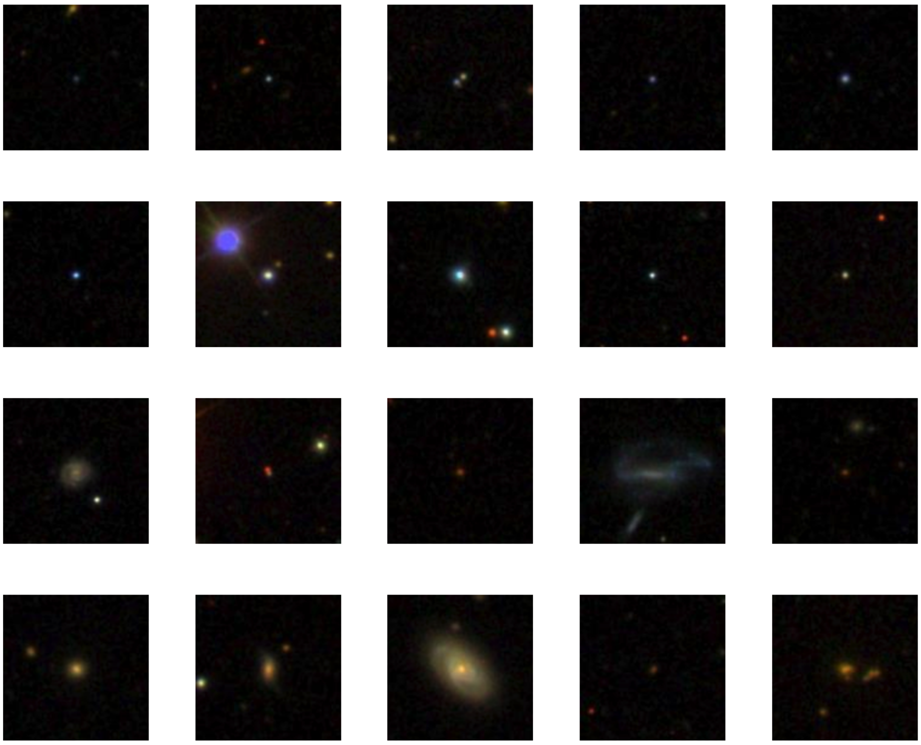
UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE



PhD thesis

Selection of PTF Sources Based on Light-curve Variability

Sofie Helene Bruun

Supervisors: Jens Hjorth and Adriano Agnello

Submitted: May 31, 2023

Dissertation submitted for the degree of Philosophiæ Doctor to the PhD School of The Faculty of Science, University of Copenhagen

Sofie Helene Bruun, DARK, Niels Bohr Institute, Faculty of SCIENCE, University of Copenhagen
Jagtvej 128, 2200 Copenhagen N, Denmark

Front-page illustration created using the Sloan Digital Sky Survey Image List,
https://skyserver.sdss.org/dr17/VisualTools/list

Publications related to this thesis

This thesis contains the following first author publications:

- VarIabiLity seLection of AstrophysIcal sources iN PTF (VILLAIN) I. Structure function fits to 71 million objects
  Sofie Helene Bruun, Adriano Agnello and Jens Hjorth
  In press at *Astronomy & Astrophysics*; accepted on 11/04/2023. arXiv: 2304.09903

- VarIabiLity seLection of AstrophysIcal sources iN PTF (VILLAIN) II. Supervised classification of variable sources
  Sofie Helene Bruun, Jens Hjorth and Adriano Agnello
  In review at *Astronomy & Astrophysics*; submitted on 11/04/2023. arXiv: 2304.09905

Abstract

The Universe is a vast and diverse place. Fortunately for our exploration of the world around us, we have access to a wide array of tools for astronomy and data analysis. For the efficient study of astrophysical objects, it is important to choose the right tools. In this context, the easiest and cheapest source of information is light. The light we gather tells us, for example, the brightness and colours of objects. Over time, we can analyse the variability of brightness. With this, we can select specific types of light sources to study them and their connection to the rest of the Universe. For large surveys, such as the Vera C. Rubin Observatory Legacy Survey of Space and Time, selection must be done automatically. This is possible through new methods in statistics and machine learning.

We aim to characterise and identify astrophysical sources using variability and colours. In Chapter 4, we study the properties of quasars, stars and galaxies in the Palomar Transient Factory (PTF). We do so by selecting objects for these three classes using simple criteria, and then we study the differences in variability and colours. In Chapter 5, we create a machine learning model for efficient classification, and we compare the roles of variability and colours.

To quantify variability, we predict how much an object will change in magnitude depending on the difference in time between observations. For this, we use a simple power law model to extract two variability parameters. With 71 million fits to PTF light curves, and matches to optical and infrared colours in Pan-STARRS1 and the Wide-field Infrared Survey Explorer for most of them, this provides a large data set for determining the common properties of different object types. We select objects for each class using colour and variability to study their photometric properties and identify inconsistencies with spectroscopic classifications ("labels") by the Sloan Digital Sky Survey. For

automatic classification, we use a histogram-based gradient boosting classification tree, which learns decision boundaries to separate the classes in a high-dimensional parameter space. We implement efficient model selection using random search with successive halving and combine input parameters for more efficient learning. For example, we subtract magnitudes in different bands to create colours.

We find the automatic classification model to perform well with a quasar completeness of 92.49 % and a purity of 95.64 %. It is fast to train, easy to implement, automatically handles missing values and does not need scaling of inputs or calibration of outputs. We create a catalogue of the 71 million objects including their predicted classes, the probabilities of belonging to each class, structure function parameters and magnitudes. Selecting subsets of the data reveals a similar performance down to 100 000 labeled samples, which we recommend for similar, future studies, although the algorithm is well suited to large data sets. With both manual and automatic selection techniques, we find that selection by 7 band colour information performs better than by monochromatic variability in both completeness and purity for PTF sources. Fitting structure functions is cheaper than taking spectroscopy, and in the future, structure function fitting might be prioritised for the most relevant sources depending on the resources available. Experiments with different feature engineering and data might reveal further performance improvements. For large, future datasets, it is key to optimise computational resources, and in that context we recommend using histogram-based gradient boosting for astrophysical object classification.

Dansk resumé

Universet er et stort og forskelligartet sted. For at udforske verden omkring os har vi heldigvis adgang til et bredt udvalg af astronomiske og dataanalytiske værktøjer. For at kunne studere astrofysiske objekter effektivt er det vigtigt at vælge de rigtige værktøjer. I denne sammenhæng er lys den nemmeste og billigste informationskilde at bruge. Lyset vi indsamler fortæller os for eksempel om lysstyrken og farverne af objekter. Over tid kan vi analysere variabiliteten af lysstyrke. Med dette kan vi udvælge specifikke typer af lyskilder for at studere dem og deres forbindelse til resten af Universet. For store undersøgelser af himlen såsom Vera C. Rubin Observatory Legacy Survey of Space and Time må udvælgelsen ske automatisk, hvilket er muligt via nye metoder inden for statistik og maskinlæring.

Vi ønsker at karakterisere og identificere astrofysiske kilder ved hjælp af variabilitet og farver. I Kapitel 4 undersøger vi egenskaber for kvasarer, stjerner og galakser i Palomar Transient Factory (PTF). Vi gør dette ved at udvælge objekter for disse tre klasser gennem simple kriterier, og derefter studerer vi forskellene i variabilitet og farver. I Kapitel 5 skaber vi en maskinlæringsmodel for at klassificere effektivt, og vi sammenligner betydningen af variabilitet og farver.

For at kvantificere variabilitet forudsiger vi hvor meget et objekts lysstyrke vil ændre sig afhængigt af tidsforskellen mellem observationer. Til dette bruger vi en simpel potenslovsmodel til at udvinde to variabilitetsparametre. Med 71 millioner fits af PTF-lyskurver – og matches til optiske og infrarøde farver i Pan-STARRS1 og Wide-field Infrared Survey Explorer for de fleste af dem – udgør dette et stort datasæt, der kan bruges til at bestemme de mest almindelige egenskaber for forskellige objekttyper. Vi udvælger objekter for hver klasse via farve og variabilitet for at studere deres fotometriske egenskaber og identificere uoverensstemmelser med spektroskopiske klassifikationer ("mærkater") fra Sloan Digital Sky Survey. Til automatisk klassifikation bruger vi et histogrambaseret gradientboostningsklassifikationstræ, som lærer beslutningsgrænser for at adskille klasserne i et højdimentionelt parameterrum. Vi implementerer effektiv modeludvælgelse gennem tilfældig søgen med successiv halvering og kombinerer inputparametre for mere effektiv læring. For eksempel trækker vi magnituder i forskellige bånd fra hinanden for at skabe farver.

Vi konstaterer, at den automatiske klassifikationsmodel præsterer godt med en kvasarsensitivitet på 92.49 % og kvasarrenhed på 95.64 %. Den er hurtig at træne, nem at implementere, håndterer automatisk manglende værdier og behøver ikke skalering af inputs eller kalibrering af outputs. Vi danner et katalog med de 71 millioner objekter inklusiv deres forudsagte klasser, sandsynlighederne for at tilhøre hver klasse, strukturfunktionsparametre og magnituder. Udvælgelse af delmængder af dataen viser en tilsvarende præstationsevne for ned til 100 000 objekter med kendte mærkater, hvilket vi anbefaler til lignende, fremtidige studier, selvom algoritmen er velegnet til store datasæt. Med både manuelle og automatiske udvælgelsesteknikker finder vi, at udvælgelse ved farveinformation i syv bånd præsterer bedre end ved monokromatisk variabilitet i både sensitivitet og renhed for PTF-kilder. At fitte strukturfunktioner er billigere end at tage spektroskopi, og i fremtiden bliver strukturfunktionsfitning måske prioriteret til de mest relevante kilder afhængigt af de tilgængelige ressourcer. Eksperimenter med anderledes inputkonstruktion og data vil måske give yderligere præstrationsforbedringer. For store, fremtidige datasæt er det vigtigt at optimere komputationelle ressourcer, og i den sammenhæng anbefaler vi at bruge histogrambaseret gradientboostning til klassifikation af astrofysiske objekter.

Acknowledgements

I would like to thank my supervisors for their honest and typically quick feedback and for the open door policy. I would also like to thank my colleagues at DARK in general for the interesting discussions, scientific and cultural, and for the inclusive work environment. I thank the administrative staff for making DARK a great place and for supplying me with coffee, snacks and ID cards. I am especially grateful for my fellow PhD students, master students and post docs doing their best to help each other through research and bureaucracy.

I am very grateful for my time at La Palma and the welcoming people working at NOTSA. It was a pleasure to experience the active student life on a small island, get to know the wonderful people, and to work on different projects with such positive and encouraging supervision.

I would like to thank Svend for his emotional and practical support right from the beginning of the project. I would also like to deeply thank Anna Liv for all her support, inspiring outlook and enthusiasm in physics. I greatly appreciate being able to rely on you, and you have both been central to the PhD process and its completion.

I am also thankful for the amazing student environment at the Niels Bohr Institute. I had the pleasure of joining the student revue, FysikRevy$^{\text{TM}}$, and through singing, acting, dancing and backstage work, it has built a fantastic community, that I hope to still be part of for the anniversary revue later this year.

I thank my old study group for making my physics studies much more enjoyable and for their continued support during the PhD. I look forward to many more nights of board games, DnD and drinks with you.

I would like to thank Lotte and Lise for being wonderful and considerate sisters with a great sense of humour. I am lucky have had you in my life for exactly as long as

I remember, and I look forward to being able to spend more time with you after this project.

It is important to remember why we do science and the sense of wonder in trying to understand the Universe. I would like to thank the Danish Youth Association of Science for the great community; a constant in my life all the way from high school and through most of the PhD. Doing physics as voluntary work is refreshing, and it is great to have the opportunity to inspire others. You have broadened my knowledge of physics (and other sciences), solidified it in the fields I have taught, and kept my knowledge of the basics fresh.

Contents

Publications 3

Abstract 5

Dansk resumé 7

Acknowledgements 9

Introduction 19

Motivation . . . 19

Thesis outline . . . 22

I Background 25

1 Variability in astrophysics 27

1.1 Stars . . . 28

1.1.1 Variable stars . . . 29

1.2 Galaxies . . . 35

1.3 Quasars and AGN . . . 35

1.3.1 AGN structure . . . 39

1.3.2 AGN types . . . 43

1.3.3 AGN variability . . . 44

1.3.4 Lensed quasars . . . 45

1.4 Coordinates and time . . . 46

2 Statistical methods 49

2.1 Bayes and likelihoods . . . 50

2.2 Markov Chain Monte Carlo . . . 52

2.2.1 Sampling algorithm . . . 55

2.3 k-d trees . . . 55

2.4 Measuring variability . . . 57

2.4.1 Structure functions . . . 59

2.4.2 PSD . . . 60

2.4.3 CARMA . . . 61

2.5 Machine learning . . . 64

2.5.1 Model selection . . . 65

2.5.2 Classification of variable sources . . . 66

II Selection based on variability 69

3 Data processing 71

3.1 High performance computing . . . 71

3.2 Querying PTF . . . 72

3.3 Assembling and matching light curves . . . 73

3.4 Fitting . . . 74

3.5 Combining and cross matching with SDSS . . . 76

4 Structure function fits to 71 million objects 79

Author contributions . . . 79

Abstract . . . 80

4.1 Introduction . . . 80

4.2 Data . . . 83

4.2.1 PTF . . . 83

4.2.2 SDSS . . . 84

4.2.3 WISE . . . 84

4.2.4 PS1 . . . 84

4.2.5 Data cleaning . . . 85

4.3 Methodology . . . 86

4.3.1 Structure functions . . . 86

4.3.2 Fitting . . . 87
4.3.3 Evaluation metrics . . . 88
4.4 Variability and colour distributions . . . 88
4.4.1 Full Palomar Transient Factory sample . . . 89
4.4.2 Spectroscopically confirmed sample . . . 89
4.5 Photometric selection . . . 90
4.5.1 Selection properties . . . 90
4.5.2 Ambiguous sources . . . 94
4.6 Discussion . . . 95
4.6.1 A-γ clusters . . . 97
4.6.2 Comparison with SDSS light-curve analysis . . . 97
4.6.3 Spectroscopic selection bias . . . 98
4.6.4 Photometric selection bias . . . 99
4.6.5 Perspectives . . . 100
4.7 Conclusion . . . 100
4.8 Appendices . . . 104
4.8.1 Appendix A: Sky maps . . . 104
4.8.2 Appendix B: SDSS matched data . . . 106
4.8.3 Appendix C: Variability selected objects . . . 109
4.8.4 Appendix D: SIMBAD statistics . . . 113
4.8.5 Appendix E: *Gaia* proper motions . . . 115

5 Supervised classification of variable sources 117

Author contributions . . . 117
Abstract . . . 118
5.1 Introduction . . . 118
5.2 Method . . . 120
5.2.1 Models . . . 121
5.2.2 Data splitting . . . 121
5.2.3 Classification algorithm . . . 123
5.2.4 Preprocessing . . . 126
5.2.5 Performance evaluation . . . 126
5.3 Results . . . 128
5.3.1 Performance . . . 128

5.3.2 Calibration 132
5.3.3 Predicted parameter distribution 133
5.3.4 Output table 135
5.3.5 Feature importance 136
5.4 Discussion and conclusions 136
5.4.1 Model comparison 136
5.4.2 SDSS comparison 139
5.4.3 Biases 139
5.4.4 Perspectives 140

III Perspectives and conclusions 145

6 Perspectives 147
6.1 Expanded analysis 147
6.1.1 Predicted classes 147
6.1.2 Variable galaxies 150
6.1.3 Split objects 150
6.2 Alternative machine learning models 161
6.2.1 Semi-supervised learning 161
6.2.2 Features 162
6.3 Different data 165

7 Conclusions 167
7.1 Part I 167
7.2 Part II 168
7.3 Part III 169

Acronyms 173

Bibliography 175

Appendix A: Illustration 189

List of Figures

1.1 Star imaging . . . 30

1.2 Star spectra . . . 31

1.3 Star light curves . . . 32

1.4 Variable star light curves . . . 33

1.5 Galaxy imaging . . . 36

1.6 Galaxy spectra . . . 37

1.7 Galaxy light curves . . . 38

1.8 QSO imaging . . . 39

1.9 QSO spectra . . . 40

1.10 QSO light curves . . . 41

1.11 AGN diagram . . . 42

2.1 MCMC chains . . . 53

2.2 MCMC histograms . . . 54

2.3 Tree structure . . . 56

2.4 Sliding midpoint method . . . 57

3.1 Tasks per patch . . . 75

3.2 Fitting time per patch . . . 76

4.1 Heat maps of fitted sources . . . 91

4.2 Heat maps of spectroscopically confirmed sources . . . 92

4.3 Variability of colour-selected objects . . . 95

4.4 Colours of variability-selected objects . . . 97

4.5 Sky maps of fitted data 105
4.6 Time spans and epochs 106
4.7 Spectroscopically confirmed sources and selection criteria 107
4.8 Misclassified QSO spectrum 108
4.9 Redshifts of galaxies and QSOs 110
4.10 Colours of variable galaxies 111
4.11 Colours of variable QSOs 111
4.12 Colours of variable stars 112
4.13 SIMBAD registrations of selected sources 114
4.14 Gaia proper motions 116

5.1 ROC curves 130
5.2 Performance and data-set size 131
5.3 Calibration plot 133
5.4 Heat maps of predicted classes 134
5.5 Predicted classes using only variability 135

6.1 VarA prediction regions 149
6.2 Colours distributions from VarA 150
6.3 Variability distributions of classified sources 151
6.4 $g - r$ versus $z - W1$ of classified sources 152
6.5 $W2$ versus $W1 - W2$ of classified sources 153
6.6 Light curves of predicted QSOs 154
6.7 Light curves of predicted stars 155
6.8 Light curves of predicted galaxies 156
6.9 Star with close neighbour 157
6.10 SIMBAD SDSS galaxies 158
6.11 PTF neighbour distances 159
6.12 Number of PTF neighbours 159
6.13 Images of objects with neighbours 160

List of Tables

4.1 Sample sizes. 86
4.2 Selection criteria. 93
4.3 Selection statistics. 96

5.1 Classifier hyperparameters. 128
5.2 Performance statistics on the test set. 129
5.3 Catalogue parameters. 137
5.4 Feature importance. 138

Introduction

Motivation

When we humans look to the sky, we may wonder about the nature of the light sources. Some are bright; some are fainter. Some appear white, and some red or blue. Observing by eye in the city, a bright red source is likely Mars, and a periodically variable source is likely an aeroplane. Looking a bit deeper in a remote place, we might notice galaxies as extended sources. With deep sky surveys, we can use variability and colour to identify stars, galaxies and the bright, accreting centres of galaxies known as quasars. Humankind's main source of information for understanding the Universe beyond the Earth is light, and so, understanding and identifying sources of light will help us explore it.

Astrophysical sources vary in multiple ways. Brightness can vary periodically or non-periodically on different timescales. Predicting the change in brightness after waiting a specific time interval provides one way of measuring variability. With the development of new algorithms in machine learning, astronomers have access to an increasing palette of tools for understanding the connection between astrophysical sources and their observational characteristics. Optimal identification of sources requires careful data processing and algorithm selection, but even with relatively simple rules and structures, we can create complex, powerful models. Efficient machine learning gives us the opportunity to automatise manually-demanding tasks and gain new insights.

The astrophysical community is collecting and managing data sets of increasing scale. Brightness is often included in the form of magnitudes. This is a measure with roots in observations by eye in ancient Greece. Since the human eye perceives brightness logarithmically, magnitudes scale logarithmically with brightness, and for historical reasons, a lower value indicates a brighter object. The Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST) will obtain 10 year light curves of 800 epochs in the $ugrizy$ bands down to r band magnitudes of ~ 24.5 for single visits. The objects include about 20 billion galaxies and 20 billion stars over 18 000 deg^2. They will also observe "deep drilling fields" of known areas down to $ugri \sim 28.5$, creating high-cadence light curves useful for studying variability of active galactic nuclei (AGN) (De Cicco et al. 2021). Combined with *Euclid* and the *Nancy Grace Roman Space Telescope*, LSST will discover tens of quasars at $z > 7.5$ (Ivezić et al. 2019), $\sim$ 2 million quasars at $z < 2$, at least 6.2 million AGN (De Cicco et al. 2021) and at least 50 million variable stars (Sesar et al. 2007). Combined with data from the Sloan Digital Sky Survey (SDSS), Pan-STARRS1 (PS1) and the Zwicky Transient Facility (ZTF), some light curves will span 35 years and allow us to better constrain the behaviour of variable objects and to classify them. Oguri & Marshall (2010) estimate that the LSST will find ∼8000 lensed quasars and $\sim$ 130 lensed supernovae. They also create a mock catalogue of lensed quasars. Taak & Treu (2023) use this mock catalogue to estimate that ∼1000 of the lensed quasars will have variability that is observable by the LSST.

Classification is important to understand the observed objects and use them to extract information about the Universe. Quasars are extremely bright objects at the centres of some galaxies, and they can probe the distant, old Universe. We can use them as tracers of structure formation (Turner 1991; Song et al. 2016) and constrain baryon acoustic oscillations (BAO) signals, which depend on cosmological parameters (Alam et al. 2021; Blomqvist et al. 2019). Secrest et al. (2021) test the cosmological principle by measuring a cosmic dipole in the distribution of quasars. We can even directly measure cosmic acceleration with redshift-drift tests, such as by measuring the change in redshift of the Lyα forest in quasar spectra (Sandage 1962; Kim et al. 2015; Alves et al. 2019; Loeb 1998). With time-delay cosmography of lensed quasars, we can estimate cosmological parameters using the time difference between multiple images of the same quasar. Greater samples of quasars, especially at high redshifts, will benefit these efforts.

Other variable sources, such as Cepheid variables and RR Lyrae, can be used as standard candles for distance estimation. Greater and fainter sets of non-variable stars will also be useful for calibration of future surveys at the Vera C. Rubin Observatory, Extremely Large Telescope (ELT), Thirty Meter Telescope (TMT) and Giant Magellan Telescope (GMT).

The variability of quasars and AGN is useful for selecting them, although the origin of variability is still debated. Variability modelling of large quasar samples will enable better constraints of the physics of quasars and their characteristic variability behaviour. Variability has been quantified in a multitude of ways; from "blinking" by van den Bergh et al. (1973) to ensemble structure functions (SFs) (Simonetti et al. 1985) and high order continuous autoregressive–moving-average (CARMA) models (Moreno et al. 2019). SFs describe how an object changes in magnitude over different time scales. The best model of this behaviour and a possible characteristic dampening timescale is unclear. Many have used damped random walk (DRW) models, but these are currently unconstrained even on 20 year timescales. Single power laws provide a simple SF model; at least on short time scales. This is sufficient for selection of variable sources.

Classification needs to be accurate and performed with increasing speed in the face of larger surveys. Manual selections are reliable but difficult to scale. Automatic classification can pick up subtle characteristics that might go unnoticed by humans, and it can quantify its confidence in class predictions. Data mining with machine learning enables new discoveries in existing databases, which is important in case of expensive astronomical data. The objects of astrophysical databases are described by a varying number of parameters. For example, some have been observed in the infrared – and some have not. We need classification models to extract as much information from the available data as possible and handle when some values are missing.

SF variability parameters can be extracted automatically using Markov Chain Monte Carlo (MCMC). But how useful is this information compared to colours? The answer might depend on the survey and the variability model.

Several approaches have been applied for automatic classification of astrophysical sources. For example, Palanque-Delabrouille et al. (2011) use neural networks, De Cicco et al. (2021) use random forests, Cunha & Humphrey (2022) use gradient boosted trees and Logan & Fotopoulou (2020) use Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). In the literature, colour and variability are both informative and include different biases as selection tools – so the best performance is achieved using both. Some find that variability performs best and some achieve better performance with colours. It is common to create models that output the probability of each object being a quasar, allowing the identification of ambiguous objects and confident quasar candidates. However, most research is based on data sets that are not fully representative of a full survey. Most also do not handle missing values, the data sets can be small and there might be problems of biased evaluation of performance.

Both Schmidt et al. (2010) and Butler & Bloom (2011) create linear criteria or "cuts" in variability parameters for the selection of quasars. The quasars are distinguished from stars, including variable stars, but the data sets are not representative of all sources in a survey. Linear criteria are simple and easy to apply, but machine learning can potentially create more accurate selection regions using nonlinear criteria.

We aim to demonstrate an effective solution for automatic classification of quasars, stars and galaxies. We will classify a large survey as accurately as possible while being mindful of biases. We choose the Palomar Transient Factory (PTF) survey with 600 million light curves in the R band (71 million after data cleaning) taken over almost six years, and match to optical and near-infrared colours in the Wide-field Infrared Survey Explorer (WISE) and PS1. We will explore the distributions of colours and Bayesian power law SF variability in the objects and in subsets – and then use the distributions for selection of quasars, stars and galaxies. The properties of the photometric selections are analysed and compared to spectroscopic classifications by the SDSS. Then, we aim to construct and optimise an efficient machine learning model for prediction of the spectroscopic classifications based on the photometric data. We will choose a model that handles missing values and include colours created by all combinations of the matched filters. We separately assess the importance of variability and colours for classification of PTF objects as quasars, stars and galaxies. We aim to analyse the performance of the model for different data set sizes and consider possible modifications for other surveys. We will create a catalogue for the resulting classifications and fitted variability parameters for use in future projects. This thesis builds on the preliminary studies of Bruun (2020).

Thesis outline

In this introductory chapter, we have discussed the motivation behind the thesis. We will now outline the structure of the rest of the thesis.

In Part I, we describe the theoretical background. Chapter 1 presents current knowledge of variable objects. We discuss properties of stars, galaxies and quasars, and the possible mechanisms behind variability. This includes different types of variable stars and AGN. Finally, we introduce systems for expressing positions and time stamps of astronomical observations. In Chapter 2, we explain the statistical background of the thesis. This includes how to sample posterior probability distributions with Markov Chain Monte Carlo, and efficiently cross-match large databases with k-d trees. We present common methods for quantifying light-curve variability in the literature and some examples of machine learning techniques for classification of variable objects.

Part II focuses on the details of the data analysis and presents the results. This part is based on Bruun et al. (2023a) and Bruun et al. (2023b). In Chapter 3, we present an outline of the data processing pipeline for variability selection and classification of the PTF survey. In Chapter 4, we fit power laws to the SFs of 71 million objects in the PTF survey to describe their variability. We then plot the variability along with colours from WISE and PS1. We discuss differences between quasars, stars and galaxies with SDSS spectroscopic classifications. To study distributions of photometrically selected sub-populations, we define selection criteria in variability and colours. In Chapter 5, we build a machine learning model for optimal classification of the PTF objects. We assign each object a class (quasar, star or galaxy) and a probability of belonging to each of the three classes. We discuss important aspects of the model and how to apply similar models to different data sets.

Part III concludes the thesis. Chapter 6 expands on part of the data analysis, and we discuss alternative approaches that might be relevant in future projects. Alternative approaches involve different machine learning methods, data and data processing. In Chapter 7, we summarise the thesis and connect the findings in preparation of future research.

I

Background

Chapter 1

Variability in astrophysics

In this chapter, we discuss the role of light-curve variability on human timescales in the field of astrophysics. Humans have mapped the sky with the tools available since ancient times. An example is Hipparchos estimating the positions and brightness of stars at ∼140–120 BC. He also might have used the geared Antikythera mechanism for predicting the positions of objects in the Solar System (Freeth et al. 2008). Analogously, this thesis will map the variability of astrophysical objects and create a model for predicting changes in brightness.

In recent times, we are creating catalogues of astrophysical sources using, for example, imaging, photometry, spectroscopy, polarimetry, neutrinos and even gravitational waves. With photometry, we study of the brightness of objects such as by measuring magnitudes. The system of magnitudes can be traced all the way back to Hipparchos. Measuring multiple magnitudes in the same band, but at different times, will give us a light curve. Light curves show us that some sources vary measurably in brightness over time.

Colours can be estimated from photometry, but for detailed colour information, we use spectroscopy. By taking spectra with high wavelength resolution, we also get the precise properties of emission and absorption lines in a spectrum. We can use these, along with the overall shape of the spectrum, to classify and describe properties of the light source. This includes redshift, chemical composition, temperature, chemical abundances, pressure, magnetic fields and motion (Ashworth 2012).

Light is redshifted when the emitting source is moving relative to the observer (Doppler shift), when the Universe is expanding (cosmological redshift) and when it moves through a gradient in the curvature of spacetime (gravitational redshift). On distance scales of more than ~ 5 Mpc, cosmological redshift dominates (Davis & Peebles 1983).

While spectroscopy is useful for studying astronomical objects, the spectral resolution requires long exposure times, and so it is expensive. When spectroscopic data is not available, redshifts can be estimated photometrically. Photometric redshifts are less precise and rely on broader spectral features that can be detected though a few bands (Baum 1962; Koo 1985; Tanaka 2015). Classification of sources is also easier by spectroscopy, but photometry is cheaper and available for more objects. However, light-curve variability can sometimes identify spectroscopic misclassifications, as we will see in Chapter 4.

The origin of light-curve variability can be *extrinsic* or *intrinsic* to an observed astrophysical object. Extrinsic variability could be from measurement errors, rotation or other objects, dust or gas within the line of sight from Earth. Intrinsically variable objects can be transients such as supernovae and kilonovae. We will focus on the long-term intrinsic variability (or the lack thereof) of objects. The three main object types we will select in Part II are stars, galaxies and quasars. Below, we review how and why their light curves may show variability.

1.1 Stars

Stars are luminous objects of plasma and are either generating energy by nuclear fusion of hydrogen or have done so previously. Stars can spend billions of years in the *main sequence* in which they fuse hydrogen into helium in their cores. Main sequence stars generally show very little variability (Sesar et al. 2007). They stay the same size by being in *hydrostatic equilibrium* with the inward force of gravity being balanced by pressure. This creates a stable star that is neither exploding or imploding (this also applies to stars outside the main sequence). The maximum luminosity an object can have and still be in hydrostatic equilibrium is called the *Eddington limit*.

1.1.1 Variable stars

Jetsu et al. (2013) argues that variable stars have been recorded since 1271–1163 B.C. in the Cairo Calender which includes the period of Algol. In 1596, stellar variability was first recorded in modern, Western history, for the star now known as Mira. Some stars have little variability – and without detectable variability they are useful as photometric standards. Some stars are pulsating with regular patterns in variability we can use to infer physical properties. If variability periods are correlated with absolute magnitudes, the stars can work as *standard candles* for distance estimation. If we know the apparent magnitude m of an object, and the absolute magnitude M we would measure at a distance of 10 pc, we can relate the two. The difference is called the distance modulus:

$$m - M = 5 \cdot \log_{10}\left(\frac{d_L}{[\mathrm{pc}]}\right) - 5. \tag{1.1}$$

With the distance modulus, we can use an object with known absolute magnitude to get its luminosity distance d_L.

Fig. 1.1 shows imaging of spectroscopically confirmed stars from SDSS. All of them are in the main sequence. They differ in colour, which also shows in the spectra in Fig. 1.2 for six of the same objects (sorted left to right and then top to bottom). In Fig. 1.3, we show light curves for the first nine of them. Most of them show few signs of variability. The light curves are created from R band PTF data and cleaned according to Sect. 4.2.5 including the selection of outliers using a weighted moving median.

In Fig. 1.4, we show light curves for variable stars, although they are not as common. They have been selected as variable using Set of Identifications, Measurements and Bibliography for Astronomical Data (SIMBAD) main type registrations. Comparing with Fig. 1.3, there is visibly more variability and more data points are selected as outliers.

A myriad of variable stellar types and subtypes exits – we will now explore some of the mechanisms behind stellar variability. See Eyer & Mowlavi (2008) for a deeper overview of variable stars including typical periods and amplitudes of different types.

Cepheids is a group of stars used as standard candles, for testing stellar evolution models and for modelling the Milky Way disk (Riess et al. 2019; Dinnbier et al. 2022; Skowron et al. 2019). They are pulsating variables, due to a connection between opacity and temperature known as the *κ-mechanism* (Zhevakin 1959; Montalban & Miglio 2008). When the stars are warm, they ionise He I, creating He II which is more opaque.

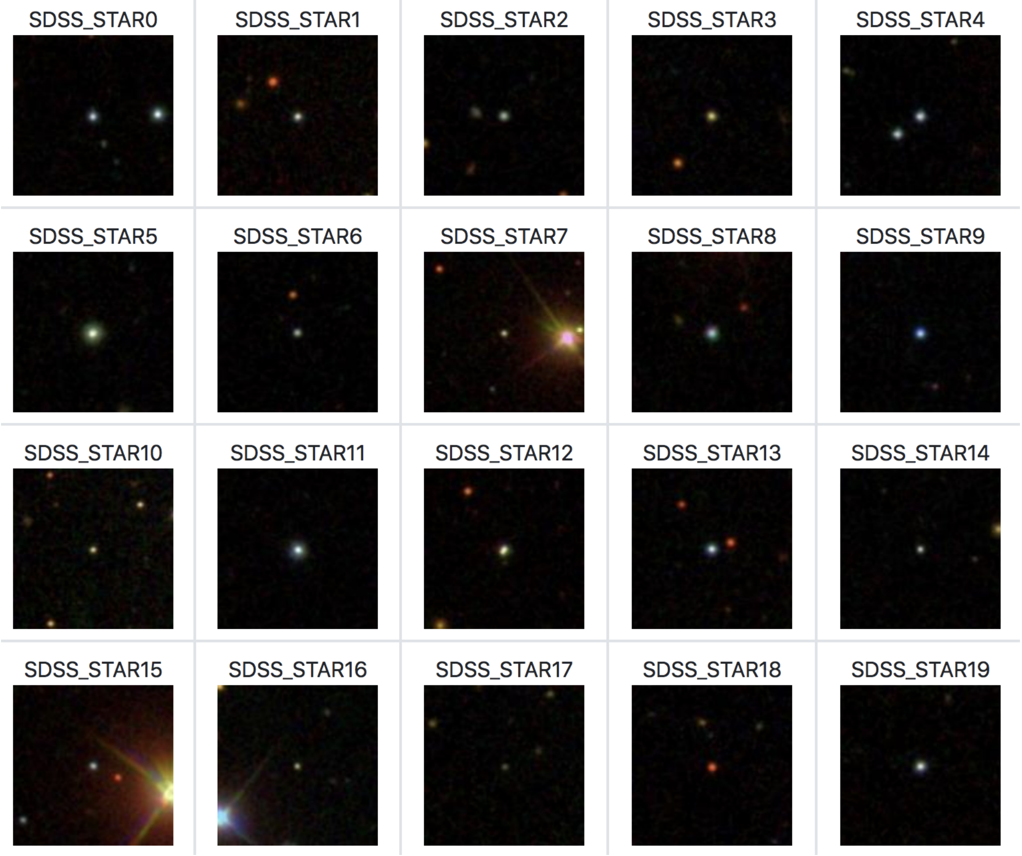


Figure 1.1: Imaging of spectroscopically confirmed stars in SDSS. They appear as point sources and are most common near the Milky Way disk. Images are selected using SDSS Image List, `https://skyserver.sdss.org/dr17/VisualTools/list`.

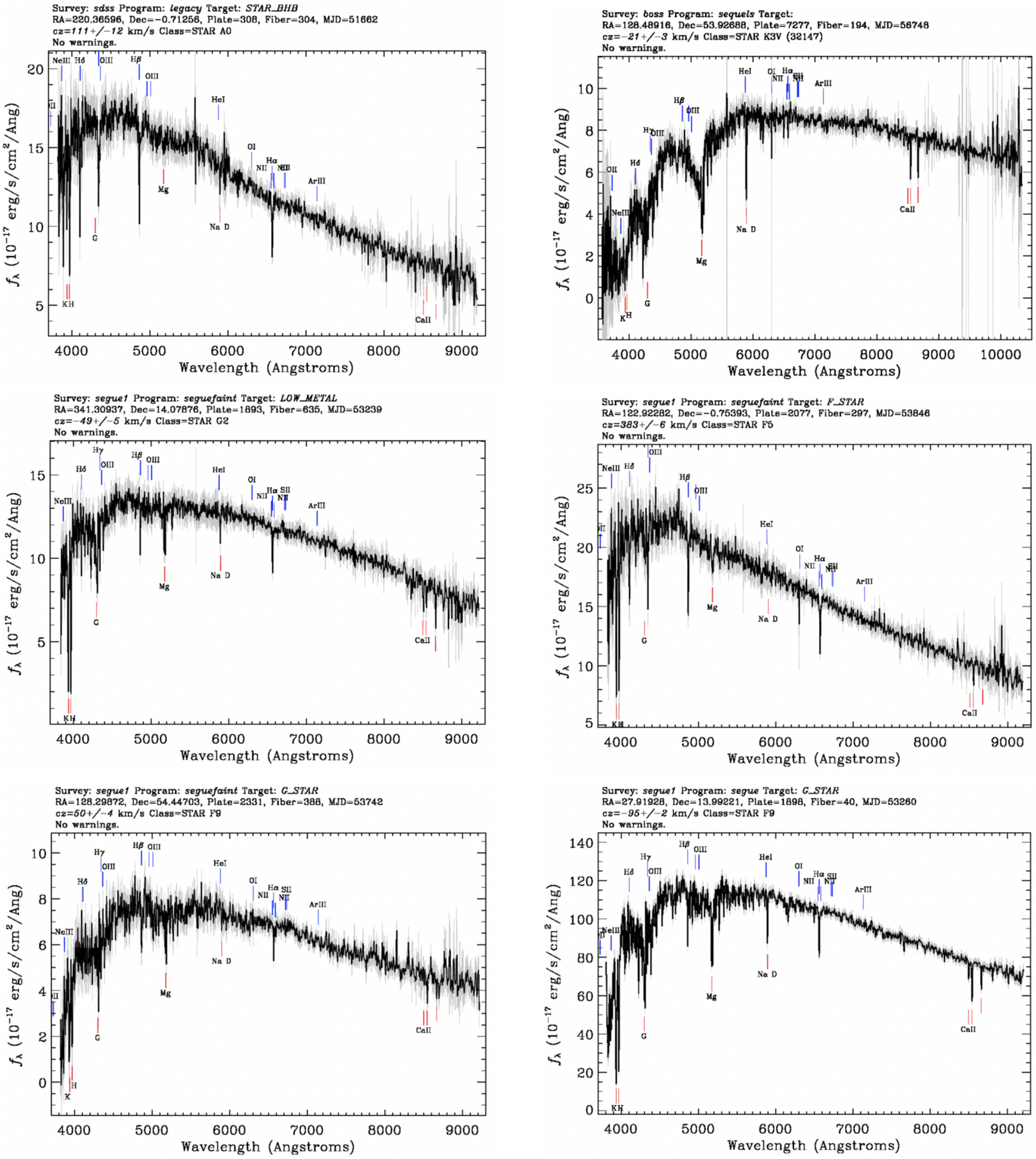


Figure 1.2: Example spectra of spectroscopically confirmed stars in SDSS. They are all in the main sequence. In SDSS they are registered with SpecObjID 3468607620749987840, 2131503125465425920, 2624579398258419712, 8193227061621708800, 2338575809775691776 and 2136969072997328896 (SDSS-IV DR17, CC-BY license, skyserver.sdss.org/dr17/VisualTools/explore/summary).

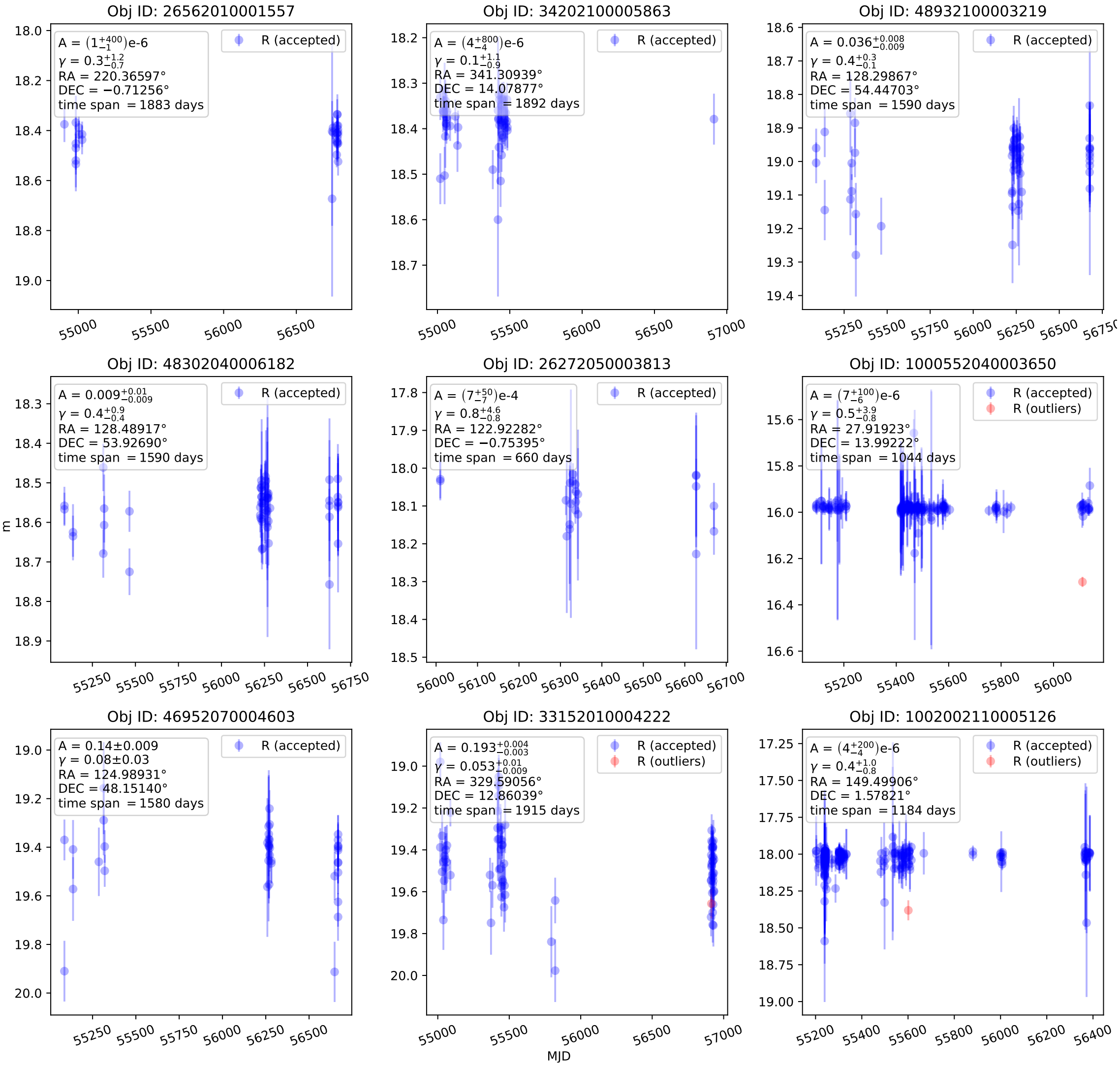


Figure 1.3: R-band PTF light curves of spectroscopically confirmed stars in SDSS. The light curves are constructed according to Sects. 3.2 and 3.3 and the outliers marked in red are selected according to Sect. 4.2.5. Some are more variable than others. For each source, we show the A and γ variability parameters fitted in Chapter 4. The variability amplitude A is extremely low and consistent with zero for most of the stars shown here.

