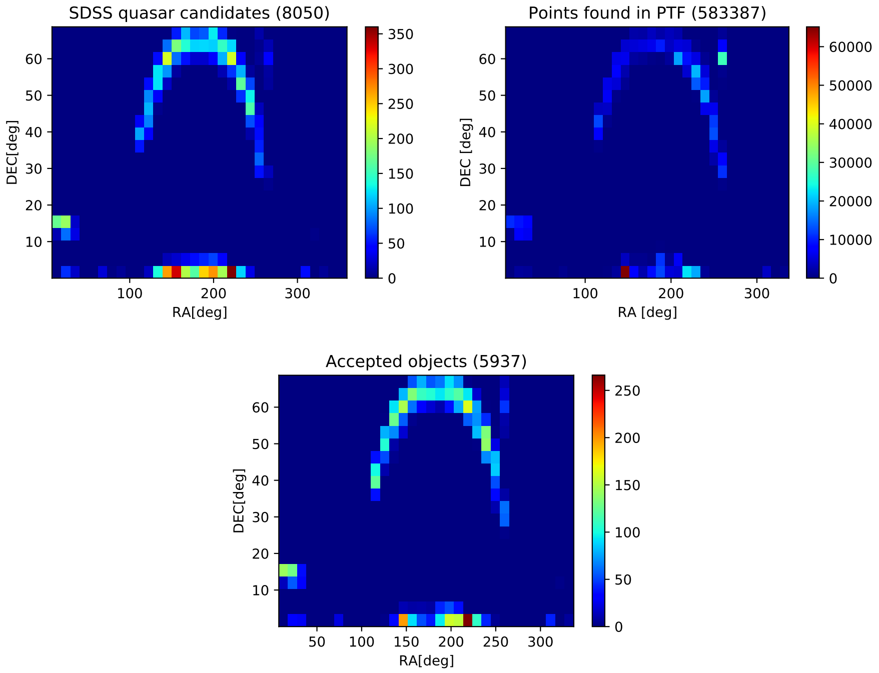

Figure 3.7: Left: Coordinates of quasars in SDSS; right: Matching points in PTF if ≥ 10 are found within a search cone of radius $5''$; and centre: The final selection of objects.

.

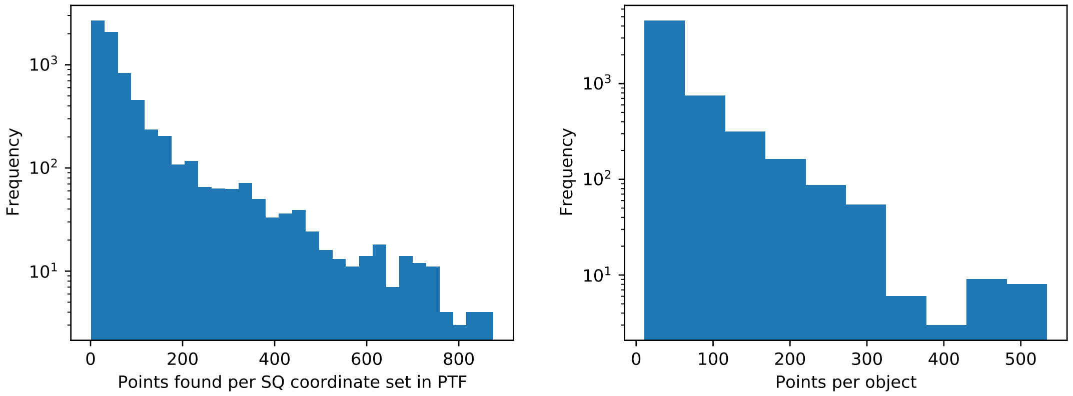

Figure 3.8: Left: Number of PTF points found per coordinate query for the SDSS quasars. Right: The number of points in each accepted object.

.

3.2 Structure functions

MCMC values for two objects are plotted in Fig. 3.9 – the first one random and the second selected to have A significantly different from 0. The median values and 1 σ differences illustrated in the plots are saved as the fit parameter values and uncertainties for those objects.

Fig. 3.10 shows the chains (including burn in) leading to the distributions for the second object.

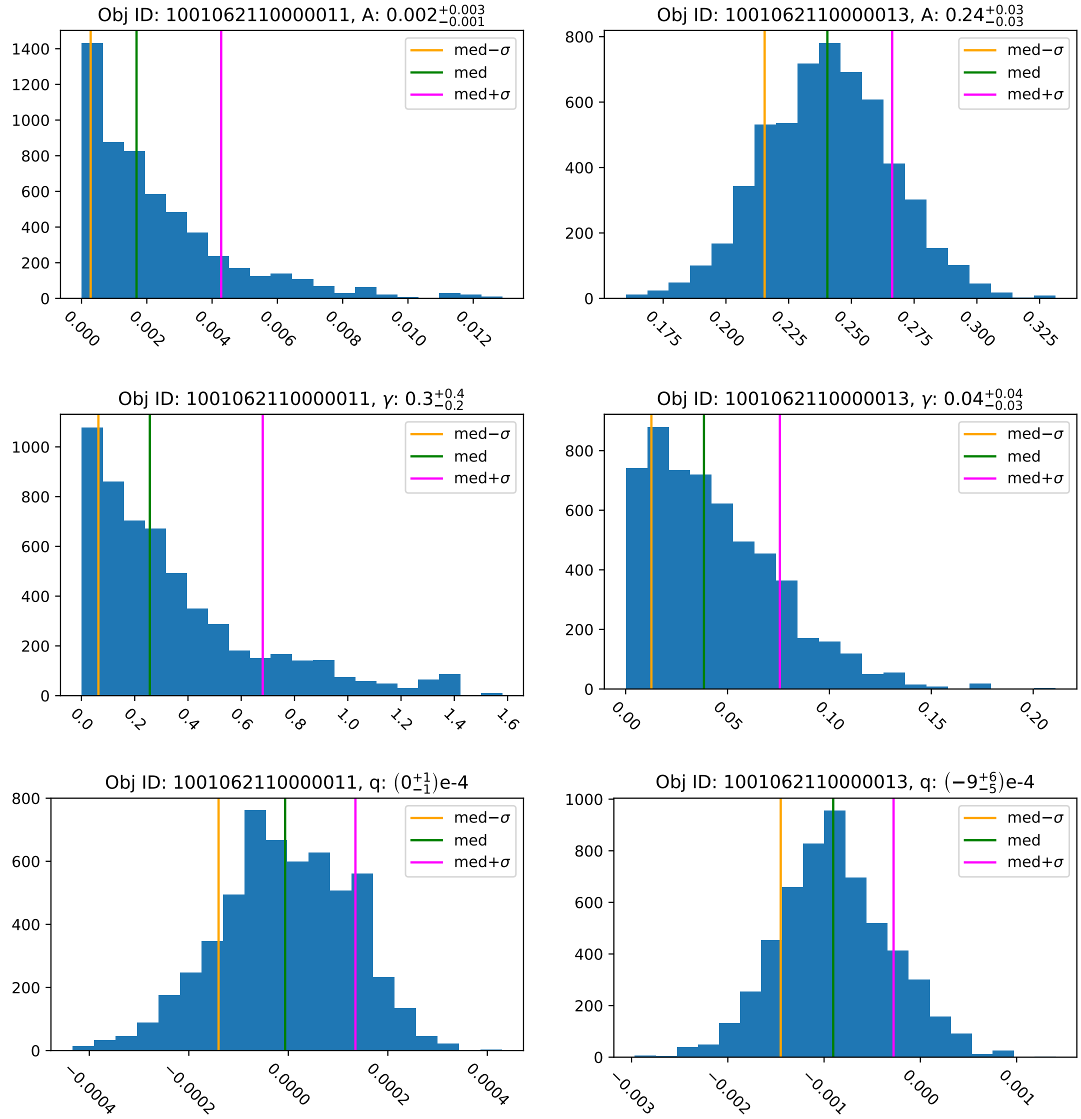

Figure 3.9: Histograms of MCMC samples after burn in for two objects. Left plots: A random object (left plots). Right plots: An object with A significantly different from 0.

.

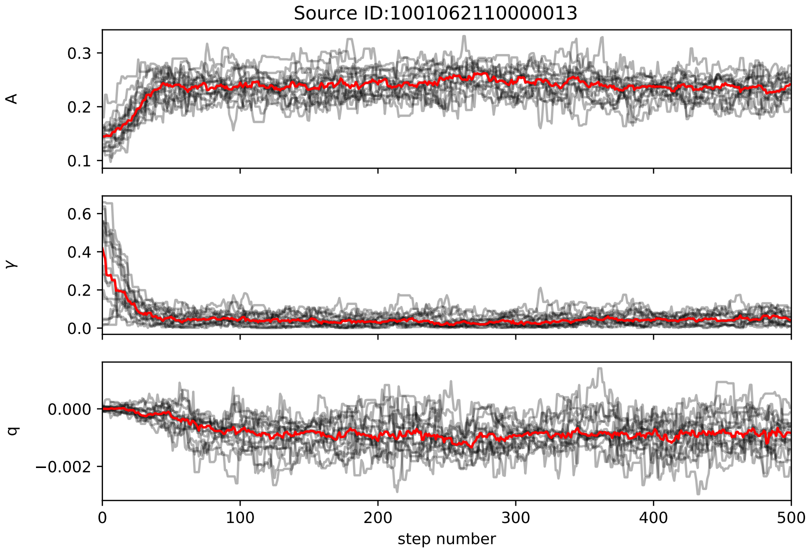

Figure 3.10: Chain plots of how the walkers explore the fit parameter space for a specific object.

.

For comparison of SF (from the MCMC fit) and the empirical structure function SF_{emp} in the following subsections, the SF_{emp} points are fitted with $SF_{emp,fit}$ as linear fits in log-logspace. These are of the form $a\log(\frac{\Delta t}{t_0}) + b$. The structure function parameters of SF are converted to a and b in the legends to have it in the same format as $SF_{emp,fit}$, and $\frac{\Delta t}{t_0}$ is abbreviated as dt in the plots. SF is converted as

$$\log\left(A^2\left(\frac{\Delta t}{t_0}\right)^{2\gamma}\right) = 2\gamma\log\left(\frac{\Delta t}{t_0}\right) + 2\log(A), \tag{3.1}$$

meaning $a = 2\gamma$ and $b = 2\log(A)$.

3.2.1 Box search variability

The objects are divided into classes as described in Section 2.4. Some examples of their lightcurves and structure functions are included in Fig. 3.11-3.16.

SF and $SF_{emp,fit}$ often turn out different, especially for the standard star candidates, where there ideally is no excess variation. But there are also big differences for some of the random sources where A and/or γ is close to 0. SF does not always match the plotted SF_{emp} well either, as we will get back to in the discussion.

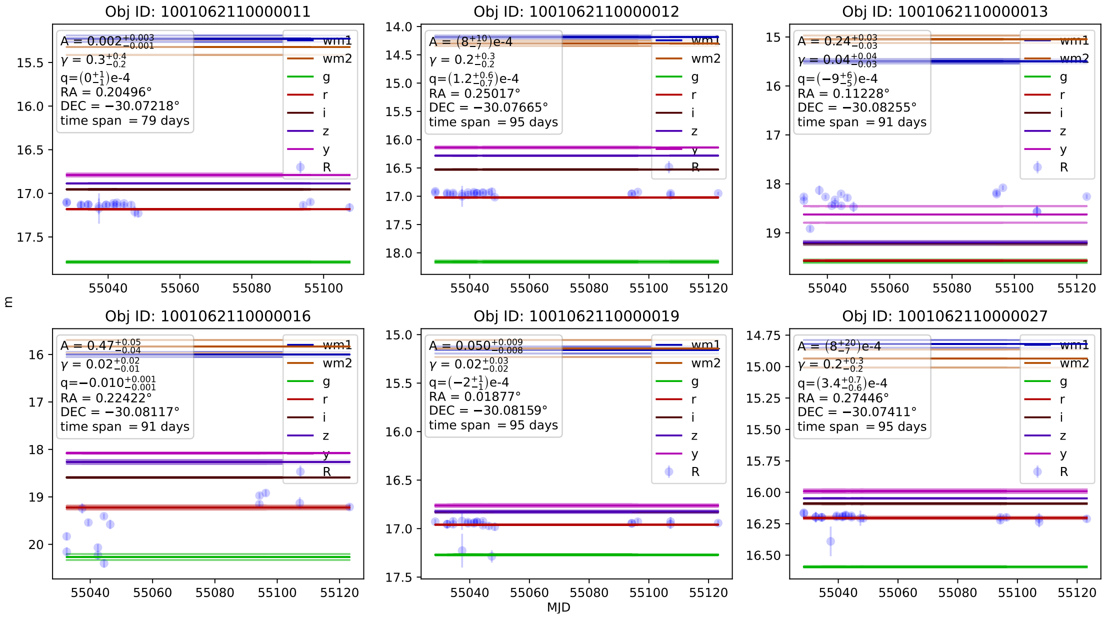

Figure 3.11: Some lightcurves from the box searches.

.

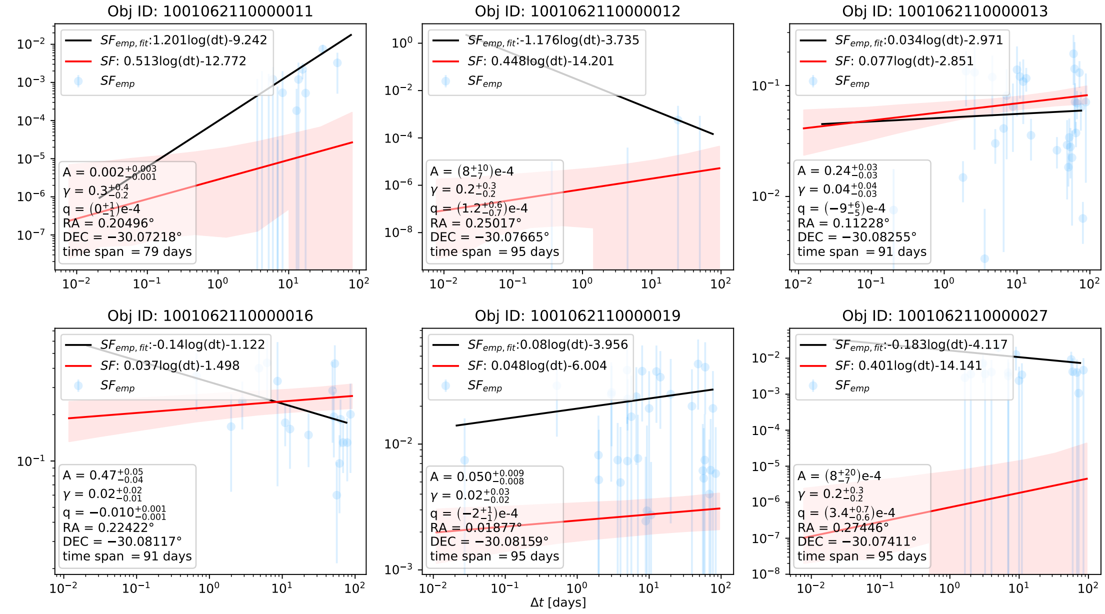

Figure 3.12: Some structure functions from the box searches. In the legend, SF fits are converted for comparison with $SF_{emp,fit}$.

.

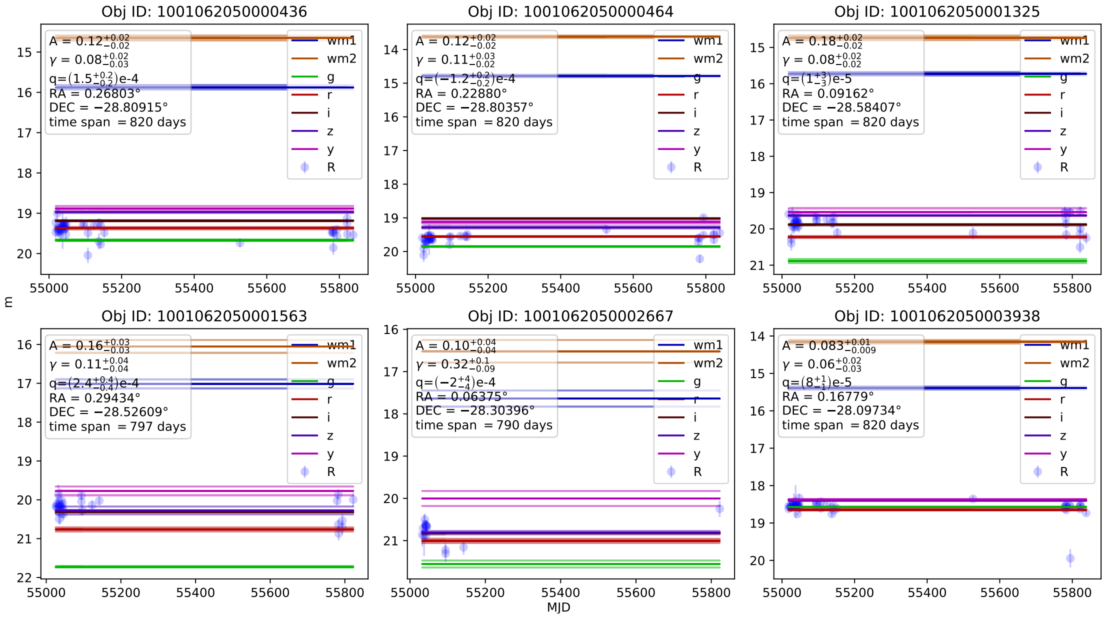

Figure 3.13: Lightcurves from the box searches consistent with quasars (high $W1 - W2$) with A and q at least two standard deviations from 0.

.

