An Automated Pipeline for Variability Detection and Classification for the Small Telescopes Installed at the Liverpool Telescope

Paul Ross McWhirter

A thesis submitted in partial fulfilment of the requirements of
Liverpool John Moores University
for the degree of Doctor of Philosophy.

July 2018

Abstract

The Small Telescopes at the Liverpool Telescope (STILT) is an almost decade old project to install a number of wide field optical instruments to the Liverpool Telescope, named Skycams, to monitor weather conditions and yield useful photometry on bright astronomical sources. The motivation behind this thesis is the development of algorithms and techniques which can automatically exploit the data generated during the first 1200 days of Skycam operation to catalogue variable sources in the La Palma sky. A previously developed pipeline reduces the Skycam images and produces photometric time-series data named light curves of millions of objects. 590,492 of these objects have 100 or more data points of sufficient quality to attempt a variability analysis. The large volume and relatively high noise of this data necessitated the use of Machine Learning and sophisticated optimisation techniques to successfully extract this information.

The Skycam instruments have no control over the orientation and pointing of the Liverpool Telescope and therefore resample areas of the sky highly irregularly. The term used for this resampling in astronomy is cadence. The unusually irregular Skycam cadence places increased strain on the algorithms designed for the detection of periodicity in light curves. This thesis details the development of a period estimation method based on a novel implementation of a genetic algorithm combined with a generational clustering method. Named GRAPE (Genetic Routine for Astronomical Period Estimation), this algorithm deconstructs the space of possible periods for a light curve into regions in which the genetic population clusters. These regions are then fine-tuned using a k-means clustering algorithm to return a set of independent period candidates which are then analysed using a Vuong closeness test to discriminate between aliased and true periods. This thesis demonstrates the capability of GRAPE on a set of synthetic light curves built using traditional regular cadence sampling and Skycam style cadence for four different shapes of periodic light curve. The performance of GRAPE on these light curves is compared to a more traditional periodogram which returns a set of peaks and is then analysed using Vuong closeness tests. GRAPE obtains similar performance compared to the periodogram on all the light curve shapes but with less computational complexity allowing for more efficient light curve analysis.

Automated classification of variable light curves has been explored over the last decade. Multiple features have been engineered to identify patterns in the light curves of different classes of variable star. Within the last few years deep learning has come to prominence as a method of automatically generating informative representations of the data for the solution of a desired problem, such as a classification task. A set of models using Random Forests, Support Vector Machines and Neural Networks were trained using a set of variable Skycam light curves of five classes. Using 16 features engineered from previous methods an Area under the Curve (AUC) of 0.8495 was obtained. Replacing these features with inputs from the pixel intensities from a 100 by 20 pixel image representation, produced an AUC of 0.6348, which improved to 0.7952 when provided with additional context to the dimensionality of the image. Despite the inferior performance, the importance of the different pixels produced relations in the trained models demonstrating that they had produced features based on well-understood patterns in the different classes of light curve.

Using features produced by Richards et al. and Kim & Bailer-Jones et al., a set of features to train machine learning classification models was constructed. In addition to this set of features, a semi-supervised set of novel features was designed to describe the shape of light curves phased around the GRAPE candidate period. This thesis investigates the performance of the PolyFit algorithm of Prsa et al., a technique to fit four piecewise polynomials with discontinuous knots capable of connecting across the phase boundary at phases of zero and one. This method was designed to fit eclipsing binary phased light curves however were also described to be fully capable on other variable star types. The optimisation method used by PolyFit is replaced by a novel genetic algorithm optimisation routine to fit the model to Skycam data with substantial improvement in performance. The PolyFit model is applied to the candidate period and twice this period for every classified light curve. This interpolation produces novel features which describe similar statistics to the previously developed methods but which appear significantly more resilient to the Skycam noise and are often preferred by the trained models. In addition, Principal Component Analysis (PCA) is used to investigate a set of 6897 variable light curves and discover that the first ten principal components are sufficient to describe 95% of the variance of the fitted models. This trained PCA model is retained and used to generate twenty novel shape features. Whilst these features are not dominant in their importance to the learned models, they have above average importance and help distinguish some objects in the light curve classification task. The second principal component in particular is an important feature in the discrimination of short period pulsating and eclipsing variables as it appears to be an automatically learned robust skewness measure.

The method described in this thesis produces 112 features of the Skycam light curves, 38 variability indices which are quickly obtainable and 74 which require the computation of a candidate period using GRAPE. A number of machine learning classifiers are investigated to produce high-performance models for the detection and classification of variable light curves from the Skycam dataset. A Random Forest classifier uses a training set of 859 light curves of 12 object classes to produce a classifier with a multi-class F1 score of 0.533. It would be computationally infeasible to produce all the features for every Skycam light curve, therefore an automated pipeline has been developed which combines a Skycam trend removal pipeline, GRAPE and our machine learned classifiers. It initialises with a set of Skycam light curves from objects cross-matched from the American Association of Variable Star Observers (AAVSO) Variable Star Index (VSI), one of the most comprehensive catalogues of variable stars available. The learned models classify the full 112 features generated for these cross-matched light curves and confident matches are selected to produce a training set for a binary variability detection model. This model utilises only the 38 variability indices to identify variable light curves rapidly without the use of GRAPE. This variability model, trained using a random forest classifier, obtains an F1 score of 0.702. Applying this model to the 590,492 Skycam light curves yields 103,790 variable candidates of which 51,129 candidates have been classified and are available for further analysis.

Publications

In the course of completeing the work presented in this thesis, the content of Chapter 3 has been accepted for publication in refereed journals:

McWhirter, P. R., Steele, I. A., Hussian, A., Al-Jumeily, D. & Vellasco, M. M. B. R., 2018, 'GRAPE: Genetic Routine for Astronomical Period Estimation', *Monthly Notices of the Royal Astronomical Society.*

In the course of completeing the work presented in this thesis, the contents of Chapter 3 & 5 have been submitted for publication in refereed journals:

McWhirter, P. R., Hussian, A., Al-Jumeily, D., Fergus, P., Steele, I. A. & Vellasco, M. M. B. R., 2018, 'Classifying Periodic Astrophysical Phenomena from non-survey optimized variable-cadence observational data', *Expert Systems with Applications.*

In the course of completeing the work presented in this thesis, the contents of Chapter 5 (Fourier Decomposition section) & 6 section 1 have been published in conference proceedings:

McWhirter, P. R., Wright, S., Steele, I. A., Al-Jumeily, D., Hussian, A. & Fergus, P., 2016, 'A dynamic, modular intelligent-agent framework for astronomical light curve analysis and classification', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9771, 820–831.

McWhirter, P. R., Steele, I. A., Al-Jumeily, D., Hussian, A. & Vellasco, M. M. B. R., 2017, 'The classification of periodic light curves from non-survey optimized observational data through automated extraction of phase-based visual features', *Proceedings of the International Joint Conference on Neural Networks (2017) 2017-May*, 3058–3065.

In the course of completeing the work presented in this thesis, the contents of Chapter 6, section 2 are in preparation for publication in refereed journals:

McWhirter, P. R., Steele, I. A., Hussian, A., Al-Jumeily, D. & Vellasco, M. M. B. R., 2018, 'Representation Learning on light curves using the PolyFit algorithm'.

Acknowledgements

I would like to thank Dr. Paul Fergus for starting me on the road to this PhD by securing my funding. This research would have been impossible without this contribution and whilst different supervisors were ultimately selected, the occasional advice was very much appreciated. I would also like to thank my Computer Science supervisors, Prof. Dhiya Al-Jumeily and Prof. Abir Hussain for managing the administration of this research and securing funding for an exchange visit to Rio de Janeiro in 2016 which substantially boosted my personal knowledge and the direction of this study. Further to this, I would like to thank Prof Marley Vellasco from the Pontifical Catholic University in Rio de Janeiro for being so welcoming and helpful in the development of the Genetic Algorithms in this work.

Very importantly I would like to thank my Astronomy supervisor Prof. Iain Steele. Without his efforts this project would have never started or progressed as efficiently as it has and the supportive environment provided in the Liverpool Telescope group has kept me focused and productive over these last few years. Additionally, I would like to thank the community at the Astrophysics Research Institute as a whole for being so welcoming to a researcher from another department and helping me fulfil my lifelong plan of producing a PhD thesis in Astronomy. I would also like to thank my colleagues in the Applied Computing Research Group who, together, have braved the PhD life with me.

More personally, I would like to thank my mother and sister, Hazel and Ellen, for their support and love through the highs and lows of the last few years. They have had to put up with some impressive rants and musings so I am extremely thankful for the support they have given even though my mum has not been a statistician for many years! I also thank my friend and flatmate Luke and my long-time friend Dane back in

Northern Ireland. Finally, I would like to thank and dedicate this thesis to my father Paul McWhirter. He passed away years before he could see my PhD dream fulfilled but I know he would be very proud that I got there in the end.

Additional thanks go to the Royal Academy of Engineering Newton Fund which supported the collaborative project with Brazil and OPTICON (Optical Infrared Coordination Network for Astronomy) for funding the NEON Observing School 2017 which I attended in La Palma in September 2017. The software produced for this thesis was developed using the language R in the integrated developer environment RStudio with mathematical extensions developed by Intel and I thank all those who maintain this open source software. Finally, thanks go again to my mother who supplied funding to assist my initial stay in Liverpool for my Masters Programme and extra money for equipment to assist in the data reduction contained within this thesis.

“ – it reaches out it reaches out it reaches out it reaches out – One hundred and thirteen times a second, nothing answers and it reaches out. It is not conscious, though parts of it are. There are structures within it that were once separate organisms; aboriginal, evolved, and complex. It is designed to improvise, to use what is there and then move on. Good enough is good enough, and so the artifacts are ignored or adapted. The conscious parts try to make sense of the reaching out. Try to interpret it.”

– James S. A. Corey, Cibola Burn

Contents

List of Figures

List of Tables

Abbreviations

AAVSO	**A**merican **A**ssociation of **V**ariable **S**tar **O**bservers
ACF	**A**uto**C**orrelation **F**unction
AGB	**A**symptotic **G**iant **B**ranch
AGN	**A**ctive **G**alactic **N**uclei
ANN	**A**rtificial **N**eural **N**etwork
AoV	**A**nalysis **o**f **V**ariance
ASAS	**A**ll **S**ky **A**utomated **S**urvey
AUC	**A**rea **U**nder the **C**urve
BGLS	**B**ayesian **G**eneralised **L**omb **S**cargle periodogram
BLAP	**B**lue **L**arge **A**mplitude **P**ulsator
BKR	**B**lum **K**iefer **R**osenblatt
BS	**B**rier **S**core
CART	**C**lassification **A**nd **R**egression **T**ree
CCD	**C**harge **C**ouple **D**evice
CE	**C**onditional **E**ntropy
CKP	**C**orrentropy **K**ernelised **P**eriodogram
CNN	**C**onvolutional **N**eural **N**etwork
CPU	**C**entral **P**rocessing **U**nit
CRTS	**C**atalina **R**eal-Time **T**ransient **S**urvey
CV	**C**ross **V**alidation
DB	**D**ata**b**ase
DBMS	**D**ata**b**ase **M**anagement **S**ystem
DFT	**D**iscrete **F**ourier **T**ransform
EB	**E**clipsing **B**inary
EROS	**E**xpérience pour la **R**echerche d’**O**bjets **S**ombres

FAP	**F**alse **A**larm **P**robability
FATS	**F**eature **A**nalysis for **T**ime **S**eries
FFNN	**F**eed **F**orward **N**eural **N**etwork
FITS	**F**lexible **I**mage **T**ransport **S**ystem
FN	**F**alse **N**egative
FoM	**F**igure **o**f **M**erit
FoV	**F**ield **o**f **V**iew
FP	**F**alse **P**ositive
FWHM	**F**ull **W**idth **H**alf **M**aximum
GA	**G**enetic **A**lgorithm
GCVS	**G**eneral **C**atalogue of **V**ariable **S**tars
GLS	**G**eneralised **L**omb **S**cargle periodogram
GPU	**G**raphics **P**rocessing **U**nit
GRAPE	**G**enetic **R**outine for **A**stronomical **P**eriod **E**stimation
HADS	**H**igh **A**mplitude **D**elta **S**cuti
HR/H–R	**H**ertzsprung–**R**ussell
IP	**I**nformation **P**otential
IR	**I**nfra**R**ed
IU	**I**ndex of **U**nion
KL	**K**ullback **L**eibler
KLIC	**K**ullback **L**eibler **I**nformation **C**riterion
KNN	**K** **N**earest **N**eighbours
LASSO	**L**east **A**bsolute **S**hrinkage and **S**election **O**perator
LBV	**L**uminous **B**lue **V**ariables
LJMU	**L**iverpool **J**ohn **M**oores **U**niversity
LPV	**L**ong **P**eriod **V**ariables
LSP	**L**omb **S**cargle **P**eriodogram
LSST	**L**arge **S**ynoptic **S**urvey **T**elescope
LT	**L**iverpool **T**elescope
MACHO	**MA**ssive **C**ompact **H**alo **O**bjects
MAD	**M**edian **A**bsolute **D**eviation
MAE	**M**ean **A**bsolute **E**rror
MBRP	**M**edian **B**uffer **R**ange **P**ercentage

MJD	**M**odified **J**ulian **D**ate
ML	**M**achine **L**earning
MUSIC	**MU**ltiple **SI**gnal **C**lassification
NB	**N**aive **B**ayes
NIR	**N**ear **I**nfra**R**ed
NSVS	**N**orthern **S**ky **V**ariability **S**urvey
OGLE	**O**ptical **G**ravitational **L**ensing **E**xperiment
OVV	**O**ptical **V**iolent **V**ariables
Pan-STARRS	**Pan**oramic **S**urvey **T**elescope and **R**apid **R**esponse **S**ystem
PC	**P**rincipal **C**omponent
PCA	**P**rincipal **C**omponent **A**nalysis
PDC	**P**hase **D**istance **C**orrelation
PDF	**P**robability **D**ensity **F**unction
PDFP	**P**ercent **D**ifference **F**lux **P**ercentile
PDM	**P**hase **D**ispersion **M**inimisation
PNNV	**P**lanetary **N**ebula **N**uclei **V**ariables
PSD	**P**ower **S**pectrum **D**ensity
QSO	**Q**uasi **S**tellar **O**bject
RBF	**R**adial **B**asis **F**unction
RCS	**R**ange of a **C**umulative **S**um
ReLU	**Re**ctified **L**inear **U**nit
RF	**R**andom **F**orest
RMSE	**R**oot **M**ean **S**quared **E**rror
RNN	**R**ecurrent **N**eural **N**etwork
roAp	**r**apidly **o**scillating **Ap** star
ROC	**R**eceiver **O**perating **C**haracteristic
RQE	**R**enyi **Q**uadratic **E**ntropy
sdB	**s**ub**d**warf **B** star
SDSS	**S**loan **D**igital **S**ky **S**urvey
SL	**S**tring **L**ength
SLLK	**S**tring **L**ength **L**afler **K**inman
SMOTE	**S**ynthetic **M**inority **O**ver-sampling **TE**chnique
SN	**S**uper**N**ova

SNR	**S**ignal to **N**oise **R**atio
SQL	**S**tructured **Q**uery **L**anguage
STILT	**S**mall **T**elescopes **I**nstalled at the **L**iverpool **T**elescope
SVM	**S**upport **V**ector **M**achine
S/N	**S**ignal-to-**N**oise
TFA	**T**rend **F**iltering **A**lgorithm
TN	**T**rue **N**egative
TP	**T**rue **P**ositive
UPSILoN	A**U**tomated Classification for **P**eriodic Variable **S**tars using Mach**I**ne **L**ear**N**ing
USNO	**U**nited **S**tates **N**aval **O**bservatory
UV	**U**ltra**V**iolet
VRP	**V**ariance **R**atio **P**eriodogram
VSI	**V**ariable **S**tar **I**ndex
WCS	**W**orld **C**oordinate **S**ystem
YSO	**Y**oung **S**tellar **O**jects
ZAMS	**Z**ero **A**ge **M**ain **S**equence

Chapter 1

Introduction

In this chapter the background of the thesis is discussed. The field of Time Domain Astronomy is introduced along with the techniques used to produce the time-series data used in this thesis in the form of photometry. The instruments used to produce the images upon which the photometry is determined are described as well as the data reduction pipeline utilised to transform these images into a catalogue of time-series data for sources identified across multiple images. The SkycamT instrument contains time-series data on hundreds of thousands of bright stars, brighter than +12 mag.

The concept of variable stars is then introduced within the context of astrophysical phenomena generating unique signatures on the time-series data and a number of the different classifiable types of variable source are defined. Variable stars are an important probe into the physics of stellar evolution, the morphology of binary stellar systems and the measurement of interstellar distances. The analysis of the time-series of the bright stars collected by the Skycam instruments allow for the production of a bright source catalogue. Bright sources are better suited to additional observations as they require less time to obtain high signal-to-noise data.

Such a bright source catalogue presents opportunities for the scientific investigation of particular classes of variable stars to test theories on stellar evolution and galaxy formation. Cepheid and RR Lyrae variables occur at specific temperatures and luminosities and changes in their variable properties over several years can reveal the timescales of their evolution. Eclipsing binaries allow the degeneracy in stellar properties as measured by photometry and spectroscopy to be broken allowing confident determination of stellar mass, radii, metallicity, surface gravity and temperature. Variable hot pulsating stars such as hot subdwarfs and white dwarfs allow for the study of the post giant-stage evolution of dying stars. These old stars contain information on the formation history of the Milky Way Galaxy and allow for investigation into the early forms of similar galaxies. The chapter concludes with the description of the aims and motivations of this thesis in the context of the background in this chapter and provides the layout for the remainder of the thesis.

1.1 Time Domain Astronomy

Time Domain Astronomy is a field of research addressing astronomical objects and phenomena which result in sky sources that exhibit photometric and spectroscopic variability with timescales detectable by telescope instrumentation. These sources can exhibit intrinsic or extrinsic variability. Intrinsic variability is due to structural changes in the light source such as stellar pulsation modes and stellar flares. Extrinsic variability is due to the obscuration of light from a background source or rotation of a source such as rotational variability, eclipses by other astronomical bodies and the formation of dust in stellar atmospheres. Analysis of these objects grants valuable information into physical processes on stellar scales up to universal scales as the properties of their variation can be used to determine the distance to the objects (through period-luminosity relations) and the chemical environment around them (through period-metallicity relations). The ability to reliably observe these light sources is rapidly improving through the development of new technological solutions, both hardware and software based.

Advances in observational, storage and data processing technologies have allowed for extended sky surveys to be conducted and exploited. These surveys range from focused observations of specific regions of the sky such as the MACHO (Alcock et al., 2000), EROS (Rahal et al., 2009) and OGLE (Udalski et al., 1997) surveys to extended sky surveys probing large swathes of the night sky such as SDSS (York et al., 2000), Pan-STARRS (Kaiser, 2002) and CRTS (Larson, 2003). This progress continues to enhance observational capability with the construction of the Large Synoptic Survey Telescope (LSST) in northern Chile due to commence operations at the beginning of the next decade (Ivezić et al., 2014). With this constant improvement in capability, the fields of Astronomy, Computer Science, Computational Intelligence and Statistics are striving to develop efficient implementations of multiple algorithms that can describe the properties of observed light sources and perform robust classification operations.

Time Domain astronomy is characterised by the large datasets generated by sky surveys containing time-series data (Vaughan, 2011). These time-series can be either photometric and/or spectroscopic data with temporal sampling over multiple epochs. This allows for the characterisation of the time-changing nature of the sources such as the determination of characteristic-timescales of variability, the periodicity of variations, amplitude changes in source brightness and changes in the signatures of chemical species allowing temperature and surface gravity variations to be analysed.

Most data-gathering exercises outside of astronomy result in a large number of observations with consistent time intervals between individual observations. In Time Domain Astronomy, it is common for these observations to have a significantly uneven distribution in time with inconsistent intervals between observations (Lomb, 1976; Scargle, 1982). Major causes of this include weather limitations that can prevent telescope operation for uncertain periods, the seasonal observability of an object which can result in unfavourable positions for part of the year and limited access time to telescopes due to the volume of astronomers requiring observations. As a result, astronomy requires data processing capable of automated analysis of time-series data on individual objects that can contain a cluster of observations over the space of days followed by no additional observations for a period of months (Lomb, 1976; Scargle, 1982).

1.2 Photometry

Photometry is a technique for the precise measurement of the electromagnetic radiation flux incident on an imaging device such as a Charge Coupled Device (CCD) from a light source. Photometry is generally conducted through the analysis of photometric images taken through a filter which prevents photons that are not of specific wavelengths from reaching the detector. Ground-based photometry is usually conducted on filter bands in the optical or near infrared wavelengths of the electromagnetic spectrum. There are a number of common filter systems such as the Johnson-Cousins UBVRI system, the five SDSS filters, u', g', r', i' and z' which have become the standard in the discipline. CCD cameras are ideal for optical and near infrared photometry as they have high quantum efficiencies at these wavelengths. This allows short exposures to detect faint objects as a small number of collected photons can trigger a charge response. These charges are then quickly conveyed across the rows and columns of the chip in a process called readout. Figure 1.1 displays a cropped frame from the IO:O CCD optical camera on the Liverpool Telescope using an SDSS-R filter. The object highlighted by the crosshairs is a pulsating variable star named OGLE-BLAP-009 which is discussed in section 1.4.1.

The comparison of the electromagnetic flux between two filters also provide colour information for target sources and having multiple band photometry allows for multiple colours to be determined. These colours, once corrected for atmospheric and interstellar extinction, allow for the determination of the temperature of target sources. Colour variability is also useful for the classification of variable sources such as Active Galactic Nuclei (AGN) and the separation of RR Lyrae overtone and contact eclipsing binaries.

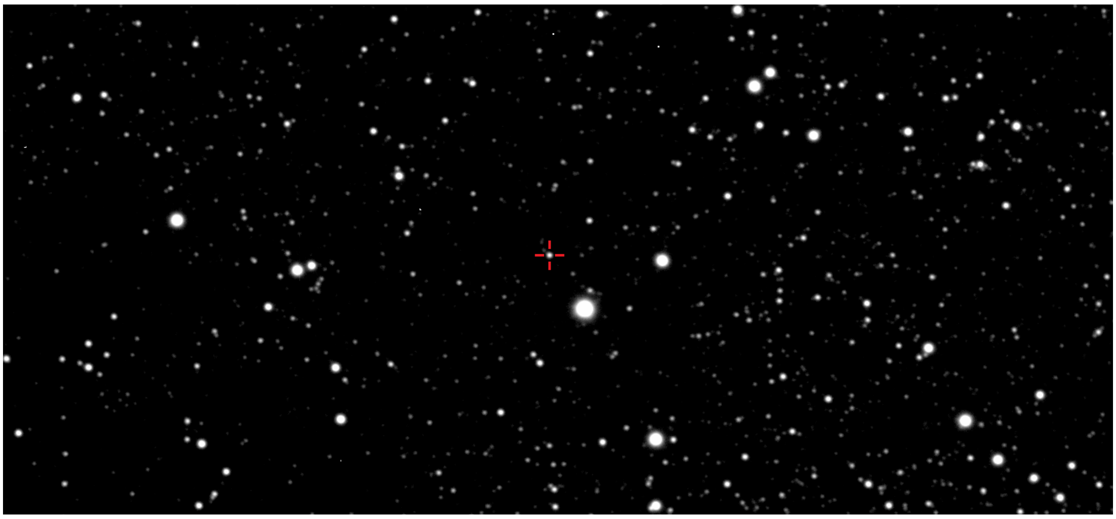


FIGURE 1.1: Cropped IO:O SDSS-R filter image from the Liverpool Telescope of the variable star OGLE-BLAP-009 highlighted in the red crosshairs.

These two variable star classes exhibit similar shaped variability and periodicity which makes these two properties difficult to disentangle them alone but RR Lyrae overtone variables exhibit colour changes throughout their period whereas contact eclipsing binaries do not (Layden et al., 2013).

Photometric data on a large number of objects can be generated through the production of wide-angle images of the sky. The intensity of the image pixels is determined by the activation of the CCD cameras pixels by incoming light from multiple astronomical objects with some background noise and detection bias from the camera (Mawson et al., 2013). As a result, each image contains important information about the brightness (magnitude) of the detected objects. By identifying objects in multiple images with different observation times, information on the change of the brightness of these objects can be determined. This task itself is non-trivial as the objects could be located in different regions of consecutive images due to the motion of the telescope. The resulting brightness-over-time data for each individual object is defined as the objects light curve (Lomb, 1976; Scargle, 1982; Huijse et al., 2012).

Light curves present a quantity of useful data on a light source in the form of a time-series. This time-series is univariate with magnitudes, magnitude error and the associated time instants of measurement. Magnitude is a logarithmic brightness scale used by astronomers as shown in Equation 1.1.

$$m - m_{\text{ref}} = -2.5 \log_{10} \frac{F}{F_{\text{ref}}} \tag{1.1}$$

Where m represents the apparent magnitude of a detected source (i.e. the magnitude of the object as it appears from Earth), m_{ref} represents the apparent magnitude of a suitably chosen reference source, F is the total flux of the detected source and F_{ref} is the total flux of the reference source.

This data can be manually manipulated by experienced astronomers to reveal a wealth of properties associated with the light source object(s). However, the number of light curves being generated by successive extended sky surveys has already passed the point where it is unfeasible for these light curves to be manually analysed. There are a number of problems associated with the extraction of useful information from light curve time-series in which computational intelligence algorithms are of extreme interest. These problems can be categorized as a parameter (feature) extraction process, an experience-based classification operation and an organizational method that attempts to identify structure across the large assortment of light curves (Richards et al., 2011b). These problems appear to be well positioned for exploitation by modern machine learning and computational intelligence methods.

The resultant databases from such extended sky surveys can be dauntingly large potentially containing the light curves of millions of individual light sources. Additionally, the data itself exhibits a number of characteristics that can prove greatly detrimental to the efficient and accurate analysis of the light curves. The dominant property of astronomical light curves is the sampling of these light curves. Whilst surveys will attempt to optimize for a specific sampling rate (a property named cadence), limitations in observational schedules and telescope limitations result in uneven sampling containing artifacts such as gaps in the dataset and non-integer deviations from the desired sampling rate. For example, Earth based observations have an unavoidable periodic one-day gap in observations due to the inability to observe during daytime hours. As well as sampling artifacts, there are also periodic light variations due to local cycles such as the orbit of the moon resulting in different phases, which periodically vary the background sky brightness through the monthly cycle. Additionally, a number of noise sources often affect astronomical data. The Earth's atmosphere can result in noise in the coordinate positioning of light sources as well as refraction and extinction resulting in variations to the measured brightness.

1.3 Small Telescopes Installed at the Liverpool Telescope

The Liverpool Telescope is a fully robotic two-metre class telescope located at the Observatorio del Roque de los Muchachos on the island of La Palma, Canary Islands. It is administered by a collaboration between Liverpool John Moores University and the Instituto de Astrofisica de Canarias (Steele et al., 2004; Mawson et al., 2013). The Small Telescopes at the Liverpool Telescope (STILT) are a set of wide field imaging devices

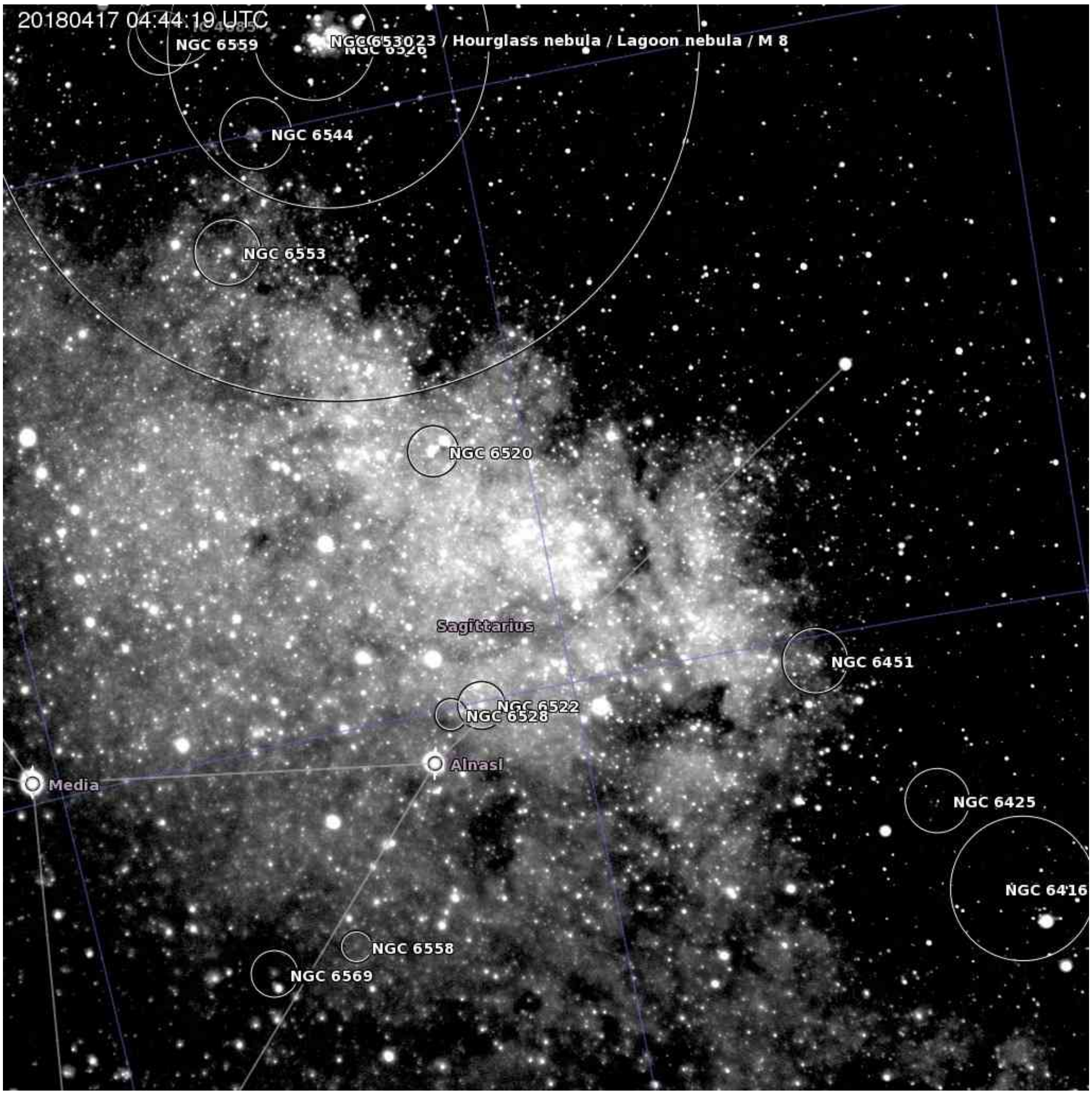


FIGURE 1.2: Annotated SkycamT image captured concurrently with figure 1.1. SkycamT has a 9° × 9° Field of View in this image capturing a large area of the constellation Sagittarius and a section of the Galactic Bulge.

that complement the instrumentation available to the Liverpool Telescope (Mawson et al., 2013). The STILT instruments consist of three instruments with Andor Ikon-M DU934N-BV CCD cameras operating at −40°C to reduce noise due to dark current which are pixel activations due to residual thermal effects (Steele et al., 2004; Copperwheat et al., 2016) detecting unfiltered optical wavelength white light (all electromagnetic radiation across the visible spectrum). They have varying field of views and are mounted directly to the body of the main Liverpool Telescope aimed co-parallel with the main telescopes focus. These instruments have no control over the motion of the Liverpool Telescope and simply take exposures as directed by a control computer. Each instrument has a dedicated small Asus eee pc-powered control unit (Mawson et al., 2013). Figure 1.2 shows the annotated SkycamT frame captured concurrently with the IO:O frame in

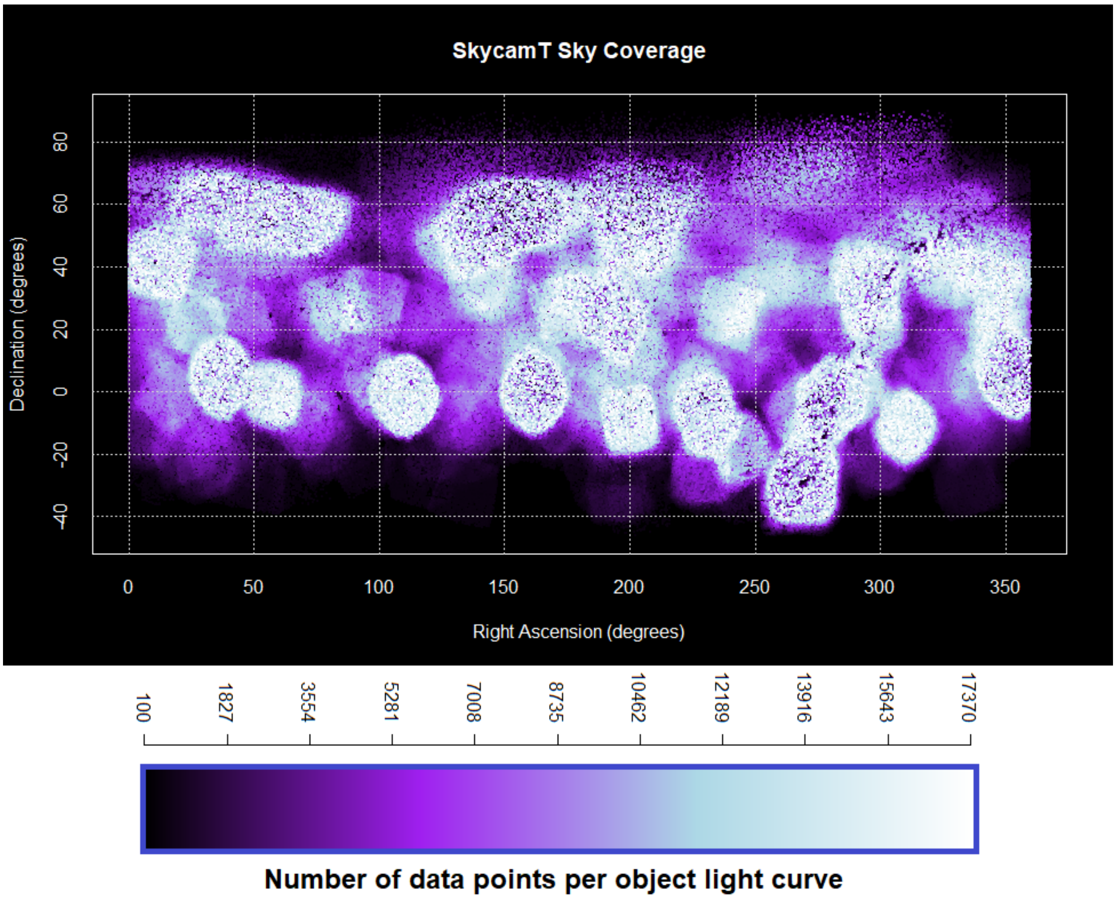


Figure 1.3: Plot of the Right Ascension and Declination of every SkycamT light source with 100 or more data points coloured according to the number of data points in their light curves from 2009–2012, the first three years of Skycam operation.

figure 1.1. This annotated frame is open source and freely available on the Liverpool Telescope website.

The sky coverage and cadence of the Skycams is highly variable and is not optimised for any particular survey or science program (Mawson et al., 2013). Their primary purpose is to provide scientific-grade complementary observations to those from the Liverpool Telescope. The secondary purpose is to conduct a full-sky survey from La Palma to identify variable and transient objects and catalogue them. Figure 1.3 shows a smoothed colour density scatterplot of all the SkycamT light sources in Right Ascension and Declination coordinates as a function of the number of observations.

1.3.1 The Skycam Instruments

The Skycams all use the same CCD device with 1024 × 1024 pixels and a chip size of 13.312 mm. Using equation 1.2, the Field of View (FoV) of the three instruments can be calculated.

$$FoV = 2\arctan\left(\frac{c_{\text{size}}}{2f}\right) \tag{1.2}$$

Where FoV is the Field of View, c_{size} is the chip size and f is the focal length of the instrument. The FoV can then be used to determine the pixel scale of the instrument, the length of each pixel in angular units (usually arcseconds), by employing equation 1.3.

$$P_s = \frac{FoV}{p_n} \tag{1.3}$$

Where P_s is the pixel scale of an instrument, FoV is the Field of View and p_n is the number of pixels across the CCD which for the Skycam instruments is 1024. These equations allow the calculation of the FoV and pixel scale of each Skycam instrument individually based on their focal length.

The first instrument, *SkycamA* is capable of imaging the entire sky from La Palma. It is primarily used for monitoring the status of weather but it can be of use in the detection of bright transient objects. It is the only instrument not attached to the telescope but instead to the inside of the Liverpool Telescope enclosure. It has a 4.5 mm fish-eye lens with a focal ratio of f/4. Using equations 1.2 and 1.3 this provides the instrument with a FoV of 111.9° × 111.9° and a pixel scale of 6.56 '/pixel. With such a large area of sky per pixel, this instrument is of very limited use in variable star photometry therefore this camera does not contribute any observations to the photometric database.

The next instrument is named *SkycamT* and is responsible for most of the photometric observations and is a 'medium field' FoV imager. It is an Andor CCD attached to a Zeiss Planar T 35 mm f/2D (set to f/4) wide-angle lens with a field of view (FoV) of 21° × 21°, a pixel scale of 75.7 "/pixel and a magnitude limit of +12 mag in the R band. In 2014 this instrument was modified to use a Zeiss Planar T 85 mm f/1.4 wide-angle lens which dropped the field of view to 9° × 9° and improved the pixel scale to 31.5 "/pixel. As the data used in this thesis was collected between 2009 and 2012, the original schematics of the instrument are relevant. As the pixel scale is reasonably large, many pixels may contain blended objects where multiple light sources are present in the same pixel. This can produce blended light curves which can appear photometrically variable despite the light sources being non-variable.

The remainder of the photometric database is constructed from observations by the *SkycamZ* instrument. This instrument contains an Andor CCD camera attached to a small Orion Optics AG8 telescope with a 200 mm primary mirror, a 760 mm focal length and a focal ratio of f/3.8. This specification gives the instrument a field of view of 1° × 1°, a pixel scale of 3.53 "/pixel and a deeper magnitude limit of +18 mag in the R band. In 2010 it was modified with an extra-length tube baffle to reduce stray light contamination. All three instruments take a 10 s exposure every minute with a readout time of 35 s whilst the telescope is in operation of their respective field of views.

1.3.2 STILT data reduction pipeline

The Small Telescopes Installed at the Liverpool Telescope (STILT) database is a photometric observation Structured Query Language (SQL) database deployed on the MySQL platform, an open source database management system (DBMS). It contains 1.24 billion separate object observations of 27.74 million independent stellar objects. The database contains time-series data on the magnitude of detected objects over a period of time from March 2009 to March 2012 for SkycamT and July 2009 to March 2012 for SkycamZ (Mawson et al., 2013; McWhirter et al., 2016).

To create this MySQL database, the time-stamped observational images undergo a data reduction pipeline (Mawson et al., 2013). The pipeline first corrects the raw images using dark and flat frames. The dark current and bias noise correction is accomplished by using a single reduction frame. This reduction frame consists of between 30 and 210 dark frames generated by obtaining exposures of the inside of the dome at midnight on nights where the weather prevents observing. These stacked frames must then be updated on a weekly basis. Upon the removal of these known sources of noise, the images are then fit to the World Coordinate System (WCS), a system that allows the location of the frame in the sky to be determined and recorded accurately. This is required as the Skycams are not linked to the Liverpool Telescopes computer systems and therefore they have no knowledge of the current coordinates of the telescopes primary field of view (Mawson et al., 2013). This necessitates the fitting of WCS information through the use of Blind Astrometric Calibration. This is accomplished through the use of two pieces of software, Source Extractor (SExtractor) and Astrometry.net. Source Extractor is capable of identifying the sources of light present in an image and outputting information about them such as their pixel coordinates, the Flux (intensity) of the sources, their ellipticity (how elliptical the light source is on the image) and properties of the size of the source such as the isophotal area (area of the same brightness) (Bertin and Arnouts, 1996). It is also capable of filtering out artifacts to maintain the purity of the sources identified. The second piece of software, Astrometry.net, uses the observing frame to determine its coordinates (Lang et al., 2010). This is accomplished by assigning each extracted light source a unique hash key generated by a quadrilateral produced by the four nearby bright sources. This hash key is generated in a specific manner such that it is invariant to the images orientation and scale allowing it to function successfully for images with various fields of view. The authors of the software claim more than a 99% success rate for contemporary near-ultraviolet (near-UV) and optical survey data with

Table 1.1: Data Summary for the SkycamT and SkycamZ photometric database (Mawson et al., 2013).

Item	SkycamZ	SkycamT
Images	272,470	315,277
Unique Objects	6,290,935	21,453,608
Data Points	332,735,320	904,033,139
Mean light curve points	53	42
Median light curve points	1	7
Max light curve points	13,802	17,370
Database disk space	66.4 GB	213 GB
Start date	29/06/2009	05/03/2009
End date	31/03/2012	31/03/2012
Min Declination	$-36.0577°$	$-51.1956°$
Max Declination	87.8928°	90°

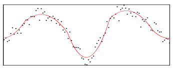
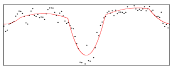
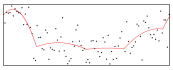
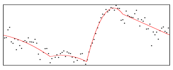
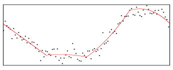
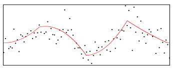
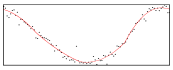
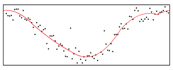
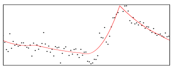
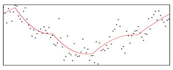
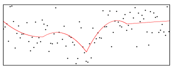
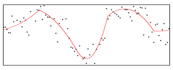
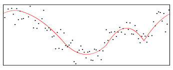
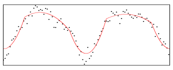
zero false positives for fields with a well matched set of reference quadrilaterals (Lang et al., 2010).

All data from these images that pass all quality control checks is then stored in one of two MySQL databases, one for SkycamT images and one for SkycamZ images (Mawson et al., 2013; McWhirter et al., 2016). This data comprises Source Extractor output as described previously, data from the Flexible Image Transport System (FITS) file header (FITS is a common file type used for astronomy images) and catalogue information from the US Naval Observatory B catalogue based on coordinate matching the sources to known stars with a tolerance of 148" for SkycamT and 9" for SkycamZ. At the end of each observing night, a check is performed to determine the quality of the observations recorded that night. The standard deviation of an objects magnitude values is then determined. A larger value of standard deviation indicates the data is of poor quality as these standard deviations are many times larger than those expected from even the most variable stars. This result is recorded into the database for each observation on an object recorded on a poor quality night. The Liverpool Telescope astronomer-in-charge identifies good quality photometric nights. This was defined as a standard deviation of less than 0.2 mag for SkycamZ data and less than 0.25 mag for SkycamT data. Of the 881 nights that STILT was operating during the three years processed into the Skycam databases, 558 of these nights were classified as photometric quality nights. Table 1.1 displays a summary of the reduced data for SkycamT and SkycamZ.

1.4 Variable Stars

The analysis of variable astronomical objects is a major element in the understanding of stellar and galactic evolution as well as the topology of the universe (Richards et al.,

Table 1.2: Summary of SkycamT variable star classes.

Class	Type	Period	Light Curve
β Lyrae	Eclipsing Binary	<1 day – few days	
β Persei	Eclipsing Binary	0.1 days – years	
Chemically Peculiar	Rotational Variable	Few days	
Classical Cepheid	Pulsating Variable	1 day – months	
δ Scuti	Pulsating Variable	0.02 – 0.3 days	
Ellipsoidal	Non-eclipsing Binary	<5 days	
Mira	Pulsating Variable	80–1000 days	
Type II Cepheid	Pulsating Variable	1–20 days	
RR Lyrae Fundamental	Pulsating Variable	0.3–1 day	
RR Lyrae Overtone	Pulsating Variable	0.1–0.5 days	
RS CVn	Non-eclipsing Binary	Few days	
RV Tauri	Pulsating Variable	20–200 days	
Semiregular Variable	Pulsating Variable	30–1100 days	
W Ursae Majoris	Eclipsing Binary	<1 day	

2011b). Many astronomical objects exhibit brightness variability due to a large number of differing physical processes that uniquely influence an objects light curve. Therefore, the light curve can be used in the classification of variable objects based on the signature of these potentially periodic physical processes and the detection of unknown candidate objects or even unknown variability phenomena that might be due to previously unrecognized astrophysical processes (Protopapas et al., 2006). In this subchapter the variable stars are placed into four distinct subgroups: Pulsating variables powered by stellar radial or non-radial shape oscillations, Variable Binaries due to the orbital dynamics of multiple gravitationally bound stars, Rotational variables with periodic changes during stellar rotation and Eruptive variables with accretion and expulsion of gas and dust from stellar objects. Table 1.2 summarises the main variable stars and time domain events present in the SkycamT database with their typical periods and demonstrations of their light curve shape.

1.4.1 Pulsating Stars

Pulsating stars are variable objects caused by periodic or quasiperiodic expansion and contraction of atmospheric surface layers (Eyer and Mowlavi, 2008; Percy, 2008). These radii oscillations produce changes to the stars temperature and brightness resulting in a measurable change upon the light curves (Lomb, 1976; Scargle, 1982; Huijse et al., 2012). The cooling of the star can also result in a reddening of the light which can also be detected through colour changes between photometric filters. This reddening is a result of temperature decreases in the outer layers of the stars atmosphere as it recedes from the nuclear core. There are also additional non-radial vibrational modes which generally produce lower amplitude variability. These non-radial modes can occur in stars which are already pulsating potentially generating multi-periodic variations.

It is suspected that most, if not all stars exhibit some amplitude of variability due to atmospheric oscillations but few produce signals detectable to current observational technology. As instrumentation improves more variability will likely be detected and the field of astroseismology is developing rapidly from the recent high cadence space-based photometry of missions like Kepler. Therefore, the term pulsating stars is usually reserved for stars of specific properties such as mass, luminosity and temperature which, due to opacity changes in specific chemical species in their atmospheres, drive a feedback mechanism as a specific oscillation mode is continuously excited by the star's own radiation. This phenomena has been named the κ Mechanism, or alternatively the Eddington Valve (Saio, 1993).

The wide-field large sky area surveys of the recent decades have shown the Galaxy to contain many pulsating stars driven by the κ Mechanism. Stellar radiation ionises specic internal layers, often a region of hydrogen and helium although this differs depending on pulsating star class and other metals such as iron have been identified as driving pulsations. The free electrons from the ionised layer dominate the opacity due to electron scattering and free-free absorption. This increased opacity results in an increased radiation pressure on the ionised layer repulsing it from the nuclear core or shell layer. The ionised layer and all those above it expand the radius of the star resulting in an increased luminosity. As the star expands the layers begin to cool until the ions in the driving layer recombine into atoms. With the electrons now bound into atoms, the opacity decreases. With the radiation pressure support lost the star collapses under gravitational free fall contracting to the initial radius. As the driving layer is heated by its proximity to the nuclear furnace it will once again ionise renewing the pulsation cycle. The amplitude variation presented by these pulsations is highly dependent on the depth of the driving layer. Deep layers have a small effect on the above layers and layers too close to the surface have a small amount of material to repulse and therefore cannot generate large

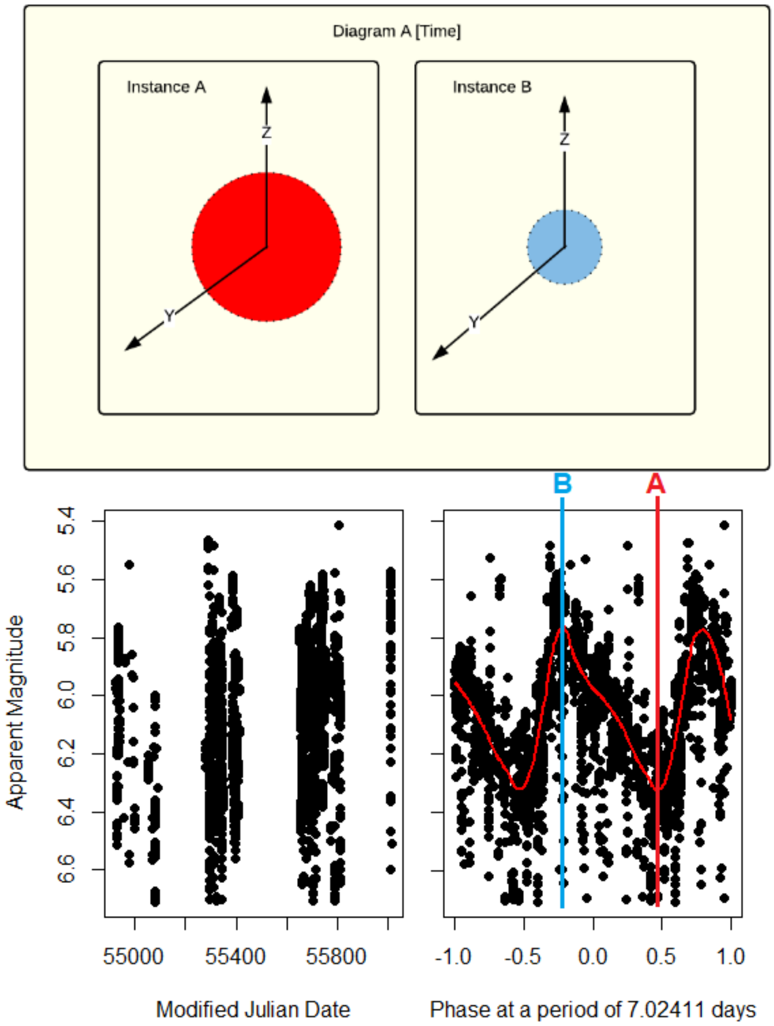


FIGURE 1.4: The light curve of the star U Aquilae, a pulsating Classical Cepheid star. The y axis shows the line of sight from the Earth.

variations (Percy, 2008). Figure 1.4 shows how the light curve of a pulsating classical cepheid star with a period of 7.02 days is influenced by the expansion and contraction of the star.

The study of the properties of these pulsations such as their period and amplitude allow for the determination of the structural parameters of these stars and of the specifics of the driving layer. This has provided data for the study of the evolutionary state of pulsating stars with notable homogeneous classes occurring as a product of stellar evolution across instability regions of the Hertzsprung–Russell (HR) diagram (Gautschy and Saio, 1996). Pulsating variables often have a period-luminosity relation that can be used as a distance estimator. With certain types of pulsating star having a high absolute magnitude, these stars have been used to calculate distances to nearby globular clusters and galaxies (Bedding and Zijlstra, 1998; Glass and Evans, 1981; Cohen and Sarajedini, 2012; Madore et al., 2009). They are an important component of the cosmological distance ladder

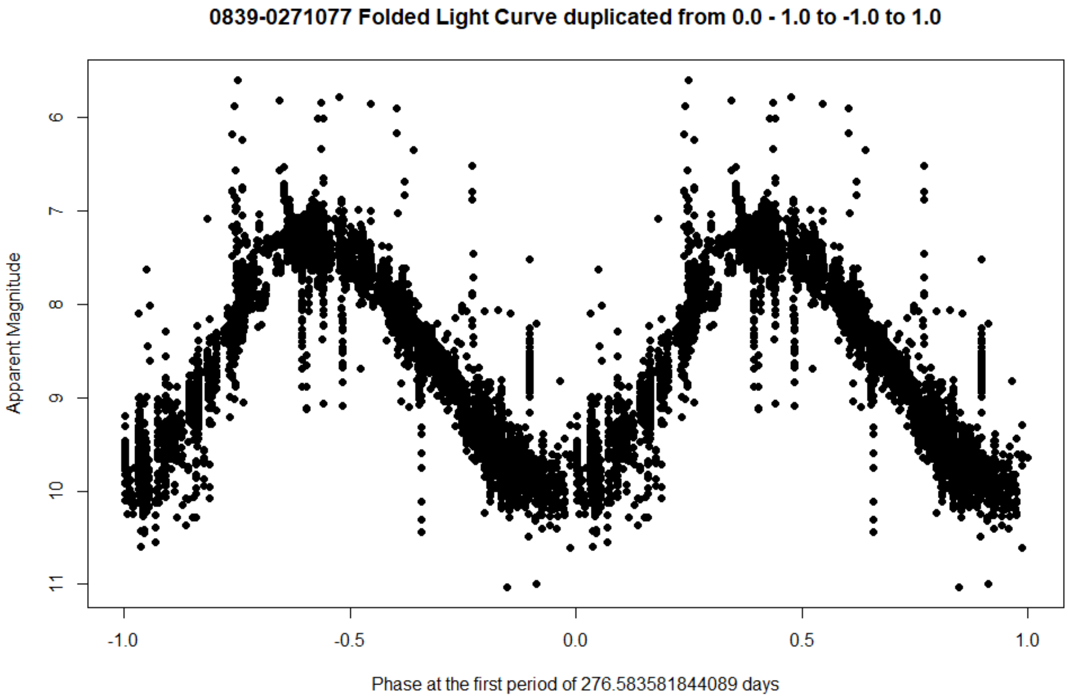


FIGURE 1.5: Phased SkycamT light curve of the Mira variable Y Librae with a period of 276.58 days.

bridging distance determinations of nearby stars with parallax measurements with those of Supernova Ia explosions used for far away galaxies (Kim and Miquel, 2007).

The analysis of pulsating variables has allowed them to be classified into a number of homogenous classes with the primary distinguishing features of each class being their *period*, *amplitude* and *colour*. Using the period, the pulsating stars can be subdivided into mid to Long Period Variables and Short Period Variables. This distinction also correlates with the evolutionary status of the pulsating stars as more compact objects near or below the Zero Age Main Sequence (ZAMS) have a higher sound speed with shorter period pulsations and evolved giant stars have a lower material density, a slower sound speed and longer periods. The well-known pulsating star classes are now discussed based on this subdivision.

Mid to Long Period Variables have periods of greater than one day up to several thousand days and are usually evolved giant and supergiant type stars. They are usually of spectral class G, K or M which gives them yellow to red colours. They often exhibit large amplitude variations and therefore have been studied for centuries.

- **Mira variables**

  Mira Variables have periods of 80-1000 days with the majority between 150-450 days although the periods can slowly vary due to evolutionary progression as well as an up to 5% cycle to cycle variation (He et al., 2016; Wood and Sebo, 1996). They are fundamental tone pulsating red giant stars of spectral class K or M with large amplitudes of greater than 2.5 mags although they can range up to 10 mags (Percy, 2008). They can be both Population I and II stars and are highly common in variable catalogues due to their ease of detection yet their evolutionary state suggests they are a rare variable. This does mean care has to be taken in regards to not introducing bias into samples. They are evolved giant stars with masses of 0.6 $M_{\odot}$ up to a few solar masses which have ended hydrogen-burning in their cores, have gone through the helium flash and are now at the end of the Asymptotic Giant Branch. They have radii of hundreds of solar radii although the computed radii are dependent on the wavelength of the filter band. This also results in their pulsation amplitude being highly dependent on filter with near infrared amplitudes being substantially smaller than V band optical amplitudes. Mira variables follow period-luminosity relations (Bedding and Zijlstra, 1998; Glass and Evans, 1981). This amplitude can also vary from cycle to cycle as well as the minimum and maximum apparent magnitude (He et al., 2016). The cooler Mira variables also have longer periods. The substantial variability is powered by pulsations and temperature changes (Wood and Sebo, 1996). As the red giant expands it cools causing less of the star's energy to be emitted in the visible spectrum. Additionally, this cooling allows Titanium Oxide to form in the outer parts of the atmosphere which absorbs additional visible light increasing the substantial reduction in luminosity (Percy, 2008). There is a period-luminosity-colour relationship for Mira variables with the most effective measures being with infrared magnitudes. They are a short-lived evolutionary stage and will soon eject their atmospheres which will light up from ultraviolet radiation from the exposed hot stellar core producing a planetary nebula. Figure 1.5 demonstrates the sinusoidal shaped phased SkycamT light curve of the Mira variable Y Librae although Mira light curves can be more asymmetric depending on the phase difference between the radii and temperature variations.

- **Semi-regular variables**

  Mira variables are an example of a rare but clearly periodic red giant variable. Photometric surveys have indicated that most if not all red giants exhibit some level of variability. Semi-regular variables are stars of spectral classes F, G, K and M which have amplitudes of below 2.5 mag and periodicity which range from persistent to undefined (Percy, 2008; Soszynski et al., 2009). Semi-regular stars are subdivided into 4 subtypes, SRa, SRb, SRc and SRd. SRa and SRb are of spectral class K or M and are all variable red giants with the subdivisions indicating the strength of the periodicity with SRa variables having somewhat periodicity

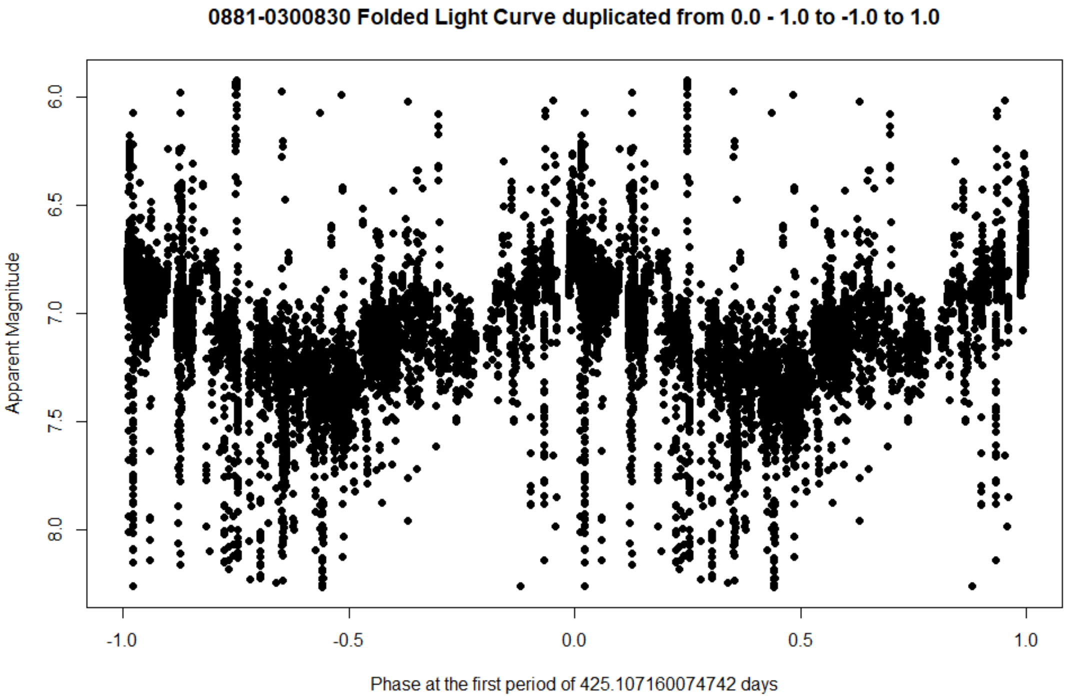


FIGURE 1.6: Phased SkycamT light curve of the Semi-regular variable variable Y Serpentis with a period of 425.11 days.

and SRb variables having very poor to no periodicity (Percy, 2008). They are lower mass, evolved older Population I or II stars pulsating in fundamental and overtone states and can have multiple pulsation periods from 50 days to several thousand days. SRc variables are spectral class M red supergiants which are massive, young, population I stars. They have semi-regular variability with periods of 250-1000 days (Percy, 2008). SRd variables are much less common and are supergiants of spectral class F, G and K having yellow to orange colours, amplitudes of up to 4 mag and semi-regular periods of 30 to 1100 days (Rosino, 1951; Giridhar et al., 1998). Semi-regular variables can also exhibit a long secondary period of thousands of days. The cause of this variability is not known but it might be due to an interaction between two pulsation modes, a pulsation mode and the stellar rotation or an unseen binary companion (Olivier and Wood, 2003). Figure 1.6 shows the phased SkycamT light curve of the periodic Semi-regular variable Y Serpentis which shows a small 'bump' in the ascending branch possibly due to a secondary pulsation.

- **Slow Irregular variables**

  Slow Irregular variables are red giants and supergiants with spectral class M. They are closely related to the Semi-regular variables and are likely variants of this class which have very poorly defined periodicity and vary over longer characteristic

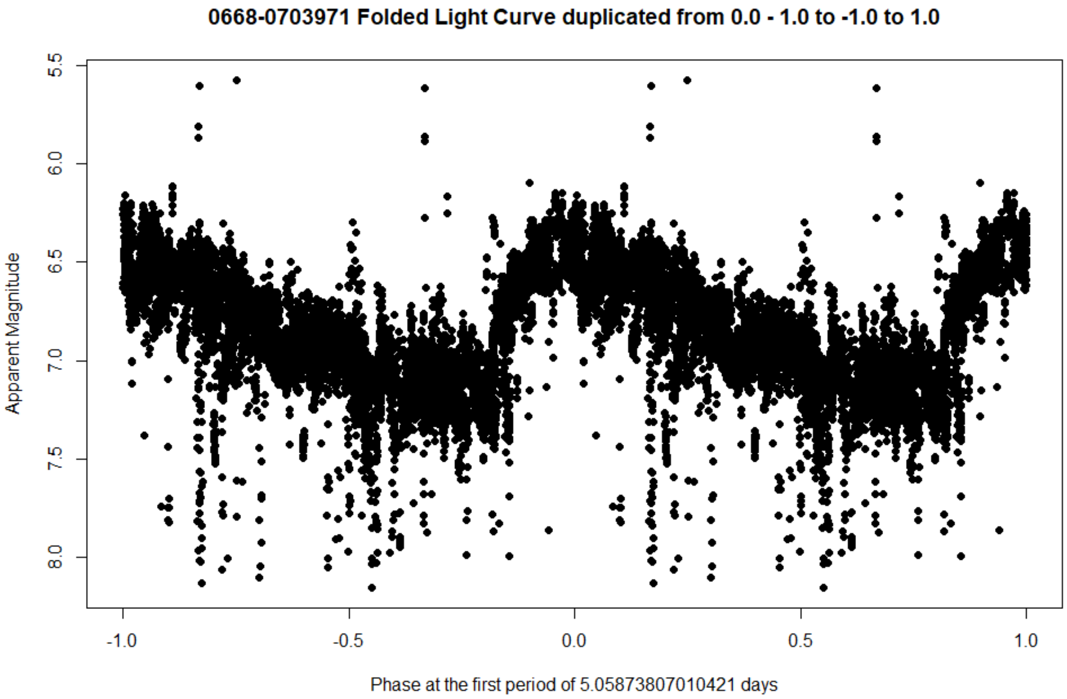


FIGURE 1.7: Phased SkycamT light curve of the Classical Cepheid variable AP Sagitarii with a period of 5.06 days.

timescales (Percy, 2008). They are also subdivided into two types, Lb and Lc. Lb variables are red giants related closely to the SRb variables and Lc variables are red supergiants closely related to the SRc variables (Percy, 2008). The light curves of these objects will identify them as variable objects but a period analysis will likely result in a spurious sampling period although likely with a strong confidence generated by the underlying variability (Richards et al., 2012).

- **Small Amplitude Red Giants**

  Small Amplitude Red Giants (SARGs) are very common and show variability on short and long timescales (Percy, 2008). They are of spectral class K and M and have semi-regular periods of 10-100 days and amplitudes of 0.005-0.13 mag in the I band (Wray et al., 2004). They are population I and II red giants which pulsate from fundamental to high overtone modes and can be multi-periodic (Xiong and Deng, 2006).

- **Classical Cepheids**

  Classical Cepheids are Population I yellow giant/supergiant stars of spectral class G. They are evolved stars which are moving off the main sequence and have begun hydrogen shell burning and helium core burning (Stetson, 1996; Percy, 2008). They are mostly fundamental mode pulsators powered by the κ mechanism on a helium

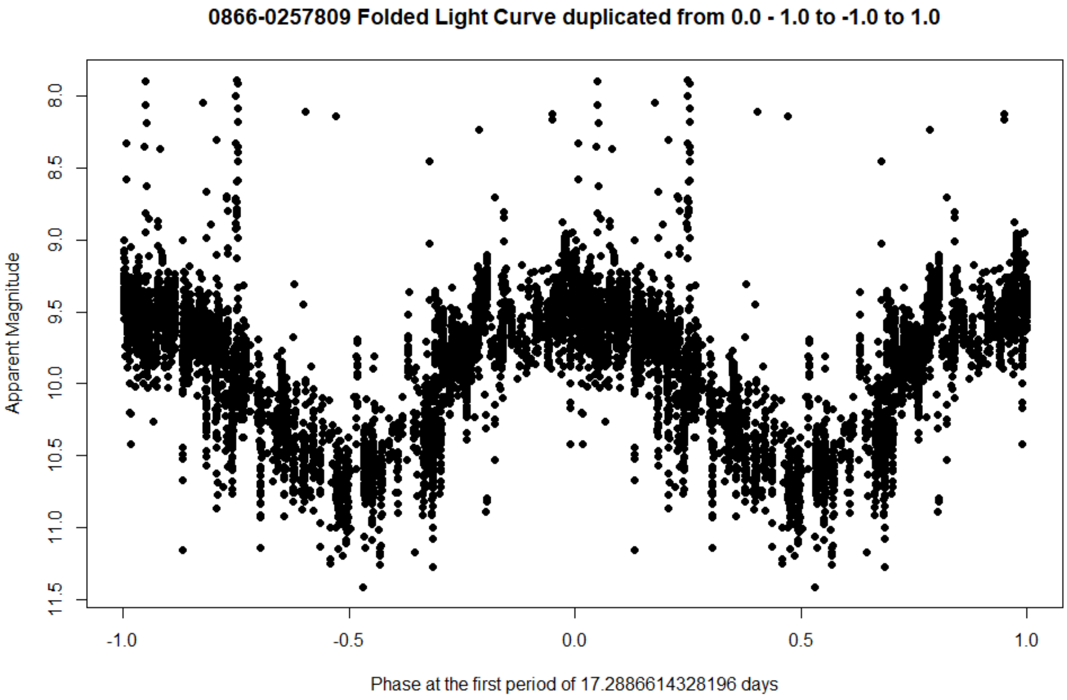


FIGURE 1.8: Phased SkycamT light curve of the Type II Cepheid W Virginis with a period of 17.29 days.

layer although they have also been observed pulsating in overtone modes separately, the Overtone or s–Cepheids or simultaneously, the Multimode Cepheids. Their physical properties places them in a region of the H–R diagram named the Cepheid Instability Strip, a region they can migrate across multiple times as they evolve (Gautschy and Saio, 1996). They have periods of 1-70 days and obey a period-luminosity relation named Leavitt's law (Percy, 2008). This has been used as an independent distance measurement due to the relatively bright absolute magnitude of Classical Cepheids (Madore et al., 2009). They have highly asymmetric light curves with a sawtooth shape. They exhibit a rapid rise to maximum brightness followed by a gradual fall back to minimum with amplitudes of 0.5-2 mag. Overtone cepheids have a more sinusoidal light curve with amplitudes of under 0.5 mag (Stetson, 1996; Yoachim et al., 2009). Classical Cepheid light curves with periods of 6-20 days also show additional features such as 'bumps' at certain phases. These are likely caused by a 2:1 resonance between the fundamental mode and the second overtone (Percy, 2008; Yoachim et al., 2009). Their amplitude and phase are highly correlated with the pulsation period of the star located on the descending branch before 10 days and on the ascending branch after 10 days before disappearing completely at longer periods. Figure 1.7 shows the sawtooth, asymmetric phased SkycamT light curve of the Classical Cepheid AP Sagitarii.

- **Type II Cepheids**

  Type II Cepheids are low mass, metal poor Population II giant/supergiant stars. They are also located within the Cepheid Instability Strip. They are of higher effective temperature than the classical cepheids and have spectral classes of mid to late F and G (Percy, 2008). They have periods of 1-20 days and bridge the short period, metal poor, slightly higher effective temperature RR Lyrae variables to the longer period RV Tauri variables (Wallerstein, 2002). They also exhibit a Period-Luminosity relationship which differs from the Classical Cepheid relation. Type II Cepheids are often subclasses into three types by period. BL Herculis variables have periods of 1-8 days and amplitudes of a few tenths of a magnitude and are crossing the instability strip post horizontal branch (Percy, 2008). They are smaller and fainter than classical cepheids with similar periods. W Virginis variables have periods of 10-20 days with amplitudes of around 1 mag and are burning hydrogen or helium in nuclear shells (Soszynski et al., 2008). They are evolving through the instability strip from the Asymptotic Giant Branch due to an increase in effective temperature from hydrogen envelope contraction caused by a helium shell thermal pulse, named a blue-loop excursion (Percy, 2008). The final subtype is named the RV Tauri variables which are discussed below. Figure 1.8 shows the phased SkycamT light curve of the star W Virginis.

- **RV Tauri variables**

  RV Tauri variables are bright, low mass, metal poor Population II giant/supergiant stars of spectral classes of late F and G. They are the longest period Type II Cepheids (Percy, 2008). They have periods greater than 20 days with amplitudes of around 4 mags and are the last excursion of Asymptotic Giant Branch stars through the instability strip before nuclear fusion ends and they transition to white dwarf stars. Their light curves exhibit a unique shape assisting in their identification. They have alternating shallow and deep minima and often have quoted periods of twice that of the variation period identifying a deep and shallow minima pair (Rosino, 1951). Over longer timescales these minima also modulate resulting in the order of the deep and shallow minima disappearing and then reversing. Over these timescales variations in period and the brightness of both the minima and maxima as well as additional chaotic behaviour can be seen (Percy, 2008). Figure 1.9 shows the phased SkycamT light curve of the RV Tauri variable AC Herculis with the clear alternating minima.

Short Period Variables have periods of less than one a few days down to a few minutes. They are dwarf, subdwarf or compact objects with the exception of the giant RR Lyrae stars. They are substantially hotter than the long period variables with spectral classes of

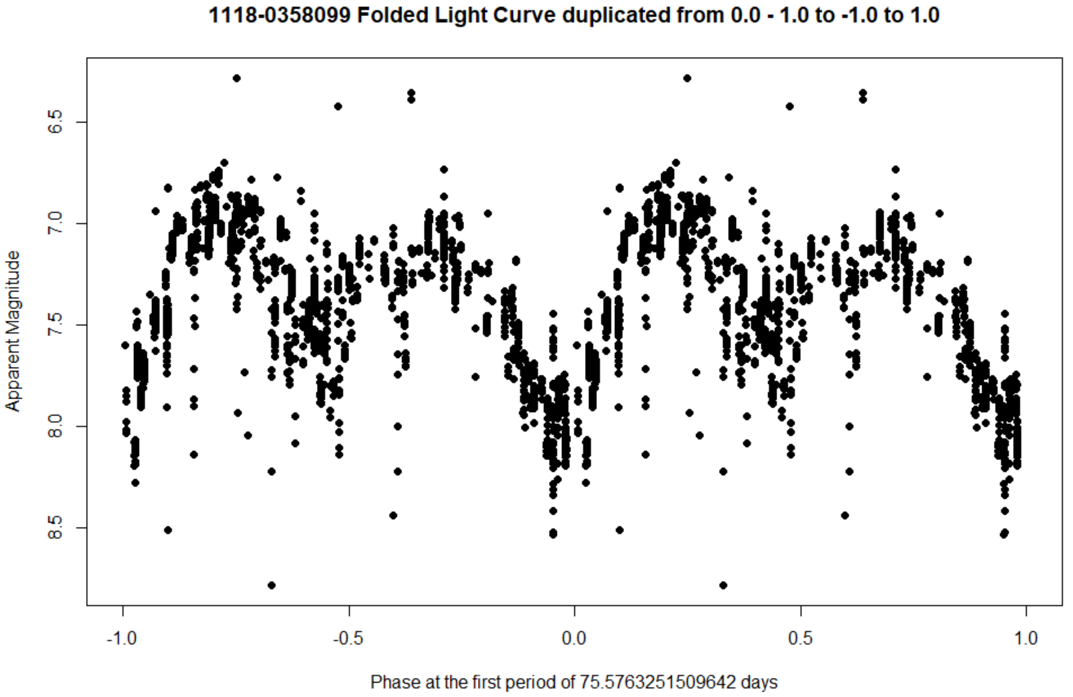


FIGURE 1.9: Phased SkycamT light curve of the RV Tauri variable AC Herculis with a period of 75.58 days.

O, B, A and early F. This gives them colours ranging from blue through white to yellow-white. As they have shorter pulsation periods they generally have smaller amplitudes than the Long Period Variables. This has resulted in the identification of many new classes of Short Period Variable within the last few decades due to improvements in the precision of observational technology.

- **RR Lyrae variables**

  RR Lyrae variables are old, population II giant stars with spectral classes late A to early F and are common in globular clusters. They have stable periods of 0.1 to 1 day and amplitudes of up to 1.5 mag in the V band (Percy, 2008; Smith, 2004). They are evolved horizontal branch stars with masses of about 0.5 $M_\odot$ and low metallicity which have begun burning helium in their cores. Many RR Lyrae variables have also been found to have a long period amplitude and/or phase modulation with a timescale of 11 days to 533 days named the Blazhko effect (Blazhko, 1907). They are subtyped into three groups based on their light curves, RRa, RRb and RRc (Percy, 2008). RRa variables have long periods, high amplitudes and asymmetrical light curves similar to the classical cepheids. RRb variables appear to be a smooth transition of the RRa type and have slightly longer periods, slightly smaller amplitudes and less asymmetry in their light curves compared to the RRa variables. Often this distinction is ignored and they are grouped as

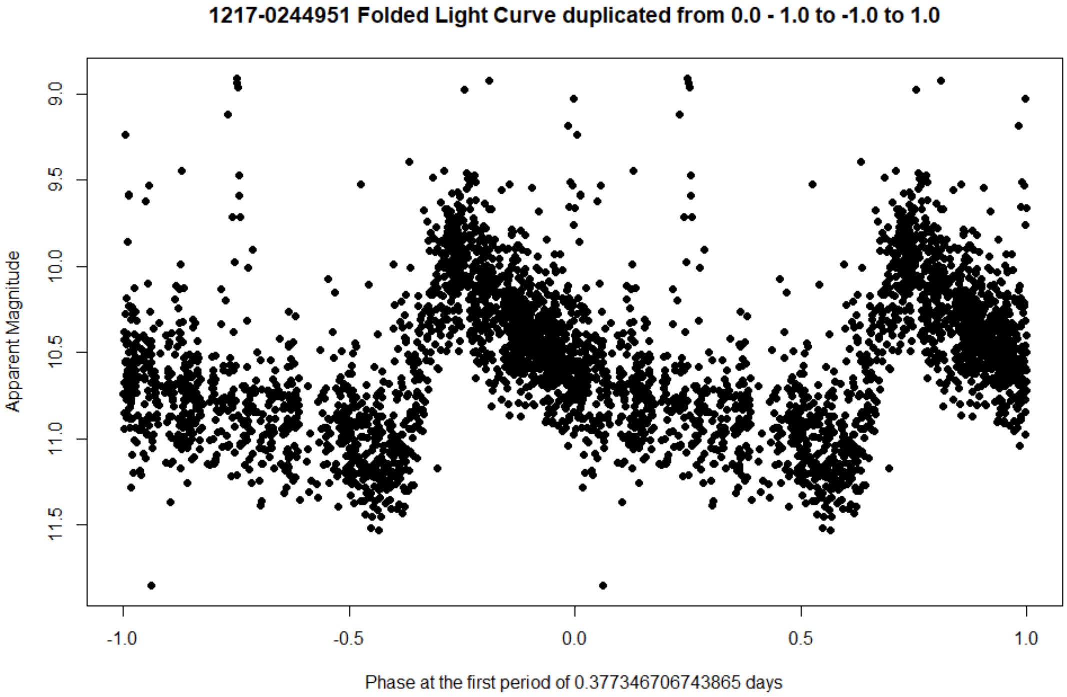


Figure 1.10: Phased SkycamT light curve of the RR Lyrae type ab variable RS Bootis with a period of 0.377 days.

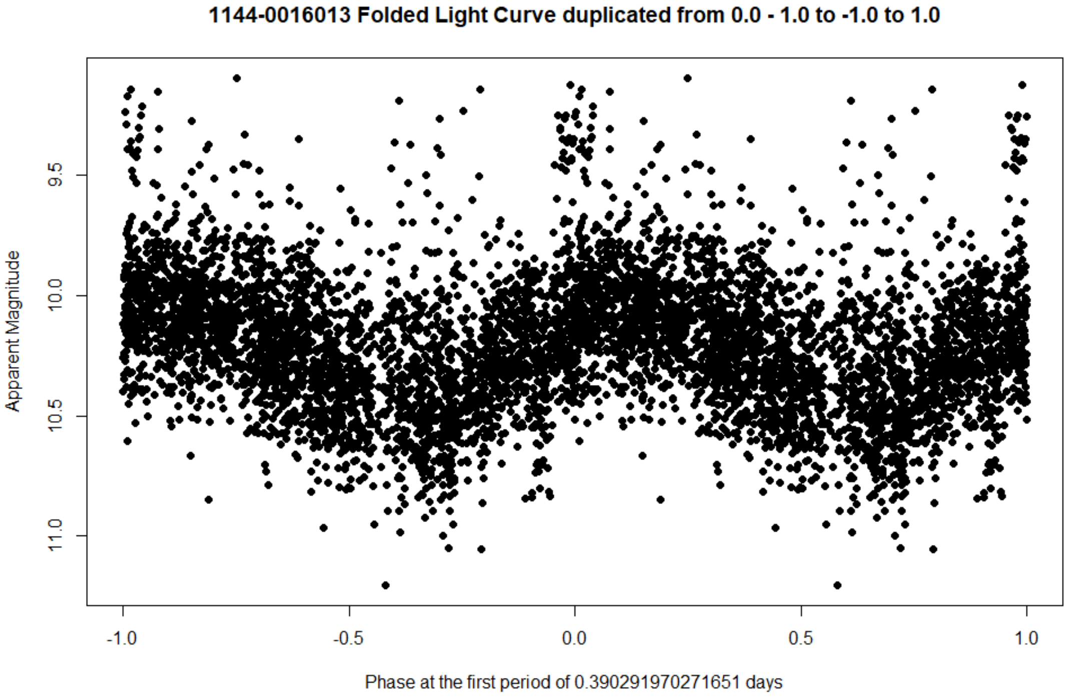


Figure 1.11: Phased SkycamT light curve of the RR Lyrae type c variable RU Piscium with a period of 0.390 days.

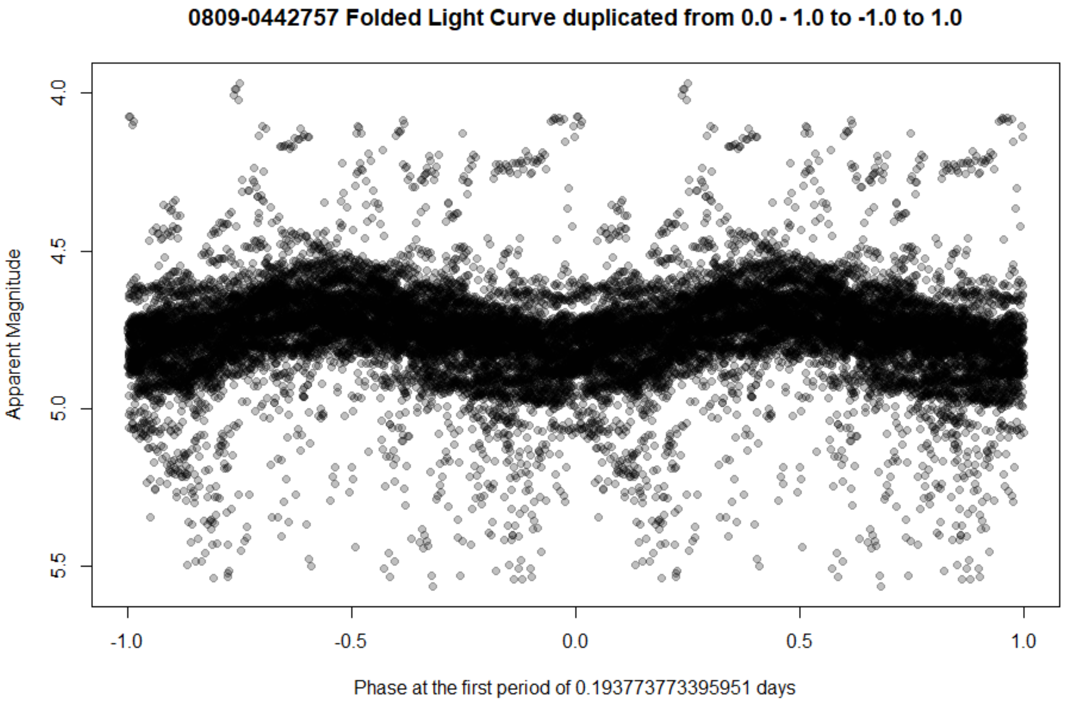


Figure 1.12: Phased SkycamT light curve of the δ Scuti variable δ Scuti with a period of 4.65 hours.

RRab variables. The RRab variables are pulsating in a fundamental mode and this is why their light curves are similar to the fundamental mode classical cepheids. RRab light curves also show a ‘bump’ near minimum light caused by shockwaves reflected upwards from layers deep in the star (Percy, 2008). The RRc variables have shorter periods, much smaller amplitudes and sinusoidal light curves as these stars are pulsating in the first radial overtone. There are also rare, double-mode RR Lyrae variables which pulsate in both the fundamental and first overtone periods and are often assigned a fourth RRd subgroup. RR Lyrae variables obey a period-luminosity-metallicity relationship and have been used to measure distances to nearby globular clusters (Smith, 2004). Their lower absolute magnitude relative to classical cepheid variables and more complicated relationship has limited their applicability in cosmological distance determinations. Figure 1.10 shows the phased SkycamT light curve of the RRab variable RS Bootis and figure 1.11 shows the phased SkycamT light curve of the RRc variable RU Piscium.

- **δ Scuti variables**

δ Scuti variables are zero age main sequence stars of spectral classification A and F. They are located in a position on the H–R diagram where the cepheid instability strip crosses the dense main sequence and are mostly core hydrogen burning stars (Percy, 2008). As a result, the δ Scuti variables are some of the most common pulsating variable stars with pulsations driven by the κ mechanism similar to cepheids. They have been subtyped by amplitude and metallicity into High Amplitude δ Scuti (HADS) variables, low-amplitude δ Scuti variables and the low metallicity, high amplitude SX Phoenicis variables (Percy, 2008; Pigulski et al., 2005; Cohen and Sarajedini, 2012). They have periodicities of 0.02 to 0.3 days usually positively correlated with amplitude. They have amplitudes of less than 0.2 mag with smaller amplitudes being more common although some of the HADS have amplitudes over 0.3 mag. They obey a period-luminosity-colour relationship although their low amplitude variability limits the use of this as a distance measurement (Percy, 2008). Many δ Scuti variables are multi-periodic and can exhibit over ten independent periods (Poleski et al., 2010). The HADS variables are evolved stars and have begun hydrogen shell burning. SX Phoenicis variables are Population II stars in an advanced evolutionary state and are known as blue stragglers (Cohen and Sarajedini, 2012). These stars have not evolved away from the main sequence as would be expected for stars of their mass and temperature and therefore are suspected to have gained mass late in their evolution through binary interaction or possibly a stellar merger. Figure 1.12 shows the phased SkycamT light curve of the prototype variable δ Scuti. A large number of observations are required to obtain this period from SkycamT data for low amplitude variables.

- **γ Doradus variables**

  γ Doradus variables are zero age main sequence stars with spectral types of early F. They are a recently identified class of pulsating variable which usually have multiple periods in the range of 0.4 to 3 days and amplitudes of 0.1 mag in the V band (Percy, 2008). They are high-order, non-radial, gravity mode pulsators pulsating in multiple modes simultaneously (Krisciunas, 1993; Percy, 2008). It is also possible for hybrid δ Scuti and γ Doradus variables to exist which pulsate in the radial δ Scuti modes and the non-radial γ Doradus modes (Guo et al., 2016).

- **β Cephei variables**

  β Cephei stars, not to be confused with cepheid variables, are hot stars with spectral classes of early B. They are often multi-periodic with periods of 0.1 to 0.3 days and with very small optical band amplitudes of 0.01 to 0.3 mag although substantially larger in the ultraviolet (Percy, 2008). The multi-periodicity is due to fundamental and overtone mode radial pulsations with close periods due to the stellar rotation. The radial pulsations are powered by the κ mechanism except instead of a helium driving layer, the pulsations are due to the ionisation of iron deep in the star at temperatures of $200,000\,K$ (Miglio et al., 2007). Whilst the light amplitudes are small, the radial pulsation is large and the stellar atmosphere expands and contracts at colossal speeds up to 200 kms^{-1} (Percy, 2008).

- **Slowly Pulsating B variables**

  Slowly Pulsating B variables, or 53 Persei variables, are hot stars of spectral class B. They are similar to the β Cephei variables but pulsate purely in non-radial modes (Percy, 2008; Miglio et al., 2007). These variables are often multi-periodic with periods ranging between 0.5 to 5 days and amplitudes of under 0.1 mag in optical bands but larger in the ultraviolet (Waelkens and Rufener, 1985).

- **Rapidly oscillating Ap (roAp) stars**

  Ap stars are chemically peculiar main sequence stars with spectral classes of late B to early F. This chemical peculiarity is due to the stars powerful magnetic field which allows elements usually present deep in the stellar interiors to migrate to the surface (Percy, 2008). A subset of these stars lying in the δ Scuti instability strip are also pulsating with periods of 5 to 23 minutes (Kurtz, 1982). These pulsations are described by the *oblique pulsator model* and are caused by high overtone pulsations from a hydrogen ionisation zone which are strongly influenced by the stellar magnetic field. These pulsations are along the magnetic axis which is offset from the rotation axis. This results in a modulation of the pulsation by the stellar rotation as the observer is presented with a different aspect of the pulsation throughout the rotational period (Kurtz, 1982; Percy, 2008).

- **Pulsating Hot Subdwarfs**

  Pulsating hot subdwarfs (sdB) are spectral class B stars with size and luminosity below main sequence dwarf stars (Jeffery et al., 2001; Percy, 2008). These subdwarfs are evolved stars on the extreme horizontal branch at the hot end (Geier, 2015). They are evolving past the instability strip as they lack a massive outer atmosphere to progress onto the Asymptotic Giant Branch. This stripping of the outer atmosphere is hypothesised to be a result of binary interaction as many hot subdwarfs are found in binary systems. They are helium core burning stars with a thin hydrogen-rich shell and the pulsations are generated by a partially ionised iron layer deep in their atmosphere (Geier, 2015). There are two major subgroups of pulsating sdB star, the EC 14026 stars and the PG1716 stars. EC 14026 stars have non-radial pulsations with periods of several hundred seconds with amplitudes of up to 0.1 mag (Kilkenny et al., 1998). The PG1716 stars are also non-radial pulsators with a similar amplitude to the EC 14026 stars but with periods of 1-2 hours (Reed et al., 2004).

- **Blue Large Amplitude Pulsators**

  Blue Large Amplitude Pulsators (BLAPs) are a recently discovered type of pulsating variable detected from years of photometry from the OGLE survey (Pietrukowicz et al., 2017). They were originally classified as δ Scuti variables however whilst their amplitude was equivalent to HADS variables at 0.25-0.3 mag in the I band, they had periods of 20-40 minutes which are shorter than the shortest δ Scuti period (Pietrukowicz et al., 2017). As δ Scuti variables usually have a positive correlation between period and amplitude, these objects appeared to be a poor fit to the class (Poleski et al., 2010). Spectroscopic follow-up identified these stars as having effective temperatures of around $30,000\,K$ making them spectral class O and B. The light curves of these variables have very similar shapes to the fundamental mode RR Lyrae stars. Spectral analysis indicates that they are driven by helium ionisation similar to the other fundamental mode pulsators and are moderately helium enriched (Pietrukowicz et al., 2017). The higher luminosity and lower surface gravity of BLAPs compared to hot subdwarf stars suggest they might have inflated envelopes. Three possible configurations have been proposed: A helium core burning star with a core mass of about $1.0\,M_{\odot}$ which requires substantial atmosphere stripping possibly by a binary companion with a fundamental mode pulsation. A red giant with a stripped atmosphere sustained by hydrogen shell burning above a $0.3\,M_{\odot}$ degenerate helium core with a fundamental mode pulsation. The final possibility is that these are hot pre-Extremely Low Mass white dwarf stars which have low-order radial pulsations including the fundamental mode at the shortest periods (Romero et al., 2018). Thus far all known BLAPs have been discovered in or near the Galactic Bulge with no BLAPs detected by OGLE surveys in the Magellanic Clouds (Pietrukowicz, 2018). Figure 1.13 shows the phased OGLE light curve of the star named OGLE-BLAP-009 with a period of 31.94 minutes.

- **ZZ Leporis variables**

  ZZ Leporis variables are pulsating stars in the centre of planetary nebulae evolving towards white dwarfs (Handler et al., 2013). They are of spectral class O and B and have semi-regular radial pulsations with periods of 3-10 hours with amplitudes of 0.4-0.6 mag. Their light curves are somewhat sinusoidal with irregular long term variability possibly due to their stellar winds. They have hydrogen-rich atmospheres with effective temperatures of $25,000 - 50,000\,K$.

- **Pulsating White Dwarfs**

  White dwarf stars are the hot, degenerate cores of dead stars which are no longer burning material via nuclear fusion and are now slowly cooling (Percy, 2008). They

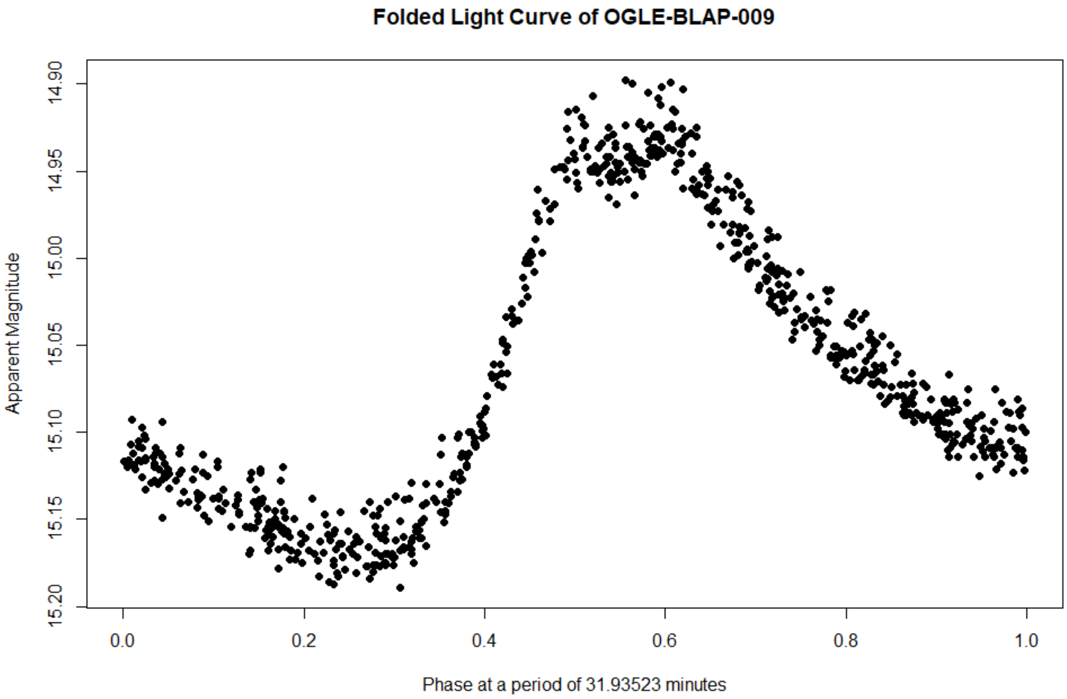


Figure 1.13: Phased OGLE light curve of OGLE-BLAP-009 with a period of 31.94 mins produced by the author from openly available OGLE data.

are the final evolutionary state of low mass stars and have an average mass of 0.6 $M_{\odot}$ and radii similar to that of the Earth. There are a number of pulsational instability zones with properties that match white dwarfs and as they cool they can pass through these instability strips (Althaus et al., 2010). These instabilities power non-radial gravity wave pulsations with a superposition of vibrational modes. There are four classes of white dwarf each with their own pulsation modes (Percy, 2008). The DAV white dwarfs are of spectral class A and have hydrogen atmospheres with periods of 1.6-23.3 minutes and amplitudes of 0.01-0.3 mag due to ionised hydrogen. They are also known as ZZ Ceti variables (Koester and Chanmugam, 1990). The DBV white dwarfs are of spectral class B and have primarily helium atmospheres with periods of 2-18 minutes and amplitudes of 0.05-0.3 mag due to helium ionisation. These variables are also known as V777 Her variables (Winget et al., 1982). DOV white dwarfs are of spectral class O and are the hottest white dwarfs. They have carbon/oxygen atmospheres with ionisation of these metals driving the pulsations. Standard DOV variables have a period of 5-43 minutes and amplitudes of 0.02-0.1 mag and are also known as PG1159 or GW Virginis variables (Cox, 2003). Similar to the DOV white dwarfs there is also the Variable Planetary Nebula Nuclei (PNNV) variables. The PNNV stars are surrounded by planetary nebulae whereas DOV white dwarfs are not. They have periods of 7-100 minutes and amplitudes of 0.01-0.15 mag (Percy, 2008). The final class of pulsating white dwarf is the recently discovered DQV class. The DQV white dwarfs have temperatures between the DAV and DBV types and are of spectral class O to B yet are rich in carbon. These stars have periods of 4-18 minutes and amplitudes of 0.005-0.015 mag (Dufour et al., 2007).

1.4.2 Variable Binaries

Variable binaries are extrinsic variables due to two or more gravitationally bound stars executing orbits around a common gravitational centre-point (LaCourse et al., 2015). The relative proximity of the stars often means that they cannot be distinguished on an image and appear as a single source of light. These variables are highly periodic in their photometric variability equal to the orbital period of the binary system (Percy, 2008). The colour of these variables is of less importance for their classification than the pulsating variables as there is a wider configuration of possible spectral classes in these systems. Variable binaries can be subset into eclipsing and non-eclipsing binaries. Eclipsing binaries are the easier to detect with variations caused by the plane of the binary orbit aligning with the view from Earth. As a result, one star periodically passes in front of another resulting in a change in either the brightness and/or the relative brightness in different colour filters of the source of light in the astronomical images (LaCourse et al., 2015). Analysis of the light curves of eclipsing binary candidate light sources can be used to determine the number and types of star present within the system as well as their mass and orbital period (Percy, 2008). Similar smaller amplitude signals can also be caused by exoplanetary transits where one of the objects is a planet orbiting a parent star.

Figure 1.14 demonstrates the dynamics behind eclipsing binary systems and the associated light curve. The light curve in this example is of the variable source HT Virginis, a W Ursae Majoris type Eclipsing Binary system consisting of two stars closely orbiting each other with an orbital period of 0.41 days. As the more luminous star is eclipsed by the dimmer star, there is a large reduction in source brightness called the primary eclipse. The primary eclipse ends when the more luminous star emerges from the eclipse. After half an orbit, the dimmer star is eclipsed by the more luminous star resulting in a smaller reduction in source brightness, the secondary eclipse. As the stars orbit they periodically execute primary and secondary eclipses with a period equal to the orbital period of the binary.

Whilst the masses and radii of the binary stars will limit the possible orbital period of these systems, eclipsing binaries can be found with a wide variety of periods from a few

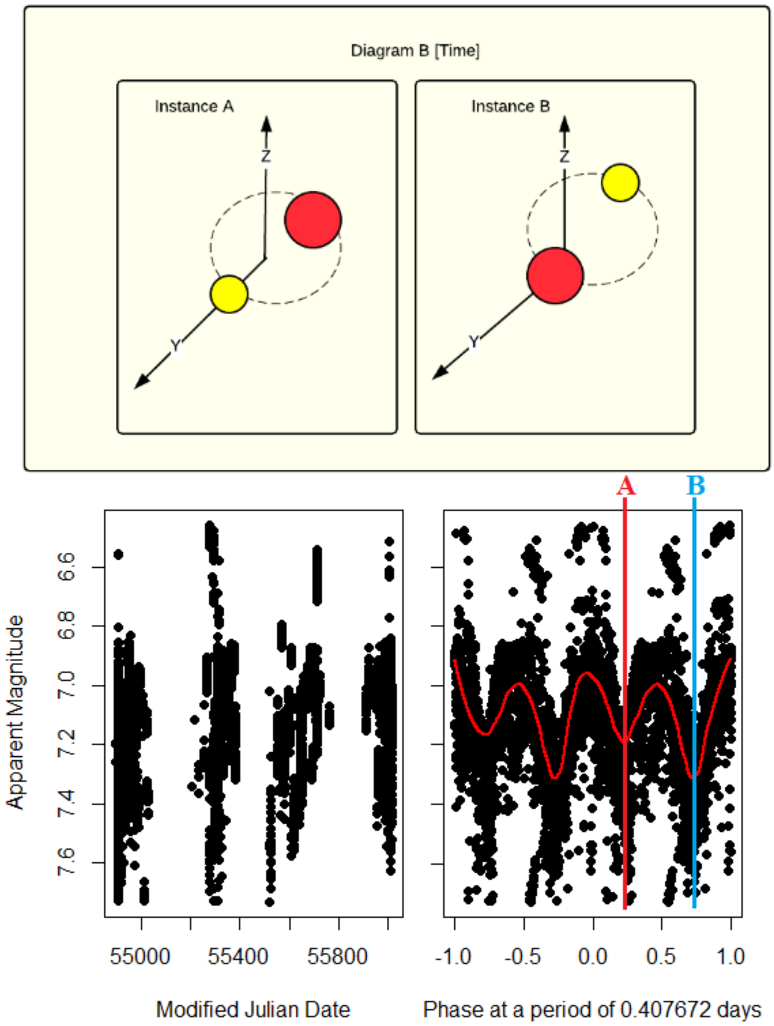


FIGURE 1.14: The light curve of the light source HT Virginis, an eclipsing binary system. The y-axis shows the line of sight from the Earth.

hours up to multiple decades (LaCourse et al., 2015). This is simply due to the wide range of possible stellar masses and radii allowing many stable configurations. Due to the relationship with orbital period and stellar mass and radii, eclipsing binaries are not subtyped by their periods or amplitudes. Instead, the light curve shape is used to classify eclipsing binaries into three common subtypes (Prsa et al., 2008). A word of caution however as classifying based on light curve shape, known as phenomenological classification, does not have a simple mapping to the astrophysical classification. Whilst there is a reasonable comparison between the three phenomenological types and the three astrophysical types, there is also substantial, well discussed overlap.

These three phenomenological classifications are now highlighted with their closest matching astrophysical type acknowledging the above warning.

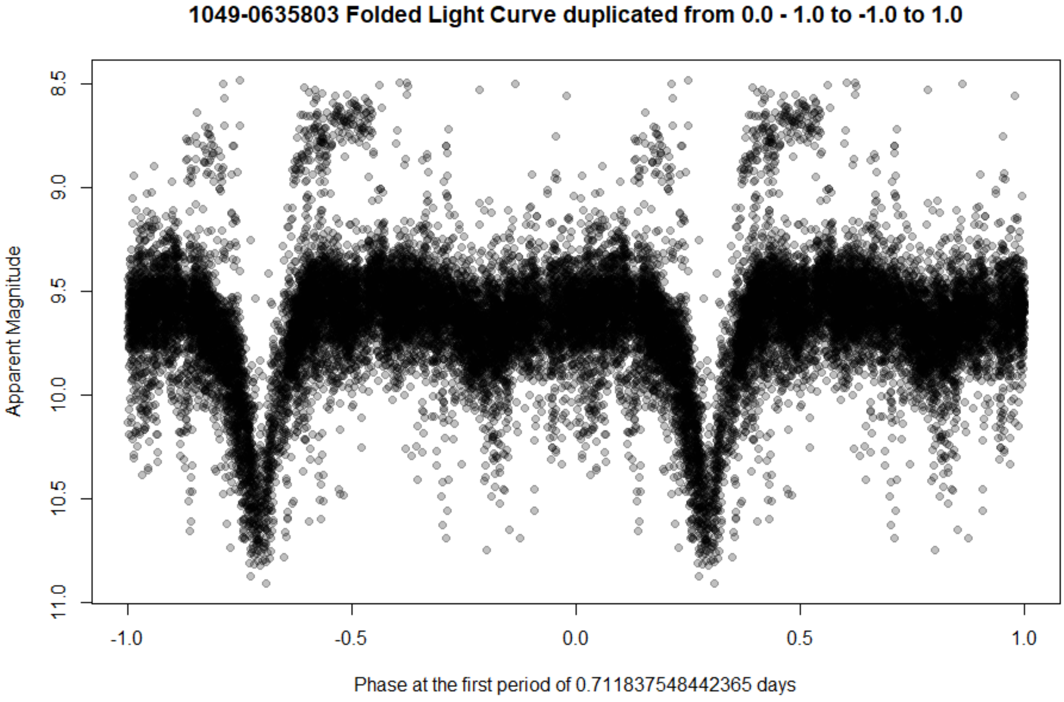


FIGURE 1.15: Phased SkycamT light curve of the β Persei eclipsing binary DI Pegasi with a period of 0.712 days.

- **β Persei eclipsing binaries**

  β Persei (Algol) eclipsing binary light curves exhibit sharp, narrow primary eclipses with a short duration relative to the period of the system. The maxima are almost flat due to the undistorted component stars and are highly non-sinusoidal (Percy, 2008; Kochoska et al., 2017). Some of these light curves contain a secondary eclipse and some do not depending on the relative radii and temperature between the two components (Prsa et al., 2008). Most β Persei light curves are detached eclipsing binaries in which neither component fills its Roche lobe and can have a wide range of orbital periods, orbital eccentricity and component spectral types. Eccentric systems result in light curves with an unequal phase difference between primary to secondary and secondary to primary eclipses (Paegert et al., 2014). In systems where there is no secondary eclipse the eccentricity is not so easily determined. Interestingly, β Persei itself is a semi-detached binary system and one of the components has overflowed its Roche lobe and is now transferring material to the other (Percy, 2008). Most β Persei systems consist of dwarf, subgiant and giant stars yet some contain subdwarf and red dwarf stars. These systems are sometimes named HW Virginis binaries and have short duration orbital periods of down to 0.1 days (Kilkenny et al., 1998). At the other extreme, detached binaries have been measured with periods of thousands of days. ϵ Aurigae is an eclipsing binary system containing a pulsating semi-regular yellow supergiant and a spectral type B main sequence star (or possibly a binary of two type B stars) surrounded by a large dust torus. It undergoes a two year eclipse every 27 years, the orbital period of the binary system (Hoard et al., 2012). The most recent eclipse was in 2009-2011. They can be difficult to detect without an extended survey campaign as few observations are likely to be performed during eclipses. This can set upper limits on the detectability of these binaries. Preliminary investigation for the Large Synoptic Survey Telescope (LSST) places a limit of periods of 10 days on the reliable recovery of β Persei systems (Wells et al., 2017). Figure 1.15 shows the phased SkycamT light curve of the eclipsing binary DI Pegasi with its narrow primary eclipses. The secondary eclipses are barely visible.

- **β Lyrae eclipsing binaries**

  β Lyrae (Sheliak) eclipsing binaries have slightly rounded eclipsing light curves as the stars are distorted into ellipsoids (Percy, 2008). They have primary and secondary eclipses which are uneven in size but are always present and wider than those in β Persei light curves. They are closer to sinusoid in shape with less time spent out-of-eclipse than the β Persei light curves (Prsa et al., 2008). They have a range of periods although not to the extreme extent that β Persei binaries do as the two components must be ellipsoidally distorted. The prototype of this type of eclipsing binary, β Lyrae, is an unusual example of this class as one of the stars is obscured by a large dust disk accreted from the other component (Percy, 2008). Most β Lyrae systems are semi-detached binary systems where one of the two stars has filled its Roche lobe and is now transferring material to the other star. Figure 1.16 shows the phased SkycamT light curve of the eclipsing binary ES Librae with clearly visible primary and secondary eclipses of different minima.

- **W Ursae Majoris eclipsing binaries**

  W Ursae Majoris eclipsing binaries have short periods, usually under 1 day and wide continuous primary and secondary eclipses with minimal time spent out-of-eclipse. The primary and secondary eclipses have similar depth (within a few tenths of a magnitude) as the two components have almost identical surface temperatures (Percy, 2008). The spectral class of the stars is usually F to G or later. These light curves indicate that the two stars are in contact and are indicative of contact binaries. Contact binaries are systems which have evolved to where both stars overflow their Roche lobes and are now sharing a common atmosphere (Ivanova et al., 2013). Over time they approach the same surface temperature as the common atmosphere reaches a thermodynamic equilibrium. The two stars are highly ellipsoidally distorted and will eventually merge due to orbital decay.

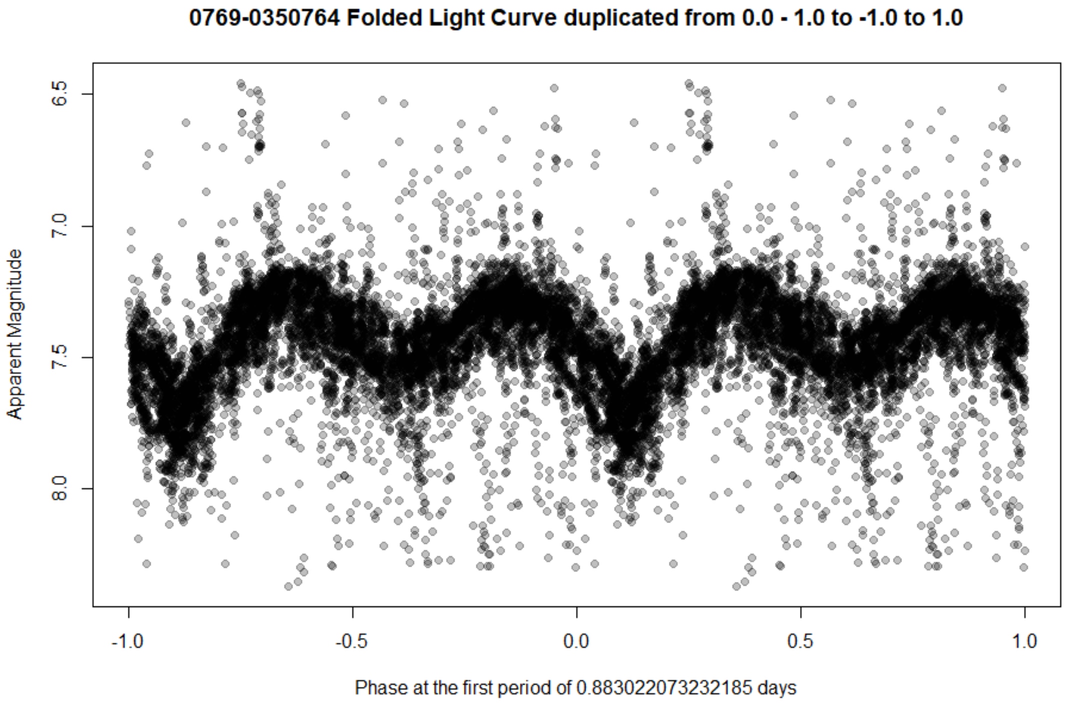


FIGURE 1.16: Phased SkycamT light curve of the β Lyrae eclipsing binary ES Librae with a period of 0.883 days.

Figure 1.17 shows the phased SkycamT light curve of the eclipsing binary AM Leonis with both primary and secondary eclipses being of similar size.

The variability of non-eclipsing binaries is a result of gravitational or magnetic effects from the interacting stars producing a photometric variation. This variability is generally of lower amplitude than the eclipsing binaries but has a larger sufficient inclination range of the orbital plane relative to Earth allowing many more variables of this type to exist.

- **Ellipsoidal variables**

  Ellipsoidal variables are gravitationally interacting binary stars with a small orbital separation. Both stars are distorted by each other's gravity into ellipsoidal shape as their radii is close to their Roche lobes. As the two stars orbit they present a different profile to the observer ranging from circular to egg-shaped (Percy, 2008). This variation in the size of these profiles results in a small amplitude photometric variation with a sinusoidal shape although if the binary stars differ in radii this can deform this shape. The amplitude of these variations are dependent on how close the stellar radius is to the Roche lobe although the amplitudes rarely exceed 0.1 mag (Morris, 1985). Binaries close enough to produce ellipsoidal variability rarely have periods greater than 5 days. Figure 1.18 shows the phased SkycamT

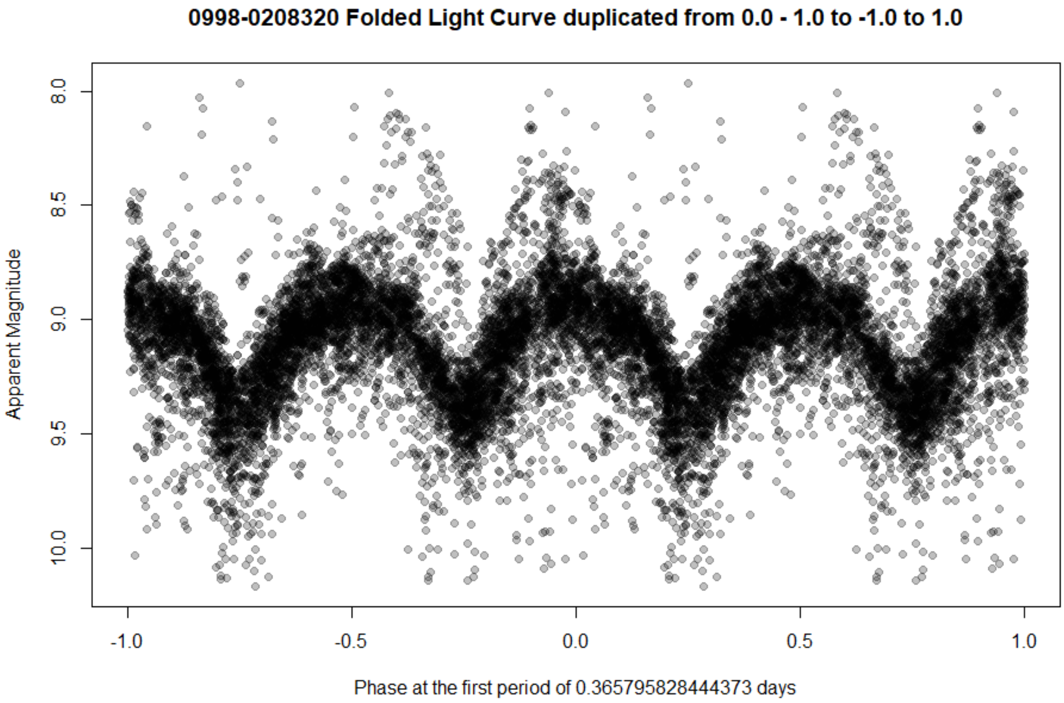


FIGURE 1.17: Phased SkycamT light curve of the W Ursae Majoris eclipsing binary AM Leonis with a period of 0.366 days.

light curve of the ellipsoidal variable NSVS 2600150. It is one of the only ellipsoidal variables with amplitude large enough to be detected in this dataset.

