Investigating Quasar Variability using Large Ensemble Studies of Optical Photometry over Seven Decades

H. Rendell-Bhatti

Doctor of Philosophy
The University of Edinburgh
May 2024

Abstract

The term 'Active Galactic Nucleus', or AGN, refers to the existence of energetic phenomena in the nuclei of galaxies which cannot be attributed directly to stars. AGN cover a great range of luminosities. In a typical Seyfert galaxy the energy emitted by the central region is comparable with the total energy output of all the stars in the galaxy, but in a typical quasar the nuclear source is 100 or more times brighter than the collective star output of the galaxy. Quasars are the most luminous AGN and are among the most powerful sources of electromagnetic radiation in the Universe. They are distinctly variable, exhibiting fluctuations in luminosity in all parts of the electromagnetic spectrum. The timescales of variability can range from minutes to decades, depending on the wavelength observed.

Quasars are rare and only found at great distances. As a result, their small angular size make them particularly difficult to study. The majority of emitted radiation of quasars is produced in a region with an apparent size on the order of microarcseconds. However, while we cannot resolve the structure of quasars spatially, it is possible to resolve temporal fluctuations in brightness at different wavelengths. By analysing variability in photometric and spectroscopic observations, we may constrain the emission region size, understand the energetics of the system and reveal what physical processes are responsible for quasar phenomena that we observe. However, our current models of quasars fall short when attempting to predict the observed variability, especially on the timescales seen in observations.

Over the last two decades, sky surveys such as Sloan Digital Sky Survey (SDSS), Panoramic Survey Telescope and Rapid Response System (Pan-STARRS) and the Zwicky Transient Facility (ZTF) have been repeatedly imaging the sky. By combining data from these surveys, as well as photometry from photographic plates, I have produced 7-DQ; a database containing light curves of over 500,000

quasars with a 7-decade baseline. In this thesis, I describe the construction, calibration, and the subsequent analysis of 7-DQ. Through large ensemble photometric studies of this database I reveal subtle systematic trends, push existing analyses to new ground in the time-domain and test existing claims with a higher precision. These studies, presented in this thesis, pave the way to understanding the physical mechanisms responsible for variability.

Lay Summary

Active Galactic Nuclei (AGN) are extreme astrophysical objects. As their name suggests, they reside at the nucleus, or centre, of galaxies and are capable of outputting immense amounts of energy. The most luminous class of AGN, known as quasars, often produce enough energy to outshine their entire host galaxy. What's even more impressive is their relatively small size; they typically occupy a region not much larger than our own solar system. To generate power on such vast scales requires a supermassive black hole. It is so-called because it is extremely massive, greater than the mass of a million suns, and black because not even light can escape its gravitational pull. It seems counter-intuitive that a black hole can be extremely luminous. However, the source of radiation is actually from in-falling matter, which the black hole is able to convert into energy through a highly efficient process.

Quasars are so distant and small (relatively) that they appear as single points of light when looking through even the largest optical telescopes. Therefore, there is some contention about the finer details regarding the exact shape and structure of quasars. However, we are confident that, in order to generate the vast quantity of energy observed, quasars must be powered by accretion disks. Additionally, when examining the light emitted over time, it is clear that there are turbulent and chaotic physical processes which give rise to variability in brightness over timescales of days to years. This flickering provide clues to the physical mechanisms occurring in the disk, and by studying it we may better understand these mechanisms.

In recent years, advancements in telescopes, detector and computer technologies has enabled astronomers to carry out massive variability studies. Such studies involve taking many observations from a large number of quasars and mining the resulting data for patterns and correlations that were not previously seen.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

At the time of writing, outcomes of this work are in preparation for submission to the academic journals:

- Rendell-Bhatti, H., Lawrence, A., *7-DQ: Seven decades of Quasar photometry*, MNRAS, in prep.
- Rendell-Bhatti, H., Lawrence, A., *The Ensemble Variability Properties of Half a Million Quasars over Seven Decades*, MNRAS, in prep.

(*H. Rendell-Bhatti, May 2024*)

Acknowledgements

Producing this thesis was a monumental task, one that would not have been possible without the unwavering support of my family, friends and colleagues.

I would like to thank my supervisors, Andy Lawrence and James Aird, for their guidance and their support. It has been a pleasure to work alongside you both.

To my parents, whose love, encouragement, and belief in me have been constant, thank you for providing me with the freedom to pursue my passions.

To my siblings, your examples have been a source of inspiration. Fred, thank you for igniting my initial passion for science, and Flo, thank you for the countless mountain bike rides and for being the glue that holds the RBs so closely together.

I have had the enormous privilege of studying in a beautiful city in the company of wonderful friends. I am grateful to Charlie for teaching me to climb and helping me develop a skill which has brought me so much joy. Thank you Rokas, the computing wizard; assembling this thesis was much easier thanks to your custom functions, may they live long in the thesis template repository. Thanks to Matt and Hector for our countless trips to the Pentlands. To Ed, thank you for being a fountain of wisdom and creating music that brightened even the darkest times. There are so many others who have supported me along the way, and although I can't name everyone here, I am deeply grateful to each of you and for the memories we have shared.

I am also grateful to the Scottish Highlands, whose vast landscapes and pure tranquillity provided a sanctuary whenever I needed to retreat and regain perspective.

And to Nisha, there are times when it felt that finishing this thesis was an impossible task, but you helped me keep going and stay sane. I could not have done this without your immense support. I am not only grateful for your encouragement, but also our many shared memories that I will treasure forever. I am incredibly lucky to have met you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The term 'Active Galactic Nucleus', or AGN, refers to the existence of energetic phenomena in the nuclei of galaxies which cannot be attributed directly to stars. AGN cover a great range of luminosities. In a typical Seyfert galaxy the energy emitted by the central region is comparable with the total energy output of all the stars in the galaxy ($\sim 10^{11} L_{\odot}$), but in a typical quasar the nuclear source is 100 or more times brighter than the collective stellar output of the galaxy. Quasars are the most luminous AGN and are among the most powerful sources of electromagnetic radiation in the Universe. They are distinctly variable, exhibiting fluctuations in luminosity in all parts of the electromagnetic spectrum. Variability has been observed on timescales ranging from minutes to decades, depending on the wavelength observed. Since their discovery in 1963 (Schmidt 1963), the study of quasars has advanced greatly. We have gained insight of the processes underlying their emission and mechanisms driving variability to produce a 'unified model of AGN'; a theory that explains the observed phenomena of different types of AGN within a single model. This model is quite successful at describing different types of emission but falls short when attempting to predict the observed variability, especially on the timescales seen in observations. There are a number of textbooks that provide an overview on these topics (see e.g., Peterson 1997; Krolik 1999; Netzer 2013).

Quasars are rare and only found at great distances. As a result, their small angular size make them particularly difficult to study. The majority of a quasar's energy output is produced from a region ($< 500 R_s$) with an apparent size on the order of microarcseconds, assuming a typical black hole mass

($10^8 M_\odot$) and distance (100 Mpc). Only in the last few years, the Event Horizon Telescope has been able to resolve objects on this scale through international collaboration utilising a global network of radio telescopes (Event Horizon Telescope Collaboration et al. 2019). By directing these telescopes to the centre of M87, a nearby AGN, this collaboration provided the first direct image of a supermassive black hole, revealing what is believed to be a gravitationally lensed accretion disk encompassing the shadow of the event horizon. This strongly supports the prevailing idea that AGN are powered by accretion from a disk onto a supermassive black hole. While M87 is not luminous enough to be classed as a quasar, it is widely accepted that quasars are powered by the same process, except at a much higher accretion rate. This was a breakthrough in imagery, although it has not led to drastic revision of the unified model.

While direct imaging may not hold the answers to the many open questions regarding AGN, clues can be found by analysing the fluctuations of photometric and spectroscopic observations over time. Through these studies we may constrain the emission region size, understand the energetics of the system, and reveal what physical processes are responsible for AGN phenomena that we observe. There are types of variability that contribute to the overall observed variability; that which is intrinsic to the quasar and that which is extrinsic, caused by something interfering with our view of the quasar. Intrinsic quasar variability is driven by various complex physical mechanisms and is different for each part of the electromagnetic (EM) spectrum. The sources of intrinsic variability are discussed in more detail in Section 1.5. Extrinsic effects include occultation of the central source by clouds or winds, and gravitational microlensing. However, obscuration is likely to happen on timescales much longer than the timescales over which we observe, and microlensing is rare. Additionally, my work focuses on a subgroup of quasars, known as type I quasars, in which we have an unobstructed view of the nucleus. Observed variability in these objects will be vastly dominated by intrinsic variability, and therefore extrinsic contributions may be considered negligible.

Quasars' immense energy output and strong gravitational fields make them relevant for other areas of astronomy as well. Their great luminosities allow them to be detected over cosmic distances, making them a key population for testing cosmological models. AGN also have a close relationship with their host galaxy and play an important role in galaxy evolution by controlling star formation rate through AGN feedback (see e.g., Ishibashi & Fabian 2012; Fabian 2012; Heckman

& Best 2014; Tadhunter 2016). This process drives out gas from the galaxy and can quench star formation, which is the likely cause of the observed M-σ relation (see Ferrarese & Merritt 2000). Quasars can also be used to study the effects of gravitational redshift, as the majority of emission is produced from regions of strongly curved space-time close to the black hole (see e.g., Tanaka et al. 1995; Fabian et al. 2000).

My research, presented in this thesis, concerns the variability of quasars seen in the optical band. Radiation observed at visible wavelengths originates from a range of rest-frame wavelengths due to cosmological redshift. For the range of quasar redshifts considered in this work, these rest-frame wavelengths span from the optical to the near ultraviolet. Emission produced at these wavelengths is generally considered to originate from the accretion disk and caused by similar physical mechanisms, compared to X-rays and Radio waves, for example. Thus, studying optical variability in the observer-frame is a proxy for investigating the accretion disk, its relationship with the black hole, and the local surrounding medium. In this thesis, optical variability will refer to the observed variability in the optical bands, but the reader should note that this corresponds to rest-frame variability in the optical (~5000 Å) to near-UV (~1000 Å), as mentioned. Rest-frame wavelengths will be stated explicitly in situations where this is not the case.

Optical radiation is often thought to originate from the accretion disk as black-body radiation. Optical variability is therefore attributed to changes in accretion rate within the disk, though it is possible that part of the variability is due to reprocessed emission from other sources. My work focuses on photometric analysis of a large sample of quasars, utilising photometric observations from multiple surveys to characterise optical variability in a large sample of quasars over unprecedented timescales.

Over the last two decades, sky surveys such as the Sloan Digital Sky Survey (SDSS; York et al. 2000), the Panoramic Survey Telescope and Rapid Response System (Pan-STARRS; Chambers et al. 2016) and the Zwicky Transient Facility (ZTF; Masci et al. 2019) have been repeatedly imaging the sky. By combining data from these surveys, as well as photometry from photographic plates, I have produced 7-DQ; a database containing light curves of over 500,000 quasars with a 7-decade baseline. A database of this size, in number of objects, volume of observations, and time-span, is unprecedented. With it, I have been able to carry out detailed ensemble studies to reveal subtle systematic trends, pushing

existing analyses to new ground in the time-domain, and test existing claims with a higher precision. These studies pave the way to understanding the physical mechanisms responsible for variability and how this variability is dependent on quasar properties.

This chapter aims to provide an overview of quasars and focuses on aspects that are key to my work. I start by describing the discovery of quasars in Section 1.1. In Section 1.2, I describe the observed properties of quasars across the EM spectrum. In Section 1.3, I outline the theory of black hole accretion and give an explanation for how the extraordinary luminosities of quasars are achieved. Section 1.4 focuses on the inner region of quasars, describing the accretion disk which is the source of the quasar's tremendous energy output, including a discussion of the different timescales involved in black hole accretion and how they manifest into what we see in observations. I then go on to describe our current understanding of the geometry and structure of quasars in Section 1.5, and suggest how the inner and outer regions give rise to observed variability at different wavelengths. Section 1.6 provides a brief overview of some of the most variable quasars observed, dubbed 'Extremely Variable Quasars' (EVQs). In Section 1.7, I give an overview of common statistical tools used to study quasar variability. One such tool is the structure function, which, being particularly important for my work, is discussed in more detail in Section 1.8. In Section 1.9, I give an overview of stochastic models that are often used to model quasar variability, and describe one such model that I use in this thesis. In Section 1.10, I give an overview of recent efforts to carry out quasar variability studies on a large scale and highlight current open questions which motivates my work. In Section 1.11, I outline the current challenges and open questions which motivates my work. This chapter concludes in Section 1.12 with an outline of the remainder of the thesis.

1.1 Discovery of Quasars

Carl Seyfert was the first person to classify a group of sources that are known today to belong to the wider category of active galaxies (Seyfert 1943). These sources, known as Seyfert galaxies, tend to be local, have a clearly resolved host galaxy and have a relatively low luminosity ($10^{9-10} L_{\odot}$) compared to quasars. Their activity originates from the galactic nucleus, making them a subset of Active Galactic Nuclei (AGN). Seyfert noticed strong, broad emission lines in some of these galaxies, which were clearly distinct from stellar emission. Although this

was unusual, it wasn't until the emergence of radio astronomy in the following decade that it became clear that Seyfert galaxies belonged to a much broader class of objects. Large radio surveys subsequently lead to the identification of quasars; the Third Cambridge (3C) catalogue (Edge et al. 1959) contained two objects which were key to their discovery. One of these objects, 3C 48, was first identified as a blue star (Matthews & Sandage 1963), although it displayed broad emission lines at unexpected wavelengths. The other object, 3C 273, appeared to be a variable blue star-like object, also exhibiting peculiar emission lines. Schmidt (1963) and Oke (1963) realised that one of these emission lines was MgIIλ2798, which had been significantly redshifted ($z \approx 0.158$). Such a redshift implied they were very distant (compared to other known distant sources at the time) and therefore extremely luminous given their apparent magnitude. Their structure could not be resolved; they appeared point-like and were therefore dubbed 'quasi stellar objects', often abbreviated to QSO.

It is worth noting that the discovery of the first quasars via the identification of radio sources led to a misunderstanding. The term 'quasar', now applied to any high-luminosity active galactic nucleus (AGN) exhibiting broad optical and ultraviolet emission lines, was coined as a contraction of 'quasi-stellar radio source'. For nearly a decade, it was assumed that the typical high-luminosity AGN (a term now considered outdated) was inherently 'radio-loud'. It wasn't until around 1970 that it became apparent that only about 10% of all quasars emit as much as 1% of their total luminosity in the radio spectrum.

Much of AGN research in the decade following their discovery was devoted to produce a single model of AGN to understand what these objects were, and explain their extreme luminosities. This single model represents the idea of AGN unification; a theory that could describe all types of AGN and explain their unusual behaviour. However, AGN unification is still a contentious topic to this day.

1.2 Observed Emission of Quasars across the Electromagnetic Spectrum

Quasars emit radiation over the full range of the EM spectrum. The spectral energy distribution (SED) for a typical quasar is shown in Figure 1.1. The

SED displays a triple hump structure: a peak in the infrared, a peak in the UV/optical, and a peak in the X-rays. Not all quasars are radio sources, resulting in the distinction between radio-loud and radio-quiet seen in Figure 1.1. In the following subsections, I will discuss observed quasar emission in each part of the EM spectrum in more detail, focussing on optical/UV which is the focus of this thesis.

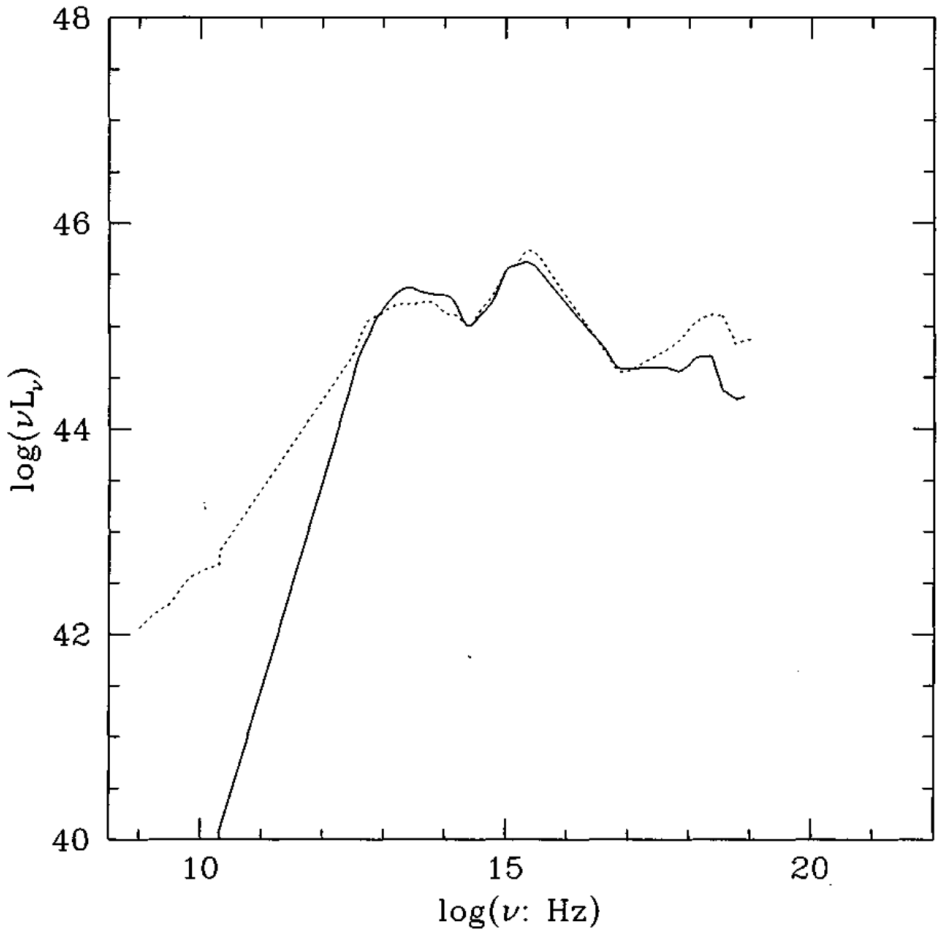

Figure 1.1: Composite SEDs for radio loud (dotted) and radio quiet (solid) AGN. Both radio loud and quiet sources display three distinct peaks in the infrared, optical/UV and X-ray. Figure from Krolik (1999)

1.2.1 Emission in the Optical and Ultraviolet

Quasars emit optical and ultraviolet radiation in the form of emission lines and a thermal continuum. While the hydrogen lines are the most prominent, lines from metallic elements such as carbon and magnesium are also commonly observed. The emission lines of AGN can be categorised into two main groups: broad and

narrow. Varying line widths is most likely due to Doppler broadening caused by bulk motion. The emission in the thermal continuum accounts for the majority of energy output in this spectral range, and its fluctuation with time is the focus of my work. A composite of ~2000 quasar spectra from Vanden Berk et al. (2001) is shown in Figure 1.2.

The 'big blue bump' at ~1000Å is the most prominent feature in the spectral energy distribution of AGN. This broad feature peaks in the UV (see e.g., Sanders et al. 1989; Elvis et al. 1994), somewhat consistent with the expectation that this emission is produced from the superposition of thermal emission from a disk at a range of different radii (Lawrence 2012). In this thesis, I explore rest-frame wavelengths ranging from 1000 Å to 5000 Å. This range enables me to investigate the emission in the region of the prominent big blue bump, where most of the interesting variability is expected to occur.

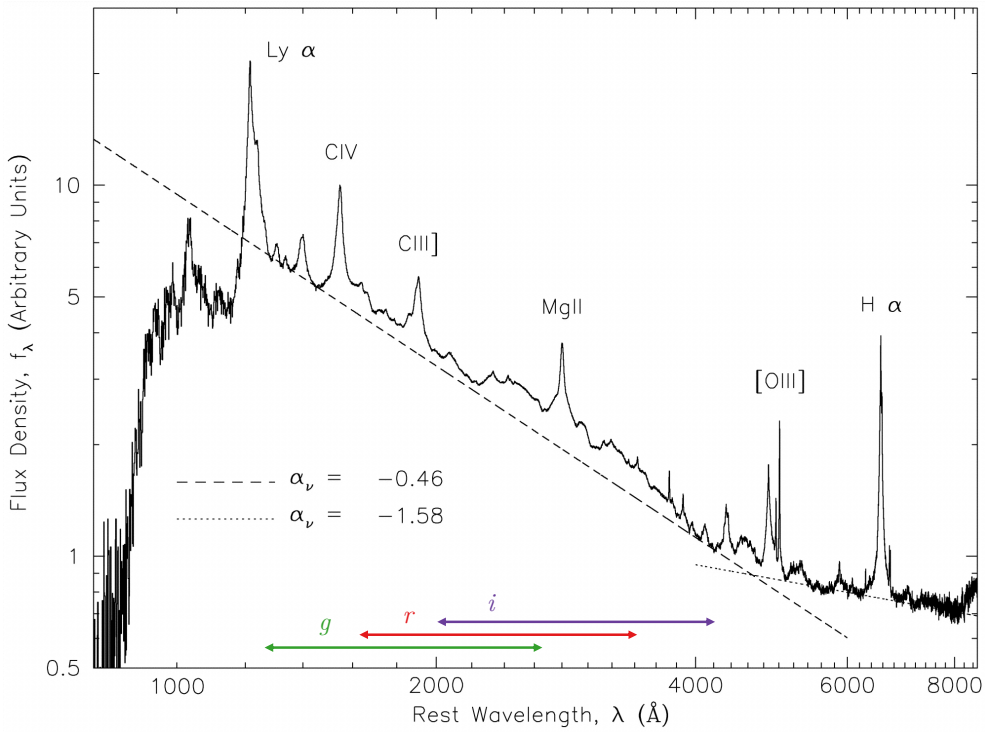

Figure 1.2: A composite spectrum from the Sloan Digital Sky Survey. The three horizontal coloured bars show the effective rest-frame wavelength ranges for the g, r and i optical bands for quasars with redshifts in the range $0.8 < z < 2.8$, which contain 80% of the DR14Q sample. Adapted from Vanden Berk et al. (2001).

1.2.2 Emission in the Radio and X-rays

Nearly all AGN produce radio emission to some degree, however, only ~10% of AGN are classified as radio-loud sources. It is possible to resolve smaller structures through Very Long Baseline Interferometry (VLBI), where milliarcsecond imaging is possible. In a number of sources, radio emission has been used to map extremely large structures extending from the nucleus, often forming double lobes either side of the galaxy. This arrangement suggests a system where energy and matter are ejected from the central region, generating emission on scales similar to that of the host galaxy. However, even in radio-loud AGNs, radio emission never accounts for more than ~1% of the bolometric luminosity.

While AGN display many phenomena, the single most common characteristic of AGN is that they are all luminous X-ray sources. It is common to observe a high energy bump in the X-rays, around 10 keV, representing approximately 10% of the bolometric luminosity. X-ray images of AGN are generally not well resolved and show a softer spectrum than that of the cosmic X-ray background, suggesting that a large fraction of AGN are obscured (see e.g., Merloni et al. 2014).

1.3 Quasar Energy Output

1.3.1 Radiation efficiency

The mechanism by which quasars generate their extreme luminosity was unclear until work by Salpeter (1964) and Lynden-Bell (1969) showed that accretion onto a compact, supermassive ($> 10^6 M_\odot$) object can liberate gravitational potential energy through a highly efficient process, thereby explaining the observed vast energy output. I will quantify this efficiency through a relativistic calculation. Starting from the GR energy equation,

$$\dot{r}^2 + \frac{h^2}{c^2 r^2}\left(1 - \frac{2\mu}{r}\right) - \frac{2\mu}{r} = (k^2 - 1), \tag{1.1}$$

where h is the specific angular momentum, k is a constant proportional to energy and $\mu = GM/c^2$ for convenience. By considering a particle at rest at infinity, i.e., $\dot{r} = 0$ and $r \to \infty$, we see that k is simply the ratio of the energy, E, to the rest

mass energy, m_0c^2,

$$k = \frac{E}{m_0c^2}. \tag{1.2}$$

Since we are interested in finding the energy as a function of radius for circular orbits, we may set $\dot{r} = 0$ and rearrange to give,

$$k^2 = \left(1 + \frac{h^2}{c^2r^2}\right)\left(1 - \frac{2\mu}{r}\right). \tag{1.3}$$

It is possible to show that the angular momentum relates to the radius of a circular orbit by

$$h^2 = \frac{\mu c^2 r^2}{r - 3\mu}. \tag{1.4}$$

Combining Equations 1.3 and 1.4 gives

$$k = \frac{1 - 2\mu/r}{(1 - 3\mu/r)^{1/2}}. \tag{1.5}$$

The trajectory closest to the black hole, where a particle can remain stable without being drawn into the singularity, is known as the 'innermost stable circular orbit' (ISCO) and is located at $r = 6\mu$ for a non-rotating Schwarzschild black hole. Evaluating the previous expression at $r = 6\mu$ and relating k back to E via Equation 1.2 gives

$$\frac{E}{m_0c^2} = \frac{2\sqrt{2}}{3} \approx 0.943. \tag{1.6}$$

Assuming that the particle liberates the difference in energy as radiation, the maximum radiation efficiency of the disk is

$$\eta_{\text{acc}} \approx 1 - 0.943 = 5.7\%, \tag{1.7}$$

slightly less than the simple Newtonian calculation which yields $\eta_{\text{acc}} = 8.3\%$. For a spinning black hole, the ISCO is even smaller, allowing for efficiencies up to 42% for a maximally spinning black hole. In contrast, nuclear fusion of hydrogen into helium yields efficiencies of $\eta_{\text{fusion}} \approx 0.7\%$. Therefore, it should come as no surprise that a black hole with a high accretion rate gives rise to the most energetic phenomena in the known universe.

1.3.2 The Eddington limit

It is intriguing to ask whether there is a limit to the rate at which a black hole can accrete. Could a quasar reach arbitrary large luminosities if more and more matter were to be supplied to it? While the gravitational pull acts radially inward, radiation pressure caused by photons emitted close to the black hole will push, on average, radially outward. As accretion rate goes up, so does radiation pressure. At some point, a balance between gravity and radiation pressure is achieved; this is known as the Eddington Limit. This approximation assumes the photons interact with free electrons whose cross-section is given by the Thompson cross-section σ_{T}, and that the electrons are bound, either as a plasma, or in atomic form to single protons (of mass m_p). Using these assumptions, the Eddington luminosity of a black hole of mass M is given by

$$L_{\mathrm{Edd}} = \frac{4\pi G m_p c}{\sigma_{\mathrm{T}}} M, \tag{1.8}$$

with a corresponding Eddington accretion rate,

$$\dot{M}_{\mathrm{Edd}} = \frac{L_{\mathrm{Edd}}}{\eta c^2}, \tag{1.9}$$

where η is the radiation efficiency. The same concept also gives the Eddington ratio n_{Edd}, which is defined as the ratio of the bolometric luminosity of the quasar L_{bol} to the Eddington luminosity,

$$n_{\mathrm{Edd}} = \frac{L_{\mathrm{bol}}}{L_{\mathrm{Edd}}}. \tag{1.10}$$

We expect $0 < n_{\mathrm{Edd}} \leq 1$ for most cases. Despite this theoretical limit, it is possible for quasars to achieve $n_{\mathrm{Edd}} > 1$, although I will not focus on such sources in this work.

1.4 The Accretion Disk

Up until this point, I have assumed that accreting matter always forms a disk. While there are numerous geometries that matter can take when accreting onto a black hole, in all of these geometries (with the exception of idealised spherical accretion) there will be a combination of non-zero angular momenta from the

in-falling particles. A disk naturally forms through collisions of the accreting matter, with the vector sum of angular momenta dictating the orientation of the disk. Therefore, it is widely accepted that the majority of black hole accretion, and luminosity generated as a result, occurs through disk accretion. In this section, I will discuss the standard simplified disk model and characteristics of such a model and where it falls short of reproducing observed phenomena.

1.4.1 Disk Models

The archetypal model for the accretion disk assumes a geometrically thin, optically thick disk with a constant accretion rate (Pringle & Rees 1972; Shakura & Sunyaev 1973). In this context, geometrically thin means that the ratio of the disk scale height, h, to the radius, r, is small ($h/r < 0.1$). Shakura & Sunyaev (1973) assumed the unknown kinematic viscosity ν took the form $\nu = \alpha c_s h$ where c_s is the local sound speed and α is a dimensionless viscosity parameter restricted to $0 < \alpha \leq 1$. This was motivated by the idea that the upper limit on the size of turbulent eddies is of the order h which propagate at speed c_s, giving the form of ν simply via dimensional analysis. Many unknowns about the dynamics of the system, including mechanisms of angular momentum transport, are absorbed into the parameter α, hence the reason why these types of simplified models are often known as 'α-disk' models. Gravitational energy of such a disk is dissipated through viscosity, causing local heating and determining the temperature profile of the disk (Peterson 1997),

$$T(r) = \left[\frac{3GM\dot{m}}{8\pi\sigma r^3}\left\{1 - \left(\frac{r_{\rm in}}{r}\right)^{1/2}\right\}\right]^{1/4}, \tag{1.11}$$

where $r_{\rm in}$ is the radius of the inner edge of the disk. For $r \gg r_{\rm in}$, this approximates to

$$T(r) \approx \left(\frac{3GM\dot{m}}{8\pi\sigma r_s^3}\right)^{1/4}\left(\frac{r}{r_s}\right)^{-3/4}, \tag{1.12}$$

where $r_s = 2GM/c^2$ is the Schwarzschild radius. This can be written in terms of the Eddington accretion rate (Equation 1.9) for quasars of appropriate mass,

$$T(r) \approx 6.3 \times 10^5 \left(\frac{\dot{M}}{\dot{M}_{\rm Edd}}\right)^{1/4}\left(\frac{M}{10^8 M_\odot}\right)^{-1/4}\left(\frac{r}{r_s}\right)^{-3/4} \text{K}. \tag{1.13}$$

For a $10^8 M_\odot$ black hole accreting at the Eddington rate, the innermost edge of an accretion disk reaches temperatures of $T(3r_s) \approx 3.7 \times 10^5$ K, resulting in peak black body emission at ~150Å in the extreme UV, sharply contrasting with the observed peak emission in the near-UV (see Section 1.2.1). The full spectrum of the disk may be calculated by integrating over the disk,

$$L_\nu = \int_{r_{\rm in}}^{r_{\rm out}} dL_\nu = \frac{8\pi^2 h\nu^3}{c^2} \int_{r_{\rm in}}^{r_{\rm out}} \frac{r dr}{\exp(h\nu/kT) - 1}, \tag{1.14}$$

where we have assumed the disk is face-on. By using $T(r)$ from Equation 1.13, this reduces to

$$L_\nu \propto \dot{M}^{2/3} M^{2/3} \nu^{1/3}, \tag{1.15}$$

revealing the important $L_\nu \propto \nu^{1/3}$ dependence.

1.4.2 Sources of Viscosity in the Disk

For accretion to occur, the infalling gas must not only dissipate energy, but it must also lose almost all of its angular momentum, which is assumed to happen through viscous torque. The disk rotates differentially, such that neighbouring rings slip past each other. Viscosity within the disk causes these rings to drag on one another, allowing angular momentum to be transferred radially outward. Viscous drag also causes local heating which, if radiated thermally on the spot, generates a temperature profile $T \propto r^{-3/4}$. The origin of this viscosity is still not fully understood. One possibility, molecular viscosity, is far too small and cannot be responsible for the typical luminosities of AGN. Balbus & Hawley (1991) showed that the magneto-rotational instability (MRI) was a possible mechanism by which turbulence could be sustained and is currently a strong candidate for explaining the effective viscosity. Another possibility is that the viscous torques and thermal dissipation do not occur on the spot, but instead act through physical mechanisms at a distance, such as large scale magnetic fields.

1.4.3 Timescales

Physical processes in quasars can occur over a wide range of different timescales; from hours to thousands of years. Therefore, it is important to quantify the

physical timescales present in the disk. The following subsections describe the timescales relevant to optical variability. The expressions for these timescales are well established and detailed in Netzer (2013).

The Light Crossing Timescale

The shortest timescale of the disk is the light crossing time. It is calculated simply by the time taken for light to traverse the disk,

$$\tau_{\text{light}} = r/c, \tag{1.16}$$

and represents the smallest physically possible interval over which a region at r can respond to changes in brightness of the central region. This timescale sets the lower temporal limit for variations in brightness which originate from the same event in the disk.

The Dynamical Timescale

If accretion flow is approximately Keplerian, the dynamical timescale at some radius r in the accretion disk around a black hole of mass M is given by the Keplerian frequency at that radius:

$$\tau_{\text{dynamical}} = 1/\Omega_K = \sqrt{\frac{GM}{r^3}}. \tag{1.17}$$

The timescale for matter to free-fall, τ_{freefall}, into the black hole, or orbit, τ_{orb}, at radius r is of the same order. $\tau_{\text{dynamical}}$ provides the lower temporal limit for structural changes in the disk. Note that spiral density waves or magnetic fields can transfer angular momentum at a distance on the dynamical timescale.

The Sound Crossing Timescale

Density perturbations in the disk will occur over the sound crossing timescale,

$$\tau_{\text{sound}} = l/v_{\text{s}} = r\sqrt{\frac{\rho}{P}}, \tag{1.18}$$

where we have expressed the sound speed v_s in terms of the density ρ and pressure P of the disk. Here, l can either be the radius, r, for radially propagating waves, or the height of the disk, h, for vertically propagating waves. Compared to $\tau_{\text{dynamical}}$, this timescale is a more realistic measure of the interval over which we are likely to see physical changes in the disk, as this is the timescale over which density perturbations travel through the disk.

The Thermal Timescale

The thermal timescale of the disk is defined as a ratio of internal energy to cooling or heating rate, which can be thought of as the time taken for stored energy to dissipate within the disk. For the simple α-disk model discussed in Section 1.4.1, this ratio is specified by the viscosity parameter α and the dynamical timescale,

$$\tau_{\text{thermal}} \sim \alpha^{-1} \tau_{\text{dynamical}}. \tag{1.19}$$

The Viscous Timescale

Lastly, we have the viscous timescale which is the characteristic timescale of mass flow in the radial direction. It is defined as the ratio of the radius to the inflow velocity. If we assume the α disk model, we can express τ_{viscous} simply as

$$\tau_{\text{viscous}} = \left(\frac{r}{h}\right)^2 \tau_{\text{thermal}}, \tag{1.20}$$

where h is the height of the disk at radius r. For a cold, optically-thick disk, h/r is expected to be so small that the viscous timescale is many orders of magnitude longer than the thermal timescale. Conversely, for super-Eddington flow, or alternatively a hot optically-thin plasma, $h/r \approx 1$ such that the viscous and thermal timescales are equal, although in this regime the infalling material no longer reassembles a disk. For an intermediate regime, a suitable value is $h/r = 0.1$ such that the viscous timescale is 100 times longer than the thermal timescale.

1.4.4 Summary of timescales for a typical black hole

Krolik (1999) gives an overview of these timescales, and how they may be expressed in terms of the angular velocity, Ω, the scale height of the disk h and the efficiency factor α,

$$\begin{aligned}
\tau_{\text{light}} &\sim \frac{r}{c} && \approx 10 \text{ hours}, \\
\tau_{\text{dynamical}} &\sim \frac{1}{\Omega} && \approx \text{ days}, \\
\tau_{\text{sound}} &\sim \frac{r}{v_s} && \approx 10 \text{ years}, \\
\tau_{\text{thermal}} &\sim \frac{1}{\alpha\Omega} && \approx 10^2 \text{ years}, \\
\tau_{\text{viscous}} &\sim \left(\frac{r}{h}\right)^2 \frac{1}{\alpha\Omega} && \approx 10^3 \text{ years}.
\end{aligned}$$

Here, I have evaluated these timescales for a typical supermassive black hole of mass $10^8 M_\odot$ at a radius $R = 50 r_g$ in the accretion disk, where the majority of optical emission is expected to originate.

1.5 Quasar Structure and Variability across the Spectrum

Variability is a fundamental aspect of AGN emission, occurs on many timescales and affects all parts of the spectrum. The majority of emission in different bands is often attributed to different parts of the quasar. The timescales and amplitude of fluctuations differ across the spectral range. In this section, I will discuss variability at different wavelengths and the possible physical mechanisms taking place in different substructures of quasars that are responsible for such fluctuations. This section focuses on variability in the optical/UV, since variability at these rest-frame wavelengths contribute to observed optical variability, which is the focus of this thesis.

1.5.1 Optical and UV Variability

Since optical and UV continuum emission is thought to emanate from the accretion disk, continuum variability is therefore expected to be the result of changes in the disk, such as density or temperature fluctuations, or changes in accretion rate. Optical and UV emission can vary on days to weeks, with larger changes occurring on timescales of years. There is a clear correlation of fluctuation amplitude with wavelength; over timescales of weeks, rms variability of the UV flux can be ~30%, while the corresponding optical flux changes by only ~3% (Lawrence, 2016). However, on the timescale of years, the optical variability can reach a similar amplitude to UV. Variations in the optical continuum and IR seem to track variations in the UV on roughly the light-travel time delays (Lawrence, 2016). An example of typical light-travel times can be found in Table 1.1. The amplitude and timescales of observed variability of quasars, particularly in the optical and UV continuum, pose challenges to the simple disk models described in Section 1.4.1. Lawrence (2012) provides an overview of these challenges. The most relevant of these to this thesis are two issues that are known as the timescale problem and the coordination problem.

Structure	size (R_s)	τ_{light}	$\tau_{\text{dynamical}}$	τ_{sound}	τ_{thermal}	τ_{viscous}
Inner disk	5	1.4 hrs	4 hrs	1 yr	19 days	100 yrs
Optical disk	50	14 hrs	6 days	23 yrs	2 yrs	2200 yrs
BLR	1000	11 days	1.4 yrs	800 yrs	-	-
Torus	10^5	3.1 yrs	1.4 kyrs	350 kyrs	-	-

Table 1.1: Typical timescales for different regions of an AGN with a typical $10^8 M_\odot$ black hole from Lawrence (2016)

The timescale problem concerns the speed at which observed changes in the optical/UV can occur. The timescales outlined in Table 1.1 and their mismatch with observations is a source of contention in the study of quasar variability, particularly in the optical band. Collier & Peterson (2001) show that the flattening of structure functions (discussed later in Section 1.8) of UV light curves imply characteristic timescales between 5 and 94 days. However, timescales of this order are inconsistent with simple disk models which would expect such changes over the viscous timescale (Abramowicz 1991), which is on the order of hundreds to thousands of years at the radius associated with UV emission. Other timescales, outlined in Section 1.4.3, may be responsible for the characteristic

timescale seen by Collier & Peterson (2001).

Another disparity is the coordination problem, which refers to the simultaneity of optical and UV emission. Disk models predict time-lags between variations in the optical and UV over the viscous timescale. This is because density fluctuations that produce primary variations in the optical are expected to propagate inward through the disk, causing secondary variations in the UV at some time later, with an expected high coherence between the primary and secondary signals. While there is clear coherence between wavelengths (see e.g., Dexter & Begelman 2019), cross-correlation studies of continuum bands show UV/optical responding to the far-UV with a lag on the order of hours-days (see e.g., Clavel et al. 1990; McHardy et al. 2014; Horne et al. 2021), once again contrasting sharply with the viscous timescale, which is thousands of years of a typical quasar.

Lawrence (2012) suggests that both the timescale and coordination problems (as well as two other problems) could be solved if some primary emission is reprocessed by dense clouds at $\sim 30 R_s$. In order to reproduce the big blue bump, this primary emission must be high energy, e.g., X-rays or extreme UV (EUV). Lawrence (2012) proposes that this primary emission is EUV radiation emitted from the inner region of the disk, which remains obscured. An alternative theory is that the primary emission is X-rays (see e.g., Shappee et al. 2014). However, Lawrence (2012) argues that the total energy output in the X-rays is not sufficient to account for the observed large amplitude of optical/UV variability.

1.5.2 X-ray Variability

As discussed, variability in the optical is often attributed to variations in accretion rate in the disk. The same is true for the UV and IR bands, except UV emission is generated by the hotter, more central, parts of the disk, while IR is produced in the cooler, further regions and reprocessed by dust on large scales. Quasars also exhibit variability in X-rays, however, the timescale and obscuration effects of these fluctuations indicate that the X-ray emitting region is close to the centre (see e.g., Maiolino et al. 2010; Merloni et al. 2014). A popular model for this emission region is the X-ray corona (Haardt & Maraschi 1991), which is thought to consist of an optically thin plasma above the central part of the disk. Coronal X-rays are then the result of inverse Compton scattering of thermal photons from the disk (see e.g., Merloni 2016; Heckman & Best 2014). The size of the corona and its geometry with respect to the disk is unclear and is an area of active

research (see Reis & Miller 2013 and references therein).

1.5.3 Outer region

Outside the accretion disk is a geometrically thick, dusty region, sometimes referred to as a 'torus', and two emission line regions; the broad-line region and the narrow-line region. The broad-line region (BLR) is a region of high velocity gas at $\sim 1000 R_s$ from the black hole. The high velocity of this gas causes relativistic Doppler broadening of emitted radiation, which is the source of broad emission lines observed in type I AGN and quasars. Broad-line widths range from $\Delta v_{\mathrm{FWHM}} \approx 500\,\mathrm{km\,s^{-1}}$ (only slightly broader than narrow lines) to $\Delta v_{\mathrm{FWHM}} \approx 10^4\,\mathrm{km\,s^{-1}}$, with typical values $\Delta v_{\mathrm{FWHM}} \approx 5000\,\mathrm{km\,s^{-1}}$ (see Peterson 1997). By analysing the line profiles and their widths, we can infer two things: Firstly, gas kinematics in the BLR are determined by the central source through radiation pressure and gravity. Secondly, the BLR reprocesses ionizing UV photons emitted by the continuum source that would otherwise not be observed. The structure of the BLR itself may be determined through reverberation mapping (see Peterson 2006 and references therein); a technique which analyses how the broad emission-line profiles and fluxes respond to changes in the continuum flux. By measuring the time delay of the response, it is possible to infer the structure and kinematics of the region (see e.g., Blandford & McKee 1982; Peterson 1993; Horne et al. 2021)

The narrow-line region (NLR) is extended in radius and lies in a shell between 10pc and 1kpc (Netzer 2013). The NLR is of interest for three reasons. First, the NLR is the largest spatial scale where the ionizing radiation from the central source dominates over other sources. Second, the NLR is the only AGN component which is spatially resolved in the optical, being illuminated non-isotropically by the central source. Finally, the NLR dynamics indicate how AGN are fuelled. Unlike the BLR, electron densities in the NLR are low enough such that many forbidden transitions are allowed. This allows us to measure the electron density and temperature of the gas. While the BLR/NLR are important to understand the full picture of AGN, they constitute a small part of the total output of the AGN in the optical/UV, which is dominated by the continuum. The unified scheme for our current understanding of AGN is shown in Figure 1.3.

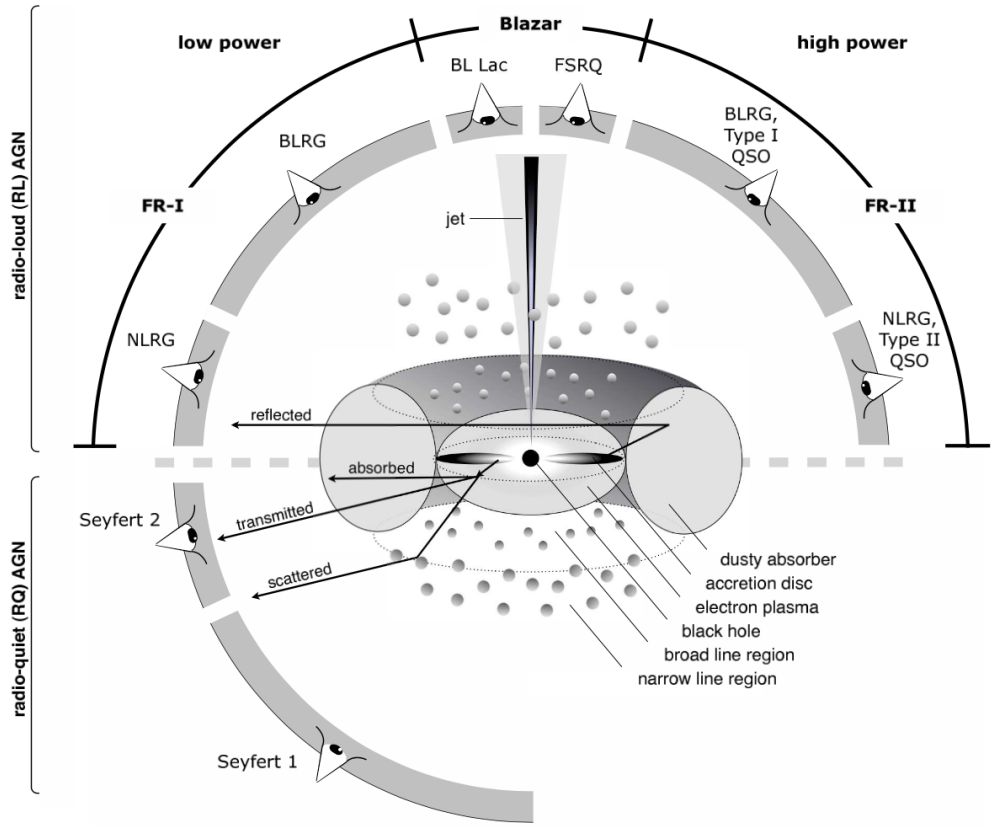

Figure 1.3: Schematic representation of our current understanding of AGN phenomenon in the unified scheme. The type of object we see depends on the viewing angle, whether the AGN produces significant jet emission, and how powerful the central engine is. The objects most relevant to my research are type I QSOs. From Beckmann & Shrader (2012).

1.6 Extremely Variable Quasars

Our understanding of the structure of the quasars has been developed over the past decades to match observations, as described in the previous sections. However, some of these observations, such as the timescale problem, highlight issues with the current unified model. Furthermore, recent observational evidence, particularly in the cases of continuous monitoring and repeat spectroscopy, has put additional strain on current models by exposing quasars with unexpected behaviour. An example of this are quasars classed as extremely variable quasars (EVQs). These quasars show large photometric changes in optical or IR magnitude, typically $|\Delta m| > 1\,\mathrm{mag}$. Closely related to EVQs are changing look quasars (CLQs), which are highly variable and display dramatic changes in

their broad emission lines. The drastic changes that EVQs and CLQs exhibit occur on timescales of years, which presents challenges to the expected timescales for such variations, based on the viscous timescale presented in Table 1.1, which is the minimum timescale over which large variations are expected to occur.

There have been a number of recent studies on EVQs and CLQs which have found interesting correlations in these objects. Rumbaugh et al. (2018) used a sample of ~1000 EVQs (defined by $|\Delta g| > 1$ mag) to show that these objects tend to exhibit stronger broad UV lines. They were also found to be more variable on all timescales and have, on average, a lower Eddington ratio than the general AGN population. After accounting for selection effects in their sample, Rumbaugh et al. (2018) argue that EVQs make up $30 - 50\%$ of the quasar population.

Lawrence et al. (2016) defined a sample of 76 large amplitude transients ($|\Delta m| > 1.5$ mag). These objects, named 'slow-blue hypervariables' were found to be particularly blue on average, and evolve on a timescale of ≈ 10 years, as their name suggests. Lawrence et al. (2016) estimate that these objects make up 1 in 10^5 of the AGN population, after accounting for selection effects.

A small fraction of quasars show large changes in Balmer broad emission lines (BELs) and are called 'changing-look' quasars (CLQs) if, over time, they significantly lose or gain BEL flux. Their spectra change so dramatically they appear to switch between type I and type II Seyfert classes. MacLeod et al. (2016) performed a systematic search for these objects using SDSS spectra. Out of 1011 objects, they found 10 CLQs, four with emerging BELs, five with disappearing BELs, and one with both emerging and disappearing BELs. Observational effects such as dust extinction were ruled out, suggesting that some dramatic physical changes are occurring in the central regions of these quasars. These objects could be an entirely separate class of quasars, or perhaps there is just a low probability of all quasars exhibiting these changes occasionally.

While there will certainly be extreme variables in my 7-DQ quasar sample, I consider them part of the quasar population and are therefore included in my general ensemble studies. While 7-DQ could be used to find new EVQs, these objects are not the focus of my research and will therefore not be studied in detail in this thesis.

1.7 Statistical tools for characterising quasar variability

The main tools for time series analysis fall into either the time domain or the frequency domain. In the frequency domain, Fourier analysis is a common tool, but does not perform well on sparse or irregularly sampled data, which is often the case for astronomical observations. Algorithms like the Lomb-Scargle algorithm (Scargle 1989) are specifically designed for analysing power spectra of unevenly spaced data. This approach involves placing zeros on an evenly sampled grid at points where there are no data in the mean-subtracted light curve, and is effective for identifying periodicities. However, when dealing with noise-like spectra such as those seen in AGN, this approach can distort the true slope of the power spectrum. This distortion arises because a light curve with zeros in place of missing data tends to exhibit more pronounced short-term fluctuations, resulting in an overemphasis of power at shorter time scales. In other words, slower harmonics in the Fourier sum are inadequately weighted due to multiplication by zeros. Alternatively, interpolating the light curve to an evenly sampled grid can be done, but this introduces artificially smoothed variations in the observational gaps, leading to a damping effect on high frequencies in the power spectrum.

A more suitable tool to characterise amplitudes and timescales of variability is the autocorrelation function (ACF), since it is robust to sparse, irregular sampling and does not suffer from aliasing (Giveon et al. 1999). Related to this is the structure function, which is the most common statistical tool for characterising quasar variability.

In this thesis, I primarily use the structure function to analyse variability and look for the clues of characteristic timescales. In particular, I will focus on characterising the long term behaviour of variability, which holds clues to the physical mechanisms governing variability.