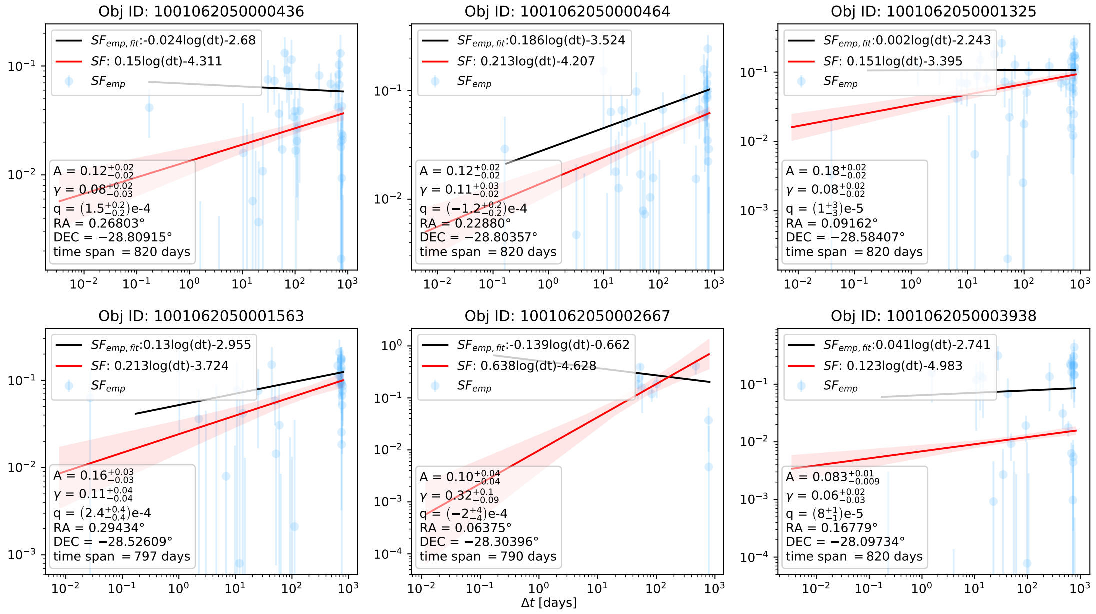

Figure 3.14: Structure functions from the box searches consistent with quasars (high $W1 - W2$) and A and q at least two standard deviations from 0. In the legend, SF fits are converted for comparison with $SF_{emp,fit}$.

.

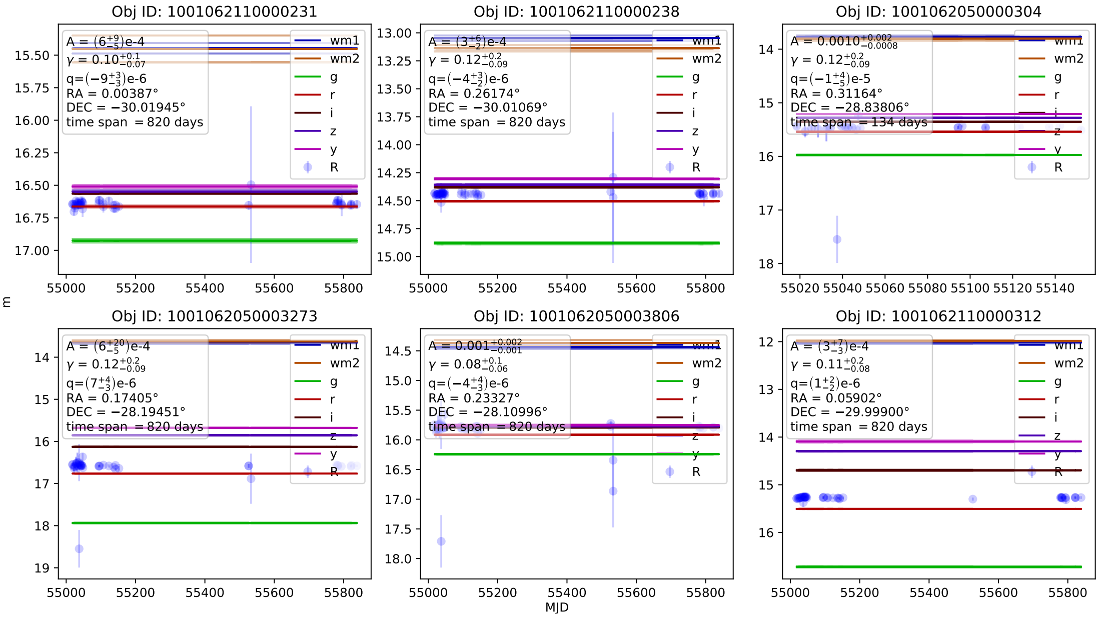

Figure 3.15: Lightcurves from the box searches consistent with being standard stars.

.

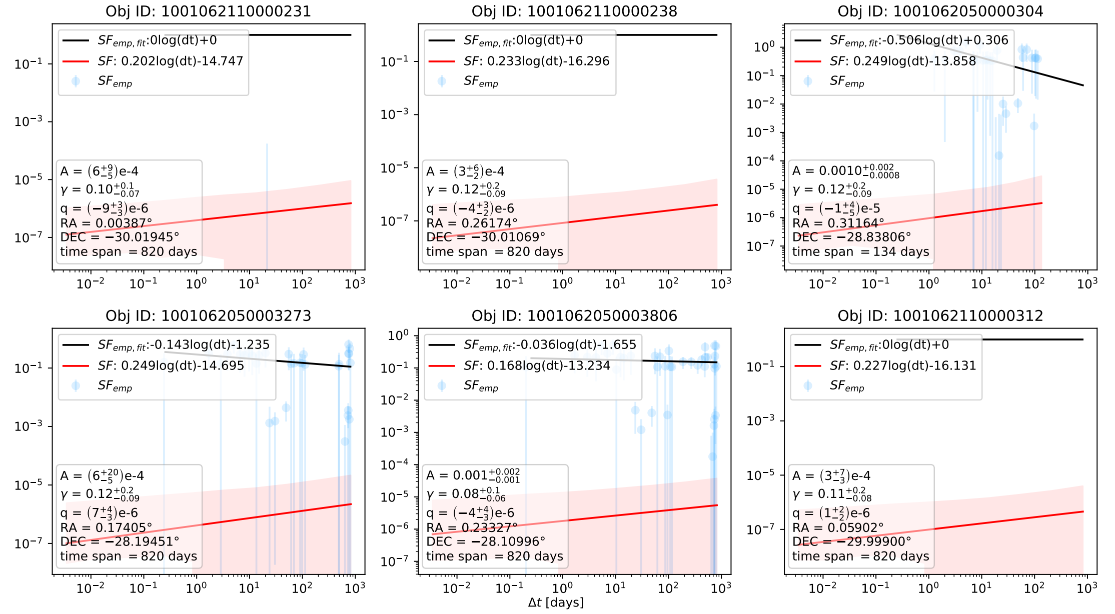

Figure 3.16: Structure functions from the box searches consistent with being standard stars. In the legend, SF fits are converted for comparison with $SF_{emp,fit}$. When $SF_{emp,fit}$ shows "0log(dt)+0", it means the fit failed due to $SF_{emp} < 0$ for all Δt_{ij}.

.

Fig. 3.17 shows A vs. γ for different selections of objects – including a version of the plot similar to plots made by Schmidt et al. [20].

2 branches of objects are observed in the data at low A, as framed by the blue and magenta lines by eye and described in Section 2.4. 9 % of the objects have low σ_m, meaning most do not have any excess variation - comparing the median error and median variation. So the lightcurves vary less than you would expect from the errors.

4 % are "well defined", meaning A and γ are significantly different from 0. Out of 693 objects with well defined fit parameters, 207 fulfilled the Schmidt quasar criteria (30 %). And only 14 of the well defined points had quasar colours in WISE (2%, or 0.09 % of the total sample). 89 sources were consistent with being standard stars.

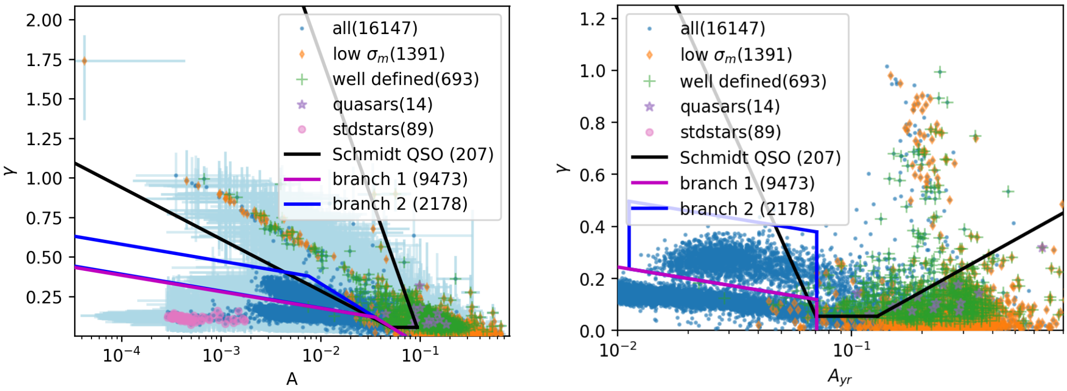

Figure 3.17: A vs. γ for different selections of objects (defined in Section 2.4) in the box searches. The plot to the left used $t_0 = 1$ day, and the one to the right has been converted to $t_0 = 1$ yr, had the errors removed and is zoomed in to look like results from Schmidt et al. [20]

.

The fit parameter distributions for all classes are found in Fig. 3.18. For each histogram bin, bars for all classes are plotted - so the histogram bars are thinner than the bins they represent. For A, in general the sources are split between high and low values with fewer in between. For high A, they have low errors and A is not consistent with 0 – and the opposite for low A. Or they are standard star candidates with even lower values than required. It makes sense that low magnitude errors lead to higher A, since if there is little excess variation to fit, A should be low or consistent with 0. They have low γ on average, so for most sources with excess variation there is little change in variability magnitude depending on the time scale. The same holds for "well defined" sources – they are generally found at low γ despite requiring that γ is 2 $\sigma_{\gamma,-}$ from 0. Lower γ than the average result. If there is no excess variation to fit in the average lightcurve, any γ is consistent with the data if A is set low enough. The colour selected quasar candidates are few and consistent with the general distribution of well defined objects.

In q, branch 2 stands out as having a wider distribution. But note that their lightcurves are shorter according to Fig. 3.19. Both branches have high measurement errors and few points per object – unlike the standard stars that are also found at low A. The standard stars have enough points to ensure that the structure functions are only consistent with very low variability. The well defined objects also mostly contain a large number of points and the colour selected quasar candidates follow the same distribution – but out of the well defined objects, the Schmidt QSO are mostly found among the few objects with few points.

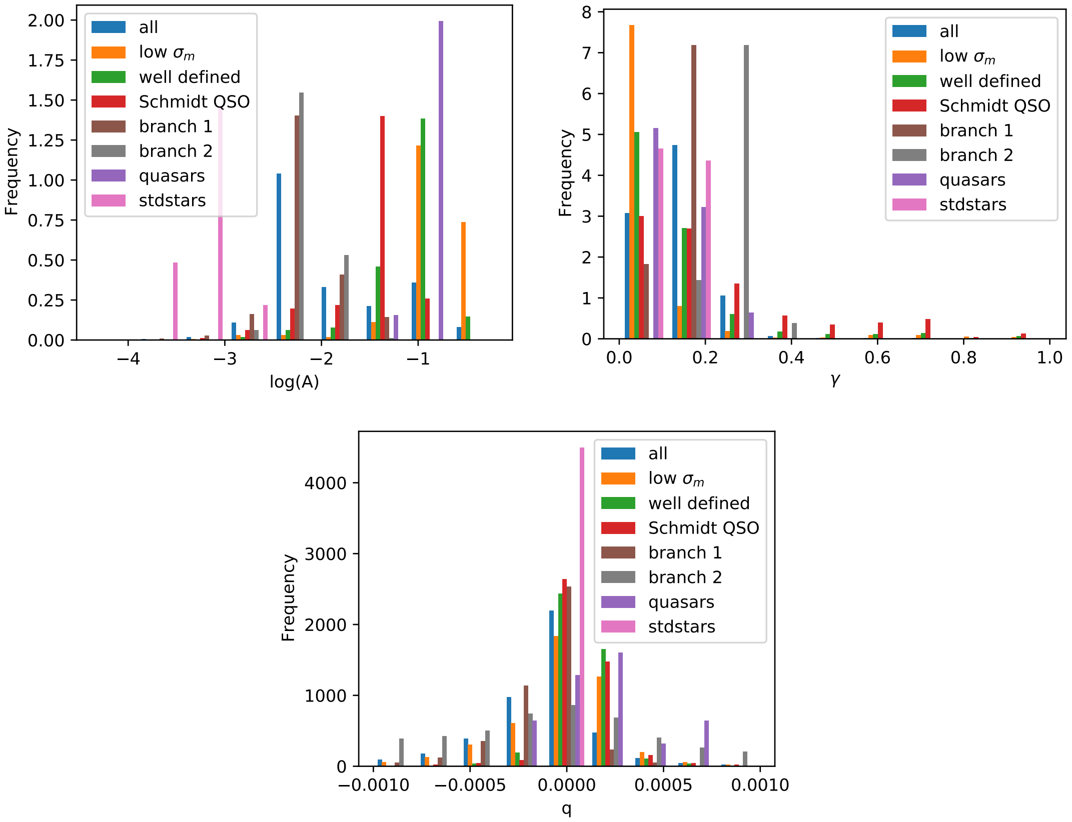

Figure 3.18: Histograms of the fit parameters for different selections of objects in the box searches. Left: A; right: γ; and centre: q. The histograms are zoomed in for γ and q.

.

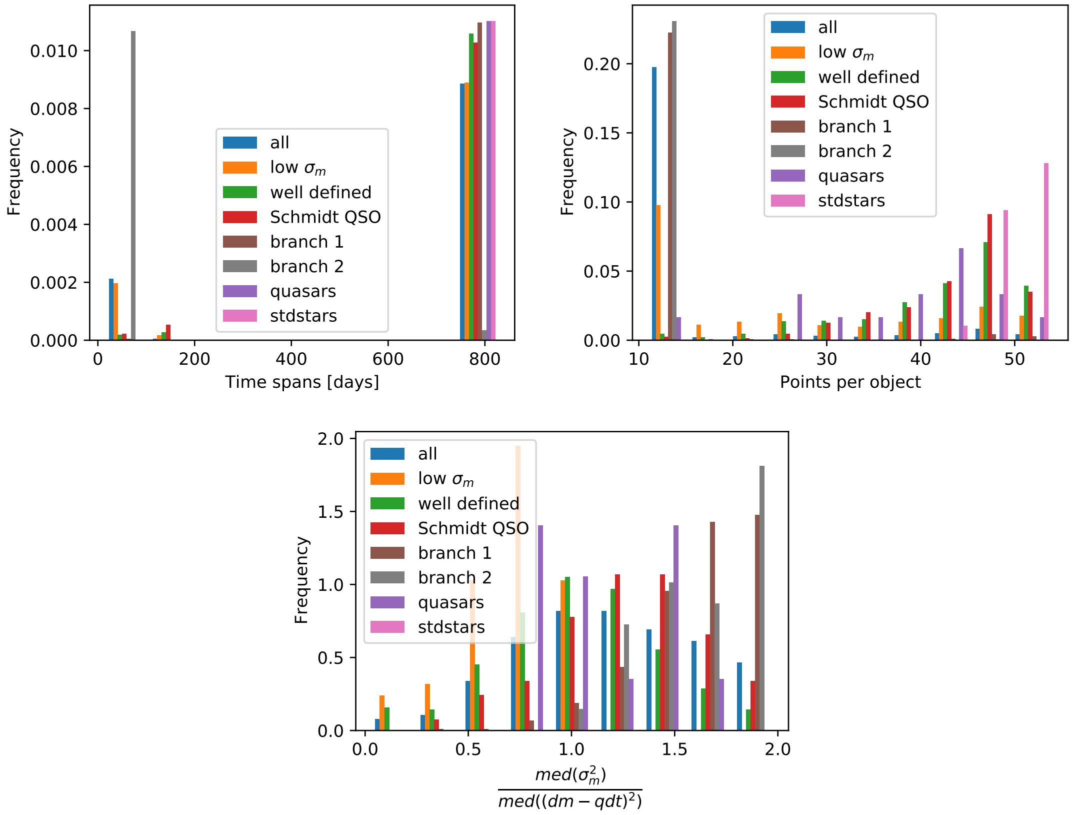

Figure 3.19: Histograms of some light curve parameters for the box searches: Left: Time spans; right: Number of points; and centre: The ratio of measurement uncertainties to observed variation. The last plot is zoomed in, since for some objects the measurement variation were of order 10^3 times larger than the observed variation.

.

The full cornerplot for all sources in Fig. 3.20 shows some of the tendencies from the previous plots, but also that there is more variation in q values at low γ. This is also where most of the data is found. But the average stays at 0. The most extreme outliers are found at high A and low γ. So if the sources vary a lot, especially on short time scales, q will sometimes detect a high linear magnitude change. q is closer to 0 for well defined sources, so large variations in a lightcurve with a structure function that is difficult to fit could lead to some of the variation being explained by a high q in addition to high, but uncertain A. Especially if the lightcurve is short, so high and low magnitudes randomly could end up giving a linear trend. In Fig. 3.22 the large $|q|$ are mostly found at short time spans, but also at some of the longest - even for some well defined sources. $|q|$=0.03 mag/day means a yearly change of 11 magnitudes which is not realistic and an unexpected find for long lightcurves that should have time to determine q well. Three sources had time spans >600 days, $|q| > 0.01$ mag/day and $q > 3$ standard deviations from 0. These are plotted in Fig. 3.25. We see that the lightcurves vary a lot at the end and could be declining very steeply – but not in a linear way over the entire lightcurve and so it cannot be explained by a single parameter q.

Comparing the corner plot of all sources in Fig. 3.20 with that of the "well defined" sources in Fig. 3.21, we see that the objects with low A were mostly sorted out, so the peak in $\log_{10} A$ changed from ∼2.4 to ∼1.0. γ decreased a bit, but the change is well within the errors. Some of the most extreme outliers in q are not included but some very high and low values are still found. The two branches also disappear and of course the standard star candidates as their definition is incompatible with being "well defined".

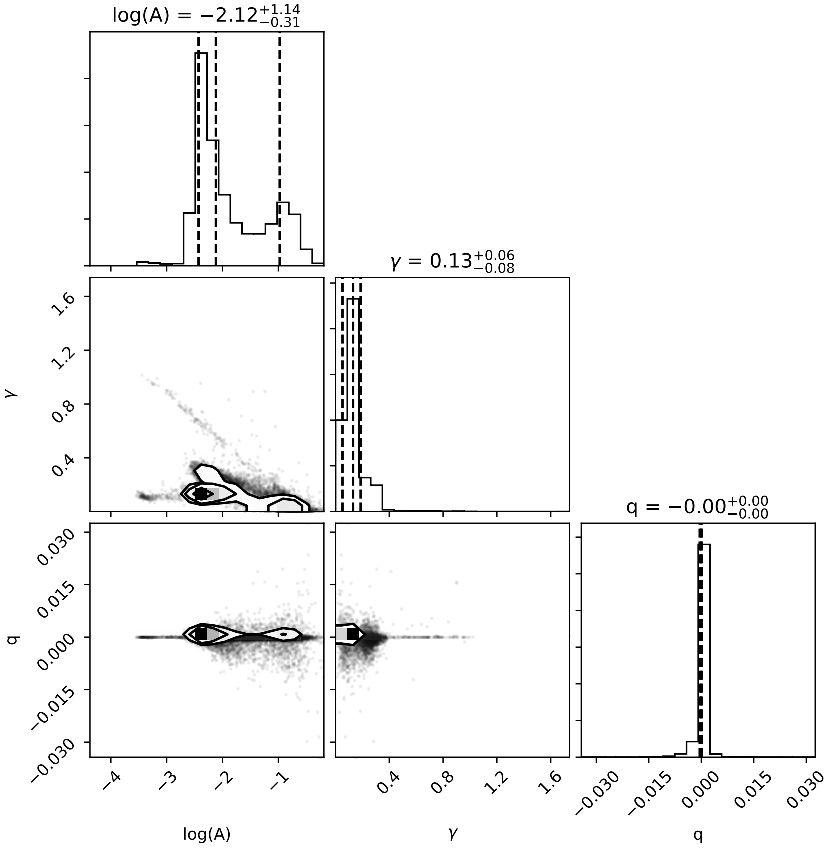

Figure 3.20: Corner plot of fit parameters for the box searches.