- **RS Canum Venaticorum variables**

  RS Canum Venaticorum stars are surface temperature-asymmetric close orbiting detached binary systems. The cooler component is a giant or subgiant star with spectral classes of G or K. The hotter secondary is a subgiant or dwarf star of spectral class F or G (Percy, 2008). There is also a subtype of RS Canum Venaticorum variables named V471 Taurus variables where the hotter component is a white dwarf. Powerful magnetic chromospheric activity on the cooler binary component produces substantial starspot activity in the photosphere of the giant star. These starspots can potentially obscure up to 50% of the stellar surface (Hall, 1981). The spots are concentrated on the surface facing the hotter companion star. Therefore, as the stars orbit each other this introduces a rotationally modulated variability with a period similar to the orbital period and amplitudes of up to 0.6 mag in the V band (Percy, 2008). RS Canum Venaticorum binaries can also contain eclipses for systems of sufficiently low inclination to allow for them (Sowell et al., 1983). The light curves exhibit a sinusoidal 'distortion wave' with an amplitude and phase

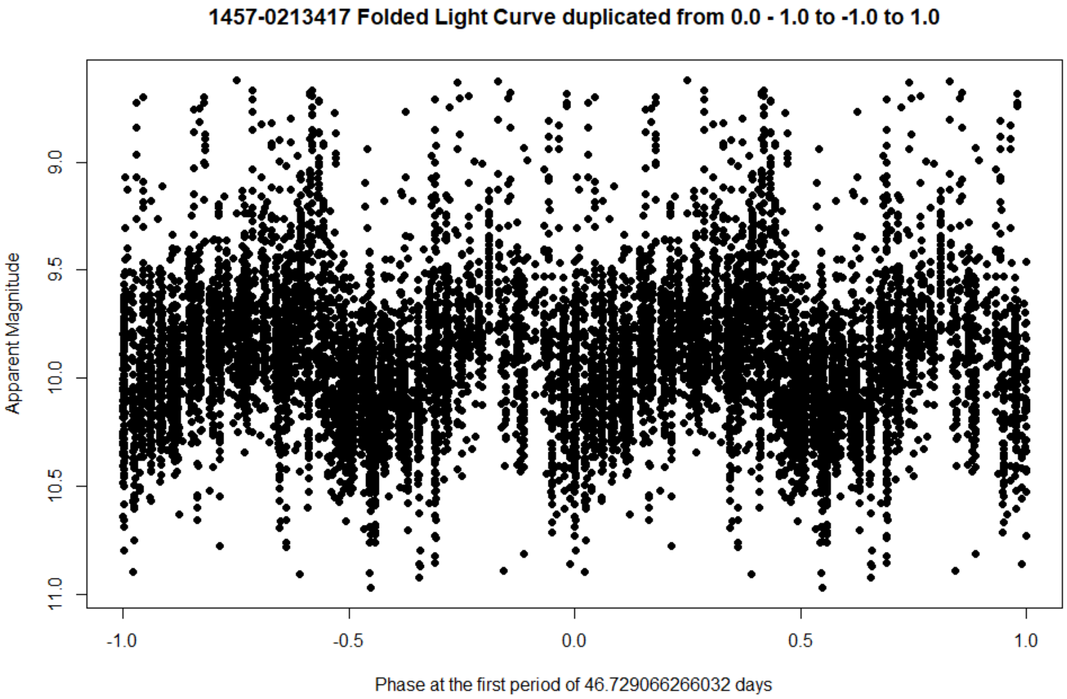


FIGURE 1.18: Phased SkycamT light curve of the ellipsoidal binary NSVS 2600150 with a period of 46.73 days.

that slowly vary relative to the orbital period and, if present, eclipses. These variations are a result of changes in the location and size of the starspots (Hall, 1981; Percy, 2008).

1.4.3 Rotational Variability

The stellar surface is full of active, energetic processes such as stellar flares, starspots, convection cells of hot gas and powerful magnetic events. These events are often localised to specific locations on the stellar surface and as the star rotates these photometrically active regions move into and out of view. This can produce periodic photometric variability equal to the rotational period of the star although as many of these processes themselves vary over characteristic timescales, the variability will fluctuate between cycles. A number of well-known variables are understood to be a result of rotational variability. These variable types are now individually discussed.

- **Sunlike stars**

  The sun has been observed to have various concentrations of sunspots which vary over an 11 year activity cycle. As the sun rotates over its 24.47 day period these sunspots would be rotated in and out of view. If viewed from interstellar distances,

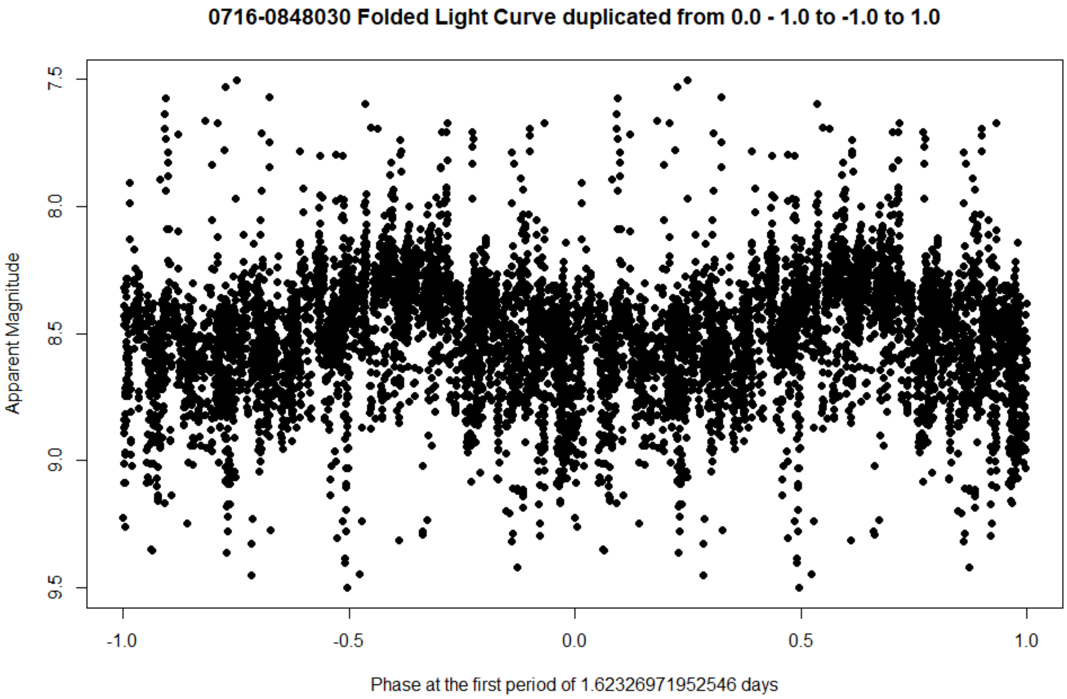


FIGURE 1.19: Phased SkycamT light curve of the α^2 Canum Venaticorum variable AR Capricorni with a period of 1.623 days.

this would produce micro-variability with an amplitude of about 0.01 mag in the V band (Percy, 2008). During the solar minimum this variability would all but disappear due to the low numbers of sunspots present. The light curves from sunlike variability would be reasonably sinusoidal in shape with possible phase changes throughout multiple cycles assuming a uniform distribution of starspots (Percy, 2008).

- α^2 **Canum Venaticorum variables**

  α^2 Canum Venaticorum variables are chemically peculiar stars with spectral classes of late B to early F. As discussed under rapidly oscillating Ap (roAp) stars, chemically peculiar stars are a result of strong magnetic fields allowing the migration of heavy elements to the surface of the star (Kurtz, 1982). Whilst few of these stars pulsate like the roAp stars, many still exhibit a periodic variability equal to the stellar rotation period. This is a result of an inhomogeneous distribution of these heavy elements across the stellar surface producing brightness changes as a function of stellar latitude and longitude (Percy, 2008). As the star rotates these differentially bright regions get rotated in and out of view. They have sinusoidal light curves with strict periods equal to the stellar rotation period, usually a few days, and amplitudes of usually 0.02-0.05 mag although some have amplitudes of up to 0.3 mag (Percy, 2008). Figure 1.19 shows the phased SkycamT light curve

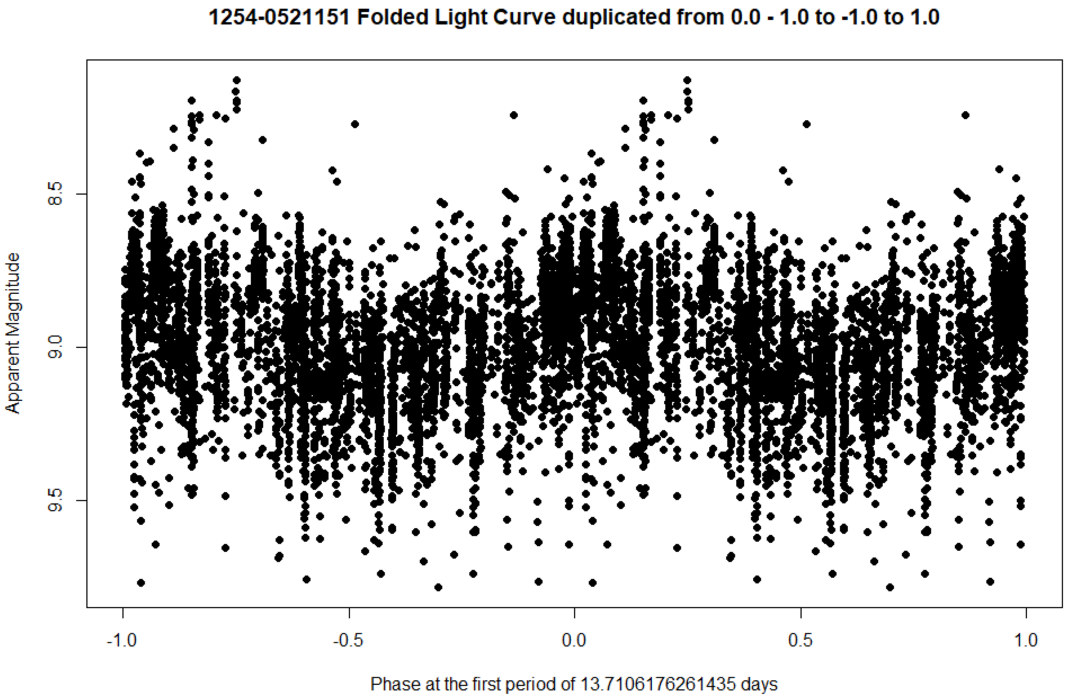


FIGURE 1.20: Phased SkycamT light curve of the BY Draconis variable V0584 Andromedae with a period of 13.71 days.

of the α^2 Canum Venaticorum variable AR Capricorni with an amplitude close to the maximum of the class of 0.25 mag.

- **BY Draconis variables**

  BY Draconis variables are cool stars of spectral type K and M. Their variability is due to large numbers of cool starspots and the stellar rotation. They are found in single and binary systems and have a relatively fast stellar rotation period (Percy, 2008). They have typical periods of 1-10 days and amplitudes of up to 0.3 mag in the V band but more typically around 0.1 mag in the V band. As the starspots appear, grow and decay the amplitude, phase and shape of the light curve slowly changes with time. They can also exhibit flare-like activity. Figure 1.20 shows the phased SkycamT light curve of the BY Draconis V0584 Andromedae.

- **FK Comae Berenices variables**

  FK Comae Berencies variables are late-type giants with spectral classifications G and K. They rotate extremely rapidly resulting in an ellipsoidal shape which results in a changing shape depending on the viewing angle of the observer relative to the rotation axis (Bopp and Stencel, 1981). The period of these variables is of the order of a few days with amplitudes of up to 0.5 mag in the V band although are more typically around 0.2 mag in the V band (Percy, 2008). Their light curves are generally sinusoidal in shape due to the ellipsoidal shape. These stars are possibly formed by a merger event from a contact binary system and are found in both single and binary systems. The fast rotation of these stars allows them to be studied with Doppler Imaging techniques (Korhonen et al., 2005).

- **Pulsars**

  Pulsars are rapidly rotating degenerate stars, either white dwarfs or more often neutron stars. Powerful magnetic fields generate a beam of radiation from their magnetic poles. As the star rotates these beams are swept through space as the magnetic axis and rotational axis are usually misaligned (Lyne and Graham-Smith, 2006). Neutron stars typically have rotational periods of 0.001 to 10 seconds (Percy, 2008). This results in a very rapid pulsing of electromagnetic energy detectable across much of the electromagnetic spectrum. These beams of energy can also induce photometric variability in a nearby companion as it is heated by the sweeping beam of radiation (Percy, 2008).

1.4.4 Eruptive Variability

Eruptive variability tends to be highly non-periodic and is the result of rapid releases of stellar energy or dust obscuration over variable timescales. They can produce large photometric changes very rapidly followed by a more gradual return to normal. These processes can be particularly destructive, disrupting the victim star or even obliterating it leaving a depleted remnant or sometimes nothing at all. Eruptive variability is not necessarily from a stellar source with the energetic supermassive black holes of distant galaxies producing substantial variability due to accretion of vast quantities of matter and the generation of powerful jets. A number of classes of eruptive variability are now discussed.

- **Flare Stars**

  Flare stars, also known as UV Ceti variables, are dwarf stars of spectral class K and M. They are a common type of variable star due to the numerous population of dwarf K and M stars in the galaxy. These stars are photometrically detectable through their rapid increase in brightness of several magnitudes over timescales of seconds to minutes (Percy, 2008). This rapid release in energy is due to large stellar flares liberating energy stored in the stars' magnetic fields. Smaller flares are more common than larger flares and occur at random (Schaefer et al., 2000).

- **Young Stellar Objects**

  Young Stellar Objects (YSO) are young, pre-main sequence stars with ages of less than ten million years old and have spectral classes of F, G, K or M. They have larger radii than main sequence stars of these spectral classes as they are still contracting. They exhibit irregular optical variability of up to 5 mag due to surface processes on the forming stars and sometimes the presence of a disc of accreting material (Percy, 2008). They are also known to generate large bipolar outflows. They are subtyped into a number of groups such as classical T Tauri stars, weak-lined T Tauri stars, Herbig AE/BE stars and FU Orionis variables. Classical T Tauri stars have a circumstellar accretion disc whereas weak-lined T Tauri stars lack this disc (Appenzeller and Mundt, 1989). Herbig AE/BE stars are similar to T Tauri stars but have a higher mass and temperature and FU Orionis stars are T Tauri stars that undergo rapid outburst events with a gradual decay where they brighten by possibly more than 6 mag and change spectral classification from cool spectral class M variables to a spectral class resembling F or G supergiants (Reipurth, 1990).

- **γ Cassiopeiae variables**

  γ Cassiopeiae variables also known as Be stars are non-supergiant stars of spectral class B. They are shell stars, they rotate so rapidly they are ellipsoidally distorted to the degree that material separates from the surface of the star and forms an equatorial disc (Slettebak, 1982). The orientation, thickness and variable presence of this disc can produce intermittent variability with amplitudes of up to 1.5 mag (Percy, 2008; Slettebak, 1982). A subset of these shell stars also exhibit small amplitude variations of up to 0.1 mag with periods of 0.3-2 days named λ Eridani variables. This variability may be rotationally driven or a result of non-radial pulsations (Carrier et al., 2002).

- **R Coronae Borealis variables**

  R Coronae Borealis variables are uncommon, low-mass yellow supergiant variables of spectral classes F and G. They exhibit sudden and dramatic reduction of brightness by up to 10 mag followed by a slow return to maximum light (Clayton, 1996). These events happen unpredictably after long periods of time of quiescence. These brightness reductions also coincide with a reddening of the stars light indicating possible extinction by clouds of dust generated by some unknown process in the stellar atmosphere. As these stars are yellow supergiants, some of them have been seen to exhibit pulsational variability (Percy, 2008).

- **S Doradus variables (Luminous Blue Variables)**

  S Doradus variables, also known as Luminous Blue Variables (LBV) are rare, massive blue supergiant variable stars of spectral classes O and B. They exhibit brightness fluctuations of several magnitudes combined with occasional outbursts and dimming of up to 5 mag (Thackeray, 1974). This can lead to these variables having a combined variability range of 10 mag. They are massive, unstable supergiant stars loosing $10^{-6}\,M_{\odot}yr^{-1}$ of atmosphere through intense stellar winds. They are stars rapidly approaching the end of their lives as a supernova explosion and undergo giant outbursts with increased luminosity and mass-loss as well as the production of surrounding nebulae (Percy, 2008). They can also exhibit short duration variability with periods of under a year and amplitudes of around 0.1 mag as well as stochastic variations. α Cygni variables are an example of Luminous Blue Variables which exhibit multi-periodic non-radial pulsation modes (Percy, 2008).

- **Wolf Rayet stars**

  Wolf Rayet stars are supermassive, evolved stars with masses of 10-25 $M_{\odot}$ and spectral class O. These objects follow the S Doradus variables and are highly luminous Population I stars (Percy, 2008). They have powerful, high velocity, high temperature stellar winds loosing considerable mass every year, up to $10^{-4}\,M_{\odot}yr^{-1}$. They have a highly unstable photometric variability possibly powered by rotation, pulsation, binary interactions and variations in the stellar wind. As these objects are highly helium enriched due to their mass loss, they are possibly the progenitors of type Ib and Ic supernovae as discussed below (Percy, 2008).

- **Cataclysmic variables**

  Cataclysmic variables are binary systems containing a degenerate star, usually a white dwarf, and an evolved companion usually of spectral class G to M which has overflowed its Roche lobe (Percy, 2008). They can exhibit complex light curves as an accretion disc or a powerful magnetic field deposits material onto the white dwarf. Small amplitude variability is common due to thermal instabilities from this deposition of material resulting in stochastic variations. Over longer timescales, more interesting activity can occur. Novae are unpredictable increases in the brightness of a cataclysmic variable source by up to 15 mag over a few days (Darnley et al., 2012). This nova is caused by a thermonuclear reaction when sufficient hydrogen rich material has been deposited on the surface of the white dwarf by the binary companion. These novae can occur over long timescales or, in the case of dwarf novae, over timescales as short as a week by a few magnitudes in close orbiting systems with a dwarf companion. Dwarf novae are not a result

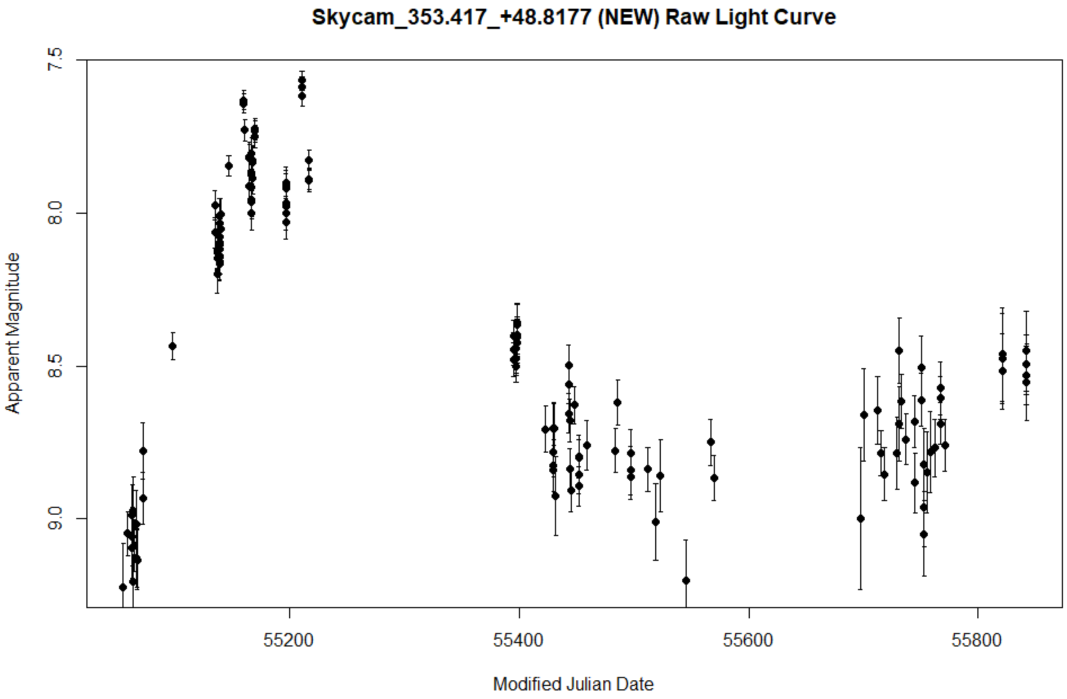


FIGURE 1.21: Raw 2009–2012 SkycamT light curve of the Symbiotic binary Z Andromedae with a bright irregular outburst during the first observation year with a peak brightness at the end of 2009.

of thermonuclear reactions but from an instability in the accretion disc (Percy, 2008). Novae can also be recurrent if they have been detected previously.

- **Z Andromedae variables**

  Z Andromedae variables, also known as symbiotic binaries are binary systems containing a hot main sequence star of spectral class O to B or a white dwarf and a cool giant of spectral class M. The binary system undergoes nova-like outbursts due to the transfer of material from the cool giant which has overflowed its Roche lobe to the hot companion (Percy, 2008). This extended envelope is excited by the hot star's radiation which can result in irregular eruptive events. They are complicated systems and can exhibit variability from nova-like outbursts, pulsational variability from the cool giant, eclipses from the orbital motion and violent eruptions which lead to irregular variations in brightness of up to 4 mag (Percy, 2008). Figure 1.21 shows the raw 2009–2012 SkycamT light curve of Z Andromedae exhibiting a major outburst event with a peak brightness in December 2009/January 2010.

- **Supernovae**

  Supernovae are amongst the most powerful events in the universe. Over a timescale of days a star brightens by 10-20 mag potentially outshining its host galaxy and slowly fading over the next few weeks. These outbursts mark the death of massive stars or the last phase of a cataclysmic variable (Percy, 2008). They irreversibly disrupt the victim star ejecting its material into an expanded shell of ejecta occasionally leaving behind an exotic remnant such as a neutron star or black hole. The initial outburst is a result of the shockwave that ejects the stellar atmosphere and the longer duration afterglow is powered by radioactive decay processes within the ejected material (Percy, 2008). These events are classified under the name 'transient' as they are bright and not expected to be a repetitive event. Supernovae are spectroscopically classified into two subtypes with a number of additional divisions depending on the features of their spectra. Supernovae type Ia are the last stage of a cataclysmic variable (Yoon and Langer, 2004). The thermonuclear detonations that power novae do not eliminate all the accreted material therefore the white dwarf gains mass with each subsequent eruption. Eventually the white dwarf reaches the Chandrasekhar limit, $1.4\,M_{\odot}$, the maximum mass a white dwarf can remain stable. A runaway thermonuclear explosion propagates through the star in less than a second ejecting its material leaving no remnant. As these explosions all have a similar mass and therefore luminosity and are extremely bright, they have been used to determine cosmological distances (Kim and Miquel, 2007). Supernovae type II are core collapse supernova where massive stars are unable to generate more energy and collapse under their own gravity forming an extremely dense core (Heger et al., 2003). A rebounding shockwave propagates out from the core of the collapsing star and ejects the stellar atmosphere in a massive explosion. Supernovae type Ib and Ic are also core collapse supernovae where the progenitor lacked hydrogen in their atmospheres due to some form of atmospheric stripping possibly by a binary companion (Woosley and Eastman, 1997).

- **Active Galactic Nuclei**

  Active Galactic Nuclei (AGN) are photometrically variable sources generated by supermassive black holes in the centre of distant galaxies. BL Lacertae is a variable source initially thought to be a variable star which undergoes stochastic large-amplitude variations of several magnitudes (Falomo et al., 2014). It is now known that BL Lacertae is a Blazar, a type of AGN where the jets generated from the supermassive black hole are orientated towards the observer. The brightest Blazars are BL Lacertae objects and Optical Violent Variables (OVVs) (Percy, 2008). The variable light from these objects is highly polarised during outbursts and is a result of the generation of synchrotron radiation from relativistic particles interacting with magnetic fields in the jet. There can also be a thermal component to this light caused by energy emitted from the supermassive black hole's accretion disc (Urry and Padovani, 1995).

1.5 Aims, Motivations and Research Objectives

In this chapter the background of the Small Telescopes Installed at the Liverpool Telescope has been introduced and a classification scheme detailing multiple variable sources has been described. The STILT data reduction pipeline has reduced the SkycamT and SkycamZ images into photometric light curves calibrated with the US Naval Observatory B catalogue. The STILT database contains the light curves of hundreds of thousands of objects with sufficient observations to determine if a source is variable and the classification of the variable sources. The goal of this thesis is the development, implementation and testing of an automated pipeline designed to process the STILT light curves into variable and non-variable sources. The light curves exhibiting sufficient variability are then classified into variable star classes as described in this chapter. This pipeline must be efficient as it is required to process over half a million light curves. This is a very important task as astronomers require a catalogue of interesting objects for their chosen field of study as follow-up observations have a cost in both time and resources and therefore a selection of targets is required prior to the collection of additional data. The key objectives are:

- Produce a complete literature review that catalogues previous statistical and computational achievements within the fields of period estimation methods and variable light curve detection and classification. This review would be combined with an investigation validating the performance of previous methods on the STILT dataset to produce a control set by which the performance of the proposed novel automated pipeline can be measured. This analysis will also characterise the strengths and weaknesses of the STILT data to assist in the application of novel methods to improve the reliability of the automated pipeline. This objective is explored in chapters 2, 4 and 5.
- Development of a period estimation method designed for the uneven cadence of the STILT light curves. The method is also required to be robust to automation and therefore capable of correcting known failure modes without intervention. This method is to be rigorously tested by applying synthetic light curves with controlled noise statistics. These light curves will have regular cadence variants and Skycam cadence variants to determine the relative performance differences with Skycam cadence. This will help clarify the potential issues with the Skycam survey method in the detection of periodic light curves. This objective is explored in chapter 3.
- Identify the multiple statistical features described in the literature review that can be extracted from the STILT dataset in order to determine those that correlate strongly with the desired machine learning classification task. Such results will

improve understanding and strongly assist in the production of novel features. Develop additional features as required to improve the performance of machine learned classifiers. This objective is explored in chapters 5 and 6.

- Develop a method to utilise machine learning algorithms for time-series analysis with the ability to adapt to uneven time-series data to generate automated features. This representation learning approach can allow the machine learning algorithm to construct an independent set of features automatically from the training data without the need for intensive feature engineering. The machine learning models will be compared to the feature engineered models for performance in variable classification tasks. This objective is explored in chapter 6.
- Use variable star catalogues to construct a high quality dataset of light curves of multiple variable star types with known periods for the training of machine learning classification models. This cross-matched dataset is required for the comparison of the period estimation methods. The variable detection and classification models are trained with a number of different machine learning algorithms and their relative performance compared using metrics such as Sensitivity and Specificity, Receiver Operating Characteristic (ROC) curves and the Area under the curve (AUC) statistic and the Brier score of the cross-validated and tested models. This objective is explored in chapter 7.
- Upon the training of the best performing variable light curve detection and classification models, apply the automated pipeline to the sufficiently sampled STILT light curves to identify candidate variable sources from the database and determine anomalous light curves for further investigation. This objective is also explored in chapter 7.

1.6 Thesis Structure

The layout of the remaining chapters of this thesis is as follows. Chapter 2 reviews the literature on the period estimation of time-series data with the goal of characterising the value and strength of periodic activity. This can be performed using a number of methods which are detailed and tested on Skycam light curves. The characterisation of the significance of periodic activity and the possible failure modes of the period estimation algorithms are also discussed.

Chapter 3 presents GRAPE: Genetic Routine for Astronomical Period Estimation, a novel method for the rapid and accurate estimation of periods tuned specifically for the Skycam data. The chapter carefully describes the operations present in this method and subjects the technique to a selection of tests to determine the computational and performance characteristics of the method using a set of synthetic data generated to mirror periodic Skycam light curves.

Chapter 4 reviews the capabilities of Machine Learning classifiers in the production of automated classification models. The performance measures used in the determination of the capability of the automated pipeline classification models are also defined.

Chapter 5 introduces the numerous engineered features, the inputs to the classification algorithms. Many of these features have been developed by previous research and are described for their inclusion into the new Skycam classification pipeline. At the end of the chapter a Random Forest classification model is trained using these features to demonstrate the strengths and weaknesses of this approach and to highlight potential improvements to the feature set.

Chapter 6 presents the use of representation learning to allow machine learning algorithms to produce novel features suited to the classification tasks. This chapter introduces the use of Machine Learning algorithms to automatically extract features from the light curves based on an epoch-folded image-based representation. The chapter also introduces a genetic algorithm enhanced polynomial interpolation procedure named the Genetic PolyFit algorithm. The PolyFit algorithm is a method design for modelling eclipsing binaries (Prsa et al., 2008) which has been enhanced with a novel genetic algorithm optimisation to accurately determine the fitted model for noisy SkycamT light curves. This algorithm allows the interpolation of the light curve for the removal of noise and the generation of more accurate statistical features of Skycam light curves based on the GRAPE estimated period. The Genetic PolyFit is described in detail and the application of novel automated shape representation features with a Principal Component Algorithm is discussed.

Chapter 7 details the construction and performance of the final classification pipeline which builds on the design of Richards et al. ASAS classifier (Richards et al., 2012) whilst implementing a novel SkycamT Trend Removal Pipeline based on the Trend Filtering Algorithm (TFA) (Kovacs et al., 2005), the features produced from the Genetic PolyFit in chapter 6 and a period estimation from GRAPE in chapter 3.

Finally, Chapter 8 presents the conclusions of this thesis and proposed future work to build on the conclusions.

Chapter 2

Period Estimation

The classification of variable stars requires the determination of important properties of their light curves. Three of the most important features, especially for pulsating variables, are *period*, *amplitude* and *colour* (Debosscher et al., 2007; Richards et al., 2011b). Colour is easily determined from multiband photometry, although in the case of white light instruments such as the Skycams, this information must be extracted from catalogue data (Mawson et al., 2013). For period and amplitude, these features are linked as amplitude is best determined by treating it as function of period to separate the signal and noise components and an incorrect determination of period will lead to an incorrect amplitude resulting in two of the three main features performing poorly. This is even an underestimate of how destructive a poorly calculated period is as many features that attempt to characterise the shape and dispersion of periodic light curves are generated within *phase space* which itself is dependent on the determined period (Richards et al., 2011b).

It is clear it is worth expending a large fraction of the computational time on a given light curve on correctly estimating the underlying primary period (McWhirter et al., 2016). As a result, multiple methods of period estimation have been developed with a variety of differing approaches. These range from fitting sets of sinusoids to the data to minimising the dispersion of the data points for different candidate periods across a dense grid (Lomb, 1976; Scargle, 1982; Stellingwerf, 1978). The majority of period estimation methods can be grouped into three subtypes, frequency domain methods, time domain methods and epoch folding methods (Huijse et al., 2012). Frequency domain methods which utilise Fourier transforms and their uneven data analogues to deconstruct a light curve into a series of frequency components with associated amplitudes and phases (Schuster, 1898; Lomb, 1976; Scargle, 1982; Zechmeister and Kürster, 2009). Time Domain methods identify groups of data points which align with others over given time lags and if enough groups share this correlation it is evidence for an underlying period Jenkins and Watts (1968). Finally, epoch folding methods 'fold' the light curves over candidate periods and characterise the resulting phased light curves for correlations and signals for these periods (Lafler and Kinman, 1965; Stellingwerf, 1978; Schwarzenberg-Czerny, 1989; Clarke, 2002). As period estimation methods produce the same object, a real-valued period, they can be used interchangeably depending on the quality of the data and the class of light curve of most interest. These algorithmic approaches may also be combined in linear and non-linear ways to produce statistics that benefit from the individual components (Saha and Vivas, 2017).

Neural Networks have also been used to extract useful information from light curves for period estimation tasks. A prominent example uses a non-linear neural network to extract the important frequencies of a light curve which is then classified using the Multiple Signal Classification (MUSIC) algorithm to reproduce the primary periods through astrophysical noise (Tagliaferri et al., 1999). Gaussian Processes have been introduced to model the secondary stochastic variability of Mira variables (He et al., 2016). They are good for fitting non-sinusoidal light curves and they require reasonably good sampling to fit the stochastic variations. Wavelets are another set of period finding algorithms suitable for light curves with quasi-periodic variability. These algorithms are utilised to extract information at multiple frequency ranges. This has been used to identify the redshift of detected Gamma Ray Burst (GRB) transient events (Ukwatta and Wozniak, 2015). The correct periodogram for a survey is not a precisely known factor although many appear to be similar in performance (Heck et al., 1985) and research has been performed to understand the relationship between these algorithms and survey data (Graham et al., 2013b).

Deep Learning presents another avenue for period estimation. Despite the multiple approaches discussed in this chapter, the 'perfect' periodogram does not yet exist. The perfect periodogram would have a high recovery fraction down to low numbers of observations and/or signal to noise whilst being independent of the shape of the periodic signal. The learning capability of modern networks can be trained on a set of data with known periods. The performance of these networks can provide a strong indication towards the optimal period estimation approach. Work in this area is only just beginning with previous deep learning light curve classification approaches still relying on a separate period estimation method prior to processing. An example of this is the deep learning classification applied to Kepler light curves which used a Box Least Squares (BLS) algorithm to process the light curves prior to training (Shallue and Vanderburg, 2018). The first attempts to use deep learning for period estimation have only recently been published in the form of a deep one-dimensional Convolutional Neural Network (CNN) designed to process simulated data from a high-cadence space telescope to search for shallow planetary transits (Zucker and Giryes, 2018; Pearson et al., 2018). The deep CNN outperformed the BLS algorithm in detecting the true simulated transits whilst minimising the false positives despite the presence of correlated noise generated by a Gaussian process. The performance of the deep CNN remained strong in the regime of low signal-to-noise light curves.

The potential for deep learning is very promising with continued research in the next few years as the next generation of surveys become available. A major difficulty of the deep learning approach is formulating a topology for entering the data into the deep learning network. The shallow transit model discussed in the previous paragraph accomplished this by using a high cadence space telescope with well sampled, evenly distributed observations. Such a method is not available for ground based surveys and the current method of phasing a light curve prior to analysis presents a clear problem as the period estimation must be computed prior to the deep learning. Until this problem is resolved, ground based surveys must depend on traditional methods. In this chapter a number of proven methods are discussed based on subtype and their performance compared using a set of STILT light curves. Methods of defining this performance and the possible failure modes are also described.

2.1 Frequency Domain Methods

Fourier analysis is based on Fourier series. A Fourier series is an infinite sum of sine and cosine functions with the frequencies, amplitudes and phases of these component sinusoids providing the required degrees of freedom to exactly represent a time-series such as a light curve (Fourier, 1878). The Fourier transform is a mathematical transformation that can transform the time-series from time-space to frequency-space where the Fourier components are represented as a set of peaks at the associated frequencies with a magnitude based on the amplitude. The difference between the amplitudes of the sine and cosine functions at the same frequency describe the phase of the original signal component (Debosscher et al., 2007). As the infinite frequencies are impossible to compute for a finite problem, the Discrete Fourier Transform (DFT) allows this computation for a set of discrete frequency components of sufficient density to closely approximate the time-series (Jenkins and Watts, 1968). The Discrete Fourier Transform of a finitely-sampled and evenly-sampled signal x_n with data points $n = 0, \ldots, N-1$ where N is the number of data points is defined in equation 2.1.

$$X_k = \sum_{n=0}^{N-1} x_n \cdot \exp\left(-\frac{2\pi i}{N}kn\right) \tag{2.1}$$

The conversion from an continuous signal into a discrete signal results in an effect named aliasing where a significant Fourier component can become indistinguishable at a number of different frequencies due to the sampling rate. An appropriate choice of sampling rate can be used to minimise this disruption. For a sampling frequency F_s, the DFT produces a set of frequencies $f = \frac{k}{N} F_s$. The Shannon-Nyquist Sampling Theorem defines a frequency around which a reflected alias frequencies become indistinguishable from the true frequencies as a function of the sampling rate of the signal (Shannon, 1949). This is named the Nyquist frequency and is shown in equation 2.2.

$$f_N = \frac{F_s}{2} \tag{2.2}$$

Where f_N is the Nyquist frequency and F_s is the sampling rate. The Fourier transform contains the full information on the frequency, amplitude and phase of the sampled signal but when attempting to identify clear frequencies in the data it is better to analyse the signal power as a function of frequency. The power spectral density is the square of the absolute values of the Fourier transform. The classical periodogram or the Schuster periodogram can estimate the power spectrum of a time-series directly from the data and present it via a metric representing the strength of periodic activity at a given frequency against a frequency (the inverse of the period) spectrum (Schuster, 1898). The strongest components result in peaks located at the dominant periods in the data. The periodogram is defined in equation 2.3.

$$P_s\left(f\right) = \frac{1}{M} \left| \sum_{n=0}^{M-1} x_N \cdot \exp\left(-\frac{j2n\pi f}{F_s}\right) \right|^2 \quad \text{where} \quad -f_N < f < f_N \tag{2.3}$$

where $P_s(f)$ is the periodogram power at frequency f, M is the number of data points in the time-series, F_s is the sampling rate and f_N is the Nyquist frequency. For an evenly sampled signal the peak of the periodogram will quickly identify the dominant period in a time-series with a peak at the associated frequency at $f = \frac{1}{P}$ where f is the frequency with the maximum periodogram value and P is the dominant period of the time-series.

Ultimately, the DFT and Fourier transform are of limited usefulness in light curve period estimation as they require the light curve to be evenly sampled (Lomb, 1976; Scargle, 1982). Applying them to a light curve with uneven data points results in unreliability in the resulting peaks as the phases of the different frequency components are not consistent. A light curve can be interpolated into evenly sampled points but this typically introduces additional noise into the periodogram and unreliable results (VanderPlas, 2017). The solution is the implementation of a method designed around the correction of the uneven sampling.

2.1.1 Lomb-Scargle Periodogram

The Lomb-Scargle periodogram (LSP) was developed independently by Lomb and Scargle for the fitting of unevenly sampled time-series (Lomb, 1976; Scargle, 1982). The LSP can be considered in a least-squares view as the minimisation of a sinusoidal model as shown in equation 2.4 (VanderPlas, 2017).

$$y\left(t; f\right) = A_f \sin\left(2\pi\left(t - \phi_f\right)\right) \tag{2.4}$$

where A_f is the amplitude of frequency f and ϕ_f is the phase of frequency f. By calculating a $\chi^2(f)$ metric at a set of candidate frequencies, shown by equation 2.5, the performance of a fitted sinusoidal model at that frequency can be determined. The power of the Lomb-Scargle periodogram is then proportional to the reciprocal of the $\chi^2(f)$ values and the minimised $\hat{\chi}^2(f)$ is located at the best fitting frequency f.

$$\chi^2(f) = \sum_N \left(y_n - y(t_n; f)\right)^2 \tag{2.5}$$

where $\chi^2(f)$ is the χ^2 statistic at frequency f, N is the total number of data points in the time-series, y_n is the univariate time-series value at the n^{th} data point and $(y_n - y(t_n; f))$ is the predicted univariate time-series value by the sinusoidal model at a frequency f at the time instant t_n.

The solution to the inconsistent phases of the frequency components is to introduce an arbitrary phase correction τ which acts at each frequency to generalise the periodogram. Scargle introduced the periodogram in the form shown in equation 2.6 (Scargle, 1982).

$$P(f) = \frac{A^2}{2}\left(\sum_n y_n \cos\left(2\pi f(t_n - \tau)\right)\right)^2 + \frac{B^2}{2}\left(\sum_n y_n \sin\left(2\pi f(t_n - \tau)\right)\right)^2 \tag{2.6}$$

where $P(f)$ is the periodogram power, y_n is the univariate value at the n^{th} data point, t_n is the n^{th} time instant and A, B and τ are functions chosen as to correct the phases at each time instant for each frequency. The solutions for these three functions lead to the Lomb-Scargle periodogram as shown in equation 2.7 with a phase correction 2.8 (Lomb, 1976; Scargle, 1982).

$$P_{\text{LS}}(f) = \frac{1}{2}\left[\frac{\left(\sum_n y_n \cos\left(2\pi f(t_n - \tau)\right)\right)^2}{\sum_n \cos^2\left(2\pi f(t_n - \tau)\right)} + \frac{\left(\sum_n y_n \sin\left(2\pi f(t_n - \tau)\right)\right)^2}{\sum_n \sin^2\left(2\pi f(t_n - \tau)\right)}\right] \tag{2.7}$$

$$\tau = \frac{1}{4\pi f}\tan^{-1}\left(\frac{\sum_n \sin(4\pi f t_n)}{\sum_n \cos(4\pi f t_n)}\right) \tag{2.8}$$

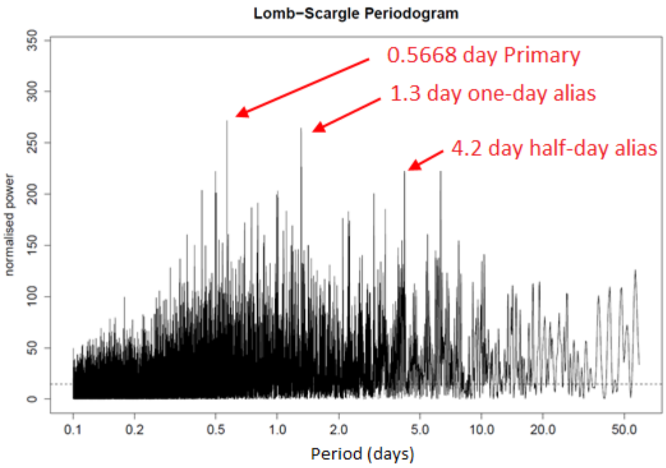


Figure 2.1: The Lomb-Scargle periodogram of the star RR Lyrae. The three main peaks are highlighted and are the true period and two sidereal day aliases due to the sampling cadence of ground based surveys.

This equation forms the basis of one of the most common period estimation methods in astronomy today (Debosscher et al., 2007; Richards et al., 2011b, 2012). Figure 2.1 shows a Lomb-Scargle periodogram computed on the variable star RR Lyrae with a period range of 0.1 days to 60 days (Ruf, 1999). The three dominant peaks are the true underlying period and two aliases. Despite the popularity of the approach, it does exhibit a number of drawbacks that inhibit its use. Like the classical periodogram, the LSP suffers from the leakage of spectral power into aliases and harmonics of the true underlying frequency (VanderPlas, 2017). It can also be contaminated by spurious peaks due to sampling periodicities within the time instants and correlated noise. In the worst cases, these two effects can superimpose resulting in a noisy, aliased peak that has a higher power value than the true result. Close analysis of the underlying sampling periodicities and frequency harmonics is recommended for this method and often requires human study to determine good candidate periods. Another difficulty presented by the LSP is the sinusoidal model it relies on for period fitting. If a periodic signal is highly non-sinusoidal such as those produced by eclipsing binary light curves, the periodogram power is substantially weakened and therefore even more at risk of aliased and spurious peaks. As a result, research continues to search for a version of this periodogram equivalent to fitting an arbitrarily shape model to satisfy all possible signals (Schwarzenberg-Czerny, 1996).

Upcoming surveys will gather light curves in a number of different filters or bands. Multiband periodograms are of interest for the processing of this data as the periodicity in each band can be modelled and uses to reinforce the confidence in detected frequencies (Vanderplas and Ivezić, 2015). When the light curves are poorly sampled, the multiband generalised Lomb-Scargle algorithm can be improved through the intelligent combination of information from each band using a penalised generalised Lomb-Scargle algorithm through penalising the phases and amplitudes of the different multiband functions (Long et al., 2016).

2.1.2 Generalised Lomb-Scargle Periodogram

The Generalised Lomb-Scargle periodogram (GLS) is an extension to the Lomb-Scargle periodogram through the addition of two main generalisations (Zechmeister and Kürster, 2009). The first is the addition of a floating mean constant term that allows the fitted sinusoids to have a non-zero mean as shown in equation 2.9.

$$y\left(t;f\right) = C(f) + A_f \sin\left(2\pi f\left(t - \phi_f\right)\right) \tag{2.9}$$

where $C(f)$ is the floating mean constant term for the sinusoidal model at each frequency. Previously the mean of a univariate time-series would be subtracted off all the data points and the LSP would assume a zero mean. In astronomy this is a risky procedure as light curves can lack data coverage such as with an object that varies above and below a given instruments sensitivity limit. When this occurs, the fainter points are potentially not recorded in the light curve. This would result in a truncated light curve displaying only the brightest observations of the variations. In this case, the mean of the recorded data points is brighter than the true mean brightness of the source. For a signal well-modelled by a sinusoid this would likely lead to the LSP underestimating the true period with the half-period harmonic by fitting a lower amplitude sinusoidal model to the bright half of the light curve (VanderPlas, 2017). The second addition is the application of an error term to the χ^2 fitting procedure allowing the periodogram to take into account poor quality data points and outliers as shown in equation 2.10.

$$\chi^2(f) = \sum_N \frac{\left(y_n - y(t_n; f)\right)^2}{\sigma_n^2} \tag{2.10}$$

where N is the number of data points and σ_n is the error in the univariate value of the n^{th} data point. The Phase Distance Correlation periodogram extends the GLS periodogram into a correlation measure using the fact that phase is a circular variable. This approach improves on the GLS performance for sawtooth type light curves such as those from Cepheid and RR Lyrae variable stars (Zucker, 2018).

2.1.3 Variance Ratio Periodogram

The Variance Ratio Periodogram (VRP) is a version of the multi-harmonic Fourier periodograms. Multi-harmonic periodograms fit a harmonic sinusoidal model with extra

Fourier components at integer multiples of the candidate frequency (Baluev, 2009; VanderPlas, 2017; Saha and Vivas, 2017). This allows the periodogram to accurately model more non-sinusoidal light curve shapes producing stronger peaks at the frequencies of these signals. Unfortunately, the additional degrees of freedom in the model allow the periodogram to better fit the noise at an incorrect frequency increasing the strength of the continuum peaks (VanderPlas, 2017). The multi-harmonic periodogram is displayed in equation 2.11 with $K - 1$ harmonic terms.

$$y\left(t;f\right) = C(f) + A_f^{(0)} + \sum_{k=1}^{K} A_f^{(k)} \sin\left(2\pi f\left(t - \phi_f\right)\right) \tag{2.11}$$

This periodogram is useful on the light curves of eclipsing binaries due to its more non-sinusoidal nature. A common failure mode of the Lomb-Scargle and Generalised Lomb-Scargle periodograms on these objects is obtaining a peak value at half the true period as the single sinusoid model cannot closely fit the primary and secondary eclipse (VanderPlas, 2017). The multi-harmonic periodogram is capable of accomplishing this but at the risk of falling victim to noisy peaks. The multi-harmonic periodogram also produces harmonic peaks at integer submultiples of the true frequency due to the harmonics achieving a reasonable performance where the primary frequency component cannot. The more harmonics in the model, the more harmonic peaks in the resulting periodogram which can result in an incorrect result in an automated period estimation task.

The variance ratio periodogram used in this work functions by fitting a harmonic model of the form shown in equation 2.11 for a set of candidate frequencies. The variance of the light curve is then measured before and after a prewhitening operation performed for the harmonic model for each frequency (Debosscher et al., 2007; McWhirter et al., 2016). The prewhitening operation involves the subtraction of the harmonic model from the light curve leaving a residual light curve. The ratio between the residual light curve and the initial light curve as a function of frequency is used as the period estimation statistic and is shown in equation 2.12.

$$P_{\mathrm{VRP}}(f) = \frac{\sigma^2_{\mathrm{res}}(\theta; f)}{\sigma^2_{\mathrm{lc}}} \tag{2.12}$$

where $P_{\mathrm{VRP}}(f)$ is the variance ratio for a frequency f, $\sigma_{\mathrm{res}}(f)^2$ is the variance of the residual light curve produced from prewhitening the light curve with the harmonic model θ at frequency f and σ^2_{lc} is the variance of the initial light curve. This statistic has the advantage of being automatically normalised between 0 and 1. A poor fitted model will remove no variance from the light curve leaving the variance ratio as $\frac{1}{1} = 1$, whereas a perfect fit will remove all the variance from the residual light curve with a variance ratio of $\frac{0}{1} = 0$. The best fitting frequencies will minimise the variance ratio periodogram.

2.2 Time Domain Methods

Time domain methods are a set of methods which compute some measure of similarity between neighbouring data points separated by a candidate time gap (a candidate period). Through the comparison of every pair of data points $[x_i, x_j]$ and their similarity as a function of their time instants $[t_i, t_j]$, a measure of periodic or semi-periodic activity is possible.

2.2.1 Autocorrelation Function

The autocorrelation function is a measure of how similar a signal is to itself when separated by a time lag τ. It is calculated using equation 2.13 for a discrete time-series (Jenkins and Watts, 1968).

$$r_x(\tau) = \frac{1}{N - \tau + 1} \sum_{i=\tau}^{N-1} x_n \cdot x_{i-\tau} \tag{2.13}$$

where τ is a time lag (related to a trial period), $r_x(\tau)$ is the autocorrelation for a candidate time lag, N is the number of data points in a time-series and x_i is the i^{th} univariate time-series value. For values of τ with multiple data points with similar univariate values x_i, the autocorrelation function $r_x(\tau)$ is maximised.

The autocorrelation function has a minimal computational cost allowing it to be used for a rapid analysis of potential periodic activity. It suffers from degraded performance in situations with multi-periodic signals as one period repositions the time lag separated data points of a second period. Additionally, correlated noise within the time-series produces a slope in the autocorrelation function which may result in false positive period detections. The autocorrelation function is not functional for time-series with unevenly sampled data and requires a modification to the discrete or slotted autocorrelation function which introduces a hyperparameter named the slot size to generate lags as a function of the slot size (Edelson and Krolik, 1988).

2.2.2 Correntropy Kernelised Periodogram

The Correntropy Kernelised Periodogram (CKP) is based on the concept of Correntropy. Correntropy or Correlated Entropy measures the similarity between the information entropy or content of the data points of a time-series separated by a time lag (Liu et al., 2006; Huijse et al., 2012). It is related to autocorrelation but with a kernel function which measures the similarity in a higher-dimension kernel space and is defined in equation 2.14.

$$V(t_1, t_2) = \mathbb{E}_{x_{t_1} x_{t_2}} \left[\kappa(x_{t_1}, x_{t_2}) \right] \tag{2.14}$$

where $\mathbb{E}$ is the expectation value of a function and κ is a positive definite kernel function. The application of a Gaussian kernel function allows the definition of the autocorrentropy of a univariate time-series using equation 2.15 with a similar appearance to the autocorrelation defintion in equation 2.13.

$$\hat{V_{\sigma_m}}[\tau] = \frac{1}{N - \tau + 1} \sum_{i=\tau}^{N} \frac{1}{\sqrt{2\pi}\sigma_m} \exp\left(-\frac{||x_i - x_{i-\tau}||^2}{2\sigma_m^2} \right) \tag{2.15}$$

where τ is a candidate time lag, $\hat{V_{\sigma_m}}[\tau]$ is the autocorrentropy value of time lag τ, N is the number of data points in the time-series and x_i is the univariate value of the i^{th} data point.

The CKP makes use of a periodic Gaussian kernel function shown in equation 2.16. This function is cyclical and repeats with an arbitrary period P.

$$G_{\sigma_t;P}(z - y) = \frac{1}{\sqrt{2\pi}\sigma_t} \exp\left(-\frac{2\sin^2\left(\frac{\pi}{P}(z - y)\right)}{\sigma_t^2} \right) \tag{2.16}$$

where σ_t is the Gaussian standard deviation and P is a candidate period. Using the correntropy function and the periodic kernel function the CKP is computed using equation 2.17.

$$\hat{V}_{P(\sigma_t,\sigma_m)}(P_t) = \frac{\sigma_m}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\frac{1}{\sqrt{2\pi}\sigma_m} \exp\left(-\frac{||x_i - x_j||^2}{2\sigma_m^2} \right) - \text{IP} \right) \cdot G_{\sigma_t;P}(t_i - t_j) \cdot \omega(t_i - t_j) \tag{2.17}$$

where $\hat{V}_{P(\sigma_t,\sigma_m)}(P_t)$ is the CKP statistic, σ_m and σ_t are hyperparameter variables named the magnitude kernel size and the time kernel size of which σ_m can be estimated from the time-series, $G_{\sigma_t;P}(t_i - t_j)$ is the periodic kernel function for a trial period P_t, IP is the Information Potential defined in equation 2.18 and $\omega(t_i - t_j)$ is a Hamming window defined in equation 2.19.

$$IP_{\sigma_m} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_m} \exp\left(-\frac{||x_i - x_j||^2}{2\sigma_m^2} \right) \tag{2.18}$$

$$\omega(t_i - t_j) = 0.54 + 0.46 \cdot \cos\left(\frac{\pi(t_i - t_j)}{T} \right) \tag{2.19}$$

where T is the time span of the time-series. The Hamming window smooths the estimation of the periodogram.

The CKP is very flexible as the correlation based metric is independent of the shape of a light curve signal (Huijse et al., 2012). It has two primary weaknesses, the determination of the hyperparameter σ_t and the computational overhead. σ_t has been determined to be primarily dependent on the sampling pattern of the light curves therefore it should maintain a small range for a set of light curves from the same survey. The computational overhead discourages its use in period estimation tasks for time-series with many observations or over a large period range. Optimisation techniques may assist in exploring the period space more efficiently than a brute-force approach although as the CKP as a statistic is usually a highly non-linear function of the trial period P_t which may limit such an approach.

2.3 Epoch Folding Methods

Epoch Folding is an astronomical analysis technique which involves the mapping of time-series data from a *time representation* to a *phase representation*. The phase representation is a function of a candidate period P and when this candidate period aligns with the period of a true underlying periodic signal, each individual oscillation of the signal is superimposed onto one waveform (Larsson, 1996). As poor sampling is a common problem in astronomy, this acts to take numerous oscillations with a minimal number of observations and combine them into one which reveals the true underlying waveform. The new phase data points $\phi_i \in \mathbb{R} \in [0, 1]$ represent the decimal periods from the time instants t_i with some arbitrary zero phase epoch defined in equation 2.20.

$$\phi_i(P) = mod \left[\frac{t_i - t_0}{P} \right] \tag{2.20}$$

where $\phi_i(P)$ is the phase value of the observation i as a function of the candidate period P, t_i is the measurement time of the observation i, t_0 is an arbitrarily chosen start time (currently defined such as to make the peak magnitude occur at a phase of 0.25), P is a candidate period for the light curve and the modulus (mod) operation retains just the decimal component of the calculated value (Lafler and Kinman, 1965; Dworetsky, 1983; Schwarzenberg-Czerny, 1989, 1996; Larsson, 1996; Clarke, 2002; Domínguez et al., 2013). Epoch Folding can be used as the basis of a family of period estimation methods. By establishing a set of trial periods P_t, a set of light curve data points with time instants t_i and observations x_i can be transformed into a set of phases and magnitudes $[\phi_i, x_i]$ using equation 2.20. Finally, the magnitudes are re-ordered by ascending phase from 0 to 1. These trial periods can be formulated from a grid search or alternatively, epoch

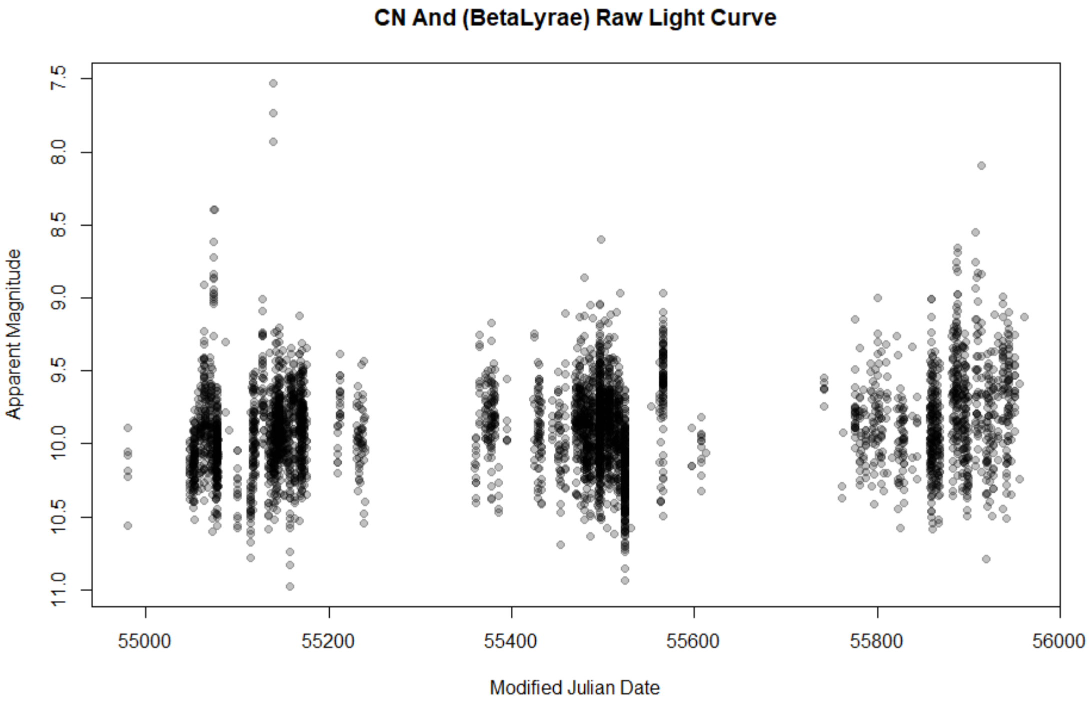


FIGURE 2.2: The raw SkycamT light curve of the eclipsing binary variable CN Andromedae over the three year 2009–2012 first observing season.

folding can be used as a follow-up to a previous periodogram and the trial periods are determined from strong candidate frequencies from those tests. Alone, this is sufficient to allow a human to visually inspect the folded light curves (the magnitudes plotted as a function of phase). For example, consider the plot in figure 2.2. This is the raw light curve of the star CN Andromedae, a β Lyrae type eclipsing binary. This object has an underlying period of 0.463 days (Van Hamme et al., 2001) and when folded around this period using equation 2.20 it produces the characteristic eclipsing binary shape in figure 2.3(a). The data points are clearly aligned into the expected shape but if folded at an arbitrary incorrect period such as 5.23 days, it produces an unaligned spread of data points devoid of structure as shown in figure 2.3(b). This analysis of the plotted folded light curve is useful, but it is insufficient for an automated period estimation method (except if some form of computer vision method is used as described in chapter 6). Therefore some form of statistical metric is required to extract some representation of the folded data points and return a value that reflects how well or poorly the data fit a trial period. In the following subsections, a set of the most popular epoch folding statistical metrics are presented and discussed as they exhibit a number of potent benefits such as often being completely independent of the light curve shape, a property that the Fourier derived methods (such as the LSP) do not share.

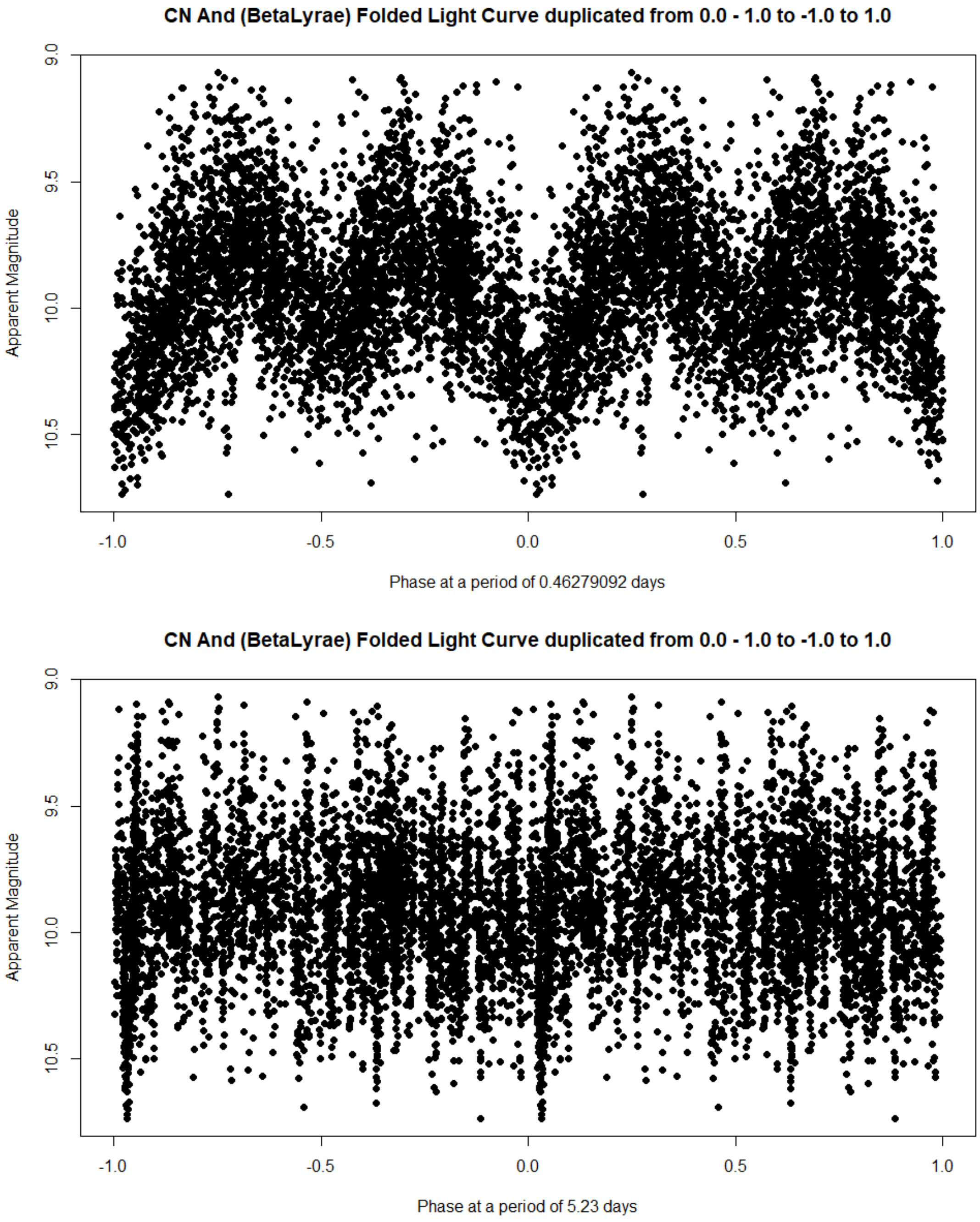


Figure 2.3: Plots of the folded SkycamT light curve of the eclipsing binary variable CN Andromedae at the true period of 0.463 days (Van Hamme et al., 2001) (a) and an incorrect period of 5.23 days (b). Any structure in the 5.23 day folded light curve are primarily a result of the sampling cadence of the SkycamT instrument on this variable.

2.3.1 String Length

String Length methods utilise a simple metric to represent the alignment of a phase folded light curve of trial period P_t (Lafler and Kinman, 1965; Dworetsky, 1983). Their name is derived from the idea that a string can be laid on top of the data points such that each phase-ordered data point $[\phi_i(P_t), x_i]$ is connected to the previous data point $[\phi_{i-1}(P_t), x_{i-1}]$ and the next data point $[\phi_{i+1}(P_t), x_{i+1}]$ with straight pieces of string. Finally, the last data point $[\phi_N(P_t), x_N]$, where N is the number of data points is connected to the first data point $[\phi_1(P_t), x_1]$ across the overlapping phase boundary. The final length of the string provides a Euclidean distance between each data point across the whole phase-space. For a set of trial periods, this produces a statistic which, when minimised, identifies a trial period which produces a well-aligned folded light curve. This string length is defined in equation 2.21 through the use of trigonometric relations for Euclidean distances (Dworetsky, 1983).

$$d(P_t) = \sqrt{(\phi_1 - \phi_N)^2 + (x_1 - x_N)^2} + \sum_{k=2}^{N} \sqrt{(\phi_k - \phi_{k-1})^2 + (x_k - x_{k-1})^2} \tag{2.21}$$

where $d(P_t)$ is the string length distance as a function of trial period P_t.

The metric in 2.21 has the disadvantage of being a function of the number of data points in a light curve. Therefore a normalisation factor was proposed to both eliminate this dependence and limit the maximum value of the function to $d(P_t) \approx 1$ at the limit of maximally unaligned phased data (Clarke, 2002). This normalised definition is shown in equation 2.22 and is named the String Length Lafler Kinman (SLLK) statistic (Lafler and Kinman, 1965; Clarke, 2002).

$$T(P_t) = \frac{N-1}{2N} \times \frac{\sum\limits_{i=1}^{N} (x_{i+1} - x_i)^2}{\sum\limits_{i=1}^{N} (x_i - \bar{x})^2} \tag{2.22}$$

Where x_i is the magnitude of data point i folded and sorted by trial period P_t, N is the number of data points in the time-series and $T(P_t)$ is the String Length Lafler Kinman (SLLK) statistic as a function of P_t (Clarke, 2002). The minimisation of $T(P_t)$ by a trial period $P_t = P_c$ identifies a candidate period P_c which produces well-aligned epoch folded data independent of the number of light curve data points. As the Fourier methods produce peaks in their associated statistics, the final operation, $\theta_{SLLK}(P_t) = 1 - T(P_t)$, re-orders the period estimation so when no signal is present $\theta_{SLLK}(P_t) = 0$ with aligned data peaking towards $\theta_{SLLK}(P_t) = 1$.

This metric struggles when the trial period P_t is close to the total timespan of the light curve $t_{\mathrm{max}} - t_{\mathrm{min}}$. At this limit, the data points no longer sample the phase space sufficiently and large sections of the signal are missing. The string length method can calculate a straight-line Euclidean distance over a large phase-range with no data points. As the minimum distance this string length can be is a straight line, it results in a minimisation of the string length statistic, not as a result of data points aligned on a signal, but due to a lack of data.

An extension to this idea of aligning data within phase space is a recently proposed periodogram based on the Blum-Kiefer-Rosenblatt (BKR) statistical independence test (Zucker, 2016). Instead of utilizing the alignment of the data such as in the string-length methods, a rank correlation test is performed on the phase-folded data within each candidate period phase space. As the phase folded light curve aligns at a strong period, the correlation between the magnitude and phase of each data point rises. The peak of this correlation is a good candidate frequency and can be determined from time-series of limited size (Zucker, 2016).

2.3.2 Analysis of Variance

Analysis of Variance (AoV) is a period estimation method based on a family of statistical tests (Schwarzenberg-Czerny, 1989). It requires the partitioning of the phase space into r equally spaced bins. For a trial period P_t, the time-series data is epoch folded and placed into the phase space. The variance of the data in each of the distinct equally spaced phase bins is determined as shown in equation 2.23. The variance is also computed between each of the phase bins using the equation 2.24.

$$s_1^2(P_t) = \frac{1}{r-1} \sum_{i=1}^{r} n_i (\bar{x_i} - \bar{x})^2 \tag{2.23}$$

$$s_2^2(P_t) = \frac{1}{N-r} \sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \bar{x_i})^2 \tag{2.24}$$

where $\bar{x}$ is the mean of the whole time-series, $\bar{x_i}$ is the mean of the data in phase bin i, n_i is the number of data points present within bin i and N is the total number of data points in the time-series. The final statistic is just equation 2.23 over equation 2.24 as shown in equation 2.25.

$$\theta_{AoV}(P_t) = \frac{s_1^2(P_t)}{s_2^2(P_t)} \tag{2.25}$$

θ_{AoV} is described as a Fisher-Snedecor statistic and can be tested with the F statistic with parameters N and r (Schwarzenberg-Czerny, 1989). This statistical test provides a value at which the null hypothesis no longer holds for the epoch folded time-series as one or more of the phase bin means differs substantially from the overall mean from every phase bin. In the event that the light curve has been folded at an incorrect period, the data points are randomly distributed within the phase space and every bin should have similar mean values.

This method is well-suited to the determination of non-sinusoidal data as it assumes no particular shape for a folded signal. Like the other period estimation methods, it is usually tested by constructing a frequency grid of trial periods across the desired period range. As the F statistic is a significant part of the test, it is trivial to determine confidence limits on the strength of any candidate period. The primary disadvantage of the Analysis of Variance periodogram is the required hyperparameters which define the size, location and number of phase bins required by the method. This therefore necessitates some form of data-driven validation analysis to determine optimal bins for a set of light curve data. Phase Dispersion Minimisation (PDM) is another period estimation method using the variance in phase bins to evaluate trial periods (Stellingwerf, 1978). This statistic is not a Fisher-Snedecor statistic and therefore is not evaluated with the F statistic and should perform similarly to AoV (Schwarzenberg-Czerny, 1989). The Plavchan algorithm is a modification of phase dispersion minimization (Stellingwerf, 1978) that discards the original requirement for phase binning (Plavchan et al., 2008). This is accomplished by running a boxcar smoothing function across the light curves and compares the smoothed light curve to the original.

2.3.3 Entropy Minimisation

Information Theory has developed the concept of information entropy which describes the amount of information held in a dataset of an underlying signal. When a time-series has been epoch folded using a trial period P_t, an aligned or ordered phase space data has a lower entropy than unaligned or disordered data (Graham et al., 2013a). Therefore, if a frequency grid of trial periods is produced and the entropy determined for each trial period, a period where the information entropy of the epoch folded data is minimised is a good candidate period for a periodic signal. Information entropy or alternatively Shannon's entropy is defined in equation 2.26.

$$H_s(P_t) = -\sum_{i=1}^{N} p(x_i)\log(p(x_i)) \tag{2.26}$$

where H_s is Shannon's entropy, $p(x_i)$ is the probability that the data point x_i has its associated univariate value given an underlying signal and N is the number of data points in the time-series. The probability $p(x_i)$ can be estimated from the epoch folded data through the computation of the ‘density’ of data points within r equally spaced phase bins upon the normalisation of the folded time-series in both phase and magnitude so as $\phi \in [0,1]$ and $x_i \in [0,1]$. With this transformation, the occupation probability of the univariate data points x within phase bins a_j can be estimated by equation 2.27 where δ is the dirac delta function (1 when x_i is in the phase bin a_j and 0 otherwise).

$$\mu_p(a_j) = \frac{1}{N} \int_{a_j} \sum_{i=1}^{N} \delta\left[(\phi, x) - (\phi_i, x_i)\right] d\phi dx \tag{2.27}$$

where $\mu_p(a_j)$ is the occupation probability. As this is a direct estimation of $p(x_i)$ across phase bins, this can be substituted into equation 2.26 as ‘phase bin data points’ producing the computable Shannon’s entropy shown in equation 2.28.

$$H_p(P_t, a) = -\sum_{j=1}^{r} \mu_p(a_j) \log(\mu_p(a_j)) \tag{2.28}$$

This computable definition of Shannon’s entropy has introduced a hyperparameter in the form of the number of phase bins used in the estimation of the occupation probability. The resulting periodogram of computing this statistic across a frequency grid will be influenced by this therefore it must be tuned and validated independent of a period estimation task. The method also struggles with aliased signals and those frequencies have similar information content as the true signal period. Spurious periods are a significant issue as, when epoch folded, they produce a phase space where all data points occupy a small phase range. As this alignment of data points is superior to any true periodic signal, spurious periods produce a much larger response from this method.

A solution to the spurious period problem was proposed in the form of Conditional Entropy minimisation (Graham et al., 2013a). This method introduces a second tuned hyperparameter in the form of magnitude bins. The Conditional Entropy minimisation statistic places an additional requirement in the form of a good alignment in the magnitude bins as well as the phase bins. Data epoch folded at spurious periods, whilst well aligned in the phase bins, are poorly aligned in the magnitude bins, whereas for true astrophysical periods and their aliases, the epoch folded data is aligned in both the magnitude and phase bins.

Through the selection of r phase bins and m magnitude bins, the new occupation probability $p(\phi_i, x_i)$ of the set of $[a_j, b_k]$ magnitude-phase bins is determined. The Conditional Entropy statistic is then defined by equation 2.29.

$$H_s(P_t) = -\sum_{i,j} p(\phi_i, x_i) \log\left(\frac{p(\phi_i)}{p(\phi_i, x_i)}\right) \tag{2.29}$$

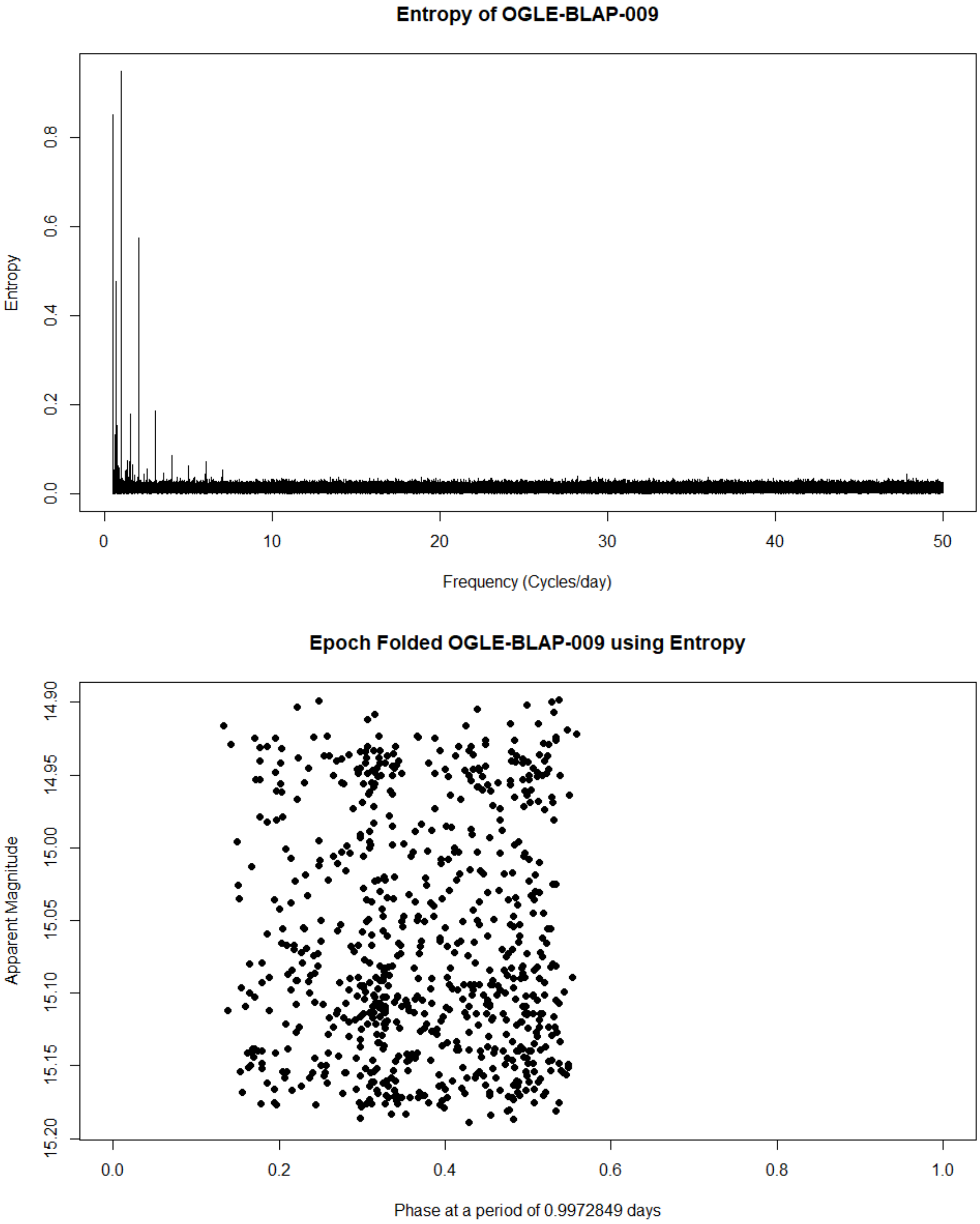


Figure 2.4: Plot (a) demonstrates the Entropy periodogram for the variable star OGLE-BLAP-009. The peak is located at 0.997 days and is a sampling period caused by the sidereal day. Plot (b) shows the epoch folded OGLE light curve of OGLE-BLAP-009 at this period showing the data points concentrated over a small phase range.

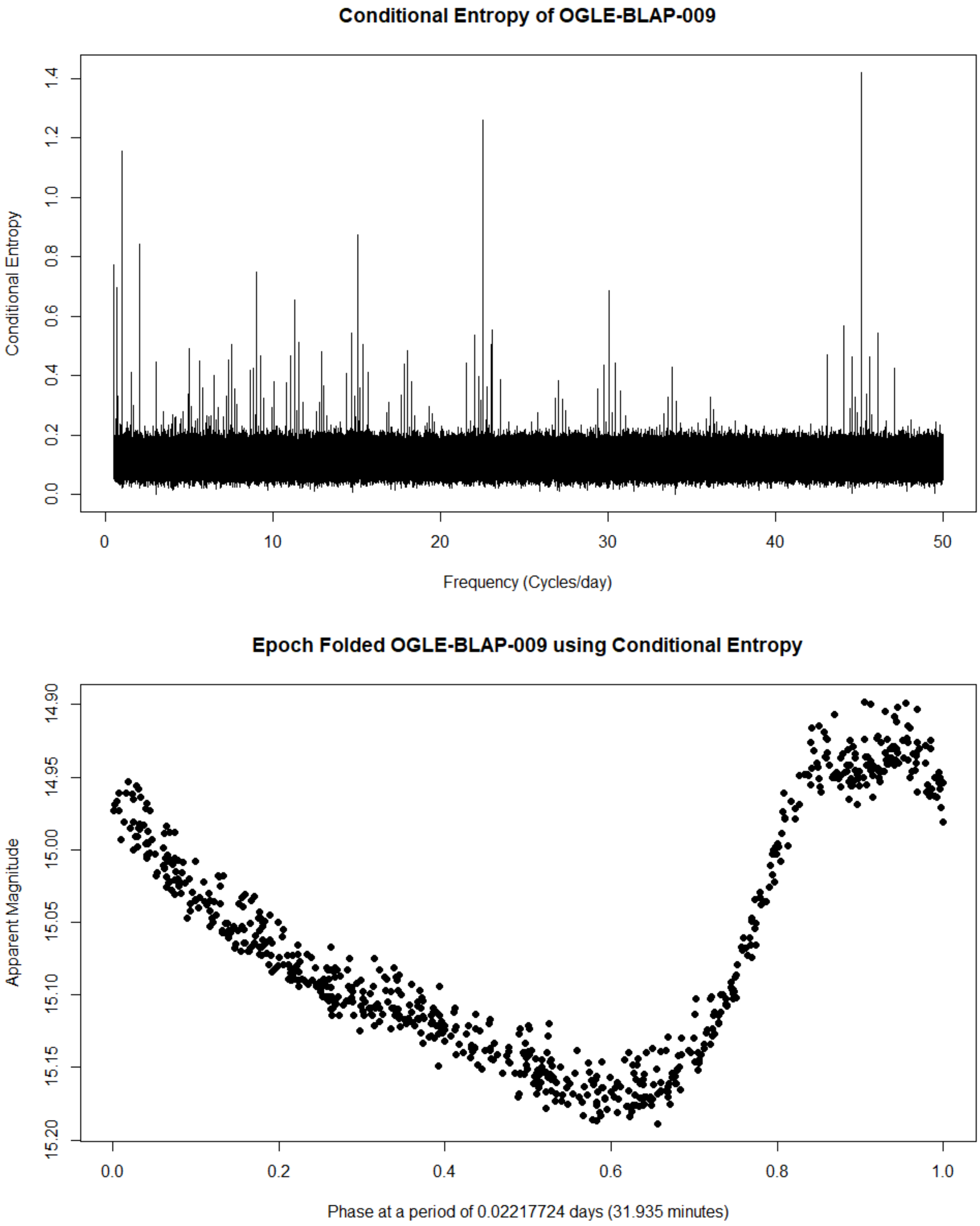


FIGURE 2.5: Plot (a) demonstrates the Conditional Entropy periodogram for the variable star OGLE-BLAP-009. The peak is located at 0.022 days and is the pulsating astrophysical signal from this star (Pietrukowicz et al., 2017). Plot (b) shows the epoch folded OGLE light curve of OGLE-BLAP-009 at this period showing the smooth pulsating modes of this star.

As entropy and conditional entropy period estimation rely on a minimisation of the target statistic we apply a transformation of $\theta_{ent}(P_t) = 1 - H_s(P_t)$ similar to the String Length method. This allows the resulting Entropy and Conditional Entropy periodograms to be analysed for peaks associated with maximised values. Figure 2.4 demonstrates the result of an Entropy minimisation period estimation on the pulsating star OGLE-BLAP-009 over a frequency grid of 0.5 cycles/day to 50 cycles/day using 20 phase bins. The estimated period of 0.997 days is clearly a spurious period produced by the alignment of the data points around a sampling period (specifically the sidereal day sampling period). Compare this figure with the one in figure 2.5 using 20×20 phase-magnitude bins where the estimated period of the variable star is 0.022 days, or 31.94 minutes which agrees closely with the period estimated by other methods such as the Generalised Lomb-Scargle periodogram (Pietrukowicz et al., 2017).

2.4 Sampling Frequency

Period estimation methods operate over a frequency range with a finite set of candidate frequencies separated by intervals. As the objects in the database can have low and high period variations, this interval is set as constant to produce a uniform sample across the full frequency range. The lowest frequency is the longest period that can be expected to be detected by the periodogram. It is defined as the reciprocal of the difference between the maximum and minimum modified Julian Date of the objects observations named the total observation time t_{tot}. The maximum frequency is an interesting discussion area and is related to the minimum periods that can be found from an objects data. Hypothetically scanning down to the minimum possible periods for variable stars is recommended. However, for some pulsating white dwarf stars, this can be as low as 1-2 minutes. The Lomb-Scargle Periodogram is not limited to a Nyquist frequency (VanderPlas, 2017). However, at very high frequencies many noisy peaks can be generated as any data can be fitted well to a model with such high frequencies. In previous methods, a Pseudo-Nyquist frequency was proposed for the determining of a maximum frequency for unevenly sampled data which approximates the Nyquist frequency by taking the mean of the individual time intervals between the observations of an object (Shannon, 1949; Debosscher et al., 2007; Percy, 2008; VanderPlas, 2017). This equation is shown in equation 2.30.

$$f_{sN} = \frac{1}{2}\left\langle \frac{1}{\Delta T} \right\rangle \tag{2.30}$$

where f_{sN} is the Pseudo-Nyquist frequency and ΔT is the vector of time intervals between observations of an object. In the Richards et al. methodology, the frequencies

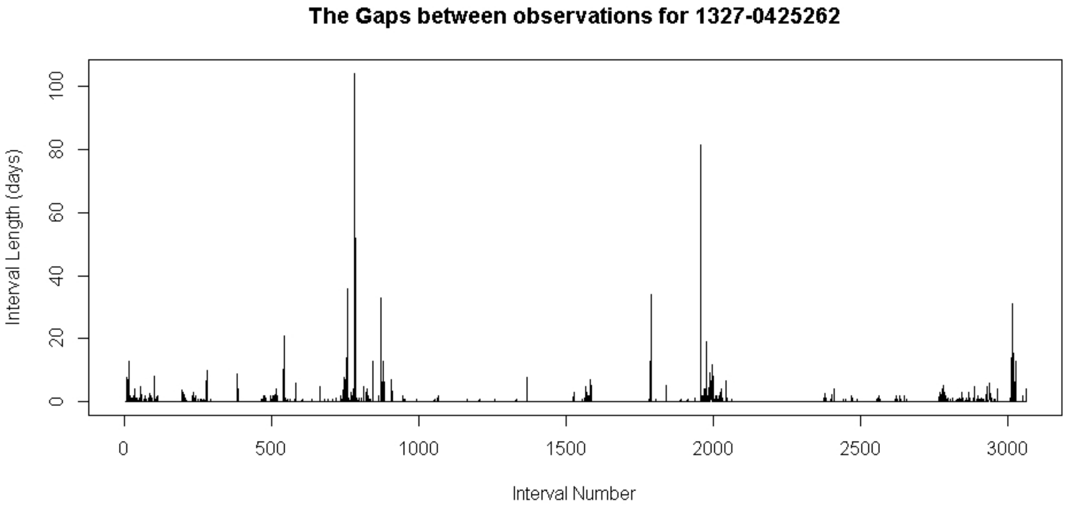


FIGURE 2.6: Plot demonstrating the gaps between data points for the variable star RR Lyrae. The plot contains two dominant peaks due to seasonal sampling but many of the smaller peaks are due to the telescope scheduler not returning to the area of the sky with RR Lyrae.

could rise beyond this value with the periodogram normalized power subjected to a soft penalty term that weakens the peaks beyond the Pseudo-Nyquist frequency based on the relative difference between this frequency and the candidate frequency (Richards et al., 2011b).

This frequency is determined by the mean intervals between observations in an uneven time-series. Gaps in observations are not considered uneven observations and instead are just considered times with a lack of observations and do not contribute to the calculation of the Pseudo-Nyquist frequency. Despite this distinction, a globally accepted definition of a 'gap' and an 'uneven sample' was not identified (VanderPlas, 2017). Theoretically, the time intervals between observations could all be considered gaps. This is not really possible as the start and stop times of Skycam observations are unlikely to be an integer number of minutes (the interval between exposures). Ignoring this, the sampling rate could be the Nyquist frequency of the evenly sampled exposure intervals which is half of the reciprocal of a minute, 720 cycles per day, equivalent to a period of 2 minutes. For the example with no gaps, every interval is included in the calculation of the Pseudo-Nyquist frequency. This results in a frequency that may result in the loss of low period signals from the periodogram. Figure 2.6 demonstrates the gaps between data points for the variable star RR Lyrae. The seasonal observational gaps of this star produce two clear peaks in the plot yet there are numerous smaller gaps due to the scheduled observations of the Liverpool Telescope not resampling the patch of sky containing this star.

Evidence suggests that neither of these options is ideal. For example, the dominant period of the star RR Lyrae has been determined as 0.5668 days (Kolenberg et al., 2010). From the SkycamT data for this star, if none of the time intervals are considered gaps, the Pseudo-Nyquist frequency results in a minimum detectable period of approximately 0.7 days. No candidate frequencies would be evaluated by the periodogram near the 0.5668 day period despite it being a strong peak. By manually allowing for larger frequency range, this period is detected as the dominant period. Conversely, if the 720 cycles per day frequency is used as the Nyquist-frequency as described above, the frequency spectrum is dense enough to result in extreme processing load from the periodogram. Ultimately, the solution is to choose a minimum period in the search range as defined by the objects detectable or of interest to a survey. Alternatively, the use of a method which does not rely on a frequency grid to operate renders this distinction unimportant as long as such method does have the required resolution.

2.5 Spurious period detection

In order to remove sampling periodicities from a light curve, the first step is the identification of the spurious periods. These periods are caused by both local periodicities due to the location of the measuring device at a ground based observatory as well as interactions between the sampling time instants and these local periodicities (Huijse et al., 2012; Protopapas et al., 2015; McWhirter et al., 2016). The strongest spurious periods for ground based surveys tend to be related to the sidereal day and a yearly seasonal period. The sidereal day is generated as a result of no observations being gathered during day time hours due to scattered sunlight as well as observations being taken when objects have returned to favourable positions in the sky each day. The seasonal periods are usually strongly related to the solar year as, depending on the location of an object in the sky and the location of the telescope on the Earth, the object may only be visible for certain months during the year. The sampling rate can also produce a spurious period in the data as the observations are always an integer multiple of this sampling rate. In the case of highly unevenly sampled data such as the Skycam cadence, these sampling periods tend to be weaker but it can be a major concern to space-based observations (Neff et al., 2014; McWhirter et al., 2016).

Through the use of the classical, Schuster periodogram or the Lomb-Scargle periodogram it is possible to detect the spurious periods for a given light curve (Protopapas et al., 2015; VanderPlas, 2017). if a classical Periodogram is utilised on data featuring the time instants of a given objects light curve, but with a constant signal with no magnitude data, the peaks received cannot be caused by variations in the object magnitude due

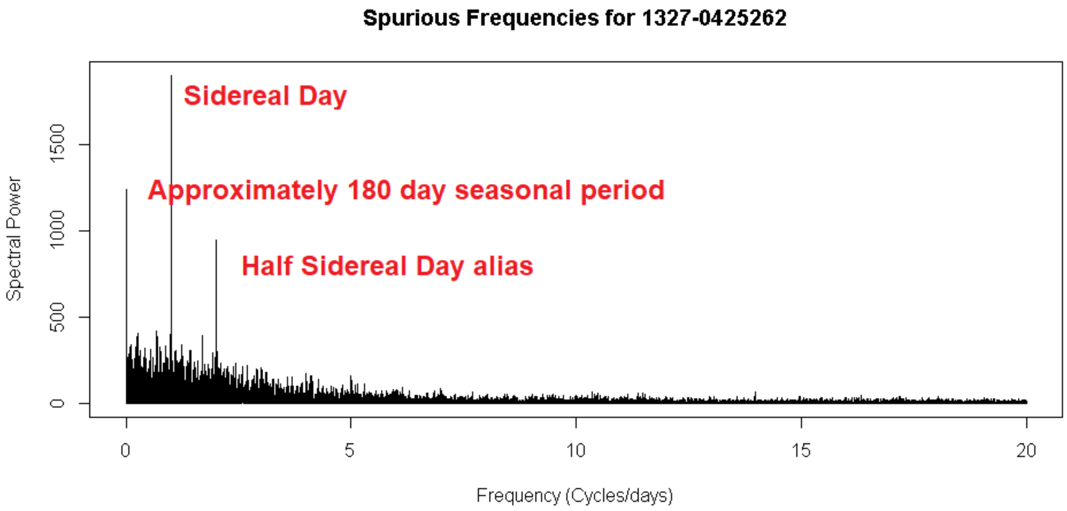


Figure 2.7: The periodogram of the time instants of the SkycamT light curve of the star RR Lyrae. The three dominant peaks are related to the sidereal day, the half sidereal day alias and an approximately 180 day seasonal periodicity likely due to the sky coordinates of the star.

to an astrophysical signal or noise and therefore must be caused by sampling periodicity. The Fourier Periodogram is defined as the square modulus of the Discrete Fourier Transform of the autocorrelation function of a given set of observations or samples, or alternatively, the square modulus of a signal's Power Spectrum Density (PSD). In this method, we do not want to provide any magnitude data as this could introduce actual signal periodicity which we do not wish to filter. Therefore, a constant signal is provided and the autocorrelation of a constant is unity. Therefore the autocorrelation component can be removed from the equation giving equation 2.31.

$$P_k = \frac{1}{N}\left|\sum_{i=1}^{N}\exp(j2\pi f_k t_i)\right|^2 \tag{2.31}$$

Where P_k is the spectral power for a candidate frequency k, f_k is the a candidate frequency from a set of candidate frequencies and t_i is the time instant value of an observation i. This produces a vector of spectral power values for each frequency k. Figure 2.7 shows the Periodogram from this spurious period removal method applied to the SkycamT light curve of the star RR Lyrae.

2.6 The bands method

The computational load from investigating a large frequency spectrum is prohibitive to realistic analysis of large astronomical light curve datasets. It would be advantageous to have access to an efficient method of reliably reducing the size of this frequency spectrum to a subset of useful candidates. This can be accomplished by using a method very similar to the one used above to remove spurious periods (Protopapas et al., 2015). By splitting the light curves into unevenly spaced horizontal bands based on percentiles, a tightly controlled amount of magnitude information can be provided to the period estimation methods generating new interesting signal related peaks whilst filtering out the sampling periodicities established from the spurious period removal method. The bands method relies on data points of similar magnitude being separated by integer multiples of the period in a periodic light curve. Therefore, if specific magnitude bands are selected, there should be a strong periodic response where these bands differ from the nominal magnitude bands of the light curve.

This method is unable to guarantee that the true astrophysical period for every object will be present in the resulting candidate frequency subset but it can achieve reasonably high accuracy. Additionally, even if the astrophysical period is present in the candidate set, this method alone is highly unlikely to place this value at the top of the predicted frequencies. Therefore, additional Periodograms are still required, but they can be more processing intensive on the smaller candidate set yet operate at similar processing load than inferior methods applied to the whole frequency spectrum (Protopapas et al., 2015). The execution of this method commences with the determination of the derivatives of the light curve. This is simply computed as the gradients between successive time-ordered data points as shown in equation 2.32.

$$d_i = \frac{x_{i+1} - x_i}{t_{i+1} - t_i} \tag{2.32}$$

The raw light curve is then split into ten unevenly spaced horizontal bands based on magnitude percentile. Figure 2.8 demonstrates the split of the raw light curve of the variable RR Lyrae into ten coloured bands. The time instants associated with a magnitude value within each of these ten bands are then binned into ten magnitude percentile bins. For each of the ten bands, equation 2.33 is computed for each frequency in the frequency spectrum using the time instants present in the associated band.

$$D_j = \sum_{i \in B_j} |d_i|, \quad \text{where } j = 1, \ldots, 10 \tag{2.33}$$

This summation describes the variability of each band with the nominal portion of the light curve having a lower value of D_j than the bands where substantial variation occurs. Therefore, only the top N_b bands are retained with the selected value of N_b providing a trade-off between the rate of correct period estimations and the number of trial frequencies. For each of the retained bands the spectral window function shown

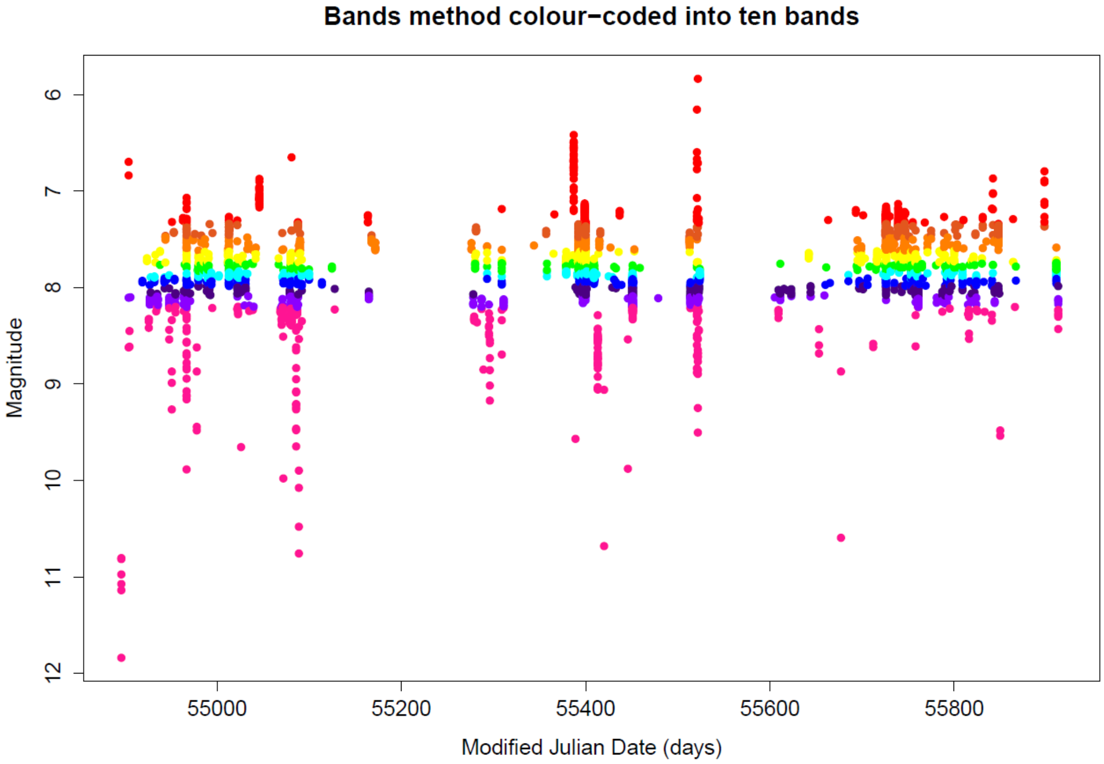


FIGURE 2.8: The raw SkycamT light curve of the star RR Lyrae from March 2009 to March 2012 with ten percentile magnitude bands each colour coded from red, through yellow, green and blue to pink.

in equation 2.34, similar to the function used to detect spurious periods, is computed across a frequency spectrum f with individual frequencies f_m.

$$P_m = \left| \sum_{k \in B_j} \exp\left(j 2\pi f_m t_k\right) \right|^2 \tag{2.34}$$

where P_m is the spectral power, j is an imaginary number, f_m is a candidate frequency from the frequency spectrum set and t_k is the time instant of an observation k whose magnitude value is present within band B_j where $j \in \mathbb{I} \in [0, 10]$. For each band, the top N_t frequencies are retained and like the N_b parameter, it determines the accuracy of the bands method against the volume of candidate frequencies.

This calculation is being performed on a constant signal like in the spurious period example, but by providing time instants within a specific band, some magnitude information is applied to the spectral window function. In a periodic light curve, the top and bottom percentile bands are likely to have groups of equally spaced time instants due to this periodic signal (Protopapas et al., 2015). This then translates into an associated peak for each band. The frequency spectrum for each band is then sorted placing by spectral power in decreasing order. Two arguments have then been supplied for this method

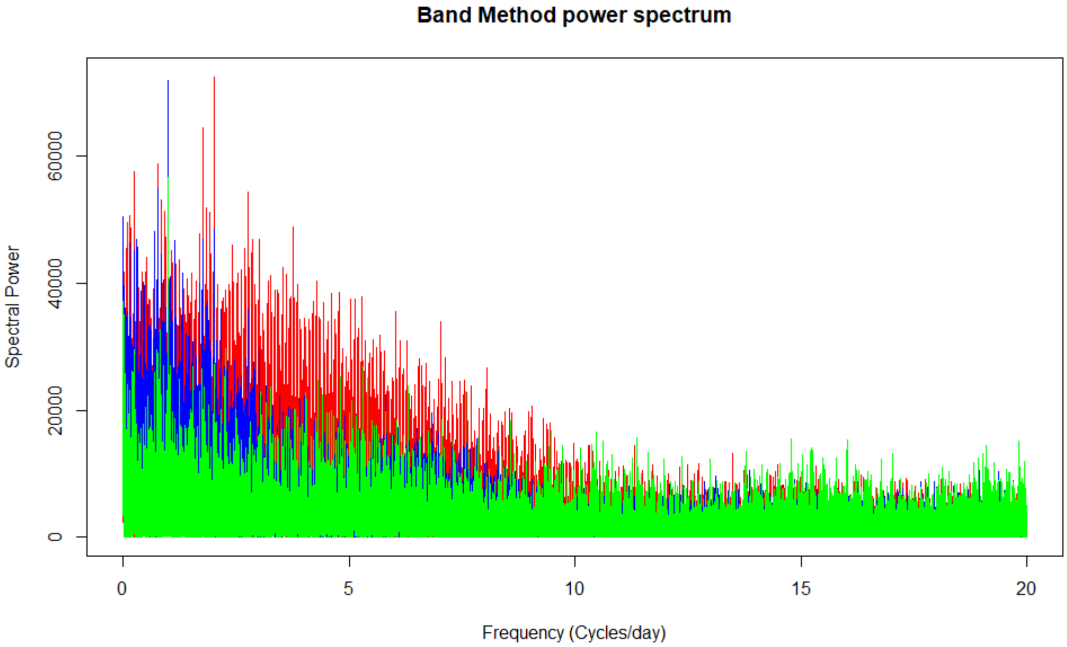


FIGURE 2.9: The power spectra across the three strongest of ten percentile bands for the star RR Lyrae from the SkycamT light curve data. The red peaks are the most variable band, followed by the blue peaks and lastly the green peaks.

named N_t and N_b and they define the number of frequencies and bands to retain. The top N_t frequencies from each band are then retained with all others rejected and the top N_b bands are then retained with the others rejected. This produces a set of candidate frequencies of size $N_t N_b$ although there are likely to be many duplicates identified within multiple bands which can be filtered out leaving this as the maximum possible set of candidate frequencies.

As this method will still generate the spurious periods identified in the previous spurious period removal method, all trial periods that match with a spurious period are removed using a Gaussian filter. It is interesting to note that for known periodic objects in the Skycam data, the number of trials tended to be greater than the number of trials for known non-periodic objects. This is not surprising as the non-periodic objects would contain more sampling periodicity duplicates as there are no astrophysical periodicities to compete for the limited number of candidate frequencies. Figure 2.9 demonstrates how the power spectrum changes across bands for the star RR Lyrae based on magnitude variations across the top three bands. The red peaks are from the calculation of the magnitude band with the highest variability, the blue peaks are from the band with the second highest variability and the green peaks are from the band with the third highest variability. This periodogram contains the primary spurious period peaks from figure 2.7 with additional peaks containing possible periodic signal components.

2.7 Correlated noise fitting routine

As previously discussed, as well as an astrophysical signal (which can be constant for a non-variable star) and the sampling periodicities, noise results in a substantial contribution to the resulting peaks in a Periodogram (VanderPlas, 2017). There is a continuum of small peaks caused primarily by noise and because the time series is not infinitely long. This means that a Periodogram on a finite time series with a zero noise component will still have a slight continuum due to patterns in the shape and length of the sampled time period. In fact, another common sampling periodicity in the Skycam object data is simply the total length of time that an object has been observed.

Consider the SLLK period estimation shown in figure 2.10. For noise with a purely white noise component, the power of the noise continuum is independent of frequency. However, in this SLLK period estimation of RR Lyrae, the strength of the noise is clearly increasing at lower frequencies. This is due to the correlated or red noise component of the noise (Vaughan, 2011). The Lomb-Scargle Periodogram does not suffer as large a performance loss as it models a signal as a sinusoid which the red noise does not fit well. Unfortunately, this means that Periodograms that are more inclined to model a signal in a non-sinusoidal form, or epoch folding methods with no defined form, this red noise component becomes significantly worse. The String Length Lafler Kinman period estimation almost always returns a peak value close to the maximum length of the observing time being the shortest frequency tested and therefore the frequency that receives the largest contribution from the red noise component.

The candidate frequency chosen from the period estimation is simply the frequency with the largest peak which has not been filtered out as a spurious period. As the red noise causes the low frequency peaks to amplify often resulting in an incorrect period, a method is needed to subtract away this red noise component flattening the noise continuum. Unfortunately, this is not easily accomplished when utilising the bands method due to the very rapid rejection of frequencies for performance purposes. In order to perfectly model the continuum, the entire frequency spectrum would need to be calculated using the period estimation algorithm defeating the purpose of the bands method. Fortunately, it is not necessary to model the red noise to an extreme accuracy, only to the point where enough of the red noise is subtracted from the period estimation to allow the peaks caused by the signal at any frequency to become dominant. Therefore, we develop a simple, but novel, method to collect a number of red noise fitting frequencies, determined to be in the noisy continuum by examining the weak end of the bands method candidate frequencies, and fit the appropriate red noise model using linear regression to these frequencies. The resulting red noise model customised for the specific period estimation method in question is then subtracted from the initial period

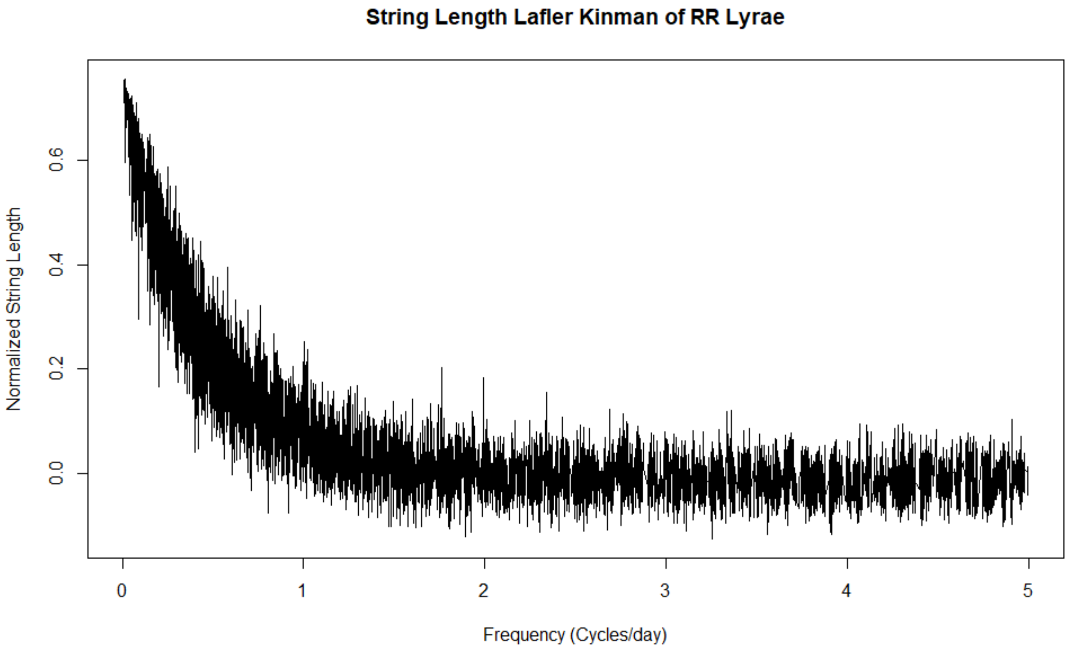


FIGURE 2.10: The SLLK period estimation of the SkycamT light curve of the variable star RR Lyrae. Without correcting for the exponential noise continuum caused by red noise, the peak period is primarily influenced by noise rather than signal. In this case the estimated period is 66.25 days which completely disagrees with the known astrophysical signal (Kolenberg et al., 2010).

estimation statistic depleting a significant amount of the red noise contribution to render it undisruptive.

Red noise is a correlated noise with an exponential decrease in power with frequency (Vaughan, 2011). The model followed by the red noise in the Periodograms is shown in equation 2.35.

$$P_{RN}(f) = k\exp(-Af) + C \tag{2.35}$$

Where $P_{RN}(f)$ is the associated Periodogram statistic contribution sourced from red noise for a given frequency f and k, A and C are coefficients based on the specific period estimation generated from a specific light curve. The A determines the rate of exponential decay of the red noise strength across the frequency spectrum. The variable k is the red noise component strength at a frequency of zero, the minimum possible frequency and C determines the strength of the flat white noise continuum.

Linear Regression can be used to model the values of the variables k and C based on candidate frequency power spectrum values extracted from the period estimation. However, A cannot be modelled this way as it is not a coefficient and instead involves a change to the models exponential shape. Therefore it must be approximated by testing a spectrum of possible A values and utilising the value associated with the best fitting model. A spectrum of possible A values is defined over a range from 0 to 10 by 0.01 per step resulting in 1001 candidates for the A value.

This method can accurately fit the appropriate model to the period estimator, but a selection of candidate frequencies which generate the power spectrum values must be selected. These frequencies must have a periodicity contribution sourced almost entirely from red noise spread sufficiently logarithmically across the frequency spectrum as the primary exponential contribution is detected at low frequencies. Therefore, first the power spectrum of a light curve for a preselected red-noise prone algorithm is generated. These algorithms include the Variance Ratio Periodogram, the String Length Lafler Kinman method and the Correntropy Kernelised Periodogram. This power spectrum is produced after selecting candidate frequencies from the initial frequency spectrum using the bands method. Then a logarithmic set of bins must be generated in which to generate candidate frequencies determined to be red noise dominant. There is an important compromise here, too many candidates and the red noise fit incurs a significant performance penalty. Whereas, with too few candidates, any frequencies chosen that do have a significant non-red noise component can have a major effect on the fitted model and therefore distort the red noise elimination.

Fits on a set of SkycamT light curves has shown that 50 red noise candidate frequencies appears to be sufficient for a good fit whilst maintaining pipeline performance. These 50 frequencies are selected as this number is a good trade-off between the computation of a good fit and the computation time required to compute the additional frequencies and fit the model. These 50 frequencies are determined by first computing 51 logarithmically separated frequencies between the minimum and maximum frequencies within the initial frequency spectrum. This produces 50 binned regions in which one red noise dominated frequency is chosen from each. A random number generator produces a randomly selected frequency from each of the 50 bins resulting in 50 candidate frequencies. In order to select good frequencies that have no signal or sampling contributions, any random frequency found to match with a frequency found using the band method, be it a spurious period or otherwise, is regenerated using the random number generator.

Upon the selection of 50 satisfactory trial frequencies, the same period estimation algorithm is utilised to calculate the associated period estimation statistic at each of these frequencies. Due to the trial frequency selection process, these are assumed to be dominated by the noise contribution. These values are then fit using 1001 linear regression operations for each value of A in the sampling grid. This returns the best fitting values of k, A and C based on the value of A for the model that minimises the mean squared error of the fit and the k and C are also taken from the fitted coefficients of this specific model. This model is then used to whiten the Periodogram determined prior to

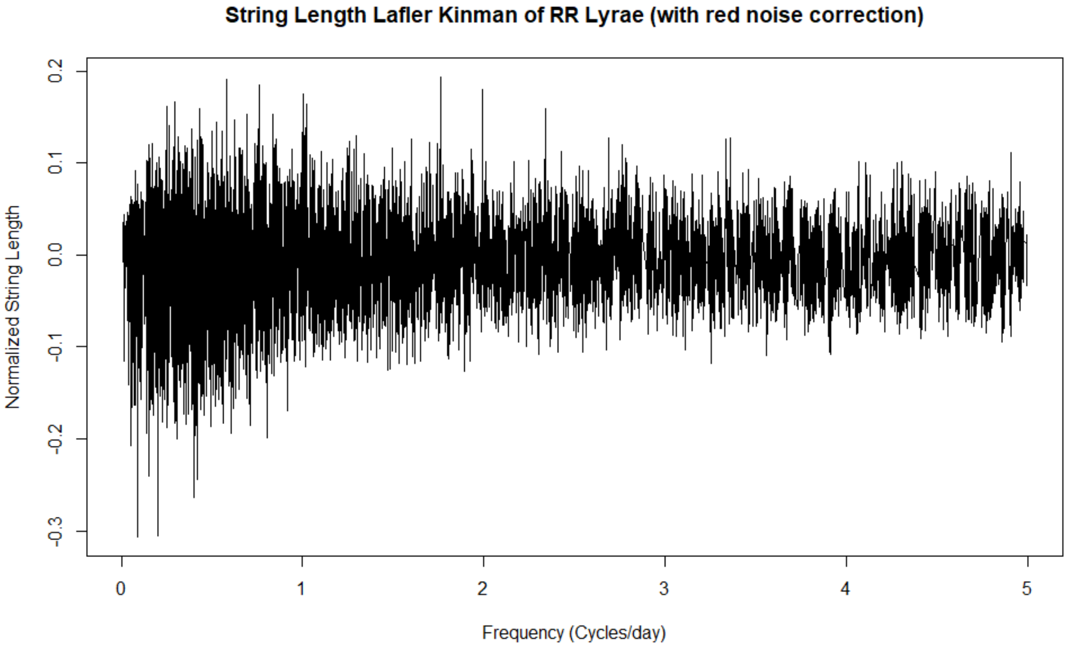


FIGURE 2.11: The SLLK period estimation of the SkycamT light curve of the variable star RR Lyrae with a continuum corrected for red noise. Now the estimated period is 0.5668 days (Kolenberg et al., 2010) which agrees with the known astrophysical signal.

noise-extraction by simply using equation 2.35 to predict the power level associated with noise for each candidate frequency and subtracting it from the original period estimators statistic for each candidate frequency. This results in a periodogram that has been mostly flattened through the removal of the exponential frequency-dependent red noise component. This technique has now allowed for the period estimation algorithms prone to red noise errors to function as intended. Applying this technique to the SLLK period estimation on RR Lyrae shown in figure 2.10 results in a fitted red noise model with coefficients $k = -0.0069$, $A = 2.16$ and $C = 0.74$. The prewhitened SLLK is shown in figure 2.11 and has a flattened noise continuum. The new highest SLLK peak occurs at 0.5668 days matching the expected astrophysical period of this variable star (Kolenberg et al., 2010).

2.8 Failure Modes

Period estimation methods are, by definition, designed to select the most likely period from an input time-series by maximising or minimising some statistical function. Despite the best efforts, it is nearly impossible to ensure that all light curves can be correctly estimated and there are a number of common failure modes that the period estimation

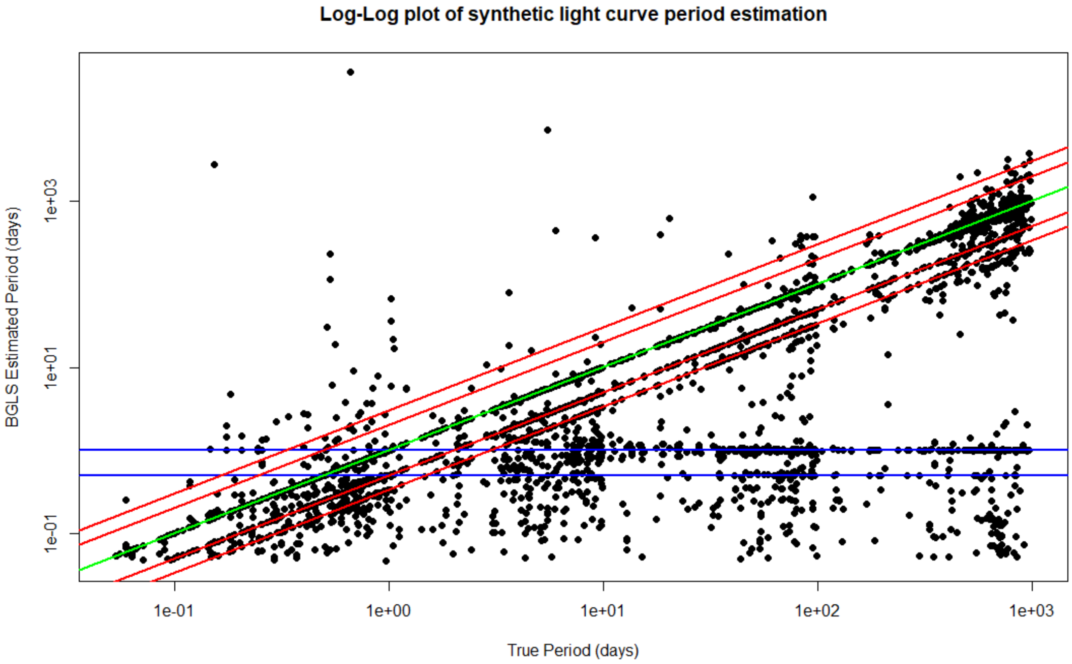


Figure 2.12: The true period against Generalised Lomb Scargle estimated period for 4000 synthetic light curves with SkycamT cadence using 4 different sinusoidal and non-sinusoidal shapes. This plot demonstrates the four most common failure modes with the red lines indicating the harmonic multiples and submultiples up to $n = 3$ and the blue lines indicating the sidereal day and half sidereal day spurious periods. The green line shows the correct period estimation region with gradient unity.

methods can produce (VanderPlas, 2017). These failure modes are often a result of directly computing some spurious period, as explained above, or some form of aliasing or pseudo-aliasing effect generated as a result of a true astrophysical signal or the interaction of this signal with a spurious period. Figure 2.12 shows the comparison of the true period against the period estimated by a Bayesian Generalised Lomb-Scargle (BGLS) periodogram for 4000 synthetic SkycamT cadence light curves of 4 different sinusoidal and non-sinusoidal shapes. This plot demonstrates the four well known failure modes, the spurious period failure mode, the aliasing failure mode, the harmonic multiple or submultiple failure mode and the symmetric failure mode.