1.8 The Structure Function

1.8.1 Overview

Structure functions are a class of functions originally developed for telecommunications engineering in the 1970s (see Lindsey & Chie 1976 and references therein). They provide a robust means to quantify variability of irregular and sparse time series, and became popular with astronomers the following decade. Simonetti et al. (1985) first used structure functions in the context of astronomy, where they applied the first-order structure function to analyse flicker of extragalactic radio sources. Press et al. (1992a,b) later redefined the first-order structure function by introducing a factor of $\frac{1}{2}$,

$$V(\Delta t) \equiv \frac{1}{2}\left\langle \big(y(t+\Delta t) - y(t)\big)^2 \right\rangle \tag{1.21}$$

where $y(t)$ represents a measurement taken at time t, Δt is a fixed lag between the two signals, and the angle brackets denote expectation. In the early 2000s, astronomers started to use the square root of $2V$, referring to it simply as the 'structure function': SF = $\sqrt{2V}$ The earliest example of this appears to be by Hawkins (2002). While there are numerous other definitions of the structure function, this is the one most commonly used in the literature. For consistency with my work I denote it $\mathrm{SF}_{\mathrm{obs}}$;

$$\mathrm{SF}_{\mathrm{obs}}(\Delta t) = \sqrt{2V(y_i, y_j)}, \tag{1.22}$$

where y_i and y_j are measurements (e.g., magnitude or flux) taken at times t, and $t+\Delta t$, respectively. The measurement y is considered to be the sum of a signal, s, plus noise, n; $y = s+n$. I introduce the subscript 'obs' to be explicit that this is the observed structure function, which is a combination of both intrinsic variability and photometric noise. This contrasts to the intrinsic structure function,

$$\mathrm{SF}_{\mathrm{int}}(\Delta t) = \sqrt{2V(s_i, s_j)}, \tag{1.23}$$

where s is the signal. $\mathrm{SF}_{\mathrm{int}}$ is the main quantity of interest, as I will use it to investigate quasar variability on different timescales. Although the signal s cannot be known exactly, I will discuss how $V(s_i, s_j)$ can be estimated in Section 1.8.2.

The structure function has become a popular tool to characterise quasar variability. It is not sensitive to aliasing problems due to discrete or sparse time sampling, making it possible to apply to light curves of astrophysical objects, which are usually irregularly sampled and often have large temporal gaps. Broadly speaking, the structure function $\mathrm{SF}(\Delta t)$ is equal to the rms of the distribution of the magnitude differences $\Delta m = m_j - m_i$ from pairs of observations (t_i, m_i) and (t_j, m_j) separated by a time-lag $\Delta t = t_j - t_i$. For small Δt, the quasar structure function is expected to be small and dominated by photometric noise. We then expect it to increase monotonically with Δt, provided there are no periodic oscillations. The structure function should plateau at arbitrarily large Δt. The converse would suggest that energy output of quasars could wander to arbitrarily large values, which is unphysical.

For a general overview on the origin and application of structure functions to stochastic processes in the context of AGN variability, see Kozłowski (2016b). In Section 1.8.2, I provide a derivation of the structure function from fundamentals, which is an explicit adaptation of the derivation provided by Kozłowski (2016b).

1.8.2 The Structure Function from First Principles

Consider a light curve composed of a collection of measured data y_i (e.g., magnitudes) at times t_i with $i = 1, \ldots, N$ points. This can be represented as a sum of the true signal s_i and noise n_i, $y_i = s_i + n_i$. We are interested in investigating the behaviour and properties of s_i over different observed time intervals $\Delta t_{\mathrm{obs}} = (0, \ldots, t_N - t_1)$ via the autocorrelation function (ACF). The ACF quantifies the relationship between the light curve and copy of itself shifted by Δt and, as we will show, contains the clues to understanding the origin of variability. In the context of quasar variability, s_i is the intrinsic variability of the quasar, n_i is photometric noise and the observed time lags are transformed to the rest-frame time by scaling by $(1 + z)^{-1}$ where z is the redshift of a particular quasar (see Chapter 3, Equation 3.1 for a more detailed explanation of this scaling). We start by first considering the covariance between the signal and shifted copy of itself, $\mathrm{cov}\big(y(t), y(t + \Delta t)\big)$. For brevity, I will write this quantity as $\mathrm{cov}(y_i, y_j)$, such that y_i and y_j are measurements within the same light curve separated by a time lag Δt. If the underlying process generating the signal s is stationary, (its mean, variance and probability distribution do not change with absolute time),

the covariance function is defined as

$$\mathrm{cov}(y_i, y_j) = \left\langle (y_i - \langle y \rangle)(y_j - \langle y \rangle) \right\rangle \tag{1.24}$$

and the covariance for $\Delta t = 0$ days ($i = j$) is

$$\mathrm{cov}(y_i, y_i) \equiv \mathrm{var}(y_i) \equiv \langle y^2 \rangle - \langle y \rangle^2 \tag{1.25}$$

where $\langle y \rangle$ and $\mathrm{var}(y_i)$ are the mean and variance of the data, respectively. We now wish to relate this quantity to the structure function.

The first order structure function $V(y_i, y_j)$, defined in Equation 1.21, may then be defined in terms of the variance and covariance of y,

$$\begin{aligned} V(y_i, y_j) &\equiv \frac{1}{2}\left\langle (y_i - y_j)^2 \right\rangle && (1.26) \\ &= \langle y^2 \rangle && - \langle y_i y_j \rangle \\ &= \langle y^2 \rangle - \langle y \rangle^2 && - \left(\langle y_i y_j \rangle - \langle y \rangle^2 \right) \\ &= \mathrm{var}(y_i) && - \mathrm{cov}(y_i, y_j) \qquad (1.27) \end{aligned}$$

where we have expressed V in similar notation for consistency. In the final line we have identified that $\mathrm{cov}(y_i, y_j) = \langle y_i y_j \rangle - \langle y \rangle^2$, which is not trivial but can be shown from Equation 1.24. By deconstructing V into the variance and covariance of y, it is possible to separate the contributions from the signal and the noise,

$$\begin{aligned} V(y_i, y_j) &= \mathrm{var}(y_i) - \mathrm{cov}(y_i, y_j) \\ &= \mathrm{var}(s_i + n_i) - \mathrm{cov}\big((s_i + n_i), (s_j + n_j)\big) \\ &= \mathrm{var}(s_i) - \mathrm{cov}(s_i, s_j) + \mathrm{var}(n_i) - 2\mathrm{cov}(s_i, n_j) - \mathrm{cov}(n_i, n_j) \\ &= V(s_i, s_j) + \mathrm{var}(n_i), \qquad (1.28) \end{aligned}$$

where $\mathrm{cov}(n_i, n_j) = \mathrm{cov}(s_i, n_j) = 0$ because data points in the signal and noise are both assumed to be uncorrelated with the neighbouring data points in the noise. We have identified the first two terms of the third line as the structure function of the signal, $V(s_i, s_j)$, and we see this may be calculated from Equation 1.28 by subtracting the variance of the noise from the structure function of the measurements,

$$V(s_i, s_j) = V(y_i, y_j) - \mathrm{var}(n_i). \tag{1.29}$$

Combining this result with 1.23 provides an expression for the intrinsic structure function,

$$\mathrm{SF}_{\mathrm{int}}(\Delta t) = \sqrt{2V(y_i, y_j) - 2\mathrm{var}(n_i)}. \tag{1.30}$$

Thus, the intrinsic variability may be estimated by calculating the structure function from the measurements and subtracting excess variance contributed by photometric noise.

To calculate $\mathrm{SF}_{\mathrm{int}}$ in practice, we may combine Equations 1.26 and 1.30,

$$\mathrm{SF}_{\mathrm{int}}(\Delta t) = \sqrt{\frac{1}{M}\sum_{\mathrm{pair}\,k}[(y_{1,p} - y_{2,p})^2 - \sigma_{1,p}^2 - \sigma_{2,p}^2]}, \tag{1.31}$$

where we have $p = (1, \ldots, M)$ unique measurement pairs denoted $(t_{1,p}, y_{1,p}, \sigma_{1,p})$ and $(t_{2,p}, y_{2,p}, \sigma_{2,p})$, separated by the time interval $\Delta t = t_{2,p} - t_{1,p}$. We have approximated $\mathrm{var}(n_i)$ as the sum of photometric variances on each measurement. It is easy to show that a series of N points yields $M = N(N-1)/2$ unique pairs. From this equation we see that the structure function represents the rms of magnitude differences, minus photometric error, of all pairs that are separated by a time Δt. Equation 1.31 is the same result arrived at by Press et al. (1992a) and is deemed the most correct way to estimate SFs by Kozłowski (2016b).

The ACF is defined by the covariance,

$$\mathrm{ACF}(\Delta t) \equiv \frac{\mathrm{cov}(s_i, s_j)}{\sigma_s^2}. \tag{1.32}$$

Using this definition, we may express $\mathrm{SF}_{\mathrm{int}}$ as

$$\mathrm{SF}_{\mathrm{int}}(\Delta t) = \sqrt{2\sigma_s^2\big(1 - \mathrm{ACF}(\Delta t)\big)}. \tag{1.33}$$

The limit $\lim_{t\to\infty} \mathrm{ACF}(\Delta t) = 0$ is a necessary condition for the signal to be real and physical.

1.8.3 Variations of the Structure Function

In this thesis, I use two variations of the structure function, which are also commonly employed by others in the literature. I briefly introduce these variations here to ensure the reader is familiar with them during the review

of variability studies in Section 1.10. However, note that I reserve a detailed discussion of these structure functions for when I utilise them in Chapter 5.

The Ensemble Structure Function

Although one of the main advantages of the structure function is its ability to handle irregular and sparsely sampled data, the structure function for any single quasar for such data is itself sparse. Therefore, a common procedure is to group all Δm measurements from many different quasars. The result is named the 'ensemble structure function' and is identical to $\mathrm{SF}_{\mathrm{int}}$ with the exception of an additional sum,

$$\mathrm{SF}_{\mathrm{int}}(\Delta t) = \sqrt{\frac{1}{N(\Delta t)} \sum_{k} \sum_{j<i} (\Delta m_{ij,k}^2 - \sigma_{i,k}^2 - \sigma_{j,k}^2)}, \tag{1.34}$$

where k denotes the index of the quasar.

Asymmetric Structure Functions

Asymmetries between the brightening and dimming parts of a signal can be investigated by separating the fluctuations into positive and negative changes, computing their respective structure functions and comparing them. This separates the structure function into two parts Kawaguchi et al. (1998):

$$\mathrm{SF}_{-} = \sqrt{\frac{1}{N(\Delta t)} \sum_{j<i} (\Delta m_{ij,-}^2 - \sigma_{i,-}^2 - \sigma_{j,-}^2)} \qquad \text{(brightening)} \tag{1.35}$$

$$\mathrm{SF}_{+} = \sqrt{\frac{1}{N(\Delta t)} \sum_{j<i} (\Delta m_{ij,+}^2 - \sigma_{i,+}^2 - \sigma_{j,+}^2)} \qquad \text{(dimming)} \tag{1.36}$$

where Δm_{ij} have been separated into $\Delta m_{ij,-}$ and $\Delta m_{ij,+}$ for brightening and dimming observations respectively.

1.9 Stochastic Models of Quasar Variability

Extracting physically meaningful and interpretable information from temporal data sets is the crux of the time-domain data analysis problem. In certain

cases, extracting such information is moderately straightforward. For example, the variability of periodic signals, such as eclipsing binaries or variable stars, may be characterised using tools such as the periodogram. The period, or spectrum of periods, may then be used to understand the physics of the system. Another example is in the analysis of relatively well-defined transient phenomena, such as supernovae and black hole tidal disruption events. Such signals can be parameterised by fitting a physically-motivated, deterministic model to the data. Some (or all) of the fitted parameters can then be directly related to real, measurable properties such as redshift or mass.

Conversely, the analysis of quasar variability does not fit into such straightforward cases. Variability exhibited by quasars across the spectrum is non-periodic, stochastic and changes over the timescale which it is observed. Currently, no physically-motivated models have been developed that are capable of parameterising the physical processes taking place in the accretion disk and surrounding regions of quasars. The primary reason for this is that the mechanisms behind these physical processes are complex and still poorly understood.

The problem of fitting models to quasar light curves to obtain meaningful information is non-trivial. However, some recent studies have used non-deterministic, stochastic models to provide simplified parameters which act as a proxy to the real, physical mechanisms responsible for producing variability. In the context of quasar variability, the most popular choice of such models are those that fall in the Continuous Autoregressive Moving Average (CARMA) family. The primary strength of CARMA processes lie in their ability to account for irregular sampling and measurement errors, making them very suitable for quasar light curves and astronomical datasets in general.

CARMA models are the continuous version of ARMA models, which themselves are comprised of two components; autoregressive (AR) and moving average (MA) processes. In the remainder of this section, I will give a brief overview of the AR, MA processes and how together they form CARMA models of different orders. I reserve a detailed discussion of one such CARMA model, known as the damped random walk, for Section 1.9.3. This model deserves a thorough explanation, as I utilise it for simulation and analysis in Chapters 5 and 6. Additionally, it is ubiquitous in recent studies of quasar variability in the optical/UV, being the most popular stochastic model in the literature. Some of these studies and their results are also discussed in Section 1.9.3.

1.9.1 CARMA processes from First Principles

As discussed, a CARMA process is the combination of three concepts. The first is the autoregressive process, the second is the moving average process, and the third is the generalisation to continuous-time. Here, I describe the basic principles of these models. Lawrence (2019) and Kelly et al. (2014) provide overviews of much of this material and are useful as general references.

Autoregressive (AR) processes

An autoregressive process of order p, denoted AR(p), is a representation of a type of random process. The output of the process at a given step, x_t, depends on a linear combination of previous values in the series, plus an independent random variable,

$$x_t = c + \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \ldots + \alpha_p x_{t-p} + \epsilon_t, \tag{1.37}$$

where ϵ_t is generated by sampling from a random white noise process with zero mean. Note that AR(1) with $\alpha_1 = 1$ gives the standard random walk.

Moving Average (MA) processes

Another key process is the moving average, denoted MA(q), where q is the order of the process. It is written as:

$$x_t = c + \beta_0 \epsilon_t + \beta_1 \epsilon_{t-1} + \ldots + \beta_q \epsilon_{t-q}. \tag{1.38}$$

Again, ϵ_t is sampled from a zero-mean white noise process. The MA process may be interpreted as a random seed passed through a linear response filter β_i. Note that AR(1) is equivalent to an MA process with exponential filter (see e.g., Lawrence 2019).

Autoregressive Moving Average (ARMA)

We can combine the two concepts above into the Autoregressive Moving Average (ARMA). We retain the autoregressive order p and moving average order q such

that the ARMA process has order (p, q). It is denoted as ARMA(p, q) and takes the form:

$$x_t = \epsilon_t + \sum_{i=1}^{p} \alpha_i x_{t-i} + \sum_{i=1}^{q} \beta_i \epsilon_{t-i} \tag{1.39}$$

Continuous Autoregressive Moving Average (CARMA)

Finally, we can transform the discretised ARMA process into a continuous version by representing it as a differential equation. Written as CARMA(p, q), it takes the form:

$$\begin{aligned} &\frac{d^p x(t)}{dt^p} + \alpha_{p-1} \frac{d^{p-1} x(t)}{dt^{p-1}} + \cdots + \alpha_0 x(t) \\ &= \beta_q \frac{d^q \epsilon(t)}{dt^q} + \beta_{q-1} \frac{d^{q-1} \epsilon(t)}{dt^{q-1}} + \cdots + \epsilon(t). \end{aligned} \tag{1.40}$$

Note that this equation has been split and aligned for readability, but should be interpreted as one single equation. In addition, I have fixed $\alpha_p = 1$ and $\beta_0 = 1$ for convenience without losing generality of the model. The parameters $\alpha_0, \ldots, \alpha_{p-1}$ are the autoregressive coefficients, and the parameters $\beta_1, \ldots, \beta_{p-1}$ are the moving average coefficients. Further details and mathematical properties of the CARMA process can be found in e.g., Jones (1981); Jones & Ackerson (1990); Brockwell (2001); Roux (2002); Koen (2005); Feigelson et al. (2018).

CARMA models as Gaussian Processes

The stochastic term, $\epsilon(t)$, in Equation 1.40, is responsible for variability in the output, $x(t)$. The CARMA process in its most general form requires only $\epsilon(t)$ to be a white noise process with zero mean, but imposes no constraints on the distribution from which it is drawn. In the special case that $\epsilon(t)$ is drawn from a Gaussian, the resulting CARMA process is known as a Gaussian process (GP). GPs are specified in terms of a kernel (also known as a covariance function), equivalent to the distribution of β_q in a CARMA process (see e.g., Yu et al. 2022; Aigrain & Foreman-Mackey 2023). This important relation between CARMA processes and GPs is not always noted in the literature. The use of GPs in astronomy is a recent emergence, and they have proven to be a powerful tool in the analysis of astronomical time series. The damped random walk, discussed in

Section 1.9.3, is actually a GP with an exponential kernel. For a general overview of GPs in the context of astronomy, see Aigrain & Foreman-Mackey (2023).

1.9.2 Modelling Quasar Variability as Higher Order CARMA Processes

Higher order CARMA models have also been used to model quasar light curves (see e.g., Yu et al. 2022; Kasliwal et al. 2017; Sheng et al. 2022). The motivation for this is that higher order CARMA models have a flexible power spectral density (PSD) and can therefore be used to model time series generated by power spectra of a wide range of shapes, including those that exhibit quasi-periodicity. However, using CARMA models that are too complex often suffers overfitting and a loss of interpretability. Therefore, in this thesis, I will only be using the lowest order CARMA process, CARMA$(1, 0)$, described in the following section.

1.9.3 The Damped Random Walk

The expression for lowest order CARMA process, i.e., CARMA$(1, 0)$, may be obtained from 1.40 by setting $p = 1$ and $q = 0$,

$$\frac{dx}{dt} + \alpha_1 x = \beta_0 \epsilon(t) \tag{1.41}$$

The value of the parameter α_1 determines the behaviour of the process:

$$\begin{aligned}
\alpha_1 < -1 &: \quad \text{the process diverges rapidly} \\
\alpha_1 = -1 &: \quad \text{simple random walk} \\
-1 < \alpha_1 < 0 &: \quad \text{damped random walk} \\
\alpha_1 = 0 &: \quad \text{random process determined by } \epsilon(t) \\
\alpha_1 > 0 &: \quad \text{the process oscillates}
\end{aligned}$$

The case of the damped random walk, defined for $-1 < \alpha_1 < 0$, has particular importance in the context of modelling quasar variability. Note that this process is often referred to by several names. The most common is the damped random walk (DRW), and is how I will refer to it hereafter. However, it is also often referred to in the literature as CARMA$(1, 0)$ and CAR(1) to emphasise its place within the CARMA hierarchy. Additionally, it is also referred to as the Ornstein–Uhlenbeck

(OU) process, though this is more common in the physics literature.

Gaussian processes with an exponential kernel fall into a special class of stochastic models, such that their computational complexity only scales linearly with the number of points in the light curve (i.e., $O(n)$) (see e.g., Rybicki & Press 1995; Kozłowski et al. 2010; Foreman-Mackey et al. 2017b). This allows for fast and efficient computation of the likelihood function, making these class of models more scalable and practical to work with compared to general CARMA models. The DRW is one such process that has fast algorithms for computing its likelihood function. Since this thesis involves working with a massive photometric dataset with $\sim 10^5$ light curves, scalability and computational efficiency are necessary attributes of models that I use; another reason why I decided to use the DRW model for my analysis in Chapter 6.

Modelling Quasar Variability as a Damped Random Walk

The damped random walk (DRW) is widely utilised to model quasar light curves. The initial application of this model to quasar light curves was conducted by Kelly et al. (2009). Subsequent studies have shown that it is a viable description of optical continuum variability on timescales > 1 year (MacLeod et al. 2010, 2012; Zu et al. 2013; Andrae et al. 2013).

The basic assumption of the DRW is that variability may be described as a random walk with a driving amplitude $\sigma_{\rm DRW}$ and damping timescale $\tau_{\rm DRW}$ (equivalent to β_0 and α_1 in Equation 1.41, respectively). Physically, $\tau_{\rm DRW}$ corresponds to the characteristic timescale for the process to 'forget' about previous perturbations, while $\sigma_{\rm DRW}$ corresponds to the amplitude of long timescale variations. This process behaves as a standard random walk for short timescales and asymptotically reaches a finite variability amplitude at long timescales.

The motivation of using such a model is to parameterise variability and obtain parameters ($\tau_{\rm DRW}$ and $\sigma_{\rm DRW}$) which act as a proxy for accretion mechanisms. Studies have shown that these parameters correlate with quasar properties such as luminosity and black hole mass (see e.g., MacLeod et al. 2010; Kelly et al. 2011; Burke et al. 2021).

Although it is possible to fit more complex CARMA models to quasar light curves, the simplicity of the DRW model enables me to better interpret fitted parameters.

Additionally, the popularity of the model allows me to compare my results with others. In this thesis, I use the DRW process in two ways. First, I simulate quasar light curves using a DRW to make comparisons against observational data in Chapter 5. Second, I fit DRW models to quasar light curves to obtain amplitude and characteristic parameters, which are correlated against quasar properties in Chapter 6.

Properties of the Damped Random Walk

The ACF of the DRW is given by

$$\mathrm{ACF}(\Delta t) = e^{-|\Delta t|/\tau}. \tag{1.42}$$

Using Equation 1.33, the SF of the DRW is therefore

$$\mathrm{SF}(\Delta t) = \mathrm{SF}_\infty (1 - e^{-\Delta t/\tau})^{1/2}, \tag{1.43}$$

where we adopt the notation $\mathrm{SF}_\infty = \sqrt{2}\sigma$ used by MacLeod et al. (2010) to explicitly highlight the long timescale behaviour of the DRW SF. The asymptotic values of the DRW SF for small and large Δt are

$$\mathrm{SF}(\Delta t \ll \tau) = \frac{\mathrm{SF}_\infty}{\sqrt{\tau}} \Delta t^{1/2}, \tag{1.44}$$

$$\mathrm{SF}(\Delta t \gg \tau) \equiv \mathrm{SF}_\infty = \sqrt{2}\sigma. \tag{1.45}$$

Thus, for small time-lags, $\mathrm{SF} \propto \Delta t^{1/2}$, which is often quoted as a 'slope of 0.5' (i.e., the visual slope on a log-log plot). In the short time-lag regime, the DRW is equivalent to an ordinary random walk, as the variations are dominated by Gaussian noise. Conversely, for long time-lags, the DRW SF plateaus to a constant value. This is the 'damped' aspect of the DRW and prevents the process from 'walking' to arbitrarily large values. This is physically motivated, since we *generally* do not see consistent drifts in apparent magnitude of quasars over the timescales which we have observed them (although this is a contentious issue and explored in more detailed in Chapter 4, Section 4.4.1). In Figure 1.4, I show an example of two DRW processes with different SF_∞ and τ, along with their respective structure functions.

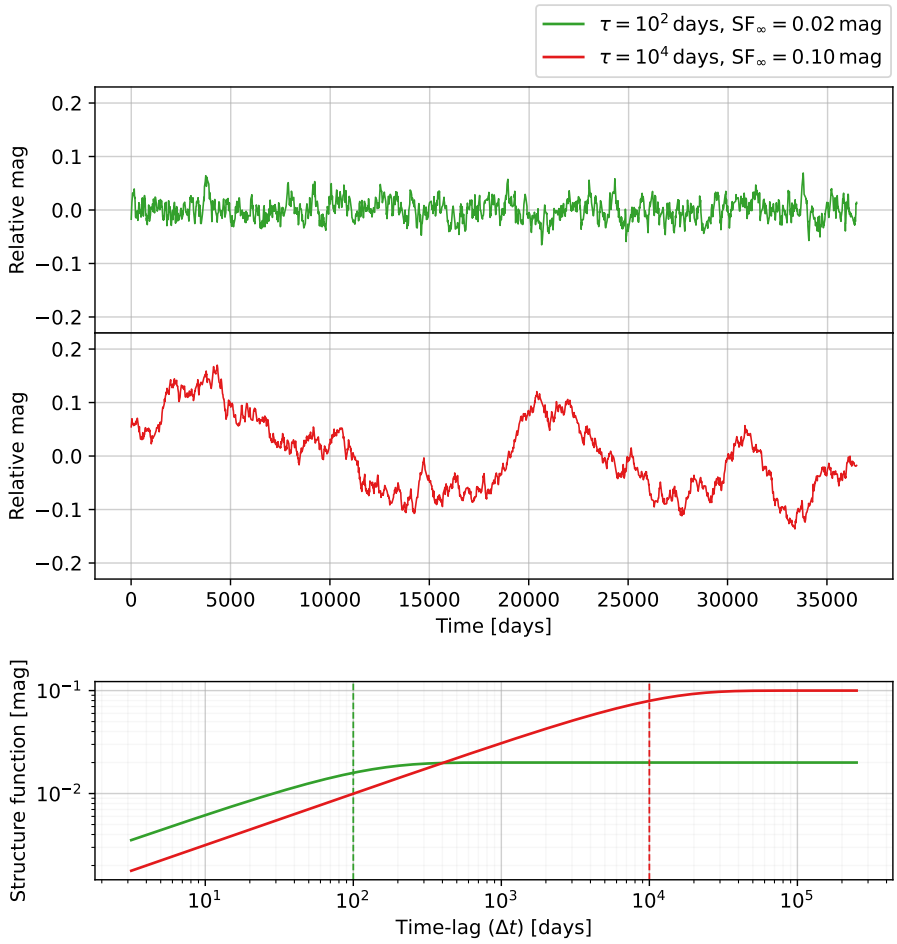

Figure 1.4: Top two panels: examples of two DRW processes with two different sets of SF_∞ and τ. Bottom panel: The respective structure functions of the two DRW processes. The characteristic timescales, τ, mark the position of the turnover in the structure function plot (dotted lines). This plot shows the long and short time-lag regimes clearly; note the 0.5 slope on short time-lags, and the plateu at SF_∞.

1.10 Massive variability studies

In recent years, astronomical data sets have undergone an exceptional and ongoing expansion in volume, quality, and complexity, driven by advancements in telescope, detector, and computer technologies. Similar to various other scientific disciplines, astronomy has become a very data rich science. This has enabled astronomers to carry out massive variability studies.

The term 'massive variability studies' refers to the practice of investigating

the variability of a large sample of astrophysical objects using photometry or spectroscopy collected from one or more sky surveys. Such a study would generally involve a wealth of data, which may be used to discover patterns and correlations between physical properties and statistics of variability calculated from the photometry.

In this thesis, I outline the construction and analysis of the largest dataset of quasar photometry to date. In this section, I discuss the history of photometric studies of quasar variability on a large scale to provide the reader with an overview of recent efforts in this field. Note that I will be focussing on the sample size, baseline and surveys used. I will reserve a detailed discussion of results from these papers to the discussion sections of Chapters 4, 5 and 6.

Many of these variability studies use the structure function to analyse quasar variability. However, the observed shape of the quasar structure function remains a contentious issue, and many of these studies have tried to constrain its form. Therefore, I will quote fits to the structure function explicitly where possible.

The first large-scale variability studies of quasar variability used repeat observations from photographic plates. One such study was carried out by Hook et al. (1994), utilising a sample of 332 quasars with photometry from 12 plates from the UKST, spanning 16 years. They were the first to use structure function analysis to characterise quasar variability. Additionally, this was the first comprehensive study of optically selected quasars to disentangle the dependence of variability on luminosity/redshift; they revealed that quasars of high luminosity exhibit significantly less variability than those with low luminosity, a trend that has been confirmed by many since.

Although the results of Hook et al. (1994) were a significant step in understanding quasar variability, the concept of massive variability studies with repeat imaging was pioneered by Hawkins and his ~200 plates of UK Schmidt Field 287 (Hawkins 1983). With these plates, Hawkins (2002) obtained ~4 measurements per object per year, over 24 years, in a sample of 400 quasars. An ensemble structure function analysis led Hawkins to conclude that microlensing is the likely cause of variability, supported by the fact that variability timescales did not show a correlation with redshift, which would be expected from cosmological time dilation. This result was quite controversial and has not been widely accepted as the source of quasar variability.

The study carried out by de Vries et al. (2003) was a milestone for massive

variability studies, as they were able to extend the temporal baseline to 50 years by combining photographic plate data and photometry from the SDSS early data release. This study was of particular importance as it was the first application of combining data from multiple surveys. Their sample consisted of 3,791 quasars, which was unprecedented at the time. By carrying out ensemble structure function analysis, they found that their results were consistent with results obtained from continuous monitoring of smaller samples of quasars, validating the application of multi-survey data sets for studying quasar variability. de Vries et al. (2005) followed on from their previous study (de Vries et al. 2003), using SDSS DR2 to extend the sample to 31,165 quasars over the same baseline of 50 years.

The first data release of SDSS, and the corresponding quasar catalogue produced by Schneider et al. (2003), enabled quasar samples to reach sizes of $\sim 10^4$. Vanden Berk et al. (2004) combined objects from this catalogue as well as matched FIRST radio sources (Stoughton et al. 2002) to create a sample of 25,710 quasars. However, their baseline was limited to 2 years as they did not include photometry from earlier surveys.

Subsequent SDSS data releases have resulted in larger sample sizes and longer baselines (see e.g., de Vries et al. (2005)). Additionally, quasar samples used in recent studies are often defined using the SDSS Quasar Catalogue. This is because SDSS provide spectra of its targets, and therefore quasars in this catalogue are spectroscopically-confirmed. The SDSS Quasar Catalogue is current in its 16^{th} edition (Lyke et al. 2020) consisting of an impressive 750,414 quasars. Additionally, recent large surveys such as Pan-STARRS, ZTF, DECaLS, HSC and DES have led to a wealth of accessible photometry. Combining data from multiple surveys is advantageous in two ways; more measurements per object enables one to better constrain statistics of variability, while a longer temporal baseline allows one to probe quasar variability on these long timescales. Examples of this include the study carried out by Morganson et al. (2014), which involved a cross-comparison of SDSS and Pan-STARRS data in a sample of 105,783 quasars over 10 years. Another example can be seen in Li et al. (2018), which is the largest variability study to date using a sample of 119,305 quasars over 15 years, using data from both SDSS and DECaLS. Table 1.2 presents a summary of several of thesse studies, focussing on the sample size and baseline.

Authors	Survey/Bands	$N_{\rm obj}$	Δt [yr] observer frame
Pica & Smith 1983	Photographic P & B	130	13
Hook et al. 1994	Photographic J	332	16
Giveon et al. 1999	B & R bands	42	7
Hawkins 2002	Schmidt plates U, B, V, R, I	400	24
de Vries et al. 2005	POSS SDSS g	31,165	50
Vanden Berk et al. 2004	SDSS g, r, i	~25,000	2
Voevodkin 2011	SDSS g	7,562	10
MacLeod et al. 2012	SDSS u, g, r, i, z	33,881	20
Morganson et al. 2014	SDSS, PS g, r, i, z	~100,000	10
Li et al. 2018	SDSS, PS g, r, z	119,305	15

Table 1.2: A summary of quasar variability studies, highlighting the sample size and baseline used for each study.

1.11 Challenges, open questions, and limitations of current studies

Both the current unified model of AGN (see e.g., Netzer 2015 for an overview) and stochastic models face many challenges presented by observations of variability in the optical and UV continuum of quasars. Currently, these models struggle to

explain observed phenomena.

In this section, I will summarise these challenges and highlight open questions. I will then briefly discuss limitations of current studies, which motivate my work.

i) Model discrepancies

Discrepancies between observations and model predictions pose prominent challenges to the unified model. Standard accretion disk models such as the α-disk are a constituent part of the unified model and face many of the same challenges. Some of these challenges include the timescale and coordination problems (discussed in Section 1.5.1) and evidence of extreme variability (see Section 1.6).

Additionally, stochastic models have been shown to have some limitations in accurately describing variability. The application of stochastic models to characterise variability is a source of contention, with some studies claiming that it accurately represents fluctuations in the disk, and others stating that these models suffer from degeneracies and biases. Furthermore, as discussed in Section 1.9, higher order CARMA models are less interpretable.

ii) Open questions

The disparities mentioned in **(i)** lead to a multitude of open questions. Quasar variability is still poorly understood because many of these questions are contentious and unresolved. This is due to both insufficient data used in studies, and a lack of understanding of the physical processes governing variability. Some of these open questions have been addressed by certain studies, while other questions remain completely unexplored. I present a list of these questions, restricting to those which I aim to answer in this thesis:

1. What is the primary driver of optical/UV variability on timescales of years, given that basic models expect this variability to occur on timescales of thousands of years?

2. What is the behaviour of variability on the longest possible timescale that we can currently observe (~70yrs)?

3. How does variability vary with quasar properties, such as luminosity, black hole mass and Eddington ratio?

4. What are the ensemble variability statistics of quasars in larger samples and longer baselines than those used by previous studies?

5. To what degree does the DRW provide an accurate description of variability?

6. To what extent do DRW parameters correlate with quasar properties?

iii) Limitations of current studies

Much of our lack of understanding of quasar variability is due to not being able to characterise variability on a wide range of timescales, including the longest possible timescales with current observations. Additionally, it is not clear how variability depend on quasar properties, and the strength of that dependence with differing timescales. However, significant progress can be made by increasing sample size, temporal baseline, and expanding the number of observations per object using existing data.

Recent studies have been limited to ~100,000 quasars over a ~50 year baseline. These limitations are somewhat self-imposed, despite the abundance of publicly accessible data. Therefore, in this thesis, I aim to answer some of the questions mentioned in **(ii)** by carrying out massive variability studies on a large database of quasar photometry. A substantial part of this thesis involves the creation of this database, which is unprecedented in sample size, temporal baseline, and number of observations. This database will not only be used for my studies, but will also benefit the wider scientific community that are researching quasar variability. This will help answer many unresolved questions and pave the way to understanding these enigmatic objects.

1.12 Thesis outline

In this thesis, I explore the topics discussed above in different projects. In Chapter 2, I outline the creation of my photometric database, 7-DQ. In Chapter 3, I describe the computational methods and algorithms used in my data reduction pipeline. In Chapter 4, I explore basic statistical properties of my

7-DQ database and its preprocessed derivatives. In Chapter 5, I calculate the structure function, and variations of it, to investigate quasar variability and characterise its relationship with physical properties. In Chapter 6, I explore and employ modelling techniques to obtain parameters which I correlate with quasar properties. Finally, I draw conclusions from my work in Chapter 7.

Chapter 2

Data sources, samples and colour transformations

2.1 Introduction

This chapter details the steps I took in the creation of the 7-Decades of Quasar light curves database (referred to as 7-DQ hereafter), a comprehensive compilation of photometric data from quasars, and a control sample of stars, sourced from multiple surveys, spanning 70 years in the observer's frame. The extended timescales presented here are unprecedented, as previous large-scale photometric studies have typically been limited to 50 years. Notably, both the sample size and the number of observations in the 7-DQ database exceed those of recent efforts (see Chapter 1, Section 1.10 for an overview of other large photometric datasets of quasar photometry). 7-DQ forms the basis for my analyses presented in subsequent chapters. The extensive temporal coverage enables me to investigate quasar variability further into the time domain, while the substantial sample size enables a detailed exploration of how variability statistics change across the population, achieved through the categorisation of the population into subensembles (i.e., subgroups). Furthermore, the large number of photometric observations in the 7-DQ database strengthens the robustness of my findings, imposing tighter constraints than prior studies and allows me to tease out subtle trends within the quasar population. The reader should note that while 7-DQ was designed to investigate quasar variability, it refers to the full dataset of photometry from both my quasar and star samples (which are defined later in

Section 2.3). Therefore, I will refer to my sample of quasars, and the photometry I have collected of them in 7-DQ, as '7-DQ quasars'. Similarly, I will refer to my control sample of stars and their corresponding photometry as '7-DQ stars'. I will refer to the two populations in this way hereafter to avoid confusion. The structure of the remainder of this chapter is as follows: in Section 2.2 I outline the surveys whose photometry I use to construct my database. In Section 2.3 I present the sample definitions for the quasar and star population. In Section 2.4 I describe the process of obtaining the photometry from the respective surveys and illustrate some sample data. Finally, each survey uses different instrumentation and filters, and therefore the magnitudes must be transformed into the same photometric system, described in Section 2.5.

2.2 All Sky Surveys

In this section, I describe four distinct all sky surveys whose photometry I use in the construction of my 7-DQ database. These four surveys are; the Sloan Digital Sky Survey, Pan-STARRS1, the Zwicky Transient Facility and the SuperCOSMOS Sky Surveys, described in the following subsections.

2.2.1 Sloan Digital Sky Survey

The Sloan Digital Sky Survey (SDSS; York et al. 2000) uses the 2.5 m Sloan Foundation Telescope (Gunn et al. 2006), which has become one of the most productive observatories in the world, producing annual data releases and populating databases used by thousands of astronomers worldwide (Raddick et al. 2014a,b). Photometric observations of over 450 million unique, primary sources across ∼14,000 deg^2 in five optical filters have been taken as of the Sixteenth Data Release (DR16; Ahumada et al. 2020). The SDSS magnitude scale, described by Fukugita et al. (1996), is based on the original AB_ν system (Oke 1969, unpublished), which was updated to the AB_{95} system by Oke & Gunn (1983). A comprehensive calibration of the SDSS photometric system using a set of standard stars has been carried out by Smith et al. (2002). For the remainder of this chapter, when comparing bands of different surveys, I denote photometry of the *ugriz* bands in this system as $u_{\rm SDSS}$, $g_{\rm SDSS}$, $r_{\rm SDSS}$, $i_{\rm SDSS}$ and $z_{\rm SDSS}$. The same applies to the subscripts of each of the other surveys in this chapter.

The imaging survey is taken in drift-scan mode; the camera continually sweeps the sky in great circles, and a given point on the sky passes through the five filters in succession, resulting in near-simultaneous imaging of the five *ugriz* filters. Photometry from SDSS has an impressive signal-to-noise, enabling reliable observations at faint magnitudes. This is quantified by the 5σ limiting magnitudes, which are summarised in Table 2.1. The imaging system has dedicated astrometric arrays which have been shown to achieve an RMS astrometric accuracy of 150 milliarcseconds in each coordinate (Becker et al. 1995, Skrutskie 1999).

The Sloan Digital Sky Survey is a core component of my multi-survey database for a few reasons. Firstly, it is used to produce the Data Release 14 Quasar catalogue (DR14Q; Pâris et al. 2018); a low-contamination (0.5%) quasar catalogue made possible by SDSS's high quality optical spectra. I define my quasar sample based on DR14Q in Section 2.3. Secondly, the near-simultaneous imaging in each of the five *ugriz* bands enables me to compute colours of each object at a single epoch. Since quasars are variable, it is not guaranteed that measurements in two bands taken at different times would yield the correct colour. Accurate colours are a necessity when applying photometric corrections to transform magnitudes into the same photometric system, discussed later in Section 2.5. Thirdly, the precision astrometry provides accurate RA/Dec coordinates of the objects in my sample, which I use to cross-match objects between SDSS and the other three surveys. Furthermore, SDSS's impressive imaging depth supplies excellent photometry by virtue of its small photometric uncertainties. Finally, repeated imaging of specific regions of sky, e.g., Stripe 82, delivers high cadence imaging for a portion of my objects.

Filter	5σ depth			
	SDSS	Pan-STARRS1	ZTF	SuperCOSMOS
u	22.15	-	-	-
g	23.13	22.0	20.8	21.2
r	22.70	21.8	20.6	20.3
i	22.20	21.5	19.9	18.9
z	20.71	20.9	-	-
y	-	19.7	-	-

Table 2.1: Single-epoch 5σ imaging depths for SDSS, Pan-STARRS1, ZTF, and SuperCOSMOS in *ugrizy* bands. Note that, since each plate in SuperCOSMOS has a different limiting magnitude, the SuperCOSMOS depths represent an average over the plate-emulsion combinations corresponding to each *gri* band.

2.2.2 Pan-STARRS1

The Panoramic Survey Telescope and Rapid Response System 1 (Pan-STARRS1, PS1, PS; Chambers et al. 2016) is the first telescope from Pan-STARRS; an innovative wide-field astronomical imaging and data processing facility. Pan-STARRS1 is a 1.8 m telescope situated at the Haleakala Observatories in Hawaii. It is equipped with a 1.4 gigapixel camera with a pixel size of 10 μm such that each pixel subtends 0.258″ on the sky. Pan-STARRS1 has conducted a set of distinct synoptic imaging sky surveys, the most prominent of which is the 3π Steradian Survey, capturing photometry for 3 billion astronomical objects across 22 million CCD images. Data from this survey have been released in two public releases; Data Release 1 (DR1) and Data Release 2 (DR2). DR1 contains stacked images, while DR2 contains individual epoch data. Since DR2 includes all the data from DR1, it contains an impressive 1.6 petabytes of data, making it the largest volume of astronomical information ever released (Flewelling et al. 2020). Pan-STARRS1 is capable of capturing imagery over 6000 square degrees of the sky every night. Objects are therefore imaged frequently, making it ideal for studying quasar variability.

The Pan-STARRS1 photometric system uses monochromatic *AB* magnitudes (Oke & Gunn 1983) as described in Tonry et al. (2012). Observations are taken through a set of five broadband filters, denoted as $g_{\rm P1}$, $r_{\rm P1}$, $i_{\rm P1}$, $z_{\rm P1}$ and $y_{\rm P1}$. These filters are similar to those used by SDSS, since they both adopt the *AB* system, but they are not identical. Transformations between these two systems are provided by Tonry et al. (2012) and further discussed in Section 2.5. The 5σ limiting magnitudes are summarised in Table 2.1. Magnier et al. (2020) re-calibrated Pan-STARRS1 astrometry using Gaia revealing that the 1σ uncertainty of the median residuals (ΔRA, ΔDec) are (3.1, 4.8) milliarcsec, which is more than sufficient for my needs. Pan-STARRS1 is another fundamental constituent of my 7-DQ database. It delivers high quality photometry on a par with SDSS with a higher cadence and provides the photometric system which I transform all other surveys to (described in detail in Section 2.5).

2.2.3 Zwicky Transient Facility

The Zwicky Transient Facility (ZTF; Bellm et al. 2019; Masci et al. 2019) is an optical time-domain survey and a successor to the Palomar Transient Factory.

ZTF employs a wide-field 600 megapixel CCD camera which utilises the entire focal plane ($\sim$47 deg^2) of the Palomar 48 inch Schmidt telescope, providing the largest instantaneous field-of-view of any camera on a telescope of aperture greater than half a meter in diameter. The camera has a pixel size of 15 μm, providing a pixel scale of 1″ per pixel. ZTF uses a complement of three custom filters, ZTF-g, ZTF-r, ZTF-i, which I denote as $g_{\rm ZTF}$, $r_{\rm ZTF}$, $i_{\rm ZTF}$ for consistency. Note that only one third of the i-band data is publicly accessible. These filters exist within the ZTF photometric system, which differs from the AB photometric system that SDSS and PS1 use. However, using a subset of photometric calibrator stars from the Pan-STARRS1 catalogue, each observation is assigned its own zeropoint (ZP_f) and colour coefficient term (c_f). The ZTF Science Data System (ZSDS) Explanatory Supplement (Masci et al., 2019) describes transformations of ZTF photometry to the Pan-STARRS1 photometric system. ZTF is an ongoing mission, with data still being collected. The current public release is Data Release 21 which includes observations up until 1 March 2024.

Photometry from ZTF constitutes the majority of the data within 7-DQ, making up 92% and 64% of the quasar and star photometry respectively. This is due to its impressive cadence of scanning the entire Northern sky every two days. This cadence holds the key to understanding short term variability and catching rapid, short-lived outbursts. Although ZTF photometry has slightly lower signal-to-noise compared to SDSS and Pan-STARRS1, the sheer number of observations and use of optimal averages mean that it does not compromise the overall quality of my 7-DQ database. The 5σ limiting magnitudes are summarised in Table 2.1.

2.2.4 SuperCOSMOS Sky Survey

The SuperCOSMOS Sky Survey (SuperCOSMOS; SSS; Hambly et al. 2001a,b,c) programme was a project aimed at digitising the entire sky in three bands via automatic scans of sky atlas photographic plates. SuperCOSMOS combines photographic plates from several subsurveys, listed in Table 2.2. Note that I use the term 'subsurvey' to avoid confusion with other surveys in this section. The three bands of the photographic plates are broadly blue, red and near-IR. The exact profiles of the bands depend on the combination of filter and emulsion, which differs between subsurvey. These combinations are explained in detail in the references listed in Table 2.2. Typically, these combinations are given a single symbol to represent the filter and emulsion, explained as follows: B_J corresponds

to the blue band, using a Kodak IIIaJ emulsion and Schott GG395 filter. R_F corresponds to the red band, using the Kodak IIIaF emulsion and Schott OG590 filter. E also corresponds to the red band, using the Kodak 103a-E emulsion and Plexiglas 2444 filter. Finally, I_N corresponds to the near-infrared band, using the Kodak IV-N emulsion and Schott RG715 filter. These blue, red and near-infrared bands correspond roughly to the g, r, and i bands used by SDSS, Pan-STARRS and ZTF.

The original photographic material comes from three Schmidt telescopes; the UK Schmidt Telescope (UKST), European Southern Observatory (ESO) Schmidt Telescope, and the Palomar Schmidt Telescope. Collectively, these three telescopes conducted 9 individual subsurveys between 1949 and 2002, each with differing filters, sky coverage, and epochs. Each subsurvey produced a set of photographic plates which were scanned using the SuperCOSMOS machine, a high-precision plate scanning facility with 0.67 arcsecond per pixel resolution operated at Edinburgh. The astrometry is precise to (0.2,0.3) arcseconds in RA and Dec, respectively.

The survey imaged the entire sky once in each of the three bands, with the exception of the red band, which contains two epochs in the Palomar subsurvey. Generally, the UKST and ESO provided photometry for the Southern Hemisphere, while the Palomar Oschin telescope provided photometry for the Northern Hemisphere. Although the SuperCOSMOS Sky Survey has the lowest signal-to-noise of the four surveys, its power is realised when calculating variability statistics on long temporal baselines; these historical observations act as a lever-arm and are the key to understanding variability on unprecedented timescales of ~70 years.

Table 2.2: Summary of individual surveys which contribute to the full SuperCOSMOS Sky Survey. Adapted from Hambly et al. (2001a).

ID	Subsurvey	Telescope	Dec. range	Band	Dates of observation	Reference
1	SERC-J/EJ	UKST	$-90 \rightarrow +3$	B_J	1974—1994	Cannon (1984)
2	SERC-R/AAO-R	UKST	$-90 \rightarrow +3$	R_F	1984—2000	Cannon (1984) and Morgan et al. (1992)
3	SERC-I	UKST	$-90 \rightarrow +3$	I_N	1978—2003	Hartley & Dawe (1981)
4	ESO-R	ESO Schmidt	$-90 \rightarrow -17$	R_F	1978—1990	West (1984)
5	POSSI-E(N)	Palomar Oschin	$+6 \rightarrow +90$	E	1949—1958	Minkowski & Abell (1963)
6	POSSII-B	Palomar Oschin	$-3 \rightarrow +90$	B_J	1987—1999	Reid et al. (1991)
7	POSSII-R	Palomar Oschin	$-3 \rightarrow +90$	R_F	1986—1999	Reid et al. (1991)
8	POSSII-I	Palomar Oschin	$-3 \rightarrow +90$	I_N	1989—2002	Reid et al. (1991)
9	POSSI-E(S)	Palomar Oschin	$-18 \rightarrow 0$	E	1949—1958	Minkowski & Abell (1963)

2.3 Sample definitions

In this section, I present two main samples that form the super-set of objects in the 7-DQ database. The first sample specifies the main objects of interest; a group of quasars defined by the SDSS Data Release 14 Quasar catalogue (DR14Q, Pâris et al. 2018). I utilise this sample as the basis for my 7-DQ quasars. The second is a sample of non-variable stars, which is a subsample of those defined by Ivezić et al. (2007). This sample acts as a control and is valuable for a number of reasons including: assessing random and systematic uncertainties, evaluating the effectiveness of the colour transformations described in Section 2.5, and estimating photometric uncertainties of the SuperCOSMOS data, whose magnitude measurements have no corresponding magnitude errors.

2.3.1 7-DQ quasar sample

The SDSS Data Release 14 Quasar catalogue (DR14Q, Pâris et al. 2018) is a catalogue of 526,356 spectroscopically confirmed quasars from the SDSS Data Release 14 (DR14). This catalogue includes all SDSS-IV/eBOSS objects that were spectroscopically targeted as quasar candidates and that are confirmed as quasars via an automated procedure. It is estimated to have ∼0.5% contamination. I primarily use the DR14Q catalogue for RA/Dec and redshift information of each quasar. In order to obtain properties such as black hole mass, luminosity and Eddington ratio for the 7-DQ quasars, I cross-matched with the DR16 catalogue of quasar properties provided by Wu & Shen (2022), discussed further in Chapter 5, Section 5.6.1.

I define the 7-DQ quasar sample to be the same as the DR14Q sample. This sample constitutes the super-set for my analysis and I use it in two ways; first, in my ensemble analysis, I use all available photometry from the full set of quasars in this sample. Second, in my subensemble analysis, I draw subsamples from this sample and compare results from analysis applied to each subsample. These subsamples are defined in such a way as to group by properties such as black hole mass, luminosity or rest-frame wavelength. In order to track quasars from this sample, I assign each of them a unique identifier, `uid`, which is simply the row number of the DR14Q sample once sorted by RA. This identifier is arbitrary and only occasionally mentioned in this thesis, but it is incredibly convenient from a

computational and analysis perspective. The apparent magnitude and redshift distributions of the DR14Q sample, as given by the DR14Q catalogue, are shown in Figure 2.1 and Figure 2.3. In Figure 2.2 I present the density map of the quasars in the luminosity-redshift ($L - z$) plane.