.

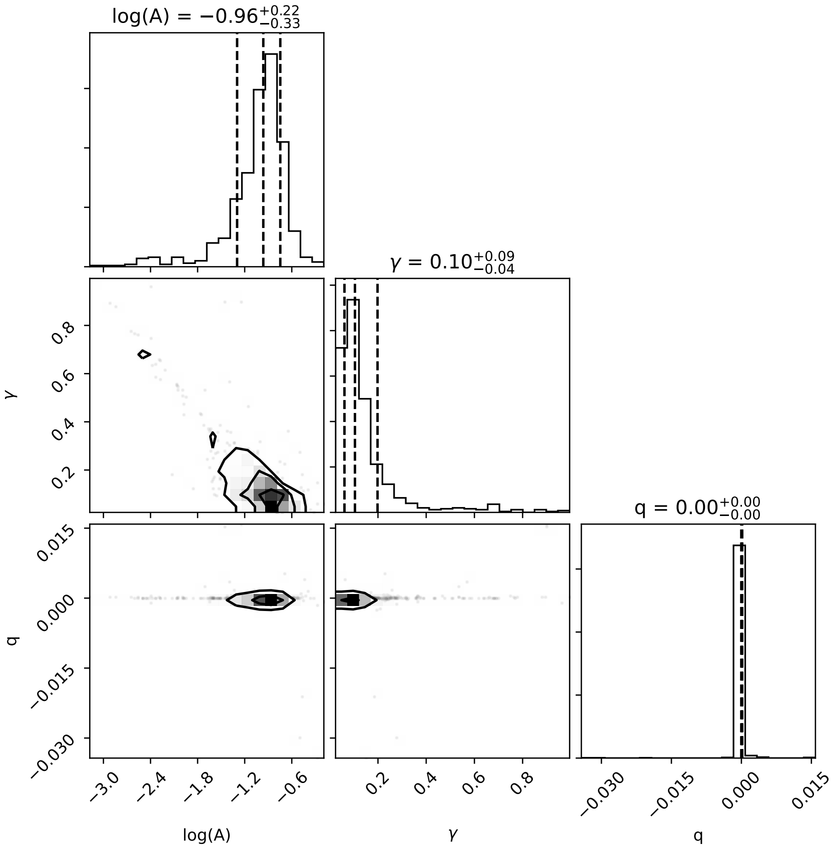

Figure 3.21: Corner plot of fit parameters for the "well defined" objects in the box searches.

.

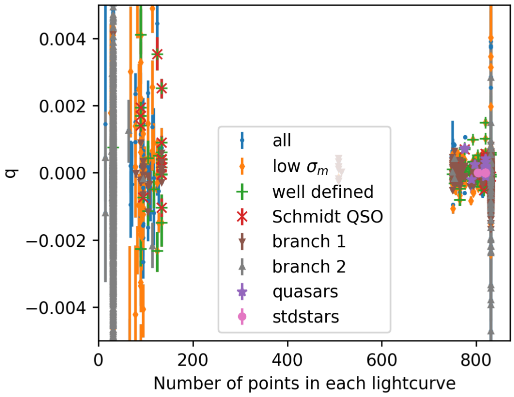

Figure 3.22: q at different time spans in the box searches.

.

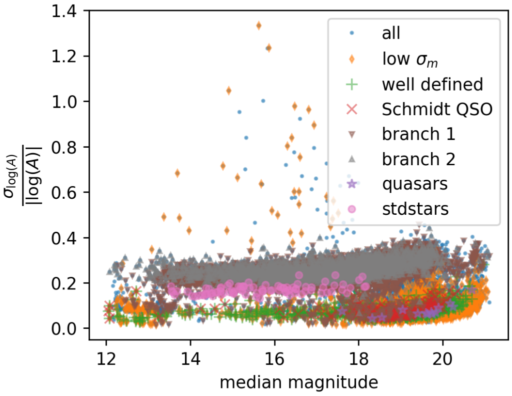

Figure 3.23: Relative uncertainty in $\log(A)$ vs. median PTF magnitude for data from the box searches. The $\sigma_{log(A)}$ are computed using averages of the upper and lower errors on A.

.

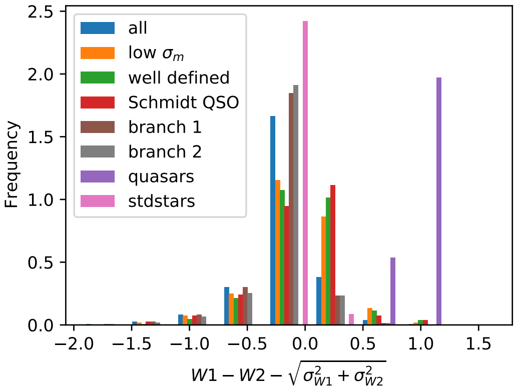

Figure 3.24: Histogram of W1-W2 for the box searches.

.

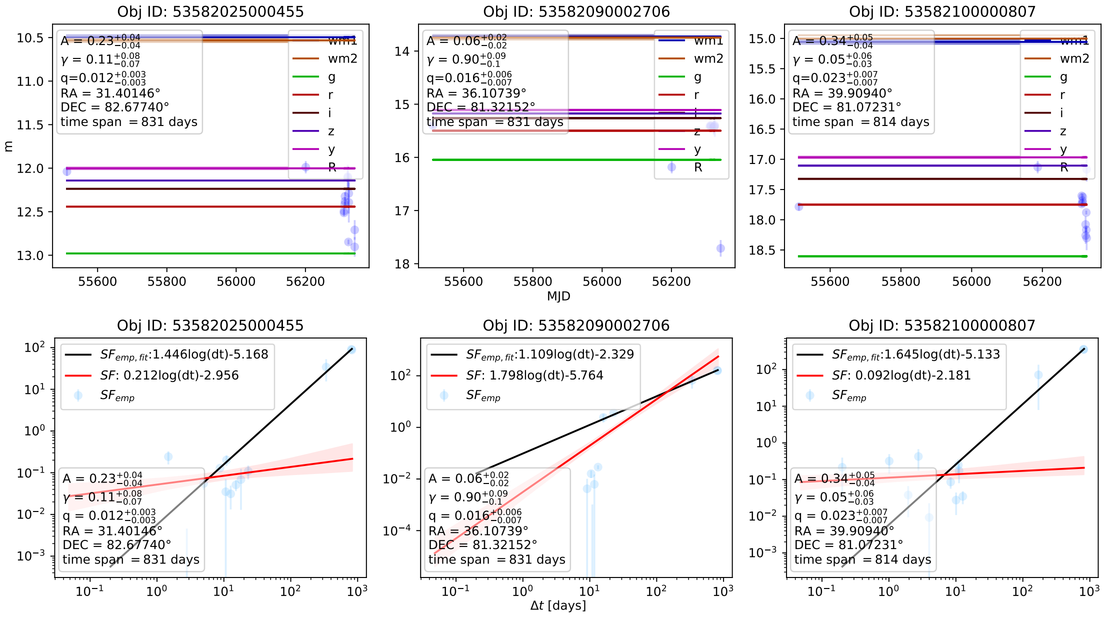

Figure 3.25: Lightcurves and structure functions for sources with high q in the box searches.
.

Fig. 3.23 illustrates how the relative uncertainty in SF amplitude A changes with brightness. It does not change much, but the spread decreases and the values are a bit lower for bright sources. Sources with well defined A and γ or low σ_m generally have lower relative errors than the rest (as required for the well defined ones).

The colour selection of stars and quasars are made using $W1-W2-\sqrt{\sigma_{W1}^2+\sigma_{W2}^2})$ as plotted in Fig. 3.24 for all classes. Most sources are close to 0, especially the uncertain ones. Schmidt QSO are are bit higher than well defined sources in general, but usually not enough to look like quasars in WISE colours.

The colours are also plotted in Fig. 3.26. Based on the variability selection by Schmidt et al. [20], one would expect more of the objects at high γ to be quasars and therefore have quasar colours in WISE. But this is not observed – the objects with high $W1-W2$ are not found to have higher γ. The objects with high or low $W1-W2$ do not have visibly different A either – independently of A most objects have $W1-W2$ close to 0. Bright sources with low $W2$ magnitude are close to 0 in $W1-W2$ and are therefore possible standard stars, but the selected standard star candidates are spread over a wide range of $W2$. So some faint standard stars have been selected. Fainter sources in $W2$ have more variation in $W1-W2$. Comparing with Ansari et al. [2], the sources mostly occupy the colour space consistent with stars, but also areas with even lower $W1-W2$ and $W2$ than explored by the paper. The colour selected quasars are of course an exception, since they were selected to have $W1-W2$.

The plots of $g-r$ vs. $z-W1$ show that the distribution of objects extend beyond the stellar locus and so cannot only be explained by stars with the behaviour described by Ansari et al. This is clearest at high $z-W1$, where the sources behave more like quasars. But this space is not populated only by quasars selected by colour or variability – though all the colour selected quasar candidates do still behave as we would expect. The standard stars are in the area most like for stars, but so are many of the variability selected quasar candidates.

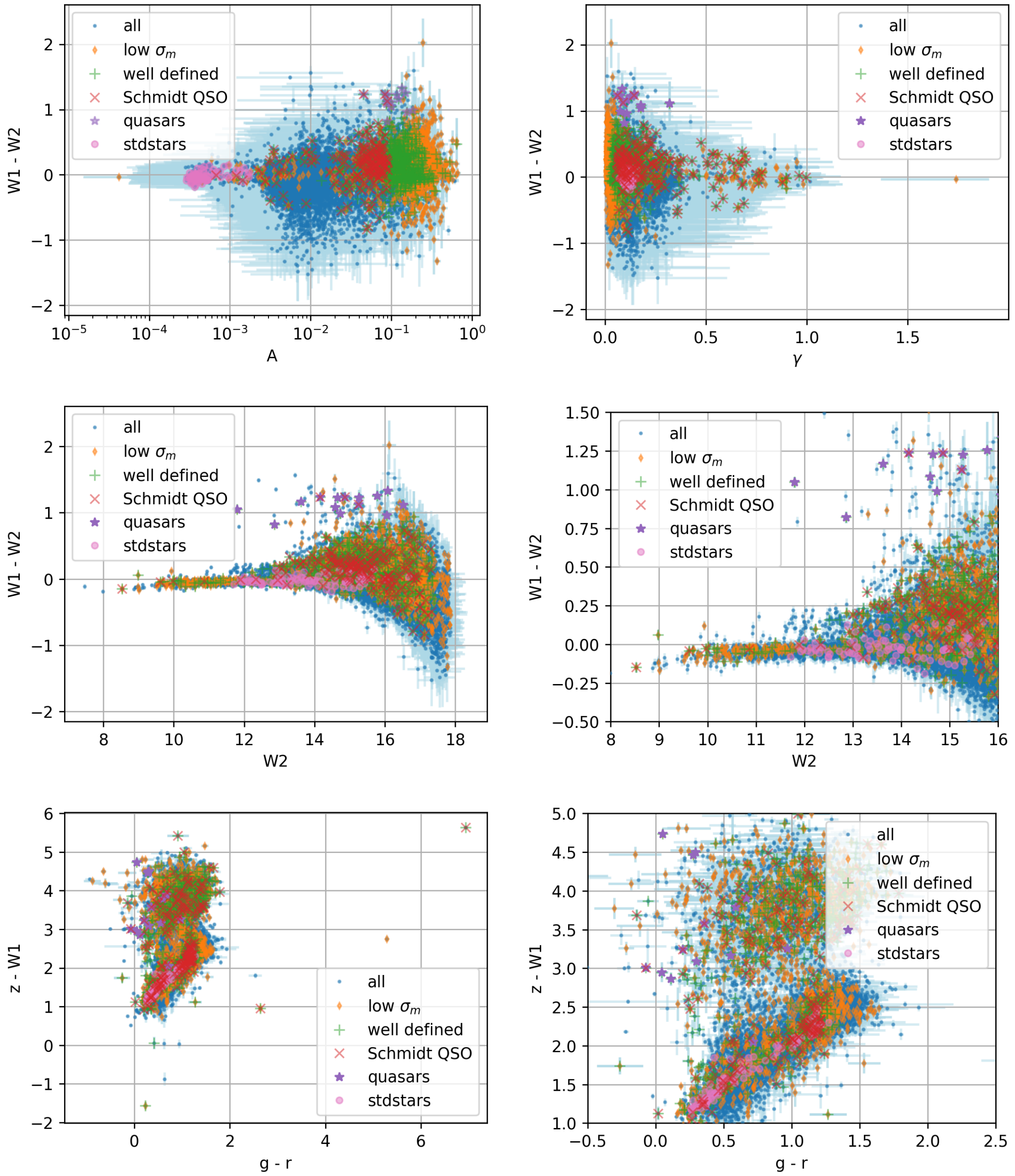

Figure 3.26: Colour plots of the sources in the box searches. Upper left: A vs. $W1-W2$; upper right: γ vs. $W1-W2$; centre left and zoomed in to the centre right: $W2$ magnitude vs. $W1-W2$;, and finally, lower left and a zoomed in version to the lower right: The full colour-colour plot of $g-r$ vs. $z-W1$. The zoomed in versions of the last two plots are for easier comparison with Ansari et al [2].

.

3.2.2 WISE quasar candidate search variability

The WISE colours of the quasar candidates are already selected for, so for them the classes "quasar" and "standard stars" are removed. The 2 branches are also not separated in these sources, so they are also removed as classes. A single point had to be removed from all plots for clarity since it had $A = \left(0.0004^{+1.6}_{-0.00043}\right) 10^{12}$. It had a time span of just 0.231 days and the lightcurve is shown in Fig. 3.27. There are no valid points in SF_{emp} since all σ_m^2 are larger than the variations. Furthermore, 5 objects were matched to the wrong WISE object, and were also removed.

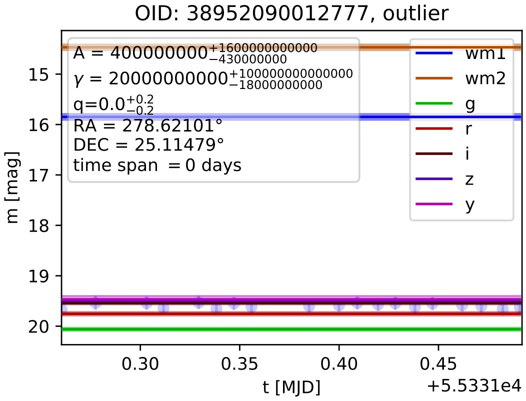

Figure 3.27: Lightcurve for an extreme outlier from the WISE quasar candidate search.
.

Some random example lightcurves and structure functions for the WISE quasar candidate search are shown in Fig. 3.28 and Fig. 3.29. Again, the degree to which SF matches SF_{emp} varies a lot. The sources with well defined A and γ in Fig. 3.30 and 3.31 still do not always agree well with $SF_{emp,fit}$.

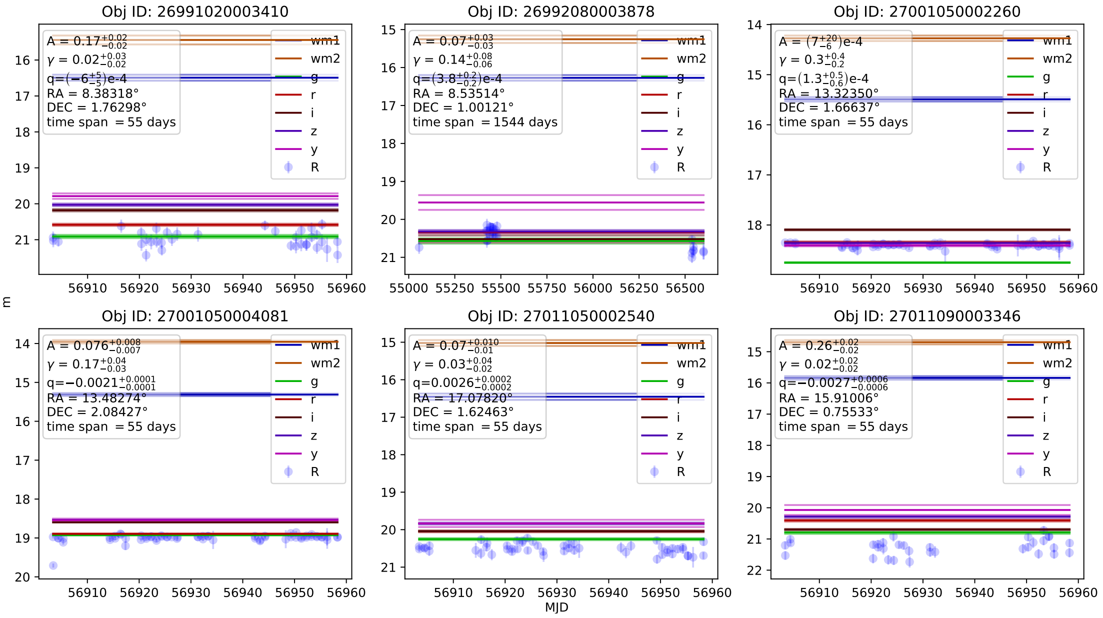

Figure 3.28: Random lightcurves at the WISE quasar coordinates.

.