Fortunately, many of these failure modes have a simple linear or non-linear relationship to the true period allowing any peak to be mapped to a potential true period. However, the true period and a failure mode of the true period may be hard to separate into the true period and the incorrect period. Using these functional relations, the periodogram performance on the 4000 light curves and with a 1% tolerance on period estimation error is: 35.125% of the estimated periods correctly matched the true period, 0.1% were harmonic multiples, 13.875% were harmonic submultiples, 7.175% were an alias of the one day spurious period and 2.3% were an alias of the half day harmonic of the spurious period. The remaining 42.05% of the remaining periods are either a spurious period failure mode or unknown.

The spurious period failure mode is marked by the blue lines in figure 2.12 for both the sidereal day and half sidereal day periods. This failure mode has no dependence on the period of the astrophysical signal (the only failure mode that does not) as it is a result of the spurious period peak becoming dominant in the periodogram. This failure mode is described by equation 2.36.

$$P_{obs} = (\delta f)^{-1} \tag{2.36}$$

where δf is the frequency associated with a spurious sampling periodicity. This may be correctable through the filtering of spurious periods using a classical periodogram. This failure mode often occurs when the signal-to-noise of the astrophysical signal is very weak.

Alias periods are caused by spectral leakage from the periodogram or in the case of epoch folding, reflections of the true signal caused by a spurious sampling period. As these reflections produce a signal with very similar shape to the true signal, they generate strong peaks in period estimation methods. Noisy data can make this effect worse as the weakened reflections are indistinguishable from the noisy true period signal. Equation 2.37 shows the functional relationship between an aliased period and the true astrophysical period through a spurious sampling frequency δf (VanderPlas, 2017).

$$P_{obs} = \left(\frac{1}{P_{true}} + n\delta f \right)^{-1} \tag{2.37}$$

where δf is the frequency associated with a spurious sampling periodicity, P_{true} is the true underlying astrophysical period and n is an integer. As the alias decreases in strength with higher values of n it is usually acceptable to stop searching for alias periods above $n = 4$ as they are highly unlikely to produce a dominant peak.

The red lines in figure 2.12 are associated with the harmonic failure mode. Integer multiples of a periodic signal also produce a strong peak in many period estimation methods as the signal aligns up well every m oscillations. Non-sinusoidal light curves can also result in harmonic failures in sinusoidal modelled period estimation methods such as the Fourier methods due to a superior sinusoidal model fit at these harmonics. The sampling of a time-series can result in undersampled regions which allow an integer submultiple period to fit the data as sufficiently as the true period. This is an example of the flaw discussed with Lomb-Scargle periodograms and Generalised Lomb-Scargle periodograms where the bottom half of a light curve has been undersampled due to instrument detection limitations resulting in a half-period fit producing a dominant periodogram peak. A modification of equation 2.37 introduces an additional positive integer value m which can also take reciprocal values m^{-1} for the submultiple periods. This modified version is displayed in equation 2.38 (VanderPlas, 2017).

$$P_{obs} = \left(\frac{m}{P_{true}} + n\delta f \right)^{-1} \tag{2.38}$$

The symmetric failure mode is a result of even symmetry around $f = 0$ frequency axis. Every possible candidate period P_c has a reflection $-P_c$ which produces an almost identical periodogram peak (VanderPlas, 2017). Usually this is not a problem as we would never search for a *negative* period as it does not make phyiscal sense. The harmonic and alias failure modes from equation 2.38 complicate this as they can act on a negative frequency to produce an aliased frequency in the positive frequency space. A positive frequencies alias positioned in the negative frequency space can also reflect along the $f = 0$ axis into positive frequency space. As a result, equation 2.38 is again modified to show no distinction between positive and negative frequencies by taking its absolute value shown in equation 2.39. It is also interesting to note that in the case of asymmetric light curves, the alias reflection will also reflect this asymmetry. In the case of a sawtooth Cepheid light curve with a rapid ascending branch and slow descending branch, a reflected alias will produce a clearly variable shape but with a slow ascending branch and fast descending branch (VanderPlas, 2017).

$$P_{obs} = \left| \frac{m}{P_{true}} + n\delta f \right|^{-1} \tag{2.39}$$

Remaining failures are due to noise or, when the true period is close to the time span of the light curve, a period related to the true period or a known failure mode but outside of the utilised tolerance limit. Employing techniques that investigate these well understood failure modes on candidate periods from the period estimation methods can potentially allow the correction of some of these effects leading to improved performance. It is important to note that some failure modes such as the alias periods are not distinguishable from the true period using the periodogram alone and additional techniques are required to correct for them.

2.9 Statistical Significance

Reporting uncertainties in period estimation tasks is unusual compared to many other tests. As period estimation methods operate in a frequency space, a convolution operation on the usual time-series, the widths of peaks are more dependent on the sampling rate and time span of the time-series than the signal-to-noise (S/N) or number of data points of the time-series (VanderPlas, 2017). Therefore techniques that are usually considered reasonable such as measuring the full-width-half-maximum (FWHM) of a Gaussian function at the peak frequency have no meaningful information on the confidence in a candidate frequency.

The uncertainty in a periodogram peak is instead associated with the strength, or height, of the peak. Higher S/N and/or a greater number of observations of the time-series will improve the quality of the fit associated with the period estimation method statistic resulting in a higher (if a maximising test) or lower (if a minimising test) statistical value (VanderPlas, 2017). Therefore the statistical significance of a peak is based on its height (value) relative to the heights of the frequency peaks in the nominal regions of the periodogram. It is important to determine this statistical significance as to confirm that the peak of a candidate frequency is significant and not just a noisy result in a non-variable time-series (Frescura et al., 2008). This also quantifies the strength of the primary periodic component in the time-series for use as a feature of the data which can influence the classification of the variable light curves.

2.9.1 False Alarm Probability

The Fourier based methods including the Lomb-Scargle periodogram utilise a False Alarm Probability (FAP) to describe the uncertainty in a periodogram peak. The FAP $\in [0,1]$ is the probability that a peak $PN(f)$ at frequency f satisfies the null hypothesis, in this case, that the peak was drawn from a signal with no periodic component (VanderPlas, 2017). The False Alarm Probability can be calculated from the value of $PN(f)$ using equation 2.40 (Scargle, 1982).

$$p(PN(f)) = 1 - (1 - \exp(-PN(f)))^M \tag{2.40}$$

where $p(PN(f))$ is the probability of obtaining a peak at frequency f with power $PN(f)$ and M is a variable depending on the number of data points, their cadence and the number of independent frequencies tested by the periodogram (Scargle, 1982; Press et al., 1994).

A confidence level must also be calculated from the time-series M variable value to provide a probability level required to obtain a statistical significance with a given type I error probability α. Equation 2.41 shows the calculation of this confidence level Λ.

$$\Lambda = -\log\left[1 - (1 - \alpha)^{M^{-1}}\right] \tag{2.41}$$

where Λ is the confidence level for a periodogram and α is the type I error probability such as $\alpha = 0.01$.

The determination of M has been shown to be best estimated by the number of data points N in a time-series and applies for both evenly and unevenly sampled time-series when the periodogram frequencies are greater than the Nyquist frequency (Horne and Baliunas, 1986; Press et al., 1994). Equation 2.42 shows the determination of M used in the Lomb-Scargle method used here (Ruf, 1999). This determination is dependent on the time-span of the time-series and the number of trial frequencies in the periodogram.

$$M = 2\left(\frac{n_{out}}{v}\right) \quad \text{where } n_{out} = v\left(f_{\max} - f_{\min}\right)\left(t_N - t_1\right) \tag{2.42}$$

where n_{out} is the number of periodogram frequencies from $f_{\min}$ to $f_{\max}$, v is the oversampling rate, t_N is the time instant of the N^{th} data point where N is the number of data points and t_1 is the initial time instant. The FAP is valid only in cases where the errors are normally distributed, or alternatively, in the absence of correlated noise (Frescura et al., 2008). The inclusion of correlated noise into the FAP requires substantial computational effort and therefore it remains uncorrected therefore this must be remembered when applying such a method.

2.9.2 Vuong-Closeness Test

The Vuong-Closeness test is an information theoretic comparison between two parameterised models using the Kullback-Leibler Information Criteron (KLIC) (Vuong, 1989; Baluev, 2012). For a model $\mu_k(t, \theta_k)$, the estimation of θ_k can be calculated using maximum-likelihood from equation 2.43.

$$\hat{\theta_k} = \arg \max_{\theta_k \in \Theta_k} \sum_{i=1}^{N} f_k\left(x_i | z_i, \theta_k\right) \quad \text{where } z_i = [t_i, \sigma_i] \tag{2.43}$$

where $\hat{\theta_k}$ is the estimated model parameters for model k, Θ_k is the set of model parameters for model k, x is the vector of univariate measurements, z is the set of univariate measurements and their errors and f_k is the probability density function of each measurement. The Kullback-Leibler Information Criterion (KLIC) is a measure of the relative information content between two discrete probability distributions, in this case the functions of two models with parameter sets θ_1 and θ_2 and is defined in equation 2.44 for a parameter set θ_k.

$$\text{KLIC}_k(\theta_k) = \int \log \frac{h(x|z)}{f_k(x|z, \theta_k)} h(x, z) dx dz = \mathbb{E}^0_{x,z} \log \frac{h(x|z)}{f_k(x|z, \theta_k)} \tag{2.44}$$

where $h(x, z)$ represents the true, unknown, joint probability density of a measurement x and the associated errors in z and $h(x|z)$ is the conditional density of x given associated errors z. For a comparison between two models θ_1 and θ_2 we perform a difference calculation as shown in equation 2.45.

$$\mathrm{KLIC}_{12}(\theta_1, \theta_2) = \int \log \frac{f_1(x|z, \theta_1)}{f_2(x|z, \theta_2)} h(x, z) dx dz = \mathbb{E}^0_{x,z} \log \frac{f_1(x|z, \theta_1)}{f_2(x|z, \theta_2)} \tag{2.45}$$

This determines the ability of the first model to produce the underlying distribution relative to the second model and is dependent on $\mathbb{E}^0$ which is the expectation of the underlying distribution. The Vuong-Closeness test will automatically process the underlying distribution so it does not need to be estimated. The log likelihood ratio for each data point x_i is used to define an empirical variance of the log likelihoods as shown in equation 2.46.

$$v^2 = \frac{1}{N} \sum_{i=1}^{N} l_i^2 - \frac{1}{N^2} \left(\sum_{i=1}^{N} l_i \right)^2 \quad \text{where } l_i = \log \frac{f_1(x_i|z_i, \hat{\theta_1})}{f_2(x_i|z_i, \hat{\theta_2})} \tag{2.46}$$

where $\hat{\theta_k}$ is the best estimated parameters for model k and N is the number of data points in the time-series. The Vuong statistic is computed from equation 2.46 by re-normalising the log-likelihood ratio as shown in equation 2.47.

$$V = \frac{\sum_{i=1}^{N} l_i}{\sqrt{\sum_{i=1}^{N} l_i^2 - \frac{1}{N} \left(\sum_{i=1}^{N} l_i \right)^2}} \tag{2.47}$$

The Vuong Closeness test functions by utilising a null hypothesis $\mathrm{KLIC}_{12} = 0$. If the value of V is large then the models are well-distinguishable and can be compared selecting the model with the better likelihood. If V is positive then model θ_1 is the better choice and if V is negative then model θ_2 is the better choice. For low values of V the models are indistinguishable and neither model can be reliably selected as a superior option. The Vuong Closeness test can also be adapted into a False Alarm Probability through the calculation $FAP = 2\Phi(|V|)$, where $\Phi(x)$ is the Gaussian tail function where a cut can be selected so that the probability of the determined value of V is comparable to a confidence level (Baluev, 2012).

2.10 Performance on STILT data

To guide the selection of an appropriate method for period estimation on SkycamT light curves, some of the period estimation methods as well as the bands method and red noise removal method are tested using a dataset of 859 SkycamT light curves consisting

Table 2.1: Data Summary for the 859 object, 12 class, SkycamT light curve dataset.

Number	Class	Type	Count
1	β Lyrae	Eclipsing Binary	57
2	β Persei	Eclipsing Binary	106
3	Chemically Peculiar	Rotational Variable	18
4	Classical Cepheid	Pulsating Variable	67
5	δ Scuti	Pulsating Variable	14
6	Mira	Pulsating Variable	369
7	RR Lyrae Fundamental Mode	Pulsating Variable	26
8	RR Lyrae Overtone Mode	Pulsating Variable	9
9	RV Tauri	Pulsating Variable	5
10	Semiregular Variable	Pulsating Variable	50
11	Spotted Variable	Rotational Variable	22
12	W Ursae Majoris	Eclipsing Binary	116

of variable stars from 12 known variability classes shown in table 2.1. They are a subset of a crossmatch with the AAVSO Variable Star Index catalogue with a tolerance of 148" which have been tested for periodicity using a number of methods and then manually vetted to confirm they are examples of the given class. They are all detected by the GRAPE algorithm discussed in chapter 3 as either a period match with the AAVSO catalogue, or a multiple or submultiple of the catalogue period. The light curves have been processed by the trend removal pipeline documented in chapter 6. The quoted correct estimations and associated failure modes are all determined using a tolerance of 5% ($\epsilon = 0.05$) on the estimated period relative to the AAVSO catalogue period as defined in the Failure Mode subsection.

2.10.1 Bands Method

The bands method allows for the rapid estimation of useful frequencies to reduce the size of the frequency space to be explored. The method returns a maximum of $N_t N_b$ trial frequencies where N_t is the number of frequencies per band and N_b is the number of bands, sorted by variability. The larger the number of trial frequencies returned, the higher likelihood that the true frequency of a given light curve is contained within the banded frequency spectrum (Protopapas et al., 2015). This does not answer which of the two input arguments is dominant in this operation. We performed an experiment using the 859 light curve dataset to evaluate the bands method performance on SkycamT light curves. For a given input of N_t and N_b, a set of trial frequencies is generated with the bands method. The frequency-inverted period which has the closest match to the AAVSO catalogue period is then selected as the 'best match' to the desired period. These objects are then measured for hits, submultiples and multiples.

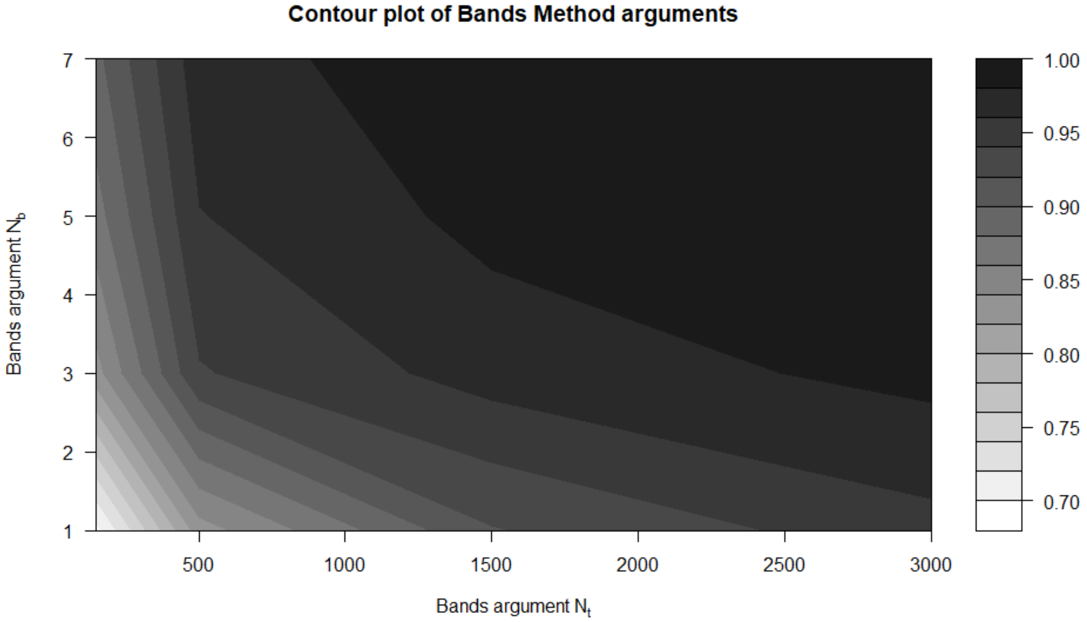


FIGURE 2.13: Contour plot of the hit rate performance of the bands method on the 859 SkycamT light curves as a function of the two input arguments N_t and N_b.

The bands method is run with a minimum period of 0.05 days to the timespan of the light curve (the difference between the maximum and minimum time instant of the light curve) in days and with an oversampling factor $v = 5$ which determines how finally sampled the period range is by defining the step between frequencies shown by equation 2.48 (Ruf, 1999).

$$f_{\text{step}} = \frac{1}{v(t_{\max} - t_{\min})} \tag{2.48}$$

where f_{step} is the step size between frequencies, $t_{\max} - t_{\min}$ is the timespan of the light curve and v is the oversampling factor. This method is used in the generation of the frequency spectrum in all following period estimation tests. We computed this performance for 16 different bands method runs with $N_t = [150, 500, 1500, 3000]$ and $N_b = [1, 3, 5, 7]$. Figure 2.13 demonstrates the contour plot produced by determining the hit rate of each experiment and extrapolating the hit rates into a surface. The N_t argument contributes more to the performance of the bands method. This is not surprising as there is only a limited number of highly variable percentile bands and therefore a value of N_b greater than this results in a performance loss. This result mirrors that found from EROS2 light curves in the original paper (Protopapas et al., 2015).

Table 2.2: Period Estimation algorithm performance on the 859 SkycamT light curve dataset. The LSP has the highest recovery of correct periods and submultiple periods (likely from eclipsing binaries). The VRP has the best performance on minimising the proportion of unknown light curves however many of the light curves generate an aliased result around the one day spurious period.

Method	Hit	Multiple	Submult	1 day alias	$\frac{1}{2}$ day alias	Unknown
LSP	**29.453%**	1.746%	**23.166%**	16.647%	12.573%	28.056%
VRP	17.737%	1.630%	13.970%	46.217%	10.943%	**23.283%**
SLLK	4.773%	1.397%	8.964%	30.966%	19.208%	42.142%
SLLK w/RN	12.806%	5.937%	1.630%	0.931%	0.116%	78.696%
CE	3.609%	2.445%	1.513%	30.617%	4.075%	59.371%

2.10.2 Lomb-Scargle Periodogram

The first row of table 2.2 demonstrates the performance of the Lomb-Scargle Periodogram (LSP) on the 859 light curve dataset. The selected period is simply the highest peak from the periodogram run with a minimum period of 0.05 days to the timespan of the light curve (the difference between the maximum and minimum time instant of the light curve) in days and with an oversampling factor $v = 5$.

The LSP performs reasonably well by recovering over 50% of the light curves as either hits (correct estimations), multiples or submultiples of the catalogue period. The LSP commonly produces a period submultiple which is half the true period for eclipsing binaries which make up a substantial part of the 859 light curve dataset (VanderPlas, 2017). The LSP also returned almost 30% of the light curves with a period equal to the one day alias of the catalogue period due to the diurnal sampling rate of Skycam. Almost 30% of the returned periods did not match with the catalogue period in a known way. The total percentages can add up to over 100% as a period can satisfy multiple criteria simultaneously, commonly a submultiple and an alias period.

2.10.3 Variance Ratio Periodogram

The second row of table 2.2 shows the results of the application of the Variance Ratio Periodogram (VRP) to the 859 object dataset. Due to the computational cost of the VRP relative to the LSP, the bands method was first used to preselect a set of candidate frequencies using the settings $N_t = 500$, $N_b = 3$. Upon the selection of these frequencies, the VRP was computed for each of them and the highest peak is the selected period. Compared to the LSP which uses a similar but simpler model, the VRP performs significantly worse with over 20% poorer recovery of hits, multiples and submultiples. However, the VRP also exhibited less unknown failure modes with a large number of one day aliased periods. On Skycam it appears that the extra flexibility of the multi-harmonic model allows the detection of a form of the correct periodicity but at the sidereal day aliased period. The improved ability of the VRP to fit more non-sinusoidal shapes appears to be overfitting the large noise component of the Skycam light curves. Alternatively, the bands method may be failing to obtain the correct period, but obtaining the aliased period limiting the options of the VRP. The VRP may be useful for fine-tuning periods over small search ranges but is not a good choice for the initial check especially with its computational expense. The single sinusoid of the LSP appears to be a better option.

2.10.4 String Length Lafler Kinman Periodogram

The third row of table 2.2 shows the results of the String Length Lafler Kinman (SLLK) statistic on the 859 object dataset (Clarke, 2002). As with the VRP, the SLLK periodogram is computationally more intensive than the LSP but not to the same extent as the VRP. Therefore the selection of a larger set of candidate frequencies by the bands method was chosen using the settings $N_t = 2500$, $N_b = 3$. The trial period where the SLLK statistic is minimised is then selected as the candidate period for a given light curve. Similar to the VRP periodogram, the models with increased non-sinusoidal model capability struggle with the aliasing structure present in the Skycam light curves. The 42.142% unknown failure mode is significantly inferior to that of the Fourier based methods suggesting that they are the better choice for processing SkycamT light curves likely due to being more resilient to the systematic noise in the data.

The SLLK periodogram is highly disrupted by red noise and therefore the procedure detailed above is used to reduce the effect of this on the frequency continuum. The fourth row of table 2.2 shows the results of the same SLLK periodogram without the application of the red noise removal. Whilst the hit rate is higher, mostly due to the Long Period Variables which suffer extremely heavily from aliasing in the red noise corrected method, as they benefit from the red noise bias at higher periods, the overall unknown failure mode is catastrophically higher. Almost every shorter period variable light curve has returned a red noise generated longer period. As these light curves have been processed by a trend removal pipeline, although limited to yearly trends, there must be additional correlations in the data which will affect the performance of many period estimation methods. Additionally, when the red noise removal method is applied to the SLLK periodogram, performance is highest on light curves with a period under one day. This suggests that the current method may *overfit* the long period noise and therefore it might be responsible for the incidence rate of aliased Long Period Variables. These results suggest that the Fourier based methods are the best option for the development of a dedicated SkycamT period estimation pipeline.

2.10.5 Conditional Entropy Periodogram

The last row of table 2.2 shows the performance of the Conditional Entropy (CE) periodogram. Initially it appears that the method has identified many one-day aliases from the data but under closer inspection these are all just the sidereal day. The cadence of the Skycam light curves results in data points having a very clumpy distribution when folded at the astrophysical period. As a result, the modification to the Entropy periodogram to produce Conditional Entropy breaks down and the results equal those of the Entropy periodogram identifying the dominant spurious period for every light curve. Most of the 859 SkycamT light curves generated a CE period near 0.5 days, 1.0 days and around the 1.0 year period, the same periods identified by computing a spurious period search. As the longer period light curves satisfy the alias requirement at the spurious period, this leads to this result. Of the light curves that satisfied the hit criteria, some were objects with catalogue periods near 0.5 days and 1.0 days whilst the remainder were well sampled Long Period Variables with periods near 1.0 year. Inspection of the periodogram plots indicates that the poor performance of this method is not due to an exponential red noise component. These results clearly show that Conditional Entropy is an extremely poor choice for use on variable cadence data like the Skycam data.

2.10.6 Periodogram Runtime

For each of the computations above, we recorded the time required to compute each statistic with the experiment being performed on a virtual machine with 4 cores of an Intel Xeon E5-2670 processor with 12 GB of RAM. Figure 2.14 shows the mean runtime of each of the methods as a function of the number of data points in each light curve binned into 10 bins. The SLLK and VRP results include the utilisation of a smaller bands method frequency spectrum as defined in their associated subsection above. All the methods have a quadratic dependency on the number of data points. The implementation used for these algorithms is purely single-threaded in the R environment. All the algorithms analysed in this experiment are natively highly parallelisable as the calculation of the associated statistic for a trial frequency f_k is independent from the calculation of any other trial frequency in the frequency spectrum. This means that with sufficient processing cores, the periodograms can be very quickly calculated. For the processing of many light curves, the period estimation task can also be parallelised by computing multiple light curves simultaneously across several processing cores.

The LSP has the largest dependency on the number of data points in the light curves and quickly rises to almost three minutes per light curve with close to 15,000 data points. The CE periodogram is the next most efficient method as it does not require the use

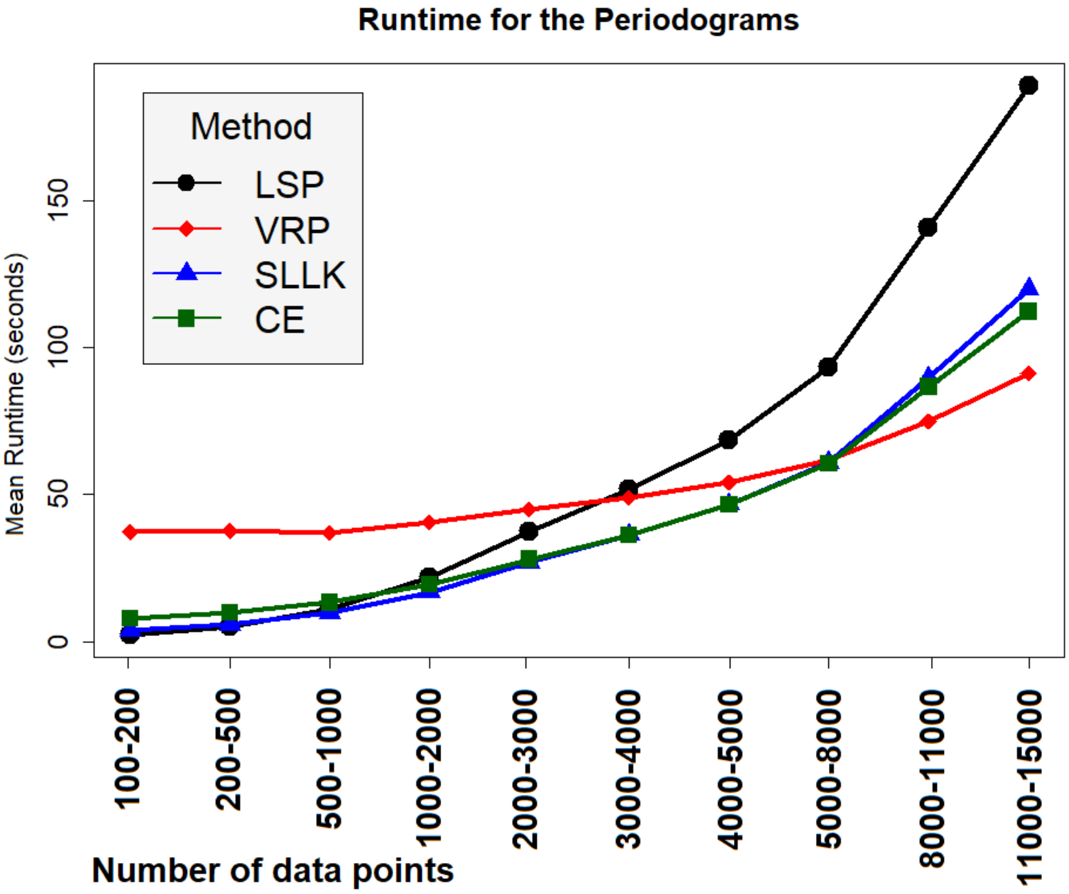


FIGURE 2.14: Plot of the mean runtime of the four periodograms as a function of the number of data points in the light curves binned into 10 bins.

of the bands method and follows a shallower, but still exponential gradient. The SLLK uses the bands method which cuts down the number of trial periods it evaluates to less than 10% of the number evaluated by the LSP and CE methods. At these settings, it follows a similar runtime to the CE method although it appears to have a slightly larger dependency on data point number. The VRP method was substantially slower when computed on low numbers of data points whilst also using the bands method to reduce its frequency spectrum to 1-2% of the LSP frequency spectrum. This is likely due to the complexity of calculating the multi-harmonic model with six learned coefficients for each trial period. However, the method also appears to have a lower dependency on the number of data points although the number of data points in which this difference becomes appreciable is near the limit of the best sampled SkycamT light curves which make up a tiny fraction of the total database.

The computational overhead of the bands method components are also examined using the mean runtimes as a function of data points for each of the combinations of N_t and N_b from the evaluation above. Figure 2.15 demonstrates the runtimes for each of the

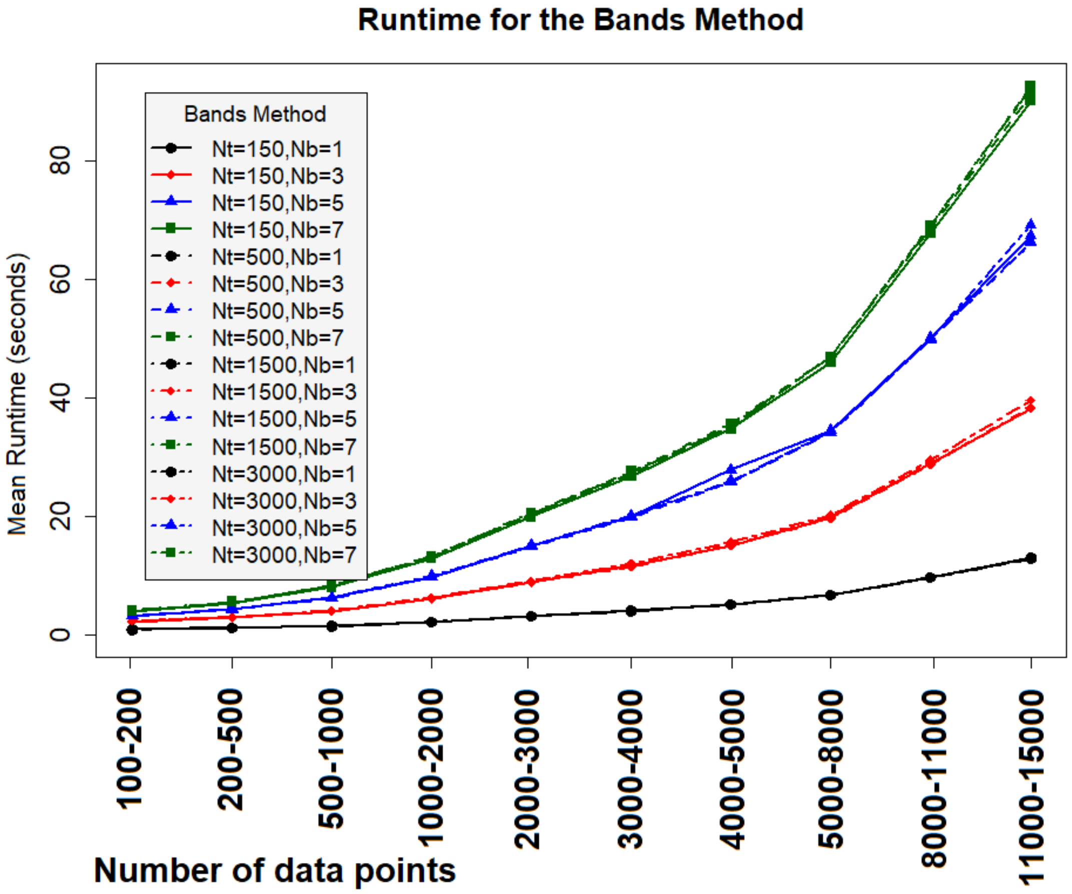


Figure 2.15: Plot of the mean runtime of the 16 bands method evaluations as a function of the number of data points in the light curves binned into 10 bins.

16 combinations. The plot clearly shows that the bands method also has a quadratic dependency on the number of data points which is to be expected as it makes use of the classical 'Schuster' periodogram. The interesting result is the large disparity between the runtime dependency of N_t and N_b with N_t exhibiting virtually zero effect on the computational runtime of the bands method. This makes sense when the algorithm is considered. The value of N_t determines how many frequencies are retained from the classical periodogram, but the whole periodogram must still be computed to decide the N_t retained frequencies. Therefore, the value of N_t has no effect on the number of frequencies computed. On the other hand, N_b does effect the number of classical periodograms which require computation, one per retained band. Therefore it does have a clear contribution to the evaluation runtime.

From the above experiment it is clear that N_t is superior to N_b in the selection of candidate frequencies for the set of 859 SkycamT light curves and it has a reduced computational cost on the calculation of the reduced frequency spectrum. Therefore, it is clear that when using the bands method, it is important to select as high a N_t as required for good performance on the period selection whilst minimising the computational cost of N_b. The bands method has significant potential and has certainly impressed when used on the noisy and poorly sampled Skycam light curves but ultimately it still has a computational overhead which can be still be significant on light curves with many observations.

2.10.7 Conclusion

In this section we have performed experiments to determine the performance and runtimes of several of the period estimation methods described in this chapter. The bands method has obtained good performance despite the limitations of the Skycam data but still requires a quadratic computational overhead even with the good selection of input arguments N_t and N_b. Many of the periodograms exhibited great difficulty in operating successfully on the Skycam light curves with substantial aliased components. The Lomb-Scargle Periodogram based methods appear to provide the best performance as they have the best performance on correctly estimated the periods of the set of known light curves. This is likely due to the resilience of the method to correlated noise in the light curves due to the simple sinusoidal fit which, whilst best tuned to sinusoidal light curves, still maintains reasonable performance on the more non-sinusoidal options such as the eclipsing binary light curves.

The LSP has the poorest performance in terms of runtime combined with a greater quadratic dependence on the number of data points. This can be mitigated by the use of the bands method or possibly a more powerful optimisation method. In conclusion, for the best performance on Skycam light curves, the utilisation of a Lomb-Scargle based method is recommended with an improved optimisation technique which can limit the quadratic dependency on well observed light curves. As the Skycam data is prone to heavily aliasing, the use of a method which can correct for this is recommended. In the next chapter these recommendations are employed in a new period estimation method using the Vuong Closeness test to correct aliased periods in a set of candidate periods determined by a genetic algorithm optimisation.

Chapter 3

Genetic Routine for Astronomical Period Estimation

This chapter has been accepted by the Monthly Notices of the Royal Astronomical Society Journal with submission ID: MN-18-0502-MJ.R1

In chapter 2 the strengths and weaknesses were determined of traditional period estimation methods on SkycamT light curves. It is clear the successful extraction of the period for a periodic source is of great importance to successful and reliable object classification. Many periodic variable types such as the Mira-type red giant long-period variables, the classical cepheid giants and the population II cepheid giants inhabit a strict range of periods (Glass and Evans, 1981; Stetson, 1996; Tanvir et al., 2005; Yoachim et al., 2009). The shape of many light curves is not sinusoidal and is closer to an asymmetrical sawtooth (Tanvir et al., 2005; Yoachim et al., 2009). This can result in difficulties for methods designed to detect sinusoidal variability. Some periodic sources such as eclipsing binaries have a large number of configurations with a wide range of periods (Prsa et al., 2008). In these cases the precise determination of the period of these eclipses is still required for the generation of the characteristic eclipse shapes (Paegert et al., 2014).

In this chapter we introduce a new approach to facilitate the rapid estimation of light curve periodicity through the use of a tuned genetic algorithm named GRAPE: Genetic Routine for Astronomical Period Estimation. This method combines genetic parameter optimisation with astronomical period estimation functions and alias correction routines allowing the quick and accurate evaluation of candidate periods across a large potential period range for light curves with many observations (Charbonneau, 1995). The fitness function for a given candidate period is determined by a Bayesian Generalised Lomb-Scargle (BGLS) periodogram capable of detecting periodic variability in light curves with substantial noise components (Mortier et al., 2015; Mortier and Cameron, 2017). Genetic algorithms, by the virtue of their ability to quickly move around large, high dimension, parameter spaces allow this period estimation task to be performed with a high speed even on light curves with thousands to tens of thousands of data points. The versatility and speed of this method is offset by a difficulty in fine-tuning the solution as genetic algorithms sacrifice absolute accuracy on a final solution with finding it in such a large parameter space.

3.1 The frequency spectrum

All period estimation methods produce a statistic which is a function of period or frequency. Of course, this makes perfect sense as determining strong periods is the result we are hoping to obtain. However, many approaches select trial periods through the creation of a frequency spectrum. This discretised vector of candidate periods limits the precision of best-fit period the algorithms can present. Whether the method looks for a minimum or a maximum of a given statistic, or perhaps something more complicated, the period associated with this optimised candidate period is still one of these initial candidates. Period estimation pipelines often use a fine-grid search to mitigate this by performing ever finer grids around strong candidates which results in increased computational effort (Debosscher et al., 2007; Richards et al., 2011b, 2012). It is also likely that many of the trial frequencies contain nothing of value and large sections of the frequency spectrum may be filtered out prior to more computationally expensive calculations. Yet, despite the capability of frequency spectrum approaches, it is clear that the true period parameter space is a continuous variable space. The discretised frequency spectrum can be oversampled to a degree where frequency estimation errors associated with the finite observational data are larger than the gaps between consecutive candidate frequencies. As true frequencies produce Gaussian shape peaks due to finite-length data (VanderPlas, 2017), a candidate frequency which is slightly out of position from the true frequency might only be evaluating the side of the Gaussian peak. This would result in a lower valued statistic than would be appropriate for the true period resulting in an increased potential for an incorrect result. This can be seen in long period variability candidates as the frequency spectrum is reciprocal to the period space. The treating of the period parameter space as a continuous variable allows for the correction of this artefact.

Trial period extraction using the bands method was used as a follow-up to the development of the Correntropy Kernalised Periodogram (CKP) (Huijse et al., 2012). Due to the increased computational complexity of this calculation, it was determined that preselecting trial periods was optimal (Protopapas et al., 2015). The bands method was introduced as a method to conduct this preselection. The approach operates on the idea that data points of similar magnitude are likely to be found at integer multiples of any underlying period within magnitude percentile bands which contain a strong signal component. A candidate variable light curve can be dismantled into a number of percentile bands. N_b percentile bands have their spectral window function computed on a linearly spaced frequency grid and the resulting N_t local maxima are retained. This results in $N_b N_t$ trial periods for further analysis with the values of N_b and N_t chosen as a trade-off between computational load and correct trial estimation (Protopapas et al., 2015). This trial period selection approach, whilst still relying on frequency grids, suggests that it is possible to rapidly evaluate sections of the candidate period parameter space. With our proposed method, we consider an approach that can treat the period parameter as a continuous variable whilst simultaneously exploring the period parameter space and isolating regions of interest to optimise to the true underlying period.

3.2 Genetic Algorithms

Period estimation tasks are best described as a global optimisation problem across a continuous parameter space. This optimisation must be solvable with a minimal computation of trial periods to identify a global minimum in a cost function. This is a function of the parameter we wish to evaluate, in this case the period of a light curve, and the trial period at the global minimum is the desired result (Charbonneau, 1995; Rajpaul, 2012). Many optimisation routines have been developed but the complexity of the underlying cost function can rapidly diminish the performance of many problems until they are no better than a brute force approach. Evolutionary algorithms are inspired by biological evolution and are capable of exploring a large, possibly high dimension, parameter space efficiently whilst also selecting good candidate results with a high precision (McWhirter et al., 2016). Genetic algorithms are the most popular subset of evolutionary algorithms and utilise computational variants of many well-known staples of biological evolution (Holland, 1975). These include natural selection (survival of the fittest), genetic recombination, inheritance and mutation (Rajpaul, 2012). Genetic algorithms have been utilised in a number of problems in astrophysics (Rajpaul, 2012) including period estimation (Charbonneau, 1995). To our knowledge, we have not seen them employed with current generation periodograms using their highly capable models in the form of a fitness function whilst allowing the genetic algorithm to optimise to the desired period result without requiring a frequency spectrum.

3.3 Bayesian Generalised Lomb-Scargle Periodogram

The Bayesian Generalised Lomb-Scargle Periodogram (BGLS) applies a Bayesian framework to the generalised Lomb-Scargle periodogram by applying a sinusoidal prior to the data (Mortier et al., 2015). This allows the transformation of the power statistic of the Lomb-Scargle periodogram to a posterior probability of frequency f given light curve data D and a sinusoidal model I producing equation 3.1.

$$P(f|D,I) \propto \exp\left(P_{\mathrm{GLS}}(f)\right) \tag{3.1}$$

where $P_{\mathrm{GLS}}(f)$ is the generalised Lomb-Scargle power and $P(f|D,I)$ is the probability of a fitted candidate frequency given the light curve data and a sinusoidal signal. This exponential transformation results in the suppression of less-significant peaks in the periodogram such as the noisy continuum. Caution must be taken when applying such a method as it can eliminate defects in the estimated power spectrum that could give insight to properties of the light curve such as the associated cadence and possible non-sinusoidal features. Mortier et al. developed the BGLS on the formalisms of the original generalised Lomb-Scargle periodogram (Zechmeister and Kürster, 2009; Mortier et al., 2015). These formalisms are shown in equations 3.2 to 3.10

$$W = \sum_{i=1}^{N} w_i \tag{3.2}$$

$$Y = \sum_{i=1}^{N} w_i d_i \tag{3.3}$$

$$\widehat{YY} = \sum_{i=1}^{N} w_i d_i^2 \tag{3.4}$$

$$\widehat{YC} = \sum_{i=1}^{N} w_i d_i \cos(2\pi f t_i - \theta) \tag{3.5}$$

$$\widehat{YS} = \sum_{i=1}^{N} w_i d_i \sin(2\pi f t_i - \theta) \tag{3.6}$$

$$C = \sum_{i=1}^{N} w_i \cos(2\pi f t_i - \theta) \tag{3.7}$$

$$S = \sum_{i=1}^{N} w_i \sin(2\pi f t_i - \theta) \tag{3.8}$$

$$\widehat{CC} = \sum_{i=1}^{N} w_i \cos^2(2\pi f t_i - \theta) \tag{3.9}$$

$$\widehat{SS} = \sum_{i=1}^{N} w_i \sin^2(2\pi f t_i - \theta) \tag{3.10}$$

where t_i is the time instant of data point i, d_i is the magnitude of data point i, w_i is the weight of data point i, f is a candidate frequency and θ ensures the orthogonality of the sine and cosine equations (as discussed in the Lomb-Scargle section in chapter 2) shown in equation 3.11.

$$\theta = \frac{1}{2}\tan^{-1}\left[\frac{\sum w_i \sin(4\pi f t_i)}{\sum w_i \cos(4\pi f t_i)}\right] \tag{3.11}$$

Combining these expressions into the least squares representation yields solutions for the coefficients of the sine and cosine components with respect to the Bayesian priors to produce the proportionality that describe the probability of the candidate frequencies as a function of the sinusoidal components shown in equations 3.12 to 3.15.

$$P(f|D,I) \propto \frac{1}{\sqrt{|K|\widehat{CC}\widehat{SS}}} \exp\left(M - \frac{L^2}{4K}\right) \tag{3.12}$$

where:

$$K = \frac{C^2\widehat{SS} + S^2\widehat{CC} - W\widehat{CC}\widehat{SS}}{2\widehat{CC}\widehat{SS}} \tag{3.13}$$

$$L = \frac{Y\widehat{CC}\widehat{SS} - C\widehat{YC}\widehat{SS} - S\widehat{YS}\widehat{CC}}{\widehat{CC}\widehat{SS}} \tag{3.14}$$

$$M = \frac{\widehat{YC}^2\widehat{SS} + \widehat{YS}^2\widehat{CC}}{2\widehat{CC}\widehat{SS}} \tag{3.15}$$

As the prior assumes the data has been generated by an underlying sinusoidal signal, important peaks are suppressed for more unusual shaped signals. Therefore this method is usually only recommended in situations where the data is highly sinusoidal as the aliased features in the generalised Lomb-Scargle method will convey more important information to the user. The approach introduces a hyperparameter named ‘white noise jitter’. This parameter is used to determine the variance of the prior sinusoidal model by additional signals and correlated noise in the light curve. If this parameter is too low, every frequency model has a vanishingly low probability and the periodogram collapses to zero whereas if it is too large, the noise is incorporated into the frequency models and noisy peaks are generated in the continuum.

This initial assessment is not complete in this case as one of the main aims of this thesis is the development of a completely automated period estimation technique. For a completely automated method, it is somewhat more important to suppress complications in the periodograms at the risk of lower performance on non-sinusoidal data. As the genetic algorithm must quickly determine the significant frequencies of a light curve, the suppression of the myriad of low-significance peaks assists in the optimisation routine. The assumption of an imperfect prior distribution is a worthy price to pay for tighter failure mode control, the suppression of aliased frequencies and more reliable optimisation. These periodograms can be further stacked to reveal fluctuations in the Signal to Noise of detected sine functions in time (Mortier and Cameron, 2017).

3.4 The GRAPE method

We introduce the method we developed named GRAPE: Genetic Routine for Astronomical Period Estimation. The method was needed to improve the performance and runtime of period estimation on the Skycam database. As a result of the Skycam instruments not controlling the motion of the Liverpool Telescope, Skycam has a unique variable cadence compared to more regular surveys (Mawson et al., 2013). There are no guarantees when an object will be resampled and often a cluster of data points with a sampling of minutes are separated by gaps many days long. This can lead to difficulties such as aliasing where sampling periods reflect signals across the parameter space. Figure 3.1 demonstrates the difference between Skycam cadence and regular cadence for a simulated sinusoidal signal with white noise.

3.4.1 Initialisation method

Genetic algorithms operate by employing a set number of individuals (or phenotypes) named the population. These individuals are a set of candidate solutions (in our case candidate periods). The individuals are distributed across the feature space in either a random or in an organised approach, such as from a previously identified underlying distribution like the bands method. Despite these varying degrees of complication, generally the improvement from a uniformly random distribution is minimal. We describe how GRAPE establishes its initial population below as it requires definitions of the parameter space. This population is evolved from generation i to generation $i + 1$ until a cut-off point has been reached, either when the best answer reaches a given precision or a predefined number of generations has been computed (Charbonneau, 1995).

The parameters of the solution, which in our case is just a single period parameter, must be encoded into a string named a chromosome with each character named a gene. GRAPE utilises base-10 chromosomes where each gene can have an integer value $I \in \mathbb{Z} \in [0, 9]$ where $\mathbb{Z}$ is the set of integers. Our encoding process is very simple and is dependent on the size of the period parameter space to be explored. GRAPE operates in frequency space and the frequency search is performed from a minimum frequency

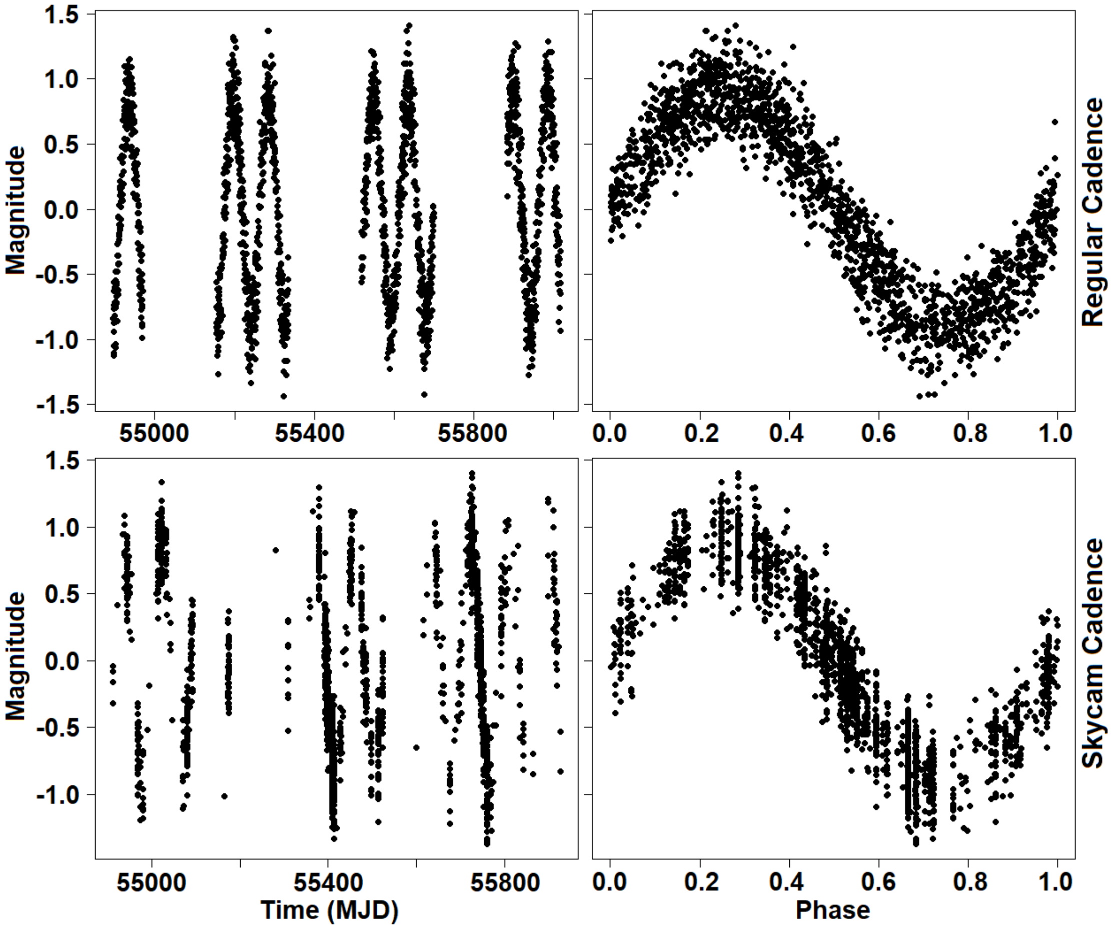


FIGURE 3.1: A simulated sinusoidal signal with white noise sampled with more traditional regular cadence and the Skycam highly variable cadence. Top left is the regular cadence raw light curve, top right is the regular cadence phased light curve, bottom left is the Skycam cadence raw light curve and bottom right is the Skycam cadence phased light curve. Skycam cadence clearly introduces new structures into the data which can produce spurious and aliased results in period estimation tasks.

f_{min} defined in Equation 3.16.

$$f_{\text{min}} = \frac{1}{t_{\text{max}} - t_{\text{min}}} \tag{3.16}$$

where t_{max} is the time instant of the last data point and t_{min} is the time instant of the first data point to a maximum frequency of $f_{\text{max}} = 20$ (Charbonneau, 1995). This f_{max} is equivalent to detecting periods down to 1.2 hours. This is sufficient to detect the variable star classes expected to be present in the Skycam data. The f_{max} can be increased, if required, to probe shorter duration periods at a cost of increased computational load as the parameter space to be explored is larger. All units of frequency are in cycles per day (d^{-1}). For any candidate frequency f_{can}, which always obeys $f_{\text{min}} \leqslant f_{\text{can}} < f_{\text{max}}$, the encoded chromosome is generated by first rescaling the frequency space so that any candidate frequency $\in [0,1]$ using Equation 3.17.

$$f_{\text{scaled}} = \frac{f_{\text{can}} - f_{\text{min}}}{f_{\text{max}} - f_{\text{min}}} \tag{3.17}$$

The frequency is rounded to 10 decimal places and the value of each decimal is encoded into the associated gene creating chromosomes containing 10 genes. Using this calculation and its inverse the genetic algorithm can perform generational updates on the chromosomes and extract new candidate frequencies for testing. Whilst we discuss our treatment of period as a continuous variable, it is important to recognise that we still have a precision limit with this method. The base-10 chromosomes can encode ten decimal places of the normalised period space. Skycam light curves have a baseline of approximately 1000 days which results in a precision of 10^{-10} days for extremely short periods and 10^{-3} days for periods close to the maximum period. Light curves with baselines of up to several hundred years should still maintain a 0.1 day precision at candidates close to this maximum. A frequency spectrum from 0.05 days to 1000 days and an oversampling factor of 10 would only achieve equivalent worst-case precision on periods under 3 days, and are as low precision as 2+ day precision above 100 days. Therefore, we believe that our precision is sufficient to justify our description of a continuous parameter space and the use of 10 decimal places. As with the f_{max} calculation, the number of genes can be increased at increased computational cost. The precision of, at worst, 10^{-3} days for long period variables using 10 decimal places is sufficient to produce folded light curves sufficient for further analysis which is why this value was selected.

The initial GRAPE population (individuals of generation 1) of number N_{pop} is generated using two distributions to sample sufficiently across the parameter space prioritising regions where periodic variability is common. The first half of the population is generated using a distribution which is base-10 logarithmic in frequency space shown in Equation 3.18.

$$\{\text{Pop}_1\}_i = 10^{\textit{runif}(\log_{10}(f_{\text{max}})-\log_{10}(f_{\text{min}}))+\log_{10}(f_{\text{min}})} \tag{3.18}$$

where $\{\text{Pop}_1\}_i$ is the i^{th} first half set candidate frequency, rerun for $i = 1, ..., 1000$ to give $\{\text{Pop}_1\}$, the set of first half set candidate frequencies, *runif* is a set of uniformly distributed random numbers generated between 0 and 1 and f_{min} and f_{max} are as above. This distribution skews the population towards lower frequencies and therefore higher periods yet it still heavily samples the low period end of the parameter space as it is a function of frequency. The second half of the population is produced by a function linear in period space and therefore a reciprocal in frequency space. This set of individuals is generated using Equation 3.19 with the same definitions as above.

$$\{\mathrm{Pop}_2\}_i = \frac{1}{runif\left(\frac{1}{f_{\min}} - \frac{1}{f_{\max}}\right) + \frac{1}{f_{\max}}} \tag{3.19}$$

where $\{\mathrm{Pop}_2\}_i$is the i^{th} second half set candidate frequency, rerun for $i = 1, ..., 1000$ to give $\{\mathrm{Pop}_2\}$, the set of second half set candidate frequencies. This distribution is highly skewed to low frequencies and therefore high periods and is required to ensure the long period end of the parameter space is sufficiently explored. Had we used a distribution that was simply linear in frequency space, it would often remain in the low period end of the parameter space resulting in a heavy loss of performance. Figure 3.2 demonstrates this candidate generation process showing a set of generated candidate periods. The candidate frequencies must then be sorted based on their performance on a fitness function which computes how well they fit the data. This sorting is then used to determine which candidate frequencies should be retained to subsequent generations. GRAPE uses the Bayesian generalised Lomb-Scargle periodogram (BGLS) as its fitness function (Mortier et al., 2015; Mortier and Cameron, 2017). This method is quickly computable due to a lack of any operation more complex than simple summations. Despite this light computational complexity, it was one of the best performing periodograms when trialled on SkycamT light curves which had been matched to known periodic variable stars through cross-matching with the American Association of Variable Star Observers (AAVSO) Variable Star Index (VSI) catalogue. We surmise this is likely due to the methods flexibility on population mean and data point weighting combined with the ability to control the signal and noise components through the careful use of the white noise jitter argument. Whilst we make use of BGLS as our fitness function, it is important to note that the genetic algorithm design is highly modular and other periodograms or combinations of periodograms may be used. Our main priority was to maintain high accuracy across the entire period space whilst limiting the processing time on the Skycam data which exhibits 136,420 light curves with more than 2000 data points which can be computationally intensive with the frequency spectrum approach (Mawson et al., 2013).

In chapter two the parallelisation options of the frequency spectrum methods were discussed. As the fitness function at a given candidate frequency is independent of the other frequencies, the frequency spectrum can be split into multiple groups and run on separate processing threads and reunited to produce the result. For a large set of light curves, the parallelisation can be applied by splitting the light curves into multiple groups across separate processors where the frequency spectrum for a given light curve is computed on one thread, but multiple light curves can be analysed simultaneously. Genetic Algorithms are also easily parallelisable for a bulk set of light curves

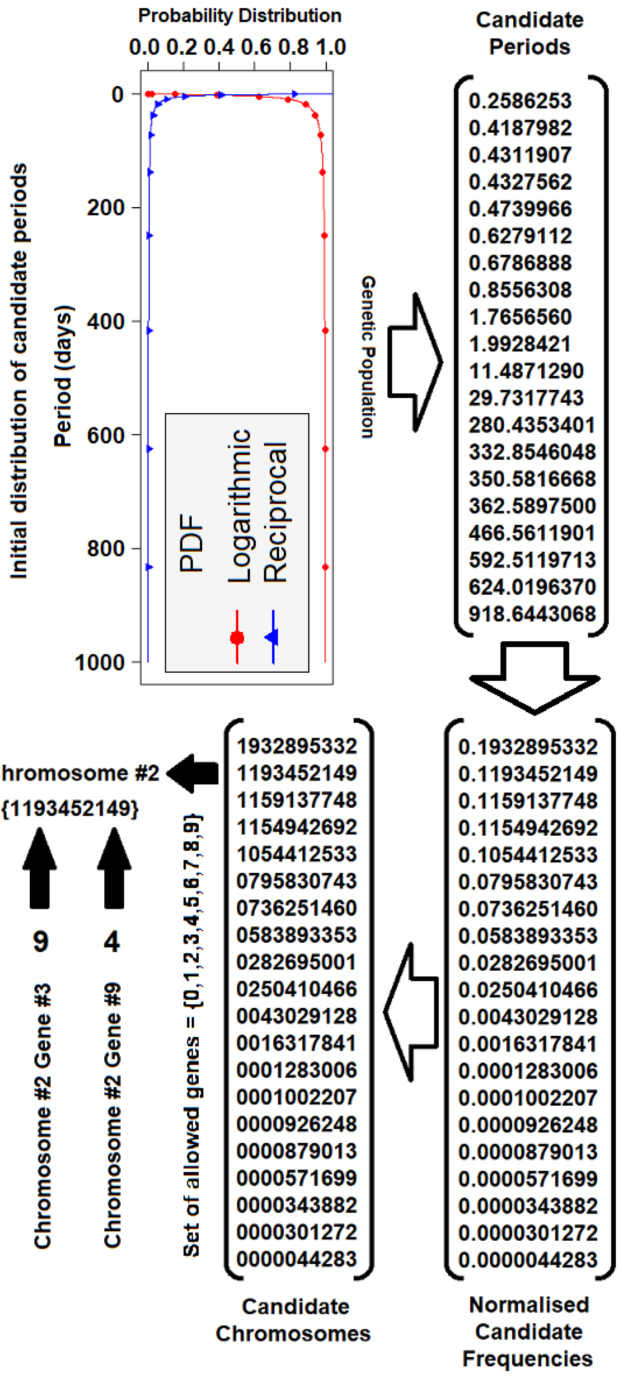


FIGURE 3.2: A demonstration of the creation of the initial population of candidate frequency chromosomes. The top left plot shows the probability distributions used to generate the randomly initialised candidates. The red line with circular points is the logarithmic distribution from equation 3.18 and the blue line with triangular points is the reciprocal distribution from equation 3.19. The top right set shows 20 candidate periods drawn from these two distributions, 10 from each distribution. The bottom right set shows these periods are transformed into frequencies and normalised based on the minimum and maximum allowed frequencies for the optimisation. In the bottom left set the candidate frequencies are rounded to 10 decimal places and these 10 decimals become the 10 genes which makes up the genetic chromosome for each candidate. The chromosomes are strings of ten digits where each digit represents a gene. This representation allows precision with our 0.05 to 1000 day period range of 10^{-3} days at the long period extreme to 10^{-10} days at the short period extreme. This precision allows the optimisation to treat the parameter space as a continuous variable.

but they are more difficult to parallelise across multiple frequencies. Whilst the fitness function remains independent across the different candidate frequencies, the evolution of the population requires the computation of each candidate frequency in a given generation before the next generation can be evolved. This places a limit on how easily the genetic algorithm can be parallelised. There are potential solutions to this issue such as the Island model used to parallelise genetic algorithms on a Graphics Processing Unit (GPU) (Cantu-Paz, 2000; Pospíchal et al., 2010). In this method, the population is split into multiple independent islands which evolve individually towards different suboptima with an additional step called migration. Migration allows the transfer of good genetic chromosomes between the islands to influence the evolution towards the best performing individuals across every island. The current implementation of GRAPE does not make use of the further parallelisation options such as the island method and therefore is only suited to parallelisation for across multiple light curves. Depending on the resources available, with a sufficiently large set of light curves such as those present in the Skycam database, GRAPE is capable of reasonable parallelisation. For an extremely large quantity of processing cores, the frequency spectrum approach may offer faster performance due to the increased potential for parallelisation across candidate frequencies and light curves. In this case, the frequency spectrum will still be limited by the oversampling precision and therefore it is possibly better to apply the additional genetic algorithm parallelisation methods to GRAPE.

3.4.2 Evolutionary Optimisation

The candidate frequencies chromosomes, sorted by the BGLS fitness function, must be used to create the next generation propagating through knowledge gained from the initial population whilst allowing flexibility to explore regions that were not initially scanned. Genetic algorithms use a mechanism analogous to sexual reproduction to generate the subsequent generation. A number of the previous generation are paired up into N_{pairups} partners. Unlike sexual reproduction, the pairups can be with two copies of the same individual if the genetic algorithm selects it. It is also important to utilise a parameter named *selection pressure* for this operation. This determines what quantity of the fittest individuals are selected to reproduce, the parents. Selection pressure can take several forms such as Tournament selection where random subsets of the individuals compete to be selected as a parent or Roulette selection where a probability of selection is assigned to each individual summing to one based on their fitness function evaluation. For GRAPE we utilise the *Roulette* selection pressure method which is superior at preserving important genetic diversity than simply selecting the top best-fitting individuals. We employ an argument named $P_{\mathrm{fdif}} \in \mathbb{R} \in [0, 1]$ which defines the contribution of the sorted fitness function to selecting the parent individuals. The argument $P_{\mathrm{fdif}} = 1$ results in a high contribution and essentially results in the best performing individual pairing up with itself N_{pairups} times. Whilst it would seem to be a good idea as it uses the best fitting frequency, the loss of all other genetic information greatly hinders the algorithms ability to explore other areas of the frequency space. The algorithm would likely get stuck at one of the intial candidate frequencies and fail to find the true frequency. On the otherhand, $P_{\mathrm{fdif}} = 0$ means that the parents are chosen randomly from the previous generation meaning very little learned knowledge of the fitness function is propagated to the next. This results in an algorithm that haphazardly jumps around the parameter space randomly requiring a 'lucky hit' in one of the final generations to produce a reasonable result. Ultimately, a value must be determined that results in a propagation of fitness information without a complete loss of genetic diversity. It is also extremely important that the rank number for the sorted chromosomes be used, not the raw BGLS statistic, or else the selection might still focus on the best performing chromosome despite the selection pressure. This is due to the large difference between BGLS statistic values for even two similarly performing candidate frequencies (Charbonneau, 1995). Selection pressure is what introduces the survival-of-the-fittest into our approach.

Upon the selection of the N_{pairups} reproduction events, the children of these events must be determined. Each two parents produce two children as part of the reproduction. These children inherit genetic information from their parents but may also be a new unique formulation of this information. This means that the reproduction will often produce two new candidate frequencies for evaluation in the next generation. These frequencies use genetic information from the previous generation whilst still exploring a new part of the parameter space. The two mechanisms used for this modification of the parents' chromosomes are called crossover and mutation. Crossover is analogous to how children inherit traits from both their parents' by splitting the parent chromosomes into sections and then recombining them into two children with combinations of both parents chromosomes. GRAPE utilises the simplest form of crossover by performing a single split on both parent chromosomes. First, an argument named $P_{\mathrm{crossover}} \in \mathbb{R} \in [0, 1]$ determines the probability that any given reproductive event will result in a crossover. If a crossover is triggered during a reproduction, a uniform random number generator selects an integer value $\in \mathbb{Z} \in [1, 10]$. This determines at what location the split will occur. Child chromosome 1 will contain the genes from parent chromosome 1 up to the split location and the genes from the parent 2 chromosome after the split location. Child chromosome 2 is the inverse with the first part being from parent 2 and the second part from parent 1. If no crossover occurs, the children chromosomes are identical to the parent chromosomes at this point. Next the mutations are calculated. Mutations are randomly selected gene changes inside the children chromosomes. For each gene in the two children chromosomes, an argument *mutation* determines the probability that a mutation occurs. A uniform random number generator $\in \mathbb{R} \in [0, 1]$ generates a value for each of the 20 genes in the two children chromosomes. Any genes for which the associated generated value is below or equal to P_{mutation} is determined to have undergone a mutation. For each mutated gene, a uniform random number generator computes an integer value between 0 and 9 and assigns this value as the new gene value at that location. Mutation allows the genetic algorithm to create new candidate frequencies in previously unvisited locations without requiring the two parents to contain this information in their chromosomes. As each child gene is subjected to this mutation probability, it is usually recommended to use a low value for the later generations to prevent the genetic algorithm from jumping away from optimised answers. Upon the completion of the reproductive events, any new child chromosomes are evaluated using the BGLS and appropriately ranked by the fitness function. Figure 3.3 shows an example of this genetic evolution using some of the chromosomes created in figure 3.2. These operations result in a generation $i + 1$ population of size $N_{\text{pop}} + 2N_{\text{pairups}}$. As each subsequent generation would produce new offspring, the population size would rapidly increase introducing extra computational complexity with zero benefits. Therefore, we must remove $2N_{\text{pairups}}$ candidate frequencies from the population in a process analogous to biological death with the exception that, if the algorithm chooses it, any individual could live forever. We employ an 'anti-selection' pressure to accomplish this which we call the *death fraction*. This is, like selection pressure, a probability $P_{\text{dfrac}} \in \mathbb{R} \in [0, 1]$ which determines what proportion of the removed candidate frequencies were from the poorest performing BGLS frequencies verses randomly eliminating chromosomes of any fitness ranking. The only difference from selection pressure is that we always retain the best performing candidate frequency from the current generation to preserve its genetic knowledge. This last step produces the individuals of generation $i + 1$ and the cycle continues as described towards the production of generation $i + 2$. This cycle is repeated until a predetermined cut off or generation N_{gen} is produced. The fittest individual from this final generation is then returned as the genetically chosen fittest chromosome. It is then decoded into a chosen frequency and returned to the user. Figure 3.4 demonstrates how the exploration of the period space evolves across the generations due to the propagation of genetic information on the performance of the fitness function for a sinusoidal light curve. It is clear that the whole period space is investigated initially and by generation 40 the only remaining regions are the true period and its multiples. By generation 80 all regions of the period space apart from that located at the best fit period are discarded. The final generations are utilised for fine tuning the ultimate result.

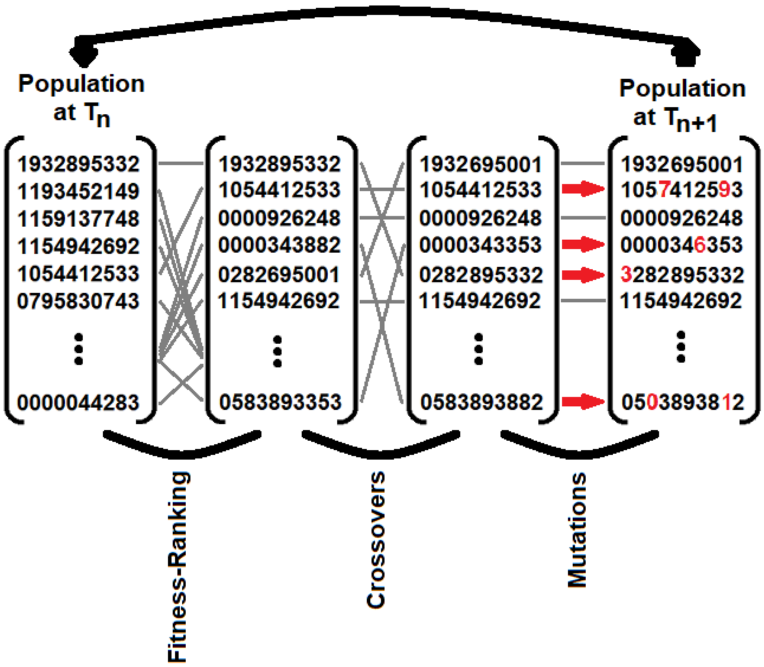


Figure 3.3: A demonstation of the primary evolutionary methods utilised by GRAPE. As seen in figure 3.2, the candidate frequencies are expressed as a set of chromosomes encoding the frequency information. A set of the population is selected to produce the next generation. The first operation utilises the fitness function, in our case a BGLS function. This function is used to rank the candidate frequencies in order of their response as the candidates selected for reproductive operations are dependent on their model performance. The crossover operation selects two chromosomes and generates two children by splitting the chromosome at a given point between two genes and placing the starting genes of parent one with the final genes of parent two and vice versa. Finally, the mutation operation can select any of the genes for a change, also known as a 'copy error'. Once complete, the newly generated chromosomes are ranked and become part of the next generation. This generational update is then repeated as many times as required to produce a population of high performing individuals based on the chosen fitness function.

3.4.3 Candidate period determination

Many period estimation methods including the Bayesian Generalised Lomb-Scargle used in the GRAPE fitness function suffer from common failure modes as a result of the cadence of a light curve and non-sinusoidal shape. This translates into the periodogram as multiple significant peaks of which any one may be the true astrophysical period. Therefore our genetic algorithm must be capable of identifying not just a global optimum but clusters of persistent local optima which are then optimised to a candidate period. Additionally, whilst it would be preferred if the chosen frequency could be trusted as the best possible result for a given set of data, genetic algorithms suffer from the disadvantage that they are somewhat a black box. The propagation of the candidate frequencies from generation to generation is heavily controlled by randomness sourced from random number generators as obvious from the method described above. This can

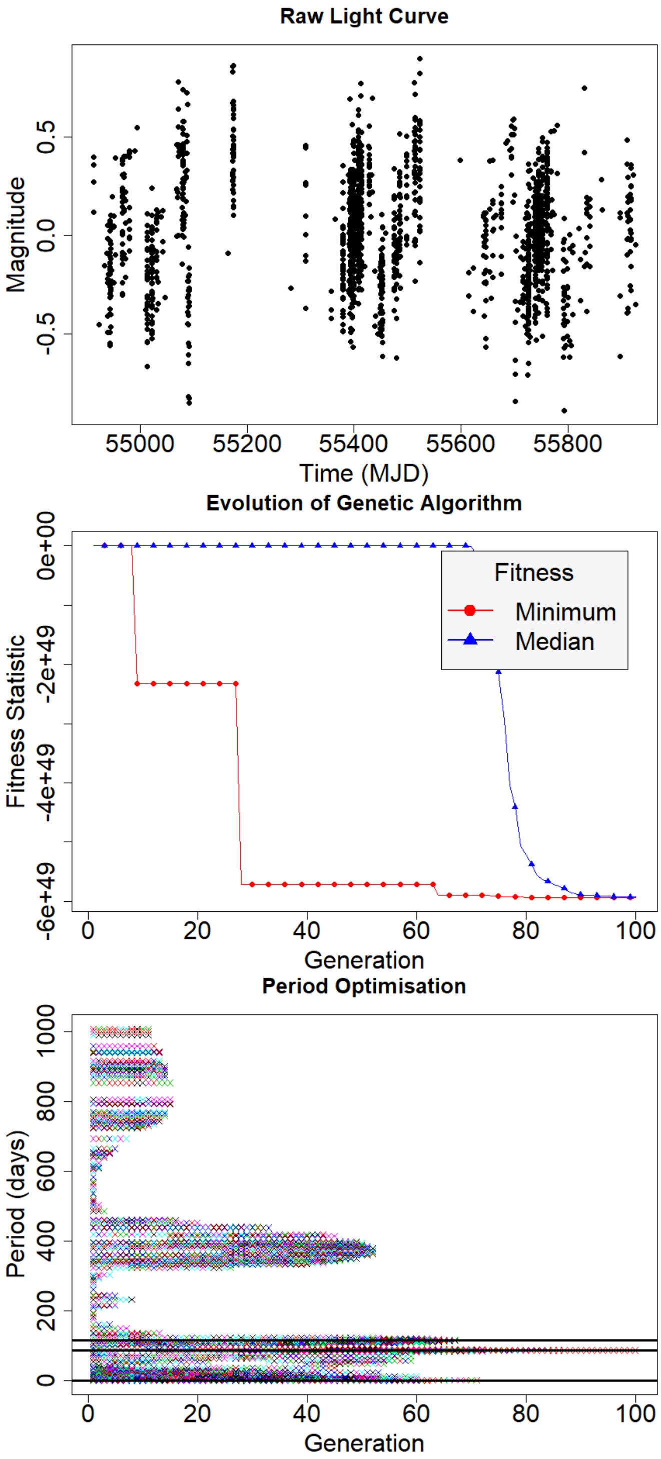


FIGURE 3.4: Plots of the exploration of the period space against generation number for a simulated sawtooth light curve. The top plot shows the simulated light curve that GRAPE is processing. It is a Skycam cadence simulated sawtooth with a period of 87.58 days and a signal-to-noise ratio (SNR) of 2. It has 1721 data points sampled in yearly seasonal windows with a strong diurnal sampling. The middle and bottom plots demonstrate the evolution of the GRAPE genetic algorithm across the 100 generations of genetic optimisation. The middle plot shows the minimum and median statistics of the BGLS fitness function for every candidate in each generation. The red line with circular points is the minimised candidate for each generation and the blue line with triangular points is the median of the fitness of every candidate for each generation. The bottom plot shows the location of each generations candidate frequencies in period space. The candidate periods selected by the genetic optimisation are demonstrated by the black horizontal lines.

result in differences in the exploration of the parameter space purely due to the use of different random number seed values in each run. Therefore, GRAPE has been designed to detect N_t trial periods through the use of k-means clustering for each generation of the routine where $k = N_t$ with a 'dominant' seed value. The clusters with a standard deviation below a critical value σ_t are recorded per generation. The routine continues until standard deviation of the cluster means for a given generation drops below σ_t. The cluster means are then rounded to 2 decimal places and analysed for repetition across multiple generations. The top $N_t - 1$ cluster means are selected as candidate periods along with the global minimum of the genetic run. Due to the randomness of the genetic algorithm this operation is performed a second time with a different 'submissive' seed value. In the event that any of the candidate periods are close repetitions of any other candidate period, the repeated candidate periods are replaced with new candidate periods from the submissive run.

Upon the production of N_t candidate periods which may or may not be similar, they are fine-tuned through the use of N_t single-period genetic runs with N_{finegen} generations to achieve an optimised result for a period range of ±10% of the candidate period. The N_t candidate periods are then tested to determine which period produces the best performing fit to the light curve. It is possible to use the BGLS to determine the best performing of these frequencies, however, GRAPE utilises a more powerful information-theoretic statistic. The Vuong closeness test is a statistical model similarity measure based on the Kullback-Leibler Information Criterion (Vuong, 1989). This method was proposed in the discrimination of aliased periods during the period estimation task (Baluev, 2009). Aliased periods are reflections of true periodic signals with a sampling period, also known as a spurious period. They are calculated by Equation 3.20 (Heinze et al., 2018).

$$P_{j,f} = f\frac{t_{sid}}{(t_{sid}/P_t) + j} \tag{3.20}$$

where $P_{j,f}$ is a set of alias periods produced by a trial period P_t, $t_{sid} = 0.99726957$ days is the sidereal day, f is an integer value from a vector of values, $f = [1, 2, 3]$ which defines the multiples of the trial period and $j = [-3, -2, -1, -0.5, 0, 0.5, 1, 2, 3]$, the set of possible alias frequencies limited to $|j| \leq 3$. This is due to higher values of j producing aliases with fitness function response of similar order to the noise continuum (VanderPlas, 2017). For each of the N_t periods, GRAPE generates 27 sinusoidal regression models for the 27 $P_{j,f}^{-1}$ frequencies with 4 parameters, an intercept, a linear trend and a sine and cosine component. These models are used to compute the Vuong Closeness statistic between the trial period P_t model and each of the $P_{j,f}$ models. If the Vuong Closeness statistic indicates one of the $P_{j,f}$ models outperforms the trial period, the trial period is replaced with the Vuong Closeness maximising candidate period as long as it is not a known spurious period calculated previously.

Upon the generation of the chosen N_t periods by the genetic algorithm, GRAPE generates another set of $N_t + 2$ sinusoidal regression models with 4 parameters, as above. The frequency of these Fourier components are defined as the N_t genetically chosen frequencies, a constant model with only an intecept and a daily model with a period of one day. Finally, the N_t models for the chosen periods are compared using the Vuong closeness test to select the final period. GRAPE offers two options for this method. In the first, the model of every chosen period from each rerun is compared to every other period model. This requires the computation of $N_t^2 - N_t$ Vuong closeness tests. Alternatively, if the value of N_t is prohibitively high (although this many reruns is unlikely), the chosen periods are all compared to a constant model with no linear or sinusoidal terms. This requires N_t Vuong closeness tests. This method does have the disadvantage that it may screen out the correct period in favour of one due to the sampling and therefore we recommend the first option. This last step completes the GRAPE routine and results in a determined periods.

The chosen period is used to compute the Vuong Closeness statistic between the chosen frequency model and the constant and one sidereal day model (a sinusoidal model with a frequency of t_{sid}^{-1} cycles/day). These statistics are used to describe the significance of the chosen period relative to a purely non-periodic model as well as comparing any periodicity detected to the one sidereal day dominant spurious period. A significant periodic signal may produce a high value against the constant model, but will have a much lower value in the one day model if it is due to a sampling periodicity. A real astrophysical signal would be expected to be significant in both of these statistics. GRAPE ultimately returns for a given light curve, an optimised period which has been checked for multiplicity and aliasing with a sinusoidal model and two Vuong closeness test statistics for the sinusoidal model of this optimised period calculated against a constant signal model and a sinusoidal signal with a period of the sidereal day. No confidence margins are applied to the Vuong closeness statistics although they can be generated from them.

3.5 Experimental Design

We presume that GRAPE should exhibit improved performance over a standard frequency spectrum approach due to the treatment of period as a continuous variable. In terms of the processing time for each light curve, GRAPE should require less calculations than a frequency spectrum as only newly computed candidate frequencies must be evaluated and low interest areas of the period space can be avoided.

We designed an experiment using simulated light curves in order to compare the performance of GRAPE against the traditional BGLS, with a dense frequency spectrum, periodogram. As our method is designed as a component of a classification pipeline for the Skycam instruments, we decided to produce two sets of light curves. The first set uses the variable Skycam cadence produced by an instrument with no control over the movement of the parent telescope. The second simulates a more traditional cadence with seasonal gaps added to reproduce light curves similar to standard surveys. We started with the generation of the Skycam cadence time instants. We accessed the SkycamT database and randomly selected 1000 light curves, of which 250 had $100 \leqslant n < 200$, 250 had $200 \leqslant n < 500$, 250 had $500 \leqslant n < 1000$ and the final 250 had $1000 \leqslant n < 2000$, where n is the number of data points in each light curve. This was chosen to generate simulated light curves with a wide range of entries but with a median much closer to low values, a statistical trait of the Skycam survey. The time instants of these light curves recorded in Modified Julian Date (MJD) are recorded as the Skycam-cadence set. To generate the regular-cadence set we took the minimum and maximum time instant value for each light curve in the Skycam-cadence set and generated a linear grid of time instants with a separation of 0.1 days for each light curve. This grid of time instants was phased with a period of 365.25 days and only time instants with an associated phase of $0.0 \leqslant ph < 0.5$ are retained. Finally, n of these time instants are randomly selected where n is the number of data points in the Skycam-cadence time instants for each of the 1000 light curves. We then generated 1000 periods for these two sets of 1000 time instant vectors. We used a uniform random number generator with the function shown in Equation 3.21.

$$P_i = 10^{runif \times (\log_{10}(P_{\max}) - \log_{10}(P_{\min})) + \log_{10}(P_{\min})} \tag{3.21}$$

where P_i is test period i, rerun 1000 times for $i = 1, ..., 1000$ to obtain $\{P\}$, the set of test periods, $runif \in \mathbb{R} \in [0, 1]$ is a uniform random number generator, $P_{min} = 0.05$, $P_{max} = 1000$ are minimum and maximum periods in the period space we wish to optimise. We use a logarithmic projection to skew the period distribution to low periods of which there are more known object classes (Debosscher et al., 2007; Richards et al., 2011b, 2012). With the two lists of 1000 light curve time instant vectors and 1000 simulated periods, we may now generate light curves of various shapes to test our method.

We chose to generate light curves of four different shapes, sinusoidal, sawtooth, symmetric eclipsing binary and eccentric eclipsing binary. This resulted in 2000 light curves of each shape, one with Skycam cadence and one with regular cadence resulting in 8000 total light curves to test with the two algorithms. The light curves are populated with

Gaussian white noise using Equation 3.22.

$$A_s = \sigma_n \sqrt{2(\mathrm{SNR})} \tag{3.22}$$

where A_s is the sinusoidal amplitude of the synthetic signal, σ_n is the standard deviation of the white noise, and SNR is the desired Signal-to-Noise ratio. For this experiment all the light curves are generated with a SNR of 2 which is determined as a 0.4 unit amplitude to a 0.2 unit standard deviation of white noise. The sinusoidal light curves were generated using the time instants and the associated simulated period using Equation 3.23.

$$y_i = A_s sin\left(\frac{2\pi t_i}{P}\right) + \sigma_n \epsilon_i \tag{3.23}$$

where y_i is the magnitude value of data point i, t_i is the time instant of data point i, ϵ_i is a normal distributed error value for data point i with a mean of 0 and a standard deviation of 1 and P is the simulated period. A_s and σ_n are as above. Sawtooth light curves use the function shown in Equation 3.24.

$$y_i = 2A_s\left(\frac{t_i}{P} - \left\lfloor \frac{t_i}{P} \right\rfloor\right) + \sigma_n \epsilon_i \tag{3.24}$$

where $\lfloor x \rfloor$ is the closest integer to x rounded down, the 'floor' of x. For the eclipsing binary light curves, we decided to keep transit duration and shape at constant phases, i.e. they are a linear function of the underlying period. We concede this is not a perfectly physical representation as the parameters of two binary stars and their orbital properties allow many different possible eclipse shapes (Prsa et al., 2008; Paegert et al., 2014) however, we wished to test GRAPE on eclipse light curve shape independent of period. Research has been conducted into the performance of eclipsing binary detection with alternate eclipse shapes and periods (Prsa et al., 2011; Wells et al., 2017). We follow a similar process to generate both symmetric and eccentric eclipsing binary simulated light curves. First we populate the light curve with a constant signal with white noise using Equation 3.25.

$$y_i = \sigma_n \epsilon_i \tag{3.25}$$

We then phase the simulated data points around the period. Data points between the phases $0.0 \leqslant ph_i \leqslant 0.1$ and $0.9 \leqslant ph_i \leqslant 1.0$ are located within the primary eclipse and have a triangular subtraction applied with depth of $2A_s$ and of base length phase of 0.2. After these operations, the symmetric and eccentric light curve method diverges. The symmetric light curves have a secondary eclipse located at $ph = 0.5$ with a triangular subtraction of depth A_s and a base length phase of 0.1. The eccentric light curves have a secondary eclipse of identical size but centred at $ph = 0.7$. Figure 3.5 demonstrates the phased light curves of these four signal shapes.

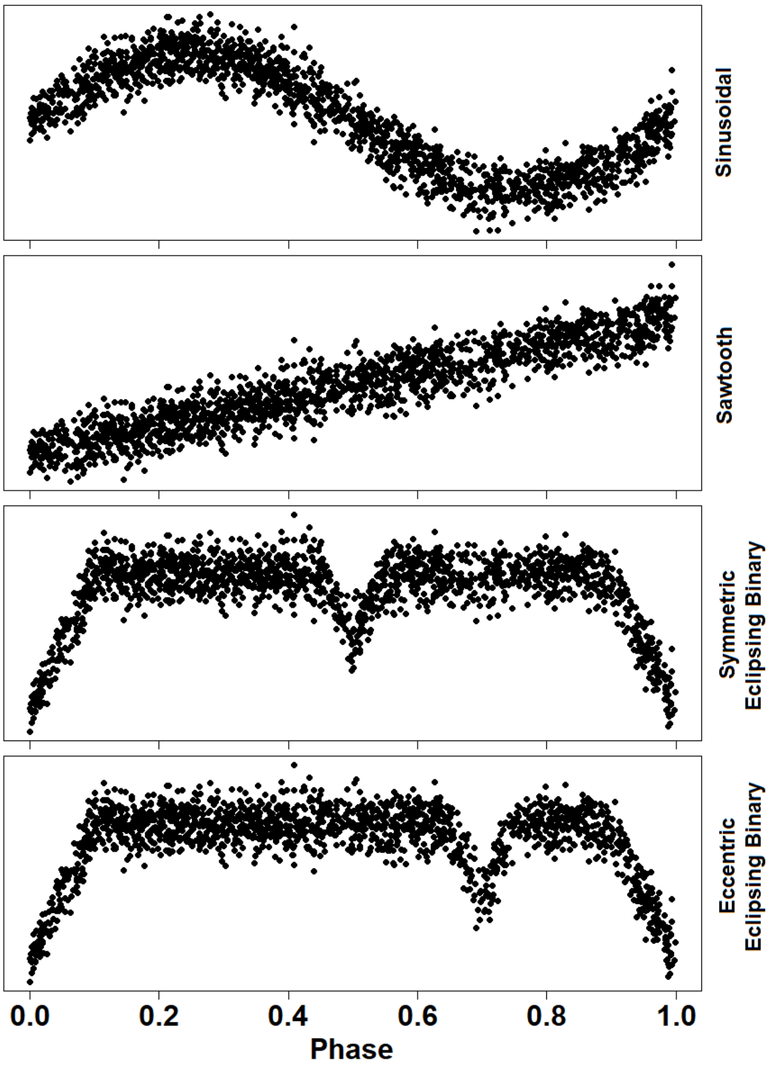


FIGURE 3.5: Phased simulated light curves of a sinusoidal, sawtooth, symmetric eclipsing binary and eccentric eclipsing binary.

The results of GRAPE and the BGLS are determined by taking the input period and the estimated period and computing if the relationship is one of six possible types: a hit, a multiple, a submultiple, a one-day alias, a half-day alias or an unknown mode. A hit is when the estimated period matches the estimated period to within a tolerance and is true if it satisfies the inequality shown in Equation 3.26.

$$|P_i - P_e| < \epsilon P_i \tag{3.26}$$

where P_i is the input period, P_e is the estimated period and ϵ is the tolerance. A multiple is a realistic integer multiple of the input period and is defined as any relationship which does not satisfy the hit inequality, satisfies $P_e > P_i$, and satisifies the inequalities in either Equation 3.27 or 3.28.

$$\left\lfloor \frac{P_e}{P_i} \right\rfloor \leqslant 3 \qquad \text{and} \qquad \left| \frac{P_e}{P_i} - \left\lfloor \frac{P_e}{P_i} \right\rfloor \right| < \epsilon \tag{3.27}$$

$$\left\lceil \frac{P_e}{P_i} \right\rceil \leqslant 4 \qquad \text{and} \qquad \left| \frac{P_e}{P_i} - \left\lceil \frac{P_e}{P_i} \right\rceil \right| < \epsilon \tag{3.28}$$

where $\lceil x \rceil$ is the closest integer to x rounded up, the 'ceiling' of x. A submultiple is similar to the multiple and is a realistic integer division of the input period and is defined as any relationship which does not satisfy the hit inequality, satisfies $P_e < P_i$, and satisifies the inequalities in either Equation 3.29 or 3.30.

$$\left\lfloor \frac{P_i}{P_e} \right\rfloor \leqslant 3 \qquad \text{and} \qquad \left| \frac{P_i}{P_e} - \left\lfloor \frac{P_i}{P_e} \right\rfloor \right| < \epsilon \tag{3.29}$$

$$\left\lceil \frac{P_i}{P_e} \right\rceil \leqslant 4 \qquad \text{and} \qquad \left| \frac{P_i}{P_e} - \left\lceil \frac{P_i}{P_e} \right\rceil \right| < \epsilon \tag{3.30}$$

If the estimated period is the one day alias of the input period, the inequality shown in Equation 3.31 is satisfied.

$$\left| \left| \frac{P_i}{1 \pm P_i} \right| - P_e \right| < \epsilon \tag{3.31}$$

The presence of a half-day alias can be determined using a similar inequality shown in Equation 3.32.

$$\left| \left| \frac{P_i}{1 \pm 2P_i} \right| - P_e \right| < \epsilon \tag{3.32}$$

Any light curve period estimation task where P_i and P_e do not satisfy any of the above inequalities are declared unknown failures.

3.6 GRAPE Evaluation

Using the synthetic data we performed a set of experiments to test the performance of GRAPE relative to the BGLS periodogram and the computational complexity of these operations. Additionally, the dependence of the performance of GRAPE on the period of the light curve is tested as is the relative performance between the regular cadence and Skycam cadence light curves. The final experiment involves the implementation of a fine-tuning frequency grid on the BGLS periodogram to attempt to replicate the period resolution of GRAPE using a more traditional frequency spectrum.

3.6.1 Period estimation performance

GRAPE ran the data set with 3 different dominant seeds $\text{Seed}_D = [1, 2, 3]$ to screen out poor convergence events with the Vuong closeness test with the submissive seed set to $\text{Seed}_S = \text{Seed}_D + 100$. The input arguments were as follows: $N_{\text{pop}} = 200$, $N_{\text{pairups}} = 50$, $N_{\text{gen}} = 100$, $N_{\text{finegen}} = 50$, $P_{\text{crossover}} = 0.65$, $P_{\text{mutation}} = 0.8 - 0.008i$, where i is the generation, $P_{\text{fdif}} = 0.6$ and $P_{\text{dfrac}} = 0.7$. These values were established by a grid-search cross-validation operation on a stratified subset of 100 synthetic light curves although the sinusoidal light curves indicated a lower gradient of $P_{\text{mutation}} = 0.8 - 0.003i$ on the mutation rate. This was determined to be a result of the selection of all N_t trial periods in the same period region as the true period. The number of candidate periods selected by the GRAPE genetic clustering method is $N_t = 5$ which are then analysed by the Vuong Closeness test. This only applies in situations where the signal is purely sinusoidal with Gaussian noise and was therefore rejected. The linear decay gradient on the mutation value has an important effect on exploring the parameter space as well as fine tuning the final period.

The BGLS periodogram is performed by selecting the top N_t independent peak periods with an oversampling factor, which determines the density of the frequency spectrum, of $N_{\text{ofac}} = 5$. The Vuong Closeness test is then applied to these $N_t = 5$ peaks as well as their multiples and aliases in a similar operation to the one performed on the GRAPE candidate periods. The best performing period model is selected as the BGLS periodogram final period. There are also a number of shared arguments for the BGLS fitness function between both GRAPE and the periodogram. The white noise jitter, which tunes the probability response for candidate periods, $\text{jit} = 0.4 \cdot A_{\text{lc}}$ where A_{lc} is the estimated amplitude of the light curve determined by Equation 3.33.

$$A_{\text{lc}} = \frac{|y_{max} - y_{min}|}{2} \tag{3.33}$$

where y_{min} and y_{max} are the minimum and maximum values of the measurement unit for a given light curve. The period space is $P_{min} = 0.05\,\text{days}$ to $P_{max} = (t_{max} - t_{min})\,\text{days}$ where t_{min} and t_{max} are the minimum and maximum time instants for a given light curve. The Gaussian filter for spurious period removal is left unused.

The confidence intervals of the GRAPE and BGLS periodogram performances on the synthetic light curves is determined using 100,000 bootstrapped samples from the 3000 GRAPE light curve period estimations (all 3 seeds on the 1000 synthetic light curves) and 1000 BGLS periodogram period estimations. In this bootstrapping operation the results are resampled with replacement 100,000 times and used to compute a mean performance and the 95% confidence intervals through the selection of the 5^{th} and 95^{th}

Table 3.1: Period estimation results on regular cadence simulated light curves with a tolerance $\epsilon = 0.01$. The confidentally best performing method for each light curve shape is emboldened.

Method	Type	Hit	Multiple	Alias
GRAPE	**Sinusoidal**	**0.831 ± 0.011**	**0.001 ± 0.001**	**0.004 ± 0.002**
BGLS	Sinusoidal	0.703 ± 0.024	0.001 ± 0.002	0.000 ± 0.000
GRAPE	**Sawtooth**	**0.741 ± 0.013**	**0.006 ± 0.002**	**0.006 ± 0.003**
BGLS	Sawtooth	0.696 ± 0.024	0.002 ± 0.003	0.001 ± 0.002
GRAPE	Symmetric EB	0.032 ± 0.005	0.637 ± 0.014	0.024 ± 0.006
BGLS	Symmetric EB	0.031 ± 0.009	0.632 ± 0.025	0.024 ± 0.009
GRAPE	Eccentric EB	0.362 ± 0.014	0.280 ± 0.014	0.011 ± 0.005
BGLS	**Eccentric EB**	**0.426 ± 0.026**	**0.251 ± 0.023**	**0.012 ± 0.007**

Table 3.2: Period estimation results on Skycam cadence simulated light curves with a tolerance $\epsilon = 0.01$. The confidentally best performing method for each light curve shape is emboldened.

Method	Type	Hit	Multiple	Alias
GRAPE	**Sinusoidal**	**0.824 ± 0.011**	**0.000 ± 0.001**	**0.026 ± 0.007**
BGLS	Sinusoidal	0.717 ± 0.023	0.001 ± 0.002	0.030 ± 0.009
GRAPE	Sawtooth	0.527 ± 0.015	0.007 ± 0.003	0.141 ± 0.004
BGLS	Sawtooth	0.508 ± 0.026	0.010 ± 0.005	0.143 ± 0.021
GRAPE	Symmetric EB	0.015 ± 0.004	0.358 ± 0.014	0.075 ± 0.010
BGLS	Symmetric EB	0.021 ± 0.008	0.382 ± 0.025	0.085 ± 0.016
GRAPE	Eccentric EB	0.149 ± 0.011	0.158 ± 0.011	0.112 ± 0.012
BGLS	Eccentric EB	0.159 ± 0.019	0.166 ± 0.020	0.121 ± 0.019

percentile performance values for the hit rate, the multiple rate (sum of the multiples and submultiples) and the aliasing rate (the sum of the one day and half day aliases).

Table 3.1 shows the results of this experiment on the regular cadence four light curve types with an accuracy tolerance of $\epsilon = 0.01$, a one percent allowed error in period. Note, it is possible for the row sum of a light curve to be above 1 as some periods can satisfy both a submultiple and an alias simultaneously and we do not presume which is the mode responsible for this error. Table 3.2 shows the performance of the two methods on the Skycam cadence light curves.

GRAPE clearly outperforms the BGLS periodogram in the correct period estimation of the sinusoidal light curves with a hit rate improvement of 10%. Sawtooth light curves performed better on GRAPE by 2% for the regular cadence light curves and similarly on the Skycam cadence light curves as well as identifying some poorer quality light curves as aliases. The symmetric and eccentric eclipsing binaries suffer from a significant submultiple failure mode. This is a well understood result of the Lomb-Scargle method and its extensions. GRAPE maintains the performance of the BGLS periodogram within the statistical confidence levels for the symmetric eclipsing binaries. The eccentric eclipsing binaries were the worst performing shape on both GRAPE and

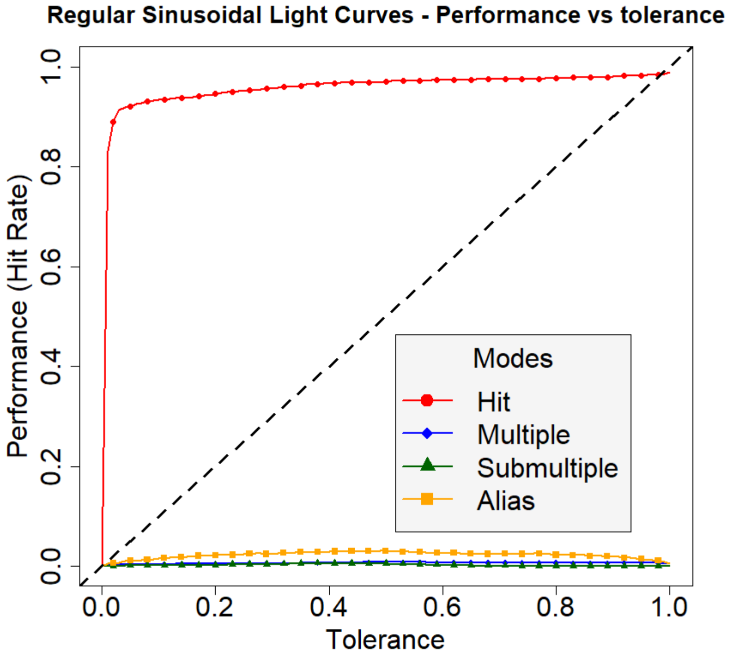


FIGURE 3.6: Performance vs Tolerance for the regular cadence sinusoidal light curves. The hit rate rapidly rises to nearly perfect very quickly showing GRAPE can effectively fine tune sinusoidal periods.

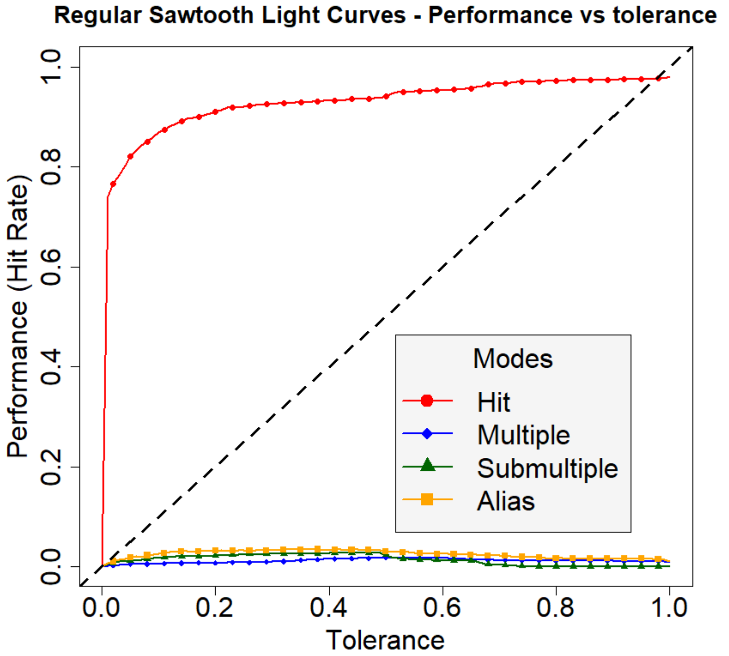


FIGURE 3.7: Performance vs Tolerance for the regular cadence sawtooth light curves. The hit rate rise is similar to the sinusoidal light curves but with an increased number of multiple and aliased periods.

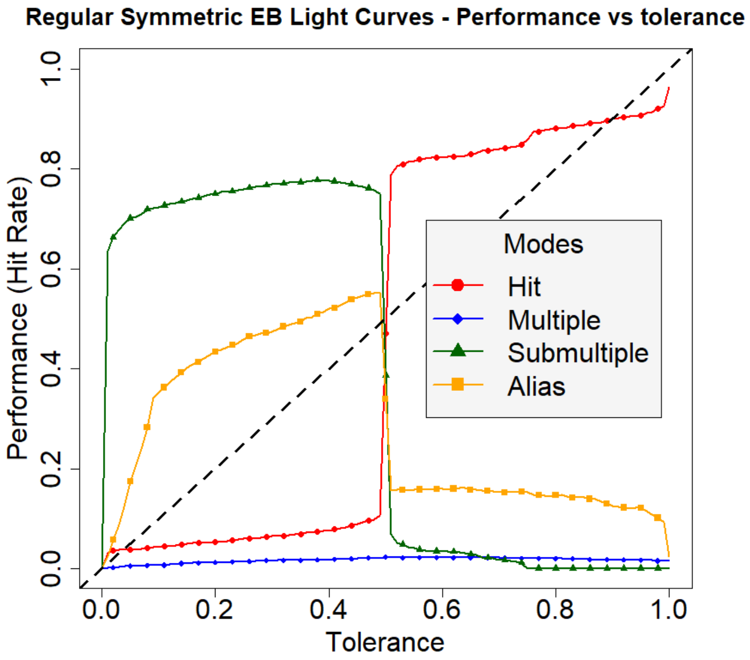


FIGURE 3.8: Performance vs Tolerance for the regular cadence symmetric eclipsing binaries. The $N = 2$ submultiple is the dominant failure mode until the tolerance reaches $\epsilon = 0.5$. Here, the submultiple satisfies the hit inequality and causes the seen performance swap. This failure mode is caused by the secondary eclipse at phase 0.5. The sinusoidal model provides a better hit at half the true period through the combined eclipses as it is unable to sufficiently model the differentially sized eclipses at the true period.

the BGLS periodogram as they were the least sinusoidal of the light curves and therefore the BGLS fitness function is not tuned for them. The periodogram slightly outperformed GRAPE for this shape of light curve. This is likely due to the importance of the fitness function in the propagation of genetic information. For eccentric binaries, the response from the correct period did not outperform a poor period by a significant margin and therefore it was never optimised heavily, whereas the frequency spectrum would sample a period close to the true period by default. These failings are not a required disadvantage of GRAPE. It is entirely possible to replace the fitness function with another more appropriate to the required task and achieve the same performance increase as the sinusoidal light curves did with the BGLS fitness function. Ultimately, this experiment shows that GRAPE exhibits a similar performance to the BGLS periodogram with Vuong Closeness for every light curve shape other than purely sinusoidal. This demonstrates that the performance of the method is primarily driven by the Vuong Closeness test and its ability to distinguish between hit, multiple and aliased models. The frequency spectrum approach did not appear to result in the expected loss of performance. It is likely that the use of the multiple models in the Vuong Closeness test corrected for this weakness.

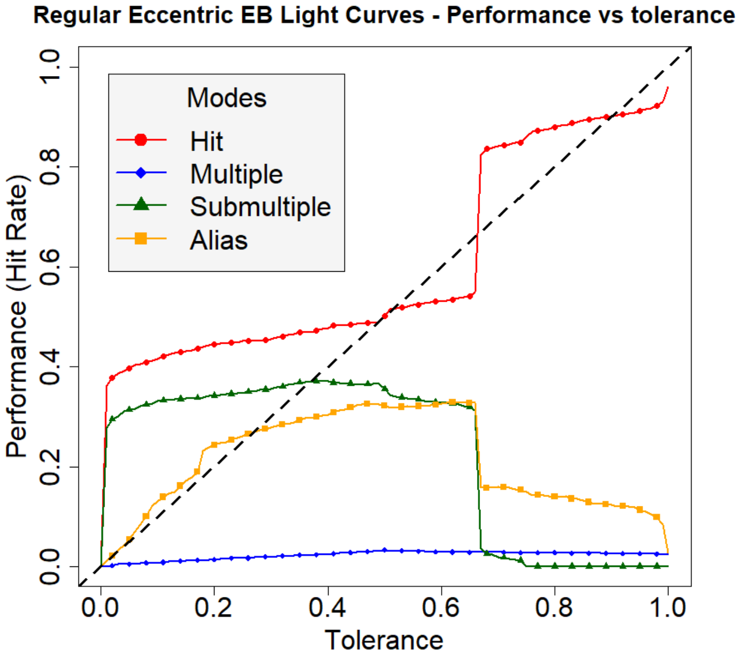


FIGURE 3.9: Performance vs Tolerance for the regular cadence eccentric eclipsing binaries. Here about half the non-aliased light curves are hits and $N = 3$ submultiples. This is a result of the eccentric secondary eclipse being at 0.7, close to 0.667. Therefore, depending on the sampling, some light curves appear as a third of the period with a missing third eclipse or else the eccentricity of the phased light curve results in the best fit being at the true period. This is only expected to happen for secondary eclipse phases near 0.333 and 0.667 as half the eclipses would be missing for the $N = 4$ submultiple variant at phases 0.25 or 0.75 which would prevent period matching at a quarter of the true period.

We also decided to investigate the tolerance of the two methods. For the previous experiment we produced results with a tolerance $\epsilon = 0.01$. It is desirable to achieve as high a tolerance as possible but as GRAPE has been designed for use in a classification pipeline, it is likely we can still generate informative features of a light curve even if the error on the period estimation is higher than expected. Figures 3.6 to 3.9 show plots of the tolerance of $0.0 \leqslant \epsilon \leqslant 1.0$ against the recovered rate of hits, multiples, submultiples and aliases for the four different light curve shapes with the regular cadence. The Skycam cadence light curves show a much higher incidence rate of aliases due to the poorer phase sampling of some periods.

Due to the random seeds used in GRAPE, we repeated the experiment with three different seed values $\text{Seed}_D = [1, 2, 3]$ to see how the performance varied with the random processes inside a genetic algorithm. Table 3.3 shows the performance of the $\text{Seed}_D = 1$ period estimation on the 1000 synthetic light curves with table 3.4 showing $\text{Seed}_D = 2$ performance and table 3.5 showing $\text{Seed}_D = 3$ performance. The reported errors are the 95% confidence intervals as computed by a bootstrapping confidence estimator with

Table 3.3: GRAPE period estimation results with a seed of 1 on regular cadence simulated light curves with a tolerance $\epsilon = 0.01$.

Type	Hit	Multiple	Alias
Sinusoidal	0.833 ± 0.019	0.001 ± 0.002	0.000 ± 0.000
Sawtooth	0.754 ± 0.023	0.004 ± 0.004	0.004 ± 0.003
Symmetric EB	0.026 ± 0.009	0.647 ± 0.025	0.025 ± 0.010
Eccentric EB	0.371 ± 0.025	0.279 ± 0.024	0.007 ± 0.005

Table 3.4: GRAPE period estimation results with a seed of 2 on regular cadence simulated light curves with a tolerance $\epsilon = 0.01$.

Type	Hit	Multiple	Alias
Sinusoidal	0.823 ± 0.020	0.001 ± 0.002	0.009 ± 0.006
Sawtooth	0.727 ± 0.023	0.009 ± 0.005	0.004 ± 0.004
Symmetric EB	0.037 ± 0.010	0.619 ± 0.025	0.026 ± 0.009
Eccentric EB	0.352 ± 0.025	0.286 ± 0.024	0.010 ± 0.007

Table 3.5: GRAPE period estimation results with a seed of 3 on regular cadence simulated light curves with a tolerance $\epsilon = 0.01$.

Type	Hit	Multiple	Alias
Sinusoidal	0.838 ± 0.019	0.001 ± 0.002	0.004 ± 0.004
Sawtooth	0.741 ± 0.023	0.005 ± 0.004	0.011 ± 0.007
Symmetric EB	0.032 ± 0.009	0.644 ± 0.025	0.021 ± 0.008
Eccentric EB	0.364 ± 0.025	0.274 ± 0.023	0.017 ± 0.009

100,000 resamples. For the modes with sufficient population to determine an accurate confidence interval, the confidence intervals indicate that the results of GRAPE are consistent over a large set of light curves. Whilst the performance of an individual light curve can vary depending on seed, the overall percentage of matched light curves should remain consistent for a large dataset. For important individual light curves, GRAPE can be rerun with different seeds or alternatively, with a larger set of returned candidate periods as this will reduce the chance that a good trial period is not detected and evaluated.

We also computed a stratified subset of 400 of the light curves, 100 for each shape. This stratified set was generated two more times in addition to the previous set resulting in three different additive noise generations to determine the consistency of the period estimation performance. This was performed for the Skycam cadence light curves. The confidence intervals are again the 95% confidence intervals as determined from a bootstrapping estimation using 100,000 resamples. The results are shown in table 3.6 with the Data column indicating the period estimation algorithm and the additive noise seed number $[10, 20, 30]$ utilised in the bootstrap. As with the GRAPE seed performance, the different noise models are consistent to the 95% confidence intervals. Therefore, we suggest that the performance of GRAPE and the BGLS periodogram are consistent

Table 3.6: GRAPE and BGLS periodogram period estimation results with three different light curve additive noise models on Skycam cadence simulated light curves with a tolerance $\epsilon = 0.01$.

Data	Hit	Multiple	Alias
GRAPE lcseed 10	0.348 ± 0.040	0.128 ± 0.028	0.073 ± 0.023
GRAPE lcseed 20	0.363 ± 0.040	0.120 ± 0.028	0.098 ± 0.023
GRAPE lcseed 30	0.365 ± 0.040	0.103 ± 0.025	0.080 ± 0.023
BGLS lcseed 10	0.310 ± 0.038	0.128 ± 0.028	0.090 ± 0.028
BGLS lcseed 20	0.313 ± 0.038	0.110 ± 0.025	0.090 ± 0.025
BGLS lcseed 30	0.335 ± 0.040	0.103 ± 0.025	0.095 ± 0.030

for large datasets independent of the random seeds used for either noise generation or genetic algorithm operation.

3.6.2 Performance vs Period

Our next experiment is to determine if the failure modes of GRAPE and the BGLS periodogram have a dependence on the period. In the frequency spectrum case of the BGLS periodogram, we would expect to see the performance of the periodogram be a function of the candidate period. We therefore plot two log-log period diagrams of the GRAPE estimated periods and the periodogram estimated periods on the sinusoidal light curves and is displayed in figure 3.10. The plots clearly show that both methods perform more poorly the closer the period is to P_{max}. This is due to poorer sampling of the phase space as there are less complete cycles viewed inside of the light curve baseline. Additionally, the BGLS periodogram suffers a much greater performance loss at this extreme. This is likely due to a combination of the frequency spectrum selecting a submultiple of the true period followed by a Vuong Closeness test correction. As this correction must be an integer multiple of the initially detected period, this introduces an error near P_{max}. This can be seen in figure 3.10 as a selection of light curves with BGLS periodogram estimated periods between 1000 and 1200 days. GRAPE performs much better in this range due to treating the parameter space as a continuous variable whereas the periodogram samples the high periods extremely sparsely and thus GRAPE does not need to rely on the Vuong Closeness test to correct as many long periods. Figure 3.11 demonstrate the same plots for the Skycam cadence sinusoidal light curves. The performance on the Skycam cadence sinusoidal light curves was similar to the regular cadence sinusoidal light curves except with a larger instance of aliased periods due to poor sampling. This is an interesting result and demonstrates that the Skycam sampling does not adversely affect the performance of the period estimation when the fitness function is a good match to the data. Figure 3.12 show the same plots for the regular cadence sawtooth light curves and figure 3.13 show the plots of the Skycam cadence

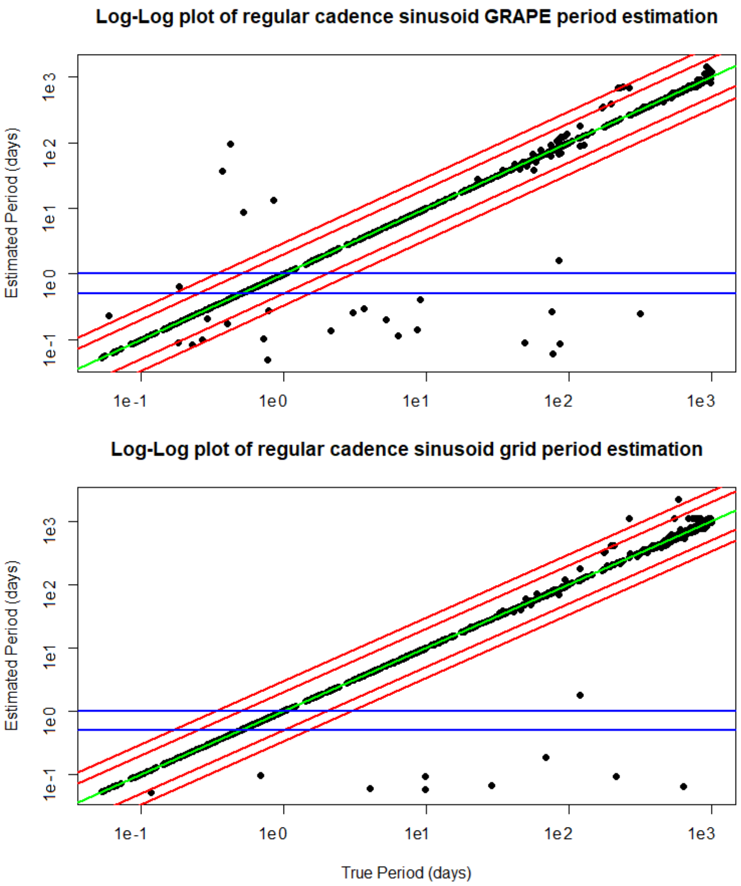


FIGURE 3.10: Log Period vs Log Period plots for the regular cadence sinusoidal light curves. For the sinusoidal light curves, performance was good from both fitness functions. GRAPE performs better at longer periods seen by the increased dispersion on the frequency grid plot due to the frequency spectrum sampling this region poorly.

sawtooth light curves. The same effects can be seen but with additional extremely long periods found by GRAPE and the BGLS periodogram due to the Vuong closeness test deciding to pick a multiple of the identified period for many light curves with periods between 850-1000 days. This effect appears worse in the regular cadence light curves with the Skycam cadence light curves showing a higher rate of aliases instead. This is likely due to the tested model being a simple sinusoid which does not fit the sawtooth signal combined with a minimal number of observed cycles. A more generalised model for the Vuong closeness test would be desirable for this event especially if a different non-sinusoidal fitness function is selected.