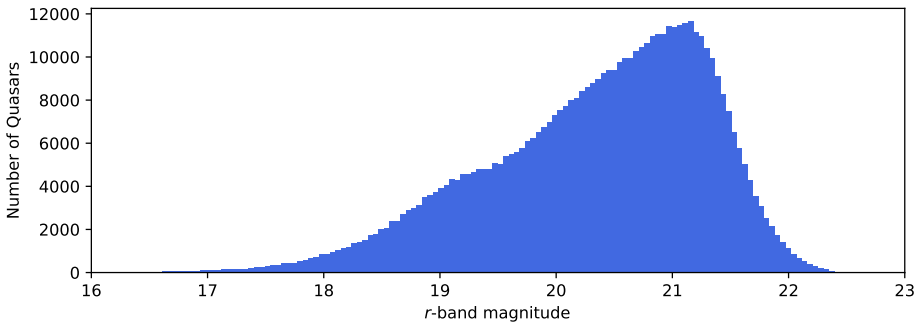

Figure 2.1: Distribution of r-band magnitudes for all DR14Q quasars, from the DR14Q catalogue.

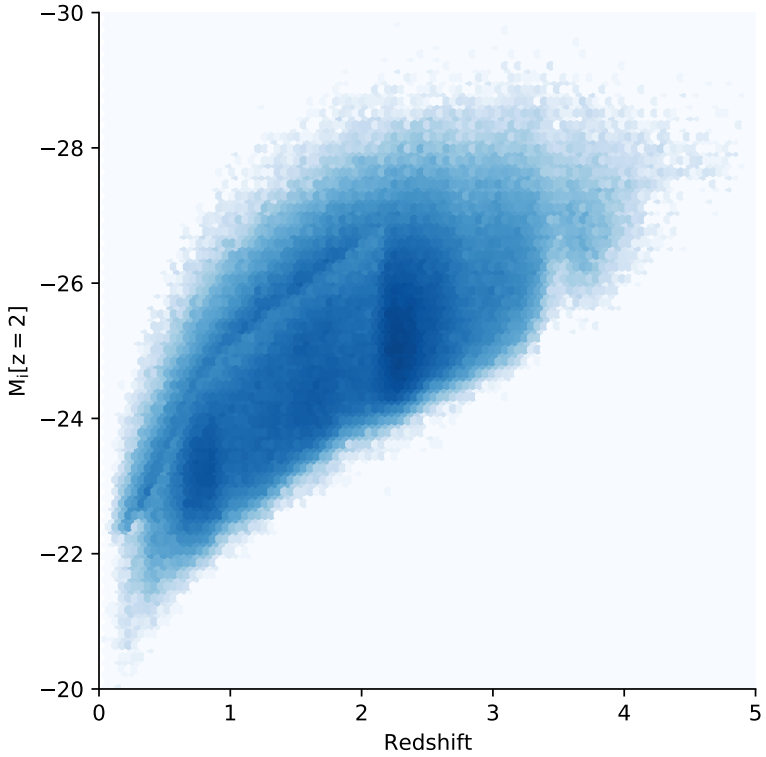

Figure 2.2: Density map of the DR14Q quasars in the L–z plane, where luminosity is expressed as absolute i-band magnitude K-corrected to $z = 2$. This absolute magnitude assumes $H_0 = 70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, as in Pâris et al. (2018).

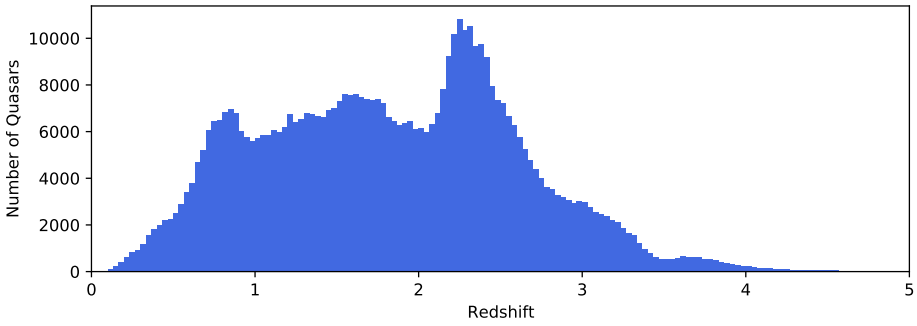

Figure 2.3: Distribution of redshifts for all DR14Q quasars, from the DR14Q catalogue. The peak at $z = 2.2$ is due to the BOSS quasar selection criteria.

2.3.2 7-DQ star sample

I take a selection of non-variable stars from the Stripe 82 Standard Star Catalogue (Ivezić et al. 2007) to use as control for the quasar sample, which makes up my 7-DQ star sample. This sample enables me to cross-calibrate photometry between surveys and to provide a comparison for variability studies of the quasar sample, as well as estimate photometric errors on the SuperCOSMOS data. This catalogue contains 1.01 million non-variable unresolved objects from the equatorial Stripe 82 ($|\delta| < 1.266\,\mathrm{deg}$) in the right ascension range $20^{\mathrm{h}}34^{\mathrm{m}}$–$4^{\mathrm{h}}00^{\mathrm{m}}$.

Using the full set of 1 million sources as my star sample would be excessive and slow down computation. Therefore, I selected a sample of 400,000 from this catalogue, ensuring a comparable size to my quasar sample. Rather than taking a purely random subsample of 400,000 stars, I aimed for the star sample to mirror the photometric characteristics of the quasar sample by matching their magnitude distributions simultaneously in all g, r and i bands. By performing this matching, I effectively balance both random and systematic uncertainties in the photometry of both datasets. This ensures that any observed noise in star photometry can serve as an appropriate measure for the baseline level of uncertainty in quasar photometry.

I achieved this by taking a weighted sample of stars from the full catalogue, with weights determined by the quasar magnitude distribution across all three bands. In practice this works as follows: First, the 7-DQ quasars are grouped into predefined magnitude bins, *separately* for the g, r, and i bands (i.e., not simultaneously in g, r, i), based on their corresponding mean magnitude in SDSS,

resulting in bin counts for each band denoted as (n_g, n_r, n_i). These bins specify the number of quasars per magnitude bin, per band. Then, for every star, I identify the corresponding three magnitude bins based on their SDSS mean magnitude in each band. By using the combined counts from the quasars $(n_g + n_r + n_i)$, I establish the weight for each star. Finally, I sample 400,000 stars from the full catalogue, using these weights. I then removed 78 objects that had not been observed in SDSS at least four times to ensure the mean magnitudes of the stars were reliable, resulting in a final sample of 399,922 stars.

The magnitude distribution showing the result of the matching process is presented in Figure 2.4, where the 'Standard Star catalogue' refers to the Stripe 82 Standard Star Catalogue (Ivezić et al. 2007), and '7-DQ star sample' refers to my 399,992 sample of stars which defines the 7-DQ star sample. The magnitude distribution is well-matched in the r-band, but not as well-matched in the g and i bands. This is because the colour distributions of the quasar and stars are significantly different. To illustrate this, colour-magnitude plots are shown for the two samples in Figure 2.5.

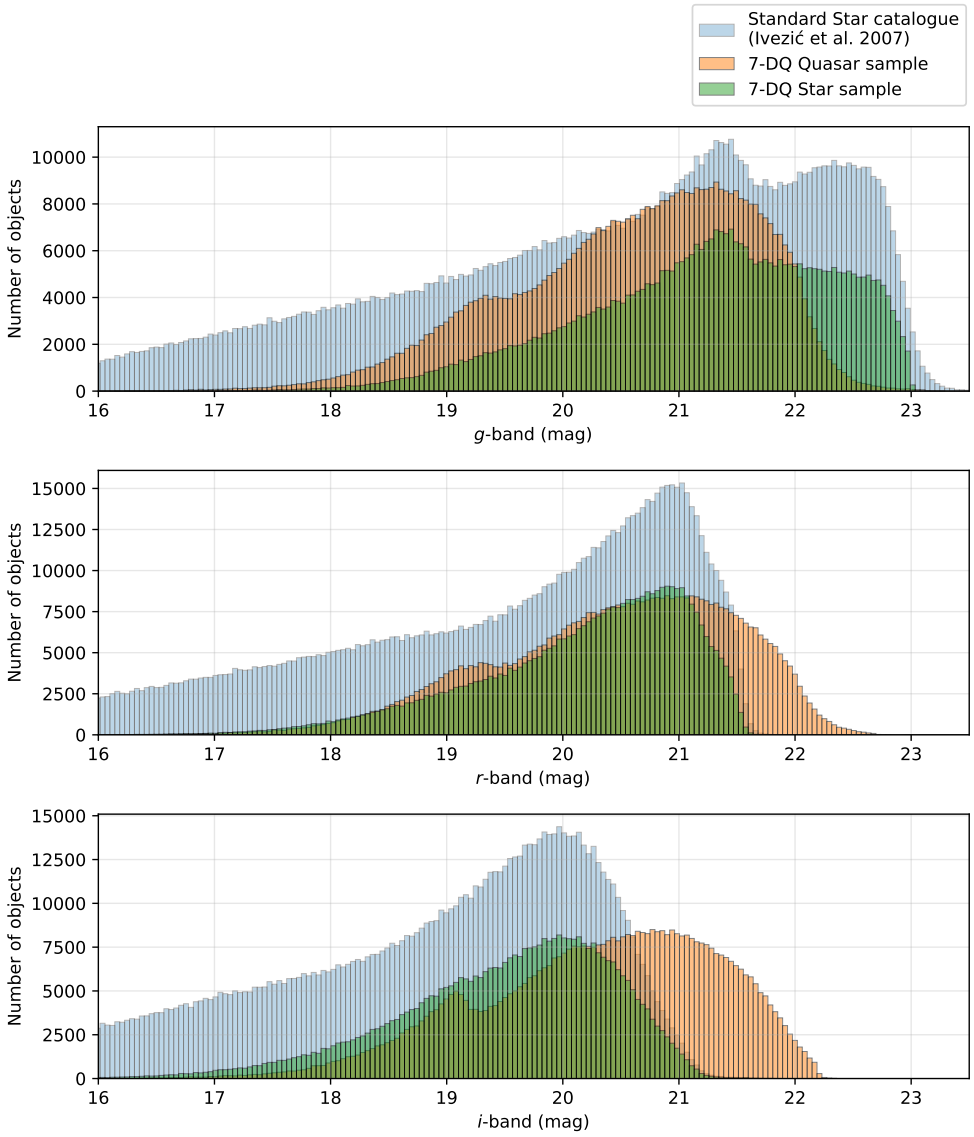

Figure 2.4: Magnitude distributions of the quasar and star samples. The full Stripe 82 Standard Star Catalogue is shown in light blue, while my matched subset is shown in green.

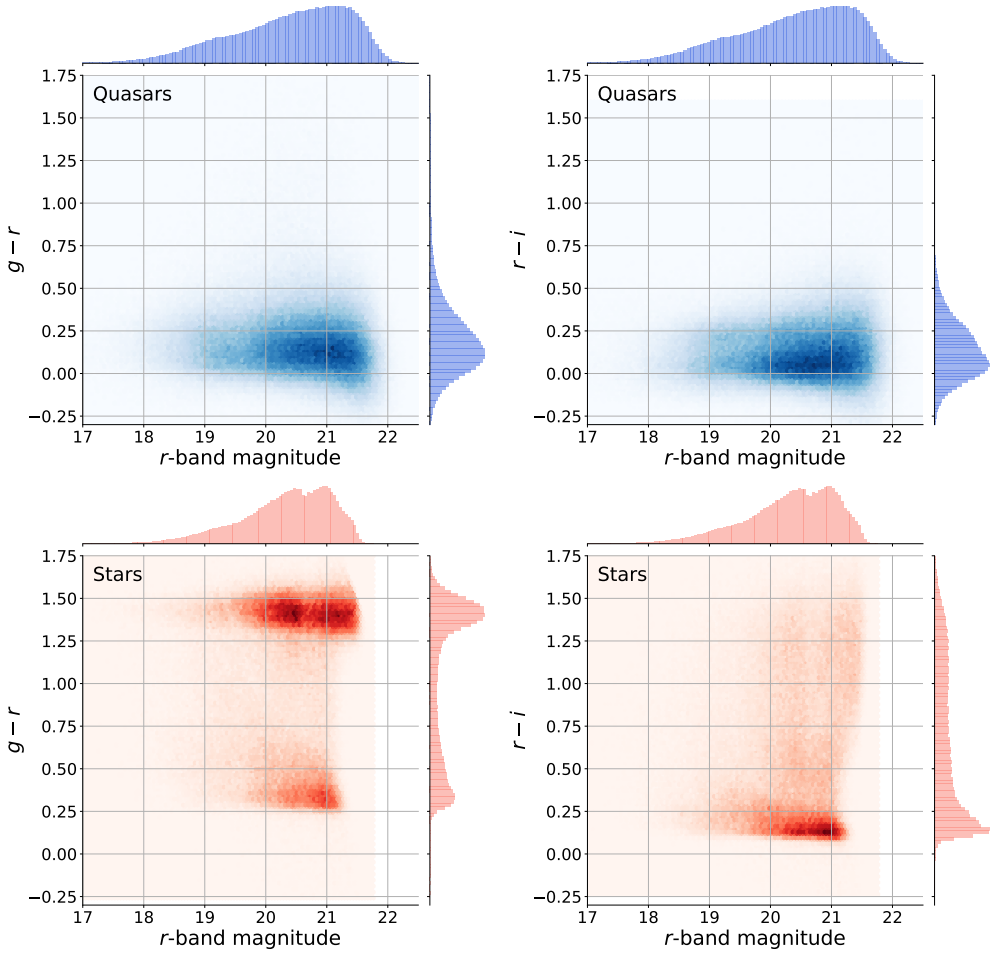

Figure 2.5: Colour-magnitude distributions of the quasars (top row, in blue) and stars (bottom row, in red), for $g-r$ and $r-i$ colors against the r-band.

2.4 Acquiring photometric data

The data used for this thesis are taken from publicly accessible data releases. PSF magnitudes are acquired via the relevant data services of each survey. Using coordinates from the quasar and star samples, photometric measurements within a radius of 1″ were obtained from the relevant data services, explained in the following sections. Note that in each case, I query a detection table and combine all photometry within 1″, assuming it comes from the same source. The alternative is to query photometry from a source table, where a particular source has assigned a unique ID within the survey database. However, I found that in many cases, this 'unique ID' was in fact not unique, and a single object could be assigned multiple IDs, even if there was no neighbouring object within 1″. This issue is particularly severe in the ZTF database, where photometry from

different bands would always be given a different ID. Therefore, I opted to collect all photometry within a given radius and assign my own unique ID to these observations. This assumes that there are no neighbouring objects within 1″ of each object in 7-DQ (quasars and stars). This assumption is generally quite safe, but will inevitably cause some contamination in the photometry. The effect of outliers introduced by this contamination will be reduced after applying outlier detection algorithms described in Section 3.2.2. Note that a radius of 1.5″ was used for SuperCOSMOS due to its less precise astrometry.

2.4.1 Choice of filter bands

Although SDSS and Pan-STARRS capture a wider spectrum by utilizing the *ugriz* and *grizy* bands respectively, ZTF and SuperCOSMOS focus solely on three bands: *gri* and *BRI* (which will later be converted to *gri* in Section 2.5), respectively. All of these surveys have *gri* in common. Hence, with the aim of combining photometric data from all four surveys, my analysis will be conducted exclusively in the g, r, and i bands for the remainder of the thesis.

Transmission curves of the g, r and i bands in the native SDSS, Pan-STARRS and ZTF photometric systems are presented in the top panel of Figure 2.6. The bottom panel shows the equivalent transmission curves for the filter-emulsion combinations in SuperCOSMOS. Note that ESO-R (a subsurvey of SuperCOSMOS) does not provide any photometry of 7-DQ quasars or stars, as the survey coverage and source footprints do not overlap. Therefore, its transmission curve is omitted from Figure 2.6. It is worth noting that the exact values of transmission in this plot are arbitrary, since they defined in different ways for each survey. However, this is not an issue since we are mostly interested in comparing the wavelength coverage of the filter profiles.

It is clear that the *gri* SDSS and Pan-STARRS filter profiles are very similar in wavelength extent and relative transmission. The ZTF g band is the most similar to the others in terms of wavelength coverage. However, r_{ZTF} and i_{ZTF} start and end between $100 - 500$ Å further redward than the corresponding bands in SDSS and Pan-STARRS. Native SDSS and Pan-STARRS magnitudes are comparable and may be used together without transformations depending on the type of analysis being done. The redder profiles of ZTF profiles, particularly in the g and i bands, necessitate colour transformations if any comparison is to be made between ZTF magnitudes and SDSS or Pan-STARRS magnitudes. The same applies to

SuperCOSMOS due to significant differences between the SuperCOSMOS filter profiles and the corresponding *gri* SDSS/Pan-STARRS filters. I discuss and apply these colour transformations in Section 2.5.

Another motivation to use the *gri* optical bands is that they sample the peak of the spectral energy distribution (SED) of a quasar with a typical redshift in 7-DQ, which is where the majority of the continuum emission from the disk is generated. By separating into three bands, I am able to make comparisons between adjacent parts of the SED and determine how variability changes with wavelength.

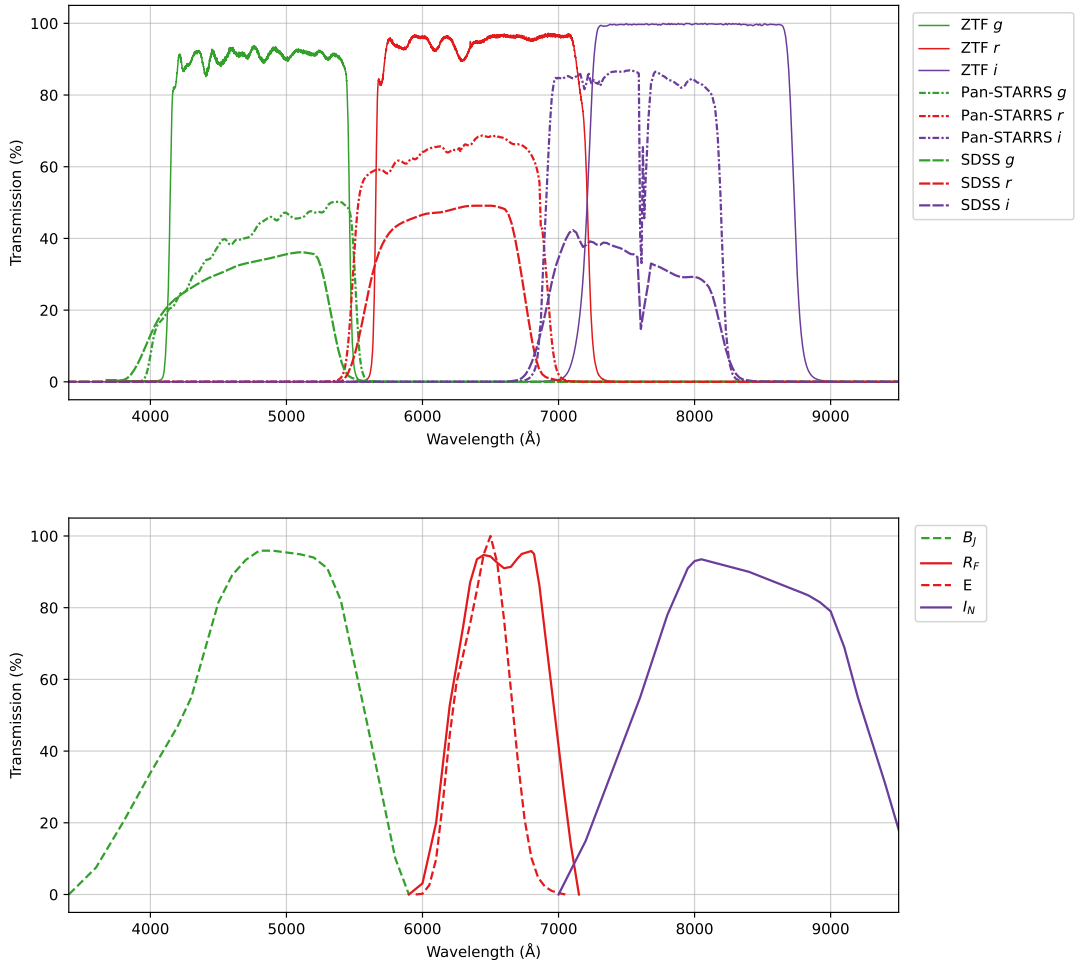

Figure 2.6: Top panel: Transmission curves of the g, r and i bands in the native SDSS, Pan-STARRS and ZTF photometric systems. Bottom panel: Transmission curves for the SuperCOSMOS filter-emulsion combinations listed in Table 2.2

2.4.2 Acquiring data from the Sloan Digital Sky Survey

SDSS photometry was obtained from the Catalogue Archive Server Jobs System (CasJobs) using the following SQL query:

```
select

q.uid as uid, p.objID,
q.ra, q.dec,
nb.distance as get_nearby_distance,

p.psfMag_u as mag_u, p.psfMagErr_u as magerr_u,
p.psfMag_g as mag_g, p.psfMagErr_g as magerr_g,
p.psfMag_r as mag_r, p.psfMagErr_r as magerr_r,
p.psfMag_i as mag_i, p.psfMagErr_i as magerr_i,
p.psfMag_z as mag_z, p.psfMagErr_z as magerr_z,
f.mjd_r as mjd

into mydb.sdss_secondary_qsos
from mydb.qsos_subsample_coords q
cross apply dbo.fGetNearbyObjAllEq(q.ra, q.dec, 1.0/60.0) as nb

join photoobj p on p.objid=nb.objid
join field f on f.fieldid=p.fieldid

ORDER BY uid ASC, mjd ASC
```

where `qsos_subsample_coords` is a list of the uid, RA and Dec of the DR14Q quasars. An example of the data retrieved by the query is shown in Table 2.3.

uid	objID	mag_g	magerr_g	...	mjd
1	1237678601842131080	21.90	0.06	...	54741.37
2	1237678663034601961	21.61	0.05	...	54747.35
3	1237656495650570598	21.22	0.03	...	52170.28
4	1237678777404358776	19.94	0.02	...	54764.19
5	1237666308022010256	21.49	0.04	...	53271.31
...	...	...	...	...	...

Table 2.3: Example of SDSS data obtained for the quasar population

2.4.3 Acquiring data from Pan-STARRS

Pan-STARRS photometry was also obtained via CasJobs using a similar query to SDSS. Since the photometry is provided in fluxes (Jy), I convert to *AB* magnitudes using the following relations,

$$m = -2.5 \log_{10}(f) + 8.9 \tag{2.1}$$

$$\sigma_m = \frac{2.5}{\ln 10}\left(\frac{\sigma_f}{f}\right). \tag{2.2}$$

Each source in the Pan-STARRS database is assigned a unique Pan-STARRS object ID, denoted `ps_objID`. There were 164 instances where multiple Pan-STARRS unique sources were mapped to a single `uid`. This is either caused by two objects in close proximity, or a mishap in the Pan-STARRS stacking, which results in the same source being assigned multiple `ps_objID`. In either case, 164 makes up a negligible fraction of our total sources, and thus we take the `ps_objID` of the object closest to the query coordinates and omit the rest. An example of the data obtained retrieved by the query is shown in Table 2.4

uid	filtercode	mjd	mag	magerr
1	g	55477.336	22.236	0.185
	g	55477.348	22.108	0.187
	i	55484.281	21.824	0.169
	r	55806.590	21.600	0.170
	i	56589.289	21.379	0.149
	...	...	...	...
2	g	55448.496	21.503	0.130
	g	55448.500	21.286	0.130
	g	55449.535	20.963	0.182
	i	55452.469	20.928	0.107
	r	55452.480	21.086	0.088
	r	55452.484	21.014	0.112
	...	...	...	...

Table 2.4: Example of Pan-STARRS data obtained for the quasar sample after converting from fluxes to magnitudes.

2.4.4 Acquiring data from Zwicky Transient Facility

To collect photometry of our samples from ZTF, the IRSA service was used to find sources which match our objects. ZTF often has multiple identifiers per object, therefore we used the source table to query ZTF sources which correspond to our `uid` unique identifier. Obtaining photometry from ZTF is less trivial than the others, since there is a limit on the number of objects when using the query form. Therefore, I wrote a script to automate the data collection. This query was rerun in April 2023 once Data Release 17 was available to obtain the most up-to-date observations for our sources. An example of the data obtained retrieved by the query is shown in Table 2.5

oid	mjd	mag	magerr	filtercode	clrcoeff
396203300009105	58289.465	20.403	0.180	zr	0.125
396203300009105	58301.457	20.040	0.145	zr	0.129
396203300009105	58312.438	20.523	0.191	zr	0.112
396303300002074	58333.426	20.300	0.183	zi	0.198
396303300002074	58351.336	20.485	0.182	zi	0.187
396303300002074	58361.363	20.179	0.179	zi	0.188
396103300005093	58318.418	20.667	0.179	zg	-0.037
...	...	...	...	...	...

Table 2.5: Example of ZTF data obtained for the quasar sample. The `oid` column contains non-unique ZTF identifiers, while `clrcoeff` column contains colour coefficients used to transform to the PanSTARRS system.

2.4.5 Acquiring data from SuperCOSMOS Sky Survey

Photometry from the SuperCOSMOS Sky Survey was acquired using the SuperCOSMOS Science Archive (SSA) (`http://ssa.roe.ac.uk`). There are several databases which I queried to get the photometry in the desired form. The Detection Table includes all observations across the SuperCOSMOS Sky Survey across multiple bands. I queried this table for all observations within 1.5″ of the quasar and star coordinates. While I used a pairing radius of 1″ for the other surveys, I decided to use 1.5″ for SSS due to its lower precision astrometry. I then joined this result to the Plate Table to find which observations belong to which plates, in order to separate observations by survey. Due to the sky coverage of each individual survey, they each return a different number of observations. Since the 7-DQ stars cover a narrow range of sky, there are only observations

from four surveys that make up the SuperCOSMOS Sky Survey. The number of observations obtained for the quasars and stars is listed in Figure 2.6

ID	Survey	Quasars	Stars
1	SERC-J/EJ	99,859	190,830
2	SERC-R/AAO-R	74,023	207,035
3	SERC-I	28,153	130,758
5	POSSI-E(N)	214,695	0
6	POSSII-B	568,369	0
7	POSSII-R	424,554	0
8	POSSII-I	183,996	0
9	POSSI-E(S)	29,514	106,322

Table 2.6: Number of observations obtained from the individual surveys which make up SSS. ESO-R has been omitted as it has no counts for either population.

2.4.6 Summary of acquired photometric data

Here I present some summary statistics of the photometric data in 7-DQ. Table 2.7 shows cumulative counts for the number of observations, $N_{\rm obs}$, and the number of unique objects, $N_{\rm uniq}$, for 7-DQ quasars and stars in the g, r and i bands, respectively. Additionally, I show magnitude error distribution of SDSS, PS and ZTF for the quasars in Figure 2.7. Differences in instrumentation result in different levels of signal-to-noise, resulting in differences in these distributions.

It is clear from Table 2.7 that there are a different number of sources for each survey and band, for both the quasar and stars. This is because I do not use a single fixed sample for my photometry. Instead, the sample is flexible, in the sense that I have data for some sources in some surveys that I do not have in other surveys. For example, SDSS has photometry for $\sim$100,000 more quasars than ZTF in the r-band. Nevertheless, I still use SDSS photometry from those additional quasars. The alternative would be to restrict the full DR14Q sample to a subset of quasars which are recorded in all 4 surveys in all 3 bands. However, this approach would involve removing the majority of photometric data, which contradicts the objective of building a comprehensive photometric database. Therefore, my full sample of 526,356 quasars defines the super-set, from which photometry is obtained from the respective surveys, where available. Since I will be primarily carrying out ensemble studies in this thesis, where photometry is combined from many sources, this is the most effective strategy to leverage the full set of available

photometric data.

The survey footprints of the matched objects for each survey are shown in Figure 2.8. To illustrate the completeness of the 7-DQ quasars, I plot the cumulative number of matched objects for increasing pairing radius in Figure 2.9, where a match is made if there is at least one observation in any band obtained within the specified radius of the DR14Q sample coordinates. I present the fraction of matched objects at the threshold radii for each survey in Table 2.8.

Survey	Band	Quasars		Stars	
		$N_{\rm obs}$	$N_{\rm uniq}$	$N_{\rm obs}$	$N_{\rm uniq}$
SDSS	*g*	1,433,578	525,776	7,017,259	399,820
	r	1,433,518	525,776	7,023,454	399,820
	i	1,433,148	525,740	7,025,247	399,819
PS	*g*	4,989,424	500,380	3,771,984	382,278
	r	6,206,943	504,321	5,847,374	399,224
	i	8,956,496	505,146	8,836,851	399,325
ZTF	*g*	113,489,045	365,086	18,232,968	121,042
	r	167,879,362	433,727	48,411,810	213,665
	i	25,859,665	380,703	7,231,088	175,960
SSS	*g*	668,163	470,649	190,830	163,821
	r	742,711	381,001	313,356	184,739
	i	212,088	166,039	130,758	114,914
Total	*g*	120,580,210	526,281	29,213,041	399,924
	r	176,262,534	526,308	61,595,994	399,927
	i	36,461,397	526,308	23,223,944	399,927

Table 2.7: Cumulative counts of the number of observations, $N_{\rm obs}$, and the number of unique objects, $N_{\rm uniq}$, for 7-DQ quasars and stars in the *g*, *r* and *i* bands

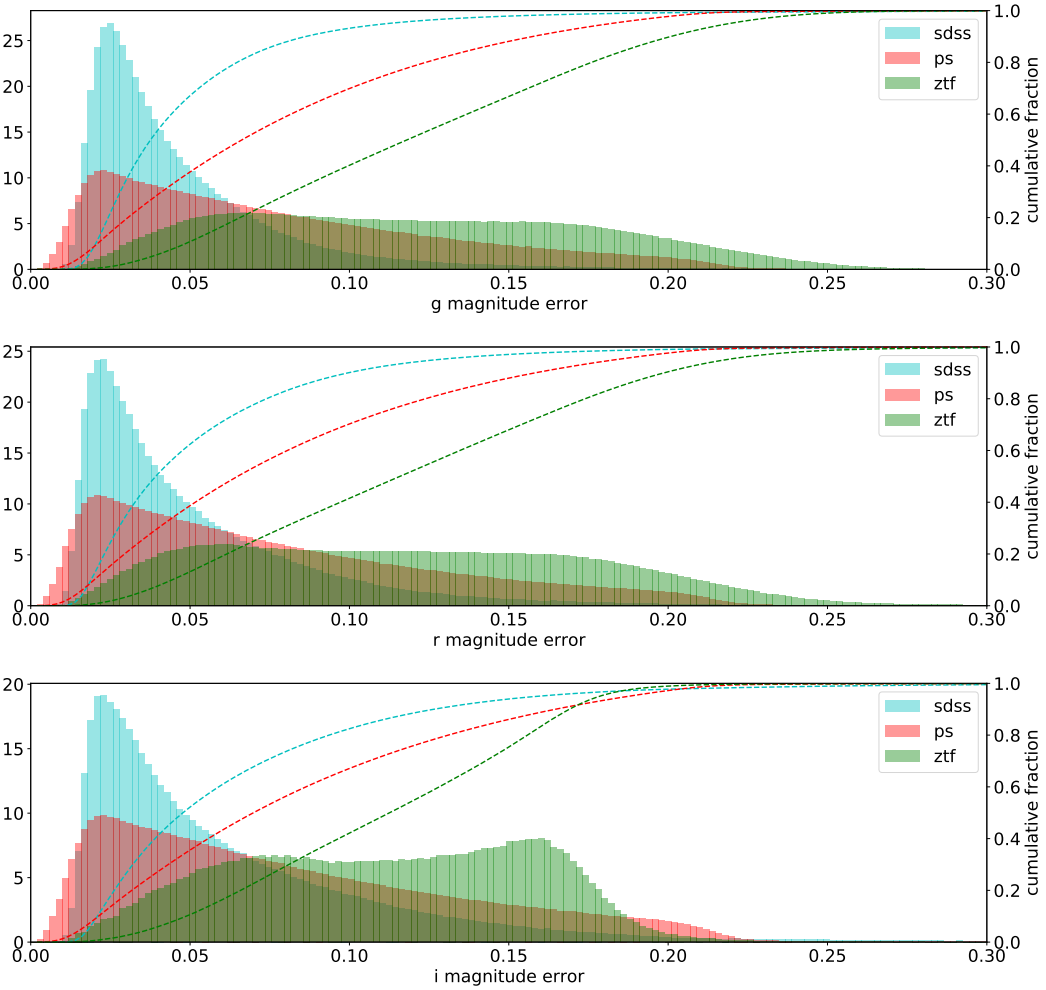

Figure 2.7: Magnitude errors distribution for the quasars in the g, r and i bands for SDSS, Pan-STARRS and ZTF. The cumulative distribution is overplotted (dotted). SuperCOSMOS is not included as the survey does not provide magnitude errors. SDSS reports the smallest errors, while ZTF has a broad noise distribution, and Pan-STARRS is between the two. This is expected given the depths of each of these surveys.

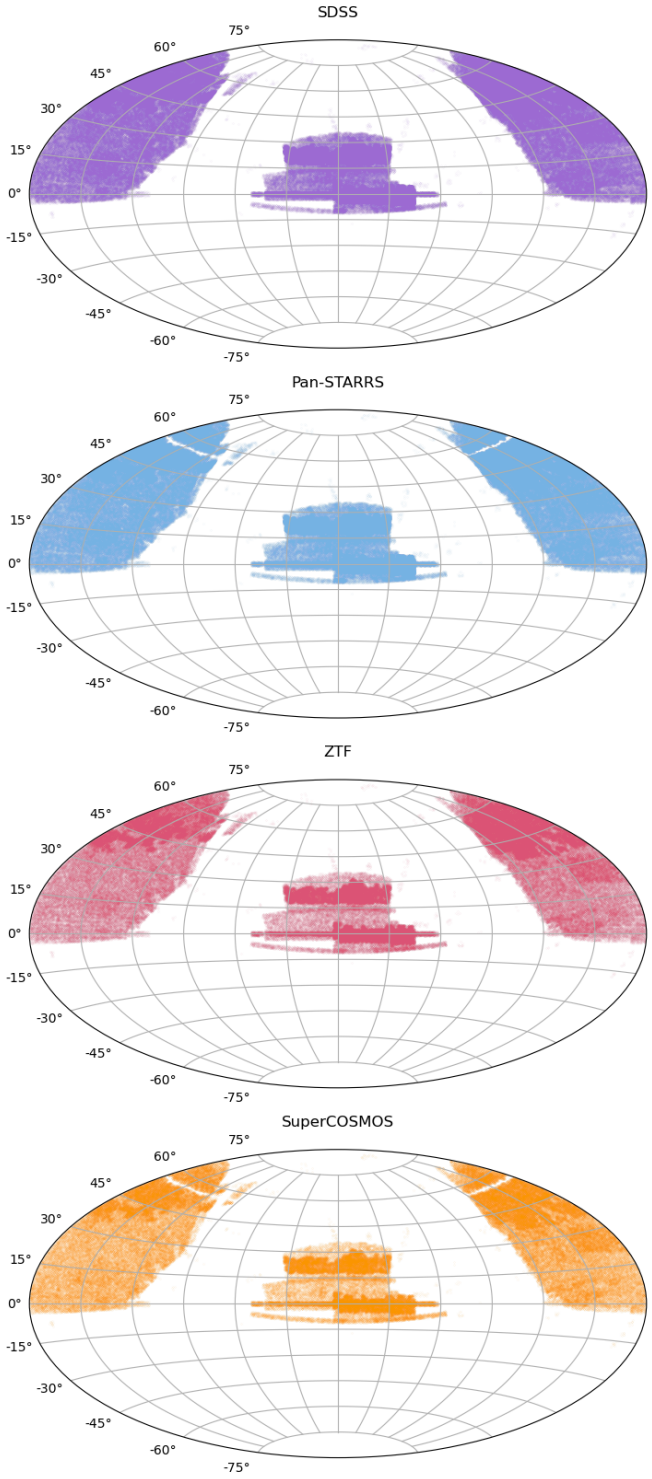

Figure 2.8: Footprint of sky observations for the quasar sample for each survey. The footprint of the star sample is not shown, but it would occupy a small region around Stripe 82 which is the narrow horizontal strip on the centre of each diagram.

Survey	Pairing radius	Completeness
SSS	1.5″	94.28%
SDSS	1″	99.91%
PS	1″	99.85%
ZTF	1″	97.77%

Table 2.8: Completeness table showing the percentage of sources cross-matched and the radius thresholds used to match photometry. SDSS has the highest completeness since it defines the DR14Q sample.

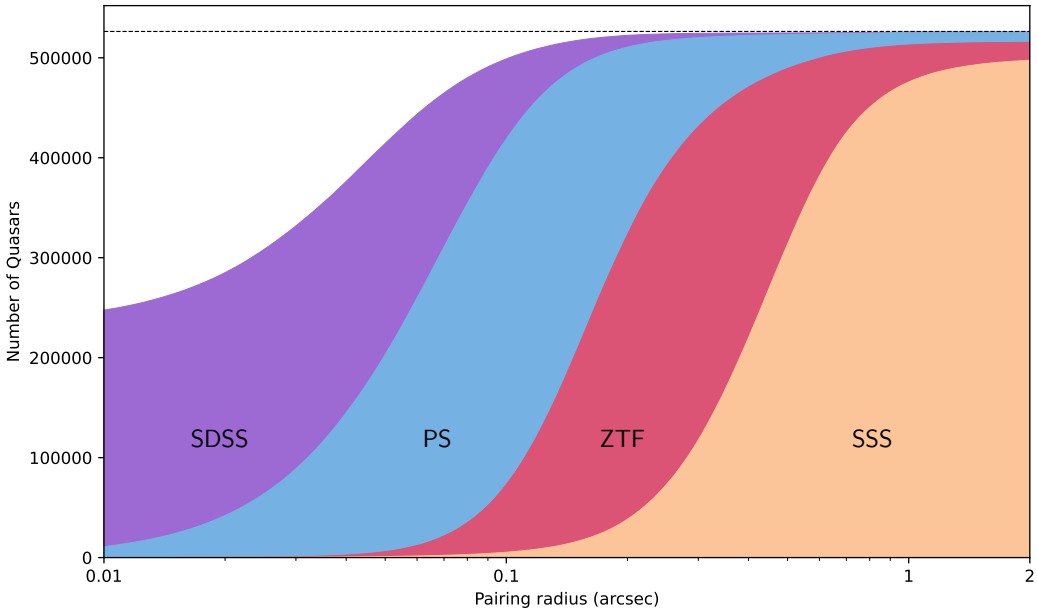

Figure 2.9: Cumulative distribution of number of objects with increasing separation from the DR14Q coordinates for each of the four surveys. The black dotted line shows the total number of quasars.

2.5 Colour Transformations

Colour transformations refer to the process of converting measurements in one set of filter bands to another set of filter bands. It is necessary because differences in instrumentation and filter profiles result in various different photometric systems for each survey, which would otherwise prevent direct comparison of raw magnitudes between surveys. A particular colour transformation requires the colour of the object to be known at the same epoch as the measurement being transformed. Unfortunately, I do not have colours for every epoch where I

wish to make a transformation, and therefore an approximation must be made. There are two possible approaches. The first involves using non-simultaneous observations to estimate the colour of an object at a particular epoch. The second approach uses the colour at one epoch (provided by SDSS) for all future transformations. The first approach is not suitable for quasars since they are inherently variable, and it is likely the quasar will have varied significantly between the two measurements, resulting in a colour that may be significantly different from the true colour at either of the two measurements. I decided to use the second approach, as it is suitable under the assumption that the colours of our objects stay roughly constant. The reason I use colours supplied by SDSS is that only this survey provides simultaneous imaging required for computing colours. Therefore, I used the mean colours from SDSS for all colour transformations. This assumes that the colour of the objects stay roughly constant throughout the light curve. Although this is a safe assumption for the stars, it will introduce some uncertainty for the quasars. To estimate this uncertainty, I investigated the spread of quasar colours in SDSS for quasars that had been observed at least twice. The mean standard deviation of quasar colours are 0.081 mag and 0.087 mag for $(g-r)$ and $(r-i)$, respectively. This was calculated using 1,167,613 measurements of each $(g-r)$ and $(r-i)$ from 258,319 quasars which had been multiply imaged by SDSS. These deviations are small enough to justify my method. In the remainder of this section, I describe the colour transformations of each survey into the Pan-STARRS system and assess the validity of each transformation using colour-magnitude plots. I decided to shift everything to the Pan-STARRS system, since it has the most reliable photometry and required the fewest total transformations.

The SDSS and SuperCOSMOS colour transformations rely on non-variable stars as calibrators. However, there are differences between stellar and quasar SEDs that will affect colour transformations between two photometric systems. To address this, I used an alternative approach for deriving colour transformations, one that accounts for the quasar spectrum, discussed in Section 2.5.1. The two methods were not compared using SuperCOSMOS data due to uncertainties in filter profiles and its lower signal-to-noise.

2.5.1 Transformation of SDSS magnitudes

The colour transformations from SDSS magnitudes to Pan-STARRS magnitudes are provided by Tonry et al. (2012). The transformation is a second order

polynomial in $g - r$,

$$m_{\mathrm{PS1}} = m_{\mathrm{SDSS}} + a_0 + a_1(g-r)_{\mathrm{SDSS}} + a_2(g-r)^2_{\mathrm{SDSS}}, \tag{2.3}$$

with coefficients are provided in Table 2.9. These coefficients were derived for stellar SEDs, however, we find that they work well for the quasar population. To illustrate the effectiveness of the transformations in each band, I produced magnitude-colour diagrams before and after the transformations were applied. An example, using the g-band, is shown in Figure 2.10 for the star and quasar population. Figure 2.10 shows that the colour dependence on magnitude is almost entirely removed in the star sample. This is also true for the quasar sample, however, it is less obvious since quasars are variable and therefore the y-axis limits are much larger. Figure 2.11 compares transformations from Tonry et al. (2012) and the method described in Section 2.5, which involves integrating the spectrum through the relevant filter profiles and estimating the difference as a polynomial function of colour. There is very little difference between the two methods over the quasar colour range. The difference in transformed magnitudes is negligible, and therefore either method appropriate. I decided to use the Tonry et al. (2012) transformations, as they work well on both the 7-DQ quasar and star photometry.

Filterband	a_0	a_1	a_2
g	−0.011	−0.125	−0.015
r	+0.001	−0.006	−0.002
i	+0.004	−0.014	+0.001

Table 2.9: Polynomial coefficients for Equation 2.3 used to transform SDSS magnitudes, from Tonry et al. (2012).

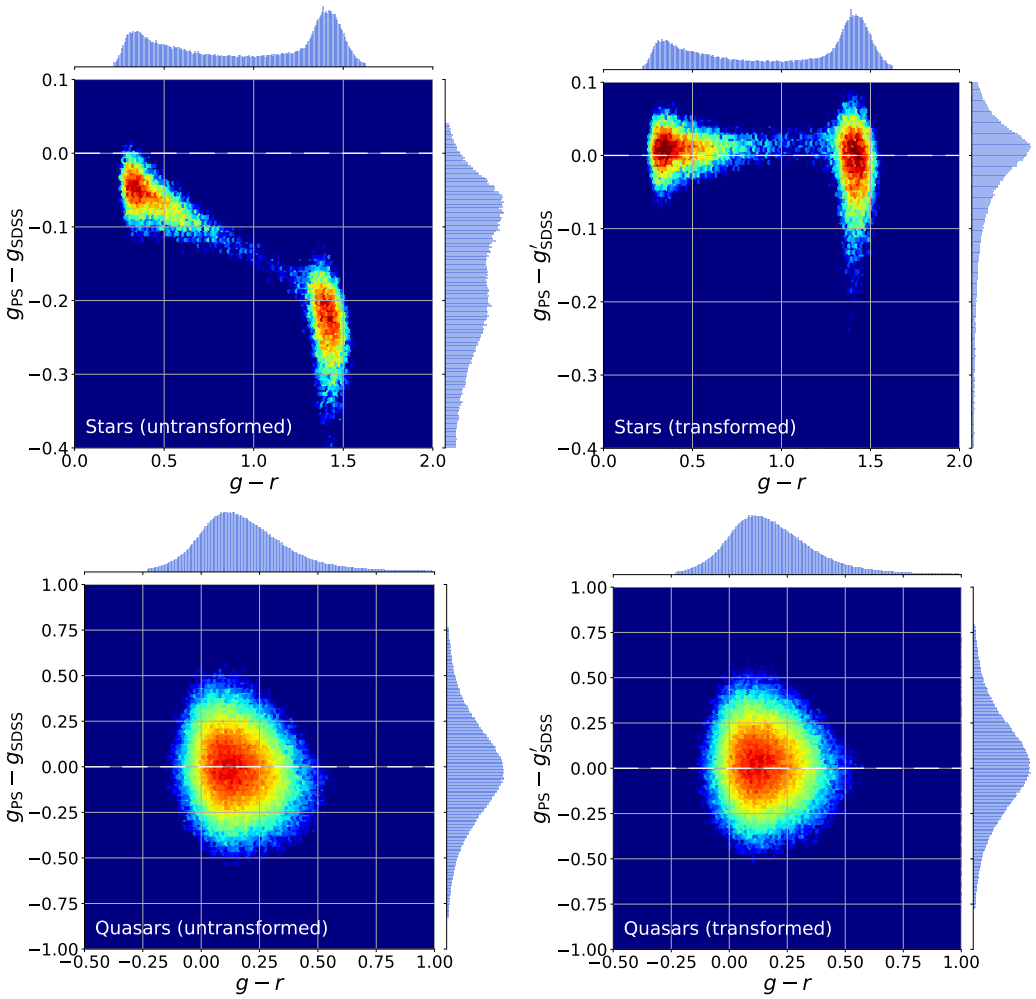

Figure 2.10: 2-D histograms illustrating the effectiveness of the colour transformations for the star sample (top row) and quasar sample (bottom row). Left and right columns are untransformed and transformed, respectively.

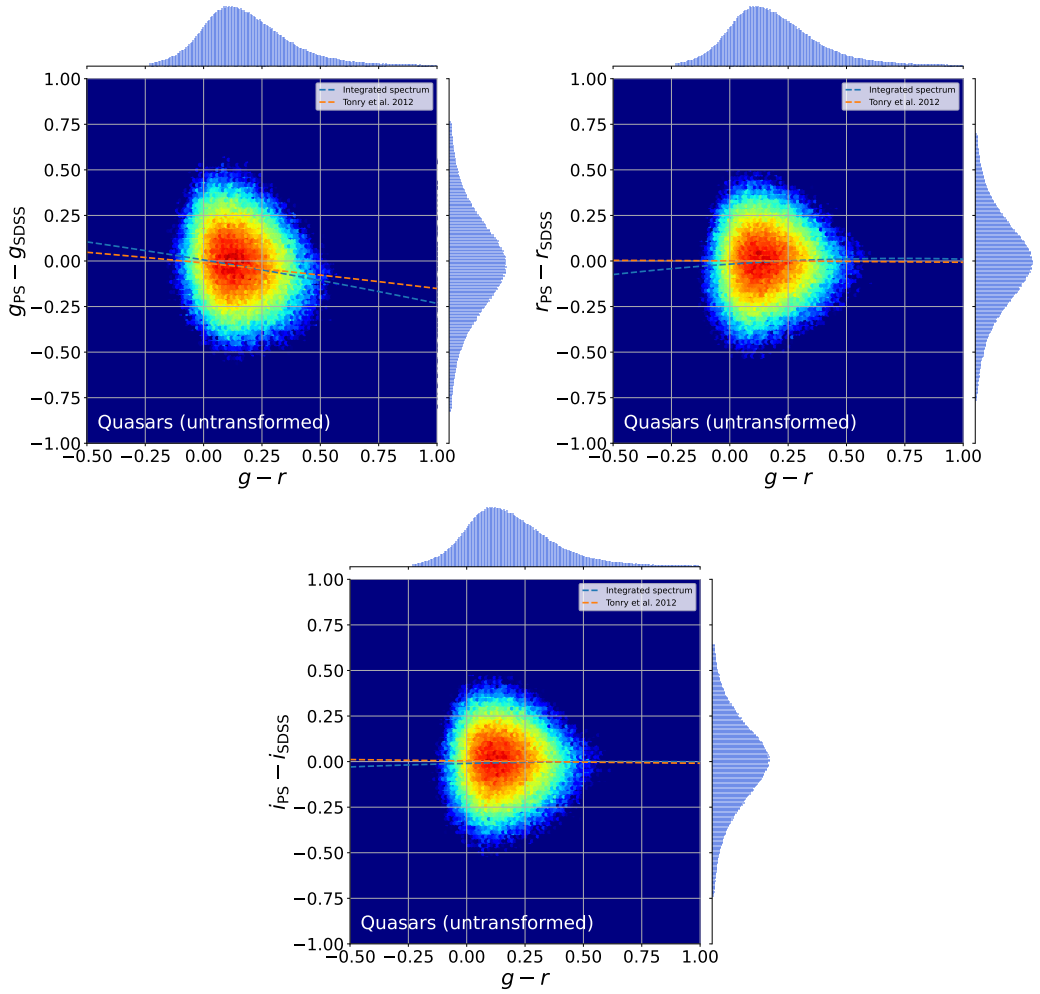

Figure 2.11: 2-D histograms comparing colour transformations on SDSS data using Tonry et al. (2012) and the integrated spectrum method, for the *gri* bands.

2.5.2 Transformation of ZTF magnitudes

The transformations from ZTF to Pan-STARRS are provided by Masci et al. (2019),

$$g_{\mathrm{PS1}} = g_{\mathrm{ZTF}} + c_g(g-r)_{\mathrm{SDSS}}, \tag{2.4}$$

$$r_{\mathrm{PS1}} = r_{\mathrm{ZTF}} + c_r(g-r)_{\mathrm{SDSS}}, \tag{2.5}$$

$$i_{\mathrm{PS1}} = i_{\mathrm{ZTF}} + c_i(r-i)_{\mathrm{SDSS}}, \tag{2.6}$$

where c_f is the colour coefficient for a given filter ($f = g, r, i$), which corresponds to the column `clrcoeff` (shown previously in Table 2.5). Each image is assigned

its own unique solution for ZP_f, c_f, which respectively correspond to the intercept and slope from a linear fit using matches of ZTF sources to a subset of calibrators in PS1. As with SDSS, I produced magnitude-colour diagrams before and after the transformations were applied. An example, using the r-band, is shown in Figure 2.12 for the star and quasar population. The transformation removes most of the magnitude-colour dependence for both populations.