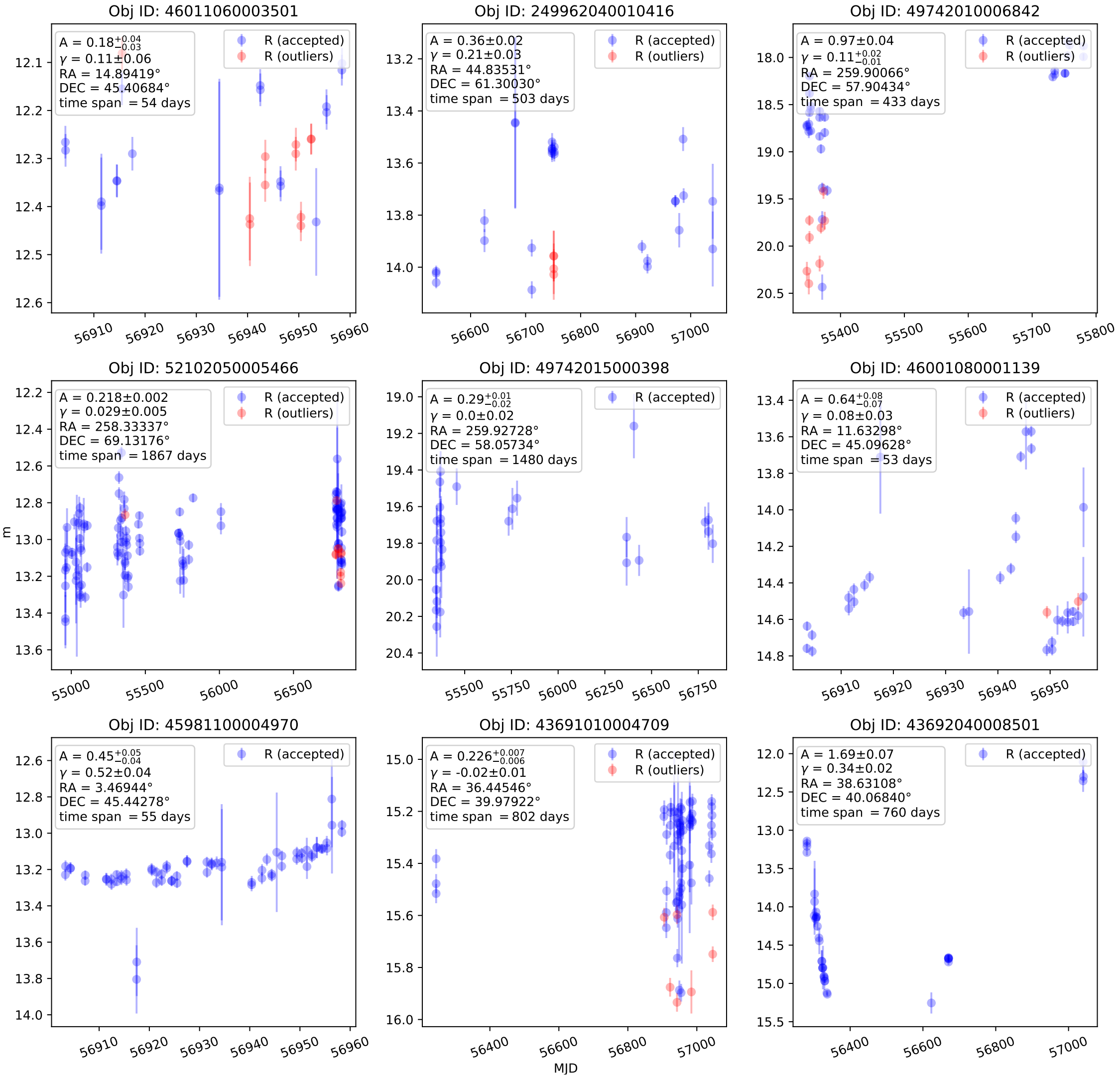


Figure 1.4: R-band PTF light curves of objects registered as variable stars in SIMBAD. The top row shows Cepheid variables, the centre row has RR Lyrae and the bottom row shows one long-period variable (left), one eclipsing binary (centre) and one Mira variable (right).

This means the helium absorbs more of the radiation, increasing radiation pressure and therefore the volume of the star. An increased volume decreases the temperature, allowing the helium to absorb electrons and become He I again. The He I is less opaque, lowering the radiation pressure, and so gravity compresses the star into a lower volume again (Cox 1963). Cepheids typically vary over days to months (Kraft 1960).

RR Lyrae were once thought to be Cepheids and can also be used as standard candles and for tracing stellar evolution. They too are pulsating due to the κ-mechanism of helium ions, but they are fainter than Cepheids and generally old and metal-poor. They can therefore be used for tracing old objects. However, metal-rich RR Lyrae can be produced through binary interactions (Bobrick et al. 2022). The variability amplitudes are similar to Cepheids but their periods are just a few hours to a day (Cabral et al. 2020).

An example of a variable star with longer-period pulsations is *Mira* variables. They are red giants that are brightest in the infrared but with variations of >2.5 magnitudes in the optical over >150 days and varying regularity. Their mass ejections play an important role in the metal enrichment of the interstellar medium (Mattei 1997; Iwanek et al. 2021).

Another type of variable source is *cataclysmic variables*. They have short, intense bursts in luminosity. Cataclysmic variables are close, binary systems with a late-type donor star that transfers mass to a white dwarf, usually through an *accretion disk* (Szkody 2021). When the white dwarf is warm enough, the hydrogen in the outer layer is ignited and fused into helium. After a burst of fusion and mass ejection, the white dwarf becomes stable again, and produces energy by accretion. As the transfered mass loses potential energy, part of that energy is converted to X-rays. X-rays are also emitted as Brehmsstrahlung when the accreted material reaches the surface of the white dwarf. Cataclysmic variables can be a source of lower energy photons depending mainly on magnetisation but also the boundary between the disk and the white dwarf (Mukai 2017).

Finally, stars can appear variable if they are eclipsing in a binary system or have large exoplanets. If they are rotating with spots of different brightness, that can also cause extrinsic variability.

1.2 Galaxies

Galaxies are large structures of stars, gas, dust and dark matter. They mostly appear as extended sources on the sky at low redshifts and depending on the telescope – example galaxies are shown with SDSS imaging in Fig. 1.5.

In Fig. 1.6 are examples of galaxy spectra. The spectra of galaxies depend on the spectra of their constituent stars and possibly AGN. We will get back to AGN in Sect. 1.3. Star forming galaxies have more young stars and therefore more blue radiation and stronger emission lines from ionised gas. As with stars, we can use the spectral lines of galaxies to estimate their redshifts. The lines of stars are broadened and shortened due to the spread in Doppler shift as the stars orbit the galaxy centre (Sparke & Gallagher 2007).

The light curves of nine galaxies are shown in Fig. 1.7. Most of them show little variability, although, for example, a starburst galaxy will slowly dim over timescales of millions and billions of years as it runs out of gas and the largest stars die (Sparke & Gallagher 2007). Salvato et al. (2009) explore the variability of galaxies using light curves from PTF. Galaxies in the Virgo cluster have higher variability close to the centre of the cluster, possibly due to higher gas contents fuelling AGN. Galaxies without AGN are not expected to be variable on human timescales.

1.3 Quasars and AGN

We know that many large galaxies contain a supermassive black hole. Most galaxies with a bulge have a supermassive black hole in the centre, but supermassive black holes can also exist without a bulge (Kormendy & Ho 2013). Some of the central black holes are accreting gas, which loses potential energy that is in part converted to radiation. Up to 10 % of the rest mass is converted, and so the objects can be extremely bright. These luminous objects are known as active galactic nuclei (AGN) and the brightest are called quasars (Sparke & Gallagher 2007).

As mentioned in the motivation, we can use quasars to trace structure formation and estimate cosmological parameters via studies of BAO signals and redshift-drift tests. Estimating the Hubble constant using quasar magnitudes at high redshifts has been attempted (Risaliti & Lusso 2019), but the accuracy is questionable (Velten & Gomes 2020).

Fig. 1.8 shows SDSS imaging of quasars. Examples of their spectra are given in Fig. 1.9

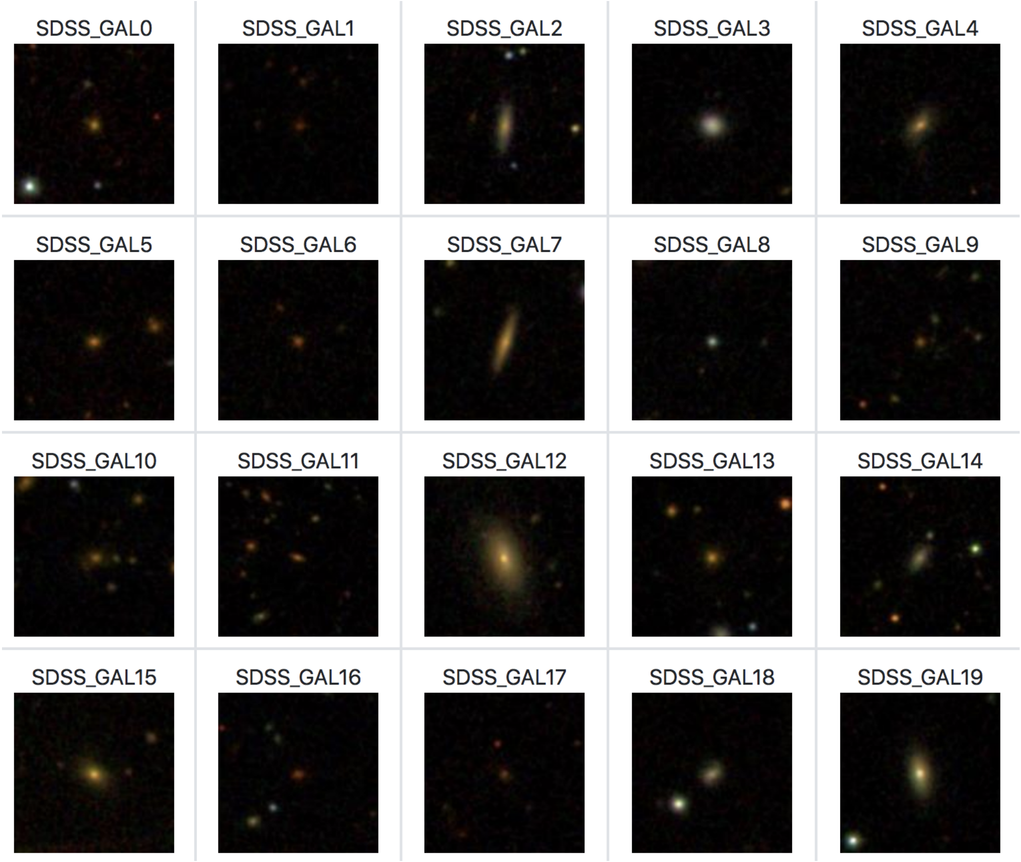


Figure 1.5: Imaging of spectroscopically confirmed galaxies in SDSS. Many of them can be recognised by appearing extended. Images are selected using SDSS Image List, https://skyserver.sdss.org/dr17/VisualTools/list.

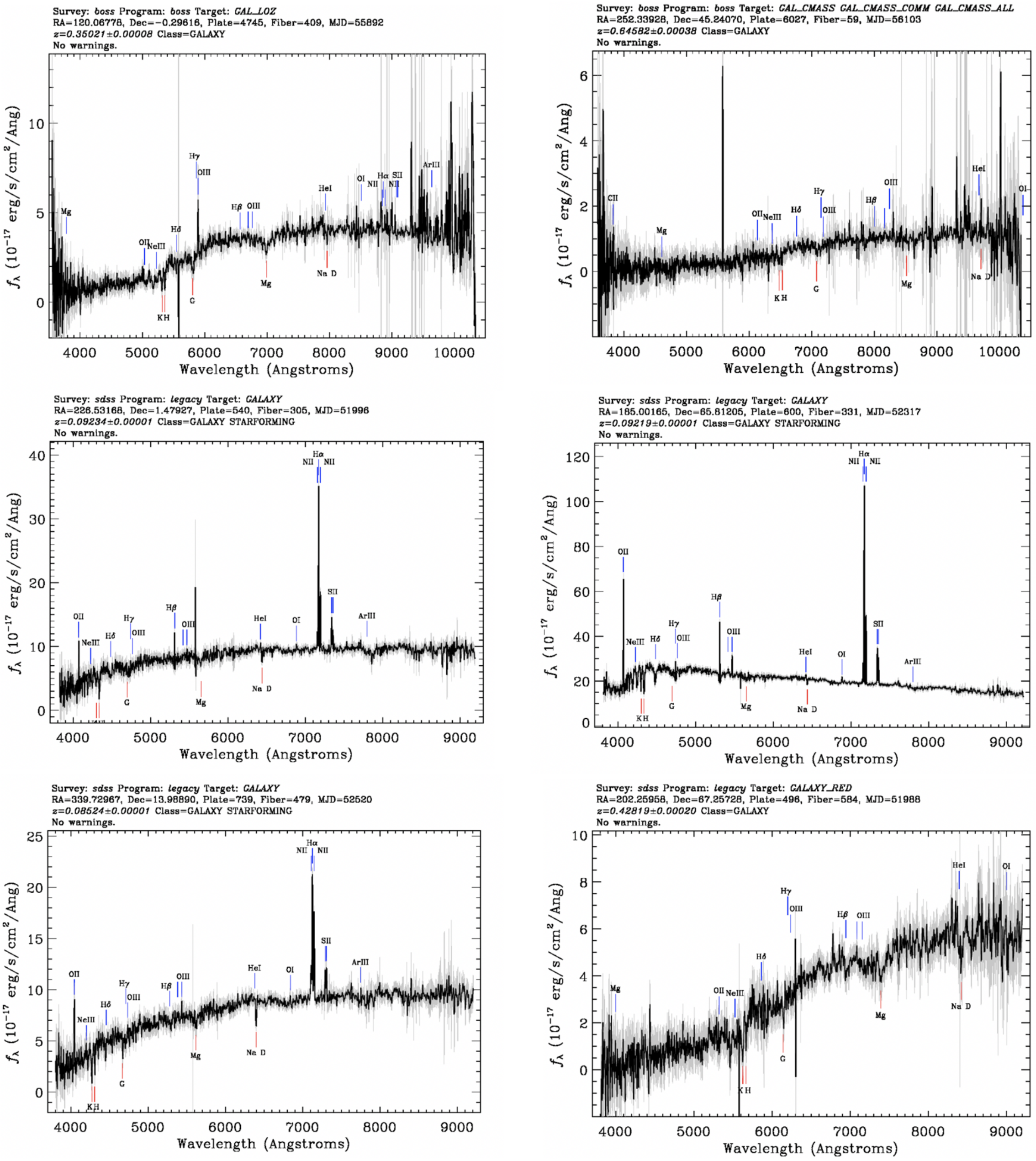


Figure 1.6: Spectra of spectroscopically confirmed galaxies in SDSS. The galaxies of the centre row and the lower right diagram are labeled as star forming. They are registered with SDSS SpecObjID 5342507581884880896, 6785815058729687040, 608069820943984640, 675630967565608960, 832171739952736256 and 558606915844728832 (SDSS-IV DR17, CC-BY license, skyserver.sdss.org/dr17/VisualTools/explore/summary).

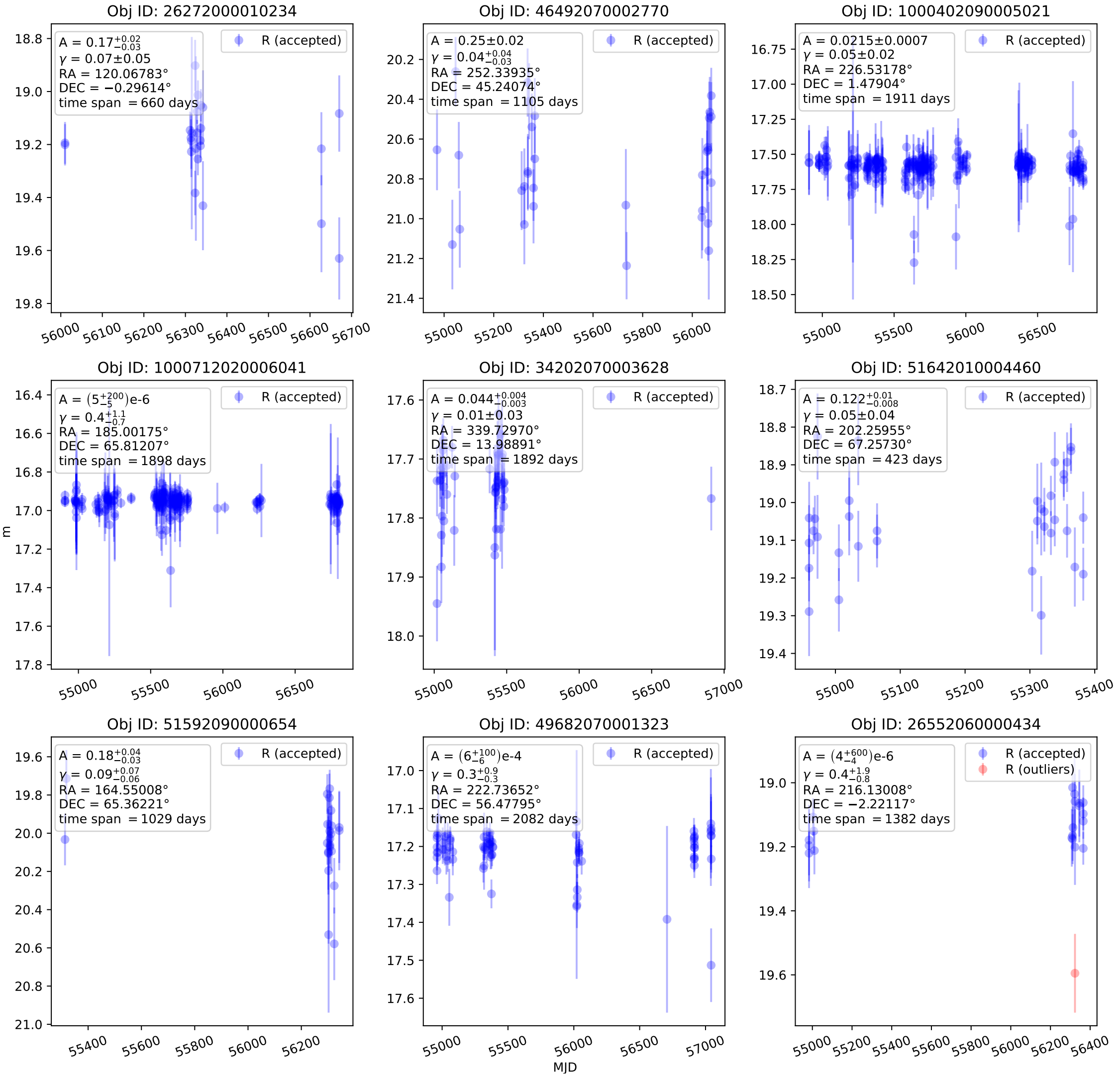


Figure 1.7: R-band PTF light curves of spectroscopically confirmed galaxies in SDSS. The light-curve construction is described in Sects. 3.2 and 3.3 and the outlier removal in Sect. 4.2.5. We see very little variability. However, photometric noise is present, and galaxies may include an AGN without being labeled as a quasar by SDSS.

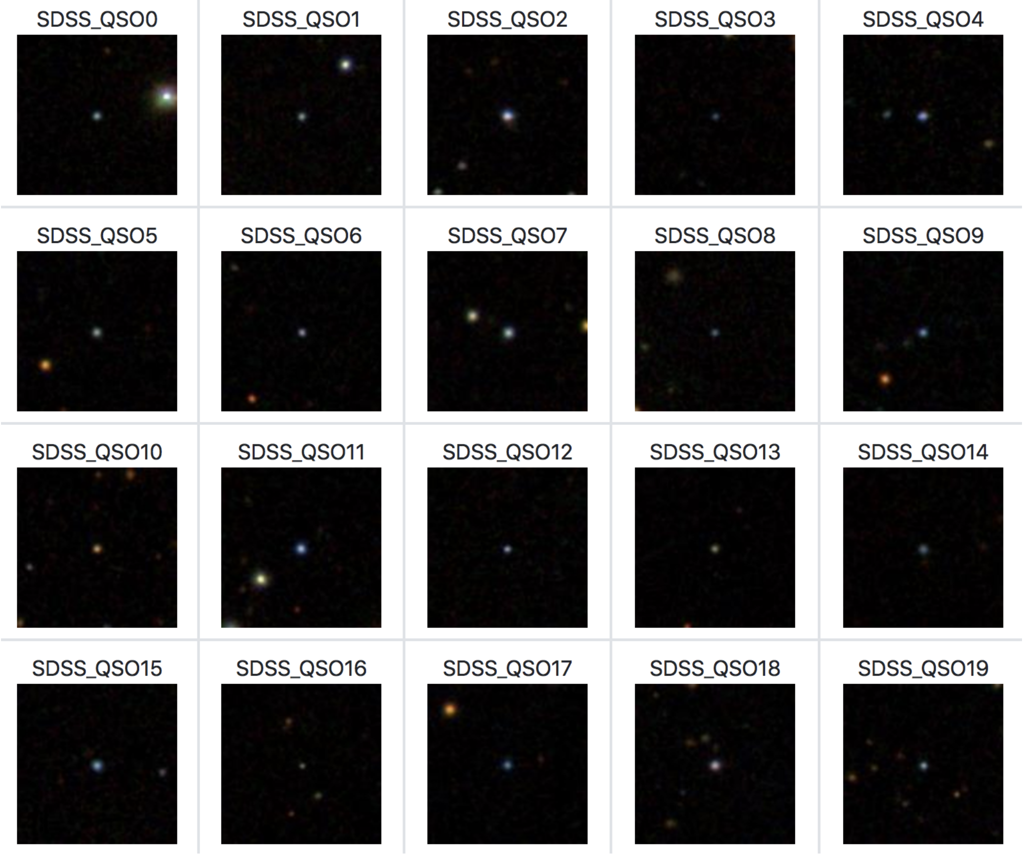


Figure 1.8: Imaging of spectroscopically confirmed quasars in SDSS. They are observed as point sources and many of them are blue. Images are selected using SDSS Image List, `https://skyserver.sdss.org/dr17/VisualTools/list`.

in the same order (left to right, then top to bottom). The sources are relatively blue and show broad emission lines in their spectra. We will come back to this in Sect. 1.3.2. Fig. 1.10 shows light curves of nine of the objects, which show more variability than the typical galaxy (see Fig. 1.7). We discuss quasar and AGN variability in Sect. 1.3.3 and how to quantify it in Sect. 2.4.

1.3.1 AGN structure

An AGN has the structure of Fig. 1.11. Close to the black hole, we find the accretion disk. The infalling matter is heated and becomes hotter closer to the black hole. This makes the disk emit thermal continuum radiation. Strong magnetic fields can produce relativistic

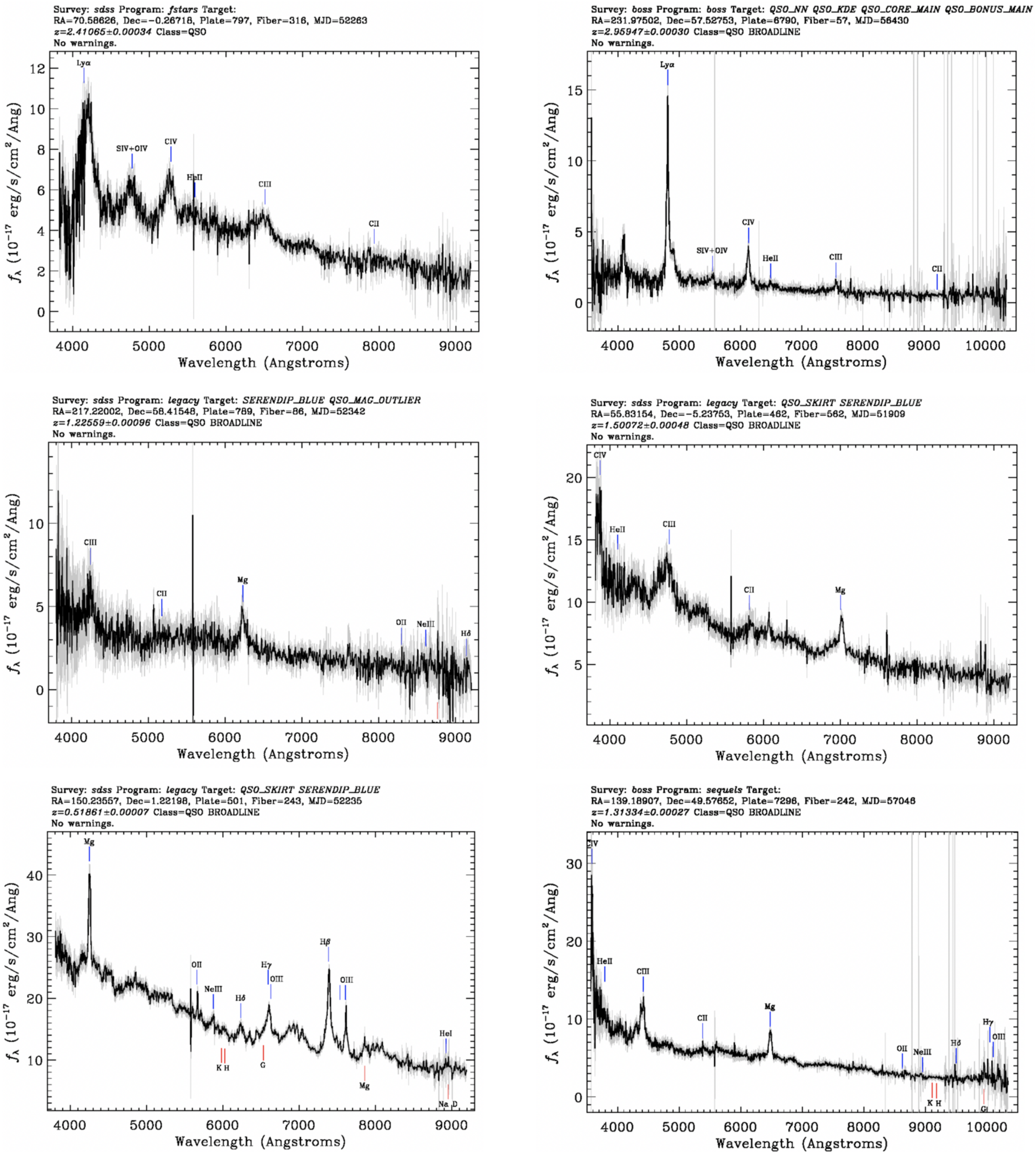


Figure 1.9: Spectra of spectroscopically confirmed quasars in SDSS. The characteristic broad lines are visible. In SDSS, they are registered with SpecObjID 897429125139032064, 888358705291094016, 5641426861566464 00, 7644876143380944896, 5203202703727267 84 and 8214632358990862336 (SDSS-IV DR17, CC-BY license, `skyserver.sdss.org/dr17/VisualTools/explore/summary`).

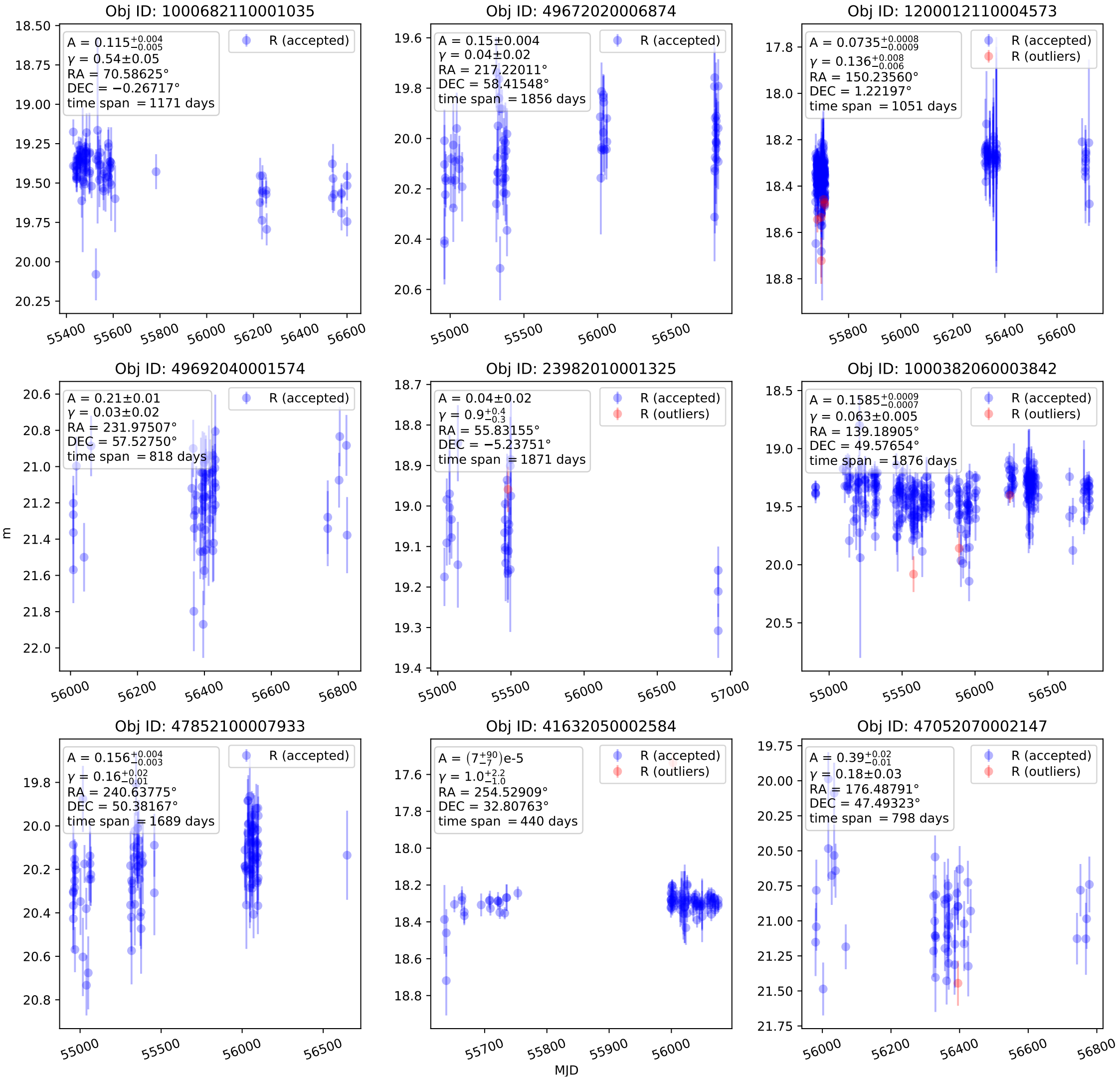


Figure 1.10: R-band PTF light curves of spectroscopically confirmed quasars in SDSS. Quasars are generally highly variable, which shows in the A and γ variability parameters (see Chapter 4). Only one of the shown sources (bottom centre) has an A value less than two $\sigma_{A,-}$ from zero.

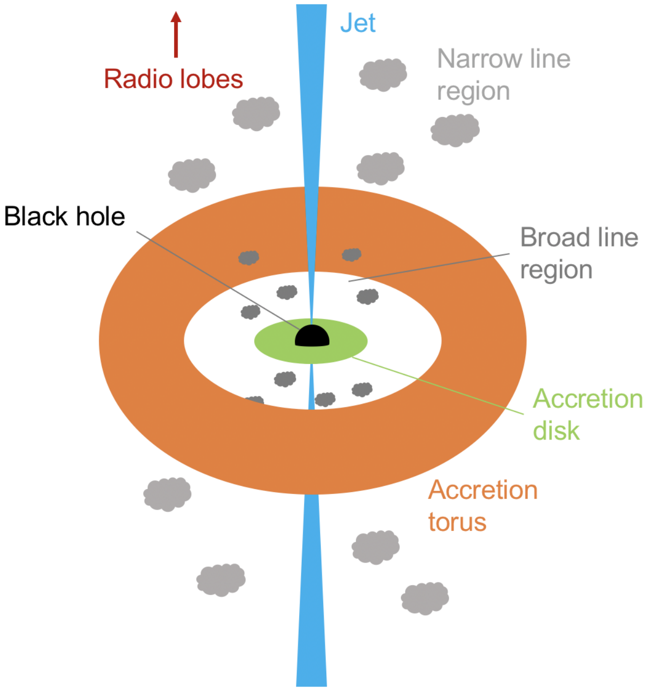


Figure 1.11: Diagram of the structure of an AGN with a jet (created by the author).

jets close to the centre. As plasma moves in the jets, it produces non-thermal polarised synchrotron radiation from the acceleration of spiralling charged particles (mainly at radio wavelengths) (Sparke & Gallagher 2007; Freedman et al. 2014). At the end of the collimated jets, the plasma has lost energy and we find *radio lobes* of material with radio emission. Some photons from the accretion disk, and sometimes the jet, are upscattered on charged particles through inverse Compton scattering. These effects produce continuum emission, including X-rays (La Mura et al. 2017).

Futher out, we have the *broad line region* and the *narrow line region*. The broad line region consists of dense clouds that absorb and re-emit radiation. The gas clouds are in close orbit around the black hole at high velocities, giving a Doppler-broadening in re-emissions. This broadens the emission lines. The narrow line region is further away from the black hole and has a density low enough for "forbidden" emission lines that are otherwise highly unlikely. In the plane of the accretion disk, the disk is surrounded by a dusty *accretion torus* that absorbs some of the emitted light – especially low energy X-rays.

1.3.2 AGN types

If the AGN has a jet that produces radio emission it is radio-loud. Other AGN are radio-quiet. AGN look different depending on viewing angle. *Seyfert galaxies*, which are radio quiet, are classified depending on the ratio of broad and narrow emission lines as Seyfert 1, Seyfert 2 or with a number in between. In Seyfert 2 galaxies, we observe the AGN in the plane of the torus which blocks the broad line region. Thus, we only see broad lines in Seyfert 1 galaxies – unless the emission has been scattered outside the torus (Antonucci 1993; Urry & Padovani 1995; Du et al. 2017). However, the appearance as type 1 or type 2 can also depend on properties of the accretion tori such as width, clumps and optical depth (Ramos Almeida et al. 2011). Seyfert galaxies are mostly spiral galaxies (Sparke & Gallagher 2007). Other AGN can also be classified as type 1 or type 2. Some quasars have high variability of broad lines and continuum emission in the optical and UV, changing their appearance between type 1 and type 1.9 (MacLeod et al. 2019).

If the jet of an AGN points directly towards us, it can be observed as a *blazar*. Blazars are radio loud, extremely bright and highly variable even on scales of days and hours. They can be categorised into two groups: BL Lacertae objects (or BL Lac), and Optically Violent Variables (OVVs) (or flat-spectrum radio quasars (FSRQ)), depending on the strength of emission lines (stronger in OVVs). The spectra are generally close to featureless for BL Lac, which makes spectroscopic redshift estimation difficult (Ajello et al. 2014; Sparke & Gallagher 2007).

Radio galaxies are elliptical galaxies with stronger radio emission of $\sim 10^8 L_\odot$ compared to Seyfert galaxies with $\sim 10^6 L_\odot$. This is emitted in the jets and large radio lobes of the AGN (Sparke & Gallagher 2007).

In the 1950's, several objects were discovered that appeared as point sources but at high redshifts and with strong radio emission. These objects become known as quasi-stellar radio sources, or *quasars* for short. However, quasar can also stand for quasi-stellar objects, and >90 % of quasars are radio quiet (Sparke & Gallagher 2007). The word quasar is now used for both object types. They are extremely bright active nuclei that can outshine their entire host galaxy. Shen (2021) has used a luminosity threshold of $L_{bol} > 10^{45}$erg/s. Extreme luminosities allows them to be observed even at redshifts of $z > 7$. The most distant confirmed quasar at time of writing is J0313−1806 at $z = 7.642$ (Wang et al. 2021). Recently, a James Webb Space Telescope spectrum of GN-z11 at $z = 10.6$ showed that the object is likely an AGN (Maiolino et al. 2023).

1.3.3 AGN variability

Over time, quasars and AGN are known to vary stocastically across the electromagnetic spectrum. Sesar et al. (2007) found that at least 90 % of all spectroscopically confirmed quasars by SDSS show variability. They generally do so over timescales of weeks to decades and can also change in colour. Some quasars may change on even shorter time scales (Schmidt et al. 2012). We will get back to why later in this section.

The physical properties of the central black hole have been correlated with parameters describing variability. Magnitude and direction of the correlations remain unclear, however, and correlations are affected by the limitations of current variability models and baselines (Kozłowski 2017b). Suberlak et al. (2021) and De Cicco et al. (2022) review predictions from theoretical models and empirical results in the literature. The results are inconsistent. De Cicco et al. (2022) only finds an anti correlation between amplitude and both the Eddington ratio, λ_E, and the bolometric luminosity, L_{bol}, but not with the mass of the black hole.

We show examples of quasar light curves in Fig. 1.10. Modelling and recognising the variability characteristics of AGN would be one way of classifying them. Many models have been used in the literature, as we will explore in Sect. 2.4.

One use of AGN variability depends on the distance between the inner accretion disk and the broad line region. When the emission of the accretion disk varies, the light scattered by the broad line region varies as well, but with a time delay of a few weeks. This allows us to measure the distance by *reverberation mapping*. The measured light curve of the continuum is mapped to the expected light curve of the broad line region and then compared to the measured light curve in the broad line region to find the time difference. The size of the broad line region is related to luminosity by the radius-luminosity relationship (Bentz et al. 2013)

$$R \propto L^{\alpha}, \alpha = 0.533^{+0.035}_{-0.033}, \tag{1.2}$$

and with luminosity estimates, we can use AGN as standard candles. Combined with redshifts, they can be used for studying cosmology even at high redshifts (Watson et al. 2011; Bruun 2017). Reverberation mapping is also used for estimating the mass of the black hole (Peterson et al. 2004).

The origin of variability in quasars and AGN is unclear. Multiple possible explanations exist such as:

- Accretion disk instabilities (Rees 1984). Changes in accretion or thermal properties change the produced emission. There are many possible explanations for these instabilities.
    - Magnetorotational instability can destabilise the disk and cause turbulence (Mushotzky et al. 2011). Magnetic torque in the inner part of the accretion disk can also create cooling and heating fronts (Ross et al. 2018).
    - If the disk is not aligned with the spin of the black hole, it can be broken into rings. This causes variability with high amplitudes and short timescales that can be quasi-periodic (Raj & Nixon 2021).
    - The torus can drive in clumps of gas and the changes in mass accretion rate changes the luminosity (Hopkins et al. 2012).
- Reprocessing of X-rays and UV in the inner disk can cause variability. Changes in emission at the centre of the disk will cause changes in re-emitted light as the light interacts with the disk (Shappee et al. 2014).
- Changes in obscuration of emitted light. For example, clumps in the torus can change the obscuration on time scales of months to years (Hopkins et al. 2012).
- Jets can cause variability on time scales of hours (Kelly et al. 2009).
- Gravitational lensing (Chang & Refsdal 1979). More on this in Sect. 1.3.4.

Other factors might affect variability timescales without causing variability themselves. For example, strong magnetic fields can increase the thickness of the accretion disk which shortens inflow times (Dexter & Begelman 2019).

1.3.4 Lensed quasars

Gravitational lensing occurs when massive objects change the curvature of spacetime. This bends light around the object, causing a lensing effect in much the same way as a convex lens of glass. To an outside observer, objects behind the lens can be amplified. They can also appear in different and even multiple positions on the sky. The lensing allows us to gather more information about both the lens and the lensed object, which is used for detection of exoplanets (Griest & Safizadeh 1998), for example.

Gravitational lensing can make a point source quasar appear in multiple positions on the sky, and if it is unresolved this can make the point source appear extended. Kochanek et al. (2006) suggests selecting lensed quasars as apparently extended and variable sources.

Lensed quasars can be used for time-delay cosmography. By studying the lensing effect, we can estimate distances, and these are tied to cosmological parameters. When the light of a quasar passes the lensing object, multiple paths can be directed towards the same observer and thus multiple images will appear. They show the quasar at different times depending on the difference in length of the paths and the gravitational potential of the lens. The time delay between images thus reveals a time delay distance. This can be related to the positions of the images and the lensing object (Refsdal 1964; Treu & Marshall 2016).

Understanding the mass and mass distributions of lensing objects allows us to study dark matter and the mass evolution of galaxies. However, lensing galaxies are not representative of all galaxies. 80 % of lenses are elliptical galaxies according to Möller et al. (2007).

1.4 Coordinates and time

Positions of the astronomical sources can be specified in multiple coordinate systems. For easy coordinate conversion, one can use a tools such as those by the Space Science Data Center [1].

We will mainly use the *equitorial coordinate system*. Using cartesian coordinates in this system, we can describe any position using right ascension (RA) and declination (Dec). RA is the angular distance along the plane of the Earth's equator from the position of the Sun during vernal equinox. It increases towards East. Dec is the angular distance perpendicular to the equator. RA has values from 0 to 360, and Dec is in the range from -90° to $+90^\circ$ with north being positive. In the J2000 frame, the equinox and equator are defined at 12:00 January 2000.

RA and Dec can also be specified in *sideral time*, as the rotation of the Earth connects

[1] Italian Space Agency, tools.ssdc.asi.it/conversionTools

time and angle. Each hour marks a change of 15 $^{\circ}$ in RA, and each minute is $1/4$ $^{\circ}$. Small angles can be measured in arcminutes ($1/60$ $^{\circ}$) or arcseconds ($1/3600$ $^{\circ}$), not to be confused with minutes and seconds of sideral time. Milliarcseconds are written "mas".

An alternative system is the *Galactic coordinate system*. Coordinates are Galactic latitude and longitude with respect to the Galactic plane, with the Galactic centre as the zero point. Galactic latitude increases to the north of the Galactic plane, and Galactic latitude increases towards the Galactic east.

The time stamps of observations (*epochs*) are typically given as Modified Julian Date (MJD). This measure counts the number of days since 00:00 17 November 1858 in Universal Time (UT). It is a modified version of Julian Date (JD), where MJD$=$JD-2400000.5. JD is zero at 12:00 1 January 4713 BC.

Chapter 2

Statistical methods

Astronomy presents particular challenges in data science. The size of the data sets may vary, uncertainties need to be handled, tabular data may be sparse and time series data have irregular sampling (as shown in e.g. Fig. 1.3). For optimal data processing, we consider several statistical techniques. Statistics and machine learning will provide essential tools for astronomical analysis. By carefully choosing and applying the methods, we can optimise computational efficiency, minimise bias and extract new information. This requires philosophical considerations about what we are truly trying to estimate and the subtle causes of overfitting.

First, we consider what we mean by *probability* and how to treat it under a Bayesian interpretation. This will lead us to the optimisation of likelihoods for parameter estimation and how to do this in practice with the chains of MCMC. The goal of this thesis is to analyse variability, and so, we will discuss various ways of quantifying it and their application in the literature. To cross match large astronomical databases, we define k-d trees. Mathematical tree structures are also a common method in machine learning for the classification of variable sources. We discuss the use of machine learning in astronomy, the fascinating prospects and the pitfalls, and provide examples of how it is used in the context of classifying variable objects.