The symmetric eclipsing binary log-log period plots in figures 3.14 and 3.15 show similar errors to the sawtooth light curves with poor Vuong Closeness estimating periods outside of the long period range. The periods are highly underestimated above 500 days due to the common $N = 2$ submultiple failure mode of symmetric eclipsing binaries. The eccentric eclipsing binaries are split almost equally between hits and the $N = 3$

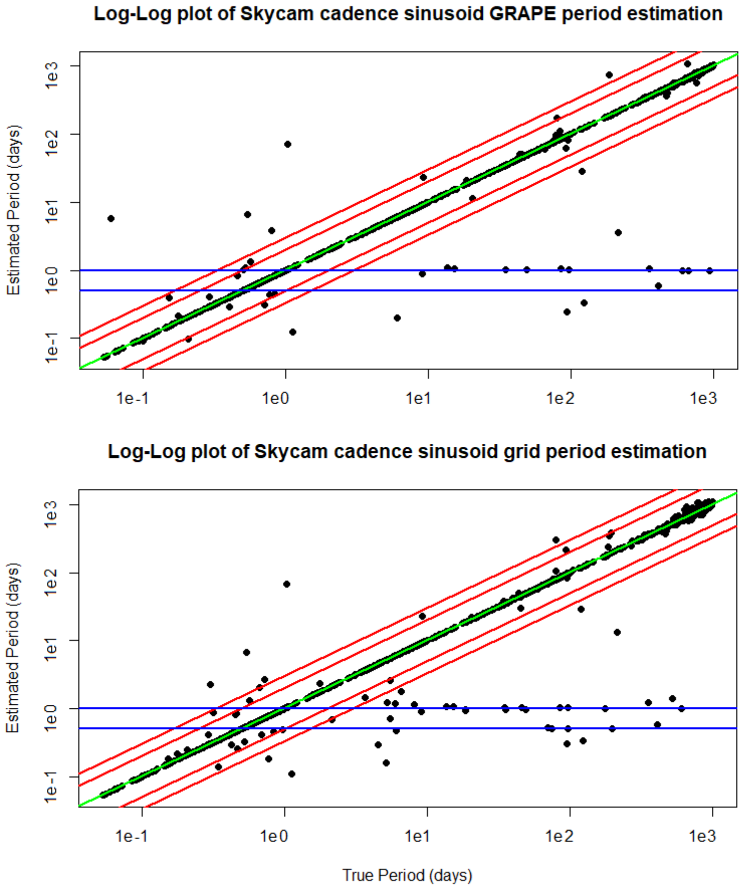


Figure 3.11: Log Period vs Log Period plots for the Skycam cadence sinusoidal light curves. The performance on the Skycam cadence sinusoids is similar to that of the regular cadence sinusoids. This is an interesting result as it indicates that the uneven sampling of Skycam may not be significantly detrimental to continuous variations.

submultiple showing a moderate long period depletion. The eccentric eclipsing binaries perform extremely poorly at long periods in both regular and Skycam cadence as seen in figures 3.16 and 3.17. This is a result of many of the long periods exhibit the $N = 3$ failure mode due to the small number of sampled eclipses. Additionally, at long periods the phase sampling often fails to measure the eclipse at all due to the seasonal sampling windows we have added to our light curves, especially for the Skycam cadence. This is a common issue in real survey eclipsing binaries (Prsa et al., 2011; Wells et al., 2017) and arguably a bigger problem as our simulated light curves have unphysically large eclipse durations at long periods which decrease the probability that the eclipse will be missed. It is interesting to note that the Vuong Closeness test long period multiplication failure mode does not seem to occur on the Skycam cadence light curves. This is possibly due to the poor long duration sampling producing poor quality models that do not improve over the initial period estimation model. Alternatively, the alias of the period might be selected instead due to the strong Skycam aliases.

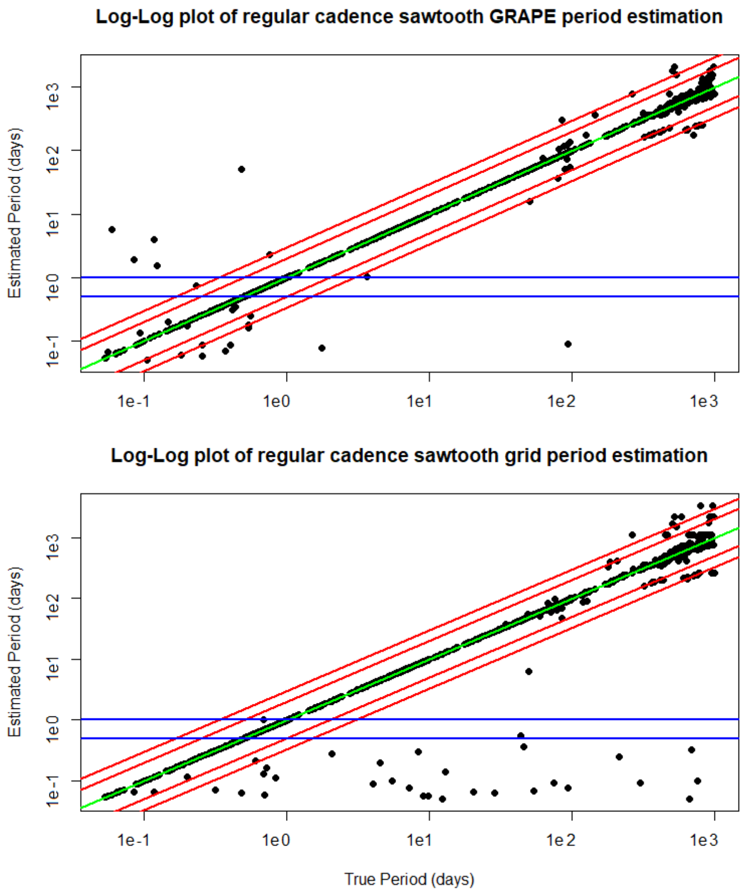


FIGURE 3.12: Log Period vs Log Period plots for the regular cadence sawtooth light curves. There are many spurious detections near the time span of the light curves for this test despite being on regular cadence as the sawtooth shape diverges from the expectations of the BGLS fitness function.

3.6.3 Regular vs STILT cadence

The error in the estimated periods against the period span is an excellent indicator of the performance of the GRAPE period estimation due to cadence against the simulated period. Period span is defined as the baseline of the light curves divided by the input period and is the number of cycles present in the light curve. It is calculated by Equation 3.34.

$$P_{span} = \frac{t_{max} - t_{min}}{P_i} \tag{3.34}$$

where t_{min} is the minimum time instant of the light curve, t_{max} is the maximum time instant of the light curve and P_i is the input period. For low values of P_{span}, the performance is expected to be poorer as there are less complete cycles and the cadence results in unsampled regions of the phase space. Large performance errors occurring where $P_{span} = t_{max} - t_{min}$ also indicate that the cadence is resulting in substantial aliasing. The estimated error is calculated by Equation 3.35 and is the fractional error

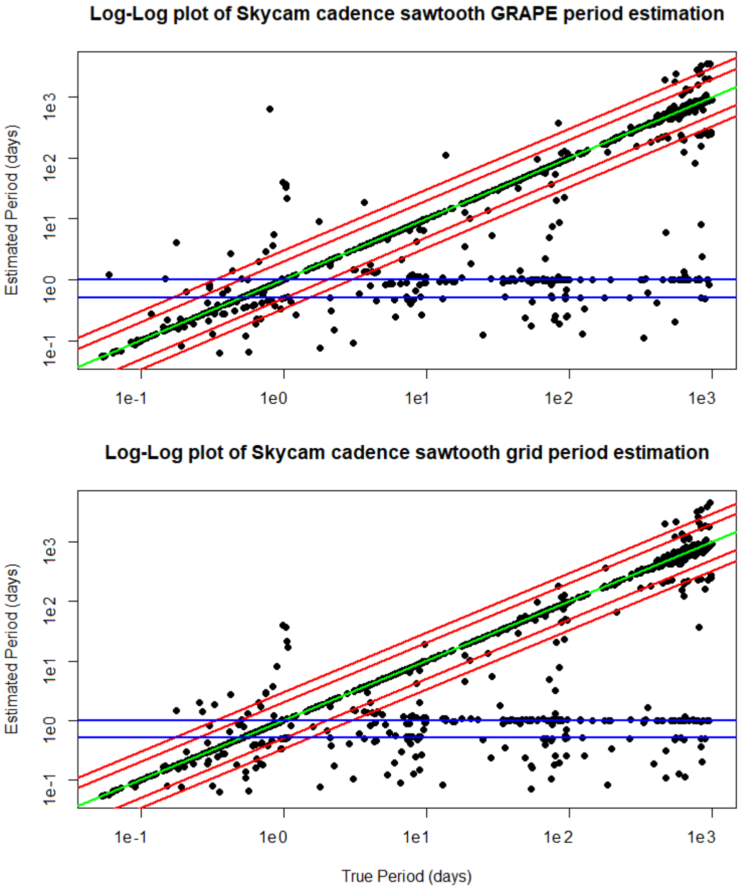


Figure 3.13: Log Period vs Log Period plots for the Skycam cadence sawtooth light curves. The performance of the Skycam cadence sawtooth light curves appears similar to the regular cadence. There is a noticeable dispersion near 1000 days likely due to sampling difficulties with a single sawtooth variation.

from the input period.

$$\xi = \frac{|P_i - P_e|}{P_i} \tag{3.35}$$

Figure 3.18 shows this plot for the regular cadence light curves of the four types and figure 3.19 shows the results of GRAPE on the Skycam cadence light curves. Symmetric eclipsing binaries estimated periods have been doubled due to the common Lomb-Scargle failure mode. The regular cadence light curves show only two main patterns. The increase in error near $\log_{10}(P_{\mathrm{span}}) = 0$ is due to the poorer sampling of the shape of the light curve. This is less of an issue in the sinusoidal light curves as the model can interpolate the missing data but it becomes progressively a bigger problem as the light curve shape becomes more non-sinusoidal. Missing the eclipses in the eclipsing binary light curves also causes smaller P_{span} values to perform poorly. There is also a number of failed period estimations near $\log_{10}(P_{\mathrm{span}}) = 1$ due to the seasonal sampling periodicity as the Vuong Closeness test incorrectly determines multiples of the period due to unsampled data.

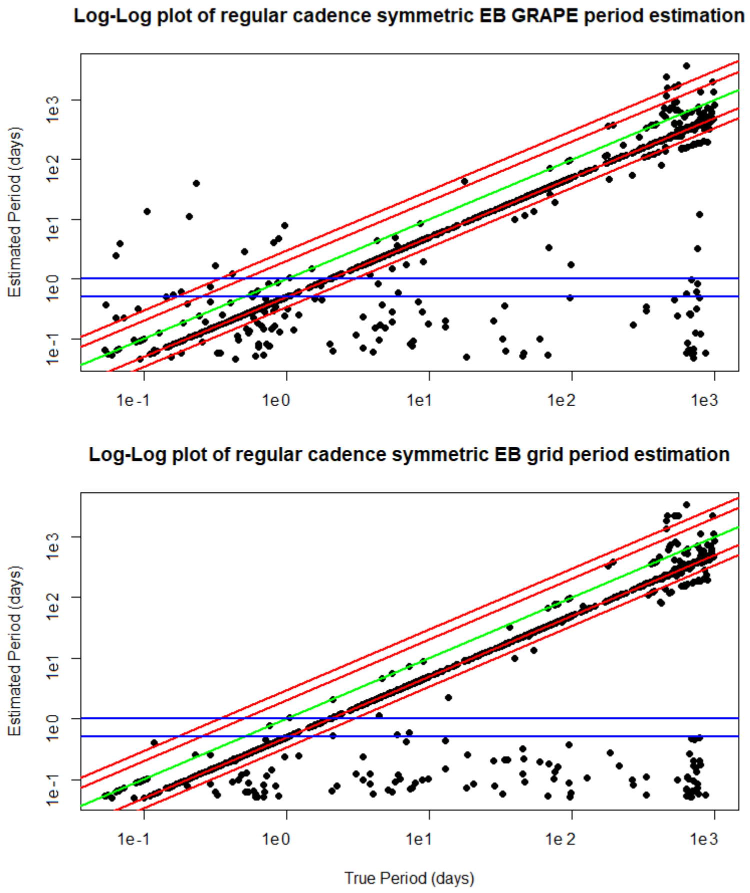


Figure 3.14: Log Period vs Log Period plots for the regular cadence symmetric eclipsing binaries. There is an expected, significant $N = 2$ failure mode due to the fitness function. Some of the longer periods are correctly estimated possibly due to sampling resulting in no data points from the secondary eclipse. In this case, the best fitting sinusoidal model is at the correct period.

The Skycam cadence plot in Figure 3.19 exhibits the same patterns but with an additional uncorrelated set of erroneous period estimations due to the sampling of these light curves being insufficient despite many cycles being present in the baseline. The dominant feature of this failure mode is the increased period estimation errors near $\log_{10}(P_{\text{span}}) = 3$ which is close to the sidereal day spurious period. The light curves of this additional failure mode also increases in number as the underlying signal becomes more non-sinusoidal. It is clear that the Skycam sampling appears inferior to the regular cadence yet this analysis can be used to augment the spurious and alias correction steps present in GRAPE. Ultimately, the advantage in Skycams capability of performing survey astronomy independently of the operations of the Liverpool Telescope is met with the significant disadvantage that many variable objects observed by the instruments will not be sampled at sufficient quality to detect.

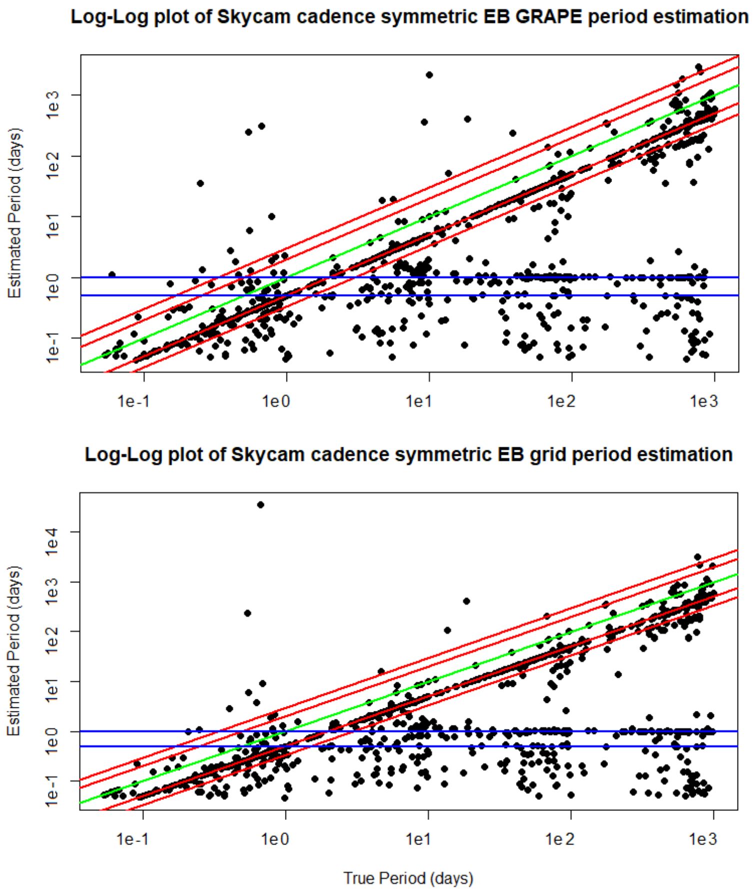
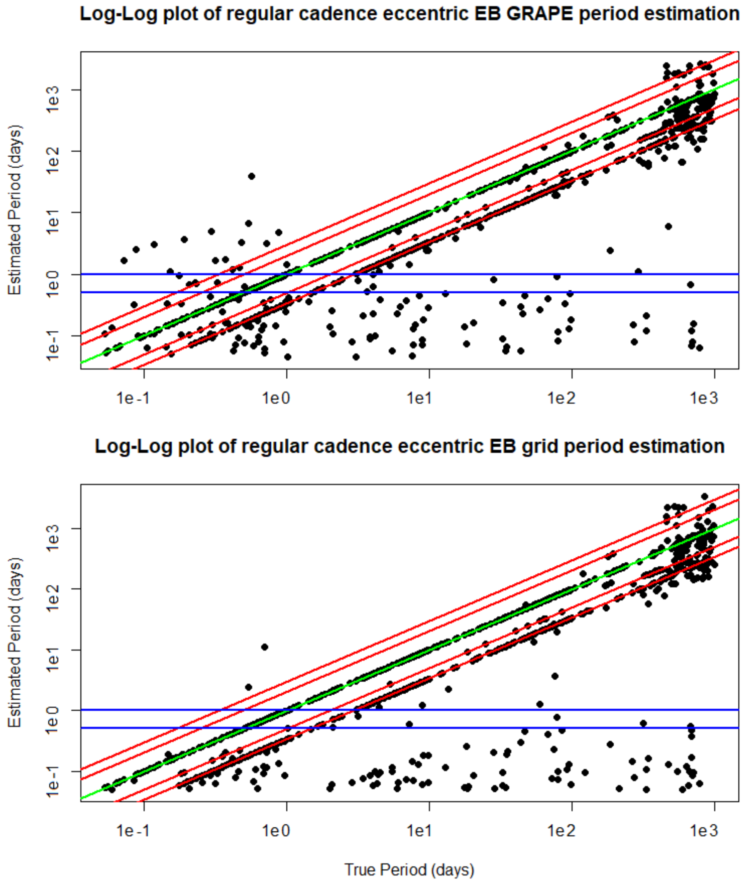


FIGURE 3.15: Log Period vs Log Period plots for the Skycam cadence symmetric eclipsing binaries. The Skycam cadence light curves have similar performance to the regular cadence light curves except for additional long period hits. This is likely due to the even greater chance of losing either eclipse at long periods due to the sampling rate of Skycam cadence compared to the regular cadence.

3.6.4 Runtime considerations

The previous experiments have shown the GRAPE performance to be consistent with the BGLS periodogram and even improved for signals close to the modelled fitness function. This fulfils the first major requirement for this method as an application to the Skycam survey data. The second requirement, which is also the original driving force behind the development of GRAPE, is the requirement for the period estimation task to be as computationally efficient as possible. Many of the important properties of the Skycam light curves are extracted from statistics which are a function of the candidate period and therefore require the estimation of a period prior to calculation. For large numbers of light curves this calculation must be as rapid as possible whilst maintaining performance. To understand the runtime requirements of GRAPE compared to a periodogram approach we calculated the average light curve processing time by calculating


Log-Log plot of regular cadence eccentric EB GRAPE period estimation

Log-Log plot of regular cadence eccentric EB grid period estimation

Figure 3.16: Log Period vs Log Period plots for the regular cadence eccentric eclipsing binaries. Many of the light curves have been underestimated into the $N = 3$ submultiple. About half of the period estimates are correct and the other half are in this $N = 3$ submulitple failure mode.

the mean runtime for the stratified set of 400 regular cadence light curves and 400 Skycam cadence light curves used in the testing of the additive noise models. These light curves are separated by the number of data points they contain based on the intervals selected when generating the synthetic data. The experiments were run on a virtual machine with 4 cores of an Intel Xeon E5-2670 CPU and 12 GB of RAM.

Figure 3.20 shows the mean runtime in seconds of the groups of light curves against the binned number of data points for the regular cadence dataset seperated by light curve shape and period estimation method. A number of interesting patterns emerges in this result. Firstly, the periodogram has an exponential dependency on the number of data points compared to GRAPE. This is expected as the periodogram algorithm is an $O(N^2)$ method whereas GRAPE, whilst having an additional overhead due to the generational updates which are not a function of data point number, is an $O(N)$ method. As both methods use the Vuong Closeness test, its contribution is present in both operations. The effect of Vuong Closeness is visible in the form of the runtime separation of the different

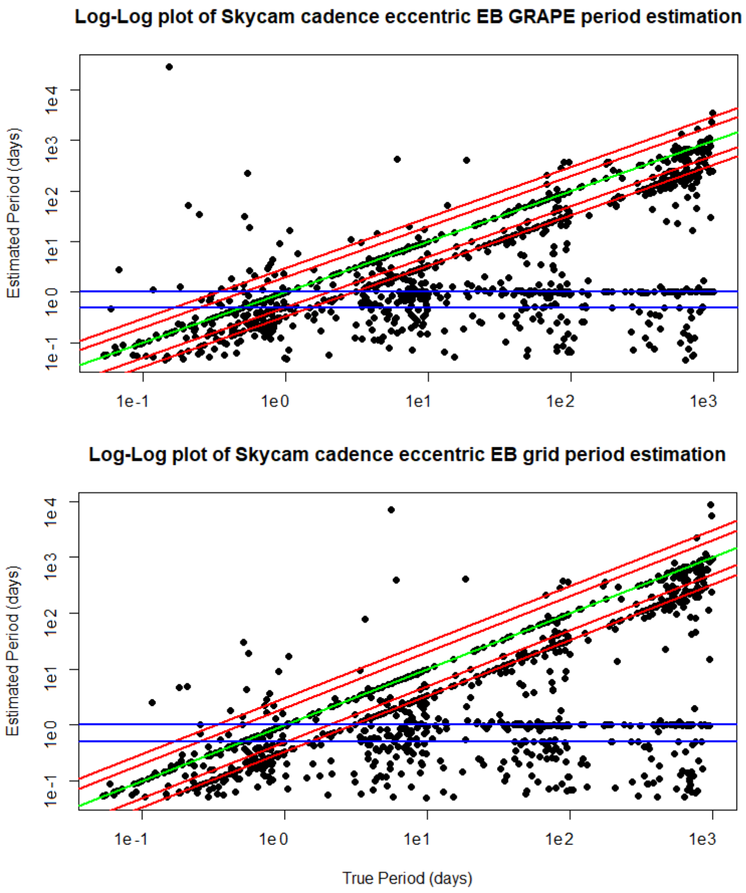


Figure 3.17: Log Period vs Log Period plots for the Skycam cadence eccentric eclipsing binaries. The poor performance at detecting many of the long period light curves is due to limits in the BGLS fitness function at fitting eccentric eclipsing binaries. Additionally, the Skycam cadence results in many of the eclipses being unobserved resulting in a spurious period estimation.

light curve shapes. For the periodogram, this results in sinusoidal and sawtooth light curves requiring less runtime at low numbers of data points compared with eclipsing binary light curves. At higher numbers of data points, the runtime required by the $O(N^2)$ periodogram becomes dominant and all the runtimes converge. For GRAPE, the effect is to also separate the sinusoidal and sawtooth runtimes from the slower eclipsing binary runtimes. However, due to the linear dependence on data points of GRAPE, this runtime difference remains to higher data point light curves.

For the Skycam cadence results shown in figure 3.21, the runtimes are longer and similar effects to the regular cadence light curves are seen but are suppressed. This suppression is likely a result of the poorer sampling quality of the light curves. This results in increased computational effort required by both GRAPE genetic optimisation and the Vuong Closeness test used in both GRAPE and the periodogram. As a result, the periodogram requires more time to process the lower data point light curves as the

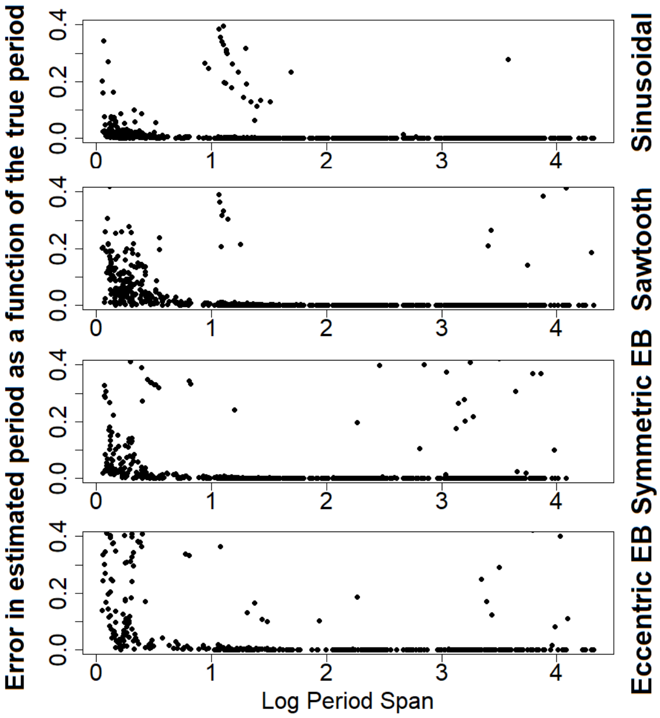


FIGURE 3.18: Plot of the base-10 logarithm of the period span of light curves with a given period as a function of the estimated period fractional error for the regular cadence light curves.

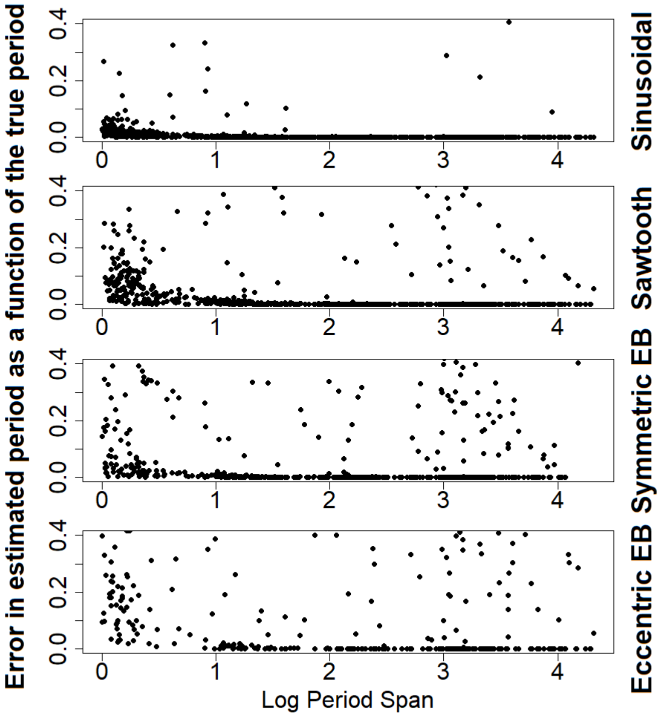


FIGURE 3.19: Plot of the base-10 logarithm of the period span of light curves with a given period as a function of the estimated period fractional error for the Skycam cadence light curves.

aliased periods are more dominant. The separation of the runtimes on the sinusoidal and sawtooth light curves compared to the eclipsing binary light curves is still apparent but reduced as the sinusoidal and sawtooth light curves require additional computation to deal with aliased periods. Ultimately, for the Skycam cadence light curves, which is of more interest due to them more closely reproducing the properties of the real survey data, GRAPE appears to maintain similar performance to the periodogram for the sinusoidal and sawtooth light curves and requires less runtime than the periodogram on light curves with greater than $500 - 1000$ data points. Whilst many of the interesting Skycam light curves do have data points in this range, there are many light curves with data points from $1000 - 15,000$ data points which would require an unacceptably long time to run using a periodogram (due to the $O(N^2)$ requirement of this method). GRAPE provides an approach to obtain the desired results in less time.

We also peformed one final experiment to determine the runtime required for the BGLS periodogram, with Vuong Closeness, to obtain similar results to GRAPE on the regular cadence and Skycam cadence sinusoidal light curves through the use of a fine-tuning step on the initial $N_t = 5$ candidate periods with a $\pm 10\%$ search radius using a boosted oversampling factor. As the ofac $= 5$ of the previous experiments was sufficient for all non-sinusoidal light curves, we perform this only on the set of 100 stratified regular cadence sinusoidal light curves and 100 stratified skycam cadence sinusoidal light curves. Table 3.7 demonstrates the results of this experiment on the regular cadence light curves using the results of the GRAPE performance contrasted with the base periodogram performance with ofac $= 5$ and with two fine-tuning variants, one with ofac $= 20$ and one with ofac $= 50$. Table 3.8 demonstrates the results of this experiment on the Skycam cadence light curves with the same oversample fine-tuning. The results indicate that the periodogram can replicate the performance of GRAPE on sinusoidal light curves through an additional oversampling fine-tuning operation. For the regular cadence light curves this is possible with a fine-tuning step with an oversampling factor of 20. For the Skycam cadence light curves the fine-tuning oversampling factor was required to be 50. These performance gains do come at a computational cost which further effects the runtime of the periodogram compared to GRAPE. The mean light curve runtime for GRAPE on the regular cadence sample was 37.32 seconds and the BGLS periodogram with no fine-tuning mean runtime was 45.43 seconds. For Skycam cadence these runtimes were increased due to the higher difficulty in identifying candidate periods with a mean runtime of 52.59 seconds on GRAPE and 57.55 seconds on the periodogram. The fine-tuning operation with ofac $= 20$ increase this runtime to 54.94 seconds on the regular cadence data and 64.25 seconds on the Skycam cadence data. For the ofac $= 50$ fine-tuning operation, the computational expense increases to 56.77 seconds on the regular cadence data and 64.46 seconds on the Skycam cadence data. Ultimately, whilst the

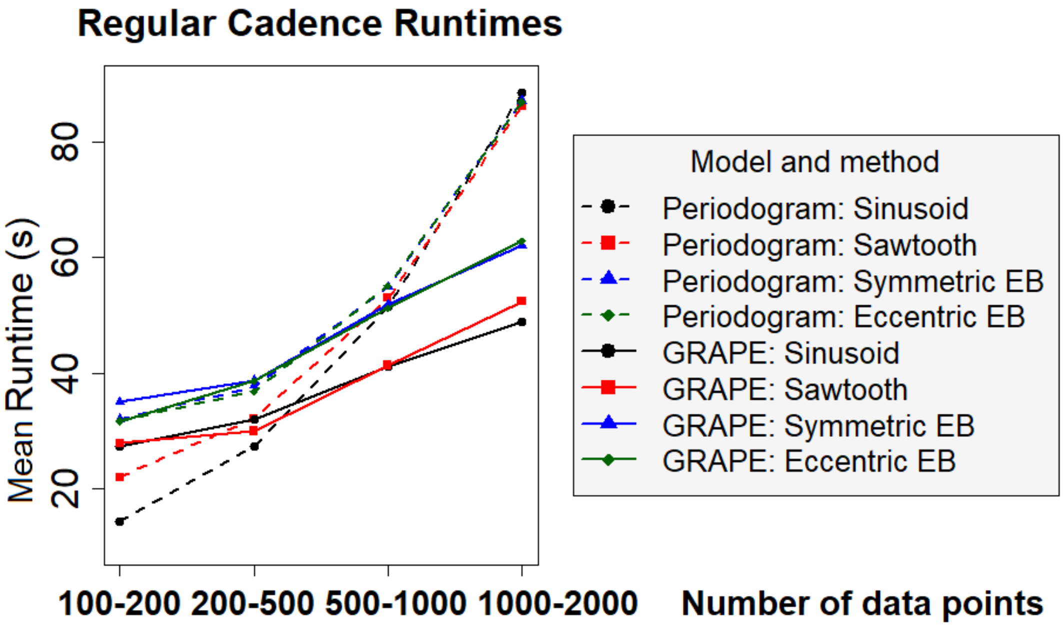


Figure 3.20: Plot of the average runtime of the different shaped light curves as a function of the number of data points for the regular cadence light curves.

Table 3.7: GRAPE and BGLS periodogram period estimation results on 100 stratified regular cadence sinusoidal light curves with various fine-tuning oversampling runs.

Data	Hit	Multiple	Alias
GRAPE	0.803 ± 0.037	0.000 ± 0.000	0.000 ± 0.000
BGLS Periodogram	0.650 ± 0.080	0.000 ± 0.000	0.000 ± 0.000
ofac = 20 finetune	0.810 ± 0.070	0.000 ± 0.000	0.000 ± 0.000
ofac = 50 finetune	**0.820 ± 0.070**	**0.000 ± 0.000**	**0.000 ± 0.000**

Table 3.8: GRAPE and BGLS periodogram period estimation results on 100 stratified Skycam cadence sinusoidal light curves with various fine-tuning oversampling runs.

Data	Hit	Multiple	Alias
GRAPE	0.780 ± 0.040	0.000 ± 0.000	0.013 ± 0.013
BGLS Periodogram	0.660 ± 0.080	0.000 ± 0.000	0.050 ± 0.040
ofac = 20 finetune	0.760 ± 0.070	0.000 ± 0.000	0.020 ± 0.030
ofac = 50 finetune	**0.790 ± 0.070**	**0.000 ± 0.000**	**0.010 ± 0.020**

fine-tuning operations allow the periodogram to match the performance of GRAPE, it is a the cost of increasing the required runtime. As the fine-tuning operation is also frequency spectrum based, this additional computational effort is also $O(N^2)$ complexity meaning for light curves with many observations, the processing time will be prohibitive for real-time analysis.

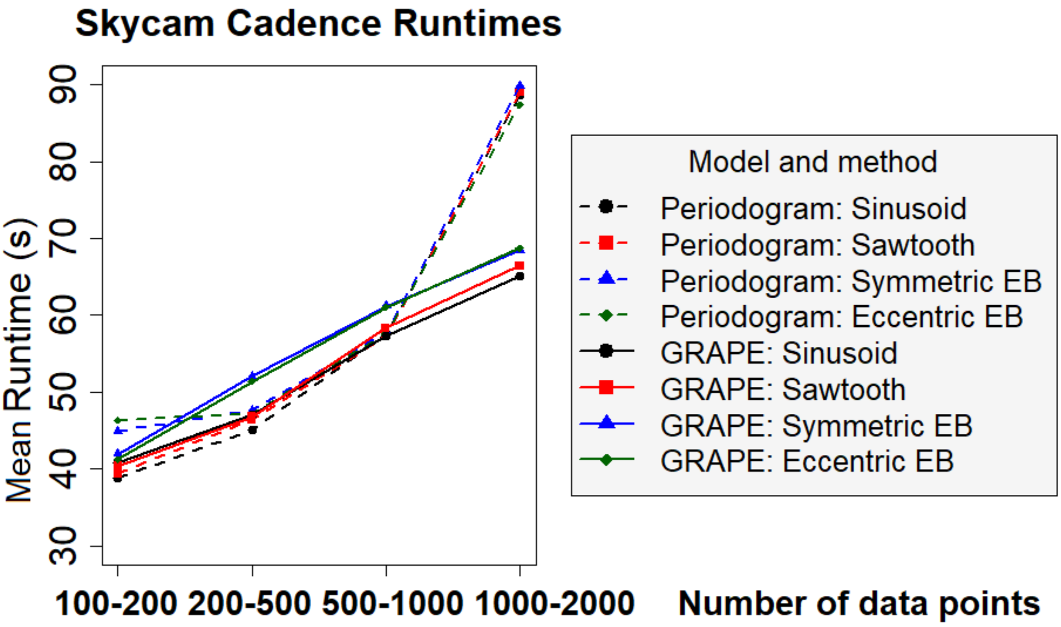


FIGURE 3.21: Plot of the average runtime of the different shaped light curves as a function of the number of data points for the Skycam cadence light curves.

3.7 Discussion and Conclusion

In this chapter we introduced GRAPE, a period estimation statistic embedded in a genetic algorithm with a Vuong closeness test based alias and multiple model discrimination procedure. BGLS was selected as the period estimation statistic to be used as the fitness function. We note that other methods can be used instead as the role of the fitness function is highly modular and different measures can be combined (Saha and Vivas, 2017). Our experiments in this chapter show that Lomb-Scargle type methods function on poorly sampled data due to sinusoidal interpolation. This does mean that non-sinusoidal signal performance degrades rapidly.

It is also important to caution the use of Bayesian periodogram methods in the search for periodicity of generic signal shapes. Bayesian periodograms output probabilistic statements based on the assumption that the data has been drawn from a sinusoidal model (VanderPlas, 2017). This results in the suppression of features in the periodogram which would normally convey information on the nature of the underlying periodicity. However, GRAPE has been designed for *automated* functionality and therefore it is important that potential failure modes be tightly controlled. As a major risk of any optimisation method is becoming trapped in local minima, the suppression of aliases is highly useful in propagating the genetic information. Therefore, BGLS has been deployed in this work despite the output being unreliable in the regime of non-sinusoidal periodicity. It is possible that the use of a quantum evolutionary algorithm would afford additional protection from populations becoming trapped at insignificant local minima and therefore sever our current dependency on the Bayesian periodogram (Abs da Cruz et al., 2007).

GRAPE outperformed the periodogram frequency grid in all datasets with sinusoidal signals which are well described by the fitness function. We suggest, as discussed earlier in the chapter, that GRAPE treating the period space as a continuous variable leads to this success as the genetic algorithm could fine tune the result. This is further supported by the fine-tuning periodogram method which used a frequency spectrum with ofac = 50 to oversample the 10% period range around the N_t candidate periods. The sinusoidal light curves had a relative hit rate improvement of 18.2% using GRAPE compared to the periodogram for the regular cadence data and 14.9% for the Skycam cadence data with both methods utilising the same BGLS fitness function. This was determined by assuming that the failed matches in GRAPE were also failures in the periodogram and calculating the percentage of additional failures in the periodogram. Using the same method, we determine that the sawtooth light curves had a 6.4% improvement on regular cadence and 3.7% on Skycam cadence when using GRAPE although this is close to the 95% confidence interval. The symmetric EB and eccentric EB light curves are too much of a departure from the sinusoidal assumptions of the Lomb-Scargle method and exhibited a GRAPE relative performance similar, possibly slightly inferior, to the periodogram when comparing the hit and submultiple rate. This is likely a result of the reliance of GRAPE on the genetic propagation of useful information about the period space during the evolution of the candidates. On a sinusoidal light curve, the genetic algorithm places additional candidates near the sinusoidal signal period due to the prevalence of superior fitting models in this region of the period space. The algorithm can then fine tune the resulting period from this region. For the eccentric EBs, the fitness function returns a substantially weaker response to candidate periods near the true underlying simulated period. As a result, it only requires the presence of a similar-strength false model (such as on a spurious or aliased signal) to 'kick' candidate periods out of this region of the period space and removing it from the fine-tuning operation. In this case, the brute force approach of the periodogram frequency grid outperformed GRAPE purely because a candidate period close to the true period would always be sampled. These results are an overall measure of the relative performance and, as can be seen in the plots in figures 3.10 to 3.16, the actual performance of the methods is strongly dependent on both the shape of the underlying light curve, the value of the true period and the baseline (total measured time) of the light curve.

Our experiments show the sampling inherent to the Skycam mode of surveying will lead to objects insufficiently sampled for successful identification however the yield looks reasonable based on the relative performance degradation between the regular cadence and Skycam cadence data for sinusoidal and sawtooth light curves. Unfortunately, the loss of eclipsing binaries will be substantial for the Skycam survey regardless of period estimation method from a sampling viewpoint. The simulated data we present in this research contains only a signal and white noise component. In reality, red (correlated) noise sources are common in real light curves. GRAPE currently makes no attempt to address the presence of red noise within candidate light curves, with the BGLS fitness function assuming purely white noise residuals. Whilst the red noise is something that can be modelled, at the moment we make use of a prewhitening technique to eliminate large correlated systematic signals when applying this method to real light curves from the Skycam database.

GRAPE has a notable computational time benefit over the frequency spectrum on light curves with more than 500-1000 data points, occasionally with increased performance likely based on the chosen fitness function. This is due to the genetic algoritms $O(N)$ dependency on the number of data points in a light curve compared with the $O(N^2)$ of the frequency spectrum approach. We found that the topology of our genetic algorithm, as defined by the arguments listed in the previous section, are close to the fastest implementation we could produce before a substantial loss in performance due to having insufficient population or generations to explore the period space. We found that the linear decay of mutation was an incredibly powerful way of maintaining performance whilst decreasing the number of generations. The mutation can be described with a thermodynamic analogy. At the beginning the mutation is extremely high and the candidate periods jump rapidly across the parameter space like hot particles escaping a nearby potential. As the mutation rate decreases, the population has less energy to climb out of the potential wells and therefore begin to fine tune the local periods compared to exploring new regions of the parameter space. Eventually the mutation rate is so low that very little to no new exploration is occuring as the population cannot stray far from the local potential. Individuals located in poor areas of the parameter space die off leaving the group near the true period to reproduce solely with the purpose of fine tuning this result. We are satisfied with the computational overhead required as many Skycam light curves do have a large number of data points which become prohibitively difficult to compute a periodogram on combined with the improved accuracy of GRAPE over the periodogram. For a Skycam light curve with 5114 data points it takes the frequency spectrum periodogram about 454.7 seconds to complete whereas GRAPE completes the same task in 93.5 seconds.

Chapter 4

Machine Learning

Machine Learning is a multidisciplinary field with a basis in statistics and computer science based on the deployment of computational algorithms which use training data to create a model which approximates an unknown function to map the inputs of training data to outputs (Samuel, 1988). This learning happens iteratively as the algorithms improve the learned model by minimising the error between the predicted output and the desired output. Machine Learning has been described as a computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E (Mitchell, 1998).

Machine Learning tasks can be subdivided into one of two groups, regression and classification (Russell and Norvig, 2009). Many of the methods already introduced such as in the form of the fitting of harmonic Fourier series and the Variance Ratio Periodogram are powered using linear regression, a form of regression machine learning. Regression methods utilise a set of features (which may just be the raw data points) and reproduce a model which is either limited such as in linear regression to a specific set of parameters or flexible, even non-linear, such as neural networks and random forests, to reproduce a real valued output. This output is then refined through the computation of an error term such as root mean-squared-error (RMSE) which is then minimised by the training process. The normal equation used for the computation of harmonic models is a form of regression training for linear regression which can refine a solution to the global optimum in one pass but other methods can accomplish the same with more iteration such as gradient descent based approaches.

Classification machine learning follows a similar set of procedures except the output is the probability of the input features being produced by an observation, given set of possible classes of which the observation may be a member (Kotsiantis, 2007). This

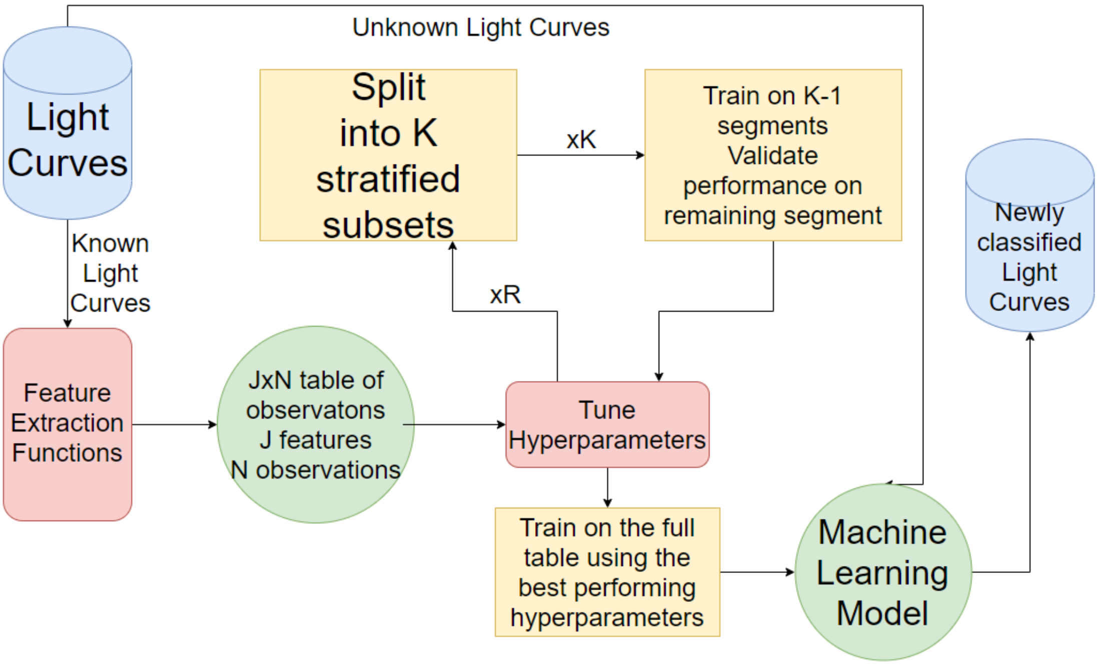

Unknown Light Curves
Light Curves
Known Light Curves
Split into K stratified subsets
xK
Train on K-1 segments Validate performance on remaining segment
xR
Newly classified Light Curves
Feature Extraction Functions
JxN table of observatons J features N observations
Tune Hyperparameters
Train on the full table using the best performing hyperparameters
Machine Learning Model

Figure 4.1: A flowchart of a traditional feature engineering machine learning pipeline. The feature extraction functions are manually built by experts requiring substantial time investment. Upon the determination of a good set of features, the best performing ones can be selected for use in a shallow machine learning model with a reduced computational effort.

probability vector of length equal to the number of possible classes can then be used to assign the observation to a given class based on probability cuts and confidence intervals. Usually this is accomplished through the application of a non-linearity function which converts the output of a machine learning algorithm into a probabilistic representation.

The training process of Machine Learning systems can be grouped into supervised, unsupervised and reinforcement learning (Russell and Norvig, 2009). In supervised learning, the labels for the classes or the regression objective value are supplied during the training phase of the algorithm. Unsupervised learning methods lack this label knowledge and instead attempt to cluster or group the training data based on the patterns in the training data. Reinforcement learning utilises a response metric to determine the performance of each classification in an attempt to learn which actions result in high performance.

Machine learning can also make use of shallow or deep topologies. Shallow topologies are easier to train but are not capable of generating a model from the raw data instead requiring a feature engineering and extraction procedure where the raw data is processed into a set of features designed to describe the desired problem. Figure 4.1 shows a flowchart demonstrating this feature engineered machine learning pipeline. Deep topologies are much harder to train due to the number of parameters in the learning

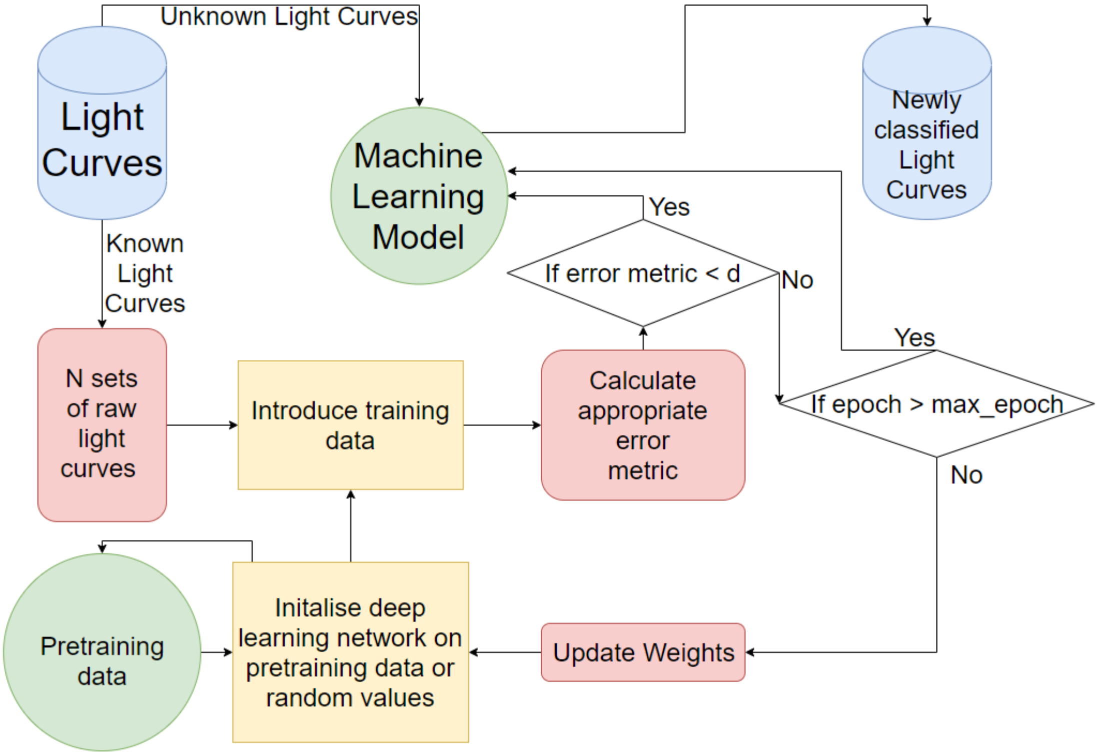


FIGURE 4.2: A flowchart of a deep learning pipeline. The deep learning training is capable of extracting features useful to the desired classification task directly from the raw light curve data due to the number of parameters available in the network. The models take substantially longer to train as they have a larger number of parameters as well as requiring a pretraining step to initialise the network. This increased training time is compensated by the reduced time requirement from human developers.

models but are capable of performing feature extraction as part of the learning process. They are supplied with the raw data in a machine-readable form and then they construct a set of non-linear features from the training data that results in the best performance for the desired task. With the new generations of modern hardware, these deep learning methods are seeing widespread utilisation. Figure 4.2 contains a flowchart demonstrating a deep learning pipeline. In this chapter we introduce the algorithms we make use of in the development of the automated classification pipeline, a predominantly supervised learning based problem. We also discuss how to evaluate the performance of a learned model through a set of performance metrics.

4.1 Principal Component Analysis

Principal Component Analysis (PCA) is a mathematical method that transforms a number of variables or features which may be correlated into a set of uncorrelated variables called principal components (Pearson, 1901; Hotelling, 1933, 1936). The principal components account for the variability of the dataset in order with the first principal component describing a large amount of the initial variance and the second principal component describing much of the remaining variance until the last principal component contains the remaining variance.

Principal components can be calculated from a design matrix X, with columns containing the variables of the dataset and the rows containing the observations. Initially the variables in the design matrix must be scaled so they have comparable values using equation 4.1.

$$\bar{x_j} = \frac{x_j - \mu_j}{\sigma_j} \quad \text{for } j = 1, 2, \ldots, N \tag{4.1}$$

where $\bar{x_j}$ is the rescaled j^{th} variable, μ_j is the mean of the j^{th} variable, σ_j is the standard deviation of the j^{th} variable and N is the number of variables in the design matrix X. The covariance matrix of the rescaled design matrix X is then computed using equation 4.2.

$$\Sigma = \frac{1}{m} \sum_{i=1}^{m} \left(X_{[i]}\right) \left(X_{[i]}^{\top}\right) \tag{4.2}$$

where $X_{[i]}$ is the row vector of the design matrix X for the i^{th} observation and m is the total number of observations (rows) in the design matrix. Using the covariance matrix, the eigenvalues and eigenvectors are computed. The sorted eigenvalues from high to low give the principal components in the solution in order of the variance contained in each principal component. The eigenvector associated with the eigenvalues can be used to determine the principal components using equation 4.3.

$$\text{PCA}_j = \Theta_j^{\top} X \tag{4.3}$$

where PCA_j is the j^{th} principal component, Θ_j is the eigenvectors associated with the j^{th} sorted eigenvalue and X is the design matrix. This produces uncorrelated principal components and can also be used for dimensionality reduction through the selection of an integer value k where $1 \leq k \leq N$. The choice of k can be decided through the computation of how much variance is described by the reduced design matrix $\hat{X}$ where $\hat{X} = \Theta_{1,\ldots,k}^{\top} X$. Equation 4.4 shows the inequality which must be satisfied with the minimum value of k to retain $1 - \epsilon$ of the variance of the design matrix X.

$$\epsilon \geq \frac{\frac{1}{m}\sum_{i=1}^{m} ||x_{[i]} - \hat{x_{[i]}}||^2}{\frac{1}{m}\sum_{i=1}^{m} ||x_{[i]}||^2} \tag{4.4}$$

where $(1 - \epsilon)$ is the required retained variance, $x_{[i]}$ is the row vector of the design matrix X for the i^{th} observation, $\hat{x_{[i]}}$ is the row vector of the reduced design matrix $\hat{X}$ for the

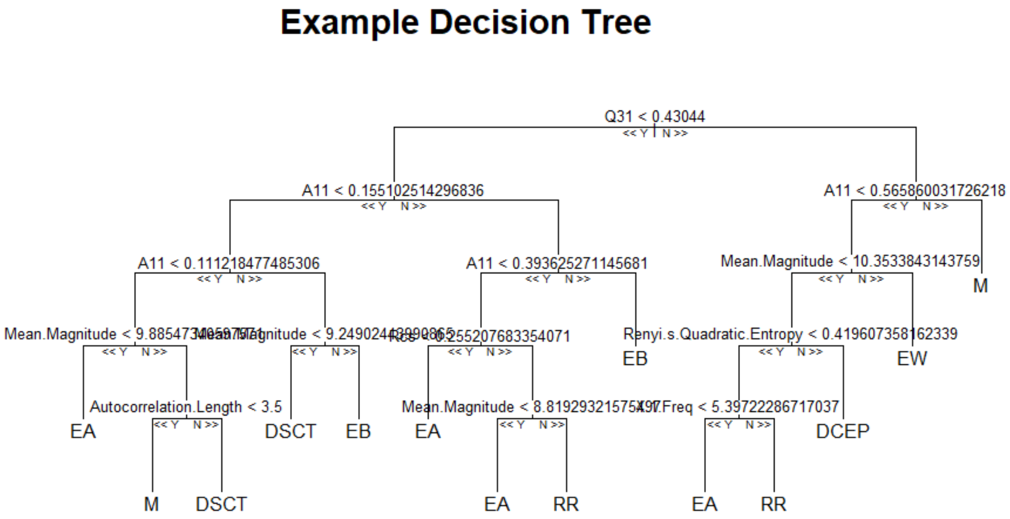


FIGURE 4.3: An example decision tree extracted from a Random Forest model trained on a subset of Skycam light curve data. Features such as the amplitude and interquartile range of the light curve are used to place the training light curves into the appropriate classes. The binary splits are clearly seen where a single feature is used to compute the split (more can be used). The leaf nodes are assigned to a class label when they contain a small enough number of training objects.

i^{th} observation reconstructed using k principal components and m is the total number of observations (rows) in the design matrix.

4.2 Random Forest classifiers

Random Forests are an example of an ensemble classification method where a collection of weak classifiers are combined to produce a single strong classifier (Breiman, 2001). For the Random Forest ensemble, the individual classifiers are Classification And Regression Trees (CARTs) (Breiman et al., 1984). A CART is constructed through performing binary splits around a variable or set of variables. These branches then undergo additional binary splits until the tree has enough depth that the branch can be ended as a leaf node where these leaf nodes have an arbitrary size subset of the initial dataset. Figure 4.3 demonstrates a decision tree extracted from a Random Forest model trained on Skycam light curves.

A number of these trees defined by an argument n_{trees} when determines the number of decision trees trained by the Random Forest ensemble. These trees are trained through the use of bootstrap aggregating where, similar to the bootstrapping using in the evaluation of GRAPE, the training set is randomly sampled with replacement to train each individual tree. This process is commonly known as Bagging (Breiman, 2001). For each split in each tree, a random sample of variables of size m_{try} is selected to decide the split. A third argument named *nodesize* allows the control of how deep or shallow the trees are grown by determining the maximum amount of training data allowed by a terminal leaf node. Any nodes with an amount of data greater than this size must execute a split. Deep trees can describe many variable interactions but readily overfit the training data whereas shallow trees limit the complexity of the decision tree model which can result in a high bias model, the model is too simple for the desired problem. The Random Forest can be formally defined as a set of classifiers $h(x|\Theta_1), \ldots, h(x|\Theta_K)$ produced based on training from a set of training data $D = [(x_i, y_i)]_{i=1}^{n}$ where h is the non-linear function of the decision tree parameters, Θ_j are the parameters of the j^{th} decision tree, x_i is the variables of the i^{th} observation and y_i is the class of the i^{th} observation.

Random Forests are not trained in the traditional sense. They do not make use of an iterative update process which improves the performance of the individual decision trees. Rather, they continue to grow new trees using the random splitting of the variables in the training data until a cohort of trees with good performance on the desired problem are produced. This cohort of trees can be identified through the use of an 'out-of-bag' error. This error is a measure of the performance of an individual tree using the training data which was not selected as part of the bagging process and will have a lower value for the better performing trees. As the performance of the model is not effected by the low performance trees, the overall performance of the ensemble improves as trees with good performance are generated. The simplicity of the Random Forest trees compared to traditional decision trees allows them to be more generalisable, one of the main limitations of CARTs, although they must be used as part of an ensemble method.

Another interesting component of Random Forests is they have automatically incorporated performance and feature importance measures. The out-of-bag error is a measure of the error rate of a trained Random Forest model by computing the performance of each decision tree using the data which was not selected during the bootstrap sampling operations. The Gini criterion is a measure of the diversity, i.e. the proportion of classes of differing types, in each leaf node. The more important features will result in a larger change in this Gini criterion if they are removed from the training. This change is named the Mean Decrease Gini of the feature given a trained Random Forest model. The Gini criterion is defined by equation 4.6 given equation 4.5.

$$g(S_j) = \sum_{i=1}^{N} \hat{P}(C_i|S_j)(1 - \hat{P}(C_i|S_j)) \tag{4.5}$$

where S_j is the set of data in the j^{th} leaf node, C_i is the i^{th} class in the data and $g(S_j)$ is the variation of the j^{th} leaf node, minimised when the child nodes S_j contains only one class C_i. $\hat{P}(C_i|S_j)$ is the proportion of data in leaf node S_j which is of class C_i. N is the number of classes in the training dataset. The Gini criteron is then determined by the weighted sum of the variations shown in equation 4.6.

$$G = \sum_{j=1}^{M} \hat{P}(S_j) g(S_j) \tag{4.6}$$

where G is the Gini criterion and $\hat{P}(S_j)$ is the proportion of training data in the j^{th} leaf node relative to the total number of training data objects. This criterion, through the mean decrease Gini feature importance measurement, can be used in the process of feature selection. Random Forests determine their final prediction based on 'votes' from each of the individual decision tree classifiers with the probability of an object being of class C_i decided by the proportion of trees which classify the object as being a member of class C_i. Random Forests also allow for the determination of the similarity between the feature vectors of two objects by determining the proportion of the decision trees where the two objects are placed in the same terminal leaf node. This similarity measure is similar to a euclidean distance between the objects in the feature space as weighted by the importance of the features in the Random Forest model.

4.3 Support Vector Machines

Support Vector Machines (SVMs) generate a model by determining the widest margin between vectors within an n dimension multidimensional feature space (Cortes and Vapnik, 1995). This decision rule is defined as a hyperplane maximising this margin of dimensionality $n-1$. The vectors responsible for the position of this hyperplane are named support vectors. This algorithm is powerful at solving classification problems based on the hyperplane generated by the training set vectors. The representation of the class boundaries is flexible and can fluctuate based on the introduction of new support vectors. This allows the SVM to improve its model based on misclassified vectors. The algorithm also implements automatic complexity control to reduce overfitting by allowing the violation of margins by vectors in order to better accommodate the remaining support vectors. Finally, the algorithm has a single global minimum which can be found in polynomial time (Cortes and Vapnik, 1995). This means the algorithm will rapidly converge to the best fitting model without the presence of local minima causing the production of a poor model. Other competing algorithms are unable to guarantee this as they are often lack a convex optimisation function.

The SVM is easy to define by considering the dynamics of the feature space the vectors operate within. The Hard Margin Support Vector Machine is the simplest form

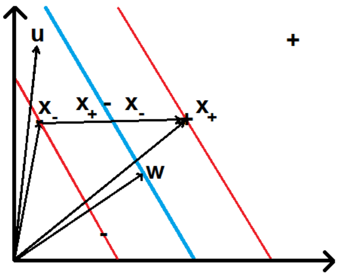


FIGURE 4.4: A two dimensional feature space containing four vectors of two different classes and their associated margins. The two red lines indicate the maximum margins determined by the closest positive and negative classes, these are the support vectors. The blue line indicates the optimal hyperplane between the two classes based on maximising the distance between the two margins. The **w** vector indicates the position of this optimal hyperplane and **u** is a test-case feature vector.

of this mathematical construct and therefore shall be described in detail here. The extended Soft Margin Support Vector Machine is then introduced which makes use of a cost penalty to allow poor performance on outlier training set vectors to improve the generalisability of the model.

Figure 4.4 represents a simple two dimensional feature space containing four feature vectors, two positively classified and two negatively classified. The vectors $\mathbf{x}_-$ and $\mathbf{x}_+$ indicate the locations of two support vectors, one classified negatively and one positively. The vector **u** is an unclassified test vector. Finally, the **w** vector is a vector normal to the blue separating hyperplane that represents this hyperplane. The two red lines represent the best fitting margins separating the negative support vectors and the single positive support vector. As the top right hand corner positive vector does not lie on or within the margin of the separating hyperplane, it is not a support vector whereas the other three vectors are. If the vector **u** lies upon the positive side of the separating hyperplane, the inner product between **w** and **u** is greater than an undefined constant c, $\mathbf{w} \cdot \mathbf{u} \geq c$. This can be converted to equation 4.7 by defining a new constant b where $c = -b$. This equation becomes the first decision rule defined by the hyperplane and defines when the unknown vector **u** is classified as positive.

$$\mathbf{w} \cdot \mathbf{u} + b \geq 0 \tag{4.7}$$

An additional variable can be introduced to generalise equation 4.7 into a single decision rule. Name this new variable y_i such that $y_i = +1$ for positive samples and $y_i = -1$ for negative samples. This produces the new decision rule shown in equation 4.8.

$$y_i \left(\mathbf{x_i} \cdot \mathbf{w} + b \right) - 1 \geq 0 \tag{4.8}$$

For x_i along the margin, the limit of the margin, $y_i\left(\mathbf{x_i}\cdot\mathbf{w}+b\right)-1=0$. The width of the margins can be defined as shown in equation 4.9 where $\frac{\mathbf{w}}{||w||}$ is the unit vector of $\mathbf{w}$.

$$M_w=(\mathbf{x}_+-\mathbf{x}_-)\cdot\frac{\mathbf{w}}{||w||}=\frac{2}{||w||} \tag{4.9}$$

The best model will maximise the size of these margins so therefore we must maximise $\frac{2}{||w||}$ which is equivalent to a minimisation of $||w||$. This is a quadratic optimisation problem with the minimisation of $||w||$ with respect to the margins and is solved using Lagrange multipliers leading to the final decision rule in equation 4.10 for a positive classification of $\mathbf{u}$.

$$\sum\alpha_i y_i\mathbf{x_i}\cdot\mathbf{u}+b\geq 0 \tag{4.10}$$

Where $\mathbf{w}=\sum\alpha_i y_i\mathbf{x_i}$ determines the values of the α_i weights as α_i provides the weighting to each training vector $\mathbf{x_i}$.

This Hard Margin SVM is very inflexible. It can only create decision rules where the vectors are never allowed to violate the margin boundaries. This can lead to hyperplane overfitting and therefore an overfitting decision rule if any of the support vectors are outliers. A better approach is to use a Soft Margin Support Vector Machine. This approach allows vectors to violate the margins at an associated penalty cost. This can result in a superior decision rule despite the possible incorrect classification of feature vectors in extreme cases. As any vector that manipulates the decision boundary is a support vector, any vectors that violate the margins are also support vectors. A new cost parameter ξ_i is introduced. This parameter identifies the cost associated with the violation of the margin by a support vector $\mathbf{x_i}$.

SVMs are natively a linear binary classification algorithm. That is, they determine a flat hyperplane as the decision boundary. Kernel functions are a method to extend this functionality to non-linear models (Shawe-Taylor and Cristianini, 2004). The SVM is dependent only on the inner product between support vectors, not the specific values of the support vectors themselves. Therefore Kernel functions express this inner product in a higher dimensional space where the non-linear problem becomes linearly separable. This is also known as the ‘Kernel Trick’. Kernel functions $K(x,y)$ must satisfy Mercer’s condition as shown in equation 4.11.

$$\int\int g(x)K(x,y)g(y)dxdy\geq 0 \tag{4.11}$$

The four Kernel functions most often utilised for non-linearity in SVMs are the linear kernel function, Radial Basis Function (RBF) kernel function, the polynomial kernel function and the sigmoid kernel function. There are also kernel functions for more specific use cases such as the family of string kernel functions for dealing with sequences of string-based data.

4.4 Artificial Neural Networks

Artificial Neural Networks are highly scalable learning methods which attempt to replicate the biological structures present in the human brain (Rosenblatt, 1958). This scalability allows them to produce state-of-the-art performance on many modern tasks at the cost of requiring substantial amounts of training data and computational resources. Neural Network algorithms use layers of neurons, a nexus of connections from a previous layer where the outputs of the previous layer are weighted, summed and then have a non-linear activation function applied to them to introduce non-linearity in the system (Vora and Yagnik, 2014; Che et al., 2011). The networks assume that the variables are normally distributed and therefore the data should be rescaled prior to use. Transforming some of the variables using logarithms may also be recommended.

Artificial Neural Networks combine multiple neurons into a connective learning network which can adopt multiple topologies. The two most common topologies are the Feedforward Neural Network and the Recurrent Neural Network (Fine, 1999; Jain and Medsker, 1999). Feedforward Neural Networks propagate information throughout the layers in one direction, from the early hidden layers to the later hidden layers until it reaches the output layer. Recurrent Neural Networks also implement links which allow information from later layers to be transmitted to earlier layers or even the same layer. These recurrent links make these networks much more difficult to train but allow them to store information in a sort of neural memory. For this reason they are attractive to data containing a temporal structure. Recurrent Neural Networks have seen recent successes in the field of variable star classification (Naul et al., 2018).

Feedforward Neural Networks contain an input layer, composed of the number of input features plus a bias unit, i.e. the input layer has $N + 1$ neurons where N is the number of features. There are then one or more hidden layers, sets of neurons which learn representations of the input features which map well to the output task. Finally, the output layer is either a single linear neuron for regression tasks or, for classification, a set of m neurons where m is the number of classes. This classification output layer can either feed into another classification method such as the Support Vector Machine or alternatively can be implemented using the softmax function, a multinomial logistic regression classification function. They can assume a number of different topologies based on the number of neurons present in the hidden layers and the number of hidden

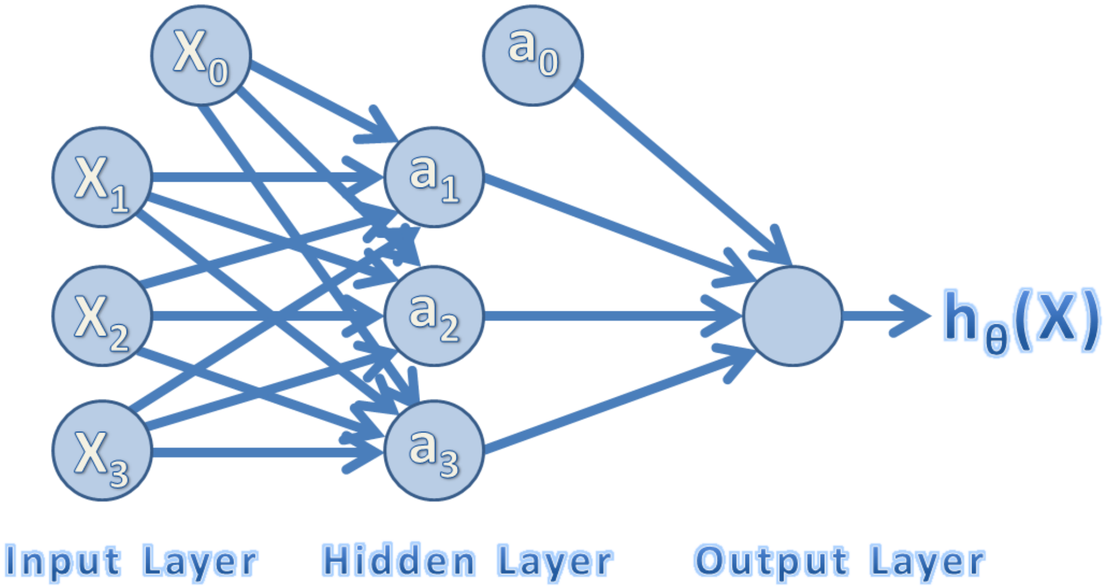


FIGURE 4.5: A single hidden layer Artificial Feedforward Neural Network. The inputs from the previous layer are weighted by learned parameters (Θ_j matrix for the j^{th} layer) and summed. This final summation has a non-linear activation function applied to it and this result is sent to the next layer.

layers. Figure 4.5 demonstrates the topology of a simple Feedforward Neural Network with three input features and three hidden layer neurons.

The Feedforward Artificial neural network can be defined using a set of activation function neurons $a_i^{(j)}$, the activation of unit i in layer j. These activations are weighted by a matrix of weights $\Theta^{(j)}$ controlling the mapping from layer j to $j+1$. Equation 4.12 and equation 4.13 demonstrate the activation of the hidden layer neurons by the inputs.

$$a^{(2)} = g\left(\left(\Theta^{(1)}\right)^\top X\right) \quad \text{for the first hidden layer} \tag{4.12}$$

$$a^{(j)} = g\left(\left(\Theta^{(j-1)}\right)^\top a^{(j-1)}\right) \quad \text{for the remaining layers} \tag{4.13}$$

where $a^{(j)}$ is the output of layer j, $\Theta^{(j-1)}$ is the weight matrix of the layer $j-1$, $X = x_1, x_2, \ldots, x_n$ is the input feature vector and $g(x)$ is the non-linear activation function of the network. There are a number of popular activation functions including the sigmoid function, shown in equation 4.14, the tanh function $g(x) = \tanh(x)$ and the rectified linear unit (relu) function $g(x) = \max(0, x)$.

$$g(x) = \frac{1}{1 + \exp(-x)} \tag{4.14}$$

This forward pass of information is named Forward Propagation and it allows the inputs to be mapped to the outputs of the Neural Network. However, this method does not allow the training of the weight matrices $\Theta^{(j)}$ based on a training dataset. The Back-Propagation algorithm was developed to accomplish this task through an iterative foreward then backward pass structure. The forward pass would compute the outputs of a set of input training data X. The output layer is then compared to the desired output through the computation of a cost or error function. The errors are then determined as a function of the neurons in the network and corrected through the calculation of an 'error derivative' for the output layer neurons. This correction then propagates backwards through the network until the weight matrices of each layer have been adjusted. A forward pass is then computed to determine the change in error from this corrective backward pass and this process is repeated iteratively until the output errors on the training set have been minimised.

Consider $\delta_j^{(l)}$ as the error of neuron j in layer l. We can compute the error of the output layer as $\delta^{(o)} = a^{(o)} - y$ where $a^{(o)}$ is the output layer after a forward pass, y is the objective ground output and $\delta^{(o)}$ is the error in the output layer after a foreward pass. Equation 4.15 defines the computation of these derivatives as they propagate backwards through the network.

$$\delta^{(j)} = \left(\Theta^{(j)}\right)^{\top} \delta^{(j+1)} \odot g' \left(\left(\Theta^{j-1}\right)^{\top} a^{(j-1)}\right) \tag{4.15}$$

where $\delta^{(j)}$ is the error vector for the j^{th} layer, $\Theta^{(j)}$ is the weight matrix of the j^{th} layer, $a^{(j)}$ is the activation function of the j^{th} layer and $g'(x)$ is the first derivative of the activation function $g(x)$. $\odot$ represents the elementwise multiplication operation. In this backpropagation method there is no $\delta^{(1)}$. Backpropagation is not the only way to train a Neural Network and the Levenberg-Marquadt method has also been used to produce trained models (Basterrech et al., 2011). Genetic algorithms are also a potential training method due to their ability to locate global minima in complex functions (Che et al., 2011).

4.5 Naive Bayes

The Naive Bayes classifier makes use of Bayesian probability to classify a set of variables into a given class. They are described as Naive as they apply the assumption that the features are strongly independent, i.e. they are not correlated (Kotsiantis, 2007; Russell and Norvig, 2009). The algorithm also assumes that the features are part of a normal distribution. Equation 4.16 demonstrates how the probability of an input feature vector x conditioned on the class y_j is a function of the probability of each feature x_i having its value conditioned on being a member of class y_j.

$$p(x|y_j) = p(x_1|y_j), p(x_2|y_j), \ldots, p(x_n|y_j) \tag{4.16}$$

where x_i is the i^{th} feature, y_j is the j^{th} class and there are n total features in the feature vector. It is more important to determine the probability of the feature vector being of class y_j given it has a set of features x. This is accomplished using Bayes' Theorem shown in equation 4.17.

$$p(y_j|x) = \frac{p(y_j)p(x|y_j)}{p(x)} \tag{4.17}$$

Assuming that the features satisfy conditional independence, equation 4.18 modifies Bayes' Theorem to show the probability of the feature vector x being a member of class y_j is determined by the probability of any feature vector being a member of class y_j times the product of the individual probabilities of the features being members of class y_j scaled by a constant produced by the probability of the feature vector having the associated features.

$$p(y_j|x) = \frac{p(y_j)\prod_{i=1}^{n} p(x_i|y_j)}{p(x)} \tag{4.18}$$

The algorithm is quick to train and can operate successfully on feature vectors with missing values by ignoring the missing features. The conditional independence assumption is a significant limitation to the method as many training sets will have some level of correlation between the features.

4.6 Performance Evaluation

The performance of a classification algorithm on a set of training and test data must be expressed mathematically or differing methods cannot be compared. There is a number of potential diagnostic functions which define a 'Figure of Merit' (FoM) for a given classification model (Brink et al., 2013). In this section a number of these FoMs utilised in the rest of the thesis are defined.

The production of a machine learning model always begins with the collection of a set of data as these algorithms are driven by learning from data. To evaluate the learned model a portion of the data must be set aside for the testing phase of a completed model. Often an additional set of data is selected for validating the hyperparameters of a model as doing this on the training data can introduce a bias into the model. Popular splits include 60% training set, 20% testing set and 20% validation set.

In some machine learning problems (such as the one addressed in this thesis) a lack of usable high quality training data can require the use of a technique named k-fold Cross-Validation where the performance of the model is determined through the splitting of the data into k sets. The machine learning algorithm then learns k distinct models using k different $k - 1$ subsamples of the data and leaving the remaining split of the data for testing. The test results on each of the models are then aggregated into a final performance measure.

Upon the successful training of a machine learned model with a set of available testing data, there are a number of performance metrics which define the success of assigning objects to the correct class relative to the objects assigned to the incorrect class. The first four metrics to highlight are the true positives (TP) which are the correctly classified members of the class C_i, true negatives (TN) which are the correctly classified non-members of the class C_i, false positive (FP) which are the incorrectly classified non-members of the class C_i assigned to this class and false negative (FN) which are the incorrectly classified members of the class C_i not assigned to this class. The TP, TN, FP and FN are often displayed in the form of a confusion matrix, a table of the predicted classes against the actual classes. These four metrics are used to produce the first FoM, the accuracy defined in equation 4.19.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{4.19}$$

This is a very simple measure of the number of TPs and TNs (correct classifications) relative to the total number of objects which is the sum of all four metrics. This metric is popular but highly limiting as it makes no assumptions about the distribution of the classes therefore it can claim a model is performing extremely well when it is ignoring a minority class completely. To address the class imbalance two new metrics are introduced named sensitivity and specificity. These two metrics are a measure of a classifiers performance at detecting all members which belong to a class whilst minimising the number of falsely classified members proportionally to the number of members in each class. The sensitivity is defined in equation 4.20 and the specificity is defined in equation 4.21 (Sokolova et al., 2006).

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{4.20}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4.21}$$

The Receiver Operating Curve (ROC) is a 2D representation of the cut-off values of sensitivity and specificity. This is a simple method of showing the possible performance of these two metrics as a function of the probability cut on a given classification model required to assign an object to a class C_i. The ROC is shown in equation 4.22 (Sokolova et al., 2006).

$$\text{ROC} = \frac{P(x|C_i)}{P(x|\bar{C}_i)} \tag{4.22}$$

where $P(x|C_i)$ is the probability of the feature vector x being of class C_i and $P(x|\bar{C}_i)$ is the probability of the feature vector x being not of class C_i. The value $i \in \mathbb{I} \in [0, N]$ is an integer value where N is the total number of classes in the problem.

The Area under the Curve metric is a measure of the area under the ROC curve. The AUC is determined by equation 4.23 which determines a compromise value where the sensitivity is maximised relative to the specificity (Sokolova et al., 2006).

$$\text{AUC} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \tag{4.23}$$

This value describes the best possible correct classification rate whilst minimising incorrect classifications where the weighting of correct and incorrect classifications are identical. In many cases this is not the case. In this pipeline, minimising the false positives tends to be more important than a full recall of all true members of a class. This is due to the computational difficulties inherent to the processing of a large dataset of light curves. It is often better to risk not detecting potential variable objects than become laboured on a large number of falsely flagged objects. Despite this fact, the AUC can still be useful in the identification of high quality models as well as selecting suitable cut probabilities for the final classification pipeline.

There is an alternative metric which automatically incorporates this weighting between detectability and the purity of the classification named the F-Score or F$^\beta$-Score (Sokolova et al., 2006). The F-Score is defined using a modification of the sensitivity and specificity named precision and recall. Precision is defined in equations 4.24 and recall is equal to the sensitivity.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.24}$$

The F-Score is then defined by equation 4.25.

$$\text{F}^\beta = \frac{(\beta^2 + 1) \times \text{Precision} \times \text{Recall}}{\beta^2 \times (\text{Precision} + \text{Recall})} \tag{4.25}$$

where the term β defines the weighting of the precision and recall. When $\beta > 1$ the precision is favoured by the evaluation metric and when $\beta < 1$ the recall is favoured by the evaluation metric. When $\beta = 1$ precision and recall have equal weighting similar to the previous metrics and the F-Score becomes known as the F1-Score and is a harmonic mean between the two measurements.

The ROC curve of a given classification model can also be used to determine the probability cuts for optimal performance on individual classes using the Index of Union (IU) measure (Unal, 2017). This is a recently proposed method which utilises the AUC, sensitivity and specificity of a classification model to determine these cuts and can be applied to multiclass problem. This is achieved through the computation of multiple ROC curves using a one-verses-many approach where for each class C_i the objects classified as a member of class C_i are placed into a positive class and all other objects are placed into the negative class regardless of which other class they were classified. This technique is of interest in producing high purity classifications in the automated pipeline and it is enhanced with an additional parameter named offset which acts similar to the β in the F-Score. The enhanced Index of Union is defined in equation 4.26.

$$\mathrm{IU}(c) = (|\mathrm{Sensitivity}(c) + \mathrm{offset} - \mathrm{AUC}| + |\mathrm{Specificity}(c) - \mathrm{offset} - \mathrm{AUC}|) \tag{4.26}$$

where $\mathrm{IU}(c)$ is the Index of Union at cut probability c, $\mathrm{Sensitivity}(c)$ is the Sensitivity at cut probability c, $\mathrm{Specificity}(c)$ is the Specificity at cut probability c. The Index of Union is computed across a fine resolution probability spectrum from 0 to 1 for each class C_i and the probability c which minimises the Index of Union function is the optimal cut probability for the C_i class. This is repeated for the N classes in the multiclass problem to determine the optimal cut probabilities for every class.

Finally, there is a metric named the Brier Score. The Brier Score is best described as a 'mean squared error of classification' and uses the ground truth class probability vector and the predicted probability vector summed over multiple observations (Brier, 1950). The Brier Score is shown in equation 4.27.

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}(y_{ij} - \hat{y}_{ij})^2 \tag{4.27}$$

where $\mathrm{BS} \in \mathbb{R} \in [0,1]$ is the Brier Score, y_{ij} is the ground truth probability of the i^{th} observation being a member of the j^{th} class, $\hat{y}_{ij}$ is the predicted probability of the i^{th} observation being a member of the j^{th} class, C is the total number of possible classes and N is the total number of observations. When the Brier Score has a value of 0 the model performance is perfect and at 1 the model performance is completely incorrect. The method does require a significant number of observations to correctly measure the performance of rare events such as the presence of an uncommon class in a dataset (Wilks, 2010).

Chapter 5

Feature Extraction

The variability indices subsection and the Training Classification Models section are in press with the Expert Systems with Applications Journal with submission ID: ESWA-D-18-00934 and the Fourier Decomposition subsection has been published in the Lecture Notes in Computer Science Journal (McWhirter et al., 2016) with title 'A dynamic, modular intelligent-agent framework for astronomical light curve analysis and classification'.

Chapter one introduced the concept of light curves, time-series photometry of light sources, as well as the variations these light curves can exhibit due to different variable star types. The STILT reduction pipeline has produced an easily queried database of Skycam objects with their associated observations (Mawson et al., 2013). To further process the STILT database observational data, a method is devised to describe individual light sources using a number of fundamental parameters that describe the photometry recorded in their light curves. The STILT database contains data from both the SkycamT and SkycamZ instruments. As discussed previously, SkycamT has a larger FoV which improves the resampling of light curves whereas SkycamZ sacrifices this resampling rate for increased depth. In this trade-off between depth and resampling, the SkycamT instrument is the clear victor. SkycamT has over three times the number of sky sources with just under three times more observations. SkycamZ has a number of well sampled fields but most of the sky is not sufficiently sampled for the production of a good set of cross-matched training objects. SkycamT has a much higher number of well sampled light curves with 590,492 individual objects with more than 100 data points. For this reason, the development of this automated pipeline makes use of only the SkycamT data as to perform the analysis on a homogeneous survey catalogue.

The determination of high quality features which describe the statistics of unevenly sampled light curves relevant for classification problems is a multiple decade old problem (Stetson, 1996; Richards et al., 2011b; Debosscher et al., 2007). Multiple research groups have taken data from inhomogeneous surveys and investigated which descriptors provide superior performance in a time-intensive procedure named feature engineering (Protopapas et al., 2006; Debosscher et al., 2007; Richards et al., 2011b; Deb and Singh, 2009; Tanvir et al., 2005; Yoachim et al., 2009; Shin et al., 2009; Prsa et al., 2008; Paegert et al., 2014; Pichara et al., 2016; Butler and Bloom, 2011; Kim and Bailer-Jones, 2016). There are currently hundreds of possible features which analyse the shape, distribution and variation of light curves of various performances depending on the target survey data and the desired classification problem (Richards et al., 2011b; Nun et al., 2015; Kim and Bailer-Jones, 2016). These features also exhibit differing computational overheads for their determination (Richards et al., 2011b). With modern astronomical surveys the volume of data continues to increase and demands ever more sophisticated yet computationally-efficient analysis techniques. This is clearly an optimisation problem for future data reduction pipelines as a good balance must be found between the computational power and time required to analysis new data combined with the ability to extract interesting signals from the deluge of data. Another concern is the automation of such reduction pipelines as human intervention must be extremely limited due to the volume of necessary decisions. In this chapter the current extent of light curve feature engineering developed by multiple research groups over the previous decade is explored with results generated from SkycamT data. The results of this experiment demonstrate that the currently available methods are insufficient to accurately classify the SkycamT light curves. Two new features are also introduced in the form of the Quick LSP Period and Quick LSP P-value features. These novel features use the strength of the Skycam aliases to quickly define the strength of periodic activity in the light curves without requiring a computationally intensive period search across the entire parameter space.

5.1 Variability Detection

Variability detection is the problem of identifying and separating candidate variable light curves from the large population of non-variable light curves (Shin et al., 2009). This can be treated as a binary classification problem or an anomaly detection problem depending on the utilised methodology (Nun et al., 2014; Pichara et al., 2016). As variable detection is an initial component to the process of identifying variable objects, it must be computationally inexpensive due to its application to many light curves (Shin et al., 2009; Nun et al., 2015). Therefore the features chosen for this problem tend to be quickly computable statistics of the light curve data points such as their skewness, dispersion and alignment in time (Richards et al., 2011b). In this section, the set of variability indices used in current generation surveys are discussed. Additionally, the Quick Lomb-Scargle features are introduced. These features are designed to improve the detection of short period variables which can be neglected by the former indices in the presence of noisy data.

5.1.1 Variability Indices

These features are in many ways easier and less computationally intensive to produce as they tend to involve the use of simple statistical functions applied to one or more of the light curves time instants, magnitude measurements and magnitude errors (Richards et al., 2011b; Nun et al., 2015).

The first of the variability indices is skewness (Nun et al., 2015). Skewness is a measure of asymmetry of the magnitude measurements around the mean magnitude measurement. Put in the context of light curves, it is a measure of whether an objects light curve features brightening events, dimming events or, like in the case of a periodic pulsating star, both brightening and dimming events as it oscillates around its mean brightness (Richards et al., 2011b). Skewness is calculated using Equation 5.1.

$$b = \frac{\frac{1}{n}\sum_{i=1}^{n}(m_i - \mu)^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(m_i - \mu)^2\right]^{\frac{3}{2}}} \tag{5.1}$$

where b is the estimated value of the skewness, m_i is the magnitude value of observation i, μ is the weighted (by the reciprocal of the magnitude errors) mean of the magnitudes and n is the total number of observations. Due to pulsating stars having skewness close to zero and eclipsing binaries with their strong dimming periods having a positive skewness, this feature is expected to be extremely useful in separating these two classes of objects (Richards et al., 2011b).

The standard deviation is expected to be an important feature in discriminating between variable and non-variable objects as the variable objects will have a larger deviation caused by a signal and noise component whereas non-variable stars will only have a noise component (although in reality all stars will have a signal component caused by Earth-based daily and seasonal light level variations). It is calculated using Equation 5.2 (Nun et al., 2015).

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(m_i - \mu)^2} \tag{5.2}$$

where σ is the estimated value of the standard deviation, m_i is the magnitude value of observation i, μ is the weighted mean of the magnitudes and n is the total number of observations.

Beyond 1 Standard Deviation is another important feature that describes the fraction of photometric magnitudes that lie above or below one standard deviation from the weighted mean of the photometric magnitudes. It is similar to the standard deviation as it is a representation of the proportion of data points whose magnitude varies far from the weighted mean (Richards et al., 2011b; Nun et al., 2015).

Small Kurtosis is a feature which measures how peaked a distribution is, or alternatively described, how closely clumped the data values are to the mean (Nun et al., 2015). This calculation is performed on a vector of magnitude values using Equation 5.3.

$$k = \left(\frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{m_i - \mu}{\sigma} \right)^4 \right) - \frac{3(n-1)^2}{(n-2)(n-3)} \tag{5.3}$$

where k is the estimated value of the small kurtosis, m_i is the magnitude value of observation i, is the weighted mean of the magnitudes, σ is the standard deviation of the magnitudes and n is the total number of observations.

The weighted mean is retained as a non-periodic feature although it is not expected to be as informative as other features (Richards et al., 2011b; Nun et al., 2015). This is due to the mean brightness not revealing much information about the brightness of the actual object (which would be useful) as all these objects are at various distances from the observer and the brightness of light decreases as a function of the distance to the light source squared. It is calculated using Equation 5.4.

$$\mu = \frac{\sum_{i=1}^{n} (m_i w_i)}{\sum_{i=1}^{n} (w_i)} \tag{5.4}$$

where μ is the weighted mean of the magnitudes, m_i is the magnitude value of observation i and w_i is the weighting of observation i.

Mean variance is a feature defined as the ratio of the standard deviation to the weighted mean magnitude (Shin et al., 2009; Richards et al., 2011b; Nun et al., 2015). This is easily calculated using Equation 5.5 as the standard deviation and weighted mean have already been established. If a light curve has a strong variability the value of this feature will generally assume a larger positive value.

$$V = \frac{\sigma}{\mu} \tag{5.5}$$

where V is the mean variance of the observed magnitudes, μ is the weighted mean of the magnitudes and σ is the estimated value of the standard deviation.

Variability Index is a feature that strongly indicates if there is a variable signal present in the data and is a robust standard deviation indicator (Shin et al., 2009; Richards et al., 2011b; Nun et al., 2015). It is defined as the ratio of the mean of the square of successive differences to the variance of the data points. This means that it is resilient to noise as only a signal propagating throughout the light curve will cause a notable change in this ratio. It checks if successive data points are independent or if they are linked by a common signal. It is capable of detected trends throughout the light curve and modified to take into account the uneven sampling of astronomical light curves. The variability index is calculated as indicated in Equation 5.6.

$$\eta^e = \bar{w}(t_{N-1} - t_1)^2 \frac{\sum\limits_{i=1}^{N-1} w_i(m_{i+1} - m_i)^2}{\sigma^2 \sum\limits_{i=1}^{N-1} w_i} \qquad w_i = \frac{1}{(t_{i+1} - t_i)^2} \tag{5.6}$$

where w is a weight term that decreases the contribution of data points that are further apart, $\bar{w}$ is the mean of these weights, m_i is the magnitude of the observation i, t_i is the time instant of the observation i and σ^2 is the variance of the magnitudes. This equation is used to generate a second variability feature called the folded variability index. This is the same calculation but on the light curve folded at the dominant period detected by the periodogram. As a result, this feature cannot be calculated until the period has been determined (Richards et al., 2011b).

A number of the features so far are designed to describe normal or near normal distributions. Many astronomical objects are unlikely to fit this description and therefore additional features are needed to describe them. The first of these is named Q31 defined as the difference between the third and the first quartile of the magnitude values (Richards et al., 2011b; Nun et al., 2015). The first quartile is a magnitude value in which 25% of the data points have lower magnitudes and 75% have higher magnitudes. The third quartile is a magnitude value in which 75% of the data points have higher magnitudes and 25% have lower magnitudes.

Range of a cumulative sum is a feature that can detect the largest change in the partial sums of the magnitudes within a light curve (Richards et al., 2011b; Nun et al., 2015). The feature is defined as the difference between the maximum and minimum value of a vector of sums defined by Equation 5.7.

$$S = \frac{1}{N\sigma} \sum_{i=1}^{l} (m_i - \mu) \quad \text{for } l = 1, 2, \ldots, N \tag{5.7}$$

Like the variability index, there is a second feature using the equation of the range of a cumulative sum but applied to magnitudes sorted by phase instead of time using phases generated by the dominant period from a periodogram on the light curve (Richards et al., 2011b).

In addition to the kurtosis measure, Stetson devises a second robust kurtosis based feature for the determination of descriptive parameter for the light curves of Cepheid variables (Stetson, 1996). Variable objects with highly sinusoidal light curves will exhibit more data points at the maximum and minimum amplitude that at the mean whereas for a light curve with a constant signal and Gaussian noise, more data points are expected near the mean compared to the minimum and maximum amplitudes (Shin et al., 2009; Richards et al., 2011b; Nun et al., 2015). Stetson kurtosis is determined as follows.

$$k_S = \frac{\frac{1}{N}\sum\limits_{i=1}^{N}|\delta_i|}{\sqrt{\frac{1}{N}\sum\limits_{i=1}^{N}\delta_i^2}} \qquad \delta_i = \sqrt{\frac{n}{n-1}}\frac{m_i-\mu}{\sigma} \tag{5.8}$$

where N is the number of observations, m_i is the magnitude of the observation i and σ is the standard deviation of the magnitudes. This feature is often referred to as Stetson-K. There are three additional features devised by Stetson which have the names Stetson-I, Stetson-J and Stetson-L (Stetson, 1996; Shin et al., 2009; Richards et al., 2011b; Nun et al., 2015).

Maximum slope is a feature that gives the value of the gradient of maximum absolute deviation between two consecutive observations (Richards et al., 2011b; Nun et al., 2015). This is a simple and quick calculation defined in Equation 5.9.

$$\max\left[\frac{m_{i+1}-m_i}{t_{i+1}-t_i}\right] \quad \text{for } i = 1, 2, \ldots, N-1 \tag{5.9}$$

Amplitude is a feature that describes the largest consistent deviation in variability from the mean (Richards et al., 2011b; Nun et al., 2015). In a pulsating object, it is interpreted to be the extent of the brightness changes between the stars brightest or dimmest state at its mean state. In order to include consistent deviations and not fit noise, which would just produce the largest difference in the magnitudes with both signal and noise components, the median of the top and bottom 5% of measurements, is used (Richards et al., 2011b). This is calculated as shown in Equation 5.10.

$$a = \frac{\text{median(top 5\%)} - \text{median(bottom 5\%)}}{2} \tag{5.10}$$

Consecutive points or *con* is defined as the proportion of consecutive data points in the light curve where 3 data points have magnitudes atleast two times the standard deviation away from the mean magnitude (Shin et al., 2009; Richards et al., 2011b; Nun et al., 2015). This is calculated as shown in Equation 5.11.

$$\text{con} = \frac{C}{n-3} \tag{5.11}$$

where C is the number of consecutive sets of 3 data points with magnitudes greater than two standard deviations and n is the number of data points.

Median Absolute Deviation is a measure of the median discrepancy of the data (Richards et al., 2011b; Nun et al., 2015). It is another method of describing the distribution of magnitudes in a distribution-generic way and is shown by Equation 5.12.

$$\text{MAD} = \text{median}(|m - \text{median}(m)|) \tag{5.12}$$

Median Buffer Range Percentage measures the fraction of photometric magnitudes within the amplitude divided by 10 of the median magnitude (Richards et al., 2011b). Therefore, it is a measure of what fraction of data points lay close to the median magnitude. It is expected that a variable object would have a lower fraction of points near this value. To simplify the calculation, the inverse is calculated, i.e. the fraction of data points lying outside of the amplitude divided by 10 above and below the median magnitude and then subtracted from one.

The pair slope trend is a feature that considers the last thirty time-sorted measurements of observed magnitudes and determines the fraction of increasing first differences minus the fraction of decreasing first differences (Richards et al., 2011b). It counts the number of intervals between the last thirty measurements where the slope is positive and divides it by the number of intervals where the slope is negative.

There are a set of five features based on the Flux Percentile Ratio. These features describe the ratio between different percentiles of the magnitudes and the difference between the 95^{th} and the 5^{th} percentiles (Richards et al., 2011b). For example, a value named $F_{a,b}$ is defined as the difference between the b^{th} percentile and the a^{th} percentile. We divide this percentile difference $F_{a,b}$ by the percentile difference between the 5th and 95th percentiles $F_{5,95}$. This is used to define five features by substituting in percentiles for the a and b variables. These features are the Flux Percentile Ratio mid20 $F_{40,60}/F_{5,95}$, the Flux Percentile Ratio mid35 $F_{32.5,67.5}/F_{5,95}$, the Flux Percentile Ratio mid50 $F_{25,75}/F_{5,95}$, the Flux Percentile Ratio mid65 $F_{17.5,82.5}/F_{5,95}$ and the Flux Percentile Ratio mid80 $F_{10,90}/F_{5,95}$. They describe the distribution of brightness across the central magnitude band of a light curve with five differently sized magnitude bands.

Flux Percentile Ratio mid20 has a small central band whereas Flux Percentile Ratio mid80 uses a large band covering most measured magnitudes.

Percentile Amplitude is a feature that determines the largest percentage difference between the minimum or maximum magnitude and the median magnitude (Richards et al., 2011b). Equation 5.13 shows the Percentile Amplitude calculation.

$$\mathrm{PA} = \frac{\max(|m - \mu|)}{\mathrm{median}(m)} \tag{5.13}$$

where m is the vector of magnitudes and μ is the mean magnitude of the light curve. Next, the Percent Difference Flux Percentile is a feature defined as the ratio of the percentile magnitude difference $F_{5,95}$ over the median magnitude as shown in Equation 5.14.

$$\mathrm{PDFP} = \frac{F_{5,95}}{\mathrm{median}(m)} \tag{5.14}$$

The Anderson-Darling statistic is a statistical test of whether a given sample of data is drawn from a given probability distribution, in this case, a normal distribution (Richards et al., 2011b). It is a powerful statistical tool for detecting most departures from normality. The statistic is rescaled using Equation 5.15 to emphasise departures from normality. Normal distributions tend to return a value of 0.3 and any departures from normality rapidly tend towards one.

$$A = \frac{1}{1 + e^{-10a-0.3}} \tag{5.15}$$

where a is the Anderson-Darling normality statistic and A is our rescaled statistic.

The final two features describe light curve correlation between points in time. The first is named the autocorrelation function length (Richards et al., 2011b). The autocorrelation function can be used to determine the linear dependence of a signal with itself at two points in time. It is a vector of values that describes the similarity of the signal with itself against time lags. These time lags are integer values based on the units of the time instants (in the case of Skycam light curves, these are units of days). For strongly periodic signals, it would be expected to see a slow fall off as the signal deviates from itself until a lag that aligns with the period where the value would rise again. Equation 5.16 shows the autocorrelation function.

$$\hat{\rho}_h = \frac{\sum_{t=h+1}^{T} (m_t - \bar{m})(m_{t-h} - \bar{m})}{\sum_{t=1}^{T} (m_t - \bar{m})^2} \tag{5.16}$$

For an observed series $m_1, m_2, \ldots, m_T$ with a sample mean $\bar{m}$ and a sample lag h. It should be noted that $\hat{\rho}_h$ is a vector of values. A feature must be a single value descriptor and therefore a mechanism of generating this value must be defined. This is where the 'length' component of autocorrelation function length is determined. It is defined as the length of the autocorrelation function as the lag value h where $\hat{\rho}_h$ drops below the value of e^{-1}.

The last feature is called the slotted autocorrelation function length (Richards et al., 2011b). The slotted autocorrelation function is an improvement on the standard autocorrelation function by implementing the lags as slots (intervals) instead of single time values. The slotted autocorrelation function at a certain time lag slot is computed by averaging the cross product between samples whose time differences fall in the given slot as shown in Equation 5.17.

$$\hat{\rho}(\tau = kh) = \frac{1}{\hat{\rho}(0)N_\tau} \sum_{t_i} \sum_{t_j = t_i + (k-0.5)h}^{t_i + (k+0.5)h} \bar{m}_i(t_i)\bar{m}_j(t_j) \tag{5.17}$$

where h is the slot size, $\bar{m}$ is the normalised magnitude, $\hat{\rho}(0)$ is the slotted autocorrelation for the first lag and N_τ is the number of pairs of observations that fall in the given slot. Again the length of the slotted autocorrelation function is defined as the lag value where $\hat{\rho}(\tau = kh)$ becomes smaller than e^{-1}.

We include the Renyi Quadratic Entropy feature (Huijse et al., 2012). This is an information theoretic estimator of the randomness of a set of data points. In this case, this is the magnitude values and their associated errors sorted by the time instants of measurement. These can be used to calculate the Information Potential shown in Equation 5.18.

$$\mathrm{IP} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left[\frac{1}{\sqrt{2\pi}\sigma_m} e^{-\frac{(m_i - m_j)^2}{2\sigma_m^2}} \right] \tag{5.18}$$

where N is the number of data points, σ_m is the median of all the magnitude errors and m_i is the magnitude value, sorted by time instant at time instant i. This is used to calculate the Renyi Quadratic Entropy using Equation 5.19.

$$\mathrm{RQE} = -\log(\mathrm{IP}) \tag{5.19}$$

Additionally, there are two features that allow the differentiation between stellar sources and quasars, short for quasi-stellar objects, the active cores of faraway galaxies. Variations in quasars can produce similar features to variable stars therefore new features are needed to distinguish them. The features are called QSO and Non-QSO (Butler and Bloom, 2011; Dubath et al., 2012). QSO is a statistic that identifies the Bayesian probability of the source being a Quasar. There have also been methods designed to classify QSO light curves from more standard features (Kim et al., 2011). Other purpose built models have also been developed that can characterise quasi-periodic variations using a damped random walk model (Zinn et al., 2017).

Characterising variability from non-periodic features is challenging, as the Period is an important property of many variable objects. This challenge can prove worthwhile, as the extraction of periodic information is computationally expensive. Neuropercolation, a family of probabilistic biological inspired models, has also been used to characterise quasi-periodic light curves (Elzo et al., 2016). Kepler light curves were used to identify periodic activity through the determination of the width of a histogram across a characteristic timescale (Neff et al., 2014). Care must also be taken in regards to the selection of features that may be of interest to the classification of astronomical transients. These light curves rarely display periodic activity and are usually of short duration therefore automated methods must have access to features that correctly describe these properties (Disanto et al., 2016).

5.1.2 Quick Lomb-Scargle Indicies

The variability indices presented above are designed to be easily computed as to be applicable to a large set of light curves for the detection of variable light curves independent of noise and sampling effects (Richards et al., 2011b; Nun et al., 2015). Due to the Skycam cadence resulting in large sampling gaps, many of these features are sensitive to longer period light curves such as the variability index and the autocorrelation function length. The Skycam database contains a large number of short period variables with sufficient sampling and signal to allow for correct period estimation and classification.

Other research has indicated that performing a period estimation procedure, even if low resolution and obtaining an incorrect answer, can still provide useful information in the identification of variability across a wide range of periods (Shin et al., 2009). Unfortunately, due to the number of data points present in the Skycam data, utilising even the fastest of period estimation methods was still unfeasible (McWhirter et al., 2016). Initial attempts to compensate for this involved subsetting the light curves into a smaller set of data points but this was unreliable due to many light curves requiring the total number of data points to have a sufficient confidence in the candidate period. Additionally, the exact selected subset of data points was not obvious and rerunning the method with multiple subsets defeats the benefit of subsetting. Computing the periodogram with a smaller frequency spectrum also produced a poor performance as many of the catalogued periods were unsampled.

Two possible solutions are proposed for the production of a low computation period estimation for this task. The first method involves applying the bands method discussed

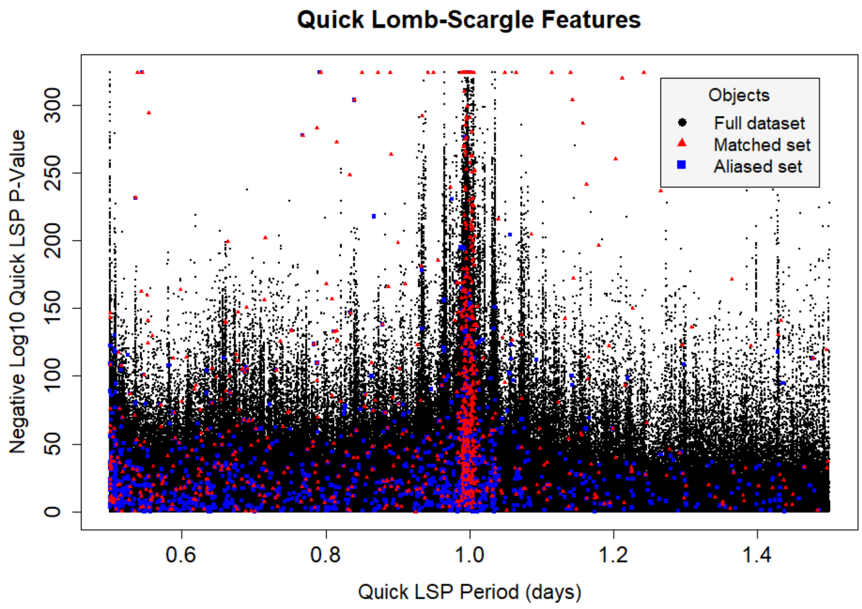


FIGURE 5.1: Plot of the two Quick-LSP features. Whilst not a definitive method of identifying variable light curves, especially with the aliasing due to the Skycam cadence, there are interesting structures inside the data of benefit to a variability detection classification model.

in chapter 2 using a Lomb-Scargle Periodogram (Protopapas et al., 2015). With a sufficiently small value of N_t and N_b, the frequency spectrum can be reduced to a very rapidly computable size whilst maintaining the performance shown in the bands method experiments in chapter 2.

The second method utilises the strong diurnal sampling window of the Skycam data. In chapter 2, the alias failure mode was defined as a ‘reflection’ of the true period around the sidereal day period (VanderPlas, 2017). As a result, most periodic light curves will produce a strong response in the periodogram at the one day alias of the true period shown by equation 5.20.

$$P_{\text{alias}} = \left(P_{\text{true}}^{-1} \pm t_{\text{sid}}^{-1}\right)^{-1} \tag{5.20}$$

where P_{alias} is the aliased period, P_{true} is the underlying true astrophysical period and t_{sid} is the sidereal day at 0.99726957 days.

This method relies on this aliasing to reduce the range of the period search from 0.5 days to 1.5 days. The Lomb-Scargle Periodogram can then maintain the required resolution to sample the period range without introducing a substantial computational runtime. Two additional variability indices are generated by this period estimation, the period estimated by the maximum peak of the Lomb-Scargle Periodogram and the negative base-10 logarithm of the p-value of the maximum peak as calculated by the False Alarm Probability (Scargle, 1982; VanderPlas, 2017). Using double precision, the precision of the p-value is limited to 10^{-323} and values below it are rounded to zero. As the negative base-10 logarithm of zero is positive infinity, all objects with p-values of zero are set to the value of 324. These two features are named the Quick Lomb-Scargle Period and Quick Lomb-Scargle p-value and indicate the presence of a strongly periodic signal regardless of the true period. Figure 5.1 demonstrates the scatterplot of the Quick LSP features determined by the 590,492 well-sampled SkycamT light curves overlaid by red data points representing the 859 catalogue period matched objects and 1186 catalogue objects with aliased periods. The identification of variability is not definitive using these two features but they do give a good indication of potential variables with strong periodicity. The many periodic peaks in the p-values across this range show the extent of the intensive aliasing caused by the Skycam cadence.

As the bands method has an associated computation time which rivals the time required for the complete computation of the Quick Lomb-Scargle features, the Quick Lomb-Scargle features are added to the variability indices defined above to produce the complete set of features used by the automated pipeline to detect variable light curves in the Skycam data.

5.2 Variable Star Classification

The variability indices have been developed for the task of discriminating variable and non-variable light curves; however, this is not the limit of their usefulness. Many of these features are useful for the classification of different classes of variable star (Richards et al., 2011b). For example, as noted above, skewness should be capable of distinguishing eclipsing binary light curves from pulsating variable light curves. Richards et al. make use of many of the above variability indices to supplement their variable star classifiers (Richards et al., 2011b, 2012).

Useful as the variability indices are for classifying variable light curves, they lack full descriptions for light curve shape and periodicity. In the previous chapters the importance of the period in the determination of variable stars was discussed in detail. The classes used by modern astronomers for variable stars are very much driven by their period and colour. As a result, the period is clearly an important feature for this task which is not present in the variability indices due to the increased computational load in its determination as shown in chapters 2 and 3 (Debosscher et al., 2007).

The period is more than just an important feature in this task as it can be used as the gateway to transforming the light curves into new representations such as using the epoch folding method defined in chapter 2. The period is used as the basis for a selection of features named ‘periodic’ features which use a candidate period returned from a period estimation algorithm to define additional properties of the light curve such as its amplitude and shape (McWhirter et al., 2017). The documented variable star classes from chapter 1 clearly show that the shape of the light curve can vary significantly based on the underlying astrophysics even if the amplitudes and periods are similar. In this section the set of periodic features developed over the last decade is introduced and described demonstrating the use of Fourier models to extract amplitude and phase information from light curves developed by Debosscher et al. (Debosscher et al., 2007) to the additions from two of the dominant sources of engineered periodic features, Richards et al. and Kim & Bailer-Jones (Richards et al., 2011b, 2012; Kim and Bailer-Jones, 2016).

5.2.1 Fourier Decomposition

Fourier decomposition makes use of the property of Fourier series. A Fourier series is a method of approximating any continuous univariate function as a sum of an infinite number of sine and cosine functions (Fourier, 1878). The amplitudes and phases of these sine and cosine components define the shape of the resulting Fourier series. A truncated Fourier series is a finite number of these infinite components which will result in an imperfect approximation with an accuracy depending on how many components are retained (Debosscher et al., 2007; Deb and Singh, 2009; Richards et al., 2011b, 2012; Kim and Bailer-Jones, 2016). For light curve modelling, the univariate time-series can be fit using regression by a truncated Fourier series with properties determined by the developer based on the quality and complexity of the light curves. The amplitudes and phases can be extracted from the regressed model and form a set of features that describe the shape of the light curve (Debosscher et al., 2007).

The method we utilise on the SkycamT light curves is a variant of the approach introduced into Machine-Learned periodic light curve classification by Debosscher et al (Debosscher et al., 2007) and improved with a regularisation technique by Richards et al. (Richards et al., 2011b, 2012). To model periodic and multi-periodic light curves, a period estimation algorithm is applied to the light curve time-series data. Any efficient and reliable period estimation method may be chosen for Fourier decomposition as it only requires the real-valued period. Upon the determination of the candidate period, a harmonic model with a linear trend is fitted with this period across the time-series data. In order for the model to have the degree of freedom required to accurately fit the data, artificial data points were bound into the time-series. These artificial data points have a uniform distribution in time and a magnitude equal to the mean magnitude of the time-series. The artificial data points must be assigned zero weight as to not contribute to the fitted model. As the weights are defined as the reciprocal of the magnitude error, the artificial data points are assigned a magnitude error of positive infinity. Weighted linear regression is performed on this new time-series to fit an m-harmonic sinusoid model using the period detected by the periodogram and is demonstrated by Equation 5.21. This model has ten coefficients in our method as we use $m = 4$ to sufficiently model the light curve.

$$y(t) = ct + \sum_{j=1}^{m} \big[a_j \sin(2\pi j f_1 t) + b_j \cos(2\pi j f_1 t)\big] + b_0 \tag{5.21}$$

Where b_0 is the mean magnitude of the light curve and c is the linear trend of the time-series. a_j and b_j are Fourier coefficients for the fitted model. f_1 is the frequency from the periodogram. This model is subtracted from the time-series in a process called pre-whitening to eliminate any periodic activity within the time-series based on the dominant period detected by the periodogram. This pre-whitened time-series is processed with another period estimation method to identify a second period independent of the first dominant period. A harmonic model is fitted for this period and subtracted off in a second pre-whitening phase. Finally, a third period is identified independent to the first two periods. The time-series is then restored to its original, archived prior to the pre-whitening operations. A harmonic best-fit is computed using weighted linear regression and a model with twenty six coefficients. This is shown in Equation 5.22 when $n = 3$. It is possible to perform additional pre-whitening operations for additional model complexity although overfitting is a concern.

$$y(t) = ct + \sum_{i=1}^{n} \sum_{j=1}^{m} \big[a_{ij} \sin(2\pi j f_i t) + b_{ij} \cos(2\pi j f_i t)\big] + b_0 \tag{5.22}$$

The linear trend of the time-series, calculated alongside the sinusoidal model is retained as a feature of the object. By calculating this linear trend in the same regression operation as the sinusoidal models, a time-series with a non-integer number of wavelengths within the sampling period (with a corresponding trend) will not interfere with the linear trend caused by a gradual brightening or dimming of the object, an important feature. The frequencies f_i and the coefficients a_{ij} and b_{ij} are retained to provide a good description of the light curve as long as it is periodic and accurately reproducible as a sum of sinusoids. These coefficients are not time-translation invariant and are transformed into better descriptors of the light curve. This was accomplished by transforming Fourier coefficients into a set of amplitudes A_{ij} and phases PH_{ij} and are determined according to Equations 5.23 and 5.24 respectively (Debosscher et al., 2007).

$$A_{ij} = \sqrt{a_{ij}^2 + b_{ij}^2} \tag{5.23}$$

$$PH_{ij} = \arctan\left(\frac{b_{ij}}{a_{ij}}\right) \tag{5.24}$$

The phases are not time-translation invariant and are defined relative to PH_{11}, the phase of the first harmonic of the dominant period using Equation 5.25.

$$PH_{ij}^{'} = \arctan\left(\frac{b_{ij}}{a_{ij}}\right) - \left(\frac{jf_i}{f_1}\right)\arctan\left(\frac{b_{11}}{a_{11}}\right) \tag{5.25}$$

The phases were then constrained between $-\pi$ to $+\pi$ by the transformation as indicated in Equation 5.26. Please note that for simplicity, the double dash is dropped through the rest of the paper.

$$PH_{ij}^{''} = \arctan\left(\frac{\sin\left(PH_{ij}^{'}\right)}{\cos\left(PH_{ij}^{'}\right)}\right) \tag{5.26}$$

For light curves that are primarily monoperiodic, this results in the production of 28 features that are time-translation invariant allowing direct comparison between light curves measured at different phases (Debosscher et al., 2007). Monoperiodic light curves are those that oscillate with one dominant period (Aerts et al., 2006). This assumption does not hold for all potential variable stars but as the primary period is usually highly dominant in multi-period variables, this assumption is a good approximation (Debosscher et al., 2007). These features include the slope of the linear trend, the three frequencies used in the final harmonic model, the twelve amplitude coefficients and eleven phase coefficients (as PH_{11} is always zero it is discarded) and the ratio of data variance (called variance ratio) between the variance before the pre-whitening of the harmonic model of the primary period and after. This statistic is a strong indicator of the importance of the primary period to the light curve relative to the other periods.