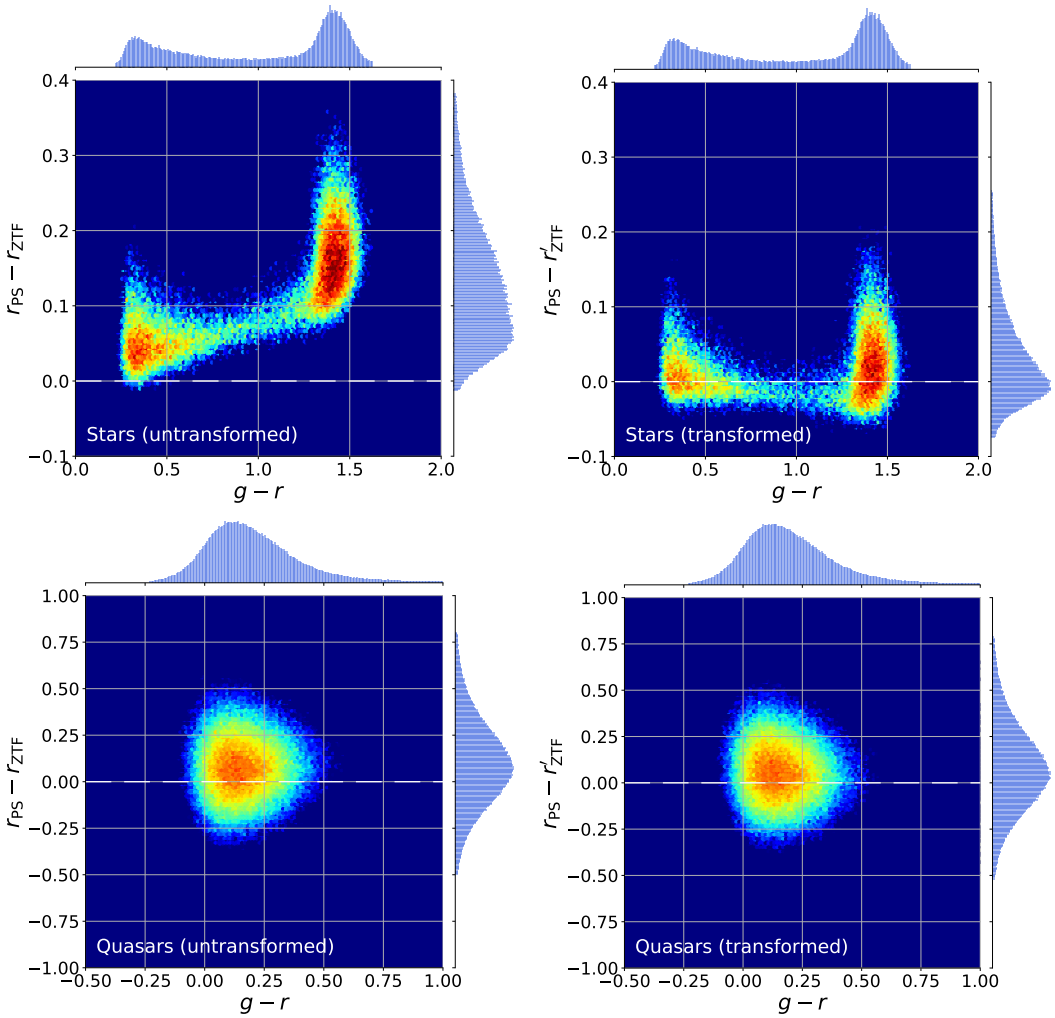

Figure 2.12: 2-D histograms illustrating the effectiveness of the colour transformations for the quasar sample. Left and right columns are untransformed and transformed, respectively.

2.5.3 Transformation of SuperCOSMOS magnitudes

The case of transforming SuperCOSMOS magnitudes requires more care. One possibility is to use transformations produced by Peacock et al. (2016). They provide transformations of B_J, R_F, and I_N bands to standard SDSS gri bands.

However, since their colour corrections are designed for galaxies, I found that they performed poorly on 7-DQ stars and quasars; the transformations shifted 7-DQ star and quasar magnitudes away from their corresponding magnitudes measured by Pan-STARRS. In the absence of any other suitable published transformations, I decided to create my own. I used the 7-DQ star photometry to produce linear transformations of B_J, R_F, E, and I_N to Pan-STARRS gri magnitudes. An additional benefit to creating my own transformations is that I can fit over a colour range that is more suitable for the quasars. For example, when creating transformations which use the $g - r$ colour term, I pick stars in the range $-0.1 < g - r < 0.7$ which is the extent of the quasar $g - r$ colours. This results in a more accurate transformation for the 7-DQ quasars.

To create my own transformations, I calculated magnitude offsets between stars in the SuperCOSMOS and Pan-STARRS data, and performed linear regression of the offsets as a function of colour. The coefficients of the linear regression were found to be:

$$g_{\mathrm{PS1}} = B_J - 0.222(g-r)_{\mathrm{SDSS}} - 0.200, \tag{2.7}$$

$$r_{\mathrm{PS1}} = R_F + 0.091(g-r)_{\mathrm{SDSS}} + 0.151, \tag{2.8}$$

$$r_{\mathrm{PS1}} = \mathrm{E} - 0.210(g-r)_{\mathrm{SDSS}} + 0.338, \tag{2.9}$$

$$i_{\mathrm{PS1}} = I_N + 0.126(r-i)_{\mathrm{SDSS}} + 0.436. \tag{2.10}$$

Note that, since the sky footprint of the 7-DQ stars is quite narrow, I was only able to use photometry from four subsurveys of SuperCOSMOS to generate these transformations (see Table 2.6). Fortunately, these 4 subsurveys covered all four bands of the photometric plates (i.e., B_J, R_F, E, and I_N). The results of the transformation for the 7-DQ stars and quasars are shown in Figure 2.13, using the g and B_J band as an example.

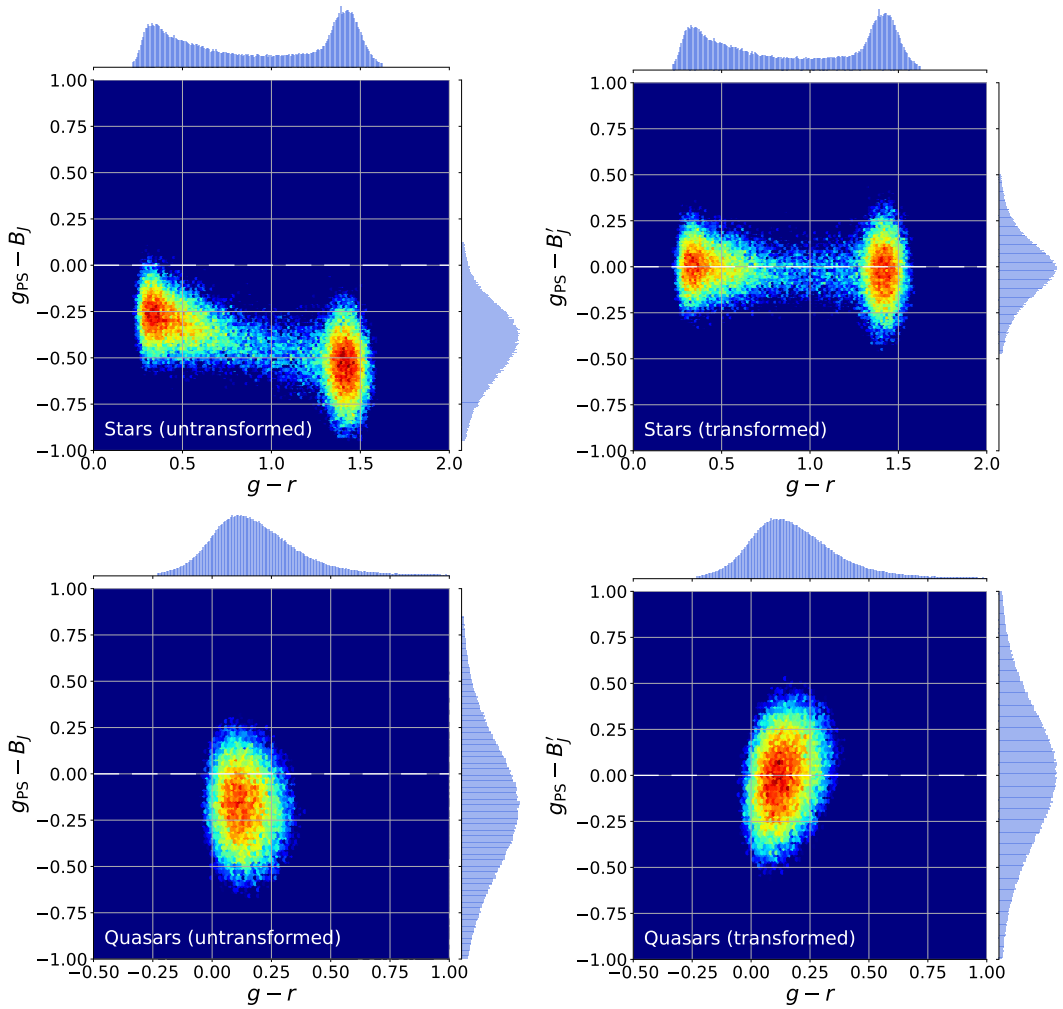

Figure 2.13: 2-D histograms illustrating the effectiveness of the colour transformations for the quasar sample. Left and right columns are untransformed and transformed, respectively.

2.5.4 Effectiveness of transformations

To evaluate the performance of colour transformations, I analyse magnitude differences, which I refer to as 'residuals', for corresponding objects (quasars and stars) between surveys. These residuals are computed by first determining the median magnitude for each object within each survey. Next, I calculate the difference between the median magnitudes obtained from each survey and the median magnitude of the same object in Pan-STARRS. This procedure is conducted separately for all three *gri* bands, and both untransformed magnitudes (without any colour corrections) and transformed magnitudes (with colour corrections applied). The residuals for the 7-DQ stars and quasars are

shown in Figures 2.14 and 2.15 respectively. Since quasars are variable, these distributions are much wider than for the star sample. Therefore, the effect of the transformation is not as clear as it is for the stars (with the exception of SSS). However, the transformations are still very important to reduce systematic bias introduced by differences in instrumentation. While it is qualitatively clear from these plots that the transformations reduce the residuals overall, I computed the means of these distributions in each case to verify, quantitatively, that the magnitudes transform as expected. The mean residuals and their errors (calculated as the standard deviation of the residuals) are presented in Tables 2.10 and 2.11 respectively. Since the mean residuals in each band are reduced in the majority of cases, it is clear that the transformations are performing as expected and are now within the same photometric system with sufficient precision.

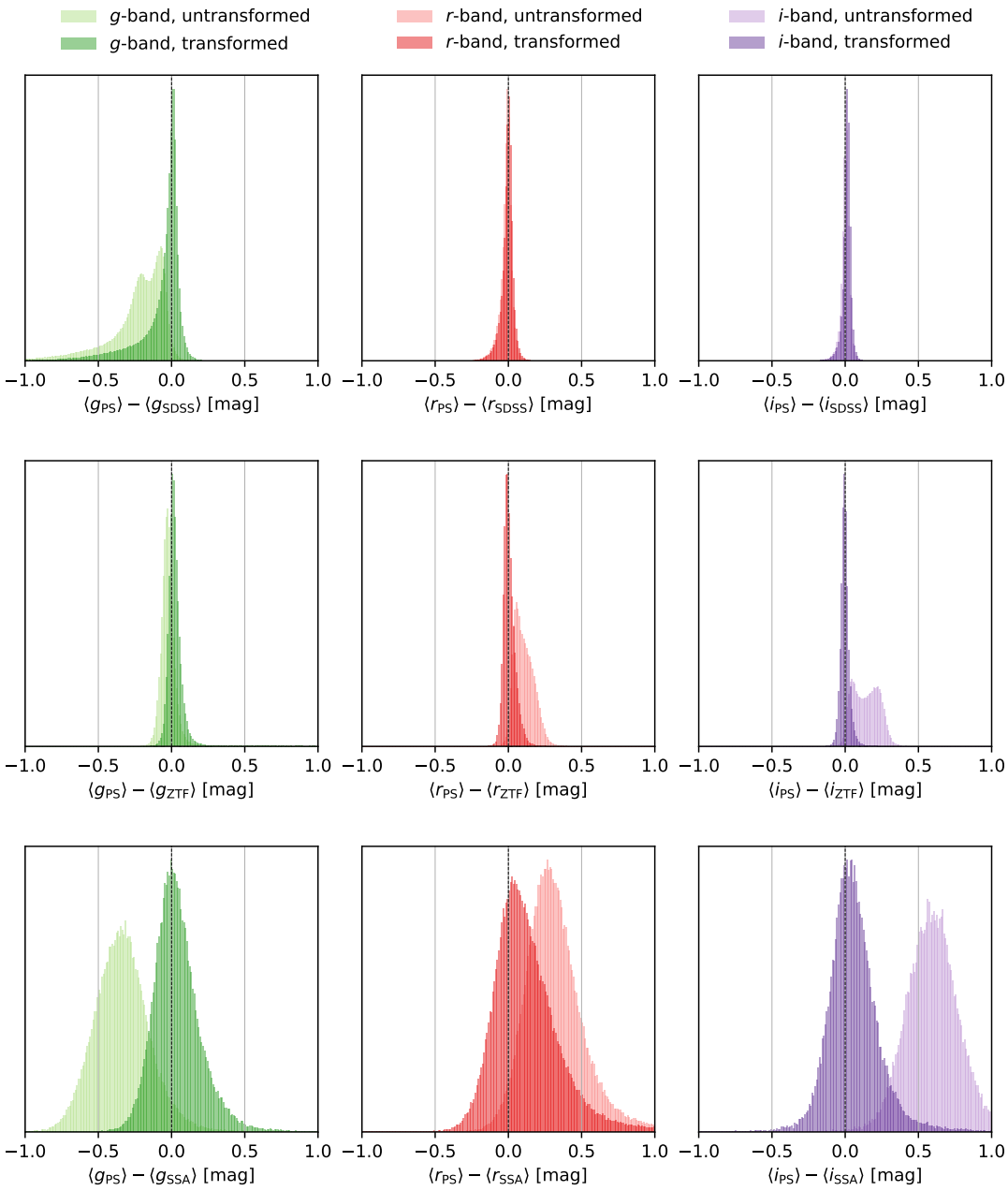

Figure 2.14: Residuals between Pan-STARRS and the other surveys for the star sample. Residuals are calculated as median magnitudes per object before and after transformation. Note that the angled brackets denote the median average.

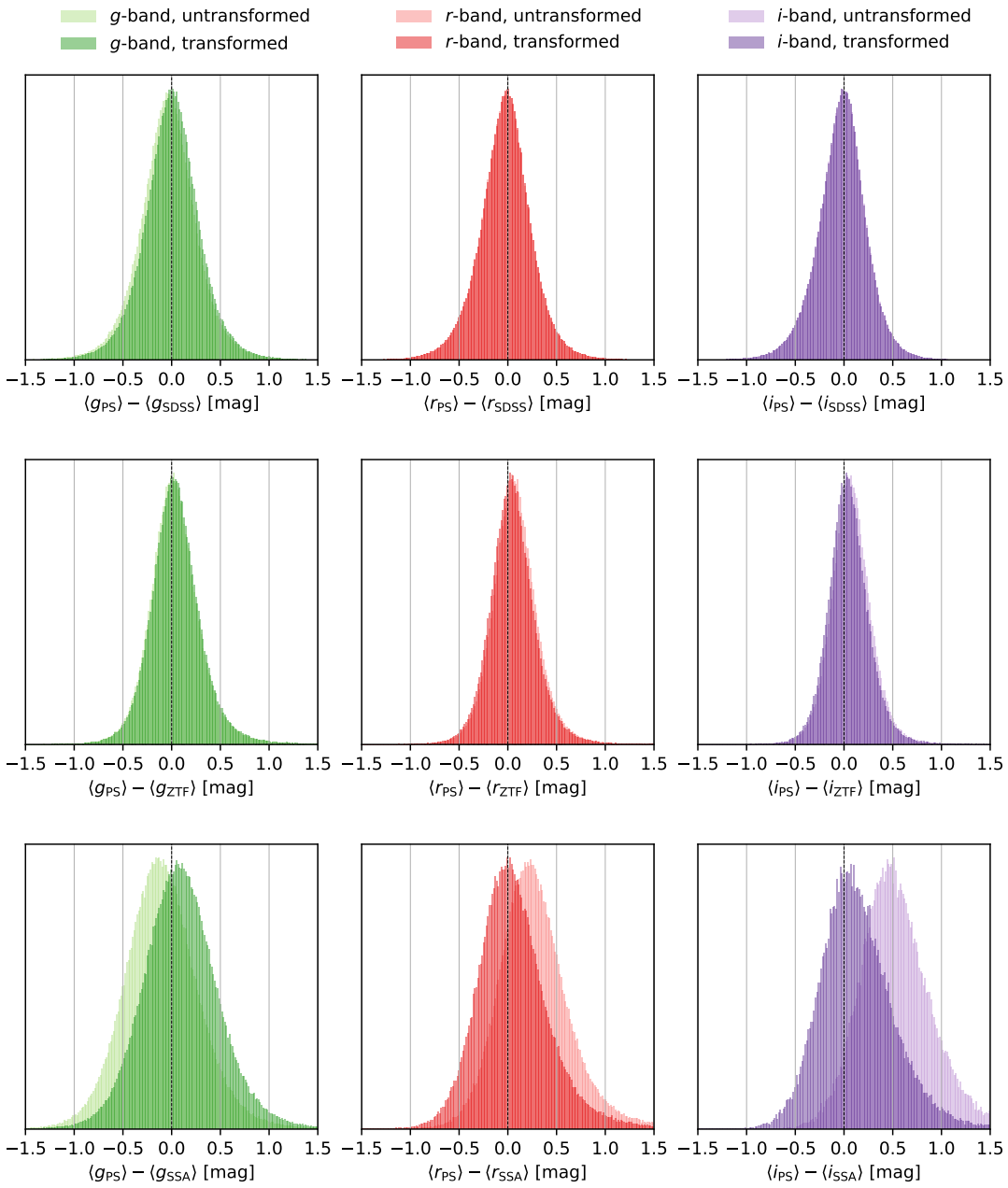

Figure 2.15: Residuals between Pan-STARRS and the other surveys for the quasar sample. Residuals are calculated as median magnitudes per object before and after transformation. Note that the angled brackets denote the median average.

Survey pair	Band	Mean residual (untransf.)	Mean residual (transf.)
PS-SDSS	g	-0.237 ± 0.207	-0.076 ± 0.170
	r	-0.015 ± 0.042	-0.007 ± 0.042
	i	$+0.001 \pm 0.030$	$+0.011 \pm 0.031$
PS-ZTF	g	-0.012 ± 0.119	$+0.034 \pm 0.121$
	r	$+0.106 \pm 0.065$	$+0.007 \pm 0.040$
	i	$+0.143 \pm 0.086$	-0.002 ± 0.031
PS-SSS	g	-0.338 ± 0.186	$+0.039 \pm 0.161$
	r	$+0.328 \pm 0.220$	$+0.134 \pm 0.239$
	i	$+0.608 \pm 0.200$	$+0.062 \pm 0.183$

Table 2.10: Mean residuals of the star sample in the g, r and i bands before and after applying colour transformations

Survey pair	Band	Mean residual (untransf.)	Mean residual (transf.)
PS-SDSS	g	-0.035 ± 0.301	$+0.003 \pm 0.296$
	r	-0.026 ± 0.275	-0.025 ± 0.275
	i	-0.021 ± 0.256	-0.022 ± 0.257
PS-ZTF	g	$+0.034 \pm 0.264$	$+0.047 \pm 0.265$
	r	$+0.075 \pm 0.235$	$+0.051 \pm 0.235$
	i	$+0.074 \pm 0.210$	$+0.047 \pm 0.210$
PS-SSS	g	-0.099 ± 0.380	$+0.097 \pm 0.388$
	r	$+0.281 \pm 0.375$	$+0.062 \pm 0.373$
	i	$+0.547 \pm 0.383$	$+0.140 \pm 0.382$

Table 2.11: Mean residuals of the quasar sample in the g, r and i bands before and after applying colour transformations

2.6 Approximating SuperCOSMOS magnitude errors

The SuperCOSMOS data does not provide magnitude errors on individual measurements. Therefore, I estimated the photometric noise of these measurements as a function of magnitude using the 7-DQ star photometry. Since the stars are non-variable, the observed variance of magnitude differences is solely due to photometric noise. Thus, the variance, $\sigma^2_{\rm tot}$, of the magnitude differences, $\Delta m_{\rm SSS\text{-}PS} = m_{\rm SSS} - m_{\rm PS}$ between the SuperCOSMOS and Pan-STARRS observations is due to the combined errors of these two surveys. By approximating photometric noise as a function of magnitude, we may write the

total combined photometric error as

$$\sigma^2_{\text{PS-SSS}}(m) = \sigma^2_{\text{PS}}(m) + \sigma^2_{\text{SSS}}(m). \tag{2.11}$$

I was able to estimate $\sigma^2_{\text{PS-SSS}}(m)$ as the width of the $\Delta m_{\text{PS-SSS}}$ distribution. There are two methods to estimate $\sigma^2_{\text{PS}}(m)$. The first involves calculating the average photometric error with increasing magnitude bins. The second involves calculating the width of the $\Delta m_{\text{SSS-SSS}}$ distribution with increasing magnitude bins. After trying both of these methods, I found that the average photometric noise was on average 50% smaller than the observed scatter of the $\Delta m_{\text{SSS-SSS}}$ distribution. I opted to take the second method in case the photometric errors were slightly underestimated. Having approximated both $\sigma^2_{\text{PS-SSS}}(m)$ and $\sigma^2_{\text{PS}}(m)$, the difference represents the average photometric noise as a function of magnitude for SuperCOSMOS. I added these errors back into each SuperCOSMOS observation depending on the magnitude of the object, using a linear fit to interpolate between magnitude bins. This process gives a good estimate of the errors and allows us to weight observations in favour of objects which have better photometry. Figure 2.16 illustrates the process of calculating $\sigma^2_{\text{SSS}}(m)$ in magnitude bins, with a linear fit.

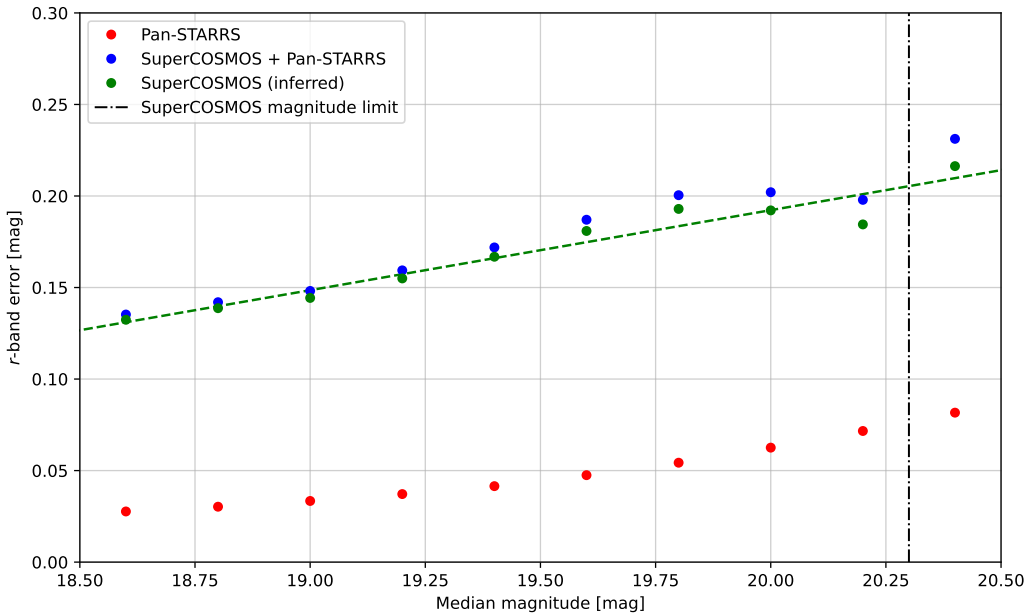

Figure 2.16: Photometric error-magnitude plot for Pan-STARRS and SuperCOSMOS for the star population

Chapter 3

Computational methods, algorithms and preprocessing

3.1 Introduction

The effectiveness of my analysis relies heavily on a sophisticated data processing pipeline that I developed. This pipeline has played a crucial role in enabling me to both clean and preprocess the 7-DQ photometry acquired in Chapter 2. Additionally, it has allowed me to and carry out large-scale data exploration and conduct large ensemble photometric studies presented in Chapters 4 and 5. 7-DQ presents computational challenges due to its large sample size and sheer volume of observations; therefore, developing efficient computational methods within this pipeline is vital in order to reduce, preprocess, and analyse it effectively. Furthermore, my pipeline involves producing data derivatives that are easier to analyse from a computational perspective.

The development of this pipeline required a large amount of work, however, I will refrain from explaining it in full technical detail; full knowledge of the entire data pipeline and its intricacies is not necessary in order to understand the results that I produce from it. Describing it fully would simply be too cumbersome for the reader. Therefore, this chapter is kept brief, and I will focus on two key techniques which outline some of the original methods that I developed and have employed in this pipeline that are relevant to the reader. The first is the data cleaning part of the pipeline, described in Section 3.2. This involves the removal of 'bad'

measurements as well as photometry below the limiting magnitude of the survey it is taken from. The second is outlined in Section 3.3 and involves the construction of a secondary dataset which is a derivative of 7-DQ. This dataset is defined in such a way as to make it convenient to study unique pairs of measurements in a given light curve, for all the light curves in 7-DQ. I refer to this secondary dataset as the 'pairwise dataset' and it facilitates my large ensemble studies. More specifically, I utilise the pairwise dataset to characterise quasar variability in terms of time-lag and magnitude differences in Chapters 4 and 5.

3.2 Data cleaning: Outlier detection and removing bad photometry

Raw observational data will almost always be contaminated with bad measurements due to a number of contributing factors. Some of these include instrumental errors (e.g., issues with photometric or astrometric calibration), environmental conditions (e.g., poor seeing conditions or cosmic rays) or data and computational errors (e.g., issues with data transfer, storage or retrieval).

All observations will suffer from these issues to some degree, however, each measurement is usually quoted with some uncertainty to account for this where possible. For example, uncertainties in the full photometric pipeline (i.e., the process from collecting light to storing data) of a survey mean that magnitudes are almost always quoted with an associated magnitude error. Nevertheless, some observational data will be so severely affected by these issues that they are considered 'bad data'. I use this term loosely, as it is a general term and its definition depends on the context in which it is applied. For example, bad data could refer to measurements which could not have been produced by a real observation, e.g., null pixel values. Alternatively, bad data could refer to an outlier which is valid data, but highly improbable that it could have been caused by a real event. A relevant example of this is the mismatching of photometry between two nearby objects, due to issues with astrometric calibration; if these objects are significantly different in apparent magnitude, the resulting light curve will contain anomalous data points. Encountering bad data is inevitable and will introduce additional uncertainty and bias when performing analysis if not removed. This effect is exacerbated for statistics that are sensitive to outliers. Additionally, the method of collection and merging of this data also suffer from

computational errors such as cross-matching errors. Therefore, the data should be properly cleaned before using it for analyses. However, it is not practical to visually inspect the hundreds of thousands of light curves in my 7-DQ database for outliers. Thus, I use a combination of magnitude cuts (discussed in Section 3.2.1) and outlier detection algorithms (discussed in Section 3.2.2) to reduce the effect of bad data on my analyses.

3.2.1 Magnitude cuts

First, I remove unphysical observations by omitting observations whose magnitudes are outside the range $15 < m < 25$ and magnitude errors are outside the range $0 < \sigma < 2$. Although these are generous bounds, they are designed to remove observations that have had measurement issues, such as those which contain NaNs or values such as -999 (a typical entry for a null measurement). This step removes a very small fraction of the photometry (<1%).

Second, I imposed magnitude constraints on each survey based on the limiting magnitude (shown previously in Table 2.1). Since the surveys have different imaging depths, some surveys are able to provide reliable photometry of fainter objects while others are not. For example, the 5-σ depths for ZTF and SDSS are 20.6 and 22.7 in the r-band respectively, therefore a quasar of magnitude 21 would have reliable photometry from SDSS, but not ZTF. Moreover, because quasars are variable, ZTF would more likely report observations of this object if it exhibited a period of brightening during the observations. The resulting effect is a systematic overestimation of brightness for dimmer quasars in shallower surveys, known as Malmquist bias. To remove this bias, I decided to remove photometry of an object from a particular survey if the median magnitude of that object across the light curve is fainter than the limiting magnitude of that survey. I opted for the median as it is more robust to outliers. Compared to other data cleaning steps outlined in this section, these magnitude constraints remove the most amount of photometry by far. Nevertheless, it is a necessary trade-off to drastically reduce the effect of Malmquist bias in my analyses. On average, these constraints remove up to 30% of quasars for SuperCOSMOS and ZTF, and up to 3% of quasars in SDSS and Pan-STARRS. The equivalent fractions for the stars are roughly half of those for the quasars.

While the process of removing photometry of an object based on limiting magnitude will reduce the effect of Malmquist bias, it does not eliminate it

completely. Therefore, I defined a 'bright subset' of 139,233 quasars whose median $g_{\mathrm{SDSS}} < 20$. This subset, being less susceptible to Malmquist bias, will serve as a control for some of my analysis to determine the extent of bias in the full sample. I also define a bright subset of 64,626 stars, specified by the same constraint, $g_{\mathrm{SDSS}} < 20$.

3.2.2 Outlier detection

I developed an outlier detection algorithm to automatically remove unphysical observations. Given a distribution of the underlying data points, an outlier may be caused by one or more of the following:

1. A legitimate but surprising and unexpected data value.
2. A data value that was measured or stored incorrectly.
3. A contaminant, i.e., an observation from some neighbouring source.

In this section, I will focus on removing outliers caused by 2 and 3. I take a few steps to clean the data; almost all quasars (96.8%) have at least two observations within the same night at some point in their light curve. However, since observed variability on timescales this short will be dominated by noise, I combine nightly observations into a single data point by taking the median of the measured magnitudes. The median was chosen over the mean as it is less sensitive to outliers. I then remove observations within a light curve whose magnitude deviates 7σ from the median magnitude of that light curve, where σ is the standard deviation of the magnitudes in that light curve. Although this is a generous constraint, quasars are known to exhibit dramatic changes in brightness and I did not want to unintentionally remove observations of extreme variability. This only removes a small fraction of obvious bad data, so I do a second pass with a more sophisticated outlier detection algorithm to further remove unphysical observations.

Bad data points are often easy to spot by eye. By leveraging contextual information from adjacent data points, examining data in different wave bands, and applying an intuitive understanding of physical plausibility, a person can confidently determine the authenticity of an observation with a relatively high degree of certainty by eye. However, producing an algorithm to carry out the

same process with similar accuracy is not a simple task. I created my own outlier detection algorithm to filter out bad points which were not removed from the first pass. My algorithm involves a rolling window of a width of 5 data points to calculate the absolute deviation of a particular point away from the median of the window. If this deviation is greater than 2.6 times the interquartile range of the distribution of magnitudes in the light curve, it is deemed an outlier. The number 2.6 seems arbitrary, but I calculated it empirically after looking at the distribution of absolute deviations across a sample of 5000 light curves which are representative of the rest of the light curves. This number roughly marks the boundary between high variability and obvious outliers. A demonstration of my algorithm is shown in Figure 3.1.

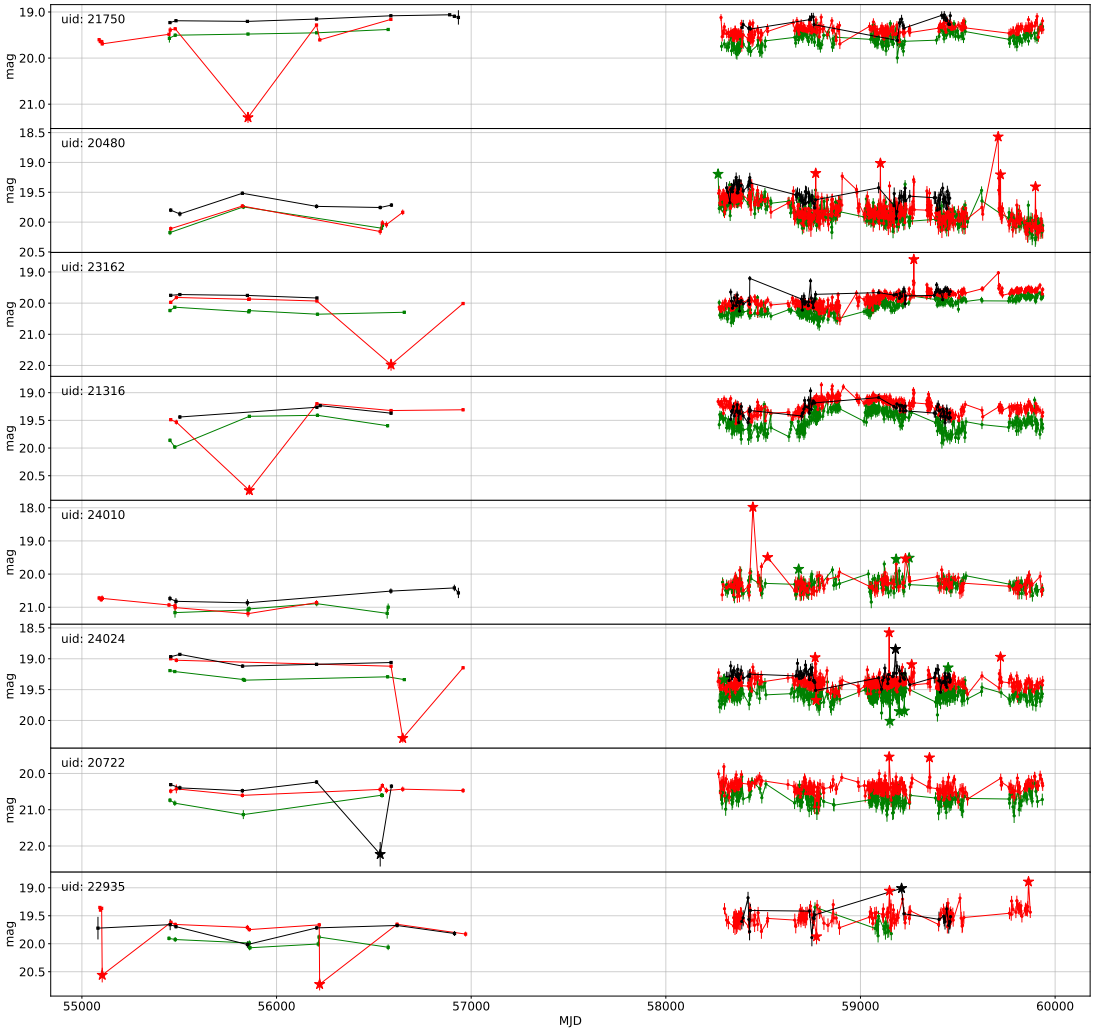

Figure 3.1: Demonstration of my outlier detection algorithm. Outlier points are marked with a 'star', and will be removed from the light curve. Here, I chose to show only parts of the light curve covering photometry from Pan-STARRS and ZTF for clarity, but the same algorithm applies to the entire light curve which includes all four surveys.

3.2.3 Data summary after cleaning

Here, I provide a concise summary of the 7-DQ data following the cleaning process. First, Table 3.1 shows an example of light curve data for a particular quasar (`uid` = 1), in the g, r and i bands. Note that I introduce a survey identifier, which I refer to as the survey ID, to distinguish between surveys once the data have been merged. This identifier takes on the integers 3, 5, 7, and 11, for SuperCOSMOS, Pan-STARRS, SDSS and ZTF, respectively. These integers

are intentionally prime, the reason for which will become clear in Section 3.3.2. Second, in Table 3.2, I show the number of observations and number of unique objects in each survey for the 7-DQ quasars and stars across the g, r and i bands. Third, in Figure 3.2, I show histograms of the number of observations per object for both the 7-DQ quasars and stars. Finally, Figure 3.3 displays a 4-way Venn diagram to illustrate the number of 7-DQ quasars with at least one observation in each combination of surveys.

uid	band	mag	magerr	mjd	mjd_rf	sid
1	g	21.917612	0.067707	54741.371094	16543.176517	5
1	i	21.933083	0.148910	54741.371094	16543.176517	5
1	g	22.171011	0.186231	55477.343750	16765.591946	7
1	i	21.863562	0.171900	55484.289062	16767.690862	7
1	r	21.584364	0.179572	55806.593750	16865.093306	7
1	g	21.753338	0.153714	55879.226562	16887.043385	7
1	i	21.352200	0.181100	56231.410156	16993.475417	7
1	r	21.763350	0.176923	56247.320312	16998.283564	7
1	g	20.977379	0.094571	56549.582031	17089.628900	7
...	...	...	...	...	...	...

Table 3.1: An example of quasar photometry from the 7-DQ database. Here, `uid` refers to my unique quasar identifier, `mjd` and `mjd_rf` are the time of observation (in MJD), except the latter is in the rest-frame, `sid` is the survey ID.

Survey	Band	Quasars		Stars	
		N_{obs}	N_{uniq}	N_{obs}	N_{uniq}
SDSS	g	1,403,558	524,966	6,925,451	399,806
	r	1,403,448	525,311	6,931,668	399,819
	i	1,383,692	522,078	6,933,457	399,819
PS	g	2,587,265	500,380	2,036,345	382,275
	r	3,023,809	504,321	2,830,731	399,223
	i	3,620,154	505,146	3,503,282	399,324
ZTF	g	62,635,876	263,221	11,573,951	71,071
	r	75,540,780	269,808	28,076,200	132,170
	i	11,203,322	144,125	4,853,779	101,747
SSS	g	464,138	324,465	112,231	95,411
	r	476,091	213,150	208,257	106,878
	i	67,467	49,703	44,375	37,792
Total	g	67,090,837	525,507	20,647,978	399,914
	r	80,444,128	525,965	38,046,856	399,926
	i	16,274,635	524,512	15,334,893	399,927

Table 3.2: Cumulative counts of the number of observations, N_{obs}, and the number of unique objects, N_{uniq}, for 7-DQ quasars and stars in the g, r and i bands after removing outliers and objects outside the magnitude thresholds

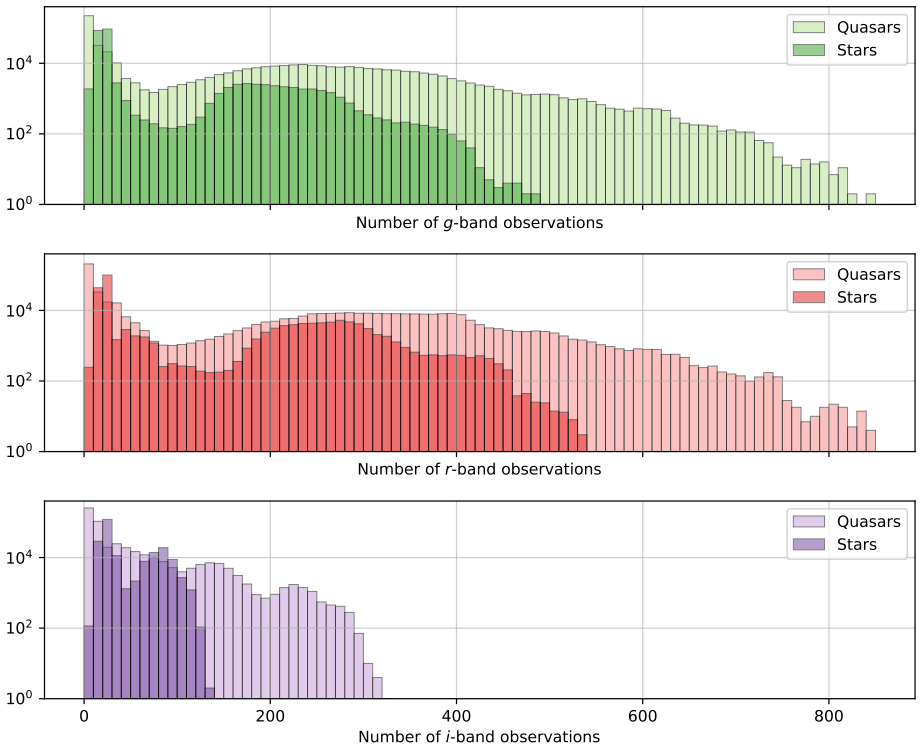

Figure 3.2: Number of observations per object for 7-DQ quasars and stars. There are fewer observations in the i-band because only 1/3 of the ZTF i-band is public.

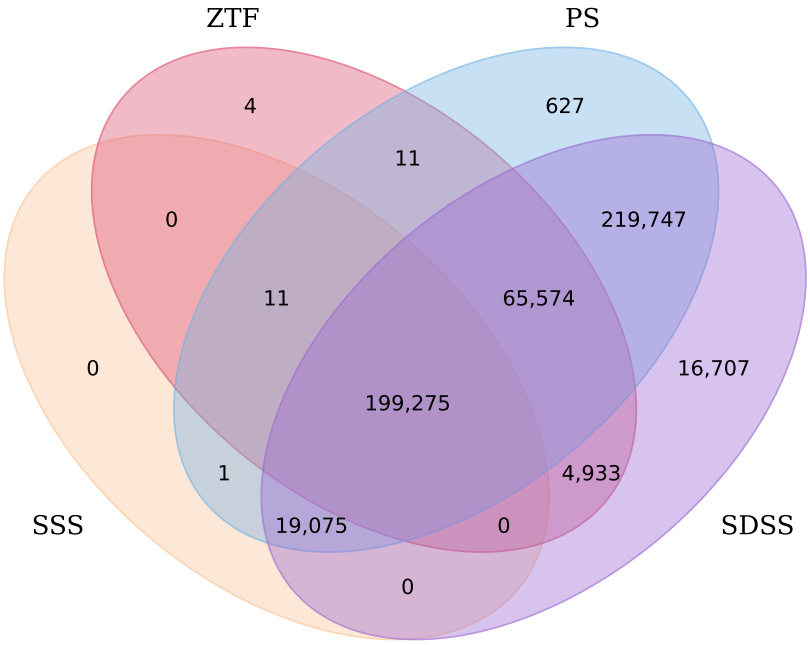

Figure 3.3: 4-way Venn diagram showing the number of 7-DQ quasars which contain at least one observation in each combination of surveys after the data cleaning steps outlined in Section 3.2. The majority of quasars are present in either all four surveys, or just Pan-STARRS and SDSS

3.3 Pairwise dataset

3.3.1 Motivation

The focus of my studies involves characterising quasar variability as a function of time-lag, Δt; I am not concerned with changes in variability over *absolute* time. Likewise, I aim to quantify this variability in terms of magnitude differences Δm; I am not concerned with how variability changes as a function of *apparent* magnitude. In Chapters 4 and 5, I will consider pairs of observations that are

separated into bins of Δt and calculate statistics from the resulting magnitude distribution, such as the structure function. By grouping together pairs of observations which are separated by small time lags, I may investigate variability on short timescales. Conversely, by grouping together pairs of observations with a large time gap, I can probe variability on long timescales. Note that I use the terms timescale and time-lag interchangeably, and will often use both of these terms when referring to Δt, depending on the context.

To speed up computation, I decided to carry out a preprocessing step to generate a secondary dataset which is a derivative of 7-DQ, whereby I take the unique pairs of observations within all light curves and compute their magnitude change, time separation and combined photometric error. This dataset, which I will refer to as the pairwise dataset, acts as the starting point for the analysis in Chapters 4 and 5. This dataset was generated for computational convenience and does not affect the results. Representing the light curve as a collection of pairs is the same approach used to calculate the structure function, explained in Chapter 1, Section 1.8.

I use the pairwise dataset for the quasars and stars in two different ways. Firstly, I use Δm pairs derived from the 7-DQ quasar photometry to quantify quasar variability. Secondly, I utilise Δm pairs derived from the 7-DQ star photometry to constrain uncertainties and systematic effects caused by the individual surveys and my merging of them. Such uncertainties include the combination of photometric error from the respective surveys that a particular Δm is derived from, while systematic effects include biases introduced during the colour transformation process.

For a light curve of length N, there are $N(N+1)/2$ unique unordered $(\Delta m, \Delta t)$ pairs. For any one quasar, the distribution of Δt will be sparse and gappy, but since Δm is mostly independent of the magnitude m of the quasar, we can combine Δm's of many quasars, for a given range of Δt, into an ensemble. These distributions carry important information about quasar variability on different timescales, which is investigated in Chapter 4.

Quasars are distant sources and are therefore embedded in the Hubble flow. Thus, we must correct the observed-frame variability to the rest-frame via

$$\Delta t = \Delta t_{\rm obs}/(1+z), \tag{3.1}$$

where redshifts are provided by DR14Q. Using rest-frame time differences has the additional benefit of smoothing the Δt distribution to produce bins with more consistent number counts. For the remainder of my thesis, in the context of quasars, Δt will refer to rest-frame time-lag unless specified otherwise. For stars, Δt remains in the observer-frame.

3.3.2 Construction

Let $\boldsymbol{L} = [\boldsymbol{l_1}, \ldots, \boldsymbol{l_K}]$ be a list of light curves for K total photometric sources. The light curve for object k, denoted $\boldsymbol{l_k}$, can be considered a vector with n_k measurements (which I refer to as its length, not to be confused with temporal length). Note that, since light curve length varies between objects, n depends on k. The i^{th} measurement in $\boldsymbol{l_k}$ contains the following quantities; time, t_i, magnitude, m_i, magnitude error, σ_i, and survey ID, v_i, for each observation in the light curve, i.e., $\boldsymbol{l_k} = [(t_{1,k}, m_{1,k}, \sigma_{1,k}, v_{1,k}), (t_{2,k}, m_{2,k}, \sigma_{2,k}, v_{2,k}), \ldots, (t_{n,k}, m_{n,k}, \sigma_{n,k}, v_{n,k})]$. $\boldsymbol{l_k}$ is sorted such that t_i is monotonically increasing.

I will now explain how to construct the relevant pairs from a particular light curve $\boldsymbol{l_k}$. Since this process is independent of k, I will omit the subscript for brevity, hereafter. In order to construct the relevant pairs from a light curve, $\boldsymbol{l}$, I find the unique combinations of $(t_i, m_i, \sigma_i, v_i)$ and $(t_j, m_j, \sigma_j, v_j)$ for $j < i$, and perform the appropriate operations on each quantity.

First, for $\boldsymbol{m}$ and $\boldsymbol{t}$, the appropriate operation is simply the difference, i.e., $m_i - m_j$ and $t_i - t_j$. Second, I defined the operation on $\boldsymbol{\sigma}$ to be a quadrature sum, i.e., $\sigma_i^2 + \sigma_j^2$. I decided that this would be the best approximation of the error on the value $m_i - m_j$, as it corresponds to summing the photometric variances (described previously in Chapter 1, Section 1.8). Finally, I defined the operation for $\boldsymbol{v}$ to be multiplication, i.e., $v_i \times v_j$. As values of survey ID are prime (see Section 3.2.3), the result of this operation is defined uniquely, which enables me to quickly identify from which surveys the pair came from. While there are other methods to achieve this, multiplication is computationally the fastest and most memory-efficient.

A naïve approach to find, then operate on, unique non-ordered pairs would be to first loop over each observation, and then over each subsequent observation. This process is described in Algorithm 1. However, this is computationally slow and unpractical due to the sheer size of my database. I developed an algorithm, implemented in `Python` using the standard `numpy` package. My algorithm employs

vector operations to greatly increase the speed of construction of pairs, shown in Algorithm 2. Its functionality is explained as follows: The operation $\boldsymbol{m} \ominus \boldsymbol{m}$ represents a subtraction between every unique combination of pairs of points in $\boldsymbol{m}$ to generate a matrix with elements $m_i - m_j$, while $\oplus$ is the equivalent operation except for addition. In practice, these two operations are performed by casting $\boldsymbol{m}$ to a matrix (i.e., adding a new axis to the `numpy.array`) and adding or subtracting a transposed copy of itself. The operation $\boldsymbol{v} \otimes \boldsymbol{v}$ has its usual definition as an outer product and generates a matrix with elements $v_i \times v_j$. These operations, $\ominus$, $\oplus$, and $\otimes$, are vectorised and therefore much faster than the two for-loops of the naïve algorithm. Since these operations generate duplicate pairs (i.e., $m_1 - m_2$ and $m_2 - m_1$, when only one of these is actually needed), unique pairs are picked using the indices of $\boldsymbol{C}$, representing the upper triangle of the matrix. I tested the speed of my algorithm on my 7-DQ quasar photometry, using the naïve approach as a baseline, and found a huge speed increase, illustrated in Figure 3.4.