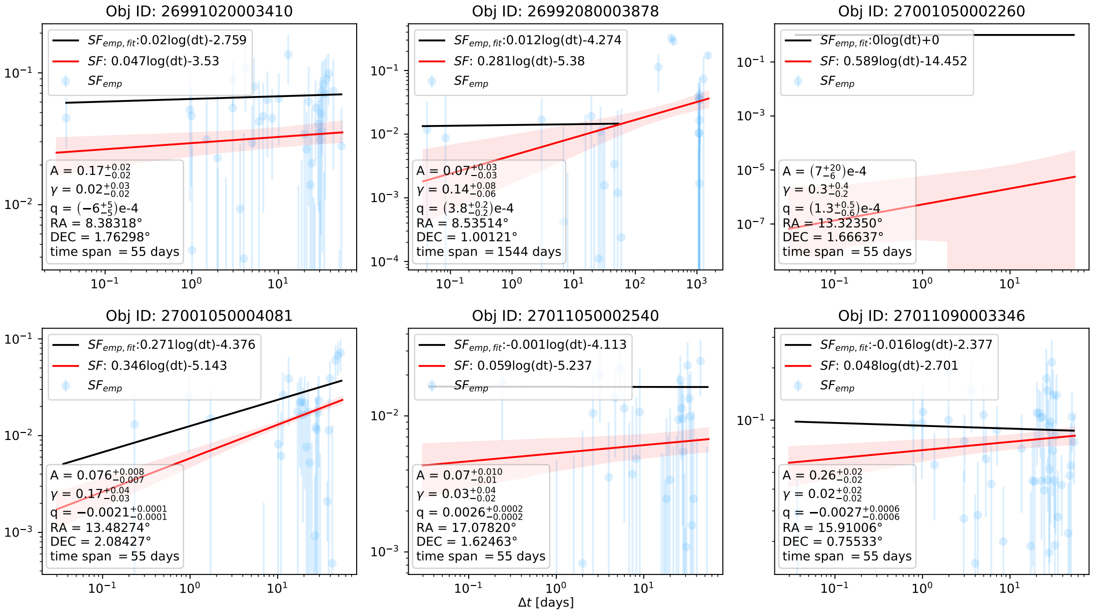

Figure 3.29: Structure functions at random WISE quasar coordinates. In the legend, SF fits are converted for comparison with $SF_{emp,fit}$. When $SF_{emp,fit}$ shows "0log(dt)+0", it means the fit failed due to $SF_{emp} < 0$ for all Δt_{ij}.

.

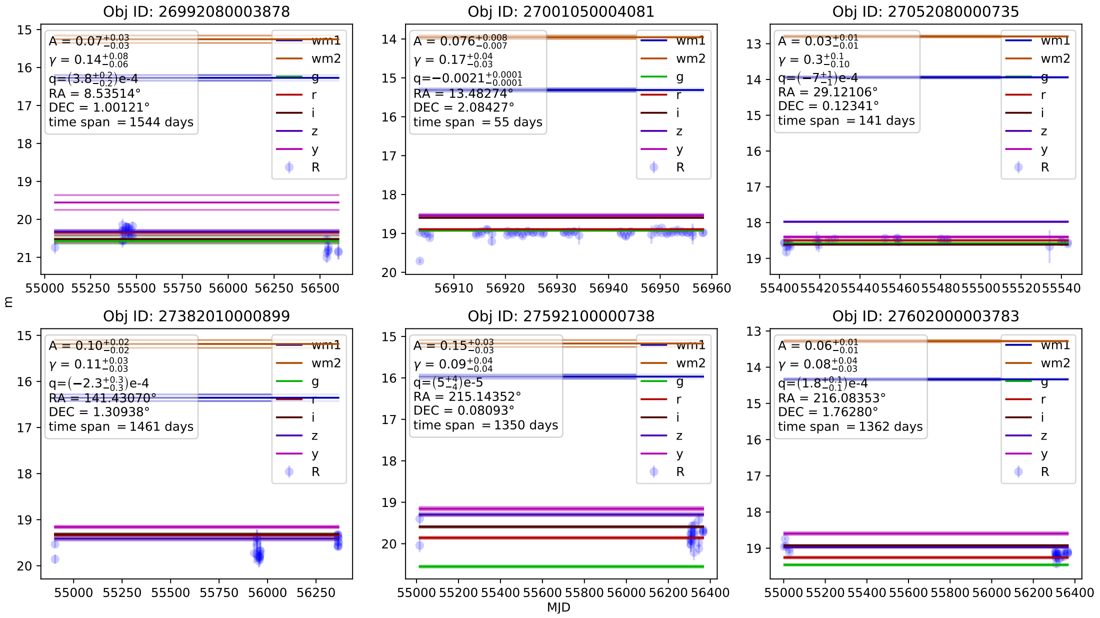

Figure 3.30: Lightcurves at the WISE quasar coordinates with A and q at least two standard deviations from 0.

.

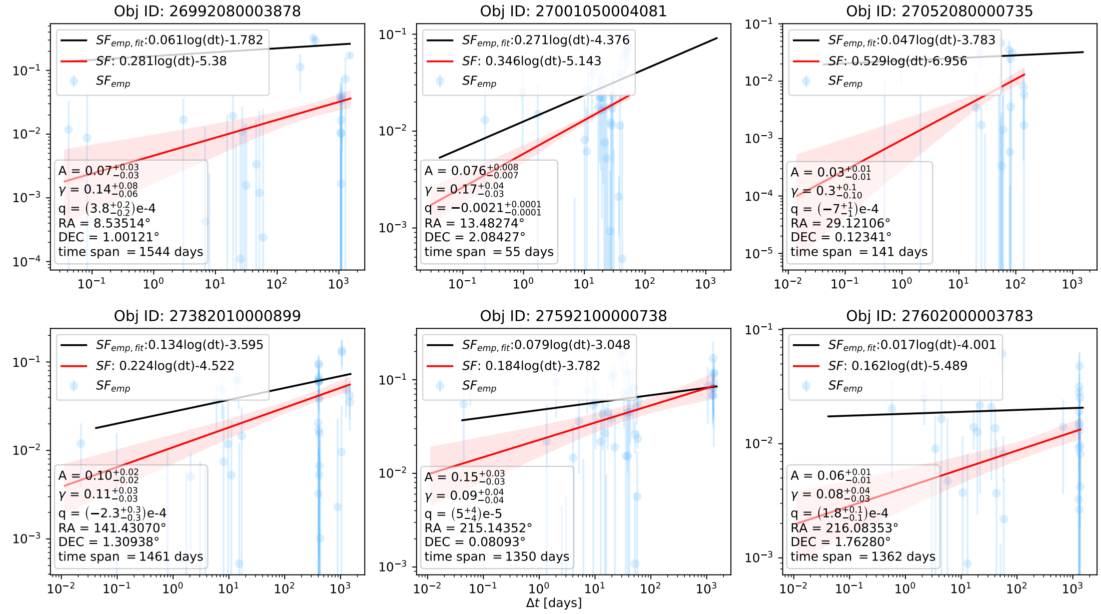

Figure 3.31: Structure functions at WISE quasar coordinates with A and q at least two standard deviations from 0. In the legend, SF fits are converted for comparison with $SF_{emp,fit}$.

.

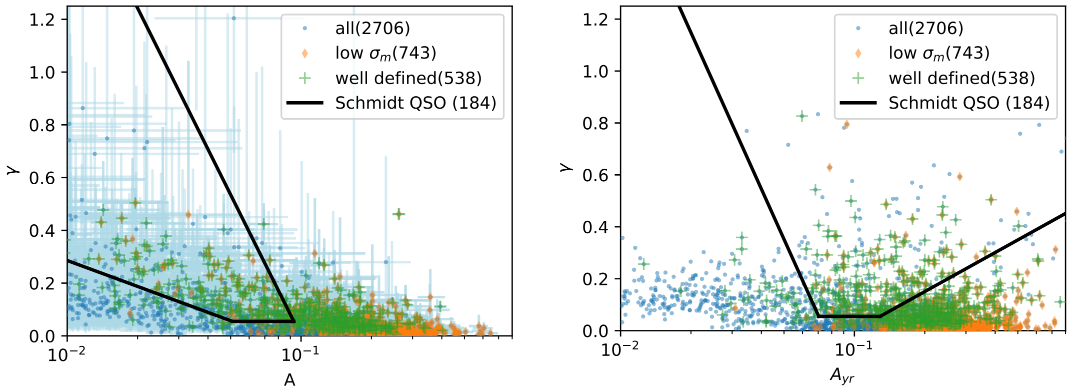

Figure 3.32: A vs. γ for different selections of objects (defined in Section 2.4) at coordinates to WISE quasar candidates. The plot to the left used $t_0 = 1$ day and the one to the right has been converted to $t_0 = 1$ yr and had the errors removed to look like the plots from Schmidt et al. [20]. Both plots are zoomed in, but several outliers exist at $2 < \gamma < 16$.

.

In the $A - \gamma$ plane illustrated in Fig. 3.32, we see some of the same patterns from Fig. 3.17. The "well defined" sources and those with low σ_m are mostly found at high A and low γ, while some other sources seem to form a branch to lower A. The Schmidt QSO region is sparsely populated, but this time the sources do cover a wider range of A values in it. 34 % of the well defined sources are classified as Schmidt QSO, which is a bit more than the 30 % from the box searches.

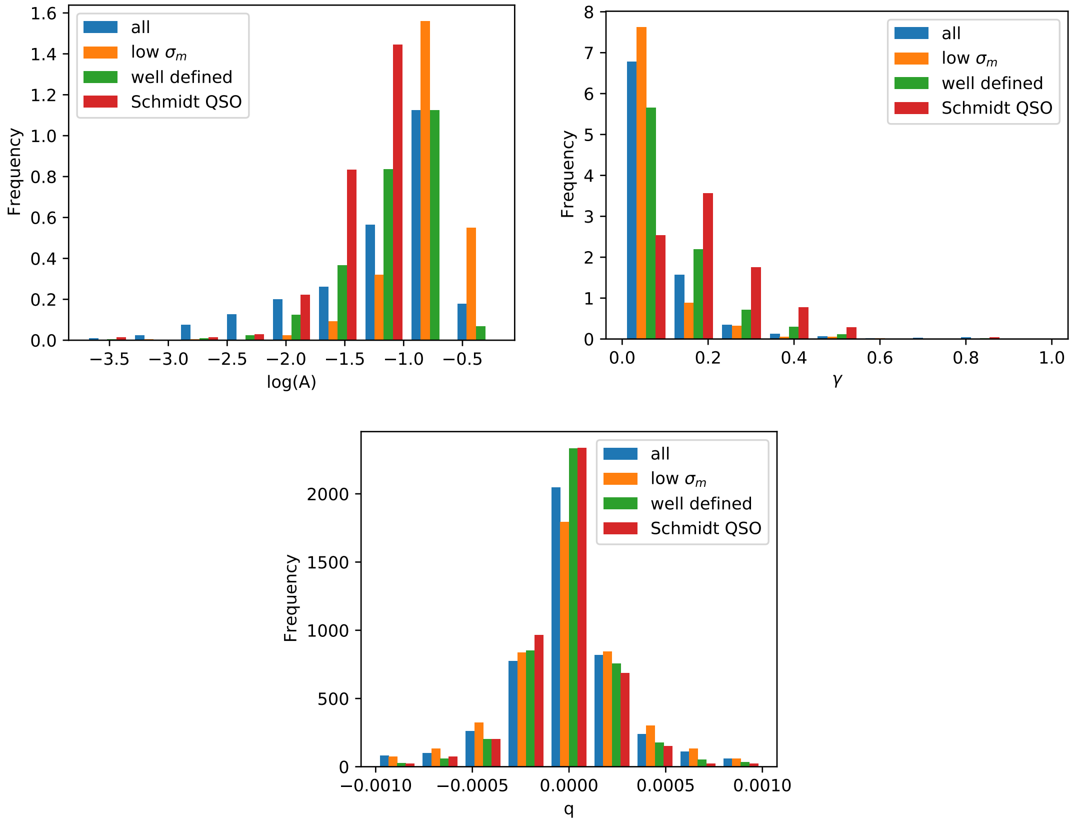

Figure 3.33: Histograms of the fit parameters for different selections of objects at coordinates to WISE quasar candidates. Left: A; right: γ; and centre: q. The histograms are zoomed in for γ and q.

.

As shown in Fig. 3.33, the sources follow the same relative distributions from Fig. 3.18. In the colour plots of Fig. 3.34, the $W1 - W2$ values are of course high, as they were selected to be. There are still no clear trends in A and γ, as the colour changes. Comparing with the results from Ansari et al. [2], the objects do mostly behave as quasars. A few are in regions of $g - r$ vs. $z - W1$ associated with stars or galaxies, but most of these sources are not "well defined" and have high σ_m.

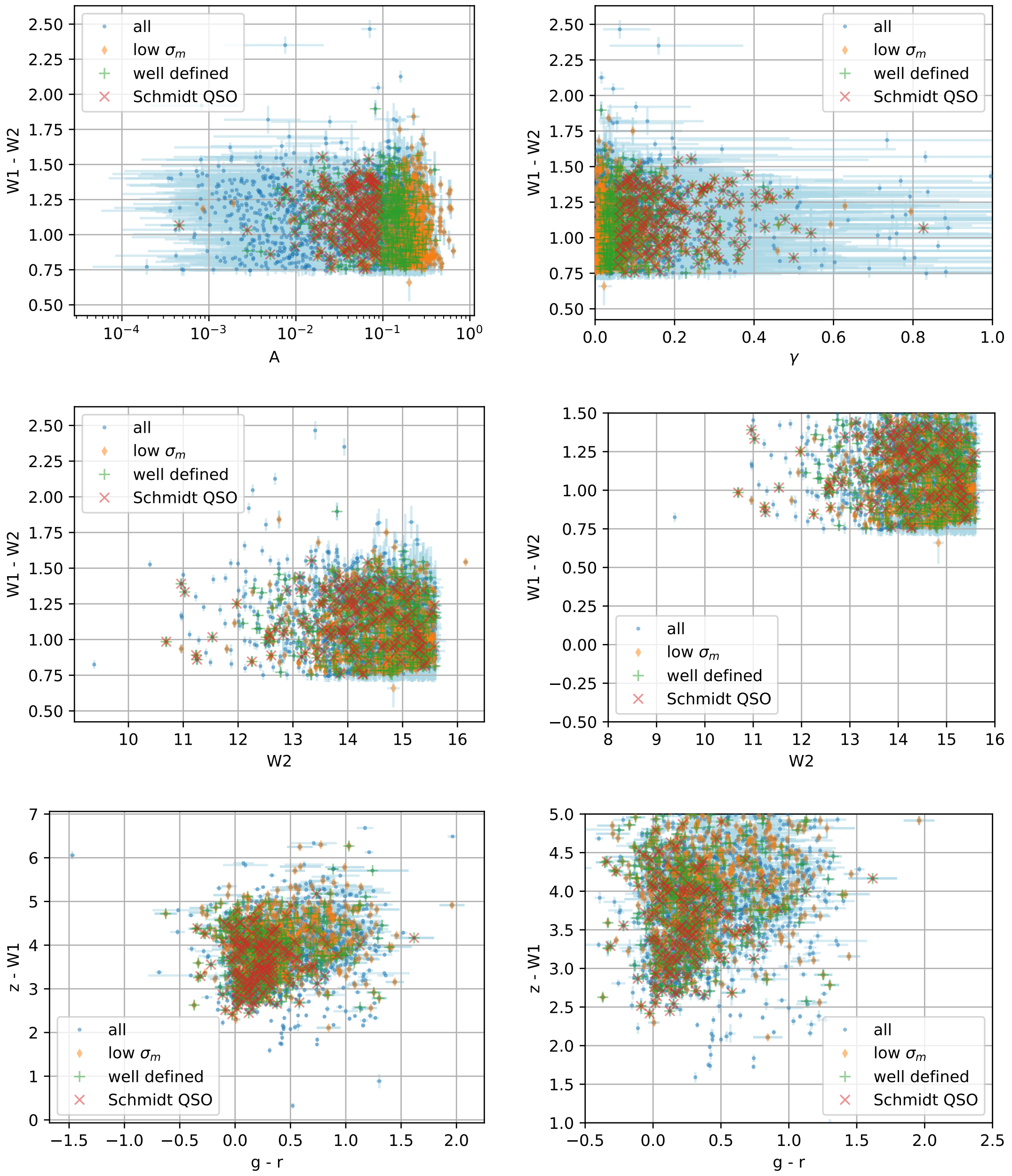

Figure 3.34: Colour plots of the sources at coordinates to WISE quasar candidates. Upper left: A vs. $W1 - W2$; upper right: γ vs. $W1 - W2$ (zoomed in); centre left and zoomed in to the centre right: $W2$ magnitude vs. $W1 - W2$;, and finally, lower left and a zoomed in version to the lower right: The full colour-colour plot of $g - r$ vs. $z - W1$. The zoomed in versions of the last two plots are for easier comparison with Ansari et al [2].

.

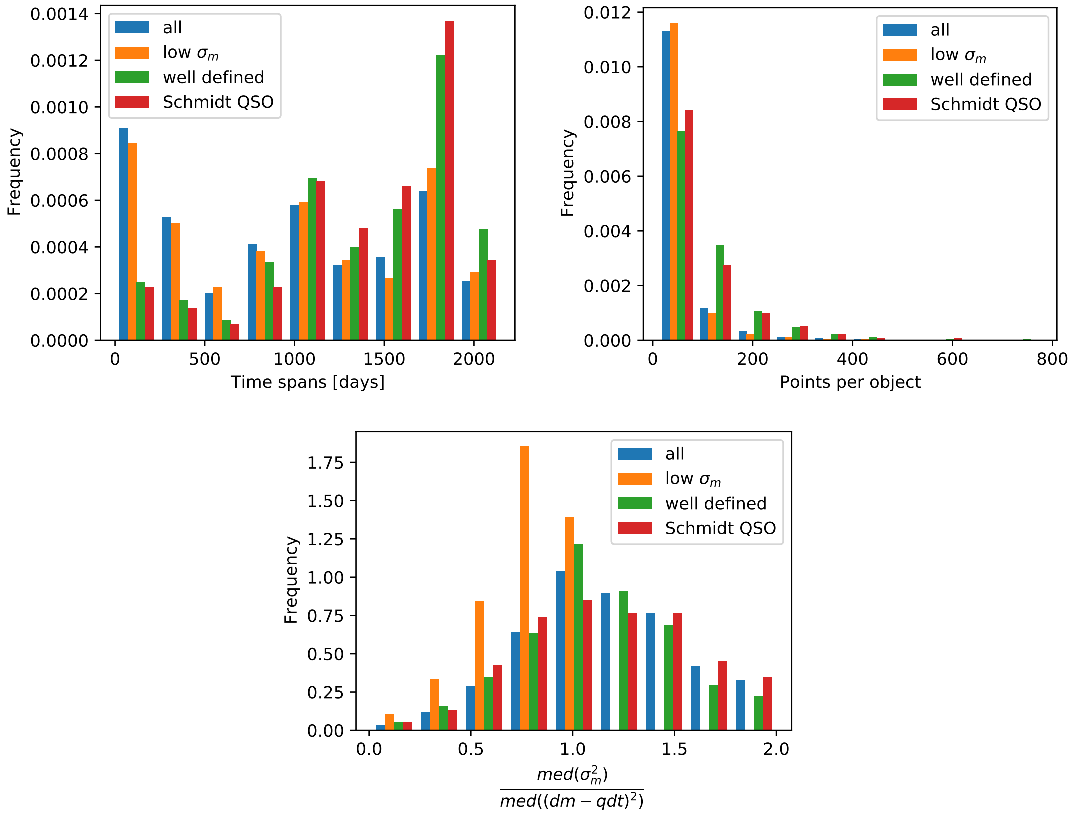

Figure 3.35: Histograms of some light curve parameters for WISE quasar candidates: Time spans (left), number of points (right) and the ratio of measurement uncertainties to observed variation (centre). The last plot is zoomed in, since for some objects the measurement variation were of order 10 times larger than the observed variation.