2.1 Bayes and likelihoods

Probability can be interpreted through two major paradigms: frequentism and Bayesianism. In the frequentist paradigm, a probability describes the frequency by which something happens. The chance of rolling a 6 with a fair 6-sided dice is 1/6, because it happens 1/6th of the time (as a limit for an infinite number of rolls). In this view, one cannot state that the probability of the Universe having some constant property is 17 %, because either it does or it does not. It must be possible to measure multiple times and get multiple different outcomes to estimate the probability of each as a frequency. Alternatively, probability can be interpreted as an objective property of the object in question, but this would still not allow us to say a hypothesis has a probability of 17 % (Barlow 1989).

In the Bayesian view, probability is no longer a property of a dice or the Universe, but it quantifies a subjective degree of belief about them. In a case of two mutually exclusive theories about the Universe and no prior knowledge of which is true, we would guess that each of them have a 50 % chance of being correct. There is no reason to prefer one over the other. When we take measurements, we update the estimate to e.g. 70 % and 30 %. In Bayesianism, when we evaluate hypotheses, we update a prior probability to get a posterior probability (Barlow 1989).

Bayes theorem, named after Thomas Bayes, was published for a special case in 1763 (Bayes & Price 1763). It explains how to update the conditional probability of an event A given another event B, $P(A|B)$:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}. \tag{2.1}$$

The prior probability of A is $P(A)$. $P(A|B)$ is the posterior probability. $P(B|A)$ is the conditional probability of B given A, or the likelihood. $P(B)$ is the marginal probability, or evidence. The more likely A is in general, the more likely it is after observing that B is true. If the probability of B is high given A is true, then A is also more probable when B is true. However, the more likely B is to be true in general, the less information it gives us about the probability of A.

When $P(B)$ is unknown, it can be computed as $P(A)P(B|A) + P(\neg A)P(B|\neg A)$, if we know probability of B being true in the cases of A being true or not true. For multiple values of A_i, estimation of $P(B)$ with a sum or integral can be complicated. Sometimes, it is enough to consider $P(A|B) \propto P(A)P(B|A)$. For hypothesis testing, the evidence $P(B)$ is constant during comparison of a hypothesis A_i with A_j.

A common challenge is the estimation of a continuous variable θ based on data x consisting of measurements $x_1, x_2, ...x_n$. To find the maximum a posteriori probability (MAP) and thereby the most probable value of θ given the data, we optimise $P(\theta)P(x|\theta)$. For $P(x|\theta)$, we need the total probability of observing all data points x_i given θ. Assuming x_i are independent measurements, the probability of observing all of them is the product of observing each one.

$$\hat{\theta}_{\mathrm{MAP}} = \mathrm{argmax}_\theta P(\theta)P(x|\theta) = \mathrm{argmax}_\theta P(\theta)\prod_i P(x_i|\theta) \tag{2.2}$$

If the prior is uniform or negligible compared to the evidence, we simply find the maximum likelihood, $P(x|\theta)$. That is, we identify the θ value that would be most likely to produce the observed data though maximum likelihood estimation (MLE) (Barlow 1989).

$$\hat{\theta}_{\mathrm{MLE}} = \mathrm{argmax}_\theta P(x|\theta) = \mathrm{argmax}_\theta \prod_i P(x_i|\theta) \tag{2.3}$$

For easier computation, it is common to optimise the log likelihood using sums instead of products. This can sometimes be done analytically by differentiating $\ln P(x|\theta)$ and setting it equal to zero.

The prior can be known from previous estimations of the posterior, or it can be estimated from background information. Without prior information, we want to choose an uninformative prior – i.e. a prior that gives equal weight to all possible outcomes when there is no reason to choose one over the other. Often, an uninformative prior would be a uniform prior, which corresponds to not using a prior. However, a uniform prior in log space would be a logarithmic prior. For variables which must be positive, the prior could be uniform for all positive values and zero for negative values. To estimate the prior of events with different multiplicities, the prior can be based on entropy or the information in each event. A common choice which is invariant under coordinate transformations (after applying a Jacobian) is Jeffreys prior (Jeffreys 1946). Jeffreys prior is based on information in all possible θ values, which can be computed from the likelihood.

2.2 Markov Chain Monte Carlo

We now have two methods for estimation of a parameter θ based on data x (Eqs. 2.2 and 2.3). Both methods depend on optimisation of the likelihood. However, it is not always possible to find an analytical solution. How do we then estimate $\hat{\theta}_{\text{MAP}}$ or $\hat{\theta}_{\text{MLE}}$ in Eqs. 2.2–2.3?

One way is to use a Monte Carlo method. Monte Carlo methods are numerical methods that rely on generating random data. We can use this to simulate events and their probabilities. For example, we can draw random samples θ_i, compute $P(x|\theta_i)$ for each and thereby we have sampled the probability density function (PDF) of the likelihood. This allows us to estimate the distribution of likelihoods and infer both the "best" $\hat{\theta}_{\text{MLE}}$ and its uncertainties even if they are asymmetric. By using a prior as well, we estimate $\hat{\theta}_{\text{MAP}}$.

We will specifically use a Markov Chain Monte Carlo (MCMC) method. These are widely used in astronomy (Sharma 2017). We create a walker to explore the parameter space of θ by suggesting values θ'. In a Markov Chain, this happens in steps, where the value θ_{i+1} of step $i + 1$ depends on the previous one, θ_i. The next step should only depend on the current step and be independent of the past – this is called the Markovian property. If the suggested θ' is accepted, it becomes θ_{i+1}. If θ' is rejected, a new value is suggested based on θ_i.

For a walker in equilibrium, there is no overall drift in the accepted values. Each step draws one random sample from a PDF of θ. We can plot them in a histogram or apply metrics to study the distribution. The starting value θ_0 is chosen manually. If it is far from the "true" θ we are trying to estimate, it will take some steps before the walker finds a probable parameter region and reaches an equilibrium. Therefore, the first steps are discarded as "burn-in".

In Fig. 2.1 we see the chains of multiple walkers exploring a two dimensional parameter space. The burn-in phase is visible during the first ~ 100 steps. Fig. 2.2 shows histograms of the generated parameter values after discarding the first 200 steps, and how the histograms reveal the shape of the PDF for each parameter. The choice of how long to run the MCMC depends on the trade off between computing resources and how well the PDF should be estimated.

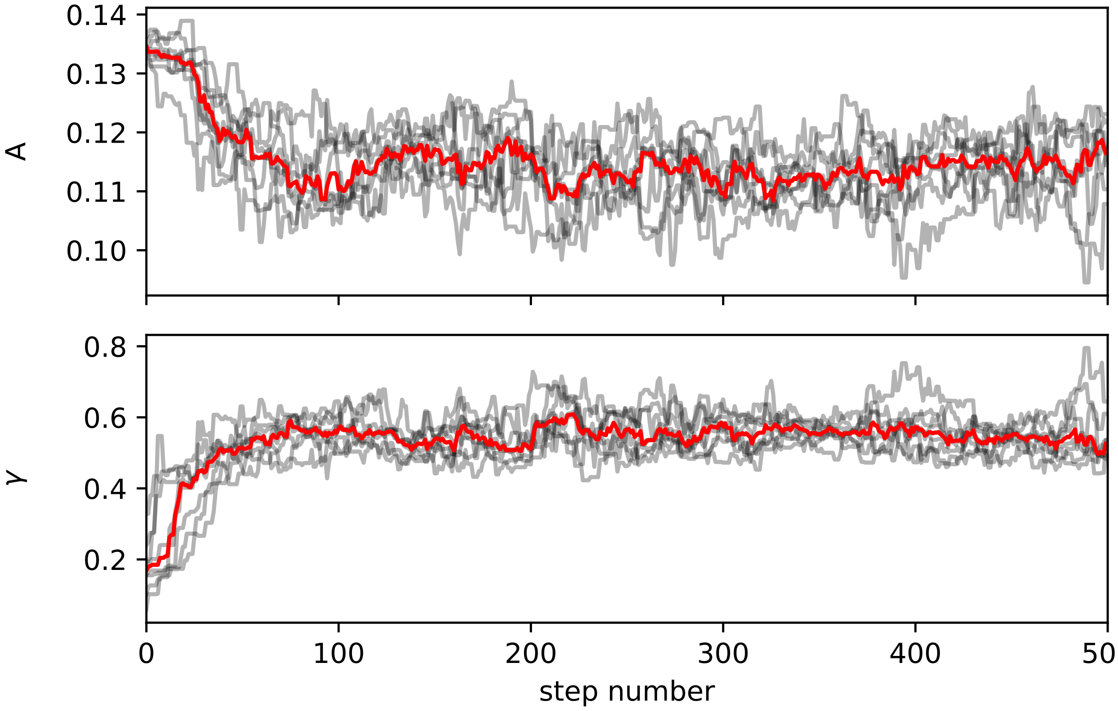


Figure 2.1: Chains of multiple walkers in a MCMC method to estimate the posteriors of two variables: A and γ. The parameters describe the SF of a light curve (see Sect. 2.4.1). The path of each walker is traced in black, and the median is marked in red. The results may vary between multiple runs of the algorithm if it falls into a local minimum. The chains above fit the SF of a spectroscopically confirmed quasar at RA=70.586255, Dec=-0.26717 with PTF object ID (OID) 1000682110001035.

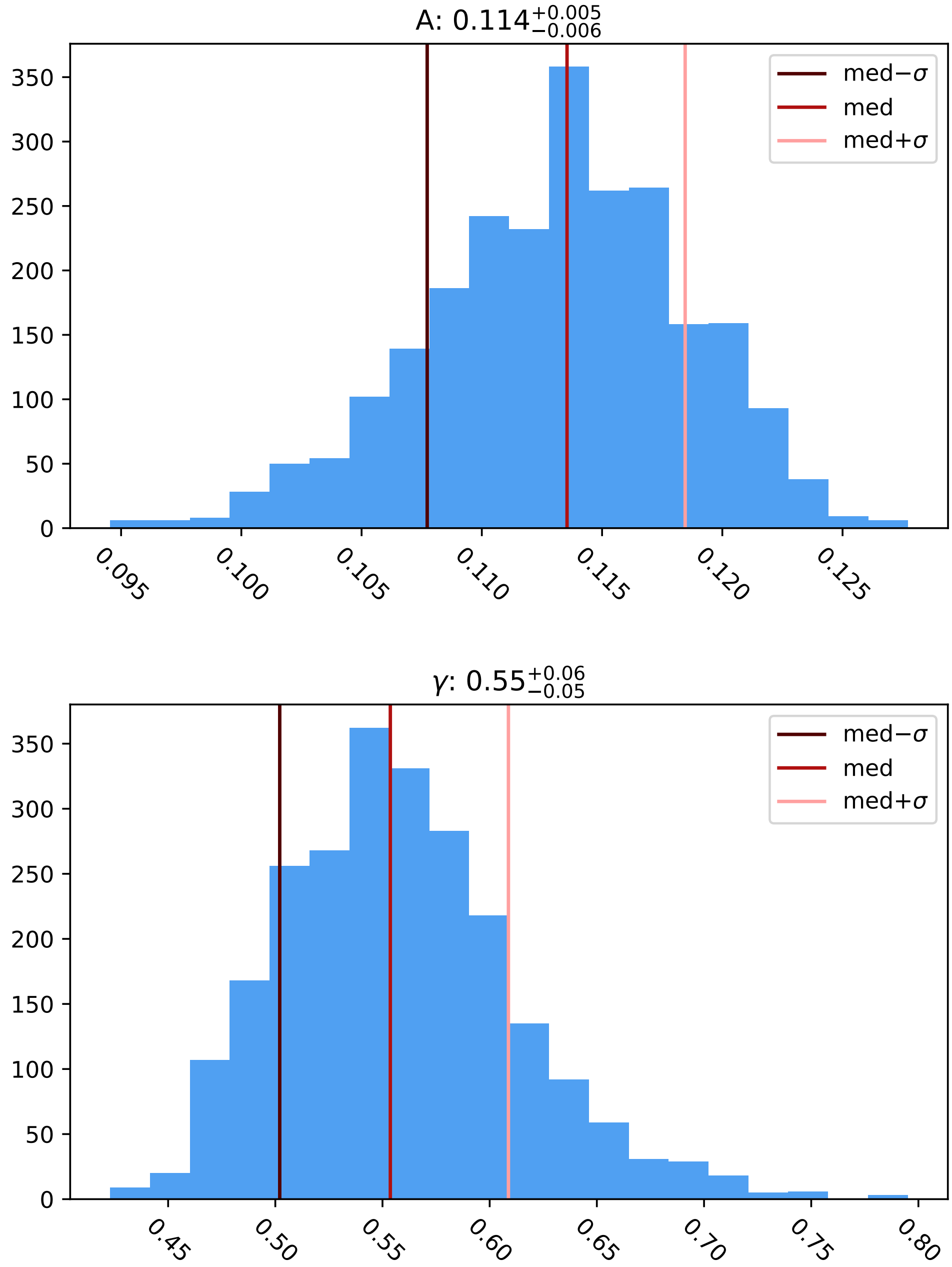


Figure 2.2: Histograms of the MCMC chains of Fig. 2.1 after burn in. They sample the PDF of the fit parameters A (top) and γ (bottom). We have marked the median and 16th and 84th percentiles.

2.2.1 Sampling algorithm

Multiple algorithms exist for the suggestion and acceptance of new steps. A common one is the Metropolis-Hastings algorithm, in which the probability of acceptance depends on the ratio of the posteriors of the current and the suggested step, $\alpha = \frac{P(\theta')P(x|\theta')}{P(\theta_i)P(x|\theta_i)}$ (Chib & Greenberg 1995).

For multivariate sampling of parameters θ and ϕ, sampling the joint distribution from $P(\phi, \theta|x)$ can be difficult. Instead, we can use their conditional probabilities. We can sample θ_{i+1} from $P(\theta|\phi_i, x)$ and then ϕ_{i+1} from $P(\phi|\theta_{i+1}, x)$. This is called Gibbs sampling (Geman & Geman 1984) and is an alternative to Metropolis-Hastings. We sample along one dimension at a time to choose the coordinates of the next step in the parameter space of θ and ϕ.

Some MCMC algorithms run multiple walkers for each step of the chain. The walkers can interact through different moves, i.e. ways of suggesting new steps. In the `emcee` package (Foreman-Mackey et al. 2013), which is used in this thesis, the default is the stretch move (Goodman & Weare 2010). In a stretch move, a walker suggests a step directly in the direction of, or away from, the position of another walker. The package only uses affine invariant moves to improve sampling of anisotropic PDFs Affine transformation invariance includes invariance under reflection, rotation, scaling, translation and shearing, but parallel lines stay parallel. The affine invariant sampler is unaffected by linear transformations that could transform the posterior into a space where it is isotropic. This is relevant for multivariate sampling of covariant parameters, and helps the walkers explore the parameter space by suggesting steps that are more likely to be accepted during Gibbs sampling.

2.3 k-d trees

To cross-match astrophysical objects, we search the coordinates for their nearest neighbours in other surveys. As the surveys are large, this needs to be done efficiently, so we divide the data into partitions using a data partitioning algorithm. This way, we can access one partition at a time, and we will not need to compare the coordinates of all objects in one survey to the coordinates of all objects in another survey.

The coordinates are two dimensional, so for efficient matching, we create two dimensional *tree structures*. A tree is a structure consisting of *nodes* that are connected hierarchically. Each node can have multiple *child nodes* but only one *parent node*. In a *binary tree*, each parent node has a maximum of two children. The end nodes (with no children) are called *leaf nodes*, and the rest are *internal nodes*. The node with no parents is called the *root node*. This is illustrated in Fig. 2.3.

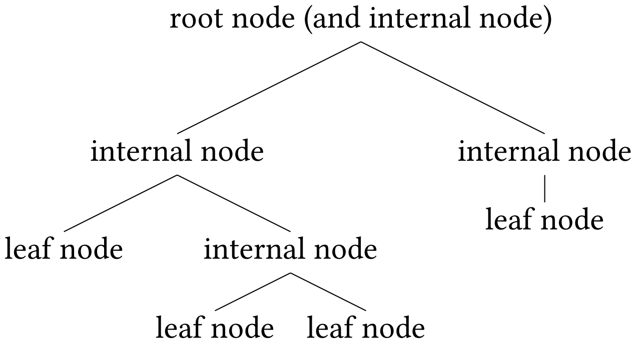


Figure 2.3: Diagram of a tree structure. Parent nodes are shown above and connected to their children with a line. This tree is binary, since no parent node has more than two child nodes.

By creating a binary tree that connects all coordinates in a survey, we can speed up the process of searching for them. A *k-d tree is* a binary tree in k dimensions (Bentley 1975). It consists of nodes that each represent a k dimensional hyper-rectangle. We call each rectangle a *cell*. All internal nodes split a cell into sub-cells and are associated with a splitting hyperplane. The hyperplane is described by an axis and a value along the axis.

While searching the k-d tree for the neighbours to a set of coordinates, we start at the root node. Then, we follow the splits to arrive at the leaf node representing the cell that would include the queried coordinates. We then search the objects in the cell represented by the leaf to find close matches. If other cells are closer to the queried coordinates than the closest matched object so far, those cells can be searched as well. This might be unnecessary if the search is simply for the *approximate nearest neighbor* within $(1+\epsilon)d$ of the distance d to the true closest match, for a specified ϵ.

To create the cells, we start with a cell including all data, represented by the root node. We then choose a hyperplane and use it to split the cell into two sub-cells. We choose new hyperplanes for those cells etc. until a set maximum leaf size is reached for all leaf nodes. Several splitting methods exist. The *standard splitting method* (Friedman et al. 1976) uses the median of the positions of the objects. For clustered data, this method has

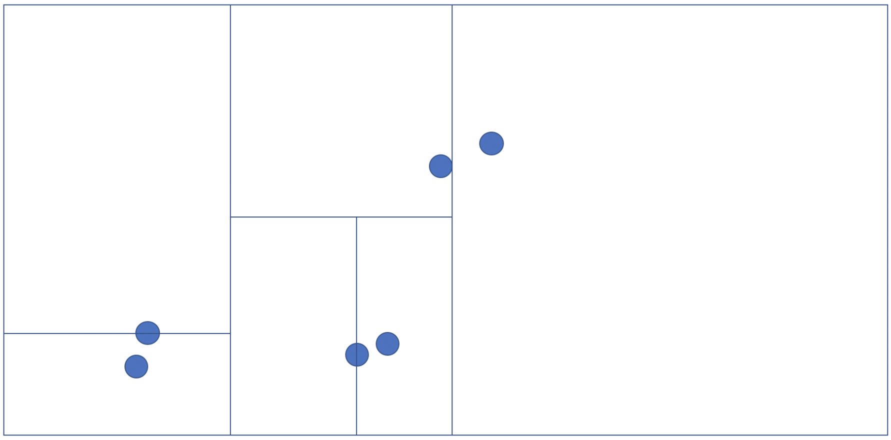


Figure 2.4: Illustration of how the sliding midpoint method would split a two dimensional space into cells depending on the positions of data points. Hyperplanes split cells into subcells starting from a cell including all data.

a tendency to produce highly elongated cells, which slow down the querying process. The *midpoint method* split hyper-rectangles in the middle, but this can produce empty cells. Empty cells means the tree is larger than necessary and that also slows querying. Arya & Mount (1998) created a method with a bounded aspect ratio for cells and no empty cells: the *sliding midpoint method*. This method is illustrated in Fig. 2.4. Cells are split based on their midpoint, but if this would result in an empty cell, the splitting hyperplane "slides" to the position of the nearest object. A detailed comparison is found in Maneewongvatana & Mount (1999). We will use the sliding midpoint method in Part II.

2.4 Measuring variability

As we have seen in Chapter 1, many types of objects show variability and for many different reasons. To identify variable sources, understand the physical mechanisms behind variability and use them in other astrophysical contexts, we need ways of quantifying variability.

Quasars have long been selected by variability. For example, van den Bergh et al. (1973) used variability to avoid a bias against highly redshifted quasars in colour selection. They measured variability as differences between at least two photographic plates. They did so using visual inspection with a blink comparator that could quickly "blink" between images.

A few years later, Usher (1978) selected objects in colour ($U - V$) and optical variability with a quasar purity of 70 %. Variability was measured using 3 statistics: χ^2 of the null hypothesis that the source is non-variable, signal-to-noise ratio of all pairs of points in a light curve, and the difference between the magnitudes and the weighted average magnitudes.

Another example is Butler & Bloom (2011) using variability for 99 % complete quasar selection and selecting new ones in SDSS Stripe 82. They study the variability of quasars in general and classify individual objects based on how well they fit the ensemble model. Quasars have correlated variability on longer timescales than most stars, so they can be recognised by studying timescales of variability. Variability is especially relevant for quasars with redshifts of $2.5 < z < 3$, where it is more difficult to use colours. Quasars are often selected by "UV excess" using $u - g$, but at $2.5 < z < 3$ the values similar to those of stars.

AGN in low mass galaxies are also easier to miss by emission line ratios, and so Baldassare et al. (2020) selected low mass AGN in PTF using a DRW model of variability. Ward et al. (2021) also selected AGN in dwarf galaxies using several variability measures. They used the Pearson correlation coefficient between binned fluxes in g and r (normalising the covariance $C_{g,r}$ between the bands),

$$r = \frac{C_{g,r}}{\sigma_g \sigma_r}, \tag{2.4}$$

where σ_g and σ_r represent the standard deviation of each band (or the root mean square (RMS) of the scatter). So σ_g is

$$\sigma_g = \sqrt{\frac{1}{N-1} \sum_i \left(g_i - \langle g \rangle\right)^2}, \tag{2.5}$$

including Bessel's correction of -1 to get an unbiased estimator when the mean is estimated from the sample (Barlow 1989). Ward et al. (2021) also quantified variability with the reduced χ^2 comparing g and r to the mean of each band,

$$\chi^2/N = \frac{1}{N} \sum_i^N \frac{\left(m_i - \langle m \rangle\right)^2}{\sigma_i^2}, \tag{2.6}$$

the reduced χ^2 compared to the ensemble SF of Butler & Bloom (2011), and the excess variance σ^2_{rms}. The excess variance relates the magnitudes to the light-curve mean after subtracting photometric uncertainties:

$$\sigma^2_{rms} = \frac{1}{N\langle m\rangle}\sum_i^N (m_i - \langle m\rangle)^2 - \sigma_i^2. \tag{2.7}$$

They found that the best selection of AGN was achieved with the Pearson correlation coefficient and χ^2/N compared to the mean magnitude.

Sesar et al. (2007) also used the standard deviation and reduced χ^2 of the mean in each SDSS band ($ugriz$). In addition, they used the minimum and maximum magnitude, light-curve skewness, median magnitudes and colours. This allowed them to select RR Lyrae (see Sect. 1.1.1) with 95 % completeness and 70 % purity (see Eq. 4.8 for definitions of purity and completeness).

The choice of informative variability measures plays a special role in source classification with machine learning. We will get back to this in Sect. 2.5.

2.4.1 Structure functions

Simonetti et al. (1985) introduced the structure function (SF) in astronomy. They are well suited for studying stocastic time series and allow for unevenly spaced data. The idea is to predict how much the magnitude of an object changes given the time between observations. We compare changes in magnitude (Δm_{ij}) with changes in time scale (Δt_{ij}) between two data points i and j.

An empirical SF can be found without assuming a specific model of variability. The data point pairs are binned in time Δt_{ij} and the measurement uncertainties subtracted to estimate the average intrinsic variability with

$$\mathrm{SF}_{\mathrm{emp}}(\Delta t) = \sqrt{\frac{1}{N}\sum_{ij}\Delta m_{ij}^2 - \sigma_i^2 - \sigma_j^2}, \qquad |\Delta t_{ij}| \sim \Delta t \text{ and } i < j. \tag{2.8}$$

When SFs are used on ensembles, the variability is studied for those objects in general (Simonetti et al. 1985). Only two points are needed per source, and so, more objects can be included. Li et al. (2018) include quasars measured with DECaLS and SDSS for combined baselines of up to 15 years. However, ensemble SFs ignore the significant differences between individual quasars (MacLeod et al. 2010). The ratio of Type I to Type II AGN affects the shape of the SF, and Type I AGN are easier to detect with optical variability, as the accretion disk is directly observed. Long baselines are especially important for detection of type II AGN (De Cicco et al. 2022).

SFs of the light curves of individual quasars can also be modelled. To preserve as much information as possible for sparsely sampled objects, this can be done without binning. (Schmidt et al. 2010) use a power-law model

$$\mathrm{SF} = A\left(\frac{|\Delta t_{ij}|}{t_0}\right)^{\gamma}. \tag{2.9}$$

with amplitude A over a timescale t_0 (one year) and a power-law index γ. We describe this simple, well-studied process in more detail in Sect. 4.3.1.

At large timescales, $\mathrm{SF}_{\mathrm{emp}}$ can reach a turning point where it stops behaving as a power law, as the variability would be too large. Instead, $\mathrm{SF}_{\mathrm{emp}}$ becomes flat and might behave as a DRW model in Sect. 2.4.3.4. Kozłowski (2016b) describe this with a turnover timescale τ, which they find to be one year. MacLeod et al. (2010) estimate it as 500 days, and Stone et al. (2022) as 750 days and unconstrained. γ decreases until it is zero, and this means long light curves can introduce a bias towards low γ. Very short timescales can also bias γ estimation depending on the noise subtraction. Kozłowski (2016b) warns of several ways the noise has been incorrectly subtracted in the literature, and they fit a single power law for $4 < \Delta t < 365$ days. Bauer et al. (2009) and De Cicco et al. (2022) argue that the turnover could be an artefact of irregular sampling.

2.4.2 PSD

The power of a light curve may depend on frequency (ν). This is measured as the power spectral density (PSD), describing the power per unit frequency as a function of frequency. The PSD of aperiodic signals follow $\mathrm{PSD}(\nu) \propto \nu^{\alpha}$. Different values of α correspond to different "colours" of noise. White noise is defined by a mean of 0, a finite variance and being uncorrelated, and so the PSD is flat (Scargle 1981). Pink noise falls of as ν^{-1}, red noise as ν^{-2}, blue noise is proportional to ν and violet noise to ν^2. Red noise is also known as random walk noise or Brownian noise. The index α is related to γ in Eq. 2.9 by $\alpha = -4\gamma$ (Kozłowski 2016b).

Kelly et al. (2011) is an example of fitting AGN light curves with a PSD (of DRW mixtures, see Sect. 2.4.3.4). Kozłowski (2016b) shows how an AGN PSD might look, but they also note that PSDs are more difficult to work with than SFs and sensitive to irregular sampling.

2.4.3 CARMA

Kelly et al. (2009) describe quasar variability by a DRW. Such a process is similar to a random walk but *mean-reverting* – the further the flux is from the mean, the greater the tendency of returning to the mean. To understand the DRW model and more advanced models, we will first inspect some related processes and terms.

In an autoregressive process (AR) model, the process remembers previous output steps of the time series, and base a new step on them with added noise. In a moving average (MA) model, the process remembers previous input noise terms and knows how they influence the next step. This is generalised as autoregressive–moving-average (ARMA) models, or CARMA in continuous time.

2.4.3.1 *AR*

In an AR model, values in a time series depend linearly on past values and a random term. This is written

$$X_t = \sum_k B_k X_{t-k} + R_t \tag{2.10}$$

for a flux X at step t with white noise R. B_k are constants describing how well the model "remembers" the past (Scargle 1981).

2.4.3.2 *MA*

A MA model describes time series data with a series of random pulses with a deterministic influence of future values. A MA process is described by

$$X_t = \sum_k C_k R_{t-k}, \tag{2.11}$$

where R is white noise and C_k are constants. The constants work as a filter determining the influence of previous pulses of white noise on the current X_t. To ensure that current pulses do not influence X_t at infinity, C_i follow the stability condition $\sum_{-\infty}^{\infty} C_i^2 < \infty$ (Scargle 1981). The MA model is not to be confused with computing the moving average of a time series similar to the moving median of Sect. 4.2.5.

2.4.3.3 *ARMA*

AR and MA models can be converted and expressed as the other – but this can turn finite order models into infinite-order models. They can also be generalised as ARMA models that depend on both the input noise and the output steps to avoid pure AR or MA of infinite order.

$$X_t = \sum_k^p B_k X_{t-k} + R_t + \sum_k^q C_k R_{t-k}, \tag{2.12}$$

is an ARMA model of order (p,q) (Scargle 1981).

2.4.3.4 *DRW*

MA, AR and ARMA models can be generalised for continuous time. A continuous version of AR is known as a continuous autoregressive process (CAR). Equivalently, CARMA is the continuous version of ARMA. The DRW is based on a first order CAR (the Ornstein–Uhlenbeck process). This can be expressed as a differential equation

$$\frac{dX}{dt} = -\frac{1}{\tau} X(t) + \sigma R(t), \tag{2.13}$$

where τ and σ are positive constants. The first term represents the mean-reverting property depending on the current step X and second term represents the noise (Gillespie 1996).

Kelly et al. (2009) describe a quasar DRW by

$$\frac{dX(t)}{dt} = -\frac{1}{\tau} X(t) + \sigma\sqrt{dt}\epsilon(t) + bdt \tag{2.14}$$

where τ is the dampening timescale, σ describes variability at short timescales, $\epsilon(t)$ is white noise with zero mean and unit variance, and b is related to the mean flux by $b =$

$\frac{\langle X(t) \rangle}{\tau}$.

The SF of a DRW model can be related to the autocorrelation function (ACF) of the light curve. The ACF is the correlation of the light curve with itself shifted in time as a function of the time lag. Kozłowski (2017a) models the ACF with a power exponential

$$\mathrm{ACF}(\Delta t) = \exp\left(-\left(\frac{|\Delta t|}{\tau}\right)^{\beta}\right). \tag{2.15}$$

$\mathrm{SF}_{\mathrm{emp}}$ and its value at infinite Δt (SF_{∞}) are related to the ACF by (ignoring photometric noise; Hughes et al. 1992; Li et al. 2018; Kozłowski 2016b) :

$$\begin{align} \mathrm{SF}_{\mathrm{emp}}(\Delta t) &= \sqrt{2}\sigma\sqrt{1 - \mathrm{ACF}(\Delta t)} \tag{2.16} \\ &= \mathrm{SF}_{\infty}\sqrt{1 - \mathrm{ACF}(\Delta t)}. \tag{2.17} \end{align}$$

When $|\Delta t| \ll \tau$, $\mathrm{SF}_{\mathrm{emp}}$ can be approximated as a single power law (using the Taylor approximation $e^x \approx 1 + x$ for small x)

$$\begin{align} \mathrm{SF}_{\mathrm{emp}}(\Delta t) &= \mathrm{SF}_{\infty}\sqrt{1 - \exp\left(-\left(\frac{|\Delta t|}{\tau}\right)^{\beta}\right)} \tag{2.18} \\ &\approx \mathrm{SF}_{\infty}\left(\frac{|\Delta t|}{\tau}\right)^{\beta/2}. \tag{2.19} \end{align}$$

Kawaguchi et al. (1998) discuss how β depends on physical models of variability (see Sect. 1.3.3). With $\beta = 1$, the model is a DRW (Kelly et al. 2009; Li et al. 2018):

$$\mathrm{SF}_{\mathrm{DRW}}(\Delta t) = \mathrm{SF}_{\infty}\sqrt{\frac{|\Delta t|}{\tau}}. \tag{2.20}$$

This is similar to the SF in Eq. 2.9 for $\gamma = 0.5$, since $\beta = 2\gamma$. The DRW can be fit to $\mathrm{SF}_{\mathrm{emp}}$ for estimation of τ and SF_{∞} (Ivezić & MacLeod 2014; Li et al. 2018).

On longer timescales than the dampening timescale τ, the magnitudes become (approximately) uncorrelated, $\mathrm{SF}_{\mathrm{DRW}}$ approaches SF_{∞} and the light curve will be dominated by white noise. At shorter timescales, the PSD falls of as ν^{-2} and the DRW behaves as red noise (Kelly et al. 2009; MacLeod et al. 2010).

The DRW model has its limitations. For example, it may fit light curves well even if

$\beta \neq 1$ (Kozłowski 2016a), and this would affect the estimated τ and $\mathrm{SF}_{\mathrm{emp}}$. Sánchez-Sáez et al. (2018) found that γ in SF models are often not 0.5, and then DRW models are not appropriate.

For DRW processes, unbiased estimates of τ and $\mathrm{SF}_{\mathrm{emp}}$ also require baselines ten times longer than the "true" τ (Kozłowski 2017b). Suberlak et al. (2021) decreased the scatter in correlations of DRW parameters with physical AGN parameters by combining light curves from SDSS and PS1 for 15 year baselines for 9248 quasars. They argue that longer light curves would further improve the fits. Stone et al. (2022) fit DRW models to 190 quasars, and find that τ is unconstrained and continues to increase with baseline for a median τ of 750 days, even with a baseline of 20 years. This indicates that either longer baselines or more complex models are needed, such as higher order CARMA models (Moreno et al. 2019) or combinations of multiple DRWs (Kelly et al. 2011). Stone et al. (2022) also find a divergence from the expected DRW PSD on scales shorter than one month.

2.5 Machine learning

Machine learning plays an increasingly important role in astrophysics. With larger datasets, manual data analysis becomes impractical and less efficient. A machine can automatically detect patterns even in high-dimensional data and optimise the analysis of it. An example of such a task is the creation of decision boundaries between classes. Each object is described by a number of "features", and a model will use these to learn where in feature space one can find different types of objects.

Ball & Brunner (2010) describe four paradigms of science: Theory, observation, simulation and data mining. Machine learning is efficient for data mining of existing astronomical databases for new insights. Djorgovski et al. (2022) discuss how machine learning is used in many branches of astronomy. Supervised machine learning models can efficiently classify sources, while unsupervised learning can detect sub-populations and rare objects in large data sets. Baron (2019) summarises important machine learning techniques for astronomy such as neural networks, random forests, support vector machines and some unsupervised algorithms for clustering, dimensionality reduction and outlier detection. Tree-based methods create decision boundaries using tree structures, similarly to the k-d trees of Sect. 2.3. We explain this in more detail in Sect. 5.2.3. Smith &

Geach (2022) discuss the use of neural networks in astronomy in particular. They predict and recommend the creation of a general, open-source "foundation" model requiring less specialist knowledge. They also point out that large, astronomical datasets can benefit the field of machine learning, as other large data sets are often proprietary.

Recently, natural language models have achieved the capacity to produce academic writing of a quality that is difficult to distinguish from humans, apart from a tendency to produce generic texts and the extent of confident misinformation. Although no language models have aided the production of this thesis, it may very soon be considered another standard machine learning tool in astronomy (Nature Astronomy editorial 2023). The abstract of Smith & Geach (2022) was generated automatically. Ciucă & Ting (2023) have developed a language model specifically for "chatting" with the astronomical literature. It can distil papers while preserving their meaning and references, give short summaries, compare papers and even generate ideas for new research.

2.5.1 Model selection

The best machine learning model depends on the type of data and task. This may be time series, images, spectral cubes, text etc. *Domain knowledge* is important to the choice of model and prepossessing strategies. The best model is selected based on performance evaluated using chosen metrics. For unbiased evaluation with the purpose of choosing the best model for unknown data, performance of a supervised classifier must be estimated from a hold out data set. This *test set* is separate from the *training set* using during model selection. Ideally, the test set has not influenced model selection at all (including preprocessing). It should not be used or even inspected before the final evaluation. Even if the test set is inspected and no prior decisions are changed, it will be biased – because it could have had other, possibly extreme, values that would have led a human to change the model. This way, some information from the test set can leak into the model, and it will no longer be independent. We discuss this in detail in Chapter 5, especially in Sect. 5.2.2.

Interpretability of the final model depends on the chosen algorithm and may even be prioritised over performance. The linear criteria of a single decision tree are easier to interpret than the complicated decisions of a deep neural network, which may appear as a black box. There are ongoing efforts to develop methods for explainable AI, which interprets machine learning models. For example, we might analyse which features are most important to a model (Lundberg & Lee 2017).

Models may also be chosen based on speed, memory or climate impact. In many cases, model complexity can be increased indefinitely with diminishing returns. For example, one may train deep neural networks taking up more resources than a tree based model with similar or even better performance (Grinsztajn et al. 2022). Schwartz et al. (2019) argue that efficiency should be an evaluation metric in large models as cheaper models will benefit both the climate and inclusion. Plotting performance as a function of, for example, the size of the training set may improve efficiency and cost of future research. Model selection can also be optimised by trying random *hyperparameter* combinations instead of creating a grid of all combinations (see Sect. 5.2.3.2).

2.5.2 Classification of variable sources

To classify the variability of astronomical sources, it is possible to simply input the full light curve in a machine learning model without preprocessing. However, this would mean the input space is extremely large, and we would suffer the curse of dimensionality (Altman & Krzywinski 2018). This requires more effort to learn from than a smaller feature space. We want preprocessing to preserve as much information in as few features as possible – but sacrificing a little information for simplicity can improve performance. Below, we describe some machine learning strategies for classification of variable objects.

Belokurov et al. (2003) use neural networks for detecting microlensing events and distinguishing them from other types of variability. They discuss the importance of domain knowledge and use this to select 5 light-curve features especially relevant to microlensing events. This model can directly output the probability of each object being microlensed.

Palanque-Delabrouille et al. (2011) also uses a neural network for distinguishing quasars from stars based on variability. They input χ^2 (Eq. 2.6) for each SDSS band ($ugriz$) and power law SF parameters A and γ (Eq. 2.9, with separate A for the g, r and i bands). The output y_{NN} quantifies quasar-like probability. For quasars, it increases up to about 40 epochs, and also performs better for bright quasars ($g \sim 18.5$). A tight cut in y_{NN} leads to worse performance for quasars at high redshifts where the model is less confident. They analyse a (non-representative) subset of SDSS to reduce the required resources of time and memory. Some objects would otherwise have been rejected as quasar candidates by colours.

Jamal & Bloom (2020) discuss the efficiency of various neural network architectures for time-series classification and test them on variable stars. They input the entire light curves and use *encoder* modules to create representations of them. These representations showed clustering at values common for different types of variable stars, which is useful for subtype classification.

Peters et al. (2015) select quasars in SDSS using nonparametric Bayesian Classification Kernel Density Estimation from a set of 916 587 objects. They form a data set of objects that have been spectroscopically confirmed as quasars and objects that have not. They conclude that a combination of colours and variability gives the best performance, judging by spectroscopically confirmed quasars. Colour alone performs a bit better than variability alone. Variability is described with power law SFs for light curves with at least 10 epochs, to show the sufficiency of this simple model for variability selection of sources. However, their "test" and "training" sets did not represent the same population of objects, and in practise, they tested and trained only on the training set. This would be fine using cross-validation (see Sect. 5.2.2.1) if it was not used for deciding on an optimal model, but it is, and so, this process leads to a classifier with unknown performance for new, unbiased data.

Sánchez-Sáez et al. (2019) and De Cicco et al. (2021) use a random forest algorithm to select optically variable AGN using colours and variability. De Cicco et al. (2021) do so for the COSMOS field with the VLT Survey Telescope. They include colours ($u-B$, $B-r$, $r-i$, $i-z$ and $z-y$) and 29 variability features for 20 670 sources over $\sim$3.3 years with at least 27 epochs. Unlike Sánchez-Sáez et al. (2019), they do not preselect by variability, to ensure the sample is more representative of the sky. They do require that all sources have data in $uBrizy$ and matches to the COSMOS ACS catalog for morphology (stellarity index by a neural network in SExtractor, Bertin & Arnouts 1996). By training the model with and without colours, they find variability (in combination with stellarity) to be more powerful for classification than colour. Purity is especially high for Type I AGN – compared to spectroscopically confirmed objects, they achieve a purity of 91 % and a completeness of 69 %. This is for a binary classifier selecting AGN and non-AGN. Sánchez-Sáez et al. (2019) studied the QUEST-La Silla AGN variability survey (Cartier et al. 2015) and included a model only trained on variability features

Cunha & Humphrey (2022) do not include variability features, but they use another powerful tree-based method, namely gradient boosted trees (for 3 497 864 sources). We explain this algorithm in detail in Sect. 5.2.3. They use an ensemble of three variants – XGBoost, LightGBM and CatBoost – and combine the outputs by voting for better performance than the random forest of Clarke et al. (2020). They emphasise the importance of combining data to create colours and photometric redshifts (feature engineering). They perform multi-class classification of SDSS spectroscopically confirmed stars, galaxies and quasars. They include RA and Dec for taking into account the relative density of each class at different galactic latitudes.

Logan & Fotopoulou (2020) take another approach, and choose unsupervised learning with HDBSCAN (McInnes et al. 2017) on a representative set of 49 181 sources. They use magnitudes and all colour combinations (no variability). HDBSCAN selects clusters in parameter space by density. They then compare the clusters with spectroscopic labels when available for an overall semi-supervised method.

II

Selection based on variability

Chapter 3

Data processing

Data processing takes in place in multiple parts. The goal is to select and classify sources based on variability and colours – but first we must gather and process this information. We start by querying the raw PTF data. We then use it to assemble light curves and match them to objects in WISE and PS1. A model is then fit to the SF of each light curve. Finally we match to SDSS data. This all needs efficient processing to be completed in a reasonable time frame. In the following sections, we detail how we have optimised the processing.

3.1 High performance computing

For increased computing power and to improve the total run time, we run multiple processes in parallel using the High Performance Computing Centre (HPC) at the University of Copenhagen. This is managed by the SCIENCE HPC Center and includes a partition for DARK. The HPC is a computer cluster that can run long scripts in parallel and with large amounts of memory and storage space. We will use it to run python 3.6 scripts. 70 computing nodes (and 5 frontend nodes) are available in the DARK partition of the cluster. The nodes have two Intel Gold 6130 each with a total of 32 cores and 64 threads.

The nodes have access to 3 GB memory per thread except two nodes with 12 GB per thread. The scripts are scheduled and controlled via the Slurm Workload Manager (Yoo et al. 2003) using *batch scripts*. We create *job scripts* with instructions and requirements for computing jobs as batch scripts. Upon completion of a job script, Slurm will generate an output file containing the python outputs and potentially error messages. The HPC allows job scripts to run for up to 14 days.

3.2 Querying PTF

First, we need to query data from the PTF light-curve database. As the database contains 598 975 024 objects, this cannot be done with a single all-sky query. We divide the sky into $10^\circ \times 10^\circ$ *patches* and query each with a Table Access Protocol (TAP)[1] query to the NASA/IPAC Infrared Science Archive (IRSA) Catalog Search Tool[2]. The resulting tables can be large, so we use the asynchronous TAP operation mode, which generates a link to obtain the results when the query has finished. In query number 32 of the patch at 10°< RA < 20°and +50°< Dec < +60°, we apply the following constraints:

- `fid=2` (we use the R band),
- `ngoodobs>=20` (at least 20 observations that have not been flagged or masked),
- `oid>0` (a valid object ID has been assigned),
- `mag_autocorr>12` (R is larger than 12),
- `mag_autocorr<22` (R is less than 22),
- `limitmag>0` (the limiting magnitude is valid),
- `ra>=10.0`,
- `ra<20.0`,
- `dec>=50.0`, and
- `dec<60.0`.