Algorithm 1 Naïve algorithm

Input: $\boldsymbol{L} = [\boldsymbol{l}_1, \ldots, \boldsymbol{l}_n]$

```
a ← []
for l ∈ L do
    b ← []
    for i ← 1, n do
        for j ← i + 1, n do
            dm ← m_i − m_j
            dt ← t_i − t_j
            dσ ← sqrt(σ_i^2 + σ_j^2)
            dv ← v_i × v_j
            b.append(dm, dt, dσ, dsid)
        end for
    end for
    a.append(b)
end for
return a
```

Algorithm 2 My algorithm

Input: $\boldsymbol{L} = [\boldsymbol{l}_1, \ldots, \boldsymbol{l}_n]$

$a \leftarrow []$
for $\boldsymbol{l} \in \boldsymbol{L}$ **do**
$\boldsymbol{m} \leftarrow [m_1, m_2, \ldots, m_n]$
$\boldsymbol{t} \leftarrow [t_1, t_2, \ldots, t_n]$
$\boldsymbol{\sigma^2} \leftarrow [\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2]$
$\boldsymbol{s} \leftarrow [s_1, s_2, \ldots, s_n]$
$\boldsymbol{C} \leftarrow$ `numpy.triu_indices`$(n, 1)$
$\boldsymbol{dm} \leftarrow \boldsymbol{m} \ominus \boldsymbol{m}$
$\boldsymbol{dt} \leftarrow \boldsymbol{t} \ominus \boldsymbol{t}$
$\boldsymbol{d\sigma^2} \leftarrow \boldsymbol{\sigma^2} \oplus \boldsymbol{\sigma^2}$
$\boldsymbol{dv} \leftarrow \boldsymbol{v} \otimes \boldsymbol{v}$
a.append(dm_{ij}, dt_{ij}, $\sqrt{d\sigma_{ij}^2}$, ds_{ij} for $i, j \in C$)
end for
return a

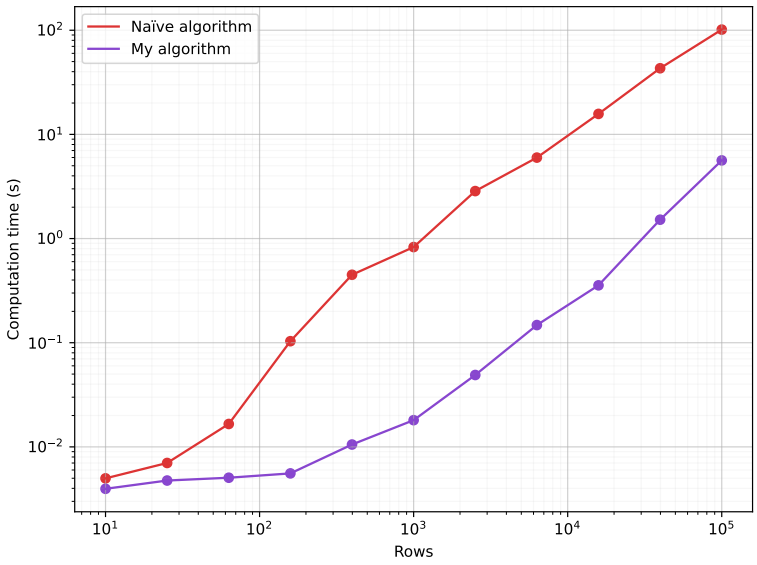

Figure 3.4: Computational speed of my algorithm, compared to a naïve approach as a baseline, as a function of rows. Here, rows refers to the number of rows of real data taken from the 7-DQ quasar photometry. My method is over an order-of-magnitude faster when processing files with > 100 rows.

3.3.3 Summary of pairs

I present an example of data from the pairwise dataset in Table 3.3. Additionally, in Table 3.4, I show a summary of the number of pairs that make up the pairwise

dataset for the 7-DQ quasars and stars in the g, r and i bands, to illustrate its immense size. Finally, in Table 3.5, I show the number of pairs formed from each combination of surveys for the g, r and i bands, for the quasar photometry.

uid	dm	dt	de	dsid
1	0.178986	133.19026	0.252087	49
3	-0.025074	112.04023	0.135006	49
3	0.002586	209.57025	0.101637	49
3	-0.250851	412.65503	0.100035	49
3	0.027660	97.53003	0.133875	49
3	-0.225777	300.61480	0.132663	49
3	-0.253437	203.08480	0.098503	49
4	0.169956	2511.08180	0.239570	21
4	0.188280	2511.46680	0.238982	21
4	0.134626	2517.16720	0.255475	21
...	...	...	...	...

Table 3.3: An example of the data from the pairwise dataset for r-band observations of the 7-DQ quasars. Each row represents a unique pair. Here, `uid` is my unique quasar identifier to specify which quasar the pair came from, `dm` is the magnitude difference (equivalent to Δm) `dt` is the time difference (equivalent to Δt). For quasars, this is transformed to the rest-frame, whereas for stars, it is left in the observer-frame. `de` is the root-sum-square of the photometric errors, and `dsid` is the product of the survey IDs.

	Nº pairs in g-band	Nº pairs in r-band	Nº pairs in i-band
Quasars	4,959,745,834	7,001,145,864	362,894,257
Stars	835,486,406	2,416,938,570	231,827,462

Table 3.4: Total count for number of pairs calculated from photometry of quasars and stars in the 7-DQ database.

Survey combination	Number of pairs		
	g-band	r-band	i-band
SSS-SSS	84,601	174,218	5,798
SSS-SDSS	405,253	397,488	32,983
SSS-PS	1,344,862	1,545,208	154,336
SSS-ZTF	42,099,734	63,621,981	1,365,843
SDSS-SDSS	2,705,746	2,848,155	2,644,237
SDSS-PS	2,709,202	3,471,226	3,787,033
SDSS-ZTF	48,529,353	61,150,318	6,613,372
PS-PS	3,424,657	4,854,153	6,913,230
PS-ZTF	168,019,874	241,167,030	39,727,966
ZTF-ZTF	4,690,422,446	6,621,915,981	301,649,353

Table 3.5: Number of pairs in each band for the quasar photometry, split into survey combinations.

3.4 Pooling statistics

The volume of data points in my pairwise dataset (see Table 3.4) exceeds 500GB for the quasars alone, and therefore it is not practical to load all this data into memory in order to calculate ensemble statistics used later in Chapter 4 and 5. Therefore, I decided to split the full ensemble into smaller subsamples and calculate statistics within these subsamples. However, in order to recover the correct full ensemble statistics I used the law of total expectation and the law of total variance, explained by the following example. If Z represents the ensemble, which may be represented by the union of two subsamples X and Y, then

$$X = \{x_1, \ldots, x_N\} \tag{3.2}$$

$$Y = \{y_1, \ldots, y_N\} \tag{3.3}$$

$$Z = \{x_1 \ldots, x_N, y_1, \ldots, y_N\}. \tag{3.4}$$

The law of total expectation states that the mean of the ensemble is simply the mean of the means of the subsamples,

$$\mu_z = \frac{\mu_x + \mu_y}{2}. \tag{3.5}$$

While the law of total variance states that the variance of the ensemble is the mean of the variances plus the variance of the means.

$$\sigma_z^2 = \frac{1}{2}(\sigma_x^2 + \sigma_y^2) + \left(\frac{\mu_x - \mu_y}{2}\right)^2. \tag{3.6}$$

In my case, I dissected the full ensemble into 106 subsamples and processed them individually before calculating statistics of the full ensemble, which are used in Chapters 4 and 5.

Chapter 4

The Δm distribution, its moments and their evolution with time

4.1 Introduction

In this chapter, I will investigate quasar variability by looking at the shape and moments of Δm distributions from the pairwise dataset (whose construction is outlined in Chapter 3 Section 3.3). I refer generally to the full set of magnitude differences as Δm. In practice, Δm is calculated for each of the $m = g, r, i$ bands individually. This is reflected in my results where analysis is usually repeated for each band, for both this chapter and Chapter 5. Therefore, the reader should note that Δm implicitly implies Δg, Δr and Δi, unless specified otherwise. By definition, the pairwise database does not hold any more information than the photometry in 7-DQ. Nevertheless, it provides a computationally convenient means to investigate quasar variability in large ensemble studies. In my case, given the size of 7-DQ, the pairwise database is essential in order to perform analysis such as ensemble structure function studies, on different timescales across various quasar properties, effectively and efficiently. The pairwise dataset forms the basis of my analysis in this chapter and Chapter 5.

The 'Δm distributions' (which is how I refer to them hereafter) are the result of grouping the full set of Δm pairs in various ways. The trivial way to group Δm pairs is into bins of Δt, similarly to how the structure function is grouped into Δt so that it is a function of time lag. Due to the sheer number of pairs within the

pairwise dataset, more complex groupings are possible, e.g., grouping into Δt and luminosity simultaneously. In this chapter, the trivial Δt groupings are considered, while analysis of complex groupings are saved for Chapter 5. An outline of this chapter is as follows: In Section 4.2 I present the Δm distributions for the quasars and stars from the pairwise dataset in the r band. In Section 4.3 I characterise the shapes of these distributions, including an original fitting technique. In Section 4.4, I compute moments of these distributions to investigate the behaviour of quasar variability over different timescales, in the g, r and i bands.

4.2 The Δm distribution of Quasars and Stars

4.2.1 Methods and Results

To obtain each Δm distribution, I grouped the full set of pairs into bins of Δt. However, even after this process, there are still a huge number of points per bin (see Table 4.1 for an example in the r-band). It was not practical to load all this data into memory for the distributions to be plotted. Therefore, I calculated the bin-counts iteratively, loading each portion of data, binning it, and summing the bin counts cumulatively. For this method, it was necessary that I predefine a set of bins for Δm and Δt.

For Δm, I opted for 200 bins over the range $[-2, 2]$ (mag). I found that a set of bins with linearly spaced edges could not simultaneously give me the fine resolution required around $\Delta m = 0$, as well as a coarse resolution in the tails of the distribution. Therefore, I used a custom set of bins which are logarithmically spaced from zero, in both positive and negative directions. This resulted in a minimum bin width of 0.00485 mag at the centre, and a maximum width of 0.05 mag at the edges. For Δt, I experimented with a various number of bins, from $10-20$, but I found that 14 was an optimal number resulting in gradual, but visible changes between the resulting Δm distributions. These bins cover the full extent of possible Δt in the data ($0 < \Delta t < 26{,}063$ days). The edges of these bins are also logarithmically spaced, as this is the natural scale for analysing quasar variability over time. Additionally, this allows for comparison with subsequent plots in Section 4.4, in addition to plots in Chapter 5, later in this thesis, all of which use a logarithmic scale of Δt. Table 4.1 shows my chosen set of Δt bins and the corresponding number of pairs in each bin, for both the quasars and stars

in the r-band. Furthermore, I use the same set of Δt bins when presenting Δm distributions in Figures 4.1, 4.3 and 4.4 for consistently, as well as to allow direct comparison between figures. I present histograms of the Δm distributions, in the r-band, grouped by Δt, from the pairwise dataset for the 7-DQ quasars and stars in Figure 4.1.

Δt bin [days]	Number of pairs in r-band	
	Quasars	Stars
$0 < \Delta t < 5$	202,803,880	20,170,313
$5 < \Delta t < 10$	181,806,672	21,542,983
$10 < \Delta t < 21$	333,966,423	44,633,221
$21 < \Delta t < 39$	398,442,175	68,995,998
$39 < \Delta t < 73$	571,683,055	107,877,353
$73 < \Delta t < 138$	1,372,414,045	128,349,932
$138 < \Delta t < 282$	2,068,294,615	182,091,211
$282 < \Delta t < 532$	1,243,693,629	522,374,221
$532 < \Delta t < 1{,}003$	382,516,418	587,229,966
$1{,}003 < \Delta t < 1{,}894$	126,514,941	360,185,279
$1{,}894 < \Delta t < 3{,}872$	77,574,618	103,748,513
$3{,}872 < \Delta t < 7{,}311$	26,052,899	216,021,542
$7{,}311 < \Delta t < 13{,}803$	12,858,969	42,061,942
$13{,}803 < \Delta t < 26{,}063$	2,523,419	11,655,990

Table 4.1: Number of pairs in the r-band for the quasars and stars in the pairwise database after grouping into bins of Δt. These Δt bins match up with the panels of Figures 4.1, 4.3 and 4.4 for a direct comparison. I have only shown the number of r-band pairs for brevity, although the number of g-band pairs is comparable. However, the number of i-band pairs is considerably less due to the limited availability of ZTF i-band observations.

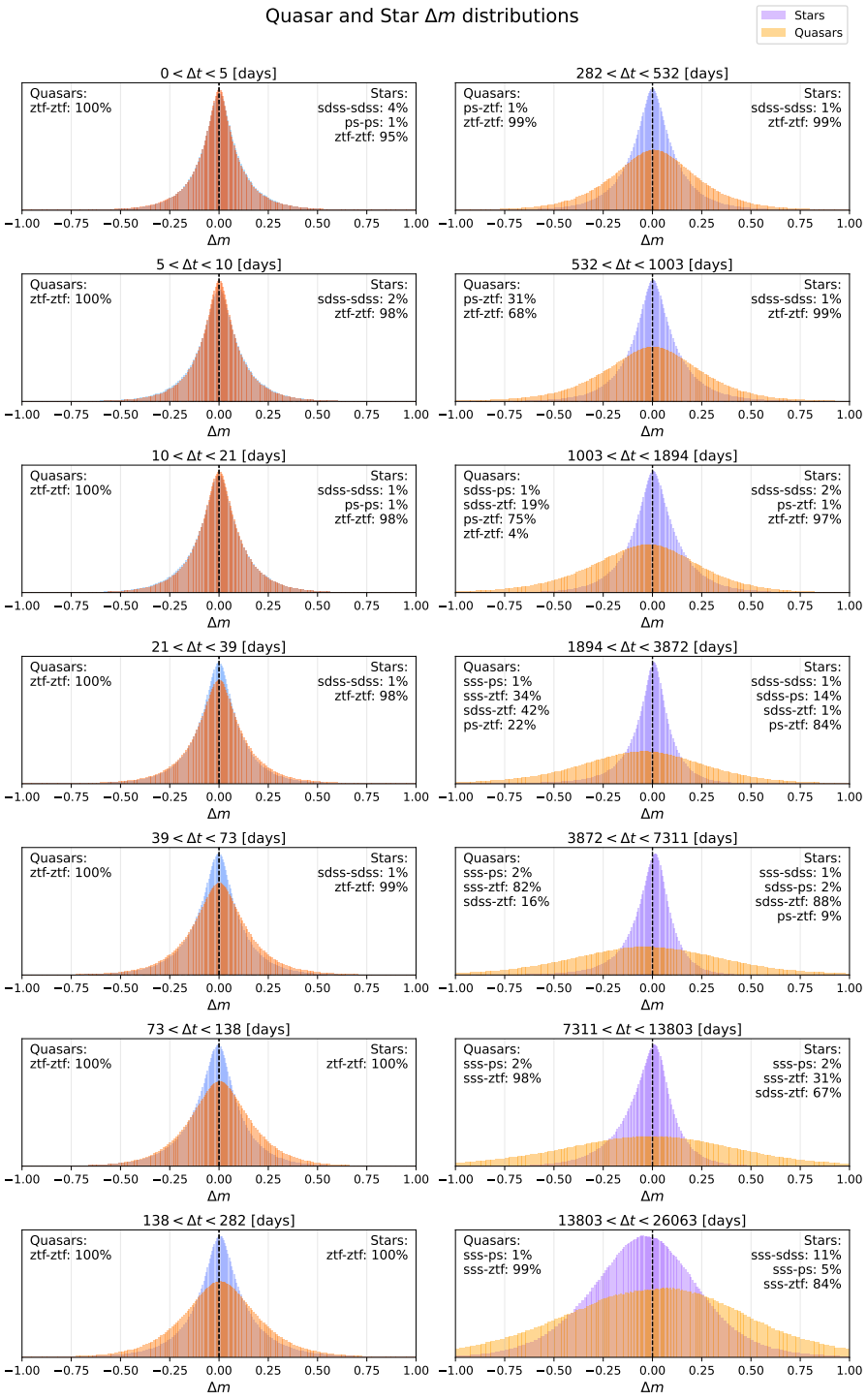

Figure 4.1: Δm distributions grouped by increasing time-lags (top to bottom, left to right) for the ensemble quasar (orange gradient) and star (blue gradient) population. The histograms are coloured using a gradient that changes progressively with Δt, to aid the eye. Each panel is annotated with the percentage of pairs from a specific survey combination compared to the total number of pairs. The two distributions have been normalised by their integrals so that their shapes may be directly compared. Note that, because the quasar Δt have been transformed to the rest frame, the survey combination fractions differ to the stars.

4.2.2 Discussion

Since the 7-DQ stars are non-variable sources, repeat measurements of a given star should return the exact same magnitudes, regardless of the time interval between observations. In this idealistic framework, the resulting Δm distributions would effectively be a δ-function. The fact that the observed star Δm distributions are clearly not δ-functions represents the accumulation of uncertainties from many sources throughout the full photometric pipeline (discussed in detail in Chapter 3, Section 3.2). Nevertheless, the total uncertainty on a particular Δm is dominated by the photometric errors of magnitude measurements that contributed to that particular Δm value. Since a particular Δm pair can either originate from two measurements in the same survey, or one measurement from each of two surveys, there are two cases: First, the pair originates from the same survey, and the uncertainty in Δm is determined by the photometric error distribution of that survey. Second, the pair is formed between two different surveys, and the uncertainty in Δm is determined by the combination of photometric error distributions for those surveys (more specifically, the error distribution of Δm is a convolution of the photometric distributions of the two surveys).

This interpretation explains the form of the star Δm distributions in Figure 4.1. Their shape, including width, skewness and zero-offset of the peak, are governed by the set of contributing pairs from each survey combination to the total distribution. Therefore, I have annotated the fraction of pairs from each survey combination on each panel of Figure 4.1 (including panels in Figures 4.3 and 4.4, shown later in Section 4.3). For example, for distributions where $\Delta t < 282$ days, the pairs are formed solely within ZTF. Therefore, the width, symmetry, and zero-offset (which is negligible in these distributions) are all the same. However, for distributions $\Delta t >282$ days, pairs are formed between several different combinations of surveys. Since each survey has its own photometric noise distribution, the widths are variable, particularly for the longest timescales which involve SuperCOSMOS−ZTF comparisons. Additionally, the combination of data from surveys with differing photometric systems, as well as uncertainties and introduced during the colour transformation process, result in a slight skewness and non-zero peak offset of the distributions. The star Δm distributions are therefore important in understanding the extent of bias and uncertainty in the data, which must be carried forward when interpreting the quasar Δm distributions. Comparison of the star and quasar distributions in Figure 4.1 illustrates clearly that quasars variations are detected on timescales $\Delta t > 20$ days.

Additionally, the amplitude of variability (width of the distribution) increases with Δt, as expected. For small time-lags, ($\Delta t < 20$ days), the quasar and star distributions are very similar, as observed variability is dominated by photometric noise.

4.3 Shape of the quasar Δm distribution

4.3.1 Exponential approximation

MacLeod et al. (2012) (M12 hereafter) calculated Δm pairs for a sample of 33,881 spectroscopically confirmed quasars using repeated SDSS and POSS imaging. They grouped their Δm values into three time bins of $365 - 730$, $730 - 1500$, and $1500 - 5844$ days, and show that the resulting Δm distributions follow an exponential distribution over these timescales. M12 claim that each quasar is accurately modelled by a damped random walk (DRW; see Chapter 1, Section 1.9.3). This process, driven by Gaussian noise, is described fully by a variability amplitude, $\sigma_{\rm DRW}$, and characteristic timescale, $\tau_{\rm DRW}$. With this reasoning, the Δm distribution of a single quasar should be a Gaussian whose width depends on $\sigma_{\rm DRW}$, and variation with timescale depends on $\tau_{\rm DRW}$. Thus, M12 claim that their exponential distributions arise naturally by combining Δm measurements of many quasars into an ensemble, that are individually well described by a Gaussian process, specified by a particular combination of $\sigma_{\rm DRW}$ and $\tau_{\rm DRW}$ within a broad distribution. However, M12 determined this qualitatively without an explanation of why an ensemble of Gaussian distributions results in an exponential, and for which regimes this is true. In this section, I will study this in more detail using Monte-Carlo simulations.

The process of 'ensembling' (i.e., combining) Δm measurements is equivalent to taking the union of sets generated by Gaussian random variables. M12 states that the Δm distributions for each quasar are represented by a Gaussian. In the case of two quasars, A and B, the ensemble distribution, $p_E(\Delta m)$ may be written

as,

$$p_A(\Delta m) \sim \mathcal{N}(0, \sigma_A^2) \tag{4.1}$$

$$p_B(\Delta m) \sim \mathcal{N}(0, \sigma_B^2) \tag{4.2}$$

$$p_E(\Delta m) = p_A(\Delta m) \cup p_B(\Delta m). \tag{4.3}$$

It should be noted that this is not the equivalent of adding the random variables p_A and p_B, which would result in p_E also being distributed normally. Instead, the probability density function (PDF) of p_E is the sum of the PDFs of p_A and p_B. I verified this using Monte-Carlo simulations in Figure 4.2. In this simulation, I took the ensemble of 10 Gaussian distributions with 10,000 samples drawn per Gaussian. Each Gaussian had a width varying linearly from a lower bound, $\sigma^2_{\rm min}$, to an upper bound $\sigma^2_{\rm max}$. I tested two regimes where $\sigma^2_{\rm min} = 0.1$ and $\sigma^2_{\rm min} = 0.5$. In each case, the upper bound was fixed $\sigma^2_{\rm max} = 1$. When there is a larger spread of Gaussian variances, e.g., for $\sigma^2_{\rm min} = 0.1$, a single exponential is a better fit than a Gaussian. Conversely, for a smaller spread of Gaussian variances, $\sigma^2_{\rm min} = 0.5$, a single Gaussian provides a better fit. However, in both cases, the true ensemble distribution is obtained by summing the individual Gaussian PDFs.

Following this analysis, I decided to fit both exponential and Gaussian PDFs to the quasar Δm distributions. The results are shown in Figure 4.3. The two regimes are clear: on short timescales, the exponential PDF is a reasonably good fit (except at $\Delta m = 0$), while on long timescales, the Δm distribution is modelled well by a single Gaussian. It is quite striking that only a single Gaussian is needed to accurately describe the ensemble of difference observations of quasars on long timescales. One might expect this result from the central limit theorem. However, as mentioned previously, taking the ensemble of many Δm values is *not* equivalent to adding random variables, (i.e., convolving the distributions of Δm from individual quasars) which would result in a Gaussian.

Outside the short and long timescale regimes, the quasar Δm distribution is not described well by either the exponential or Gaussian distributions. The following section will introduce an alternative approach that has not yet been applied to these distributions, and provides a far better fit.

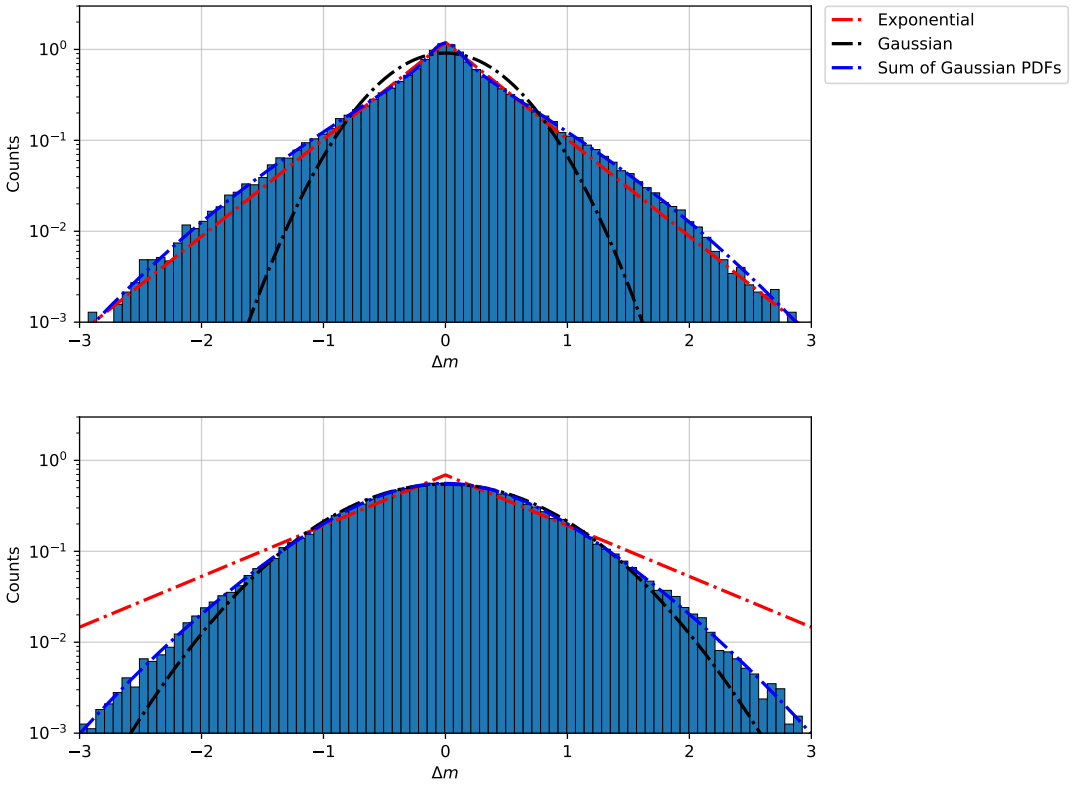

Figure 4.2: Ensemble Δm distributions simulated by combining individual Δm distributions, each modelled as a Gaussian of varying widths from σ^2_{min} to σ^2_{max}. $\sigma^2_{\text{min}} = 0.1$, 0.5 for top and bottom panel, respectively.

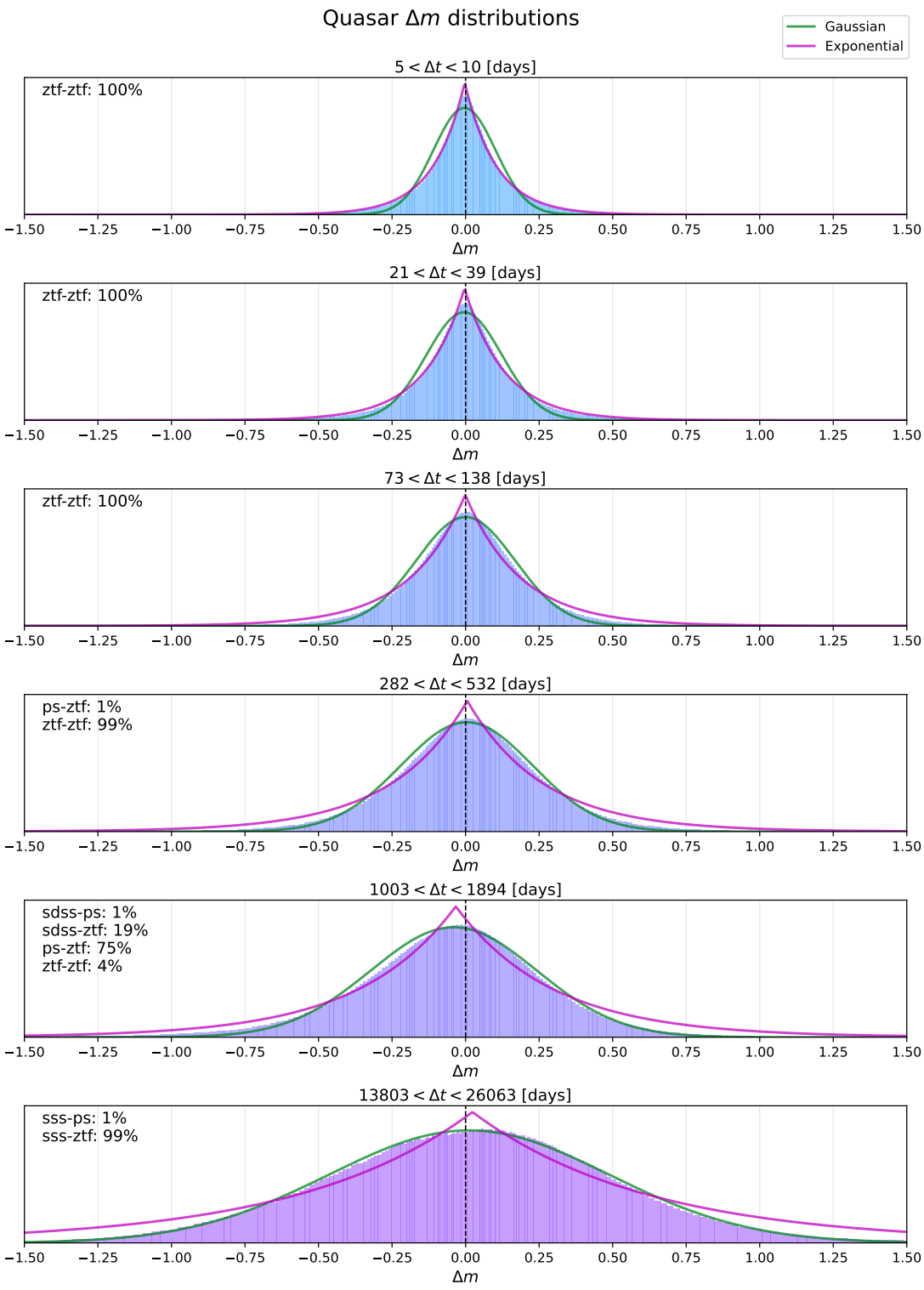

Figure 4.3: Δm distributions for pairs of observations with increasing time-lags for the ensemble quasar population, fitted with exponential and Gaussian probability density functions. Note that the usual colour scheme (orange for quasars) has been changed, as I found it to be clearer when over-plotting the fits. I selected a ranging subset of Δt bins to illustrate the fits.

4.3.2 Gaussian Mixture Models: a better fit

Previously, I demonstrated that an exponential and a Gaussian are an adequate fit for short and long timescales, respectively. However, I found that a Gaussian mixture model provided a much better fit to the data. Mixture models, in general, combine multiple components into a single probability density function. They are a natural statistical model for many situations in astronomy, particularly in cases with complicated distribution functions that cannot be parameterised with a specific model. Gaussian Mixture Models (GMMs) are a special case of mixture models, and are usually the first choice when few priors are known about the data distribution, as is the case here. The likelihood of a data point x_i for a GMM is given by

$$p(x_i \mid \boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \sum_{j=1}^{M} w_j \mathcal{N}(\mu_j, \sigma_j), \tag{4.4}$$

where dependence on x_i comes via a Gaussian $\mathcal{N}(\mu_j, \sigma_j)$. Here, w_j, μ_j, and σ_j are the weights, means, and widths (i.e., standard deviations) of the mixture model, respectively. The weights are defined such that

$$0 \leq w_j \leq 1, \quad \sum_{j=1}^{M} w_j = 1. \tag{4.5}$$

The log-likelihood for the whole dataset is then

$$\ln \mathcal{L} = \sum_{i=1}^{N} \ln \left[\sum_{j=1}^{M} w_j \mathcal{N}(\mu_j, \sigma_j) \right]. \tag{4.6}$$

To calculate the parameters of the mixture model, I used the `sklearn.mixture.GaussianMixture` routine within the `scikit-learn` Python package, which implements the expectation-maximisation algorithm for fitting. I used this routine to fit each Δm distribution independently and was able to achieve an almost perfect fit with only 3 components (i.e., $M = 3$) on all observed timescales. This is a novel application of GMMs to the Δm distributions from observations of quasars. The Δm distributions of the r-band and their fits are shown in Figure 4.4. Since the fits of the g and i band show the same qualitative result, they are not shown.

The fits in Figure 4.4 are clear evidence that difference observations, in the

optical/UV continuum of quasars, are accurately modelled by the mixture of only a few Gaussian distributions (3 in my case), over timescales from 5 days to 70 years in the rest-frame. By looking at the components of the GMM for each panel, it can be seen that at least 3 Gaussians with contrasting widths are necessary to fit the short timescales. Conversely, fewer Gaussians (or multiple with similar widths and weights) are required to fit on long timescales. In fact, I found that as few as two Gaussians would provide an accurate fit, of a similar quality to three, on timescales > 1 year.

However, when performing this fit, I found that the GMM model is degenerate in certain regimes; the fit is not unique and the weights and widths of each component can vary while still providing a similar quality of fit. This was evident by the fact that repeat fits would sometimes result in different weights and widths of the Gaussian components. However, despite degeneracies in the components, the overall GMM would be nearly identical. These regimes are determined by the number of components, M, and the average timescale, $\langle \Delta t \rangle$, of the distribution. By testing some of these cases, I found that the GMM fit is degenerate in any of the following regimes:

1. $M \geq 5$ and $\langle \Delta t \rangle < 100$ days
2. $M \geq 4$ and 100 days $< \langle \Delta t \rangle < 10$ years
3. $M \geq 3$ and $\langle \Delta t \rangle > 10$ years

Therefore, this effect is only significant when performing a fit with more than 3 Gaussian components, although it is exacerbated for distributions with longer timescales. The cause of these degeneracies is overlap in the Gaussian components ($\mu_j \approx 0$ for all j).

As mentioned previously, M12 claimed that their exponential distributions can be explained by the ensemble of Δm measurements from many quasars, if individual quasars have Gaussian Δm distributions of varying widths. My results are consistent with this, and generalises further by showing that the range of widths becomes smaller with time-lag such that ensemble difference observations are distributed as a single Gaussian beyond ~50 years. Neither I nor M12 are able to say for certainty that the Δm distribution of each quasar is normally distributed. However, I have shown that only a maximum of 3 Gaussian components are required to model the ensemble on timescales $0 < \Delta t < 70$ years, which is a

surprisingly few given that I am combining measurements from ~500,000 quasars. It is likely that each component represents a broad group of quasars with similar properties, which causes the apparent difference in variability.

While my analysis of the GMM fits has been mostly qualitative, I believe that it is possible to carry out a quantitative analysis of the weights and widths of the Gaussian components of the GMM fit to extract physically meaningful information relating to variability. However, to do this, it would be necessary to constrain the degeneracies mentioned above. This would be possible using algorithms such as MCMC to explore the parameter space effectively. For example, using MCMC, one could obtain posterior distributions for the weights, widths, and means of the GMM components. These distributions could then be used to quantify errors on the parameters, which is essential to interpret them in the context of quasar variability. Furthermore, exploring joint probability distributions between parameters could help break the degeneracies mentioned earlier.

I attempted to fit the pre-binned Δm distributions using MCMC, however, the sheer number of points in the distribution (see Table 4.1) resulted in the process being extremely computationally expensive, resulting in slow iterations which prevented me from reaching convergence. For this idea to be explored further, additional techniques must be developed for reducing the data in order to achieve convergence on the posterior distributions. While beyond the scope of this thesis, it presents new and promising avenues for exploration, either by myself or the wider scientific community.

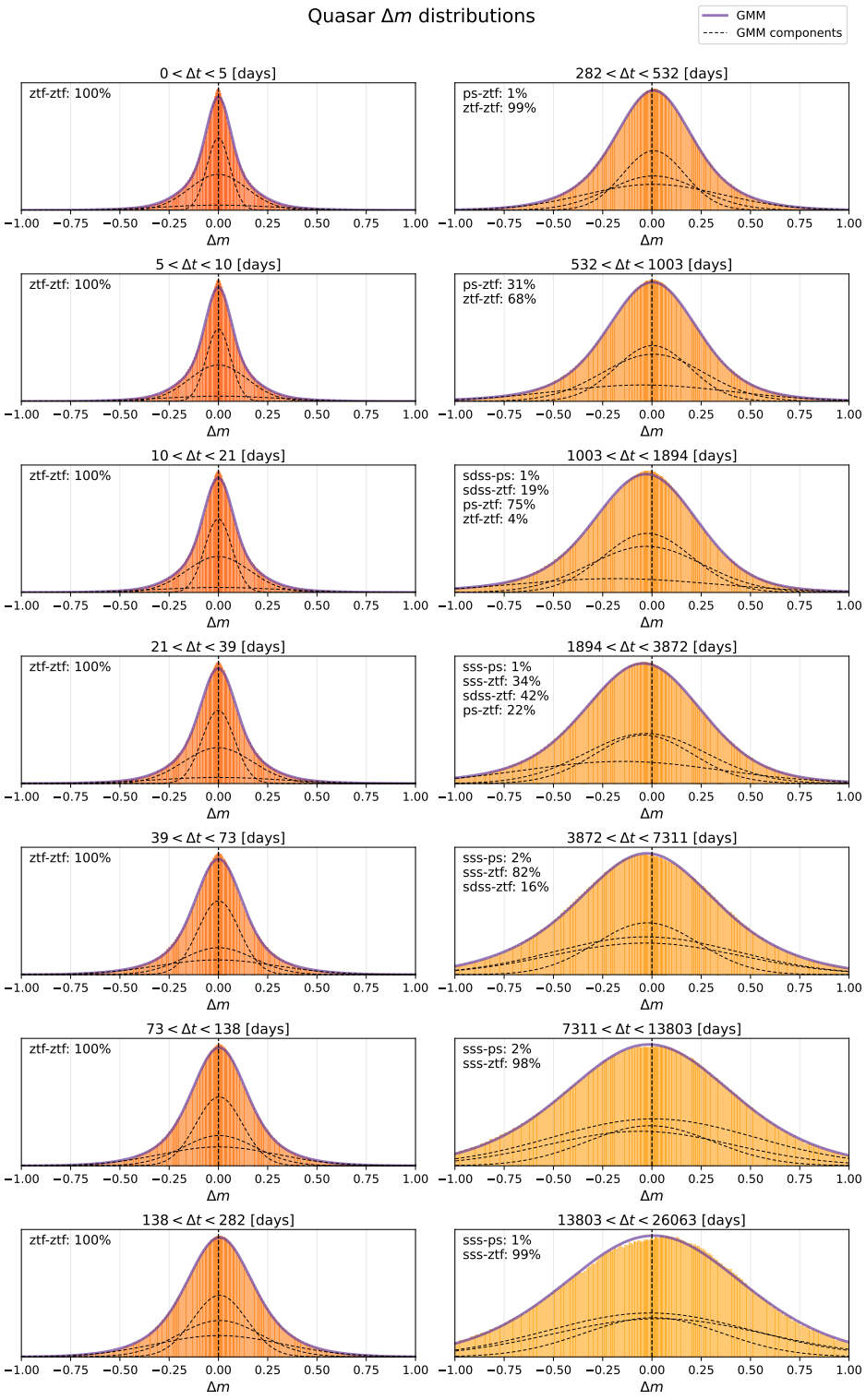

Figure 4.4: Δm distributions for pairs of observations with increasing time-lags for the ensemble quasar population, fitted with a Gaussian mixture model (green). The three components making up the Gaussian mixture are overplotted (dotted black lines). Panels should be read from top to bottom, right to left, which represents monotonically increasing bins of Δt.

4.4 Moments of the Δm distribution

Analysing the moments of Δm distributions is a novel method to investigate quasar variability. Other recent efforts to characterise long term quasar variability usually involves skipping straight to the structure function calculation, without explicitly looking at the distributions of Δm. The first moment, the mean, reveals bulk changes in apparent magnitude. The second moment, the variance, quantifies the amplitude of variability. The third moment, skewness, reveals the asymmetry of difference observations. Finally, the fourth moment, kurtosis, may be used to quantify the degree of non-Gaussianity of the difference measurements. In this section, I analyse the mean, skewness and kurtosis of the Δm distributions from the 7-DQ quasars and stars as a function of time. The variance, which is analogous to the structure function, is a common tool used by many others to characterise quasar variability and therefore its analysis is reserved for Chapter 5 where it will be studied in detail.

4.4.1 Mean

Long-term studies of AGN activity suggest large changes over Myr scales. Evidence for this can be seen in the large-scale extended structures emitted from AGN, such as the so-called "Voorwerp" objects (see e.g., Sartori et al. 2016 and references therein), which clearly show that some AGN exhibit order-of-magnitude changes in their luminosity over $10^4 - 10^5$ year timescales. However, there has been comparatively little research to investigate average luminosity changes in large samples of quasars, over timescales for which we have continuous monitoring. Quasar variability is often assumed to be a weakly stationary process, i.e., that the mean luminosity of large ensembles of AGN do not change with time. A few initial studies looked at AGN that were brightening or dimming during the period of observation, but found little or no evidence of differences in their variability properties (see e.g., de Vries et al. 2003, de Vries et al. 2005, Bauer et al. 2009, Voevodkin 2011). Studies that have observed bulk differences in mean brightness over a sample of quasars often attribute the result to bias; MacLeod et al. (2012) combined data from the Palomar Observatory Sky Surveys (POSS) and SDSS data and noted that objects from POSS are dimmer when observed in SDSS. They concluded that this may be explained by a Malmquist bias, i.e., the fact that a flux limited sample of variable objects will necessarily be dimmer in

the later survey even if there is no change in the mean brightness of the underlying sample. Rumbaugh et al. (2018) came to a similar conclusion when studying a sample of extreme variability quasars (EVQs). Morganson et al. (2014) found decrease in mean brightness on decade timescales when comparing SDSS and Pan-STARRS data, but attributed this to filter differences.

One of the most precise study of mean brightness changes to-date was conducted by Caplar et al. (2020). By comparing observations of ~6000 quasars from SDSS and Hyper Suprime-Cam (HSC), they claimed a consistent dimming of ~0.2 mag over a timescale of 10 years in the rest-frame. However, Shen & Burke (2021) claim that the result obtained by Caplar et al. (2020) is biased due to using a flux limited sample of objects. Clearly, this topic is contentious and there is no clear consensus of the behaviour of stationarity of luminosity in AGN.

If quasar variability is not a weakly stationary process, as is often assumed, we would expect the ensemble mean magnitude to drift with time as Caplar et al. (2020) observed. Using 7-DQ, I have tested this assumption and measured the mean magnitude drift over the ensemble population by calculating the first moment of the Δm distributions. If quasar variability is non-stationary, it would manifest in a significant non-zero mean of my Δm distributions, provided that the result cannot be attributed to bias.

I calculated the mean using two methods: First, I used all Δm pairs from every combination of surveys. I refer to the mean using this method as $\mu_{\rm all}$. Second, I used Δm pairs within surveys only, such that I can avoid biases introduced when comparing magnitudes between different photometric systems. I refer to the mean using this method as $\mu_{\rm inner}$. Note that there are two exceptions for the inner pairs: First, I included observation pairs between SDSS and Pan-STARRS because their photometric systems are very similar. Second, I omitted pairs within SuperCOSMOS. This was done because the photometry of SuperCOSMOS does not use the same photometric system; each survey within SuperCOSMOS has its own filter profile and plate emulsion combination. Furthermore, photographic plate data is inherently more noisy than the other surveys and does not have the required precision to effectively constrain the mean. The survey combinations used for the inner pairs are therefore: $\Delta m_{\rm SDSS-SDSS}$, $\Delta m_{\rm PS-PS}$, $\Delta m_{\rm SDSS-PS}$ and $\Delta m_{\rm ZTF-ZTF}$.

In order to obtain a precise measurement of the mean, I use an optimal weighted average of Δm values,

$$\widehat{\mu}(\Delta t) = \frac{\sum_k \Delta m(\Delta t)_k \cdot w_k}{\sum_k w_k} \tag{4.7}$$

with weights defined as the variance of pair k, defined as sum of photometric variances from measurements i and j,

$$w_k = \sigma_k^{-2} = (\sigma_i^2 + \sigma_j^2)^{-1}. \tag{4.8}$$

The rms uncertainty on $\widehat{\mu}$ is then,

$$\sigma_{\widehat{\mu}} = \left(\sum_k w_k \right)^{-1/2}. \tag{4.9}$$

This process was repeated in the g, r and i bands, and the results are shown in Figure 4.5. $\mu_{\rm all}$ and $\mu_{\rm inner}$ for the stars are both consistent with zero on all timescales (with the exception of a few noisy points), which is reassuring, and confirms the effectiveness of my colour transformations. Note that $\mu_{\rm inner}$ for the stars is not shown in Figure 4.5 as it is visually identical to $\mu_{\rm all}$, except that it does not extend beyond $\Delta t > 5\times10^3$ days. For $\Delta t < 5\times10^2$ days, $\mu_{\rm all}$ and $\mu_{\rm inner}$ for the quasars are both consistent with zero, however, on longer time-lags, $\mu_{\rm all}$ starts to deviate from zero, then fluctuate erratically at the longest timescales of 10^4 days. Conversely, $\mu_{\rm inner}$ does not show these large variations on long timescales. Since $\mu_{\rm all}$ and $\mu_{\rm inner}$ only differ in the pairs used during calculation, differences between them can only be caused by bias introduced when comparing magnitudes from different surveys.

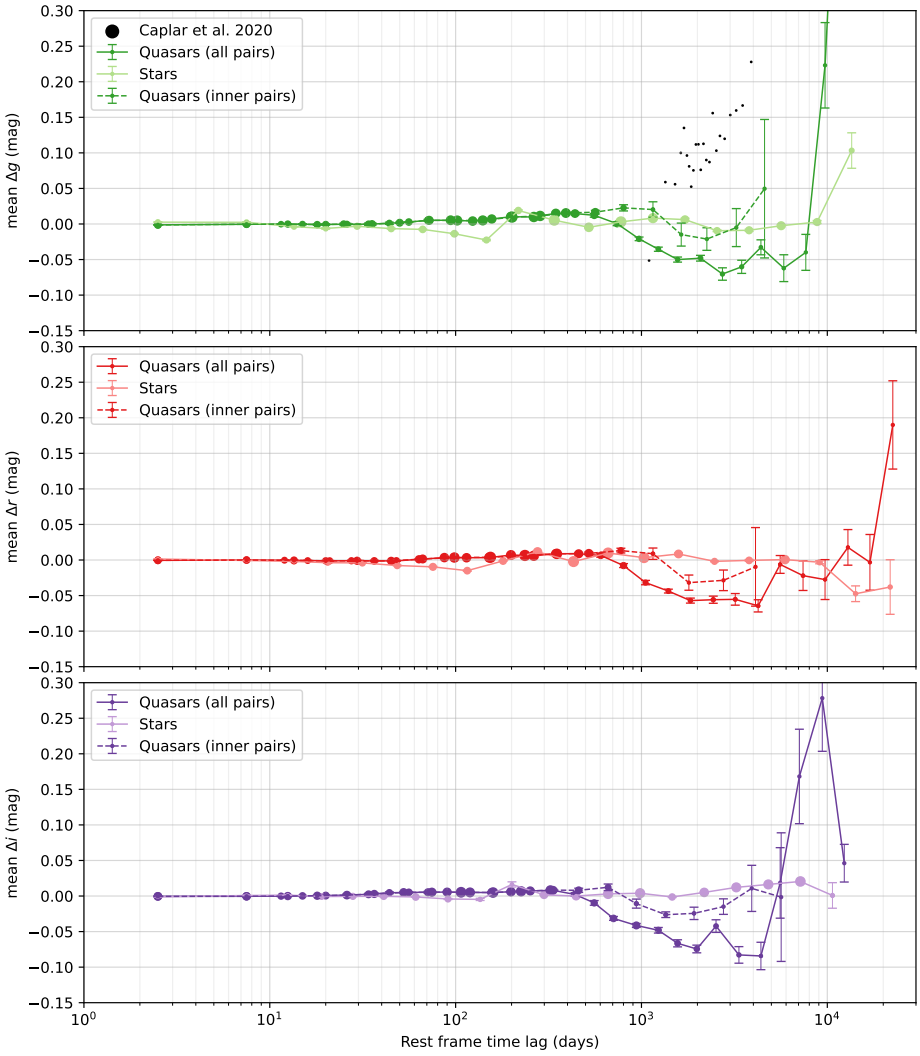

Figure 4.5: Mean magnitude drift for g, r and i bands for 7-DQ quasars (dark line) and stars (light line). The mean calculated using inner pairs for the quasars is also shown (dotted line). Data from Caplar et al. (2020) is overplotted on the top panel (black points). Note that the size of the circular markers illustrates the relative number of points in that bin. Note that $\mu_{\rm inner}$ is not shown for the stars to prevent overcrowding. Negative magnitude change corresponds to brightening.

The behaviour of my result $\mu_{\rm all}$ (for the quasars), along with the result obtained by Caplar et al. (2020) (overplotted on Figure 4.5), can be attributed to biases described by Shen & Burke (2021). Caplar et al. (2020) compared quasar photometry from SDSS to Hyper Suprime-Cam (HSC), with HSC having a fainter limiting magnitude, than SDSS. Their quasar sample was defined and observed using SDSS, with HSC data of the same objects being taken at a later date. Shen & Burke (2021) claim that, by using a deeper survey at a later date, the

mean of a flux-limited sample is biased such that dimming is more likely to be seen. My result, using $\mu_{\rm all}$, behaves in the opposite sense. Every survey combination (ignoring inner pairs) involves a shallower survey following a deeper survey, with the exception of SuperCOSMOS. For example, SDSS-PS, SDSS-ZTF, and PS-ZTF comparisons are all deep-to-shallow comparisons, which explains the apparent brightening implied by negative $\mu_{\rm all}$ on timescales of 3×10^3 days. However, pair combinations using SuperCOSMOS photometry involve shallow-to-deep comparisons and explains the sudden apparent dimming on the longest timescale (final data points in each g, r and i bands).

As a second test, to demonstrate that these effects are indeed caused by bias, I re-evaluated $\mu_{\rm all}$ using the bright subsample of quasars (defined in Chapter 3, Section 3.2.1). The result is shown in Figure 4.6, with $\mu_{\rm all}$ overplotted for comparison. I have only plotted the g-band for brevity, but the same effect applies in all g, r and i bands.

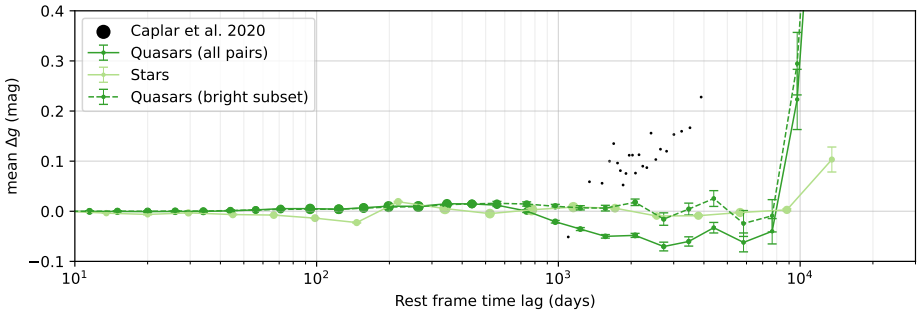

Figure 4.6: Comparing the mean magnitude drift calculated for all quasars, and the bright subset of quasars.