.

The time spans in Fig. 3.35 are more smoothly distributed than for the box searches. Many of them are also much longer. This indicates that the lightcurves in the few selected boxes were not representative of the entire sky. One could explore this by plotting the average time spans on a sky map. Again, the "well defined" objects are underrepresented at short time spans and overrepresented for the long ones.

The number of points per lightcurve also includes much higher values than for the box searches. The ratios of median σ_m^2 to median variations are similar to those of the box searches. In general, they peak around 1 corresponding to no excess variation.

The corner plot in Fig. 3.36 is for "well defined" sources, as only those are relevant when looking for quasars. q is still sometimes found at very high values – but that is usually for short light curves, as shown in Fig. 3.37. The variations at longer time spans are not for "well defined" objects.

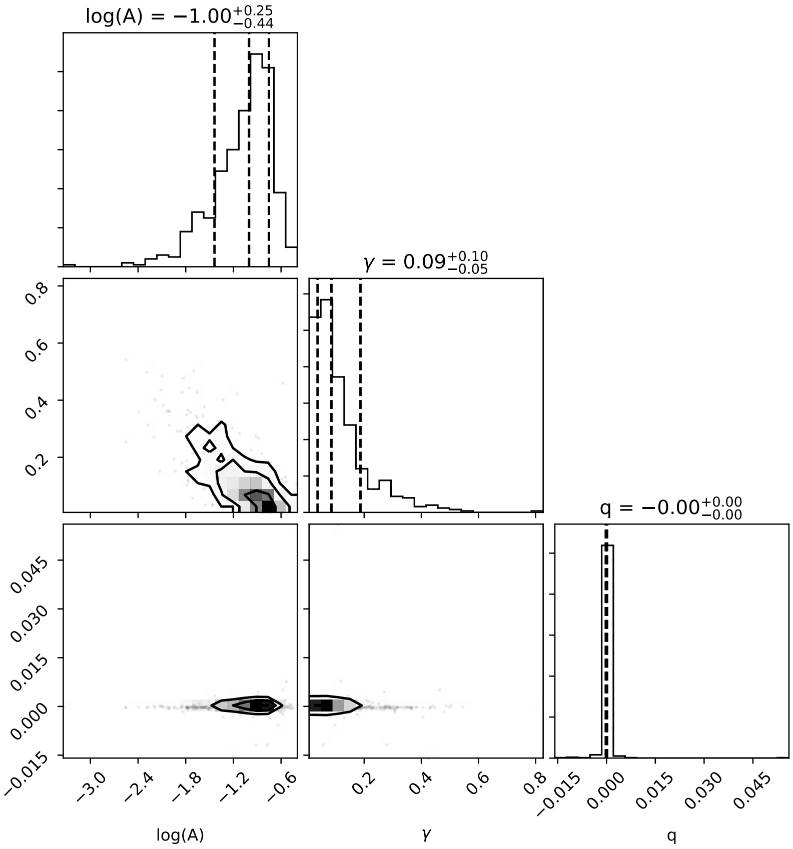

Figure 3.36: Corner plot of fit parameters for the "well defined" objects from the WISE quasar candidate search

.

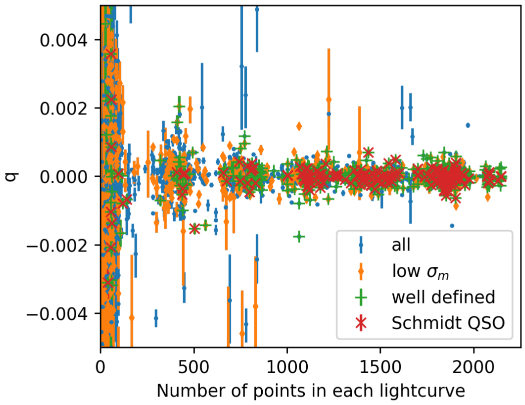

Figure 3.37: q at different time spans for WISE quasar candidates.

.

3.2.3 Lensed quasar search variability

Examples of lensed quasar search lightcurves and structure functions are found in Fig. 3.38-3.41 for random objects and objects with A and γ more than two standard deviations from 0 and quasar colours in WISE. Again we see that SF and $SF_{emp,fit}$ do not always match well.

In the $A-\gamma$ plane of Fig. 3.42, most objects are found at high A and low γ. 46 % of the "well defined" points fall into the Schmidt QSO region and 68 % have quasar colours in WISE, but these numbers are based on low number statistics. Most objects at low A are not "well defined" and have high σ_m, as we also see from Fig. 3.43 . Relatively fewer sources are found at $\gamma \sim 0.2$ compared to the box searches, but the distributions are similar for the "well defined" sources. The distributions are also similar in q.

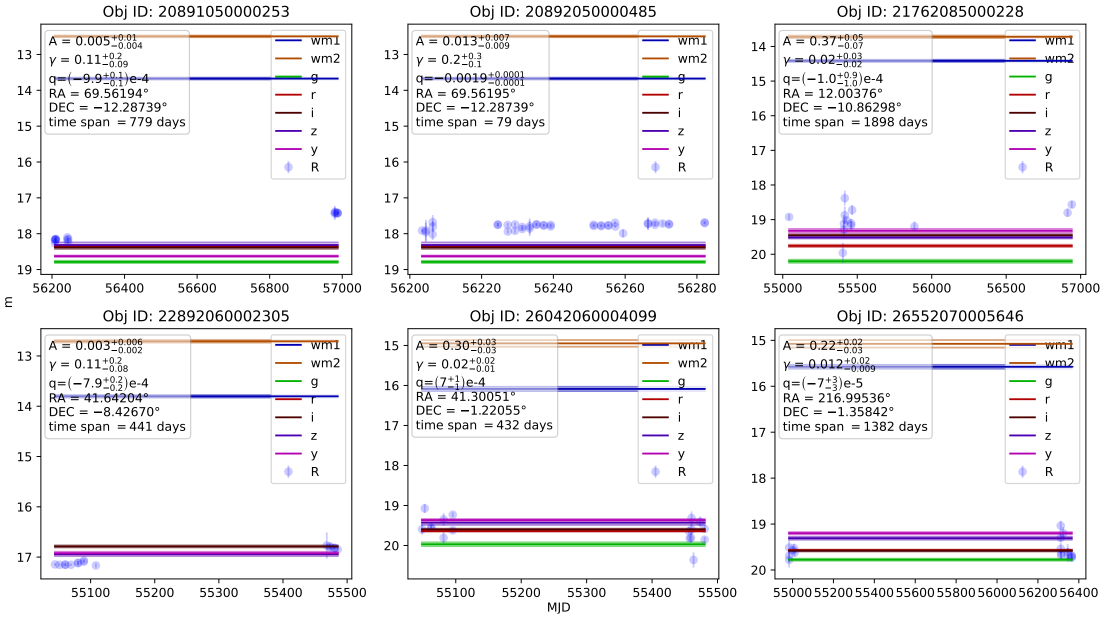

Figure 3.38: Random lightcurves at the lensed quasar coordinates.

.

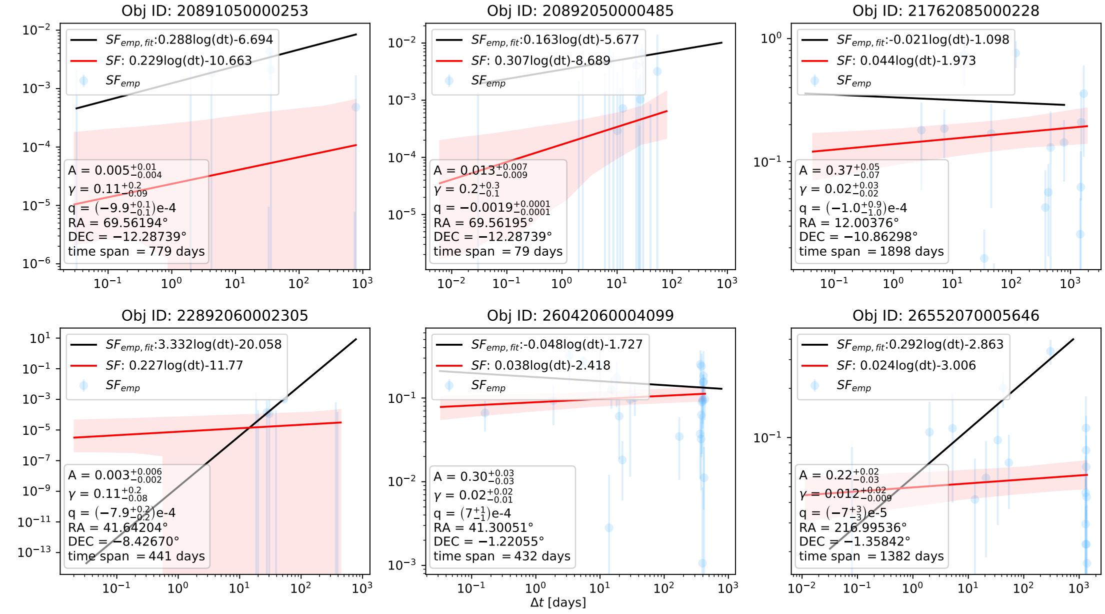

Figure 3.39: Structure functions at random lensed quasar coordinates. In the legend, SF fits are converted for comparison with $SF_{emp,fit}$.

.

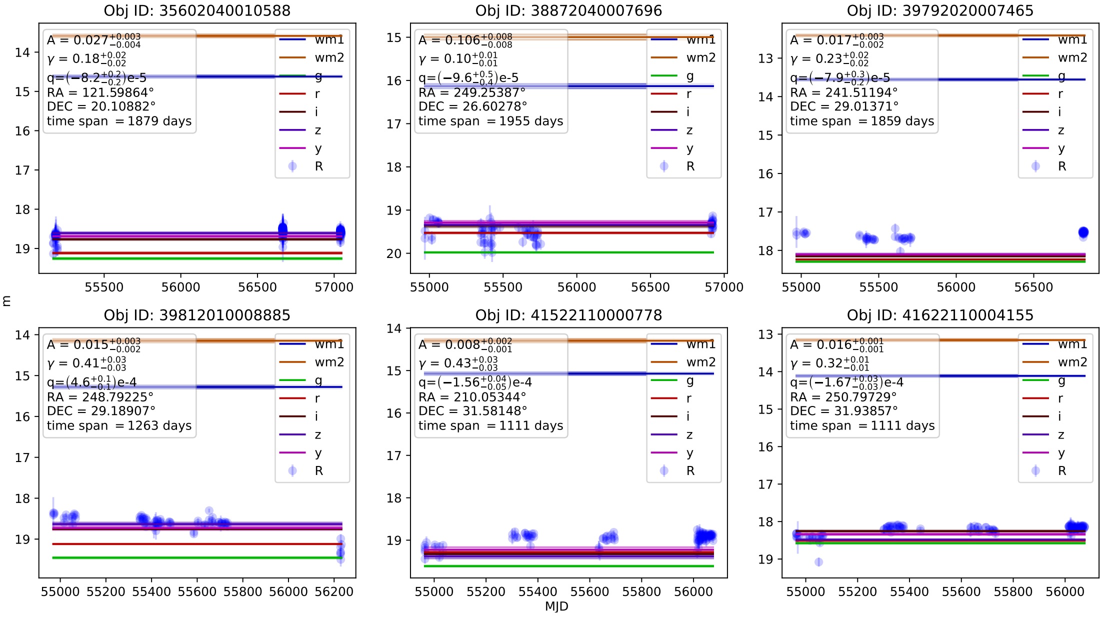

Figure 3.40: Lightcurves at the lensed quasar coordinates with A and q at least two standard deviations from 0.

.

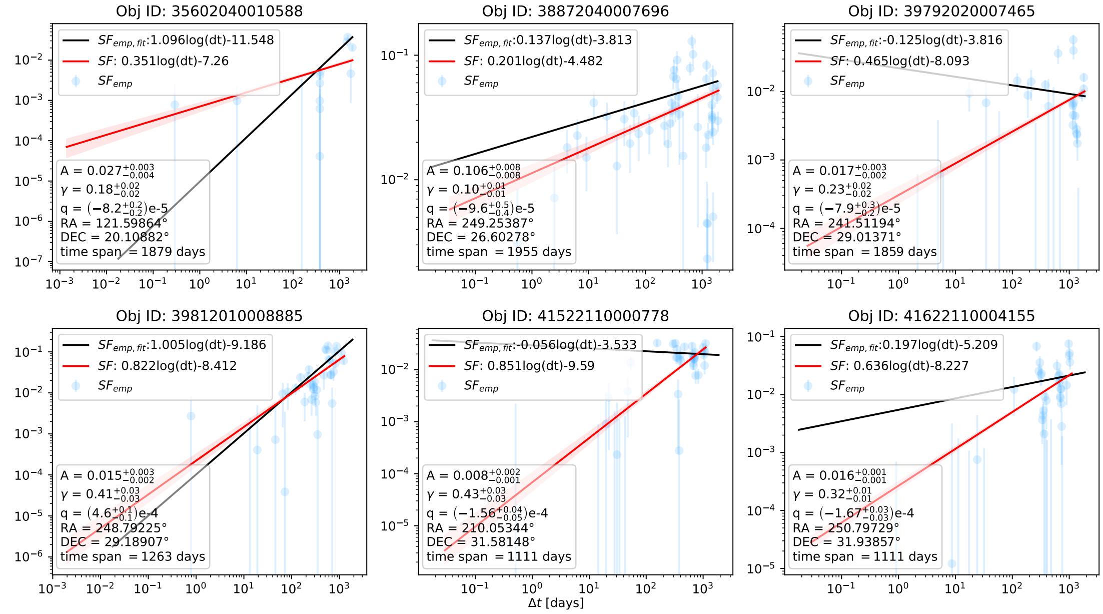

Figure 3.41: Structure functions at lensed quasar coordinates with A and q at least two standard deviations from 0. In the legend, SF fits are converted for comparison with $SF_{emp,fit}$.

.

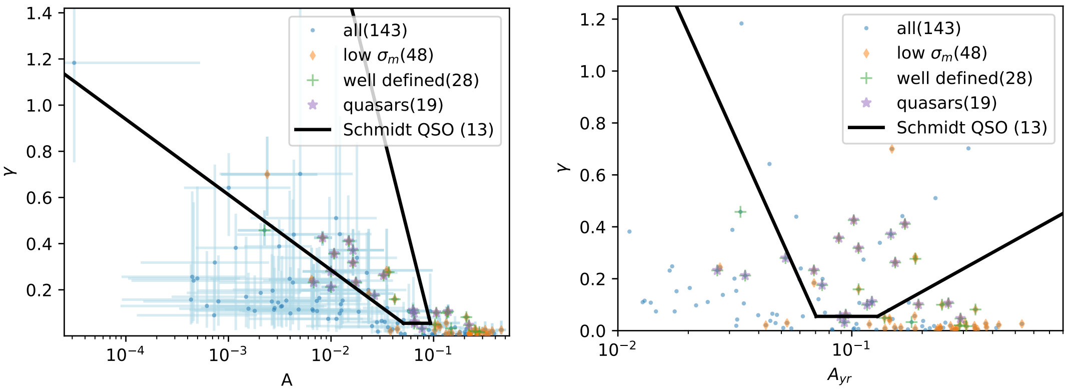

Figure 3.42: A vs. γ for different selections of objects (defined in Section 2.4) at the lensed quasar coordinates. The plot to the left used $t_0 = 1$ day and the one to the right has been converted to $t_0 = 1$ yr and had the errors removed to look like the plots from Schmidt et al. [20].

.

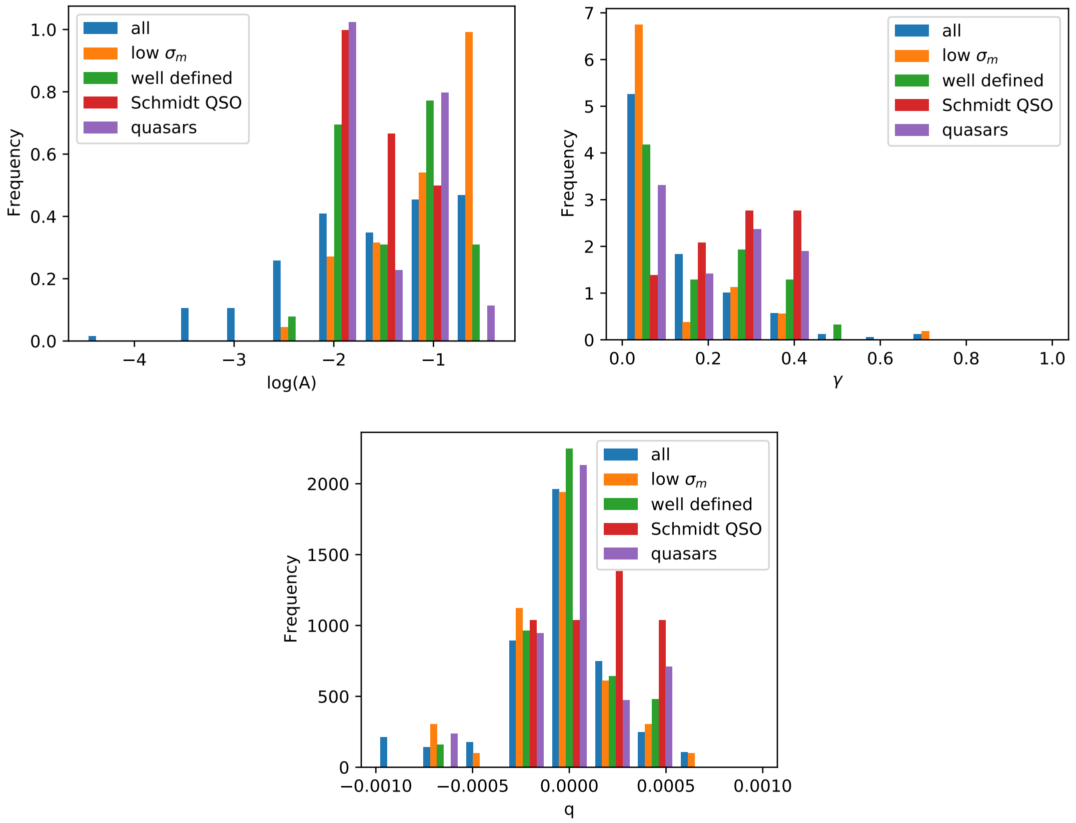

Figure 3.43: Histograms of the fit parameters for different selections of objects at lensed quasar coordinates. Left: A; right: γ; and centre: q. The histograms are zoomed in for γ and q.