[1] International Virtual Observatory Alliance, http://www.ivoa.net/documents/TAP/

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

We apply further data cleaning later in the process (see Sect. 4.2.5). Below is an example query for patch number 32 with the output saved as `ptf_32_10p0_20p0_50p0_60p0.csv`:

```
curl -v -o /storage/dark/shbruun/output_all/ptf_32_10p0_20p0_50p0_60p0.csv "https://irsa.ipac.caltech.edu/TAP/async?QUERY=SELECT+LC.ra,+LC.dec,+LC.obsmjd,+LC.mag_autocorr,+LC.magerr_auto,+LC.oid,+LC.bestMedianMag+FROM+ptf_lightcurves+AS+LC+WHERE+LC.ra>=10.0+AND+LC.ra<20.0+AND+LC.dec>=50.0+AND+LC.dec<60.0+AND+LC.fid>1+AND+LC.ngoodobs>=20+AND+LC.oid>0+AND+LC.limitmag>0+AND+LC.mag_autocorr>12+AND+LC.mag_autocorr<22+ORDER+BY+LC.oid&FORMAT=CSV&PHASE=RUN"
```

We create and run a python script (`Generate_script_query.py`) for the generation of 648 queries (one for each patch), distributed on 21 job scripts with up to 32 queries in each. For each job script, slurm gives an output file with the TAP links to the results. We create another python script (`check_queries.py`) to check all query links in a slurm output file and download the results if they are ready. The script also saves lists of queries that have been completed, are being executed or failed and need to be restarted.

3.3 Assembling and matching light curves

A row in a table of query results from the PTF light-curve database each describe a single *data point* in a light curve. We will assemble light curves and match them to data in WISE and PS1. We do this with a script (`Classfication_PartI.py`).

First, we load a table of query results. Each row contains a value of RA, Dec, MJD, R, σ_R, OID and median R. We now group the data points of the patch. We identify unique OIDs and how many data points are associated with each. We then search the list of OIDs to get the indices to the data points of each light curve. Objects can be split at the edge of a query patch, but this is unlikely. The mean RA and Dec are saved as the coordinates to each astrophysical object.

For querying WISE and PS1, we make a shorter list of PTF coordinates in the patch after rounding to the nearest arcsecond. This rounding speeds up the process without affecting the results much, since we match within five arcseconds. First, we query to get a list of all nearby data points in another survey (WISE or PS1). We use Astroquery (Ginsburg et al. 2019) to access the CDS X-Match Service (Boch et al. 2012) for 'vizier:II/328/allwise' and 'vizier:II/349/ps1'(Cutri et al. 2021; Chambers et al. 2016; Ochsenbein et al. 2000). We then group the data points by OID and compute the mean coordinates, like for PTF, if a source has been observed multiple times. We cross match the full list of mean PTF coordinates in the patch to the list of mean coordinates to objects in another survey using a k-d tree (see Sect. 2.3).

The number of data points per patch vary across the sky. This results in some empty query files and some that will take longer than 14 days to process. This is especially important during the fitting of Sect. 3.4, which is slower than the work of this section. Therefore, we split 310 large query files into 6340 files with a python script (`Split_files.py`). They are split based on size and actual processing time to avoid reaching the 14 day limit per job script. Each file represents a computing *task* for the HPC. In Fig. 3.1, we show the number of tasks per patch. We keep data points with the same OID in the same query file (unless it was split across different patches to begin with). We generate 50 job scripts for light-curve assembly and matching with max 128 tasks per file (plus an extra job script to rerun failed tasks).

3.4 Fitting

We will now perform the fitting detailed in Sect. 4.3.2 with a python script `Classification_PartII.py`. This loads the light curves from one task in Sect. 3.3 and fits their SFs with MCMC. But first, we must clean the light curves. We check again that $12 < R < 22$ and remove outliers, as we will get back to in Sect. 4.2.5. If the light curve still has > 20 data points and the time span is still > 365 days, we fit the object.

We initiate the MCMC sampler with emcee (Foreman-Mackey et al. 2013) and define the log likelihood function. We will get back to this in Sect. 4.3.2. Fitting with the MCMC is the slowest part of the data processing. The sampler needs to evaluate the log likelihood for every suggested step (see Sect. 2.2), so we simplify the computations as much as possible. We do this by pre-computing some expressions for data points i and

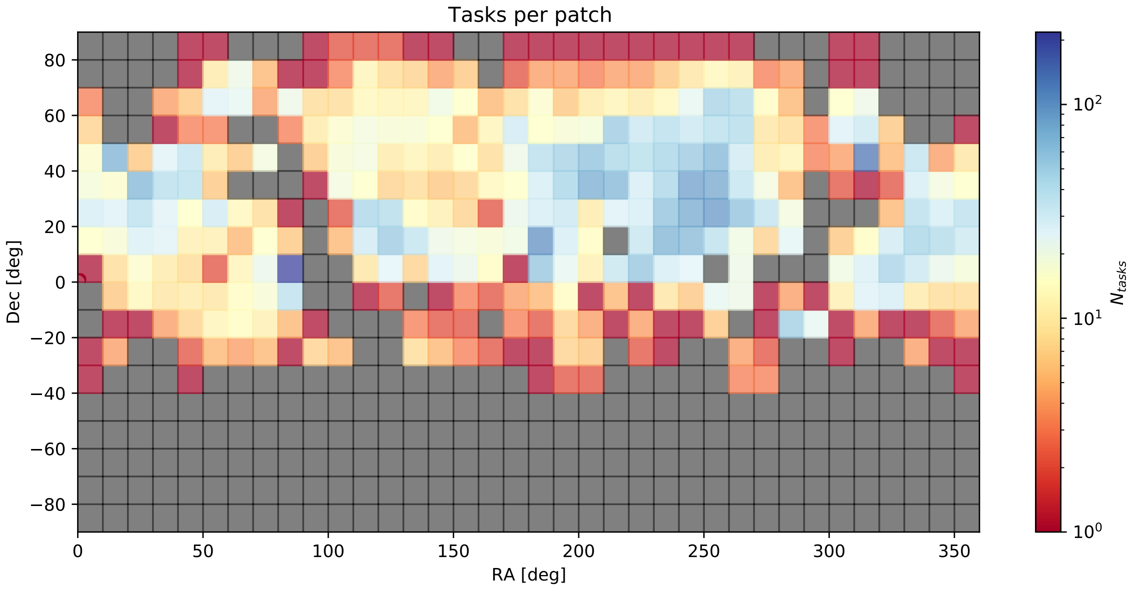


Figure 3.1: A map of the number of fitting tasks per patch.

j, where $i \neq j$ and $i > j$. We define Δt_{ij}, Δm_{ij}, $\Delta m^2/2$, $\sigma^2_{m,i} + \sigma^2_{m,j}$ and $|\Delta t_{ij}/t_0|$.

Due to the long computing times, the 6340 tasks must be distributed more carefully to job scripts than in Sect. 3.3. We sort the sizes of files containing the light curves and generate 473 job scripts processing a similar amount of data. Note that this is not an ideal way of estimating processing time, since e.g. the number of data points per light curve has not been taken into account. The first script has 90 tasks and the last only 7. However, patch number 153 (80 °< RA < 90 °, 0 °< Dec < +10 °) has especially well sampled light curves, so we separate most of the 217 tasks processing this patch and dedicate a job script to each. Some light curves have over 5000 epochs. We also split most job scripts with more than two tasks processing patch 31 (10 °< RA < 20 °, +40 °< Dec < +50 °) as we find them to be slow. Memory is not an issue – we only need 128 MB per thread.

To automatically start new job scripts when the previous ended, we create 14 *tracks* of job scripts with identical names and use the slurm `dependency=singleton` option. Processing the PTF footprint takes approximately six months. To keep track of the progress, we create a python script, `Classification_ProcessPlots.py`. This estimates the remaining processing time and creates plots to visualise progress. At the end, the estimated total fitting time for each patch is illustrated in Fig. 3.2. It is

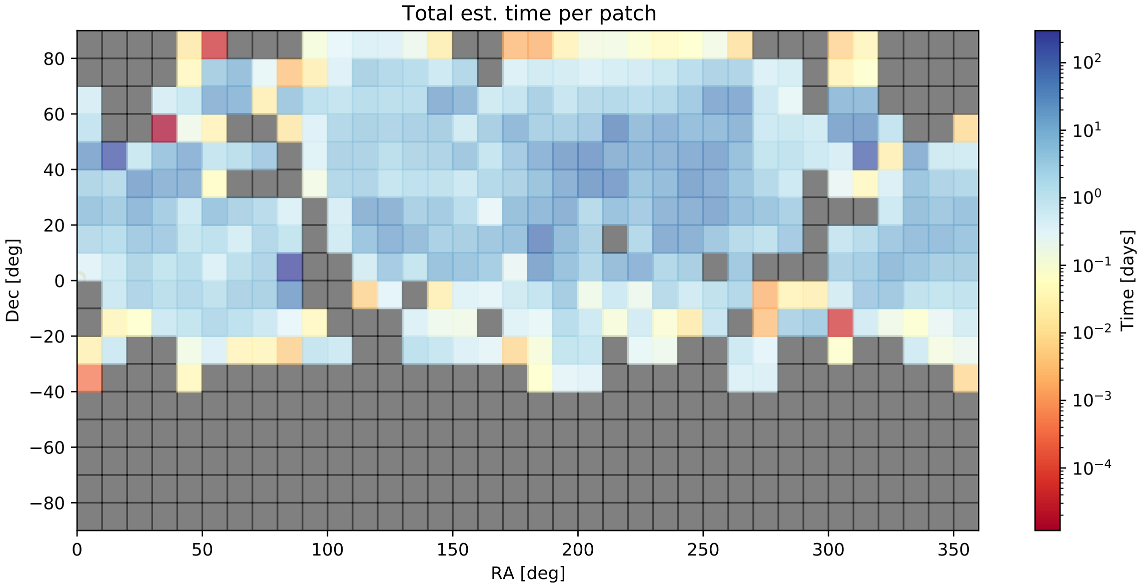


Figure 3.2: A map of estimated fit processing time in each patch. The slowest patch would take over 200 days to process if we did not split it into smaller tasks that can be processed in parallel.

clear from the map that patch 153 and 31 are among the slowest to process. We also use the script to keep track of which tasks are complete and to automatically read and identify errors in the slurm output files. This allows us to quickly check the error types and create job scripts for rerunning failed tasks. For the fitting process, we need 62 extra job scripts for failed tasks.

3.5 Combining and cross matching with SDSS

Results of the fitting and matching are now spread across thousands of files. To combine them, we create a python script (`Classification_Combine.py`) and a single job script using 2990 MB per CPU. With this, we load each type of output file for all patches (and all tasks in each), combine them and save them. When we have all data in a few files, we can load it to make selections. For example, sources with specific fit parameter values or colours.

To cross match with SDSS objects, we generate files of data for all spectroscopic quasars, stars and galaxies in SDSS PhotoObj and SpecObj DR17 (Abdurro'uf et al. 2022).

We use CasJobs on SciServer (Taghizadeh-Popp et al. 2020) to query – the script below shows the quasar query:

```
SELECT ALL
  p.ra,p.dec,p.objID,s.z,s.zErr
INTO mydb.QSO_table_DR17
FROM PhotoObj AS p
  JOIN SpecObj AS s ON s.bestobjid = p.objid
WHERE
  s.class='QSO'
```

For each class, we compare the SDSS coordinates to the PTF coordinates using a k-d tree with a maximum distance of two arcseconds. The data is now ready for selection and classification.

Chapter 4

Structure function fits to 71 million objects

This chapter is based on the following article:

"VarIabiLity seLection of AstrophysIcal sources iN PTF (VILLAIN) I. Structure function fits to 71 million objects"

Authors: Sofie Helene Bruun, Adriano Agnello and Jens Hjorth

In press at Astronomy & Astrophysics; accepted on 11/04/2023. arXiv: 2304.09903

Author contributions

SB led the project, queried the data and performed the data analysis of this chapter. SB wrote and revised the draft which has benefited from multiple rounds of comments and suggestions from all authors. The main idea was proposed by AA and JH. SB was the corresponding author with support from AA and JH.

Abstract

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

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

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

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

4.1 Introduction

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

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

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

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

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

(2021) found variable AGN in the Zwicky Transient Facility (ZTF; Bellm et al. 2019) using a combination of variability measures.

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

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

In this paper, we aim to analyse the results and performance of a similar method but applied to the entire PTF survey. We show where variability is most effective in breaking degeneracies and which forms of variability are common among different types of objects. Specifically, matching with spectroscopic SDSS classifications of QSOs, stars, and galaxies, which we assume as ground truth, we can evaluate new variability selection criteria and compare with those of Schmidt et al. (2010)[1]. By including as many objects from the PTF survey as possible, the properties of the total light-curve sample are more representative of PTF sources than if we only included specific stellar subtypes, for example.

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

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

We describe the datasets (and the data-cleaning process) used in this work in Sect. 4.2. We introduce the methodology in Sect. 4.3. In Sect. 4.3.1, we define models of the PTF light-curve SFs, and we fit them in Sect. 4.3.2. The metrics used for evaluation of the selection of objects are defined in Sect. 4.3.3. The results of the SF fitting are plotted and described in Sect. 4.4. Based on the variability parameters and matched colour information, we choose photometric selection criteria and examine the properties of selected objects in Sect. 4.5. We discuss these properties in Sect. 4.6 and summarise our findings in Sect. 4.7.

4.2 Data

4.2.1 PTF

The Palomar Transient Factory was a project with the Palomar 48 Schmidt telescope at Palomar Observatory in California. This includes imaging in DR1, DR2, and DR3[2] from 1 March 2009 to 28 January 2015 covering 20 000 deg^2 (Law et al. 2009; Rau et al. 2009)[3]. The PTF light-curve database is based on a subset of these images with data from 598 975 024 objects in R and g.

As most data points are in R, these are chosen for the analysis of this paper (as `mag_autocorr` and `magerr_auto`). The R band is adapted from the 658 nm Mould-R filter described in Cuillandre et al. (2000), has a limiting magnitude of 20.6, and photometry is in the AB magnitude system (Law et al. 2009). From this database, we also consider the timestamps, PTF object ID, RA, Dec and median R, all queried in 648 patches of $10^\circ \times 10^\circ$ via the IRSA Catalog Search Tool[4]. We split 310 large patch files into 6340 files, keeping data points with the same object ID together. File by file, we group data points by ID to assemble light curves. These are cleaned and fitted in Sects. 4.2.5 and 4.3.1 in parallel using 14 computing nodes for six months.

[2] as the Intermediate Palomar Transient Factory for DR3

[3] www.ptf.caltech.edu/iptf

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

4.2.2 SDSS

The Sloan Digital Sky Survey DR17 (Blanton et al. 2017; Abdurro'uf et al. 2022) includes 5 789 200 spectra (Smee et al. 2013; Wilson et al. 2019; Drory et al. 2015) from the Apache Point Observatory in New Mexico, USA (Gunn et al. 2006). Based on DR17 from February 2021 (with data both in PhotoObj and SpecObj) 866 338 QSOs, 962 162 stars, and 2 790 253 galaxies are spectroscopically confirmed. Their coordinates, redshifts, and spectroscopic classes are queried via CasJobs on SciServer (Taghizadeh-Popp et al. 2020). We cross-match by creating a k-d tree (Maneewongvatana & Mount 1999) of the PTF coordinates. We then query the tree to find the nearest SDSS object within two arcseconds, if it exists. By inspecting SDSS images[5], we find that this radius best avoids spurious matches.

4.2.3 WISE

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

We query WISE objects with the CDS X-Match Service (Boch et al. 2012) within five arcseconds of the mean PTF coordinates to each object. From WISE ('vizier:II/328/allwise'; Cutri et al. 2021; Ochsenbein et al. 2000), we get $W1$ (3.4 µm), $W2$ (4.6 µm), their errors, and coordinates. $W3$ and $W4$ are not included, as they are not deep enough to include most of the PTF sources. WISE magnitudes are in the Vega magnitude system (Wright et al. 2010). For each PTF object, only the closest WISE source is selected via a k-d tree, as described in Sect. 4.2.2. If this source contains multiple data points, we choose the lowest magnitude, as this is the brightest detection, and save the mean RA and Dec.

4.2.4 PS1

The Panoramic Survey Telescope and Rapid Response System includes two telescopes in Hawaii: Pan-STARRS1 and Pan-STARRS2. We use the g (481 nm), r (617 nm), and z (866 nm) bands (Tonry et al. 2012) from the Pan-STARRS1 survey DR1 with 1.92 billion

[5] using the Image List tool at skyserver.sdss.org/dr17/VisualTools/list

objects (‘vizier:II/349/ps1’; Chambers et al. 2016; Ochsenbein et al. 2000). This covers the sky at J2000.0 declination > -30 °, including the PTF footprint. The PS1 magnitudes are in the AB system. As in WISE, in PS1 we query within five arcseconds of PTF objects and select one set of magnitudes, magnitude errors, and coordinates per source.

4.2.5 Data cleaning

Data cleaning is needed for reliable fitting results. For PTF, extreme magnitudes outside the range of $12 < R < 22$ are removed. If a light curve contains too few data points, fit parameters are difficult to constrain, and if the time span is too short, we cannot detect variability over long timescales. Therefore, we require that each light curve must contain at least 20 data points and span at least one year. At 80°< RA < 90° and 0°< Dec < +10°, light curves are so well sampled that in order to speed up the analysis in Sect. 4.3.2 we select a maximum of 1500 random data points for some sources.

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

$$\frac{|R - \mathrm{MWM}|}{\sqrt{\sigma_R^2 + \mathrm{MAD}^2}} > 3, \tag{4.1}$$

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

$$\mathrm{MAD} = \text{weighted median}\left(|R_{close,i} - \mathrm{MWM}|\right). \tag{4.2}$$

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

Table 4.1: Sample sizes.

Sample	Counts
Full PTF light curve database	598 975 024
Cleaned PTF R-band sample	70 920 904
Matched	
SDSS spectra	1 748 047
WISE	57 007 069
PS1	70 891 378
WISE & PS1	56 991 591
All surveys	1 613 916

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

to base rejection on.

After data cleaning, the remaining light curves are distributed on the sky according to Fig. 4.5, mostly covering the northern hemisphere and avoiding low Galactic latitudes. Table 4.1 contains sample statistics including objects matched in SDSS, WISE, and PS1. 70 920 904 objects are analysed, spanning $365-2147$ days and containing $20-5579$ data points.

4.3 Methodology

We describe the variability of each object by defining power-law models of their structure functions. The models are fit to structure functions of individual objects to extract variability descriptors as fit parameters. These are used for exploring relations in the data and photometric selection of object classes in Sects. 4.4–4.5; but first, we define metrics for evaluation of selection quality in Sect. 4.3.3.

4.3.1 Structure functions

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

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

$$V^2_{\text{eff},ij} = \text{SF}^2_{ij} + \sigma^2_{m,i} + \sigma^2_{m,j} \approx (\Delta m_{ij})^2, \tag{4.3}$$

$$\text{SF}_{ij} = A\left(\frac{|\Delta t_{ij}|}{t_0}\right)^{\gamma}, \; A \geq 0. \tag{4.4}$$

Equation 4.4 has two free parameters, A and γ; the former quantifies the amplitude of the variations, in units of magnitude, on the timescale t_0, and the power-law index γ describes how the amplitudes depend on timescale. We choose $t_0 = 1$ year (observed frame). One could correct for the factor of $1 + z$ to find rest frame Δt_{ij}, but redshift information is limited, and SDSS redshifts are not independent of the spectroscopic labels we use for comparison.

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

4.3.2 Fitting

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

$$L_{ij} = \frac{1}{\sqrt{2\pi V^2_{\text{eff},ij}}} \exp\left(-\frac{(\Delta m_{ij})^2}{2V^2_{\text{eff},ij}}\right) \tag{4.5}$$

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

$$\ln(p) = \ln\left(\frac{1}{1+\gamma^2}\right) + \ln\left(\frac{1}{A}\right), \tag{4.6}$$

$$\mathrm{LP} = \sum_{i,j} \ln(L_{ij}) + \ln(p). \tag{4.7}$$

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

4.3.3 Evaluation metrics

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

$$\mathrm{Completeness} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \qquad \mathrm{Purity} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}. \tag{4.8}$$

4.4 Variability and colour distributions

We present plots and statistics for two datasets: all fitted objects and those with spectroscopic classifications. In Sect. 4.5, we also assemble photometric selections. The datasets allow us to compare distributions of three classes: QSOs, stars, and galaxies. The plots include variability parameters (A and γ) and colour diagrams of $W2$ versus $W1-W2$ and $g-r$ versus $z-W1$. We know stars, galaxies, and QSOs to have good separation in optical versus IR colours (Maddox & Hewett 2006; Lang et al. 2016). $W1-W2$ was also used by Wright et al. (2010), and Assef et al. (2013) combined $W2$ and $W1-W2$ for selecting AGN.

We query and fit the data in batches as described in Sect. 4.2.1. For 6340 files with approximately 1 million light-curve data points per file, the batch processing time is about 1 hour on a computing node with 32 cores. The maximum time is 11 days for one file.

4.4.1 Full Palomar Transient Factory sample

In Fig. 4.1, we show parameter distributions for the full PTF sample. The top left panel is a 2D histogram of A and γ values, illustrating the variability of the 70 920 904 fitted light curves. We see two large clusters, at $\log A \sim -0.8$ and $\log A \sim -5$ (base 10). Variations of just 10^{-5} magnitudes over timescales of one year indicate that the values of the latter cluster are spurious.

The top right and bottom panels of Fig. 4.1 display a colour–colour and a colour–magnitude diagram. As these are based on parameters in WISE and PS1, they show distributions for the 80 % (56 991 591) of the PTF objects with matches in both of these surveys (see Table 4.1). Matches in the two surveys are found for 93 % of QSOs, 75 % of stars, and 98 % of galaxies.

4.4.2 Spectroscopically confirmed sample

Figure 4.2 contains similar plots to Fig. 4.1, but only for data with spectroscopic SDSS classes. In the variability parameter plot, this is 2.5 % of the fitted sources (1 748 047), and the colour diagrams show the 2.276 % (1 613 916) with matches in SDSS, WISE, and PS1. The colours denote different classes: red for QSOs, green for stars, and blue for galaxies. These are scaled based on relative density, with full saturation for areas of parameter space with one class and maximum relative density, white for areas without data and black or grey for areas with high relative densities of multiple classes. We note that this means a space that is more green than blue can still have more galaxies than stars. The full dataset includes more galaxies than stars and QSOs combined (see Sect. 4.2.2).

In the top left panel of Fig. 4.1, we notice the two clusters of A-γ space. The one at $\log(A) \sim -0.8$ has high relative densities of all classes, but they are spread out along different axes in the A-γ plane; more galaxies have $\log(A) < -1$ and more QSOs have $\gamma > 0.1$. Stars are mostly in the same areas as galaxies, but with more spread in γ and higher relative density at $\log(A) < -3$. In the top right panel, the stellar locus is immediately recognisable, as are the main distributions of QSOs and galaxies (Ansari et al. 2021; Lang et al. 2016). The latter two include an overlap most apparent at $g - r \sim$

0.4 and $z - W1 > 0.3$. The bottom panel shows that the three classes have different distributions in $W1 - W2$ versus $W2$ as well, but with a large overlap between stars and galaxies at $W2 > 15$.

4.5 Photometric selection

We next explore how variability relates to broad-band colours through photometric selections using colours, variability, or both. Simple photometric selections of QSOs, stars, and galaxies allow us to inspect differences between parameter distributions of photometrically and spectroscopically selected objects. Based on the distributions of Fig. 4.2, we define regions of parameter space dominated by QSOs, stars, or galaxies. Schmidt et al. (2010) used a similar approach in A and γ. For reliable selections, we focus on purity and accept a low completeness. We avoid areas of high degeneracy between classes and achieve separate selection criteria for variability parameters and for colours. All criteria are listed in Table 4.2. While these are simple, linear criteria in parameter space that can be used to study relations between colour and variability, more advanced techniques exist and have been deployed for example by Ansari et al. (2021) in colour space. A study of this kind is presented in a companion paper (Bruun et al. 2023b).

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

$$\frac{A}{\sigma_{A,-}} > 2, \qquad \frac{\gamma}{\sigma_{\gamma,-}} > 2. \tag{4.9}$$

4.5.1 Selection properties

The selection criteria of Table 4.2 are based on purity and completeness of the labelled objects in Fig. 4.2 and are then applied to the larger, unlabelled datasets of Fig. 4.1. The criteria are illustrated in Fig. 4.7 with those of Schmidt et al. (2010) for comparison.

Table 4.3 lists sample statistics of candidates selected within these parameter space regions of variability, colour, or both. The Schmidt criteria are included for comparison. To estimate completeness, colour-based selections are given as percentages of the total

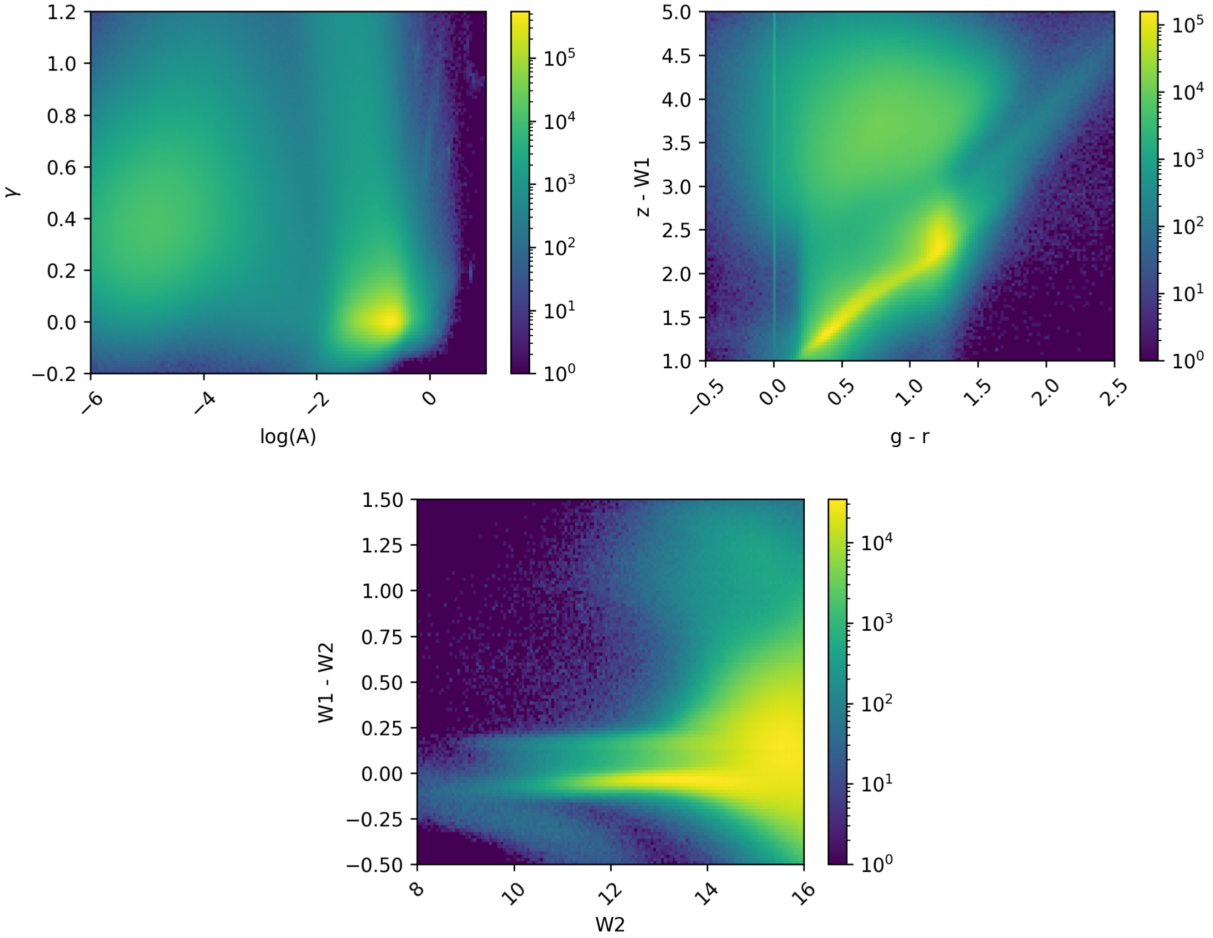


Figure 4.1: Heat maps of variability parameters (top left), $g - r$ vs. $z - W1$ (top right) and $W2$ vs. $W1 - W2$ (bottom). These are the full parameter distributions of objects with light curves in PTF after cleaning and matching with other surveys as necessary. Two large clusters are observed in A-γ-space. Selection criteria are applied to this data for analysis of candidate QSOs, stars and galaxies. $\log(A)$ is in base 10 and based on A units of in magnitudes.

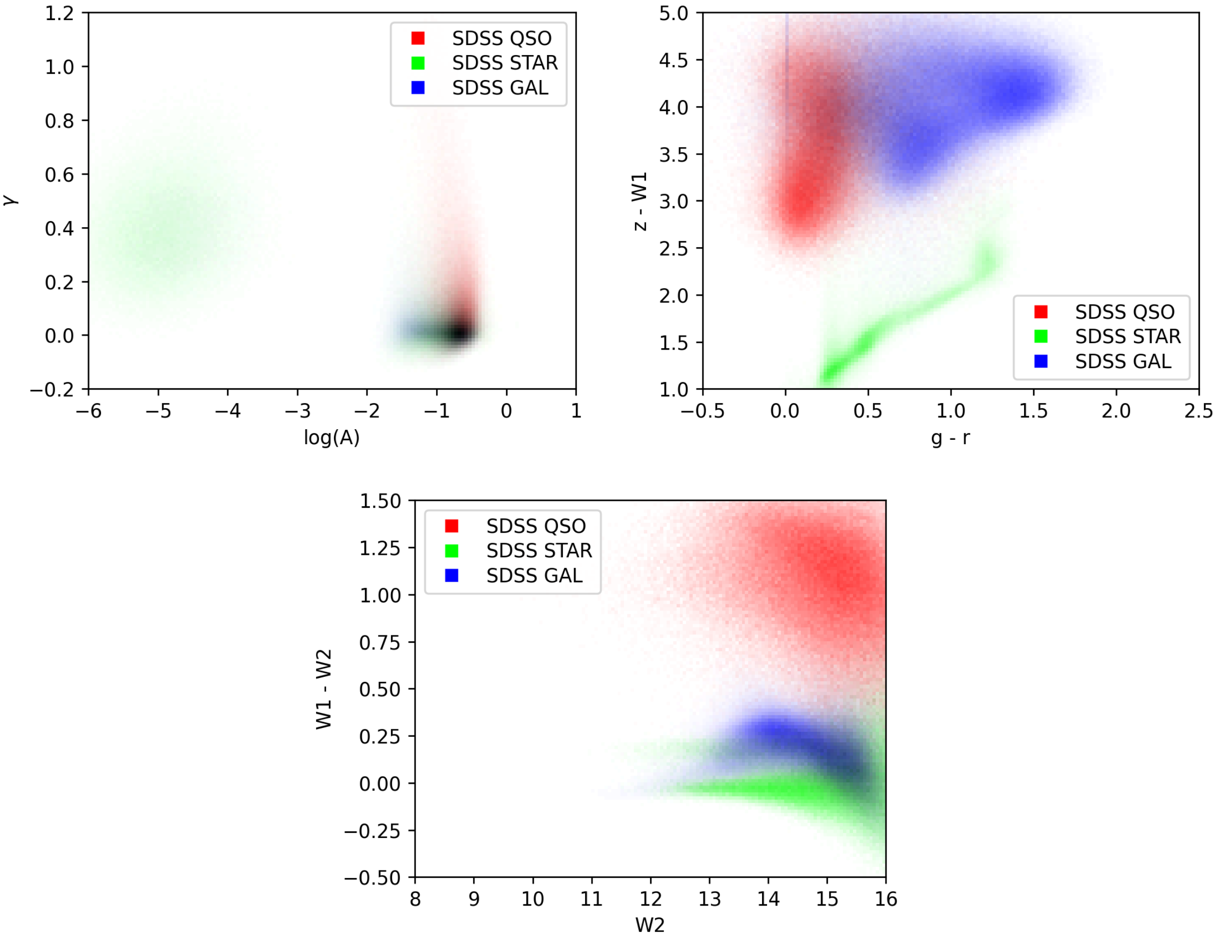


Figure 4.2: All spectroscopically confirmed stars (green), QSOs (red) and galaxies (blue) from SDSS plotted in variability fit parameters (top left) and colours: $g - r$ vs. $z - W1$ (top right) and $W2$ vs. $W1 - W2$ (bottom). Marker colours show the object class and blend to grey or black when multiple classes occupy the same parameter region. A heat map of all spectroscopically confirmed objects is found in Fig. 4.7 for comparison with Fig. 4.1.

Table 4.2: Selection criteria.

Parameter	QSO candidates	Star candidates	Galaxy candidates
γ	> 0.13	$\ldots$	< 0.1
$\log_{10}(A)$	> 0.11	> 0.01	$< 0.1 \wedge > 0.01$
$g - r$	< 0.2	$\ldots$	> 1
$z - W1$	> 2.8	$< 2 \vee < 1 + 1.25(g - r)$	$> 2.3 + 1.25(g - r)$
$W1 - W2$	> 0.75	< -0.05	$> 0.25 \wedge < 0.35$
$W2$	$\ldots$	< 15	< 15

Notes. Criteria for selection of candidate QSOs, stars, and galaxies. These are based on the SDSS distributions of Fig. 4.2 and the parameter regions are plotted in Fig. 4.7.

set of objects with colour matches in WISE and PS1. Using both colour and variability criteria, the selections have low completeness and are very pure. However, selection bias in the spectroscopic sample affects purity estimates (see Sect. 4.6.3). If we disregard the bias, a naive purity estimate for colour-selected QSO candidates is 98.7 %, assuming 15.5 % of all PTF light curves with colour matches are QSOs. We obtain a purity of 99.3 % for colour- and variability-selected QSOs, and of 59.3 % with just variability criteria (assuming 15.3 % QSOs in the latter case because WISE and PS1 matches are not required). Most candidates selected with colour and variability are registered with the same label as main type in SIMBAD; Fig. 4.13 shows statistics for each class for comparison.

Gaia DR3 (Gaia Collaboration et al. 2016, 2022) sources within one arcsecond of PTF sources have lower proper motions for QSO candidates. This is especially clear when colour information is included in QSO selection, as shown in Fig. 4.14. Objects with stellar colours and variability have the highest proper motions, as expected. Assuming a maximum of one *Gaia* match per source, only 3.7% of SDSS galaxies have a match in *Gaia* with a measured proper motion, and the fraction decreases to 0.5% for galaxy candidates selected by colour and variability. Further details are given in Appendix 4.8.5.

If instead we use the Schmidt criteria and require the sources to have A and γ significantly different from zero (Eq. 4.9), we find the most common SDSS class to be QSOs. However, these criteria lead to more contamination from stars and galaxies than the QSO criteria of Table 4.2.

4.5.1.1 Colour selection

Figure 4.3 shows distributions of colour-based selections from the large dataset of fitted sources in Fig. 4.1. The overall pattern is the same as in Fig. 4.2, but the stars are more spread out. More of them have very low A or high γ, and relatively few are found at $\log(A) \sim -0.8$ and $\gamma \sim 0$. According to Table 4.3, the colour-selected candidates have similarly high purities of $98.6\ \%{-}99.7\ \%$ for all classes, but with varying completeness, of namely 6.82 % for galaxies, 7.92 % for stars, and 38.3 % for QSOs.

4.5.1.2 Variability selection

We plot the variability-selected sources in Fig. 4.4. Relying solely on A and γ is challenging according to the SDSS labels. For example, while the variability criteria for stars catch 53.4 % of stars and 19.8 % of galaxies, they select more galaxies due to different population sizes. Naturally, criteria based on overlapping classes in A-γ-space lead to overlapping selections in colour space.

4.5.2 Ambiguous sources

Given the purity of the photometric selections in Table 4.3, the contaminating sources are expected to have a high rate of misclassification by SDSS. Checking the most confident photometric predictions with differing spectroscopic classifications, we find examples of spectra that appear to be more typical of the photometric candidate class. We inspect the spectra of the 32 photometric QSO candidates with stellar spectroscopic classifications, and judge $\sim$ 20 % to be QSOs and $\sim$ 50 % to be possible QSOs. Of these objects, 16 are in SIMBAD, and are registered as 8 QSOs, 4 BL Lacertae objects, 1 blazar, and just 3 stars. For example, SDSS J120429.34+495814.4, with the spectrum of Fig. 4.8, has the characteristic broad lines of a QSO. This object is part of the Data Release 12 Quasar Catalog from SDSS (Paris et al. 2017), but is spectroscopically confirmed as a star in SDSS DR17. This misclassification analysis shows the power of the photometric selection criteria of this work, although we do not expect high rates of misclassification in the full SDSS spectroscopically confirmed sample.

Variable galaxies in the Table 4.2 QSO region of A and γ have higher redshifts, $W2$, and $z - W1$ than other galaxies (typically redshift 0.5 vs. 0.1; see Figs. 4.9 and 4.10) . In SIMBAD, variable galaxies are more often registered as the brightest galaxy in a

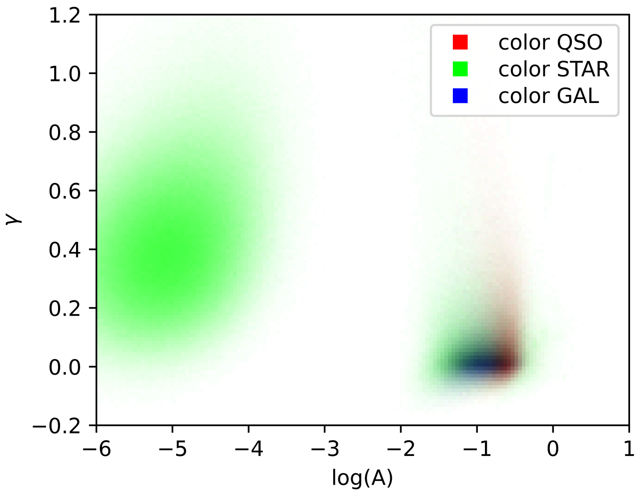


Figure 4.3: Variability of objects of Fig. 4.1 selected by the colour criteria from Table 4.2. These are based on the colour distributions of Fig. 4.2.

cluster than other galaxies are. The fraction increases from 9% to 15 %. There is also a low but relatively much higher fraction of radio sources, which goes from 1.6% to 3.7 %. SDSS QSOs have similar redshifts independently of variability, except more non-QSO-like variability at $z < 0.3$ (see Fig. 4.9).

Spectroscopic QSOs variability-selected as stars or galaxies are more often found at $W1 - W2 < 0.75$ and $W2 < 14$ than other QSOs. However, this difference in WISE colours is smaller than for spectroscopic galaxies and stars in different photometric classes, as illustrated in Figs. 4.10 – 4.12.

4.6 Discussion

Quasars, stars, and galaxies are distributed differently in A and γ and in colours, but with overlaps limiting the quality of simple selections. We see this in differences between the SDSS class distributions of Fig. 4.2 and the corresponding candidate class plots in Figs. 4.3 and 4.4 based on colour and variability, respectively. The overall patterns can still be recognised. Class information in the SF variability parameters is especially interesting for objects without spectroscopy and limited colour measurements. Variability alone does not give candidates that are as reliable as those from colour selection, but colour and variability combined can break degeneracies and select classes

Table 4.3: Selection statistics.

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

Notes. Sample statistics of selections in variability and colour, including counts, completeness, and purity (see Eq. 4.8). In row 1 (All), we compare spectroscopically classified objects to the full fitted source count and to the full set with spectroscopic classes to get relative frequencies in each set. In rows 5-7 and 11, for computing purity and completeness, selection counts are compared to the total spectroscopic counts. In rows 2-4 and 8-10, comparisons also require matches in WISE and PS1, because colours are used.

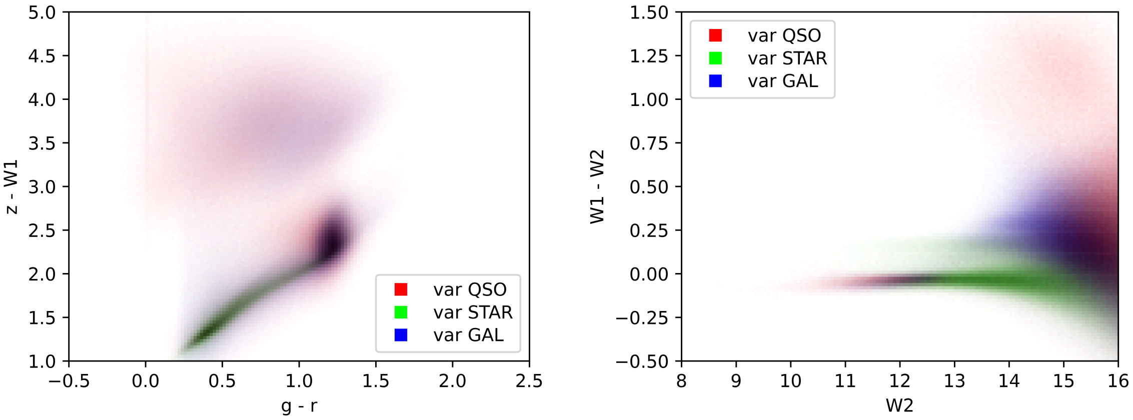


Figure 4.4: Colour distributions of variability-selected objects from Fig. 4.1. These fulfil the variability criteria of Table 4.2 based on the upper diagram in Fig. 4.2.

better than each of them separately.

4.6.1 A-γ clusters

In A-γ space, we see two large clusters. One of them is likely an artefact from objects with undetectable variability, leading the MCMC to suggest arbitrarily small A and a large range of γ values. There is only one clear cluster at higher A ($\log(A) \sim -0.8$), but the QSOs and galaxies are spread along different axes. This is most apparent in Fig. 4.2. The galaxies cover a broad range of A, but their variability is rarely timescale dependent, at least not on most relevant scales. This is shown by the low γ values. In contrast, QSOs are even more variable (high A) with a clearer timescale dependence. Stars are spread out more evenly in A and γ, limiting selection purity and reflecting the diverse nature of stellar variability.

4.6.2 Comparison with SDSS light-curve analysis

The selection criteria of Table 4.2 are simple and focus on purity. For QSOs, they differ from those by Schmidt et al. (2010). Table 4.3 shows a purer QSO set with slightly lower completeness, than if we apply the Schmidt criteria. This may be due to slightly different fitting and outlier removal or differences in noise and measurements between PTF and SDSS. The dataset presented in this paper is more representative of all observed object types, as it includes all sources from the PTF survey that pass through the data cleaning of Sect. 4.2.5, whereas Schmidt et al. (2010) introduced specific types of contaminants and in specific ratios.

4.6.3 Spectroscopic selection bias

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

There are differences in the distributions of sources with and without spectroscopic classes. This is especially clear if we compare Fig. 4.1 to Fig. 4.7. For example, SDSS is missing spectra for objects at low $W2$, including a band of sources at low $W1 - W2$. At $10 < W2 < 12$ and $-0.5 < W2 - W1 < -0.25$, we have a very mixed group, with sources typically marked as stars (21 %) or binary star systems in SIMBAD.

SDSS also has a bias in favour of sources with $z - W1 > 3$. This is part of the reason why in Fig. 4.4, QSOs and galaxies are relatively infrequent at those values compared to Fig. 4.2. If more objects are included in star-dominated areas, they also include a larger fraction of the galaxies and QSOs. However, many of these are actually stars incorrectly selected according to variability. Variable stars late in the main sequence can be especially difficult to distinguish from QSOs and galaxies; the parameter region at $2.1 < z - W1 < 2.6$ and $1.1 < g - r < 1.4$ is dominated by stars for all three variability selections. Variability-selected star candidates in the area are at least 86 % stars judging by the most common stellar classifications registered in SIMBAD ('Star', 'low-mass*', and 'PM*'). Galaxy and QSO candidates are at least 79 % and 72 % stars, respectively, showing a small difference in the nature of these objects. For spectroscopic stars, those with QSO- or galaxy-like variability are more spread out in $z - W1$ and found at higher values of $W1 - W2$. This is illustrated in Fig. 4.12. The variability does indicate a physical difference in these cases.