$\mu_{\rm inner}$ shows no significant change in mean magnitude on timescales $\Delta t < 15$ years. Since this is my most precise measurement of the mean, I conclude that there is no bulk change in quasar magnitude up to these timescales, and that the results of Caplar et al. (2020) are due to bias, consistent with Shen & Burke (2021). I am unable to constrain the mean effectively beyond these timescales, as $\mu_{\rm all}$ also suffers from the same bias.

4.4.2 Skewness

The skewness, S is the third moment of the Δm distribution and represents the overall asymmetry of the distribution. It is calculated via

$$S = \frac{\mu_3}{\mu_2^{3/2}}, \tag{4.10}$$

where μ_n is the n^{th} central moment of the distribution,

$$\mu_i = \frac{1}{N}\sum_{n=1}^{N}(x[n] - \bar{x})^i. \tag{4.11}$$

The standard error on the skewness depends only on the sample size:

$$\sigma_{\text{skew}} = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}}, \tag{4.12}$$

and is used for the error bars in Figure 5.3. For a Gaussian, $S = 0$. Distributions with a positive skew, $S > 0$, implies that the $\Delta m > 0$ tail is longer and therefore a higher probability of large dimming changes, while $S < 0$ implies that the $\Delta m < 0$ tail is longer and therefore a higher probability of large brightening changes. However, because the skewness is a dimensionless parameter, it is hard to physically interpret the exact value of S over a given timescale.

I found that skewness was particularly sensitive to differences in photometric systems of the surveys, evidenced by erratic fluctuations in the skewness of the star sample, when using the full set of pairs. Therefore, as with the mean, I decided to compute the skewness using restricted pairs ($\Delta m_{\text{SDSS–SDSS}}$, $\Delta m_{\text{PS–PS}}$, $\Delta m_{\text{SDSS–PS}}$ and $\Delta m_{\text{ZTF–ZTF}}$), which offers the most precise measurement of S. In Figure 4.7, I present the skewness of the quasar and star Δm distributions, in the g, r and i bands. This is a novel result, however, since this statistic has not yet been studied for quasar variability I present it tentatively as a reference that can be used when comparing results from different photometric studies or simulations. We see that the skewness of the 7-DQ quasars and stars are comparable for time-scales up to 1 year. This suggests fluctuations in brightness for quasar variability are equally probably for dimming and brightening over this timescale. For time-scales greater than 1 year, the data is not conclusive, and it is likely that inaccuracies of the colour transformations cause systematic errors. In Chapter 5 I use a different tool, the asymmetric structure function, to better constrain

asymmetry of quasar variability.

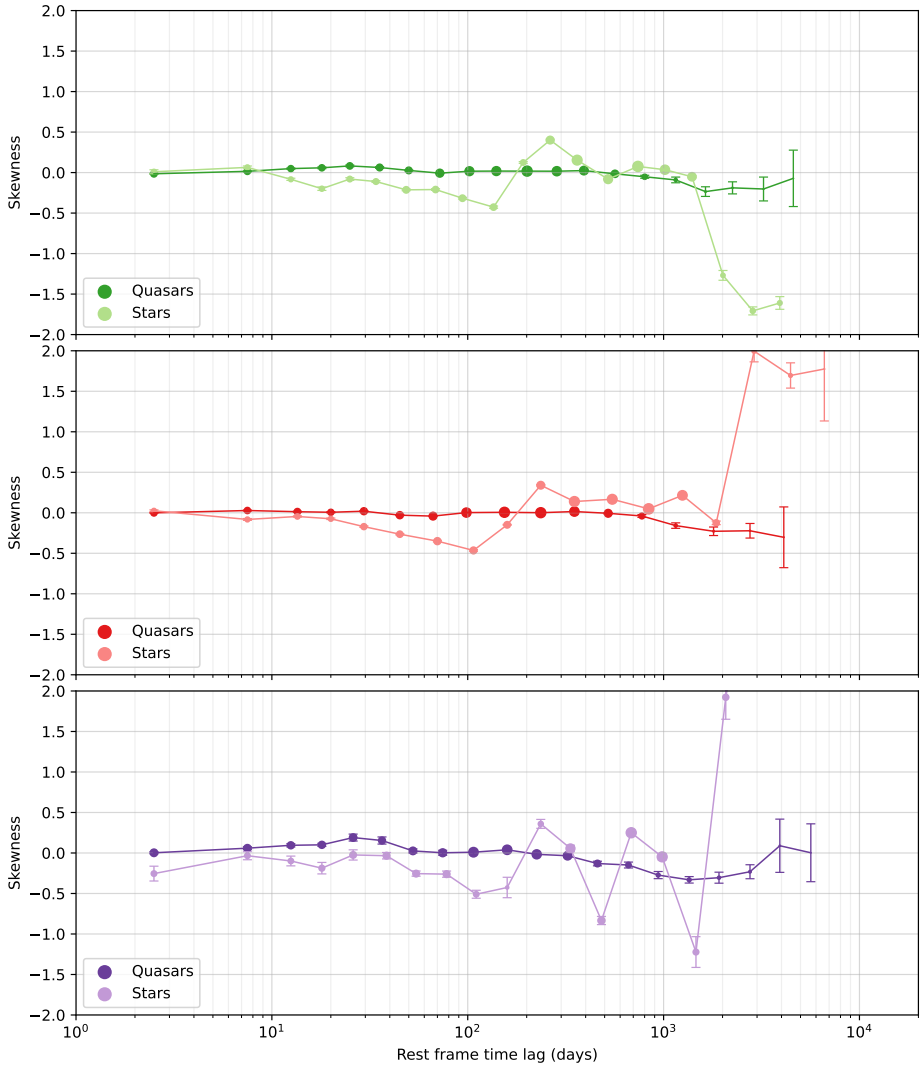

Figure 4.7: Skewness of Δm of 7-DQ quasars and stars in the g, r and i bands. Note that the size of the circular markers illustrates the relative number of points in that bin.

4.4.3 Kurtosis

The kurtosis, K, is the fourth moment of the Δm distribution, defined as:

$$K = \frac{\mu_4}{\mu^2}, \tag{4.13}$$

defined again in terms of the n^{th} central moment,

$$m_i = \frac{1}{N}\sum_{n=1}^{N}(x[n] - \bar{x})^i. \tag{4.14}$$

K quantifies the presence of tails in a distribution. For a Gaussian distribution, we expect $K = 3$. Excess kurtosis is defined as $K_{\text{ex}} = K - 3$, which is used to show deviation from a normal process. Therefore, we expect $K_{\text{ex}} = 0$ for a Gaussian process, whereas a distribution with heavy tails and many outliers will have $K_{\text{ex}} \gg 0$. As with skewness, the standard error on kurtosis depends only on sample size:

$$\sigma_{\text{kurt}} = \sqrt{\frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}}, \tag{4.15}$$

I show the excess kurtosis for the 7-DQ quasars and stars in Figure 4.8. For the stars, K_{ex} seems to fluctuate around the average with no clear trend with Δt, which is reassuring. The average K_{ex} is significantly non-zero in all three bands, which implies that the combination of noise distributions results in a distribution with heavier tails than that of a Gaussian, on all timescales.

The excess kurtosis of the quasars shows a gradual decline with Δt, ranging from $K_{\text{ex}} \approx 4$ on the shortest timescales to $K_{\text{ex}} \approx 0$ on the longest timescales of $\Delta t = 10^4$ days. This is consistent with my result obtained in Section 4.3.2, which showed that the Δm distribution is well described by a single Gaussian on these timescales. Again, this is a dimensionless parameter and therefore the exact value of kurtosis is difficult to interpret physically.

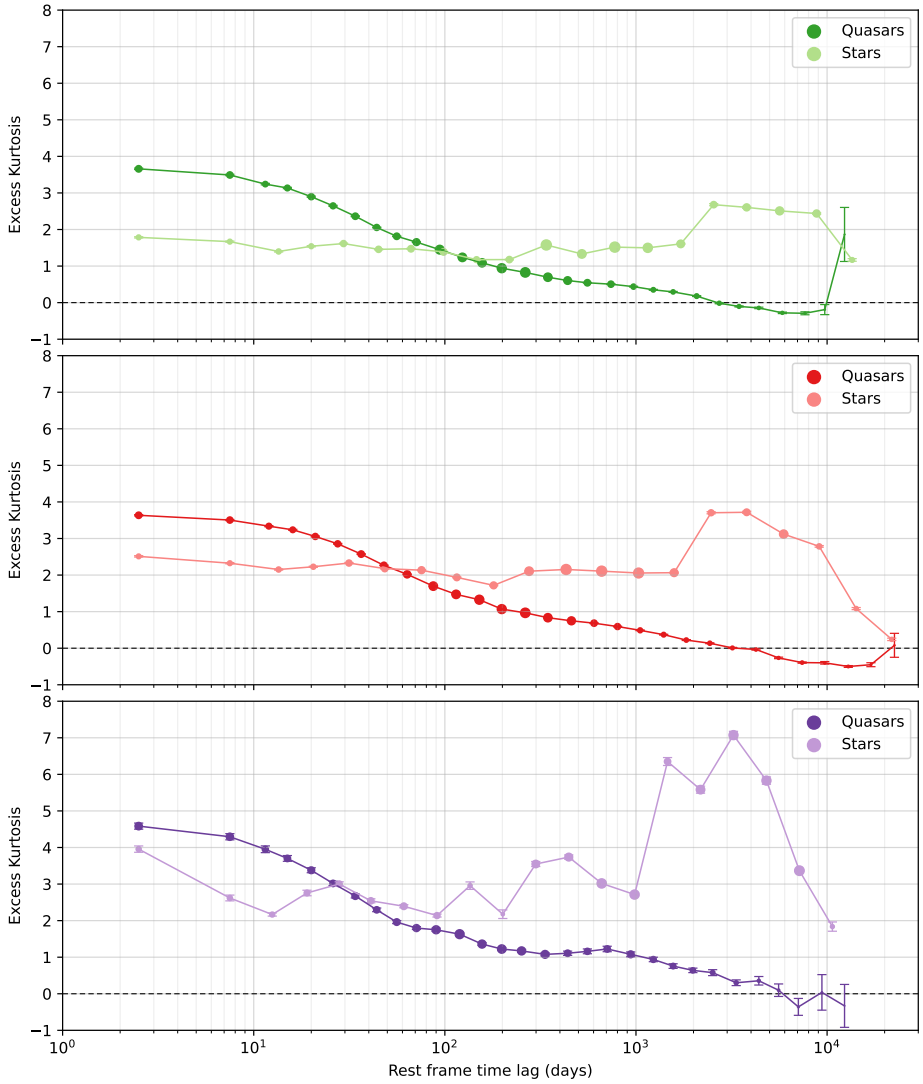

Figure 4.8: Kurtosis of Δm of 7-DQ quasars and stars in the g, r and i bands. Note that the size of the circular markers illustrates the relative number of points in that bin.

4.5 Summary

In this chapter, I have explored the quasar Δm distribution, analysed its shape and computed its moments. In Section 4.2, I presented the Δm distributions for the 7-DQ quasars and stars. Furthermore, in Section 4.3, I investigated the shape of the Δm distributions and presented a novel fitting technique using Gaussian mixtures. I also demonstrated that quasars exhibit similar amplitudes of variability at timescales > 50 years. In Section 4.4, I calculated different moments

of the Δm distributions, including the mean, skewness and kurtosis. By utilising the high volume of observations and extended baseline of 7-DQ, I showed that the mean magnitude drift of quasars is not significant on timescales up to 20 years, ruling out claims from other studies. For the first time, I present the skewness and kurtosis of the Δm distribution as a function of time-lag, which is a novel result and can be used as a benchmark for future studies.

Chapter 5

Structure Function Analysis

5.1 Introduction

In this chapter, I use the structure function, and variations of it, to characterise variability of the 7-DQ quasar photometry over different timescales. First, in Section 5.2, I provide an overview of the structure function, including a comparison of definitions commonly used in the literature. Subsequently, in Section 5.3, I introduce a novel variation of the structure function that incorporates photometric uncertainties to calculate an optimal average. Furthermore, in Section 5.4, I present the ensemble structure function for the 7-DQ quasars and stars, contrasting it with the expectation of a damped random walk. Section 5.5 details my application of a variation of the structure function to probe time asymmetries in the light curves. Additionally, in Section 5.6, I investigate how the structure function varies amongst groups of quasars with differing properties such as black hole mass, luminosity, and Eddington ratio. In Section 5.7, I discuss the dependence between quasar properties and structure function amplitudes as a function of observed time-lag, quantifying this effect in terms of a correlation coefficient. Furthermore, in Section 5.8, I narrow down Δm pairs to a specific rest-frame wavelength range and compute the corresponding structure functions to investigate how variability changes with wavelength. Finally, I conclude this chapter with a brief summary in Section 5.9.

5.2 Overview and Definitions of the Structure Function

5.2.1 Comparison of Structure Function Definitions

In Chapter 1, Section 1.8, I explained how the structure function is constructed by considering the covariance between points within a light curve. Structure function studies are the traditional method for quantifying quasar variability as a function of time-lag. It is a powerful tool robust against sparse sampling and may be applied to either a single object or an ensemble. Over its 40-year history, different interpretations of the structure function have resulted in an abundance of definitions. In this section, I state and compare some of the most common definitions. A review of structure function definitions in the literature can be found in Kozłowski (2016b).

2In its most general form, the structure function of a light curve is the RMS of the magnitude differences $\Delta m = m(t_j) - m(t_i)$ evaluated at times t_j and t_i, separated by a time-lag $\Delta t = t_j - t_i$. Simonetti et al. (1985) defines the structure function as:

$$\mathrm{SF}_{\mathrm{obs}}(\Delta t) = \sqrt{\frac{1}{N(\Delta t)} \sum_{j<i} \Delta m_{ij}^2}, \tag{Si85}$$

where I have used the 'obs' subscript for consistency, since this is the definition I use for the 'observed structure function' defined in Chapter 1, Section 1.8. Sumi et al. (2005) and Hook et al. (1994) instead define the SF as the median of magnitude changes,

$$\mathrm{SF}(\Delta t) = \mathrm{med}\left(\Delta m_{ij}^2\right), \tag{Su05}$$

with the latter using the modulus instead of the square of Δm pairs. Vanden Berk et al. (2004) estimate intrinsic variability by subtracting photometric noise,

$$\mathrm{SF}(\Delta t) = \left\langle \sqrt{\frac{\pi}{2}} \left|\Delta m_{ij}\right| - \left\langle \sigma^2 \right\rangle \right\rangle, \tag{VB04}$$

where angled brackets denote a mean average. Bauer et al. (2009) used a similar

definition,

$$\mathrm{SF}(\Delta t) = \sqrt{\left\langle \Delta m_{ij}^2 \right\rangle - \left\langle \sigma^2 \right\rangle}, \tag{B09}$$

however, both definitions subtract $\langle \sigma^2 \rangle$ instead of the expected $2\langle \sigma^2 \rangle$, which leads to a flatter SF. The equation

$$\mathrm{SF}(\Delta t) = \left\langle \sqrt{\frac{\pi}{2}} \left| \Delta m_{ij} \right| - \sqrt{\sigma_i^2 + \sigma_j^2} \right\rangle \tag{Sc10}$$

from Schmidt et al. (2010) seems to subtract too much noise, since σ has not also been scaled by $\pi/2$. MacLeod et al. (2012) uses the interquartile range (IQR) between 25% and 75% of the Δm distribution, in order to be robust against outliers,

$$\mathrm{SF}(\Delta t) = 0.741 \times \mathrm{IQR}(\Delta m), \tag{M12}$$

however, they do not subtract photometric noise. Kozłowski (2016b) attempts to remove this photometric noise with the definition:

$$\mathrm{SF}(\Delta t) = 0.741 \times \sqrt{\mathrm{IQR}(\Delta m) - \mathrm{IQR}(n)}, \tag{K16}$$

where $\mathrm{IQR}(n)$ is the IQR of the Δm distribution of the shortest observed time-lags, i.e., the IQR of the photometric noise.

A plethora of definitions causes problems with interpretations and comparisons, and therefore it is important to adopt one system. I use the following definition to approximate the intrinsic structure function,

$$\mathrm{SF}_{\mathrm{int}}(\Delta t) = \sqrt{\frac{1}{N(\Delta t)} \sum_{j<i} (\Delta m_{ij}^2 - \sigma_i^2 - \sigma_j^2)}, \tag{5.1}$$

where the sum runs over the $N(\Delta t)$ epochs such that $t_j - t_i = \Delta t$. As discussed in Chapter 1, Section 1.8.2, this definition was originally proposed by Press et al. (1992a) and is deemed the most correct way to estimate structure functions by Kozłowski (2016b).

5.2.2 Regimes of the structure function

Ensemble structure functions of quasars are often modelled as a single power law (SPL). However, this cannot persist for $\Delta t \rightarrow \infty$, as this would imply very large or infinite power at arbitrarily long timescales (Kozłowski 2016b). Therefore, there must be a break in the SPL, which is often referred to as the 'turnover' in the structure function, marked by some characteristic timescale, τ. The quasar structure function, whether calculated from observational data or computed analytically from a model, can be categorised into two or three regimes for different ranges of Δt. In the first regime, $\Delta t \ll \tau$, the structure function can be modelled using an SPL of the form

$$\mathrm{SF}(\Delta t) = \alpha \Delta t^{\beta}, \tag{5.2}$$

where β is often referred to as the 'slope', as it corresponds to the linear slope on a log-log plot. α is a constant of proportionality and determines the amplitude of the structure function if β is fixed. In this regime, the structure function slope of a damped random walk (DRW) is fixed at $\beta = 0.5$ (derived in Chapter 1, Section 1.9.3). The second regime is for $\Delta t \gg \tau$, where the structure function *should* ultimately plateau, as justified earlier. Therefore, between the first and second regime, marked by $\Delta t \approx \tau$, we expect a transition (i.e., a turnover) where the structure function starts to deviate from the SPL. The third regime is for small Δt, though its extent is dictated by photometric noise. For variability studies using typical surveys (e.g., SDSS, Pan-STARRS), this regime is around $\Delta t < 15$ days, and corresponds to a second flattening of the structure function at the amplitude of photometric noise in the data. The regime exists because quasars, on average (which is the case in the ensemble), exhibit little intrinsic variability in the optical/UV continuum on these timescales. If the photometric noise is removed correctly, the third regime vanishes, illustrated in Figure 5.1 by the red and black lines. Figure 5.1 also shows comparisons between the different structure functions calculated using the same data, defined in Section 5.2.1.

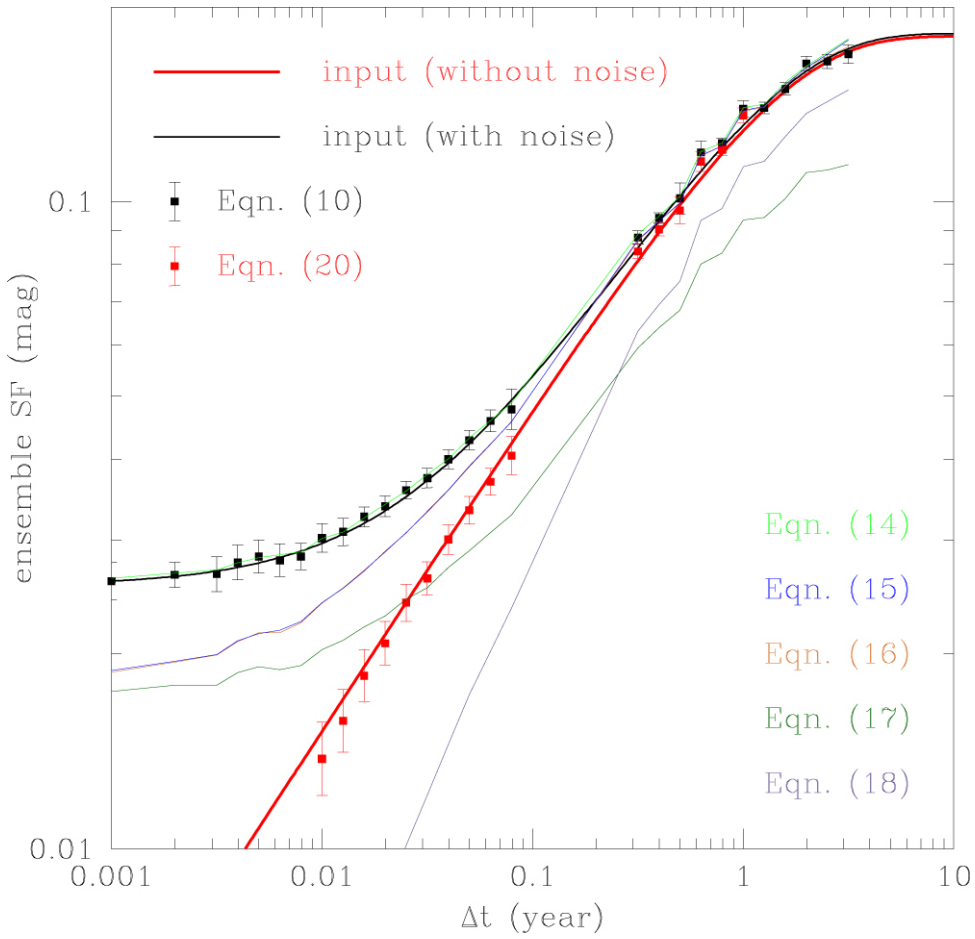

Figure 5.1: Example structure function calculations for 1000 simulated AGN light curves using a DRW model. The input signal (signal and noise) structure function is shown as the thick red (black) line. Structure functions from other definitions, calculated on the same data, are included for comparison. Red and black points are structure function measures calculated in Kozłowski (2016b). Figure adapted from Kozłowski (2016b)

5.2.3 Structure Function: Magnitudes or Fluxes?

When the structure function was first introduced into the context of astronomy by Simonetti et al. (1985), it was defined in terms of fluxes. Hook et al. (1994) later defined it in terms of magnitudes, without justification for the switch, which was adopted by all subsequent studies in the field. Since magnitudes are logarithmic, symmetric variations in flux will lead to asymmetric variations in magnitude. For small perturbations in magnitude, this asymmetry is negligible. However, large magnitude changes could introduce significant bias. Since there do not seem to be any studies which measure this effect, I simulated a simple toy model to

approximate the extent of the bias.

I used a DRW process ($SF_\infty = 5 \times 10^{-6}$ Jy, $\mu = 10^{-5}$ Jy, $\tau_{DRW} = 10^3$ days) to simulate the flux output of a quasar and calculated the structure function using fluxes. I chose these values for SF_∞ and μ as they correspond to a 21.4 mag quasar with $SF_\infty \approx 0.5$ mag which are typical values for my 7-DQ quasars. I found that τ_{DRW} had little effect on the bias, so I picked a reasonable value of 10^3 days. I converted flux, f, to magnitudes, m, using the relation,

$$m = -2.5 \log_{10}(f) + 8.9, \tag{5.3}$$

and calculated the corresponding structure function using magnitudes. I computed the fractional difference between these two structure functions for each Δt bin in a set of 30 logarithmically spaced Δt bins over the range $10 < \Delta t < 3 \times 10^4$ days. I computed the mean and standard deviation of these fractional differences and repeated the process for 50 runs. I found the average of these means and standard deviations to be 5.51% and 5.53%, respectively. While this is quite a rudimentary calculation, the average fractional difference of ~5% is small enough to conclude that the use of magnitudes in structure function calculations does not introduce significant asymmetric bias.

5.3 The Variance-weighted Structure Function

The definition of the intrinsic structure function (SF_{int}; Equation 5.1) involves subtracting photometric noise, which is equivalent to removing the third regime discussed in Section 5.2.2. However, this correction can cause the sum inside the square root of SF_{int} to be negative: $\sum_{ij} \Delta m_{ij}^2 - \sigma_i^2 - \sigma_j^2 < 0$. This occurs predominantly at very small time lags where we expect, on average, the true variability of the quasar population to be small compared to noise. One possibility to overcome this issue is to omit observations with high photometric errors. However, this removes information and introduces systematic biases, since the signal and noise of an observation are correlated. I developed my own solution to this issue by creating a variation of SF_{int}, which I have named 'variance-weighted structure function', denoted $\widehat{SF}$. I designed this to favour high signal-to-noise observations, while still including those with low signal-to-noise. I achieved this by weighting Δm^2 values by their inverse-variance, which is common practice for optimal averages.

First, we compute the variance of Δm^2. If we assume Δm values for a single quasar follow a normal distribution with zero mean and a width σ^2, then Δm^2 follows a χ^2 distribution with one degree of freedom, ($\kappa = 1$), scaled by σ^2,

$$\Delta m^2 \sim \sigma^2 \chi^2(\kappa = 1). \tag{5.4}$$

The variance is therefore:

$$\mathrm{Var}[\Delta m^2] = \mathrm{Var}[\sigma^2 \chi^2] \tag{5.5}$$

$$= \sigma^4 \mathrm{Var}[\chi^2], \tag{5.6}$$

and since $\mathrm{Var}[\chi^2(\kappa = 1)] = 2$, we arrive at the result:

$$\mathrm{Var}(\Delta m^2) = 2\sigma^4. \tag{5.7}$$

In practice, the variance σ^2 represents the sum of variances corresponding to photometric uncertainties,

$$\sigma^2 = \sigma_i^2 + \sigma_j^2, \tag{5.8}$$

and therefore the expected variance of $\mathrm{SF}^2(\Delta t)$ is:

$$\mathrm{Var}[\mathrm{SF}^2(\Delta t)] = 2(\sigma_i^2 + \sigma_j^2)^2. \tag{5.9}$$

I define the variance-weighted structure function, $\widehat{\mathrm{SF}}$, as the square root of the optimal average of SF^2:

$$\widehat{\mathrm{SF}}(\Delta t) = \sqrt{\frac{\sum_k \mathrm{SF}_k^2(\Delta t) \cdot w_k}{\sum_k w_k}}, \tag{5.10}$$

where the weights, w_k, are defined as the inverse-variance of SF^2, calculated previously (Equation 5.9),

$$w_k = \mathrm{Var}[\mathrm{SF}^2(\Delta t)]^{-1} = \frac{1}{2}(\sigma_i^2 + \sigma_j^2)^{-2}. \tag{5.11}$$

The rms uncertainty on $\widehat{\mathrm{SF}}$ is then,

$$\sigma_{\widehat{\mathrm{SF}}} = \left(\sum_k w_k \right)^{-1/2}. \tag{5.12}$$

I calculated the ensemble structure function for the 7-DQ quasar r-band photometry using this definition, and two other variations of the structure function. I present them in Figure 5.2 for comparison. We see that SF observed (Equation Si85) plateaus at small time-lags around 0.18 mag, corresponding to the third regime discussed in Section 5.2.2, suggesting the entire observed structure function is overestimated due to photometric noise. 0.18 mag is much larger than the average magnitude error across photometric observations, suggesting that this offset is likely dominated by a fraction of observations with large errors. Furthermore, SF intrinsic (defined in Equation 5.1) displays more erratic behaviour especially at small time-lags, where it becomes 'negative'. Note that the structure function should not be negative, however, I took the sign out of the square root in Equation 5.1 for the purpose of illustrating how large photometric errors can lead to invalid measurements of the structure function. Of the definitions considered, my variance-weighted structure function, defined in Equation 5.10, clearly provides the best estimate of the structure function; it approaches zero amplitude for small time-lags illustrating the correction for the photometric noise variance, and increases smoothly with Δt without kinks or excessive fluctuations. I will use this weighting method for all subsequent calculations of the structure function (and variations of it) hereafter, unless stated otherwise. Similarly, subsequent notation of SF implicitly refers to $\widehat{\mathrm{SF}}$, unless stated otherwise.

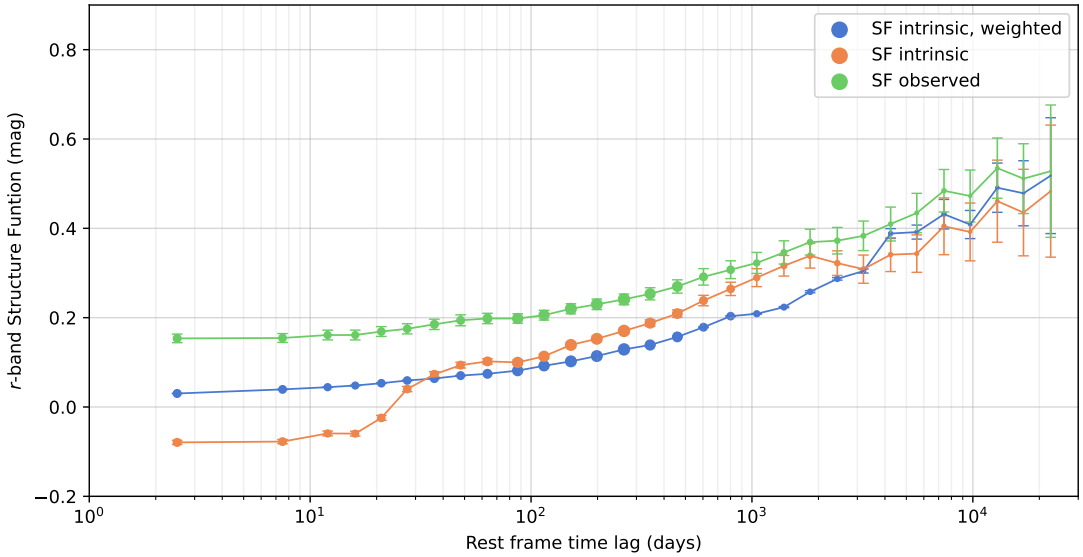

Figure 5.2: Comparison between different versions of the structure function using 7-DQ quasar photometry. SF observed represents the total observed variance in the Δm distribution and therefore is an overestimate of the true structure function. SF intrinsic represents the total variance minus the photometric variance, calculated on a per-observation basis. Low signal-to-noise measurements causes the intrinsic structure function to become negative at small time-lags and also causes erratic wiggles. My definition, the weighted intrinsic structure function, is the best behaved of all three definitions, and is a better representation of the underlying intrinsic variability of the quasar population.

5.4 The Ensemble Structure Function

5.4.1 Methods

Although one of the main advantages of the structure function is its ability to handle irregular and sparsely sampled data, the structure function for any single quasar for such data is itself sparse. Therefore, a common procedure is to group Δm measurements from many different quasars. The result is referred to as the 'ensemble structure function' and is identical to $\mathrm{SF}_{\mathrm{int}}$ with the exception of an additional sum,

$$\mathrm{SF}_{\mathrm{int,ensemble}}(\Delta t) = \sqrt{\frac{1}{N(\Delta t)} \sum_k \sum_{j<i} (\Delta m_{ij,k}^2 - \sigma_{i,k}^2 - \sigma_{j,k}^2)}, \tag{5.13}$$

where k denotes the index of the quasar. Although the ensemble structure function was previously defined in Chapter 1, Equation 1.34 it is repeated here for readability. As mentioned in Section 5.3, it is possible for the sum inside the square-root to be negative. This is rarely the case for quasars, as $\Delta m \gg \sigma$ provided $\Delta t > 10\,\text{days}$. However, since the stars are non-variable, Δm is on the order of σ for all Δt and therefore a negative sum is possible. For bins of Δt where this is the case, I do not have a reliable estimate of the structure function and therefore these points are omitted. My subsequent calculations of the structure function make use of all the techniques of Equations 5.1, 5.10, and 5.13, which correspond to subtracting photometric noise variance, weighting by inverse-variance and summing over the ensemble. The resulting structure function could be called the 'intrinsic, variance-weighted, ensemble structure function', however, this is cumbersome, and since I will employ all these techniques for subsequent calculations of the structure function (unless explicitly stated otherwise) there is no need to differentiate. Therefore, I will refer to this intrinsic, variance-weighted, ensemble structure function simply as the 'structure function' or 'ensemble structure function'.

I also applied a maximum likelihood estimation to estimate the ensemble structure function, using Gaussian errors, which I will refer to as $\text{SF}_{\text{int,MLE}}$. This method estimates the mean, μ, and the excess root-mean-square (RMS) scatter, σ_{excess}, of pairs of Δm from a light curve with normally distributed errors through maximum likelihood estimation (MLE). This is done by iteratively computing the values of μ and σ_{excess} which maximises the probability of observing the data. Given independent data points, Δm_k, with Gaussian errors, σ_k, the log-likelihood function is

$$\log L = -\sum_{k=1}^{n} \frac{(\Delta m_k - \mu)^2}{2(\sigma_k^2 + \sigma_{\text{excess}}^2)} - \frac{1}{2}\sum_{k=1}^{n} \log\left(\sigma_k^2 + \sigma_{\text{excess}}^2\right). \tag{5.14}$$

The values of μ and σ_{excess} that maximise this likelihood are those that minimise the sum of squared residuals weighted by the effective variances, $(\sigma_k^2 + \sigma_{\text{excess}}^2)$. In this method, we achieve this by computing weights w_k based on both measurement errors and the excess RMS scatter, σ_{excess}, as

$$w_k = \frac{\sigma_{\text{excess}}^2}{\sigma_{\text{excess}}^2 + \sigma_k^2}. \tag{5.15}$$

These weights effectively scale each data point's contribution based on both observed scatter and measurement error, ensuring that data points with larger measurement errors contribute less to the estimate of μ and $\sigma_{\rm excess}$. This weighting is crucial for MLE, as it allows data points with greater intrinsic variability to have a proportionate influence.

Through iterative refinement, μ and $\sigma_{\rm excess}$ are updated by recalculating these weights in each iteration and adjusting the estimates until convergence. Convergence is defined as the point at which further updates to μ and $\sigma_{\rm excess}$ are smaller than a specified threshold, implying that the likelihood function has reached its maximum. The final value of $\sigma_{\rm excess}$ represents the "excess" RMS scatter that maximises the likelihood, capturing any additional variability in the data that is not accounted for by the measurement errors alone. Therefore, $\sigma_{\rm excess}$ is another statistic which maybe used to calculate the intrinsic structure function, $\mathrm{SF}_{\rm int}$, and will be compared against the variance-weighted structure function in Section 5.4.2.

5.4.2 Results

In Figure 5.3, I present the ensemble structure functions of my 7-DQ quasar and star populations for the g, r and i bands. This study represents the most precise investigation of the ensemble quasar structure function to date. Compared to previous studies, this analysis uses the largest sample of quasars, with the greatest volume of observations in three optical bands, extending to time-lags that have been previously unexplored: by using a sample of $\sim$500,000 quasars, it is a representative average of the majority of currently observable spectroscopically-confirmed quasars. By leveraging over 10^9 Δm difference observations, I can constrain the shape of the ensemble structure function in the g, r and i bands with high precision. Additionally, unprecedented baselines enable me to constrain the shape of the structure function on timescales of up to 70 years in the rest-frame.

The structure function of the stars is flat, fluctuating around a median value of SF = 0.05, 0.03, and 0.02 mag for the g, r and i bands, respectively. This can be interpreted as residual variability due to photometric noise, and SF values below these values in their respective bands are unreliable. This is only the case for the quasar structure function on timescales $\Delta t < 10$ days, suggesting that there is minimal intrinsic variability exhibited by the ensemble quasar population on

these timescales.

On timescales $\Delta t > 10$ days, the quasar ensemble structure functions appear to follow a single power law (SPL) up to the longest timescales of $\Delta t \sim 10^4$ days. Therefore, I fitted SPLs of the form $\alpha \Delta t^{\beta}$ to each band, over these timescales, giving slopes of $\beta_g = 0.361 \pm 0.003$, $\beta_r = 0.355 \pm 0.004$, and $\beta_i = 0.418 \pm 0.007$, where the subscript denotes the band. Additionally, amplitudes were found to be $\alpha_g = 0.019 \pm 0.003$, $\alpha_r = 0.018 \pm 0.004$, and $\alpha_i = 0.011 \pm 0.007$. The r and i band structure functions continue to follow an SPL (within error) even on the longest timescales, indicated by the right-most points. The final point of the g band structure function deviates from an SPL, however, relatively fewer pairs are used in this final bin (10^4, compared to a mean 10^8 pairs for all bins) and therefore it is less reliable. Ensemble structure functions calculated by MacLeod et al. (2012), de Vries et al. (2005), and Morganson et al. (2014) are overplotted and discussed in the next section.

To test the effect of Malmquist bias on my ensemble structure function, I re-calculated the r-band structure function using the quasar and star bright subset (defined in Chapter 3, Section 3.2.1). The result is shown in Figure 5.4. For the quasars, the structure function of the bright subset shows minimal deviation from that of the full sample across all time-lags. Note that the clear visual difference at short timescales is exaggerated by the logarithmic scale. The absolute difference between the two structure functions, averaged over Δt, is on the order of 0.01 mag. This result is very reassuring and indicates that Malmquist bias has negligible impact on my structure function analysis in general. The structure function of the bright subset of stars is, on average, slightly lower than that of the full sample. This is as expected, as the structure function for the stars measures residual photometric noise, which is lower for the brighter stars.

Figure 5.5 shows the results obtained using MLE fitting. At shorter timescales, this method yields a smaller structure function amplitude, as it accounts for the excess variance despite the underreported ZTF errors. Over intermediate timescales of 10^2 to 10^3 days, both structure functions exhibit a similar, relatively shallow slope (not consistent with a DRW). However, the MLE method produces a slightly larger amplitude. Beyond these timescales, the $\mathrm{SF}_{\mathrm{int,MLE}}$ begins to plateau, while the variance-weighted structure function continues to follow a simple power law (SPL). This suggests that the MLE structure function has a characteristic timescale around 3000 days (approximately 8 years).

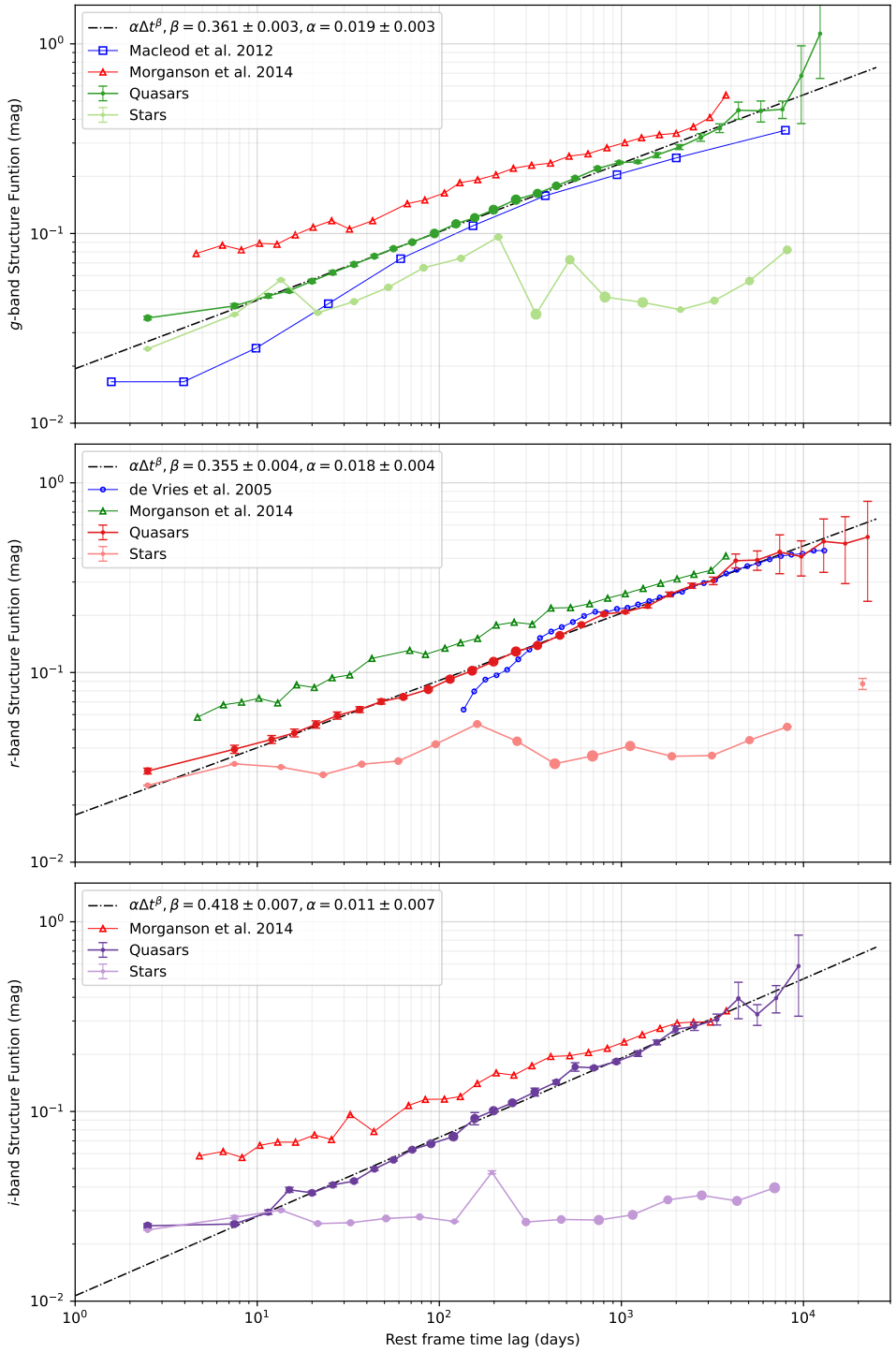

Figure 5.3: Ensemble structure function for the quasars and stars in the g, r and i bands. The size of the points represents the relative number of points in the bin. The black dot-dashed line represents an SPL fit to the quasar structure function data points for $\Delta t > 10$ days, with the slope shown in the legend. In the top panel, comparison data from MacLeod et al. (2012), de Vries et al. (2005), and Morganson et al. (2014) are overplotted. A few points are omitted for the ensemble star structure function, as the photometric errors are greater than the observed variability.

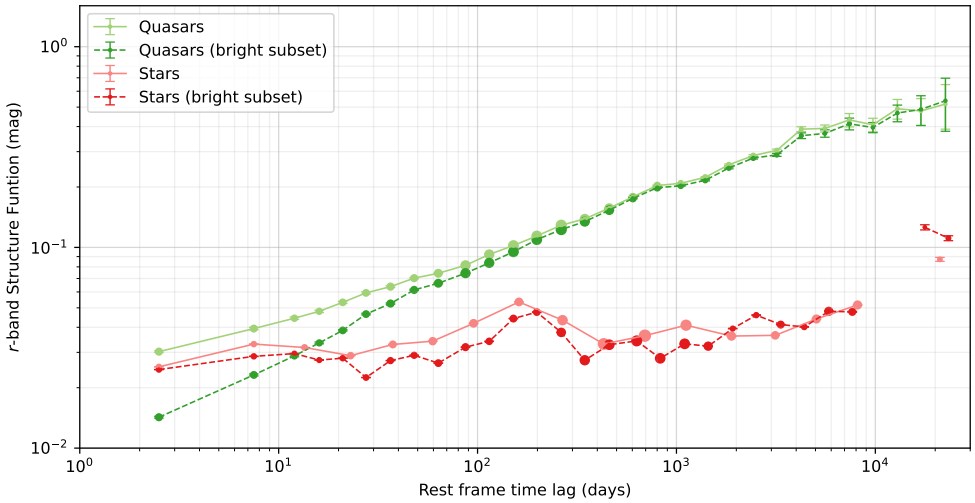

Figure 5.4: The same r-band ensemble structure function plotted in Figure 5.3, except with the bright subset overplotted.

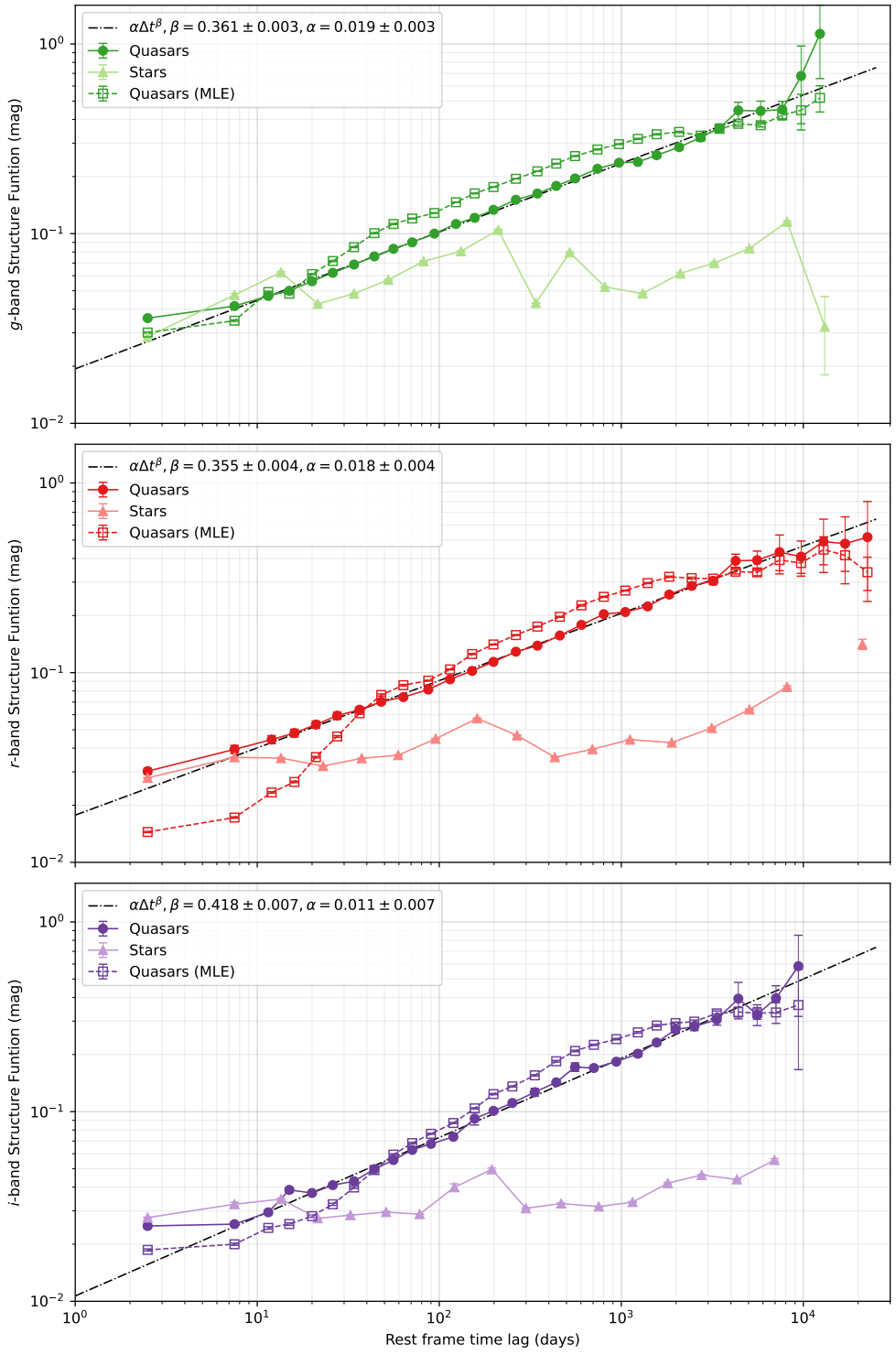

Figure 5.5: The same *gri* ensemble structure functions plotted in Figure 5.3, except with the maximum likelihood estimation of the structure function overplotted (labelled 'MLE').

5.4.3 Discussion

I will compare my ensemble structure function with two other massive variability studies, which calculate the ensemble structure function on long timescales. These studies are de Vries et al. (2005), MacLeod et al. (2012), and Morganson et al. (2014) which I will refer to as dV05, M12, and Mo14, respectively, hereafter. Their results are plotted over my results in Figure 5.3 for comparison. M12 used a sample of 33,881 quasars over a baseline of 50 years in the observer frame by using POSS and SDSS plate data. dV05 used a similar sized sample, 35,165 quasars, with photometry again from POSS and SDSS such that their baseline was also 50 years. Mo14 compared SDSS and Pan-STARRS photometry for 10^5 over a baseline of 10 years. These studies have some of the longest baselines to date, and while more recent studies have expanded their sample size or number of observations per object, none have compared the most recent observations against historic plate photometry to push to new temporal baselines.

dV05 found that their ensemble structure function follows an SPL with a slope of $\beta = 0.30 \pm 0.01$ and no sign of a turnover (i.e., no break in the SPL) up to ~40 years in the rest-frame. Consequently, they concluded that there is no single preferred characteristic timescale for the quasars, but instead most likely a continuum of timescales. M12 covered a much larger range of Δt by using repeat imaging of quasars in SDSS Stripe 82. Given the overall shape of their structure function, in particular the hint of a turnover at long time-lags, they concluded that their structure function is consistent with that of a damped random walk (DRW; see Chapter 1, Section 1.9.3). The baseline of Mo14 does not extend as far as the other two studies, but they show a consistent SPL on all observed timescales, with slopes of $\beta_g = 0.251 \pm 0.004$, $\beta_r = 0.275 \pm 0.004$, and $\beta_i = 0.277 \pm 0.005$.

Comparing my result with the results of dV05 and M12, we see that all structure functions are in agreement for $10^2 - 10^3$ days. Remarkably, my r-band structure function is strikingly similar to dV05 $\Delta t > 10^3$, despite using mostly different data. However, I estimate significantly more variability than M12 on timescales $\Delta t < 10^2$ days. I suspect that it is because I am more likely to catch short-lived outburst events, given that ZTF has a high cadence. These outburst events often involve dramatic changes in magnitude and therefore increase the structure function on these timescales. Additionally, M12 uses the interquartile range (IQR) to calculate their structure function (see Equation M12), which reduce the effect of such outbursts on their results. Mo14 report a higher structure

function amplitude in all three bands compared to my results and the other studies, which could be due to residual photometric noise that has not been properly subtracted. This also leads to a flatter structure function, which is the cause of their significantly flatter SPL slopes.

The long-term behaviour of the structure function (i.e., on timescales $\Delta t >$ 20 years) is a contentious topic, but the literature involving its calculation tend to follow a common theme: the structure function initially grows as an SPL on timescales up to 1–3 years, then does one of two things based on the subsequent data points, which are often only a few. Either the structure function breaks from the SPL and starts to turnover at some characteristic timescale, often denoted τ (see e.g., MacLeod et al. 2012; Sesar et al. 2006; Stone et al. 2022; Voevodkin 2011), or, the structure function continues to follow the SPL on all observed timescales (see e.g., Hawkins 2002; de Vries et al. 2005; Vanden Berk et al. 2004; Morganson et al. 2014). A few studies claim that their data is unable to discern between the two, usually due to uncertainties (see e.g., Li et al. 2018).