.

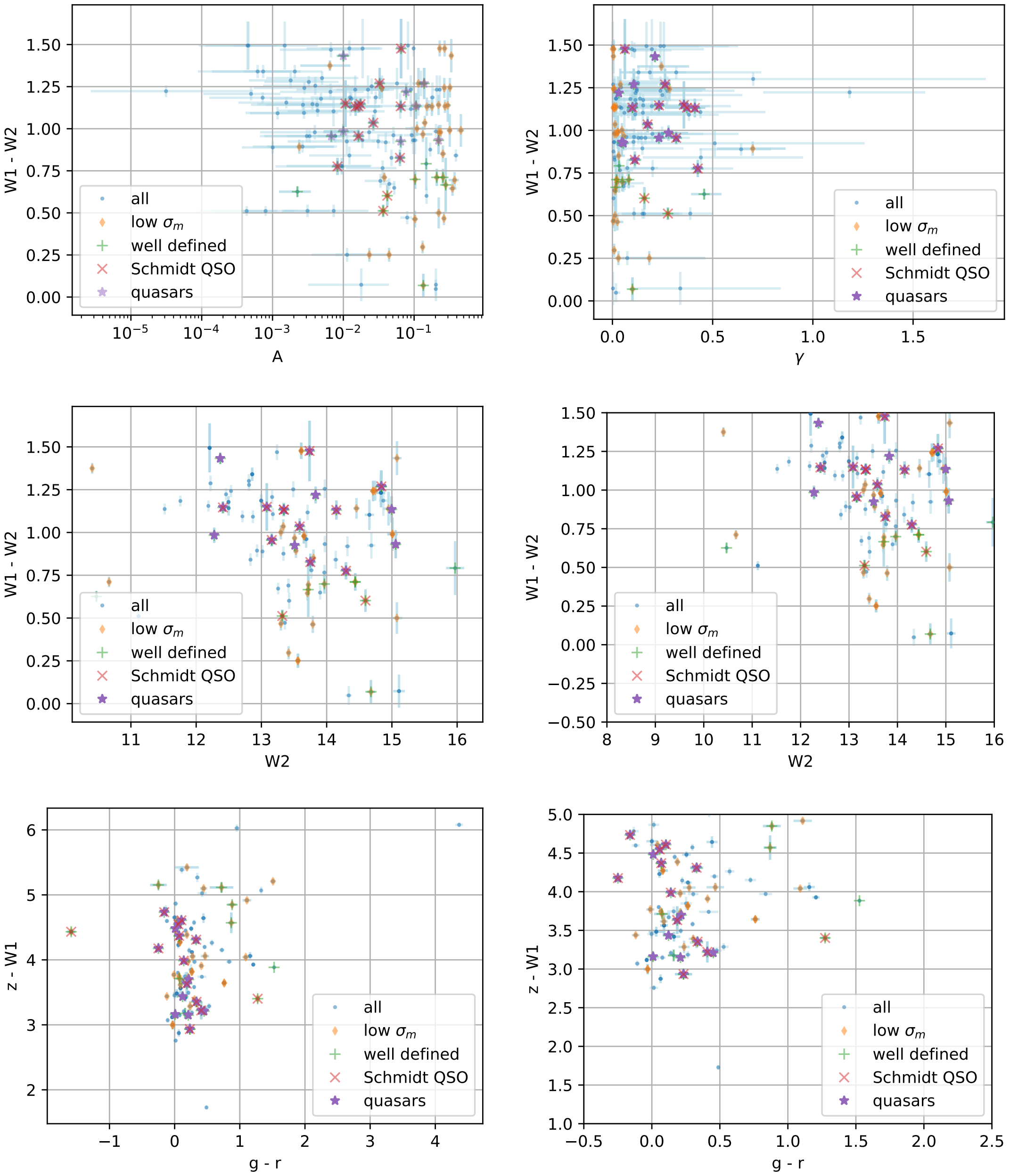

Figure 3.44: Colour plots of the sources at lensed quasar coordinates. Upper left: A vs. $W1-W2$; upper right: γ vs. $W1-W2$; centre left and zoomed in to the centre right: $W2$ magnitude vs. $W1-W2$;, and finally, lower left and a zoomed in version to the lower right: The full colour-colour plot of $g-r$ vs. $z-W1$. The zoomed in versions of the last two plots are for easier comparison with Ansari et al [2].

.

Most sources are found at high $W1-W2$ in the colour plots of Fig. 3.44. Not all of these were identified as quasar candidates, however, due to A and γ being too close to 0. It is unclear whether sources with high $W1-W2$ tend to have other values of A or γ as there are very few low $W1-W2$ sources to compare with. In $W2$ vs. $W1-W2$ and $g-r$ vs. $z-W1$, the sources fill regions most associated with quasars [2].

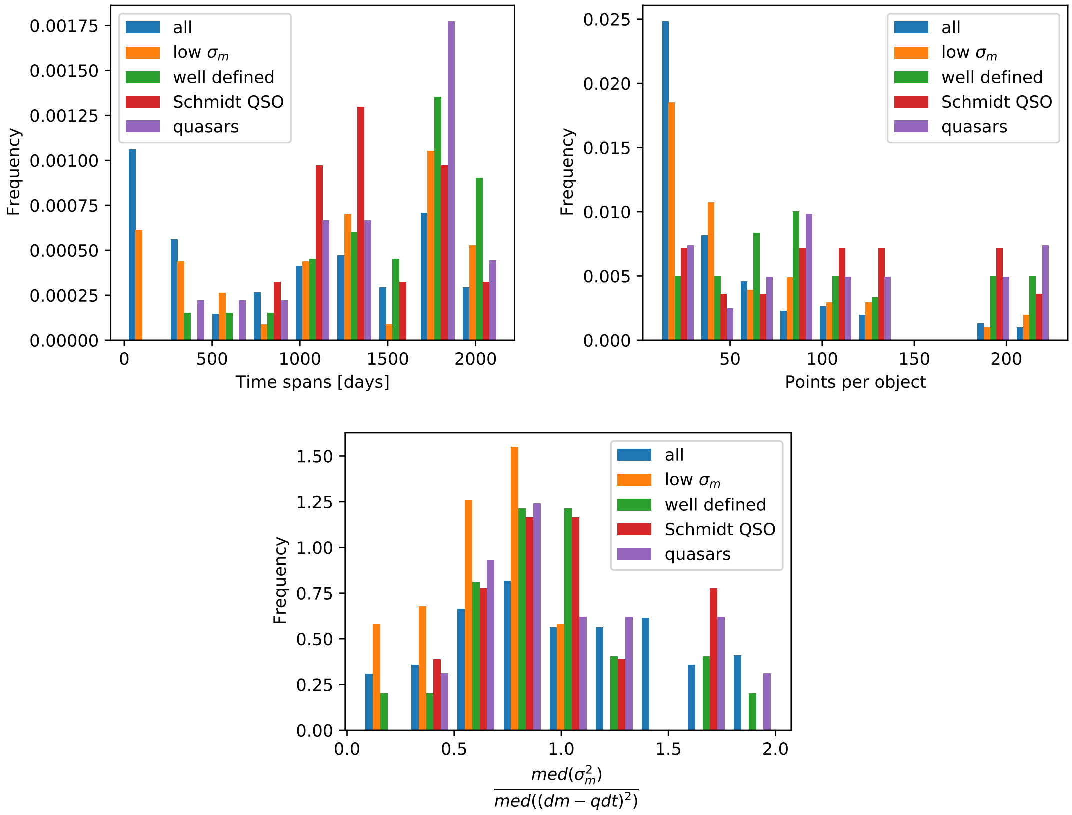

Figure 3.45: Histograms of some light curve parameters for objects at coordinates to lensed quasars: Left: Time spans; right: Number of points; and centre: The ratio of measurement uncertainties to observed variation. The last plot is zoomed in, since for some objects the measurement variation were of order 10 times larger than the observed variation.

.

Fig. 3.45 shows that the lightcurves span up to ~ 2000 days and "well defined" and especially colour selected quasars are overrepresented at long time spans. None of them are found among the objects with the shortest lightcurves, but they are included among objects with few lightcurve points (but underrepresented there). The plot of ratios of median σ_m^2 to observed variations shows similar trends to that from the box searches.

There are too few well defined points for a meaningful corner plot of those points. But the q values are close to 0 for all long light curves as shown in Fig. 3.46.

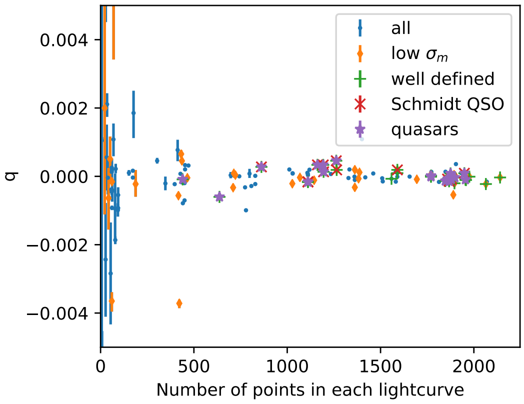

Figure 3.46: q at different time spans

.

3.2.4 SDSS quasar search variability

Some example lightcurves and structure functions from the SDSS spectroscopic quasar coordinate search are plotted in Fig. 3.47-Fig. 3.50. The first two figures are for random objects, and the last two from "well defined" objects with quasar colours in WISE.

Fig. 3.51 shows the distribution of objects in the $A-\gamma$ plane. 5181 out of the 5937 (87 %) spectroscopically selected quasars are also colour selected as quasars using WISE colours (ignoring if they have well defined variability or not). Out of the well defined ones, it is 90 %. But only 831 of the 1631 (51 %) well defined ones fullfilled the Schmidt QSO criteria for variability. Most sources are found at $A_{yr} \sim 0.1$, and the ones at low A are generally not "well defined". 2 objects were even identified as standard star candidates.

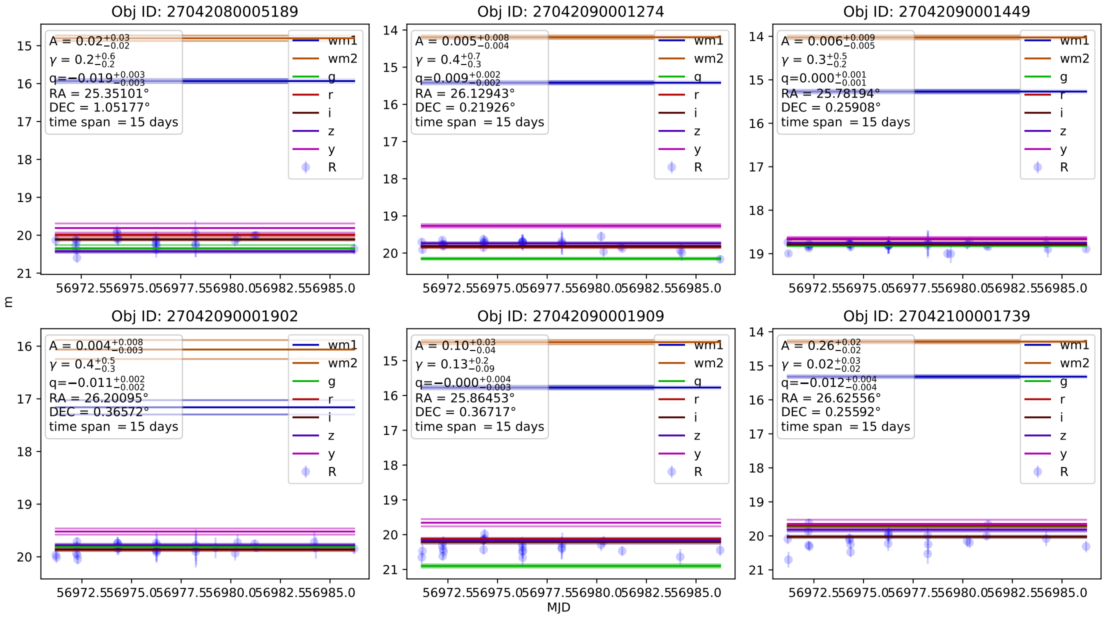

Figure 3.47: Random lightcurves at the SDSS quasar coordinates.

.

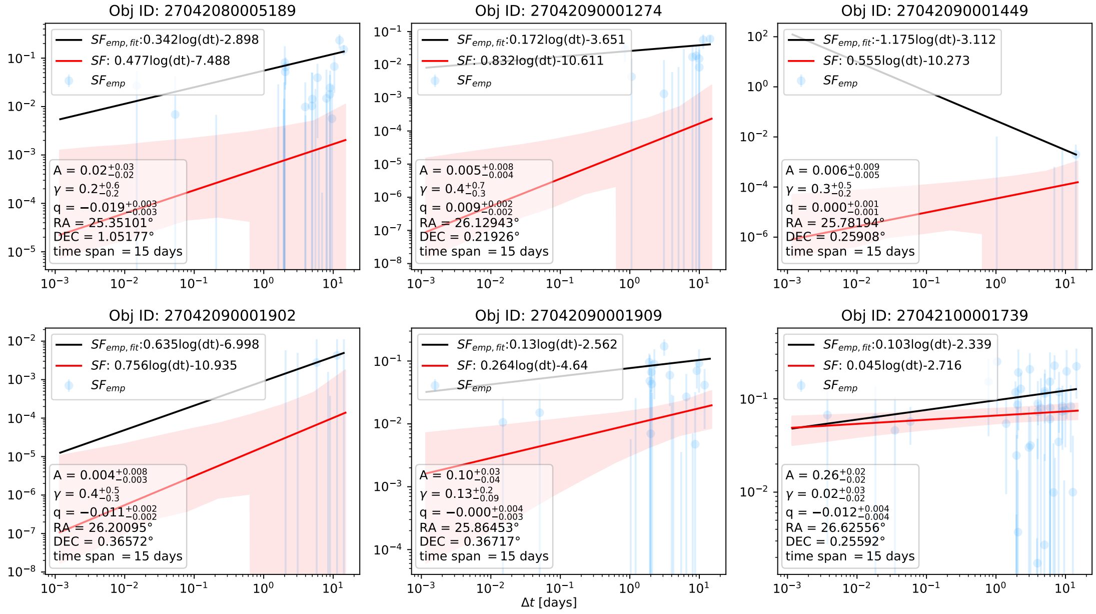

Figure 3.48: Structure functions at random SDSS quasar coordinates. In the legend, SF fits are converted for comparison with $SF_{emp,fit}$.

.

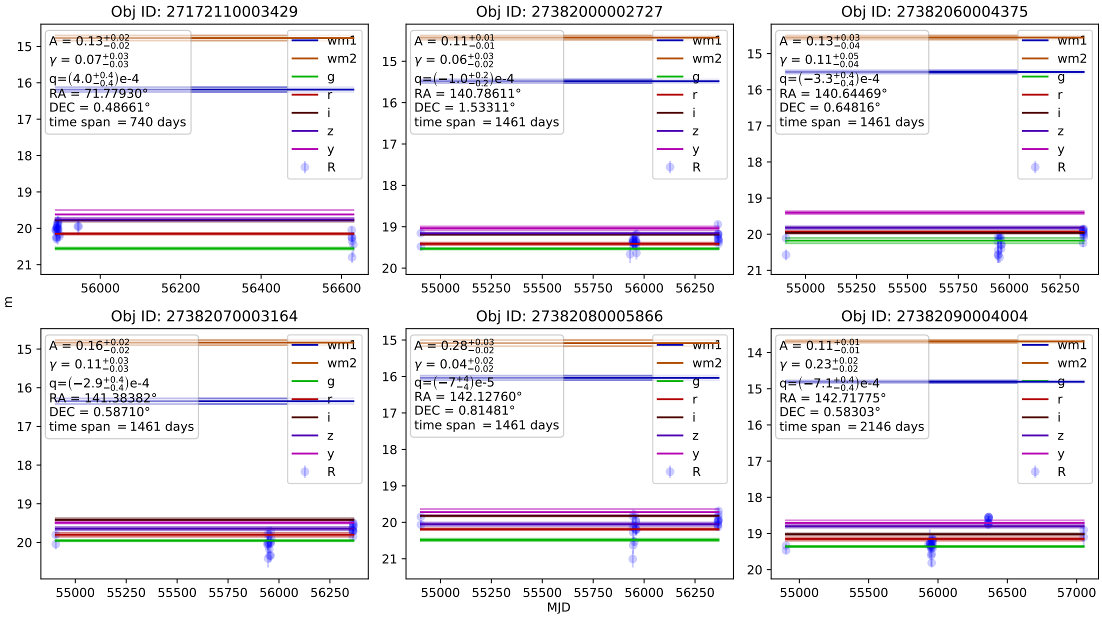

Figure 3.49: Lightcurves at the SDSS quasar coordinates with A and q at least two standard deviations from 0.

.

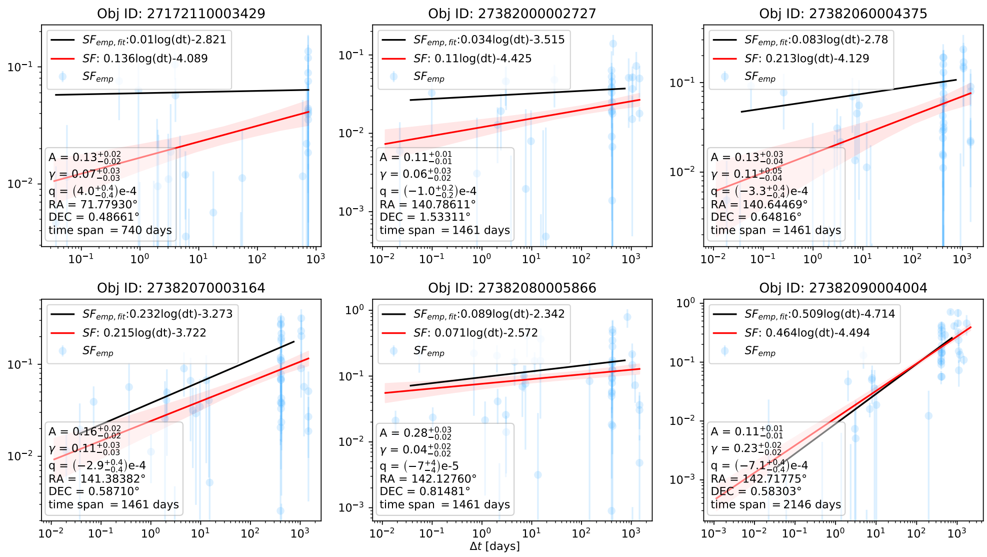

Figure 3.50: Structure functions at SDSS quasar coordinates with A and q at least two standard deviations from 0. In the legend, SF fits are converted for comparison with $SF_{emp,fit}$.