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

SDSS-matched sources have relatively long time spans and more epochs per light curve, as shown in Fig. 4.6. Hence, variability estimation of the full sample might be less accurate, spreading out sources in A-γ space. Therefore, removing the most sparsely sampled sources is important. To both include large datasets and be confident in the results, the balance of data cleaning will also be important in future surveys such as the LSST. Even for sources with >100 epochs over >5 years, the spectroscopic classes still have an overlap at $\log A \sim -0.8$ and $\gamma \sim 0$, but it is about 50 % smaller in both $\log A$ and γ. Longer timescales may change the selected populations, for example by increasing the fraction of type II to type I AGN (De Cicco et al. 2015, 2019).

4.6.4 Photometric selection bias

We select pure sets of each class, but with low completeness and a bias for sources with specific parameters. With variability, we only select stars with low A; we know variable stars exist, but they are difficult to isolate, and so reduce completeness for stars and purity for galaxies and QSOs. Type I and Type II AGN differ in colour and SF, and so are not best selected by one simple set of criteria either (De Cicco et al. 2022). The most densely populated variability region, at $\log A \sim -0.8$ and $\gamma \sim 0$, is not covered by the criteria at all. The same goes for the dense colour diagram area at high $W2$. In Fig. 4.2, galaxies dominate a triangular area of $g - r$ versus $z - W1$ with two large clusters. The criteria only cover one of them. In SIMBAD, the cluster at high $g - r$ has more galaxies labelled as being part of a cluster, and especially as the brightest galaxy in a cluster. A more advanced selection method could solve these issues (Bruun et al. 2023b).

The surveys are not completely representative of the sky, which is not observed uniformly (see Fig. 4.5). We include fewer stars at low Galactic latitudes, where Galactic extinction has a greater affect on colours and there is a higher risk of mismatches with nearby sources. Stars have the fewest colour matches, indicating an under-representation in WISE or PS1. A change in ratios of object types and stellar subtypes affects purity. The reason is that the number of true or false positives depends on the selection of objects that are not equally difficult to distinguish from or as stars. For example, including more variable stars could lower the QSO selection purity, which is due to both the overlap with QSO variability and the increased prior probability of a classified object being a star. However, including more data would also mean there is more data to learn from. The balance of object-type frequencies is relevant if we apply Table 4.2 criteria to other datasets, and in evaluation of criteria of Schmidt et al. (2010) on PTF data. Completeness is computed independently for each class, but could also be affected by a change in the fractions of subtypes.

4.6.5 Perspectives

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

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

4.7 Conclusion

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

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

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

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

Acknowledgements

This work was supported by a Villum Investigator grant (project number 16599). SHB was also supported by a grant from the Danish National Research Foundation. AA was also supported by a Villum Experiment grant (project number 36225).

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

Aeronautics and Space Administration.

The Pan-STARRS1 Surveys were made possible through contributions by the Institute for Astronomy, the University of Hawaii, the Pan-STARRS Project Office, the Max-Planck Society and its participating institutes, the Max Planck Institute for Astronomy, Heidelberg and the Max Planck Institute for Extraterrestrial Physics, Garching, The Johns Hopkins University, Durham University, the University of Edinburgh, the Queen's University Belfast, the Harvard-Smithsonian Center for Astrophysics, the Las Cumbres Observatory Global Telescope Network Incorporated, the National Central University of Taiwan, the Space Telescope Science Institute, and the National Aeronautics and Space Administration under Grant No. NNX08AR22G issued through the Planetary Science Division of the NASA Science Mission Directorate, the National Science Foundation Grant No. AST-1238877, the University of Maryland, Eotvos Lorand University (ELTE), and the Los Alamos National Laboratory.The Pan-STARRS1 Surveys are archived at the Space Telescope Science Institute (STScI) and can be accessed through MAST, the Mikulski Archive for Space Telescopes. Additional support for the Pan-STARRS1 public science archive is provided by the Gordon and Betty Moore Foundation.

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

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

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

SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, Center for Astrophysics | Harvard & Smithsonian, the Chilean Participation Group, the French Participation Group, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU) / University of Tokyo, the Korean Participation Group, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatário Nacional / MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University, and Yale University.

This research made use of the cross-match service provided by CDS, Strasbourg.

This research made use of Astropy,[6] a community-developed core Python package for Astronomy (Astropy Collaboration et al., 2013; Astropy Collaboration et al., 2018).

This research has made use of the NASA/IPAC Infrared Science Archive, which is funded by the National Aeronautics and Space Administration and operated by the California Institute of Technology.

This research makes use of the SciServer science platform (www.sciserver.org).

SciServer is a collaborative research environment for large-scale data-driven science. It is being developed at, and administered by, the Institute for Data Intensive Engineering and Science at Johns Hopkins University. SciServer is funded by the National Science Foundation through the Data Infrastructure Building Blocks (DIBBs) program and others, as well as by the Alfred P. Sloan Foundation and the Gordon and Betty Moore Foundation.

This research has made use of the VizieR catalogue access tool, CDS, Strasbourg, France (DOI : 10.26093/cds/vizier). The original description of the VizieR service was published in 2000, A&AS 143, 23.

This work has made use of data from the European Space Agency (ESA) mission *Gaia* (`https://www.cosmos.esa.int/gaia`), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, `https://www.cosmos.esa.int/web/gaia/dpac/consortium`). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement.

[6] http://www.astropy.org

4.8 Appendices

4.8.1 Appendix A: Sky maps

The sources with PTF light curves are mainly from the northern hemisphere. These are distributed according to the top diagram of Fig. 4.5 after the data cleaning of Sect. 4.2.5. When we also match to SDSS spectroscopic classifications, the sources in the bottom diagram are left. The differences indicate a bias in photometry between the surveys, as mentioned in Sect. 4.6.4.

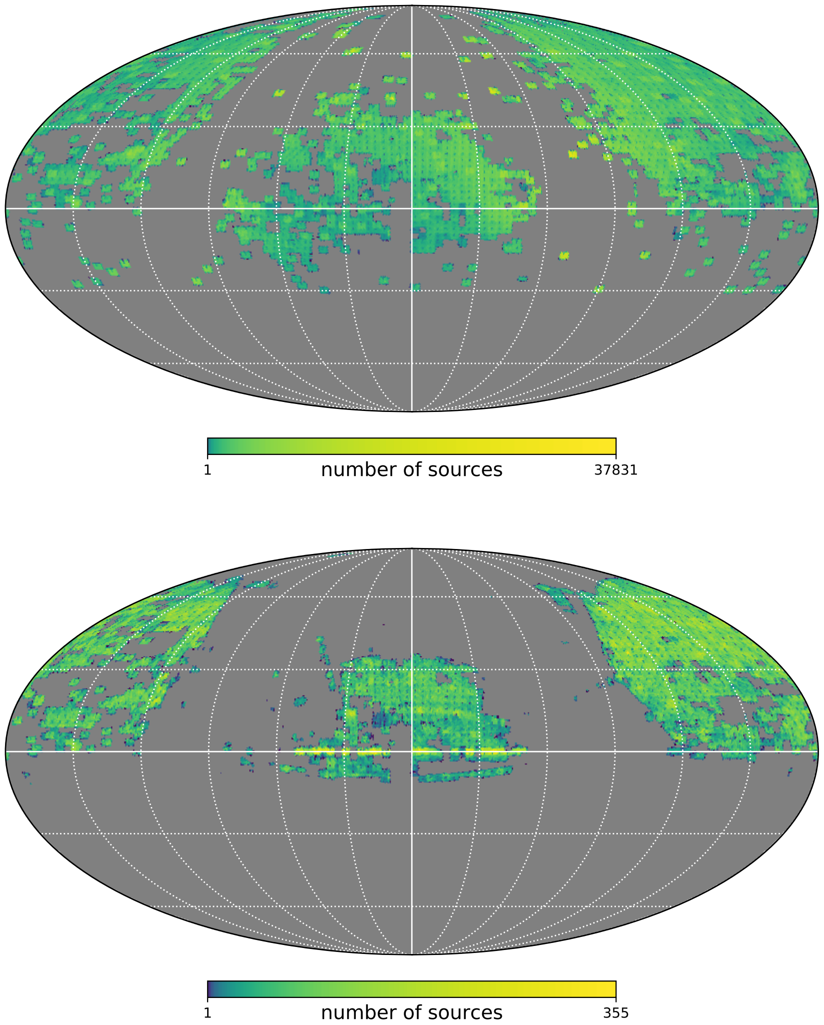


Figure 4.5: Sky distributions of fitted and accepted PTF light curves (top) and of those also matched to spectroscopic classifications in SDSS (bottom). The coordinates are equatorial, with RA increasing to the right. In grey areas, no data exist or remain after the data cleaning of Sect. 4.2.5.

4.8.2 Appendix B: SDSS matched data

Of the fitted PTF sources, 2.5 % are matched to spectroscopic classifications in SDSS. This subset has different parameter distributions from the full PTF sample. On average, the time spans are longer and the number of epochs is higher, as shown in Fig. 4.6. The distributions of variability parameters A and γ, and colours in the optical and mid-IR are shown in Fig. 4.7 for SDSS matched data for comparison with the full sample in Fig. 4.1. The differences are discussed in Sect. 4.6.3. Figure 4.7 also illustrates the selection criteria for stars, galaxies, and QSOs in Table 4.2 and the criteria of Schmidt et al. (2010):

$$\gamma > 0.055 \tag{4.10}$$

$$\gamma > 0.5\log_{10} A + 0.5 \tag{4.11}$$

$$\gamma > -2\log_{10} A - 2.25. \tag{4.12}$$

These criteria can point to potential misclassifications in SDSS, as discussed in Sect. 4.5.2, and an example of this is SDSS J120429.34+495814.4. This object is registered as a star in SDSS, but the selection criteria of this paper point to it being a QSO. The spectrum in Fig. 4.8 has the broad emission lines characteristic of QSOs.

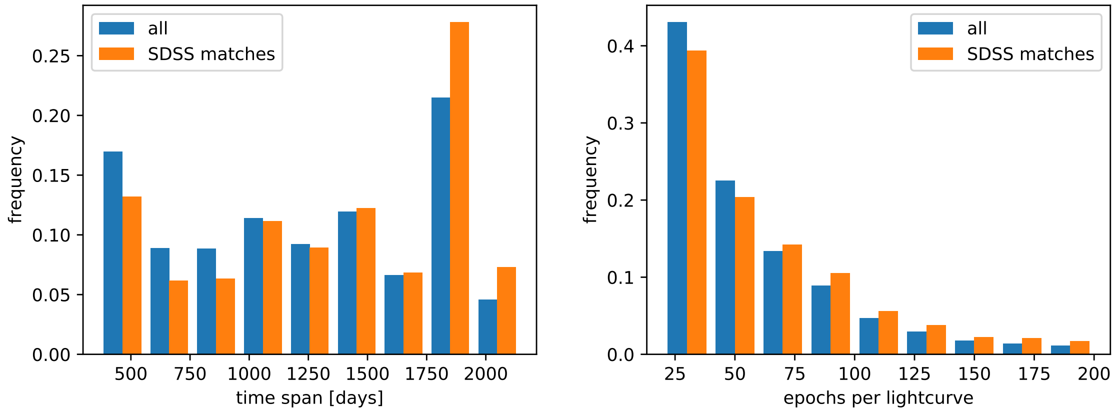


Figure 4.6: Histograms of all PTF sources and sources matched to spectroscopic classifications in SDSS. The differences in time span and number of epochs show two ways in which the SDSS matched data are not representative of the entire PTF sample.

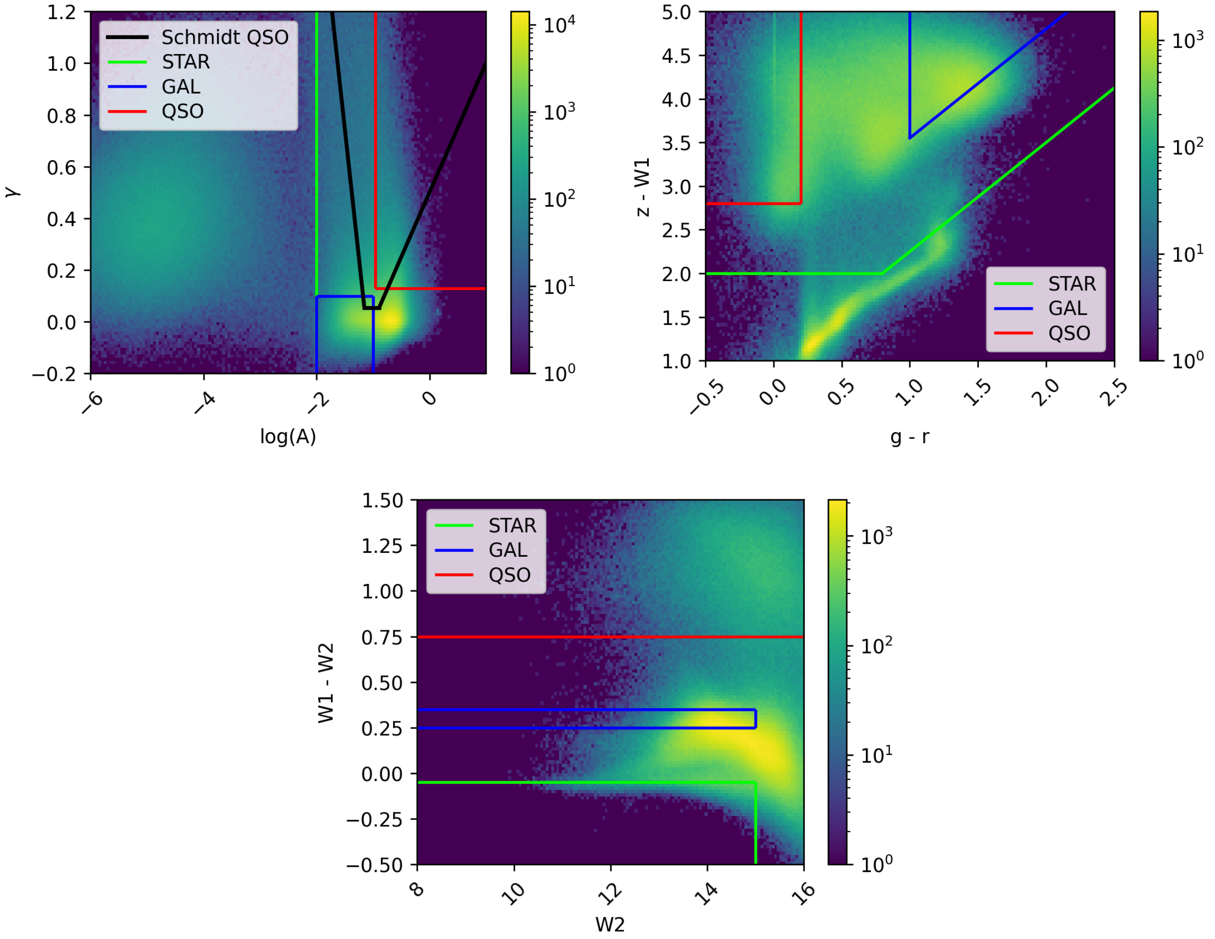


Figure 4.7: Heat maps of all data with spectroscopic classifications in SDSS. Comparing this with Fig. 4.1, we see SDSS has spectroscopic data focused on specific parts of the parameter space. Strict criteria for pure selection of stars (green), QSOs (red), and galaxies (blue) are overplotted. These are listed in Table 4.2. The fit-parameter panel also includes a black line showing the Schmidt et al. (2010) QSO selection.

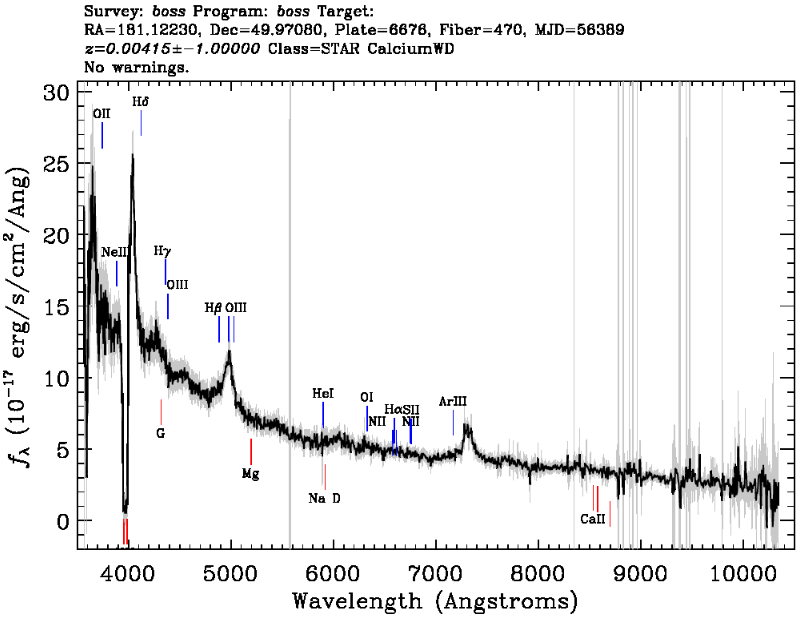


Figure 4.8: Spectrum of SDSS J120429.34+495814.4 from the BOSS spectrograph. The source is classified as a star in SDSS-IV DR17, but it has the variability and colour parameters of a QSO candidate. The broad emission lines support the latter classification. This is one of 32 objects selected as a QSO candidate and whose SDSS spectra were classified as stars, of which 20% – 50 % are QSOs based on visual inspection of the spectra. (SDSS-IV DR17, CC-BY license, skyserver.sdss.org/dr17/VisualTools/explore/summary?objId=1237658613058109587.)

4.8.3 Appendix C: Variability selected objects

The photometric selections are performed based on either the variability criteria, the colour criteria, or both of Table 4.2. Some spectroscopically confirmed objects have conflicting photometric parameters, as discussed in 4.5.2. In Fig. 4.9 the redshifts for spectroscopic galaxies are typically higher for variability-selected galaxy candidates than for star candidates. QSOs have similar variability except at $z < 0.3$ where more appear like stars or galaxies in A and γ.

Figures 4.10–4.12 show colour diagrams for all combinations of spectroscopic and variability selected classes. This illustrates how the variability selection criteria are picking objects with different colour distributions, even when the objects are spectroscopically confirmed to belong to the same class. This is discussed further in Sects. 4.5.2 and 4.6.3.

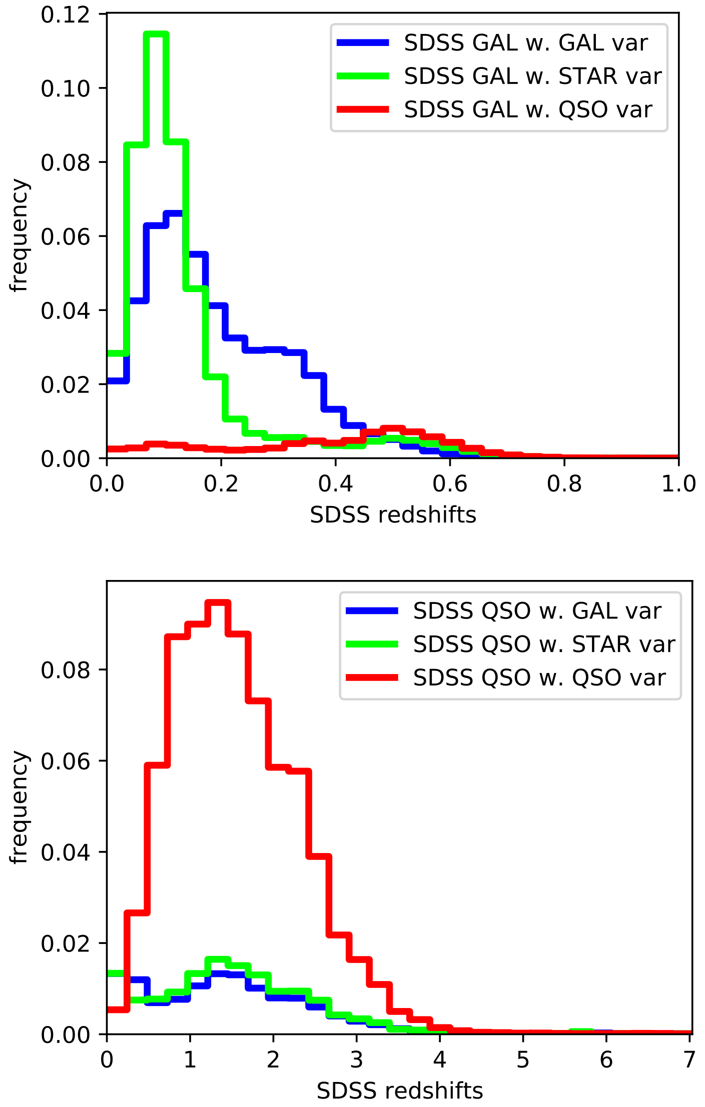


Figure 4.9: Redshift distributions of galaxies (top) and QSOs (bottom) that appear as candidates of different classes based on variability criteria. The galaxies with QSO-like high variability are typically found at high redshifts of ~ 0.5, and those with star-like low variability have lower redshifts than other galaxies. SDSS QSOs are dominated by QSO-like variability except at $z < 0.3$. The bins of each diagram are normalised with respect to the total samples of spectroscopic SDSS galaxies and QSOs, respectively.

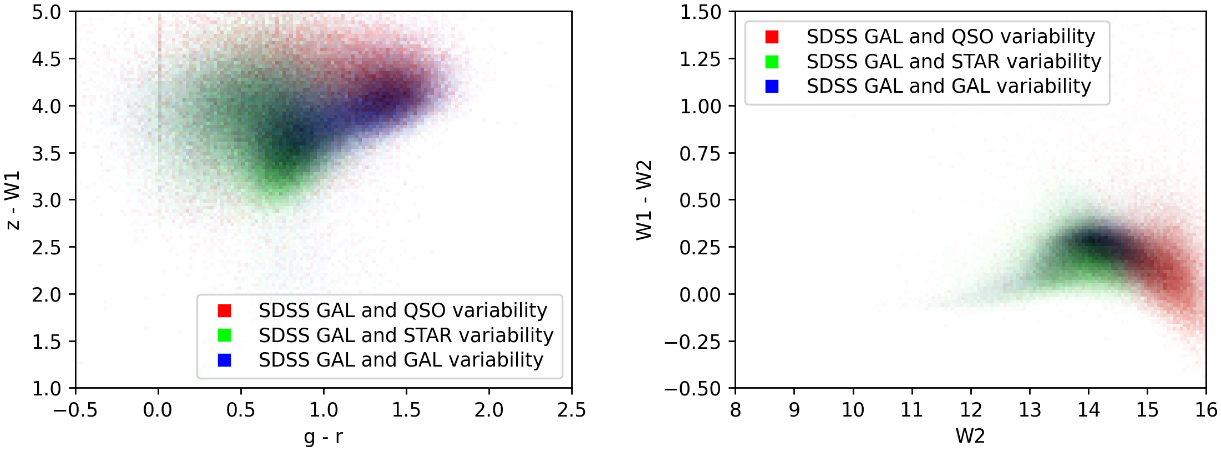


Figure 4.10: Galaxies with variability parameters typical of different types of sources. We see that the more variable galaxies (with QSO variability) have high $W2$ and $z - W1$.

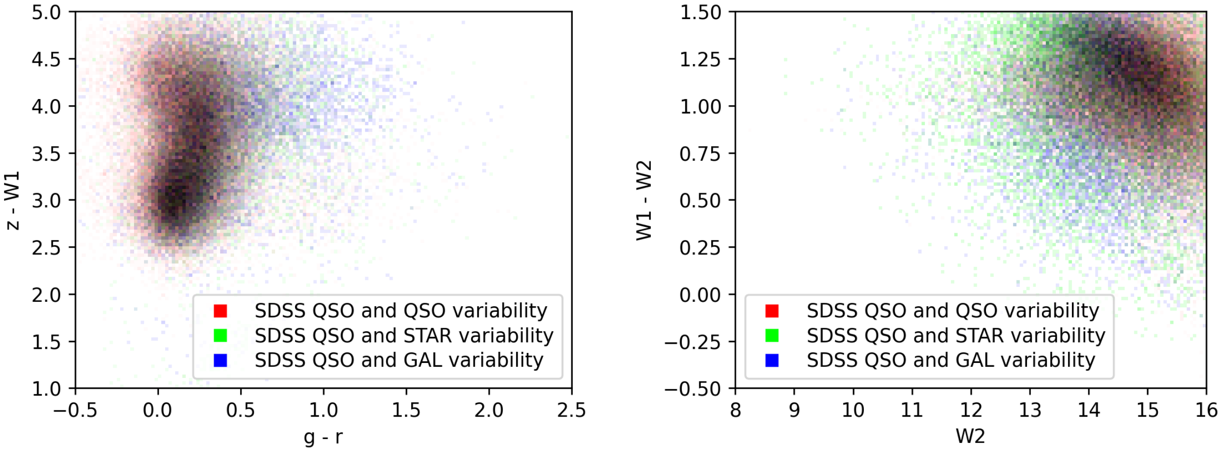


Figure 4.11: Quasars with different variability parameters are still mostly found in the same regions of the colour diagrams.

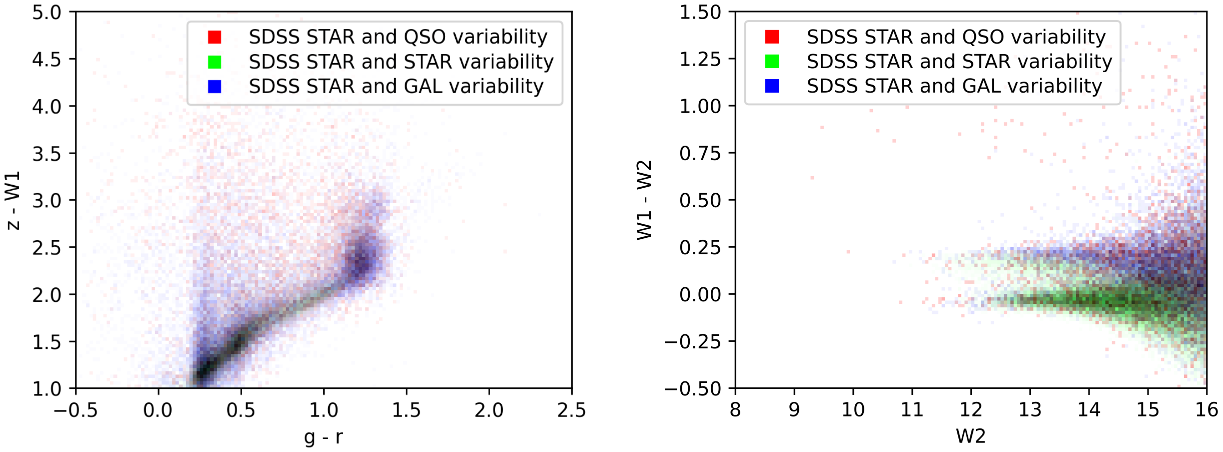


Figure 4.12: Stars are in the stellar locus of the upper panel, but with more spread for those with variability that is more typical of QSOs and galaxies. Galaxy-like variability is also found at higher $W1 - W2$ than for most stars. Based on inspection of SDSS imaging, the 73 SDSS stars with galaxy-like colours mostly resemble galaxies or a star–galaxy chance alignment. SDSS stars with $0.2 < W1 - W2 > 0.3$ and $W2 < 13$ are also often not isolated. We do not see this for random subsets of all of the 12 470 SDSS stars with galaxy-like variability.

4.8.4 Appendix D: SIMBAD statistics

Candidate stars, QSOs, and galaxies are selected based on the criteria in Table 4.2. To understand the physical nature of the objects, we looked up the statistics of their 'main type' in SIMBAD. Most photometrically selected objects were registered with the an expected label in SIMBAD, as shown in Fig. 4.13.

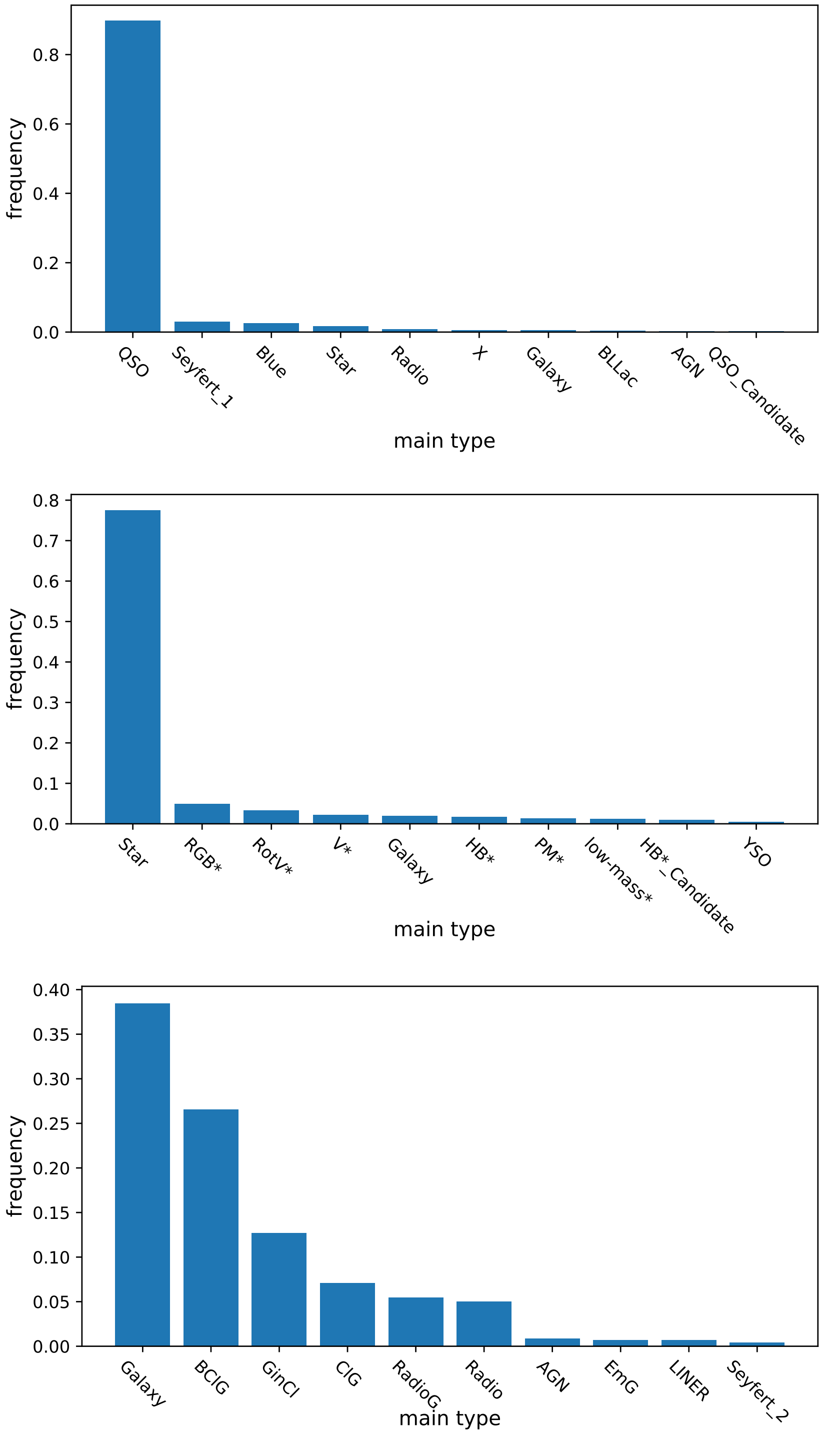


Figure 4.13: SIMBAD 'main type' statistics for candidates by colour and variability. This shows the subtypes and contaminating sources for QSO (upper panel), star (middle), and galaxy (lower) candidates.

4.8.5 Appendix E: *Gaia* proper motions

We cross match with *Gaia* DR3 (Gaia Collaboration et al. 2016, 2022) proper motions using the CDS X-Match Service (Boch et al. 2012). This is done for all sources within one arcsecond of the PTF coordinates. For those selected through variability, we limit the search to 1 million random sources of each class. In Fig. 4.14 we show histograms of the sources with measured proper motions. As expected, these are generally lower for QSOs and higher for stars. The differences are smaller for variability-selected objects and they follow the distribution of spectroscopically confirmed stars almost perfectly, indicating that most of the objects with non-zero proper motions are stars. However, the objects with QSO-like variability have a higher spectroscopic purity (59.3 %) in the SDSS. Sources selected by both colour and variability have distributions closer to those of the spectroscopically confirmed classes, especially for QSOs. We note that the sample sizes are also different, as not all sources have a match in *Gaia* and multiple *Gaia* sources can be detected per PTF source. For variability-selected objects, the *Gaia* sets are 60.6 %, 92.0 %, and 72.0 % of the sizes of the QSO, star, and galaxy sets, respectively. For sources selected based on variability and colour, these numbers are 84.7 %, 99.6 %, and 0.5 %, and for spectroscopically confirmed sources, they are 82.3 %, 95.9 % and 3.7 %. *Gaia* has proper motions for almost all stars and very few galaxies, as its design is optimised for detecting stars and the limit for proper-motion measurements is not as faint as the PTF or Pan-STARRS source-detection depth. As is shown by the comparison of SDSS and *Gaia* purity of the variability-selected QSOs, atypical proper motions may indicate misclassifications and mismatches between surveys, or interesting subclasses or flaring objects. The sum of the residuals between histograms of variability-selected QSOs and variability-selected stars (-0.01 for stars to calibrate the curves at high proper motions), when the total area under the curve is normalised to 1, is 0.29. This gives a lower bound on the QSO purity, in addition to the higher spectroscopic purity, as it does not take into account differences in how many objects are matched, chance alignments of stars, and so on. (Even SDSS QSOs usually have a proper motion measured nearby in *Gaia*.)

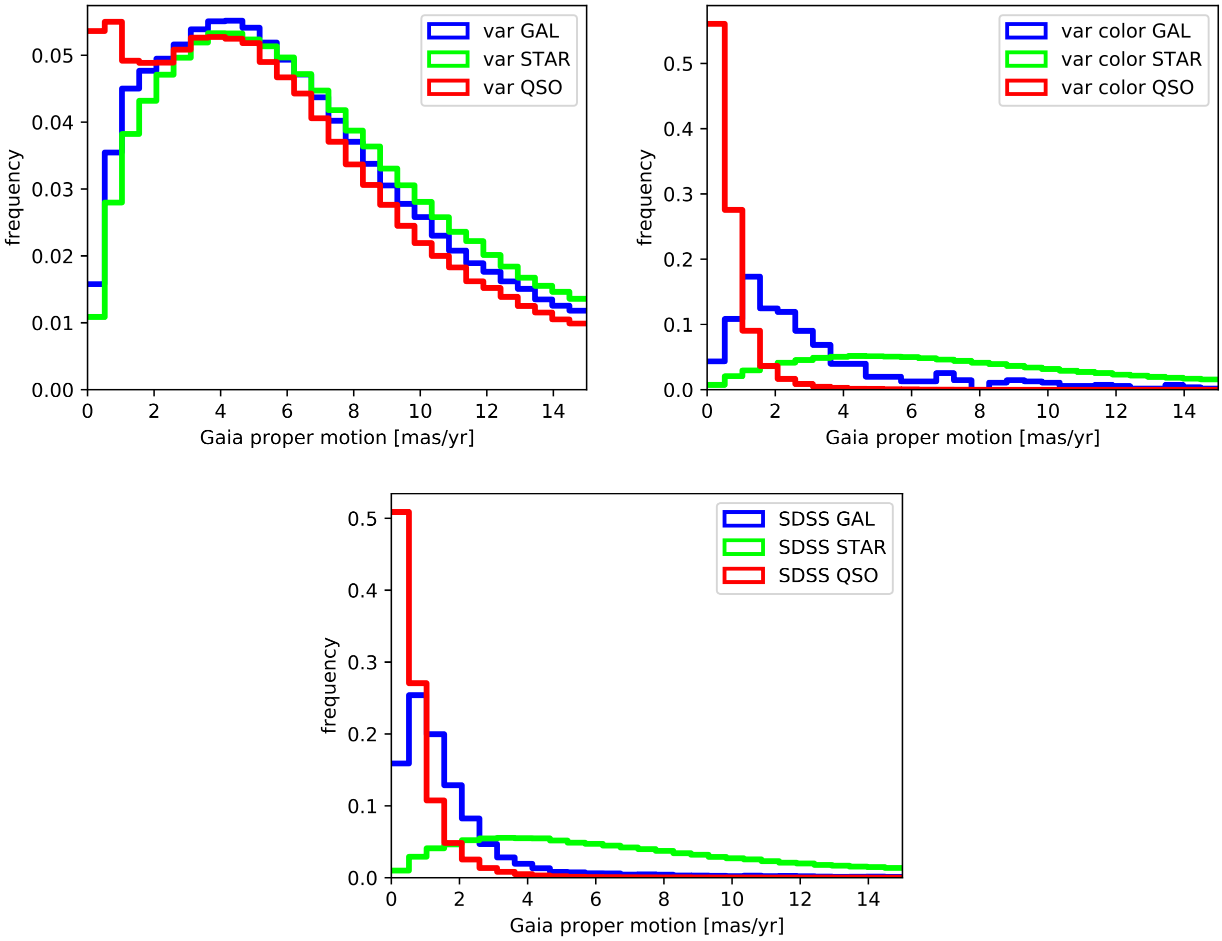


Figure 4.14: *Gaia* proper motions for sources selected according to variability (top left), colour and variability (top right), and spectroscopy (bottom). The histograms are normalised individually. All variability-selected sources approximately follow the distribution of SDSS stars, but when we also select by colour, the results are generally close to those of the spectroscopically confirmed objects. A high-purity and high-completeness cut at $\lesssim 2$ mas/y, justified by the bottom panel, would yield $\approx 50\%$ purity in a sample of QSO candidates selected based on variability only.

Chapter 5

Supervised classification of variable sources

This chapter is based on the following article:

VarIabiLity seLection of AstrophysIcal sources iN PTF (VILLAIN) II. Supervised classification of variable sources

Authors: Sofie Helene Bruun, Jens Hjorth and Adriano Agnello

In review at Astronomy & Astrophysics; submitted on 11/04/2023. arXiv: 2304.09905

Author contributions

SB led the project, performed the data analysis and chose the methodology of this chapter. SB wrote and revised the manuscript which has benefited from multiple rounds of comments and suggestions from all authors. The main idea of classifying sources with variability and colours was proposed by AA and JH. SB was the corresponding author with support from AA and JH.

Abstract

Context. Large, high-dimensional astronomical surveys require efficient data analysis. Automatic fitting of light-curve variability and machine learning may assist in identification of sources including candidate quasars.
Aims. We aim to classify sources from the Palomar Transient Factory (PTF) as quasars, stars or galaxies, and to examine model performance using variability and colours. We determine the added value of variability information as well as quantifying the performance when colours are not available.
Methods. We use supervised learning in the form of a histogram-based gradient boosting classifier to predict spectroscopic SDSS classes using photometry. For comparison, we create models with structure function variability parameters only, magnitudes only and using all parameters.
Results. We achieve highly accurate predictions for 71 million sources with light curves in PTF. The full model correctly identifies 92.49 % of spectroscopically confirmed quasars from the SDSS with a purity of 95.64 %. With only variability, the completeness is 34.97 % and the purity is 58.71 % for quasars. The predictions and probabilities of PTF objects belonging to each class are made available in a catalogue, VILLAIN-Cat, including magnitudes and variability parameters.
Conclusions. We have developed a method for automatic and effective classification of PTF sources using magnitudes and variability. For similar supervised models, we recommend using at least 100 000 labeled objects, and we show how performance scales with data volume.

5.1 Introduction

Machine learning has gained increasing importance in astronomy (Smith & Geach 2022; Ball & Brunner 2010; Djorgovski et al. 2022). In the era of large astronomical surveys, automatic classification is necessary for fast and reliable processing of sources, and for detecting patterns in high-dimensional data. Detailed observations including spectroscopy are expensive for large datasets, so high-quality photometric classification is needed for future surveys such as the Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST; Ivezić et al. 2019).

Quasars and active galactic nuclei (AGN) can be identified based on their variability. Variability is especially important for classification of AGN in low mass galaxies, as they are often missed using emission line ratios. This is due to low metallicities (weakening the [N II] line), different AGN fuelling mechanisms or "star formation dilution" – dilution of the optical emission line signature of AGN by H II region emission lines in galaxies with high star formation rates (Baldassare et al. 2020; Trump et al. 2015; Groves et al. 2006). Treiber et al. (2022) selected candidate AGN using optical variability of 142 061 light curves from the Transiting Exoplanet Survey Satellite (Ricker et al. 2014) using parameter cuts and visual inspections. Out of the 29 AGN candidates, 8 are in low mass galaxies, but the method would be impractical for larger datasets.

Various machine learning models have been applied to photometric data in search of quasars and AGN candidates. De Cicco et al. (2021) selected optically variable AGN candidates with a Random Forest algorithm (Breiman 2001) on measures that will available with the LSST: light curves and colours in the optical and near-infrared. Palanque-Delabrouille et al. (2011) used neural networks to distinguish high-redshift quasars and stars in Stripe 82 from the Sloan Digital Sky Survey (SDSS; Abazajian et al. 2009) and in simulated Palomar Transient Factory (PTF; Law et al. 2009; Rau et al. 2009) data. Incorporating variability parameters improved the selection purity for quasars at high redshifts and quasars with broad absorption lines.

Cunha & Humphrey (2022) applied gradient boosting algorithms for SDSS class prediction on SDSS and WISE photometry. The model performed much better using combinations of magnitudes to create colours and photo-z predictions than without the combinations. The photo-z model was trained to predict spectroscopic redshifts from SDSS.

In this paper, we aim to create a large catalogue of candidate quasars, stars and galaxies, and to analyse the importance of monochromatic light-curve variability compared to the importance of optical and infrared magnitudes and colours using a machine learning model. We use the full set of light curves from PTF fitted with power-law variability models by Bruun et al. (2023a); paper I of the VILLAIN project.

In Sect. 5.2 we define the machine learning models, how they are selected and the training strategy. The section includes definitions of model inputs, preprocessing to create colours from magnitudes etc., and we define the metrics for model evaluation. An overview of the results including model performances is presented in Sect. 5.3. In Sect. 5.4, we discuss the results, biases and perspectives for the use of the model predictions and adaptions of the method for different contexts.

5.2 Method

To define a model that can classify objects, we need a set of objects to learn from. In the present paper, we assume spectroscopically confirmed classifications in SDSS to be the ground truth. Each object is classified by SDSS as either a quasar, star or galaxy, which are the "labels" of the model. To learn how to reproduce the labels without spectroscopy, we build a model to guess the labels using data from PTF, Wide-field Infrared Survey Explorer (WISE; Wright et al. 2010) and Pan-STARRS1 (PS1; Chambers et al. 2016). We use a set of 70 920 904 PTF sources with power law fits of their structure functions. The PTF sources have been cross-matched to sources in the WISE, PS1 and SDSS, as described by paper I.