The Fourier decomposition method is powerful but is sensitive to noise and outliers. Reducing the number of model parameters is a solution to preventing overfitting on noise but can also result in a loss of information on complicated astrophysical signals. An alternative method is the use of L2 regularisation (Richards et al., 2012). L2 regularisation is a technique applied to optimisation methods such as the harmonic regression used in fitting the Fourier models. It is a second term to the least squares minimisation used in the harmonic regression which places a penalisation on the higher harmonic amplitudes and phases. If the higher harmonic components attempt to fit on noise, the penalty will cause their parameters to asymptotically approach zero eliminating their effect. However, in the case of fitting a complicated signal with sufficient supporting evidence in the light curve, the higher harmonic parameters will resist the regularisation and maintain their contribution. Equation 5.27 demonstrates the optimisation function

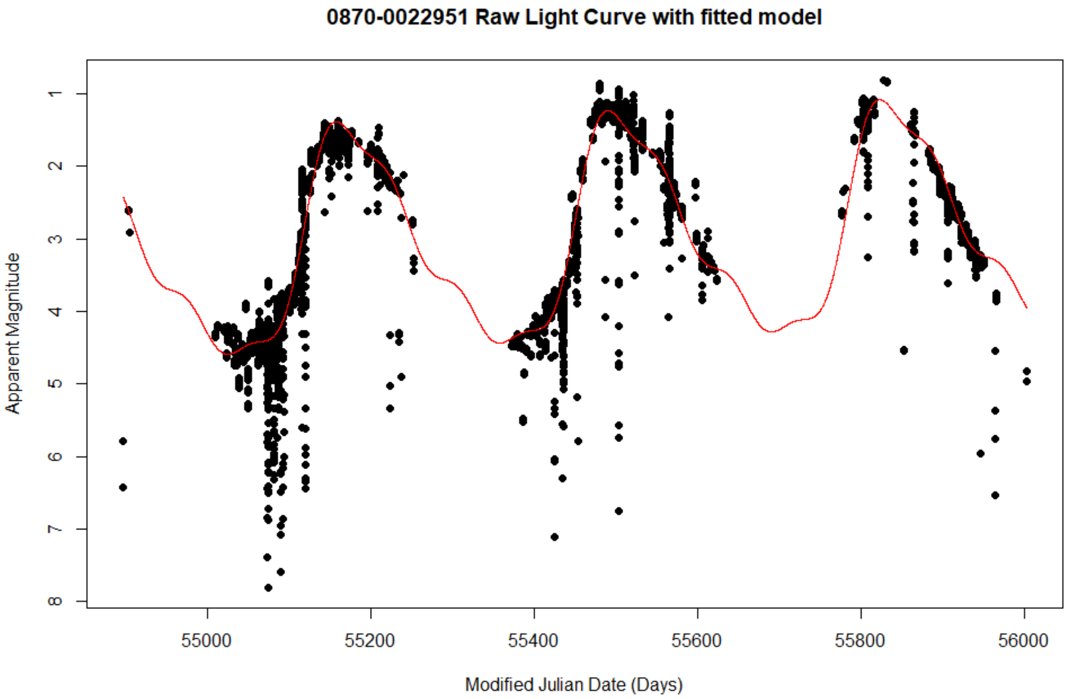


FIGURE 5.2: The light curve of the star Mira with a regularised harmonic fit with a primary period of 332.57 days determined by the GRAPE method.

to be minimised in this regularised regression fitting.

$$R(\theta, \lambda) = \sum_{i=1}^{N} \frac{(d_i - m_i)^2}{\sigma_i^2} + N\lambda \sum_{n=1}^{4} n^4 \left(A_n^2 + B_n^2\right) \tag{5.27}$$

Where θ is the model parameters, λ is the regularisation parameter, N is the number of light curve points, d_i are the photometric data points, m_i are the model data points and $\sqrt{A_n^2 + B_n^2}$ is the amplitude of the n^{th} Fourier harmonic. The value of the regularisation parameter allows the control of the smoothing of the model with small values allowing the modelling of high frequency structure and large values smooth this structure out. For the SkycamT light curves the regularised fit is applied with a regularisation parameter of 0.01 determined by manual inspection of the resulting models across a set of known variable light curves. In the event that the regularisation procedure fails, the light curves are fit with a non-regularised procedure at the risk of overfitting. The regularisation is performed using the normal equation shown in equation 5.28.

$$\theta = \left(X^\top X + \lambda W\right)^{-1} X^\top y \tag{5.28}$$

Where X is the 'design matrix' of the problem, a matrix with the Fourier components in columns and the data points in rows, λ is the regularisation parameter, W is the regularisation weights determined from equation 5.27, y is the vector of data point values and $\top$ is the transpose operator. This also provides an additional benefit of making the $X^\top X + \lambda W$ invertible as the normal equation cannot be solved if $X^\top X$ is non-invertible, also known as a *singular* matrix.

Figure 5.2 demonstrates the model produced using described method for the SkycamT data collected on the star Mira, the prototype of the Mira class variables showing a clear sinusoidal oscillation. The period of Mira has been widely reported as 332 days, verified by surveys such as Hipparcos (Bedding and Zijlstra, 1998). In the event of the periodogram returning a result similar to the stars correct period, linear regression can produce an accurate model despite the prevalence of noise within the time-series data.

5.2.2 Richards et al. features

Richards et al. utilised a large number of features in their machine learned classification models for the All-Sky Automated Survey (ASAS) (Richards et al., 2012). Whilst most of their features have already been discussed in the variability indices and Fourier decomposition subchapters, they did introduce a number of novel features to characterise the performance of period estimation methods in addition to descriptors of the phase positions of eclipse features for eclipsing binary classification. Eclipsing binaries are not well modelled by sinusoidal models therefore the Fourier decomposition is of limited use as higher harmonic components are required for good fits yet these parameters are heavily regularised.

There are five eclipsing binary features which have been calculated from the Skycam light curves by making use of the PolyFit algorithm discussed in chapter 6. This allows the phases of the maxima and minima to be more carefully isolated than if extracted from the Fourier harmonic model. They are defined as:

- Eclipse Max Delta Mags

  This feature determines the absolute value of the magnitude difference between the two maxima of the phase-folded light curves at 2× the candidate period. These maxima correspond to the primary and secondary eclipses for an eclipsing binary. For most eclipsing binaries this value will be non-zero as the eclipses have differing heights whereas for a light curve folded at 2× the true period, the two maxima should be the same producing a feature of zero.

- Eclipse Min Delta Mags

  This feature determines the absolute value of the magnitude difference between the two minima of the phase-folded light curves at 2× the candidate period. This feature measures the relative out-of-eclipse brightness difference between the primary and secondary eclipse and the secondary and primary eclipse. For most eclipsing binaries this is expected to be zero unless an additional effect is present such as the distortion wave of an RS Canum Venaticorum variable.

- Eclipse Phase Ratio

  This feature is used to define the eccentricity of an eclipsing binary based on the phases of the primary and secondary eclipses. It is defined as the ratio between the phase difference of the first minimum and first maximum associated with the primary eclipse and the phase difference between the second minimum and second maximum associated with the secondary eclipse. This feature takes on values near unity for highly symmetrical eclipsing binaries. Values diverging from unity suggest the presence of either an eccentric eclipsing binary or some other class of non-eclipsing variable.

- Reference Phase

  This last feature is present to define the location of the reference phase, the phase associated with the pre-primary eclipse out-of-eclipse brightest magnitude. This feature defines the performance of the PolyFit algorithm at fitting the phase-folded light curve.

- Period Double Ratio

  The Period Double Ratio is a new feature we have defined which is also designed for assisting in the identification of light curves where the period estimation method has produced a period at half of the true astrophysical period. The period estimated by the GRAPE method is fine-tuned through the use of the Variance Ratio Periodogram, a multi-harmonic period estimation which can correctly model non-sinusoidal signals such as the eclipsing binary light curve. The Period Double Ratio is defined as the ratio between the variance ratio of the candidate period and the variance ratio of twice the candidate period. This calculation is shown in equation 5.29.

$$P_{\text{rat}} = \frac{V_{\text{rat}}(P)}{V_{\text{rat}}(2P)} \tag{5.29}$$

  where P_{rat} is the Period Double Ratio and $V_{\text{rat}}(x)$ is the variance ratio calculated for a four harmonic Fourier fit determined by a candidate period x. If this feature is greater than unity, the better fitting model is the one generated by the candidate period indicating the light curve is either a pulsating star or a very close contact binary. On the other hand, if the feature is less than unity, the light curve is likely an eclipsing binary with a better fitting multi-harmonic model at twice the candidate period.

Richards et al. also introduced a set of features designed to describe the performance of the Fourier harmonic model in fitting the light curves (Richards et al., 2012). Using statistics that quantify the normality and scatter of the residuals of the fit calculated by equation 5.30, light curves with complexity beyond the modelling capability of the harmonic models can be identified. These features have been adopted from the work of Dubath et al. and Kim et al. from their classification methods (Dubath et al., 2011; Kim et al., 2011).

$$r_i = m_i - \hat{m_i} \tag{5.30}$$

where r_i is the residual of the i^{th} data point, m_i is the magnitude of the i^{th} data point and $\hat{m_i}$ is the predicted magnitude of the i^{th} data point as determined by the harmonic Fourier model. There are two features that quantify the statistics of the residuals of the harmonic Fourier model defined as:

- Residual Normality

  This feature is produced by applying the Anderson-Darling test discussed above to the set of residuals calculated by equation 5.30. For light curves with signal and Gaussian noise, the application of a good-fitting harmonic model to the light curve should leave purely Gaussian noise which will return a highly normal distribution. For real light curves such as the SkycamT light curves, the correlated noise does reduce the effectiveness of this test but it might still be of some use to the classifiers.

- Residual Raw Scatter

  The Residual Raw Scatter determines the ratio of the range of the spread of magnitudes of the residuals to the initial amplitude. If the residuals have values across a large range of magnitudes it suggests either a poor fit by the harmonic model or the light curve has a large noise contribution. As the light curves are assumed to have similar noise statistics, the classifier will interpret the relative values of this feature as a measure of the goodness-of-fit of the harmonic model. Residual Raw Scatter is defined by equation 5.31.

  $$r_{\text{MAD}} = \frac{\text{med}\left(|r - \text{med}(r)|\right)}{\text{amp}} \tag{5.31}$$

  where r_{MAD} is the Residual Raw Scatter, med is the median operation, r is the vector of residual magnitudes and amp is the amplitude of the initial light curve.

There is one final feature named Squared Differences over Variance. This feature is defined as the sum of the squared magnitude differences in successive measurements divided by the variance of the light curve (Kim et al., 2011). It is the equivalent to the unnormalised variability index so it is much more sensitive to the changes in well sampled light curves. The normalised variability index is likely more useful for light curve classification but this is implemented regardless and can be removed by feature selection operations. Equation 5.32 demonstrate the calculation of this feature.

$$\text{sdv} = \frac{1}{\sigma^2} \sum_{i=1}^{N-1} (m_{i+1} - m_i)^2 \tag{5.32}$$

where sdv is the Squared Diferences over Variance feature, σ is the standard deviation of the magnitudes of the light curve, N is the total number of data points in the light curve and m_i is the magnitude value of the i^{th} data point.

5.2.3 Kim & Bailer-Jones et al. features

In 2016 Kim & Bailer-Jones published a new package for the classification of variable stars named UPSILoN: A**U**tomated Classification for **P**eriodic Variable **S**tars using Mach**I**ne **L**ear**N**ing (Kim and Bailer-Jones, 2016). This package uses a machine learning classification model trained using Random Forest classifiers from a high quality set of OGLE and EROS-2 data. The classification performance was then tested on a set of data from the MACHO, LINEAR and ASAS surveys achieving reasonable cross-survey performance. The periodic light curves were defined using a set of 16 features including period and Fourier decomposition components. A number of the features have been already implemented such as period and a set of variability indicies such as skewness, kurtosis, the Stetson-K index, the interquartile range and the folded variability index.

The remaining features range from modifications of the Fourier decomposition components to novel features which describe a previously unmeasured light curve property. In the Fourier Decomposition method there are multiple features describing amplitude and phase. Kim & Bailer-Jones identified that the ratio or difference between these features was sufficient to describe the primary properties of the light curves of the differing classes. Their method uses a Fourier model with the primary period and four harmonics of which they retain the period and two harmonics to produce two amplitude ratios and two phase differences as well as retaining the amplitude of the primary period fit.

These are defined as:

- R_{21}, Ratio between the 2nd and 1st amplitudes from the Fourier decomposition.
- R_{31}, Ratio between the 3rd and 1st amplitudes from the Fourier decomposition.
- ϕ_{21}, Difference between the 2nd and 1st phases from the Fourier decomposition.
- ϕ_{21}, Difference between the 3rd and 1st phases from the Fourier decomposition.

There are two more features of interest defined by a previous study (Long et al., 2012):

- m_{p10}, 10% percentile of slopes for 2× candidate period, phase-folded light curve.
- m_{p90}, 90% percentile of slopes for 2× candidate period, phase-folded light curve.

These features are calculated by calculating the slopes between each set of data points phased at a candidate period and sorted by phase as shown in equation 5.33.

$$\delta_i = \frac{\phi_{i+1} - \phi_i}{m_{i+1} - m_i} \quad \text{for } i = 1, 2, \ldots, N - 1 \tag{5.33}$$

where ϕ_i is the phase of the i^{th} data point and m_i is the magnitude of the i^{th} data point. mp_{10} and mp_{90} are then simple generated by taking the 10% percentile and 90% percentile respectively.

The work by Kim & Bailer-Jones also makes use of a different normality statistic to the Anderson-Darling method used as one of the variability indices. Their normality statistic is named the Shapiro-Wilk normality test. This method has advantages over the Anderson-Darling method as it has been shown to perform slightly better using a Monte Carlo analysis (Razali and Wah, 2011). Despite this, the implementation we made use of was limited to a maximum of 5000 data points and therefore was not applicable to many of the Skycam light curves. We therefore do not implement the Shapiro-Wilk test and remain with our original statistic. Due to the correlated noise in the Skycam data, this feature is unlikely to be of much use due to no light curve approximating a normal distribution.

5.3 Training Classification Models

To demonstrate the capability of these sets of features for classification, we selected 2326 variable star light curves from the STILT database to compute and evaluate machine learned models on. These light curves have over 100 observations per object and have been cross-matched to the AAVSO variable star catalogue with a tolerance of 3.6”. These light curves have been manually classified into eight classes as shown in Table 5.1 along with the number of light curves. There are four primary classes, with the eclipsing binaries split into three subtypes, contact, semi-detached and detached. The Long Period Variables have also been split into Mira class variables and Semiregular variables. The utilised features are heavily influenced by the work of Richards et al. (Richards et al., 2011b) which have been integrated into a new Feature Analysis for Time Series (FATS) library produced by the Institute for Applied Computational Science at

TABLE 5.1: The class distribution of the STILT 2326 variable light curves.

Class	Type	Acronym	Count
1	Delta Cepheid Variables	DCEP	132
2	Delta Scuti Variables	DSCT	499
3	Eclipsing Binaries – Detached	EA	683
4	Eclipsing Binaries – Semi-detached	EB	242
5	Eclipsing Binaries – Contact	EW	291
6	Mira-Type Long Period Variables	M	149
7	RR Lyrae Variables	RR	114
8	Semiregular Long Period Variables	SR	216

Harvard University (Nun et al., 2015). They are used in this light curve classification problem. This experiment utilizes 62 features designed to differentiate between the eight classes in Table 5.1. The methodology used to create these models is based on traditional shallow machine learning methods using the Random Forest algorithm with a layout identical to the flowchart shown in figure 4.1.

5.3.1 Feature Selection

There is likely an amount of redundancy of information in the 62 features for the optimal discrimination of these eight classes. Machine Learning classifiers have been found to suffer from a performance penalty because of this redundancy although Random Forest classifiers are resilient due to their nature as a set of weak classification trees (Brink et al., 2013). Therefore, features that contain less useful information for this classification task need to be eliminated (Donalek et al., 2013). We created a backwards-selection algorithm based on a previous method that uses a Random Forest classifier on different subsets of the features to iteratively deselect them (Brink et al., 2013). This algorithm makes use of a cross-validation step of variable size and repeats. In this experiment, we used a 5-fold cross-validation with two repeats to train a set of models to discriminate between the features. The 5-fold cross-validation involves a randomisation in the order of the 2326 light curves. They are separated into five sets. These sets are stratified, meaning they have approximately the same proportion of each class in each subset as in the complete collection. These sets are used to validate five random forest models. One of the subsets is selected as a validation set and the other four are used to train the learned model. This is performed five times for each subset. The 2326 light curves are re-randomised for the second repeat and the same process is repeated.

Each validation step is used to produce a figure of merit (FoM), a metric that defines the success of each model on the successful classification of all eight classes. We chose a multi-class Receiver Operating Characteristic (ROC) Area under the Curve (AUC) statistic that reflects the overall performance on all the classes. By maximizing this metric, we

Table 5.2: The 16 selected features from the Mean Decrease Gini method.

Number	Name	Mean Decrease Gini
1	Mean Magnitude	146.66110
2	Range of a Cumulative Sum	102.65558
3	Folded Range of a Cumulative Sum	99.57324
4	Amplitude of first frequency, first harmonic	99.07806
5	Q31 Interquartile range	96.34338
6	Percent Difference Flux Percentile	96.26259
7	Variance Ratio	95.24125
8	Median Absolute Deviation	92.05929
9	Mean Variance	90.49209
10	Folded Variability Index	85.90732
11	Stetson Kurtosis Measure	82.42626
12	Skewness	81.37099
13	First identified (dominant) Frequency	78.97057
14	P-Value of first frequency	78.68567
15	Autocorrelation Length	77.47186
16	Renyi Quadratic Entropy	76.93127

can select the best model for overall performance. Cross-validation is performed for a set number of times, based on the number of features, with each step having one of the features removed. The cross-validation that returns the best FoM is then used to determine which feature to remove from this iteration, namely the feature removed from that particular cross-validation. We have selected a cut-off number of features of three in this experiment.

For 62 features, this process can be computationally expensive. A solution to this is to drop the number of models generated in each cross-validation but this was found to decrease the performance of the algorithm due to variance in the AUC of the trained models. Instead we opted for a second rougher feature selection method to initialise the feature selection by quickly eliminating a large number of the features in a single cross-validation step. This is accomplished through a single 10-fold cross-validation with three repeats on all 62 features. This produced 30 Random Forest models. Each model has a statistic called the Mean Decrease Gini. It is a measure of how each feature contributes to the nodes and leaves in the numerous trees of the random forest model. It can be used as a simple method of describing the importance of the features in the learned models. For each of these thirty models, which all differed slightly due to their individual dataset sampling, we took the mean of all thirty individual Mean Decrease Gini measures for each of the 62 features. We selected 16 features that have the highest mean value. Sixteen features are arbitrarily chosen due to the computational expense in the backwards selection algorithm and was not a required value. These 16 features are displayed in Table 5.2. The backwards selection algorithm is performed on these 16 features with the iterative removal of one feature at a time down to three final features.

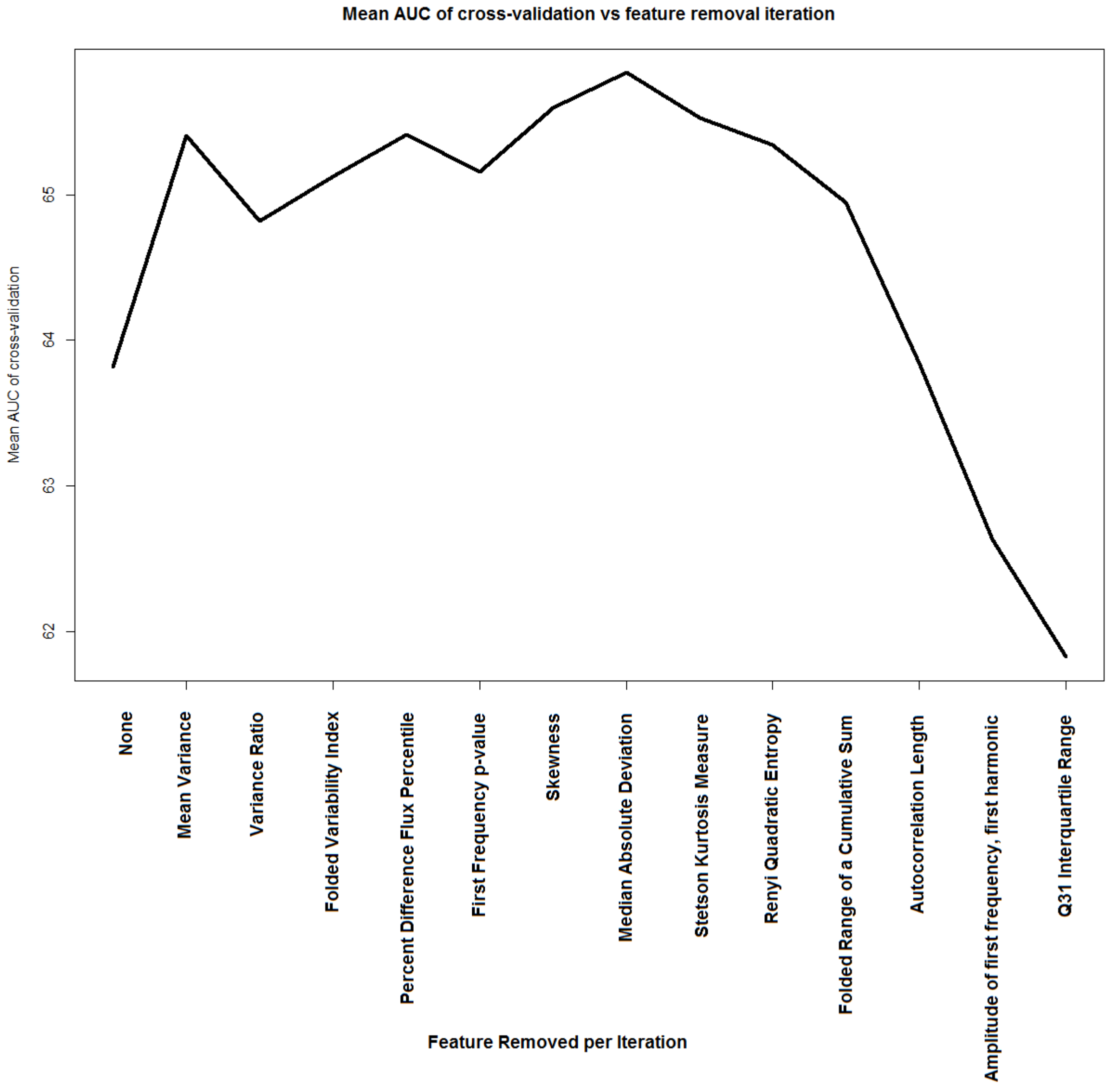


Figure 5.3: The 16 selected features from the Mean Decrease Gini method removed iteratively using the backwards selection random forest algorithm. Peak performance attained after the removal of the Median Absolute Deviation feature.

The performance of these models is expected to improve for the first number of steps whilst poorer features are removed from the learned models until a threshold where useful information starts being lost too resulting in a performance decrease. This was observed after 7 of the 16 features (iteration 8) had been eliminated as shown in Figure 5.3.

From this feature selection, we retained 9 features from the original 62 that appear to be the most informative for our dataset. The features selected by this method exhibit a few interesting results. Many of our folded-light curve representations have been rejected. These are usually expected to strongly identify the individual signals of the different classes and carry information on additional periodicities. The variance ratio statistic performs a similar function. However, for many classes, the period detected by the

Lomb-Scargle Periodogram fails to align with the reference AAVSO period. For many objects it fails to identify a period inside the normal range for their associated class of object. Only the Mira, Delta Cepheid and RR Lyrae classes have a reasonable fraction of correct period estimations. This explains why the dominant frequency and dominant amplitude features remain in the selected set although clearly with importance metrics that are substantially weaker than would be expected.

For many of the classes in our dataset, a good performance is expected from these two features alone (Richards et al., 2011b). Therefore, many of our surviving features are from the non-periodic (independent) set, which do not require correct period estimation. The mean magnitude feature exhibited the highest Mean Decrease Gini value when tested with the initial feature selection method. This feature is more dependent on the distance between the light source and the Earth than on the actual luminosity of the object. However, the CCD instrument does have an associated performance based on the apparent magnitude of a light source. It is possible that combined with one or more of the other features, the mean magnitude is communicating useful information about the quality of the light curves. The performance of the Mean Magnitude could also indiciate there is a selection bias present in our training data. This feature is strong at identifying our Mira-type Long Period Variable training objects. These stars are large, bright stars (Percy, 2008) and are easily detectable due to their large brightness changes. Their increased brightness allows them to be detectable at larger distances from Earth allowing this class to become dominant at the brighter magnitudes in our training dataset.

5.3.2 Random Forest classification models

With the selection of nine features for our classification task, we then performed a 10-fold cross validation with three repeats on the 2326 STILT light curves. We made use of a random forest classifier as it was found to produce state-of-the-art results using these features (Richards et al., 2011b). We used the same random forest algorithm utilised in the backwards feature selection to explore the hyper-parameters more thoroughly. The Random Forest algorithm uses three primary parameters in order to generate the classification trees used for the learning task. mtry defines the number of variables that are randomly sampled as candidates for each branching in the classification trees. ntrees defines the number of decision trees generated in the random forest. Finally, nodesize is a parameter that adjusts the size of the trees terminal node in terms of the number of training objects populating each node. These tuning parameters smooth the learning problem by affecting the complexity of the decision boundaries estimated by the random forest ensemble (Brink et al., 2013).

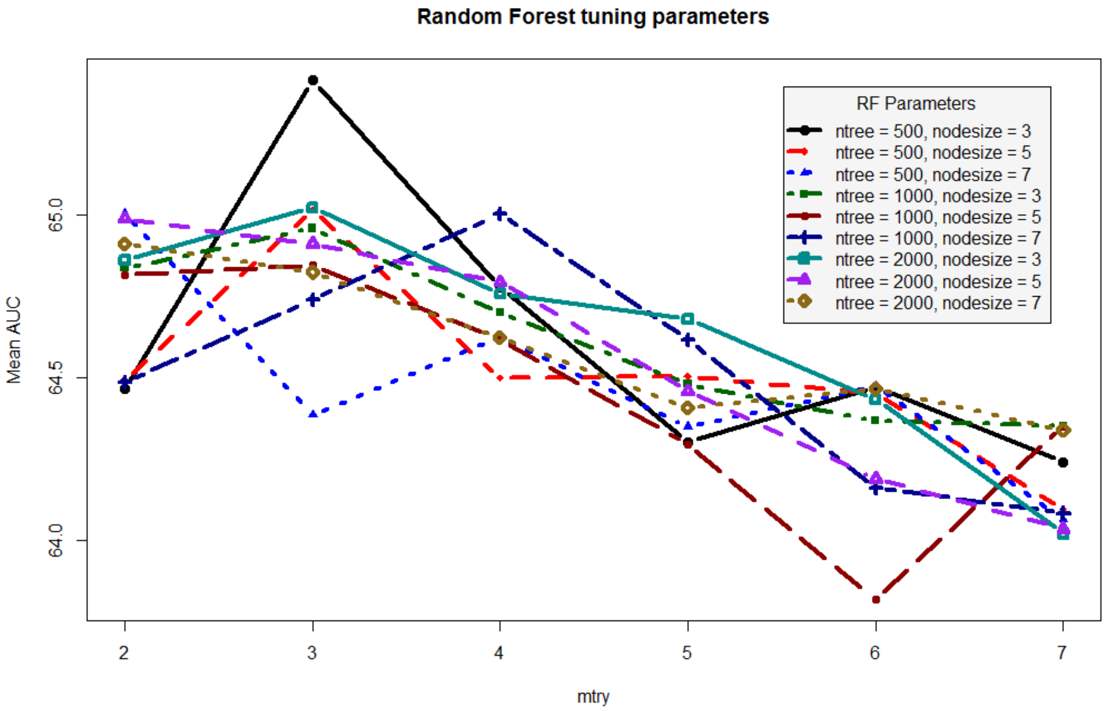


FIGURE 5.4: The 9 selected features utilised in a cross-validation grid-search tuning of the random forest hyper-parameters. Higher data-points represent superior performing models. Our best models were generated with ntree = 500, mtry = 3 and nodesize = 3.

Maintaining our figure-of-merit from the feature selection, we utilise the mean AUC of our cross-validation to estimate the performance of each hyper-parameter in a grid-based search. We utilised 10-fold cross-validation with three repeats for each hyper-parameter combination and deployed 108 combinations. This encompasses every possible configuration of mtr from two to seven, nodesize from two to seven and three configurations of ntree, one with 500 trees, one with 1000 trees and the final with 2000 trees. We find that our optimal performance is found at ntree = 500, mtry = 3 and nodesize = 3. All of our cross-validation models, regardless of hyper-parameter value are stable in terms of the standard deviation of the individual cross-validated model AUCs. Figure 5.4 demonstrates the performance of nine different combinations of the ntree and nodesize hyper-parameters when plotted on a graph of the cross-validation mean AUC against the mtry hyper-parameter. This performance is found to be fairly stable with similar performance at slight deviations from the optimal hyper-parameters.

With the hyper-parameter tuning completed, we recreate 10-fold cross-validation with three repeats of our optimal hyper-parameter setup. We generate confusion-matrices and one-verses-many ROC curves for the eight classes in our dataset to understand the individual performances of each class. We demonstrate one of the thirty, trained models here that is typical of the standard performance of the cross-validated models. This specific model had an overall multiclass AUC of 61.74%. The confusion matrix

Table 5.3: The confusion matrix of the specific 61.74% AUC model on the associated validation set. Mira type stars (M) perform well but many of the other classes are heavily miss-classified likely due to the class-assignment probabilities being undefined and therefore defaulting to whatever class has the highest probability for each test candidate.

Prediction/Reference	DCEP	DSCT	EA	EB	EW	M	RR	SR
DCEP	4	1	3	1	1	0	0	2
DSCT	1	17	15	7	4	0	1	3
EA	3	17	26	9	12	0	2	5
EB	0	3	2	1	1	1	1	1
EW	0	4	11	3	9	0	3	0
M	0	1	0	0	0	12	0	0
RR	0	2	3	1	2	1	3	0
SR	6	5	9	3	0	1	2	11

produced for the validation set for this model is shown in Table 5.3. Only the Mira type variable stars exhibited a high sensitivity and specificity, statistics that reflect the rate of false positives and false negatives for each class. This is due to the class-assignment metric being purely the highest-class probability for a given test candidate. This is not the optimal method for assigning class to a candidate as it risks many false positive and false negatives, as it can be noticed from the confusion matrix in Table 5.3. The method to correct these errors is to accept that for some light curves, it might not be possible to assign a class label and they are best left as unknown. This is accomplished by studying the ROC curves of each individual class and deciding an appropriate false positive rate that minimizes these miss-classifications at a cost of reducing the true positive rate, i.e. the number of objects of that class in our dataset that we do identify instead of assigning them unknown. Inspecting the ROC curves reveals that a number of our classes have reasonable performance as long as a reasonable false positive rate is determined. Figure 5.5 shows the one-verse-many ROC curve for the 61.74% AUC model. These curves reflect the performance of each class weighted against the miss-classification of every other class. Some interesting patterns are present. Firstly, the Mira type light curves are well classified by the model. This is not surprising as they have larger variability amplitudes than the other classes and the selected features tended to notably diverge from the remaining classes. Five of the eight classes had reasonable results on the selected set of features.

The more common misclassifications also share some interesting patterns. Four of these five classes, Delta Cepheids, RR Lyraes, Mira type variables and the contact eclipsing binaries all exhibit light curve signals that can be reasonably well modelled by the sinusoidal Lomb-Scargle Periodogram. Therefore, it is likely that the dominant frequency feature for these classes is substantially better. The fifth class with reasonable performance, the semi-regular variables, can have sinusoidal signals too but tend to be a bit

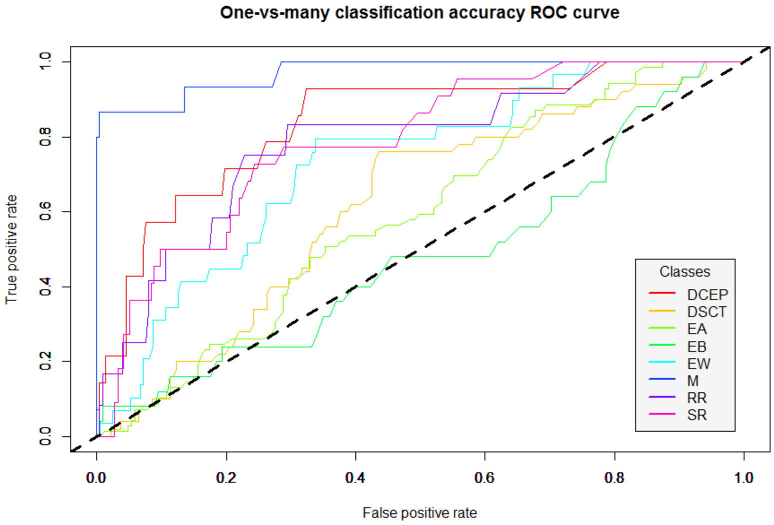


FIGURE 5.5: One-vs-many ROC curves for the 61.74% AUC model. Mira type variable stars (M) perform well with slightly poorer performance from Cepheids, Semi-regular variables and RR Lyraes. Eclipsing Binaries performed extremely poorly.

more variable in their periodicity too. In this case, they are have the second highest amplitude of variability in the set too which will clearly differentiate out of many of the other classes.

The detached and semi-detached binaries exhibited significantly worse performance. The signals of these variables tend to be highly non-sinusoidal resulting in a very poor performance of the dominant frequency feature. Additionally, they can have much wider ranges of period and amplitude than the pulsating variable stars resulting in greater difficulty for the learned model to identify a typical parameter set for these classes.

Finally, the Delta Scuti variable type light curves show poor performance. This is due to the amplitude of the brightness variations in these classes being minor, likely below the noise threshold, making these signals extremely difficult to detect. This results in a complete failure for the Periodogram to identify the underlying signal, even if it is highly sinusoidal and therefore the amplitude and period features end up producing incorrect results.

Chapter 6

Representation Learning Features

The 'Automated Extraction of Visual Representation Features' section has been published in the International Joint Conference on Neural Networks 2007 (IJCNN '07) proceedings (McWhirter et al., 2017) with title 'The classification of periodic light curves from non-survey optimized observational data through automated extraction of phase-based visual features'.

Chapter 5 presented a set of carefully engineered features for the description of light curves. The process of developing these features was powered by over a decade of work performed by experts in the analysis of light curves. The features they developed were designed for the current generation of survey data present at the time with many systems relying on OGLE (Udalski et al., 1997; Richards et al., 2011b), MACHO (Alcock et al., 2000; Kim et al., 2011), EROS2 (Rahal et al., 2009; Protopapas et al., 2015) and Kepler (LaCourse et al., 2015; Kugler et al., 2016; Matijevic, 2012; Parvizi et al., 2014; Neff et al., 2014) data. As a result, despite the best of intentions, biases have been introduced into the classification process. Firstly, the experts involved in this process will have individual and combined opinions on the properties that define light curves. These objects are often part of a continuous range of types with fairly ambiguous boundaries such as the Type II Cepheids which extend from the short period BL Herculis objects to the long period RV Tauri stars (Percy, 2008). Whilst period ranges have been used to define these classes, it can be difficult to distinguish between a long period BL Herculis variable and a short period W Virginis variable. Secondly, features that are designed for and perform well at classification tasks for one set of surveys does not guarantee good performance in later surveys with differing statistics (Benavente et al., 2017).

In regards to the classifiers, research has been conducted in mapping the models trained by one survey to the unclassified data of another such as considering different survey statistics to be a rotation in a higher dimensional feature space (Benavente et al., 2017). Other approaches aim to produce a set of highly capable classification models on a subset of object types with high performance and then combine them into a meta-classification model for improved multi-survey capability (Pichara et al., 2016). The experiment at the end of chapter 5 demonstrates that, whilst the features derived from the works of Richards et al. (Richards et al., 2011b) and Kim & Bailer Jones (Kim and Bailer-Jones, 2016) are useful in the classification of Skycam light curves, many of the statistical features are heavily under-represented in the resulting mean decrease Gini importance measurements. The physical reason for these features to be important for classification remains intact but the considerable noise in the Skycam data heavily poisons these features. As a result, research was conducted into the computation of a set of new features tuned for performance on the Skycam light curves. Representation Learning is a machine learning technique which extracts useful non-linear representation features of the raw data based on their performance at a given task, such as the classification of the variable star light curves (Bengio et al., 2013).

In this chapter we introduce two novel methods designed for the automatic extraction of light curve classification features specifically designed for performance on the Skycam database. The first method involves transforming the light curves in a 2D image representation which is then processed by a number of machine learning classifiers to identify groups of useful pixels. This method identifies features useful for the classification, but inferior to the previously engineered features therefore the second method was implemented using an interpolated PolyFit light curve model applied to an epoch folded light curve at the GRAPE period (Prsa et al., 2008). The application of a Principal Component Analysis method to the interpolated PolyFit model allows for the production of a finite set of features that describe the shape of the model. Additionally, the interpolated PolyFit model is used to produce replacements for the engineered features which suffer poor performance due to the noise by using the interpolation to ‘clean’ the noise from the light curve based on the assumption that the GRAPE estimated period is correct.

6.1 Automated Extraction of Visual Representation Features

In this section we introduce the initial results and problems from the application of the methods from these previous studies to the STILT observations and propose a method of automatically extracting shape-based features from the phase-folded light curves through the use of multiple learning algorithms trained to recognize visual features mirroring the methodology employed manually by astronomers. In the future additional topologies will

Table 6.1: Data Summary for the 2519 object SkycamT light curve dataset.

Class	Type	Acronym	Count
1	δ Cepheid (Classical Cepheids)	DCEP	132
2	δ Scuti	DSCT	499
3	Eclipsing Binaries	EB	1409
4	Long Period Variables	LPV	365
5	RR Lyrae Variables	RR	114

be introduced to further power this feature extraction and allow for the classification of super and sub-classes in a large hierarchical multiclass problem.

We present Random Forest, Support Vector Machine and Feedforward Neural Network models to classify 2519 SkycamT variable star light curves. We extract 16 features found to be highly informative in previous studies (Kim and Bailer-Jones, 2016) and achieve an area under the curve of 0.8495 using a feedforward neural network with 50 hidden neurons trained with stratified 10-fold cross-validation with 3 repeats. We propose using an automated visual feature extraction technique by transforming bin-averaged phase-folded light curves into image based representations. This eliminates much of the noise and the missing phase data, due to sampling defects, should have a less destructive effect on these shape features as they still remain at least partially present. There is also no need for feature engineering as the learning algorithms can learn shape features directly from the light curves. We produced a set of scaled images based on a threshold of data points in each pixel. Training on the same feedforward network, we achieve an area under the curve of 0.6348. By introducing the Period and Amplitude as features into this dataset therefore giving meaning to the dimensions of the image we show this improves to 0.7952. Our current models lack translational-invariance and the method may be better suited to specific sub-classification problems common in the variable object hierarchical multi-class problem.

Reliable class information is required for a subset of objects in the SkycamT database in order to test classification methods on these light curves. The optimal method to extract this class information is through a comparison between the SkycamT object data to a variable star catalogue. The American Association of Variable Star Observers (AAVSO) operates one of the largest and best-updated catalogues of nearby bright variable stars in the world, The AAVSO International Variable Star Index. This catalogue does not contain any of the AAVSO gathered light curves but it does contain data on 373,565 known variable stars including their name, coordinates in right ascension and declination and their currently identified class. The coordinates of these variable stars were matched to objects in the STILT database with a tolerance of 3.6" (seemingly sufficient to avoid detection collisions between nearby stars) and a minimum of 100 individual observations. This resulted in the production of 12,461 variable stars of various types. Five variable

Table 6.2: Kim et al. (Kim and Bailer-Jones, 2016) 16 variability features.

Feature	Description	Reference
Period	Period derived by the Lomb-Scargle Periodogram	Kim et al. 2014
Ψ^n	Variability Index of a phase-folded light curve	Kim et al. 2014
Ψ^{cs}	Cumulative sum index of a phase-folded light curve	Kim et al. 2014
R_{21}	Amplitude ratio of the 2nd and 1st Fourier Harmonics	Kim et al. 2014
R_{31}	Amplitude ratio of the 3rd and 1st Fourier Harmonics	Kim et al. 2014
Φ_{21}	Difference between 2nd and 1st phase	Kim et al. 2014
Φ_{31}	Difference between 3rd and 1st phase	Kim et al. 2014
γ_1	Skewness	Kim et al. 2014
γ_2	Kurtosis	Kim et al. 2014
K	Stetson K index	Kim et al. 2014
Q_{3-1}	Difference between 3rd and 1st quartiles	Kim et al. 2014
A	Ratio of magnitudes brighter or fainter than the average	Kim et al. 2016
H_1	Amplitude from Fourier decomposition	Kim et al. 2016
W	Shapiro-Wilk normality test	Kim et al. 2016
m_{p10}	10th percentile of slopes of a phase-folded light curve	Long et al. 2012
m_{p90}	90th percentile of slopes of a phase-folded light curve	Long et al. 2012

star super-classes were selected which describe a large number of known periodic variable object types. They were all well-represented in this dataset leaving 2519 corresponding objects. Table 6.1 demonstrates the class by class breakdown of this dataset.

6.1.1 Epoch-Folded Image Representation method

Following the selection of these 2519 variable light curves, the performance of the features used in previous high-performance general purpose classifiers was established. The 16 features used by Kim et al. (Kim and Bailer-Jones, 2016) were chosen for this operation as they had been shown to be capable of reliably separating super-classes as well as achieving respectable inter-class accuracy. These features are shown in Table 6.2.

The Lomb-Scargle Periodogram (Lomb, 1976; Scargle, 1982; Ruf, 1999) utilised in this method operated over a linear frequency grid from the reciprocal of the total observation time of a light curve up to 20 (cycles/day). The interval between candidate frequencies is shown in equation 6.1 where t_{max} and t_{min} are the last and first observation times respectively.

$$f_{\text{step}} = \frac{0.25}{t_{\text{max}} - t_{\text{min}}} \tag{6.1}$$

When calculated using the Lomb-Scargle Periodogram using light curves from the STILT dataset, the Period feature appears to have a correct match rate of under 5% relative to the AAVSO reference period (which is treated as the ground truth in this study) for many of the classes. As the period, calculated by the Lomb-Scargle Periodogram, is the basis in which phase-folded light curves are generated, this inaccuracy heavily pollutes an additional 9 features. This is over half the number of features used in this analysis. As for the non-folded features, the distribution of these features amongst the classes appears to centre at or near their expected means. However, the range is much greater than expected increasing the overlap between classes. This is likely a result of the larger-than-usual noise threshold in the STILT data (McWhirter et al., 2016). Classifiers trained using these polluted features resulted in models of accuracies only slightly better than the no-information accuracy, the expected result of a completely randomly trained model. This is primarily due to the Long Period Variable class as it exhibits long period sinusoidal variations with a high amplitude signal. Therefore, for the following analysis, the periodogram-derived period was replaced with the AAVSO reference period purifying the features shown in figure 6.1.

These 16 features have been shown to be fully capable of training useful classifiers in previous work however they required the fitting of Fourier models to the light curves. These Fourier models are used to generate features such as the amplitude ratios, the phase information and the amplitude of the first harmonic. In the method of Kim et al., utilised in this work, a five harmonic Fourier model has been fit to the light curves (Kim and Bailer-Jones, 2016). These models are very versatile but can suffer from three serious drawbacks on our dataset. Firstly, the sums of sinusoids used to assemble Fourier models fit some classes of astrophysical signal poorly such as the sharp dips associated with eclipsing binary stars (Richards et al., 2011b). Secondly, the more harmonics used in a Fourier model, the more complex the signal it can fit. Unfortunately this can also result in overfitting the light curve data causing the noise to have an unwanted contribution to any resulting features. Finally, as these Fourier models are being fit in the time dimension, the poorly sampled regions can cause the fit model to deviate outside of expected ranges again resulting in a non-signal contribution to the Fourier coefficients. The cadence concerns raised in the STILT dataset can cause this to become a considerable source of poor results.

It is important for the analysis method to address the above dangers as the extracted features are important. We propose, as the classes we are attempting to train models on are highly periodic, to transform the representation of the light curves into an epoch-folded representation and extract features. These features describe the shape of brightness changes through the dominant periodic variation by 'folding' all the gathered data points into one waveform. This is very useful in astronomy due to the limitations in gathering data. In fact, this is one of the most powerful techniques in eliminating sampling issues as long as the light curve does have a dominant period (Paegert et al., 2014). For non-periodic variable objects in astronomy, such as transient light sources, other approaches must be considered. In the case of many periodic variable objects, the

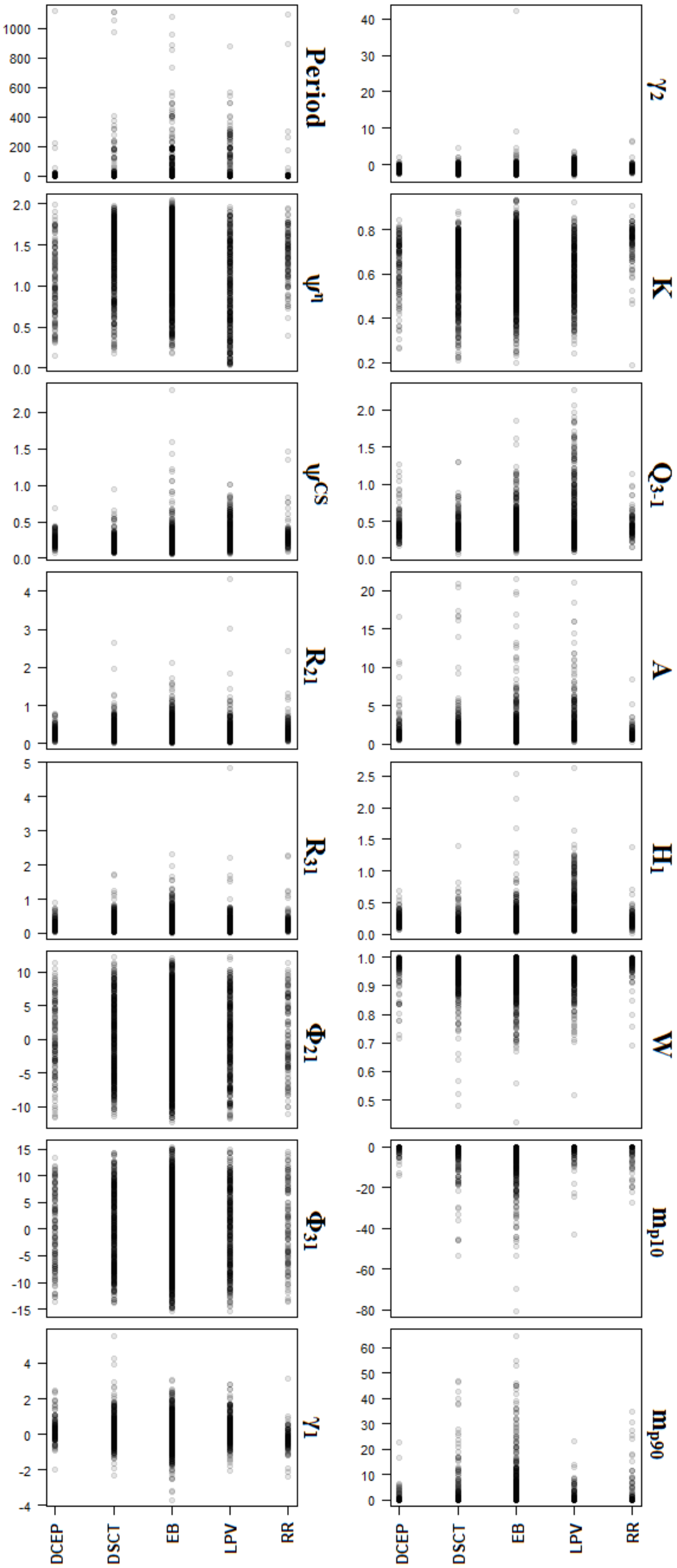


Figure 6.1: Plot of each of the 16 features against the five super-classes in the order shown in table 6.1. Many features appear to poorly differentiate the super-classes with light curves from the STILT dataset.

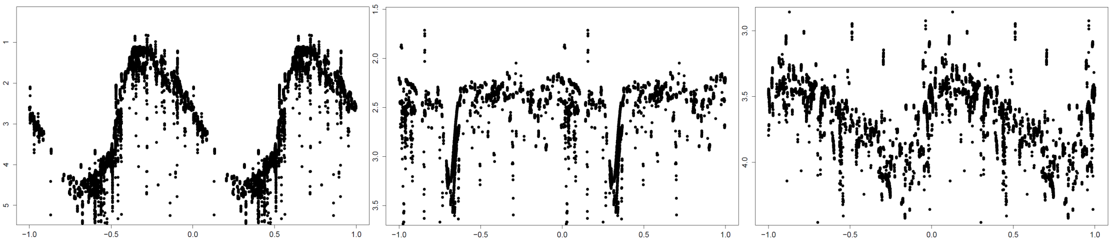


Figure 6.2: STILT dataset folded Light curves of the star Mira (Mira class) with a 332 day period, Algol (Algol-type eclipsing binary) with a 2.86 day period and Eta Aquilae (Delta Cepheid class) with a period of 7.18 days. The shape of each light curve is distinctive to the associated class.

shapes of the light curves in these phase-folded representations carry significant information about the class of the light source. Figure 6.2 shows an example of three of the light curves in this dataset, a Mira-type Long Period Variable, an Algol-type eclipsing binary and a Delta Cepheid. These light curves have been folded at the AAVSO period of the associated objects. Therefore, a light curve must clearly demonstrate these shape features in order for the 16 Kim et al. features (and many more) to extract enough of this information from any noise.

In this form, the poorly sampled regions due to cadence limitations in the original observations can be removed with the restriction of the dominant period must be a fraction of the complete observed time. The 16 features of the previous research does contain a number of features extracted from the phase-folded light curve however the Fourier model fit is performed in the time-space. We propose to replace the shape features from the Fourier model with new shape features automatically extracted from the folded light curves through applying machine learning algorithms directly to a visualised image-based representation of this folded light curve. In essence, allowing the learning of class-specific shapes.

The light curves in figure 6.2 are fairly typical of the better sampled light curves from the STILT database yet they do exhibit potential issues. Firstly, there are a lot of points with significant noise. This is possibly instrumental in nature but is much more likely due to a number of simplifications made to reduce the computational load of the pre-processing pipeline. This noise is likely to be the cause of both the larger range on the non-periodic features as well as the cause of the very low period match rate from the Lomb-Scargle Periodogram. Secondly, whilst the examples in figure 6.2 are well sampled across the whole phase space, there are other light curves that lack this due to the highly variable cadence of the STILT observations. This means that important shape features may only be partly present and not to the level required for the extracted features in previous studies.

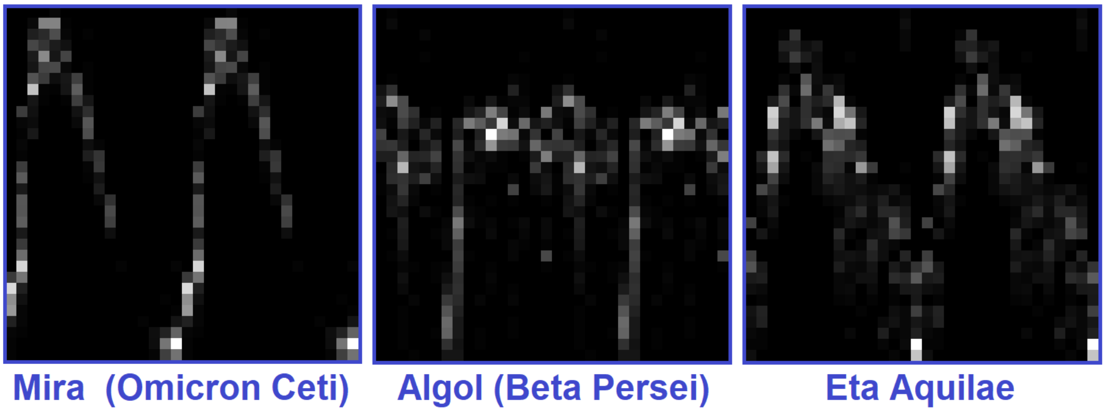


Figure 6.3: STILT dataset folded Light curves of the stars Mira (Mira class), Algol (Algol-type eclipsing binary) and Eta Aquilae (Delta Cepheid class) transformed into 28×28 pixel images. These images are in a two dimensional data format which can be accessed by machine learning algorithms designed for image recognition.

Yet, despite these obvious limitations, human astronomers can still look at these light curves and recognise the main shape patterns. Therefore it seems reasonable to conclude that even in the more poorly sampled, noisy STILT light curves, there are still features that have not yet been extracted which are being gathered for manual classification. Ideally the models used to fit the light curves should attempt to parameterise the shape of the actual variable object classes rather than some predefined or abstract form. This can be done by determining the specific form of different astrophysical signals directly from the astrophysics driving the variability of these object types. This is quite an undertaking further complicated by a lack of consensus about the dominant physical processes shaping these variabilities in many classes. Therefore, it would seem to be more appropriate to have a model that can identify the patterns in the shape of the light curves without requiring an underlying physically produced model. This can be accomplished through a learning process applied to visualized examples of light curves. Over the last decade, neural networks have been developed into platforms for visual reasoning (Krizhevsky et al., 2012). The ImageNet classification is a good example, a large dataset of images collected into 1000 classes. Respectable classification accuracy has been found through the use of deep networks with convolutional layers for visual feature extraction (Krizhevsky et al., 2012). We attempt to replicate these visual feature extraction layers through the construction of hidden layers tuned to find visual features. As this is just the initial investigation, convolutional layers have not yet been utilised and this does result in limitations to the models ability to detect variable light curves which have been translated (i.e. at a different phase) to the training data. To minimise this potential issue the epoch-folded light curves are phase-normalised to have maximum brightness occur at a phase of 0.25. Figure 6.3 demonstrates a set of 28×28 pixel images generated on the SkycamT light curves shown in figure 6.2.

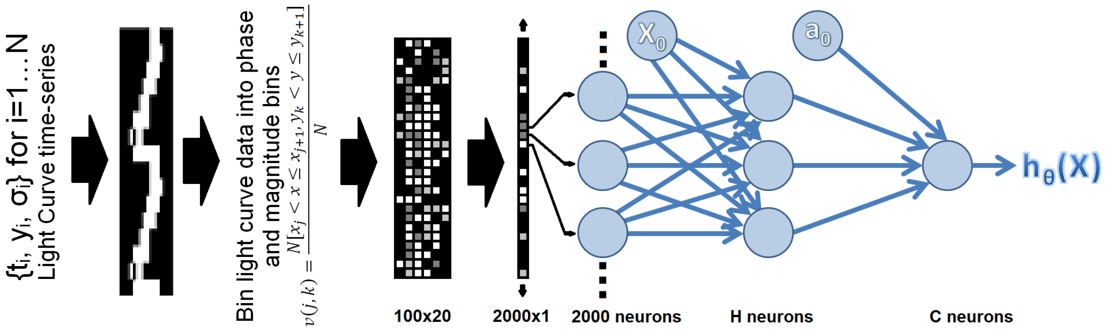

{ti, yi, σj} for i=1...N
Light Curve time-series
Bin light curve data into phase and magnitude bins
$v(j,k) = \frac{N[x_j < x \leq x_{j+1}, y_k < y \leq y_{k+1}]}{N}$
100x20
2000x1
2000 neurons
H neurons
C neurons
X0
a0
hθ(X)

FIGURE 6.4: Demonstration of the transformation of the 100x20 pixel phased light curve images into a flattened feature vector of 2000 real values of the pixel activations. These 2000 pixel activations are then used as input into a machine learning algorithm (in this example, a neural network) for training or prediction into one of C classes after non-linear modelling in the H neuron hidden layer.

We first phase-folded the light curves for each of the STILT dataset light sources. This task required a candidate period. As the Lomb-Scargle Periodogram is performing poorly on our data, we instead used the AAVSO period as we had when the 16 features used in Kim et al. were produced. This phase space exists from a phase of 0 to 1 with the brightest data point defined as 0.25. Outliers were also eliminated from the light curves by defining the brightness range of the folded light curve from the mean brightness as the amplitude of the light curve. This amplitude is defined as the difference between the median of the maximum 5% of data points and the median of the minimum 5% of data points divided by two. In order to emphasise shapes present at the edge of the folded light curve (as shapes can be split as phase values over 1 loop around to 0), the folded light curve was duplicated operating over a new phase space of -1 to 1. In order to reduce the noise we applied a bin-averaging process to the folded light curve data. The phase-space was binned into 100 phase-bins each bin having a phase range of 0.02. All observed data points in each phase bin are mean averaged and retained. Empty phase bins are removed. The bin-averaged phase-folded light curves were then used to generate pixelated images. This has a number of important uses. First we can guarantee an identical number of inputs into our neural network regardless of the sampling of the light curve. Second, it can be minimized to a level which optimizes for computational cost. For this task we decided to transform the light curves into 100x20 pixel images giving 2000 input 'feature' pixels shown in figure 6.4. Each light curve produced a magnitude-scaled image of the amplitude of the light curve centred on its weighted mean.

6.1.2 Results: Kim et al. Features

Previous studies found that the Kim et al. features performed best when trained using the random forest algorithm (Richards et al., 2011b; Kim and Bailer-Jones, 2016). Therefore, in this study we make use of Random Forest models with a number of different parameters, a Feedforward Neural Network with a single hidden layer of appropriate size and a Support Vector Machine with a linear kernel as the radial basis function kernel was unable to extract usable information from the features resulting in all predictions being assigned to the dominant class in the dataset, the eclipsing binaries. All the results were obtained through training on a 3.4 GHz Intel Core i7-3770 processor with 16 GB of memory. RStudio was used as the running environment. The STILT data was stored on a separate 1 TB hard drive within a MySQL database.

The 2519 light curves are evaluated through a process of stratified 10-fold cross-validation with 3 repeats. In this procedure the dataset is split into ten sections whilst maintaining the ratio of each class in the subsets relative to the whole dataset. Each subset is then used as a validation set for models trained using the other nine subsets. This validation involves the prediction of the classes of the validation set light curves followed by the computation of the Area under the Curve (AUC) statistic from the computed multi-class Receiver operating characteristic (ROC) curve. This validation is performed ten times for each data subset and mean averaged to produce the validation statistic for that repeat. This process is repeated three times with differing random seed values and again mean averaged to produce the final validation statistic. As the AUC is expected to vary around a mean value due to slight difference in the quality of the light curves being used for training and validation in each cycle, this procedure is hoped to be sufficient to eliminate any potential variation in the trained models performance.

The 10-fold cross-validation was first applied to the full 16 features dataset determined for each of the 2519 light curves. The random forest model was tuned with a hyper-parameter that defines the number of predictors sampled for splitting at each node. The best performance was obtained with this parameter set to 4 features. The Neural Network was trained using backpropagation on a single hidden layer feedforward neural network with 16 input neurons, 50 neurons in the hidden layer and 5 neurons in the output layer using a softmax classifier. The Hyperbolic-Tangent function was used for non-linearity and complexity control was introduced through a momentum term valued at 0.9. All neurons are initialised with a uniform random number between 0 and 0.07. The learning rate was set at 0.005. The network was trained using backpropagation for 600 iterations. The Support Vector Machine was tuned using a grid based search for the best performing cost value which was found to be 32 for this evaluation. The results of this evaluation are shown in table 6.3.

Table 6.3: AUC of the three algorithms on the 16 Kim et al. Features.

Model	Mean AUC
Random Forest, mtry = 4	0.84195528
Support Vector Machine, linear kernel, Cost = 32	0.80900461
Feedforward Neural Network, layers: 16-50-5	**0.84946005**

Table 6.4: AUC of the three algorithms on the scaled 16 Kim et al. Features.

Model	Mean AUC
Random Forest, mtry = 4	**0.84195528**
Support Vector Machine, linear kernel, Cost = 32	0.80897909
Feedforward Neural Network, layers: 16-50-5	0.76962864

Additionally, due to neural networks often performing better with scaled data, the 16 features were scaled to a mean of 0 and a standard deviation of 1 and all three machine learning methods validated on this scaled dataset with results shown in table 6.4. The hyper-parameters were re-tested but did not change from the previous validation.

Surprisingly, the scaled dataset resulted in a small drop in performance for the Support Vector Machine and Neural Network. The results show that our 50 hidden layer feedforward neural network achieved the best performance on the STILT light curves. The random forest model had a similar performance due to not assuming normally scaled data, a result expected by previous studies (Debosscher et al., 2007) with the no information rate being an AUC statistic value of 0.5.

By using the probabilities predicted by the random forest algorithm for each of the light curves to determine five binary one verses all ROC curves, a form of multi-class ROC curve can be plotted. These curves are a measure of a class' true positive rate against the false positive rate with the ideal classifier maximising the true positive rate whilst minimising the false positive rate. Therefore, the better performing a class, the closer it will deviate towards the top left corner from the random-state as a straight line with gradient unity shown by the dotted black line in the figures. Figure 6.5 shows the ROC curve generated by one of the validation 16 feature random forest models with a multiclass AUC of 0.8102. Each line is related to a one-vs-many prediction on a specific class given by the line colour in the legend. This is performed by assigning a class label of 1 to the appropriate class and a label of 0 to all other classes.

These results are reasonable considering the cadence limitations of this survey with all the classes detectable with greater than 80% retrieval rate at a cost of at worst a 10% false retrieval rate. The Delta Scuti and Long Period Variable classes achieve the best AUC with values of 0.9959 and 0.9877 respectively. This is likely a result of the two classes exhibiting clear periodic and amplitude features. Long Period Variables tend to have highly sinusoidal variations with periods of the order of years and amplitudes of

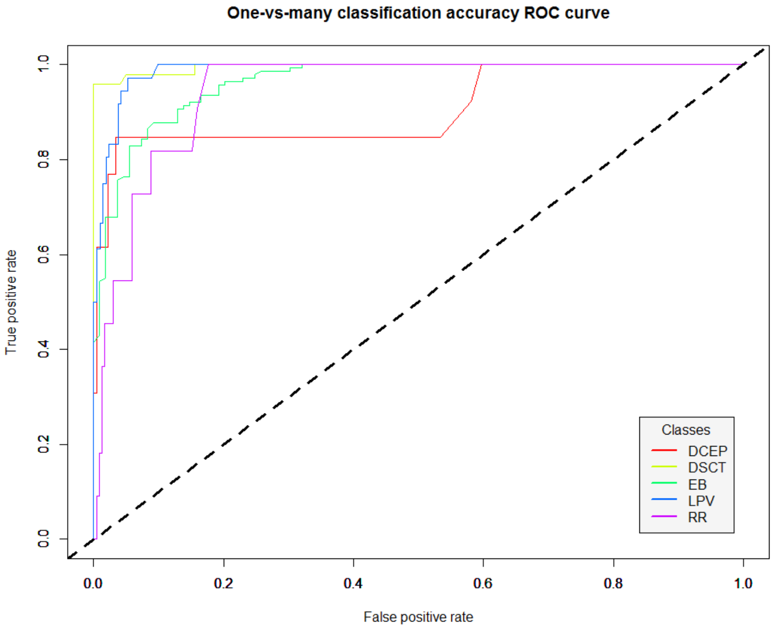


FIGURE 6.5: ROC curve for the 16 feature trained random forest model. The Long Period Variable (LPV) and Delta Scuti (DSCT) classes exhibit the best performance.

multiple magnitudes whereas Delta Scuti variables have periods of only a few hours and variations on the order of a tenth of a magnitude. This conclusion can be reinforced through obtaining the feature importance of the random forest model defined by the Mean Decrease Gini statistic. Table 6.5 demonstrates this importance statistic for the model used to generate the ROC curve in figure 6.5. This shows that the Period, variability in the folded light curve and the amplitude of the Fourier model are dominant.

Overall, the most important features are the period and the amplitude of the primary harmonic of the Fourier model. This amplitude is superior to the range of magnitude for a light curve as the range can be prone to noisy observations. The Fourier model does have the disadvantage of poor fits due to sampling as discussed previously however the models have still selected it as a strong feature in these 2519 light curves. Additional features of interest are the Kurtosis and slope gradient features m_{p10} and m_{p90}. These features are strong at identifying light curves with sharp peaks or dips in brightness which is a feature commonly associated with a large number of eclipsing binary light curves. Finally, the feature Ψ^n shows how strongly aligned the data points are for a given period. This can be useful for Long Period Variables due to secondary periods.

6.1.3 Results: Image Representation Features

Like the previous 16 feature models, the images produced from the bin-averaged phase folded light curves are used to produce a new validation dataset. The same machine

TABLE 6.5: Mean decrease Gini coefficients for the 16 Kim et al. Features.

Feature	Mean Decrease Gini
Period	2019.2553583
Ψ^n	549.73631507
Ψ^{cs}	192.76109940
R_{21}	209.11562924
R_{31}	136.55228001
Φ_{21}	69.433471117
Φ_{31}	71.555076485
γ_1	97.304429700
γ_2	107.25834895
K	142.53335421
Q_{3-1}	150.26993955
A	81.838996126
H_1	413.76605741
W	88.665784819
m_{p10}	425.10419581
m_{p90}	320.10894756

learning algorithms from the previous validation were applied to this dataset generated by the new method. The primary difference from the previous models was there were now 2000 input units where each one is the value of a specific pixel from a concatenated 100x20 image representation vector, -0.5 for an off (black) pixel and +0.5 for an on (white) pixel.

The random forest model was tuned with the number of predictors sampled for splitting at each node hyper-parameter valued at 4 like the previous validation. The Neural Network was trained using backpropagation on a single hidden layer feedforward neural network with 2000 input neurons, 200 neurons in the hidden layer and 5 neurons in the output layer using a softmax classifier. The number of neurons in the hidden layer was increased in order to model more complex patterns expected to be present in the input features. Potentially more might be required but this was limited by the available resources. The Hyperbolic-Tangent function was used for non-linearity and complexity control was introduced through a momentum term valued at 0.9. All neurons are initialised with a uniform random number between 0 and 0.07. The learning rate was set at 0.005. The network was trained using backpropagation for 600 iterations. The Support Vector Machine was tuned using a grid based search for the best performing cost value which was found to be $C = 1$ for this evaluation. The results of this evaluation are shown in table 6.6.

Whilst the AUC results are notably inferior to the previous results from Period and Amplitude features from the previous study, the result does show that features were automatically extracted by the machine learning algorithms and used to train to recognise

Table 6.6: AUC of the three algorithms on the Visual Features.

Model	Mean AUC
Random Forest, mtry = 4	**0.63483958**
Support Vector Machine, linear kernel, Cost = 1	0.58861276
Feedforward Neural Network, layers: 2000-200-5	0.61050239

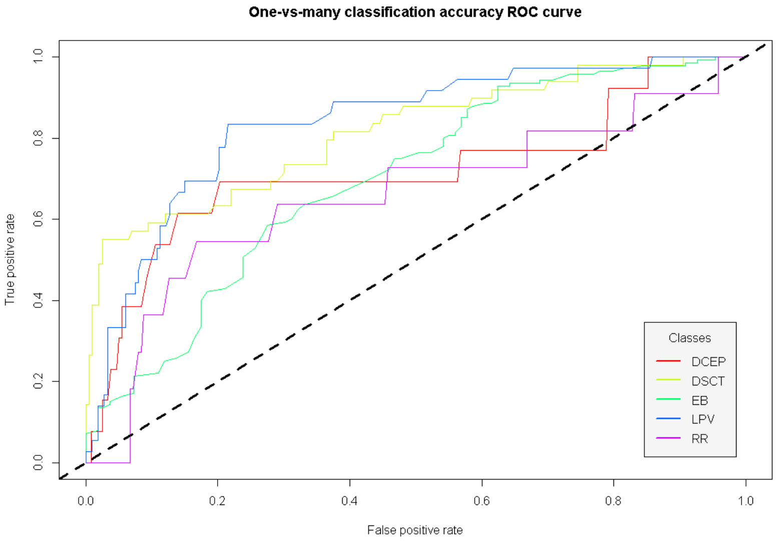


Figure 6.6: ROC curve for the 0.6386 AUC image representation model. Global performance is poorer than the 16 features models with the best resolved classes remaining the Long Period Variables and the Delta Scuti variables.

visual shapes for use in a classification task. It is also worth noting that this approach may prove better at discriminating between two similar subclasses than on an overall superclass problem. This network is also extremely limited in the visual features it can extract. For example, despite attempts to position certain magnitude features at specific phases, noise quite often causes these features to be placed at slightly different phases. This results in the requirement of any visual feature layer to implement translation invariance. This can be accomplished by neural networks using convolutional layers (Krizhevsky et al., 2012) but this has not been implemented in these models, which is a big limitation. Figure 6.6 shows the ROC curves from one of the image representation random forest models with a multi-class AUC of 0.6386.

The image representation performance can be augmented through the recognition that significant information is lost through the lack of scaling in the images. The Delta Cepheid and Long Period Variable folded waveforms can look very similar until the realisation is made that the amplitude of the Long Period Variables is significantly larger. Therefore we included two features that describe the two axes of the images. As the horizontal direction shows the phase of the folded light curve, the period describes the length of time this phase covers. As for the vertical direction, this is by definition the amplitude of the light curve as defined above. Including these two features along

TABLE 6.7: AUC of the three algorithms on the Visual Features with Period and Amplitude.

Model	Mean AUC
Random Forest, mtry = 4	0.66047666
Support Vector Machine, linear kernel, Cost = 1	0.76388919
Feedforward Neural Network, layers: 2000-200-5	**0.79524852**

with the 2000 input pixel values produced the AUC cross-validation results displayed in table 6.7. The neural network model can also be used to plot the weights between the 2000 input neurons and the H hidden layer neurons to produce a set of H 100×20 'weight images' which show the areas of the input image-representation light curves which produce strong responses in the hidden layer neurons.

Figure 6.7 demonstrates the importance of the individual image pixels for the classification task on three of the super-classes using their mean decrease Gini coefficients. Whilst each individual pixel has a minimal weighting, together they can identify interesting structures. All three plots show a clear structure across the phase space. This is the importance varying from low weight at the exterior of the image to higher weighting close to the centre which is where important signal structures are expected to be found. In the Delta Cepheid and RR Lyrae classes we can see a clear secondary structure where the importance rises to peaks at pixels near the phase of 0.25 and 0.75 where major peaks and dips are expected given we set the maximum brightness to occur near phase 0.25 when the light curves were epoch-folded. Finally, the Eclipsing Binary plot shows the weighting is a fraction of that from the other two classes. It appears clear that our new proposed method was struggling to resolve usable detail from the eclipsing binary light curves. Whilst the reason for this is unclear it may be a result of the observational cadence of these objects. If the characteristic dip in the light curve due to the transit event is not sufficiently observed the resulting folded light curve will exhibit reduced amplitude and without this characteristic dip feature except for possibly a gap near the expected light curve dip but it would be very unlikely for this gap to occur at the same phase location.

6.1.4 Conclusion

In these experiments we built a number of models based on features from Kim et al. and our own image-representation approach using stratified 10-fold cross-validation on 2519 STILT variable light curves from five object super-classes. We showed these features contained important information when the light curves were noisy and poorly sampled. Our method initially struggled to compete with features engineered from previous studies attaining a best AUC of 0.6348 compared to the 16 Kim et al. features with a best

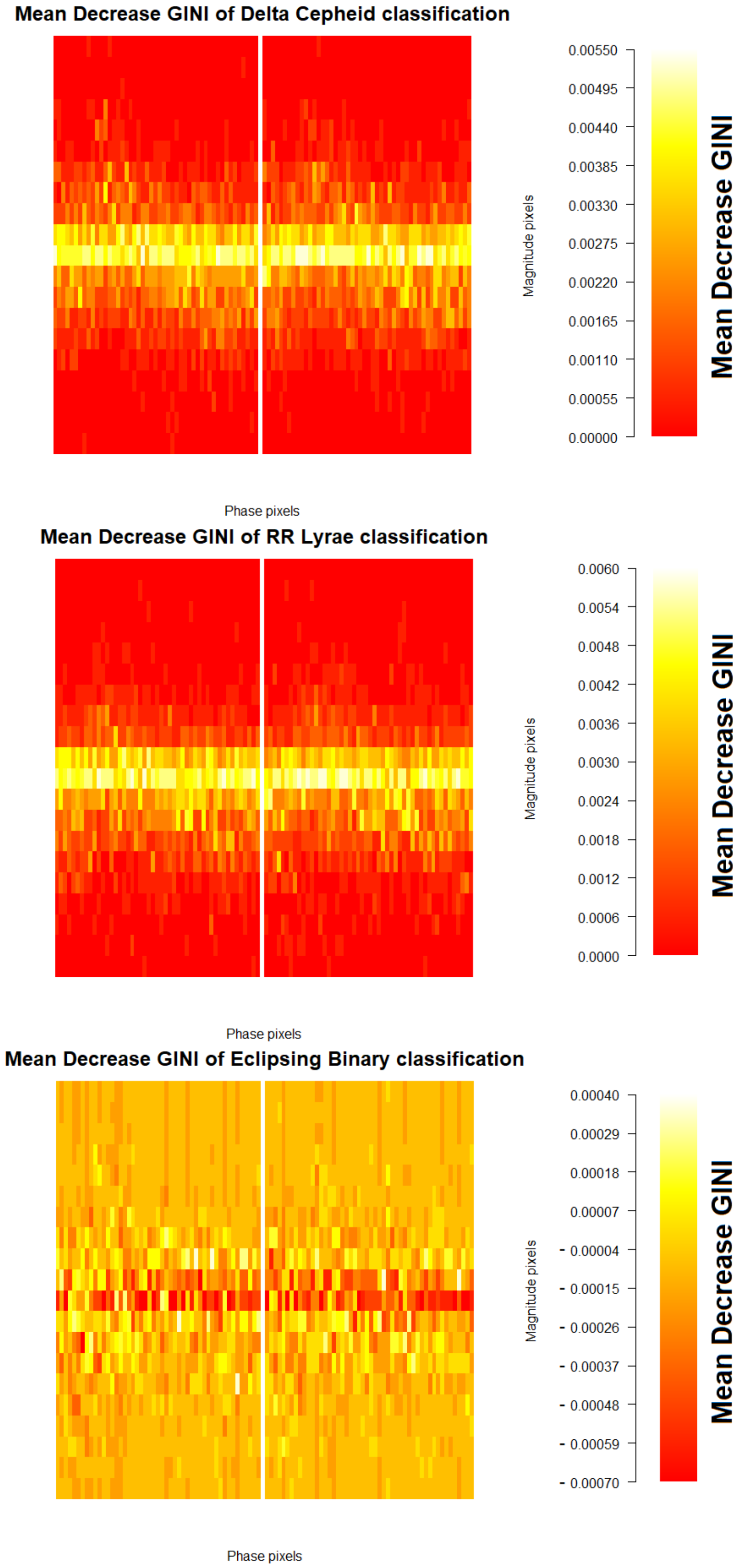


Figure 6.7: Heatmaps of the Mean Decrease Gini feature importance as a function of pixel position for three of the main super-classes. The Delta Cepheid variables at the top, RR Lyrae variables at the middle and Eclipsing Binaries in the bottom plot. Each pixel individually has a low weighting but together can communicate important class knowledge.

AUC of 0.8495 until we introduced the period and amplitude features into the training phase. These features give context to the two dimensions on the image representations allowing for an improvement in the best AUC to 0.7952. These strengths were offset by limitations in the machine learning algorithms we used when applied to image based representations especially in the lack of translational and scale invariance. This caused test light curves of a well-known class but with a different phase alignment to be misclassified. By implementing convolutional layers in our feedforward neural network topology, this should introduce these invariances (Krizhevsky et al., 2012). The neural networks used in this study are of the shallow variety having only one hidden layer. Extending this network using deep learning methods into a deep topology with multiple layers possibly implementing convolutional layers may offer an improvement to the performance and constitute an important future work for this development.

There are also a number of hyper parameters that havent been fully investigated such as the optimal pixel ‘resolution’ for the light curve images. The horizontal pixels contain the majority of the sampling noise and the vertical pixels carry a lot of the magnitude noise which is heavily instrumental and data-reduction limited. A superior method of determining the candidate period to phase-fold the light curve at must also be determined or the resulting image carries no useful information on the class of the light curve. These efforts will improve light curve classification and potentially redefine the limitations of survey cadence required for scientific analysis. Ultimately, until this method matches the performance of more traditional feature extraction methods, the automated classification pipeline development will continue with the Richards et al., Kim et al. and interpolated PolyFit features presented below.

6.2 PolyFit Feature Representation

The poor performance of the image representation learning method required another approach to perform a similar task for the Skycam classification pipeline. Deep Learning is an option based on the image representation shown previously. However, in this section, an alternative approach is suggested using unsupervised learning methods to construct a representation of the data. The goal of the representation learning is to produce features that model the shape of the folded light curve. This can be accomplished through the interpolation of a fitted model on the data points. This model would remove much of the light curve noise and produce a fit which has the flexibility to correctly fit any possible phase-folded light curve shape whilst not overfitting on the noise. The chosen model for this interpolation on the Skycam light curves is the PolyFit model. This model was developed for the fitting of eclipsing binary light curves and therefore is specifically designed to accurately reproduce the thin primary and secondary eclipses of detached binaries whilst still maintaining good performance on other light curve shapes such as pulsating variables (Prsa et al., 2008; Paegert et al., 2014; Parvizi et al., 2014).

6.2.1 PolyFit Algorithm

The PolyFit algorithm is designed to outperform Fourier and Spline models when applied to any eclipsing binary light curve. The PolyFit algorithm is a method of fitting a polynomial chain $P(x)$ of smooth, piecewise n^{th} order polynomials which connect at a set of knots (Prsa et al., 2008). The algorithm has two main additions compared to normal piecewise polynomial methods to achieve the desired performance. First, unlike spline models, the polynomials are not required to be differentiable (although they remain continuous) at the knots allowing the modelling of sharp, narrow features such as eclipses. The second requirement is that the model cycles across the phase boundary between 0.5 and -0.5 when centred on zero. Our implementation of PolyFit utilises 4 knots with 4 2^{nd} order piecewise polynomials fit using regularised polynomial regression through the implementation of the normal equation on the 4 subsets of data defined between each pair of consecutive knots. This produces a set of 16 parameters which fully describe the fitted PolyFit model, 4 knot locations with phases of $[-0.5, 0.5]$ and 12 polynomial parameters, a intercept, first order and second order for each of the four polynomials. This is substantially less free parameters than a Fourier model would require for similar eclipse fitting performance (Debosscher et al., 2007). In addition to these features, the PolyFit model is used to interpolate 99 magnitude values across the $[-0.5, 0.5]$ phase range for further analysis.

Figure 6.8 demonstrates the PolyFit algorithm applied to the Skycam light curve of the eclipsing binary RS Sagitarii. The black points indicate the light curve observations phased by a candidate period and phase binned into 100 bins, the red line indicates the fitted PolyFit model and the green crosses indicate the phase locations of the four knots. Figure 6.9 shows the capability of this method to accurately fit narrow eclipse features compared with spline and Fourier models. The top plot is the PolyFit model which fits the primary and secondary eclipses without substantially overfitting on the out-of-eclipse noise. The middle plot demonstrates a spline model with a *span* of 0.2 where the span defines the smoothness of the fitted spline polynomials. This shorter span results in a spline model which performs well on the deep eclipse but overfits the noise in the out-of-eclipse light curve. Increasing the span improves the out-of-eclipse performance at a cost of poorer eclipse modelling. Each eclipsing binary light curve will have an optimal value of span which compromises between eclipse and noise fitting yet it is unlikely that this optimal span value will perform as well as the PolyFit model. The

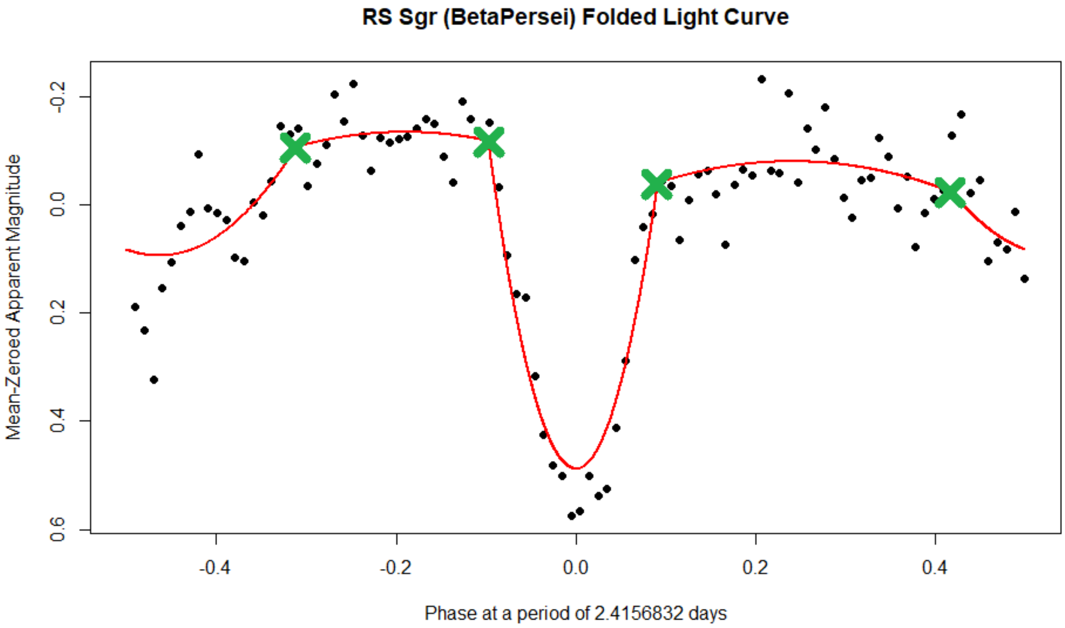


Figure 6.8: The PolyFit algorithm applied to the Skycam light curve of the Eclipsing Binary RS Sagitarii. The light curve has been phase-folded and phase binned into 100 bins. The red line indicates the fitted PolyFit model and the green crosses indicate the optimal knot points found by the optimisation algorithm. This method produces a superior fit to the narrow eclipse feature than the Fourier or spline models in figure 6.9.

bottom plot demonstrates a Fourier fit with eight harmonics and an intercept with 17 parameters to the 16 PolyFit model parameters. As with the spline model with a low span argument, the eight harmonic Fourier model correctly models the deep primary eclipse but also overfits the out-of-eclipse noise. Reducing the number of harmonics in the Fourier model will result in a similar effect to increasing the span argument value. The PolyFit model is the only method of the three to correctly fit the eclipses without overfitting on the noise.

Another possible method for reconstructing the astrophysical signal from the observed data point is using a convex optimisation procedure using the robust L1-norm (Candès and Romberg, 2005). Most astrophysical signals produce a sparse frequency spectrum when Fourier transformed as they are a sum of a small number of frequency components. Therefore, this signal can be rebuilt from a small number of measurements as long as there are more observations than frequency components. Using a non-linear L1-norm regularisation parameter, the degeneracy of the parameter selection can be probed for a model which closely reconstructs the original sparse signal. This method has been used to create 'superresolution' sequences from poorly sampled periodic data (Chan et al., 2016). Such a technique has not been applied to unevenly sampled data in astronomy and may struggle with a limited set of measurements due to the observations being substantially fewer than the sampling rate for the SkycamT instrument. The other concern relative to models such as PolyFit is the presence of correlated noise in the data. These noise sources will introduce additional components in the Fourier transform of the observations which are unwanted for modelling the underlying astrophysical signal. In this case, it may be better to apply models which are not designed to reconstruct signals in the Fourier space.

The PolyFit algorithm is implemented by first selecting an initial state for the knots either by random or by a controlled method such as where the difference between the magnitudes of two data points crosses the mean magnitude. Using this initial set of knots x_k, $k = 1, \dots, 4$, the phase range of $[-0.5, 0.5]$ is partitioned into 4 intervals as shown in equation 6.2 (Prsa et al., 2008).

$$I_1 = [x_1, x_2), \quad I_2 = [x_2, x_3), \quad I_3 = [x_3, x_4), \quad I_4 = [x_4, x_1) \tag{6.2}$$

For the first phase interval I_1, use a regularised least-squares regression fit using the data points in this phase interval with 3 free parameters as shown in equation 6.3.

$$P_1(x) = a_0^{(1)} + a_1^{(1)}(x - x_1) + a_2^{(1)}(x - x_1)^2 \tag{6.3}$$

where $P_1(x)$ is the first polynomial as a function of phase x, x_1 is the phase of the first knot and $a_j^{(1)}$ are the fitted polynomial parameters where $j = 1, \dots, 3$. With the first three parameters computed, the next requirement is to compute $p_2(x)$ with respect to $p_1(x)$ and $p_3(x)$ with respect to $p_2(x)$ as shown in equation 6.4.

$$P_k(x) = a_0^{(k)} + a_1^{(k)}(x - x_k) + a_2^{(k)}(x - x_k)^2 \tag{6.4}$$

where $P_k(x)$ is the k^{th} polynomial of interval I_k. This must be computed whilst satisfying the constraint that the polynomial must connect with the previous polynomial at the knot x_k. This is shown in equation 6.5 and results in the computation of $p_2(x)$ and $p_3(x)$ being for 2 free parameters as the intercept $a_0^{(k)}$ where $k = [2, 3]$ has already been computed.

$$P_k(x_k) = p_{k-1}(x_k): \quad a_0^{(k)} = a_0^{(k-1)} + a_1^{(k-1)}(x_k - x_{k-1}) + a_2^{(k-1)}(x_k - x_{k-1})^2 \tag{6.5}$$

For the final phase interval I_4 there are two constraints to be satisfied. The polynomial must connect with the third interval I_3 at the knot location x_4 (connectivity) and the phase space wrapping from 0.5 to -0.5. As the connectivity has constrained the intercept of the 4th polynomial, $a_0^{(4)}$, calculated as before from equation 6.5 the phase wrapping

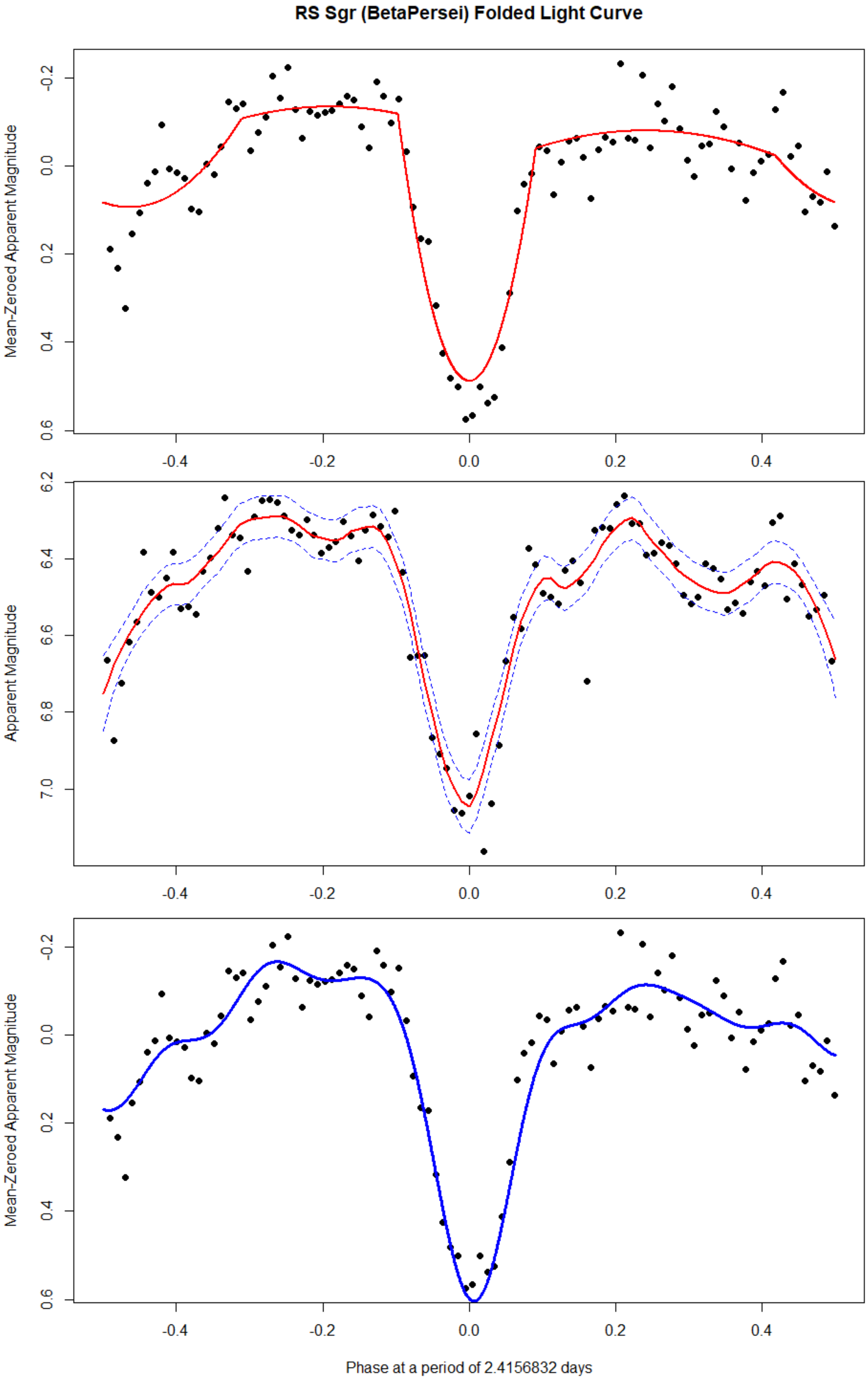


Figure 6.9: The PolyFit algorithm (top) fitted to a Skycam eclipsing binary light curve. The model provides a much more satisfactory fit than the spline model (middle) or the Fourier model (bottom) despite the Fourier model utilising more fitted parameters. The PolyFit model can accurately reproduce narrow eclipsing binary features whilst still providing good performance on pulsating and rotational light curves.

constraint constrains the first order parameter of the polynomial fit $a_1^{(4)}$ through constraint equation 6.6 revealing the remaining free parameter $a_2^{(4)}$.

$$P_4(x) = a_0^{(4)} + a_2^{(4)}(x - x_4)(x - x_1) \tag{6.6}$$

The original PolyFit implementation placed the four knots where the light curve data points crossed the mean magnitude of the light curve, randomly perturbed the knots using a random Gaussian 'kick' and then allowed them to relax into a minimum χ^2 state over a small number of iterations (Paegert et al., 2014). Each iteration must be carefully checked as the phase intervals must have an appropriate number of data points to prevent degeneracy in the polynomial fits. This means that the set of knots x_k must be rejected if the interval I_1 lacks 5 data points, I_2 and I_3 lack 4 data points each and I_4 lacks 3 data points. Finally, to prevent knots from adopting values which place two or more knots too close to each other, the fitting function χ^2 must have an additional penalty term which disincentives this undesirable outcome. This is accomplished by using a quadratic repulsion term as shown in equation 6.7 which decreases the performance of a given fit by the square of the distance between each pair of knots with the size of this repulsion defined by an argument ϵ (Prsa et al., 2008).

$$r_{\text{cost}}(x_k;\epsilon) = \epsilon\left[(x_2 - x_1)^{-2} + (x_3 - x_2)^{-2} + (x_4 - x_3)^{-2} + (x_1 + 1 - x_4)^{-2}\right] \tag{6.7}$$

Due to the noise in the Skycam light curves this approach was not sufficient to produce good models as the nearest χ^2 minimum was highly dependent on where the noise settled the initial state of the knots. To initially limit the noise from the Skycam light curves, the data points are phase binned into 100 mean averaged bins. This reduces the high frequency noise from affecting the PolyFit model and is particularly effective on Skycam due to the large number of observations in many light curves as well as a substantial reduction on computation time as the regression has less data points to compute. Despite this binning operation, the white noise in the light curves was still of sufficient amplitude to produce many local minima in the χ^2 minimisation procedure. As a result, the fitting procedure was insufficient for reliably determining the optimal PolyFit model for a given light curve.

6.2.2 Genetic Optimisation for PolyFit

The poor performance of the PolyFit algorithm's original fitting routine was a substantial problem in the use of this method on the Skycam light curves. Fortunately, the genetic algorithm optimisation developed for use in the GRAPE method provide a novel solution to the issues with PolyFit on noisy light curves. As discussed in chapter 3, Genetic

Algorithms are highly capable at the identification of the global optimum of a highly non-linear fitness function with many local optima (Charbonneau, 1995). As stated above, the fitness function of the PolyFit algorithm exhibits these properties.

The Genetic Algorithm method from GRAPE was modified to identify the optimal knot locations for the set of 4 knots x_k through the computation of the χ^2 fitness function augmented by the repulsion term in equation 6.7. This fitness function is showed in equation 6.8.

$$\chi^2(x_k;\epsilon) = \sum_{j=1}^{N} w_j \left(p(x_j) - y_j\right)^2 + r_{\text{cost}}(x_k;\epsilon) \tag{6.8}$$

where $p(x_j)$ is the PolyFit interpolated magnitude of phase point x_j, y_j is the magnitude of the phase binned data point j at a phase of x_j, w_j are the weights of each phase bin which are kept at $w_j = 1$ and $r_{\text{cost}}(x_k;\epsilon)$ is the knot repulsion from equation 6.7 which is a function of a repulsion strength ϵ and the knot positions x_k and N is the number of binned data points in the light curve.

Where GRAPE utilises a one-dimensional feature space, the genetic PolyFit method requires the use of a four-dimensional feature space, the phase locations of the 4 knots. As the 12 polynomial parameters are generated through regularised regression as a function of the 4 knot positions x_k, they do not need to be determined by the genetic algorithm leaving just the 4 knots. The initial population of size N_{pop} is established by a uniform random number generator which creates N_{pop} sets of $x_k = [-0.5, 0.5]$ sorted from -0.5 to 0.5. This population is then encoded into chromosome strings by rescaling the phases to between 0 and 1 (which is simply performed by computing $\hat{x}_k = x_k + 0.5$) followed by the recording of the top 5 decimal places for each of the 4 knots into a concatenated string of 20 base-10 numerals. Similar to the GRAPE method, these chromosomes undergo a genetic update process where knots which minimise the χ^2 fitness function are bred into children and have crossover, mutation and fitness selection operations applied for a set number of generations N_{gen} where the knots have converged to the global optimum.

The arguments of the genetic algorithm are selected through a grid cross-validation procedure on a set of 859 light curves with the limitation that the PolyFit routine must complete in under two seconds. The input arguments were as follows: $N_{\text{pop}} = 100$, $N_{\text{pairups}} = 20$, $N_{\text{gen}} = 100$, $P_{\text{crossover}} = 0.65$, $P_{\text{mutation}} = 0.03$, $P_{\text{fdif}} = 0.6$ and $P_{\text{dfrac}} = 0.7$. For further information on these genetic algorithm arguments we recommend reading the chapter 3 Evolutionary Implementation section. We found that this genetic optimisation routine produced more reliable optimal knot locations x_k regardless of the distribution of the initial knot candidates.

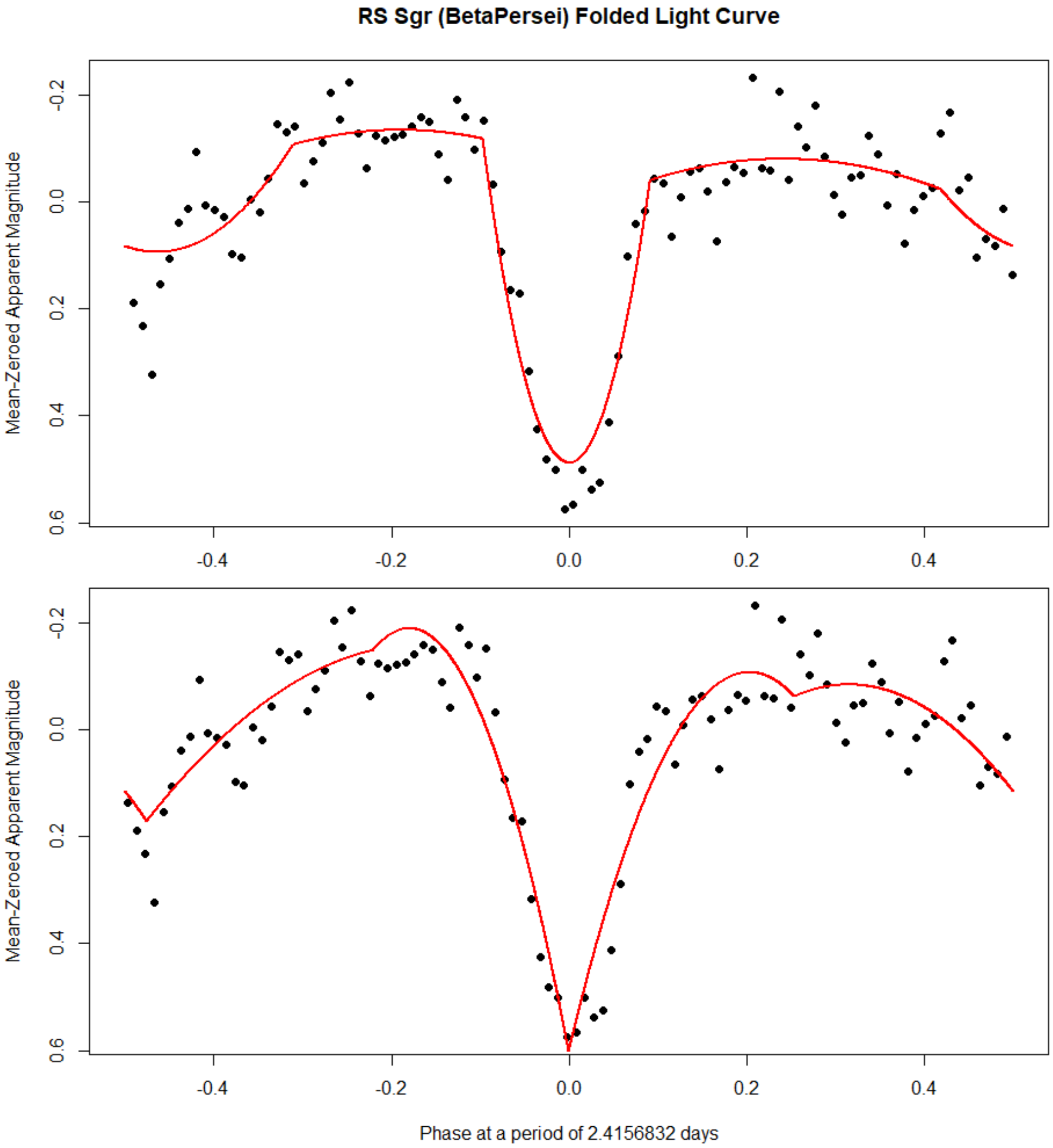


Figure 6.10: The same Skycam light curve with two different PolyFit optimisations. The top plot demonstrates the desired PolyFit model where the knots are located either side of the eclipse at the beginning and end of the dimming event. The bottom plot demonstrates a PolyFit model with a superior fit according to the fitness function in equation 6.8. This is not an ideal model as the knot has been located at the base of the eclipse, not the intended location and the second knot has been used to overfit on noise. Whilst the bottom fit minimises the fitness function, it is not the desired model and therefore the fitness function requires an additional corrective penalty term.

There remained one limitation relative to the original expected PolyFit performance on Eclipsing Binaries. The eclipse features are intended to be modelled by two knots at the beginning and end of the eclipse with a highly quadratic polynomial modelling the narrow eclipse dip. Unfortunately, the increased noise in the Skycam light curves resulted in the genetic PolyFit algorithm determining an optimal knot as the base of the eclipse and modelling the two sides of the dip in two separate phase intevals. This allows the PolyFit to use the second knot elsewhere in the phase space usually overfitting on the noise. This defect is shown in figure 6.10 where the top plot demonstrates the desired PolyFit model and the bottom plot demonstrates a model with a superior χ^2 performance but overfit on noise by placing the knot at the bottom of the peak. Our solution is to employ a second additional term to the χ^2 fitness function shown in equation 6.8 which introduces a penalty to selecting knot phase locations x_k where the resulting polynomial fit interpolated magnitude at that phase is far from the median magnitude of the light curve. This penalty incentivises the genetic PolyFit algorithm to place the knots at locations with interpolated magnitudes near the out-of-eclipse magnitude. This penalty is shown in equation 6.9 and is weighted by an argument δ which defines the relative cost of placing knots far from the median.

$$m_{\text{cost}}(x_k;\delta) = \delta \sum_{j=1}^{4} \left(a_0^{(j)} - \text{median}(y)\right)^2 \tag{6.9}$$

where $m_{\text{cost}}(x_k;\delta)$ is the new cost term, δ is an argument that defines the strength of the penalty, $a_0^{(j)}$ is the polynomial intercept term for phase interval I_j and median(y) is the median magnitude of the phase binned light curve. Equation 6.10 defines the final fitness function of the genetic PolyFit algorithm with the two penalty terms.

$$\chi^2(x_k;\epsilon;\delta) = \sum_{j=1}^{N} w_j \left(p(x_j) - y_j\right)^2 + r_{\text{cost}}(\epsilon;x_k) + m_{\text{cost}}(\delta;x_k) \tag{6.10}$$

This method, whilst designed to correctly fit narrow feature such as eclipses, does not adversely affect smooth continuous light curves such as eclipsing binaries. The $m_{\text{cost}}(x_k;\delta)$ penalty term does apply an increased cost to the knots which may be correctly located far from the median magnitude for these light curve shapes. In this case, the substantial number of data points far from the median magnitude in these phase locations allow the initial χ^2 component to 'outweigh' this penalty term allowing for the correct PolyFit model to be applied.

To verify the performance of our genetic algorithm approach to optimising the PolyFit algorithm, we implement an experiment to compare the knots fit by the genetic algorithm against those fit using a random perturbation approach across a set of 50 initial knot positions. This is applied to the 859 SkycamT light curves selected for use by a correct period match or submultiple match with the AAVSO catalogue period using the GRAPE period estimation method. These light curves are phased around the AAVSO catalogue period in order to generate a set of pulsating, rotational and eclipsing light curve shapes for the PolyFit algorithm to model.

Figure 6.11 shows the results of this experiment with the two distinct distributions of standard deviation performance for the two optimisation method. For every knot the genetic algorithm produced more consistent knot locations regardless of the initial knot positions with phase standard deviations of 0.01 to 0.1 for most of the 859 light curves. The random perturbation method was substantially less stable with the standard deviation of the final knots of the 50 initial states varying by a standard deviation of 0.1 for most of the light curves with many performing more poorly than this out to 0.2 phase standard deviation. Both these methods require approximately 1.5 seconds of runtime to determine the optimal knots and produce a final PolyFit model. The genetic algorithm based optimisation has a larger range of standard deviations to the random perturbation optimisation. This is likely a result of the quality of the light curve having more effect on the resulting fit as the very noisy light curves can still vary substantially with the genetic generational updates due to the minimal difference between the fitness function values for many potential models and the optimal model. The variability in the genetically optimised PolyFit model is still not ideal; however, it is satisfactory as the possible optimised models are all acceptable for the extraction of classification features for a given light curve.

6.2.3 PolyFit Principal Component Analysis

Despite the efforts of implementing the δ cost and using the genetic algorithm to optimise the knot locations, there is still substantial variation on the 16 parameters of the PolyFit model. Therefore they cannot reliably be used as features as, at best, the relationship between these parameters for multiple classes is highly non-linear and difficult to learn. Therefore, a different method of representing the important features of the interpolated PolyFit model shape is required which does not rely on the parameter values but simply the magnitudes of the fitted model. Principal Component Analysis has been previously used to learn a set of features from interpolated Fourier models applied to light curves (Deb and Singh, 2009; Tanvir et al., 2005; Yoachim et al., 2009). The extraction of features from fitted models has also been accomplished using other methods such as echo-state-networks, a form of recurrent neural network (Kugler et al., 2016) and local linear embedding (Matijevic, 2012) with positive results. The method used is not dependent on the model used to interpolate the light curves and simply on the distribution of

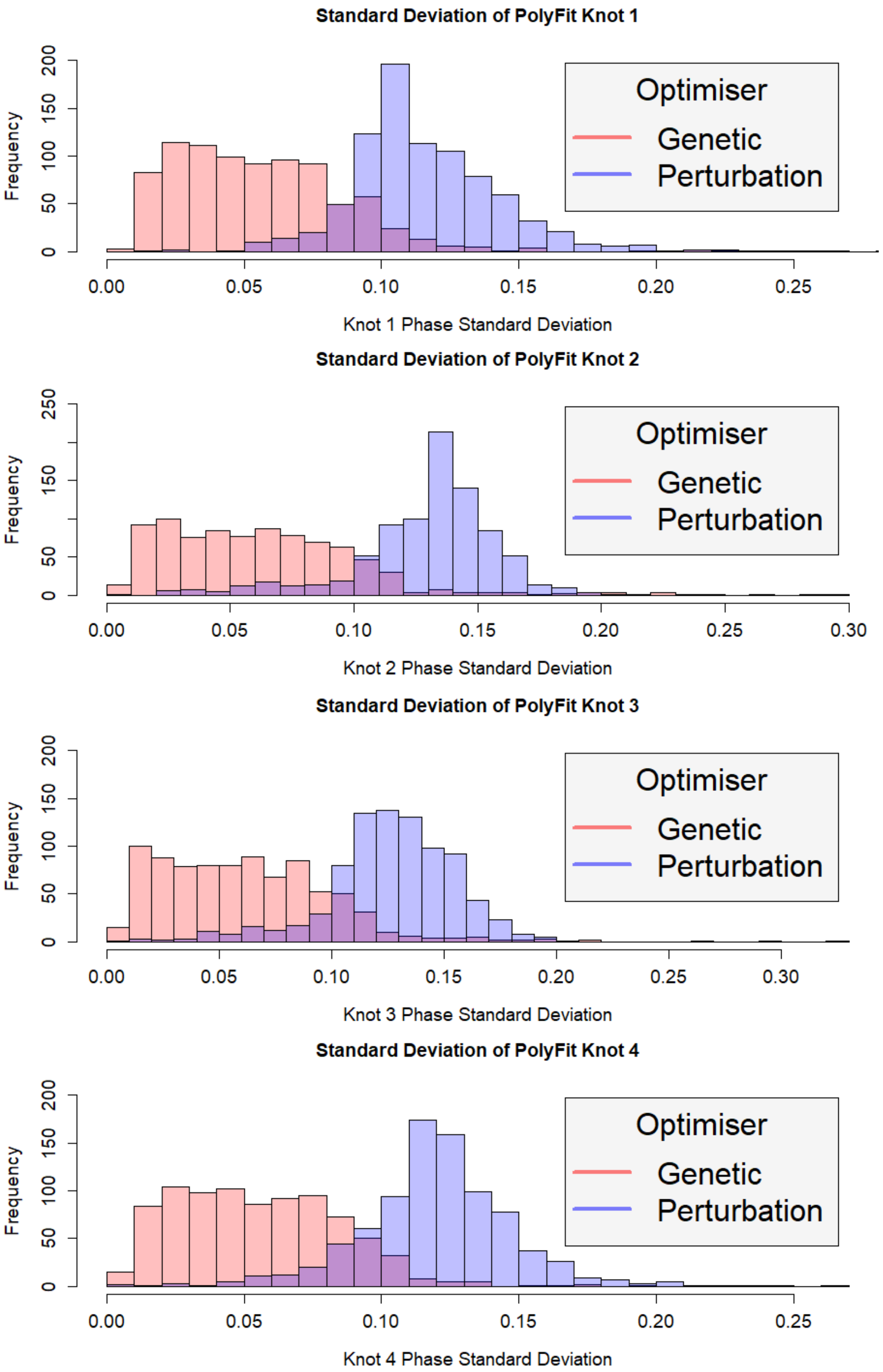


FIGURE 6.11: Histograms of the measured standard deviation of the 4 PolyFit knots on the set of 859 SkycamT light curves from table 2.1. The knots produced by each light curve over 50 initial states are recorded and the standard deviation measured from this set. Each histogram contains two distributions. The red distribution was produced by the knots optimised by the genetic algorithm method and the blue distribution was produced by the knots optimised by the random perturbation method from the original PolyFit algorithm (Prsa et al., 2008).

TABLE 6.8: The class distribution of the STILT 6897 variable light curves used for the PCA training.

Number	Class	Type	Count
1	β Lyrae	Eclipsing Binary	412
2	β Persei	Eclipsing Binary	1518
3	Chemically Peculiar	Rotational Variable	477
4	Classical Cepheid	Pulsating Variable	195
5	δ Scuti	Pulsating Variable	453
6	Ellipsoidal	Non-eclipsing Binary	131
7	Mira	Pulsating Variable	1256
8	Pop II Cepheid	Pulsating Variable	29
9	R Coronae Borealis	Eruptive Variable	3
10	RR Lyrae Dual Mode	Pulsating Variable	3
11	RR Lyrae Fundamental Mode	Pulsating Variable	99
12	RR Lyrae Overtone Mode	Pulsating Variable	60
13	RS Canum Venaticorum	Non-eclipsing Binary	528
14	RV Tauri	Pulsating Variable	36
15	Small Amplitude Red Giant	Pulsating Variable	4
16	S Doradus (Luminous Blue Variable)	Eruptive Variable	1
17	Semiregular Variable	Pulsating Variable	989
18	W Ursae Majoris	Eclipsing Binary	703

interpolated magnitudes. It is therefore suitable to investigate the performance of such an approach on sets of interpolated magnitudes extracted from SkycamT light curves by the PolyFit algorithm. This method can potentially generate learned features useful for Machine Learning classification in the automated pipeline.

Principal Component Analysis (PCA) defined in chapter 4 can be used as a dimensionality reduction technique to define the interpolated magnitudes of the PolyFit model as a set of features determined by a training set of light curves determined by cross-matching the SkycamT database on the set of AAVSO catalogue objects with period information and with types present in the BigMacc All-Sky Automated Survey (ASAS) light curve classification pipeline (Richards et al., 2012). This produces 6897 light curves across 18 variable star classes shown in table 6.8. The set of light curves is much bigger in this set as the trained PCA must be capable of modelling any possible PolyFit interpolation on the Skycam database. Many of these light curves are of dubious quality and GRAPE does not agree with the catalogue period for many of the objects as they are not in the 859 light curve dataset. Despite this, the learned PCA components using the larger light curve dataset generalises better than the smaller dataset.