.

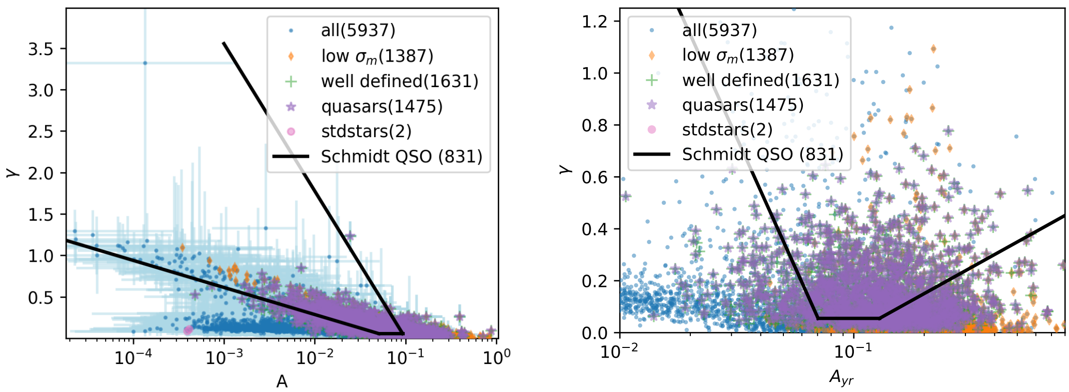

Figure 3.51: A vs. γ for different selections of objects (defined in Section 2.4) at SDSS quasar coordinates. The plot to the left used $t_0 = 1$ day and the one to the right has been converted to $t_0 = 1$ yr and had the errors removed to look like the plots from Schmidt et al. [20].

.

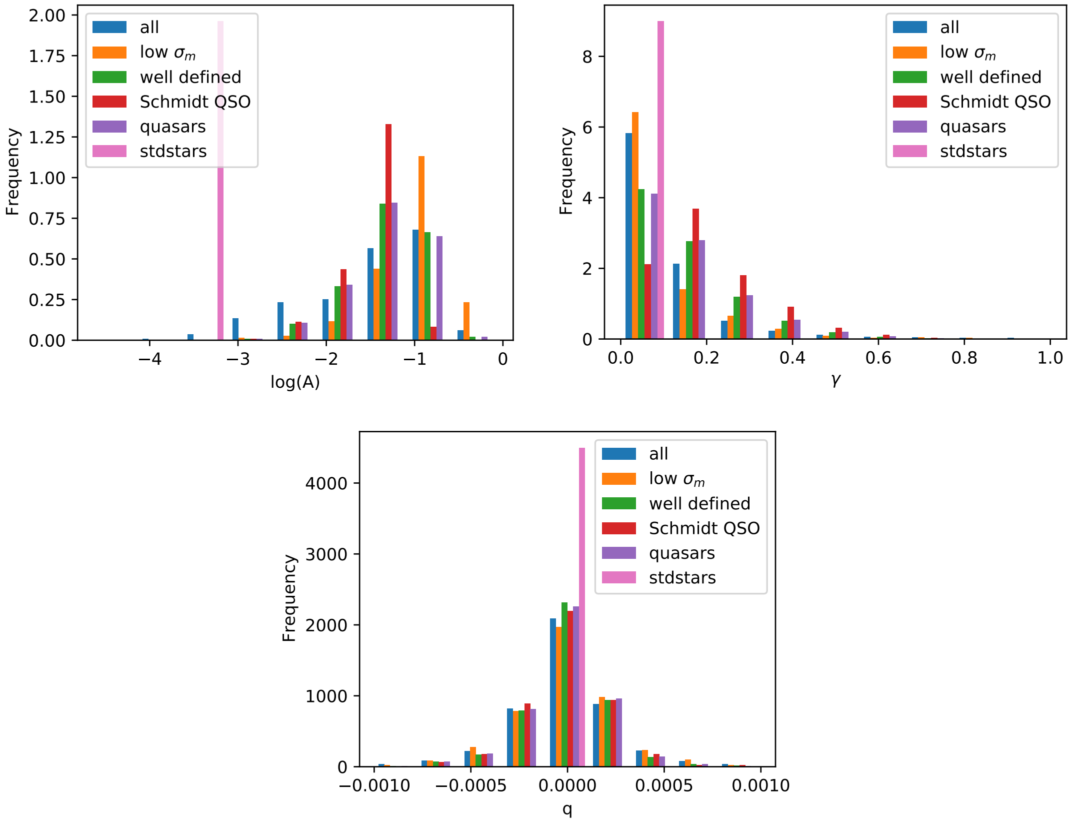

Figure 3.52: Histograms of the fit parameters for different selections of objects at SDSS quasar coordinates. Left: A; right: γ; and centre: q. The histograms are zoomed in for γ and q.

.

According to the histograms in Fig. 3.52, more of sources are found at high A than for the box searches. For γ, more sources are found at low values, and now the "well defined" objects are overrepresented at $\gamma \sim 0.2$ instead of 0.1. But this did not change for colour selected quasar candidates – they were already underrepresented at $\gamma \sim 0.1$. This means that if the goal is to select quasars, in the box search it would be better to focus on the "well defined" sources at higher γ, but in the SDSS sample, the fraction seems constant and it would be unnessecary to make a cut in γ (or A) instead of using all the "well defined" sources.

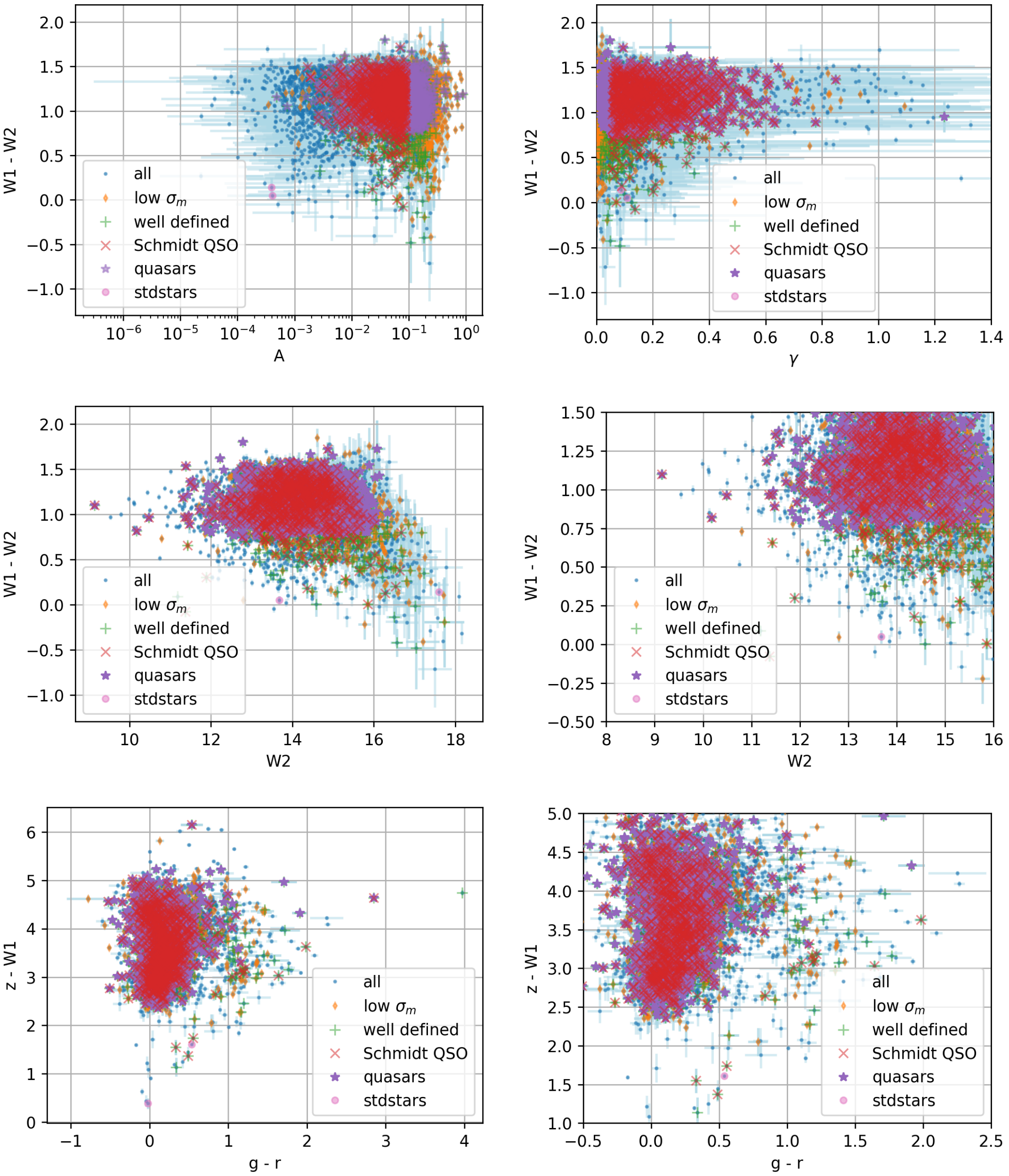

Figure 3.53: Colour plots of the sources at SDSS quasar coordinates. Upper left: A vs. $W1-W2$; upper right: γ vs. $W1-W2$ (zoomed in); centre left and zoomed in to the centre right: $W2$ magnitude vs. $W1-W2$;, and finally, lower left and a zoomed in version to the lower right: The full colour-colour plot of $g-r$ vs. $z-W1$. The zoomed in versions of the last two plots are for easier comparison with Ansari et al [2].

.

From the colour plots in Fig. 3.53, we see that most sources have high $W1-W2$, and these values are associated with a wider spread in A and γ, but there is no clear difference in the median. In $W2$ vs. $W1-W2$ and $g-r$ vs $z-W1$ the points fill regions associated with quasars, especially the quasars with low $g-r$ and high $W1-W2$ [2].

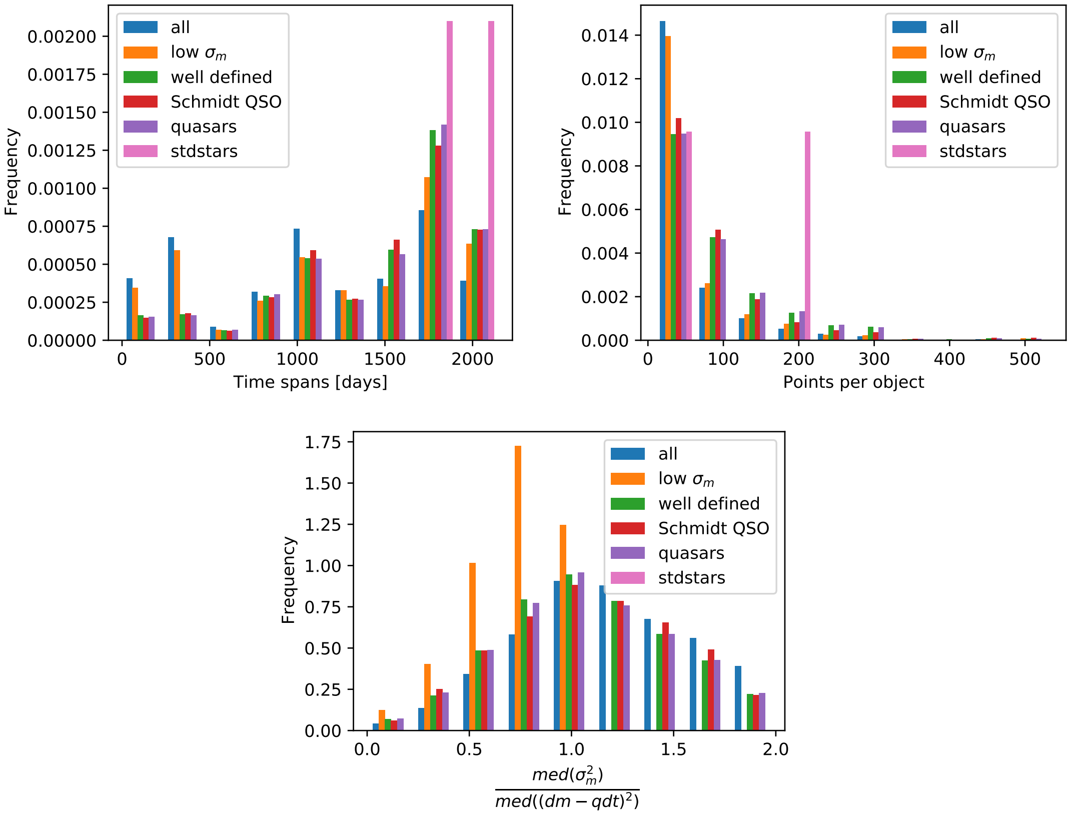

Figure 3.54: Histograms of some light curve parameters for objects at SDSS quasar coordinates: Left: Time spans; right: Number of points; and centre: The ratio of measurement uncertainties to observed variation. The last plot is zoomed in, since for some objects the measurement variation were of order 10 times larger than the observed variation.

.

The time spans are up to $\sim$ 2000 days in Fig. 3.54, and the "well defined" objects are underrepresented at short lightcurves and overrepresented for long lightcurves. The Schmidt QSO and colour selected quasars follow almost the exact same distribution as the "well defined" objects in general. Most of the SDSS quasar coordinates are associated with PTF lightcurves with few points but some have up to $\sim$ 500. Most objects have median magnitude errors close the median variability and the classes follow almost the same distribution (except the low σ_m class of course).

A corner plot for the "well defined" objects is found in Fig. 3.55. There is some correlation between A and γ for these sources, but it seems to disappear when converting to A_{yr} in Fig. 3.51. q is usually below 0.001 mag/day, but sometimes varies for short lightcurves as shown in Fig. 3.56.

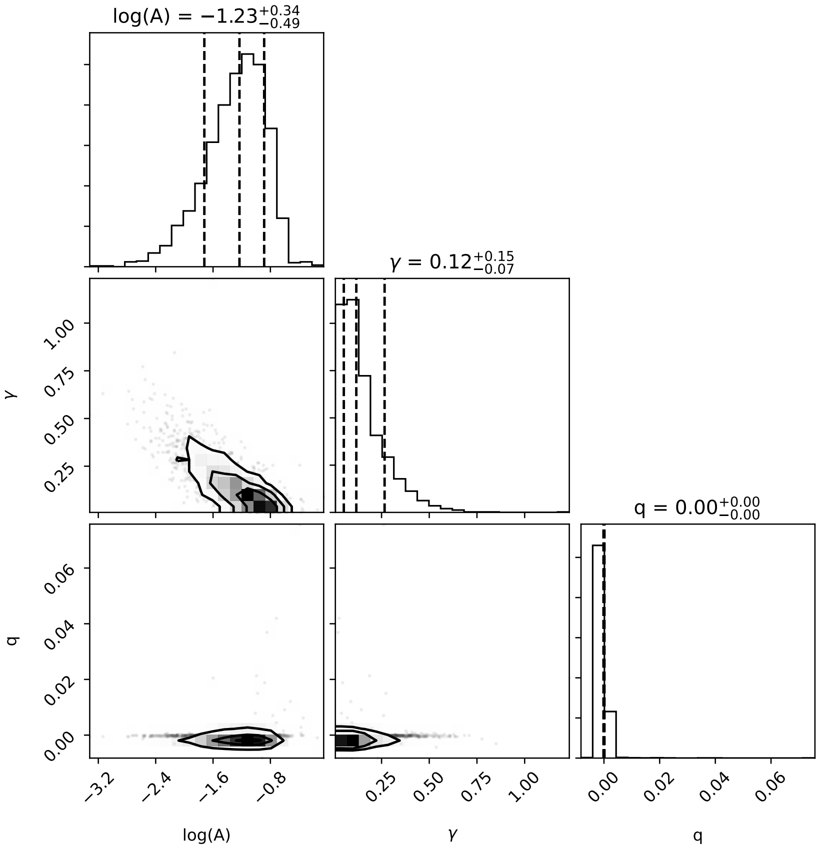

Figure 3.55: Corner plot of fit parameters for the "well defined" objects from the SDSS spectroscopically selected quasar search.

.

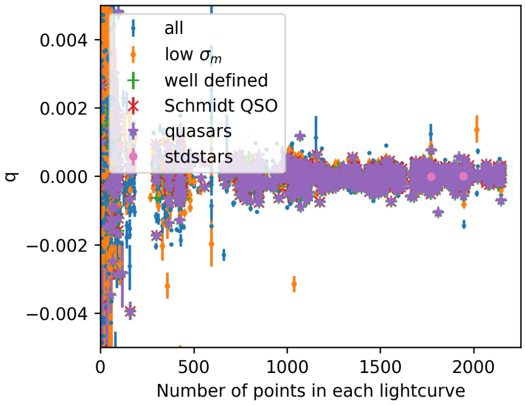

Figure 3.56: q at different time spans for objects at SDSS quasar coordinates.

.

Chapter 4

Discussion

Four different searches have been performed of sources with points in PTF, WISE and PS1: Box searches, a WISE colour selected quasar candidate search, a lensed quasar search and a SDSS spectroscopically selected quasar search. After fitting structure functions to the PTF lightcurves, the lightcurves from the box searches were most often found at $\log_{10} A \sim -2.4$ and $\gamma \sim 0.1$, but requiring that A and γ were at least two standard deviations from 0 changed the distributions, so $\log_{10} A \sim -1.0$. So when the MCMC is unsure of the fit parameters, they usually occupy a different region in the $A - \gamma$ plane. A small fraction of these were classified as standard star candidates – in fact, only 5 % of the objects from the box searches are either "well defined" or standard star candidates. So for most of these candidates, their variability is very uncertain. But it is not random for which sources this happens – sources with some properties are more difficult to classify.

The minimum number of points in each lightcurve was set to 10, and the searches are very sensitive to this number. The vast majority of accepted objects from the box searches have 10-15 points in PTF, 1 point in WISE and all of them have 1 point in PS1. But we see in Fig. 3.19 from the box search that when the number of points is below ~ 15, the objects are rarely "well defined". Few quasars are found in this sample, but they all have $\gtrsim 25$ points, and the standard star candidates are mostly found among the longest lightcurves with ~ 50 points, but there is also one with ~ 30 points. So the lightcurves with few points have A and γ that tend not to be significantly different from 0, but also not well determined enough to classify the objects as standard stars. So for the purpose of making pure and complete classifications, a more useful cut would be at least 15 points.

The box search finds objects with much fewer PTF points than the coordinate searches. This does make sense since the analysed boxes were selected as the 3 10×10 deg^2 boxes with the fewest data points to cover more of the sky without analysing too many points to begin with. So when they do contain a source it usually has few points. For the same reason, only selecting the box with the most data would not necessarily fix the problem, as it could have sources more points than the average lightcurve, but it could be interesting for comparison. A representative data set based on random boxes or objects would be better suited for further exploration of the variability of PTF lightcurves.