We refer to each object in the dataset – and all parameters associated with it – as a "sample", and each sample corresponds to a PTF object. A sample is described by a number of properties called "features" which are used as inputs for the machine learning model, e.g. magnitudes. We can split the set of samples into a labeled set and an unlabeled set. 2.5 % of PTF sources have matches in SDSS PhotoObj and SpecObj, resulting in a labeled set of 1 747 471 sources.

A supervised machine learning algorithm takes input features X and predicts labels Y_{pred} after learning on "true" labels Y_{true}. The input "true" labels are spectroscopic classifications from SDSS. The model only learns from the labeled subset, but is able to predict labels for every source. While fitting (also called "training") the model to data, model performance is quantified by a loss function, which measures dissimilarity between the true labels and predicted labels. The loss is minimised to improve classification.

It is possible to learn from the unlabled subset as well. An unsupervised machine learning model would learn patterns in X and could group sources with similar parameters into "clusters". To learn from both unlabeled data and the labels, when they are available, semisupervised learning is preferable. One form of semisupervised learning is self training which iteratively learns from the labeled samples and assigns probable pseudo-labels to unlabeled samples. Due to the large dataset of this project, self training did not improve performance (F1 scores as described in Sect. 5.3.1). We therefore proceed with supervised learning.

5.2.1 Models

Machine learning models learn parameters from the input data, but they also have hyperparameters to control the learning process. To choose an algorithm and its hyperparameters, testing multiple models is beneficial. To compare the performance of models with and without variability and magnitude information, we also create multiple final models for different datasets using different features. The datasets include samples with missing values in WISE and PS1, so for fair evaluation of the importance of colours, we create different models for the full dataset and for the "matched" dataset. The six models use:

1. AllA: **All** features and **A**ll data (only PTF light-curve variability and SDSS matches are required)
2. AllM: **All** features and only data with **M**atches in WISE and PS1 (so their magnitudes are always included)
3. VarA: Only **Var**iability features and **A**ll data
4. VarM: Only **Var**iability features and only data with **M**atches in WISE and PS1
5. MagA: Only **Mag**nitude features and **A**ll data
6. MagM: Only **Mag**nitude features and only data with **M**atches in WISE and PS1

We create additional models for AllA using random subsets of varying size to analyse the importance of data volume.

5.2.2 Data splitting

The goal is to create a model that can reliably classify data it has not seen before. It should give predictions based on general trends in the data and avoid overfitting to random patterns in a specific dataset even if it can perfectly replicate SDSS labels in data it was trained on. To get an unbiased measure of model performance, we set aside part of the labeled data as a test set for evaluation. The model is only tested on the test set once, and we do not examine its properties prior to this test, except size and number of matches in WISE and PS1, to avoid any leakage of information from the test set into the model.

During model optimisation, we train the model multiple times and determine when it is ready to run on the test set. We evaluate performance on the training set, but this is prone to overfitting. For a less biased estimate, we use a validation set. While trying different models and tuning hyperparameters, we evaluate on the validation set. Generally, the performance will seem best on the training set and worst on the test set. Using the test set multiple times would likely give seemingly better scores as well, but it would be a biased estimate of performance on new data. For further discussion of best practices in machine learning, see Lones (2021) and Kapoor & Narayanan (2022).

Before testing, we could combine the training and validation sets to train one last time on more data. We choose not to combine them, because the new model randomly could be worse (and we would not know without a separate validation set) and because some models are sensitive to the size of the dataset. For small datasets, the advantage of learning from more data would outweigh this concern.

We shuffle the data and use a 60-20-20 split for creating a training set, a validation set and a test set. This is done both for the full dataset and the matched subset. We make sure that the training set for the matched subset, is also a subset of the training set for the full dataset, and likewise for validation and test sets.

5.2.2.1 *Cross-validation*

An alternative splitting technique for training and validation is cross-validation (Stone 1974), which allows the use of all data for both purposes. In n-fold cross validation, the data is split into n folds. Alternately, $n-1$ folds are used for training and one for validation. The average score of the validation folds is then used for evaluation – a process which reduces variance compared to a single validation set. We use cross-validation within the training set for less overfitting, but still keep extra validation and test sets, since we run the cross-validation multiple times for adjusting hyperparameters. The balance of class frequencies (fraction of samples of each class) in the dataset will affect the model. During cross-validation, we ensure all folds are representative of the full dataset by using stratified folds, meaning that the class balance of the full set is preserved in each fold.

5.2.2.2 Early stopping

Reducing overfitting on the training set will likely lead to better performance on the test set, because when the model mistakenly thinks it perfectly understands the training data, it will stop learning. We use several methods to avoid overfitting. One of them involves data splitting, namely early stopping.

To stop the algorithm before it overfits, we take out part of the training set as a validation set for evaluating performance after every step. If performance does not improve by more than a predefined tolerance, training is stopped. We implement the early stopping and cross-validation in Sect. 5.2.3.2.

5.2.3 Classification algorithm

Astronomical datasets often include objects with missing measurements of some properties. To analyse a dataset with missing feature values, one could drop objects with missing features or drop features that are not available for all objects. Another option is to guess the missing features (imputation). To learn from all data without dropping or guessing information, we use a model that by construction accepts and learns from missing values (Josse et al. 2019; Twala et al. 2008).

For fast classification with low memory usage and including non-linear patterns, we choose `HistGradientBoostingClassifier`, which is a tree based model from scikit-learn (Pedregosa et al. 2011). Tree based models classify data using hierarchical tree-structures. A decision tree places "nodes" which split the feature space, and the nodes are used one by one to decide which class a sample belongs to. The final nodes, which are not split further, are called leaf nodes, while the internal nodes are called branch nodes. Sources are classified based on which leaf they belong to.

A single decision tree is simple and easy to interpret, however, ensemble tree models will usually perform better. One way to improve the outputs of a tree f_i, is to add another tree h_{i+1} that predicts how the first one could be improved, and create a new ensemble model f_{i+1}. Gradient Boosting Machines sequentially add trees in this manner. The loss function L is minimised to improve classification, using the gradients $G = \nabla L(Y, f(X))$ and hessians $H = \nabla^2 L(Y, f(X))$ of L with respect to the model $f(X)$. First, constant initial predictions C are chosen. Then, a tree $h_1(X)$ fits the gradients of the constant model $f_0 = C$, thereby predicting a correction that can be added to the initial predictions.

The next tree h_2 predicts gradients of the updated model f_1 and so forth. This continues for max_iter boosting iterations with the final predictions $F(x)$ including all correcting trees multiplied by the learning rate, η (Friedman 2001):

$$F(X) = C + \eta \sum_{i}^{max\_iter} h_i(X). \tag{5.1}$$

The learning rate is a regularisation parameter that prevents overfitting by preventing the model from learning too much from a single tree.

5.2.3.1 Histogram-based Gradient Boosting Classification Tree

`HistGradientBoostingClassifier` is a scikit-learn implementation of gradient boosting similar to LightGBM (Ke et al. 2017). In this tree-based ensemble method, each sample is processed by a series of trees. In each tree, the sample is assigned to a leaf which has a leaf weight. The leaf weights can then be summed for prediction of the class of the sample. To convert the sum to a class for multi-class prediction, a tree is created for each class during each iteration. The values can then be compared following a one versus rest (OvR) approach in which every class is compared against all other classes. Softmax normalisation converts the values to probabilities.

The trees are estimating corrections to previous trees, so the leaf weights depend on gradients and hessians of the loss function of a tree i:

$$w_i = -\frac{G_i}{H_i + \lambda}. \tag{5.2}$$

Here λ is a regularisation parameter used for penalising complex models to avoid overfitting, similarly to η. It is also part of the loss function. For multiclass classification, we define the loss function L as the categorical crossentropy between a tree model f and "true" labels Y plus a regularisation term Ω for each tree h_j. L is computed for N samples as:

$$L(Y, f(X)) = -\frac{1}{N} \sum_i \sum_k Y_{i,k} \ln(f_{i,k}(X_i)) + \sum_j \Omega(h_j), \tag{5.3}$$

$$\Omega(h_j) = \frac{1}{2}\lambda ||w_j||^2, \tag{5.4}$$

where $f_{i,k}$ is the predicted probability for a sample i of belonging to class k, and $Y_{i,k}$ is the "true" one-hot-encoded value (0 or 1) we are trying to predict. For regularisation, λ penalises trees with high $||w_j||^2$ to avoid large contributions from individual trees (Chen & Guestrin 2016).

When the model decides where to split the data (creating nodes) for each tree, `HistGradientBoostingClassifier` speeds up the process by binning the feature space instead of sorting all feature values. The algorithm creates histograms of G and H for all samples in the feature bins, and uses the histograms to decide at which bin edge to split. A split optimises the gain of the left and right nodes,

$$\text{Gain} = \frac{1}{2}\left(\frac{G_{\text{left}}^2}{H_{\text{left}}+\lambda} + \frac{G_{\text{right}}^2}{H_{\text{right}}+\lambda} - \frac{(G_{\text{left}}+G_{\text{right}})^2}{H_{\text{left}}+H_{\text{right}}+\lambda}\right). \tag{5.5}$$

To include samples with missing values, one bin is dedicated to this, which is a "missingness incorporated in attributes" strategy (Twala et al. 2008). At every split, the model learns which nodes to assign sources with missing values to. This strategy takes into account the information on whether data is missing or not, and allows a model trained on all features of Sect. 5.2.4 to classify sources even if they have no colour information.

5.2.3.2 *Hyperparameter tuning*

For fast hyperparameter tuning, we use successive halving iterations (Jamieson & Talwalkar 2016; Li et al. 2016). The `HalvingRandomSearchCV` class in scikit-learn randomly chooses hyperparameter values in specified ranges and examines performances using 5-fold cross-validation. The model is evaluated with a low number of samples at first. Then the best performing hyperparameters are selected, and they are evaluated with a higher number of samples until the best hyperparameter combination is determined. Performance in `HalvingRandomSearchCV()` is evaluated with the macro averaged F1 score as defined in Sect. 5.3.1.

We use `HistGradientBoostingClassifier` with early stopping, which sets aside 10 % of the data for validation, and stops the fitting if the last 30 models did not improve the loss function by more than 10^{-7}. With max_iter at 200, we test learning rates from 0.01 to 0.3, maximum number of leaf nodes from 11 to 81, minimum samples per leaf from 5 to 200 and l2 regularisation parameters (λ in Eq. 5.4) from 0 to 5.

5.2.4 Preprocessing

The SDSS labels are quasar, star and galaxy, and we choose to encode them as 0, 1 and 2, respectively, in the Y_{true} vector. The model converts Y_{true} to $Y_{i,k}$ of Eq. 5.3. We choose input features that are suitable for class prediction by being measured independently of the spectroscopic classes (unlike e.g. spectroscopic redshift). We include variability parameters A and γ, $W1$ and $W2$ from WISE, g, r, i, z and y from PS1 and median R from PTF. In tree based models, scaling the features would also scale the positions of tree nodes in feature space. This does not affect performance, so the models are insensitive to monotonic feature transformations, and the features will not need scaling.

5.2.4.1 *Feature engineering*

Features may interact in ways that are useful for classification. To make it easier for a model to learn such interactions, we construct additional features by combining existing ones. We create colour features and features based on the selection criteria of paper I and Schmidt et al. (2010).

In paper I, manually chosen selection criteria are effective in selecting pure sets of each class. The criteria are applied in $\log(A)$ vs. γ, $z - W1$ vs. $g - r$ and $W2$ vs. $W1 - W2$. Here $\log(A)$ is in base 10. The criteria are linear in said parameter spaces. For easier separation of classes, we construct four features $z - W1 - 1.25(g - r), W1 - W2 - 0.017W2, \gamma + 0.5\log(A)$ and $\gamma - 2\log(A)$. We also combine all eight magnitude features for 28 colour features and thereby use a total of 42 features. For non-linear feature construction, the features could be combined by multiplication as well, but this did not give an improvement compared to models with only linear constructed features.

5.2.5 Performance evaluation

To measure multiple qualities of the predicted outputs, the model performance is evaluated with multiple metrics. Mean accuracy (the fraction of correct predictions) is simple, but highly dependent on class ratios. The macro-averaged scores are the unweighted means of the scores for each class and assume equal importance of each class.

5.2.5.1 *F1 score*

F1 is the harmonic mean of completeness and purity (Dice 1945; Sørensen 1948), which are assumed equally important. They are computed from the number of true positives (TP), false negatives (FN), and false positives (FN):

$$\text{Completeness} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{5.6}$$

$$\text{Purity} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{5.7}$$

$$\text{F1} = 2\frac{\text{Completeness} \times \text{Purity}}{\text{Completeness} + \text{Purity}}. \tag{5.8}$$

5.2.5.2 *ROC AUC*

ROC AUC is the Area Under the Curve (AUC) for the Receiver Operating Characteristic (ROC; Fawcett 2006). A ROC curve evaluates the trade-off between a high true positive rate (TPR, completeness) and low false positive rate for a binary classifier. The false positive rate, FPR, is

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \tag{5.9}$$

where TN are the true negatives. The classifier of this project is multinomial, so we compute the ROC AUC for each class with an OvR approach.

For the average ROC AUC, to avoid the influence of class imbalances, the macro average is taken following a one versus one approach (OvO; Hand & Till 2001):

$$M = \frac{2}{c(c-1)} \sum_{i<j} \hat{A}(i,j), \tag{5.10}$$

where c is the number of classes and $\hat{A}$ is a measure on i and j – in this case $\hat{A}$ is ROC AUC. This takes into account that each class can have better separation from one class than the other.

Table 5.1: Classifier hyperparameters.

Model features	Variability		Colours		All features	
Samples	All	Matched	All	Matched	All	Matched
λ	0.918	2.38	3.06	2.41	4.03	2.63
η	0.0462	0.0215	0.0498	0.0415	0.142	0.0653
Max leaf nodes	11	23	45	63	58	61
Min samples per leaf	41	37	7	19	6	19

Notes. Hyperparameters of the histogram-based gradient boosting classifier are selected as described in Sect. 5.2.3. Here "colours" include all magnitudes and colours.

5.3 Results

We tune the models in Sect. 5.2.1 to the training set and validate on the validation set with the chosen hyperparameters listed in Table 5.1. Afterwards, we run them once on the test set for each final model. This is to test performance, classify all samples and predict probabilities of being a quasar, star or galaxy.

5.3.1 Performance

In Table 5.2, the models are listed along with several test statistics described in Sect. 5.2.5. Performance is measured on the test set and is slightly lower than on the training and validation sets due to overfitting, as expected. The "matched" samples are matched in both PS1 and WISE, so some "unmatched" samples do have colours and magnitudes.

Fig. 5.1 shows test set ROC curves for each class in two models using all samples. The model using all features (AllA) shows almost perfect performance, while the model with just variability (VarA) has a greater trade off between a high TPR and low FPR. This is also reflected in the macro averaged ROC AUC scores of 0.9945 for AllA and 0.7671 for VarA. Both models are much better than random guessing, since the ROC AUC are larger than 0.5 for all classes in Table 5.2.

5.3.1.1 *Baseline*

In Table 5.2 we include a baseline model that randomly assigns labels in a uniform manner. Random labels result in completenesses of 0.33 and purities equal to the frequencies of each class (but sensitive to differences between the training and test sets). The macro

Table 5.2: Performance statistics on the test set.

Model features	Variability		Colours		All features		Uniform baseline	
Samples	All	Matched	All	Matched	All	Matched	All	Matched
ROC AUC								
OvO macro avg	0.7671	0.7995	0.9945	0.9962	0.9945	0.9962	0.50	0.50
Quasars	0.8182	0.8251	0.9932	0.9951	0.9931	0.9951	0.50	0.50
Stars	0.7274	0.7864	0.9969	0.9976	0.9969	0.9976	0.50	0.50
Galaxies	0.6997	0.7268	0.9954	0.9956	0.9953	0.9957	0.50	0.50
F1								
macro avg	0.5431	0.5693	0.9624	0.9733	0.9639	0.9738	0.3039	0.2966
Quasars	0.4383	0.4462	0.9379	0.9570	0.9404	0.9580	0.2094	0.2086
Stars	0.4195	0.4596	0.9642	0.9755	0.9656	0.9758	0.2679	0.2361
Galaxies	0.7714	0.8022	0.9853	0.9875	0.9857	0.9878	0.4345	0.4423
Completeness								
Quasars	0.3497	0.3515	0.9220	0.9505	0.9249	0.9511	0.33	0.33
Stars	0.3308	0.3864	0.9702	0.9722	0.9716	0.9725	0.33	0.34
Galaxies	0.8682	0.8769	0.9871	0.9900	0.9875	0.9903	0.33	0.33
Purity								
Quasars	0.5871	0.6107	0.9543	0.9636	0.9564	0.9650	0.1527	0.1528
Stars	0.5732	0.5671	0.9582	0.9788	0.9597	0.9790	0.2240	0.1822
Galaxies	0.6940	0.7392	0.9834	0.9850	0.9839	0.9852	0.6232	0.6613
Mean accuracy	0.6687	0.7064	0.9734	0.9806	0.9744	0.9810	0.33	0.33

Notes. Classification performance of on the test set including all samples or only those matched in WISE and PS1. Performance improves with more input features, especially "colours" including magnitudes. The baseline is described in Sect. 5.3.1.1.

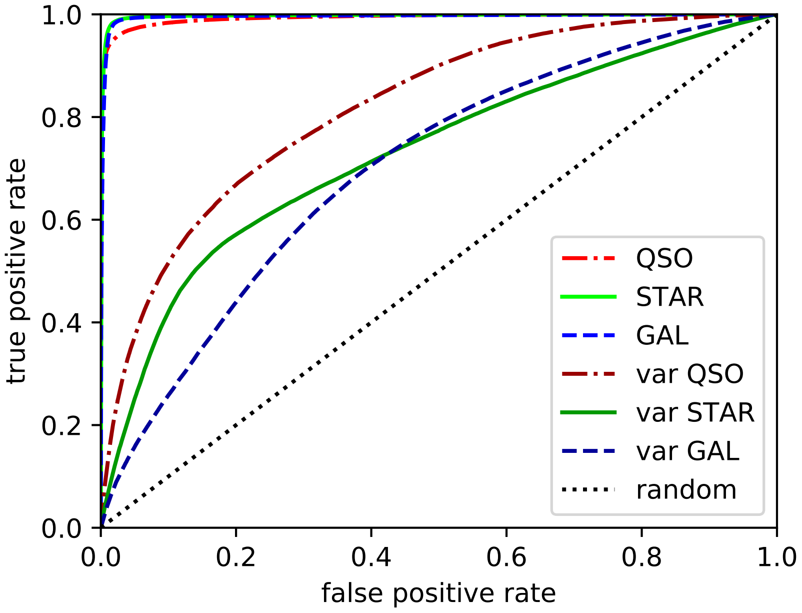


Figure 5.1: ROC curves for the model with all features and the model with just variability ("var" in the legend) on all data. Including colours makes it possible to achieve both very high true positive rates and low false positive rates for both quasars (QSO), stars (STAR) and galaxies (GAL). The dotted line shows the performance of random guessing which is worse than for all trained models.

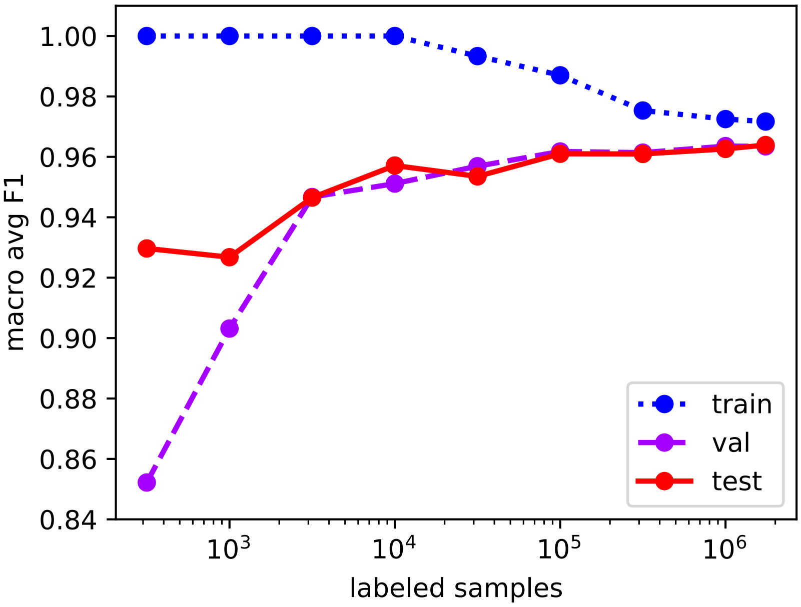


Figure 5.2: The macro averaged F1 score rises with the number of labeled samples for the validation and test sets. We see a clear increase up to ∼100 000 labeled samples, and gap between performance on training and test data continues to narrow until we reach best model on the full labeled dataset of 1.7 million samples.

averaged F1 score is 0.30, which is much lower than for all trained models.

Another baseline could be labelling everything as the most frequent class: galaxies. Completeness is then 0 for quasars and stars, and 1 for galaxies. Galaxy purity is 0.62 (their frequency), and this gives an F1 score of 0.77 for galaxies. We get undefined purities and F1 scores for quasars and stars, but if we set the undefined F1-scores to 0, the macro averaged F1 score is 0.26 using all data.

All models perform better than the baseline models. For some metrics of some classes, VarA only matches the baselines – but it simultaneously performs better for other classes with the same metrics.

5.3.1.2 *Number of samples*

To evaluate how many samples are needed in future surveys, we create new models using all features on random subsets of the full dataset. We estimate performance using macro averaged F1 scores. We still use a 60-20-20 split and choose random samples within the original training, validation and test sets.

In Fig. 5.2, we see how a larger training set leads to better performance. For less than 10^4 labeled samples, the training set (<6 000 samples) is completely overfitted with a perfect F1 score of 1. For larger samples, F1 decreases for the training set and increases for the validation and test sets, except for a few cases due to randomness in which sources are included. Randomness also leads to much better scores on the test set than on the validation set for 316 and 1000 samples. For 316 samples, the main difference is that test set includes stars that are easier for the model to classify. With 1 000 samples, the model performed better on both the galaxies and stars in the test set and it contained fewer quasars. Larger partitions for validation and test could reduce the uncertainty at the cost of less training data. Very similar scores on the validation and test set for the full model indicates that the amount of overfitting is similar. It indicates that the validation set could have been used more during model selection and hyperparameter tuning without becoming too biased.

The greatest improvements are for seen up to $\sim 10^5$ labeled samples. Even with just 316 labeled samples, the F1 score on the test set is 0.93. We created a model with 100 labeled samples as well, but it only predicted galaxies although the training set included all classes.

5.3.2 Calibration

Fig. 5.3 shows residuals of the calibration curves for each class for the models using all features or only variability on the full dataset with SDSS labels (models AllA and VarA in Sect. 5.2.1). The probabilities of belonging to each class sum to one for every object, as the classes are mutually exclusive. Well calibrated classifiers result in probabilities close to the dotted line where a predicted probability of e.g. 60 % of being a quasar means that 60 % of sources with that probability prediction are actually labeled quasars by SDSS. The bins contain 5 % of the samples each and are not of equal width. The largest residual shows predicted stellar probabilities of ∼50 % being just 1.5 percentage points too low. As the models are well-calibrated, we do not apply further calibration. The other models predict probabilities of similar quality to AllA and VarA.

For model AllA, the predicted probabilities of being a quasar are in the range 0.00003 −99.994 % (over both the train, val and test set). They are 0.0007−99.987 % for stars and $0.007 - 99.9993$ % for galaxies. With just variability, the VarA model is not as confident,

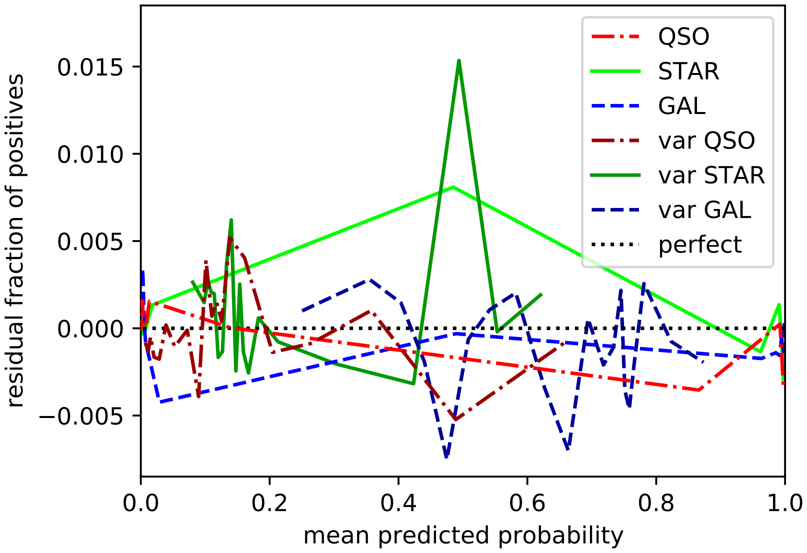


Figure 5.3: Calibration plot for the model using all features and all data. As the predictions are very well calibrated, the y axis shows the residuals after subtracting fractions of a perfect model. With only variability, the model is less confident so the predicted probabilities are closer to 50 % than for models including colours.

predicting quasar probabilities of 3—80 %, 5—90 % for stars and 13—90 % for galaxies.

5.3.3 Predicted parameter distribution

In Fig. 5.4, the final predictions of AllA are colour-coded with saturation based on the relative density of each class in the shown parameter space. The plots include the full dataset of 70 920 904 samples. The classifier of Sect. 5.2.3 selects sources using positions in parameter space in a manner similar to that of paper I – but optimised automatically to include even the samples with parameters that are not typical of a single class which are therefore difficult to classify (found in grey areas of Fig. 5.4).

5.3.3.1 *Variability classification*

Classifying with just the variability parameters A and γ (and features combining them) allows us to plot the complete feature space of a model. Fig. 5.5 shows $\log(A)$ vs. γ and how VarA separates the classes. Most of the feature space is predicted to contain galaxies. The regions assigned to each class can be compared to those of paper I – but

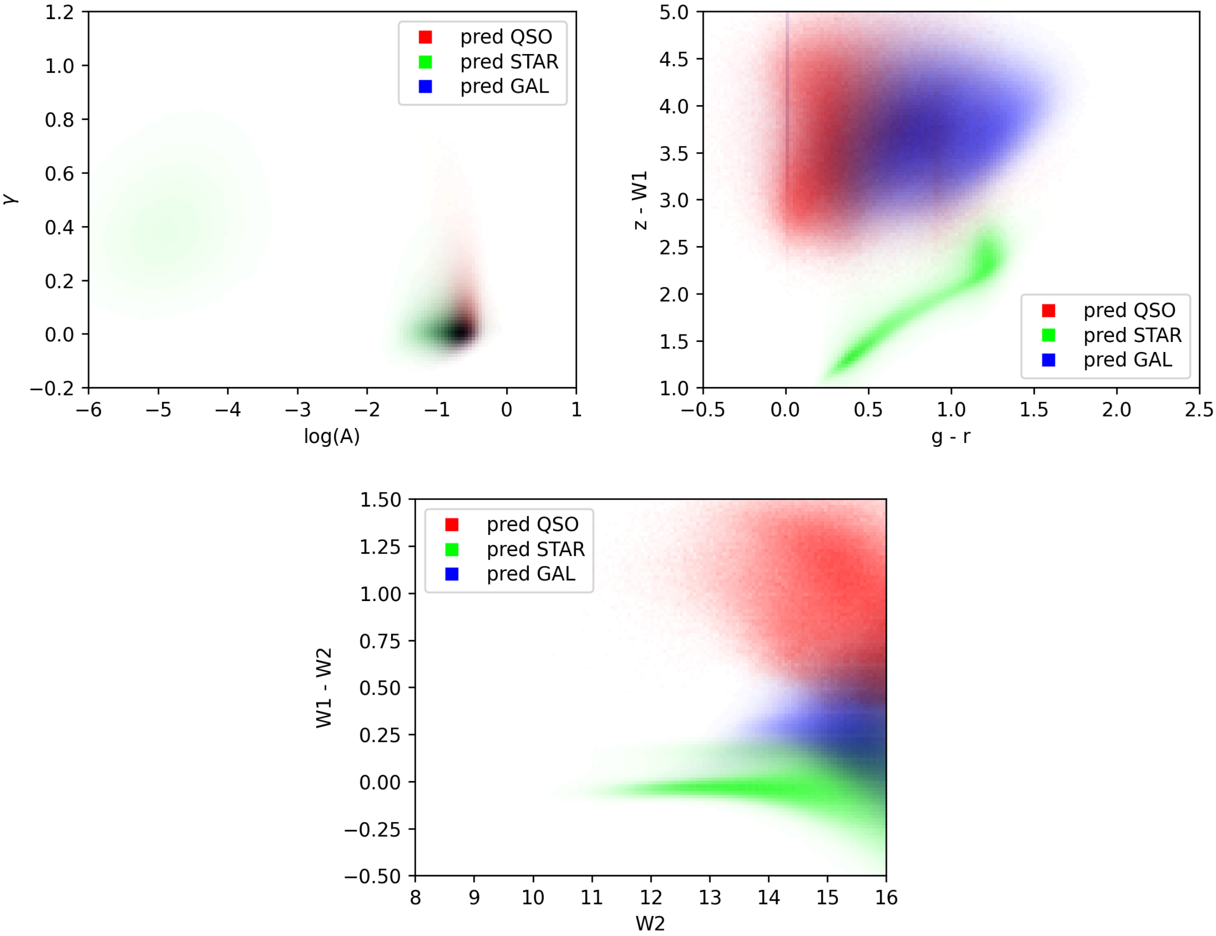


Figure 5.4: Predicted classes of all samples using all features, projected in three different parts of the input feature space. With the variability parameters A and γ (top left) an area at $\log_{10}(A) \sim -0.8$ and $\gamma \sim 0$ has high degeneracy, while especially stars and quasars are easier to distinguish in other areas. In $g - r$ vs $z - W1$ (top right) and $W2$ vs. $W1 - W2$ (bottom), the model shows greater separation of the classes. These plots include both training, validation and test data.

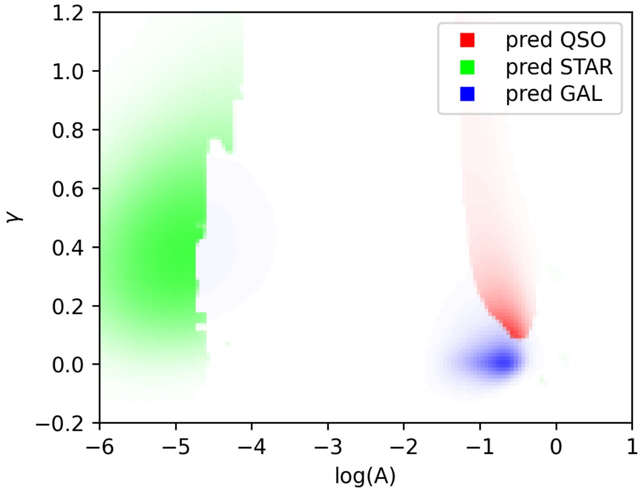


Figure 5.5: The model on all samples using only variability predicts classes as illustrated by the diagram above. The model only uses A and γ plus combinations of them described in Sect. 5.2.4, so the diagram illustrates how the full input feature space maps to model outputs. In most of the feature space, the model predicts the objects to be galaxies, which is also the most frequent class of the dataset.

now, they must cover the entire feature space. The regions are still relatively simple due to regularisation and they approximately match the selection regions of paper I apart from the higher prioritisation of galaxies in sparsely populated regions ($-4.6 < \log(A) < -2$ and $\log(A) > 0$). Areas of low relative frequency of predicted galaxies touch areas of high relative frequencies of predicted stars and quasars, due to the higher total frequency of galaxies in the labeled dataset.

5.3.4 Output table

Information on outliers, time spans etc. are saved in the final output table as described in Table 5.3, along with the resulting predictions and probabilities of each class. The results are produced by model AllA and are published as VILLAIN-Cat in CDS. We also select nonvariable sources from two criteria:

- A close to 0: $A + 3\sigma_{A,+} < 0.01$, and
- A consistent with 0: $A/\sigma_{A,-} < 3$.

5.3.5 Feature importance

We measure the contribution of features by dropping some features in models 3 − 6 of Sect. 5.2.1 and with permutation importance. With permutation importance, each feature is evaluated by randomly permuting its values and measuring the macro averaged F1 loss. This measure has a bias towards correlated features (Strobl et al. 2007; Nicodemus et al. 2010), so important but highly correlated features will likely all get low rankings. It is, however, less resource intensive than computing Shapley values (Shapley 1951) which can be implemented in ways that minimise the bias (Mase et al. 2021; Amoukou et al. 2021).

The top 20 permutation importances of AllA are listed in Table 5.4. The top 13 features are all engineered, including two based on selections in paper I. The highest ranking variability feature is $\gamma - 2\log(A)$ in the 16th place. The sum of the losses is much smaller than the total F1 score of 0.9639, showing that most information lies in highly correlated features. $R - r$ might be ranked high because it can be used for estimation of stellarity since the image quality of PTF and PS1 is different.

5.4 Discussion and conclusions

Using PTF variability, magnitudes from PTF, WISE and PS1 and their combinations, we have created a model (AllA in Sect. 5.2.1) that identifies 1 330 412 quasar, 48 618 737 star and 20 971 755 galaxy candidates with a macro averaged F1 score of 0.9639 on spectroscopically classified sources from SDSS. The model detects 92.49 % of SDSS quasars and has an overlap with SDSS quasars of 95.64 %. We are able to assess the quality of the classifications using predicted probabilities, which are highly accurate, as described in Sect. 5.3.2, and allow for accurate discrimination of true and false positives, as described in Sect. 5.2.5.2.

5.4.1 Model comparison

According to Table 5.2, performance is slightly improved by including variability, especially for quasars. The quasar F1 score improves from 0.9570 to 0.9580 on the matched samples (MagM vs. AllM). The importance of colour is much greater, shown by the lower performance of variability-only models, giving a quasar F1 score of 0.4462 on the

Table 5.3: Catalogue parameters.

Parameter	Description
PTF_RA	Mean J2000 RA in PTF [deg]
PTF_Dec	Mean J2000 Dec in PTF [deg]
PTF_ID	Object ID in PTF
A	Structure function amplitude of variations on time scales of one year [mag]
dA_m	Lower error on A [mag]
dA_p	Upper error on A [mag]
gamma	Power law index
dgamma_m	Lower error on γ
dgamma_p	Upper error on γ
Nepochs	Number of R band epochs in PTF after outlier removal
t_span	Time span in the R band from PTF after outlier removal [days]
median_R	Median R magnitude from PTF [mag]
outlier_fraction	Fraction of R band detections classified as outliers
max_magdiff	Maximum magnitude difference from the R band moving median [mag]
mean_magdiff	Mean magnitude difference from the R band moving median [mag]
W1	WISE $W1$ band [mag]
dW1	Error on $W1$ [mag]
W2	WISE $W2$ band [mag]
dW2	Error on $W2$ [mag]
g	PS1 g band [mag]
dg	Error on g [mag]
r	PS1 r band [mag]
dr	Error on r [mag]
i	PS1 i band [mag]
di	Error on i [mag]
z	PS1 z band [mag]
dz	Error on z [mag]
y	PS1 y band [mag]
dy	Error on y [mag]
SDSS_RA	J2000 RA to SDSS match [deg]
SDSS_Dec	J2000 RA to SDSS match [deg]
redshift	Spectroscopic SDSS redshift
redshiftErr	Redshift error
SDSS_class	Spectroscopic classification from SDSS
nonvariable	1 if A is low and consistent with zero (see Sect. 5.3.4)
P_QSO	Probability of being a quasar
P_GAL	Probability of being a galaxy
P_STAR	Probability of being a star
pred	Predicted class

Notes. Parameters of the output catalogue published as VILLAIN-Cat. For SDSS_class and pred, quasars are labeled 0, stars 1 and galaxies 2. Unknown values are included as NaN.

Table 5.4: Feature importance.

Feature	Loss
$R - r$	0.045
$z - W1$	0.032
$z - W1 - 1.25(g - r)$	0.029
$r - i$	0.013
$i - z$	0.010
$R - W2$	0.008
$z - W2$	0.007
$R - i$	0.004
$R - g$	0.004
$y - W1$	0.002
$W1 - W2 - 0.017W2$	0.002
$r - z$	0.002
$y - W2$	0.002
$W2$	0.001
i	0.001
$\gamma - 2\log(A)$	0.001
$\gamma + 0.5\log(A)$	0.001
γ	0.001
g	0.001
z	0.001

Notes. Permutation importance of the top 20 features. Importance is evaluated with the model using all features and all data, on the test set. All uncertainties are smaller than 0.0005. The losses are highly biased due to feature correlations, but indicate that engineered features are important. $R - r$ could work as a measure of stellarity.

matched samples (VarM). This is still better than the baseline for random guessing of 0.33. Using all features and data (AllA), we are able to give confident predictions of up to 99.994 % as quasar probability – which is also much better than the 80 % with only variability (VarA). By including variability in models, however, we identify more candidates, as more samples can be used. Fig. 5.5 shows how classification is performed with only variability.

For objects with variability parameters, one could always include a feature with the average magnitude in the light curve to improve performance with e.g. a variability + R model. This would likely identify more bright objects as stars.

This study is on PTF sources with measured variability. If variability was not required, including colours would also enable the classification of more sources, i.e. sources without variability information. Although the presented models are classifying all samples, we still see higher purities, completenesses and clearer visible separation of classes in Fig. 5.4 than with the limited manual selections in Figs. 2–3 of paper I.

5.4.2 SDSS comparison

The candidate parameter distributions of Fig. 5.4 are close to the distributions in SDSS labeled objects in Fig. 2 of paper I, but with different total distributions as the sources only have to be found in PTF (Fig. 1 in paper I). We find relatively more galaxies at $W1-W2 > 0.5$ and $W2 > 15$ and a more unified cluster of galaxies in $g-r$ vs. $z-W1$ compared to SDSS. The small structures at $\gamma > 0.5$ and $\log A > -0.2$ in Fig. 1 of paper I are mostly classified as stars. We judge the structure at $g-r=0$ to be an artefact.

5.4.3 Biases

Each survey has selection effects, affecting ratios of the object types and their parameters. Performance is better for sources matched in PS1 and WISE, indicating that matched sources are generally easier to classify. Variability might perform differently compared to colours, if we change the outlier removal and constraints on e.g. light-curve length by paper I.

We also have a bias from the manual exploration of the same dataset in paper I, including the test set of this paper. The models are therefore not created entirely independently of the test set, but they have only been tested on it once.

Each PTF light curve is constructed in paper I by grouping data points with the same PTF ID. All light curves have different PTF ID's, so all samples should correspond to independent astrophysical objects. However, some PTF objects have near-identical coordinates. For these ID's, PTF light curves are therefore correlated, and the objects are likely matched to the same object in WISE and PS1. This gives a small bias, since correlated feature values appear across the training and test set. For labeled data, all SDSS matches are unique, but 0.02 % (268) of the sources have a duplicate match in WISE and 0.004 % (70) have a duplicate in PS1. 22 % (25 329 144) of PTF objects have a neighbour within two arcseconds, but only 0.01 % (158) within the labeled set.

5.4.4 Perspectives

The catalogue described in Sect. 5.3.4 includes predictions and probabilities of 70 920 904 PTF objects. It greatly expands on the set of 4 618 756 DR17 SDSS spectroscopic classifications out of which the labels are from 1 747 471. The catalogue includes confident quasar candidates (high P_QSO) and standard star candidates (flagged nonvariable) for use in future research. Subsets could be interesting to observe and analyse further, such as quasar candidates or variable galaxies. Using the included probabilities, one can extract a set of e.g. quasar candidates of a chosen minimum P_QSO for a preferred trade-off between completeness and purity.

Monochromatic variability is not enough for confident classification of most sources into the three macro-classes. However, it does add information within each class and could be used for selection of rare subtypes or distinction of e.g. type I and type II AGN (De Cicco et al. 2022). It is also useful in cases where colours are not available. For simple variability selection without machine learning, we suggest selecting by regions similar to those of Fig. 5.5.

For machine learning tasks on similar data, we suggest datasets of at least ∼100 000 labeled samples for an expected macro averaged F1 score of 0.9610 for new data. The required survey size for a given score can be estimated from Fig. 5.2. For small datasets, we suggest adjusting class ratios by oversampling (copying or generating synthetic samples) or more advanced techniques combining oversampling and undersampling (removing samples) such as SMOTETomek or SMOTEENN (Batista et al. 2004) on the training set (not on the test set). We also suggest combining the training and validation sets before testing on small sets and using cross-validation. Stratification can ensure that even small folds are representative of all classes. `HistGradientBoostingClassifier` is only faster than `GradientBoostingClassifier` from scikit-learn for datasets of $\geq$10 000 samples, so alternative algorithms may be considered for speed. For large numbers of classes, both algorithms are inefficient since they create a tree for each class during each boosting iteration. If a large unlabeled set is available, semi-supervised learning can be applied to learn from it. In that case, the unlabeled samples can all be added to the training set.

In areas of parameter space with few SDSS labels, performance can be improved using active learning (Settles 2009). By identifying where in feature space new labels would be most useful, and expanding the labeled set accordingly, fewer labeled samples are needed. This is especially relevant in areas with higher relative population densities in PTF than SDSS or, in the future, in the LSST. Performance would of course also be improved by adding additional information or well-designed feature engineering such as more bands, photometric redshifts, proper motions, stellarity etc. The joint survey processing of LSST, *Euclid* (Laureijs et al. 2011) and the *Nancy Grace Roman Space Telescope* (Akeson et al. 2019) will include deep, multi-band information in the optical and near-infrared, which can be used similarly to the WISE and PS1 bands in this project (Chary et al. 2020). Another approach to automatic classification with variability is using time-series layers in neural networks. More information can be captured by the model by directly using the full light curves instead of, or in addition to, manually selected summarising features like A and γ (Jamal & Bloom 2020).

We used a histogram based gradient boosting classification tree, which is fast, performs well, learns from missing values, produces probabilities, detects nonlinear patterns and is easy to implement with scikit-learn. Feature engineering further improves performance. No other astronomical papers in the SAO/NASA Astrophysics Data System (ADS)[1] mention this implementation to date, but we recommend further use in astronomy as an alternative to XGBoost and LightGBM. With models like the ones of this paper, future sources can quickly and automatically be classified.

[1] https://ui.adsabs.harvard.edu

Acknowledgements

This work was supported by a Villum Investigator grant (project number 16599). SHB was also supported by a grant from the Danish National Research Foundation. AA was also supported by a Villum Experiment grant (project number 36225).

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

The Pan-STARRS1 Surveys (PS1) and the PS1 public science archive have been made possible through contributions by the Institute for Astronomy, the University of Hawaii, the Pan-STARRS Project Office, the Max-Planck Society and its participating institutes, the Max Planck Institute for Astronomy, Heidelberg and the Max Planck Institute for Extraterrestrial Physics, Garching, The Johns Hopkins University, Durham University, the University of Edinburgh, the Queen's University Belfast, the Harvard-Smithsonian Center for Astrophysics, the Las Cumbres Observatory Global Telescope Network Incorporated, the National Central University of Taiwan, the Space Telescope Science Institute, the National Aeronautics and Space Administration under Grant No. NNX08AR22G issued through the Planetary Science Division of the NASA Science Mission Directorate, the National Science Foundation Grant No. AST-1238877, the University of Maryland, Eotvos Lorand University (ELTE), the Los Alamos National Laboratory, and the Gordon and Betty Moore Foundation.