Upon the computation of the PolyFit model for any given light curve, the model is used to interpolate 99 magnitude data points on an evenly distributed grid of phase values from -0.49 to 0.49 with intervals of 0.01. The interpolation does not go to -0.5 or 0.5 due to limitations in the spline fitting code as the edge values could be out of bounds. Of course this is not a limitation of the PolyFit algorithm as the entire phase space is mapped by the polynomial chain $P(x)$ yet in the event of the PolyFit algorithm failing to converge on a light curve the spline algorithm may still produce an acceptable (but likely inferior) fit to collect some useful information. The interpolated magnitudes are zeroed by having the mean magnitude of the phase-binned data points subtracted from their values. Using the phase-binned mean magnitude instead of the interpolated mean magnitude preserves the interpolated skewness of the light curve. The PolyFit interpolated magnitudes are then used to recalculate the phase zero-point through the identification of the maximum magnitude y_i data point (which corresponds to the minimum brightness). The phase space is then adjusted so that the minimum brightness of the light curve occurs at phase 0.0. For eclipsing binaries this is the primary eclipse and for other variables it simply corresponds with the minimum brightness of the variability. This is performed so the Richards et al. features discussed in chapter 5 for determining eclipsing binary light curve eccentricity and shape are directly computable from the interpolated magnitudes (Richards et al., 2012). The interpolated magnitudes are then normalised by equation 6.11 which means that the amplitude of the light curve is not a component of the learned PCA components. The mean zero and scaling operation corresponds to the operation in the PCA method in chapter 4 shown by equation 4.1.

$$\hat{y}_i = \frac{y}{|\max(y) - \min(y)|} \tag{6.11}$$

The genetic PolyFit algorithm was then applied to the 6897 light curves in table 6.8 producing a training matrix of 99 columns (the 99 interpolated magnitudes across the phase space) and 6889 rows (6889 light curves generated a PolyFit model, for 8 light curves the PolyFit did not converge and were discarded). Using equations 4.2 and 4.3 from chapter 4, the PCA operation is computed producing 99 Principal Components sorted by their variance where the first Principal Component contains most of the information of the system. Figure 6.12 shows the top 10 principal components of the resulting PCA model. The top ten principal components describe 96% of the variance in the PolyFit modelled light curves whereas the remaining 89 principal components only add an additional 3% variance which is likely noise dominated.

The top ten Principal Components (PCs) can be used to closely reconstruct the original PolyFit model as each learned component contains information about a specific transformation of the light curve weighted by the specific value for a given light curve. Figure 6.13 demonstrates this reconstruction on the eclipsing binary RS Sagitarii. The black line shows the original PolyFit model as seen previously in figure 6.8. Each coloured line represents the reconstruction of the original model by adding on an additional principal component. From this reconstruction it is clear that the first principal component

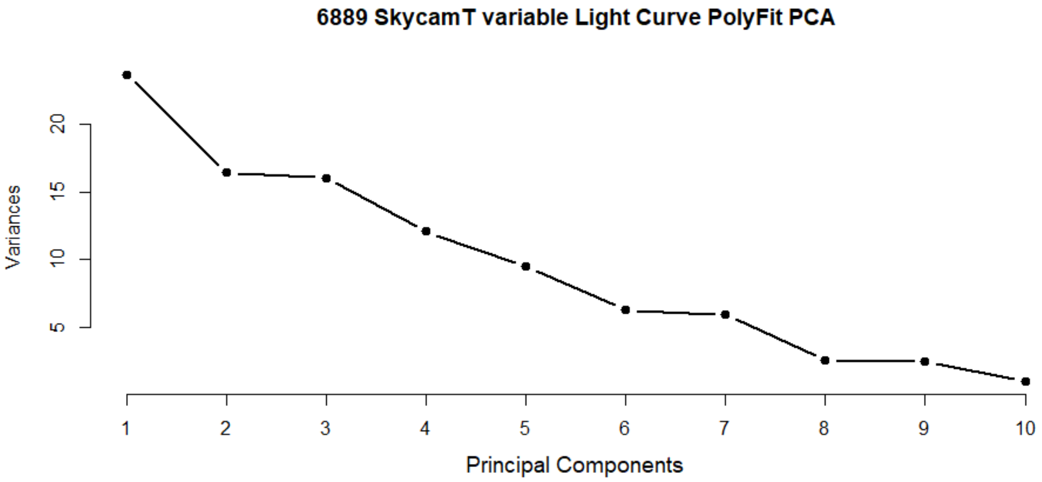


FIGURE 6.12: Plot showing the variance described by each principal component determined from the 6889 SkycamT light curves. The first principal component describes 23.7% of the variance of the light curves and the second principal component an additional 16.4%. The first ten principal components describe 96% of the total variance of the light curves. This allows these ten values to describe the shape of the light curves as the remaining 4% is likely noise related.

PC1 models the general shape of the light curve determining the range between the minimum and the maximum magnitudes of the light curve but does not model the secondary eclipse. PC2 and PC3 are responsible for modelling the asymmetries between the $[-0.5, 0]$ and $[0, 0.5]$ phase intervals as well as the presence of a secondary eclipse. This is important in the modelling of eccentric eclipsing binaries as well as containing the weighting that distinguishes eclipsing binary light curves from pulsating light curves. PC4 to PC9 are used to reconstruct variations in the smooth continuum of the light curve such as the ‘bumps’ common to some Cepheid and RR Lyrae type pulsating variables although there is likely a large noise component to these principal components. The final principal component, PC10 appears specifically for modelling the very narrow primary eclipses seen in the longer period β Persei variables. As there are few examples of this type of variable in the set of Skycam light curves, this explains why this important principal component has a lower variance as the variance describes how informative it is to the given training dataset. The most distinguishing feature of the principal components is the narrow spike feature located near phase zero. This is the component of the learned representation that models the width of the eclipses and is weaker for objects lacking a narrow eclipse. This is why PC1 to PC9 contain this spike; it is there to cancel out the narrow eclipse in the majority of light curves which lack this feature. PC10 is used to apply this feature when it is required for a given light curve.

The features utilised from this PCA are the ten weights which, in conjunction with the PCA model, allow the approximate reconstruction of the PolyFit model. The different

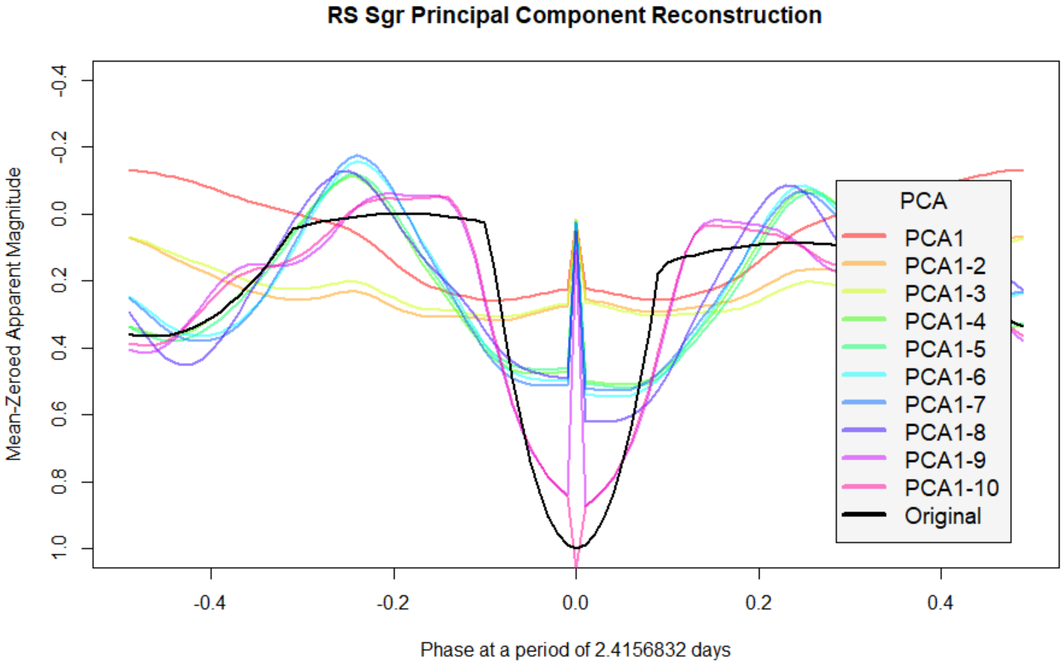


FIGURE 6.13: Plot showing the reconstruction of the SkycamT light curve of the eclipsing binary RS Sagitarii from the top ten principal components of the learned PCA model.

shapes of light curve are expected to produce different sets of the ten weights and therefore can be used in the classification task. Figure 6.14 shows the principal component reconstruction for U Aquilae a Classical Cepheid and CN Andromedae, a β Lyrae eclipsing binary. Figure 6.15 shows this reconstruction for S Pegasi, a Mira-type Long Period Variable and RS Bootis, a fundamental mode RR Lyrae variable. Each variable light curve contains shapes distinctive to each of their classes and therefore they have unique PCA weights which can be used to determine the class of an unknown light curve.

Comparing the PCA features for the set of 859 light curves across 12 classes allow for the identification of features which may be of potential interest to a classification method. The best feature for discriminating the classes is the PC2 feature which is not surprising as PC2 and PC3 are the strongest indicators of the differences between the eclipsing binary ‘double dip’ light curve to the pulsating variable ‘single dip, single peak’ light curves. Figure 6.16 demonstrates the plot of the base-10 logarithm of the Period (the best feature for distinguishing the classes) against the second principal component feature. Whilst there is substantial overlap between the classes, the eclipsing binary classes tend to adopt lower values of P.PCA2 around 1 to 2 whereas the pulsating variables have larger values of P.PCA2 closer to 3 and 4. The feature appears very useful for separating the short-period eclipsing binaries from the RR Lyrae pulsating variables which means it could be of important use in the machine learning classifiers.

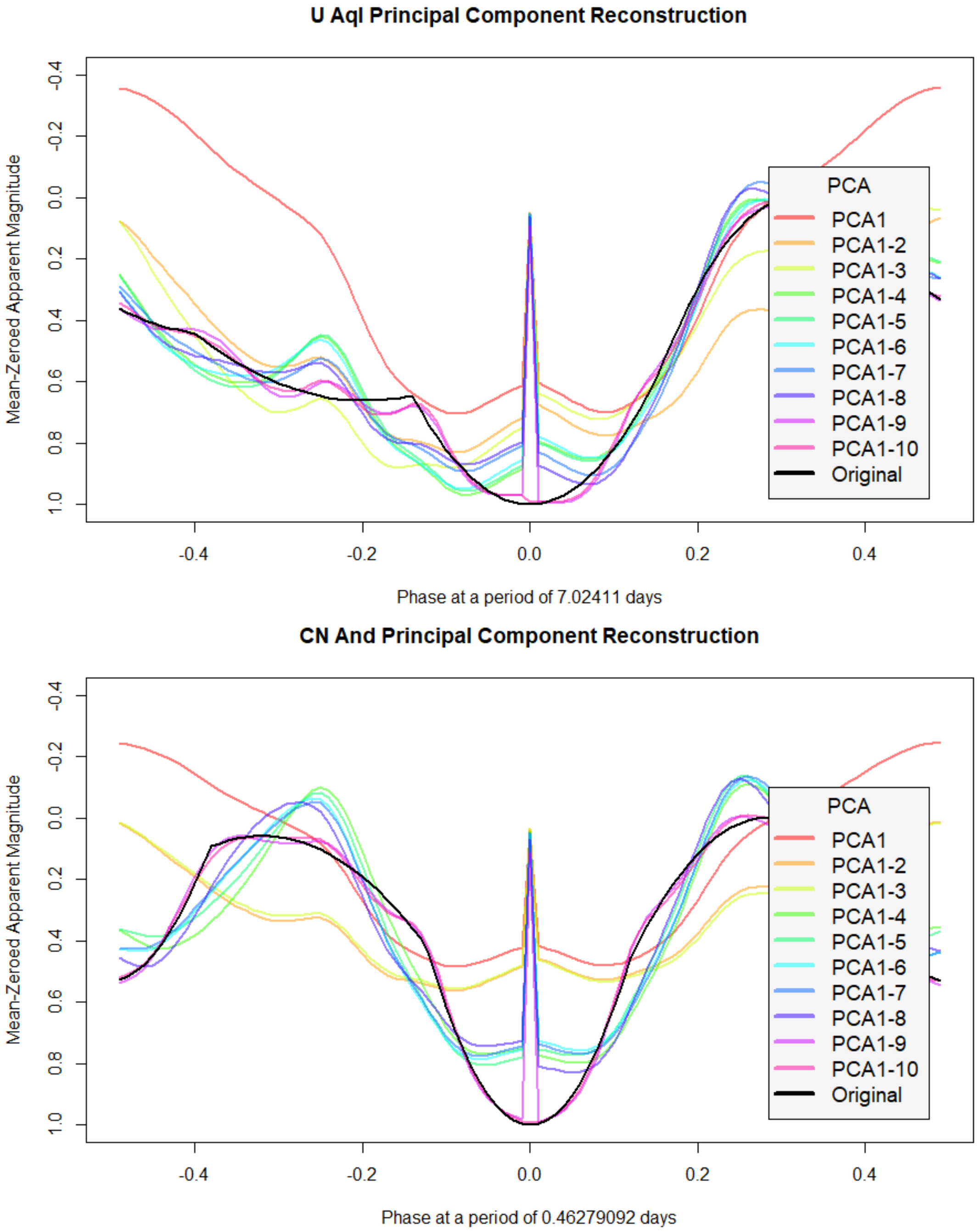


FIGURE 6.14: Plot showing the reconstruction of a selection of SkycamT light curves using the learned PCA model. The top light curve is of U Aquilae, a Classical Cepheid pulsating variable. The bottom light curve is of CN Andromedae, a β Lyrae eclipsing binary.

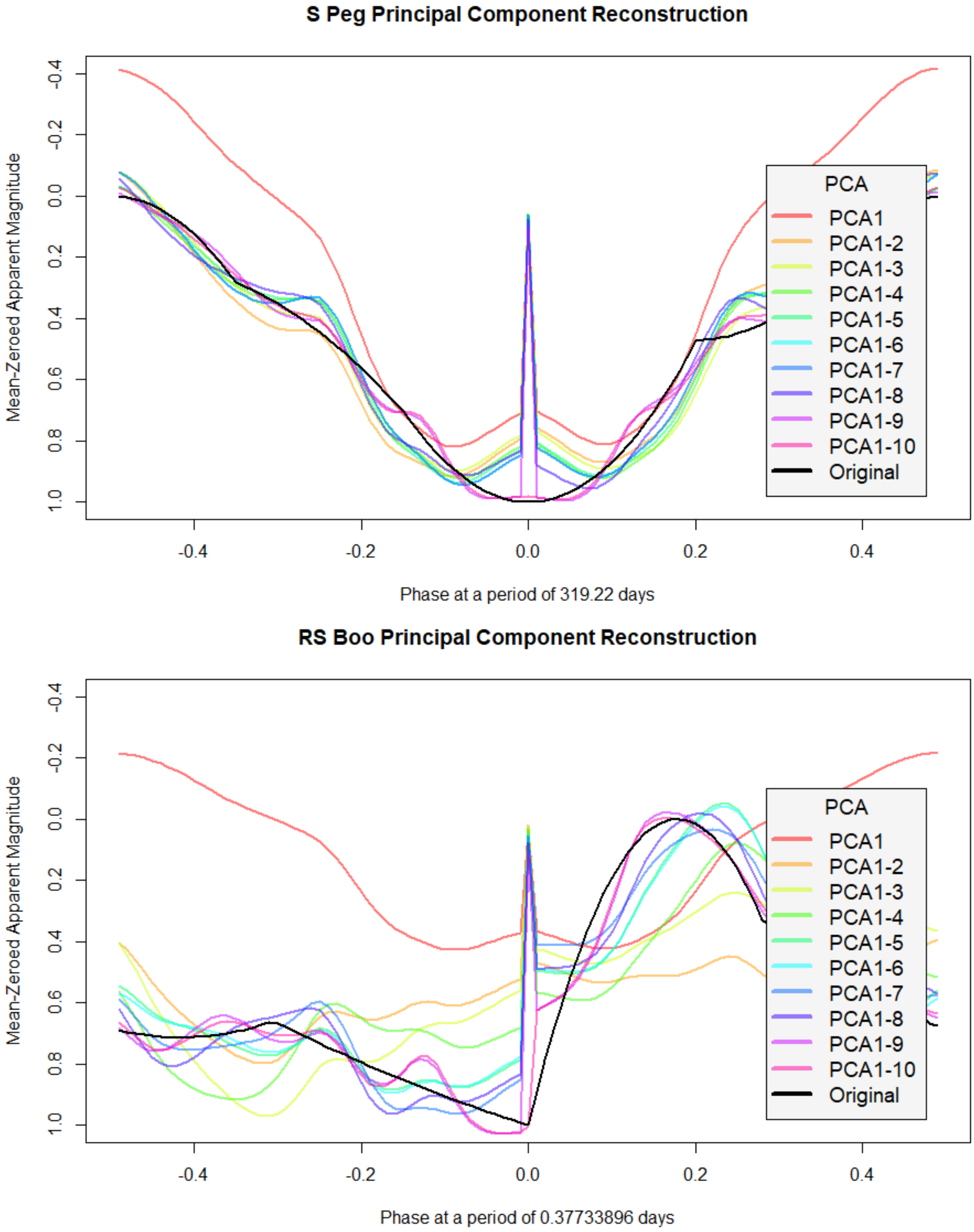


FIGURE 6.15: Plot showing the reconstruction of a selection of SkycamT light curves using the learned PCA model. The top light curve is of S Pegasi, a Mira-type Long Period Variable. The bottom light curve is of RS Bootis, a fundamental mode pulsating RR Lyrae variable.

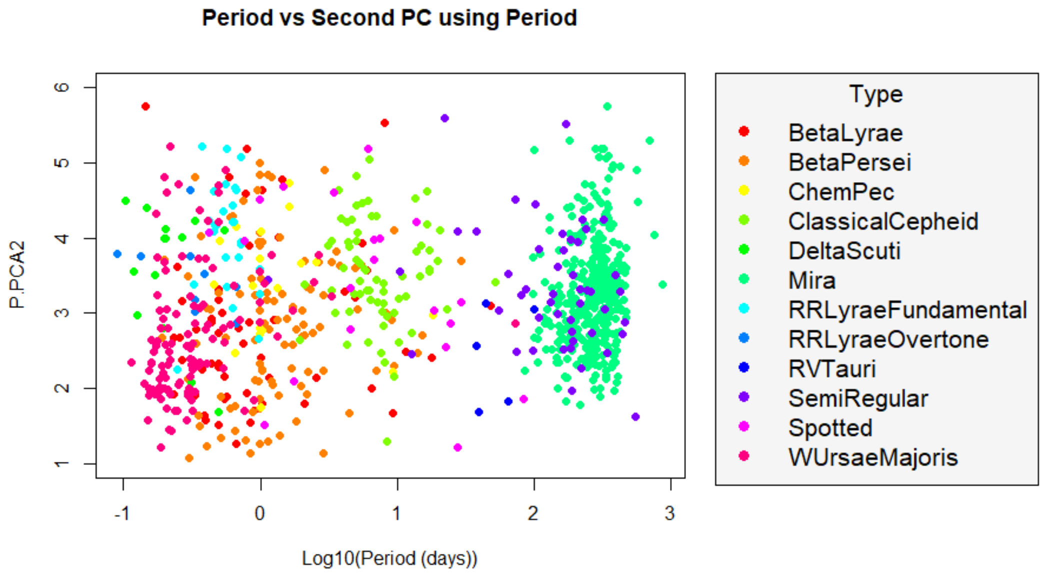


FIGURE 6.16: Plot showing the base-10 logarithm of the GRAPE-determined Period feature against the P.PCA2 feature, the second principal component calculated by the PolyFit model and the PCA model at the estimated GRAPE period. The feature partially seperates the pulsating variables and eclipsing binaries with a specific stength at classifying RR Lyrae fundamental mode and overtone mode pulsators from the short-period eclipsing binaries of similar period range.

6.2.4 Interpolated Statistical Features

The features extracted directly from the PolyFit interpolation are not the only set of useful information extractable from the PolyFit method. The Random Forest classification models trained in chapter 5 on the set of variability indices and Fourier decomposition features assigned a low importance to many of the statistical properties of the light curve such as standard deviation, kurtosis and amplitude with skewness and the Fourier decomposed amplitude of the sinusoidal fit to the primary period remaining the only useful features. It is possible that these features are not useful in the classification task. Alternatively, the loss of performance might be a result of a substantial noise component propagating into the features. As a result, the larger noise causes a larger overlap between the features of different variable star classes leading to poorer discrimination.

The interpolated PolyFit model provides an alternative method to define these statistics by computing them directly from the interpolated data. A number of features are produced to potentially replace the original variability indices and Fourier components as the binned genetic PolyFit algorithm has reduced the high frequency noise and fit models which are capable of good quality modelling of any light curve shape. The 99 interpolated magnitude data points at the 99 evenly split phase positions $[x_i, y_i]$ replace the $[\phi_i, m_i]$ data points from the phase-binned light curve in the computation of the following features:

- PolyFit.Phase.Binned.Ratio

  The PolyFit phase binned ratio is a measure of how well sampled a light curve is. It is the ratio of phase bins containing at least one phased data point from the light curve n to the total number of phase bins N_{bins}. It is calculated by equation 6.12.

  $$\text{PBR}_{\text{PF}} = \frac{n}{N_{\text{bins}}} \tag{6.12}$$

  This feature is not directly related to the classification of variable star light curves but it is a measure of how well sampled the light curve is when epoch-folded around the estimated period. Poorly sampled light curves have less reliable fitted models.

- PolyFit.Goodness.of.Fit

  The PolyFit Goodness of Fit feature is a measure of how well the PolyFit model matches the phase-binned light curve data points and is defined as the χ^2 value of the fitted model prior to the addition of the penalty terms r_{cost} and m_{cost} as shown in equation 6.10. This feature is expected to be of moderate usefulness as it indicates light curves with multi-periodic signals as they have significant variance remaining after fitting the dominant period. The feature is limited by the presence of noise which also increases the χ^2 statistic along with poorly selected candidate periods although this would also disrupt the rest of the PolyFit features.

- PolyFit.Interpolated.Amplitude

  The PolyFit Interpolated Amplitude is calculated by computing equation 6.13 on the interpolated PolyFit magnitude data.

  $$a_{\text{PF}} = \frac{|\max(y) - \min(y)|}{2} \tag{6.13}$$

  where a_{PF} is the interpolated amplitude and y is the vector of interpolated magnitudes. This feature is expected to be very important as many variable star types are classified by the amplitude of their light curves and the interpolated amplitude is expected to be less distorted by noise than the other amplitude features.

- PolyFit.Interpolated.StD

  The PolyFit Interpolated Standard Deviation is calculated by computing equation 6.14 on the interpolated PolyFit magnitude data.

  $$\sigma_{\text{PF}} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \mu_{\text{PF}})^2} \tag{6.14}$$

where μ_{PF} is the mean of the y_i magnitudes computed by equation 6.15.

$$\mu_{\mathrm{PF}} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{6.15}$$

- PolyFit.Interpolated.Skewness

  The PolyFit Interpolated Skewness is calculated by computing equation 6.16 on the interpolated PolyFit magnitude data.

$$b_{\mathrm{PF}} = \frac{\frac{1}{n}\sum\limits_{i=1}^{n}(y_i - \mu_{\mathrm{PF}})^3}{\sigma_{\mathrm{PF}}^3} \tag{6.16}$$

- PolyFit.Interpolated.Small.Kurtosis

  The PolyFit Interpolated Small Kurtosis is calculated by computing equation 6.17 on the interpolated PolyFit magnitude data.

$$k_{\mathrm{PF}} = \left(\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{y_i - \mu_{\mathrm{PF}}}{\sigma_{\mathrm{PF}}}\right)^4\right) - \frac{3(n-1)^2}{(n-2)(n-3)} \tag{6.17}$$

- PolyFit.Interpolated.Beyond.1.StD

  The PolyFit Interpolated Beyond 1 Standard Deviation feature calculates the ratio of interpolated data points $[x_i, y_i]$ which have magnitude values outside of plus or minus the standard deviation of the mean of the interpolated magnitudes. This feature is calculated by equation 6.18.

$$(\mathrm{beyond}1\sigma)_{\mathrm{PF}} = \frac{n_{>\sigma_{\mathrm{PF}}}}{n} \tag{6.18}$$

where $n_{>\sigma_{\mathrm{PF}}} = \sum_{i=1}^{n}$ if $|y_i - \mu_{\mathrm{PF}}| > \sigma_{\mathrm{PF}} = 1$, otherwise 0.

- PolyFit.Interpolated.Range.Cumulative.Sum

  The PolyFit Interpolated Range of a Cumulative Sum is calculated by first computing the vector of cumulative sums S_{PF} using equation 6.19 on the interpolated PolyFit magnitude data.

$$S_{\mathrm{PF}} = \frac{1}{n\sigma_{\mathrm{PF}}}\sum_{i=1}^{l}(y_i - \mu_{\mathrm{PF}}) \quad \text{for } l = 1, 2, \ldots, n \tag{6.19}$$

The range of the cumulative sum $R(S_{\mathrm{PF}})$ is then determined using equation 6.20.

$$R(S_{\mathrm{PF}}) = \max(S_{\mathrm{PF}}) - \min(S_{\mathrm{PF}}) \tag{6.20}$$

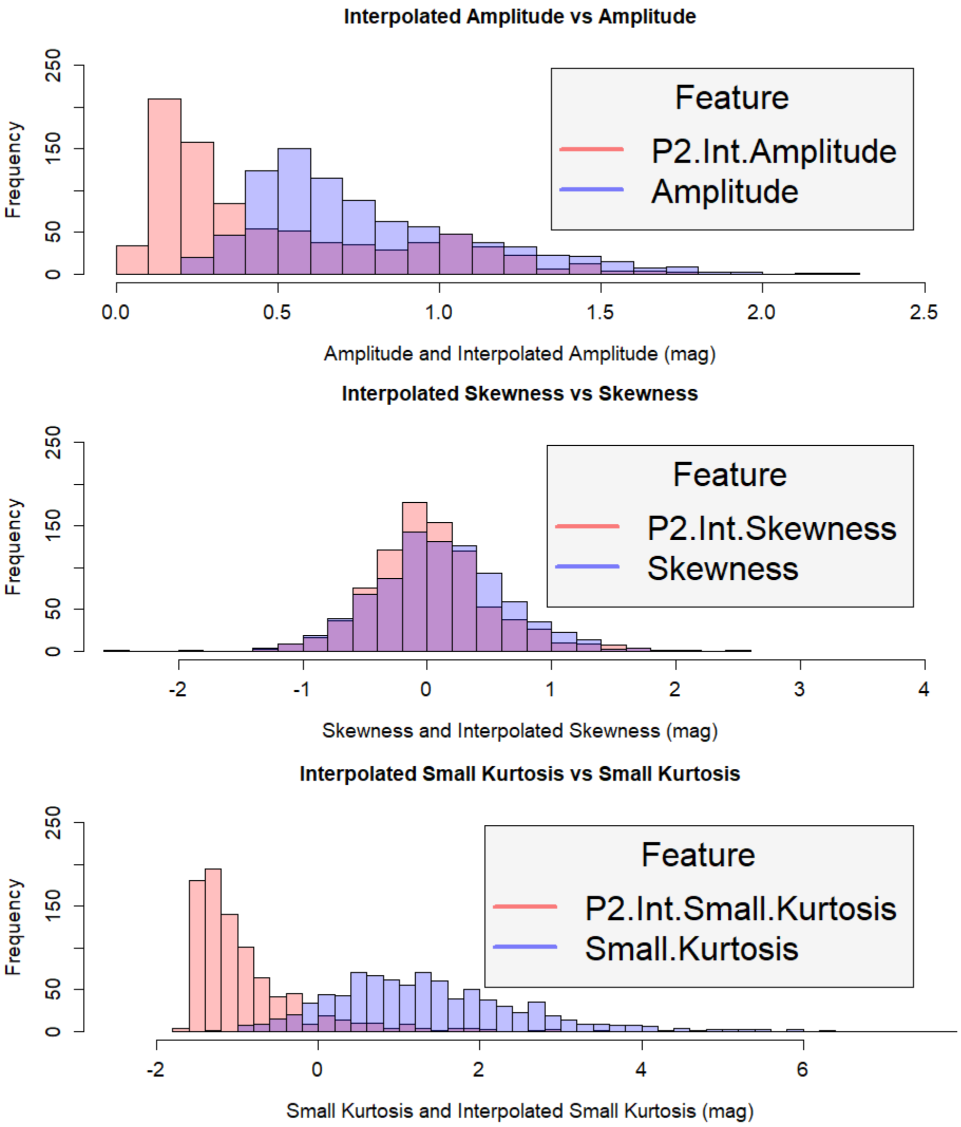


Figure 6.17: Histogram showing the distribution of the Interpolated features compared to the variability index features for the 859 SkycamT light curves. Many of the features exhibit superior performance due to their resilience to the noise in the data. This explains the interpolated amplitude being closer to zero and the narrower distribution of the interpolated skewness and interpolated small kurtosis. This figure contains Interpolated Amplitude (top), Interpolated Skewness (middle), and the Interpolated Small Kurtosis (bottom).

Using the 859 SkycamT light curves selected by GRAPE shown in table 2.1, we compute a PolyFit model for each light curve phased at 2× the GRAPE estimated period and calculate these interpolated features. These distributions of features are shown in a histogram relative to the equivalent feature from the non-interpolated data. Figure 6.17 (top) demonstrates the Interpolated Amplitude relative to the variability index Amplitude. The Interpolated Amplitude distribution is closer to zero than the Amplitudes for the 859 light curves. This is due to the interpolated model reducing the high frequency noise allowing the fit to more accurately reflect the true amplitude of the variability.

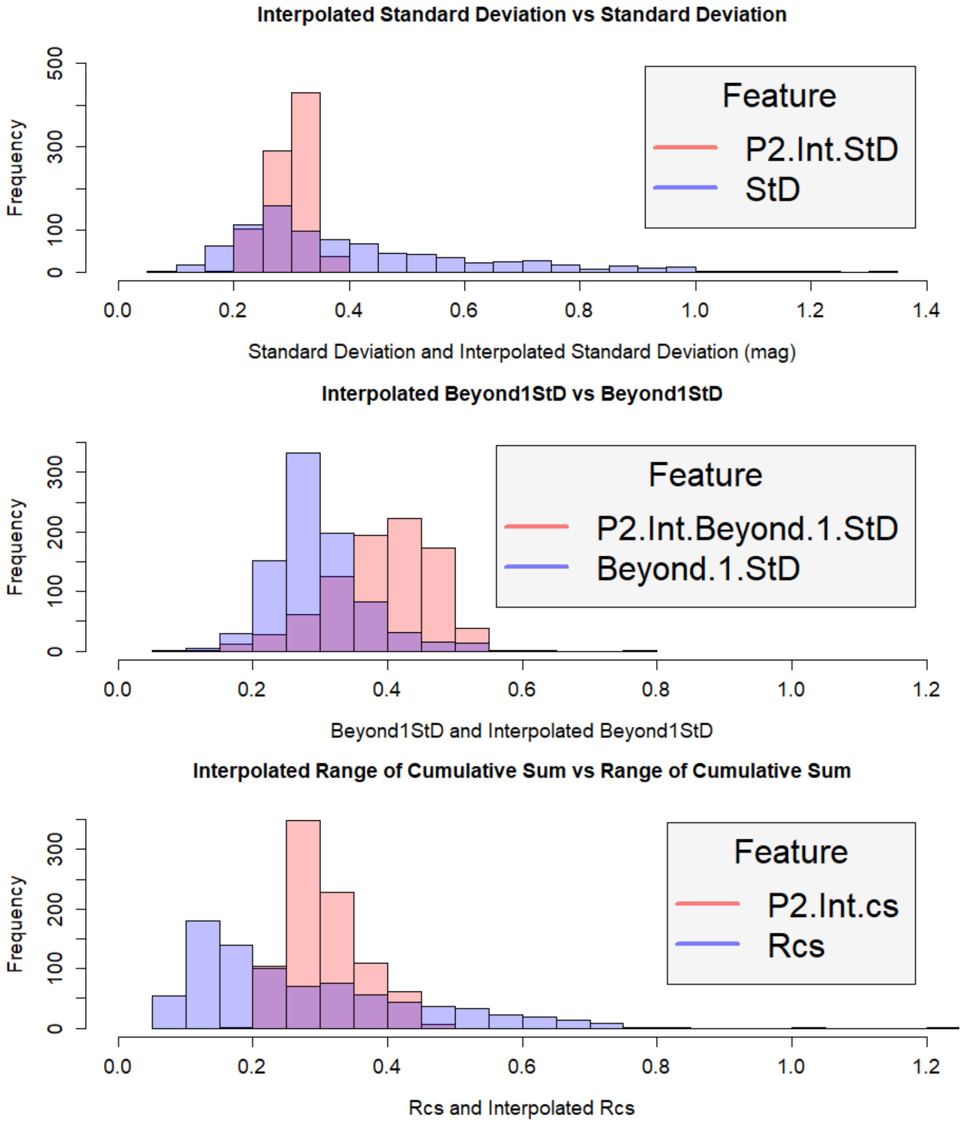


FIGURE 6.18: Histogram showing the distribution of the Interpolated features compared to the variability index features for the 859 SkycamT light curves. This figure contains Interpolated Standard Deviation (top), Interpolated Beyond 1 σ (middle), and the Interpolated Range of a Cumulative Sum (bottom).

Figure 6.17 (middle) shows the Interpolated Skewness feature relative to the variability index Skewness. The distribution of these two features is very similar which suggests that the skewness feature is not distorted by the noise. This is surprising as the skewness feature was not as dominant as expected in the feature selection in chapter 5 shown in figure 5.3. The conclusion that this is due to noise may not be accurate and it is possibly due to a sample bias due to lack of narrow eclipses in our set of β Persei eclipsing binaries. This bias is not due to noise and is a result of poor sampling of this class of object by the Skycam cadence. This is a problem which has been discussed for other surveys such as Kepler and the Large Synoptic Survey Telescope (LSST) (Prsa et al., 2011; Wells et al., 2017; Parvizi et al., 2014; LaCourse et al., 2015).

Figure 6.17 (bottom) demonstrates the distribution of the Interpolated Small Kurtosis feature relative to the variability index Small Kurtosis. This feature is interesting as the Interpolated Small Kurtosis feature is primarily negative for this set of light curves whereas the small kurtosis feature is positive. The Interpolated Small Kurtosis feature has a narrower distribution which is an effect of the reduced noise component in the interpolated feature which likely causes the difference in the distribution centres. Figure 6.18 (top) shows the interpolated standard deviation feature relative to the variability index standard deviation. The interpolated standard deviation has a much narrower distribution due to a lower noise component. Figure 6.18 (middle) demonstrates the interpolated Beyond 1 Standard Deviation feature. This feature adopts higher values as the approximately 0.2-0.25 mag white noise component suppresses the non-interpolated feature in the low amplitude variable classes. Figure 6.18 (bottom) shows the interpolated Range of a Cumulative Sum feature which also exhibits a narrower distribution relative to the variability index Range of a Cumulative Sum feature due to the noise reduction from the interpolated PolyFit model.

The performance of these features on the 859 SkycamT light curves appear superior to the associated variability indices. This demonstrates the strength of this approach although as the period is an important component of the generation of these models, failure of the period estimation method will disrupt these features substantially more than the variability indices. As the period is the dominant feature in the variable star classification task, the poorer interpolated features are unlikely to be the dominant source of error in the classification of poorly detected variability.

6.2.5 Training on PolyFit Features

Using the features described in the sections above, a dataset was generated from the 859 SkycamT light curves from the GRAPE period match operation. A set of 38 features are produced and they are displayed in table 6.9. These features include the GRAPE estimated period, and the PolyFit features, both PCA and interpolated variability indices generated by a PolyFit model phased at the GRAPE estimated period. This process is repeated to produce a second set of PolyFit features at two times the GRAPE estimated period, the ‘double period’. The final feature is the ratio of the variances for the period and the double period.

A Random Forest classifier was selected to determine the individual importance and overall performance of these features in the classification task of assigning the 859 SkycamT light curves to the correct variability class out of 12 possible classes. We perform a hyper-parameter optimisation on the three Random Forest arguments using the method

Table 6.9: The 38 features used in the classification of the 859 SkycamT light curves using the PolyFit algorithm.

Feature	Description
Period	Period (P) estimated by the GRAPE method
P.Binned.Ratio	PolyFit phase binned ratio at P
P.Goodness.of.Fit	χ^2 of PolyFit model at P
P.Int.Std	PolyFit interpolated Standard Deviation at P
P.Int.Skewness	PolyFit interpolated Skewness at P
P.Int.Small.Kurtosis	PolyFit interpolated small Kurtosis at P
P.Int.Amplitude	PolyFit interpolated Amplitude at P
P.Int.Beyond.1.StD	PolyFit interpolated Beyond 1 StD at P
P.Int.cs	PolyFit interpolated Range of a Cumulative Sum at P
P.PCA1	PolyFit interpolated PC1 at P
P.PCA2	PolyFit interpolated PC2 at P
P.PCA3	PolyFit interpolated PC3 at P
P.PCA4	PolyFit interpolated PC4 at P
P.PCA5	PolyFit interpolated PC5 at P
P.PCA6	PolyFit interpolated PC6 at P
P.PCA7	PolyFit interpolated PC7 at P
P.PCA8	PolyFit interpolated PC8 at P
P.PCA9	PolyFit interpolated PC9 at P
P.PCA10	PolyFit interpolated PC10 at P
P2.Binned.Ratio	PolyFit phase binned ratio at 2P
P2.Goodness.of.Fit	χ^2 of PolyFit model at 2P
P2.Int.Std	PolyFit interpolated Standard Deviation at 2P
P2.Int.Skewness	PolyFit interpolated Skewness at 2P
P2.Int.Small.Kurtosis	PolyFit interpolated small Kurtosis at 2P
P2.Int.Amplitude	PolyFit interpolated Amplitude at 2P
P2.Int.Beyond.1.StD	PolyFit interpolated Beyond 1 StD at 2P
P2.Int.cs	PolyFit interpolated Range of a Cumulative Sum at 2P
P2.PCA1	PolyFit interpolated PC1 at 2P
P2.PCA2	PolyFit interpolated PC2 at 2P
P2.PCA3	PolyFit interpolated PC3 at 2P
P2.PCA4	PolyFit interpolated PC4 at 2P
P2.PCA5	PolyFit interpolated PC5 at 2P
P2.PCA6	PolyFit interpolated PC6 at 2P
P2.PCA7	PolyFit interpolated PC7 at 2P
P2.PCA8	PolyFit interpolated PC8 at 2P
P2.PCA9	PolyFit interpolated PC9 at 2P
P2.PCA10	PolyFit interpolated PC10 at 2P
Period.Double.Ratio	Ratio of the VRP at P and 2P

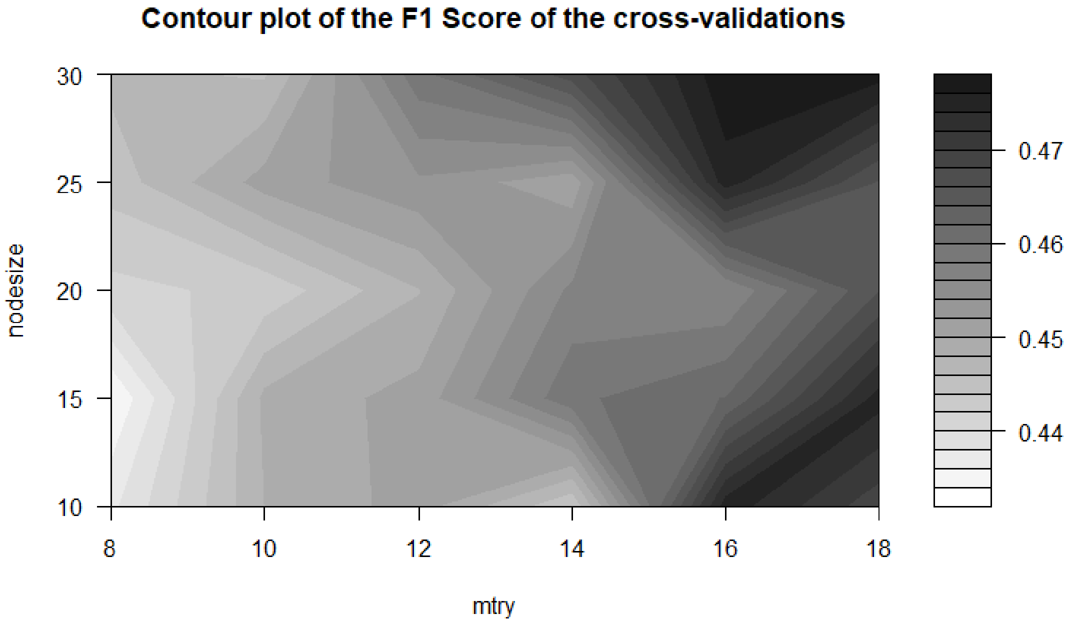


FIGURE 6.19: Contour plot of the F1 score performance of the 5-fold cross-validation using a Random Forest classifier with 500 trees on the 859 SkycamT light curves as a function of the m_{try} and nodesize hyperparameters. The optimal hyperparameters are $m_{try} = 16$ and nodesize = 30.

described in chapter 5. We found that the number of trees in the Random Forest model did not heavily influence the performance of the classification task therefore we kept this value at ntree = 500. The m_{try} and nodesize parameters are determined using a grid-search from 8 to 18 with intervals of 2 for the m_{try} parameter and 10 to 30 with intervals of 5 for the nodesize parameter. Figure 6.19 demonstrates the surface plot generated from the F1 Score of a 5-fold cross-validation with 2 repeats, our figure of merit (FoM) in this experiment as a function of the Random Forest arguments m_{try} and nodesize. This hyperparameter optimisation procedure selects the optimal values as $m_{try} = 16$ and nodesize = 30 for 500 trees in the Random Forest with a12-class mean F1 score of 0.477 with a standard deviation of 0.087.

Figure 6.20 demonstrates the Receiver Operator Characteristic (ROC) Curve of the model trained with the optimal hyperparameters. The poorest performance is found on the β Lyrae eclipsing binaries, overtone mode RR Lyrae variables and the Semiregular Long Period Variables. The PolyFit features are primarily descriptors of the shape of the folded light curves. The β Lyrae eclipsing binary folded light curves appear similar to the β Persei eclipsing binary light curves which leads to misclassifications between these two classes from purely light curve shape features (Malkov et al., 2007; Hoffman et al., 2008). The overtone RR Lyrae light curves are very similar in shape to the much more common W Ursae Majoris eclipsing binaries which results in a poor performance of this

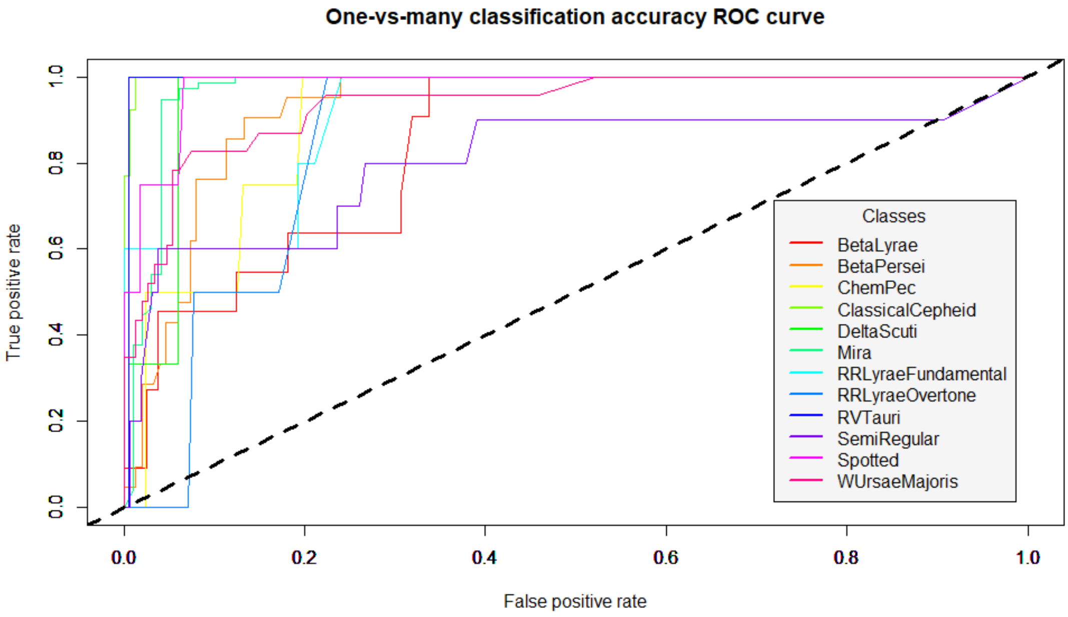


FIGURE 6.20: ROC curve of the model trained at the optimal hyperparameters. Many classes exhibit good performance as they obtain high recall of the true light curves of their respective classes whilst minimising the false positives. The poorest performing classes are the β Lyrae eclipsing binaries, overtone mode RR Lyrae variables and the Semiregular Long Period Variables. This failure is due to limitations in the description of these classes from a purely epoch-folded light curve shape position.

class. The poor performance of the Semiregular Long Period Variables is easily explained as this class of variable does not have a consistent period or amplitude which results in long term modulation of the light curve. This important class-specific information is discarded when the light curve is epoch-folded and averaged in the phase bins.

The mean decrease Gini of the Random Forest can be used to display the importance of the features in the trained classification model. Figure 6.21 demonstrates the importance of the top 20 features in the classification of the 859 light curves. The period is the dominant feature as expected by the definition of many variable star types being based on this property of the variable. The interpolated amplitude features are the next most important which again relates to the amplitude being an important part of the definition of the variability classes. The interpolated Range of a Cumulative Sum and interpolated Skewness features are also of higher importance than many other features. The interesting selection is the use of the second principal component of the PolyFit model folded at the period. In the previous subsection this was highlighted as a possible discriminator between a number of classes which overlap strongly in the Period and Amplitude feature space. The mean decrease Gini feature can also be determined for a specific class and for the two RR Lyrae variable types and the W Ursae Majoris eclipsing

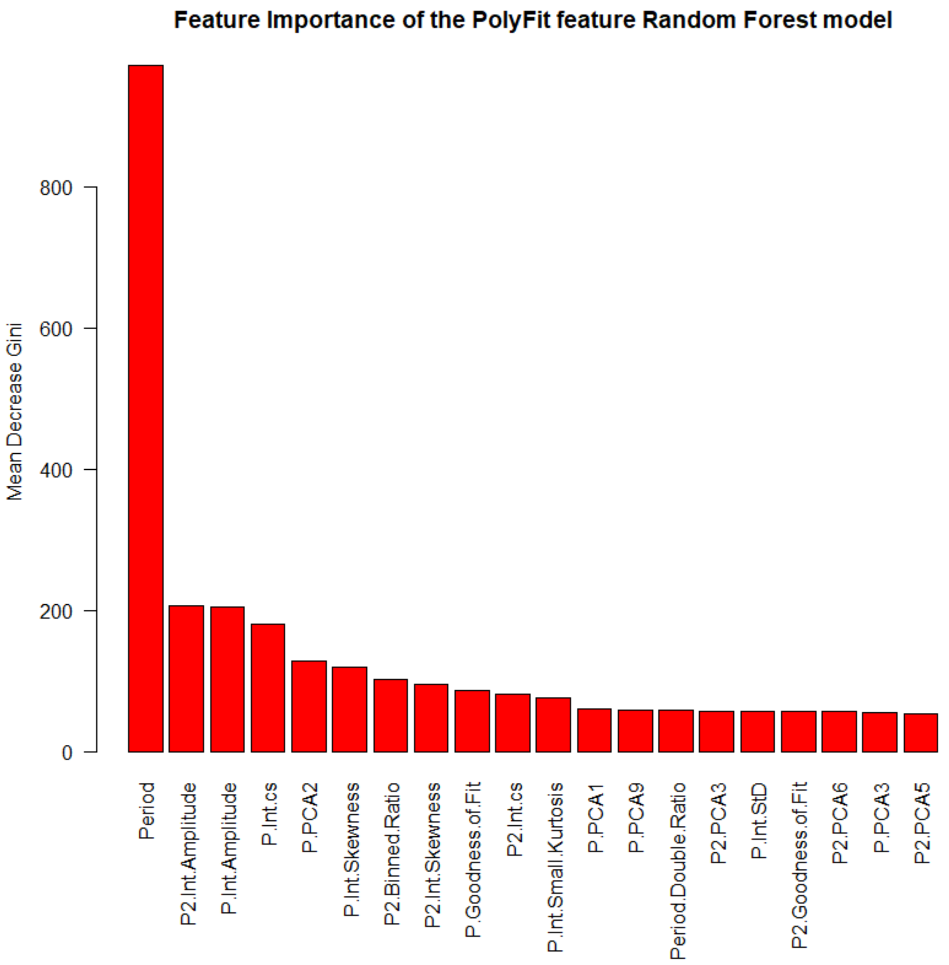


FIGURE 6.21: Bar Plot of the Mean Decrease Gini of the top 20 features in the PolyFit feature Random Forest model. As expected, the Period feature is dominant in the classification task followed by the interpolated amplitudes. The second principal component of the PolyFit model folded at the period is also important as was suspected earlier. Other important features include the interpolated Range of a Cumulative Sum feature and the interpolated Skewness. Skewness is normally a more important feature in the detection of eclipsing and dimming variable stars yet it appears diminished in the Skycam light curves possibly due to a more robust description from the PCA features.

binaries this feature became the second or third most important feature replacing the interpolated amplitude features although period still retained the top position.

These trained models indicate that the PolyFit derived features contain significant knowledge on the shape and distribution of the variable light curves. These features allow the discrimination of the twelve chosen variability classes with reasonably strong performance using the SkycamT light curve database. The features are also limited as they are specifically tuned to detect shaped based information without taking into account the suitableness of the initial epoch-folding operation. This means the features rapidly loose importance and meaning in the event of a light curve being semi-periodic or non-periodic such as was seen in the performance loss on Semiregular Long Period Variables.

The other primary limitation is the dependency on period. If an incorrect period is estimated, the interpolated features will be poor and possibly inferior to the variability indices they were designed to replace. Ultimately, due to the importance of period, an incorrect period estimation is likely to have wider ranging problems than those caused by a poorly generated PolyFit. The PolyFit features have been shown to be powerful on the noisy SkycamT light curves yet they cannot be applied to the light curve classification task alone and should be used in concert with the features discussed in chapter 5. Regardless of the strengths and weaknesses of this approach, this investigation has shown that representation learning methods can generate new and informative features for learned light curve classification models.

The model produced by the PolyFit interpolation clearly outperforms the model trained using the traditional engineered features in chapter five. The ROC curve shown in figure 6.20 shows that the 12 variability classes have higher AUCs than the 8 variability classes in 5.5. This shows that the representation features are conveying superior information to the machine learning algorithm than the traditional features which appear to have substantial difficultly with the noisy light curves in the SkycamT data. The use of principal components is also useful in decorrelating the features as each principal component is orthogonal to the others. Correlated features can cause difficulties in training well performing models and therefore this decorrelation can improve the final classification.

It is possible that the traditional features may still contain information not described by the PolyFit approach and therefore it is wise to offer both sets of features to the machine learning algorithms and let them select the best performing features for the task from a large selection of statistics. This is demonstrated in the following chapter as the techniques discussed in chapters five and six are combined to produce a SkycamT classification pipeline.

Chapter 7

Automated Classification Pipeline

The previous chapters have discussed the importance of accurate and efficient period estimation for the variable classification task combined with additional features, both periodic and non-periodic which describe many properties of a variable light curve. The introduction of GRAPE and the PolyFit features are designed around improving the performance of Machine Learning classifiers in the identification of variable light curves in the Skycam data. The Skycam database can be explored for interesting objects by integrating these novel techniques into an automated pipeline designed to select cross-matched variable light curves, generate feature data for the light curves and train machine learning classifiers for the detection and classification of candidate variable light curves.

In this chapter the techniques utilised in the development and deployment of the Skycam Automated Classification Pipeline are documented, a system which interfaces with the SkycamT SQL database (Mawson et al., 2013) to train variability detection and classification models using a set of 109 features which describe the periodicity, shape, variability and sampling of the Skycam light curves. The light curves are corrected for long-term seasonal trends in the data which can result in incorrect period estimation and poor generated features. Outliers are removed using an iterative σ–k clipping routine which cuts poor quality data points on a light curve by light curve basis. The addition of a feature to describe the colour of the SkycamT objects is developed through the cross-matching to another catalogue as the Skycam data lacks colour information. Matched objects are then used to determine the colours of unmatched objects through a Random Forest imputation method.

The 109 features are utilised with a training set of cross-matched variables to train a variable star classifier capable of classifying variable light curves into one of 12 variability classes. The high-confidence classifications from the variable star classifier are then used to train a variability detection classifier, a binary classifier designed to use a subset of the features, the 35 variability indices, to classify all other light curves with more than 100 data points as variable or non-variable. The light curves classified as variable are then selected for the computation of the full set of 109 features for novel variable star classification. Figure 7.1 shows a flowchart demonstrating the components of the SkycamT automated classification pipeline where the SkycamT light curve database undergoes a trend removal operation followed by a feature extraction process. The set of feature extracted light curves are then connected to a cross-matched catalogue dataset of known variable objects for ground truth label information to train a machine learned classification model which is then used to create a machine learned variability detection model.

7.1 Trend Removal

In previous chapters the definition and cause of spurious periods was discussed. These periods are a result of sampling periodicity and are common to ground-based surveys due to the motion of the Earth, both rotationally and in orbit around the Sun combined with the tilt of the Earth's axis. These motions can result in more than just a sampling periodicity as many of the effects that prevent observations can also produce a sinusoidal variation in the light levels received from an object depending on the time instant of the measurement. This is easily understood by considering the light changes throughout the day. The telescope is shut during the daytime which produces the spurious period in the light curve data yet the telescope opens shortly after sunset and closes shortly before sunrise. The sky brightness is higher at this time of the day due to scattered solar light from higher in the atmosphere. Either side of this 'civil' twilight there is additional sky brightness changes due to interactions between solar radiation and the atoms in the upper atmosphere named 'nautical' and 'astronomical' twilight.

High quality photometry takes this into account by using various techniques to construct background maps around sources (Newell, 1983; Bijaoui, 1980; Beard et al., 1990; Almoznino et al., 1993) although this can be difficult in dense fields (an area of the sky with multiple closely packed sources) (Bertin and Arnouts, 1996). There are also monthly and yearly variations due to the proximity between a target source and the moon and sun respectively. In many of the SkycamT light curves there is a clear yearly periodicity generating a sinusoidal signal within the data. This is not simply a spurious period as many light curves when epoch-folded at 365 days reveal this low amplitude sinusoidal shape. In light curves with low amplitude astrophysical periodicity, this signal was not detected due to the presence of this yearly variability.

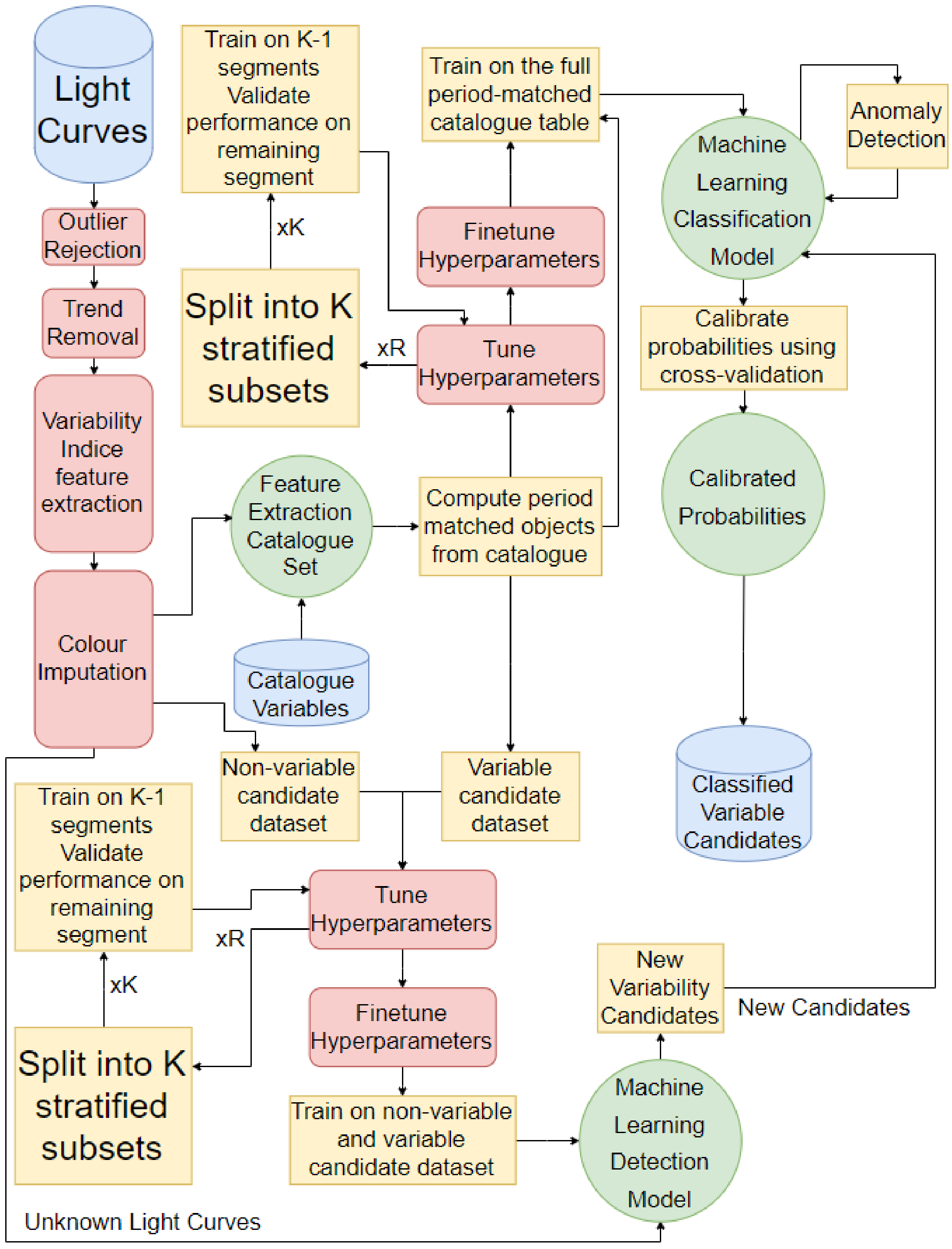

Light Curves
Outlier Rejection
Trend Removal
Variability Indice feature extraction
Colour Imputation
Train on K-1 segments Validate performance on remaining segment
xK
Split into K stratified subsets
xR
Train on the full period-matched catalogue table
Finetune Hyperparameters
Tune Hyperparameters
Machine Learning Classification Model
Anomaly Detection
Calibrate probabilities using cross-validation
Calibrated Probabilities
Classified Variable Candidates
Feature Extraction Catalogue Set
Compute period matched objects from catalogue
Catalogue Variables
Non-variable candidate dataset
Variable candidate dataset
Train on K-1 segments Validate performance on remaining segment
xR
Tune Hyperparameters
xK
Finetune Hyperparameters
Split into K stratified subsets
Train on non-variable and variable candidate dataset
New Variability Candidates
New Candidates
Machine Learning Detection Model
Unknown Light Curves

FIGURE 7.1: Flowchart demonstrating the components of the SkycamT automated classification pipeline. The pipeline uses a cross-matched catalogue of variable objects to automatically train models for variability detection and classification. The pipeline automatically tunes the hyperparameters and selects appropriate machine learning algorithms based on a cross-validation performance measure. Newly detected variables are detected and classified with calibrated probabilities to match the expected distribution of these variable objects. The final light curves are returned as a data table of unknown variable light curves.

There are two options to attempt to correct this yearly trend. The first is to go back to the software which performed the initial source extraction and photometry on the raw SkycamT images. This is likely the ideal method but it requires significant re-engineering of the original image processing pipeline which was not addressed during this thesis. As a result, the second option is adapted to attempt to remove the yearly sinusoidal trend at the light curve level. In this section we describe the method utilised to reliably remove this yearly periodicity without causing significant changes to the light curves of true long period variables through the determination of seasonal variability on large samples of SkycamT light curves.

7.1.1 Trend Filtering Algorithm

Our method is adopted from the Trend Filtering Algorithm (TFA) which utilises the statistics of a large number of light curves to determine variability present in a bulk of the data in a field (Kovacs et al., 2005). The TFA was designed to replace filtering using a low-pass filter such as a spline or low-order polynomial fit as these filters are tuned specifically for one period and more importantly, they are unable to distinguish the signal due to a trend and the signal due to true variability. The algorithm utilises the premise that a large number of stars from a survey will have similar systematic effects which introduce these trends across many light curves. A trend template can be modelled by inspecting the common variability traits for a subset of the light curves named the template set. This light curve subset is chosen randomly from the total set of light curves with care taken not to introduce biases from low quality and poorly sampled light curves. The original paper detailing this trend filtering algorithm utilised 50 light curves in the template training set (Kovacs et al., 2005).

The algorithm first assumes that the light curves contain the same number of data points all sampled at the same time instants. This is a poor assumption on the Skycam dataset but we will address a method to correct this shortly. A template filter is produced from a subset of zero-averaged light curves of size M. The filter $F(i)$ for $i = 1, 2, \ldots, N$, where N is the number of data points in the light curves, is a weighted linear combination of the light curves present in the template training subset of light curves $X_j(i)$ for $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, M$ as shown in equation 7.1.

$$F(i) = \sum_{j=1}^{M} c_j X_j(i) \tag{7.1}$$

where c_j are the coefficients which weight the contribution to the template filter by the light curve X_j. These coefficients are determined through the minimisation of equation

7.2.

$$\mathbb{D} = \sum_{j=1}^{N} (Y(i) - A(i) - F(i))^2 \tag{7.2}$$

where $\mathbb{D}$ is the detrending error, Y is the target light curve being filtered, A is the best estimate of the detrended light curve and F is the template filter. Using a frequency analysis or signal reconstruction method, A can be iteratively tuned after initialisation as a constant at the mean magnitude of the target light curve. This iterative method for determining c_j is computationally heavy and is not adopted for the Skycam light curves leaving $c_j = 1$ (Kovacs et al., 2005). The corrected time series is then simply computed using equation 7.3.

$$\hat{Y}(i) = Y(i) - \sum_{j=1}^{M} c_j X_j(i) \tag{7.3}$$

where $\hat{Y}(i)$ is the i^{th} magnitude of the detrended light curve, $Y(i)$ is the i^{th} magnitude of the initial trended light curve and the summation is $F(i)$, the magnitude value of the template filter for the i^{th} data point.

7.1.2 Light Curve interpolation

This method of removing trends is powerful as it does not depend on any specific period or even regular periodicity to the removed trends. Any common variations amongst the subset of template training light curves are modelled and removed. Unfortunately, this method requires there to be an equal number of data points in the light curves and the i^{th} data points for every light curve must be sampled at the same time instant.

Our solution to this problem is to perform a large-scale time instant interpolation operation on the template training light curves in order to align their data points onto a grid. Whilst only a few light curves will contain a data point at a specific interpolated time instant, for a sufficiently large subset of template training light curves there should still be 50 data points from 50 independent light curves in many of the interpolated time instant bins. This interpolation routine will also destroy the low period trend information and therefore these trends cannot be corrected. This is acceptable as an additional step is performed to limit the trend removal to yearly periods.

The interpolated grid is defined by a starting time instant, an ending time instant and a number of bins across the time instant space. The starting time instant was determined by querying the SkycamT database for the time instant with the minimum Modified Julian Date (MJD) value. This value is subtracted by 10 to provide an empty border to the time instant space, rounded to the nearest complete day and assigned as the starting time instant. This MJD value is 54884 days and corresponds to the calendar date of the 22$^{\text{nd}}$ of February, 2009. A similar procedure is used to select the ending time instant by querying for the maximum MJD value, adding 10 to it and rounding it to the nearest day. This MJD value is 56079 days and corresponds to the calendar date of the 1$^{\text{st}}$ of June, 2012. This corresponds to a maximum timespan of 1195 days. For the number of bins on the grid, 56,000 bins were selected corresponding to approximately 30 minutes per bin. This is a very coarse resolution and is much lower than the resolution required to determine the correct period on interpolated variable light curves yet the resolution required for the period estimation task is computationally infeasible for the trend removal task.

This low resolution interpolation means that the Trend Filtering Algorithm cannot be applied as originally intended (Kovacs et al., 2005). The resolution is sufficient to remove trends down to under a day but at the complete destruction of most interesting periodic signals. Our solution is to specifically target the yearly seasonal trends which have resulted in the greatest difficulties through the modelling of this yearly trend using a harmonic sinusoidal model using equation 7.4.

$$y(t) = b_0 + ct + \sum_{j=1}^{2} \left[a_j \sin\left(\frac{2\pi j t}{365.24217} \right) + b_j \cos\left(\frac{2\pi j t}{365.24217} \right) \right] \tag{7.4}$$

where b_0 is the intercept magnitude which should be near zero for the zero-averaged training light curves, c is the linear component of the trend and the a_j and b_j coefficients identify the amplitude and phase of the yearly sinusoidal trend.

An initial template model was produced by randomly selecting 1000 SkycamT light curves to act as the training set for the algorithm. 1000 light curves were chosen as this was around the number needed for there to be sufficient interpolation bins with 50 data points from 50 different light curves for the harmonic fitting procedure based on our interpolation grid defined above. The x_i data points from the light curves are placed in the appropriate bin and the empty bins are assigned a null data value. Each interpolation bin is mean-averaged by all the data points present in that bin from the 1000 light curves and the mean value is recorded along with the number of data points used in the calculation. Each bin which had a number of aggregated light curve data points $n < 50$ are discarded leaving a template model with each data point being a mean aggregate of at least 50 light curves. The harmonic sinusoidal model from equation 7.4 is then fit to the template light curve using normal equation regression shown in equation 5.28. The resulting fit has 6 parameters which define the shape, amplitude and phase of the yearly trend. This model can then be subtracted from candidate light curves to remove this yearly trend. In addition to this model, we use a regularised random forest regression model to determine if there is any minor seasonal change in the mean magnitude of the objects due to noise statistics. Whilst this fault is not common, these variations can prevent a correct period estimation and must be prewhitened from the light curve.

This initial template exhibited two major faults. Firstly, the coefficients produced from the fitted model indicated an extremely low amplitude to the yearly trend, significantly below that seen in the individual light curves. Recomputing the template with a number of different training sets made little difference to this result. This is a result of the yearly trend not being homogenous across the entire Skycam sky which causes the trend in one part of the sky being reduced to near zero by the lack of this trend in all the other parts of the sky. The solution to this is to partition the sky into distinct regions using the Right Ascension and Declination coordinate system used by the Skycam light curves.

Using the minimum and maximum Right Ascensions and Declinations of the SkycamT objects with greater than 100 observations, the sky was partitioned into 64 (8×8) regions from 0° to 360° Right Ascension (in degrees) and $-40°$ to 80° Declination (in degrees). Each of these 64 regions has a distinct trend removal template model trained using the method defined above. This requires the computation of $64,000$ interpolated training light curves compared to 1000 therefore it requires additional time to complete but the amplitudes of the templates are a closer match to the trends seen in candidate light curves.

The second fault is due to some light curves lacking the yearly trend for an unknown reason. These light curves are not common yet it is important that the template not induce variability were there was previously none. To determine if it is appropriate to prewhiten a candidate light curve with the trend template model, the harmonic sinusoidal model in equation 7.4 is fit to the candidate light curve and the Pearson correlation coefficient between this model and the trend template model is calculated. If the Pearson correlation coefficient between these two models is negative, the candidate light curve is not prewhitened by the trend template model.

7.1.3 Trend Removal performance

The trend removal method should reduce the number of light curves which have periods near 365 days or half this period due to the yearly trend being the dominant periodic signal in the data. To evaluate if this is occurring we computed the GRAPE estimated period for the 6897 light curves used for training the PolyFit principal component analysis model shown in table 6.8. This was performed twice, once without use of the trend removal method and the second time with the trend removal. Figure 7.2 shows the

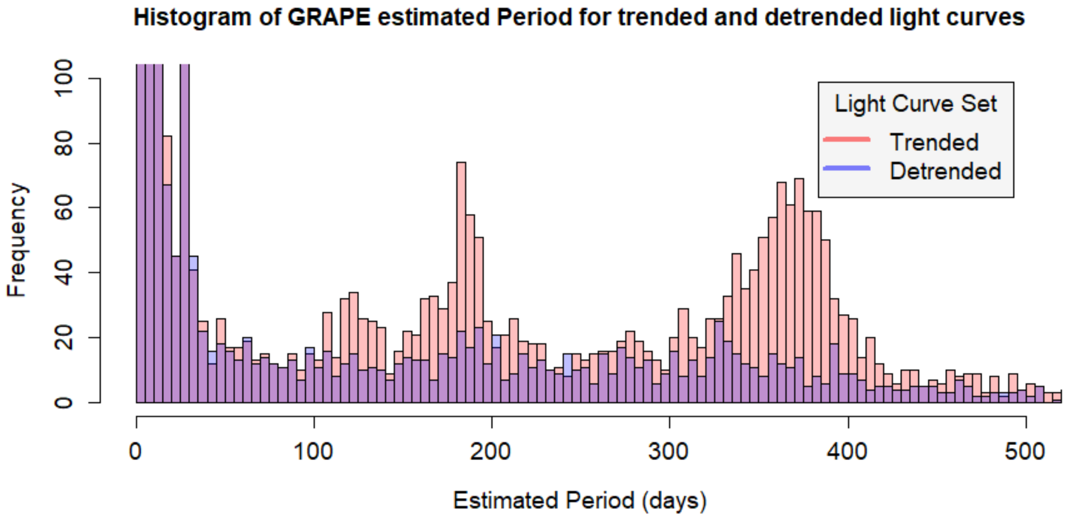


FIGURE 7.2: Histogram demonstrating the GRAPE estimated period for the set of SkycamT light curves shown in table 6.8. This estimated period is computed twice, once without applying the trend removal method and once after applying the trend removal. The red dataset shows the estimated periods of the trended light curves with a notable overestimation of periods near the yearly period and half yearly period. The blue dataset demonstrates the estimated periods of the same set of data after the trend removal method has been applied demonstrating a highly successful removal of the trends. This period range has not been completely reduced to zero as there are true periodic variables with these periods which shows the trend removal pipeline does not disrupt the light curves of these stars. The purple colour is due to an overlap between the two distributions by objects unaffected by the trend removal.

TABLE 7.1: Period estimation results on the 6897 light curves from the pre-trend removal trended dataset and the trend removed detrended dataset with a tolerance $\epsilon = 0.05$.

Data	Hit	Multiple	Alias	Unknown
Trended	0.076 ± 0.005	0.046 ± 0.004	0.118 ± 0.016	0.783 ± 0.008
Detrended	0.076 ± 0.005	0.047 ± 0.004	0.186 ± 0.018	0.720 ± 0.009

histogram of the estimated periods of the trended and detrended light curves. The overabundance of periods at the yearly period, and the $n = 2$ and $n = 3$ submultiples of this period has been corrected by the detrended light curves. There are still some objects with estimated periods at the yearly period and its submultiples due to objects with astrophysical periods at this location. As these objects have unique signals independent of the trend determined across the multiple template training light curves, the detrending method does not remove these signals which cannot be guaranteed if a low-pass filter method was utilised.

Table 7.1 shows the changes in the hit, multiple, alias and unknown rates of the trended and detrended datasets with a tolerance of 5%, $\epsilon = 0.05$. The error rates have been computed using the 95% confidence limits as computed by a bootstrapping method

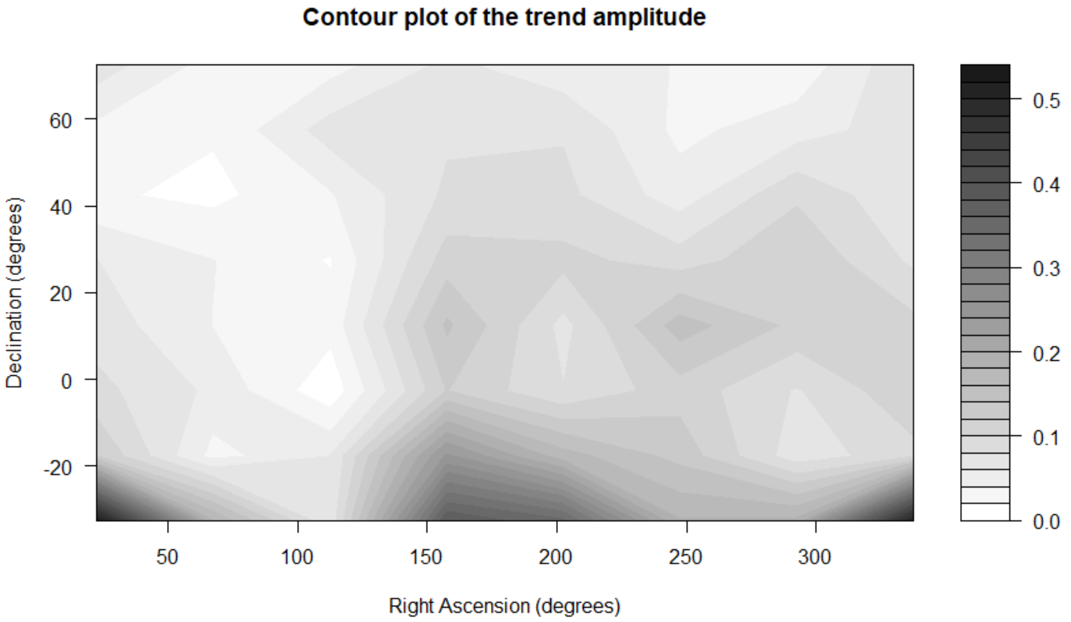


Figure 7.3: Contour plot showing the amplitude of the sinusoidal yearly trend model as a function of Right Ascension and Declination. The amplitude of the yearly trend is fairly uniform with a number of spikes at low declinations which is more likely a result of the poor quality of light curves at these positions as they never rise far above the horizon in La Palma. Most of the regions have a trend amplitude of 0.1 mag to 0.2 mag which is what is expected from inspecting candidate light curves.

with 10,000 runs. Most of the light curves from both sets have unknown periods relative to the catalogue periods yet the detrended light curves produce more aliases which may be due to the formerly yearly trend period light curves being identified as sidereal day periods. If this is the case it suggests that almost 6% of the light curves in this dataset were affected by the yearly trend. The detrending had little effect on the hit and multiple rate indicating that the light curves which are most effected by the yearly trend is still not of sufficient quality to identify the true period or alternatively have been poorly cross-matched to the AAVSO variable star catalogue.

The discovery of the yearly trend varying as a function of sky position suggests that the variation may be a result of a local sky brightness variation through the year. To investigate this phenomenon we study the amplitudes and phases of the 64 trend template models to discover the relationship between these model parameters and the right ascension and declination of the sky region. Figure 7.3 shows the contour plot of the amplitude of the sinusoidal template model across the La Palma sky. The amplitude is fairly constant across the sky with amplitudes of 0.1 mag to 0.2 mag which is similar to the amplitudes seen in individual candidate light curves exhibiting this trend. There is a spike at low declinations up to 0.5 mag which is likely a result of poor sampling and poor data quality of objects in this region of the sky as they are heavily influenced by

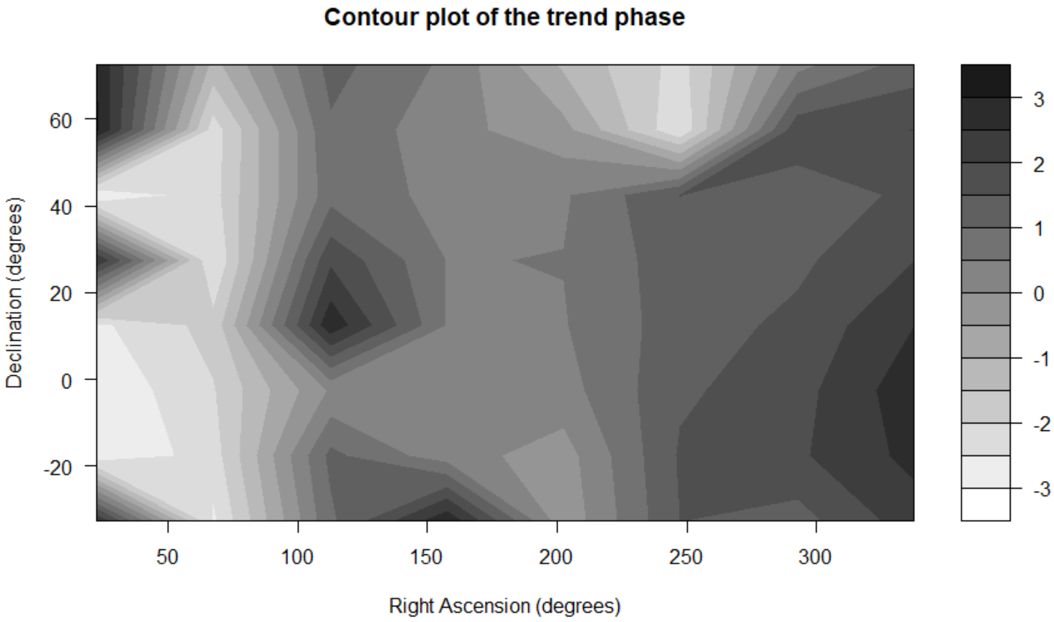


Figure 7.4: Contour plot showing the phase of the sinusoidal yearly trend model as a function of Right Ascension and Declination. There appears to be a trend in phase as a function of right ascension which indicates the time of the year that the maximum sky brightness occurs with the minimum phase occurring at approximately 0° to 60° right ascension. There is a lower relation to declination which suggests that the yearly variation is caused by sky brightness changes throughout the year due to seasonal variation because of the Earth's axial tilt. The phases overlap from $-\pi$ to π.

seasonal sampling as they are only visible for a small number of months each year as well as being low to the horizon when visible.

Figure 7.4 demonstrates the contour plot of the phase of the sinusoidal model across the La Palma sky where the phase has values $\phi = [-\pi, \pi]$ as well as looping at the boundaries similar to the epoch-folded phases. There appears to be a trend in the right ascension axis with a minimum value around 0° to 60° where the boundary loop occurs. This is more clearly visible in Figure 7.5 which projects the phase of the trend model along the right ascension axis aggregating the different declination values using a median function. There is a lower relation between the trend model phase and the declination which suggests that the yearly variation is caused by sky brightness changes throughout the year. When tested, there does not appear to be a correlation between sun distance and the brightness variations of this trend indicating it is not directly related to sun proximity. This trend is likely a result of a seasonal variation due to the Earth's axial tilt causing summer observations to have a higher sky brightness than the winter observations which have not been corrected by the photometry. Ultimately, the explanation for this trend is an interesting conundrum but is not important for the

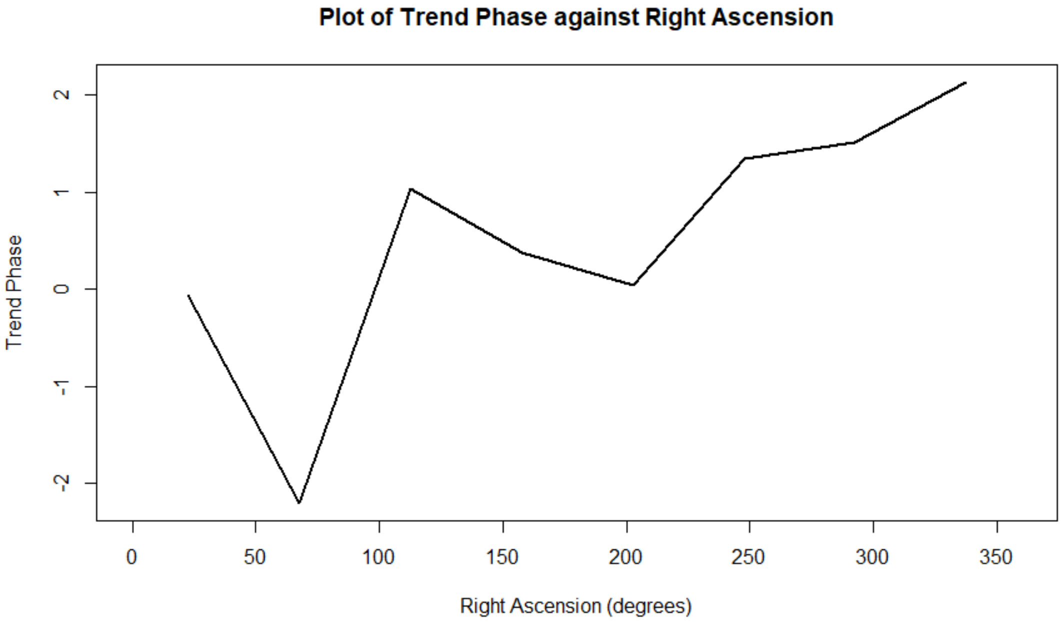


FIGURE 7.5: Projection of the phase of the trend template model along the right ascension axis with the declination values aggregated using a median function. This emphasises the claimed trend between the model sinusoidal phase and the right ascension as well as the minimum phase near 60°.

application of the trend removal method and thus we apply this method prior to the feature extraction of the light curves to improve the feature statistics.

7.2 σ–k clipping

In the initial Skycam image reduction pipeline, a percentile cut is used to remove the outer 1% of the maximum and minimum magnitude data points as outliers independent of the statistics of the light curves. The Skycam light curves are plagued by outliers and similar out-of-position data points. Many of these outliers have short duration jumps in magnitude which suggests a short duration dimming event. These groups of data points are difficult to deal with as they can appear similar to a short duration eclipse. The most common phenomenon to produce these artefacts are simply clouds causing a short duration localised extinction in the images. Despite the simplicity of this explanation, clouds have caused astronomers difficulty for a long time, they are quite difficult to correct in the light curves using reliable automated methods.

As these undesired events can alter the proportion of outliers on a light curve by light curve basis, simply performing percentile cuts across the entire database is not helpful. A cut which might remove outliers on one light curve might leave imperfections in another light curve. Additionally, cutting too aggressively will reduce the performance of the period estimation and feature extraction methods. We therefore employ a σ–k clipping algorithm which uses the effect of outlier data points on the median of the light curve to determine which data points to remove.

The σ–k clipping algorithm is a computationally efficient method of selecting outliers using an iterative procedure where the data points furthest from the median are removed and the median is recalculated. The process continues until the standard deviation of the data points changes by less than a given tolerance ξ. The procedure is defined as:

1. Calculate the standard deviation (σ) and median (m) of the light curve.
2. Remove all data points with magnitudes not within the range $m \pm k\sigma$.
3. Determine the change in the standard deviation from this removal $d = \frac{\sigma_{\text{old}} - \sigma_{\text{new}}}{\sigma_{\text{new}}}$.
4. If the value of d is greater than the tolerance $\xi < d$ then return to step 1.
5. If $\xi \geq d$ then exit identifying the data points identified as outliers for removal.

This method functions as the median is robust to outliers yet the standard deviation is highly influenced by them. This algorithm has two free arguments, the ξ tolerance for ending the iterations and the k parameter which determines the boundary for the outlier removal. On the Skycam light curves a low value of k will result in the flagging of many of the data points as outliers whereas a high value of k will result in minimal to zero data points being flagged as outliers. The optimal value of k must be determined based on the performance of a task highly influenced by the presence of outliers. We perform an experiment using a subset of the AAVSO cross-matched light curves with a cross-match distance tolerance of 10" on 12 different variability classes to produce a reliable set of objects. This identifies 3948 SkycamT light curves with a cross-matched catalogue class and period. 20 light curves were randomly selected from the 12 different variable star classes producing a dataset of 240 light curves.

For each of the 240 light curves we compute the GRAPE estimated period after the light curves have been processed by the trend removal method and the σ–k clipping outlier removal method with multiple values of k from 2 to 6 with intervals of 0.5, producing 9 distinct period estimation datasets. The tolerance adopted for this test is $\xi = 10^{-6}$ which was found to work well on the SkycamT light curves. The GRAPE estimated periods are then compared to the catalogue periods and the hit, multiple and alias rates are compared to the value of k utilised in their calculation.

Table 7.2 demonstrates the results of this experiment. There are two maximums in hit rate at $k = 3.5$ and $k = 5.5$ which have corresponding minimums in the aliasing rate.

Table 7.2: Period estimation results on the 240 subset light curves using the σ–k clipping algorithm for 9 values of k with a period estimation tolerance $\epsilon = 0.05$.

k	Hit	Multiple	Alias	Unknown
2.0	0.108 ± 0.033	0.063 ± 0.025	0.171 ± 0.046	0.692 ± 0.050
2.5	0.150 ± 0.038	0.108 ± 0.025	0.208 ± 0.054	0.588 ± 0.050
3.0	0.167 ± 0.042	0.079 ± 0.029	0.196 ± 0.046	0.592 ± 0.054
3.5	0.175 ± 0.042	0.075 ± 0.029	0.183 ± 0.046	0.600 ± 0.050
4.0	0.158 ± 0.038	0.075 ± 0.029	0.192 ± 0.050	0.613 ± 0.050
4.5	0.163 ± 0.042	0.079 ± 0.029	0.204 ± 0.050	0.588 ± 0.054
5.0	0.171 ± 0.042	0.071 ± 0.029	0.188 ± 0.050	0.604 ± 0.050
5.5	0.175 ± 0.042	0.079 ± 0.029	0.175 ± 0.046	0.608 ± 0.050
6.0	0.158 ± 0.038	0.096 ± 0.033	0.183 ± 0.046	0.600 ± 0.050

This is seen more clearly in figure 7.6 plotted using the data in table 7.2. This suggests that there might be two different distributions of light curve noise, those which require the minimal clipping of $k = 5.5$ and those with outliers closer to the mean requiring $k = 3.5$. Further analysis to determine the source of these two distributions would be advantageous although it is possibly due to a binary split between light curves heavily affected by clouds and the light curves which are clear of clouds. There is a little variation in the overall unknown rate other than a significant rise at $k = 2$ as many useful data points have been rejected as outliers at this value. This suggests that the σ–k clipping algorithm is modifying the strength of the alias relative to the true period of a signal rather than fully correcting or removing it. There is also an increase in the multiple rate at $k = 2.5$ although this does not mean that the period estimation is performing any better on eclipsing binaries.

The choice of an optimal value of k is not immediately clear from this experiment. Whilst there are the peaks in hit rate at $k = 3.5$ and $k = 5.5$, the 95% confidence limits indicate that these are not significant variations and therefore cannot be used to claim the existence of these distributions. For the pipeline we have adopted a value $k = 4$. Whilst this value is not optimal according to this experiment and can be changed in later revisions, it was selected from its performance during the pipeline development and the conclusions are not strong enough to reject this value.

7.3 Colour Imputation

The application of the Trend Removal method and the σ–k clipping algorithm to the 590,492 SkycamT sources with 100 or greater observations produces 590,492 light curves of sufficient quality to generate a set of 34 variability indices. These indices describe many properties of the light curves of the sources but they lack colour information and therefore the introduction of a feature to define the colour of the sources would be ideal.

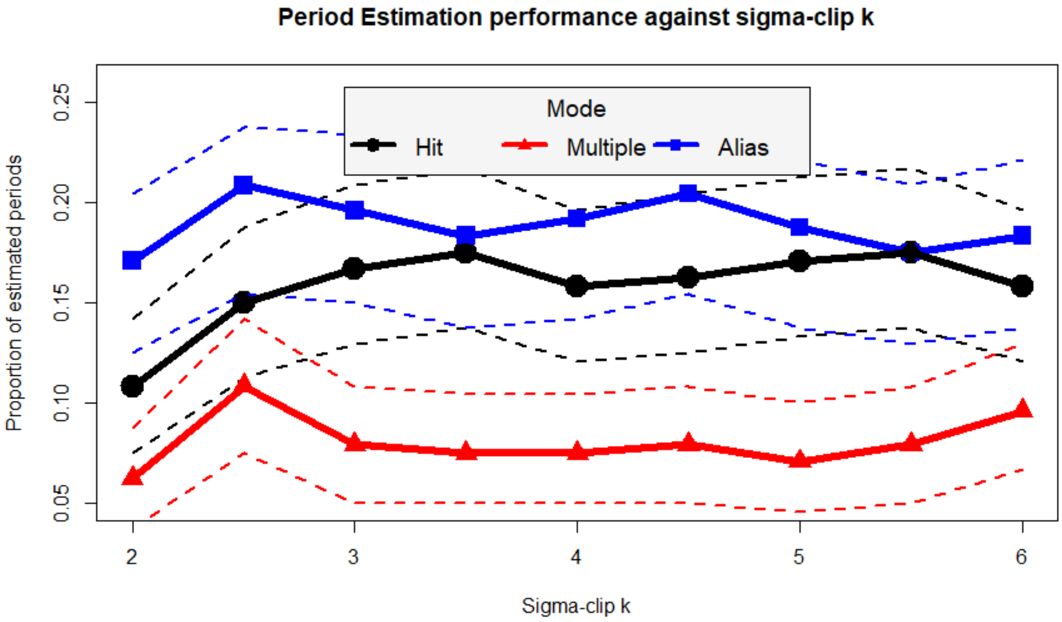


FIGURE 7.6: Performance of the σ–k clipping algorithm for 9 values of k with a period estimation tolerance $\epsilon = 0.05$ on the 240 subset light curves. The dashed lines indicate the 95% confidence limits for these results. The best performance appears to be near two locations $k = 3.5$ and $k = 5.5$ as the Hit rate is locally maximised and the alias rate is locally minimised without significant effect on the multiple rate. This suggests that there might be two different distributions of light curve noise, those which require the minimal clipping of $k = 5.5$ and those with outliers closer to the mean requiring $k = 3.5$. Attempting to determine which of these distributions a light curve is a member of would be advantageous.

The SkycamT instrument is strictly an achromatic instrument as its measurements are unfiltered optical white-light calibrated to an R-Band catalogue. Without multiband information the colours of objects located in the SkycamT images cannot be determined. This is problematic as colour is an important feature in the discrimination of some variable star classes. There is colour information available in the Skycam database due to the image reduction pipeline (Mawson et al., 2013). The sources detected in the Skycam images were cross-matched to the US Naval Observatory B (USNO-B) catalogue and those sources with a match have a selection of catalogue information recorded into the database. This selection includes a B-R colour, the difference between the magnitude of the object in the B band and the R band. Of the 590,492 SkycamT sources, 440,443 sources were successfully cross-matched to a source in the USNO-B catalogue. This is approximately three quarters of the database, 74.6% of the sources. The remaining sources do not have a corresponding catalogue entry and therefore lack colour data.

It is not optimal to remove the sources with missing B-R colour as they still describe large sections of the parameter space with their other light curve features and removing them could introduce a bias into the dataset. The lack of B-R colour data does not prevent the application of techniques such as the Naive Bayes classifier as it does not require every feature to have a real value for every source yet limiting the classification model to this method is certainly not ideal (Kotsiantis, 2007; Russell and Norvig, 2009). Richards et al. propose an alternative approach in the development of their All Sky Automated Survey (ASAS) classification pipeline named imputation (Richards et al., 2012). Imputation is the process of replacing missing data with substituted values produced by an alternative method (Barnard and Meng, 1999).

There are a number of methods for the imputation of the missing B-R colour feature through the regression modelling of the remaining light curve features. These include K-nearest neighbours where the imputed colour is the mean value of the closest k light curves with known B-R colour (Altman, 1992) and Least Absolute Shrinkage and Selection operator (LASSO) methods using regularised linear or non-linear regression methods (Tibshirani, 1996). Random Forests can also be used to produce forests of regression trees utilising both real and categorical valued features for the imputation task and through the *missForest* technique and have been shown to outperform the above methods (Stekhoven and Bühlmann, 2012). This is an iterative method which first makes an initial guess for the missing B-R colours such as the mean of the known B-R colours. Each feature vector has its B-R colour imputed using a Random Forest regression model trained on the matched source feature vectors. The iterations continue until the prediction error of the Random Forest on the feature vectors with known B-R colour is minimised.

We utilise the missForest method to impute the colours of the 25% of the SkycamT dataset lacking B-R colour by training on the 75% of the sources which have the cross-matched USNO-B B-R colour. Our feature vectors input into this method are the 35 variability indices computed on all light curves with 100 or more observations. This is performed using Random Forests with 100 trees and the method is allowed up to 10 iterations to minimise the prediction error (in practice we find it only takes 3 to 6 iterations to meet the stopping criterion). Unfortunately, we quickly found that the method has a quadratic computational complexity and the runtime and memory requirements on imputing the entire set of 590,492 light curves is prohibitive.

We need to split the dataset into smaller components without introducing bias or loosing useful imputation features. Our solution is to partition the dataset according to the position in the sources in the night sky, in coordinates of right ascension and declination. We justify this choice by considering that there is a good coverage across the sky of stars with the full range of possible B-R colours. For more unique star types which, for example, may be confined to the galactic disk, these objects will be primarily located in a small number of the sky partitions and therefore can influence the imputation in

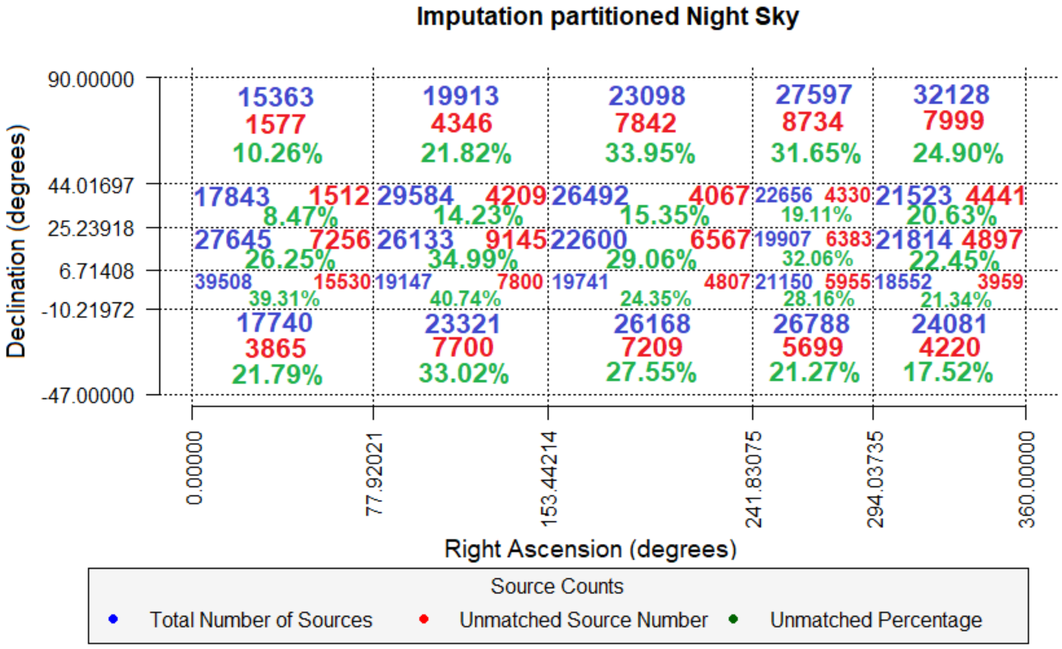


Figure 7.7: The 25 partitions used by the Colour Imputation missForest procedure on the La Palma sky. The regions are smaller near the galactic plane (between 241° and 294° right ascension) and around the best viewed areas of the sky from La Palma (between −10° and 44° declination) due to the abundance of well-sampled sources.

the other stars in those locations which may share similar properties. We found that partitioning the sky into 25 regions was sufficient to allow the imputation to complete in a satisfactory amount of time whilst still providing the Random Forest method with 10^4 to 10^5 training light curves per region. Unlike the partitioning of the sky in the trend removal method, the cuts are performed in percentiles so that right ascension and declination band have similar number of sources. Figure 7.7 shows this sky partition and the number of matched and unmatched objects in each region. There is a clear abundance of sources near the galactic plane which is to be expected. There is also many sources around −10° to 44° declination due to these objects being well positioned high in the La Palma sky at certain times of the year and therefore being well sampled. There is some colour variability for pulsating variable stars but as most of our dataset is non-variable this colour variability should not be a major concern during the imputation.

The main problem with the colour imputation procedure is that the machine learning algorithms depending on this data as a source of information are unaware of the origin of the data, be it a true catalogue match colour or the result of an imputation which is more likely to be erroneous. By applying imputation the dataset is not gaining any new information as it is being manufactured out of the already present information. As a result this procedure cannot be recommended for use in any finalised pipeline. A better alternative is to invest effort into improving the cross-matching with other astronomical catalogues which have similar detectability to the SkycamT instrument. With a high quality cross-match, any remaining unmatched sources are likely to be spurious detections (or transients) and can be processed accordingly. Therefore, the rest of this thesis makes use of these imputed colours in order to test the prototype pipeline but future efforts should be focused on replacing the colour imputation with true cross-matched catalogue colours.

7.4 Variable Classification Model

The application of the above methods using the variability indices discussed in chapter 5 have generated a dataset of 35 light curve features for the 590,492 selected SkycamT sources. Whilst these features are descriptive they are insufficient for the discrimination of specific classes of variable star. With the implementation of the feature extraction processes discussed in the previous chapters using a period estimated by the GRAPE method and fine-tuned by the multi-harmonic Variance Ratio Periodogram, it is possible to generate additional features capable of a full description of the light curves. Unfortunately, the computational effort required to generate the full set of features is prohibitive and of limited usefulness as there is no point in generating variable star light curve features from non-variable sources. Variable light curves must be isolated from the full SkycamT dataset for additional variable star classification.

Logically, it would make sense to produce a variability detection model prior to the variable star classification model. This is not an option as the training of a variable detection model requires a high quality dataset of variable light curves which can only be determined through use of a variable star classification model. The most confident classifications from this model for a given set of variable star types can then be used to train a variability detection model. The pipeline must make use of the initial data available to it which includes the 590,492 light curves from the SkycamT dataset and the list of variables in the AAVSO Variable Star Index catalogue. Using the variable star catalogue right ascension and declination positions, the 590,492 SkycamT sources are cross-matched with a tolerance radius of 148" which is sufficient to determine the cross-match given the SkycamT pixel scale. The cross-matches are also limited to a set of 12 classes which are detectable given the large white-noise component of the SkycamT data (Mawson et al., 2013). For this class-based subset of AAVSO sources with multiple Skycam objects within the 148" tolerance radius, the closest object is selected as the matched Skycam source. This method is unreliable as there is no guarantee that the closest source to the AAVSO catalogue position is the correct match therefore a further quality control is required. Our quality control method is to select objects where the

Table 7.3: The class distribution of the 6721 SkycamT light curves cross-matched with the AAVSO Variable Star Index.

Number	Class	Type	Count
1	β Lyrae	Eclipsing Binary	412
2	β Persei	Eclipsing Binary	1518
3	Chemically Peculiar	Rotational Variable	477
4	Classical Cepheid	Pulsating Variable	195
5	δ Scuti	Pulsating Variable	453
6	Mira	Pulsating Variable	1253
7	RR Lyrae Fundamental Mode	Pulsating Variable	99
8	RR Lyrae Overtone Mode	Pulsating Variable	60
9	RV Tauri	Pulsating Variable	36
10	Semiregular Variable	Pulsating Variable	988
11	Spotted Variable	Rotational Variable	528
12	W Ursae Majoris	Eclipsing Binary	702

period estimated by the GRAPE method matches, or is a multiple/submultiple of, the AAVSO catalogue period. We also construct a second dataset with the cross-matched SkycamT sources with GRAPE periods which are aliases of the catalogue period. This cross-match identifies 8041 SkycamT sources of which 6721 light curves have sufficient quality to successfully generate the full classification 109 features. This dataset has a class distribution as shown in table 7.3. Richards et al. recommend the use of a machine learning technique to test the statistics of each of the nearby matched sources relative to the catalogue data to determine which, if any, of the nearby sources are a good candidate which may prove a better method (Richards et al., 2012).

The period estimation performance of this dataset of 6721 light curve feature vectors is determined relative to the cross-matched AAVSO catalogue source with a tolerance of 5%, $\epsilon = 0.05$, as described in chapter 2. The result of this period estimation task is shown in table 7.4. This produces a data set of 859 SkycamT light curves expected to be exemplars of the 12 classes of variable star for use in training the variable star classification model as shown in table 2.1. A set of light curves with aliased periods is also produced containing 1186 light curves which still contain some useful information on the light curves as the aliased light curve produces the characteristic light curve shape of the associated class. The production of these two datasets of light curves, which are likely to be correct matches due to the strong relationship between the GRAPE estimated period and the catalogue period, allow the training of machine learning classifiers to identify the 12 classes of variable star from their 109 light curve features. Figure 7.8 demonstrates the base-10 Log-Log plot of the GRAPE period against the AAVSO catalogue period. The green line highlights the hit rate margin and the red lines the $n = 2$ and $n = 3$ multiple and submultiple margins. There are four blue lines which denote the four main failure periods which are independent of the catalogue period. These are the half

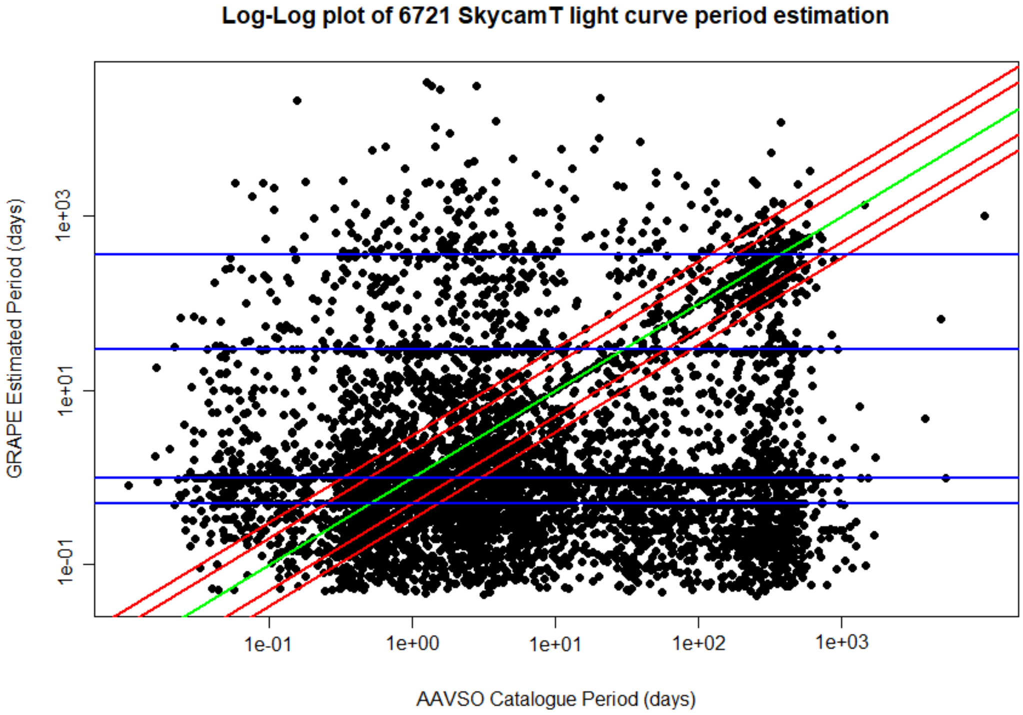


FIGURE 7.8: Base-10 Log-Log scatterplot of the GRAPE estimated periods against the AAVSO catalogue periods for the 6721 cross-matched variable stars. The green line shows the hit rate margin, the red lines the $n = 2$ and $n = 3$ multiple and submultiple margins and the blue lines the half sidereal day, sidereal day, lunar month (29.5 days) and year spurious periods. Only 7.8% of the data is correctly estimated with multiple objects producing spurious period results or an alias between 0.1 days and 10 days due to systematics from the Skycam cadence.

TABLE 7.4: Relation between the GRAPE estimated period and the AAVSO Variable Star Index period for the 6721 cross-matched SkycamT light curves.

Period Mode	Matching Percentage
Hit	7.826%
Multiple	0.818%
Submultiple	4.136%
One-day alias	8.526%
Half-day alias	10.073%
Unknown	71.448%

sidereal day, the sidereal day, the lunar month (29.5 days) and the 365 day year. The trend removal method has heavily depleted the yearly and half-year spurious periods but the lunar month still remains potent. There is also a large quantity of poor estimations around the sub-day to few-day region due to unexplained systematics in the Skycam data producing aliased signals in these locations.

TABLE 7.5: AUC Performance of the Machine Learning algorithms trained on the 859 matched SkycamT light curves.

Classifier	Hyper-parameters	Mean AUC	StD AUC
RF	$m_{\text{try}} = 40$, ntree = 500, nodesize = 75	**0.760**	0.021
SVM	linear kernel, C = 2^4	0.710	0.043
SVM	radial kernel, C = 1, $\gamma = 2^{-4}$	0.731	0.040
SVM	polynomial kernel, C = 2^3, $\gamma = 2^{-6}$, d = 3	0.717	0.057
FFNN	layers: 109-200-12, $a(x)$ = sigmoid	0.651	0.037
FFNN	layers: 109-200-12, $a(x)$ = tanh	0.582	0.038
FFNN	layers: 109-200-12, $a(x)$ = ReLU	0.508	0.022
FFNN	layers: 109-500-12, $a(x)$ = sigmoid	0.646	0.032
FFNN	layers: 109-500-12, $a(x)$ = tanh	0.605	0.048
FFNN	layers: 109-500-12, $a(x)$ = ReLU	0.512	0.031
FFNN	layers: 109-1000-12, $a(x)$ = sigmoid	0.641	0.033
FFNN	layers: 109-1000-12, $a(x)$ = tanh	0.561	0.048
FFNN	layers: 109-1000-12, $a(x)$ = ReLU	0.503	0.009
NB	None	0.706	0.048

7.4.1 Classifier Selection

In chapter 4 we described a number of machine learning classification algorithms. These algorithms each have strengths and weaknesses for the variable star classification task and must be investigated for this problem using the training dataset of 859 variable light curves of 12 variable star classes. The performance of the classifiers can be limited due to the correlation between the features so these correlations are determined using Spearman's correlation coefficient. These correlations are plotted into three correlation plots shown in figures 7.9, 7.10 and 7.11. Many of the variability indices are highly correlated as they contain multiple statistics designed to evaluate similar properties of the light curves in different ways. The amplitude and phase features from the Fourier decomposition also appear highly correlated which explains why they performed poorly in the Random Forest model documented in chapter 5 as the harmonic components are not describing the important non-sinusoidal features in the data. The interpolated PolyFit features are reasonably independent which explains why the model trained using them in chapter 6 performed well as the features describe independent properties of the light curves. This is not a surprising result as the principal components used to produce many of these features are, by definition, independent due to the PCA orthogonal rotation of the feature space. It is also interesting to see that the P.PCA2 feature which proved important in the discrimination of short period pulsating variables and eclipsing binaries is strongly negatively correlated with the interpolated skewness feature. The skewness is expected to be useful in this task yet it had thus far seemed lacklustre in this task. The P.PCA2 feature appears to be acting as a 'robust skewness' measure fulfilling the role originally designed for the skewness feature.

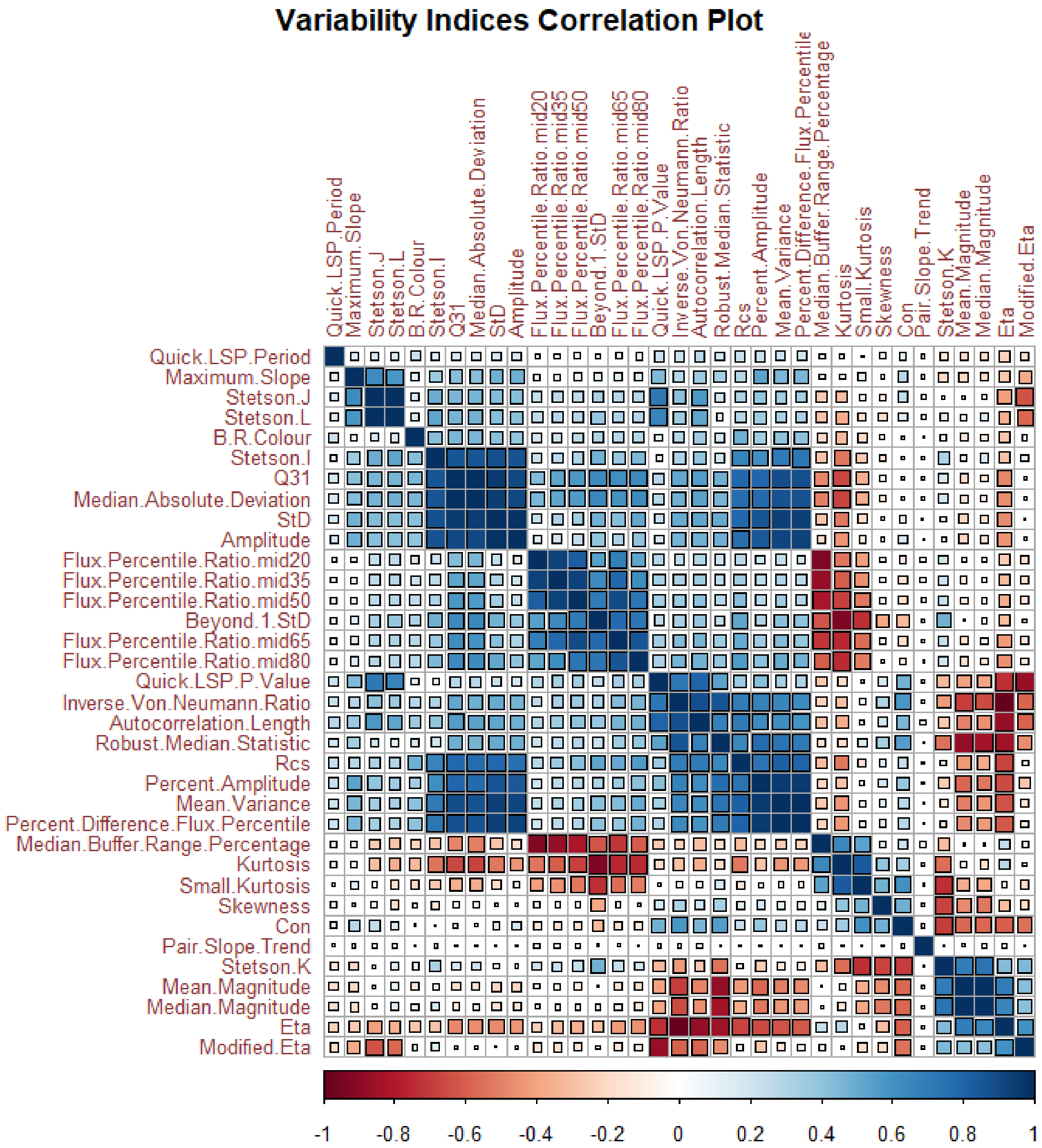


FIGURE 7.9: Correlation Plot of the variability indices computed from the 859 variable light curves. Many of the variability indices are highly correlated with eachother as there are multiple features which measure a similar aspect of the light curve in differing ways. An interesting interaction is that of the Quick LSP P value feature which does strongly correlate with the variability indices and the Stetson indices indicating these statistics are functioning on the variable cadence light curves ubiquitous in the Skycam database.