As discussed in Section 5.2.2, the quasar structure function cannot continue to follow an SPL for all time-lags. Therefore, variability studies often claim a characteristic timescale indicated by turnover, which would solve the issue of 'infinite power at infinite time-lags'. Since the shape of the structure function is not usually well constrained on the upper bound of time-lags of any particular study, one can convince oneself that a turnover is visible on those time-lags if the final few data points start to drop. Therefore, it is important to remain agnostic and acknowledge that a characteristic timescale could be orders of magnitude longer than the timescales we are currently able to observe.

Many studies that claim a turnover often fit the DRW structure function to their data (see e.g., Sesar et al. 2006; MacLeod et al. 2012; Stone et al. 2022). The DRW model is attractive because it plateaus for $\Delta t \gg \tau_{\mathrm{DRW}}$, where τ_{DRW} is the characteristic DRW timescale, effectively preventing the light curve from arbitrary magnitudes. However, for $\Delta t \ll \tau_{\mathrm{DRW}}$, the DRW has a slope of $\beta = 0.5$ which is much too steep for the majority of slopes seen in observational data.

My ensemble structure function suggests that quasar variability is consistent with an SPL on timescales ranging from 10 days to 70 years. There is no evidence of a turnover in the structure function, or characteristic timescale, even on the longest timescales of 70 years. M12 claim there is indeed a turnover consistent with a DRW model, with a characteristic timescale of ∼2 years. However, my structure

functions in all g, r and i bands show this is not the case on timescales up to ~60 years. Only my r band structure function extends beyond 60 years (due to comparisons between ZTF and the POSS-I r band photometry). The behaviour of the final few data points in this band might suggest a subtle turnover, however, the errors of these points are still consistent with an SPL. Therefore, I am not able to claim a plateau of the r band structure function on these timescales.

The best fit SPLs to my ensemble structure functions have slopes of $\beta = 0.361 \pm 0.003$, 0.355 ± 0.004 and 0.418 ± 0.007, in the g, r and i band, respectively. Note that the steeper slope in the i band can be partly attributed to the smaller number of ZTF i band observations. Given that ZTF predominantly contributes short time-lag Δm points, the reduced quantity of ZTF i photometric data points leads to the suppression of the structure function on these timescales. My structure function slopes, particularly in the g and r bands, are significantly less steep than $\beta = 0.5$. This suggests that a single DRW is not an appropriate model for the characterising variability of the ensemble population. While it is possible to combine multiple DRWs to generate a flatter slope (see Figure 5.6), it would imply that $\tau_{\rm DRW} < 5$ days for at least ~30% of the quasar population in order to account for observations, which is unphysical and not observed in light curves. Although mixing timescales is likely the reason we do not see a turnover, my shallower slope at small timescales suggests that the physical processes responsible for variability are not well modelled by a DRW.

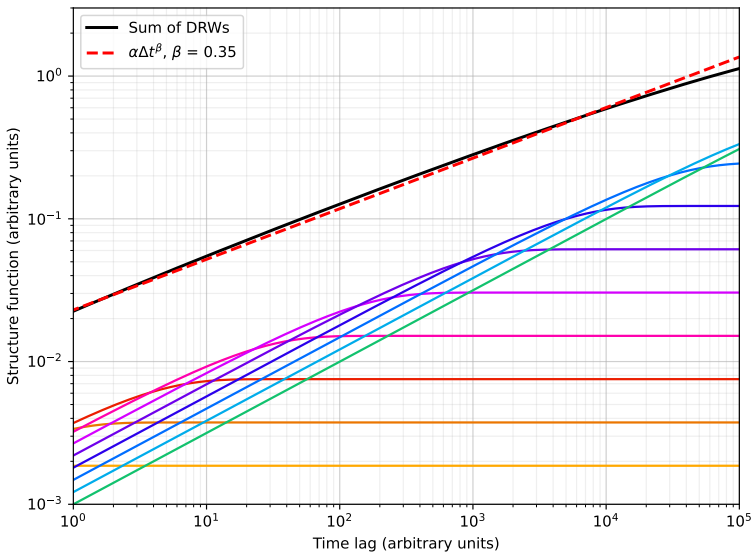

Figure 5.6: An analytical demonstration showing that a combination of DRW structure functions can approximate a single power law (SPL) over a range of timescales. By using 10 DRWs were used with a set of timescales and amplitudes that were logarithmic spaced and proportional to each other, I was able to reproduce a typical slope of ($\beta \sim 0.35$)

Many other massive variability studies also have estimates for the structure function, and I have summarised their slopes in Table 5.1. For example, Morganson et al. (2014) used a sample of 105,783 quasars over 10 years. They claim an SPL model fits the data well with a slope $\beta = 0.2457 \pm 0.0025$. Another example of a massive variability study was carried out by Li et al. (2018); they conducted the largest variability study to date using a sample of 119,305 quasars over 15 years, using photometry from SDSS and DECaLS. Their structure function was consistent with an SPL of slope $\beta = 0.254 \pm 0.011$.

In summary, my ensemble quasars structure functions are consistent with an SPL, on all observed timescales up to 10^4 days in the g and i bands, and up to 2×10^4 days in the r band. Notably, my results are most consistent with dV05. Furthermore, I rule out the statement of M12 that there is a turnover on these timescales. I agree with the findings of Mo14 that the ensemble structure function follows an SPL, however, I suspect that their slopes are underquoted due to improper subtraction of photometric noise. My i band is steeper than my g and r band structure functions, however, I suspect this is due to an imbalance of ZTF observations in this band, as both Vanden Berk et al. (2004) and Mo14 report similar slopes across these three bands. By comparing with other studies listed

in Table 5.1, the majority of studies find that slopes of $\beta \sim 0.3 - 0.4$, irrespective of band and baseline. This is significantly shallower than the predicted $\beta = 0.5$ of a DRW, which suggests that this model is not appropriate to recreate the quasar structure function. This is even the case if many DRWs are combined to produce a shallower slope, as demonstrated in Figure 5.6. While a break in the SPL must occur at some point, it is not clear when this characteristic timescale will be. It could be orders of magnitude longer than the timescales we are currently able to observe, or perhaps it will be seen with the next generation of telescopes. The only way to know is to continue monitoring, and studying, the variability of quasars.

Table 5.1: Summary of recent massive variability studies that report ensemble structure function slopes

Authors	Bands	$N_{\rm obj}$	Δt [yr] observer frame	SF slope β
Hawkins 2002	Schmidt plates U, B, V, R, I	400	24	0.20 ± 0.01
de Vries et al. 2005	POSS SDSS g	31,165	50	0.30 ± 0.01
Vanden Berk et al. 2004	SDSS g, r, i	~25,000	2	g: 0.293 ± 0.030 r: 0.336 ± 0.033 i: 0.303 ± 0.035
Voevodkin 2011	SDSS g	7,562	10	~0.33 (for $\Delta t > 42$ days)
MacLeod et al. 2012	SDSS u, g, r, i, z	33,881	20	0.40 (for $\Delta t > 42$ days)
Morganson et al. 2014	SDSS, PS g, r, i, z	~100,000	10	g: 0.251 ± 0.004 r: 0.275 ± 0.004 i: 0.277 ± 0.005
Li et al. 2018	SDSS, PS g, r, z	119,305	15	0.254 ± 0.011
This work	UKST & POSS plates SDSS, PS, ZTF g, r, i	~500,000	70	g: 0.361 ± 0.003 r: 0.355 ± 0.004 i: 0.418 ± 0.007

5.5 Investigating Structure Function Asymmetries

5.5.1 Methods

While the structure function measures the absolute overall variance, asymmetries between the rising and falling parts of the light curve can be investigated by separating the fluctuations into positive and negative changes, computing their respective structure functions and comparing them. The asymmetric structure functions were previously defined in Chapter 1, Equations 1.35 and 1.36, repeated here for readability:

$$\mathrm{SF}_{-} = \sqrt{\frac{1}{N(\Delta t)} \sum_{j<i} (\Delta m_{ij,-}^2 - \sigma_{i,-}^2 - \sigma_{j,-}^2)} \qquad \text{(brightening)} \qquad (5.16)$$

$$\mathrm{SF}_{+} = \sqrt{\frac{1}{N(\Delta t)} \sum_{j<i} (\Delta m_{ij,+}^2 - \sigma_{i,+}^2 - \sigma_{j,+}^2)} \qquad \text{(dimming)} \qquad (5.17)$$

where Δm_{ij} have been separated into $\Delta m_{ij,-}$ and $\Delta m_{ij,+}$ for brightening and dimming observations respectively.

5.5.2 Results

The asymmetric structure functions (Equations 5.16 and 5.17) for the quasar and star populations have been plotted in the g, r and i bands in Figure 5.7. Note that these have been weighted in the same manner as $\widehat{\mathrm{SF}}$. I have plotted these on linear y-axes for clarity when comparing SF_{+} and SF_{-}. Additionally, I computed the asymmetric structure functions for a DRW simulation, which is presented in the bottom panel of Figure 5.7, and is discussed in the following section.

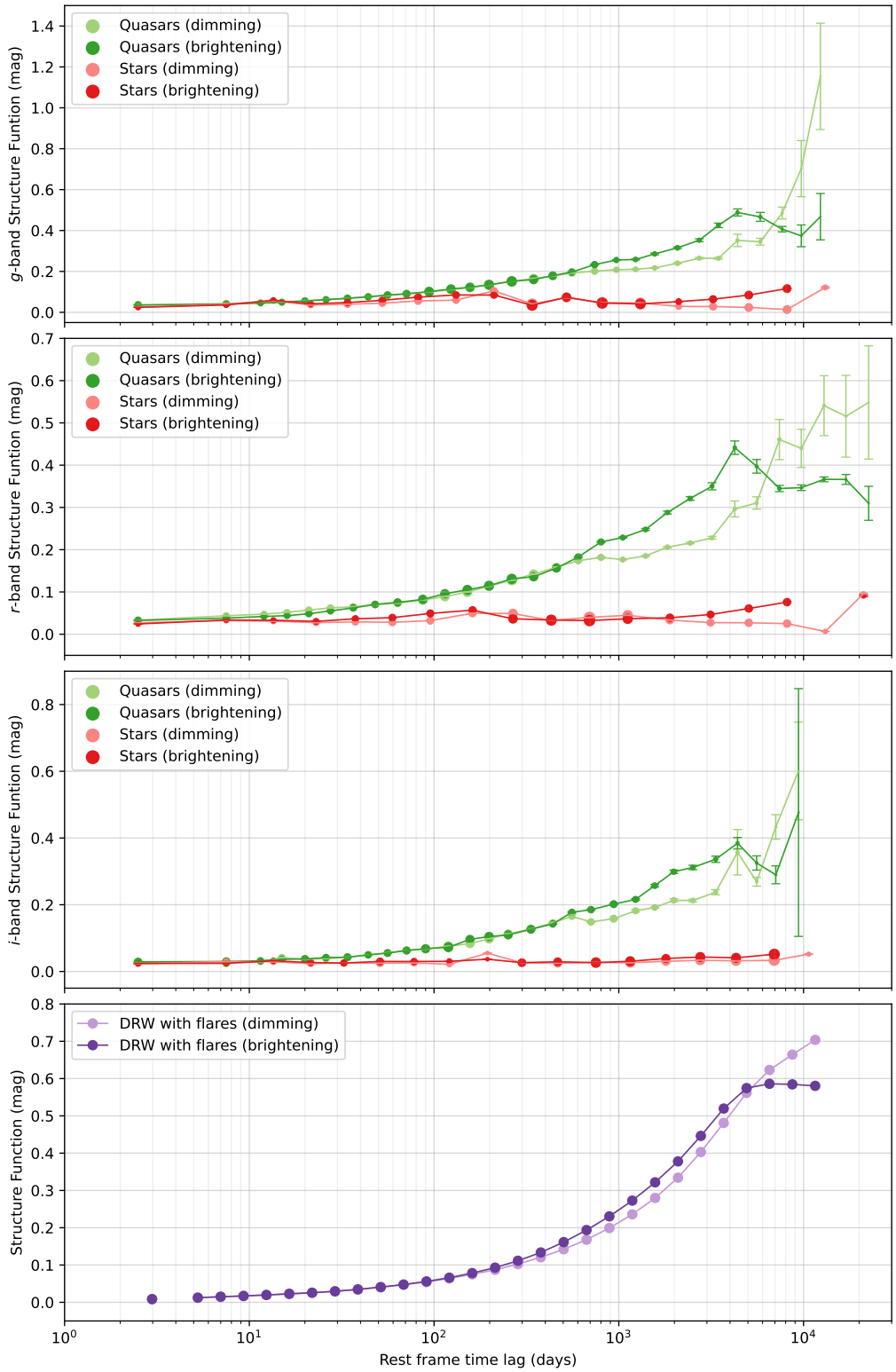

Figure 5.7: Positive and negative structure functions plotted for the quasar and star population in the g, r and i bands. Fluctuations that involve brightening are illustrated with opaque points, while those representing dimming are in translucent points. The bottom panel shows the structure function of a DRW simulation, discussed in Section 5.5.3

5.5.3 Discussion

The quasar structure functions in Figure 5.7 show SF_- (brightening) dominating at short timescales and then plateauing at 5.5 years. Conversely, SF_+ (dimming) is weaker at short timescales, but takes over from SF_- around 20 years rest-frame. This behaviour is consistent across all three bands, and implies that the timescales for brightening and dimming are different, with brightening favouring timescales ~10 years, while dimming favouring timescales > 30 years. Since SF_+ and SF_- are an ensemble of the entire quasar sample, it is likely that there are a range of timescales, and the ones I have quoted are an average of these. The stars do not show a significant difference between SF_+ and SF_-, which is reassuring.

dV05 also measured light curve asymmetry in their data. They found that variability within their sample is consistent with a shot noise model using an exponentially decaying profile with a half-life of 2 years, and that outbursts (also referred to as flares) occur on typical timescales of 200 years. By comparing SF_+ and SF_- they find asymmetries in the light curves with a statistical significance of 6σ, consistent with their proposition that variability is driven by outbursts/flares with an asymmetric profile.

Simulation of a Damped Random Walk with Flares

In order to test whether a light curve with fast-rise, slow-decline features could produce the behaviour of the observed asymmetric structure functions presented in Figure 5.7, I simulated quasar light curves using a DRW process with such features. I generated 100 light curves using a DRW process and superposed flares with a fast-rise, slow-decline profile. An example of one of the light curves is shown in Figure 5.8. I computed the asymmetric structure functions of these light curves in the same manner as the quasars and stars. The result is shown in the bottom panel of Figure 5.7. When comparing this against the observed asymmetric structure functions of the quasars, there are some similarities and differences. On timescales $\Delta t <\sim 5 \times 10^3$, the shape of the DRW-flare SF_+ and SF_- are almost identical, with slopes $\beta \sim 0.5$, as expected from a DRW process. However, the added flares cause SF_+ and SF_- to differ significantly beyond 10^3 days, with $SF_- > SF_+$ until $\sim 5 \times 10^3$. For time lags beyond this, $SF_- < SF_+$. This is the same behaviour seen in the observed asymmetric structure functions of the quasars. While this is not evidence that quasar light curves must have

flares of this kind, it demonstrates that such a light curve can reproduce the observed asymmetry in variability, and suggests that quasar light curves behave in a similar manner to light curves with fast-rise slow-decline features.

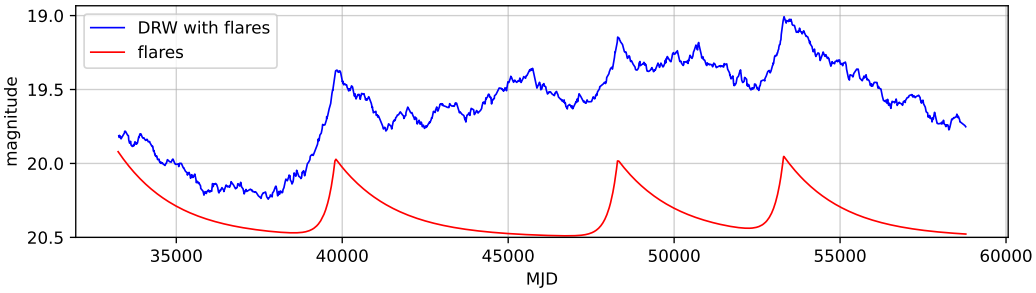

Figure 5.8: A damped random walk with flares. The blue line represents the DRW with flares imposed, while the red line shows the flares by themselves.

5.6 The effect of Quasar Properties on the Structure Function

Many simple correlations between variability and various physical quasar parameters have been known for several decades. These relationships from early studies are summarised by Giveon et al. (1999) and Helfand et al. (2001). Some of these early studies show a clear anticorrelation between variability and luminosity (see e.g., Angione & Smith 1972; Hook et al. 1994; Cristiani et al. 1997; Vanden Berk et al. 2004). However, luminosity is itself dependent on black hole mass and accretion rate (quantified as the Eddington ratio). Recent studies have shown a statistically significant anticorrelation between variability and Eddington ratio (see e.g., Kelly et al. 2009; MacLeod et al. 2010; Zuo et al. 2012; Kelly et al. 2013; Simm et al. 2016; Rakshit & Stalin 2017; Sánchez-Sáez et al. 2018; Lu et al. 2019), but whether there is a relationship between variability and black hole mass is an open discussion. Positive correlations were found by Wilhite et al. (2008); MacLeod et al. (2010), and Lu et al. (2019), while negative correlations were found by Kelly et al. (2009) and Kelly et al. (2013). No clear correlations were reported by Zuo et al. (2012); Simm et al. (2016); Rakshit & Stalin (2017), and Li et al. (2018). Arévalo et al. (2023) state that these conflicting results can be reconciled when considering that these correlations depend on the timescale of observed variability.

In this section, I present a study to investigate the effect of quasar properties on the structure function. I will focus on three key properties: bolometric luminosity, $L_{\rm bol}$, black hole mass, $M_{\rm BH}$, and Eddington ratio, $n_{\rm Edd}$. It has long been debated which of these properties is the main driver of variability. By grouping quasars by these properties and computing their respective structure functions, I am able to characterise the shape of the structure function with quasar properties, and determine which of these have the greatest impact on the amplitude of variability at different timescales.

5.6.1 Grouping by Quasar Properties

The grouping of Δm from different quasars can be done in several ways. The simplest is to combine measurements of our entire quasar sample into an ensemble structure function. However, my large quasar sample enables me to calculate the structure function for more complex groupings, such as by quasar properties.

I utilise the power of my immense sample size by repeating the analysis in Section 5.4 for quasars in bins of different bolometric luminosity, black hole mass and Eddington ratio. These properties were taken from the DR16 catalogue of quasar properties provided by Wu & Shen (2022) and cross-matched with our quasar sample to obtain properties for 296,868 quasars. For each property, I calculate the Z-score (also known as the standard score) for each quasar:

$$Z_{q,p} = \frac{x_{q,p} - \mu_p}{\sigma_p} \tag{5.18}$$

for property p and quasar q. Using the Z-score, I proceeded to bin the quasars into 8 groups based on values of $Z_{q,p}$. An example using bolometric luminosity is shown in Figure 5.9. The central bin edge is on the mean of the distribution, and successive bin edges lie at multiples of 0.5σ. Note that the extremal bins have outer edges placed at $\pm 3.5\sigma$ such that their width is instead 2σ. This was done to ensure that the two bins covering the tail ends of the distribution have a similar number of quasars to those closer to the mean. This produces a total of 8 groups, which are numbered in Figure 5.9. The same process is repeated for $M_{\rm BH}$ and $n_{\rm Edd}$.

I then compute the (variance-weighted) ensemble structure functions for the quasars in each of the 8 groups. This analysis was done for bolometric luminosity, black hole mass and Eddington ratio in g, r and i bands. I name these ensemble

groups ‘subensembles’ and their structure functions are shown in Figures 5.10, 5.11 and 5.12.

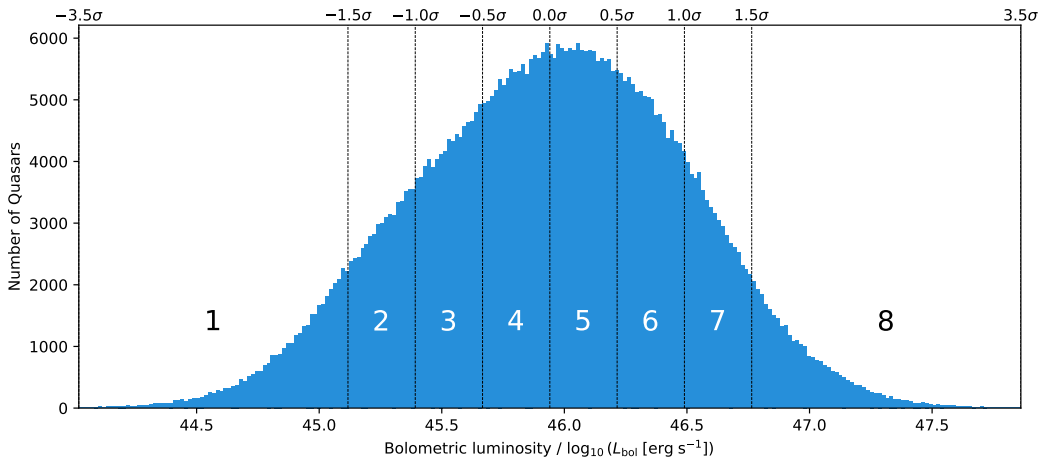

Figure 5.9: Distribution of bolometric luminosities within our quasar sample. The numbers in each rectangle denote which group the quasars belong to.

5.6.2 Subensemble Structure Functions

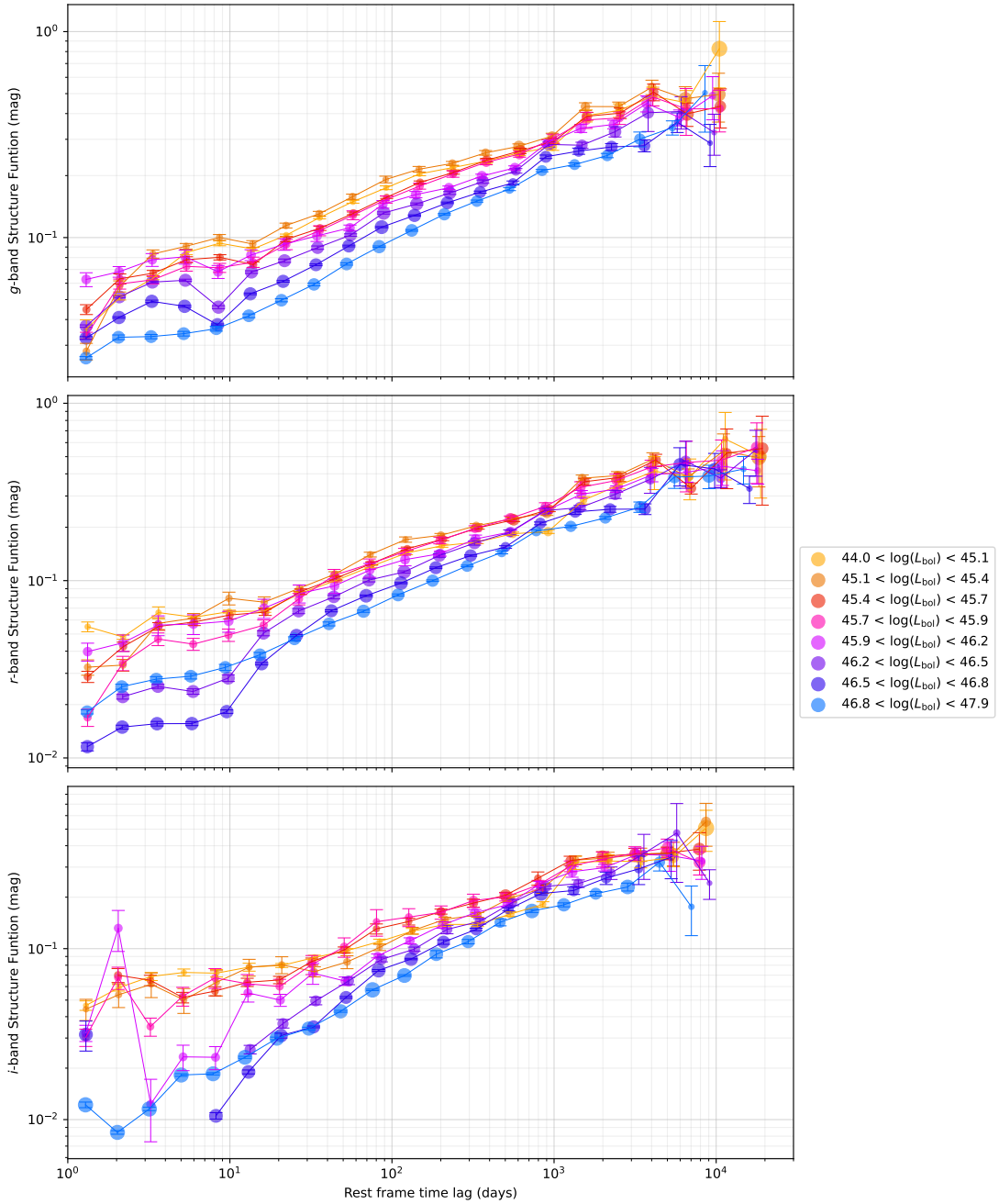

Figure 5.10: Structure functions of subensembles grouped by bolometric luminosity in the g, r and i bands. The shaded region represents timescales $\Delta t < 10$ days and is mostly dominated by noise.

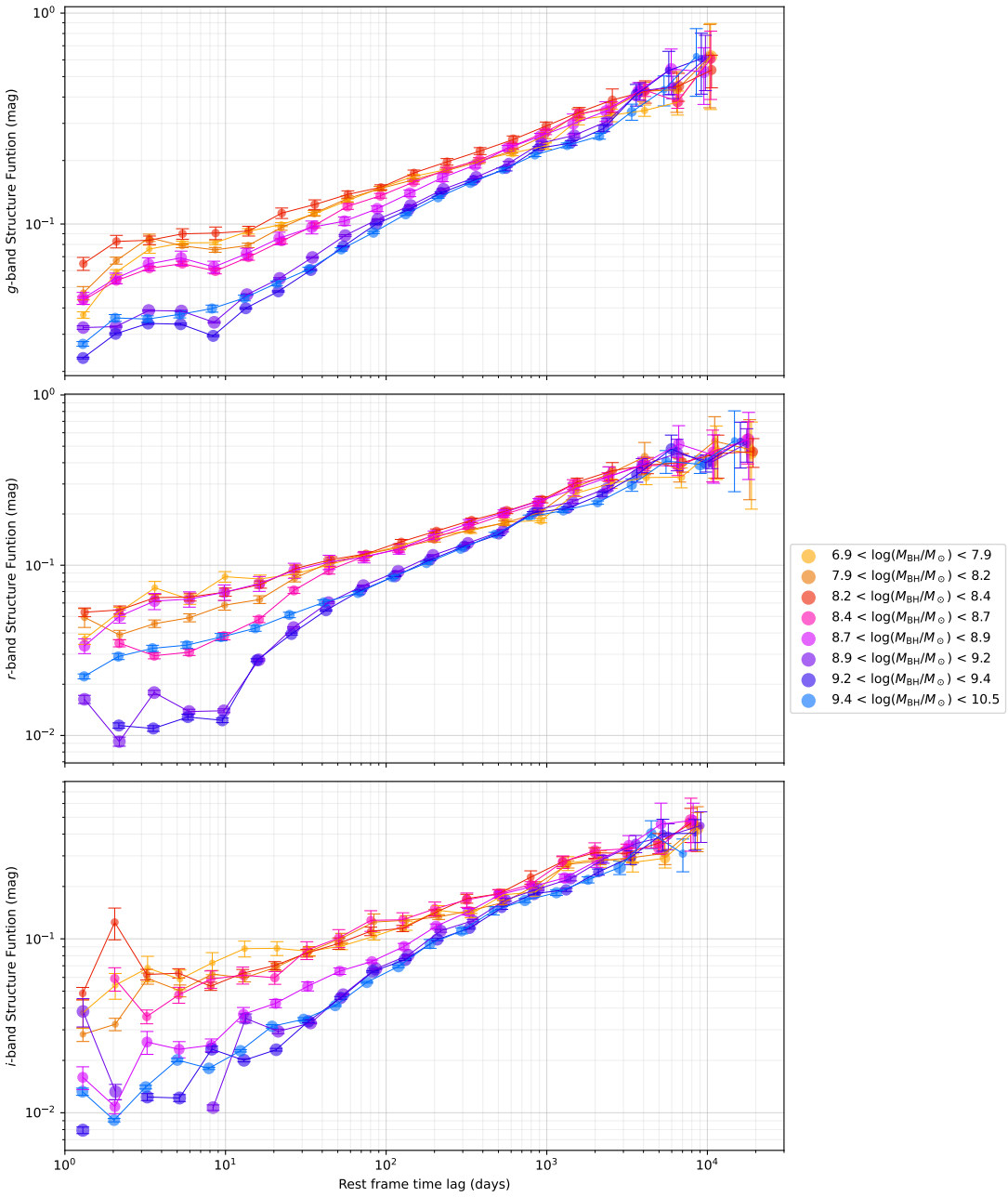

Figure 5.11: Structure functions of subensembles grouped by black hole mass in the g, r and i bands. The shaded region represents timescales $\Delta t < 10$ days and is mostly dominated by noise.

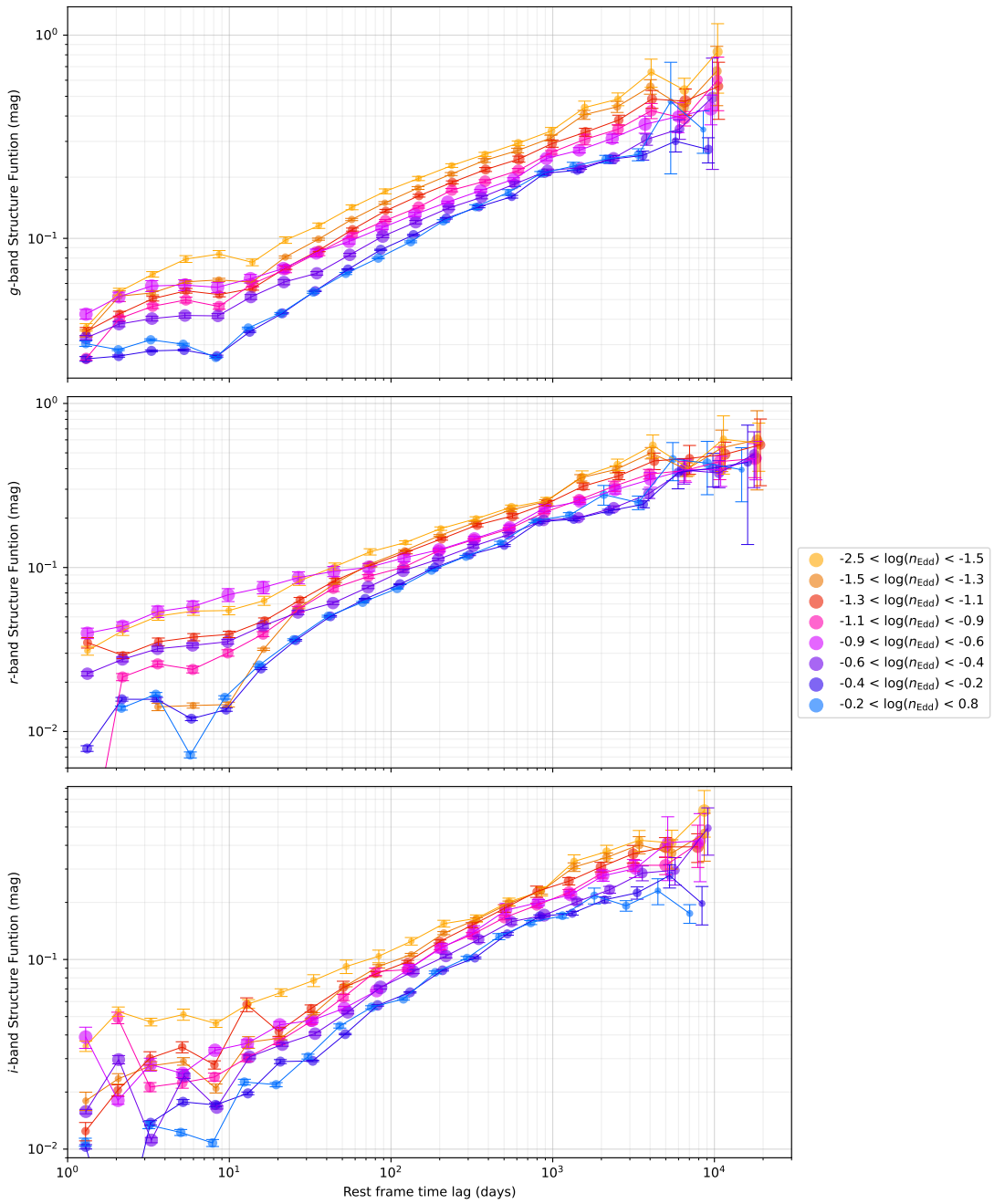

Figure 5.12: Structure functions of subensembles grouped by Eddington ratio in the g, r and i bands. The shaded region represents timescales $\Delta t < 10$ days and is mostly dominated by noise.

5.6.3 Single power law fits

In order to quantify the effect of quasar properties on the shape of the subensemble structure functions, I fitted single power laws (SPLs) of the form $\alpha \Delta t^{\beta}$ to the

structure functions in Figures 5.10, 5.11 and 5.12. I plotted the resulting values of α and β in Figure 5.13, and plotted the amplitude and slope of the resulting fits in Figure 5.13.

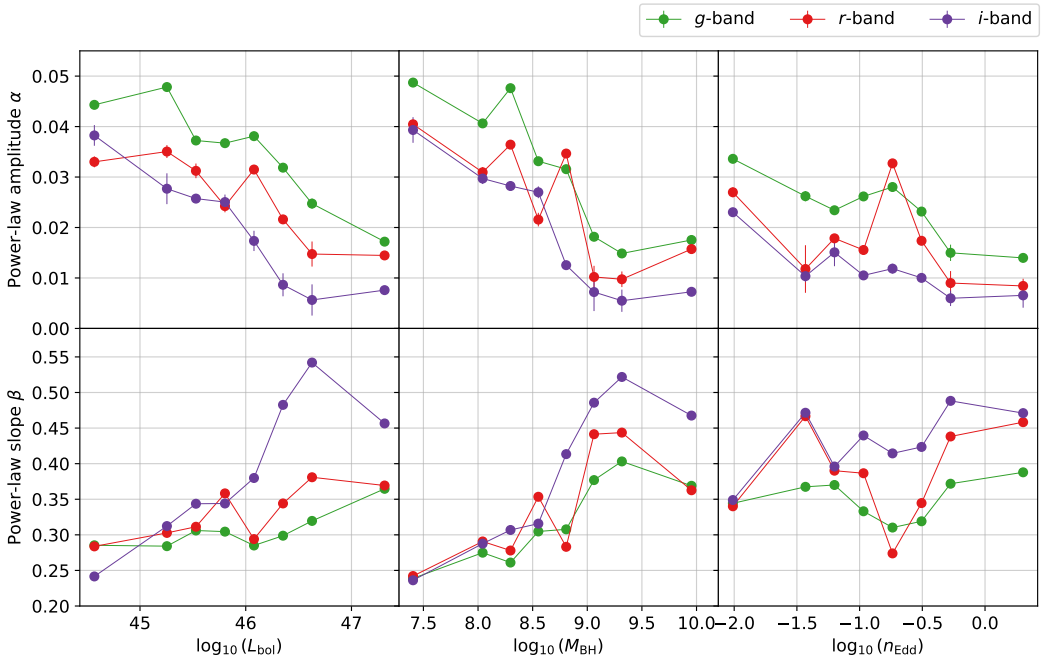

Figure 5.13: Dependence of SPL amplitude and slope on $L_{\rm bol}$, $M_{\rm BH}$, and $n_{\rm Edd}$, in the g, r and i bands

5.6.4 Discussion

The subensemble structure functions, presented in Figures 5.10, 5.11 and 5.12, along with their SPL fits presented in Figure 5.13, show clearly that the shape and amplitude of the structure function is dependent on luminosity, black hole mass and Eddington ratio to varying degrees. I have reproduced the commonly reported anticorrelation between variability and both luminosity and Eddington ratio on timescales $\Delta t < 5 \times 10^3$ days, however, the correlation is weaker on timescales beyond this (particularly in the g band), evidenced by the 'merging' of subensemble structure functions on the longest timescales. Whether there is a correlation between variability and black hole mass is an open debate in the literature. The shape of the $M_{\rm BH}$ subensemble structure functions, including their slope and amplitude, are clearly dependent on black hole mass. In particular, my SPL fits (Figure 5.13) reveal a negative correlation with variability amplitude, and a positive correlation with slope, as with luminosity.

The shape of all my subensemble structure functions are consistent with SPLs, and the steepness of the SPL slope changes with luminosity and black hole mass. dV05 saw a similar effect when splitting a population of ∼40,000 quasars into two groups of low and high luminosity. They found a difference in structure function amplitude between the two samples, consistent with my findings. Additionally, their high luminosity sample exhibits a steeper SPL slope than the low luminosity sample. This is consistent with my result shown in Figure 5.10.

The subensemble structure functions still follow an SPL, with no clear sign of a turnover. This is surprising, because M12 suggested that quasars are well represented by a DRW process, with DRW parameters determined by the properties of quasars. Their conclusions imply that my subensemble structure functions should be described by a DRW structure function. The absence of a turnover or characteristic timescale up to 10^4 days suggest that a DRW process is not a good model for groups of quasars with similar properties. My results do not rule out the presence of a characteristic timescale; it is possible that we have not been monitoring quasars for long enough to detect a characteristic timescale, which would be the case if these timescales are > 70 years. Alternatively, there could be a continuum of characteristic timescales (which do not scale simply with luminosity, black hole mass or Eddington ratio) such that the combination of these timescales produces an SPL structure function shown by my results.

On the longest timescales of $\Delta t \sim 10^4$ days, the subensemble structure functions are not well differentiated by luminosity or black hole mass, which is evidenced by the subensemble groups having a similar structure function amplitudes on these timescales. This is also the case for Eddington ratio in the g band. These results suggest that the dependence of variability amplitude on black hole properties depends on the rest-frame timescales of variability. My findings are consistent with those of Arévalo et al. (2023), who found that the anti-correlation of variability with black hole mass and Eddington ratio is stronger at short timescales. I investigate this effect further in Section 5.7

5.7 Dependence of properties depends on timescales

It is clear from the subensemble structure functions presented in Figures 5.10, 5.11 and 5.12 that the relation between quasar properties and the amplitude of variability depends on the rest-frame timescale of that variability. This effect has been observed by Arévalo et al. (2023) but a correlation between structure function amplitude and quasar properties for varying timescales remains unexplored. In this section, I will investigate this phenomenon.

5.7.1 Methods

To calculate the correlation between quasar properties and variability amplitude, I use an alternative method to the subensemble approach. I calculated the structure function for each quasar with the same set of Δt bins used previously. This results in a very sparse structure function for each quasar. However, this is not a problem, since I will be looking for correlations between values of the structure function on different timescales and quasar properties, rather than plotting the structure function as I have done previously. For a particular Δt bin, I calculate the Spearman correlation coefficient, ρ, between quasar properties and the values of the structure function for that bin, $\mathrm{SF}(\Delta t)$. I repeat this process over a set of Δt bins and different quasar properties (luminosity, black hole mass and Eddington ratio). Using luminosity as an example, the Spearman correlation coefficient with $\mathrm{SF}(\Delta t)$ may be expressed as

$$\rho(\Delta t) = \rho\left(L_{\mathrm{bol}}, \mathrm{SF}(\Delta t)\right), \tag{5.19}$$

with a standard error,

$$\sigma = \sqrt{\frac{1-\rho^2}{n-2}}, \tag{5.20}$$

where n is the number of points in the bin. The Spearman correlation coefficient can take values $-1 < \rho < 1$, where -1 implies a perfect monotonic negative relationship, $+1$ implies a perfect monotonic positive relationship, and 0 implies no correlation. For each ρ, I also calculate the p-value for a two-sided hypothesis test, whose null hypothesis is that of no ordinal correlation. The p-value roughly

indicates the probability of an uncorrelated system that has the same (or greater) value of ρ. The quantity $\rho(\Delta t)$ enables me to quantify the dependence of quasar properties on variability as a function of timescale.

5.7.2 Results

Figure 5.14 shows the Spearman correlation coefficient, $\rho(\Delta t)$, plotted against Δt for luminosity, black hole mass and Eddington ratio. I included the p-value for each calculation of ρ, using a second axis. Note that most of these values are well below $p = 10^{-4}$ and are therefore not present on the plot. For each property, there are an average of $\sim 50,000$ points per bin, however, there are significantly fewer points in the final few bins which is the cause of the larger error bars and higher p-values. The number of points in the final three Δt bins are 9276, 2506, 302, respectively.

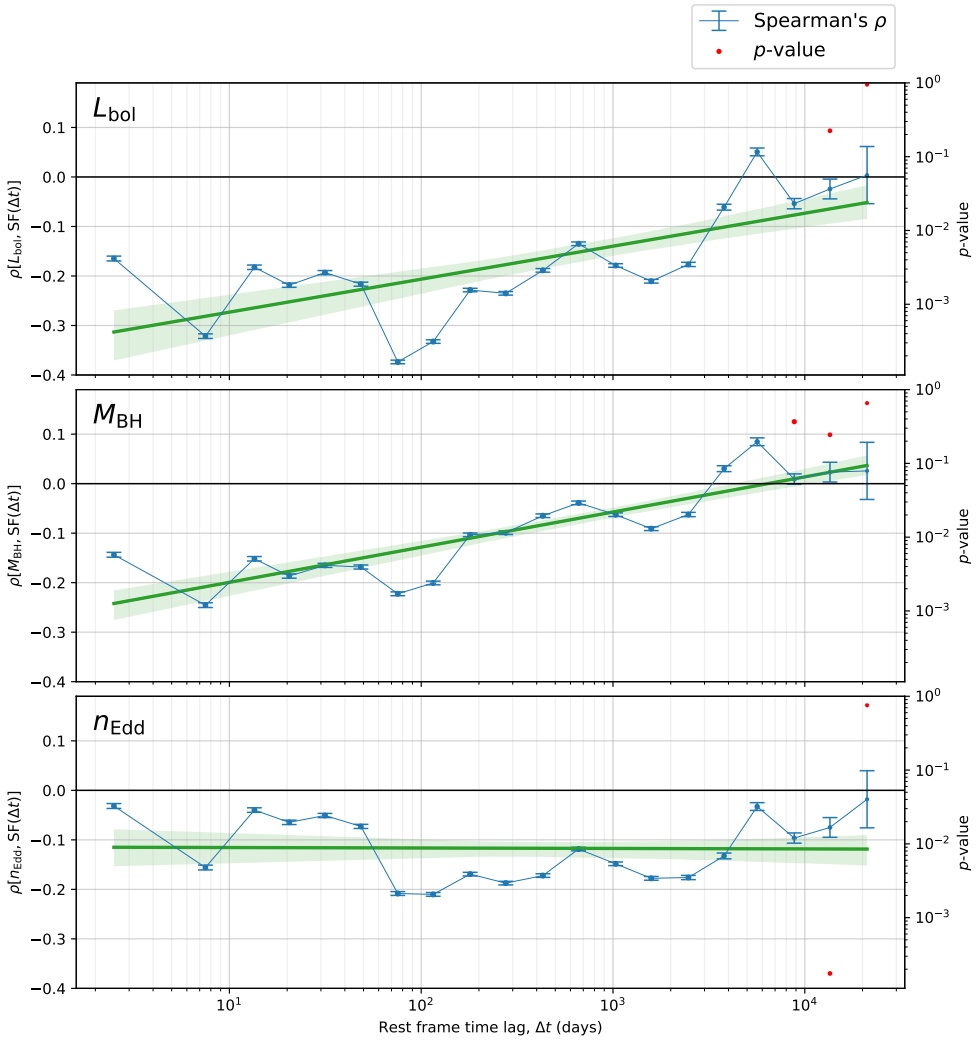

Figure 5.14: Spearman correlation coefficient between variability amplitude and $L_{\rm bol}$, $M_{\rm BH}$ and $n_{\rm Edd}$ for varying timescales (blue). A line of best fit is overplotted on each panel (green) showing a 2σ confidence interval (shaded green). p-values are shown using the right-hand axis, however, the lower axis limit is set to 10^{-4} and the majority of p-values are well below this limit.

5.7.3 Discussion

My results show that the anticorrelation between bolometric luminosity and variability amplitude depends strongly on the timescale of such variability. On timescales $\Delta t < 10^3$ days, the anticorrelation is strong at $\rho \approx -0.25$, but becomes much weaker on longer timescales and even disappears on timescales of 2×10^4 days. The same anticorrelation is true for black hole mass, but to a lesser degree on shorter timescales. Interestingly, the correlation becomes positive

for $\Delta t > 2\times10^4$ days, which could explain the positive correlation seen by Wilhite et al. (2008); MacLeod et al. (2010), and Lu et al. (2019). However, since $L_{\rm bol}$ and $M_{\rm BH}$ are themselves correlated, a more detailed analysis is required to decouple these correlations. Such an analysis is certainly possible with 7-DQ, but it is beyond the scope of this thesis and will be reserved for a future publication.

The correlation between variability and Eddington ratio is remarkably constant with time-lag, and is scattered around an average value of $\rho = -0.11$. Overall, the correlation is weak, but statistically significant. This persists until timescales of 2×10^4 days, at which point the correlation disappears, as with luminosity and black hole mass. These are striking results and suggest that on timescales of > 50 years, the amplitude of variability exhibited by all quasars, irrespective of their properties, is the same.

5.8 Effect of Rest-Frame Wavelength on the Structure Function

The effect of variability on rest-frame wavelength in quasars has been reported in the literature, with numerous studies showing that bluer wavelengths are generally more variable. The best demonstration of this was by Vanden Berk et al. (2004), showing a smooth change of structure function amplitude from $\sim$0.15 mag at 6000 Å to $\sim$0.4 mag at 1000 Å. To test this phenomenon, I split the photometry into groups of rest-frame wavelength. The rest-frame wavelength is calculated from the redshift and band that the observation was taken from. I used the following transformations to move the quasar photometry to the rest-frame,

$$\lambda_g = \frac{4866\text{Å}}{1+z}, \qquad \lambda_r = \frac{6215\text{Å}}{1+z}, \qquad \lambda_i = \frac{7545\text{Å}}{1+z}, \tag{5.21}$$

where I have scaled the mean wavelength of each band in the Pan-STARRS system (given by Tonry et al. 2012) by the redshift, z, of the quasar. The result is shown in Figure 5.15. There is some clear wavelength dependence on timescales $10^2 - 10^3$ days, however, the effect is very weak.

My method does not use the same grouping technique as I have done previously for luminosity, black hole mass and Eddington ratio. This is because my pairwise database and analysis pipeline has been aimed at performing analysis separately for the g, r and i bands. Therefore, in order to calculate the structure function

in groups of rest-frame wavelength, I calculated the value of the structure function for each quasar and combined them after grouping into bins of rest-frame wavelength. Using this method, I am not able to calculate the variance-weighted structure function, such that the wavelength grouped structure functions are overweighted by noisy measurements. While it would be possible to study the rest-frame wavelength dependence on amplitude using 7-DQ, it would require a refactor of my analysis pipeline, which is beyond the scope of this thesis and reserved for future work.

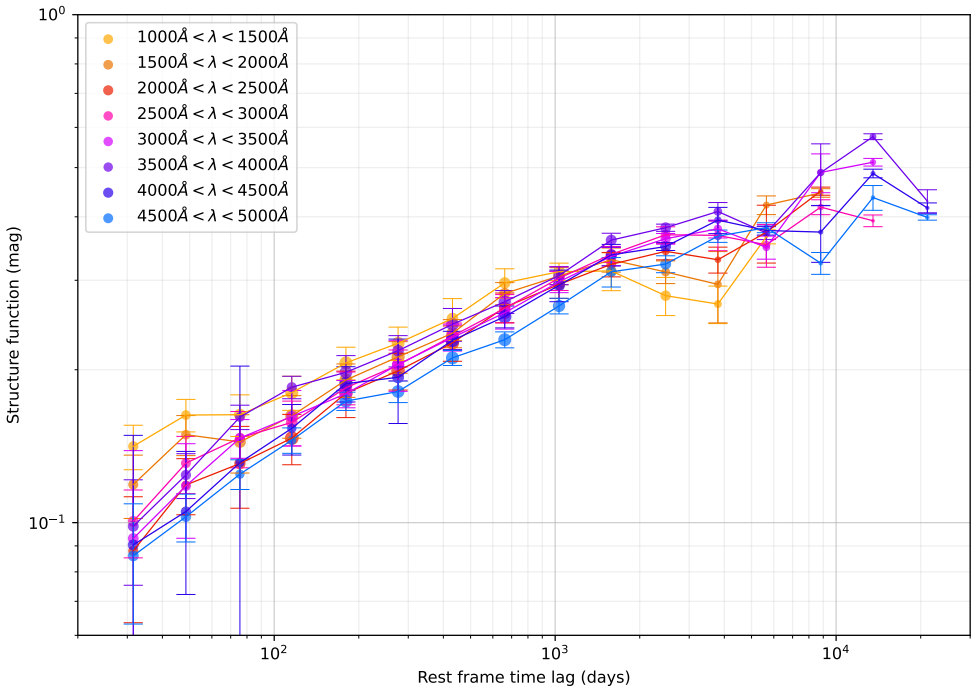

Figure 5.15: Structure functions for quasars split into groups of rest-frame wavelength over the range 1000 Å – 5000 Å.