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

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

SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, Center for Astrophysics | Harvard & Smithsonian, the Chilean Participation Group, the French Participation Group, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU) / University of Tokyo, the Korean Participation Group, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatário Nacional / MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University, and Yale University.

This research made use of the cross-match service provided by CDS, Strasbourg.

This research made use of Astropy,[2] a community-developed core Python package for Astronomy (Astropy Collaboration et al., 2013; Astropy Collaboration et al., 2018).

This research has made use of the NASA/IPAC Infrared Science Archive, which is funded by the National Aeronautics and Space Administration and operated by the California Institute of Technology.

This research makes use of the SciServer science platform (www.sciserver.org).

SciServer is a collaborative research environment for large-scale data-driven science. It is being developed at, and administered by, the Institute for Data Intensive Engineering and Science at Johns Hopkins University. SciServer is funded by the National Science Foundation through the Data Infrastructure Building Blocks (DIBBs) program and others, as well as by the Alfred P. Sloan Foundation and the Gordon and Betty Moore Foundation.

This research has made use of the VizieR catalogue access tool, CDS, Strasbourg, France (DOI : 10.26093/cds/vizier). The original description of the VizieR service was published in 2000, A&AS 143, 23.

This research has made use of NASA's Astrophysics Data System.

[2] http://www.astropy.org

III

Perspectives and conclusions

Chapter 6

Perspectives

We have analysed distributions of variable objects and created a machine learning model for their classification in the previous chapters. We will now expand parts of this analysis with extra diagrams of their parameter distributions and light curves. We will also go into more depth on the issue of objects split by the PTF. Then, we discuss some alternative machine learning techniques, including how the raw data might be used for the creation of different model input features. In addition to changing the method of extracting information from the data, we could have included different data. We end this chapter on the discussion of alternative or additional astronomical data for future models.

6.1 Expanded analysis

6.1.1 Predicted classes

In Fig. 6.1, we show the distributions of predicted classes from a model using only variability (VarA from Sect. 5.2.1). The distributions of each class are also shown in Fig. 5.5 of the previous chapter, but in Fig. 6.1 the logarithmic colour mapping allows us to inspect the more sparsely populated areas and see clearer boundaries between classification regions. Sources in most of the A-γ space are classified as galaxies. Quasars are mostly in $\gamma > 0.1$ and $-1.3 < \log(A) < 0$ and stars generally have $\log(A) < 4.5$, but sources in a few areas at $\log(A) \sim 0$ are also classified as stars.

In Fig. 6.2, we show the colour distributions of VarA predictions. Comparing the colour distributions of variability classified sources in Fig. 6.2 to the sources selected with manual variability criteria in Fig. 4.4, the classes have poor separation in both cases. However, VarA maintains a similar separation while including all sources compared to the 50 % of fitted PTF sources being selected in Fig. 4.4.

For easier comparison, especially for the colour blind, Figs. 6.3 – 6.4 re-illustrate the distributions of Figs. 4.2 and 5.4 with one class per plot. In Figs. 6.3 – 6.4, AllA must predict classes in parts of the feature spaces that are not covered by the SDSS labels. This *domain shift* of distributions between the labeled and unlabeled set makes the model less confident.

For example, in Fig. 6.4 some sources in the stellar main sequence of $z - W1$ versus $g - r$ are predicted to be quasars even though sources with these values are not present in the SDSS labeled quasar set. The predicted quasars with main sequence colours could be sources with extreme values in other features or based on overfitting in AllA. The overfitting could stem from sources misclassified by SDSS or sources with erroneous parameters.

Examples of light curves of predicted quasars, stars and galaxies are shown in Figs. 6.6—6.8. We still observe clearer variability in quasars than in most of the included stars or galaxies. Interestingly, the star in the upper right corner of Fig. 6.7 appears to include powerful eruptions – however, the two distinct magnitude "modes" could also be a sign of mixing of the light curves of two different objects. This object has a neighbour about 5 arcseconds away as shown in Fig. 6.9, and so, PTF has likely assigned the same OID to both. The combination of light curves from multiple astrophysical objects can artificially increase variability parameters. It also obscures which object a classification refers to. Compared to split light curves, it is more challenging to establish how common the problem of combined light curves is. However, it only decreases the performance of models without increasing overfitting, meaning the performance metrics can be trusted.

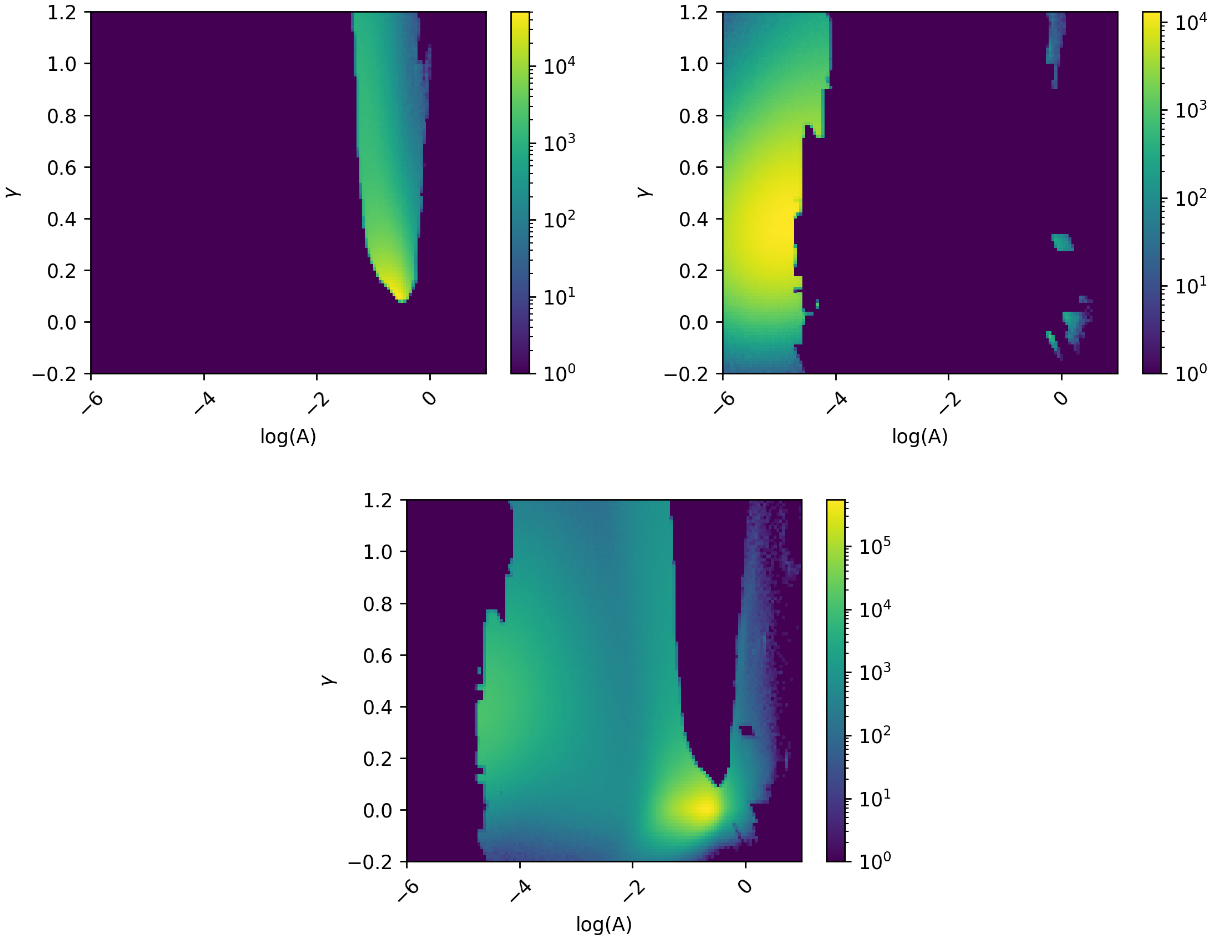


Figure 6.1: Predicted A-γ regions from VarA. A and γ cover the full decision space of the variability model, and so, we can observe the full decision boundaries in two dimensions. The top left diagram shows predicted quasars, the top right shows stars and the bottom diagram is for galaxies. The rules of the model predict objects in most of the parameter space to be galaxies.

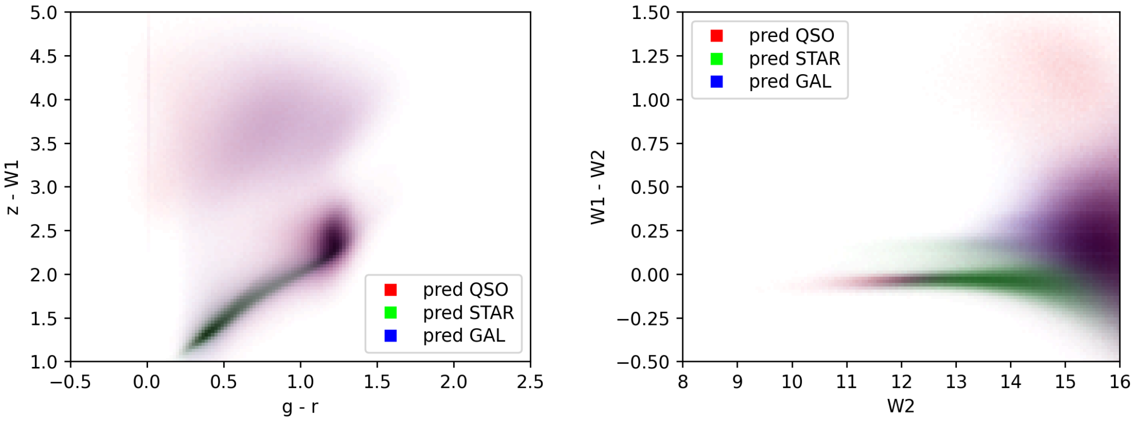


Figure 6.2: Colours of objects classified by VarA. We see little seperation, similar to Fig. 4.4, but including the whole data set of light curves.

6.1.2 Variable galaxies

A subset of the galaxies in Fig. 6.3 are variable. To further analyse these, we check SIMBAD registrations for spectroscopically confirmed galaxies in general and for spectroscopic galaxies selected as quasars by the manual variability criteria in Chapter 4. Fig. 6.10 shows histograms of the SIMBAD registrations. Most are simply registered as galaxies and relatively few as AGN or similar in both data sets. Cartier et al. (2015) also analyse structure functions and find some objects otherwise classified as "normal" galaxies to include some variable objects which are likely to be low-luminosity AGN. Their emission lines could be diluted by the host galaxies, preventing spectroscopic classification as AGN. They find broad line AGN to have $A \sim 0.1$ and $\gamma \sim 0.025$.

6.1.3 Split objects

In Sect. 5.4.3, we discussed the bias split PTF objects. This is mainly an issue in the full dataset, and can be ignored in the subset with SDSS matches. For future analysis of a larger labeled set or for unsupervised learning, identifying split objects could both improve accuracy and reduce bias of the results. Combining light curves of the same physical object would improve variability estimates and fewer duplicates could reduce overfitting during classification.

We identify close neighbours using a k-d tree (see Sect. 2.3). Fig. 6.11 shows that

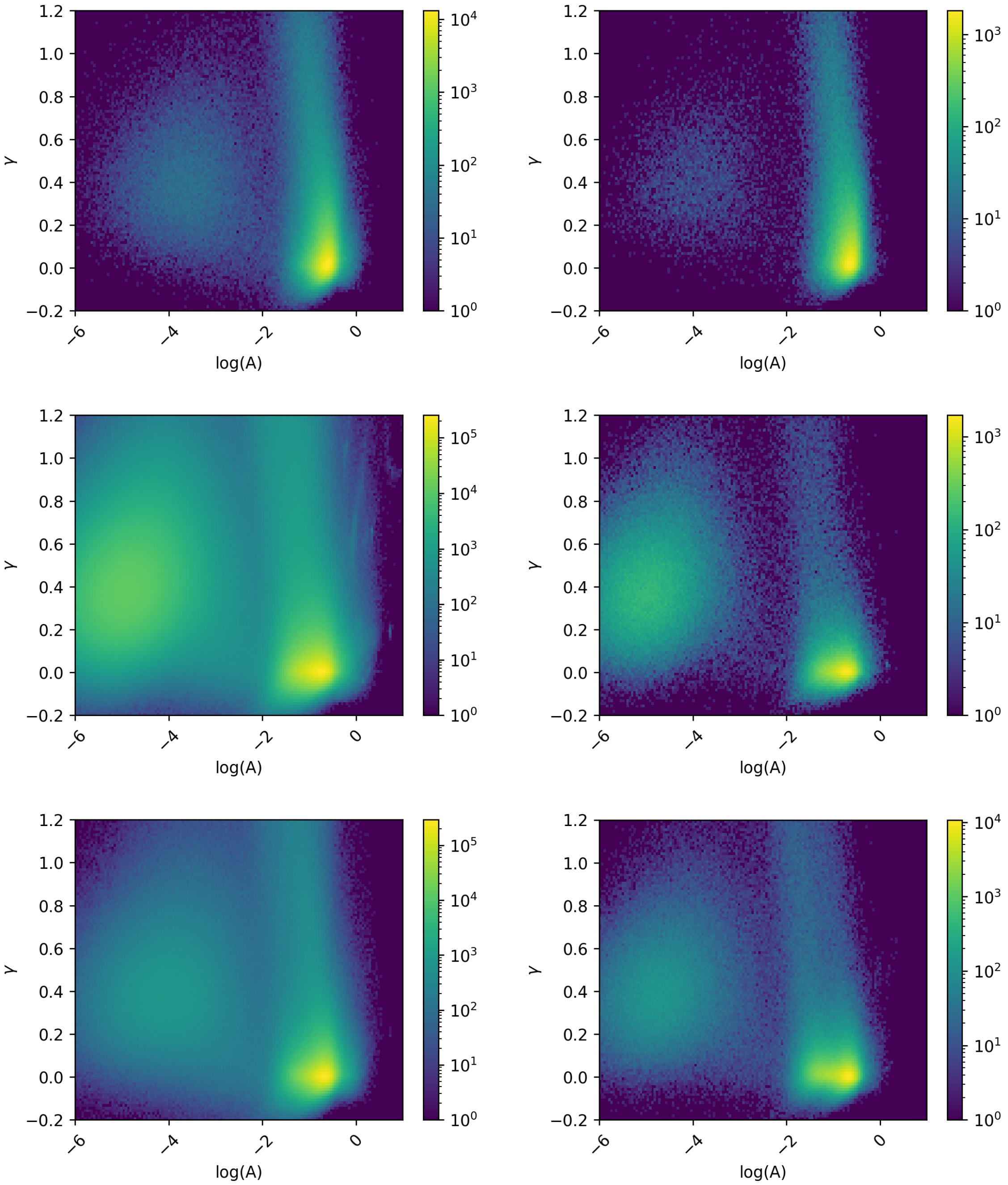


Figure 6.3: Variability distributions of classified sources from AllA (left) and SDSS spectroscopy (right). The top panels show quasars, the middle row stars, and galaxies are in the bottom row. The differences between AllA and SDSS classifications are affected by the classifiers and the full PTF data set covering a broader range of sources than the SDSS labeled subset.

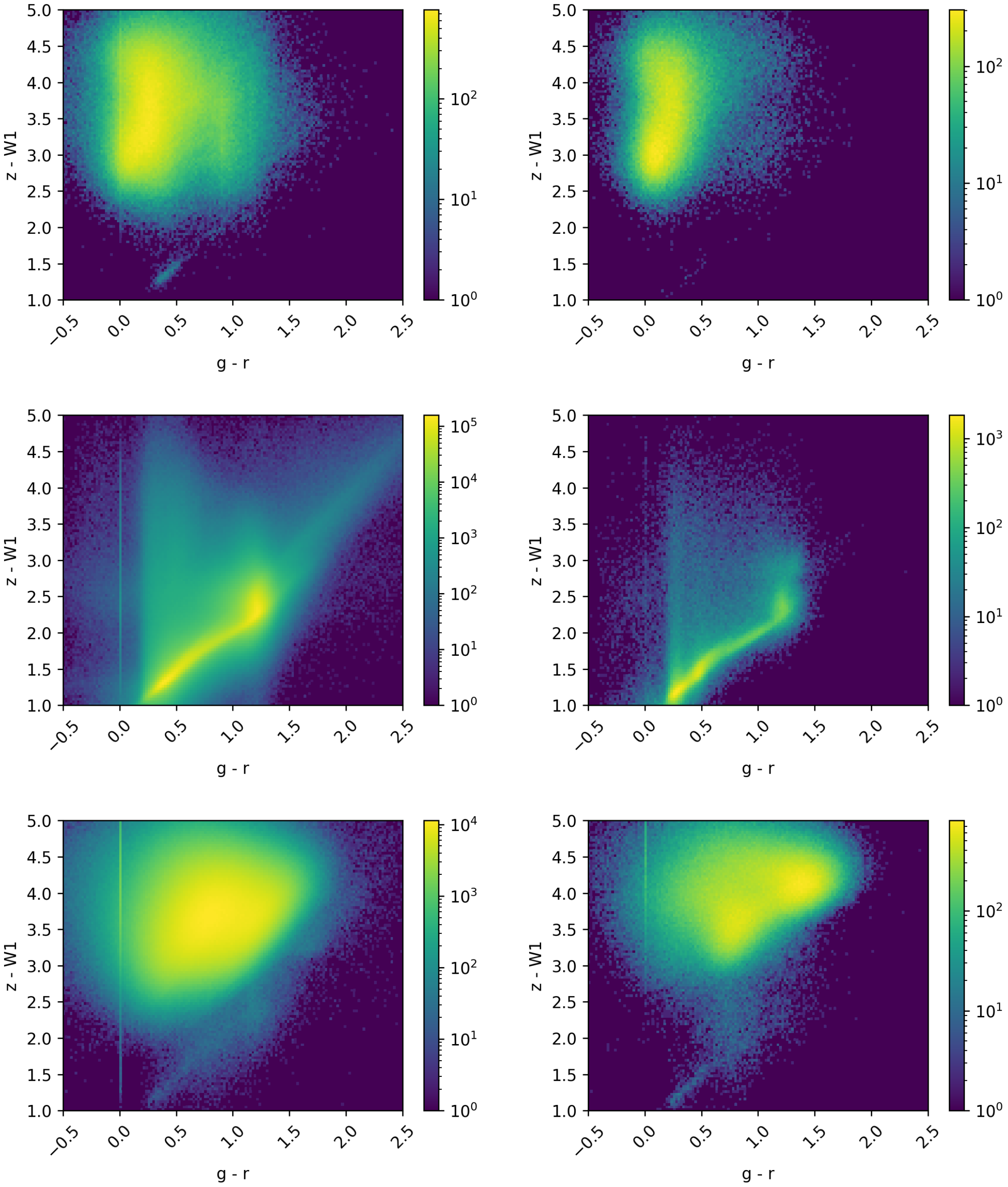


Figure 6.4: Distributions of $g - r$ versus $z - W1$ for sources classified with AllA (left) and SDSS spectroscopy (right) for quasars (top), stars (middle) and galaxies (bottom). The full PTF data set covers a broader range of colours than SDSS spectroscopic classes.

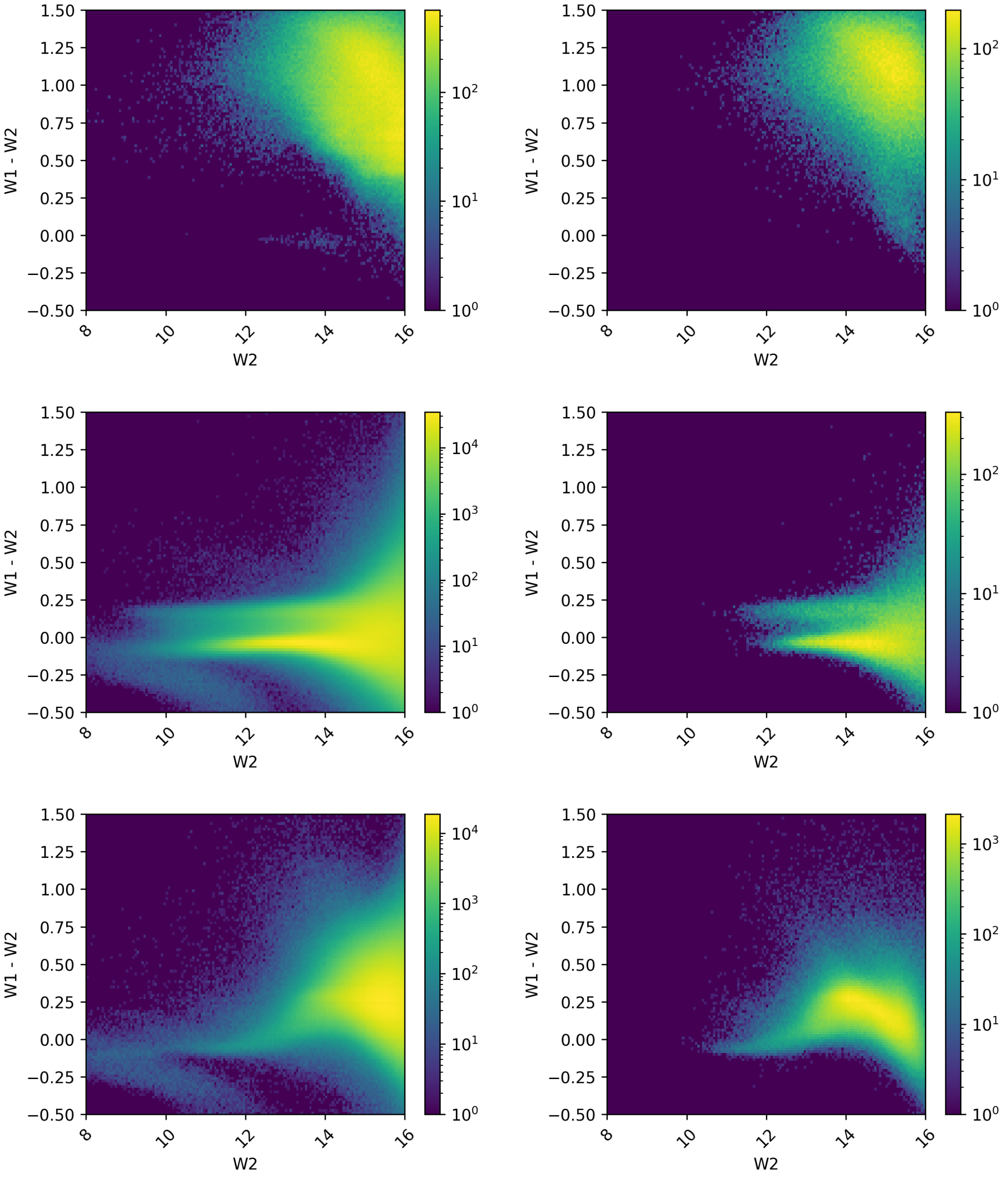


Figure 6.5: $W2$ versus $W1 - W2$ of sources classified by AllA (left) and SDSS (right) for quasars (top), stars (middle) and galaxies (bottom). The distributions are generally similar across the two columns, but again we see that AllA classifications cover more colour and magnitude values.

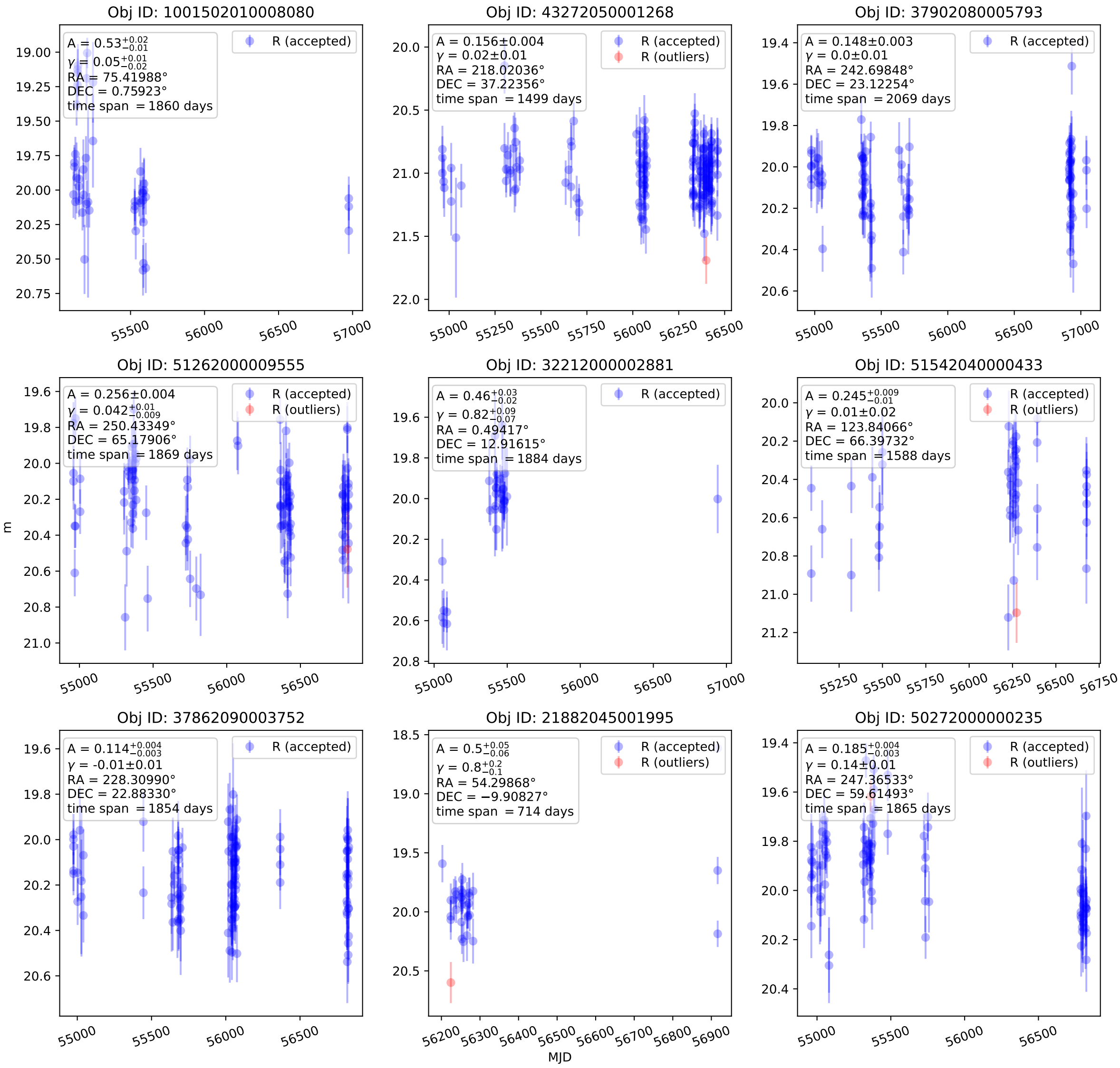


Figure 6.6: Light curves of predicted quasars from AllA. They all have A significantly different from 0, with a larger spread in γ values.

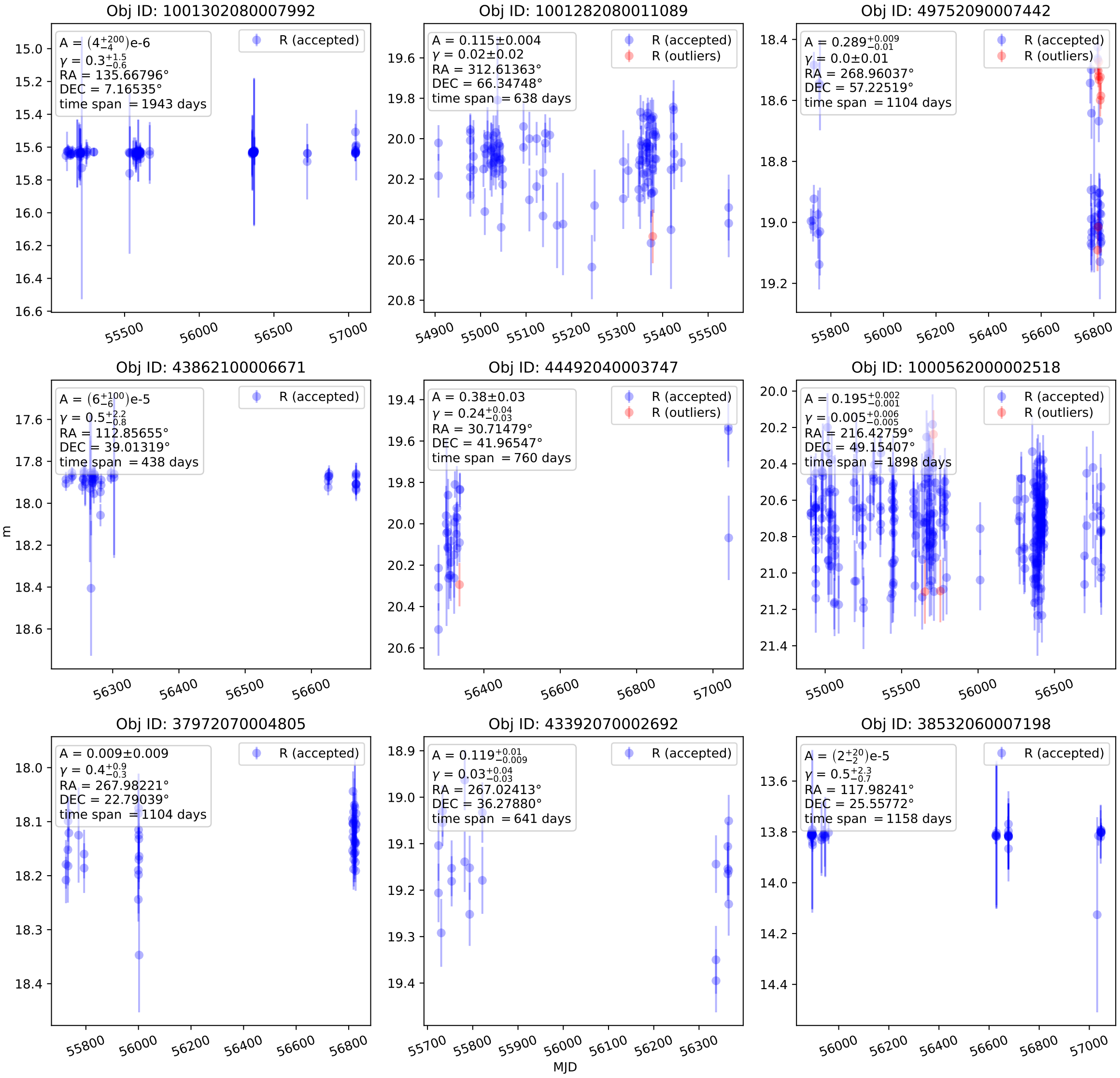


Figure 6.7: Light curves of predicted stars from AllA. The model has included both stars that do show power law variably and some that do not. The diagram in the top right appears to combine the light curves of two different objects given the same OID in PTF.

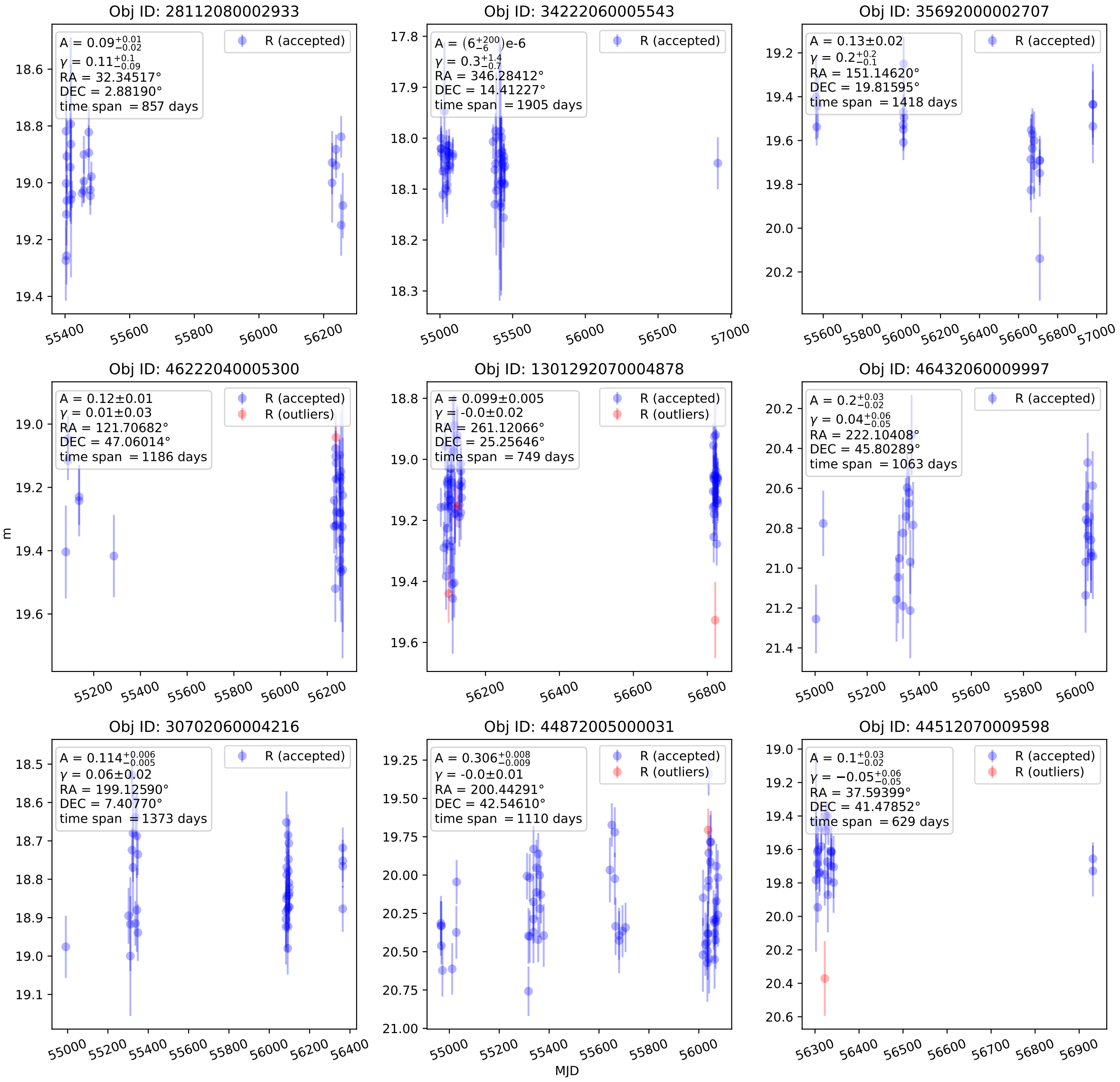


Figure 6.8: Light curves of predicted galaxies from AllA. Some show variability, perhaps due to AGN.

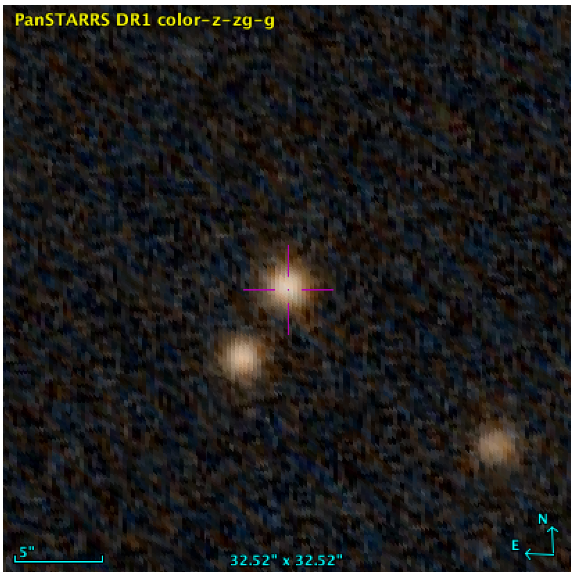


Figure 6.9: PS1 imaging of a star with a light curve that has likely been combined with a close neighbour. The full image is 32.5×32.5 square arcseconds. The image is accessed using Aladin sky atlas, developed at CDS, Strasbourg Observatory, France (Bonnarel et al. 2000).

PTF neighbours are typically either closer than an arcsecond or further away than five arcseconds. 22 % of the fitted PTF sources have at least one neighbour within two arcseconds, and one even has eight. Before data cleaning, it had 14. Most sources have zero or one neighbour as illustrated in Fig. 6.12. The neighbours of the source with eight neighbours are 0.9 – 1.9 arcseconds away from the queried coordinates at RA 83.769474 and Dec -4.898451. They are located in the Orion molecular clouds (Tsujimoto et al. 2003), and there are no close sources visible to PS1 in the area, as shown in Fig. 6.13. It has been observed during a time frame of a few days with $15.7 < R < 18.0$, and then twice in following years. Other PTF sources have much closer neighbours such as the one at RA 241.258777 and Dec 22.479685 with six neighbours 0.02 – 0.09 arcseconds away. Fig. 6.13 shows that they are all located in a bright star with g ∼14.9 (WISEA J160502.10+222846.7). De Cicco et al. (2021) also matched their data set with itself to avoid analysing blended sources, leading them to exclude 4.7 % of the sources.

Instead of only relying on PTF OIDs, better clustering of nearby points could be achieved with unsupervised learning. Algorithms such as HDBSCAN can identify clusters using differences in density while not including points it identifies as noise (McInnes

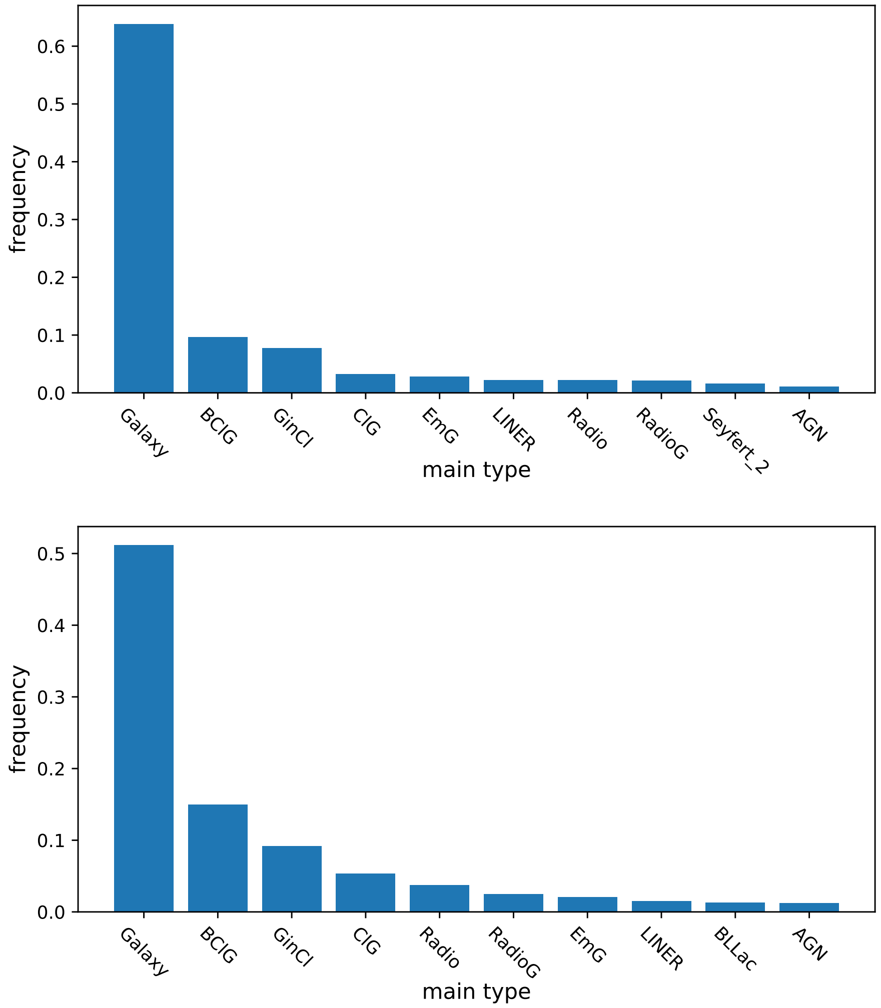


Figure 6.10: SIMBAD registrations of the spectroscopically confirmed galaxies. The upper diagram includes all galaxies and the lower diagram only includes those fulfilling the variability criteria of Chapter 4.

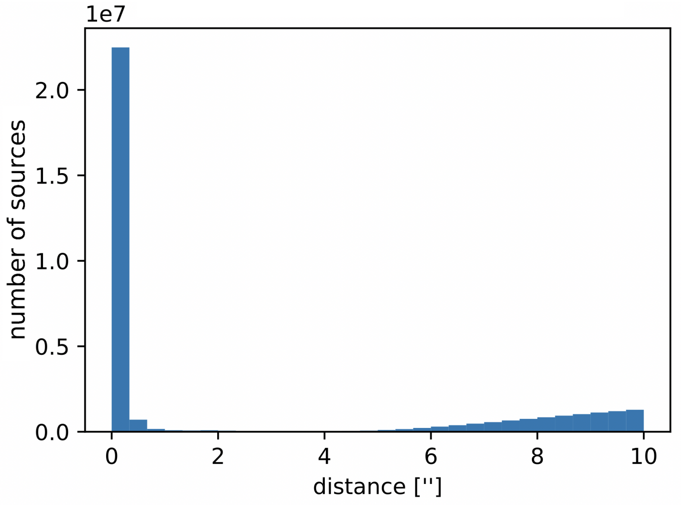


Figure 6.11: Histogram of the distance of PTF neighbours within 10 arcseconds of another PTF source. Most neighbours are closer than an arcsecond and likely represent the same physical object.

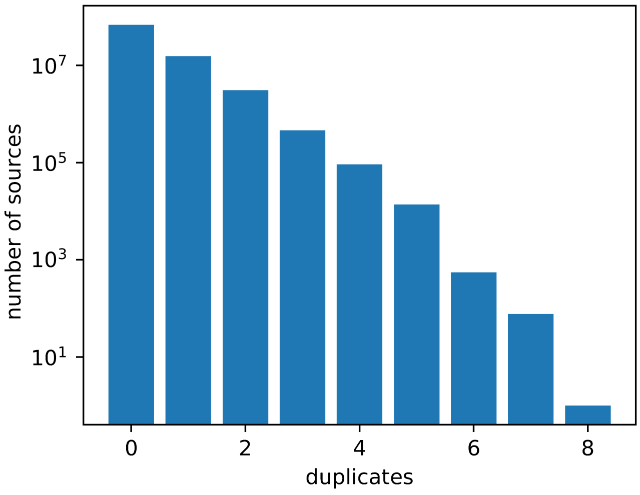


Figure 6.12: Histogram of the number of fitted neighbours to PTF sources within two arcseconds. It indicates that assigning multiple PTF OIDs to the same source is common. Only 0.01 % of PTF sources in the labeled set have a close neighbour that is also in the labeled set.