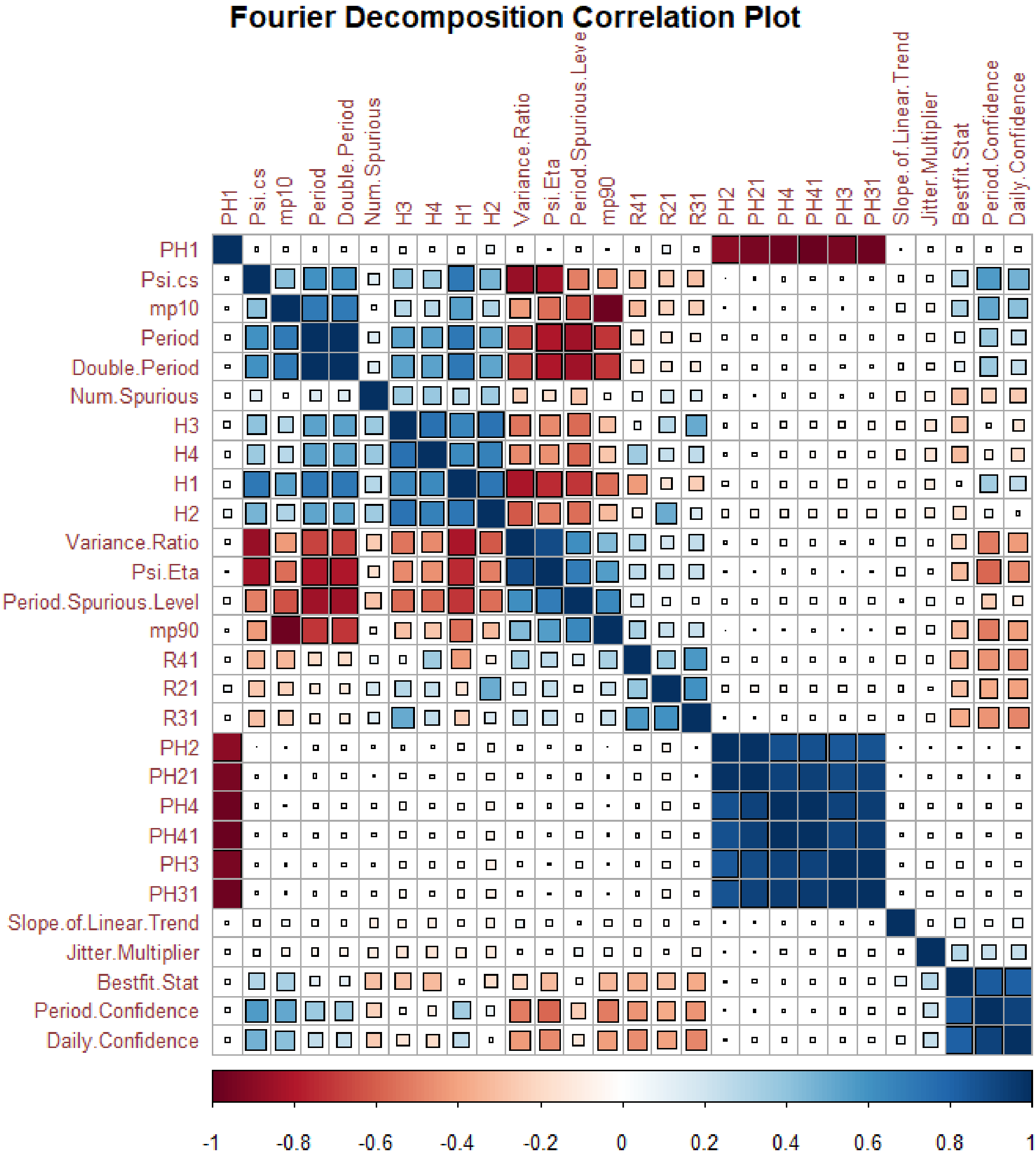


Figure 7.10: Correlation Plot of the Fourier Decomposition features computed from the 859 variable light curves. These features are less correlated than the variability indices although the harmonic amplitude and phase features are not independent which explains why they are not utilised in previous classification methods on the Skycam data. The features indicating the performance of the GRAPE method are also highly correlated as they are all quantifying the confidence of the estimated period.

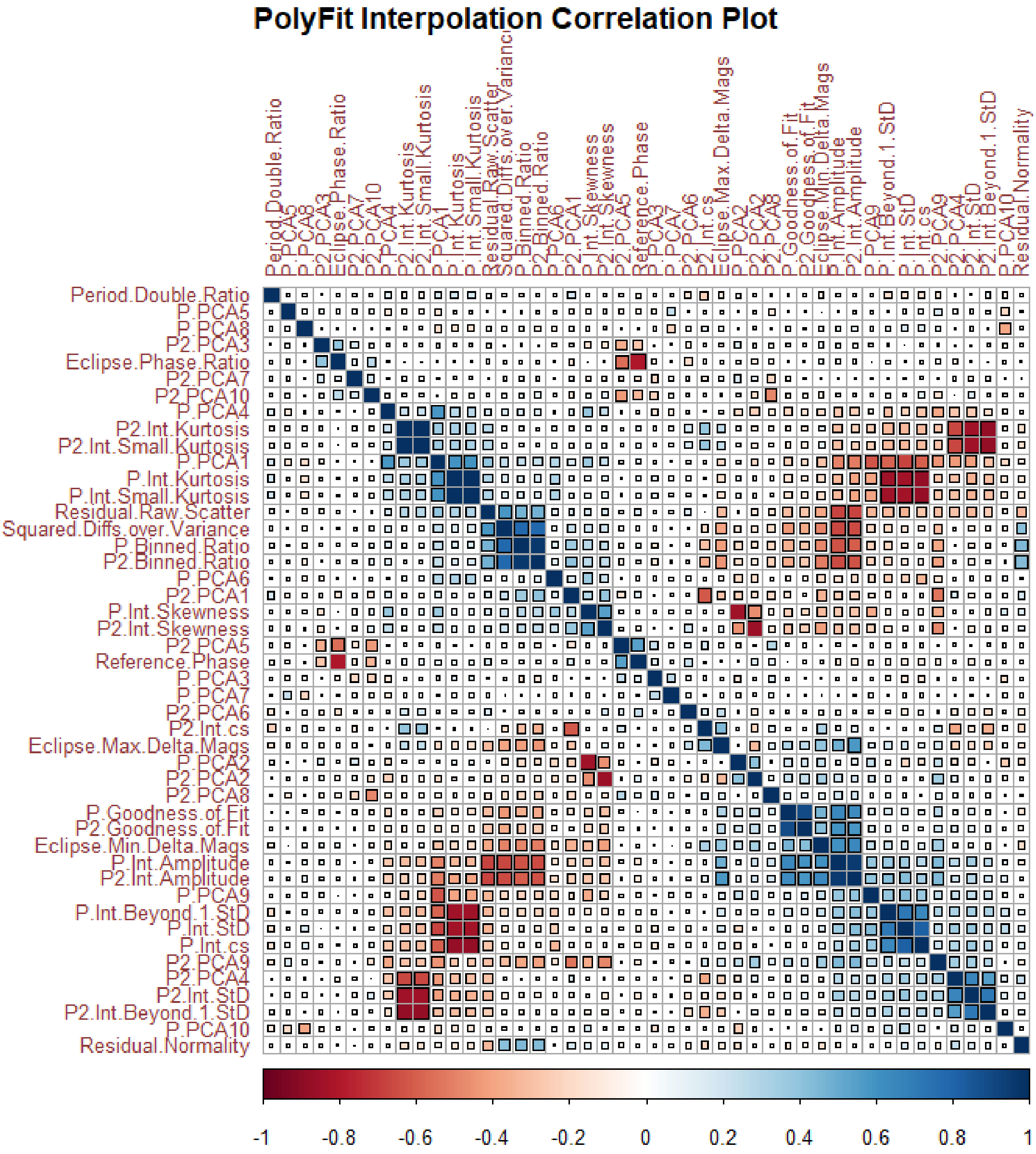


Figure 7.11: Correlation Plot of the interpolated and PCA PolyFit features computed from the 859 variable light curves. The correlation in this data is very limited which is a good indicator that they will perform well for classification. The low correlation is expected as many of the features are computed from the principal components of the PolyFit model which are all independent by definition.

TABLE 7.6: F1 Score Performance of the Machine Learning algorithms trained on the 859 matched SkycamT light curves.

Classifier	Hyper-parameters	Mean F1	StD F1
RF	m_{try} = 40, ntree = 500, nodesize = 75	**0.533**	0.073
SVM	linear kernel, C = 2^4	0.346	0.033
SVM	radial kernel, C = 1, $\gamma = 2^{-4}$	0.349	0.037
SVM	polynomial kernel, C = 2^3, $\gamma = 2^{-6}$, d = 3	0.324	0.022
FFNN	layers: 109-200-12, $a(x)$ = sigmoid	0.079	0.054
FFNN	layers: 109-200-12, $a(x)$ = tanh	0.062	0.036
FFNN	layers: 109-200-12, $a(x)$ = ReLU	0.017	0.011
FFNN	layers: 109-500-12, $a(x)$ = sigmoid	0.081	0.047
FFNN	layers: 109-500-12, $a(x)$ = tanh	0.060	0.040
FFNN	layers: 109-500-12, $a(x)$ = ReLU	0.017	0.008
FFNN	layers: 109-1000-12, $a(x)$ = sigmoid	0.081	0.049
FFNN	layers: 109-1000-12, $a(x)$ = tanh	0.047	0.021
FFNN	layers: 109-1000-12, $a(x)$ = ReLU	0.013	0.009
NB	None	0.278	0.032

These correlations indicate that the Naive Bayes classifier may perform poorly due to features not being independent (Kotsiantis, 2007; Russell and Norvig, 2009). The Random Forest method may benefit as it is capable of processing the correlated features (Breiman et al., 1984; Breiman, 2001). To determine the best classifier we perform a 5-fold stratified cross validation training with 2 repeats using the 109 features in the 859 light curve dataset. The training objects for each cross validation split are oversampled by randomly adding copies of the minority classes until every class equals the size of the largest class. This prevents the classifier from assigning too much weighting to the most common variable star types. We test the Random Forest (RF), Feedforward Neural Network (FFNN) with 1 hidden layer, the Support Vector Machine (SVM) and the Naive Bayes (NB) algorithms. As a number of these algorithms assume a normal distribution we run each classifier with the original features and with the features rescaled to a mean of 0 and a standard deviation of 1. The performance of the resulting models is determined from the mean AUC and F1 score of the ten cross validation models with the variation in the model performance determined by the standard deviation in the performance of the ten models.

The Random Forest hyperparameters were determined using a grid search for the three main arguments, $m_{\mathrm{try}} = [5, 10, 15, 20, 25, 30, 35, 40, 45, 50]$, ntrees $= [500, 1000, 1500]$ and nodesize $= [25, 50, 75]$. The cross validation is performed on each combination and the best performing hyperparameters using the F1 score are selected and used to train the final classification model. The Random Forest hyperparameter tuning results are shown in figure 7.12 with the best performing cross validation using the F1 score as a discriminator is ntree = 500, m_{try} = 40 and nodesize = 75 with a mean F1 score

Table 7.7: AUC Performance of the Machine Learning algorithms trained on the 859 matched SkycamT light curves with rescaled features.

Classifier	Hyper-parameters	Mean AUC	StD AUC
RF	m_{try} = 40, ntree = 500, nodesize = 75	0.765	0.021
SVM	linear kernel, C = 2^4	0.710	0.043
SVM	radial kernel, C = 1, γ = 2^{-4}	0.731	0.040
SVM	polynomial kernel, C = 2^3, γ = 2^{-6}, d = 3	0.717	0.057
FFNN	layers: 109-200-12, $a(x)$ = sigmoid	0.742	0.028
FFNN	layers: 109-200-12, $a(x)$ = tanh	0.759	0.034
FFNN	layers: 109-200-12, $a(x)$ = ReLU	0.756	0.033
FFNN	layers: 109-500-12, $a(x)$ = sigmoid	0.749	0.016
FFNN	layers: 109-500-12, $a(x)$ = tanh	0.750	0.039
FFNN	layers: 109-500-12, $a(x)$ = ReLU	0.747	0.036
FFNN	layers: 109-1000-12, $a(x)$ = sigmoid	0.748	0.025
FFNN	layers: 109-1000-12, $a(x)$ = tanh	0.740	0.036
FFNN	layers: 109-1000-12, $a(x)$ = ReLU	**0.773**	0.032
NB	None	0.706	0.052

Table 7.8: F1 Score Performance of the Machine Learning algorithms trained on the 859 matched SkycamT light curves with rescaled features.

Classifier	Hyper-parameters	Mean F1	StD F1
RF	m_{try} = 40, ntree = 500, nodesize = 75	**0.538**	0.070
SVM	linear kernel, C = 2^4	0.346	0.033
SVM	radial kernel, C = 1, γ = 2^{-4}	0.349	0.037
SVM	polynomial kernel, C = 2^3, γ = 2^{-6}, d = 3	0.324	0.022
FFNN	layers: 109-200-12, $a(x)$ = sigmoid	0.254	0.092
FFNN	layers: 109-200-12, $a(x)$ = tanh	0.258	0.070
FFNN	layers: 109-200-12, $a(x)$ = ReLU	0.249	0.102
FFNN	layers: 109-500-12, $a(x)$ = sigmoid	0.252	0.094
FFNN	layers: 109-500-12, $a(x)$ = tanh	0.238	0.104
FFNN	layers: 109-500-12, $a(x)$ = ReLU	0.210	0.101
FFNN	layers: 109-1000-12, $a(x)$ = sigmoid	0.268	0.083
FFNN	layers: 109-1000-12, $a(x)$ = tanh	0.228	0.096
FFNN	layers: 109-1000-12, $a(x)$ = ReLU	0.263	0.105
NB	None	0.287	0.029

of 0.533 and a mean AUC of 0.760. The Support Vector Machine utilises a tuning function implemented into the algorithm which performs a grid search across the three main hyperparameters prior to the cross validation, cost $C = 2^{[-2,-1,0,1,2,3,4,5]}$, the radial basis function and polynomial kernel $\gamma = 2^{[-8,-7,-6,-5,-4,-3,-2,-1,0]}$ and the polynomial kernel degree $d = [1, 2, 3]$. Following the hyperparameter tuning the cross validation is computed. Three kernel functions are applied individually, the linear kernel, the radial basis function kernel and the polynomial kernel. Each kernel has its hyperparameters trained individually.

The Feedforward Neural Network is implemented using the Keras tensorflow library with a GPU implementation for parallelisation. A single hidden layer was used only (a second hidden layer was tried but found that the model excessively overfit the training data) of 200, 500 and 1000 neurons in size with a final softmax (multinomial logistic regression) classification layer. The complexity of the model was controlled through the addition of a dropout layer which nulls the weights of 20% of the neurons per training iteration. This prevents complex co-adaption from allowing neurons to overfit aspects of the dataset. We also tested three popular activation functions for the neural network non-linearity, the sigmoid function, the tanh function and the Rectified Linear Unit (ReLU). The Feedforward Neural Network results are not exactly reproducible due to the lack of seed number control in the parallelised GPU implementation used in the Keras package. This is a known limitation of the package and it is being addressed although the only way to accomplish reproducibility at the moment is to disable parallelisation which would make our neural network model training computationally prohibitive. The Naive Bayes has no hyperparameters and thus was computed using a single 5-fold cross validation using the 859 light curves.

The best performing classification algorithm on the original features is the Random Forest with the Support Vector Machine using the radial basis function kernel a close second. The Naive Bayes is weaker than these two algorithms as expected and the Neural Network models perform terribly with the sigmoid and tanh networks achieving small performance and the ReLU network achieving no better than random results. This is likely a result of the neural network classifier assumptions which do not apply to the initial features in the dataset. Applying the rescaling operation to the features in the training data results in a substantial improvement in the neural networks with a small improvement in the Random Forest and Naive Bayes classifiers. The Support Vector Machine performance was unchanged by the feature rescaling resulting in the neural networks outperforming them in the AUC metric although the SVMs still had a higher F1 score. With this dataset the feedforward neural network with 100 neurons using the ReLU activation function outperformed the Random Forest classifier in the AUC metric yet was still substantially inferior in the more desirable F1 score. The standard deviations of the models mean that it is difficult to conclusively state which model is performing the best overall as they all exhibit similar performance (except for the neural networks on the unscaled data). The Random Forest F1 score on both the unscaled and scaled data is confidently outperforming the other models and thus we have selected this classifier for the production of our final classification model. These results are similar to those found by Richards et al. as on their data, which was unscaled, the neural networks were outperformed by the other algorithms and the Random Forest classifier was the top performing algorithm (Richards et al., 2011b). Figure 7.13 shows the ROC curves produced using this Random Forest model on a one-vs-many classification of the 12

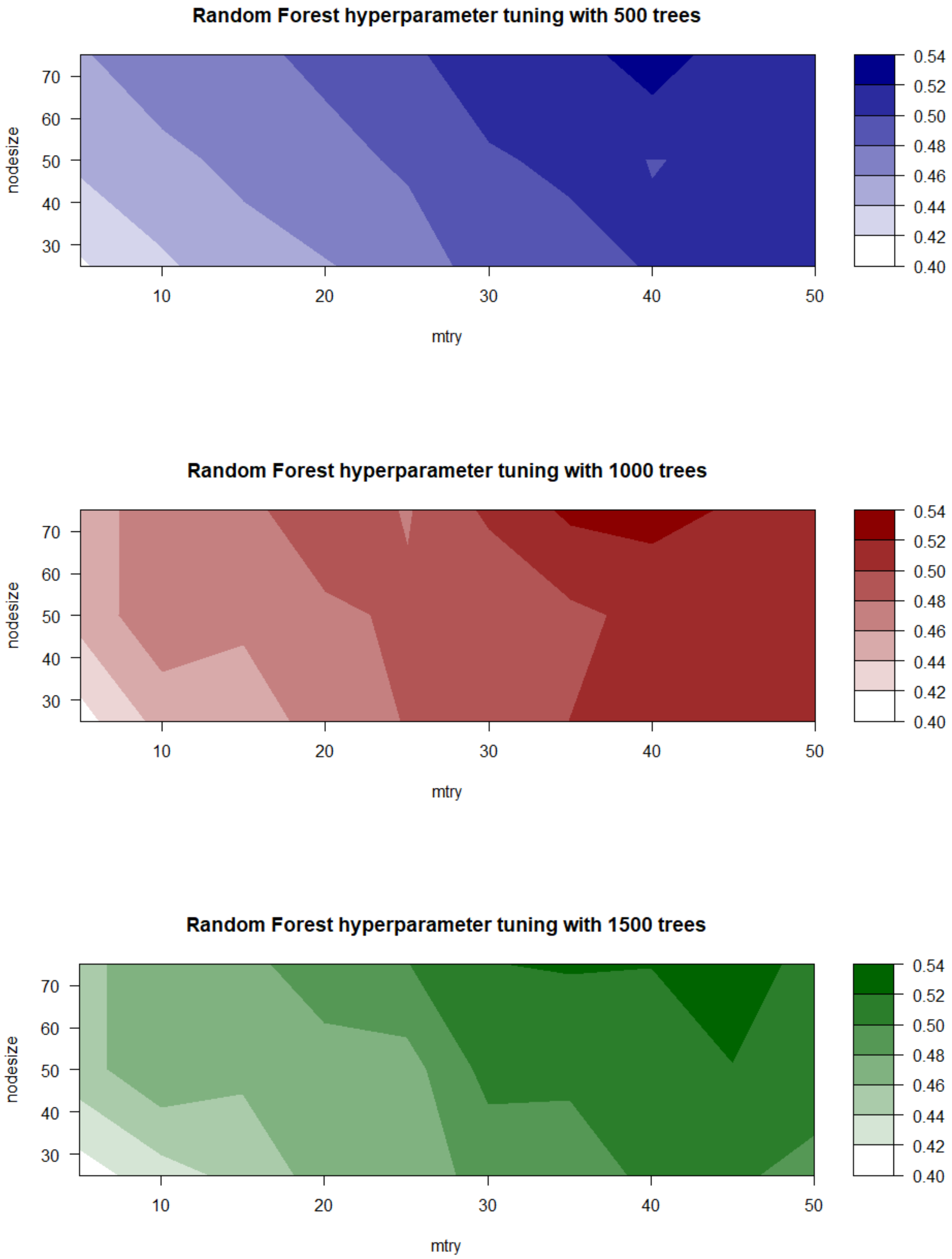


FIGURE 7.12: Contour plots of the hyperparameter tuning of the Random Forest 5-fold cross validation with 500, 1000 and 1500 trees. The model performance is primarily dependent on m_{try} and nodesize as the 500 trees are sufficient to model the required parameters of the 12 class variable star classification on the 859 SkycamT light curves. The best performing cross validation using the F1 score as a discriminator is ntree = 500, m_{try} = 40 and nodesize = 75 with a mean F1 score of 0.533 and a mean AUC of 0.760.

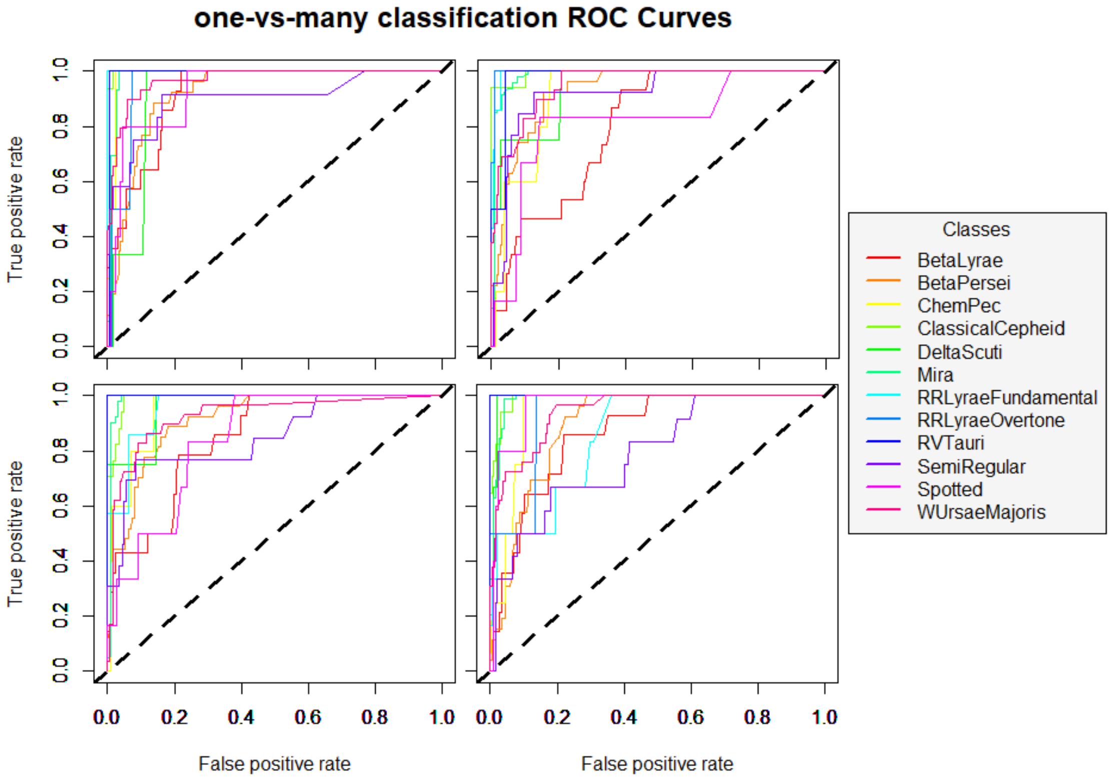


FIGURE 7.13: ROC curves of a Random Forest 4-fold cross validation with ntree = 500, m_{try} = 40 and nodesize = 75. The performance on many of the classes is strong with the weakest classes being the β Lyrae eclipsing binaries, the spotted variables and the semiregular variables. The poorer β Lyrae performance is a result of subclass confusion with the other eclipsing binaries. The spotted variables have low amplitude signals which are difficult to detect in the SkycamT light curves and the semiregular variables are poorly defined by a strict period which limits the performance of useful feature sets such as the PolyFit model features.

training classes. Many of the classes have high confidence with the poorest performing classes suffering classification difficulty due to subclass misclassification for the eclipsing binaries (Malkov et al., 2007; Hoffman et al., 2008), low amplitude variability in the spotted variables and poor epoch-folded features from the semiregular long period variables (Richards et al., 2011b).

7.4.2 Aliased Object Classifier

The set of 1186 aliased light curves are also of potential interest in the classification of candidate variables. In the event of the GRAPE estimated period of these variables being aliased by the sidereal day spurious period, a model trained on the aliased light curves could produce a more confident classification of the true class of the object than relying on the previous hit and multiple variable star classification model. We selected the Random Forest classifier to train this model and performed a hyperparameter tuning

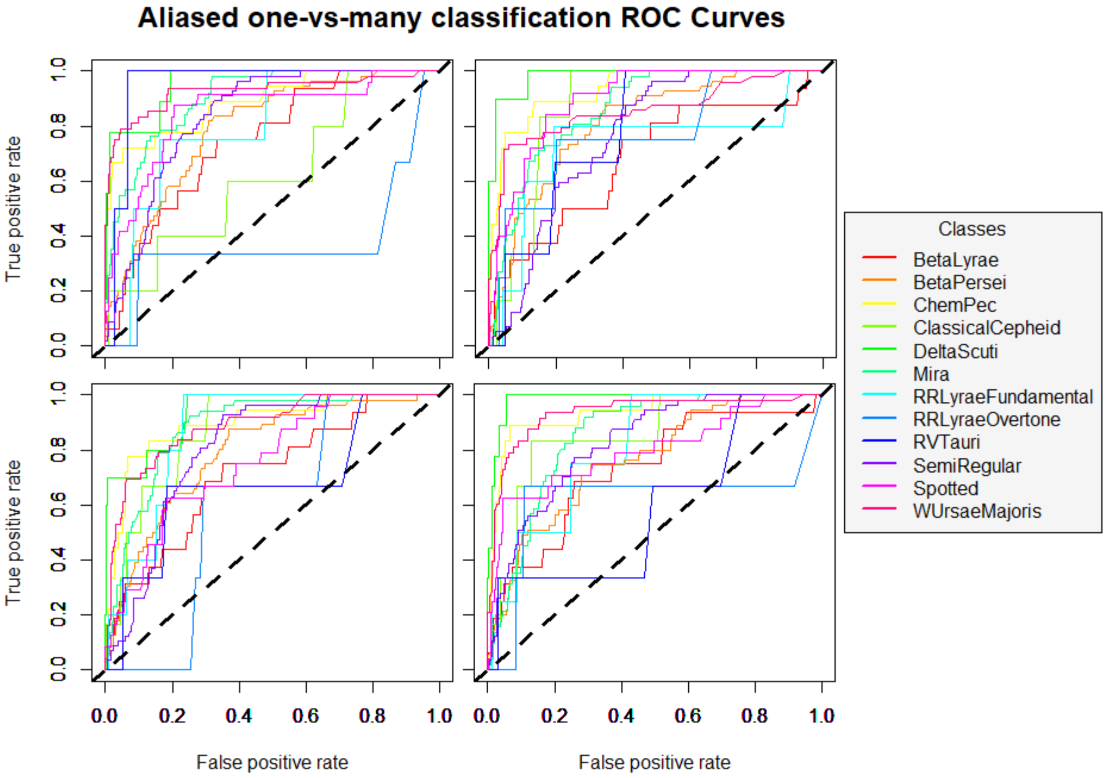
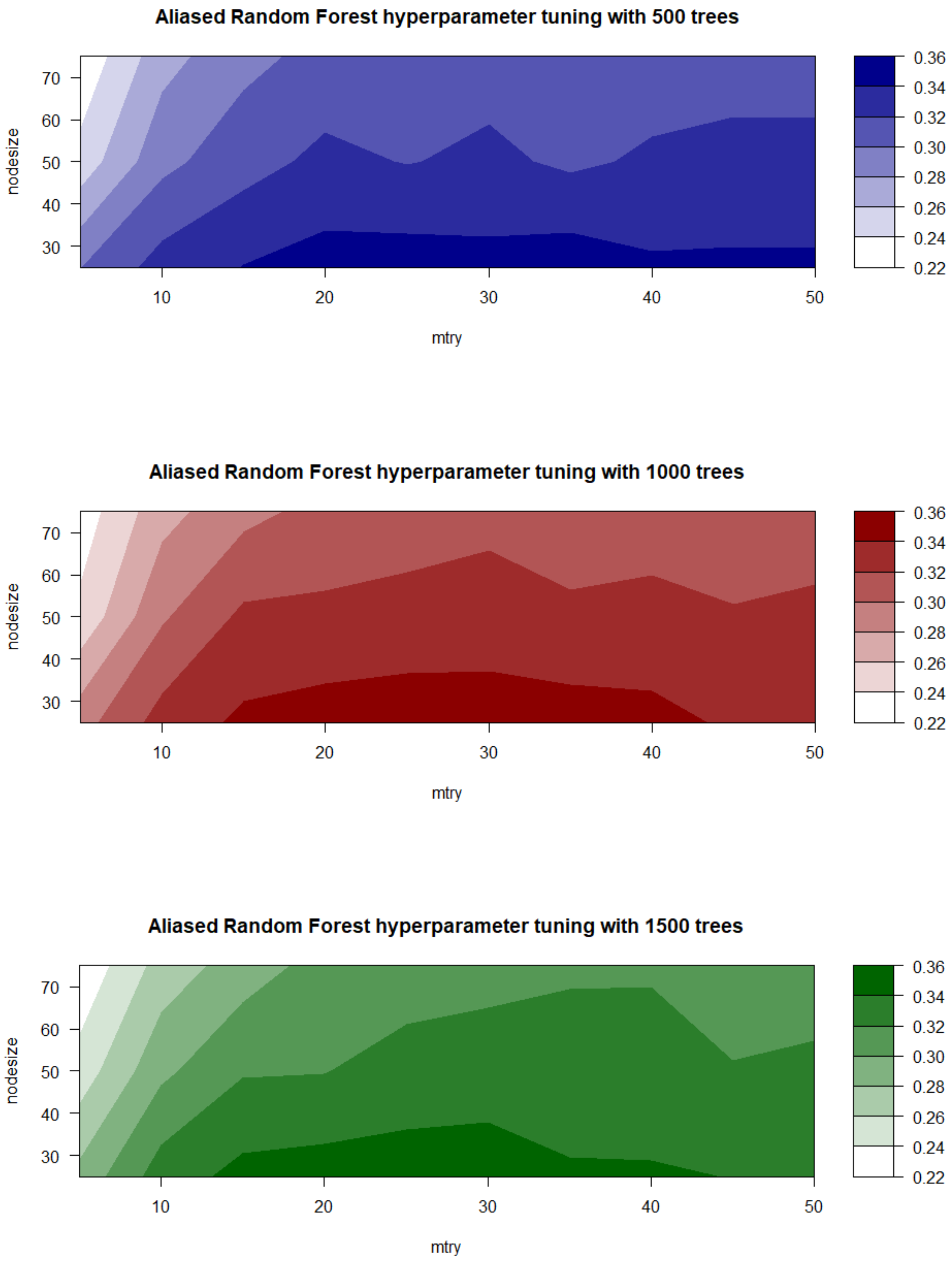


FIGURE 7.14: ROC curves of a Random Forest 4-fold cross validation on the aliased dataset with ntree = 1500, m_{try} = 30 and nodesize = 25. The performance on the individual classes is substantially weaker than the hit and multiple light curve dataset. This is due to the period feature, the most important for variable star classification, being a much poorer discriminator due to the aliased result. The classifier must therefore make use of the remaining features to try and develop a satisfactory model. The poorer ROC of the classes shows that the confidence of the classifier will be low. This means that the model can be used on potentially aliased candidate light curves as confident classifications on the aliased model over the 859 light curve model are a strong indicator that the candidate light curve is aliased.

using 5-fold cross validation on the 1186 aliased light curves to determine the optimal arguments for the final classification model. Figure 7.15 shows the contour plots of the hyperparameter tuning and the aliased data performs best when there are more, deeper trees with additional features and branches. This allows the classifier to probe the feature thoroughly for information to replace the poorly estimated period feature. The resulting best performing cross validation using the F1 score as a discriminator is ntree = 1500, m_{try} = 30 and nodesize = 25 with a mean F1 score of 0.353 and a mean AUC of 0.688. This model is clearly outperformed by the model trained on the 859 light curves with superior GRAPE estimated periods but it is better suited to the classification of other aliased light curves. Figure 7.14 shows the ROC curves produced using this Random Forest model on a one-vs-many classification of the 12 training classes. The confidence in the classifications for all the classes is lower than the 859 light curve model due to the loss of period as an important feature. Therefore, a confident classification using



Aliased Random Forest hyperparameter tuning with 500 trees

Aliased Random Forest hyperparameter tuning with 1000 trees

Aliased Random Forest hyperparameter tuning with 1500 trees

Figure 7.15: Contour plots of the hyperparameter tuning of the aliased dataset on a Random Forest 5-fold cross validation with 500, 1000 and 1500 trees. The best performing cross validation using the F1 score as a discriminator is ntree = 1500, m_{try} = 30 and nodesize = 25 with a mean F1 score of 0.353 and a mean AUC of 0.688. The aliased data is unable to depend on period therefore requires more trees and a deeper forest to produce the best possible model from the other features.

this model is a strong indicator that the candidate light curve has an aliased GRAPE estimated period.

7.4.3 Feature Importance

The introduction of the Random Forest model in chapter 4 described the Gini criterion, a measure of the diversity of the leaf nodes in the forest. The Gini measure can be used to produce a feature importance metric through computing the mean change in the Gini criterion due to the removal of a feature from the forest. This feature importance metric is named the Mean Decrease Gini and is computed directly from the final Random Forest model. We compute this statistic for each feature in our two Random Forest variable star classification models to determine which features have the most influence on the decision boundaries for this classification.

Figure 7.16 demonstrates the mean decrease Gini of the top 30 features from the 859 period-matched light curve classification model. The models trained on previous survey light curves indicate that the period, amplitude, colour and skewness features are the most important in the classification of variable stars (Debosscher et al., 2007; Richards et al., 2011b, 2012). The model trained on the initial feature set in chapter 5 did not reflect this as many of these features were not performing adequately on the SkycamT light curves. The application of the PolyFit features appears to have restored this capability with the new Random Forest model reflecting the results of the previous research. The period and double period are both dominant in the Random Forest model as the period is the basis of the initial classification hierarchy of many of these variable star types. The strong correlation between the two features suggests that one of them should perhaps be removed but the Random Forest algorithm still performs well.

The B-R colour is the second most important feature after the period features which indicates it was worth the difficulty in implementing the colour imputation routine as many giant red stars are natively variable (Percy, 2008). The Period Spurious Level feature is designed to indicate if the GRAPE estimated period is close to a known spurious period and is assigned a high importance in this model. This is unlikely due to the feature containing important information for the classification task and it is probably a result of the high correlation between this feature and period as shown in figure 7.10. The slope of a linear trend is a stronger feature than expected and it is primarily driven by the RR Lyrae overtone class which is possibly due to a selection bias or an unknown effect. The next several features are highly important to this research as they indicate our novel PolyFit interpolation and PCA features are selected over their equivalent variability indices and Fourier decomposition features. The Interpolated Amplitude of

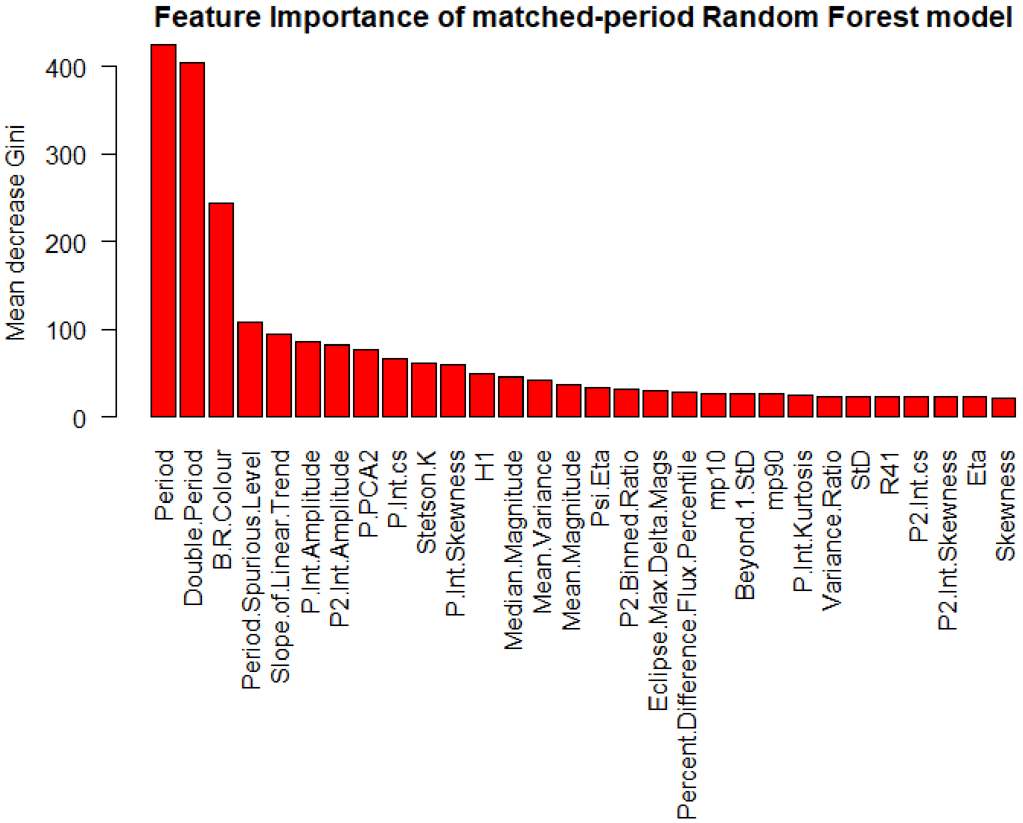


Figure 7.16: Bar plot of the mean decrease Gini of the top 30 features in the 859 period-matched light curve Random Forest model. The period feature is dominant in the classification and the PolyFit interpolated amplitude has been selected as the amplitude measure. The P.PCA2 feature has been selected as the most robust indicator of the light curve skewness. The B-R colour has also been identified as an important indicator of variable star type. Few of the variability indices are considered important for this task and the Fourier decomposition features have been almost completely replaced by their PolyFit interpolation equivalents.

the light curves calculated from PolyFit models folded at both the GRAPE estimated period and 2× this period contain the most robust estimation of the amplitude of the variable SkycamT light curves. The next highest mean decrease Gini feature is the P.PCA2 feature determined from the second principal component of the PolyFit model. This feature discriminated the short period eclipsing binaries from the short period pulsating variables in the experiment in chapter 6. The correlation plot in figure 7.11 also show that this feature is strongly negatively correlated with the interpolated skewness feature. Skewness performed this task in the Richards et al. classifiers but is not as clear in the SkycamT data (Richards et al., 2011b, 2012). The remaining features have similar importance after the interpolated range of a cumulative sum, Stetson K index and interpolated skewness features.

Figure 7.17 demonstrates the mean decrease Gini of the top 30 features from the 1186

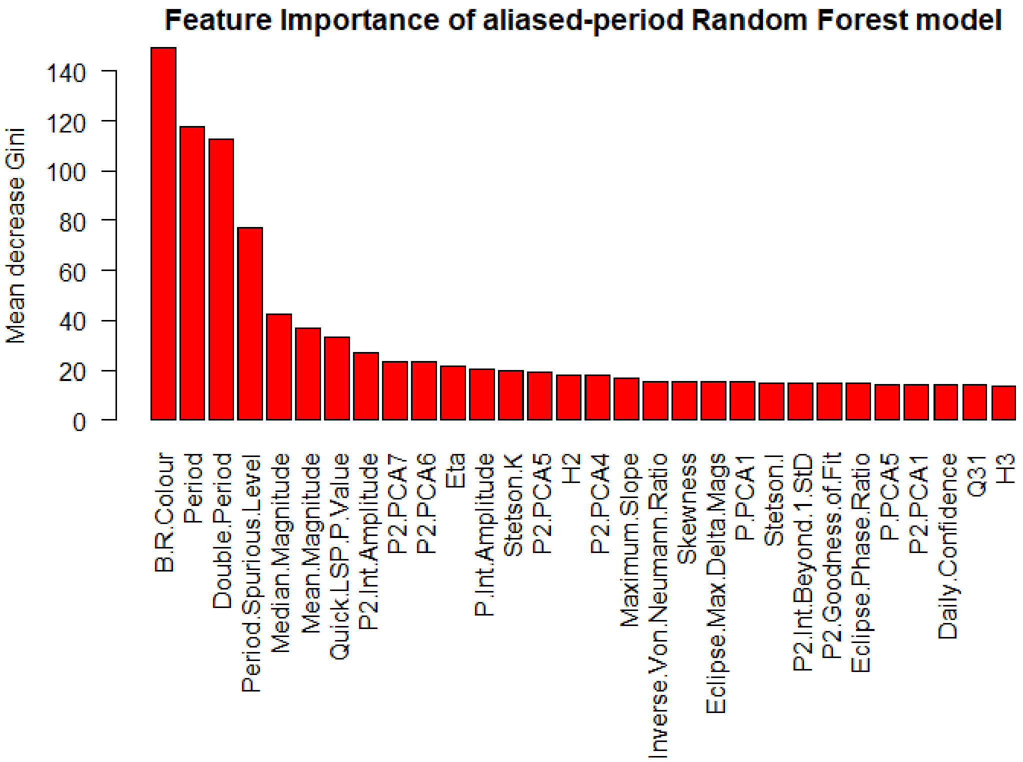


FIGURE 7.17: Bar plot of the mean decrease Gini of the top 30 features in the 1186 aliased period light curve Random Forest model. The period feature suffers a substantial reduction in feature importance due to many of the variable star classes having aliased periods near the sidereal day. The period still has some class discrimination ability therefore it is still one of the more important features, second only to the B-R colour. The Quick LSP P-Value becomes much more important in this model as the strength of the alias is a big indicator of the amplitude of the original signal, stronger than the poorer quality PolyFit features due to the poor period estimation.

aliased period light curve classification model. The period feature has dropped in importance to below the B-R colour. This is due to the B-R colour being independent of the computation of the period and the aliased periods for the 12 classes are substantially less discriminatory than the matched periods as many of them are close to the sidereal day. The Quick LSP P-Value becomes much more important in this model as the strength of the alias is a big indicator of the amplitude of the original signal, stronger than the poorer quality PolyFit features due to the poor period estimation. The light curve epoch-folded at an aliased period should still produce a light curve shape similar to the true period yet the differences produce an alternative set of PolyFit features with have higher feature importance than the matched-period model.

7.4.4 Probability Calibration

The output of the machine learned classification models are a probability vector of length 12, the probability for each of the 12 classes, and the sum of probabilities for a light curve across all classes is unity. The probabilities output by the Random Forest model are the confidence in the prediction of light curve x as a member of class C_i where $i = 1, \ldots, 12$. These probabilities do not necessarily match the actual posterior probability of a light curve x being a true member of each class as the Random Forest may be too conservative or confident in a given class probability. It would be useful to have the classification model probabilities *calibrated* to the expected posterior probabilities for the dataset. This would mean if all objects with a probability $p(x|C_i) = 0.8$ of being a member of class C_i are determined, 80% of those objects would truly be a member of that class.

We perform a probability calibration on the matching period and aliased period datasets through the use of the probability adjustment transformation defined by Boström as utilised by Richards et al. in their ASAS variable star classifier (Boström, 2008; Richards et al., 2012). This transformation is shown in equation 7.5.

$$\hat{p}_{ij} = \begin{cases} p_{ij} + r(1 - p_{ij}), & \text{if } p_{ij} = \max[p_{i1}, p_{i2}, \ldots, p_{iC}] \\ p_{ij}(1 - r), & \text{otherwise.} \end{cases} \tag{7.5}$$

where $[p_{i1}, p_{i2}, \ldots, p_{iC}]$ is the vector of class probabilities for the i^{th} light curve and $r \in \mathbb{R} \in [0, 1]$ is a scalar variable which determines the required probability calibration across the dataset for a given machine learning model. This transformation is ideal as it always sums the calibrated probability vector to unity regardless of the value of r as long as the input probability vector sums to unity (which it always should). In this form, r is very restrictive as it will apply the same rescaling to any probability vector regardless of the confidence of the classification model for any class (Richards et al., 2012). We follow the approach used by Richards et al. by parameterising r using a sigmoid function based on the classifier margin between the highest probability class and the second highest probability class $\Delta_i = p_{i,\max} - p_{i,2^{\text{nd}}}$ for each classification light curve. This model which is shown in equation 7.6, has two free parameters A and B which must be learned from the classifier training set (Boström, 2008; Richards et al., 2012).

$$r(\Delta_i) = \frac{1}{1 + \exp(A\Delta_i + B)} \tag{7.6}$$

The optimal values of A and B to generate the optimal value of r given Δ_i are tuned using the Brier Score (as shown in equation 4.27) of the dataset after the probability calibration. The parameters that minimise the Brier Score for a given training set of light curves are the optimal A and B for that classifier. In the SkycamT pipeline this calibration is performed on both the matched period model and the aliased period model using a genetic algorithm. A uniform population of n_{pop} randomly generated values of $A \in \mathbb{R} \in [-20, 20]$ and $B \in \mathbb{R} \in [-20, 20]$ populate the two dimension feature space. The input arguments were as follows: $N_{\text{pop}} = 100$, $N_{\text{pairups}} = 20$, $N_{\text{gen}} = 50$, $P_{\text{crossover}} = 0.65$, $P_{\text{mutation}} = 0.03$, $P_{\text{fdif}} = 0.6$ and $P_{\text{dfrac}} = 0.7$. For further information on these genetic algorithm arguments we recommend reading the chapter 3 Evolutionary Implementation section.

Utilising this genetic algorithm optimisation for the probability calibration on the period-match Random Forest model produces optimal values of $A = -2.100$ and $B = 2.285$ which indicates that this model is too conservative in its classifications as it underestimates the true class probabilities. The aliased-period Random Forest model produces optimal values of $A = -3.685$ and $B = 2.395$ which indicates that this model also underestimates the true class probabilities. To verify that the classifier probabilities are better calibrated to the posterior probability we use 5-fold cross validation with the optimal Random Forest hyperparameters to classify the matching period dataset and the aliased period dataset. For 13 disjoint probability bins, the proportion of objects which are true members of the specified class is determined. The classifier probabilities are plotted against the posterior probabilities along with the calibrated classifier probabilities against the posterior probabilities. Figure 7.18 demonstrates the two plots generated by performing this experiment on the matching period dataset Random Forest model and the aliased period dataset Random Forest model. The calibrated probabilities are closer to the true posterior probability across the probability space.

In addition to this probability calibration we also determine optimal probability cuts for confident classifications from the ROC curves of the classification models. To determine this probability cut vector, we compute the Index of Union as shown in equation 4.26 on the ROC curve produced from each k cross validation set (Unal, 2017). The offset term is chosen by the pipeline user as the offset between the purity (minimise the false positives at the cost of more false negatives) of the classification and the recall (minimise the false negatives at the cost of more false positives) of the classification. The Index of Union determines an optimal probability cut for each class for each k cross validation set. All k probability cuts are then mean aggregated to produce the final probability cut vector for the model for a given offset. During classification these probability vectors are also propagated through the same probability calibration procedure as the classifier probability vectors creating a unique probability cut for each light curve based on the value r and Δ_i utilised in the probability calibration. For a given light curve, the class with the highest value of $\hat{p}(x|C_i) - \hat{p}(x; C_{i,\text{cut}})$ where $\hat{p}(x|C_i)$ is the calibrated probability of the light curve x being a member of i^{th} class C_i and $\hat{p}(x; C_{i,\text{cut}})$ is the calibrated probability cut of the light curve x for the i^{th} class, is the classification pipeline

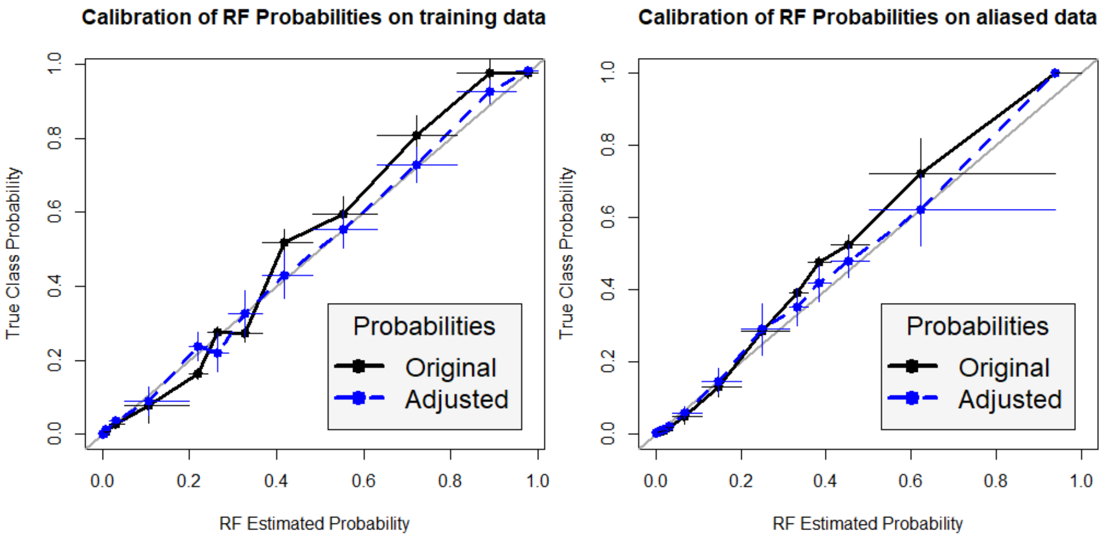


FIGURE 7.18: Before and after calibration Classifier probabilities against the Posterior probabilities for the matched period light curves RF model (left) and the aliased period light curves RF model (right). Both original models shown by the black lines initially underestimate the classification probability. When calibrated using the method developed by Boström, the classifier probabilities are much better at matching the true class probabilities shown by the dashed blue lines (Boström, 2008).

predicted class of light curve x. If every class has a negative value after this calculation, the classification model is not confident enough to assign any class to the light curve x and it is given an unknown predicted class, a 13th possible class.

7.5 Variability Detection Model

The variable star classification model allows the selection of a set of variable light curves for the training of a further variability detection model. This model is a binary classifier which separates candidate light curves into two classes, variable and non-variable. It is trained using the variability indices only as the GRAPE and PolyFit methods are too computationally intensive to apply to every sufficiently-sampled SkycamT light curve. The SkycamT subset of variable candidate light curves can then be processed by GRAPE and PolyFit and classified using the variable star classification models. The variability detection model can be trained in a supervised or unsupervised manner on one or two classes such as treating the problem as an anomaly detection problem where the light curve variability indices are clustered and the main clusters are assumed to be non-variable light curves (Shin et al., 2009, 2012). There are a number of methods which have been successfully used in astronomy to cluster variable and non-variable light curves using their variability indices. An Infinite Gaussian Mixture Model (GMM) can be used to cluster the data into a set of clusters with Gaussian noise depending on the distribution of the features (Shin et al., 2009). Unlike other clustering methods, the infinite GMM does not require the number of expected clusters to be provided prior to operation removing any prior assumptions on the distribution of the data. Whilst training the model will assume there is an infinite number of possible clusters of which only a finite number are populated by the dataset. Gibbs sampling then determines the optimal set of populated clusters. An alternative approach is to make use of the one-class Support Vector Machine (SVM) (Schölkopf et al., 2000). The one-class SVM is similar to the multiclass variant discussed in chapter four except only one class is present in the training dataset and the optimisation changes to determining only the members of this class. The SVM accomplishes this by performing a mapping to a higher dimensional space using a kernel function and then selecting a hyperplane that maximises the margin between the training data and the origin. New candidates which are located near the training data are classified as members of the one training class otherwise they are declared to be too anomalous to be classified as a member of the training class. This algorithm can be used for detection of known object types or for anomaly detection. This method has been used to detect anomalous sources from the Wide-field Infrared Survey Explorer (WISE) spacecraft (Solarz et al., 2017). These methods have limitations in noisy data like the SkycamT light curves as the clusters are all heavily convolved and it is difficult to make confident decisions based on single light curves.

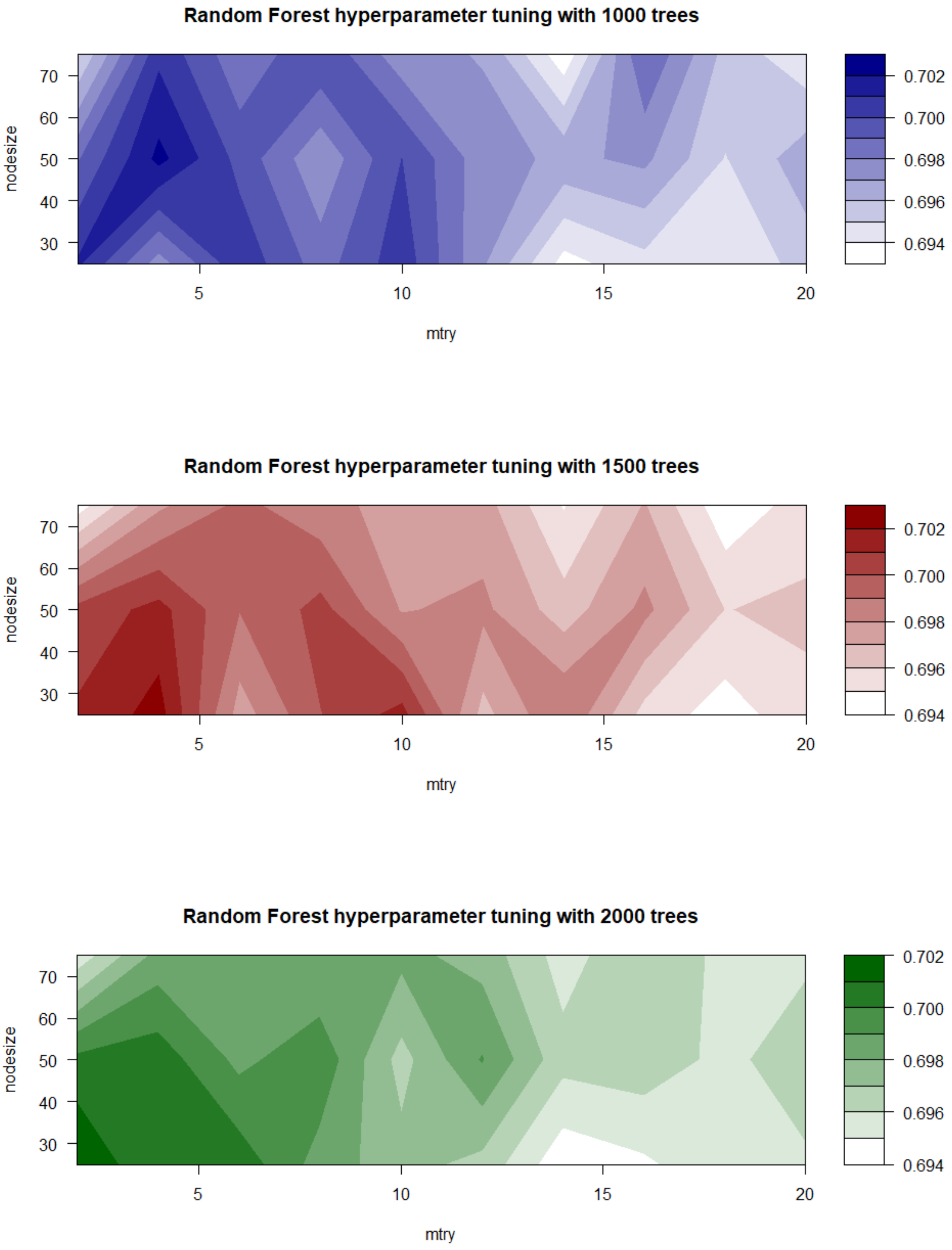
The SkycamT variability detection model is treated as a supervised binary classifier with two distinct variable and non-variable datasets. We selected the Random Forest classifier to train the variability detection model due to the correlation in the variability indices. The variable star light curves used to train both the matched-period and aliased-period classifiers in the previous section were selected as the variable light curve training set and constitutes 1895 light curves. We found that including the aliased light curves provided a training set which better identified variable light curves from the variability indices. The aliased features are only computed after the GRAPE period estimation and should have minimal effect on the training using only the variability indices. The non-variable training light curves are randomly selected from the remaining 588,597 sufficiently sampled SkycamT light curves not present in the training set. This creates a set of non-variable training light curves equal in size to the variable light curve training dataset. There is a risk in this random selection of light curves as we do not want to introduce a sampling bias through the random selection and more importantly, it is expected that a small percentage of these light curves are potentially variable candidates. In the following section we demonstrate the hyperparameter tuning results on the Random Forest variability detection model and analyse the importance of the selected non-variable training data, the use of aliased light curves in the variable training data and the features in the identification of variable candidate sources.

7.5.1 Hyperparameter Tuning

The Random Forest hyperparameters were determined using a grid search for the three main arguments, $m_{\text{try}} = [2, 4, 6, 8, 10, 12, 14, 16, 18, 20]$, $\text{ntrees} = [500, 1000, 1500]$ and $\text{nodesize} = [25, 50, 75]$ using a single set of the non-variable training light curves. The m_{try} parameters cover a smaller range due to the reduced number of features in this classification task. The cross validation is performed on each combination and the best performing hyperparameters using the F1 score are selected and used to train the final classification model. The Random Forest hyperparameter tuning results are shown in figure 7.19 with the best performing cross validation using the F1 score as a discriminator is $\text{ntree} = 1500$, $m_{\text{try}} = 4$ and $\text{nodesize} = 25$ with a mean F1 score of 0.702. As with the aliased period variable star classification model, the variability detection model requires a larger forest with deeper trees to discriminate between the two classes as they are heavily convolved due to the noise in the SkycamT light curves.

7.5.2 Variable Detection Performance

The decision to randomly select the non-variables from the SkycamT light curves which were not cross-matched with the AAVSO variable star index catalogue must be shown to not introduce significant variability to the classifier performance depending on the non-variable training set. We designed an experiment to confirm that this random selection does not substantially influence the trained model by generating 20 different non-variable training samples to compare the performance across the models. Each of the twenty training sets is evaluated using a 5-fold cross validation with 2 repeats with the mean F1 score indicating the performance of that training set and the standard deviation of the F1 score indicating the variability in the trained models. The hyperparameters determined from the tuning procedure are used for the Random Forest classifiers. The results of this experiment are shown in figure 7.20 indicating that the variance in the twenty models are similar to the variance of each specific 5-fold cross validation. The mean F1 score of this experiment is 0.713 with a standard deviation of the means of 0.007. Only one of the cross validations has a mean significantly outside of the $\pm\sigma$ margin indicating a possibly biased non-variable training set. Whilst this experiment cannot make a conclusion on how much the presence of candidate variables in the randomly selected non-variable training set influences the final model, it does show that the overall quality of the non-variable training set is similar regardless of the sampling. Overall, the stratified sampling of the training set during the cross validations have a higher variance than the random sample of non-variable training light curves. Figure 7.21 demonstrates the ROC curves of a 4-fold cross validation applied to the 12th training dataset as the



Random Forest hyperparameter tuning with 1000 trees

Random Forest hyperparameter tuning with 1500 trees

Random Forest hyperparameter tuning with 2000 trees

FIGURE 7.19: Contour plots of the hyperparameter tuning of the variability detection Random Forest 5-fold cross validation with 1000, 1500 and 2000 trees. The best performing cross validation using the F1 score as a discriminator is ntree = 1500, $m_{\mathrm{try}} = 4$ and nodesize = 25 with a mean F1 score of 0.702 although it is not very significant relative to the other models with a higher standard deviation between the 5-folds than between the hyperparameters. The hyperparameter space is highly non-linear due to the difficulty of this classification on the poor quality variability indices.

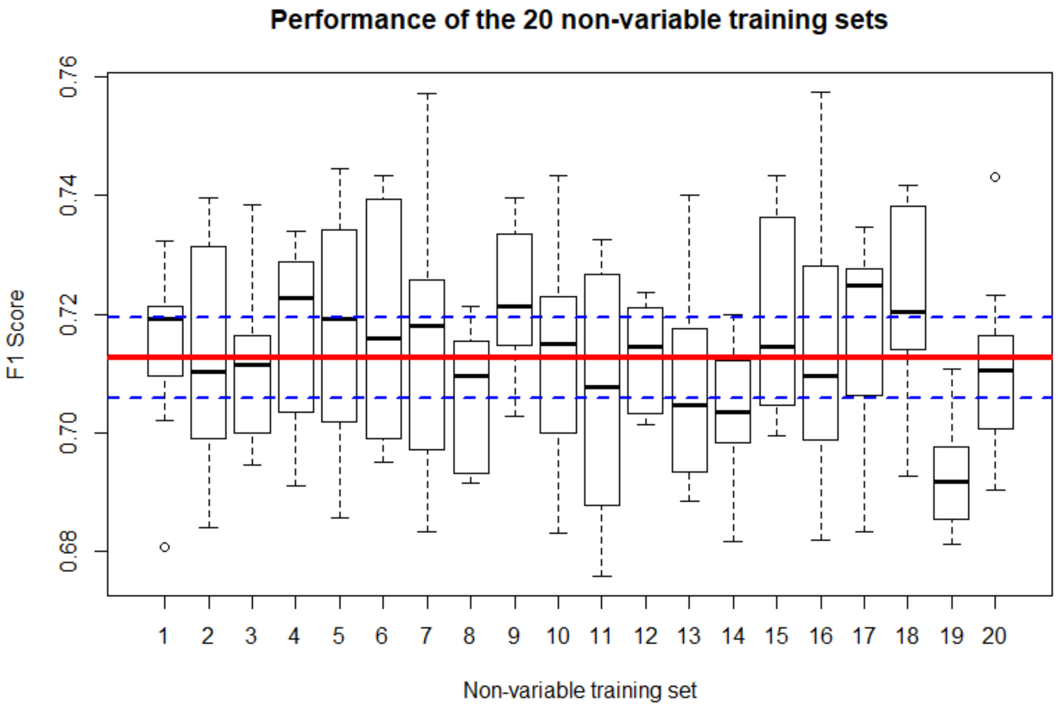


FIGURE 7.20: Box Plots of the F1 score performance of the 20 different non-variable training sets tested using 5-fold cross validations with 2 repeats (10 models per cross validation). The red line indicates the mean F1 score of every training set and the dashed blue lines indicate the $\pm\sigma$ where σ is the standard deviation of the means of each 5-fold cross validation. Despite randomly selecting the non-variable set from the non-cross matched SkycamT light curves with no guarantee of preventing unknown variable candidates from influencing the dataset, the performance is similar across most of the twenty cross validations.

mean F1 score of this cross validation most closely matches the overall mean of the 20 datasets. The model is slightly better at classifying variable light curves as variable verses rejecting non-variable light curves. This is likely a result of the spurious and seasonal trends in the SkycamT light curves which can induce signals which appear similar to astrophysical variable signals making it more difficult for the model to reject them especially with the limitations of the variability indices.

We also trained a variability detection model using the above method but with only the variable light curves with a hit or multiple/submultiple GRAPE estimated period. This model has less training light curves and is therefore at risk of additional bias problems from the non-variable training set selection. With a higher quality set of variable light curves the model may outperform the full variable set model. We use the same hyperparameters established in the previous tuning test although we concede that these may no longer be optimal due to the change in the training conditions. Figure 7.22 demonstrates the F1 score of the 20 non-variable training sets using the subset of variable light curves in this experiment. The mean F1 score performance of this new

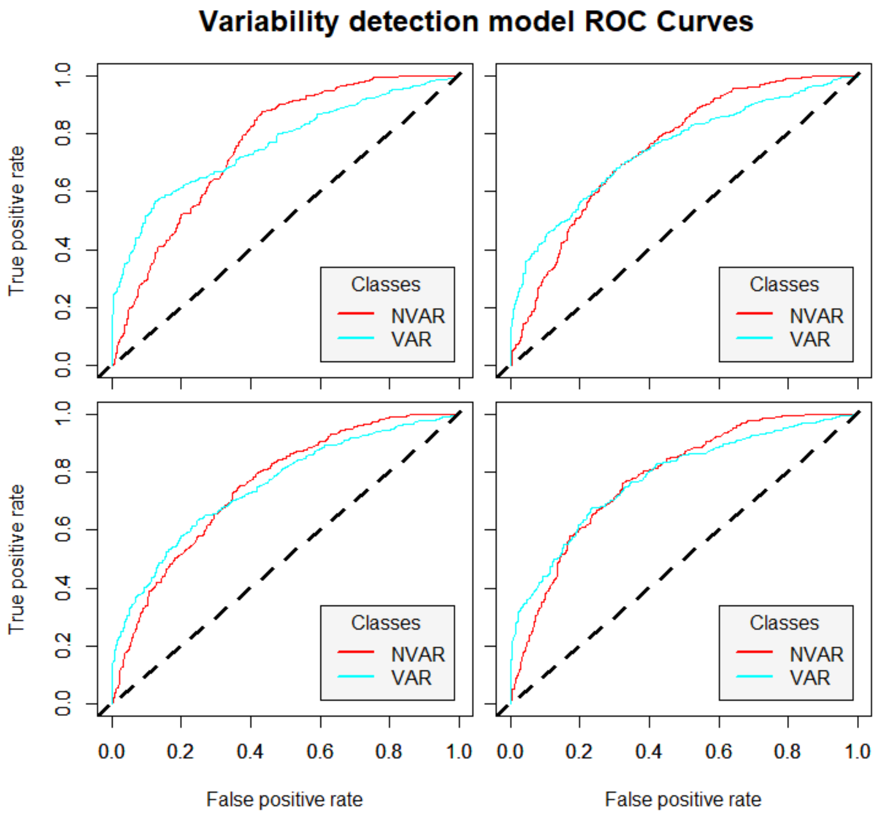


FIGURE 7.21: ROC curves of a Random Forest 4-fold cross validation with ntree = 1500, $m_{\text{try}} = 4$ and nodesize = 25 on the 12th training dataset. The performance of the model in detecting variable light curves is superior to the performance in rejecting non-variable light curves. This is likely a result of the spurious and seasonal trends in the SkycamT light curves.

training dataset is substantially improved with a mean F1 score of 0.799 with a standard deviation of 0.007. Despite the assumption that the variance between the different non-variable training sets would increase, the overall statistics appear similar to the previous full variable dataset cross validations with only one dataset having a mean significantly outside the $\pm\sigma$ margin. There do appear to be more outliers but an aggregation of these models should correct for the poor classifications of individual models. Figure 7.23 shows the ROC curves of a 4-fold cross validation applied to the 12th training dataset as the mean F1 score of this cross validation most closely matches the overall mean of the 20 datasets. These ROC curves indicate that, as with the full dataset, the matched period subset variability detection models are better at detecting variable light curves than rejecting non-variables supporting the previous conclusion as it is independent of the training set split.

Ultimately, it would appear that the best performance is gained from using the subset variable training set models although there is still a question over whether the full dataset may still identify more poor quality variables at the cost of more false positives. It is also

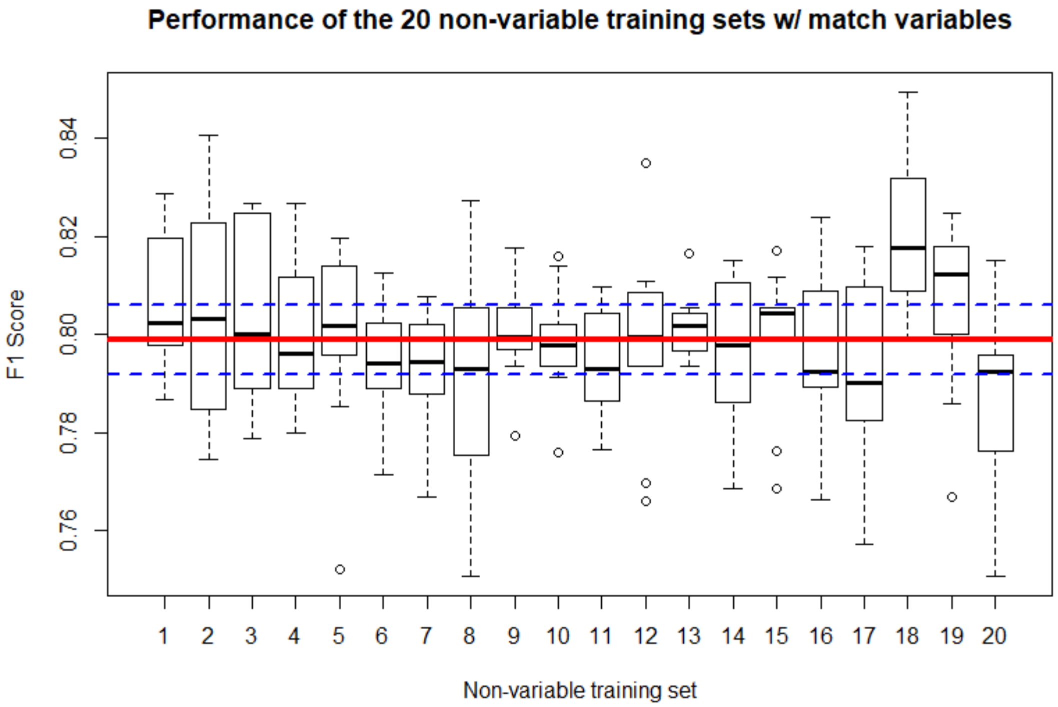


FIGURE 7.22: Box Plots of the F1 score performance of the 20 different non-variable training sets tested using 5-fold cross validations with 2 repeats (10 models per cross validation). The variable training set has been subset to the hit and multiple/submultiple estimated period sources. The red line indicates the mean F1 score of every training set and the dashed blue lines indicate the $\pm\sigma$ where σ is the standard deviation of the means of each 5-fold cross validation. Despite randomly selecting the non-variable set from the non-cross matched SkycamT light curves with no guarantee of preventing unknown variable candidates from influencing the dataset, the performance is similar across the twenty cross validations.

possible to improve the quality of these models by performing periodic retraining of the variability detection model as additional variable light curves are confidently classified. Adding these light curves to the variable training set whilst potentially removing variable light curves from the non-variable training set whilst simultaneously increasing the size of the non-variable training set will generate a training sample which better represents the statistics of the SkycamT database. This should reduce the false positives whilst providing more confident classification of acceptable quality variable light curves.

We utilise our initial model selected by the hyperparameter optimisation to classify the 590,492 SkycamT light curves from their 35 variability indices. We use a probability cut-off with an offset of 0.2 to improve the purity of the classification over recall as we want to minimise the number of false positives due to limited computational power to process a large dataset. We obtain 88,849 variable candidates from this classification with a variability cut off of 0.587. We also applied the model produced from the training subset of hit and multiple/submultiple period estimation variable light curves with an

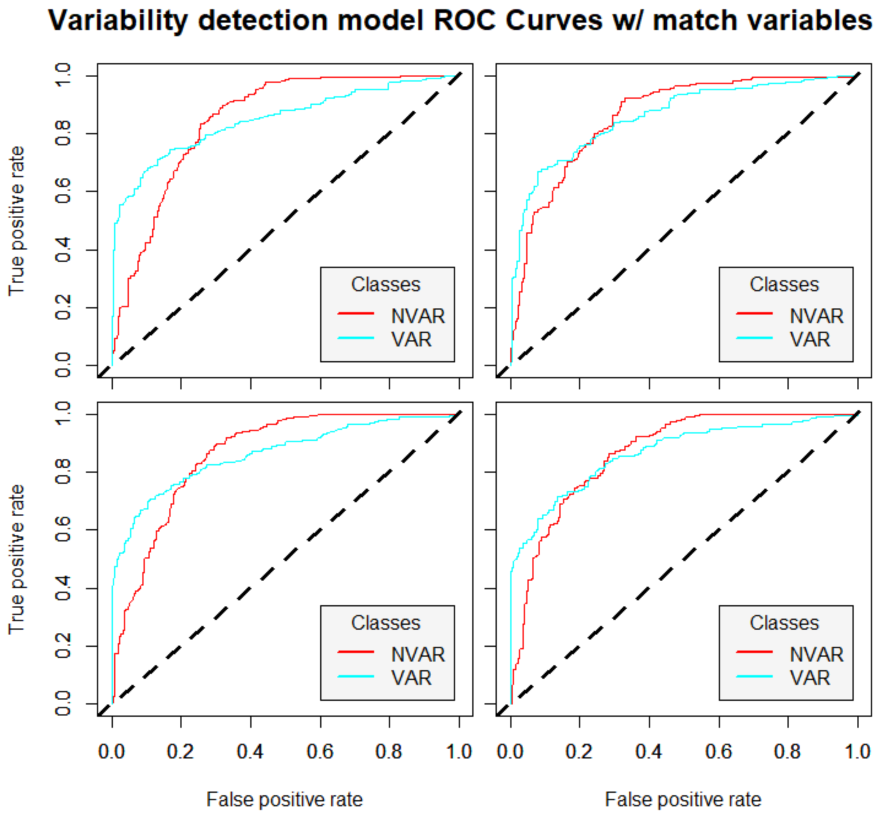


FIGURE 7.23: ROC curves of a Random Forest 4-fold cross validation with ntree = 1500, $m_{\text{try}} = 4$ and nodesize = 25 on the 12th training dataset with the subset variable training set. As with the full dataset models, variables are easier to classify than non-variables. The higher AUC of the better performing models using this subset of variable light curves is clear relative to the full dataset.

offset of 0 to identify the number of candidate variables proposed by this model. This model selects 103,790 light curves as variable candidates. This selection is equivalent to 17.58% of the SkycamT light curves which is higher than the 10% of the Northern Sky Variability Survey (NSVS) light curves which were found to be variable using a similar approach (Shin et al., 2012). This overestimation of variable candidates is likely a continuation of the difficulties of discriminating between variable light curves and non-variable light curves plagued with systematic noise. We selected the top 52,000 light curves from the initial 88,849 variable candidates sorted by the detection model probability of being a variable light curve. These light curves were processed by the GRAPE period estimation method and the genetic PolyFit algorithm to produce a full set of features for variable star classification and anomaly detection.

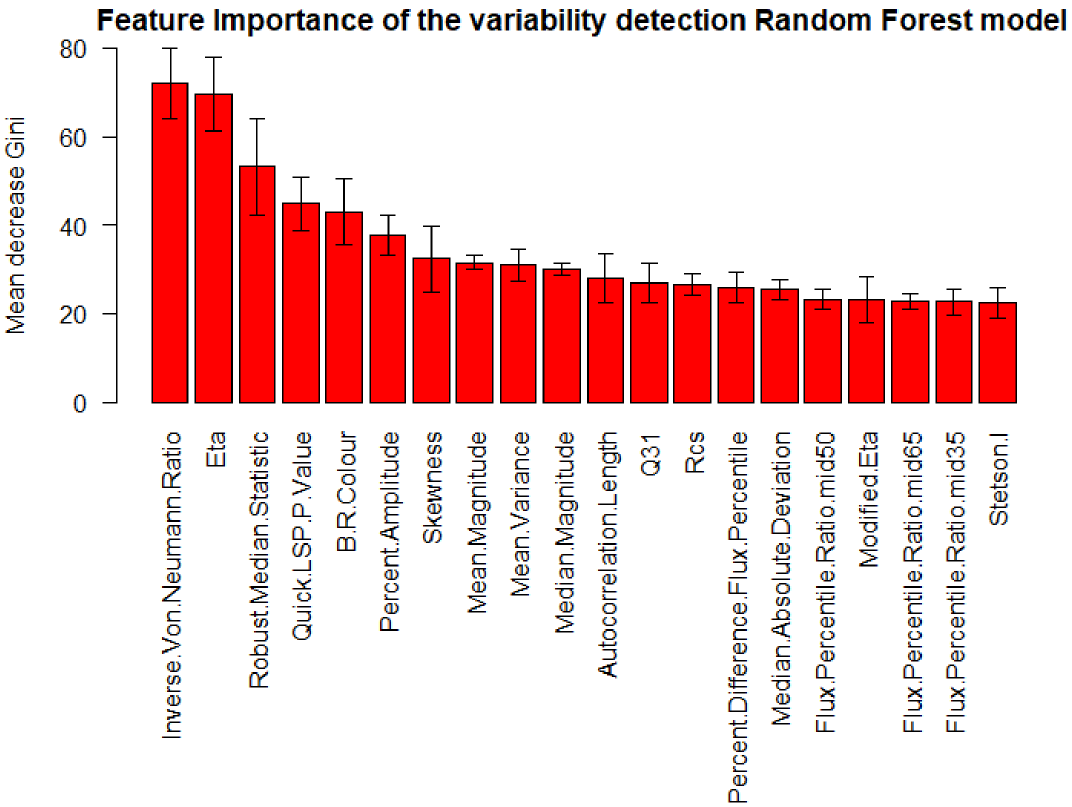


FIGURE 7.24: Bar plot of the mean decrease Gini of the top 20 features in the variability detection Random Forest model. The error bars demonstrate the $\pm 2\sigma$ where σ is the standard deviation in each of the features across the 20 different models using the 20 different non-variable light curve training sets. The top 4 features are strong indicators of the presence of periodic variability in the light curves. The B-R colour also contains important information on potential variability as some of the variable star types are in specific colour regions.

7.5.3 Feature Importance

The mean decrease Gini statistic is extracted from the trained Random Forest models for each of the 20 non-variable training sets. The mean and standard deviation of the mean decrease Gini for each feature across the 20 models (each model is trained on the entire training dataset) is computed. Figure 7.24 demonstrates the result of this computation where the bars represent the mean aggregate of the 20 mean decrease Gini values for each feature and the error bars represent the $\pm 2\sigma$ of the mean decrease Gini where σ is the standard deviation. The Inverse Von Neumann ratio and η variability index features are the dominant indicators of variability as they measure the differences between consecutive data points. A variable light curve has a continuous signal influencing these differences which clearly influences the feature. The Robust Median Statistic is a robust measure of the light curve standard deviation which is a strong indicator of variability when combined with the mean magnitude feature as variable light curves have standard deviations above those expected for the associated magnitude. The Quick LSP P-Value

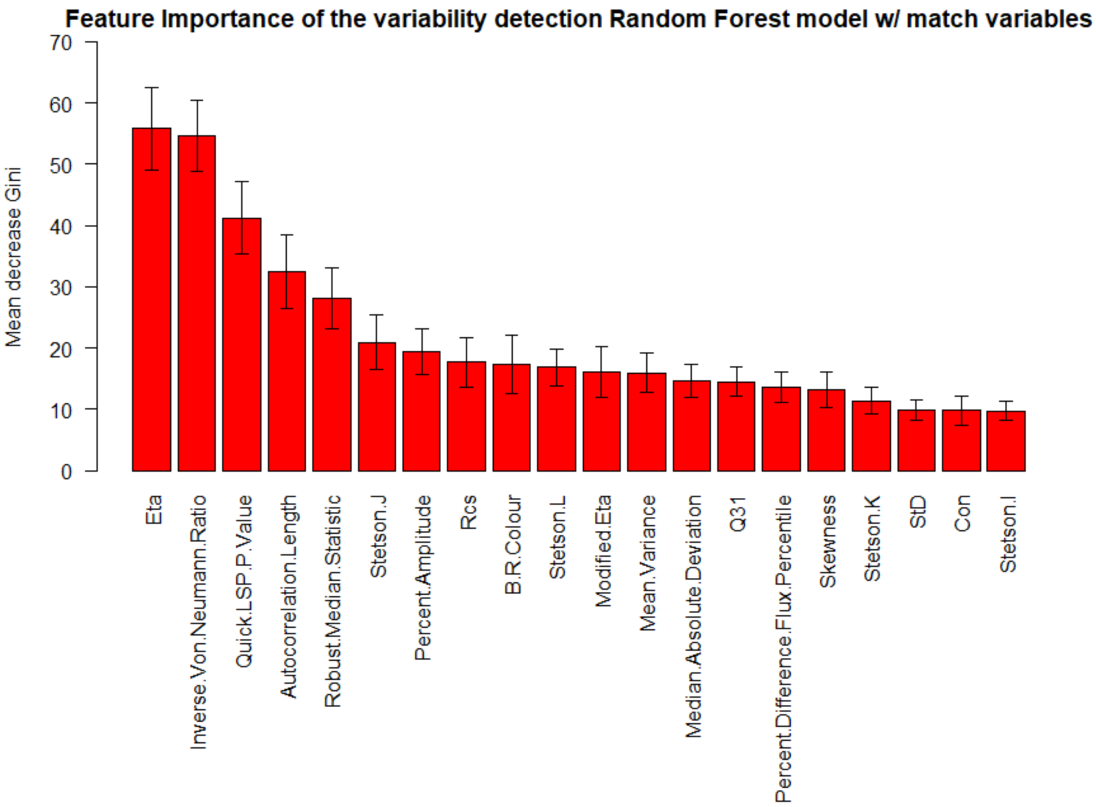


Figure 7.25: Bar plot of the mean decrease Gini of the top 20 features in the variability detection Random Forest model using the training subset of period estimated hit and multiple/submultiple variable light curves. The error bars demonstrate the $\pm 2\sigma$ where σ is the standard deviation in each of the features across the 20 different models using the 20 different non-variable light curve training sets. The top features are similar to the full training set model except the Quick LSP P-Value is stronger along with improvements in the importance of the Autocorrelation length and the Stetson J Index.

feature is also highly useful as it demonstrates the strength of the periodicity of a light curve using the sidereal alias. Even though this is not the true period, the strength of this alias still provides important information to the model and many of the classes trained on the variable star classifier are highly periodic (Shin et al., 2009, 2012). The B-R colour also contains important information on potential variability as some of the variable star types fall in specific colour regions due to the stellar properties required to drive the variability (Percy, 2008).

We also compute the mean decrease Gini measures for the 20 non-variable training sets with the subset of training variables with hit and multiple/submultiple estimated periods. The feature importance results of these models are shown in figure 7.25. The Inverse Von Neumann ratio and η variability index features remain the strongest indicators of variability in these models however the Quick LSP P-Value has improved to a clear second place behind these two correlated features. The other two features with a strong improvement in their importance are the autocorrelation length and the Stetson

J index. The autocorrelation length is a strong indicator of long period variability which compliments the Quick LSP P-Value as the Quick LSP period for a long period variable is very close to the sidereal day which can cause difficulties in estimating the strength of the periodic component of the true period. The Stetson J index is a robust variability measure which computes the sum of the weighted magnitude change between pairs of data points (Stetson, 1996). In this pipeline we define the pairs as each consecutive group of two data points. Figure 7.26 demonstrates the SkycamT light curves associated with the minimum, median and maximum of the top 6 performing variability indices, η variability index, Quick LSP P-Value, Autocorrelation Length, Robust Median Statistic, Stetson J and the Percent Amplitude. Some of these light curves are interesting such as the maximum Stetson J and the minimum η caused by the blending of two sources. The maximum Quick LSP P-Value light curve corresponds to the long period variable Mira. As this feature has a limit of 324, there are multiple light curves with this maximum value. The light curve of the maximum Autocorrelation Length feature also corresponds to a known variable star, the Semiregular variable V521 Ophiuchi.

Despite the performance difference between the variable subset detection model and the full dataset detection model, the most important features are very similar in both position and in relative mean decrease Gini value. This suggests that the performance difference is a result of a higher quality training dataset rather than any definitive differences in the models. This assumption can be examined by computing the number of objects in the 590,492 SkycamT light curves selected by the two models given the same input arguments. Using an offset of 0 to decide the probability cuts for both models, the full dataset variability detection model selects 165,573 light curves with a variable probability cut of 0.481 whereas the subset hit and multiple/submultiple variables training set selects 103,790 light curves with a probability cut of 0.456. These results indicate that our assumption is false, despite the similarity between the feature importance in the two models, the full dataset classifies many more light curves as variable candidates. Many of these additional variable candidates are like false positives due to the number of aliased light curves from the training set which are not true aliases, but spurious periods. This indicates that the subset of variable training data is the better method to train the variability detection model as the reduced size of the training dataset can be augmented through additional training cycles and additional non-variable training sets.

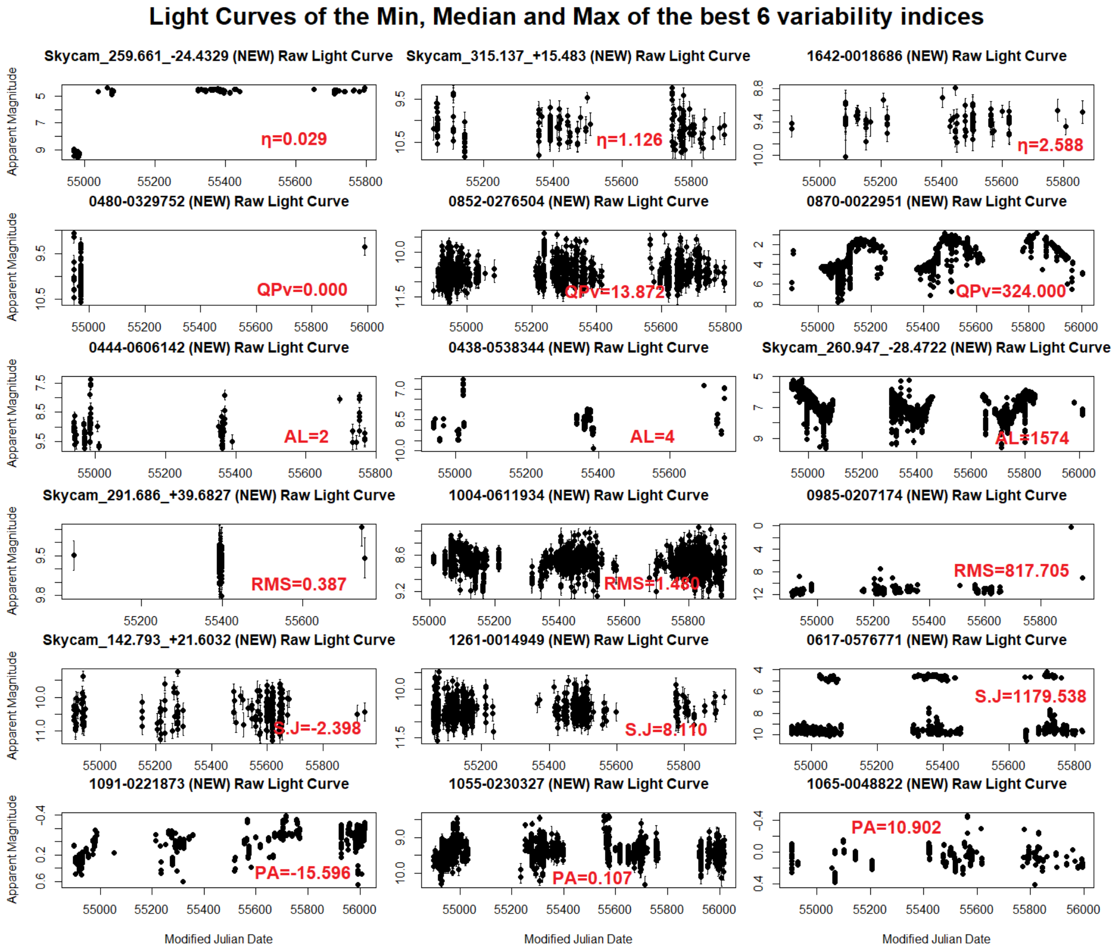


FIGURE 7.26: A set of SkycamT light curves corresponding to the minimum, median and maximum values of the 6 most important variability indices, η variability index, Quick LSP P-Value, Autocorrelation Length, Robust Median Statistic, Stetson J and the Percent Amplitude. These light curves exhibit interesting effects such as blending and two of them are known long period variable stars.

7.6 Random Forest based Anomaly Detection

The variable star classification models are trained using a set of light curves of 12 classes. These 12 classes were chosen as they fulfilled two requirements for the SkycamT light curves: They need to have characteristic variability detectable over the, at best, 0.2 mag amplitude noise present in every light curve and there needed to be sufficient AAVSO cross-matched and GRAPE estimated period matched light curves to perform the 5-fold cross validation tuning operations. As a result, there are likely light curves in the SkycamT database of sufficient quality to be detectable as a variable light curve and yet the true class of these objects is not a trained class of the variable star classification model. These light curves, when applied to the models described above, would likely produce a classification probability vector which is matched to the most similar training set class or, for objects very dissimilar to all the training classes, the classification

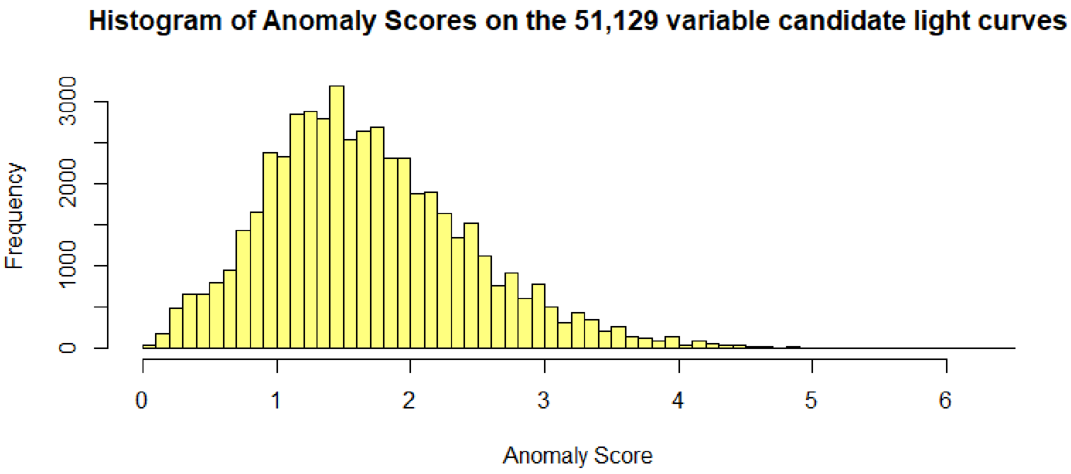


FIGURE 7.27: Histogram of the Anomaly scores of the 51,129 variable candidate light curves classified by the period-match variable star Random Forest classification model. The bulk of the light curves in this dataset have an anomaly score relative to the variable star classification training set of 1.14 to 2.13. The most anomalous light curve has a value of 6.46 and the 5% most anomalous light curves (2556 sources) in this dataset have values greater than 3.07.

probability vector would have a low, nearly equal probability for every class. This will result either in an unknown classification, or a low-confidence classification into the closest matching class. These light curves are known as anomalous light curves as they do not match the classifiers learned knowledge. The identification of anomalous light curves is of interest as it may identify interesting objects of additional classes to the current 12 trained classes.

Anomalous light curves can be determined through computing a distance (similarity) metric between the known light curves and the unknown candidate. This can be done through the computation of distance features directly from the light curve (Protopapas et al., 2006; Rebbapragada et al., 2009). Fortunately, it is possible to compute similarity metrics directly from the machine learning classification models using the training light curve feature vectors as the machine learning models weight the different features according to their importance (Richards et al., 2012; Bhattacharyya et al., 2011). The Random Forest classification models can be made to output a proximity matrix for a set of light curves. This matrix is an $n \times n$ matrix of similarities between the n feature vectors. For two feature vectors x_i and x_j, the similarity measure is computed by determining the proportion of the trees in the random forest where the two feature vectors are located in the same leaf node at the end of the classification. Similar feature vectors will end up in the same leaf node in every tree and produce a similarity $\rho_{ij} = 1$ whereas highly dissimilar feature vectors will never be in the same leaf node and therefore $\rho_{ij} = 0$. This is a very potent method of determining similarity as it is equivalent to a Euclidean

Table 7.9: Predicted Class labels after probability cuts and calibration on the 51,129 SkycamT light curves selected as variability candidates by the detection model. A 13th class has been added to identify light curves which did not secure sufficient classification confidence to obtain a class label. These may include non-variable false positives, poor period estimations or anomalous feature vectors.

Number	Predicted Class	Type	Count
1	β Lyrae	Eclipsing Binary	5924
2	β Persei	Eclipsing Binary	16544
3	Chemically Peculiar	Rotational Variable	4419
4	Classical Cepheid	Pulsating Variable	491
5	δ Scuti	Pulsating Variable	2392
6	Mira	Pulsating Variable	1469
7	RR Lyrae Fundamental Mode	Pulsating Variable	398
8	RR Lyrae Overtone Mode	Pulsating Variable	536
9	RV Tauri	Pulsating Variable	71
10	Semiregular Variable	Pulsating Variable	4733
11	Spotted Variable	Rotational Variable	983
12	W Ursae Majoris	Eclipsing Binary	6887
13	Unknown classification	Below Threshold	6282

distance between the two feature vectors as weighted by the importance of each feature in the classification model (Richards et al., 2012). The desired anomaly score must use this similarity in a method which produces a real valued measure of anomalousness. The similarity can be easily converted to this scaling using equation 7.7.

$$d(x_i, x_j) = \frac{1 - \rho_{ij}}{\rho_{ij}} \tag{7.7}$$

where $d(x_i, x_j)$ is the anomaly score of x_i based on a similarity measurement feature vector x_j. Richards et al. used this methodology in their ASAS classification pipeline through defining x_j as the second nearest neighbour to the test feature vector x_i in the training dataset utilised on the associated Random Forest model. Utilising this methodology on the 51,129 variable candidate light curves classified by the match-period variable star classification model, we find the anomaly scores for the variable candidate SkycamT light curves peak at 1.59 with a tail out to 6.46 with a top 5% percentile of 3.07. Figure 7.27 demonstrates this distribution of anomaly score for these 51,129 conservative probability cut variable candidate light curves.

7.7 Skycam bright variable source catalogue

Applying the variability detection and classification models to the 590,492 SkycamT light curves with greater than 100 data points, using probability cuts determined using the best performing sensitivity and specificity of the variable star class, 103,790 variable

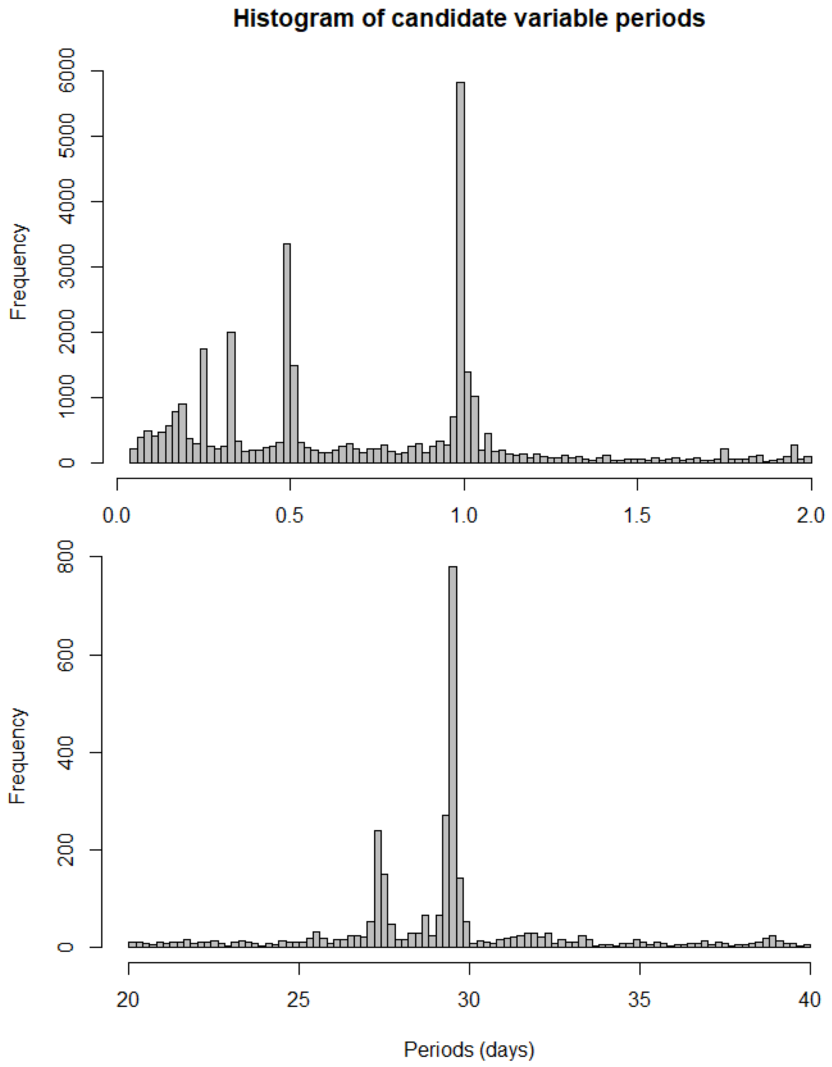


FIGURE 7.28: Histogram of the estimated period overabundance data structures near the known spurious periods for the 51,129 candidate light curves. Light curves in these overabundances cannot be trusted as true variables and must be removed.

candidates are selected. This is 17.6% of the full dataset which is a slightly higher estimation the 10% of the Northern Sky Variability Survey (NSVS) most likely due to the increased systematics in the Skycam light curves producing ‘phantom’ variability (Shin et al., 2012). Half of these candidates (51,129 sources) have had a full set of period and PolyFit features computed and classified using the variable star classification model. The probabilities have been calibrated to the posterior probability of the classes and probability cuts established from the Index of Union of the training data. The table of predicted classes is shown in table 7.9 as computed after the probability calibration and applied probability cuts.

Figure 7.28 shows the histogram produced by this set of variable candidates and it reveals that the variability detection model is not sufficient to filter all the spurious light curves

TABLE 7.10: Number of light curves filtered from the 51,129 candidate light curves due to proximity to known spurious data structures.

Spurious Period	Removed Period Range	Removed Light Curve Count
0.997 days	±0.04 days	8897
0.499 days	±0.02 days	4808
0.332 days	±0.01 days	2034
0.249 days	±0.01 days	1733
0.199 days	±0.01 days	1038
0.166 days	±0.01 days	812
27.5 days	±0.25 days	407
29.5 days	±0.25 days	1164

from the dataset. The diurnal alias is very strong in the SkycamT light curves likely due to the simplification of the airmass extinction corrections applied to the light curve photometry (Mawson et al., 2013). There are two additional significant overabundances located near 27.5 days and 29.5 days. The 29.5 day period is a result of light curves with a substantial contamination by moonlight across the lunar month. The 27.5 day is close to the seasonal alias of the lunar month and is also likely a result of moonlight contamination. To improve the quality of the final dataset, the light curves near these spurious overabundances are removed. This action does eliminate some potentially real candidate variables and therefore it would be advantageous to apply a superior method to attempt to retain significant signals. Richards et al. develop a method using the period and confidence produced by the Lomb-Scargle periodogram to determine the true variables based on the strength of the periodogram peaks against the proximity of the associated frequency to the known spurious frequencies (Richards et al., 2012). This method is not yet used in the pipeline due to the stronger spurious signals in the SkycamT data relative to the ASAS data used in their pipeline however it may be useful in the future if superior airmass corrections reduce the strength of the diurnal alias. The overabundances are manually determined through the histograms similar to those in figure 7.28 by selecting the significantly overabundant bins with a bin size of 0.02 days for the diurnal aliases and 0.05 days for the lunar spurious periods. Table 7.10 shows the number of light curves removed in this operation as well as the period range removed around each of the associated spurious overabundances.

The removal of these light curves leave 30,236 light curves with periods which cannot be explained by known spurious structures in the SkycamT data and are much more likely to be a result of true variability. Including the training dataset and assuming a similar proportion of the unclassified variable candidates are spurious, this suggests approximately 62,200 variable candidates which equals 10.5% of the SkycamT light curves with 100 or more data points which is very similar to the proportion of candidate variable sources found by the NSVS analysis. Table 7.11 demonstrates the predicted

Table 7.11: Predicted Class labels after probability cuts and calibration on the 30,236 SkycamT light curves remaining as variability candidates after removal of potential spurious light curves.

Number	Predicted Class	Type	Count
1	β Lyrae	Eclipsing Binary	2551
2	β Persei	Eclipsing Binary	8817
3	Chemically Peculiar	Rotational Variable	2827
4	Classical Cepheid	Pulsating Variable	436
5	δ Scuti	Pulsating Variable	1117
6	Mira	Pulsating Variable	1469
7	RR Lyrae Fundamental Mode	Pulsating Variable	251
8	RR Lyrae Overtone Mode	Pulsating Variable	239
9	RV Tauri	Pulsating Variable	71
10	Semiregular Variable	Pulsating Variable	4733
11	Spotted Variable	Rotational Variable	619
12	W Ursae Majoris	Eclipsing Binary	2904
13	Unknown classification	Below Threshold	4202

classes for the 30,236 light curves. The classes which have been depleted relative to table 7.9 are dominated by the short period and low amplitude variables especially the eclipsing binaries as the airmass extinction can produce eclipse-like features at diurnal alias periods. The removal of the potentially spurious light curves has also resulted in fundamental mode RR Lyrae variables outnumbering the overtone mode variants. Proportionally, a similar number of variable light curves have unknown classification after the removal operation compared to before the removal operation.

All the candidate variables in this dataset were not present in the AAVSO Variable Star Index as of the 28th of April 2016 at a tolerance of 148". Recently the Gaia mission released their Data Release 2 (DR2) for public consumption. The Gaia spacecraft has been making precision measurements of over one billion sources and DR2 contains data from the first 22 months of scientific observations. The Gaia DR2 contains 363,969 photometrically identified and classified variable sources which have been reduced to 24,984 variable sources with Gaia G-band magnitude brighter than +12. The right ascension and declination coordinates of these variable stars were downloaded along with the classification result from the Gaia machine learning classification pipeline. The 30,236 variable candidates from the above operations were then checked against this Gaia variable source catalogue. 540 of our variable candidates were within 148" distance of a Gaia variable source suggesting that approximately 1.8% of the candidate sources are due to a bright variable source with the remaining candidates likely a result of the SkycamT systematic noise and variables too faint to have been sufficiently sampled by Gaia in the first 22 months. Table 7.12 shows the predicted classes of these 540 variable candidates present in the Gaia variable catalogue.

Table 7.12: Predicted Class labels after probability cuts and calibration on the 540 SkycamT light curves matched to Gaia DR2 variable sources.

Number	Predicted Class	Type	Count
1	β Lyrae	Eclipsing Binary	23
2	β Persei	Eclipsing Binary	100
3	Chemically Peculiar	Rotational Variable	11
4	Classical Cepheid	Pulsating Variable	11
5	δ Scuti	Pulsating Variable	10
6	Mira	Pulsating Variable	113
7	RR Lyrae Fundamental Mode	Pulsating Variable	3
8	RR Lyrae Overtone Mode	Pulsating Variable	1
9	RV Tauri	Pulsating Variable	2
10	Semiregular Variable	Pulsating Variable	153
11	Spotted Variable	Rotational Variable	4
12	W Ursae Majoris	Eclipsing Binary	32
13	Unknown classification	Below Threshold	77

Table 7.13: Gaia classifier variable star types with the number of variables for each class in the 24,984 downloaded classified bright variable sources. The Long Period Variables appear to be dominant in the bright sources due to their high luminosity allowing them to be viewed across greater distances.

Number	Predicted Class	Type	Count
1	CEP	Classical Cepheid variables	496
2	DSCT_SXPHE	δ Scuti and SX Phoenicis variables	126
3	MIRA_SR	Mira and Semiregular variables	23,907
4	RRAB	Fundamental mode RR Lyrae variables	205
5	RRC	Overtone mode RR Lyrae variables	14
6	RRD	Dual-mode RR Lyrae variables	3
7	T2CEP	Type II Cepheid variables	233

Finally, using the classifier result from the Gaia machine learning pipeline, the classifier result from the SkycamT pipeline can be compared to the Gaia result. A direct comparison is difficult as the Gaia classifier is specifically trained to detect a strict set of pulsating variables only. There are seven classes shown in table 7.13 out of the nine total Gaia classes as the omitted classes have no known bright variables. As a result of the limitations of the Gaia classification, many of our variable candidates have been classified as Mira or Semiregular long period variables. Whilst this suggests that our variable candidates may be simply a result of blending with a nearby high amplitude variable source, the lack of eclipsing binary classification on the Gaia sources results in an inconclusive result. Ultimately, despite the lack of diurnal aliases in the Gaia data, the limited resampling of the Gaia DR2 data limits the potential for finding low amplitude variables and eclipsing binaries. Figure 7.29 demonstrates the results of comparing the Gaia classification to the SkycamT classification pipeline result. Assuming that Gaia DR2 has a sufficient baseline to reliably classify the long period variables, the successfully classified SkycamT long period variables is 51.7% of the expected amount.

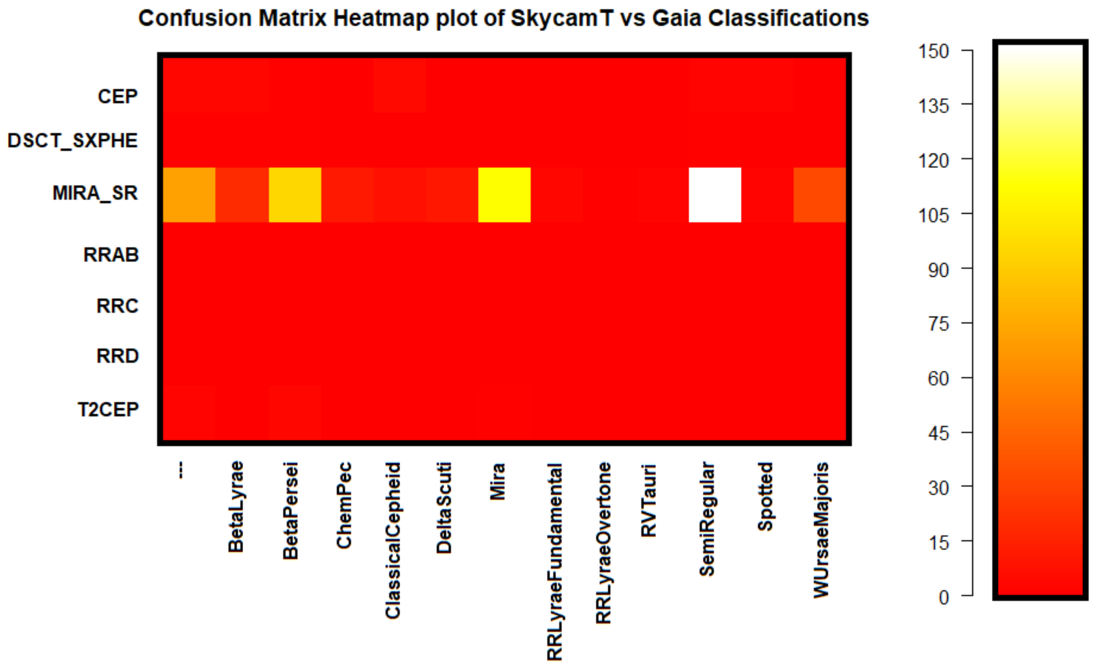


FIGURE 7.29: Confusion Matrix heatmap plot of the Gaia classifier result compared with the SkycamT classifier result for the 540 matched variable candidates. The only sources which agree well between the two sets are the Mira and Semiregular variables which is not unexpected due to the Gaia classifier being optimised for pulsating variables.

The SkycamT sources which are classified as long period variables agree strongly with the Gaia classifications. The remaining sources are primarily candidate rotational and binary variables which lack a class in the Gaia classification and may be either real variables undetected by Gaia or the result of blending, possibly aliased, with the Gaia long period variable source. For the RR Lyrae variables, SkycamT overestimates the number of both fundamental and overtone mode RR Lyrae variables relative to the Gaia classifications indicating that many of these classifications may be due to a systematic variability likely due to diurnal aliases given their characteristic periods of under a day or possibly blending with a nearby bright variable source. The Gaia data suggests that all the RR Lyrae variables are spurious results as there are no RR Lyrae matches in the Gaia classification model. As expected, these results suggest that many of the long period variable classifications are reliable detections benefitted by the long period seasonal trend removal method but many of the shorter period variables are still plagued with systematic noise which generate considerable numbers of false positive classifications on the SkycamT light curves.

Taking the 540 candidate variable sources from this analysis, interesting objects are selected for comment. The candidate variable source [0940-0328275] is classified as a W Ursae Majoris eclipsing binary with a probability of 0.8988, an anomaly score of 0.4124

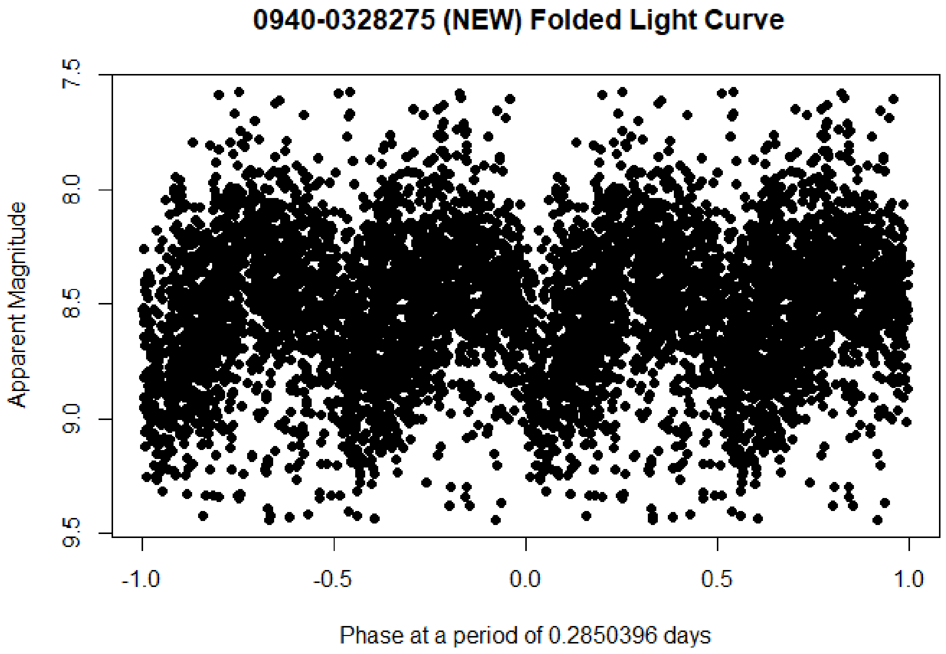


Figure 7.30: Folded light curve of the candidate variable source [0940-0328275] classified as a W Ursae Majoris eclipsing binary by the SkycamT classification pipeline. The GRAPE period has been doubled to reproduce the characteristic variability.

and a period (doubled from the GRAPE result) of 0.285 days shown in figure 7.30. The General Catalogue of Variable Stars (GCVS) identifies this source as V1089 Ophiuchi, an uncertain RR Lyrae variable of unknown subtype with an unreported period. The Gaia variable source catalogue classifies this source as a Mira or Semiregular long period variable which is a result not supported by the SkycamT light curve.

[Skycam_129.308_-16.3625] is a variable candidate light curve which is not matched to a USNO-B catalogue source shown in figure 7.31. The SkycamT classification pipeline has classified this object as a Classical Cepheid with a calibrated probability of 0.8135 and an anomaly score of 0.3477. Matching this source to the GCVS identifies it as a long period variable candidate with the name V370 Hydrae and Gaia identifies it as a Mira or Semiregular variable. GRAPE detects a period of 21.17 days for this source which suggests it may be a Small Amplitude Red Giant variable as it is too red in colour to be a Classical Cepheid. The ASAS classifier developed by Richard et al. identifies the source as a Long Secondary Period variable with a period of 386.969 days. Long Secondary Period variables are usually semiregular variables with a secondary longer period generated by an unknown process, possibly the coupling of a pulsation mode to a rotation mode (Percy, 2008). Epoch folding the light curve at this period reveals a possible variation although it is of poor quality due to seasonal sampling as shown in figure 7.31.

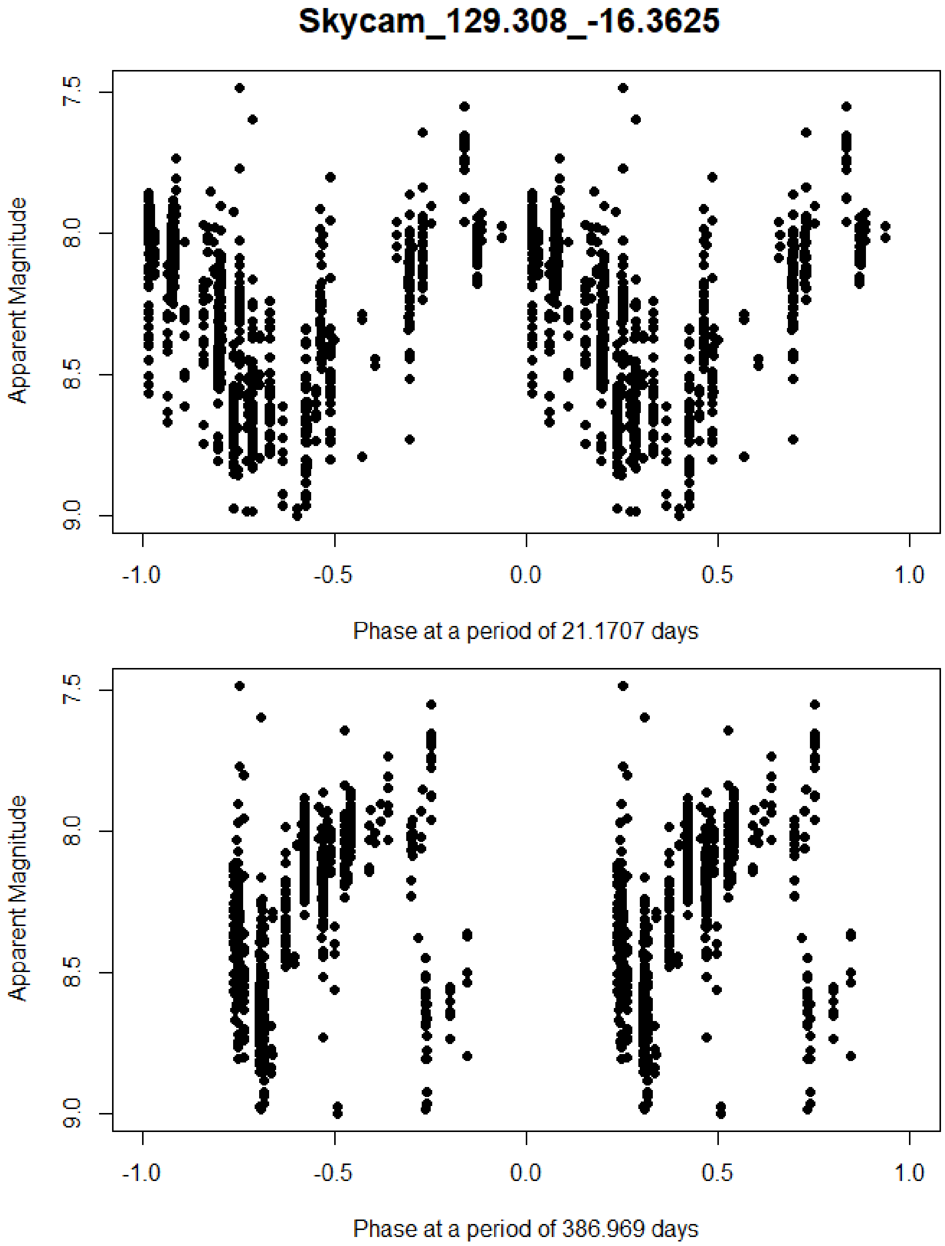


FIGURE 7.31: Folded light curve of the candidate variable source [Skycam_129.308_-16.3625] classified as a Classical Cepheid by the SkycamT classification pipeline. The top plot shows the epoch folding at the GRAPE period of 21.17 days and the bottom plot shows the light curve folded at the ASAS period of 386.969 days.