The emitted spectrum of a simple accretion disk model depends on wavelength (Netzer, 2013) and follows:

$$L_\nu \propto \lambda^{-1/3}. \tag{5.22}$$

I tested this using my structure function amplitudes for timescales within a set of 5 time bins, shown in Figure 5.16. My results do not show a significant correlation with wavelength. The discrepancy is likely to be a combination of two factors. First, the broad filters lack the spectral resolution needed to accurately test the expected correlation. Second, this correlation assumes a simple accretion disk

model, relying on assumptions that are unlikely to hold for my 7-DQ quasars.

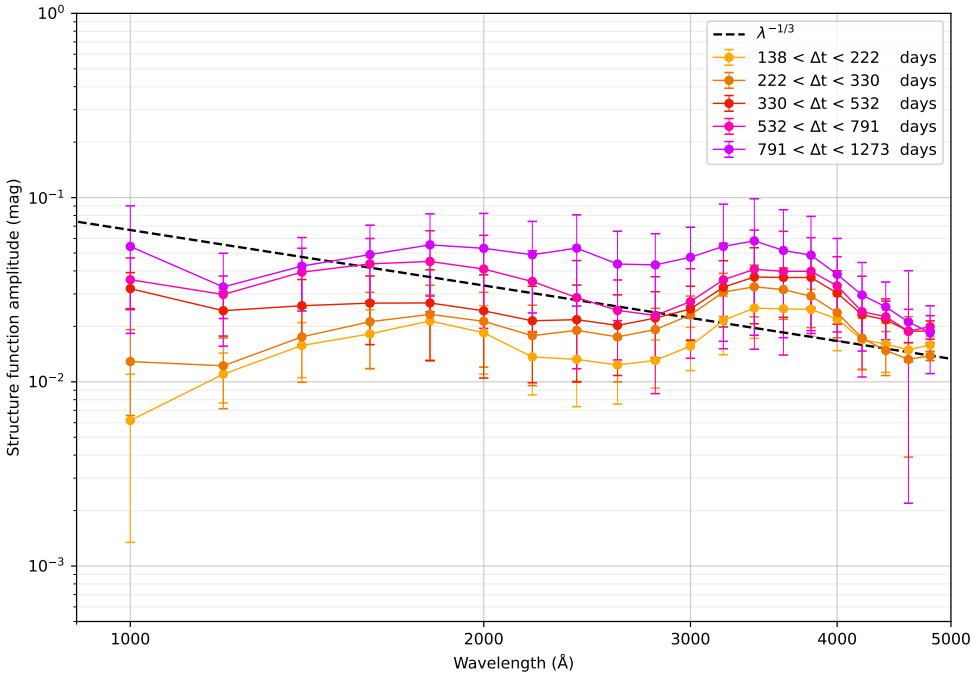

Figure 5.16: Structure function amplitudes versus wavelength for five time bins, with the expected slope of a simple accretion disk model, $\lambda^{-1/3}$, overplotted (black, dotted)

5.9 Summary

Since the bulk of my analysis falls under the same theme of ensemble structure function studies, this chapter is extensive and represents a large majority of the work in this thesis. In Section 5.3, I developed a new definition of the structure function which is robust to outliers and enables the inclusion of noisier data by underweighting it. Subsequently, in Section 5.4, I presented the most precise quasar ensemble structure function to date, using an unprecedented sample size and volume of photometry. This result extends the quasar structure function to a new time domain of 70 years, previously unexplored, and demonstrates that the ensemble structure function follows a single power law on all observed timescales. In Section 5.5, I revealed asymmetries in the behaviour of quasar variability. By comparing with a DRW flare simulation, my results suggest that fluctuations in the optical continuum of quasars favour a fast-rise slow-decline profile. Furthermore, in Section 5.6, I split my quasar sample into groups of

similar properties and studied their respective structure functions. The results show that luminosity, black hole mass and Eddington ratio affect the slope and amplitude of the structure function. In Section 5.7, I presented a novel study on the correlation between quasar properties and structure function amplitude for varying timescales, quantifying these relationships as Spearman correlation coefficients, confirming the commonly reported anticorrelation with luminosity and Eddington ratio, and showing that all quasars exhibit similar levels of variability on timescales > 50 years, irrespective of luminosity, black hole mass and Eddington ratio. Finally, in Section 5.8, I conducted a brief study to compute the structure function in ranges of rest-frame wavelength, and while my results are consistent with other studies, my analysis pipeline is not suited to perform a detailed analysis of this phenomenon, which is reserved for future work.

Chapter 6

Modelling quasars as a Damped Random Walk

6.1 Introduction

Kelly et al. (2009) first proposed the DRW as a model for quasar optical light curves, and subsequent work has confirmed that it is a viable description of optical continuum variability on timescales > 1 year (see e.g., MacLeod et al. 2010, 2012; Kozłowski et al. 2010; Zu et al. 2013; Andrae et al. 2013).

Some studies show deviations from the DRW model, although these are usually on short timescales (< 1 year). For example, slopes of the power spectral distributions (PSDs) of AGN from the *Kepler* mission are too steep to be consistent with predictions of the DRW (see e.g., Mushotzky et al. 2011; Kasliwal et al. 2015; Smith et al. 2018). Additionally, Kozłowski (2016a) claims a degeneracy in the DRW model, stating that a good DRW fit does not necessarily mean that AGN variability is driven by DRW stochastic processes.

Despite these challenges, the DRW model remains a popular tool and has been used for a number of applications, including developing AGN selection techniques (Butler & Bloom 2011; MacLeod et al. 2011; Ruan et al. 2012; Choi et al. 2014) and producing realistic simulated quasar light curves (Suberlak et al. 2021). However, its primary application is to parameterise variability and generate parameters which act as a proxy for accretion mechanisms. Studies have shown that these parameters correlate with quasar properties such as luminosity and

black hole mass (see e.g., MacLeod et al. 2010; Kelly et al. 2011; Burke et al. 2021).

In this chapter, I explore the damped random walk (DRW) model and assess its suitability for modelling quasar light curves. I fit the 7-DQ quasar light curves with a (DRW) model and correlate the DRW model parameters with quasar physical properties. The DRW process is defined by the equation,

$$\frac{dx}{dt} + \tau_{\rm DRW} x = \sigma_{\rm DRW} \epsilon(t), \tag{6.1}$$

and is discussed in detail in Chapter 1, Section 1.9.3.

6.2 Fitting DRW parameters

I employ the `EzTao` package (Yu & Richards 2022) to achieve a fast, precise fit of my quasar light curves using the DRW model. This Python toolkit specializes in time-series analysis using CARMA processes. `EzTao` uses `celerite` (a fast Gaussian process regression library; Foreman-Mackey et al. 2017a) as its backend, which enables fast evaluation of the likelihood function.

During the fitting process, I used `EzTao` to select the maximum a posterior (MAP) estimation of $\tau_{\rm DRW}$ and $\sigma_{\rm DRW}$ for each light curve. `EzTao` achieves this via maximum likelihood estimation (MLE) of the DRW model. After I obtained the MAP for each object, I used `emcee` (Foreman-Mackey et al. 2013), a Python implementation of Goodman & Weares Affine Invariant Markov Chain Monte Carlo (MCMC) Ensemble sampler (Goodman & Weare 2010), to sample the posterior distribution with the MCMC walkers initiated at the MAP position. An additional benefit of MCMC is that it enables me to explore the parameter space and refine the DRW parameters to obtain the best-fit. Additionally, MCMC provides a posterior distribution of $\tau_{\rm DRW}$ and $\sigma_{\rm DRW}$ from which I can draw uncertainties on the parameters. I illustrate this process for a particular light curve, shown in Figure 6.1; Figure 6.2 shows a corner plot of the posterior distribution of $\sigma_{\rm DRW}$ and $\tau_{\rm DRW}$ for this light curve. Additionally, I show the structure function for the same light curve, along with the DRW structure function for the median values of $\tau_{\rm DRW}$ and $\sigma_{\rm DRW}$, in Figure 6.3.

Many of the quasar light curves in 7-DQ are sparse, and some only have data from certain surveys. To provide suitable DRW fits, I only included those which

satisfy the following constraint:

$$(n_{\rm SDSS} + n_{\rm PS} > 11) \quad \text{AND} \quad (n_{\rm ZTF} > 100), \tag{6.2}$$

resulting in a subset of 37,445 light curves. This subset has an average of 289 observations per light curve, which is an adequate number of measurements to provide reliable DRW fits. I plot the joint distribution of $\tau_{\rm DRW}$ and $\sigma_{\rm DRW}$ in Figure 6.4. This distribution shows that $\tau_{\rm DRW}$ and $\sigma_{\rm DRW}$ are highly degenerate. Interestingly, there is a second bulge of $\tau_{\rm DRW}$ at short timescales. However, these timescales, being less than 10 days, are unlikely to correspond to any physical phenomena in the light curves. Further investigation of light curves which yielded small $\tau_{\rm DRW}$ showed that some of these light curves have underquoted errors on the ZTF observations, such that the DRW is interpreting photometric noise as genuine intrinsic variability. Since $\tau_{\rm DRW}$ affected by this are unreliable, I will focus my analysis on fits with $\tau_{\rm DRW} > 10^2$ days.

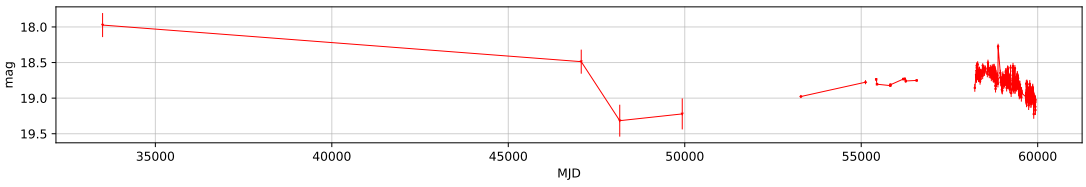

Figure 6.1: A typical quasar light curve from the 7-DQ database.

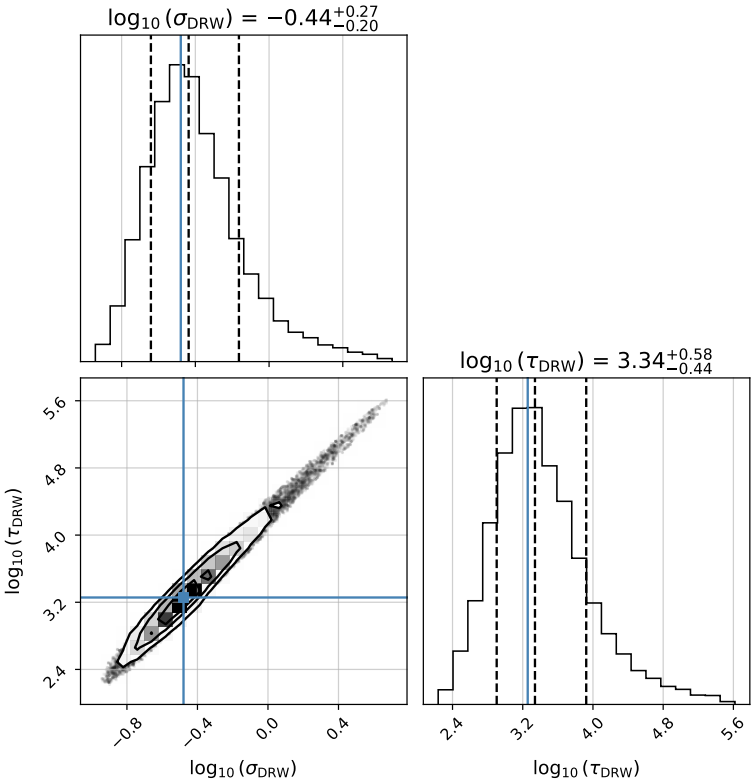

Figure 6.2: MCMC corner plot showing the distribution of $\tau_{\rm DRW}$ and $\sigma_{\rm DRW}$ for the light curve in Figure 6.1. The maximum a posterior (MAP) estimate is overplotted in blue. The black dotted lines on the histograms are the 16th, 50th and 84th percentiles. The 50th percentile is used as the best estimate of the parameter, while the 16th and 84th correspond to the -1σ and $+1\sigma$ uncertainties, respectively.

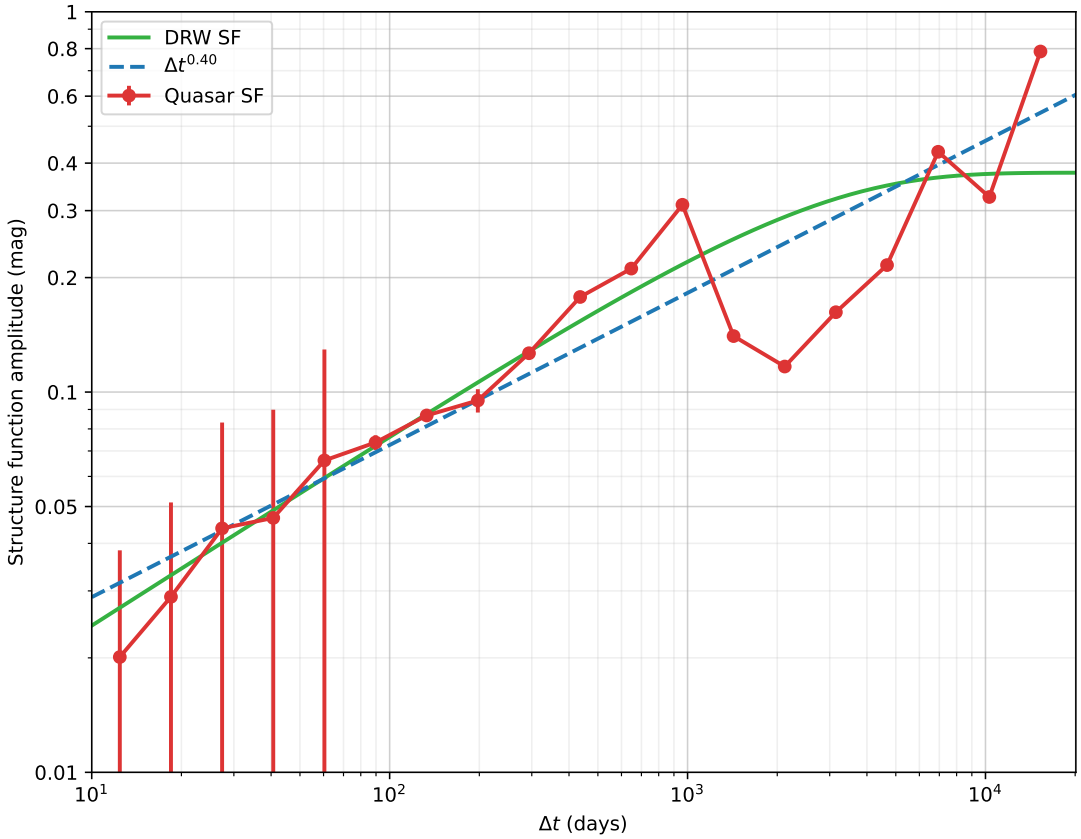

Figure 6.3: Structure function of the light curve shown in Figure 6.1 (red), and the DRW structure function using the median values of $\tau_{\rm DRW}$ and $\sigma_{\rm DRW}$ from the MCMC fit shown in Figure 6.2 (green). The best fit power law to the quasar structure function is overplotted and has a slope of 0.40 (blue, dotted).

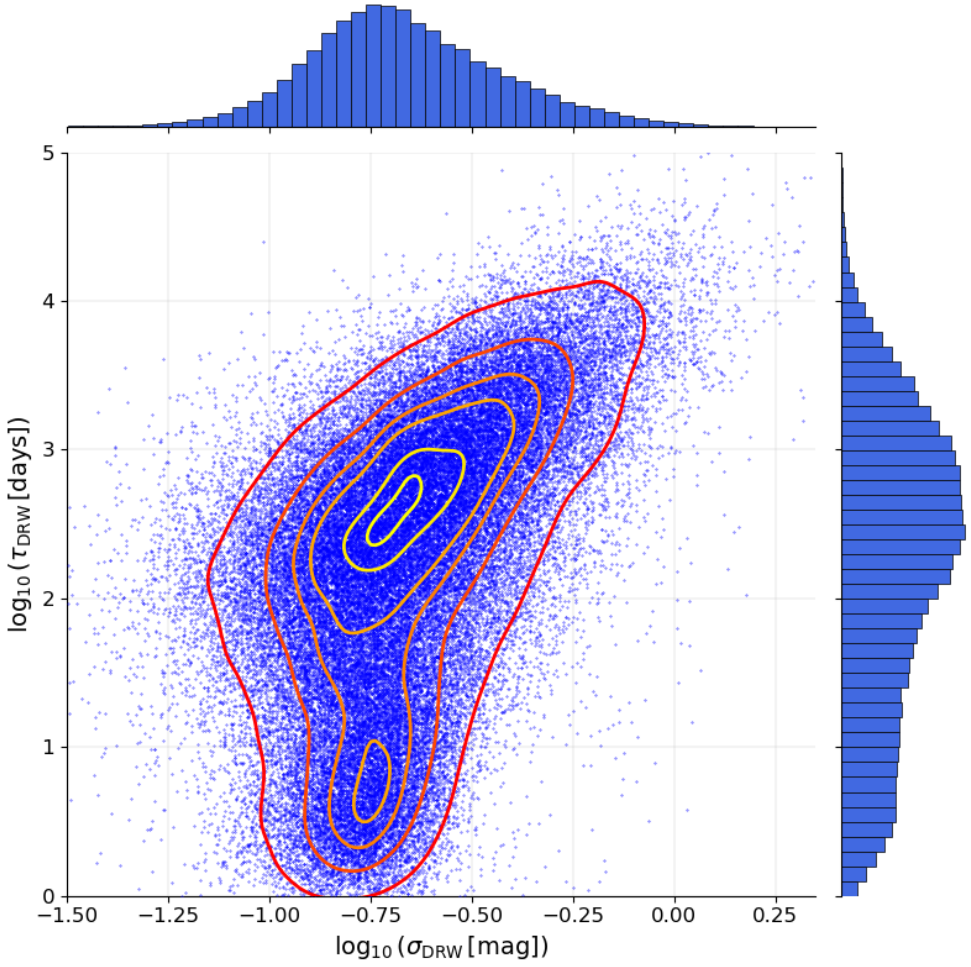

Figure 6.4: Scatter plot showing the distribution of $\tau_{\rm DRW}$ against $\sigma_{\rm DRW}$ for my DRW fits. Density contours are drawn for 10%–90% with 10% increments, with the highest contour drawn at 98%.

6.3 Searching for long characteristic timescales

Some studies have obtained DRW parameters on long baselines (see e.g., Kelly et al. 2009; MacLeod et al. 2010; Simm et al. 2016; Stone et al. 2022). By comparing SDSS to POSS plate data, MacLeod et al. (2012) obtained DRW parameters over the longest baseline to date, using light curves spanning 50 years.

Kozłowski (2017) show that, in order to constrain the characteristic timescale $\tau_{\rm DRW}$, the duration of the light curve must be substantially longer than $\tau_{\rm DRW}$. Their analysis shows that $\tau_{\rm DRW}$ is only reliable up to ~10% of the baseline of the light curve. Therefore, with my extended baseline of 70 years, I am able to fit

DRW parameters to my longest light curves to search for $\tau_{\rm DRW}$ on timescales of 7 years, significantly longer than the previous longest study by MacLeod et al. (2012).

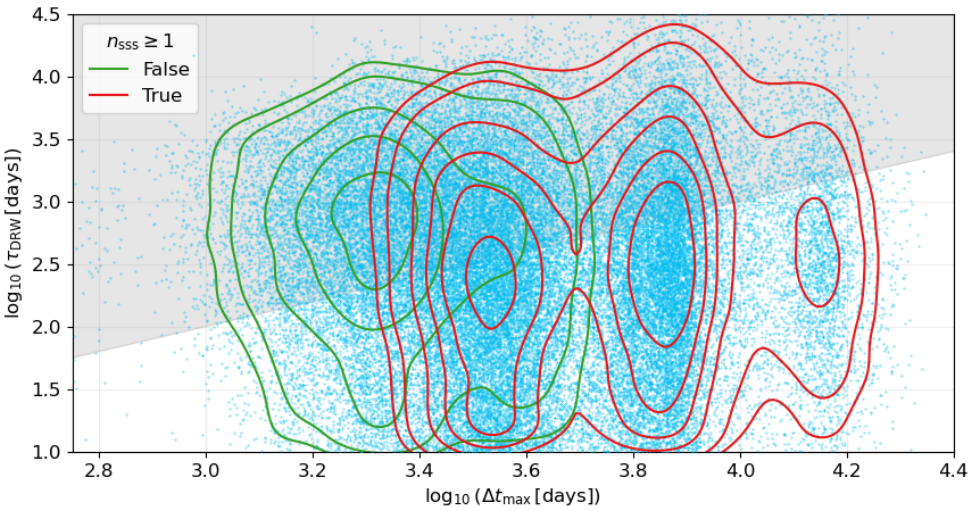

Figure 6.5: Scatter plot showing the distribution of $\tau_{\rm DRW}$ against length of the light curve, $\Delta t_{\rm max}$. Overplotted are density contours for light curves with at least one observation (no observations) in SuperCOSMOS, plotted in red (green). Points within the shaded grey region show points which have unreliable $\tau_{\rm DRW}$, i.e., $\tau_{\rm DRW} > 0.1 \times \Delta t_{\rm max}$.

Figure 6.5 shows that long baselines are needed in order to achieve reliable $\tau_{\rm DRW}$. Additionally, the contours illustrate the importance of the plate photometry. Recent studies do not usually include historic photometry (see e.g., Stone et al. 2022). However, only the longest reliable $\tau_{\rm DRW}$ can be obtained by including this data. The SuperCOSMOS data increases the max Δt in the dataset from 19 to 63 years.

6.4 Correlation of DRW parameters with quasar properties

DRW parameters are believed to be a simplified representation for real, physical mechanisms responsible for producing variability. Therefore, correlations between quasar properties and DRW parameters can give clues to which of these properties has the greatest effect on variability.

In Figure 6.6, I plot $\tau_{\rm DRW}$ and $\sigma_{\rm DRW}$ against $L_{\rm bol}$, $M_{\rm BH}$, $n_{\rm Edd}$, and rest-frame wavelength, $\lambda_{\rm rf}$. These plots show density contours for the g, r and i bands separately, with a best-fit regression line to the combined gri data. Since there are more points in the r band, the regression is weighted toward these data and does not necessarily pass through the centres of the contours in each band. As these are log-log plots, the slope represents the index between each pair. For example, the anticorrelation between $\sigma_{\rm DRW}$ and $L_{\rm bol}$ is best described by the power law $\sigma_{\rm DRW} \propto L_{\rm bol}{}^{-0.121\pm0.002}$. The best-fit slopes are presented in Table 6.1.

Many of these slopes are flat, suggesting no correlation. However, there is a significant dependence $\sigma_{\rm DRW} \propto L_{\rm bol}{}^{-0.121\pm0.002}$, confirming the commonly reported phenomenon that variability is anticorrelated with luminosity. There is a similar dependence with Eddington ratio: $\sigma_{\rm DRW} \propto n_{\rm Edd}{}^{-0.162\pm0.002}$, as expected. Furthermore, there is a clear positive correlation $\tau_{\rm DRW} \propto \lambda_{\rm rf}{}^{0.339\pm0.009}$, and I am able to provide tighter constraints on the same correlation seen by Stone et al. (2022), who report a very similar dependence $\tau_{\rm DRW} \propto \lambda_{\rm rf}{}^{0.34\pm0.10}$.

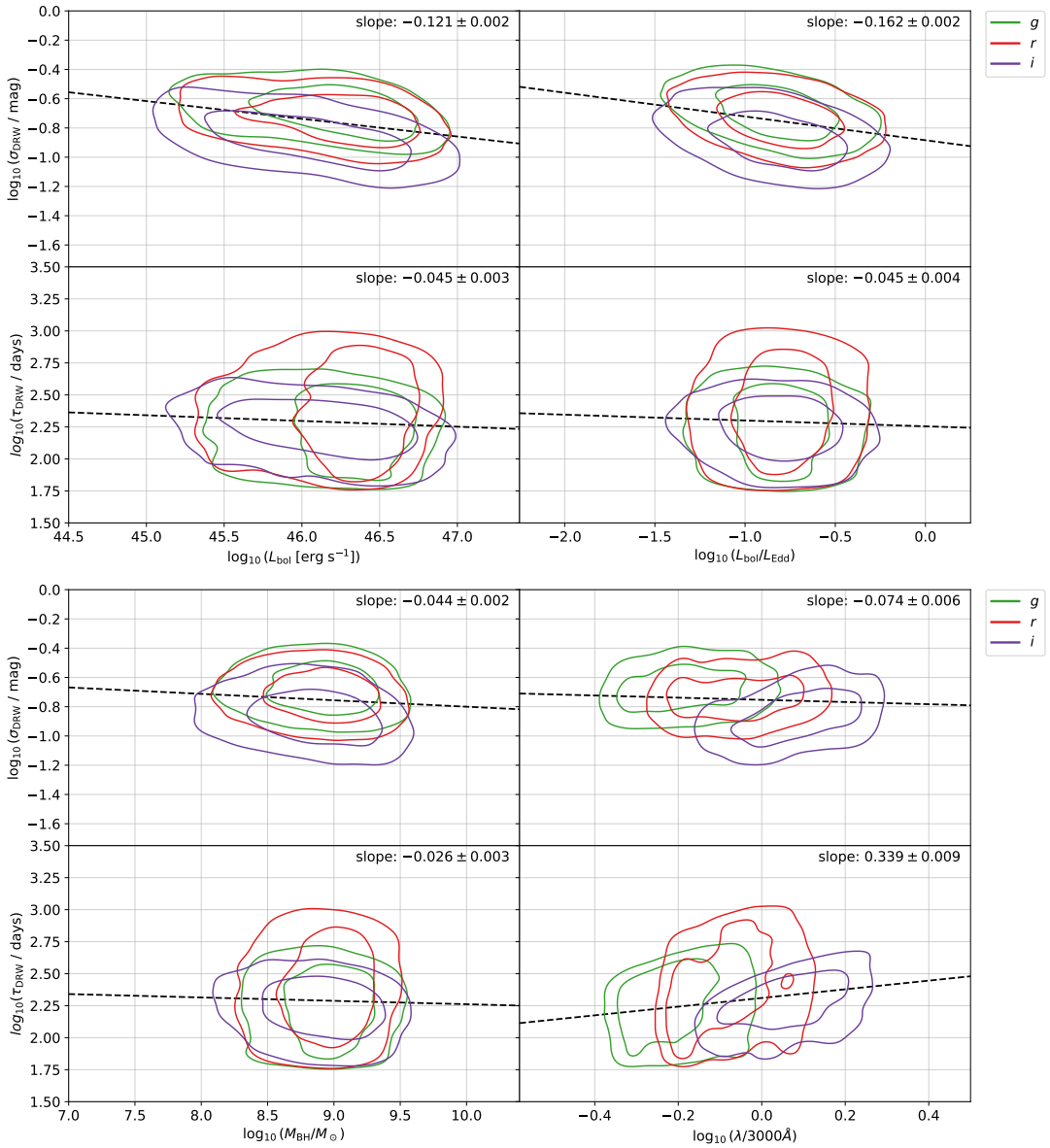

Figure 6.6: Joint distributions of DRW parameters, $\sigma_{\rm DRW}$ and $\tau_{\rm DRW}$, against $L_{\rm bol}$, $n_{\rm Edd}$ (top panel), and $M_{\rm BH}$, $\lambda_{\rm rf}$ (bottom panel). Contours show 30% and 70% of the data, coloured to distinguish data from each of the g, r and i bands. The best-fitting linear regression is shown as the black dotted line, with its slope marked in each panel.

DRW parameter	Property	Slope
$\sigma_{\rm DRW}$	$L_{\rm bol}$	-0.121 ± 0.002
$\sigma_{\rm DRW}$	$n_{\rm Edd}$	-0.162 ± 0.002
$\sigma_{\rm DRW}$	$M_{\rm BH}$	-0.044 ± 0.002
$\sigma_{\rm DRW}$	$\lambda_{\rm rf}$	-0.074 ± 0.006
$\tau_{\rm DRW}$	$L_{\rm bol}$	-0.045 ± 0.003
$\tau_{\rm DRW}$	$n_{\rm Edd}$	-0.045 ± 0.004
$\tau_{\rm DRW}$	$M_{\rm BH}$	-0.026 ± 0.003
$\tau_{\rm DRW}$	$\lambda_{\rm rf}$	$+0.339 \pm 0.009$

Table 6.1: Summary of the slopes from the best-fit regression lines of Figure 6.6

6.5 Ensemble DRW structure function

A natural question to ask is whether the ensemble quasar structure function can be reproduced by summing the DRW structure functions obtained from fitting a sample of individual quasars. The structure function for a single DRW is given by:

$$\mathrm{SF}_{\rm DRW}(\Delta t) = \sqrt{2}\sigma_{\rm DRW}(1 - e^{-\Delta t/\tau_{\rm DRW}})^{1/2}. \tag{6.3}$$

To create an ensemble DRW structure function, I combined $\mathrm{SF}_{\rm DRW}$ in quadrature for the values of $\tau_{\rm DRW}$ and $\sigma_{\rm DRW}$ obtained in Section 6.2. I did this for two sets of parameters; those which contain all fitted $\tau_{\rm DRW}$, and those which satisfy $\tau_{\rm DRW} < 0.1 \times \Delta t_{\rm max}$ to ensure reliable $\tau_{\rm DRW}$. The results are shown in Figure 6.7.

There are clear discrepancies between the shape of the DRW structure function, and the observed quasar structure function. At short timescales, the DRW structure function is much too steep, with a slope of ~0.5, while at long timescales, the DRW structure function plateaus.

While the DRW parameters may be used as a representation for variability amplitude and its associated timescale, this result demonstrates that the DRW process is not an accurate representation for physical mechanisms governing variability in the optical continuum.

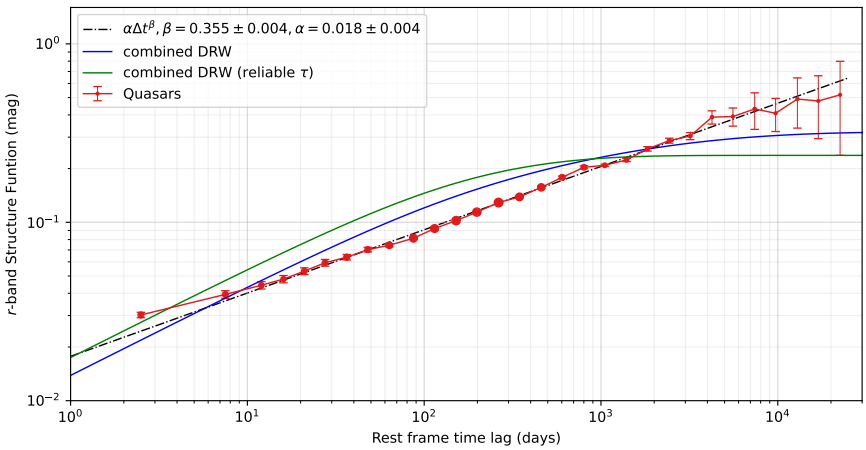

Figure 6.7: The ensemble DRW structure function for DRW parameters obtained from a subset of 7-DQ quasars. Overplotted is the ensemble quasar structure function in the r-band with its best-fit power law.

6.6 Summary

In this chapter, I fitted a DRW model to my 7-DQ quasar light curves to obtain reliable DRW parameters on longer timescales compared with previous studies. By searching for correlations between these parameters and quasar properties, I show that $\tau_{\rm DRW}$ is anticorrelated with both luminosity and Eddington ratio. Additionally, my results show a clear correlation between $\tau_{\rm DRW}$ and $\lambda_{\rm rf}$, and provide tighter constraints compared to the same result reported by Stone et al. (2022). Furthermore, I tested the suitability of the model by comparing the ensemble DRW structure function against the ensemble structure function of my 7-DQ quasars. There is a clear disparity between the shapes of these structure functions, implying that the DRW is not an accurate representation of the physical mechanisms driving variability.

The DRW model is popular for its simplicity; however, it lacks the flexibility to accurately reproduce quasar structure functions from observations. Although the parameters $\tau_{\rm DRW}$ and $\sigma_{\rm DRW}$ offer insights into the timescales and variability amplitudes of quasar light curves, it is ambitious to expect these two parameters to capture all the complex physical processes involved in black hole accretion. To effectively parameterise quasar light curves, more physically motivated models are needed. Developing such models first requires advancing our understanding of these physical processes, which is the primary aim of this thesis.

Chapter 7

Conclusion

7.1 Summary

This thesis represents a comprehensive exploration of quasar variability, addressing key issues within the field.

In Chapter 2, I presented the creation of 7-DQ; a comprehensive database of quasar photometry that is unprecedented in sample size, number of observations and temporal baseline. Furthermore, in Chapter 3, I presented the steps I took to preprocess 7-DQ, including a novel, fast computational algorithm for calculating observation pairs. 7-DQ holds many clues that will lead to further understanding quasar variability. My analysis of it is not exhaustive, and there is still plenty of scope for exploration. Therefore, 7-DQ serves as a valuable resource that benefit the wider scientific community that are researching quasar variability.

In Chapter 4, I presented a set of Δm distributions from observation pairs of quasar photometry. I demonstrated that Gaussian mixtures are a viable model for fitting these distributions. I showed that the mean magnitude of quasars do not change significantly over timescales up to 70 years, contrasting to recent claims. Additionally, by computing higher order moments, I presented the skewness and kurtosis of these distributions, which is a novel result and can be used as a benchmark for future studies.

In Chapter 5, I presented the ensemble structure function of ~500,000 quasars in the g, r and i bands, and showed that they are consistent with single power

laws up to timescales of 70 years. Additionally, I found asymmetries in the ensemble structure function which are dependent on timescale, suggesting that quasar variability favours fast-rise slow-decline profiles. Furthermore, by splitting my sample into groups of quasars of similar properties, I demonstrated that the slope and amplitude of the structure function is dependent on these properties. Finally, I showed that the dependence of quasar properties on the amplitude of variability depends on the timescale, with the exception of Eddington ratio, which has a constant relationship with variability. A consistent theme across this thesis is that all quasars exhibit similar levels of variability on timescales > 50 years, irrespective of luminosity, black hole mass and Eddington ratio. This was evidenced by results in this chapter, as well as the GMM fits in Chapter 4.

In Chapter 6, I fitted a DRW model to my 7-DQ quasar light curves, obtaining reliable parameters on longer timescales than previous studies. By searching for correlations between these parameters and quasar properties, I found that $\tau_{\rm DRW}$ is anticorrelated with both luminosity and Eddington ratio, and correlated with $\lambda_{\rm rf}$, providing tighter constraints than Stone et al. (2022). Comparing the ensemble DRW structure function with the ensemble structure function of 7-DQ quasars revealed significant differences, indicating that the DRW model does not accurately represent the physical mechanisms driving variability, motivating the development of more physically motivated models.

Collectively, the research presented in this thesis represent a significant contribution to our understanding of quasar variability. In addition, I open new avenues for possible exploration and lay down a robust foundation for future research in this field.

7.2 Future work

The 7-DQ database offers numerous possibilities for new studies, several of which have been initiated in this thesis. For example, my Gaussian Mixture Modelling (GMM) of the Δm distributions (see Chapter 4) could be further explored through a quantitative analysis of these GMM fits. In particular, one could examine the weights and widths of Gaussian components to derive physically meaningful quantities of variability.

Chapter 5 of this thesis focused on structure function analysis, a valuable tool for

investigating quasar variability across different timescales. However, theoretical and simulated models currently lack structure function predictions that agree with observed data. While the Damped Random Walk (DRW) model is a useful prescription for quasar variability, it does not capture the underlying physical processes occurring within the accretion disk. Therefore, it is crucial to develop physically-motivated models that predict statistics of variability based on quasar properties, such as black hole mass, luminosity, and Eddington ratio. The predictions of such models could then be directly compared with the structure function results in Chapter 5.

Ultimately, the most important requirement for future studies of quasar variability is the continuation of variability surveys and the combination of data across multiple surveys to maximise the observational baseline during analysis. A longer baseline serves as a lever, offering greater insight into long-term quasar variability and enabling clearer discrimination between models, such as the DRW model.

Bibliography

Abramowicz M. A., 1991, in Duschl W. J., Wagner S. J., Camenzind M., eds, , Vol. 377, Variability of Active Galaxies. p. 255, doi:10.1007/BFb0030069

Ahumada R., et al., 2020, ApJS, 249, 3

Aigrain S., Foreman-Mackey D., 2023, ARA&A, 61, 329

Andrae R., Kim D. W., Bailer-Jones C. A. L., 2013, A&A, 554, A137

Angione R. J., Smith H. J., 1972, in Evans D. S., Wills D., Wills B. J., eds, Proceedings of the International Astronomical Union Symposium Vol. 44, External Galaxies and Quasi-Stellar Objects. p. 171

Arévalo P., Lira P., Sánchez-Sáez P., Patel P., López-Navas E., Churazov E., Hernández-García L., 2023, MNRAS, 526, 6078

Balbus S. A., Hawley J. F., 1991, ApJ, 376, 214

Bauer A., Baltay C., Coppi P., Ellman N., Jerke J., Rabinowitz D., Scalzo R., 2009, ApJ, 696, 1241

Becker R. H., White R. L., Helfand D. J., 1995, ApJ, 450, 559

Beckmann V., Shrader C., 2012, in Proceedings of "An INTEGRAL view of the high-energy sky (the first 10 years)" - 9th INTEGRAL Workshop and celebration of the 10th anniversary of the launch (INTEGRAL 2012). 15-19 October 2012. Bibliotheque Nationale de France. p. 69 (arXiv:1302.1397), doi:10.22323/1.176.0069

Bellm E. C., et al., 2019, PASP, 131, 018002

Blandford R. D., McKee C. F., 1982, ApJ, 255, 419

Brockwell P., 2001, in Handbook of Statistics, Vol. 19, Stochastic Processes: Theory and Methods. Elsevier, pp 249–276, doi:https://doi.org/10.1016/S0169-7161(01)19011-5

Burke C. J., et al., 2021, Science, 373, 789

Butler N. R., Bloom J. S., 2011, AJ, 141, 93

Caplar N., Pena T., Johnson S. D., Greene J. E., 2020, ApJ, 889, L29

Chambers K. C., et al., 2016, arXiv e-prints, p. arXiv:1612.05560

Choi Y., Gibson R. R., Becker A. C., Ivezić Ž., Connolly A. J., MacLeod C. L., Ruan J. J., Anderson S. F., 2014, ApJ, 782, 37

Clavel J., et al., 1990, MNRAS, 246, 668

Collier S., Peterson B. M., 2001, ApJ, 555, 775

Cristiani S., Trentini S., La Franca F., Andreani P., 1997, A&A, 321, 123

Dexter J., Begelman M. C., 2019, MNRAS, 483, L17

Edge D. O., Shakeshaft J. R., McAdam W. B., Baldwin J. E., Archer S., 1959, Mem. RAS, 68, 37

Elvis M., et al., 1994, ApJS, 95, 1

Event Horizon Telescope Collaboration et al., 2019, ApJ, 875, L1

Fabian A. C., 2012, ARA&A, 50, 455

Fabian A. C., Iwasawa K., Reynolds C. S., Young A. J., 2000, PASP, 112, 1145

Feigelson E. D., Babu G. J., Caceres G. A., 2018, Frontiers in Physics, 6, 80

Ferrarese L., Merritt D., 2000, ApJ, 539, L9

Flewelling H. A., et al., 2020, ApJS, 251, 7

Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, PASP, 125, 306

Foreman-Mackey D., Agol E., Angus R., Ambikasaran S., 2017a, ArXiv

Foreman-Mackey D., Agol E., Ambikasaran S., Angus R., 2017b, AJ, 154, 220

Fukugita M., Ichikawa T., Gunn J. E., Doi M., Shimasaku K., Schneider D. P., 1996, AJ, 111, 1748

Giveon U., Maoz D., Kaspi S., Netzer H., Smith P. S., 1999, MNRAS, 306, 637

Goodman J., Weare J., 2010, Communications in Applied Mathematics and Computational Science, 5, 65

Gunn J. E., et al., 2006, AJ, 131, 2332

Haardt F., Maraschi L., 1991, ApJ, 380, L51

Hambly N. C., et al., 2001a, MNRAS, 326, 1279

Hambly N. C., Irwin M. J., MacGillivray H. T., 2001b, MNRAS, 326, 1295

Hambly N. C., Davenhall A. C., Irwin M. J., MacGillivray H. T., 2001c, MNRAS, 326, 1315

Hawkins M. R. S., 1983, in Swings J.-P., ed., Liege International Astrophysical Colloquia Vol. 24, Liege International Astrophysical Colloquia. pp 31–36

Hawkins M. R. S., 2002, MNRAS, 329, 76

Heckman T. M., Best P. N., 2014, ARA&A, 52, 589

Helfand D. J., Stone R. P. S., Willman B., White R. L., Becker R. H., Price T., Gregg M. D., McMahon R. G., 2001, AJ, 121, 1872

Hook I. M., McMahon R. G., Boyle B. J., Irwin M. J., 1994, MNRAS, 268, 305

Horne K., et al., 2021, ApJ, 907, 76

Ishibashi W., Fabian A. C., 2012, MNRAS, 427, 2998

Ivezić Ž., et al., 2007, AJ, 134, 973

Jones R. H., 1981, in FINDLEY D. F., ed., , Applied Time Series Analysis II. Academic Press, pp 651–682, doi:https://doi.org/10.1016/B978-0-12-256420-8.50026-5

Jones Richard H., Ackerson L. M., 1990, Biometrika, 77, 721

Kasliwal V. P., Vogeley M. S., Richards G. T., 2015, MNRAS, 451, 4328

Kasliwal V. P., Vogeley M. S., Richards G. T., 2017, MNRAS, 470, 3027

Kawaguchi T., Mineshige S., Umemura M., Turner E. L., 1998, ApJ, 504, 671

Kelly B. C., Bechtold J., Siemiginowska A., 2009, ApJ, 698, 895

Kelly B. C., Sobolewska M., Siemiginowska A., 2011, ApJ, 730, 52

Kelly B. C., Treu T., Malkan M., Pancoast A., Woo J.-H., 2013, ApJ, 779, 187

Kelly B. C., Becker A. C., Sobolewska M., Siemiginowska A., Uttley P., 2014, ApJ, 788, 33

Koen C., 2005, MNRAS, 361, 887

Kozłowski S., 2016a, MNRAS, 459, 2787

Kozłowski S., 2016b, ApJ, 826, 118

Kozłowski S., 2017, A&A, 597, A128

Kozłowski S., et al., 2010, ApJ, 708, 927

Krolik J. H., 1999, Active galactic nuclei : from the central black hole to the galactic environment

Lawrence A., 2012, MNRAS, 423, 451

Lawrence A., 2016, Clues to the Structure of AGN Through Massive Variability Surveys. p. 107

Lawrence A., 2019, Probability in Physics - An Introductory Guide, 1 edn. Undergraduate Lecture Notes in Physics, Springer

Lawrence A., et al., 2016, MNRAS, 463, 296

Li Z., McGreer I. D., Wu X.-B., Fan X., Yang Q., 2018, ApJ, 861, 6

Lindsey W. C., Chie C. M., 1976, IEEE Proceedings, 64, 1652

Lu K.-X., et al., 2019, ApJ, 877, 23

Lyke B. W., et al., 2020, ApJS, 250, 8

Lynden-Bell D., 1969, Nature, 223, 690

MacLeod C. L., et al., 2010, ApJ, 721, 1014

MacLeod C. L., et al., 2011, ApJ, 728, 26

MacLeod C. L., et al., 2012, ApJ, 753, 106

MacLeod C. L., et al., 2016, MNRAS, 457, 389

Magnier E. A., et al., 2020, ApJS, 251, 5

Maiolino R., et al., 2010, A&A, 517, A47

Masci F. J., et al., 2019, PASP, 131, 018003

Matthews T. A., Sandage A. R., 1963, ApJ, 138, 30

McHardy I. M., et al., 2014, MNRAS, 444, 1469

Merloni A., 2016, in Haardt F., Gorini V., Moschella U., Treves A., Colpi M., eds, , Vol. 905, Lecture Notes in Physics, Berlin Springer Verlag. p. 101, doi:10.1007/978-3-319-19416-5_4

Merloni A., et al., 2014, MNRAS, 437, 3550

Morganson E., et al., 2014, ApJ, 784, 92

Mushotzky R. F., Edelson R., Baumgartner W., Gandhi P., 2011, ApJ, 743, L12

Netzer H., 2013, The physics and evolution of active galactic nuclei. Cambridge University Press, Cambridge

Netzer H., 2015, ARA&A, 53, 365

Oke J. B., 1963, Nature, 197, 1040

Oke J. B., Gunn J. E., 1983, ApJ, 266, 713

Pâris I., et al., 2018, A&A, 613, A51

Peacock J. A., Hambly N. C., Bilicki M., MacGillivray H. T., Miller L., Read M. A., Tritton S. B., 2016, MNRAS, 462, 2085

Peterson B. M., 1993, PASP, 105, 247

Peterson B. M., 1997, An Introduction to Active Galactic Nuclei. Cambridge University Press, Cambridge

Peterson B. M., 2006, in Gaskell C. M., McHardy I. M., Peterson B. M., Sergeev S. G., eds, Astronomical Society of the Pacific Conference Series Vol. 360, AGN Variability from X-Rays to Radio Waves. p. 191

Press W. H., Rybicki G. B., Hewitt J. N., 1992a, ApJ, 385, 404

Press W. H., Rybicki G. B., Hewitt J. N., 1992b, ApJ, 385, 416

Pringle J. E., Rees M. J., 1972, A&A, 21, 1

Raddick M. J., Thakar A. R., Szalay A. S., Santos R. D. C., 2014a, Computing in Science and Engineering, 16, 22

Raddick M. J., Thakar A. R., Szalay A. S., Santos R. D. C., 2014b, Computing in Science and Engineering, 16, 32

Rakshit S., Stalin C. S., 2017, ApJ, 842, 96

Reis R. C., Miller J. M., 2013, ApJL, 769, 5

Roux A., 2002, PhD thesis, University of Pretoria

Ruan J. J., et al., 2012, ApJ, 760, 51

Rumbaugh N., et al., 2018, ApJ, 854, 160

Rybicki G. B., Press W. H., 1995, Phys. Rev. Lett., 74, 1060

Salpeter E. E., 1964, ApJ, 140, 796

Sánchez-Sáez P., Lira P., Mejía-Restrepo J., Ho L. C., Arévalo P., Kim M., Cartier R., Coppi P., 2018, ApJ, 864, 87

Sanders D. B., Phinney E. S., Neugebauer G., Soifer B. T., Matthews K., 1989, ApJ, 347, 29

Sartori L. F., et al., 2016, MNRAS, 457, 3629

Scargle J. D., 1989, ApJ, 343, 874

Schmidt M., 1963, Nature, 197, 1040

Schmidt K. B., Marshall P. J., Rix H.-W., Jester S., Hennawi J. F., Dobler G., 2010, ApJ, 714, 1194

Schneider D. P., Fan X., Hall P. B., Jester S., Richards G. T., Stoughton C., Strauss M. A., 2003, AJ, 126, 2579

Sesar B., et al., 2006, AJ, 131, 2801

Seyfert C. K., 1943, ApJ, 97, 28

Shakura N. I., Sunyaev R. A., 1973, A&A, 24, 337

Shappee B. J., et al., 2014, ApJ, 788, 48

Shen Y., Burke C. J., 2021, ApJ, 918, L19

Sheng X., Ross N., Nicholl M., 2022, MNRAS, 512, 5580

Simm T., Salvato M., Saglia R., Ponti G., Lanzuisi G., Trakhtenbrot B., Nandra K., Bender R., 2016, A&A, 585, A129

Simonetti J. H., Cordes J. M., Heeschen D. S., 1985, ApJ, 296, 46

Skrutskie M. F., 1999, in American Astronomical Society Meeting Abstracts. p. 34.01

Smith J. A., et al., 2002, AJ, 123, 2121

Smith K. L., Mushotzky R. F., Boyd P. T., Malkan M., Howell S. B., Gelino D. M., 2018, ApJ, 857, 141

Stone Z., et al., 2022, MNRAS, 514, 164

Stoughton C., Lupton R. H., Bernardi M., Blanton M. R., Burles S., 2002, AJ, 123, 485

Suberlak K. L., Ivezić Ž., MacLeod C., 2021, ApJ, 907, 96

Sumi T., et al., 2005, MNRAS, 356, 331

Tadhunter C., 2016, Astronomische Nachrichten, 337, 159

Tanaka Y., et al., 1995, Nature, 375, 659

Tonry J. L., et al., 2012, ApJ, 750, 99

Vanden Berk D. E., et al., 2001, AJ, 122, 549

Vanden Berk D. E., et al., 2004, ApJ, 601, 692

Voevodkin A., 2011, arXiv e-prints, p. arXiv:1107.4244

Wilhite B. C., Brunner R. J., Grier C. J., Schneider D. P., vanden Berk D. E., 2008, MNRAS, 383, 1232

Wu Q., Shen Y., 2022, ApJS, 263, 42

York D. G., et al., 2000, AJ, 120, 1579

Yu W., Richards G. T., 2022, EzTao: Easier CARMA Modeling, Astrophysics Source Code Library, record ascl:2201.001 (ascl:2201.001)

Yu W., Richards G. T., Vogeley M. S., Moreno J., Graham M. J., 2022, ApJ, 936, 132

Zu Y., Kochanek C. S., Kozłowski S., Udalski A., 2013, ApJ, 765, 106

Zuo W., Wu X.-B., Liu Y.-Q., Jiao C.-L., 2012, ApJ, 758, 104

de Vries W. H., Becker R. H., White R. L., 2003, AJ, 126, 1217

de Vries W. H., Becker R. H., White R. L., Loomis C., 2005, AJ, 129, 615