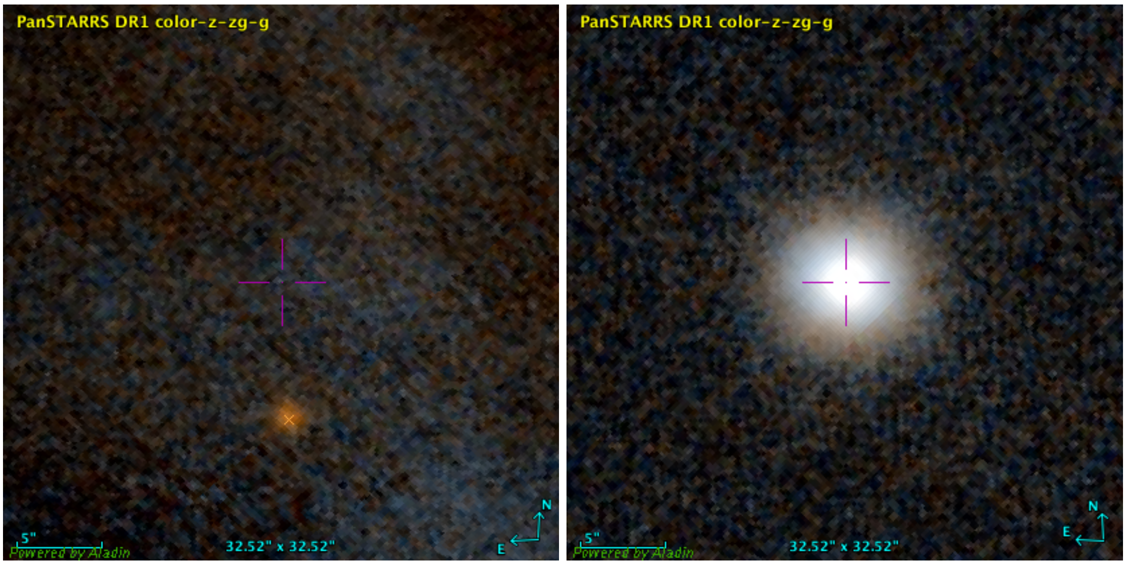


Figure 6.13: Two images showing 32.52×32.52 square arcseconds each from PS1 centred on PTF sources with multiple fitted PTF sources within two arcseconds. The source to the left (83.769474, −4.898451) has eight neighbours and the source to the right (241.258777, +22.479685) has six. The right source is also centred on WISEA J160502.10+222846.7. The images are accessed using Aladin sky atlas, developed at CDS, Strasbourg Observatory, France (Bonnarel et al. 2000).

et al. 2017). Unlike with Density-Based Spatial Clustering of Applications with Noise (DBSCAN), HDBSCAN clusters can have variable densities. Similarly to Schmidt et al. (2010), a simple solution could be to select all points within a certain distance. Fig. 6.11 suggests that this distance should be between 1 and 4 arcseconds.

6.2 Alternative machine learning models

As discussed in Sect. 2.5, a wide array of machine learning models have been applied to variable astrophysical objects. The chosen model of Chapter 5 performs well, but it is likely that other models would do the same. With different techniques and depending on the data set, performance could even be improved. Undersampling, oversampling or a combination could improve class balance and make the model focus less on galaxy prediction. Different hyperparameter selection methods could also be tested for model optimisation, although `HalvingRandomSearchCV` is already effective. FLAML (Wang et al. 2019) is an example of a library for computationally efficient hyperparameter tuning.

6.2.1 Semi-supervised learning

Especially on smaller datasets, learning overall distributions from the unlabeled data can be useful. It is possible to learn from both the labeled and unlabeled data via semi-supervised learning. We test semi-supervised learning in the form of `SelfTrainingClassifier` by sci-kit learn (Pedregosa et al. 2011). This uses a supervised model as its base estimator to predict pseudolabels for the unlabeled set in an iterative manner. The semi-supervised model, however, never gave better performance than its base estimator (the supervised model), even for a wide range of hyperparameters. With semisupervised learning, we include all unlabeled data in the training set and validate using labeled data only, because labels are needed for performance evaluation. Setting aside unlabeled data would not help during validation or testing, but it can improve model training.

We tested semi-supervised learning with `HistGradientBoostingClassifier`, but also with `LogisticRegression` implemented by sci-kit learn (Pedregosa et al. 2011). Logistic regression is a fast machine learning model, but its decision boundaries are linear. To identify non-linear patterns, we include nonlinear combinations of features with `PolynomialFeatures`. As the decisions of logistic regression depend on scale, we standardise the data with `RobustScaler`, which subtracts the median and scales based on the first and third quartiles, making the standardisation robust to outliers. This pre-processing is based on the training set to avoid overfitting. To capture patterns on similar feature scales, we also use $\log A$ (in base 10) instead of simply A as the input for `RobustScaler`.

In `SelfTrainingClassifier`, it is important that the base model is well calibrated, meaning that the predicted probabilities of belonging to each class correspond to the actual frequencies of the classes in the data. Unlike the outputs of `HistGradientBoostingClassifier` in Fig. 5.3, the `LogisticRegression` model was not well calibrated. Therefore, we applied a calibration model to learn how to adjust the probabilities. We used `CalibratedClassifierCV`, which learns via cross validation (see Sect. 5.2.2.1). However, the outputs of `SelfTrainingClassifier` are not well-calibrated either. Therefore, `CalibratedClassifierCV` is used again after applying the model. In both cases, we optimise it with a OvR approach. This results in a model with nested cross validation, which is most suitable to large datasets where splitting will not affect performance too much. Stratification helps for small data sets.

The full model becomes `CalibratedClassifierCV(SelfTrainingClassifier(CalibratedClassifierCV(LogisticRegression(RobustScaler(X,Y)))))`. This is all possible in sci-kit learn (Pedregosa et al. 2011), but more complicated to implement and optimise than `HistGradientBoostingClassifier`, and we achieve poorer performance and a slower model. Therefore, we recommend just using `HistGradientBoostingClassifier` for data sets similar to the one of this thesis.

6.2.2 Features

In the machine learning models of Chapter 5, we have created features according to Sect. 5.2.4.1. It could be interesting study how performance is affected by creating different features from the same data. We might also be interested in adjusting existing features. For example:

- Photometric redshifts could be modelled similar to Cunha & Humphrey (2022).

- If photometric redshifts are used, the observed-frame can be corrected to the rest-frame by $\Delta t = \Delta t_{obs}/(1+z)$ like by Kozłowski (2016b).

- A and γ could be computed with different outlier removal to compare classification results and find the most informative measure (separation of classes). For example, Schmidt et al. (2010) used a looser outlier criterion by removing all data points more than 0.25 mag from a moving median. We generally remove very few outliers – up to about 1 % in a source. Most objects contain data points with an absolute distance from the moving median of more than 0.25 mag, especially those we predict to be quasars or galaxies.

- If quasar and AGN SFs include characteristic time scale after which they no longer follow a power law, it can be beneficial not to fit the full SF with a power law model. If the simple power law model is still used, one could limit the fitted range to a maximum Δt depending on the expected turnover timescale, similar to Kozłowski (2016b).

- Especially in linear models, including polynomial features can improve performance by making it easier for the model to detect patterns.

- There are alternatives to creating new features manually. they can be selected automatically through Deep Feature Synthesis (Kanter & Veeramachaneni 2015) or Genetic Feature Generation (Mamontov et al. 2022). Another option is to let an encoder model learn representations of the light curves similar to the method of Jamal & Bloom (2020).

- Different variability measures from Sect. 2.4 could be included. These might detect different types of variability and have different computational costs.

- Imputation could be used for sources with missing values in some features. `HistGradientBoostingClassifier` handles missing values, but well implemented imputation such as with NeuMiss (Le Morvan et al. 2020) might perform even better.

6.2.2.1 Feature evaluation

In AllA and AllM, we use 42 features, and we might get a simpler model with similar or even better performance by removing some features. Evaluation of feature importance is also important for model interpretability. We use permutation importance which is biased for correlated features – and many of the features are highly correlated. A highly correlated feature gets a low permutation importance because the information it shares with other features is not permuted in them. Permutation will also often place samples outside the feature space regions of the training set, leaving the model to extrapolate patterns to unknown regions. As an example, a permutation of the feature "$W1 - W2$" will make its value different from the difference between the features "$W1$" and "$W2$". This is not something the model has seen before, and it is therefore difficult to interpret. Furthermore, for highly correlated features there is little permutation loss from removing one of them as the information is still included in other features, so they will be low-ranking despite the importance of not removing all.

Alternatively, we suggest using Shapley values (Lundberg & Lee 2017; Shapley 1951) with modifications for handling correlated features (Mase et al. 2021; Amoukou et al. 2021). This has a high computational cost. See Table B.1 of Ansari et al. (2022) for an example of how processing time scales with data set size.

Another approach is to re-run the model once per feature in a leave-one-out scheme (Lei et al. 2016). This is also computationally heavy. We apply a similar method by training a model with variability parameters only and colours only to evaluate changes in performance.

Tree-based models offer feature importance evaluation with Mean Decrease in Impurity (MDI) (Louppe et al. 2013). This measure is based on the positions of tree nodes which use the feature – features that are used closer to the root node will be used for classification of more sources. Features used closer to the leaf nodes are perhaps only used for a small subset of all leaves. Removing the splitting node using the evaluated feature would decrease the purity. However, the MDI is evaluated on the training set and gives higher scores to features with many unique values. It is also not currently supported by `HistGradientBoostingClassifier`.

6.3 Different data

In the previous section, we have looked into alternative classification models. Another another way of improving classification is to supply a model with additional or different information to learn from. Some options are:

- An alternative to AllWISE is the unWISE Catalog (Meisner et al. 2017) with better depth (0.7 mag fainter) and modelling of sources with close neighbours. It can be queried as II/363 in VizieR (Ochsenbein et al. 2000).
- Gaia proper motions and parallaxes would make it easier to distinguish the classes. The proper motions of each class are shown in Fig. 4.14, and we see that the distributions are different.
- Additional colours could be included. Some can be extracted via the catalogues used in this thesis. $W3$ and $W4$ were useful for classification of spectroscopic SDSS classes by Cunha & Humphrey (2022), although not as much as $W1$ and $W2$. One could also include fewer colours for better comparison with other surveys, such as the LSST.
- Variability measurements from other studies could be included when available, such as A and γ computed from SDSS light curves (Pâris et al. 2017).
- Stellarity is useful for identifying galaxies as extended objects, but other information is necessary to distinguish quasars from stars.
- Light curves from different surveys can be combined to create longer light curves with more epochs. This comes at the cost of possible artefacts of photometric transformation to the same filter, unless the combined surveys already have very similar filters. Suberlak et al. (2021) combined PS1 and SDSS light curves and suggest adding data from ZTF and LSST in the future. They do not include PTF, ZTF and the Catalina Real-time Transient Survey (CRTS) due to lower limiting magnitudes (and a broad filter for CRTS), but they have calculated the photometric offsets for their inclusion. Fig. 2 of Suberlak et al. (2021) illustrates the limiting magnitudes, baselines and sky coverage of each survey. Another example is Stone et al. (2022) combining light curves from PS1, SDSS, the Dark Energy Survey and follow-up monitoring with Blanco 4m/DECam.

- Labels can be extracted from other surveys than SDSS. This would avoid the selection biases of SDSS (see Sect. 4.6.3) and enable learning from fainter sources. One could use different labels than quasars, stars and galaxies or include labels for subtypes. This could be different types of variable stars, Type I and II AGN or lensed quasars. For large numbers of classes, `HistGradientBoostingClassifier` will be less effective. Ideally, labels from multiple surveys would be included for robustness. It is worth noting that SDSS is biased in favour of including quasars that are in the variability selection region of Table 4.2, since their target selection process includes a set of PTF light curves with $\gamma > 0$ and $\gamma > -30A + 1.5$ (Myers et al. 2015).

  Instead of adding labels, we could use one less by grouping stars and galaxies as "non-quasars". This would result in a machine learning model focusing more on the classification of quasars instead of also separating stars and galaxies.

- To include more stars, one could change the constraint of $R > 12$. Bright stars are, however, more easily distinguished from quasars. Regarding standard stars, we are mostly interested in the faint candidates. It could be interesting to compare performance on bright and faint sources.

- On the other hand, one could place stricter constraints to fit only the sources that are otherwise difficult to classify to reduce computational costs. Perhaps, the constraints could be in magnitude, colour or simple variability measures (see Sect. 2.4).

- Light curves could be constructed differently to avoid splitting light curves of the same source or combing light curves of different sources, as discussed in Sects. 6.1.1 and 5.2.2. In other surveys than PTF, the importance of this could be different depending on how OIDs are assigned. Instead of combining by OID, Schmidt et al. (2010) queried all points within 0.5 arcseconds. Based on Fig. 6.11, this seems to be on the low side for PTF. Alternatively, a clustering algorithm such as HDBSCAN could be applied.

Chapter 7

Conclusions

In this thesis, we have explored variability and colours of quasars, stars and galaxies. We have demonstrated methods of analysing the properties of these objects, and how to select them in a large astronomical survey. Below we connect the findings of each part of the thesis.

7.1 Part I

In Chapter 1, we explored the astrophysics of variable objects. The main focus is quasars and AGN, which vary non-periodically over a wide range of time scales. The main mechanism behind the variability is unknown, but many explanations of the literature focus on accretion disk instabilities. Most galaxies are not expected to be variable, and so, light-curve variability on human time scales is one way of identifying quasars and AGN in their centres. Stars are more diverse in their variability properties, and many do not have detectable variability at all.

In Chapter 2, we discussed relevant statistical measures. In order to estimate the best fit parameter values, we use MCMC to sample the PDFs of the parameter likelihoods. That is, the parameter values that would make observation of the given data most probable. This is combined with parameter priors in accordance with Bayes theorem. To search large astronomical data bases, we define mathematical tree structures for efficient querying. We discuss different methods for quantifying variability including simple power law SFs and advanced CARMA models. Neither can fully explain observations in the literature, but they are sufficient for selecting variable sources. Sources can be classified automatically through a wealth of machine learning algorithms. We stress the importance of selecting, training and testing models in an unbiased way to get reliable estimates of performance. Then, we discuss some machine learning techniques applied to variable objects in the literature.

7.2 Part II

Chapter 3 addresses the practical challenges of the data processing to obtain the variability and colour information of this thesis. The large data set requires high performance computing facilities, especially during MCMC fitting, which took about six months. Some parts of the sky took 6—7 orders of magnitude more computational resources to process than others, so we split the data into 6340 files for parallel processing.

Chapter 4 is based on Bruun et al. (2023a) in which we fit power laws to the SFs of 71 million sources in PTF. We apply data cleaning to ensure more reliable variability parameters. This includes removing outliers and using light curves with at least 20 epochs. We analyse the parameters of this data set as a whole for the analysis to be as representative of the survey as possible. This gives us the parameter distributions of Fig. 4.1. Then, we inspect the variability and colours of objects selected as quasar, star and galaxy candidates. The selection criteria are based on distributions of spectroscopically classified sources in SDSS. We find quasars to show the most power law variability (highest values of A and γ). Seven band colour information in the optical and infrared is more effective than monochromatic variability in selecting quasars, but with both we select them with a purity of 99.3 % for a completeness of 12.53 %. Colour and variability also enables the identification of spectroscopic misclassifications.

To optimise the selections with a balance of purity and completeness, we create a machine learning model in Chapter 5. This chapter is based on Bruun et al. (2023b). The model classifies the sources using the same variability and colour information as Chapter 4, but uses the information more effectively. The choice of `HistGradien` tBoostingClassifier, which is an implementation of histogram based gradient boosting, enables fast model selection and training, and the resulting classifications are well performing and well calibrated. We found no other astronomical papers using this implementation to date in the SAO/NASA Astrophysics Data System, although for example Cunha & Humphrey (2022) used similar models. The model handles missing values, enabling us to classify the whole data set including objects without colour information. We achieve a purity of 95.64 % and a completeness of 92.49 % for quasars with this photometric model compared to the spectroscopic classifications by SDSS. The trade off between purity and completeness can be adjusted using the predicted probabilities of objects belonging to each class. The trade off will then change as illustrated in Fig. 5.1. The probabilities and predictions are included in a catalogue along with variability parameters, colours, positions etc. Models that only include variability or colours do not achieve the same performance, especially models based only on monochromatic variability. We also analyse how performance depends on the number of labeled samples (objects with known classes to learn from), and this is shown in Fig. 5.2. We find that the performance is relatively stable for more than 100 000 labeled samples. We recommend the use of HistGradientBoostingClassifier and the inclusion of at least 100 000 labeled samples for optimal performance in future, similar projects.

7.3 Part III

In this final part of the thesis, we use Chapter 6 to inspect the parameter distributions of photometrically classified sources in more detail. A model can only be as good as the data you train it on. We find that although it is rare to find objects in the labeled data set with different PTF OIDs that correspond to the same physical object (0.01 % within two arcseconds), objects in the full set of fitted objects have up to eight neighbours within two arcseconds. This is a problem that not only decreases performance, as the parameters are fitted using fewer epochs – it also complicates performance evaluation because a model can accidentally be tested on the same astrophysical sources that it was trained on. This makes it too optimistic when we estimate how well the model would perform on unknown data. The bias is negligible in this thesis, but it would be important to account for in future studies with larger labeled sets. We also see signs of blended sources where PTF assigns the same OID to multiple objects. It might help to exclude sources with close neighbours or create light curves using unsupervised learning. In PTF, we find most objects (fullfilling the criteria of the data cleaning) to be either within 1 arcsecond or further away than 4 arcseconds of another source.

We also discuss the use of different machine learning models. Semi-supervised learning might improve performance for smaller data sets. We experimented with `LogisticRegression`, but this resulted in a slower model with poorer performance. It was also more complicated to implement in a way that identified non-linear patterns and gave calibrated probabilities.

Regarding input features, there are many possible changes to test in future research. We provide a list of suggestions in Sect. 6.2.2 including the prediction of photometric redshifts, changes to the computation of variability parameters and automatic feature engineering. With a large number of features, feature reduction could be important to avoid the curse of dimensionality. Different evaluation of feature importance with less bias for correlated features would also be interesting.

Finally, studies can be performed on different data with different potential for distinguishing variable objects. Observing fainter objects will influence class balance and using different colours or multi-colour light curves changes the relative importance of variability and colour features. Different types of information could be added such as proper motions and stellarity. A combination of low proper motion and high stellarity would increase the probability of being a quasar, for example. The combination of light curves from multiple surveys is especially interesting for the application of advanced variability models. Fitting the characteristic timescale of DRW models requires long baselines. Depending on the purpose of the study, using more labels could create a model with the ability to classify interesting subclasses. Using labels from multiple surveys or methods would make the model more robust to the biases of each one. One might be able to detect low luminosity AGN as predicted galaxies with high A and γ in the created catalogue, which could be tested using a set of known AGN.

In large future surveys such as the LSST, it is important to optimise computational resources. The most resource intensive part of this thesis is the MCMC fitting of power law structure functions. This is, however, still cheap compared to taking spectroscopy. Fitting 71 million light curves took 6 months, so by a naive estimate, fitting 40 billion LSST sources would require 563 times as many resources. That is, ~ 300 years with the same setup or ~ 50 years using all HPC nodes available at DARK; leaving none for other research. However, the light curves would also contain more epochs. To limit an increase in run time from longer light curves, and perhaps achieve more reliable power law SF parameters, an upper limit could be placed on $|\Delta t_{ij}|$ during fitting. One could also limit the science by stricter data cleaning constraints to only estimate the variability of sources where this is expected to be most useful. We would recommend testing different variability measures to compare and evaluate the trade off between performance and resources. Different fitting techniques also could lead to faster parameter estimation of sufficient quality. The 10-year, 800-epoch LSST light curves will provide high quality variability information – we just need to consider how to best extract and use the information given the size of the data set. The machine learning model of this thesis is already highly efficient and suitable for large surveys.

Acronyms

ACF	autocorrelation function	63
AGN	active galactic nuclei	20
AR	autoregressive process	61
ARMA	autoregressive–moving-average	61
AUC	Area Under the Curve	127
BAO	baryon acoustic oscillations	20
CAR	continuous autoregressive process	62
CARMA	continuous autoregressive–moving-average	21
CRTS	Catalina Real-time Transient Survey	165
DBSCAN	Density-Based Spatial Clustering of Applications with Noise	161
Dec	declination	46
DRW	damped random walk	21
ELT	Extremely Large Telescope	20
FPR	false positive rate	127
FSRQ	flat-spectrum radio quasars	43
GMT	Giant Magellan Telescope	20
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise	21
HPC	High Performance Computing Centre	71

IRSA	NASA/IPAC Infrared Science Archive	72
JD	Julian Date	47
LSST	Vera C. Rubin Observatory Legacy Survey of Space and Time	20
MA	moving average	61
MAD	median absolute deviation	85
MAP	maximum a posteriori probability	51
MCMC	Markov Chain Monte Carlo	21
MDI	Mean Decrease in Ímpurity	164
MJD	Modified Julian Date	47
MLE	maximum likelihood estimation	51
MWM	moving weighted median	85
OID	object ID	53
OvR	one versus rest	124
OvO	one versus one	127
OVV	Optically Violent Variable	43
PDF	probability density function	52
PS1	Pan-STARRS1	20
PSD	power spectral density	60
PTF	Palomar Transient Factory	22
QSO	quasar	80
RA	right ascension	46
RMS	root mean square	58
ROC	Receiver Operating Characteristic	127
SDSS	Sloan Digital Sky Survey	20
SF	structure function	21
SIMBAD	Set of Identifications, Measurements and Bibliography for Astronomical Data	29

TAP	Table Access Protocol	72
TMT	Thirty Meter Telescope	20
TPR	true positive rate	127
UT	Universal Time	47
WISE	Wide-field Infrared Survey Explorer	22
ZTF	Zwicky Transient Facility	20

Bibliography

Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2009, ApJS, 182, 543

Abdurro'uf, Accetta, K., Aerts, C., et al. 2022, ApJS, 259, 35

Ajello, M., Romani, R. W., Gasparrini, D., et al. 2014, ApJ, 780, 73

Akeson, R., Armus, L., Bachelet, E., et al. 2019, arXiv e-prints, arXiv:1902.05569

Alam, S., Aubert, M., Avila, S., et al. 2021, Phys. Rev. D, 103, 083533

Altman, N. & Krzywinski, M. 2018, Nature Methods, 15, 399

Alves, C. S., Leite, A. C. O., Martins, C. J. A. P., Matos, J. G. B., & Silva, T. A. 2019, MNRAS, 488, 3607

Amoukou, S. I., J-B. Brunel, N., & Salaün, T. 2021, arXiv e-prints, arXiv:2106.03820

Ansari, Z., Agnello, A., & Gall, C. 2021, A&A, 650, A90

Ansari, Z., Gall, C., Wesson, R., & Krause, O. 2022, A&A, 666, A176

Antonucci, R. 1993, ARA&A, 31, 473

Arya, S. & Mount, D. M. 1998

Ashworth, S. H. 2012, Contemporary Physics, 53, 275

Assef, R. J., Stern, D., Kochanek, C. S., et al. 2013, ApJ, 772, 26

Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, AJ, 156, 123

Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33

Baldassare, V. F., Geha, M., & Greene, J. 2020, ApJ, 896, 10

Ball, N. M. & Brunner, R. J. 2010, International Journal of Modern Physics D, 19, 1049

Barlow, R. 1989, Statistics. A guide to the use of statistical methods in the physical sciences

Baron, D. 2019, arXiv e-prints, arXiv:1904.07248

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. 2004, SIGKDD Explor., 6, 20

Bauer, A., Baltay, C., Coppi, P., et al. 2009, ApJ, 696, 1241

Baum, W. A. 1962, in Problems of Extra-Galactic Research, ed. G. C. McVittie, Vol. 15, 390

Bayes, M. & Price, M. 1763, Philosophical Transactions of the Royal Society of London Series I, 53, 370

Bellm, E. C., Kulkarni, S. R., Graham, M. J., et al. 2019, PASP, 131, 018002

Belokurov, V., Evans, N. W., & Du, Y. L. 2003, MNRAS, 341, 1373

Bentley, J. L. 1975, Commun. ACM, 18, 509–517

Bentz, M. C., Denney, K. D., Grier, C. J., et al. 2013, ApJ, 767, 149

Bertin, E. & Arnouts, S. 1996, A&AS, 117, 393

Blanton, M. R., Bershady, M. A., Abolfathi, B., et al. 2017, AJ, 154, 28

Blomqvist, M., du Mas des Bourboux, H., Busca, N. G., et al. 2019, A&A, 629, A86

Bobrick, A., Iorio, G., Belokurov, V., et al. 2022, arXiv e-prints, arXiv:2208.04332

Boch, T., Pineau, F., & Derriere, S. 2012, in Astronomical Society of the Pacific Conference Series, Vol. 461, Astronomical Data Analysis Software and Systems XXI, ed. P. Ballester, D. Egret, & N. P. F. Lorente, 291

Bonnarel, F., Fernique, P., Bienaymé, O., et al. 2000, A&AS, 143, 33

Breiman, L. 2001, Machine Learning, 45, 5

Bruun, S. H. 2017, BSc thesis, University of Copenhagen, Copenhagen, available upon request

Bruun, S. H. 2020, MSc thesis, University of Copenhagen, Copenhagen, available upon request

Bruun, S. H., Agnello, A., & Hjorth, J. 2023a, A&A, in press, arXiv:

Bruun, S. H., Hjorth, J., & Agnello, A. 2023b, A&A, submitted, in review, arXiv:

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

Cabral, J. B., Ramos, F., Gurovich, S., & Granitto, P. M. 2020, A&A, 642, A58

Cartier, R., Lira, P., Coppi, P., et al. 2015, ApJ, 810, 164

Chambers, K. C., Magnier, E. A., Metcalfe, N., et al. 2016, arXiv e-prints, arXiv:1612.05560

Chang, K. & Refsdal, S. 1979, Nature, 282, 561

Chary, R., Helou, G., Brammer, G., et al. 2020, arXiv e-prints, arXiv:2008.10663

Chen, T. & Guestrin, C. 2016, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’16 (New York, NY, USA: Association for Computing Machinery), 785–794

Chib, S. & Greenberg, E. 1995, The American Statistician, 49, 327

Ciucă, I. & Ting, Y.-S. 2023, arXiv e-prints, arXiv:2304.05406

Clarke, A. O., Scaife, A. M. M., Greenhalgh, R., & Griguta, V. 2020, A&A, 639, A84

Cox, J. P. 1963, ApJ, 138, 487

Cuillandre, J.-C., Luppino, G. A., Starr, B. M., & Isani, S. 2000, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 4008, Optical and IR Telescope Instrumentation and Detectors, ed. M. Iye & A. F. Moorwood, 1010–1021

Cunha, P. A. C. & Humphrey, A. 2022, A&A, 666, A87

Cutri, R. M., Wright, E. L., Conrow, T., et al. 2021, VizieR Online Data Catalog, II/328

Davis, M. & Peebles, P. J. E. 1983, ApJ, 267, 465

De Cicco, D., Bauer, F. E., Paolillo, M., et al. 2021, A&A, 645, A103

De Cicco, D., Bauer, F. E., Paolillo, M., et al. 2022, A&A, 664, A117

De Cicco, D., Paolillo, M., Covone, G., et al. 2015, A&A, 574, A112

De Cicco, D., Paolillo, M., Falocco, S., et al. 2019, A&A, 627, A33

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

Dice, L. R. 1945, Ecology, 26, 297

Dinnbier, F., Anderson, R. I., & Kroupa, P. 2022, A&A, 659, A169

Djorgovski, S. G., Mahabal, A. A., Graham, M. J., Polsterer, K., & Krone-Martins, A. 2022, arXiv e-prints, arXiv:2212.01493

Drory, N., MacDonald, N., Bershady, M. A., et al. 2015, AJ, 149, 77

Du, P., Wang, J.-M., & Zhang, Z.-X. 2017, ApJ, 840, L6

du Mas des Bourboux, H., Rich, J., Font-Ribera, A., et al. 2020, ApJ, 901, 153

Eyer, L. & Mowlavi, N. 2008, in Journal of Physics Conference Series, Vol. 118, Journal of Physics Conference Series, 012010

Fawcett, T. 2006, Pattern Recognition Letters, 27, 861, rOC Analysis in Pattern Recognition

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

Freedman, R. A., Geller, R. M., & Kaufmann, III, W. J. 2014, Universe, 10th edn. (New York: W.H. Freeman and Company)

Freeth, T., Jones, A., Steele, J. M., & Bitsakis, Y. 2008, Nature, 454, 614

Friedman, J. H. 2001, The Annals of Statistics, 29, 1189

Friedman, J. H., Bentley, J. L., & Finkel, R. A. 1976, ACM Trans. Math. Softw., 3, 209

Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., et al. 2016, A&A, 595, A1

Gaia Collaboration, Vallenari, A., Brown, A. G. A., et al. 2022, arXiv e-prints, arXiv:2208.00211

Geman, S. & Geman, D. 1984, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6, 721

Gillespie, D. T. 1996, American Journal of Physics, 64, 225

Ginsburg, A., Sipőcz, B. M., Brasseur, C. E., et al. 2019, AJ, 157, 98

Goodman, J. & Weare, J. 2010, Comm. App. Math. Comp. Sci., 5, 65–80

Griest, K. & Safizadeh, N. 1998, ApJ, 500, 37

Grinsztajn, L., Oyallon, E., & Varoquaux, G. 2022, arXiv e-prints, arXiv:2207.08815

Groves, B. A., Heckman, T. M., & Kauffmann, G. 2006, MNRAS, 371, 1559

Gunn, J. E., Siegmund, W. A., Mannery, E. J., et al. 2006, AJ, 131, 2332

Hand, D. J. & Till, R. J. 2001, Machine Learning, 45, 171

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

Hopkins, P. F., Hayward, C. C., Narayanan, D., & Hernquist, L. 2012, MNRAS, 420, 320

Hughes, P. A., Aller, H. D., & Aller, M. F. 1992, ApJ, 396, 469

Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111

Ivezić, Ž. & MacLeod, C. 2014, in Multiwavelength AGN Surveys and Studies, ed. A. M. Mickaelian & D. B. Sanders, Vol. 304, 395–398

Iwanek, P., Kozłowski, S., Gromadzki, M., et al. 2021, ApJS, 257, 23

Jamal, S. & Bloom, J. S. 2020, ApJS, 250, 30

Jamieson, K. & Talwalkar, A. 2016, in Proceedings of Machine Learning Research, Vol. 51, Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, ed. A. Gretton & C. C. Robert (Cadiz, Spain: PMLR), 240–248

Jeffreys, H. 1946, Proceedings of the Royal Society of London Series A, 186, 453

Jetsu, L., Porceddu, S., Lyytinen, J., et al. 2013, ApJ, 773, 1

Johns, M., McCarthy, P., Raybould, K., et al. 2012, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 8444, Ground-based and Airborne Telescopes IV, ed. L. M. Stepp, R. Gilmozzi, & H. J. Hall, 84441H

Josse, J., Prost, N., Scornet, E., & Varoquaux, G. 2019, arXiv e-prints, arXiv:1902.06931

Kanter, J. M. & Veeramachaneni, K. 2015, in 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 1–10

Kapoor, S. & Narayanan, A. 2022, arXiv e-prints, arXiv:2207.07048

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

Ke, G., Meng, Q., Finley, T., et al. 2017, in Advances in Neural Information Processing Systems, ed. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett, Vol. 30 (Curran Associates, Inc.)

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

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

Kim, A. G., Linder, E. V., Edelstein, J., & Erskine, D. 2015, Astroparticle Physics, 62, 195

Kochanek, C. S., Mochejska, B., Morgan, N. D., & Stanek, K. Z. 2006, ApJ, 637, L73

Koo, D. C. 1985, AJ, 90, 418

Kormendy, J. & Ho, L. C. 2013, Annual Review of Astronomy and Astrophysics, 51, 511

Kozłowski, S. 2016a, MNRAS, 459, 2787

Kozłowski, S. 2016b, ApJ, 826, 118

Kozłowski, S. 2017a, ApJ, 835, 250

Kozłowski, S. 2017b, A&A, 597, A128

Kraft, R. P. 1960, ApJ, 131, 330

La Mura, G., Busetto, G., Ciroi, S., et al. 2017, European Physical Journal D, 71, 95

Landolt, A. U. 1992, AJ, 104, 340

Lang, D., Hogg, D. W., & Schlegel, D. J. 2016, AJ, 151, 36

Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv e-prints, arXiv:1110.3193

Law, N. M., Kulkarni, S. R., Dekany, R. G., et al. 2009, PASP, 121, 1395

Le Morvan, M., Josse, J., Moreau, T., Scornet, E., & Varoquaux, G. 2020, arXiv e-prints, arXiv:2007.01627

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. 2016, arXiv e-prints, arXiv:1604.04173

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. 2016, arXiv e-prints, arXiv:1603.06560

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

Loeb, A. 1998, ApJ, 499, L111

Logan, C. H. A. & Fotopoulou, S. 2020, A&A, 633, A154

Lones, M. A. 2021, arXiv e-prints, arXiv:2108.02497

Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. 2013, in

Lundberg, S. & Lee, S.-I. 2017, arXiv e-prints, arXiv:1705.07874

MacLeod, C. L., Green, P. J., Anderson, S. F., et al. 2019, ApJ, 874, 8

MacLeod, C. L., Ivezić, Ž., Kochanek, C. S., et al. 2010, ApJ, 721, 1014

Maddox, N. & Hewett, P. C. 2006, MNRAS, 367, 717

Maiolino, R., Scholtz, J., Witstok, J., et al. 2023, arXiv e-prints, arXiv:2305.12492

Mamontov, D., Minker, W., & Karpov, A. 2022, in Speech and Computer, ed. S. R. M. Prasanna, A. Karpov, K. Samudravijaya, & S. S. Agrawal (Cham: Springer International Publishing), 464–476

Maneewongvatana, S. & Mount, D. M. 1999, arXiv e-prints, cs/9901013

Mase, M., Owen, A. B., & Seiler, B. B. 2021, arXiv e-prints, arXiv:2105.07168

Mattei, J. A. 1997, , 25, 57

McInnes, L., Healy, J., & Astels, S. 2017, The Journal of Open Source Software, 2

Meisner, A. M., Lang, D., & Schlegel, D. J. 2017, AJ, 154, 161

Möller, O., Kitzbichler, M., & Natarajan, P. 2007, MNRAS, 379, 1195

Montalban, J. & Miglio, A. 2008, Communications in Asteroseismology, 157, 160

Moreno, J., Vogeley, M. S., Richards, G. T., & Yu, W. 2019, PASP, 131, 063001

Mukai, K. 2017, PASP, 129, 062001

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

Myers, A. D., Palanque-Delabrouille, N., Prakash, A., et al. 2015, ApJS, 221, 27

Nature Astronomy editorial. 2023, Nature Astronomy, 7, 1

Nicodemus, K. K., Malley, J. D., Strobl, C., & Ziegler, A. 2010, BMC Bioinformatics, 11, 110

Ochsenbein, F., Bauer, P., & Marcout, J. 2000, A&AS, 143, 23

Oguri, M. & Marshall, P. J. 2010, MNRAS, 405, 2579

Palanque-Delabrouille, N., Yeche, C., Myers, A. D., et al. 2011, A&A, 530, A122

Pancino, E., Marrese, P. M., Marinoni, S., et al. 2022, A&A, 664, A109

Paris, I., Petitjean, P., Ross, N. P., et al. 2017, VizieR Online Data Catalog, VII/279

Pâris, I., Petitjean, P., Ross, N. P., et al. 2017, A&A, 597, A79

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825

Peters, C. M., Richards, G. T., Myers, A. D., et al. 2015, ApJ, 811, 95

Peterson, B. M., Ferrarese, L., Gilbert, K. M., et al. 2004, ApJ, 613, 682

Raj, A. & Nixon, C. J. 2021, ApJ, 909, 82

Ramos Almeida, C., Levenson, N. A., Alonso-Herrero, A., et al. 2011, ApJ, 731, 92

Rau, A., Kulkarni, S. R., Law, N. M., et al. 2009, PASP, 121, 1334

Rees, M. J. 1984, ARA&A, 22, 471

Refsdal, S. 1964, MNRAS, 128, 307

Ricker, G. R., Winn, J. N., Vanderspek, R., et al. 2014, in Space Telescopes and Instrumentation 2014: Optical, Infrared, and Millimeter Wave, ed. J. M. O. Jr., M. Clampin, G. G. Fazio, & H. A. MacEwen, Vol. 9143, International Society for Optics and Photonics (SPIE), 914320

Riess, A. G., Casertano, S., Yuan, W., Macri, L. M., & Scolnic, D. 2019, ApJ, 876, 85

Risaliti, G. & Lusso, E. 2019, Nature Astronomy, 3, 272

Ross, N. P., Ford, K. E. S., Graham, M., et al. 2018, MNRAS, 480, 4468

Salvato, M., Hasinger, G., Ilbert, O., Zamorani, G., & COSMOS Team. 2009, in American Astronomical Society Meeting Abstracts, Vol. 213, American Astronomical Society Meeting Abstracts #213, 612.03

Sánchez, P., Lira, P., Cartier, R., et al. 2017, ApJ, 849, 110

Sánchez-Sáez, P., Lira, P., Cartier, R., et al. 2019, ApJS, 242, 10

Sánchez-Sáez, P., Lira, P., Mejía-Restrepo, J., et al. 2018, ApJ, 864, 87

Sandage, A. 1962, ApJ, 136, 319

Scargle, J. D. 1981, ApJS, 45, 1

Schmidt, K. B., Marshall, P. J., Rix, H.-W., et al. 2010, ApJ, 714, 1194

Schmidt, K. B., Rix, H.-W., Shields, J. C., et al. 2012, ApJ, 744, 147

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. 2019, arXiv e-prints, arXiv:1907.10597

Secrest, N. J., von Hausegger, S., Rameez, M., et al. 2021, ApJ, 908, L51

Sesar, B., Ivezić, Ž., Lupton, R. H., et al. 2007, AJ, 134, 2236

Settles, B. 2009, Active Learning Literature Survey, Computer Sciences Technical Report 1648, University of Wisconsin–Madison

Shapley, L. S. 1951, Notes on the N-Person Game — II: The Value of an N-Person Game (Santa Monica, CA: RAND Corporation)

Shappee, B. J., Prieto, J. L., Grupe, D., et al. 2014, ApJ, 788, 48

Sharma, S. 2017, Annual Review of Astronomy and Astrophysics, 55, 213

Shen, Y. 2021, ApJ, 921, 70

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

Skidmore, W., TMT International Science Development Teams, & Science Advisory Committee, T. 2015, Research in Astronomy and Astrophysics, 15, 1945

Skowron, D. M., Skowron, J., Mróz, P., et al. 2019, Science, 365, 478

Smee, S. A., Gunn, J. E., Uomoto, A., et al. 2013, AJ, 146, 32

Smith, M. J. & Geach, J. E. 2022, arXiv e-prints, arXiv:2211.03796

Song, H., Park, C., Lietzen, H., & Einasto, M. 2016, ApJ, 827, 104

Sørensen, T. 1948, A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons, Biologiske skrifter (in commission by E. Munksgaard)

Sparke, L. S. & Gallagher, III, J. S. 2007, Galaxies in the Universe: An Introduction, 2nd edn. (Cambridge University Press)

Stetson, P. B. 2000, PASP, 112, 925

Stetson, P. B., Pancino, E., Zocchi, A., Sanna, N., & Monelli, M. 2019, MNRAS, 485, 3042

Stone, M. 1974, Journal of the Royal Statistical Society: Series B (Methodological), 36, 111

Stone, Z., Shen, Y., Burke, C. J., et al. 2022, MNRAS, 514, 164

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. 2007, BMC Bioinformatics, 8, 25

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

Szkody, P. 2021, Frontiers in Astronomy and Space Sciences, 8, 184

Taak, Y. C. & Treu, T. 2023, arXiv e-prints, arXiv:2304.02784

Taghizadeh-Popp, M., Kim, J. W., Lemson, G., et al. 2020, Astronomy and Computing, 33, 100412

Tamai, R., Cirasuolo, M., González, J. C., Koehler, B., & Tuti, M. 2016, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9906, Ground-based and Airborne Telescopes VI, ed. H. J. Hall, R. Gilmozzi, & H. K. Marshall, 99060W

Tanaka, M. 2015, ApJ, 801, 20

Tonry, J. L., Stubbs, C. W., Lykke, K. R., et al. 2012, ApJ, 750, 99

Treiber, H. P., Hinkle, J. T., Fausnaugh, M. M., et al. 2022, arXiv e-prints, arXiv:2209.15019

Treu, T. & Marshall, P. J. 2016, The Astronomy and Astrophysics Review, 24, 11

Trump, J. R., Sun, M., Zeimann, G. R., et al. 2015, ApJ, 811, 26

Tsujimoto, M., Koyama, K., Kobayashi, N., et al. 2003, AJ, 125, 1537

Turner, E. L. 1991, AJ, 101, 5

Twala, B., Jones, M., & Hand, D. 2008, Pattern Recognition Letters, 29, 950

Ulrich, M.-H., Maraschi, L., & Urry, C. M. 1997, ARA&A, 35, 445

Urry, C. M. & Padovani, P. 1995, PASP, 107, 803

Usher, P. D. 1978, ApJ, 222, 40

van den Bergh, S., Herbst, E., & Pritchet, C. 1973, AJ, 78, 375

Velten, H. & Gomes, S. 2020, Phys. Rev. D, 101, 043502

Wang, C., Wu, Q., Weimer, M., & Zhu, E. 2019, arXiv e-prints, arXiv:1911.04706

Wang, F., Yang, J., Fan, X., et al. 2021, ApJ, 907, L1

Ward, C., Gezari, S., Nugent, P., et al. 2021, arXiv e-prints, arXiv:2110.13098

Watson, D., Denney, K. D., Vestergaard, M., & Davis, T. M. 2011, ApJ, 740, L49

Wilson, J. C., Hearty, F. R., Skrutskie, M. F., et al. 2019, PASP, 131, 055001

Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868

Yang, Q., Wu, X.-B., Fan, X., et al. 2017, AJ, 154, 269

Yoo, A. B., Jette, M. A., & Grondona, M. 2003, in Job Scheduling Strategies for Parallel Processing, ed. D. Feitelson, L. Rudolph, & U. Schwiegelshohn (Berlin, Heidelberg: Springer Berlin Heidelberg), 44–60

Zhevakin, S. A. 1959, Soviet Ast., 3, 389

Appendix: Illustration

To compare parameter distributions of three classes in a single plot, we create 2D histograms for each class, scaled according to the maximum value in parameter space (see for example Fig. 4.2). The output values are used as inputs to an RGB matrix with the density matrix of each class corresponding to a colour channel. The values in each channel are then rescaled to span from 0 to 1 for the maximum value in the density matrix. We could stop here by rescaling to 0 to 255, but this would leave the blue channel difficult to distinguish from the black background. Inverting the colours is easy, but then the yellow is difficult to distinguish from white. So for each pixel, we convert the background from black to white. We do this using the `colormath` package in Python [1]. We first convert the RGB colours to CMYK colours. We want to increase the saturation of CMY in CMYK to keep K (black) separate. So we invert the colours and increase the saturation with

$$c_{new} = \log_{10}(c + 0.1) - \log_{10}(1.1) * c_{old} - 1.027 \tag{7.1}$$

where c is the colour channel. We convert back to RGB and multiply by 255. Increasing the saturation makes it easier to spot colour differences in areas with low relative frequencies of all classes.

Using primary colours makes interpretation easier, as we use a large colour space, unless one is colour blind. One way to check the colour accessibility is to upload the diagrams to a website dedicated to this purpose. We recommend Coblis[2].

[1] Taylor, Gregory, 2014, `python-colormath.readthedocs.org/`

[2] Wickline, Matthew, and the Human-Computer Interaction Resource Network, 2000, `www.color-blindness.com/coblis-color-blindness-simulator/`