[Skycam_285.316_+37.5594] is another USNO-B unmatched source and has been assigned a unique Skycam identifier shown in figure 7.32. This candidate variable has not been classified as the probabilities of each class did not meet the required cut-off threshold. The variable class with the highest probability is the Mira class. This object is too far north to have been detected in the ASAS data. It is present in the Gaia classification as a Mira or Semiregular variable. It has a GRAPE period of 124.19 days and an anomaly score of 0.3850. Manual inspection of the folded light curve reveals this light curve to be a Mira-type variable at the limit of detectability where the source is only visible when the Mira is at its brightest. The GCVS identifies this source as the Mira variable RT Lyrae. The poor sampling of the Mira light curve due to the detection limit results in the GRAPE period being half of the GCVS catalogue period of 253.7 days.

The candidate variable source [0972-0706308], like the above source, is also unclassified by the SkycamT pipeline although the Semiregular probability is extremely close to the threshold. The light curve has a GRAPE period of 238.61 days and an anomaly score of 0.4085. The source has been classified from its ASAS light curve as a Semiregular variable with a period of 207.067 days. On the SkycamT data, the GRAPE period has a superior fit relative to the ASAS period as shown in figure 7.33. This variable is present in the GCVS and has been assigned the name EV Pegasi and has been classified as a long period variable star, a fairly general classification.

The candidate variable source [0901-0406646] is classified as a Classical Cepheid in the SkycamT pipeline with a probability of 0.8294, a GRAPE period of 7.30 days and an anomaly score of 0.4535. This source is very close to the magnitude limit of the SkycamT instrument and exhibits increased scatter as a result. Both ASAS and Gaia also identify this variable source as a Classical Cepheid, with this object being in the training dataset for the ASAS classifier. GCVS identifies this variable star as the Classical Cepheid V336 Aquilae and GRAPE, ASAS and GCVS agree on the period of the light curve. Figure 7.34 demonstrates the noisy SkycamT light curve of this Cepheid variable.

The candidate variable sources [0936-0585159] and [0866-0257809] are classified as Classical Cepheids in the SkycamT pipeline with respective probabilities of 0.8638 and 0.6633 and respective anomaly scores of 0.6026 and 0.6835. These anomaly scores differ substantially from the other Classical Cepheids classified by the pipeline. The Gaia classifier also identifies these two sources as Classical Cepheids but GCVS identifies them as the variable stars TX Delphini and W Virginis, Type II Cepheid variables. Due to a lack of sufficient training examples as there are few Type II Cepheids bright enough to be detected by SkycamT, the SkycamT pipeline has not been trained to identify this class of variable and therefore the Classical Cepheid class is chosen as the most similar class with a notable increase in anomaly score. Gaia is trained with a Type II Cepheid class

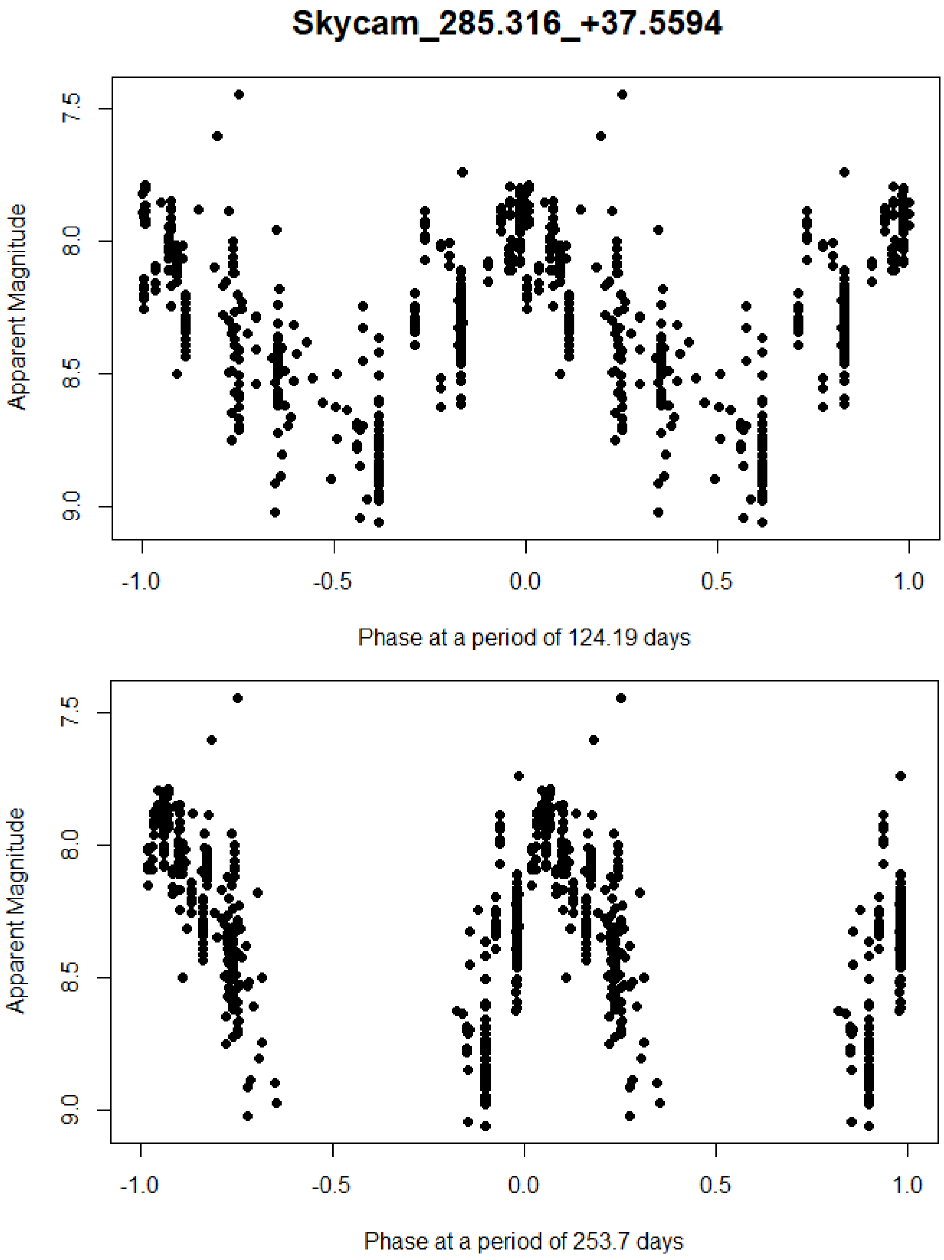


FIGURE 7.32: Folded light curve of the candidate variable source [Skycam_285.316_+37.5594] unclassified by the SkycamT classification pipeline. The top plot shows the epoch folding of this Mira variable at the GRAPE period of 124.19 days due to sampling artefacts produced by the detection limit of SkycamT and the bottom plot shows the GCVS catalogue period of 253.7 days.

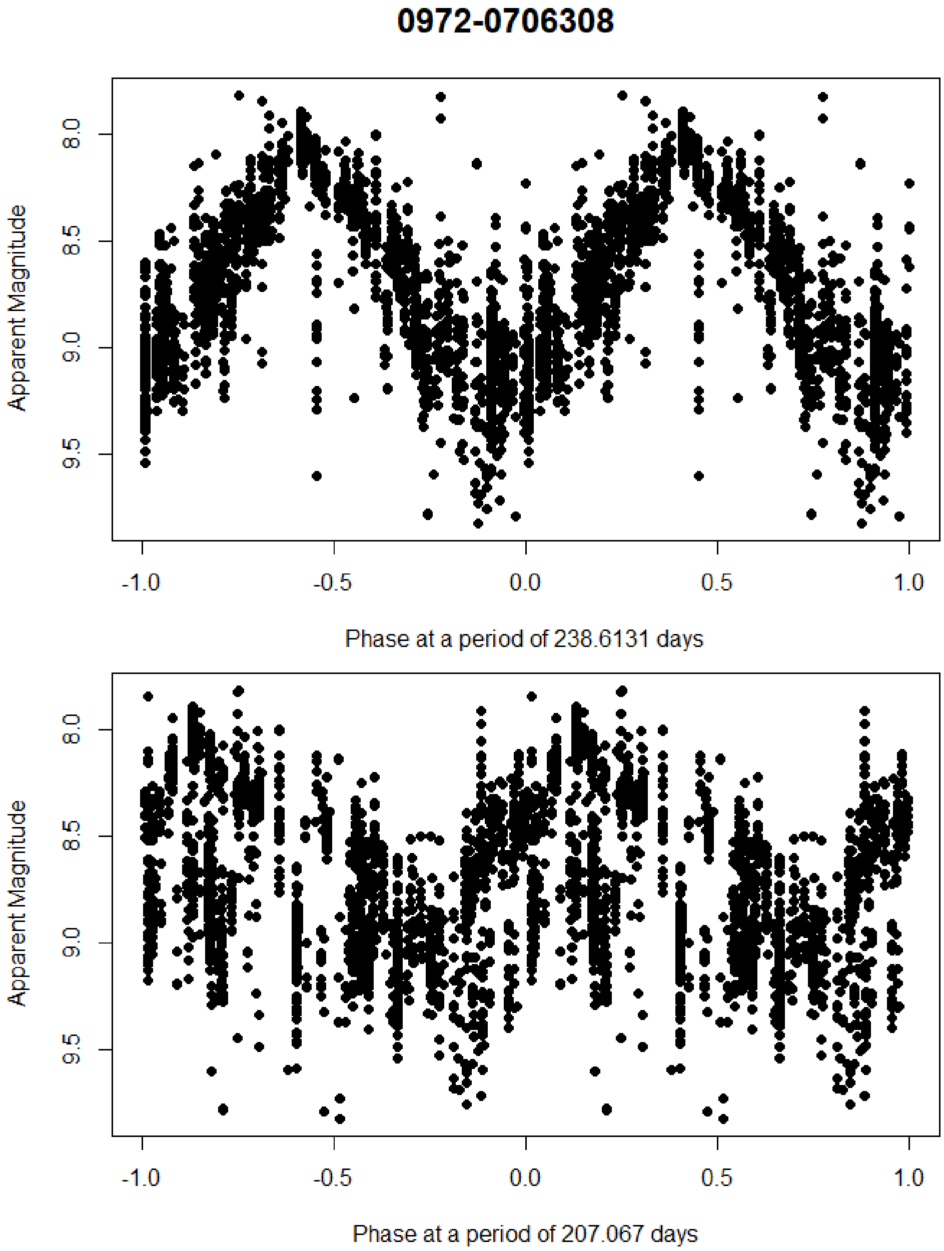


FIGURE 7.33: Folded light curve of the candidate variable source [0972-0706308] unclassified by the SkycamT classification pipeline. The top plot shows the epoch folding of this Mira variable at the GRAPE period of 238.61 days and the bottom plot shows the ASAS light curve period of 207.07 days. The GRAPE period is a much better fit on the SkycamT light curve.

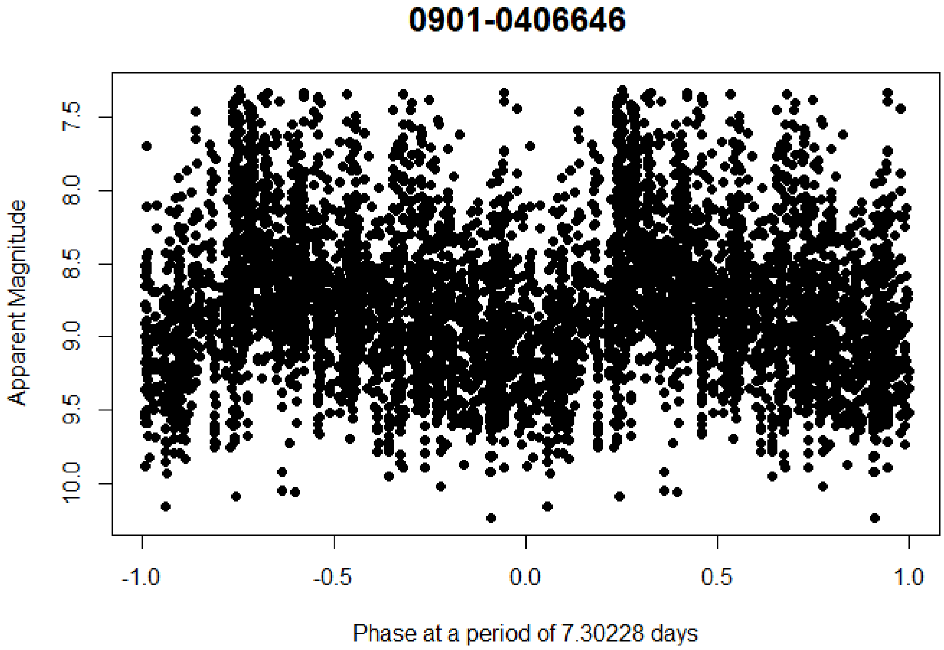


FIGURE 7.34: Folded light curve of the Cepheid variable [0901-0406646] classified as a Classical Cepheid by the SkycamT classification pipeline. The period of this variable star is 7.30 days which agrees with the periods from the GCVS and ASAS data.

and therefore it has misclassified these two variable stars. Figure 7.35 shows the light curves of these two variable stars folded at their GRAPE period which agrees with the GCVS period.

[Skycam_70.389_+40.1975] is shown in figure 7.36. This candidate source is classified as a Classical Cepheid with a probability of 0.7148 and an anomaly score of 0.6949. Gaia identifies this source as a Semiregular variable and the GCVS classifies the source as a long period variable star named HO Persei. This variable star is too far north to appear in the ASAS classifications. GRAPE identifies a period of 22.46 days but manual inspection of the folded light curve reveals this is likely an incorrect period due to many data points poorly aligning with the central sawtooth signal. The raw light curve of this variable star also shown in figure 7.36 reveals the star has a significant long term irregular variability which suggests that the object is a type SRb non-periodic semiregular variable star.

The candidate variable source [0683-0807889] is classified as a Chemically Peculiar variable in the SkycamT pipeline with a probability of 0.4962, a GRAPE period of 11.06 days and an anomaly score of 1.3256. The folded and raw light curves of this source are shown in figure 7.37. This candidate variable source is within 148" of a Gaia classified Mira or Semiregular variable star but has a spectral class A which is a strong property of chemically peculiar stars combined with the 11.06 day period which leads to this classification.

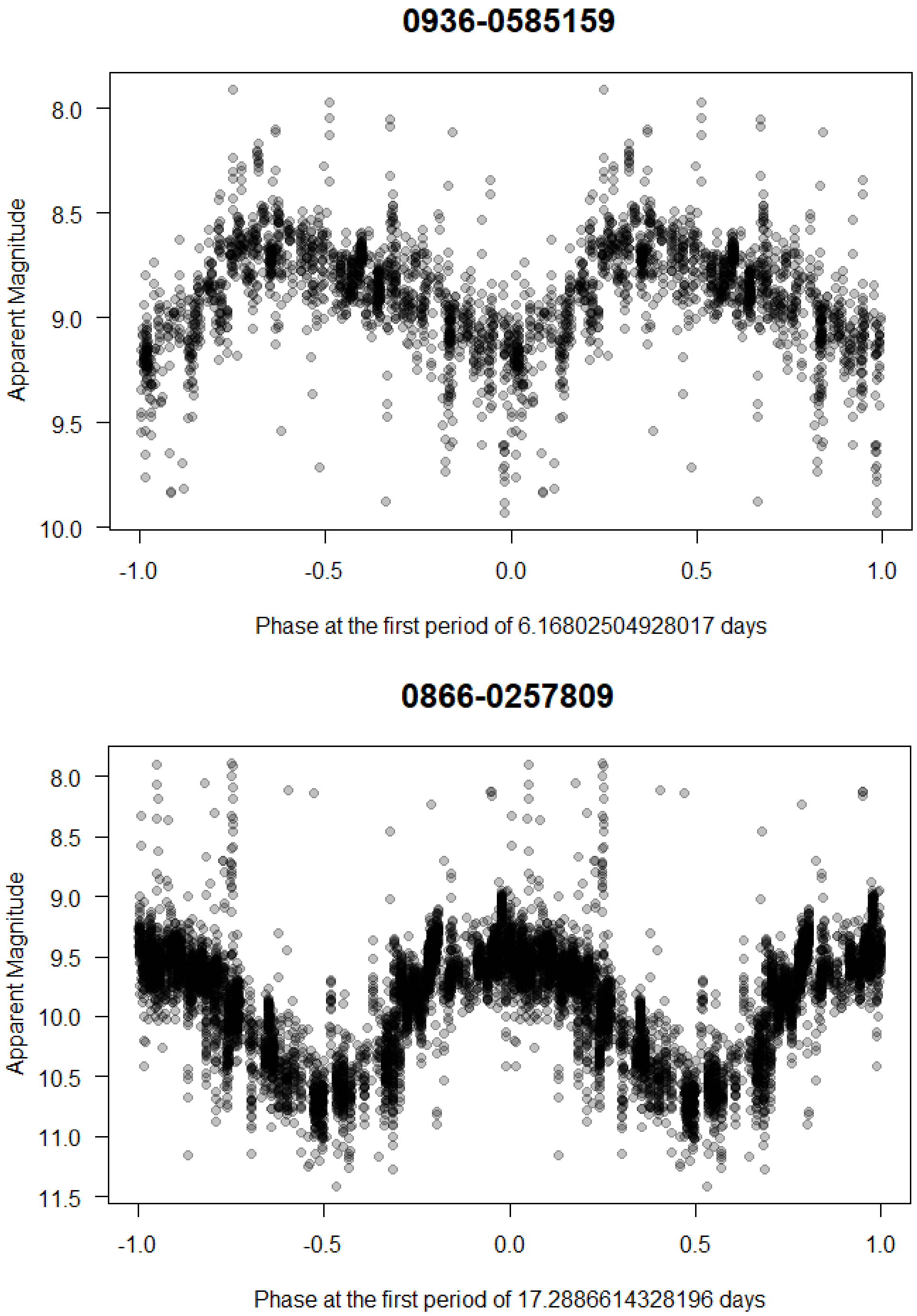


Figure 7.35: Folded light curves of the Type II Cepheid variables [0936-0585159] and [0866-0257809] classified as Classical Cepheids by the SkycamT classification pipeline. The top plot shows the light curve of TX Delphini with a period of 6.17 days and the bottom plot shows the light curve of W Virginis with a period of 17.29 days.

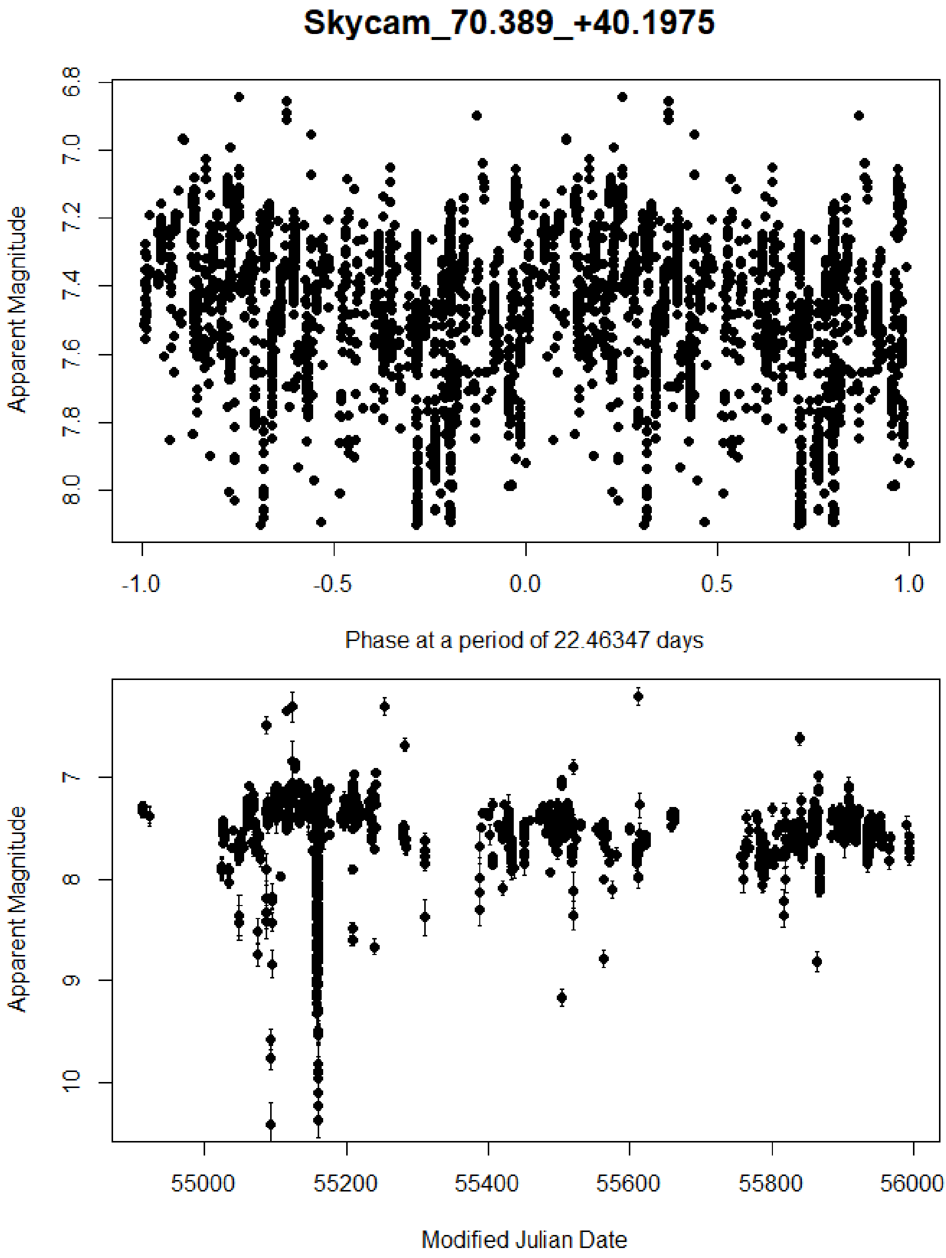


FIGURE 7.36: Folded light curve of the semiregular variable [Skycam_70.389_+40.1975] classified as a Classical Cepheid by the SkycamT classification pipeline. The top plot shows the folded light curve of this variable with a poorly defined period of 22.46 days and the bottom plot shows the raw light curve of this semiregular variable with long-term irregular variability.

This source is not present in the GCVS suggesting it may be a result of a blending with the nearby long period variable. The period at 11.06 days does not appear to be an especially clear choice for this light curve as there appears to be additional structures in the folded light curve separate to the main sawtooth signal. The raw light curve of this source reveals a long period variability in the light curve which also indicates that this variable candidate might be a result of blending with the nearby bright variable star.

The candidate variable source [0691-0647659] is classified as a RV Tauri variable in the SkycamT pipeline with a probability of 0.3192, a GRAPE period of 62.34 days and an anomaly score of 1.778. The Semiregular probability is higher than the RV Tauri probability but as the cut-off threshold for the RV Tauri is substantially smaller than the Semiregular cut-off threshold, the RV Tauri classification becomes dominant. The anomaly score indicates that this source is unlike the RV Tauri variables in the training data. Gaia classifies this source as a Mira or Semiregular variable star and the ASAS classifier identifies this source as a Long Secondary Period variable with a period of 811.514 days. Figure 7.38 shows the folded light curve of this variable source at the GRAPE period of 62.34 days and the ASAS period of 811.514 days. The GRAPE period is very clear and the asymmetry of the variations skews the light curve into a RV Tauri-like shape. This object is likely a Semiregular variable star with a period of 62.34 days. Figure 7.38 shows the folded light curve at the Long Secondary Period suggested by the ASAS data. This epoch folding reveals a possible secondary period although it is not clear due to extensive sampling gaps and a period close to the SkycamT baseline. The GCVS identifies this source as a variable star of unknown classification.

The candidate variable source [1088-0133551] is classified as a fundamental mode RR Lyrae variable in the SkycamT pipeline with a probability of 0.4064, a GRAPE period of 0.575 days and an anomaly score of 2.817. This anomaly score is relatively large and close to the outlier region discussed in chapter 7 and the object only just surpassed the cut-off threshold required for classification. Gaia classifies this source as a Mira or Semiregular variable and the ASAS classifier identifies this candidate variable as a Semiregular variable with a poorly defined period producing an alias of 0.995 days in the ASAS period estimation. Figure 7.39 shows the folded light curve of this object at the GRAPE period of 0.575 days showing a sawtooth signal with significant ‘clumping’ due to the sampling of this light curve. This figure also shows the raw light curve of this source which exhibits poorly sampled non-periodic variability over longer timescales which is not characteristic of RR Lyrae variables. This variable star is likely a SRb type Semiregular variable with a poorly defined period which has produced an alias result due to the diurnal alias. This is a good example of a source which can benefit from a secondary classification model trained on aliased light curves as the 0.575 day

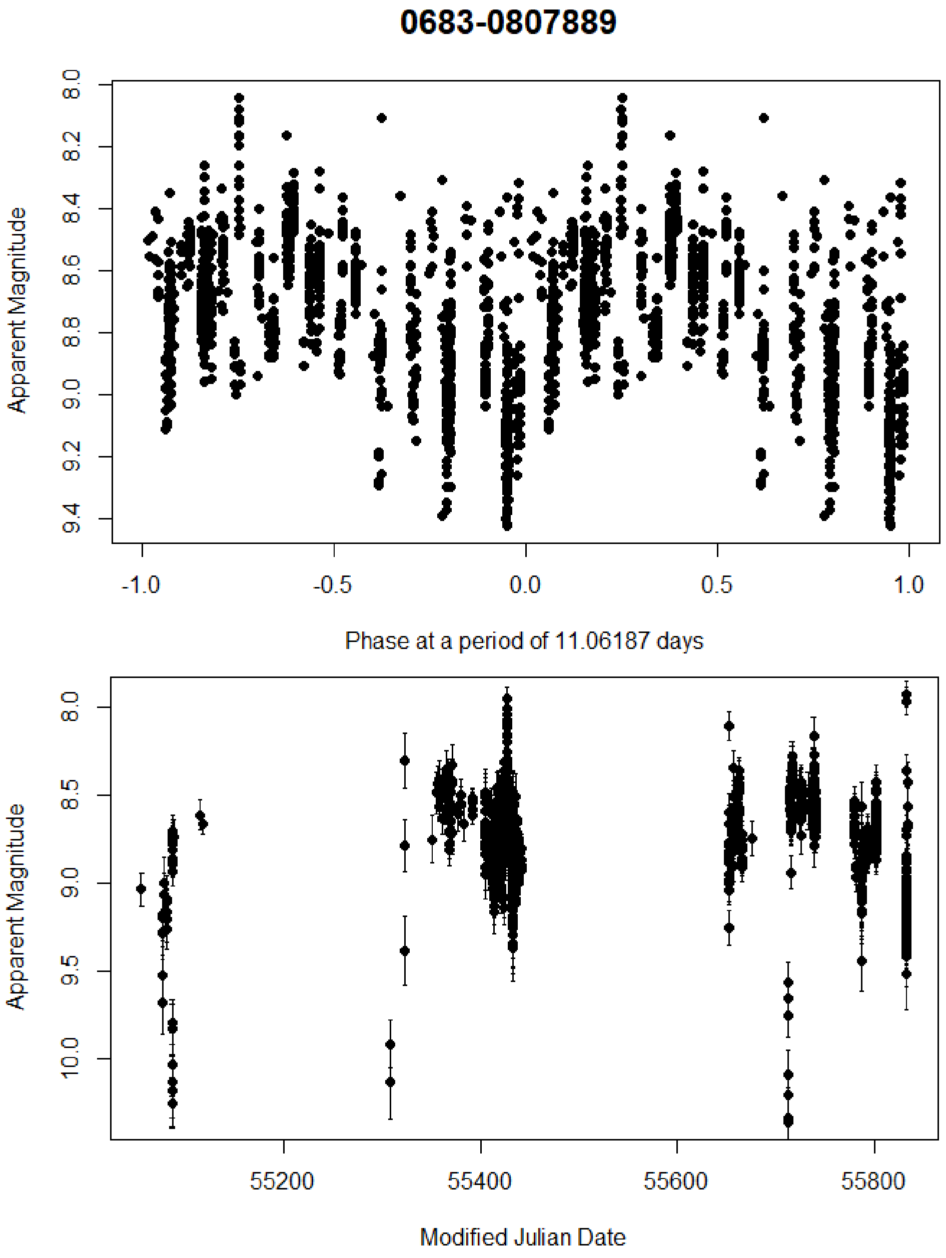


Figure 7.37: c

lassified as a Chemically Peculiar star by the SkycamT classification pipeline.]Folded light curve of the candidate variable [0683-0807889] classified as a Chemically Peculiar star by the SkycamT classification pipeline. The top plot shows the folded light curve of this variable with an unconvincing period of 11.06 days and the bottom plot shows the raw light curve of this candidate variable with long-term irregular variability potentially caused by blending with a nearby long period variable star.

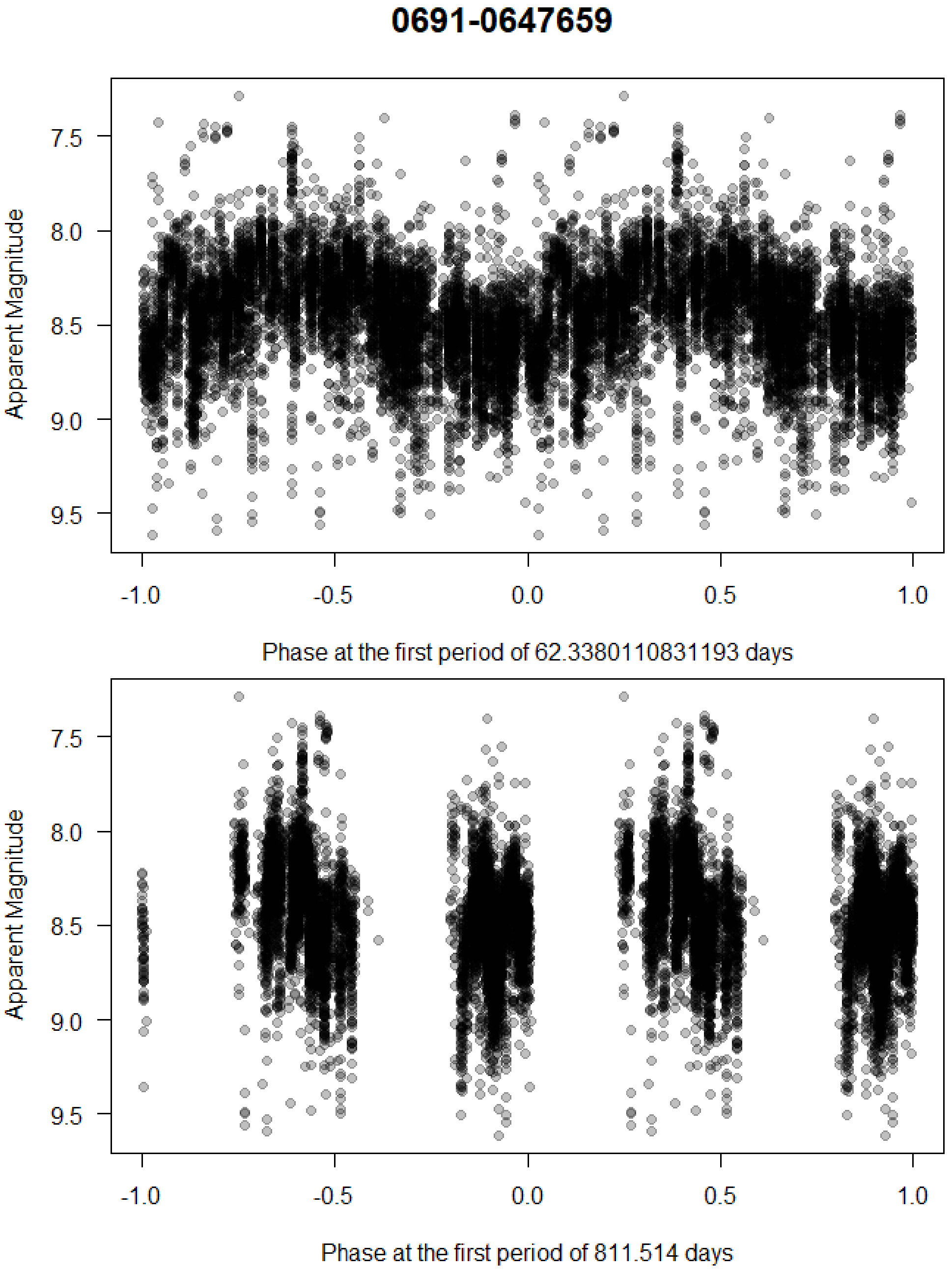


FIGURE 7.38: Folded light curve of the candidate variable [0691-0647659] classified as a RV Tauri variable by the SkycamT classification pipeline. The top plot shows the folded light curve of this variable with a GRAPE period of 62.34 days and the bottom plot shows the folded light curve of this candidate variable at a potential long secondary period of 811.514 days. This source is likely a semiregular variable star possibly with a long secondary period.

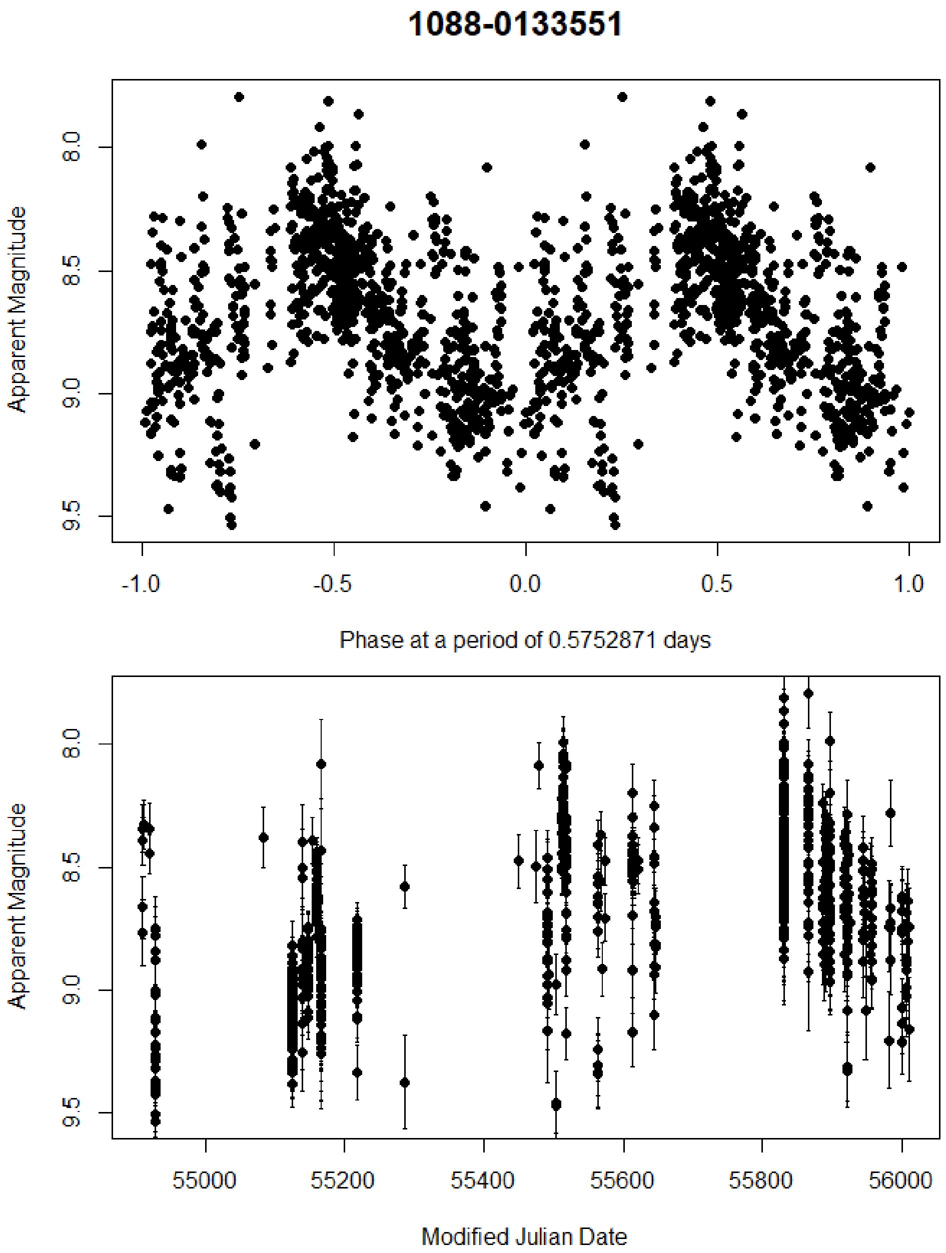


Figure 7.39: Folded light curve of the candidate variable [1088-0133551] classified as a fundamental mode RR Lyrae variable by the SkycamT classification pipeline. The top plot shows the folded light curve of this variable with a GRAPE period of 0.575 days and the bottom plot shows the raw light curve of this candidate variable revealing possible irregular long term variability indicating it is may be a SRb type Semiregular variable star.

GRAPE period suggests a shorter period variable class for the hit and multiple trained classification model.

Ultimately, most of the true variable sources with magnitudes brighter than +12 have been successfully detected although the period is often left undefined in the main catalogues such as GCVS. The classifications can also be too general such as ‘long period variable' and many sources are still listed as candidates. The SkycamT data can be used to constrain the period of these bright variables due to the increasing length baseline of the survey as it continues to collect data. The short period variables appear to be extremely unreliable due to the strong diurnal aliasing in the data. Further corrections must be applied to the diurnal feature such as airmass extinction to improve the performance of the classification pipeline for light curves with fewer than one day periods.

Chapter 8

Conclusions and Future Work

In this chapter the conclusions of this thesis are discussed. The aims and objectives of this research was the development of novel feature extraction methods and machine learning models to produce a Skycam automated variability classification pipeline designed to process the noisy and variable cadence data as effectively as those methods designed for other surveys. The analysis of the light curve data was required to determine the main difficulties to the production of the completed pipeline and the bias-free selection of training light curves. This analysis also suggested possible alternative approaches using representation learning to exploit the useful components of the data. The final result of the pipeline is a catalogue of variable candidates identified and classified into 12 detectable variable star types using statistics robust to the Skycam data systematics, which can be exploited for future scientific study. The chapter ends with a consideration of possible future work directions to improve performance in specific areas of difficulty in the pipeline and to continue to tune the machine learning models for a higher quality set of results from the final classifications.

8.1 Conclusions

There are a substantial number of variable star classes due to photometric changes from internal atmospheric changes driving pulsations to stellar rotation, with stars with surfaces of uneven brightness, and orbital motion causing stars to obscure other stars. These different classes produce different shapes of light curves due to the various timescales and processes of the underlying astrophysics. Many of these classes have subtypes which describe relatively minor differences within a class. These subtypes often have fuzzy boundaries as they are representations of a continuous distribution of variable sources (Percy, 2008). This can be seen in the misclassifications between the three eclipsing binary subtypes, the β Persei subtype, the β Lyrae subtype and the W Ursae Majoris subtype with the first two subtypes being the hardest to distinguish (Malkov et al., 2007; Hoffman et al., 2008).

The development of this pipeline required the implementation of a robust period estimation method as the period is a very important feature and is a dependency of many other features. The investigation of the literature revealed there are three primary groups of period estimation measure in astronomy, the frequency domain algorithms using Fourier decomposition of the time-series, time domain methods identifying patterns in the positions of similar data points and epoch folding methods where the light curve is phased around a candidate period and used to determine a performance measure of the phased data (Larsson, 1996). The Lomb-Scargle based frequency domain methods were the best performing on a subset of SkycamT variable cross-matched light curves (Lomb, 1976; Scargle, 1982). There is potential in the epoch folding methods yet they are plagued by the poor noise and unusual cadence of the Skycam data. We also analysed the bands method for the subsetting of the frequency space into useful trial frequencies and found that the frequency space could be reduced from 200,000 trial frequencies to fewer than 10,000 and still maintain a correct period recall rate of greater than 0.95. This is a useful result but still substantially inferior to the EROS-2 light curves which obtained these results from under 1,000 returned trial frequencies (Protopapas et al., 2015).

Many of the epoch folding methods have computational runtimes which necessitate the use of the bands method. Unfortunately, these methods are also plagued by a frequency-dependent noise continuum. This is due to correlated noise in the light curves. To attempt to address this difficulty we proposed a method of using the bands method to determine likely 'uninteresting' frequencies in the frequency spectrum. These frequencies are assumed to be noise dominated and are used to quickly and approximately fit this continuum allowing the peaks due to signals to become detectable. Most of these algorithms were shown to have a quadratic complexity with the number of light curve data points (VanderPlas, 2017). This is acceptable for most surveys as they have light curves with a few hundred high quality data points. However, for the SkycamT light curves there are enough with thousands of data points to significantly slow down the processing. Subsetting the data points is not a good or reliable solution as the light curves are noisy enough that every data point matters for the required signal to noise to consistently identify the true period.

The proposed solution to this problem is to replace the frequency spectrum grid-based method in traditional periodograms with a genetic algorithm optimisation. The genetic algorithm uses a Bayesian Generalised Lomb-Scargle (BGLS) fitness function, a variant of the Lomb-Scargle periodogram which performed well on the SkycamT cross-matched dataset (Mortier et al., 2015; Mortier and Cameron, 2017). The evolution of the candidate periods allows for an exploration of the period space from which it selects a number of candidate periods. This optimisation method is shown to have a longer runtime for the light curves with 100 to 200 data points and equivalent runtimes for 200 to between 500 and 1000 data points before becoming the fastest method over 1000 data points. This is a result of the substantially lower number of required candidate period computations for optimisation compared to a traditional periodogram although the genetic algorithm does have some other computational overhead such as the generational update method.

The method also introduces the Vuong closeness test as an alias correction method (Baluev, 2012). This information theoretic test compares the performance of two fitted sinusoidal models to select if the candidate period is better fit by an aliased or multiple/submultiple period. This method results in improved performance over selecting the period with the best fitness function for both the genetic algorithm and periodogram optimisation methods. The improved runtime of the genetic algorithm optimisation is complimented by equivalent to superior performance on the correct period estimation of a set of synthetic light curves with well-known true periods depending on how well modelled the light curve is by the BGLS fitness function. The period estimation performance was matched by a periodogram with a follow-up fine tuning optimisation around the top candidate periods from the periodogram at a cost of even longer runtimes. The completed period estimation method is optimised for time-series with many data points of poor sampling and high noise. There are potentially multidisciplinary fields such as climate science where the method may be well suited.

The development of a robust period estimation method for the SkycamT data allows the computation of sets of features identified in the literature to be good descriptors of important aspects of light curve variability. Many of these features were found to be virtually worthless for the Skycam light curves, corrupted beyond usefulness by the noise component of the data. The introduction of representation learning provides machine learning algorithms which learn to extract the components of the data least corrupted by the noise and therefore the strongest at discriminating between the different classes in the training data. The first approach of encoding the light curves into a two dimensional image-based representation prior to machine learning feature extraction showed promise but was ultimately inferior to the previous features (McWhirter et al., 2017). This is a limit of the machine learning models applied and returning to this approach with additional methods could potentially yield important results. After the development of this method an alternative approach to the image-based data encoding has shown substantially more promise (Mahabal et al., 2017).

The second approach was to interpolate the epoch-folded light curve using a parameterised model designed to fit variable light curves of sinusoidal and non-sinusoidal shape. The selected parameterised model is the PolyFit algorithm (Prsa et al., 2008; Paegert et al., 2014). Employing a novel improvement on this method designed to fit noisy light curves using a genetic algorithm on the PolyFit model, a piecewise polynomial chain applied to the binned epoch-folded light curves using a genetic algorithm and regularised polynomial regression, the light curves are reduced to a set of 99 interpolated magnitudes. These interpolated magnitudes, assuming the period estimation selected a good period for the epoch folding, represent the light curve shape with a large quantity of the noise removed. As a result, the original variability indices computed on the interpolated data have substantially improved statistics for discriminating different shapes of light curve. This technique is further extended using a Principal Component Analysis to extract a number of uncorrelated features from a large set of light curves which represent different aspects of the light curve shapes. This is an interesting result as many of the light curves in the training set were of low quality yet had minimal effect on the performance of the analysis as the model just replicated the shape of the epoch-folded training object with little response to whether it appeared to be a true variable star. The second Principal Component was found to perform well in discriminating the short period eclipsing binaries from δ Scuti and RR Lyrae pulsating variables. This feature is found to be heavily anti-correlated with the interpolated skewness feature suggesting it is a form of 'robust skewness' which is less influenced by the noise and sampling of the light curve. Training machine learning models on this set of PolyFit features performed reasonably well and are only a few percent off the overall final variable star classification model used in this pipeline.

With an improved set of training features produced by the representation learning methods, these features were calculated for a set of cross-matched variable stars for the training of machine learning classifiers. Prior to the calculation of these features, the histogram of estimated periods indicated an overabundance of light curves at the yearly period and its submultiples. This was found to be a result of a seasonal trend in the data. A modification of the Trend Filtering Algorithm (TFA) is used to fit sinusoidal yearly models to templates of the seasonal variability produced by large sets of interpolated, binned, mean aggregated light curves. This utilises the assumption that the systematic trends should occur over many of the light curves. The initial application of this method was found to be a poor fit as the phase of the yearly trend was a function of the sky position of the light curves suggesting it is related to yearly changes in sky brightness in different parts of the sky. This was incorporated into the templates by partitioning the sky into 64 regions using right ascension and declination and computing a template trend model for each region requiring a 64× larger training dataset. This partitioning improved the fit performance of the yearly trend sinusoidal model for the accurate filtering out of this trend. Plotting the histogram after this trend removal demonstrates the overabundance has been eliminated without creating an under-abundance relative to the expected number of variable sources in the period bins.

A σ–k clipping algorithm is also used to remove outliers dependent on the specific light curve. This method shows there is a bimodal distribution of light curves which have peak period estimation performance at two different values of k, $k = 3.5$ and $k = 5.5$. The possible cause of this bimodality is the light curves with and without significant cloud-related outliers as cloudy data points will need a smaller value of k for reliable removal without damaging a true variable signal. Colour is also determined to be a useful feature yet only 75% of the light curves have been cross-matched to a catalogue to input the B-R colour feature. The remaining 25% requires an imputed colour and therefore a Random Forest based regression method is applied to predict and impute the colour of the unknown sources using the common features of the sources with known B-R colour. Due to the computational effort required to compute this imputation across the whole database, the sky is partitioned into 25 percentile bins where each bin has approximately a similar number of sources for the training of the imputation models. This is not a recommended procedure and is used purely to test the initial set of SkycamT light curves. A better method of determining the colours of the detected sources must be produced.

The cleaned light curves generate a set of 109 features of which 35 are variability indices not requiring a candidate period for their computation. The full set of features is used to select a set of 859 variable light curves of 12 classes to train a machine learning variable star classifier. Of the classifiers tested, the Random Forest performed the best with the Support Vector Machine using a radial basis function kernel slightly behind. The neural network has the worst performance. This agrees with the results produced from a similar set of features on OGLE and Hipparcos light curves although with a smaller number of classes and worse performance due to the much lower quality of the SkycamT data compared to these two surveys (Richards et al., 2011b; McWhirter et al., 2016). The low performance of the neural networks is likely a result of neither Richards et al. nor our feature vectors being rescaled to reflect a normal distribution. The training is recomputed with rescaled features and the neural network performance improved significantly in regards to the AUC measure although the F1 score still show that the models had the worst classification performance. This discrepancy suggests that the probability vectors of the neural networks could use calibration. A model is also trained on the light curves with aliased estimated periods in an attempt to determine confident classifications from the poorer quality data although the performance was extremely poor. The Random Forest model was selected as the final pipeline model with hyperparameters of ntree = 500, m_{try} = 40 and nodesize = 75 with a mean F1 score of 0.533 ± 0.070. Analysing the importance of each of the features in the Random Forest models indicates that the dominant variable star discriminators are period, colour, amplitude and skewness. These results agree with the models trained on OGLE and Hipparcos (Richards et al., 2011b). It is exciting to see that the amplitude and skewness measures selected by the machine learning classifiers are from the novel set of PolyFit features proposed in this thesis.

The confident variable star classifications are used to train a variability detection model using the Random Forest classifier on a set of randomly selected non-variable candidates. This was performed for a dataset consisting of both aliased and matched period light curves and a training set of just the matched period light curves. The model with the aliased light curves has poorer performance which, like the aliased variable star model performance, shows that the aliased objects likely have many objects at the sidereal day spurious period. As a result, both the aliased variable star classifier and the full dataset variability detection model are discarded. There is a risk of unwittingly selecting variable candidates in the randomly selected non-variable dataset. To compensate, the non-variable training dataset is randomly selected a number of times and the family of variability detection models trained by the method acts as an aggregated Random Forest. As the Random Forest is a collection of weak-classifier decision trees, adding additional Random Forests is just a linear combination of these models. The final variability detection model has hyperparameters of ntree = 1500, m_{try} = 4 and nodesize = 25 with a mean F1 score of 0.799 ± 0.007. Using this model, with a probability cut determined using the best performing sensitivity and specificity of the variable star class, 103,790 variable candidates are selected. This is 17.6% of the full dataset which is a slightly higher estimation the 10% of the Northern Sky Variability Survey (NSVS) most likely due to the increased systematics in the Skycam light curves producing 'phantom' variability (Shin et al., 2012). Half of these candidates (51,129 sources) have had a full set of period and PolyFit features computed and classified using the variable star classification model. The probabilities have been calibrated to the posterior probability of the classes and probability cuts established from the Index of Union of the training data. The bright variable source catalogue will be made available for scientific follow-up.

8.2 Future Work

In the processing of the light curves, there has been a number of systematics which have resulted in poor performance in the reliable period estimation of variable stars and the detection of variable light curves. The trend removal method is used to remove a subset of these systematics in the form of yearly sinusoidal variability in a large set of the SkycamT light curves and was reflected in the elimination of the overabundance of light curves with yearly periods as shown in figure 7.2. A close inspection of this histogram reveals that there is a second overabundance near 30 days. This is a result of the lunar month causing sky background magnitude variations over a 29.46 day period. It would be advantageous to expand the trend removal method to remove this overabundance and therefore reduce the false positives due to lunar-induced light curve variability. There are other systematics generating non-astrophysical variability in the light curves as evidenced by the overabundance of variable sources in the Skycam data. The trend removal method makes use of a correlation test on interpolated harmonic models to determine if a light curve should be prewhitened. This can occasionally fail due to a poor interpolated fit due to seasonal gaps in the data therefore the method should be improved to ignore the gaps by removing the interpolated trend template and test light curve data from these unsampled regions. The trend removal method identified a sky position dependence on the phase of the seasonal trend. It is possible there is also a magnitude dependence on the trend which should be investigated to further improve the accuracy of the light curve prewhitening although this will result in the requirement of additional training light curves.

The systematics also have an undesirable performance cost in the period estimation of known variable stars for the training data. Figure 7.8 shows the base-10 log-log plot of the GRAPE estimated period against the catalogue period of this dataset of SkycamT light curves. As well as demonstrating the spurious lunar month results mentioned previously, there are other unusual structures in the plot such as a square-shaped region of heavy misclassifications between 0.1 days and 10 days on both axes. This is likely a result of the variable Skycam cadence generating undefined aliases with an undetermined cadence relationship. The analysis of the structure of these aliases would be extremely beneficial to correcting these artefacts and improving the purity of the variable light curve sample. More powerful machine learning methods might be able to model the systematics directly through training on a large set of Skycam light curves and treating the systematics like a background to the foreground variability due to variable stars. This may be possible using matrix decomposition and Principal Component Analysis techniques (Mahabal et al., 2017).

Some objects also exhibit a clear alias at periods unrelated to their true periods by an identifiable spurious period. A good example is the light curve of the variable star [0634-0892788] as shown in figure 8.1. GRAPE detects a period of 10.05 days for this object and the variable classification model predicts it is a classical cepheid with a 0.857 probability. The light curve exhibits a clear sawtooth shaped light curve with a period in the range of the cepheid variables. The only feature which indicates that it may

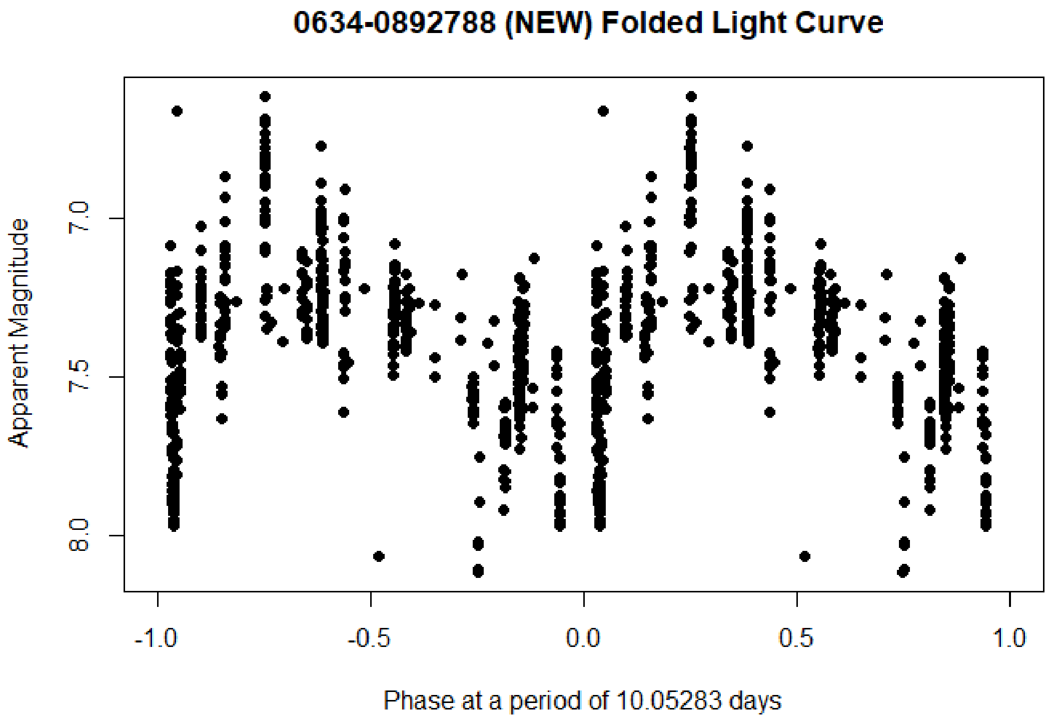


FIGURE 8.1: Folded light curve of the variable star [0634-0892788] folded at a GRAPE detected period of 10.05 days. The SkycamT classifier identifies this object as a classical cepheid with a sawtooth light curve although it is actually a small amplitude red giant, a class not present in the trained classification models.

not be a cepheid is the catalogue B-R colour suggesting it is a red star, outside of the normal yellow for this class of star. Searching the All Sky Automated Survey (ASAS) data confirms this star is a small amplitude red giant, a class not trained in the classifier and as such, the classification appears reasonable and at 10.05 days, this would be the shortest period small amplitude red giant known. On closer inspection of the ASAS light curve, the red giant has a 27.7 day period, much more typical for this class of star. There is no indication of the 10.05 day signal detected by GRAPE in the SkycamT light curve. Analysis of the light curve shows that it does have a 27.7 day periodicity as well with a signal of similar strength to the 10.05 day signal. Using the alias equation 3.20 (Heinze et al., 2018), a search for the spurious period was conducted with no candidates found. It is likely many other light curves are exhibiting similar artefacts which can lead to incorrect conclusions and the cause must be identified such as a possible beat period with a systematic variation. There is an additional airmass (brightness reduction due to atmospheric extinction) component in the image reduction data products utilised by this automated classification pipeline. A new reduction has been produced with improved computation of airmass corrections and a high priority task is to migrate the classification pipeline to operate on this new and improved dataset.

Colour imputation is a poor method of generating colour information for the machine learning classifiers as they assume that the imputed data is of the same quality as a truly measured colour quantity. This flaw can only be solved by obtaining true colour information from improvements to the instruments or superior source matching to the US Naval Observatory catalogue, or an alternative catalogue. It would be advantageous to collect data in multiple bands using the Skycam instruments but this would come at extra cost and complexity and may not necessarily be easily applied to previously gathered data. An alternative solution is improving the cross-matching to the survey data. The cross-matching to the catalogues is performed using a rudimentary distance minimisation. Due to the pixel scale of the SkycamT camera, there may be multiple sources contributing light to the photometric data points in an identified source. In fact there are sources in the database where a collection of sources with a small number of data points have been generated as the statistics of each of the data points results in a rejection of the hypothesis that the data points are from the same source. There are also examples of a false positive in this hypothesis producing blended light curves of which a number can be seen in figure 7.26 as they produce clear signatures in the variability index features. The presence of multiple catalogue objects in the same SkycamT pixels means that the distance minimisation method is not ideal as the dominant star may not be the closest one to the coordinates of the detected source. Richards et al. proposed a solution using a machine learning classifier to determine the best catalogue object for a detected source using the statistics of the source data (Richards et al., 2012). Using a number of features such as colour and magnitude to describe the detected source, a machine learning classifier is trained on a set of manually confirmed cross-matches to produce a model which predicts the probability of a catalogue object producing the detected source. If a catalogue object exists with greater than 0.5 probability of matching the detected source it is selected as the candidate (and if multiple, the highest is selected). If no catalogue object achieves this probability threshold, they are all rejected and the source is considered unmatched. Implementation of a similar method on the Skycam data will improve the light curves produced by the image reduction pipeline by reducing blended light curves and also ensure that the colour of the cross-matched catalogue object is appropriate for the training of the variable star classification models.

The GRAPE period estimation produces the best performance on the Skycam light curves but could benefit from some further development. The current clustering algorithm used to identify independent candidate periods is k-means clustering. This is the only clustering method which has been employed for this task and it is strongly recommended that additional methods be investigated especially since k-means clustering is one of the simpler methods. This must be done with caution as a significant computational cost in the clustering operation will substantially increase the runtime of the

GRAPE method. The initial GRAPE method had an additional component designed to filter out spurious periods using a Gaussian filter which modified the fitness function to disincentivise the genetic evolution from proceeding at these values. Unfortunately, it proved extremely difficult to reliably control with static parameters for the Gaussian function height and width. There is still potential in this approach and investigating an approach to produce flexible Gaussian functions for individual light curves as required would be a potent addition to this method. Use of machine learning algorithms for this task is a strong possibility.

GRAPE is currently specifically designed to identify purely periodic phenomena. There are many quasi-periodic and semiregular variable objects in astronomy. These objects suffer a significance degradation in a method such as this as amplitude and phase changes cannot be mapped across the baseline. GRAPE uses only a single dimensional parameter space despite genetic algorithms being highly functional at the exploration of high dimensional space (Charbonneau, 1995; Rajpaul, 2012). This method can be extended using a multi-dimensional parameter space, which can stack different combinations of the data points into multiple period spaces simultaneously and use the genetic methods to optimise to a single answer across all the combinations. Alternatively, depending on the combinations of data points stacked by the algorithm, this single answer could instead be a set of results expressed as a function of amplitude and phase allowing for quasi-periodic signals to be expressed in the output. This multi-dimensional approach can also be used to implement a multi-band light curve variant of GRAPE which can determine a set of candidate periods from their simultaneous performance at successfully fitting data in multiple bands. This is accomplished through a modification to the fitness function where the initial single-dimension statistic is weighted against the other dimensions as the chosen candidates should provide satisfactory fits in every band. The Vuong closeness test can also be modified to evaluate multi-dimensional hyperplanes constructed from multiple sinusoidal models in each dimensional band. This method is currently in development and is named ‘Bunch of GRAPES’. Expanding genetic algorithms into multiple period spaces has very interesting applications in upcoming next generation astronomical surveys.

This pipeline makes use of many features drawn from a variety of research projects as well as machine learning methods. There are still more features which have not been investigated such as the slotted autocorrelation function and the quasi-stellar object features which have been shown to be of high importance in other classification models (Richards et al., 2011b). There are also some limitations with the features that have been employed in the pipeline. The Stetson-I, J, K and L features have been shown to be important in their current implementation however, they are sensitive to outliers and poorly sampled data (Shin et al., 2009, 2012). As these features are formulated from pairs of consecutive data points, there are a number of variants which limit the allowed pairings by the magnitude difference between the data point pairs named the Stetson-I, J, K and L clip features and the difference between the time instants of the data point pairs named the Stetson-I, J, K and L time features. These features can be used to eliminate light curves which trigger a high or low response due to the distribution and noise of the light curves instead of the desired astrophysical signal. The optimal segmentation of the light curves can be determined using Bayesian Block Representations (Scargle et al., 2013). Another subset of variability features which demonstrates an unusual artefact is the percent amplitude and mean variance features. These features determine the variance of a light curve proportional to its mean magnitude. These features begin to generate infinities as the mean magnitude approaches zero and can go negative for negative magnitudes. This is not a problem for the surveys used in the development of these methods as they have greater depth than STILT meaning that bright objects near zero saturate the detector and are removed. This does create problems with SkycamT as it can photometrically observe objects up to -2 in the R band. A possible, yet easy to implement solution is to add a constant to the mean magnitude in the formulation of these features so their statistics act similar to the surveys with sources up to ten magnitudes fainter than SkycamT sources.

The representation learning features have been identified as highly important measures of variability such as the PolyFit interpolation results. The utilisation of machine learning techniques to learn the best features from a dataset for a given task automatically and without bias cannot be understated. The interpolated PolyFit features are extremely useful the classification of the variable stars by providing features which are more robust to systematics and noise. The variability indices recalculated on the interpolated light curve were considered of higher importance by the machine learning classifiers. Using this interpolation procedure, additional variability indices such as the Median Absolute Deviation (MAD) should be computed to replace the very noisy and correlated features. The image-representation approach did not produce a robust set of features primarily due to the extremely limited machine learning classifiers employed in the research. The utilisation of more powerful machine learning techniques such as Convolutional Neural Networks is strongly recommended to better process image based data as translational and rotational invariance can be described by the convolutional layers. As well as more capable machine learning algorithms, the technique of encoding the light curves into an image representation is of equal importance to this task. The epoch-folded representation used here is limited as it must have a pre-computed period for the light curve which limits the approach on poor quality data similar to the performance of the PolyFit interpolated features being strongly tied to the performance of the period estimation. Recent research has proposed an alternative approach by computing the magnitude and time instant differences of pairs of data points somewhat reminiscent of the Stetson variability indices (Mahabal et al., 2017). Both Convolutional Neural Networks and Random Forest classifiers demonstated good performance on these image representations. The application of this approach to the SkycamT light curves, combined with the additional component capable of isolating the systematic components, show great potential and is strongly recommended. This thesis has heavily discussed the research areas of period estimation and machine learned classification individually. The intersection between these methods through employing techniques such as this can produce the next generation of period estimation algorithms which are computationally efficient and very robust to survey systematics whilst making no prior assumptions on the shape of any potential periodic variability.

The learned features from the application of the PCA algorithm to the training set of known variable light curves peformed well and automatically learnt properties of the light curves such as skewness. The PCA method is quite simplistic given the breadth of available clustering and unsupervised learning algorithms. It is limited to linear interactions between the input interpolated magnitudes which restricts the complexity of patterns the method can establish. Methods such as t-distributed Stochastic Neighbor Embedding (t-SNE) offer the ability to propose a subset of dimensions where the data is projected with reduced dimensionality using non-linear relations (van der Maaten and Hinton, 2008). Further to this, the interactions can be modelled directly using a deep learning architecture, such as non-linear PCA and deep Autoencoders which are also capable of determining similar non-linear interactions (Baldi, 2011). By applying these more powerful techniques, the PolyFit model can better model the noisy light curves with a set of features which describe the important properties of the light curves required for classification.

The variable star classifier performed well considering the limitations of the data and was heavily tested with multiple classification methods. There is one major uninvestigated component of the models due to the assigned task. The model was trained to classify the variable light curves into a set of 12 classes. In chapter one, the variable star taxonomy is discussed and it follow a hierarchical classification structure. This hierarchical structure can be implemented into the variable star classification task through the computation of a *superclass* model such as between pulsating variables, eclipsing binaries and rotational variables. The classifier can then focus on the features which distinguish these major groups of objects before further models are trained to specifically classify objects assigned to a given superclass into the appropriate subclasses. This has been employed in previous studies to achieve superior automated classification performance at the risk of 'catastrophic' misclassifications where an object is predicted as the wrong superclass and therefore cannot be classified into the correct subclass (Richards et al., 2011b). Many of the misclassifications are of more minor error such as confusion between the eclipsing binary subtypes. There is also potential for the use of fuzzy logic based methods as the classification boundaries of the various variable star types are not hard limits and reflect the nature of these objects as a continuous distribution of astrophysical phenomena with a set of continuous properties. The capability of meta-classification models is also worth investigating as these models use a set of high performance models trained to perform extremely well at limited classification tasks named *experts* where the meta-classification model guides the selection of specific experts for the given classification task (Pichara et al., 2016).

The variability detection models also require additional investigation. The method applied here focused on the Random Forest classifier due to its strong performance on the variable star classifier. It is highly recommended that other machine learning methods be tested with this data as the minimisation of false positives whilst maintaining good true positive performance is a must to maintain the minimum computational effort for maximum reward from this pipeline. The investigation of the selection of a good training set indicated that the best performance was from the GRAPE period-matched variable light curves combined with multiple subsets of randomly selected non-variable candidates. The investigation of the rates of training set corruption due to the integration of variable candidates is an important task which was not fully explored in this thesis. Additionally, the best performance might be produced through the combination of the individual models produced by the different training sets. The multiple models can be treated as an ensemble method where the overall classifications are a linear combination of votes from the individual models, similar to how the Random Forest predicts using the votes from all the individual classification and regression trees. The implementation of an ensemble method for the variability detection models is strongly recommended for the future development of this pipeline.

The machine learned models trained and employed in this pipeline are extremely static. Upon the completion of the training phase they are no longer updated using additional data processed by the pipeline. This is a waste of the available resources and the capability of machine learning to continue to learn and improve from new experiences. Self-training, also known as co-training is the process of training a classification model, using it to classify a set of unknown data and then utilising the unknown data which has been confidently classified and verified as of sufficient quality to produce an updated model adding this new knowledge to the pipeline. This can assist in the training of under-represented classes where the model may overfit the specific training examples resulting in lower classification probabilities on objects of the same class which do not precisely match the training object feature vectors. The technique of Active Learning has also been successfully applied to variable star classifiers resulting in a drop of error rate of up to 20% (Richards et al., 2011a, 2012). In Active Learning the classifier will request further information on unknown objects which lie close to important decision boundaries. Through an expert manual classification of a minimal subset of the dataset, the overall automated classification performance can be substantially improved. This can also be implemented using a citizen science based method where the active learning objects can be uploaded to an online repository for manual classification by a large set of individuals. These classifications can then be used to update the training probability vector using manual confidences for the active learning examples allowing the machine learning model to produce state-of-the-art results for automated classification. A similar technique was used in the training of Convolutional Neural Networks for galaxy classification using Galaxy Zoo for unprecedented performance capability (Dieleman et al., 2015). The continuous improvement of the machine learning models in this pipeline will improve the purity of the variable light curve sample and the accuracy of the final variable star classification produced in the final data products.

The class imbalance of the training sets for these classifiers required the use of an oversampling technique. This technique results in the machine learning classifiers correctly using the performance on every class to determine the optimal model. Oversampling is extremely simplistic and can result in training issues such as those mentioned above with an overemphasis on specific feature vectors resulting in poorer performance on objects of the same class with small differences in their feature vectors. There are alternative methods for balancing a dataset such as the Synthetic Minority Over-sampling Technique (SMOTE) which uses imputation on the features of the minority class to generate a set of additional class examples which represent the underlying statistical distribution of the features given a particular class whilst producing unique feature vectors (Chawla et al., 2002). The unique feature vectors then prevent the machine learning classifiers from overfitting a given minority class feature vector whilst still learning a good representation of the class from the true training set objects. The employment of SMOTE or a similar technique may prevent some of the difficulties with classifying the minority classes such as the RV Tauri variables.

Using the proximity measure computed by the Random Forest variable star classifier an anomaly score has been implemented for the non-training set light curves. Using the classifications of a set of 51,129 light curves, the distribution of these anomaly scores has been determined. Whilst these anomaly scores are available in the data products, the anomaly score threshold which defines the outlier light curves is not established. Richards et al. employ a method using cross-validation to determine the misclassification error rate for their machine learning model for multiple anomaly score thresholds where each light curve with an anomaly score above the candidate outlier threshold as an error. This process is repeated until a threshold is determined which results in minimal performance decrease indicating any light curve with an anomaly score higher than this value is often misclassified regardless (Richards et al., 2012). Implementation of a similar method in the Skycam automated pipeline would assist in the identification of interesting outlier light curves for potentially novel sources.

The final recommendation for future work on this automated pipeline is the deployment of front end software for scientific investigators and others to inspect the data products from the pipeline. Currently the results are output into a table of classifications and feature vectors which are not particularly user friendly. The deployment of a software interface will allow users to inspect the light curves and study the associated feature vectors by performing a variety of queries and sorting operations on the classification table. An example of the form of interface well suited to this project is the ASAS classification pipeline results from Richards et al. found online at http://www.bigmacc.info (Richards et al., 2012). Use of such a technique will allow active engagement in the STILT project by outside researchers and the general public which has been primarily utilised by the Liverpool Telescope group thus far. The advantages offered by studying the methodology used in the ASAS classification pipeline certainly benefitted the research in this thesis greatly.

Bibliography

Abs da Cruz, A. V. A., Vellasco, M. M. B. R., and Pacheco, M. A. C. (2007). Quantum-inspired evolutionary algorithm for numerical optimization. *Hybrid Evolutionary Algorithms, Studies in Computational Intelligence*, 75.

Aerts, C., Marchenko, S. V., Matthews, J. M., Kuschnig, R., Guenther, D. B., Moffat, A. F. J., Rucinski, S. M., Sasselov, D., Walker, G. A. H., and Weiss, W. W. (2006). Delta ceti is not monoperiodic: Seismic modeling of a beta cephei star from MOST space-based photometry. *The Astrophysical Journal*, 642(1):470–477.

Alcock, C., Allsman, R. A., Alves, D. R., Axelrod, T. S., Becker, A. C., Bennett, D. P., Cook, K. H., Dalal, N., Drake, A. J., and Freeman, K. C. (2000). The MACHO project: Microlensing results from 5.7 years of large magellanic cloud observations. *The Astrophysical Journal*, 542(1):281–307.

Almoznino, E., Loinger, F., and Brosch, N. (1993). A procedure for the calculation of background in images. *Monthly Notices of the Royal Astronomical Society*, 265(3):641.

Althaus, L. G., Córsico, A. H., Isern, J., and García-Berro, E. (2010). Evolutionary and pulsational properties of white dwarf stars. *The Astronomy and Astrophysics Review*, 18(4):471–566.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.

Appenzeller, I. and Mundt, R. (1989). T tauri stars. *The Astronomy and Astrophysics Review*, 1:291–334.

Baldi, P. (2011). Autoencoders, unsupervised learning and deep architectures. *UTLW'11 Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning workshop*, 27:37–50.

Baluev, R. V. (2009). Detecting non-sinusoidal periodicities in observational data using multiharmonic periodograms. *Monthly Notices of the Royal Astronomical Society*, 395(3):1541–1548.

Baluev, R. V. (2012). Distinguishing between a true period and its alias, and other tasks of model discrimination. *Monthly Notices of the Royal Astronomical Society*, 422(3):2372–2385.

Barnard, J. and Meng, X. L. (1999). Applications of multiple imputation in medical studies: from AIDS to NHANES. *Statistical Methods in Medical Research*, 8(1):17–36.

Basterrech, S., Mohammed, S., Rubino, G., and Soliman, M. (2011). Levenberg–marquardt training algorithms for random neural networks. *The Computer Journal*, 54(1):125–135.

Beard, S. M., MacGillivray, H. T., and Thanisch, P. F. (1990). The cosmos system for crowded-field analysis of digitized photographic plate scans. *Monthly Notices of the Royal Astronomical Society*, 247:311–321.

Bedding, T. R. and Zijlstra, A. A. (1998). *Hipparcos* period-luminosity relations for mira and semiregular variables. *The Astrophysical Journal*, 506(1).

Benavente, P., Protopapas, P., and Pichara, K. (2017). Automatic survey-invariant classification of variable stars. *The Astrophysical Journal*, 845(2):147.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Bertin, E. and Arnouts, S. (1996). SExtractor: Software for source extraction. *Astronomy and Astrophysics Supplement Series*, 117(2):393–404.

Bhattacharyya, S., Richards, J. W., Rice, J., Starr, D. L., Butler, N. R., and Bloom, J. S. (2011). Identification of outliers through clustering and semi-supervised learning for all sky surveys. *Lecture Notes in Statistics, Statistical Challenges in Modern Astronomy V*, 902:483–485.

Bijaoui, A. (1980). Sky background estimation and application. *Astronomy and Astrophysics*, 84(1–2):81–84.

Blazhko, S. (1907). Mitteilung über veränderliche sterne. *Astronomische Nachrichten*, 175:327.

Bloom, J. and Richards, J. (2011). Data mining and machine learning in time-domain discovery and classification. *Advances in Machine Learning and Data Mining for Astronomy Chapman & Hall/CRC Data Mining and Knowledge Discovery Series.*

Bopp, B. W. and Stencel, R. E. (1981). The fk comae stars. *Astrophysical Journal, Part 2 - Letters to the Editor*, 247:L131–L134.

Boström, H. (2008). Calibrating random forests. *Machine Learning and Applications, 2008. ICMLA '08. Seventh International Conference on.*

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees.* Monterey Publishing.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3.

Brink, H., Richards, J. W., Poznanski, D., Bloom, J. S., Rice, J., Negahban, S., and Wainwright, M. (2013). Using machine learning for discovery in synoptic survey imaging data. *Monthly Notices of the Royal Astronomical Society*, 435(2):1047–1060.

Butler, N. R. and Bloom, J. S. (2011). Optimal time-series selection of quasars. *The Astronomical Journal*, 141(3):93.

Candès, E. and Romberg, J. (2005). l1-magic : Recovery of sparse signals via convex programming.

Cantu-Paz, E. (2000). *Efficient and Accurate Parallel Genetic Algorithms.* Kluwer Academic Publishers.

Carrier, F., Burki, G., and Burnet, M. (2002). Search for duplicity in periodic variable be stars. *Astronomy and Astrophysics*, 385(2):488–502.

Chan, K. G., Streichan, S. J., Trinh, L. A., and Liebling, M. (2016). Simultaneous temporal superresolution and denoising for cardian fluorescence microscopy. *IEEE Transactions on Computational Imaging*, 2(3):348–358.

Charbonneau, P. (1995). Genetic algorithms in astronomy and astrophysics. *The Astrophysical Journal Supplement Series*, 101:309.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Che, Z.-G., Chiang, T.-A., and Che, Z.-H. (2011). Feed-froward neural networks training: A comparison between genetic algorithm and back-propagation learning algorithm. *International Journal of Innovative Computing Information and Control*, 7(10):5839–5843.

Clarke, D. (2002). String/rope length methods using the lafler-kinman statistic. *Astronomy & Astrophysics*, 386(2):763–774.

Clayton, G. C. (1996). The r coronae borealis stars. *Publications of the Astronomical Society of the Pacific*, 108(721).

Cohen, R. E. and Sarajedini, A. (2012). Sx phoenicis period-luminosity relations and the blue straggler connection. *Monthly Notices of the Royal Astronomical Society*, 419(1):342–357.

Copperwheat, C. M., Steele, I. A., Piascik, A. S., Bersier, D., Bode, M. F., Collins, C. A., Darnley, M. J., Galloway, D. K., Gomboc, A., and Kobayashi, S. (2016). Liverpool telescope follow-up of candidate electromagnetic counterparts during the first run of advanced ligo. *Monthly Notices of the Royal Astronomical Society*, 462(4):3528–3536.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Cox, A. N. (2003). A pulsation mechanism for gw virginis variables. *The Astrophysical Journal*, 585(2):975–982.

Darnley, M. J., Ribeiro, V. A. R. M., Bode, M. F., Hounsell, R. A., and Williams, R. P. (2012). On the progenitors of galactic novae. *The Astrophysical Journal*, 746(1).

Deb, S. and Singh, H. P. (2009). Light curve analysis of variable stars using fourier decomposition and principal component analysis. *Astronomy & Astrophysics*, 507(3):1729–1737.

Debosscher, J., Sarro, L. M., Aerts, C., Cuypers, J., Vandenbussche, B., Garrido, R., and Solano, E. (2007). Automated supervised classification of variable stars. *Astronomy & Astrophysics*, 475(3):1159–1183.

Dieleman, S., Willett, K. W., and Dambre, J. (2015). Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society*, 450(2):1441–1459.

Disanto, A., Cavuoti, S., Brescia, M., Donalek, C., Longo, G., Riccio, G., and Djorgovski, S. G. (2016). An analysis of feature relevance in the classification of astronomical transients with machine learning methods. *Monthly Notices of the Royal Astronomical Society*, 457(3):3119–3132.

Domínguez, A. B., Chini, R., Nuez, F. P., Haas, M., Hackstein, M., Drass, H., Lemke, R., and Murphy, M. (2013). Eclipsing high-mass binaries. *Astronomy & Astrophysics*, 557.

Donalek, C., Djorgovski, S. G., Mahabal, A. A., Graham, M. J., Drake, A. J., Kumar, A. A., Philip, N. S., Fuchs, T. J., Turmon, M. J., and Yang, M. T.-C. (2013). Feature selection strategies for classifying high dimensional astronomical data sets. *2013 IEEE International Conference on Big Data.*

Dubath, P., Lecoeur-Tabi, I., Rimoldini, L., Sveges, M., Blomme, J., Lpez, M., Sarro, L. M., Ridder, J. D., Cuypers, J., and Guy, L. (2012). Hipparcos variable star detection and classification efficiency. *Astrostatistics and Data Mining*, pages 117–125.

Dubath, P., Rimoldini, L., Süveges, M., Blomme, J., López, M., Sarro, L. M., De Ridder, J., Cuypers, J., Guy, L., Lecoeur, I., Nienartowicz, K., Jan, A., Beck, M., Mowlavi, N., De Cat, P., Lebzelter, T., and Eyer, L. (2011). Random forest automated supervised classification of hipparcos periodic variable stars. *Monthly Notices of the Royal Astronomical Society*, 414(3):2602–2617.

Dufour, P., Liebert, J., Fontaine, G., and Behara, N. (2007). White dwarf stars with carbon atmospheres. *The Astrophysical Journal*, 450:522–524.

Dworetsky, M. M. (1983). A period-finding method for sparse randomly spaced observations or how long is a piece of string?. *Monthly Notices of the Royal Astronomical Society*, 203(4):917–924.

Edelson, R. A. and Krolik, J. H. (1988). The discrete correlation function - a new method for analyzing unevenly sampled variability data. *The Astrophysical Journal*, 333:646–659.

Elzo, C., Estevez, P. A., and Kozma, R. (2016). Modeling quasi-periodic lightcurves using neuropercolation. *2016 International Joint Conference on Neural Networks (IJCNN).*

Eyer, L. and Mowlavi, N. (2008). Variable stars across the observational hr diagram. *Journal of Physics: Conference Series*, 118:012010.

Falomo, R., Pian, E., and Treves, A. (2014). An optical view of bl lacertae objects. *The Astronomy and Astrophysics Review*, 22:73.

Fine, T. L. (1999). *Feedforward Neural Network Methodology.* Springer-Verlag.

Flewelling, H. A., Magnier, E. A., Chambers, K. C., Heasley, J. N., and et al. (2016). The pan–STARRS1 database and data products. *arXiv:1612.05243v2.*

Fourier, J. B. J. (1878). *The Analytical Theory of Heat.* Cambridge University Press.

Frescura, F. A. M., Engelbrecht, C. A., and Frank, B. S. (2008). Significance of periodogram peaks and a pulsation mode analysis of the beta cephei star v403 car. *Monthly Notices of the Royal Astronomical Society*, 388(4):1693–1707.

Gautschy, A. and Saio, H. (1996). Stellar pulsations across the hr diagram: Part II. *Annual Review of Astronomy and Astrophysics*, 34:551–606.

Geier, S. (2015). Why do hot subdwarf stars pulsate? *Astronomy in Focus*, 1.

George, S. V., Ambika, G., and Misra, R. (2015). Effect of data gaps on correlation dimension computed from light curves of variable stars. *Astrophysics and Space Science*, 360(1).

Giridhar, S., Lambert, D. L., and Gonzalez, G. (1998). The chemical compositions of the srd variable stars. I. XY Aquarii, RX Cephei, AB Leonis, and SV Ursae Majoris. *Publications of the Astronomical Society of the Pacific*, 110:671–675.

Glass, I. S. and Evans, T. L. (1981). A period-luminosity relation for mira variables in the large magellanic cloud. *Nature*, 291(5813):303–304.

Graham, M. J., Drake, A. J., Djorgovski, S. G., Mahabal, A. A., and Donalek, C. (2013a). Using conditional entropy to identify periodicity. *Monthly Notices of the Royal Astronomical Society*, 434(3):2629–2635.

Graham, M. J., Drake, A. J., Djorgovski, S. G., Mahabal, A. A., Donalek, C., Duan, V., and Maker, A. (2013b). A comparison of period finding algorithms. *Monthly Notices of the Royal Astronomical Society*, 434(4):3423–3444.

Guo, Z., Gies, D. R., and Matson, R. A. (2016). Kepler eclipsing binaries with delta scuti/gamma doradus pulsating components I: KIC 9851944. *The Astrophysical Journal*, 826.

Hall, D. S. (1981). The rs canum venaticorum binaries. *Solar Phenomena in Stars and Stellar Systems*, pages 431–447.

Handler, G., Prinja, R. K., Antoci, V., and Urbaneja, M. A. (2013). The ZZ Leporis Stars: Wind-variable central stars of young planetary nebulae. 18^{th} *European White Dwarf Workshop*.

He, S., Yuan, W., Huang, J. Z., Long, J., and Macri, L. M. (2016). Period estimation for sparsely sampled quasi-periodic light curves applied to miras. *The Astronomical Journal*, 152(6):164.

Heck, A., Manfroid, J., and G., M. (1985). On period determination methods. *Astronomy & Astrophysics Supplement Series*, 59:63–72.

Heger, A., Fryer, C. L., Woosley, S. E., Langer, N., and Hartmann, D. H. (2003). How massive single stars end their life. *The Astrophysical Journal*, 591(1).

Heinze, A. N., Tonry, J. L., L., D., H., F., Stadler, B., Rest, A., Smith, K. W., Smartt, S. J., and Weiland, H. (2018). A first catalog of variable stars measured by the asteroid terrestrial-impact last alert system (ATLAS). *arXiv:1804.0213v1*.

Hoard, D. W., Ladjal, D., Stencel, R. E., and Howell, S. B. (2012). The invisible monster has two faces: Observations of epsilon aurigae with the herschel space observatory. *Astrophysical Journal Letters*, 748(2).

Hoffman, D. I., Harrison, T. E., Coughlin, J. L., McNamara, B. J., Holtzman, J. A., Taylor, G. E., and Vestrand, W. T. (2008). New β lyrae and algol candidates from the northern sky variability survey. *The Astronomical Journal*, 136(3):1067–1078.

Holland, J. H. (1975). *Adaption in Natural and Artificial Systems*. The University of Michigan Press.

Horne, J. H. and Baliunas, S. L. (1986). A prescription for the analysis of unevenly sampled time series. *The Astrophysical Journal*, 302:757–763.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441 & 498–520.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28:321–377.

Huijse, P., Estevez, P. A., Protopapas, P., Principe, J. C., and Zegers, P. (2014). Computational intelligence challenges and applications on large-scale astronomical time series databases. *IEEE Computational Intelligence Magazine*, 9(3):27–39.

Huijse, P., Estevez, P. A., Protopapas, P., Zegers, P., and Principe, J. C. (2012). An information theoretic algorithm for finding periodicities in stellar light curves. *IEEE Transactions on Signal Processing*, 60(10):5135–5145.

Ivanova, N., Justham, S., Chen, X., and et al. (2013). Common envelope evolution: where we stand and how we can move forward. *The Astronomy and Astrophysics Review*, 21:59.

Ivezić, Z., Tyson, J. A., Abel, B., and et al. (2014). LSST from science drivers to reference design and anticipated data products. *arXiv:0805.2366v4*.

Jain, L. C. and Medsker, L. R. (1999). *Recurrent Neural Networks: Design and Applications*. CRC Press.

Jeffery, C. S., Starling, R. L. C., Hill, P. H., and D., P. (2001). Cyclic and secular variation in the temperatures and radii of extreme helium stars. *Monthly Notices of the Royal Astronomical Society*, 321(1):111–130.

Jenkins, G. M. and Watts, D. G. (1968). *Spectral analysis and its applications.* Holden-day.

Johnston, K. B. and Peter, A. M. (2017). Variable star signature classification using slotted symbolic markov modeling. *New Astronomy*, 50:1–11.

Kaiser, N. (2002). Pan-STARRS: A large synoptic survey telescope array. *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, 4836:154–164.

Kilkenny, D., O'Donoghue, D., Koen, C., Lynas-Gray, A. E., and van Wyk, F. (1998). The ec 14026 stars - VIII. PG 1336-018: a pulsating sdB star in an HWVir-type eclipsing binary. *Monthly Notices of the Royal Astronomical Society*, 296(2):329–338.

Kim, A. G. and Miquel, R. (2007). Measuring type Ia supernova distances and redshifts from their multi-band light curves. *Astroparticle Physics*, 28.

Kim, D.-W. and Bailer-Jones, C. A. L. (2016). A package for the automated classification of periodic variable stars. *Astronomy & Astrophysics*, 587.

Kim, D.-W., Protopapas, P., Byun, Y.-I., Alcock, C., Khardon, R., and Trichas, M. (2011). Quasi-stellar object selection algorithm using time variability and machine learning: Selection of 1620 quasi-stellar object candidates from macho large magellanic cloud database. *The Astrophysical Journal*, 735(2):68.

Kochoska, A., Mowlavi, N., Prsa, A., Lecoeur-Tabi, I., Holl, B., Rimoldini, L., Sveges, M., and Eyer, L. (2017). Gaia eclipsing binary and multiple systems. a study of detectability and classification of eclipsing binaries with gaia. *Astronomy & Astrophysics*, 602.

Koester, D. and Chanmugam, G. (1990). Physics of white dwarf stars. *Reports on Progress in Physics*, 53(7).

Kolenberg, K., Bryson, S., Szabo, R., Kurtz, D. W., Smolec, R., Nemec, J. M., Guggenberger, E., Moskalik, P., Benko, J. M., and Chadid, M. (2010). Kepler photometry of the prototypical blazhko star RR Lyr: an old friend seen in a new light. *Monthly Notices of the Royal Astronomical Society*, 411(2):878–890.

Korhonen, H., Berdyugina, S. V., and Tuominen, I. (2005). Surface differential rotation on fk com. *Proceedings of the* 13*th* *Cambridge Workshop on Cool Stars, Stellar Systems and the Sun*, page 719.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, 31(3):249–268.

Kovacs, G., Bakos, G., and Noyes, R. W. (2005). A trend filtering algorithm for wide field variability surveys. *Monthly Notices of the Royal Astronomical Society*, 356:557–567.

Krisciunas, K. (1993). A new class of pulsating stars. *American Astronomical Society, 183rd AAS Meeting*, 25:1422.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *in Advances in Neural Information Processing Systems 25.*

Kugler, S. D., Gianniotis, N., and Polsterer, K. L. (2016). An explorative approach for inspecting Kepler data. *Monthly Notices of the Royal Astronomical Society*, 455(4):4399–4405.

Kurtz, D. W. (1982). Rapidly oscillating ap stars. *Monthly Notices of the Royal Astronomical Society*, 200(3):807–859.

LaCourse, D. M., Jek, K. J., Jacobs, T. L., Winarski, T., Boyajian, T. S., Rappaport, R., Sanchis-Ojeda, R., Conroy, K. E., Nelson, L., Barclay, T., Fischer, D. A., Schmitt, J. R., Wang, J., Stassun, K. G., Pepper, J., Coughlin, J. L., Shporer, A., and A., P. (2015). Kepler eclipsing binary stars - vi. identification of eclipsing binaries in the k2 campaign o data set. *Monthly Notices of the Royal Astronomical Society*, 452(4):3561–3592.

Lafler, J. and Kinman, T. D. (1965). An RR Lyrae star survey with ihe lick 20-inch astrograph ii. the calculation of RR Lyrae periods by electronic computer. *The Astrophysical Journal Supplement Series*, 11:216.

Lang, D., Hogg, D. W., Mierle, K., Blanton, M., and Roweis, S. (2010). Astrometry.net: Blind astrometric calibration of arbitrary astronomical images. *The Astronomical Journal*, 139(5):1782–1800.

Larson, S. (2003). The CSS and SSS NEO surveys. *AAS/Division for Planetary Sciences Meeting Abstracts*, 35:982.

Larsson, S. (1996). Parameter estimation in epoch folding analysis. *Astronomy and Astrophysics Supplement Series*, 117:197–201.

Layden, A., Anderson, T., and Husband, P. (2013). Colors of c-type rr lyrae stars and interstellar reddening. *in 40 Years of Variable Stars: A Celebration of Contributions by Horace A. Smith.*

Leroy, B. (2012). Fast calculation of the lomb-scargle periodogram using nonequispaced fast fourier transforms. *Astronomy & Astrophysics*, 545.

Liu, W., Pokharel, P. P., and Principe, J. C. (2006). Correntropy: A localized similarity measure. *The 2006 IEEE International Joint Conference on Neural Network Proceedings.*

Lomb, N. R. (1976). Least-squares frequency analysis of unequally spaced data. *Astrophysics and Space Science*, 39(2):447–462.

Long, J. P., Chi, E. C., and Baraniuk, R. G. (2016). Estimating a common period for a set of irregularly sampled functions with applications to periodic variable star data. *The Annals of Applied Statistics*, 10(1):165–197.

Long, J. P., Karoui, N. E., Rice, J. A., Richards, J. W., and Bloom, J. S. (2012). Optimizing automated classification of variable stars in new synoptic surveys. *Publications of the Astronomical Society of the Pacific*, 124(913):280–295.

LSST Science Collaborations, Abell, P. A., Allison, J., and et al. (2009). LSST science book, version 2.0. *arXiv:0912.0201v1.*

Lyne, A. G. and Graham-Smith, F. (2006). *Pulsar Astronomy.* Cambridge University Press.

Mackenzie, C., Pichara, K., and Protopapas, P. (2016). Clustering-based feature learning on variable stars. *The Astrophysical Journal*, 820(2):138.

Madore, B. F., Rigby, J., Freedman, W. L., Persson, S. E., Sturch, L., and Mager, V. (2009). The cepheid period-luminosity relation (the leavitt law) at mid-infrared wavelengths. III. cepheids in ngc 6822. *The Astrophysical Journal*, 693(1).

Mahabal, A., Sheth, K., Gieseke, F., Pai, A., Djorgovski, S. G., Drake, A., and Graham, M. (2017). Deep-learnt classification of light curves. *2017 IEEE Symposium Series on Computational Intelligence (SSCI).*

Malkov, O. Y., Oblak, E., Avvakumova, E. A., and Torra, J. (2007). A procedure for the classification of eclipsing binaries. *Astronomy and Astrophysics*, 465(2):549–556.

Matijevic, G. (2012). Kepler eclipsing binary stars. III. classification of Kepler eclipsing binary light curves with locally linear embedding. *The Astronomical Journal*, 143:123–128.

Mawson, N. R., Steele, I. A., and Smith, R. J. (2013). STILT: System design and performance. *Astronomische Nachrichten*, 334(7):729–737.

McWhirter, P. R., Steele, I. A., Al-Jumeily, D., Hussain, A., and Vellasco, M. M. B. R. (2017). The classification of periodic light curves from non-survey optimized observational data through automated extraction of phase-based visual features. *2017 International Joint Conference on Neural Networks (IJCNN).*

McWhirter, P. R., Wright, S., Steele, I. A., Al-Jumeily, D., Hussain, A., and Fergus, P. (2016). A dynamic, modular intelligent-agent framework for astronomical light curve analysis and classification. *Intelligent Computing Theories and Application Lecture Notes in Computer Science*, pages 820–831.

Miglio, A., Montalban, J., and Dupret, M.-A. (2007). Revised instability domains of spb and beta cephei stars. *Communications in Astroseismology*, 151.

Mitchell, T. (1998). *Machine Learning.* McGraw Hill.

Morris, S. L. (1985). The ellipsoidal variable stars. *The Astrophysical Journal*, 295:143–152.

Mortier, A. and Cameron, A. C. (2017). Stacked bayesian general lomb-scargle periodogram: Identifying stellar activity signals. *Astronomy & Astrophysics*, 601.

Mortier, A., Faria, J. P., Correia, C. M., Santerne, A., and Santos, N. C. (2015). BGLS: A bayesian formalism for the generalised lomb-scargle periodogram. *Astronomy & Astrophysics*, 573.

Naul, B., Bloom, J. S., Pérez, F., and van der Valt, S. (2018). A recurrent neural network for classification of unevenly sampled variable stars. *Nature Astronomy*, 2:151–155.

Neff, J. E., Wells, M. A., Geltz, S. N., and A., B. (2014). Automated variability classification and constant stars in the Kepler database. *18th Cambridge Workshop on Cool Stars, Stellar Systems, and the Sun.*

Newell, E. B. (1983). Crowded-field stellar photometry. *Proceedings of the Workshop on Astronomical Measuring Machines.*

Nun, I., Pichara, K., Protopapas, P., and Kim, D.-W. (2014). Supervised detection of anomalous light-curves in massive astronomical catalogs. *The Astrophysical Journal*, 793(1).

Nun, I., Protopapas, P., and et al. (2015). FATS: Feature analysis for time series. *arXiv:1506.00010v2.*

Olivier, E. A. and Wood, P. R. (2003). On the origin of long secondary periods in semiregular variables. *The Astrophysical Journal*, 584:1035–1041.

Paegert, M., Stassun, K. G., and Burger, D. M. (2014). The eb factory project. i. a fast, neural-net-based, general purpose light curve classifier optimized for eclipsing binaries. *The Astronomical Journal*, 148(2):31.

Parvizi, M., Paegert, M., and Stassun, K. G. (2014). The eb factory project. II. validation with the Kepler field in preparation for K2 and TESS. *The Astronomical Journal*, 148(6):125.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572.

Pearson, K. A., Palafox, L., and Griffith, C. A. (2018). Searching for exoplanets using artificial intelligence. *Monthly Notices of the Royal Astronomical Society*, 474(1):478–491.

Percy, J. R. (2008). *Variable stars*. Cambridge University Press.

Pichara, K., Protopapas, P., and Len, D. (2016). Meta-classification for variable stars. *The Astrophysical Journal*, 819(1):18.

Pietrukowicz, P. (2018). On the properties of blue large-amplitude pulsators. no blaps in the magellanic clouds. *The RR Lyrae 2017 Conference. Revival of the Classical Pulsators: from Galactic Structure to Stellar Interior Diagnostics.*

Pietrukowicz, P., Dziembowski, W. A., Latour, M., Angeloni, R., Poleski, R., di Mille, F., Soszynski, I., Udalski, A., Szymanski, M. K., Wyrzykowski, L., Kozlowski, S., Skowron, J., Skowron, D., Mroz, P., Pawlak, M., and Ulaczuk, K. (2017). Blue large-amplitude pulsators as a new class of variable stars. *Nature Astronomy*, 1(0166).

Pigulski, A., Kolaczkowski, Z., Ramza, T., and Narwid, A. (2005). High-amplitude delta scuti stars in the galactic bulge from the OGLE-II and MACHO data. *Proceedings of the "Stellar Pulsation and Evolution" Conference.*

Plavchan, P., Jura, M., Kirkpatrick, J. D., Cutri, R. M., and Gallagher, S. C. (2008). Near-infrared variability in the 2MASS calibration fields: A search for planetary transit candidates. *The Astrophysical Journal Supplement Series*, 175(1):191–228.

Poleski, R., Soszyński, I., Udalski, A., Szymański, M. K., Kubiak, M., Pietrzyński, G., Wyrzykowski, L., Szewczyk, O., and Ulaczyk, K. (2010). The optical gravitational lensing experiment. the OGLE-III catalog of variable stars. VI. delta scuti stars in the large magellanic cloud. *Acta Astronomica*, 60(1):1–16.

Pospíchal, P., Jaroz, J., and Schwarz, J. (2010). Parallel genetic algorithm on the cuda architecture. *Applications of Evolutionary Computation, Lecture Notes in Computer Science*, 6024:442–451.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1994). *Numerical recipes in C: the art of scientific computing, 2nd edn.* Cambridge University Press.

Protopapas, P., Giammarco, J. M., Faccioli, L., Struble, M. F., Dave, R., and Alcock, C. (2006). Finding outlier light curves in catalogues of periodic variable stars. *Monthly Notices of the Royal Astronomical Society*, 369(2):677–696.

Protopapas, P., Huijse, P., Estvez, P. A., Zegers, P., Prncipe, J. C., and Marquette, J.-B. (2015). A novel, fully automated pipeline for period estimation in the eros 2 data set. *The Astrophysical Journal Supplement Series*, 216(2):25.

Prsa, A., Guinan, E. F., Devinney, E. J., Degeorge, M., Bradstreet, D. H., Giammarco, J. M., Alcock, C. R., and Engle, S. G. (2008). Artificial intelligence approach to the determination of physical properties of eclipsing binaries. i. the ebai project. *The Astrophysical Journal*, 687(1):542–565.

Prsa, A., Pepper, J., and Stassun, K. G. (2011). Expected large synoptic survey telescope (LSST) yield of eclipsing binary stars. *The Astronomical Journal*, 142(2):52.

Rahal, Y. R., Afonso, C., Albert, J.-N., Andersen, J., Ansari, R., Aubourg, ., Bareyre, P., Beaulieu, J.-P., Charlot, X., and Couchot, F. (2009). The EROS2 search for microlensing events towards the spiral arms:the complete seven season results. *Astronomy & Astrophysics*, 500(3):1027–1044.

Rajpaul, V. (2012). Genetic algorithms in astronomy and astrophysics. *Proceedings of SAIP2011, the 56th Annual Conference of the South African Institute of Physics*, pages 519–524.

Razali, N. and Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics*, 2(1):21–33.

Rebbapragada, U., Protopapas, P., Brodley, C. E., and Alcock, C. (2009). Anomaly detection in catalogs of periodic variable stars. *Astronomical Data Analysis Software and Systems XVIII ASP Conference Series*, 411:264.

Reed, M. D., Green, E. M., Callerame, K., Seitenzahl, I. R., White, B. A., Hyde, E. A., Giovanni, M. K., Østensen, R., Bronowska, A., Jeffery, E. J., Cordes, O., Falter, S., Edelmann, H., Dreizler, S., and Schuh, S. L. (2004). Discovery of gravity-mode pulsators among Subdwarf B stars: PG 1716+426, the class prototype. *The Astrophysical Journal*, 607:445–450.

Reipurth, B. (1990). FU Orionis eruptions and early stellar evolution. *in: Flare stars in star clusters, associations and the solar vicinity*, pages 229–251.

Richards, J. W., Starr, D. L., Brink, H., Miller, A. A., Bloom, J. S., Butler, N. R., James, J. B., Long, J. P., and Rice, J. (2011a). Active learning to overcome sample selection bias: Application to photometric variable star classification. *The Astrophysical Journal*, 744(2):192.

Richards, J. W., Starr, D. L., Butler, N. R., Bloom, J. S., Brewer, J. M., Crellin-Quick, A., Higgins, J., Kennedy, R., and Rischard, M. (2011b). On machine-learned classification of variable stars with sparse and noisy time-series data. *The Astrophysical Journal*, 733(1):10.

Richards, J. W., Starr, D. L., Miller, A. A., Bloom, J. S., Butler, N. R., Brink, H., and Crellin-Quick, A. (2012). Construction of a calibrated probabilistic classification catalog: Application to 50k variable sources in the all-sky automated survey. *The Astrophysical Journal Supplement Series*, 203(2):32.

Romero, A. D., Córsico, A. H., Althaus, L. G., Pelisoli, I., and Kepler, S. O. (2018). On the evolutionary status and pulsations of the recently discovered blue large-amplitude pulsators (blaps). *Monthly Notices of the Royal Astronomical Society: Letters*, 477(1):L30–L34.

Rosenblatt, F. (1958). The perceptron: A probalistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408.

Rosino, L. (1951). The spectra of variables of the RV Tauri and yellow semiregular types. *The Astrophysical Journal*, 113:60.

Ruf, T. (1999). The lomb-scargle periodogram in biological rhythm research: Analysis of incomplete and unequally spaced time-series. *Biological Rhythm Research*, 30(2):178–201.

Russell, S. and Norvig, P. (2009). *Artificial Intelligence - A Modern Approach 3rd ed.* Prentice Hall.

Saha, A. and Vivas, A. K. (2017). A hybrid algorithm for period analysis from multiband data with sparse and irregular sampling for arbitrary light-curve shapes. *The Astronomical Journal*, 154(6).

Saio, H. (1993). An overview of stellar pulsation theory. *Astrophysics and Space Science*, 210(1–2):61–72.

Samuel, A. L. (1988). Some studies in machine learning using the game of checkers. I. *Computer Games I*, pages 335–365.

Scargle, J. D. (1982). Studies in astronomical time series analysis. ii - statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal*, 263:835.

Scargle, J. D., Norris, J. P., Jackson, B., and Chiang, J. (2013). Studies in astronomical time series analysis. vi. bayesian block representations. *The Astrophysical Journal*, 764(2):167.

Schaefer, B. E., King, J. R., and Deliyannis, C. P. (2000). Superflares on ordinary solar-type stars. *The Astrophysical Journal*, 529(2).

Schölkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J., and Platt, J. (2000). Support vector method for novelty detection. *Proceedings of the 12th International Conference on Neural Information Processing Systems, NIPS'99*, pages 582–588.

Schuster, F. A. F. (1898). On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena. *Terrestrial Magnetism*, 3(1):13–41.

Schwarzenberg-Czerny, A. (1989). On the advantage of using analysis of variance for period search. *Monthly Notices of the Royal Astronomical Society*, 241(2):153–165.

Schwarzenberg-Czerny, A. (1996). Fast and statistically optimal period search in uneven sampled observations. *The Astrophysical Journal*, 460(2).

Shallue, C. J. and Vanderburg, A. (2018). Identifying exoplanets with deep learning: A five-planet resonant chain around kepler-80 and an eighth planet around kepler-90. *The Astronomical Journal*, 155(2):94.

Shannon, C. E. (1949). Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis.* Cambridge University Press.

Shin, M.-S., Sekora, M., and Byun, Y.-I. (2009). Detecting variability in massive astronomical time series data I. application of an infinite gaussian mixture model. *Monthly Notices of the Royal Astronomical Society*, 400(4):1897–1910.

Shin, M.-S., Yi, H., Kim, D.-W., Chang, S.-W., and Byun, Y.-I. (2012). Detecting variability in massive astronomical time-series data. II. variable candidates in the northern sky variability survey. *The Astronomical Journal*, 143(3).

Slettebak, A. (1982). Spectral types and rotational velocities of the brighter Be stars and A-F type shell stars. *Astrophysical Journal Supplement Series*, 50:55–83.

Smith, H. A. (2004). *RR Lyrae Stars.* Cambridge University Press.

Sokolova, M., Japkowicz, N., and Szpakowicz, S. (2006). Beyond accuracy, f-score and roc: A family of discriminant measures for performance evaluation. *Advances in Artificial Intelligence (Lecture Notes in Computer Science)*, 4304.

Solarz, A., Bilicki, M., Gromadzki, M., Pollo, A., Durkalec, A., and Wypych, M. (2017). Automated novelty detection in the wise survey with one-class support vector machines. *Astronomy and Astrophysics*, 606:A39.

Soszynski, I., Udalski, A., Szymanski, M. K., Kubiak, M., Pietrzynski, G., Wyrzykowski, L., Szewczyk, O., Ulaczyk, K., and Poleski, R. (2008). The optical gravitational lensing experiment. the OGLE-III catalog of variable stars. II. type II cepheids and anomalous cepheids in the large magellanic cloud. *Acta Astronomica*, 58:293.

Soszynski, I., Udalski, A., Szymanski, M. K., Kubiak, M., Pietrzynski, G., Wyrzykowski, L., Szewczyk, O., Ulaczyk, K., and Poleski, R. (2009). The optical gravitational lensing experiment. the OGLE-III catalog of variable stars. IV. long-period variables in the large magellanic cloud. *Acta Astronomica*, 59:239.

Sowell, J. R., Hall, D. S., Henry, G. W., Burke Jr., E. W., and Milone, E. F. (1983). MM Herculis: An eclipsing binary of the RS CVn type. *Astrophysics and Space Science*, 90(2):421–435.

Steele, I. A., Smith, R. J., Rees, P. C., Baker, I. P., Bates, S. D., Bode, M. F., Bowman, M. K., Carter, D., Etherton, J., and Ford, M. J. (2004). The liverpool telescope: performance and first results. *Ground-based Telescopes.*

Stekhoven, D. J. and Bühlmann, P. (2012). Missforest–non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

Stellingwerf, R. F. (1978). Period determination using phase dispersion minimization. *The Astrophysical Journal*, 224:953–960.

Stetson, P. B. (1996). On the automatic determination of light-curve parameters for cepheid variables. *Publications of the Astronomical Society of the Pacific*, 108:851.

Tagliaferri, R., Ciaramella, A., Milano, L., Barone, F., and Longo, G. (1999). Spectral analysis of stellar light curves by means of neural networks. *Astronomy and Astrophysics Supplement Series*, 137(2):391–405.

Tanvir, N. R., Hendry, M. A., Watkins, A., Kanbur, S. M., Berdnikov, L. N., and Ngeow, C. C. (2005). Determination of cepheid parameters by light-curve template fitting. *Monthly Notices of the Royal Astronomical Society*, 363(3):749–762.

Thackeray, A. D. (1974). Variations of s dor and hde 269006. *Monthly Notices of the Royal Astronomical Society*, 168(1):221–233.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–288.

Townsend, R. H. D. (2010). Fast calculation of the lomb-scargle periodogram using graphics processing units. *The Astrophysical Journal Supplement Series*, 191(2):247–253.

Udalski, A., Kubiak, M., and Szymanski, M. (1997). Optical gravitational lensing experiment. OGLE-2 – the second phase of the OGLE project. *Acta Astronomica*, 47:319–344.

Ukwatta, T. N. and Wozniak, P. R. (2015). Integrating temporal and spectral features of astronomical data using wavelet analysis for source classification. *2015 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*.

Unal, I. (2017). Defining an optimal cut-point value in roc analysis: An alternative approach. *Computational and Mathematical Methods in Medicine*, 2017.

Urry, C. M. and Padovani, P. (1995). Unified schemes for radio-loud active galactic nuclei. *Publications of the Astronomical Society of the Pacific*, 107(715).

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2608.

Van Hamme, W., Samec, R. G., Gothard, N. W., Wilson, R. E., Faulkner, D. R., and Branly, R. M. (2001). Cn andromedae: A broken-contact binary? *The Astronomical Journal*, 122(6):3436–3446.

VanderPlas, J. T. (2017). Understanding the lomb-scargle periodogram. *arXiv:1703.09824v1*.

Vanderplas, J. T. and Ivezić, Z. (2015). Periodograms for multiband astronomical time series. *The Astrophysical Journal*, 812(1):18.

Vaughan, S. (2011). Random time series in astronomy. *Philosophical Transactions of the Royal Society*, 371.

Vora, K. and Yagnik, S. (2014). A survey on backpropagation algorithms for feedforward neural networks. *International Journal of Engineering Development and Research*, 1(3):193–197.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2):307–333.

Waelkens, C. and Rufener, F. (1985). Photometric variability of mid-b stars. *Astronomy and Astrophysics*, 152(1):6–14.

Wallerstein, G. (2002). The cepheids of population II and related stars. *Publications of the Astronomical Society of the Pacific*, 114(797).

Wells, M., Prsa, A., Jones, L., and Yoachim, P. (2017). Initial estimates on the performance of the lsst on the detection of eclipsing binaries. *Publications of the Astronomical Society of the Pacific*, 129(976):065003.

Wilks, D. S. (2010). Sampling distributions of the brier score and brier skill score under serial dependence. *Quarterly Journal of the Royal Meteorological Society*, 136(1):2109–2118.

Winget, D. E., van Horn, H. M., Tassoul, M., Fontaine, G., Hansen, C. J., and Carroll, B. W. (1982). Hydrogen-driving and the blue edge of compositionally stratified zz ceti star models. *Astrophysical Journal, Part 2 - Letters to the Editor*, 252:L65–L68.

Wood, P. R. and Sebo, K. M. (1996). On the pulsation mode of mira variables: evidence from the large magellanic cloud. *Monthly Notices of the Royal Astronomical Society*, 282(3):958–964.

Woosley, S. E. and Eastman, R. G. (1997). Type Ib and Ic supernovae: Models and spectra. *Thermonuclear Supernovae.*

Wozniak, P. R. (2000). Difference image analysis of the OGLE-II bulge data. i. the method. *Acta Astronomica*, 50:421–450.

Wray, J. J., Eyer, L., and Paczyński, B. (2004). OGLE small amplitude red giant variables in the galactic bar. *Monthly Notices of the Royal Astronomical Society*, 349(3):1059–1068.

Xiong, D. R. and Deng, L. (2006). Small amplitude variable red giants. *Proceedings IAU Symposium No. 239.*

Yoachim, P., Mccommas, L. P., Dalcanton, J. J., and Williams, B. F. (2009). A panoply of cepheid light curve templates. *The Astronomical Journal*, 137(6):4697–4706.

Yoon, S.-C. and Langer, N. (2004). Presupernova evolution of accreting white dwarfs with rotation. *Astronomy and Astrophysics*, 419(2):623–644.

York, D. G., Adelman, J., and Anderson, J. E. (2000). The sloan digital sky survey: Technical summary. *The Astronomical Journal*, 120(3):1579–2000.

Zechmeister, M. and Kürster, M. (2009). The generalised lomb-scargle periodogram. *Astronomy & Astrophysics*, 496(2):577–584.

Zinn, J. C., Kochanek, C. S., Kozlowski, S., Udalski, A., Szymanski, M. K., Soszynski, I., Wyrzykowski, L., Ulaczyk, K., Poleski, R., and Pietrukowicz, P. (2017). Variable classification in the lsst era: exploring a model for quasi-periodic light curves. *Monthly Notices of the Royal Astronomical Society*, 468(2):2189–2205.

Zucker, S. (2016). Detection of periodicity based on independence tests – II. improved serial independence measure. *Monthly Notices of the Royal Astronomical Society: Letters*, 457(1).

Zucker, S. (2018). Detection of periodicity based on independence tests – III. phase distance correlation periodogram. *Monthly Notices of the Royal Astronomical Society: Letters*, 474(1):L86–L90.

Zucker, S. and Giryes, R. (2018). Shallow transits – deep learning. i. feasibility study of deep learning to detect periodic transits of exoplanets. *The Astronomical Journal*, 155(4):9.