To classify sources based on variability over different timescales, it is not enough that the lightcurves have many points. The points also need to be spread over different time scales. As the most extreme outlier in the WISE quasar coordinate search results showed, a lightcurve that spans less than a day can lead to extreme, and very uncertain, descriptions of the variability. In the box searches, "well defined" objects and standard star candidates were underrepresented at time spans <100 days. For the coordinate searches, the long lightcurves also contained most of the "well defined sources". Some of them were found even in short lightcurves, but a large change in the fraction is observed in lightcurves spanning ~ 500 days. The fraction also continues to increase after that point.

Furthermore, q often finds extreme changes of >10 mag per year when the lightcurves are short. This is found consistently in all the searches, but also for a few longer lightcurves with large time gaps in the box search. Short lightcurves are more vulnerable to q fitting changes that would only occur on short time scales and cannot be extrapolated months or years into the future (or even just days). This can also happen in longer lightcurves if e.g. a sudden, steep variation is well described in one end of the lightcurve, and the points well-seperated from these in time are few with large errors. q can try to fit it all as one linear change, when really, it is a temporary variation or maybe a nonlinear overall change. Most lightcurves have small q if they span $\gtrsim 250$ days. So it could be interesting to try this as a minimum lightcurve length.

To avoid overfitting with q, instead of always including it in the MCMC, it could be fitted separately first, and only kept if the fit describes the data much better than a constant. Nonlinear fits with more parameters could also be used but with increased risk of fitting some of the variability.

If there are intrinsic magnitude changes over large time scales, the simple outlier removal in this project could be problematic. When points are >3 mag away from the median of the entire lightcurve, the point furthest away is removed. This is to avoid cases where a single outlier dominates the result of the MCMC. But a lot of outliers could still be kept and increase the apparent variability. And if the lightcurve has several points indicating that it has actually changed >3 mag, though that would be rare, the point should not be removed. A better approach could be removing outliers based on a moving median and only allowing smaller magnitude changes from this, such as the 0.25 mag used by Schmidt et al. [20]

It would be interesting to study the effects of gaps in the lightcurves. For example by simulating lightcurves with different gaps in time or selecting some real lightcurves and removing points from them. One could also measure the largest gap in each lightcurve and see how it correlates with fit parameters, but this method could be biased since e.g. variability could influence the gaps. If a source is sometimes very faint, it might not be detected and that would create a gap.

Branch 1 and 2 in the $A-\gamma$ plane had very different time spans, and in general the box search lightcurves did not have a smooth distribution of time spans. This does not seem representative of the lightcurves from the whole sky, considering the distributions found for the coordinate searches. And no separate branches were seen in the coordinate searches – so they could be artefacts of the box selection. A more representative sample or simulations could reveal whether the "branches" are really part of a smooth change in γ depending on time span. But these sources contain few points and are very uncertain in A and γ, and so they are not the objects we are most interested in.

Most lightcurves have lower squared median error than median observed variability, meaning there is no excess variation to fit. The errors could be overestimated - either the contributions from random error or systemic error. Systematic error could come from calibration – if the calibration is very uncertain that would add a large error contribution to the whole lightcurve, making it difficult to use the magnitude errors for structure function fitting. So finding more standard stars for calibration could actually help classify even more standard stars through variability classification.

Sometimes, PTF points are very close but have different OID's. To avoid trusting the OID's, one could cluster sources with an algorithm such as HDBSCAN instead [6] that uses variable density clusters and can separate noise points from objects with enough points.

The "well defined" points appeared to all fall into the same cluster in all coordinate searches. It is possible this would change with larger datasets or by selecting different objects, in case some types of variability are rare.

Further analysis of variability of PTF lightcurves could include unsupervised clustering with an algorithm such as extreme deconvolution gaussian mixture modelling (XDGMM) [11] that seeks to describe the data with multidimensional gaussian clusters and takes errors into account.

Ideally, the whole sky would be queried and divided into a training set and a test set to analyse the performance. Maybe excluding galactic latitudes close to 0 to avoid bias from the Milky Way. Purity and completeness could be computed for some classes by comparing with "true labels" from e.g. spectroscopic SDSS classification. And it could be tested how the classifier performs on just variability information, just colours from WISE and PS1 and a combination. It would answer the question of how much value the structure function fitting adds to the classifications. But for now, it looks like the points at least have different A and γ, although the box search and quasar searches find the same distributions of "well defined" sources. There are some differences in how many sources are classified as "well defined", though. It is 4 % in the box search, 20 % in the WISE colour selected quasar search, 20 % in the lensed quasar search and 27 % in the SDSS spectroscopically selected quasar search. So if a source is "well-defined" that can in itself be an indication that it might be a quasar – or at least probably a variable source where the variability can be described by a power law. The RR Lyrae in Schmidt et al. [20] have high A but low γ, so they would not necessarily be selected as "well defined". But with stricter requirements to e.g. the number of points in each lightcurve, they should still show up as the classifier would just be more certain of their properties. Some objects are in the RR Lyrae region of the $A - \gamma$ plane but this happens even in the quasar searches.

4.0.1 Bias of the structure functions

The structure function fits SF from Eq. 1.1 often seem to miss the data described by the empirical structure functions SF_{emp} in Eq. 1.12. This is especially clear when comparing to the fits of standard stars. Usually, the structure function appears a bit underestimated. There are several reasons this might be:

- Often, the squared magnitude errors are close to or larger than the total variability ($\sigma_m^2 > (\Delta m - q\frac{\Delta t}{t_0})^2$), meaning there is little or no excess variability to fit. These points are not included in $SF_{emp,fit}$ as it is a fit in logspace and can only include points of $SF_{emp} > 0$. So if SF includes lower points, the resulting fit will be lower, though it cannot be negative. SF from the MCMC cannot describe negative variability, but it can say that A must be low for these sources and in some cases with enough certainty that they can be classified as standard stars.
- SF only allows $A \geq 0$ and $\gamma \geq 0$ so it can never agree with $SF_{emp,fit}$ on negative slopes.
- $SF_{emp,fit}$ is fitted to binned data, so SF included more information even for the points with low σ_m^2.
- SF is affected by uniform priors and these are not uninformative. That could give a bias towards explaining as little as possible with SF when σ_m can already explain at lot of the variation. In the case of large errors, Jeffreys prior [14] can be derived as

$$p = \sqrt{\det(I)} = \sqrt{\det\left(\begin{bmatrix} F_{AA} & F_{A\gamma} & F_{Aq} \\ F_{\gamma A} & F_{\gamma\gamma} & F_{\gamma q} \\ F_{qA} & F_{q\gamma} & F_{qq} \end{bmatrix}\right)}, \tag{4.1}$$

where I is the Fisher information matrix, and

$$F_{\theta\phi} = \left\langle \frac{\partial \ln L}{\partial \theta} \frac{\partial \ln L}{\partial \phi} \right\rangle . \tag{4.2}$$

- SF could be sensitive to hyperparameters in the MCMC and could get stuck in local minima. The walkers often converge on values close to SF_{emp} but still a bit off, so this is probably not the primary problem.

- Both SF and $SF_{emp,fit}$ use data from as many time scales as possible. But data with the largest separations in time, will be very affected by random changes at the end of the lightcurves, since only these can contribute. A solution could be to only fit up to some Δt smaller than the total time span.

4.0.2 Comparison to Schmidt et al.

The colour plots show that in the box searches, most sources are in the stellar locus, but there are also many that look like quasars. In the coordinate searches, most sources clearly have quasar colours as intended for those searches. So they do seem to be relatively pure quasar samples and would, hopefully for the purpose of variability classification, be found in a different region of the $A - \gamma$ plane than random objects from the box searches. But such a difference is not observed between the well defined objects. There are more points with uncertain fit parameters in the box searches with different distributions, but some could be removed by changing the cut in number of lightcurve points, minimum time span etc. as described above.

The criteria for variability selection of quasars from Schmidt et al. [20] only selects 51 % of the "well defined" sources from the SDSS search in this project. Even though the sources were all spectroscopically selected as in the paper the criteria are based on. And while they found quasars to typically have $0.07 < A_{yr} < 0.25$ and $0.15 < \gamma < 0.5$, the spectroscopically selected quasar sample in this project found 68 % of the "well defined" points between $0.07 < A_{yr} < 0.23$ and $0.05 < \gamma < 0.27$ – or $0.02 < A_{yr} < 0.20$ and $0.15 < \gamma < 0.19$ for the full sample. So the primary difference in these results is that this project finds much lower γ.

But there are some key differences between the method and sources used in the projects.

- Schmidt et al. used lightcurves from SDSS whereas this project uses PTF lightcurves. They mentioned that A decreased going from g to z, but γ did not change. There are several steps that might improve the analysis in this project, but the PTF lightcurves in the R band seem to give lower A and γ.

- They removed outliers by comparing to a moving median as discussed earlier. So this project probably includes more outliers that would increase A. The exact ways it would affect A and γ can be simulated by introducing outliers to some lightcurves. If the outliers make the lightcurves look more like white noise that would decrease γ.

- They did not cluster points into lightcurves by object ID's, but instead queried all points within $0.5''$ of the coordinates to each object. This project searched for points within $5''$ and identical OID's, so there is a risk of mixing lightcurves if some points have been misidentified. But it seemed to happen rarely in the WISE quasar candidate search given the high $W2 - W1$ of the results. On the other hand, a strict radius of $0.5''$ has probably led to missing points in some lightcurves used by Schmidt et al.

- Other differences in lightcurve construction. They selected only objects with good flags in SDSS, whereas this project does not check for any flags but it removes points containing NaN, unrealistic magnitudes (e.g. 60 or -99) etc. as described in Section 2.

- In the box searches, the variability selected quasars have few points compared to the rest of the objects with well defined A and γ. This suggests that their properties could change if the minimum number of points in a lightcurve was higher. Note that in Schmidt et al. they found an average of 60 epochs per SDSS lightcurve, but while the classifier still had a high purity after downsampling to 6 epochs over 3 years (~ 1100 days), the completeness fell to 44 %. ~ 60 epochs are need to get both high completeness and purity. So the low number of points in PTF lightcurves can give a lower completeness of the Schmidt QSO class.

- Schmidt et al. described their priors as $\frac{1}{A}$ and $\frac{1}{1+\gamma^2}$. But these are only uninformative priors in the case of small errors. Therefore, they are not used in this project where the uncertainties are sometimes even larger than the variations in the PTF R band magnitude values.

- In this project, A is computed as an estimator of the intrinsic variability ampitude over a time scale of 1 day. γ is included in the conversion to A_{yr}, and therefore the errors on A_{yr} has contributions from the errors of both A and γ. So for the objects with high γ, as is required for the variability selected quasars, it is more uncertain whether they have the A_{yr} required for the variability selection.

- The structure functions SF from the MCMC of this project do not always seem to fit the empirical structure functions SF_{emp} well as discussed in Section 4.0.1, and therefore the A and γ cannot always be trusted.

Chapter 5

Conclusion

This exploration of variability based classification with PTF lightcurves analysed samples of both random objects, lensed quasars and quasars in general. The quasars were selected from SDSS spectroscopy or more roughly by WISE colours, the lensed quasars from the CASTLES and SQLS databases and the random objects from box searches of 3 10×10 deg^2 patches of sky. 16147 lightcurves from the box searches, 2712 WISE quasar candidates, 593 SDSS quasars and 143 lensed quasars were analysed. Based on colours, the box searches mostly find sources consistent with being stars, but some do have quasar-like colours, whereas the other searches mostly find objects with quasar colours.

Variations in magnitudes over different timescales in the lightcurves were fitted with a power law model of the structure functions with amplitude A and power law exponent γ. An MCMC estimates these fit parameters and their errors to allow for classification based on the positions of sources in their fit parameter space. We find that the samples do include sources in different parts of the parameter space, and more of the sources are identified as varying in the quasar samples. This information could be used to separate quasars from e.g. standard stars. But for the sources where A and γ are more than 2 standard deviations from 0, the search samples find very similar distributions. So it appears difficult to separate quasars from other sources with power law variability based on PTF lightcurves. Some of the varying sources of the box search are in the stellar locus of colour diagrams, so they are not all quasars.

We also see that many spectroscopically selected quasars do not fulfill the quasar variability selection criteria from Schmidt et al. [20], meaning their selection is not as suited for the PTF lightcurves analysis of this project compared to their analysis of SDSS lightcurves. Many of the selected PTF lightcurves contained fewer points – but even out of the sources where A and γ were not 0, γ was usually found to be smaller. But e.g. the priors or a less strict the outlier removal could have biased the variabilities.

Many PTF sources had magnitude errors that could explain even more variability than what is observed in total – making it very difficult to estimate intrinsic variability. Either this means the source randomly looks very stable in magnitude – or the errors are larger than the random noise affecting the measurements.

The Schmidt quasar variability criteria do perform better on the spectroscopically selected quasar sample than the quasar candidates from WISE colours. This does make sense since they were developed for SDSS quasars and the colour selection is relatively rough given that many SDSS quasars do not have such blue colours in WISE – but it is difficult to select them based on colours without more contamination from stars.

In the future, it could be interesting to make a stricter selection of lightcurves in terms of epochs and time span to make more well defined characterisations of variability including properties that show up on long time scales. This should preferably be done with an uninformative prior taking large errors into account and on objects from the whole PTF lightcurve database. Unsupervised machine learning could be used on the fit parameters and colours to see if clusters would automatically represent e.g. quasars or separate the sources in another way. Since PTF lightcurves do have different variabilities there is potential for such further analysis. Simulations can show how e.g. time spans, gaps etc. affect lightcurves, so we understand their biases, and can remove sources that would be too difficult to fit. With cleaner samples, maybe different variability selection criteria can be created that perform better in PTF.

Bibliography

5.1 Articles

[1] Abazajian, K.N., Adelman-McCarthy, J.K. et al., 2009, ApJS, 182, 543

[2] Ansari, Z., Agnello, A. and Gall, C., 2020, submitted to A&A

[3] Bañados, E., Venemans, B. et al., 2018, Nature 553, 473–476

[4] Bellm, E.C., Kulkarni, S.R. et al., 2018, PASP, 131, 018002

[5] Bentley, J.L., 1975, Communications of the ACM, 18, 509–517

[6] Campello, R. J. G. B, Moulavi, D. and Sander, J., 2013, Springer, Berlin, Heidelberg

[7] Fabian, A.C., 1999, PNAS, 96, 4749-4751

[8] Foreman-Mackey, D., Hogg, D.W., et al., 2013, PASP, 125, 306

[9] Goicoechea, L. J., Shalyapi, V.N. et al., 2008, AA,492,411–417

[10] Hastings, W. K., 1970, Biometrika, 57, 97–109

[11] Holoien, T. W. -S., Marshall, P. J. and Wechsler, R. H., 2017, ApJ, 153, 249

[12] Inada, N., Oguri, M. et al., 2012, ApJ, 143, 119

[13] Ivezić, Ž., Kahn, S.M. et al., 2019, ApJ, 873, 111

[14] Jeffreys, H., 1946, Proc. R. Soc. Lond. A, 186, 453–461

[15] Kochanek, C.S., Mochejska, B. et al., 2006, ApJ, 637, L73–L76

[16] Kozłowski, S., 2017, AA, 597, 128

[17] Law, N. M., Kulkarni, S. R. et al, 2009, PASP, 121, 1395

[18] Metropolis, N., Rosenbluth, A.W. et al., 1953, J. Chem. Phys., 21, 1087

[19] Muñoz, J.A., Falco, E.E. et al, 1999, Astrophys.Space Sci., 263, 51-54

[20] Schmidt, K. B., Marshall, P. J. et al., 2010, AJ, 714, 1194 (Erratum 2010, ApJ, 721, 1941)

[21] Sharma, S., 2006, ARAA, 55, 213-259

[22] Simonetti, J.H., Cordes, J.M. and Heeschen, D.S., 1985, ApJ, 296, 46-59

[23] Song, H., Park, C. et al., 2016, ApJ, 827, 104

[24] Tewes, M., Courbin, F. et al., 2013, AA, 556, A22

[25] Tonry, J.L., Stubbs, C.W. et al., 2012, ApJ, 750, 99

[26] Wright, E.L., Eisenhardt, P.R.M. et al., 2010, AJ 140 1868

5.2 Websites

[27] astroquery.readthedocs.io/en/latest/irsa/irsa.html, *Accessed 16-Oct-2020*

[28] astroquery.readthedocs.io/en/latest/xmatch/xmatch.html, *Accessed 31-Oct-2020*

[29] cdsarc.unistra.fr/viz-bin/cat/II/328, *Accessed 30-Oct-2020*

[30] cdsarc.unistra.fr/viz-bin/cat/II/349, *Accessed 30-Oct-2020*

[31] hpc.ku.dk/index.html, *Accessed 16-Oct-2020*

[32] irsa.ipac.caltech.edu/applications/Gator/, *Accessed 16-Oct-2020*

[33] www.ptf.caltech.edu/iptf, *Accessed 12-Oct-2020*

[34] www.sdss3.org/dr8/scope.php, *Accessed 30-Oct-2020*

[35] skyserver.sdss.org/dr8/en/tools/search/sql.asp, *Accessed 24-Oct-2020*

[36] slurm.schedmd.com/documentation.html, *Accessed 16-Oct-2020*

Acknowledgements

This work made use of an HPC facility funded by a grant from VILLUM FONDEN (project number 16599)

This publication makes use of data products from the Wide-field Infrared Survey Explorer, which is a joint project of the University of California, Los Angeles, and the Jet Propulsion Laboratory/California Institute of Technology, funded by the National Aeronautics and Space Administration.

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, K.C., et al. 2016, arXiv:1612.05560 Pan-STARRS Data Processing System, Magnier, E. A., et al. 2016, arXiv:1612.05240 Pan-STARRS Pixel Processing: Detrending, Warping, Stacking, Waters, C. Z., et al. 2016, arXiv:1612.05245 Pan-STARRS Pixel Analysis: Source Detection and Characterization, Magnier, E. A., et al. 2016, arXiv:1612.05244 Pan-STARRS Photometric and Astrometric Calibration, Magnier, E. A., et al. 2016, arXiv:1612.05242 The Pan-STARRS1 Database and Data Products, Flewelling, H. A., et al. 2016, arXiv:1612.05243

Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions.

SDSS-IV acknowledges support and resources from the Center for High Performance Computing at the University of Utah. The SDSS website is www.sdss.org.

SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, Center for Astrophysics | Harvard & Smithsonian, the Chilean Participation Group, the French Participation Group, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU) / University of Tokyo, the Korean Participation Group, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für

Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatário Nacional / MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University, and Yale University.

This thesis made use of the NBI frontpage package developed by Karen A. Voigt.